Chapter 12

Linear Regression and Correlation

Chapter Objectives

- Compute and interpret a linear (Pearson) correlation.
- Carry out a t test for a linear correlation.
- Compute and interpret a confidence interval for a population correlation.
- Compute and interpret a monotone nonlinear (Spearman rank) correlation.
- Carry out a t test for a monotone nonlinear correlation.
- State and interpret the simple linear regression model.
- Obtain and interpret estimates of model coefficients.
- Obtain and interpret fitted values and residuals associated with a fitted regression model.
- Interpret sums of squares, degrees of freedom, and mean squares.
- Interpret the R^2 associated with a fitted regression model.
- Carry out a t test for the slope in a regression model.
- Obtain a t confidence interval for the slope in a regression model.
- Carry out a regression model F test.
- Decide whether the t test (and F test) associated with a linear regression analysis are appropriate for a given set of data.

Key Takeaways

- The (Pearson) correlation measures the strength of a linear relationship between two variables. The Spearman rank correlation measure the strength of a monotone nonlinear relationship.
- A linear regression analysis is used to estimate the equation of a linear relationship between a response variable and a numerical explanatory variable. Non-linear patterns in data can be transformed to linear ones prior to conducting the analysis. Both the t test for the slope and the model F test are tests for whether there's a linear relationship. They require either that the response variable is normally distributed or the sample size is large. A log transformation can make a right-skewed response variable more normal prior to conducting a t or F test.
- A linear regression model describes variation in a response variable in terms of a numerical explanatory variable. It contains two parts: one representing non-random variation due to the linear relationship to the explanatory variable and another representing random variation (random error).
- Sums of squares in linear regression are statistics that measure variation in the observed values of a response variable due to the linear relationship to the explanatory variable and due to random error.
- Mean squares are another way to measure variation. They're obtained by dividing sums of squares by their degrees of freedom. The degrees of freedom associated with a sum of squares is determined by how many of its squared deviations are "free to vary." The values of two mean squares are directly comparable, but the values of two sums of squares aren't necessarily comparable.

- The R^2 value is a statistic that measures how well a fitted linear regression model fits the data. Expressed as a percent, it's interpreted as the percent of variation in the response variable that's explained by variation in the explanatory variable.
- The regression model F test statistic is a ratio of two mean squares. Its numerator measures variation due to the explanatory variable and its denominator variation that's due to random error.
- The t test statistic for the slope indicates how many standard errors the estimated slope is away from zero.

12.1 Introduction

Environmental studies are often designed to investigate the relationship between two variables, an ***explanatory variable*** and a ***response variable***. In this chapter, we look at methods of analyzing such data when the explanatory variable is *numerical* as opposed to categorical (the latter situation having been the subject of Chapters 10 and 11). Numerical explanatory variables are sometimes called ***predictors*** because, as we'll see, they can be used to predict the response.

Example 12.1: Bivariate Data

Overstory trees are ones whose heights extend well above the canopy (dense ceiling of tightly packed trees and branches). In a study of the recent decline in the number of overstory aspen trees in Yellowstone National Park, Wyoming, data on the ages (years) and diameters (cm) at breast height (1.4 m) of $n = 49$ aspen trees were collected and are shown below [15].

Ages and Diameters of Trees

Tree Number	Age	Diameter
1	24	5.0
2	17	6.9
3	30	8.0
4	10	10.0
5	14	10.0
6	12	10.5
7	22	11.0
8	30	10.4
9	16	14.0
10	20	13.4
11	36	12.5
12	39	13.0
13	26	16.4
14	35	16.0
15	36	15.5
16	38	14.9
17	40	20.0
18	27	20.5
19	39	20.5
20	42	15.0
21	50	13.5
22	42	18.0
23	72	25.5
24	79	21.0
25	50	30.5
26	78	31.0
27	76	32.5
28	72	39.0
29	90	28.4
30	108	28.9
31	83	38.0
32	86	35.0
33	92	31.0
34	108	31.9
35	116	37.4
36	117	38.0
37	109	48.0
38	114	46.0
39	126	27.6
40	130	29.0
41	124	31.0
42	122	31.5
43	121	39.0
44	159	35.0
45	126	36.0
46	128	37.0
47	129	38.0
48	124	42.0
49	123	46.0

Because the diameter of a tree grows larger as the tree ages, we'll consider diameter as the *response* and age as the *explanatory variable*, or *predictor*.

When two variables are measured on each of $\boldsymbol{n}$ individuals, as in the last example, the data are said to be ***bivariate***. In bivariate data, we'll denote the explanatory and response variables by X and Y, respectively, and store them (for use with statistical software) in columns as below.

Observation	X variable	Y variable
1	X_1	Y_1
2	X_2	Y_2
3	X_3	Y_3
$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n

We use the notation

$\boldsymbol{n}$ = The number of individuals upon which X and Y are measured, or sample size.
$\boldsymbol{X_i}$ = The value of the explanatory (predictor) variable for the ith individual.
$\boldsymbol{Y_i}$ = The value of the response variable for the ith individual.

Each X_i, Y_i pair is called a ***bivariate observation***. The sample size n refers to the number of bivariate observations (rows in the data file).

In this chapter we look at ways of graphing and summarizing bivariate data, testing for relationships between the variables, and fitting linear models and assessing how well they fit the data.

12.2 Graphing Bivariate Data

Graphical displays are used to *explore* patterns of variation and relationships between variables in bivariate data and to *communicate* these aspects of the data to others.

12.2.1 Scatterplots

Creating Scatterplots

The most useful way of displaying bivariate numerical data is with a scatterplot. To construct a ***scatterplot***,

1. Designate one variable as the explanatory variable and the other as the response.
2. Plot the observation pairs as points in an x, y coordinate system, with the explanatory variable on the x-axis and the response on the y-axis.
3. Label the axes and add a title.

Example 12.2: Scatterplots

A scatterplot of the data on ages and diameters of aspen trees from Example 12.1 is shown below, with age on the x axis, and diameter on the y axis.

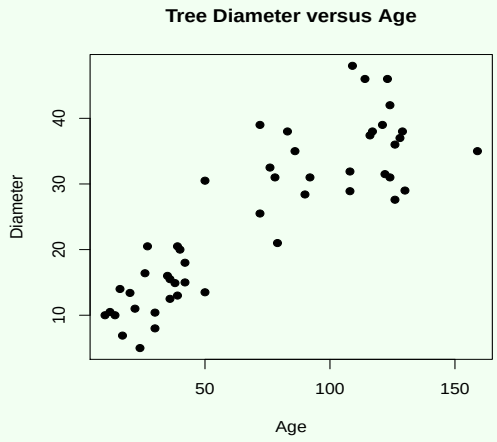


Figure 12.1: Scatterplot of the diameters versus ages of aspen trees in Yellowstone National Park.

Examining scatterplots

Scatterplots can reveal various features of the relationship between X and Y, among them:

- The "overall pattern" or "form" of the relationship, for example whether it's ***linear*** (following a straight line pattern) or curved.
- The direction of the relationship:
  - ***Positive relationship***: When X is large, Y tends to be large, and when X is small, Y tends to be small (the points in the scatterplot slope upward from left to right).
  - ***Negative relationship***: When X is large, Y tends to be small, and when X is small, Y tends to be large (the points slope downward from left to right).
- The ***strength*** of the relationship, in other words, how clear the pattern is or how closely the points in the scatterplot conform to straight line or smooth curve.
- Other interesting features, for example outliers, separate clumps of points, or unusual patterns.

Figure 12.2 below illustrates some of these features.

Example 12.3: Scatterplots

We'd describe the relationship between the ages and diameters of the aspen trees in Fig. 12.1 to be a moderately strong, positive, approximately linear relationship.

12.2.2 Time Series Plots

A ***time series*** is a set of data collected at regular intervals of time (for example hourly, daily, monthly, or yearly). They're bivariate data in which the explanatory variable is time, and are primarily used to identify trends. When graphing the data, we usually connect the points in the scatterplot by lines to highlight their sequential nature. This type of plot is called a ***time series plot***.

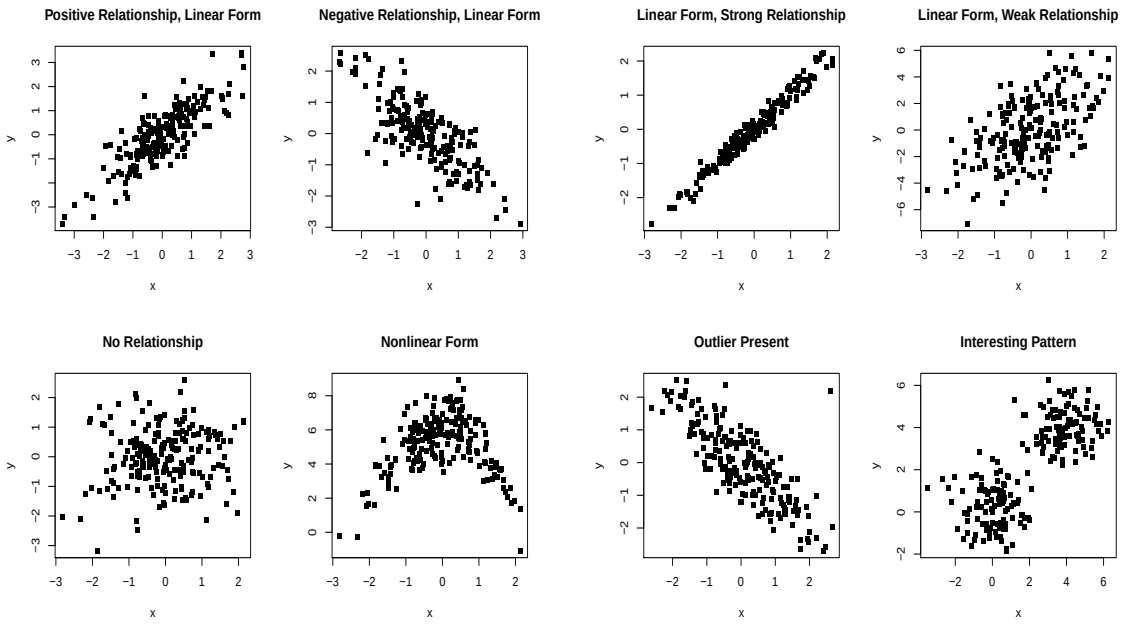


Figure 12.2: Scatterplots showing various patterns and features. First row, left to right: moderately strong positive and negative linear relationships, very strong and weak positive linear relationships. Second row, left to right: no relationship, nonlinear curved relationship, negative relationship with outlier, two separate clumps.

Example 12.4: Time Series Plots

The Alaska Climate Research Center at the University of Alaska, Fairbanks, reported yearly average temperatures (°F) for the years 1930 - 2008 in Fairbanks, Alaska. A scatterplot and a time series plot of the data are below.

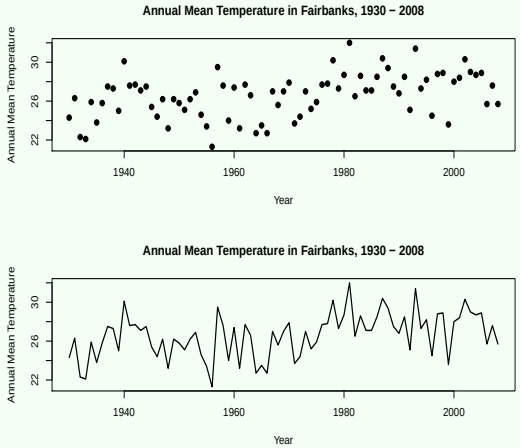


Figure 12.3: Scatterplot (top) and time series plot (bottom) of annual mean temperatures (°F) in Fairbanks, Alaska for the years 1930 to 2008.

Notice that it's visually much easier to follow the year-to-year changes in temperature in the time series plot than in the scatterplot.

12.3 Summarizing the Strength of a Linear Relationship

It's often desirable to report a statistic that summarizes the strength and direction of the relationship between two variables. We'll look at two statistics that accomplish this:

1. The Pearson correlation
2. The Spearman rank correlation

The first of these is appropriate when the relationship between the two variables is (at least approximately) *linear*. The second is more general, and is appropriate even when the relationship is *nonlinear*.

12.3.1 The Pearson Correlation

Computing *r*

When two variables in a bivariate data set exhibit, at least approximately, a linear relationship, we summarize that relationship by the ***sample correlation*** (also called the ***Pearson correlation***), denoted r, and defined as follows.

Correlation: The Pearson correlation between two variables $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S_x}\right)\left(\frac{Y_i - \bar{Y}}{S_y}\right) \tag{12.1}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of the X_i's and Y_i's, respectively, and S_x and S_y are their sample standard deviations.

Thus r is computed by *standardizing* each X_i and each Y_i, taking the products of these, and "averaging" those products (by dividing their sum by $n-1$).

Properties and Interpretation of *r*

The following properties of the correlation r provide insight into its interpretation.

1. The value of the correlation will always be between -1.0 and 1.0.
2. The correlation tells us the *direction* of the relationship between X and Y:
   - Positive correlation values indicate a positive relationship.
   - Negative correlation values indicate a negative relationship.
3. The correlation also tells us how *strong* the relationship between X and Y is:
   - Correlation values near zero imply a very weak relationship or none at all.
   - Correlation values close to -1.0 or 1.0 imply a very strong linear relationship.
   - The extreme values $r = -1.0$ and $r = 1.0$ occur only when there's a *perfect linear* relationship, that is, when the points in the scatterplot lie *exactly* along a straight line.
4. The correlation doesn't depend on which variable is labeled X and which is labeled Y. It's a measure of association *between* the two variables.
5. The correlation has no units of measure (because the X and Y observations are standardized in the computation of r). It's merely a number between -1.0 and 1.0.

6. The value of the correlation is unaffected by *linear transformations* of either X or Y. In other words, if we convert each X_i to a new measurement scale using a conversion of the form $aX + b$, and each Y_i to a new scale using $cY + d$, then the correlation after making the conversions will be the same as it was before.
7. A correlation only measures the strength of the *linear* relationship between X and Y. In particular, curved relationships often lead to correlations near zero.
8. The correlation is not resistant to outliers.
9. A correlation doesn't imply a cause and effect relationship – there may be *confounding variables* "lurking" in the background and driving both X and Y up and down together (see Chapter 2).

The scatterplots below illustrate the correspondence between the value of r and the degree of linear association between X and Y.

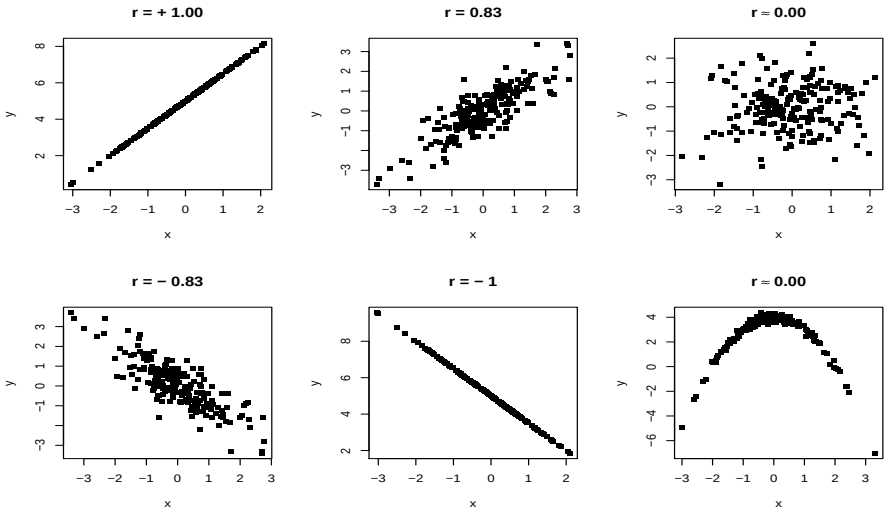


Figure 12.4: Values of r for six bivariate data sets.

Example 12.5: Correlation

The data below are the lengths (cm) and weights (g) of $n = 9$ prairie rattlesnakes sampled from the Pawnee National Grassland in northeastern Colorado as part of a study to investigate the use of snakes as a pollution bioindicator [1].

Lengths and Weights of Snakes

Snake	Length	Weight
1	85.7	331.9
2	64.5	121.5
3	84.1	382.2
4	82.5	287.3
5	78.0	224.3
6	81.3	245.2
7	71.0	208.2
8	86.7	393.4
9	78.7	228.3

We'll consider length as the explanatory variable and weight as the response. Here's the scatterplot.

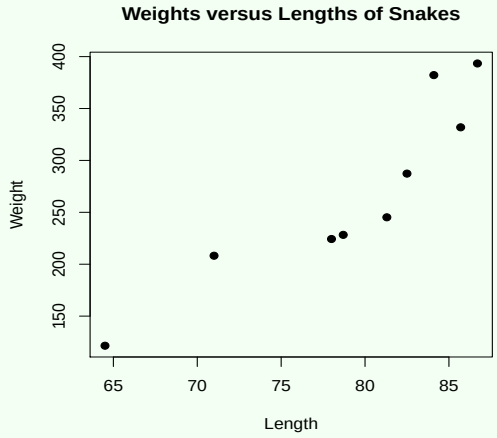


Figure 12.5: Scatterplot of weights versus lengths of prairie rattlesnakes.

The summary statistics for the lengths (x) and weights (y) are:

$$\bar{x} = 79.2 \qquad \bar{y} = 269.1$$
$$s_x = 7.3 \qquad s_y = 88.2$$

Quantities needed to calculate the correlation are given in the table below.

Snake	$x =$ Length	$y =$ Weight	$\frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y}$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
1	85.7	331.9	0.89	0.71	0.63
2	64.5	121.5	-2.01	-1.67	3.36
3	84.1	382.2	0.67	1.28	0.86
4	82.5	287.3	0.45	0.21	0.09
5	78.0	224.3	-0.16	-0.51	0.08
6	81.3	245.2	0.29	-0.27	-0.08
7	71.0	208.2	-1.12	-0.69	0.77
8	86.7	393.4	1.03	1.41	1.45
9	78.7	228.3	-0.07	-0.46	0.03
					$\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) = 7.19$

The correlation between lengths and weights is

$$\begin{aligned} r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) \\ &= \frac{1}{8}(7.19) \\ &= 0.90, \end{aligned}$$

which summarizes the strong, positive, approximately linear relationship seen in the scatterplot.

Comment: If we let

$$\mathrm{S}_{xx} = \sum_{i=1}^{n}(X_i-\bar{X})^2\,, \quad \mathrm{S}_{yy} = \sum_{i=1}^{n}(Y_i-\bar{Y})^2\,, \quad \text{and} \quad \mathrm{S}_{xy} = \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})\,, \tag{12.2}$$

and recall the definition of a sample standard deviation (Chapter 3), then it can be seen that an equivalent formula for the correlation is

$$r = \frac{\mathrm{S}_{xy}}{\sqrt{\mathrm{S}_{xx}\mathrm{S}_{yy}}}\,. \tag{12.3}$$

The quantities $\mathbf{S_{xx}}$, $\mathbf{S_{yy}}$, and $\mathbf{S_{xy}}$ are referred to as the ***X* sum of squares***, the ***Y* sum of squares**, and the ***sum of XY cross products***, respectively.

12.3.2 *t* Test for a Correlation

When there's *no relationship* between two variables X and Y, we expect their sample correlation r to be near zero, but it's unlikely to equal zero exactly because of random sampling error. We'll sometimes be interested in deciding whether an observed correlation is statistically significantly different from zero.

If a bivariate data set is a random sample from a ***bivariate population*** (one in which two variables X and Y can be measured on each population unit), we can use the sample correlation r to estimate the ***population correlation***, which is denoted $\boldsymbol{\rho}$. If in reality there's no relationship between X and Y, then the true correlation ρ would be zero and we'd expect r to be near zero too.

The null hypothesis of no relationship is written as

$$H_0 : \rho = 0$$

Values of r near -1.0 or 1.0 provide strong evidence against H_0. It turns out, though, to be difficult to determine p-values based on the value of r itself because the sampling distribution of r is rather complicated.

Instead, we use the ***correlation t test statistic***.

Correlation *t* Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \tag{12.4}$$

Notice that when r is close to 1.0, t will be large and positive, but when r is close to -1.0, t will be large and negative. When r is close to zero, t will be close to zero too. It follows that

1. *Large positive* values of t provide evidence in favor of $H_a : \rho > 0$.
2. *Large negative* values of t provide evidence in favor of $H_a : \rho < 0$.
3. *Both large positive and large negative* values of t provide evidence in favor of $H_a : \rho \neq 0$.

To decide if an observed t value provides statistically significant evidence in favor of the alternative hypothesis, we'll need to know its sampling distribution under the null.

Sampling Distribution of *t* Under H_0: Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is a sample is from a bivariate population whose correlation is ρ. Suppose also that the Y_i's follow a *normal* distribution (but not necessarily the X_i's). Then when

$$H_0 : \rho = 0$$

is true,

$$t \sim t(n-2)$$

(approximately), the t distribution with $n-2$ degrees of freedom.

P-values (and critical values for the rejection region approach) are obtained from the tail (or tails) of the $t(n-2)$ distribution in the direction specified by H_a, as summarized below.

t Test for a Correlation ρ

Assumptions: Data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are a random sample from a bivariate population in which the X and Y variables are both *normal* or n is large ($n \geq 20$), or, for each given value of X (not necessarily randomly selected), Y is a random variable that follows a *normal* distribution whose mean may depend on the value of X but whose standard deviation doesn't depend on X.

Null hypothesis: $H_0 : \rho = 0$.

Test statistic value: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or t is in rejection region.

Alternative hypothesis	P-value = area under t-distribution with $n-2$ d.f.:	Rejection region = t values such that:*
$H_a : \rho > 0$	to the right of t	$t > t_{\alpha,n-2}$
$H_a : \rho < 0$	to the left of t	$t < -t_{\alpha,n-2}$
$H_a : \rho \neq 0$	to the left of $-\lvert t\rvert$ and right of $\lvert t\rvert$	$t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$

* $t_{\alpha,n-2}$ is the $100(1-\alpha)$th percentile of the t distribution with $n-2$ d.f.

Example 12.6: t Test for a Correlation

Countries in sub-Saharan Africa have experienced high rates of urbanization in recent decades. To determine if this urbanization is associated with development, data from the United Nations Development Program's (UNDP) Human Development Report and the World Bank's World Development Report were analyzed [13].

The data are shown below. They include, for each of 40 sub-Saharan countries, the population, urbanization rate (percentage of the country's population living in cities), and human development index (HDI) value, which measures the country's health, education level, and standard of living. (Data for the three remaining sub-Saharan countries, Liberia, Rwanda, and Somalia were unavailable.)

Urbanization and Development in Africa

Country	Population	HDI	Urbanization
Angola	9.40	0.344	34.20
Benin	4.40	0.378	42.30
Botswana	1.20	0.678	50.30
BurkinaFaso	8.50	0.219	18.50
Burundi	5.10	0.241	9.01
Cameroon	11.20	0.481	48.90
CoteDIvoire	11.20	0.368	46.40
C.AfricanRepublic	2.90	0.347	41.20
Chad	5.40	0.318	23.80
P.R.Congo	2.10	0.519	62.50
D.R.Congo	33.40	0.383	30.30
Ethiopia	47.40	0.252	17.60
Eq.Guinea	0.40	0.465	48.20
Gabon	1.10	0.568	81.40
Gambia	0.80	0.291	32.50
Ghana	14.00	0.473	38.40
Guinea	5.40	0.277	32.80
GuineaBissau	1.60	0.295	23.70
Kenya	22.40	0.463	33.10
Lesotho	1.70	0.469	28.00
Madagascar	10.90	0.348	29.60
Malawi	8.00	0.344	24.90
Mali	8.00	0.236	30.00
Mauritania	1.90	0.361	57.70
Mauritius	1.10	0.833	41.30
Mozambiq	14.90	0.281	40.20
Namibia	1.70	0.644	30.90
Niger	7.30	0.207	20.60
Nigeria	110.10	0.691	44.00
Senegal	7.00	0.342	47.40
SieraLeone	3.90	0.185	36.60
Seychelles	0.08	0.845	63.80
SouthAfrica	34.00	0.717	50.40
Sudan	23.80	0.343	36.10
Swazilan	0.10	0.597	26.40
Tanzania	24.70	0.358	32.90
Togo	3.40	0.380	33.30
Uganda	16.20	0.340	14.23
Zambia	7.60	0.378	39.60
Zimbabwe	9.30	0.507	35.30

A scatterplot of the HDI versus urbanization is below.

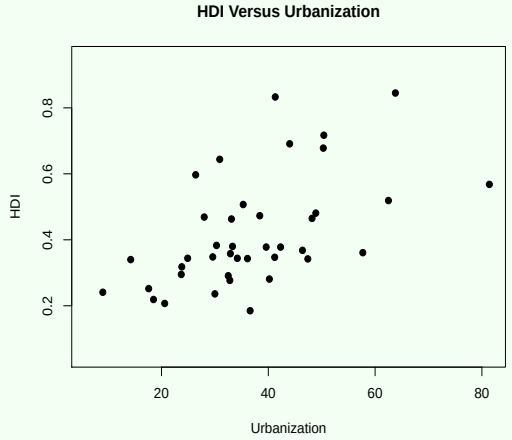


Figure 12.6: Scatterplot of human development index (HDI) versus urbanization.

The correlation between HDI and urbanization is $r = 0.54$, so for the t test of

$$\begin{aligned} H_0 : \rho &= 0 \\ H_a : \rho &> 0, \end{aligned}$$

the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.54 \cdot \sqrt{38}}{\sqrt{1-0.54^2}} = 3.95.$$

The p-value is the tail area to the right of 3.95 under the t distribution with $n - 2 = 38$ degrees of freedom. From a t distribution table, it's found to be 0.0003. Thus we reject H_0 and conclude that the observed positive correlation between urbanization and development is statistically significant, not just due to chance.

12.3.3 Confidence Interval for a Population Correlation

A confidence interval for an (unknown) population correlation ρ will give a range of plausible estimates for ρ. The sampling distribution of r is rather complicated, and the distribution of the statistic t in (12.4) is also complicated when $\rho \neq 0$. For these reasons, it turns out to be easier to base the confidence interval on the statistic

$$V = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right), \tag{12.5}$$

(where the log is the *natural log*). It can be shown that if n is large ($n \geq 25$ is large enough), the statistic V follows (approximately) a *normal* distribution with mean and standard error

$$\mu_V = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) \tag{12.6}$$

$$\sigma_V = \frac{1}{\sqrt{n-3}}, \tag{12.7}$$

Thus the random variable

$$Z = \frac{V - \mu_V}{\sigma_V} \tag{12.8}$$

follows a N(0, 1) distribution.

The point estimate for μ_V is just V, given by (12.5). A $100(1-\alpha)\%$ confidence interval for μ_V is

$$V \pm z_{\alpha/2}\,\sigma_V, \tag{12.9}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the N(0, 1) distribution. Commonly used $z_{\alpha/2}$ values are $z_{0.05} = 1.645$, $z_{0.025} = 1.96$, and $z_{0.005} = 2.58$ corresponding to 90%, 95%, and 99% levels of confidence, respectively. Substituting the right sides of (12.5) and (12.7) for V and σ_V in (12.9), the confidence interval for μ_V can be rewritten as

$$\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \pm z_{\alpha/2}\frac{1}{\sqrt{n-3}}. \tag{12.10}$$

We can be $100(1-\alpha)\%$ confident that μ_V will lie in this interval.

But we want a confidence interval for the population correlation ρ, not μ_V. So once endpoints of the confidence interval (12.10) for μ_V have been obtained, they are "backtransformed" via (12.6) to obtain the desired endpoints of the confidence interval for ρ. In other words, if we let ℓ and u be the lower and upper endpoints of the interval (12.10),

$$\ell = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) - z_{\alpha/2}\frac{1}{\sqrt{n-3}} \tag{12.11}$$

and

$$u \;=\; \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) + z_{\alpha/2}\frac{1}{\sqrt{n-3}}, \tag{12.12}$$

then we can be $100(1-\alpha)\%$ confident that

$$\ell \;<\; \mu_V \;<\; u,$$

which is to say we can be $100(1-\alpha)\%$ confident that

$$\ell \;<\; \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) \;<\; u.$$

Solving $\ell < \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$ for ρ gives

$$\frac{e^{2\ell}-1}{e^{2\ell}+1} \;<\; \rho\,,$$

and solving $\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) < u$ for ρ gives

$$\rho \;<\; \frac{e^{2u}-1}{e^{2u}+1}.$$

Therefore, the confidence interval for ρ is as follows.

Confidence Interval for ρ: Suppose we have a bivariate sample from a (bivariate) population whose correlation is ρ. Suppose also that the sample size n is large ($n \geq 25$). Then a **$100(1-\alpha)\%$ *confidence interval for ρ*** is

$$\left(\frac{e^{2\ell}-1}{e^{2\ell}+1},\quad \frac{e^{2u}-1}{e^{2u}+1}\right),$$

where ℓ and u are given by (12.11) and (12.12).

We can be $100(1-\alpha)\%$ confident the true (unknown) population correlation ρ will lie within this interval somewhere.

Example 12.7: Confidence Interval for a Correlation

For the study of urbanization and development in sub-Saharan Africa (Example 12.6), the sample size is $n = 40$ and the sample correlation between urbanization rate and human development index value is $r = 0.54$. For a 95% confidence interval for the true underlying correlation ρ, the ℓ and u values (12.11) and (12.12) are

$$\ell \;=\; \frac{1}{2}\log\left(\frac{1+0.54}{1-0.54}\right) - 1.96\frac{1}{\sqrt{40-3}} \;=\; 0.28$$

and

$$u \;=\; \frac{1}{2}\log\left(\frac{1+0.54}{1-0.54}\right) + 1.96\frac{1}{\sqrt{40-3}} \;=\; 0.93\,.$$

Thus the confidence interval is

$$\left(\frac{e^{2\ell}-1}{e^{2\ell}+1},\quad \frac{e^{2u}-1}{e^{2u}+1}\right) \;=\; \left(\frac{e^{2(0.28)}-1}{e^{2(0.28)}+1},\quad \frac{e^{2(0.93)}-1}{e^{2(0.93)}+1}\right) \;=\; (0.27,\, 0.73).$$

We can be 95% confident the true (unknown) underlying correlation between urbanization and human development, ρ, is between 0.27 and 0.73. Note that the interval doesn't contain zero, which is consistent with the result of the hypothesis test of Example 12.6.

12.4 Summarizing the Strength of a Nonlinear Relationship

12.4.1 Spearman Rank Correlation

Introduction

When the relationship between X and Y is nonlinear, the Pearson correlation doesn't adequately summarize the strength and direction of that relationship. Consider, for example, the scatterplots below.

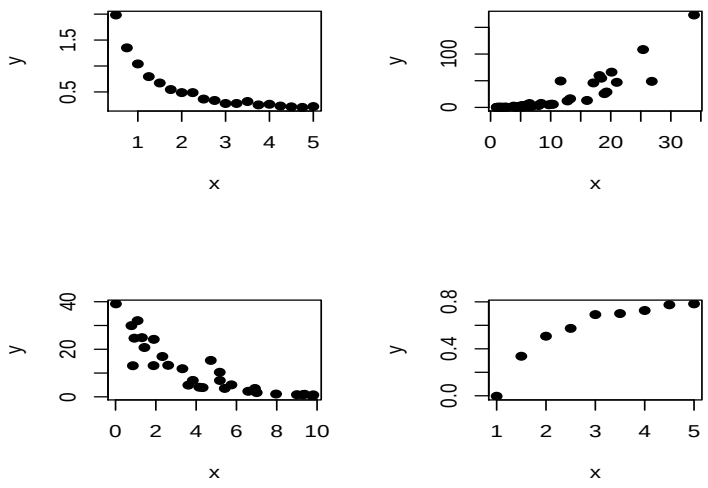


Figure 12.7: Scatterplots showing examples of monotone, nonlinear relationships.

The relationships in all four scatterplots are nonlinear and wouldn't adequately be summarized by the Pearson correlation. But all four are ***monotone***, meaning they're either always increasing or flat from left to right (a positive monotone relationship) or always decreasing or flat (a negative one), but never changing from increasing to decreasing or vice versa. An example of a relationship that's *not* monotone is the up-down curved relationship in the bottom right of Fig. 12.4.

Computing the Spearman Rank Correlation r_{sr}

To summarize the the strength and direction of a *monotone curved* relationship, we use the correlation between the *ranks* of the X's and the *ranks* of the Y's. This is called the the ***Spearman rank correlation***, denoted $\boldsymbol{r_{sr}}$.

Spearman Rank Correlation:

1. Determine the ranks of $X_1, X_2, \ldots X_n$, and denote these by $\boldsymbol{R_{X_1}}, \boldsymbol{R_{X_2}}, \ldots, \boldsymbol{R_{X_n}}$. Thus R_{X_i} is the rank of the ith observation X_i. If two or more observations are tied, assign to each of them the *average* of the ranks they would've been assigned if they hadn't been tied.
2. Determine the ranks of $Y_1, Y_2, \ldots Y_n$, and denote these by $\boldsymbol{R_{Y_1}}, \boldsymbol{R_{Y_2}}, \ldots, \boldsymbol{R_{Y_n}}$. Thus R_{Y_i} is the rank of the ith observation Y_i. If observations are tied, assign them the *average* rank.
3. The Spearman rank correlation, r_{sr}, is the correlation (12.1) between the ranks

$R_{X_1}, R_{X_2}, \ldots, R_{X_n}$ and $R_{Y_1}, R_{Y_2}, \ldots, R_{Y_n}$. In other words,

$$r_{sr} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{R_{X_i} - \bar{R}_X}{S_{R_X}} \right) \left(\frac{R_{Y_i} - \bar{R}_Y}{S_{R_Y}} \right) \tag{12.13}$$

where $\bar{R}_X$ and S_{R_X} are the sample mean and standard deviation of $R_{X_1}, R_{X_2}, \ldots, R_{X_n}$, and $\bar{R}_Y$ and S_{R_Y} are the sample mean and standard deviation of $R_{Y_1}, R_{Y_2}, \ldots, R_{Y_n}$.

Example 12.8: Spearman Rank Correlation

Nuclear weapons testing and nuclear accidents such as the one at Chernobyl in 1986 can discharge the radioactive contaminant radiocesium (^{137}Cs), which can then accumulate in forest ecosystems.

In a study to find out if concentrations of stable elements such as rubidium (Rb) could be used to predict the concentration of ^{137}Cs, both Rb and ^{137}Cs were measured in each of 29 mushrooms in a Japanese forest. The data are below.

Radioactivity in Mushrooms

Mushroom	^{137}Cs	Rb
1	42.4	75.4
2	449.0	115.0
3	179.0	88.2
4	182.0	110.0
5	230.0	105.0
6	28.7	56.5
7	8.2	31.0
8	36.0	47.7
9	34.0	49.8
10	5.4	59.8
11	55.6	93.1
12	127.0	85.4
13	65.2	87.6
14	317.0	137.0
15	675.0	133.0
16	44.5	29.5
17	45.9	76.5
18	27.4	73.9
19	1150.0	159.0
20	246.0	98.4
21	356.0	95.2
22	287.0	68.5
23	75.3	31.2
24	58.1	81.7
25	3110.0	214.0
26	598.0	125.0
27	602.0	98.5
28	135.0	101.0
29	143.0	78.6

Here's a scatterplot of the data.

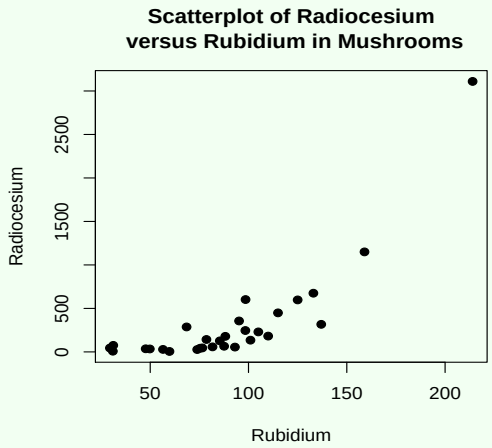


Figure 12.8: Scatterplot of radiocesium ^{137}Cs versus rubidium in mushrooms from a Japanese forest.

Notice from the scatterplot that the ^{137}Cs tends to increase as Rb increases, so the relationship between the two is monotone, but it's not linear. We'll summarize the strength of the relationship by the Spearman rank correlation.

To calculate r_{sr}, we first rank the ^{137}Cs values and then (separately) rank the Rb values. The resulting ranks are shown below along with the original values.

Mushroom	^{137}Cs	Rank of ^{137}Cs	Rb	Rank of Rb
1	42.4	7	75.4	10
2	449.0	24	115.0	24
3	179.0	17	88.2	16
4	182.0	18	110.0	23
5	230.0	19	105.0	22
6	28.7	4	56.5	6
7	8.2	2	31.0	2
8	36.0	6	47.7	4
9	34.0	5	49.8	5
10	5.4	1	59.8	7
11	55.6	10	93.1	17
12	127.0	14	85.4	14
13	65.2	12	87.6	15
14	317.0	22	137.0	27
15	675.0	27	133.0	26
16	44.5	8	29.5	1
17	45.9	9	76.5	11
18	27.4	3	73.9	9
19	1150.0	28	159.0	28
20	246.0	20	98.4	19
21	356.0	23	95.2	18
22	287.0	21	68.5	8
23	75.3	13	31.2	3
24	58.1	11	81.7	13
25	3110.0	29	214.0	29
26	598.0	25	125.0	25
27	602.0	26	98.5	20
28	135.0	15	101.0	21
29	143.0	16	78.6	12

Next we calculate the correlation between the ranks of ^{137}Cs and the ranks of Rb using (12.13). We get (using software)

$$r_{sr} = 0.84,$$

indicating a fairly strong positive monotone relationship. It's interesting to note that the Pearson correlation for this data set is $r = 0.81$, which is lower than r_{sr} because r is measuring the strength of the *linear* relationship.

Properties and Interpretation of r_{sr}

The Spearman rank correlation has properties similar to those of the Pearson correlation, but it measures the strength of the possibly *nonlinear* relationship between X and Y, not necessarily the *linear* one.

1. The value of the Spearman correlation will always be between -1.0 and 1.0.
2. The Spearman correlation tells us the *direction* of the relationship between X and Y:
   - Positive correlation values indicate a positive relationship.
   - Negative correlation values indicate a negative relationship.
3. The Spearman correlation also tells us how *strong* the relationship between X and Y is:
   - Spearman correlation values near zero imply a very weak relationship or none at all.
   - Spearman correlation values close to -1.0 or 1.0 imply a very strong *monotone* relationship, but *not necessarily* a linear one.
   - The extreme values $r_{sr} = -1.0$ and $r_{sr} = 1.0$ occur only when there's a *perfect monotone* relationship, that is, when the ranks of the X_i's are equal to those of their corresponding Y_i's.
4. The Spearman correlation doesn't depend on which variable is labeled X and which is labeled Y. It's a measure of association *between* the two variables.
5. The Spearman correlation has no units of measure. It's merely a number between -1.0 and 1.0.
6. The value of the Spearman correlation is unaffected by *order preserving transformations* of either X or Y. In other words, if we convert each X_i to a new measurement scale using a conversion that leaves their *ranks* unchanged, and each Y_i to a new scale that leaves their *ranks* unchanged, then the Spearman correlation after making the conversions will be the same as it was before.
7. The Spearman correlation only measures the strength of the *monotone* relationship between X and Y. In particular, curved *non-monotone* (up-down or down-up) relationships can lead to Spearman correlations near zero.
8. The Spearman correlation is somewhat (but not entirely) resistant to outliers.
9. A Spearman correlation doesn't imply a cause and effect relationship because there may be *confounding variables* "lurking" in the background and driving both X and Y up and down together (see Chapter 2).

12.4.2 t Test for a Monotone Relationship

Consider a bivariate population for which two *continuous* variables X and Y can be measured on each individual. In this section we'll see how to carry out a test to decide if X and Y follow a monotone relationship. The null hypothesis to be tested is

$$H_0 : \text{There's no relationship between } X \text{ and } Y$$

The test will be based on the value of the Spearman rank correlation between the observed X_i and Y_i values.

The ***Spearman rank correlation t test statistic*** is the same as the t test statistic for a correlation described in Subsection 12.3.2, but with r_{sr} used in place of r.

Spearman Rank Correlation t Test Statistic:

$$t = \frac{r_{sr}\sqrt{n-2}}{\sqrt{1-r_{sr}^2}}. \tag{12.14}$$

When r_{sr} is close to 1.0, t will be large and positive, and when r_{sr} is close to -1.0, t will be large but negative. When r_{sr} is close to zero, t will be close to zero too. Therefore,

1. *Large positive* values of t provide evidence in favor of

   H_a : There's a positive monotone relationship between X and Y

2. *Large negative* values of t provide evidence in favor of

   H_a : There's a negative monotone relationship between X and Y

3. *Both large positive and large negative* values of t provide evidence in favor of

   H_a : There's a monotone relationship between X and Y

To decide if an observed value of the test statistic provides statistically significant evidence in favor of H_a, we'll use its sampling distribution under H_0.

Sampling Distribution of t Under H_0: Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is a sample is from a bivariate population for which X and Y are *continuous*. Suppose also that n is large ($n \geq 20$ is sufficient). Then when

H_0 : There's no relationship between X and Y

is true,

$$t \sim t(n-2)$$

(approximately), the t distribution with $n-2$ degrees of freedom.

Because values of t that differ from zero in the direction specified by H_a count as evidence in favor of H_a, P-values (and critical values for the rejection region approach) are obtained from the corresponding tail (or tails) of the $t(n-2)$ distribution, as summarized below.

t Test for a Monotone Relationship

Assumptions: Data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are a random sample from a bivariate population in which the X and Y variables are both *continuous* and n is large ($n \geq 20$), or, for each given value of X (not necessarily randomly selected), Y is a random variable that follows *any continuous* distribution whose mean may depend on the value of X and n is large ($n \geq 20$).*

Null hypothesis: H_0 : There's no relationship between X and Y.

Test statistic value: $t = \frac{r_{rs}\sqrt{n-2}}{\sqrt{1-r_{rs}^2}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or t is in rejection region.

Alternative hypothesis	P-value = area under t-distribution with $n-2$ d.f.:	Rejection region = t values such that:**
H_a : There's a positive monotone relationship between X and Y	to the right of t	$t > t_{\alpha,n-2}$
H_a : There's a negative monotone relationship between X and Y	to the left of t	$t < -t_{\alpha,n-2}$
H_a : There's a monotone relationship between X and Y	to the left of $-\lvert t\rvert$ and right of $\lvert t\rvert$	$t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$

* When n is small, the test can still be carried out, but the t distribution shouldn't be used for obtaining p-values and critical values. Instead, the exact sampling distribution of the test statistic should be used. More information can be found in [7].

** $t_{\alpha,n-2}$ is the $100(1-\alpha)$th percentile of the t distribution with $n-2$ d.f.

Example 12.9: Test for a Nonlinear Monotone Relationship

For the data on radiocesium (^{137}Cs) and rubidium (Rb) in $n = 29$ mushrooms, the Spearman rank correlation was found in Example 12.8 to be $r_{sr} = 0.84$. Thus to test

H_0 : There's no relationship between ^{137}Cs and Rb

H_a : There's a positive monotone relationship between ^{137}Cs and Rb

the test statistic is

$$t = \frac{r_{sr}\sqrt{n-2}}{\sqrt{1-r_{sr}^2}} = \frac{0.84 \cdot \sqrt{27}}{\sqrt{1-0.84^2}} = 8.04.$$

The p-value is the area to the right of 8.04 under the the t distribution with $n-2 = 27$ degrees of freedom, and is found to be 0.0000. Thus we reject H_0 and conclude that the observed positive monotone relationship between ^{137}Cs and Rb seen in Fig. 12.8 is statistically significant, not just due to chance.

12.5 Linear Regression

12.5.1 Introduction

When two variables in data set have a linear relationship, we often want to find the equation of the straight line describing that relationship. Here are some ways in which we might use the line and its equation:

1. The line can enhance the appearance of the scatterplot of the two variables.
2. It can be used to quantify the amount by which Y changes, typically, for a given change in X.
3. It can be used to predict the value of Y from a given value of X.

The next two examples illustrate.

Example 12.10: Linear Regression

A scatterplot of the data on ages and diameters of aspen trees in Yellowstone National Park (Example 12.1) is shown again below, this time with a straight line that captures the overall positive linear relationship between the two variables.

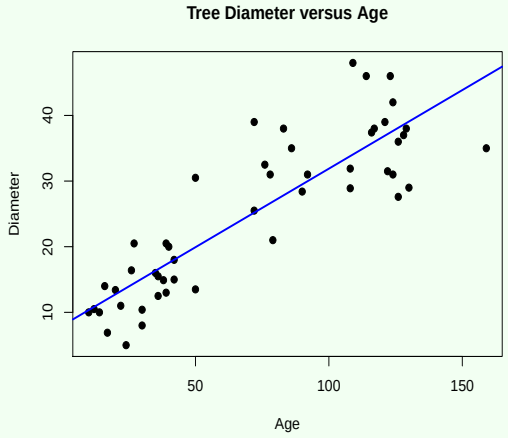


Figure 12.9: Scatterplot of ages and diameters at breast height with fitted line.

The line shown is called the *fitted regression line*. Its equation (obtained using software) is

$$\hat{Y} = 7.95 + 0.24X,$$

where $\hat{Y}$ = diameter and X = age. The "hat" (caret symbol) over the y is used to indicate that it's the equation of the *fitted regression line*. We'll see later how the equation was determined.

The slope of the line, 0.24, tells us that a tree's diameter increases by about 0.24 cm, on average, for each additional year of growth.

To predict the diameter of a tree that's, say, 100 years old, we plug $X = 100$ into the equation, which gives

$$\hat{Y} = 7.95 + 0.24(100) = 31.95.$$

Thus we predict the 100-year-old tree's diameter to be 31.95 cm.

Example 12.11: Linear Regression

For the data on lengths and weights of snakes given in Example 12.11, the scatterplot with the *fitted regression line* is shown below.

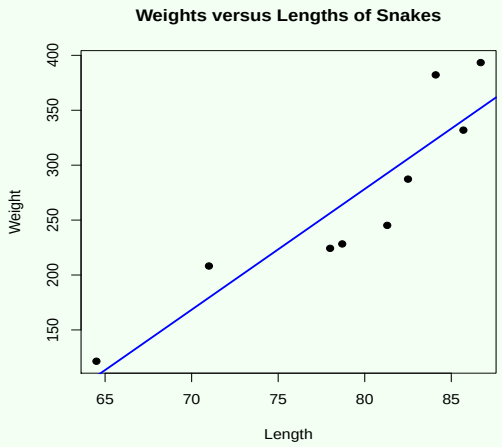


Figure 12.10: Scatterplot of lengths and weights of snakes with fitted line.

The equation of the line is

$$\hat{Y} = -601.1 + 11.0X,$$

where $\hat{Y}$ = weight and X = length.

The slope of the line, 11.0, says that on average, a snake's weight increases by about 11.0 grams for each additional centimeter of elongation.

The predicted weight of a snake that's, say, 62 cm long is obtained by plugging $X = 62$ into the equation of the line, which gives

$$\hat{Y} = -601.1 + 11.0(62) = 80.9.$$

Thus we predict that the 62-centimeter-long snake will weigh 80.9 grams.

12.5.2 The Simple Linear Regression Model

When two variables exhibit a relationship, there's usually an underlying natural process driving that observed pattern. For example, the positive relationship between diameters and ages of trees is driven by the biological process of tree growth, and the positive relationship between weights and lengths of snakes is a consequence of the physical properties of snakes and the laws of physics. A common goal of bivariate studies is to use data to draw inferences about those underlying processes. To this end, a *statistical model* that describes the data via the underlying process is employed, and values of unknown constants in the model (*parameters*) are *estimated* from the data.

For *linear* relationships, the ***simple linear regression model*** reflects an underlying linear process, but it also allows for "deviations" away from that straight line "overall pattern."

Simple Linear Regression Model: A statistical model for describing bivariate data that exhibit a linear relationship is:

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \epsilon_i, \tag{12.15}$$

where

- Y_i is the observed value of the response variable for the ith individual ($i = 1, 2, \ldots, n$).
- X_i is the observed value of the predictor variable for the ith individual.
- $\boldsymbol{\beta_0}$ is the ***y-intercept*** of the underlying ***true regression line***.
- $\boldsymbol{\beta_1}$ is the ***slope*** of the ***true regression line***.
- ϵ_i is a random error term following a N(0, σ) distribution, and the ϵ_i's are independent of each other.

The model relates the response variable Y to the predictor X by way of the so-called ***true regression line***, $\boldsymbol{\beta_0 + \beta_1 X}$, that represents the underlying process driving the linear relationship in the data. The (unkown) ***parameters*** of the model are the ***coefficients*** β_0 and β_1, representing the *y-intercept* and *slope* of the true regression line, and σ, the error distribution's standard deviation. In practice, their values will be estimated from the data.

The random error ϵ represents the deviation of Y above or below the line due to the net effect of all *other* factors, *besides* the X variable, and also measurement error. For a snake of a given length, its weight will deviate above or below the line due to factors such as its caloric intake, its metabolic rate, the density of its bones, and so on. For a tree of a given age, its diameter will deviate above or below the line due to factors such as weather conditions as it grew, spatial heterogeneity in soil nutrients and moisture, and so on.

The standard deviation σ represents the size of a typical error. In the model, its value doesn't depend on X, so the amount of variation of Y above or below the line is assumed to be the same regardless of the value of X. The model is depicted graphically below.

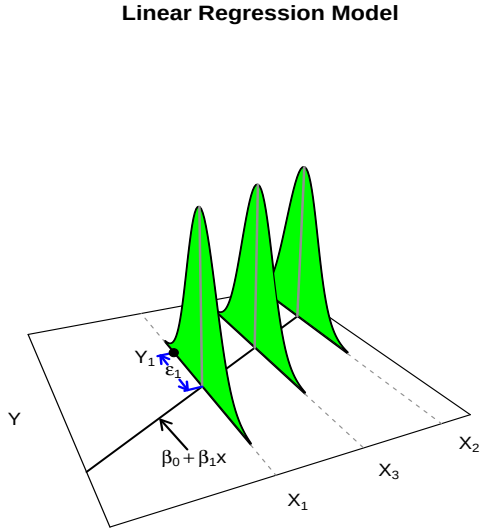


Figure 12.11: Graphical depiction of the simple linear regression model.

Comments:

- The linear regression model can be thought of as describing separate, distinct populations of Y values, one for each value of X, where the population means all lie on the line $\beta_0 + \beta_1 X$. Three

such populations are depicted in Fig. 12.11. In the study of tree diameters, each population would correspond to trees of a given age, and in the study of snake weights, each would correspond to snakes of a given length. The true regression line $\beta_0 + \beta_1 X$ is sometimes called the ***true mean response line***.

- The aforementioned Y populations are assumed to be *normal* and to all have the same standard deviation σ.
- No assumptions are made about the X variable. It's values don't even have to be randomly selected – they can be hand-picked, as would be the case, say, for dose levels in a toxicity experiment.
- The assumption of normality (and independence) of the ϵ's is only needed for the purpose of testing hypotheses about β_0 and β_1 and (constructing confidence intervals for them). If hypothesis testing isn't going to be carried out (nor confidence intervals constructed), there's no need to assume normality (or independence).

12.5.3 Least Squares Estimation of Model Parameters

The Method of Least Squares

When we estimate the slope and intercept of the true regression line, we say that we've ***fitted*** the regression model to the data. We fit the model using the ***method of least squares***, which is based on the principle that the line that "best fits" the points in a scatterplot is the one whose y-intercept $\boldsymbol{b_0}$ and slope $\boldsymbol{b_1}$ result in the smallest possible value for sum of squared vertical deviations of the Y_i values away from the line,

$$\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2 \,. \tag{12.16}$$

For the data on lengths and weights of snakes, the deviations whose sum of squares is given by (12.16) are the vertical lines in the scatterplot below.

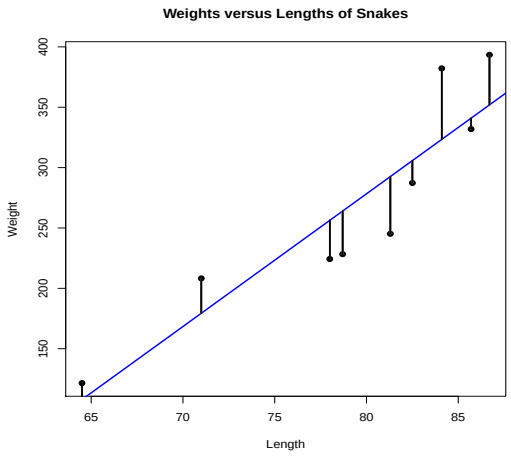


Figure 12.12: Scatterplot of lengths (X) and weights (Y) of snakes and the vertical deviations of the weights away from the regression line.

The line shown was fitted using the method of least squares, and the sum of squared deviations turns out to be 11,165. For *any other* line, the sum of squared deviations would be larger.

A line fitted by least squares is called a ***fitted regression line*** and denoted

$$\hat{Y} \;=\; b_0 + b_1 X, \tag{12.17}$$

The symbol $\hat{\boldsymbol{Y}}$ is used (instead of just Y) to indicate that the line is the *fitted regression line*. The y-intercept b_0 and slope b_1 are called the ***least squares estimates*** of the true (unknown) model parameters β_0 and β_1. The following fact tells us how they're computed from the data.

Fact 12.1 The y-intercept b_0 and slope b_1 of the fitted least squares regression line are computed from the data using

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}}, \tag{12.18}$$

where S_{xy} and S_{xx} are as given by (12.2), and

$$b_0 = \bar{Y} - b_1\bar{X}. \tag{12.19}$$

Comment: The slope of the fitted regression line and the correlation will *always* have the *same sign*. This is because, it can be shown, another form for the slope is

$$b_1 = r\frac{S_y}{S_x}, \tag{12.20}$$

where r is the correlation and S_x and S_y are the X and Y sample standard deviations (which are always positive).

Example 12.12: Linear Regression

For the data on ages and diameters of aspen trees (Example 12.1), the fitted regression line given in Example 12.10,

$$\hat{Y} = 7.95 + 0.24X\,,$$

was computed using statistical software, which reported the values of (12.18) and (12.19) as

$$\begin{aligned} b_1 &= 0.24 \\ b_0 &= 7.95. \end{aligned}$$

Some Cautionary Notes About Least Squares Regression

Linear regression analysis has its limitations, a few of which are listed below.

1. Linear regression should only be used if either the data exhibit, at least approximately, a linear relationship or there are theoretical grounds for assuming X and Y are linearly related. If they *aren't* linearly related, a few courses of action are suggested in Section 12.6.
2. Be cautious of ***extrapolation***, which means using the fitted regression line to predict Y for values of X outside the range that the line was fitted to. Extrapolation can lead to erroneous predictions because the linear relationship may not continue outside that range. See Example 12.13.
3. Beware of ***influential points*** in the data, outliers that have a strong influence on the fitted regression line. Outliers in the horizontal (X) direction can be particularly influential. See Example 12.15.

Example 12.13: Beware of Extrapolation

Although new waste disposal sites limit leakage by using clay and synthetic liners, most older ones don't have such bottom seals, and pollutants can leak into downstream groundwater used for drinking.

Environmental scientists and policy makers are interested, therefore, in exploring their options for remediation of former waste sites. One option that sometimes works is to simply rely on natural attenuation, that is, the downstream decrease of pollutant concentrations as a result of natural retention processes.

The data below, from a study of the efficacy of natural attenuation, show chlorofluorocarbons (CFCs, pmol/L) measured in groundwater monitoring wells at various distances (m) downstream from an abandoned waste disposal site near Berlin, Germany [11].

CFCs in Groundwater

Monitoring Well	Distance From Disposal Site	CFCs
1	20	530.00
2	10	160.00
3	60	270.00
4	120	55.00
5	200	39.00

The fitted regression line, using CFCs as the response and distance from the waste site as the predictor, is

$$\hat{Y} = 354.59 - 1.75X\,.$$

This line is shown in the scatterplot of the data below.

If we were to use the line to predict the CFCs in groundwater 300 m from the disposal site, it would be an *extrapolation* because the line was fitted to data for which the distances only go to 200 m. The predicted CFCs would be

$$\hat{Y} = 354.59 - 1.75(300) = -170.4\,,$$

which is obviously unreasonable because CFC values can't be negative. The problem arises because even though CFCs follow approximately a linear relationship to distance up to 200 m from the disposal site, that linear relationship doesn't continue beyond 200 m. Instead, the CFC values level off at larger distances, as depicted by the dotted curve in the scatterplot below.

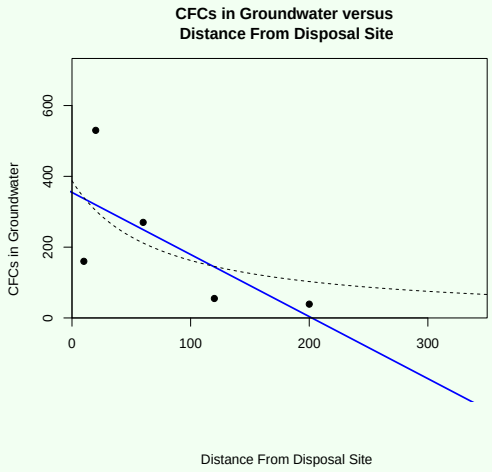


Figure 12.13: Scatterplot of CFCs in groundwater versus distance from disposal site with fitted regression line (blue) and a curve depicting a more realistic relationship for distances greater than 200 m (dashed).

Example 12.14: Influential Points in Linear Regression

In the original data set on lengths and weights of snakes reported in [1], a gopher snake was included along with the nine prairie rattlesnakes. In the scatterplot below, the gopher snake is represented by the red outlier on the right. To demonstrate the *influence* it would have on the fitted regression line, the line was fitted with and without the outlier included in the data. Both lines are shown in the scatterplot.

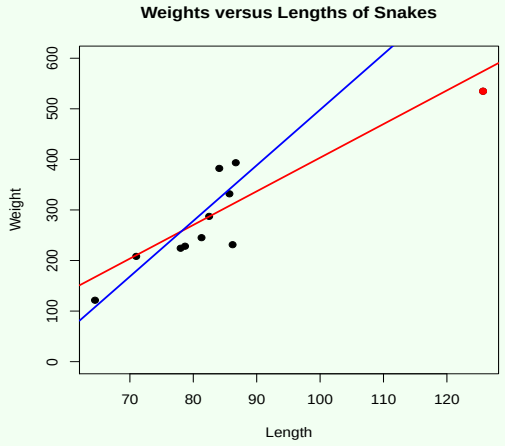


Figure 12.14: Scatterplot of weights versus lengths of snakes with regression line fitted to the data with (red) and without (blue) the outlier (red point) included.

12.5.4 Fitted Values and Residuals

Fitting a model to a set of bivariate data provides an estimate of the (unknown) true regression line. For each of the n individuals in a bivariate data set, we define the individual's ***fitted value*** (also called ***predicted value***), denoted $\hat{Y}_i$, as follows.

Fitted Value: For the ith individual in the data set,

$$\hat{Y}_i \;=\; b_0 + b_1 X_i, \tag{12.21}$$

where X_i is the value of the predictor variable for that individual.

The fitted values are just values we'd predict for Y by plugging the observed X_i's into the equation of the fitted line. *They all lie on the fitted line* and correspond to the nonrandom, linear "overall pattern" in the data. There will be n fitted values, one for each individual in the data set. For the snakes data in Fig. 12.14, they're the points along the fitted line from which the vertical deviations emanate.

The next facts says that the average of the fitted values is equal to the average of the Y_i's.

Fact 12.2 The mean of the fitted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$ in a regression analysis is equal to the mean of the observed responses $Y_1, Y_2, \ldots, Y_n$, that is,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i \;=\; \bar{Y}.$$

We'll also be interested in evaluating the random "deviations" away from the overall pattern, that is, values of the error term ϵ in the model. A ***residual***, denoted $\boldsymbol{e_i}$, is defined as the difference between the ith individual's observed Y value and the fitted value for that individual.

Residual: For the ith individual in the data set,

$$e_i \;=\; Y_i - \hat{Y}_i, \tag{12.22}$$

where Y_i is the observed response for that individual and $\hat{Y}_i$ is the fitted value.

For the snakes data, the residuals are the vertical line segments Fig. 12.14. A residual will be positive if Y_i lies above the line, and negative if it lies below the line. There will be n residuals in total, one for each individual in the data set.

By the definitions of fitted values and residuals, we can write a residual as

$$e_i \;=\; Y_i - (b_0 + b_1 X_i),$$

and rearranging this, we can write an observation Y_i as

$$Y_i \;=\; b_0 + b_1 X_i + e_i.$$

Comparing this to the model (12.15), it's apparent that *the residual e_i approximates the random error term ϵ_i*, and therefore, like the errors, correspond to the net effect of all other factors *besides* X on the response variable. In Section 12.5.9, we'll use the residuals to estimate the standard deviation of the $\mathrm{N}(0, \sigma)$ error distribution, and in Section 12.5.15 we'll use them to check the normality assumption.

It turns out that the residuals sum to zero.

Fact 12.3 The residuals in a regression analysis sum to zero, that is,

$$\sum_{i=1}^{n} e_i \;=\; 0.$$

12.5.5 Two Sources of Variation in Y

In one-factor ANOVA (Chapter 10), we partitioned the total variation in the response variable into two parts, between-groups variation due to the effect of the factor (the *treatment sum of squares*) and within-groups variation due to random error (the *error sum of squares*). We'll do something similar in regression, but in this case the two sources of variation are the X variable and random error. The following example illustrates.

Example 12.15: Partition of Variation in Linear Regression

Consider the data on lengths and weights of snakes (Example 12.5). The snakes are of different lengths, which explains some of the variation in their weights but not all of it. If it did explain all of their weight variation, the points in the scatterplot would lie *exactly* on a straight line.

Because the points *don't* all lie on a line, we know that there are other factors, *besides* length, that determine a snake's weight. These other factors (metabolic rate, caloric intake, bone density, and so on) show up as residuals. The larger their contribution to variation in weights is, the larger the residuals will be.

Thus we can think of variation in snakes' weights as arising from two sources:

1. Differences in their lengths (the X variable).
2. Differences in the values of all the *other* factors (*besides* length) that affect weight.

In general, for any regression analysis, there will be two sources of variation in the responses:

1. Variation due to differences in the value of the predictor X from one individual to the next.
2. Variation due to differences in the values of all other factors (besides X) from one individual to the next.

These correspond, respectively, to the nonrandom linear "overall pattern" in the data and the random "deviations" away from that pattern, that is, errors.

12.5.6 Sums of Squares

Introduction

We'll measure the contributions of the two sources of variation in the response variable using sums of squares, in much the way we used them to measure between- and within-groups variation in one-factor ANOVA. In the context of regression, we'll use sums of squares to:

1. Assess *how well* the regression model fits the data.
2. Estimate the standard deviation σ of the error distribution.
3. Test a hypothesis to decide if the predictor variable X explains any of the variation in the responses.

Variation Due to Random Error

Variation in Y due to random error, that is, due to all other factors besides X, is measured by the ***error sum of squares***, denoted **SSE** and defined as follows.

Error Sum of Squares:

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \,. \tag{12.23}$$

The error sum of squares is just the ***sum of squared residuals***. The SSE will be large when the variation in Y due to random error is large.

Variation Due to X

Variation in Y due to differences in the value of the predictor variable X from one individual to the next is measured by the ***regression sum of squares***, denoted **SSR** and defined as follows.

Regression Sum of Squares:

$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \,. \tag{12.24}$$

The regression sum of squares is the sum of squared deviations of the fitted values away from the overall mean $\bar{Y}$ of the responses. The SSR reflects variation in the fitted values which, recall, lie on the fitted line. It will be large when the line has a *steep slope*, that is, when variation in Y due to differences in the values of $X_1, X_2, \ldots, X_n$ is large, and small otherwise.

12.5.7 ANOVA-Like Partition of the Total Variation in Y

Introduction

In a regression analysis, we can think of the variation in Y as arising either from the nonrandom linear relationship to X or from random error. We'll see in a bit that the two types of variation account for *all* of the variation in the responses.

Total Variation

To see what this means, we'll first need a measure of the *total variation* in the response variable. We use the same one that's used in one-factor ANOVA, namely the ***total sum of squares***, denoted **SSTo** and defined as follows.

Total Sum of Squares:

$$\text{SSTo} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2. \tag{12.25}$$

Because SSTo measures total variation in the responses, it will be large if either the fitted line has a steep slope or the deviations away from the line are large. Thus SSTo reflects both variation due to X and variation due to error.

Partition of the Total Variation

The sums of squares in a regression analysis satisfy the following ***ANOVA-like partition*** of the total variation in the responses.

Fact 12.4 The sums of squares defined above satisfy the following relation.

$$\text{SSTo} \;=\; \text{SSR} + \text{SSE}\,. \tag{12.26}$$

This decomposes the variation in the responses as:

$$\text{Total Variation} \;=\; \text{Variation Due to } X \;+\; \text{Variation Due to Error}$$

Example 12.16: ANOVA-Like Partition

For the data on lengths and weights of snakes (Examples 12.5), the sums of squares (obtained using statistical software) are

$$\text{SSTo} \;=\; 62,255, \qquad \text{SSR} \;=\; 51,090, \qquad \text{and} \qquad \text{SSE} \;=\; 11,165.$$

As expected, SSTo = SSR + SSE since

$$62,255 \;=\; 51,090 + 11,165.$$

This shows that the majority of the variation in weights (51,090 out of 62,255) is due to differences in their lengths, and only a smaller portion (11,165) due to random error.

12.5.8 Degrees of Freedom

As was the case for ANOVA, each sum of squares has associated with it a corresponding degrees of freedom.

Degrees of Freedom: For linear regression, the degrees of freedom are:

$$\begin{aligned} df \text{ for SSTo} &= n-1 \\ df \text{ for SSR} &= 1 \\ df \text{ for SSE} &= n-2 \end{aligned}$$

Degrees of freedom will be used later to determine which t and F distributions p-values are obtained from when performing hypothesis tests related to the regression analysis.

The degrees of freedom are additive in the following sense.

Fact 12.5 The degrees of freedom given above satisfy the following relation.

$$df \text{ for SSTo} \;=\; df \text{ for SSR} + df \text{ for SSE}.$$

12.5.9 Mean Squares

Introduction

As for ANOVA, a ***mean square*** in linear regression is a sum of squares divided by its degrees of freedom. The two mean squares, the ***mean square for regression***, or **MSR**, and the ***mean squared error***, or **MSE**, will be used later to test for a linear relationship between Y and X.

Mean Squares: For linear regression, the mean square for regression and mean squared error are

$$\text{MSR} = \frac{\text{SSR}}{1} = \text{SSR} \tag{12.27}$$

$$\text{MSE} = \frac{\text{SSE}}{n-2}. \tag{12.28}$$

Estimating σ

The MSE can be thought of as an $(n-2)$ "average" *squared* residual, so its square root measures the size of a typical residual. Thus because the residuals are approximations of the random errors ϵ in the regression model, we use $\sqrt{\text{MSE}}$ as an estimator of the standard deviation of the N(0, σ) error distribution.

Estimator of σ: In a linear regression analysis, the estimator of σ, denoted $\hat{\sigma}$, is

$$\hat{\sigma} = \sqrt{\text{MSE}}.$$

Example 12.17: Estimating σ

For the data on lengths and weights of $n = 9$ snakes (Example 12.5), the mean squares are

$$\text{MSR} = 51,090 \quad \text{and} \quad \text{MSE} = 1,595$$

and so $\sqrt{\text{MSE}} = 39.9$. This is the size of a typical deviation of a snake's weight above or below the fitted line in Fig. 12.14, and serves as an estimate of the standard deviation σ of the N$(0, \sigma)$ distribution of the error term ϵ in the regression model.

12.5.10 Assessing the Fit of the Regression Line

Introduction

After fitting a regression model, we usually want to know how successful the predictor variable is at explaining variation in the response. Knowing how well a given predictor explains Y variation is especially useful when we need to decide *which* of *two* predictors does the job better. It's convenient, therefore, to have a statistic that tells us how well each line fits the data. Two commonly used statistics that serve this purpose are:

1. The mean squared error, MSE, or its square root.
2. The coefficient of determination, or R^2.

The MSE as a Measure of Fit of the Regression Line

Because the mean squared error (and its square root) reflect the sizes of the residuals, and a line "fits" the data better when the residuals are small, we can use the MSE (or $\sqrt{\text{MSE}}$) as a measure of how well a given line fits the data. A *smaller* value of the MSE (or $\sqrt{\text{MSE}}$) indicates a *better* fitting line.

The Coefficient of Determination R^2

One criticism of the MSE as a measure of how well a line fits the data is that its value depends on the units of measure for the response variable. For example, changing the measurement scale of Y from inches to centimeters will change the value of the MSE.

The ***coefficient of determination***, denoted $\boldsymbol{R^2}$ (and usually just called "R squared"), is an alternative measure of fit whose value *doesn't* depend on the units of measure for Y.

Coefficient of Determination:

$$R^2 \;=\; \frac{\text{SSR}}{\text{SSTo}} \;=\; 1 - \frac{\text{SSE}}{\text{SSTo}}. \tag{12.29}$$

Because SSR measures variation in the responses due to differences in the value of the predictor X, and SSTo measures *total* variation in the responses, we can think of R^2 as

$$R^2 \;=\; \frac{\text{Variation in } Y \text{ Due to } X}{\text{Total Variation in } Y}.$$

In other words, R^2 can be interpreted as the *proportion* of variation in the response variable that can be explained by differences in values of the predictor X (and the linear relationship of Y to X).

Properties and Interpretation of $\boldsymbol{R^2}$: The following properties of R^2 provide insight into its interpretation.

1. The value of R^2 will always be between zero and one (because it's a proportion).
2. R^2 tells us *how well* the regression line fits the data:
   - An R^2 value near zero means the line *doesn't* fit very well (because only a small fraction of the Y variation is explained by X).
   - An R^2 value near one means the line fits the data very well (because a large fraction of the Y variation is explained by X).

For more insight, several scatterplots and their corresponding R^2 values are shown below.
Perhaps not surprisingly, R^2 is related to the correlation r from Section 12.3, as stated in the following fact.

Fact 12.6 It can be shown that the coefficient of determination R^2 is equal to the square of the correlation r, that is,

$$R^2 \;=\; r^2.$$

Example 12.18: Coefficient of Determination $\boldsymbol{R^2}$

For the data on lengths and weights of snakes shown in Fig. 12.10, the total sum of squares and

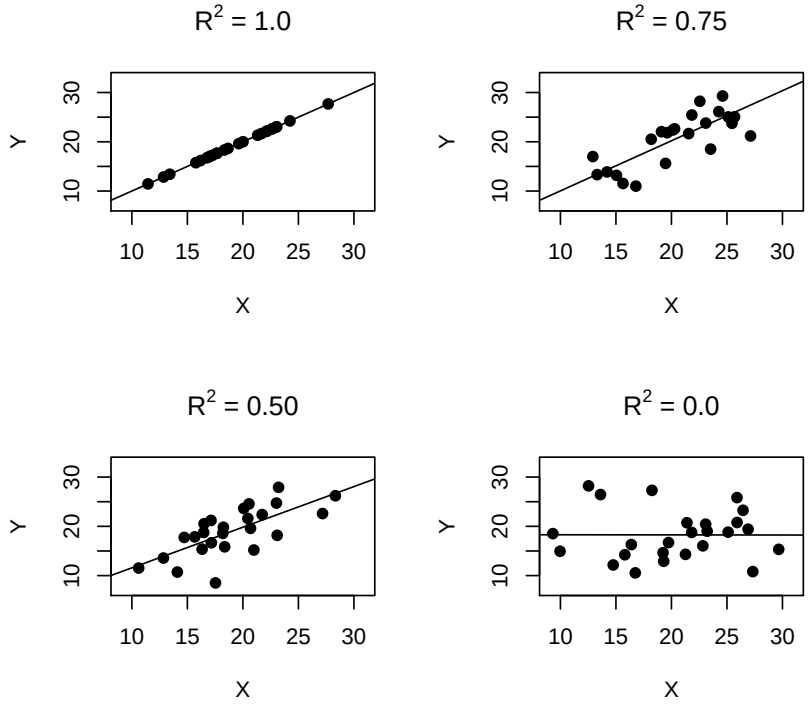


Figure 12.15: Scatterplots showing bivariate data sets with different values of R^2.

regression sum of squares, from Example 12.17, are

$$\text{SSTo} \;=\; 62,255 \quad \text{and} \quad \text{SSR} \;=\; 51,090,$$

so the coefficient of determination is

$$R^2 \;=\; \frac{\text{SSR}}{\text{SSTo}} \;=\; \frac{51,090}{62,255} \;=\; 0.821.$$

Thus about 82.1% of the variation in snakes' weights is attributable to differences in their lengths (and the linear relationship of weight to length). The other 17.9% is due to the combined effects of all *other* factors (metabolic rate, caloric intake, bone density, and so on).

The correlation between length and weight, from Example 12.5, is $r = 0.90$, and its square is $0.90^2 = 0.81$ which, up to roundoff error, is equal to R^2 as expected.

12.5.11 *t* Tests for the Slope and Intercept of the Regression Model

Introduction

In the linear regression model

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \epsilon_i\,, \tag{12.30}$$

the slope parameter β_1 specifies the average change in Y associated with a one-unit increase in X. If β_1 was zero, there'd be no change in Y as X changes (the true regression line would be horizontal), in which case there'd be *no relationship* between Y and X. A slope different from zero means there's a linear

relationship

We'll be interested, therefore, in testing the null hypothesis of *no relationship*,

$$H_0 : \beta_1 = 0$$

Sampling Distribution of b_1

Because the estimate b_1 of the true (unknown) slope β_1 is computed from the data, and the response values $Y_1, Y_2, \ldots, Y_n$ vary from one sample to the next, b_1 is a random variable that varies from sample to sample. To get a sense of its variability, fitted regression lines are shown below for 50 sets of response data randomly generated using the model (12.30) (with the ϵ values varying from one sample to the next but the X values kept the same).

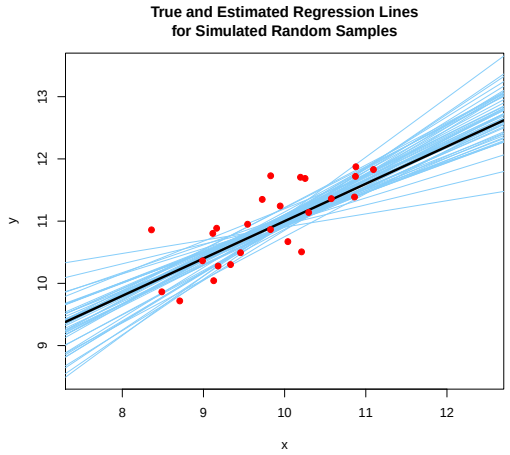


Figure 12.16: Fitted regression lines (light blue) for 50 sets of data (only one of which is shown) randomly generated from the regression model (12.30) with $\beta_0 = 5$, $\beta_1 = 0.6$, and the ϵ's varying according to a $\mathrm{N}(0, \sigma)$ distribution with $\sigma = 0.5$. The X values are the same for the 50 sets of data. Only the Y values vary. The black line is the true regression line.

The hypothesis test for β_1 will be based on how different b_1 is from zero, so to carry out the test, we'll need to be able to recognize when an observed difference from zero is larger than can be explained by chance. For this, we'll need the sampling distribution of b_1.

Fact 12.7 Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are bivariate observations described by the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where the ϵ_i's are independent and follow a $\mathrm{N}(0, \sigma)$ distribution.

Then b_1 follows a *normal* distribution with mean $\boldsymbol{\mu_{b_1}}$ and standard error $\boldsymbol{\sigma_{b_1}}$

$$\mu_{b_1} \;=\; \beta_1$$

and

$$\sigma_{b_1} \;=\; \frac{\sigma}{\sqrt{S_{xx}}},$$

which is to say,

$$b_1 \;\sim\; \mathrm{N}\left(\beta_1,\; \sigma_{b_1}\right),$$

where S_{xx} is the X sum of squares given in (12.2).

It follows that if we standardize b_1, the resulting random variable Z follows a standard normal distribution, that is,

$$Z = \frac{b_1 - \beta_1}{\sigma_{b_1}} \sim \mathrm{N}(0, 1).$$

t Test Statistic for a Slope

In practice, we estimate the standard error of b_1 by replacing σ in Fact 12.7 by its estimate, $\sqrt{\mathrm{MSE}}$. This gives the ***estimated standard error of b_1***, denoted $\boldsymbol{S_{b_1}}$.

(Estimated) Standard Error of b_1:

$$S_{b_1} = \frac{\sqrt{\mathrm{MSE}}}{\sqrt{\mathrm{S}_{xx}}}, \tag{12.31}$$

where S_{xx} is the X sum of squares given by (12.2).

This standard error represents the size of a typical sampling error when b_1 is used to estimate the true value β_1.

The next fact says that when we standardize b_1 using an *estimated* standard error, the resulting standardized variable follows a t distribution. It will be used to develop the t test procedure for a slope.

Fact 12.8 Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are bivariate observations described by the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where the ϵ_i's are independent and follow a $\mathrm{N}(0, \sigma)$ distribution. Then

$$\frac{b_1 - \beta_1}{S_{b_1}} \sim t(n-2),$$

the t distribution with $n-2$ degrees of freedom.

The ***t test statistic for a slope***, denoted $\boldsymbol{t}$, is obtained by replacing β_1 in Fact 12.8 by it's null hypothesized value zero.

***t* Test Statistic for a Slope**:

$$t = \frac{b_1 - 0}{S_{b_1}}. \tag{12.32}$$

Because b_1 is an estimator of the true slope β_1, if H_0 was true, and β_1 equal to zero, we'd expect b_1 to be close to zero, in which case t would be close to zero too. But if H_a was true, we'd expect b_1 to differ from zero in the direction specified by H_a, in which case t would differ from zero in that direction too. Therefore,

1. *Large positive* values of t provide evidence in favor of $H_a : \beta_1 > 0$.
2. *Large negative* values of t provide evidence in favor of $H_a : \beta_1 < 0$.
3. *Both large positive and large negative* values of t provide evidence in favor of $H_a : \beta_1 \neq 0$.

Furthermore, t measures (approximately) how many standard errors the estimate b_1 is away from zero, and in what direction (positive or negative). To decide if an observed value of t provides statistically significant

evidence against the null hypothesis, we'll need its sampling distribution under H_0, which, from Fact 12.8, is the following.

Sampling Distribution of t Under H_0: Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are bivariate observations described by the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where the ϵ_i's are independent and follow a $N(0, \sigma)$ distribution. Then when

$$H_0 : \beta_1 = 0$$

is true,

$$t \sim t(n-2).$$

t Test for a Slope Procedure

P-values and critical values (for the rejection region approach) for the t test for a slope are obtained from the tails of the $t(n-2)$ distribution, as summarized below.

t Test for β_1

Assumptions: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are bivariate observations described by the simple linear regression model (12.15), where the ϵ_i's are independent and either they follow a $N(0, \sigma)$ distribution or n is large ($n \geq 20$).

Null hypothesis: $H_0 : \beta_1 = 0$.

Test statistic value: $t = \frac{b_1}{S_{b_1}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or t is in rejection region.

Alternative hypothesis	P-value = area under t distribution with $n-2$ d.f.:	Rejection region = t values such that:*
$H_a : \beta_1 > 0$	to the right of t	$t > t_{\alpha,n-2}$
$H_a : \beta_1 < 0$	to the left of t	$t < -t_{\alpha,n-2}$
$H_a : \beta_1 \neq 0$	to the left of $-\lvert t\rvert$ and right of $\lvert t\rvert$	$t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$

* $t_{\alpha,n-2}$ is the $100(1-\alpha)$th percentile of the t distribution with $n-2$ d.f.

Note: Statistical software packages always report the results of the *two-sided* test of

$$H_0 : \beta_1 = 0 \tag{12.33}$$
$$H_a : \beta_1 \neq 0 \tag{12.34}$$

when they perform regression analyses. To carry out a *one-sided* test, when the observed t value differs from zero in the direction specified by H_a, we divide the reported p-value by two.

t Test for an Intercept

Although not usually of interest, statistical software packages also report the results of a t test for β_0, the true (unknown) y-intercept in the linear regression model, when they perform a regression analysis. The

hypotheses tested are

$$\begin{aligned} H_0 : \beta_0 &= 0 \\ H_a : \beta_0 &\neq 0 \end{aligned} \tag{12.35}$$

The ***t test statistic for an intercept*** is

t Test Statistic for an Intercept:

$$t = \frac{b_0 - 0}{S_{b_0}},$$

where b_0 is the least squares estimate of β_0 and $\mathbf{SE(b_0)}$ is the ***estimated standard error of b_0***,

(Estimated) Standard Error of b_0:

$$S_{b_0} = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{X}^2}{\text{S}_{xx}}\right)}, \tag{12.36}$$

where S_{xx} is given by (12.2).

When the null hypothesis (12.35) is true, the test statistic t follows a t distribution with $n-2$ degrees of freedom, from which p-values are obtained.

Carrying Out the t Tests for the Slope and Intercept

When a regression analysis is carried out using statistical software, the software summarizes the results of the tests for slope and intercept in a table of the form below.

Predictor	Estimated Coefficent	Standard Error	t	P-value
Intercept	b_0	S_{b_0}	$t = b_0/S_{b_0}$	p
X	b_1	S_{b_1}	$t = b_1/S_{b_1}$	p

Example 12.19: t Test for a Slope

A scatterplot of the human development index (HDI) values versus urbanization rates for the $n = 40$ sub-Saharan countries (Example 12.6), with fitted regression line, is below.

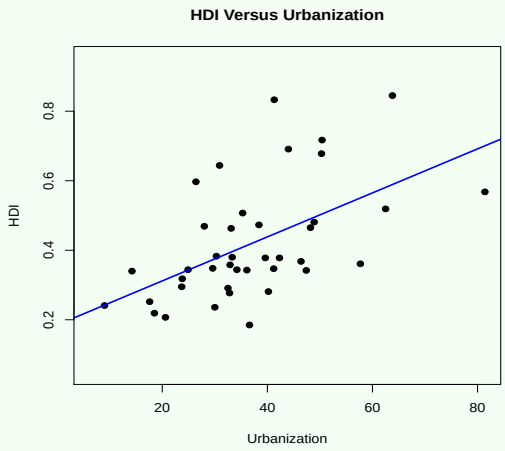


Figure 12.17: Scatterplot of human development index (HDI) versus urbanization.

We want to decide if the observed increase in HDI value with urbanization is statistically significant. The null hypothesis

$$H_0 : \beta_1 = 0$$

says that a country's HDI value isn't related to its urbanization rate, and the alternative hypothesis (tested by statistical software)

$$H_a : \beta_1 \neq 0$$

says it is. The t test results (obtained using software) are below.

Predictor	Estimated Coefficent	Standard Error	t	P-value
Intercept	0.1852	0.0629	2.942	0.0055
Urbanization	0.0063	0.0016	3.979	0.0003

Also shown are the results of the t test for the intercept.

The observed test statistic value for the test for the slope is $t = 3.979$ and the p-value, from a t distribution with $n - 2 = 38$ degrees of freedom, is 0.0003. Thus we reject H_0 and conclude that the observed linear relationship between HDI value and urbanization is statistically significant.

Equivalence of the t Test for a Slope and the t Test for a Correlation

It can be shown that the t statistic (12.32) for the test for the slope β_1 is the same as the t statistic (12.4) for the test for the population correlation ρ, and since both are compared to the $t(n-2)$ distribution, the p-values for the two tests will be the same too. As an example, compare the results of the t test for the slope in Example 12.19 to those of the t test for a correlation in Example 12.6.

12.5.12 t Confidence Intervals for the Slope and Intercept of the Regression Model

A **$100(1-\alpha)\%$ *confidence interval for* β_1**, the true (unknown) slope parameter of the regression model, is given by the following.

Confidence Interval for a Slope:

$$b_1 \pm t_{\alpha/2,n-2} S_{b_1},$$

where b_1 is the least squares estimate of β_1, $t_{\alpha/2,n-2}$ is the $100(1-\alpha/2)$th percentile of the t distribution with $n-2$ degrees of freedom, and S_{b_1} is the estimated standard error (12.31) of b_1.

We can be $100(1-\alpha)\%$ confident that the true (unknown) slope β_1 will be contained in this confidence interval somewhere.

Example 12.20: Confidence Interval for the Slope β_1

For the data on human development index (HDI) values and urbanization rates for $n = 40$ sub-Saharan countries (Example 12.6), the regression analysis of Example 12.19 produced

$$b_1 = 0.0063 \qquad \text{and} \qquad S_{b_1} = 0.0016.$$

For a 95% confidence interval for β_1, the critical value (from a t distribution table using $n-2 = 38$ degrees of freedom) is $t_{\alpha/2,n-2} = 2.024$. Thus the confidence interval is

$$\begin{aligned} 0.0063 \pm 2.024\,(0.0016) &= 0.0063 \pm 0.0032 \\ &= (0.0031,\ 0.0095). \end{aligned}$$

We can be 95% confident that the true slope β_1 is somewhere in this range. In other words, we can be 95% confident that each one-percent increase in a country's urbanization rate results in an HDI increase of between 0.0031 and 0.0095.

Although generally not of much interest, we can also compute a ***$100(1-\alpha)\%$ confidence interval for β_0***, the true (unknown) intercept parameter of the regression model.

Confidence Interval for an Intercept:

$$b_0 \pm t_{\alpha/2,n-2} S_{b_0},$$

where b_0 is the least squares estimate of β_0, $t_{\alpha/2,n-2}$ is the $100(1-\alpha/2)$th percentile of the t distribution with $n-2$ degrees of freedom, and S_{b_0} is the estimated standard error (12.36) of b_0.

We can be $100(1-\alpha)\%$ confident that the true (unknown) intercept β_0 will be contained in this interval.

12.5.13 Regression Model *F* Test

Another way to test for the slope parameter β_1 is to perform the so-called called ***regression model F test***. The null and alternative hypotheses are exactly the same as for the t test, namely

$$H_0 : \beta_1 = 0 \tag{12.37}$$
$$H_a : \beta_1 \neq 0 \tag{12.38}$$

The ***regression model F test statistic*** is

***F* Test Statistic for the Regression Model**:

$$F = \frac{\text{MSR}}{\text{MSE}}. \tag{12.39}$$

The numerator measures variation in Y due to the X variable, and will be large when the fitted regression line has a steep slope (either positive or negative). The denominator measures variation due to random error. Thus F will be large when the variation in Y due to X is large relative to the variation due to random error. It follows that

Large values of F provide evidence *against* H_0 in favor of H_a.

To decide if an observed value of F is large enough to provide statistically significant evidence against the null hypothesis, we'll need its sampling distribution under H_0.

Sampling Distribution of F Under H_0: Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are bivariate observations described by the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where the ϵ_i's are independent and follow a $\text{N}(0, \sigma)$ distribution. Then when

$$H_0 : \beta_1 = 0$$

is true,

$$F \sim F(1, n-2)\,,$$

the F distribution with numerator degrees of freedom 1 and denominator degrees of freedom $n-2$.

Because *large* values of F provide evidence against the null hypothesis, p-values (and critical values for the rejection region approach) are obtained from the *right* tail of the $F(1, n-2)$ distribution.

The F test procedure is summarized below.

Regression Model F Test for β_1

Assumptions: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are bivariate observations described by the simple linear regression model (12.15), where the ϵ_i's are independent and either they follow a $\text{N}(0, \sigma)$ distribution or n is large ($n \geq 20$).

Null hypothesis: $H_0 : \beta_1 = 0$

Test statistic value: $F = \dfrac{\text{MSR}}{\text{MSE}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or F is in rejection region.

Alternative hypothesis	P-value = area under F-distribution with 1 and $n-2$ d.f.:	Rejection region = F values such that:*
$H_a : \beta_1 \neq 0$	to the right of F	$F \geq F_{\alpha,1,n-2}$

* $F_{\alpha,1,n-2}$ is the $100(1-\alpha)$th percentile of the F distribution with 1 and $n-2$ d.f.

Most statistical software packages will report the results of this F test along with those of the t tests of Section 12.5.11. The F test result will be equivalent to the result of the t test for the slope, according to the following fact, so it's somewhat redundant.

Fact 12.9 It can be shown that the F test statistic (12.39) is equal to the square of the t test statistic (12.32), that is,

$$F = t^2,$$

and that the p-value for the F test will be the same as that of the two-sided t test for β_1.

12.5.14 The Regression ANOVA Table

The degrees of freedom, sums of squares, mean squares, observed F test statistic value, and p-value from a regression analysis are usually summarized in a ***regression ANOVA table*** having the form shown below.

Source	DF	SS	MS	F	P-value
Regression	1	SSR	MSR = SSR/1	F = MSR/MSE	p
Error	$n-2$	SSE	MSE = SSE/$(n-2)$		
Total	$n-1$	SSTo			

Example 12.21: Regression ANOVA Table and Model *F* Test

For the data on the human development index (HDI) values and urbanization rates for the $n = 40$ sub-Saharan countries (Example 12.6), the regression ANOVA table, produced by statistical software, is shown below.

Source	DF	SS	MS	F	P-value
Regression	1	0.320	0.320	15.83	0.0003
Error	38	0.768	0.020		
Total	39	1.088			

From the regression ANOVA table, the test statistic for the model F test of

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$

is $F = 15.83$, which is also the square of the t test statistic in Example 12.19, and the p-value is 0.0003, indicating a statistically significant relationship between the HDI and urbanization rate.

If we calculate the R^2 value directly, we get

$$R^2 = \frac{\text{SSR}}{\text{SSTo}} = \frac{0.320}{1.088} = 0.294.$$

Thus 29.4% of the variation in HDI values can be explained by differences in urbanization rates.

Finally, from the table, the mean squared error is MSE = 0.020, so its square root, $\sqrt{\text{MSE}} = 0.141$, corresponds to the size of a typical deviation of an HDI value above or below the fitted regression line in Fig. 12.17, and is also our estimate of σ in the N(0, σ) error distribution in the regression model.

12.5.15 Using Residuals to Check the t and F Test Assumptions

The t tests for the regression model slope and intercept and the F test for the slope rely on three assumptions:

1. The errors ϵ_i in the regression model follow a normal distribution.
2. The the standard deviation σ of the error distribution doesn't change with the value of the predictor variable.
3. The responses Y_i are independent of each other, or equivalently, the errors ϵ_i are independent.

The third assumption (independence) is usually addressed in the study design by separating observations sufficiently in space and time. The other assumptions (normality and common σ) are checked via plots of the residuals.

Checking the Normality Assumption

To check the normality assumption, we look at a normal probability plot or a histogram of the residuals.

Example 12.22: Checking Assumptions

For the data on human development index (HDI) values and urbanization rates for the $n = 40$ sub-Saharan countries (Example 12.6), the residuals after fitting the regression model (with HDI as the response) are plotted below in a normal probability plot and a histogram.

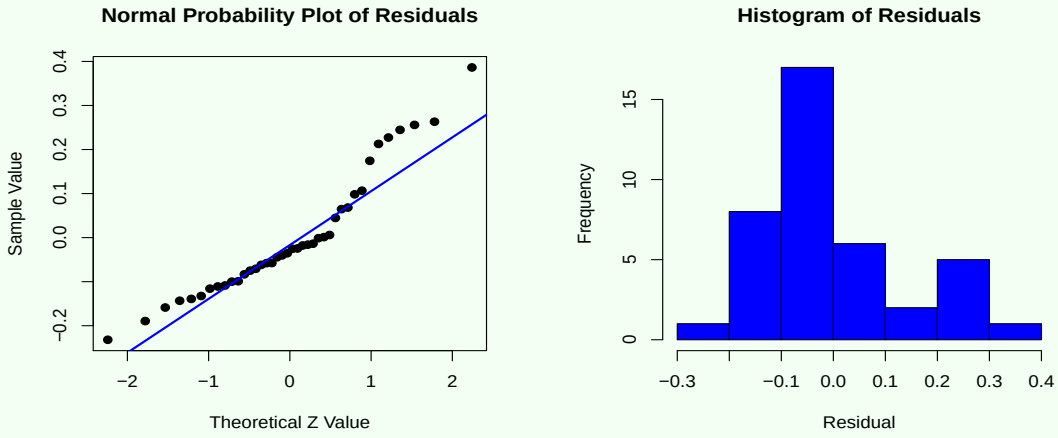


Figure 12.18: Normal probability plot (left) and histogram (right) of the residuals from the regression analysis of HDI values and urbanization rates for sub-Saharan countries.

The plots suggest that the assumption of normality of the error term ϵ in the linear regression model is approximately met. The slight hint of right-skewness isn't a concern because the sample size is fairly large.

Checking the Constant σ Assumption

There are a few ways to check the assumption that the error standard deviation σ doesn't change with the value of X

1. **Plot the residuals versus the predictor variable X**: We can look at a plot of the residuals versus the values X_i of the predictor variable, with a horizontal line at $y = 0$. The amount of vertical spread above and below the line should be roughly the same from left to right, and in particular, it shouldn't increase (or decrease) as X increases.
2. **Plot the residuals versus the fitted values**: We can look at a plot of the residuals versus the fitted values, with a horizontal line at $y = 0$. Because the fitted values are a linear function of the X_i's, this is equivalent to plotting the residuals versus the X_i's, except the units on the horizontal scale will be different. The amount of vertical spread above and below the line should be roughly the same from left to right, and in particular, it shouldn't increase with the fitted value.

Example 12.23: Checking Assumptions

For the data on human development index (HDI) values and urbanization rates for sub-Saharan countries, a plot of the residuals versus fitted values is below.

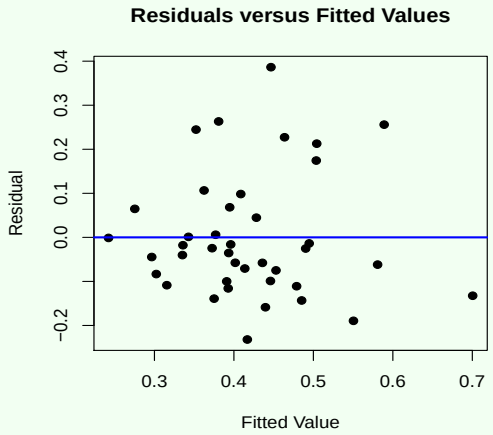


Figure 12.19: Plot of the residuals versus the fitted values from the regression analysis of HDI values and urbanization rates for sub-Saharan countries.

The sizes of the vertical deviations of the points above and below the horizontal line remain roughly the same size as we move from left to right in the plot, so the assumption that the standard deviation σ is constant (doesn't depend on the urbanization rate) appears to be met.

12.6 Dealing With Non-Linear Relationships: Transformations and Non-Linear Regression

When the relationship between the response Y and the predictor X is curved, a linear regression analysis isn't appropriate. Here are a few possible courses of action:

1. **Transform the data to linearity**: Often we can "straighten out" a curved pattern by making a transformation of the Y observations or the X observations, for example taking their logs or using some other transformation in the Ladder of Powers.
2. **Fit a curve to the data**: Another option is to fit a curve to the data, such as a polynomial or some other curved function, instead of a straight line. Methods for fitting polynomials and other curves are discussed in Chapter 13.

Example 12.24: Log Transformation to Linearity

For the data from the study of radiocesium (^{137}Cs) and rubidium (Rb) in mushrooms (Example 12.8), the left scatterplot below shows a curved relationship between these two variables. To "straighten out" the relationship, we can take the log of the values of the response variable (^{137}Cs). The right scatterplot shows the log of ^{137}Cs versus Rb. A linear regression analysis or correlation analysis could now be performed using log ^{137}Cs and Rb.

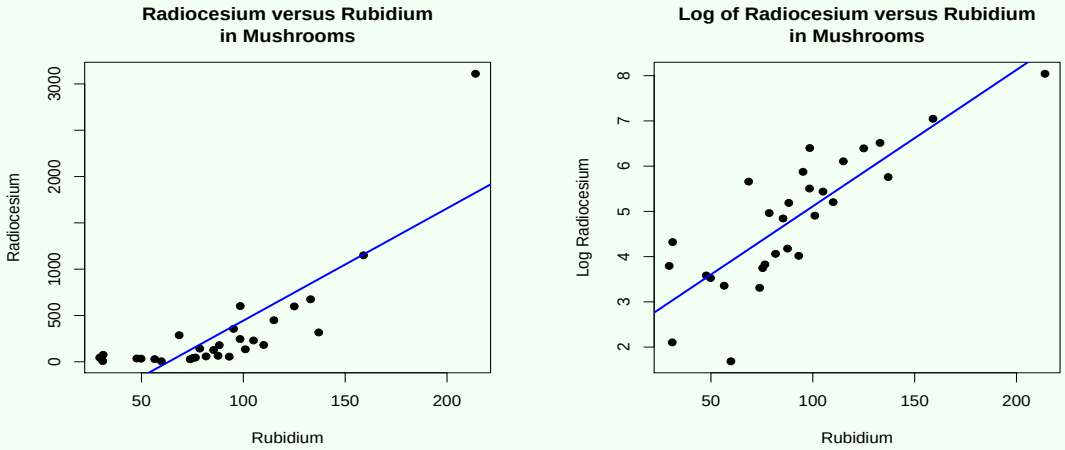


Figure 12.20: Scatterplot of radiocesium ^{137}Cs versus rubidium in mushrooms from a Japanese forest with fitted regression line (left). The scatterplot of the natural logs of the ^{137}Cs measurements versus rubidium with fitted regression line (right).

12.7 Dealing With a Non-Constant Standard Deviation: Transformations

When the constant standard deviation assumption isn't met, that is, when σ changes with X, it's sometimes possible to stabilize it by transforming the Y observations. The most commonly used transformation for this purpose is to take the logs of the Y_i's, but other transformations in the Ladder of Powers might also be used. Often a standard deviation that increases with the value of X is accompanied by an upward bending curved pattern, and taking logs of the Y_i's both "straightens out" the curved pattern *and* stabilizes the standard deviation.

12.8 Problems

12.1 For each of $n = 24$ rainfall events, cadmium (Cd, μg/L) was measured in stormwater runoff and the data used in a regression analysis, with Cd as the response and rain depth (cm) as the predictor. The coefficient estimates, their standard errors, and partial results of the t tests are shown below.

Predictor	Estimated Coefficent	Standard Error	t	P-value
Constant	2.329	0.645	3.608	0.000
Rain Depth	-0.064	0.044	?	?

Fill in the values that are missing from the table.

12.2 Refer to Problem 12.1 showing the results of the regression analysis with Cd as the response and rain depth (cm) as the predictor.

Compute a 95% confidence interval for the true (unknown) amount β_1 by which the Cd increases, on average, for each one-cm increase in rain depth.

12.3 Mercury (Hg) has been used in a variety of household and industrial products including thermometers, appliance switches, fluorescent lights, alkaline batteries, and latex paint. But Hg is also toxic, and its release into the environment can contaminate the food chain. It is important, therefore, to have efficient methods for detecting and measuring Hg in the environment.

A study was carried out to assess the performance of a new, faster method for measuring Hg in soil and plants [5]. Soil specimens were treated with known concentrations of Hg and placed in plastic pots. Chinese brake ferns (*Pteris Vittata*) were then transplanted into the pots allowed to grow for 23 days. The soil and plants were then analyzed using the both new method, inductively coupled plasma atomic emission spectrometry, and the older, more established method, cold vapor atomic absorption spectrometry.

The table below shows measured Hg concentrations (mg/L) using both methods on the same soil and plant specimens.

Hg Measurements in Soils

New Method	Established Method
0.30	0.33
0.77	0.90
2.41	2.56
2.60	3.04
2.66	2.88
3.38	3.85
3.89	4.19
3.90	4.31
3.91	4.29
4.69	5.24
6.02	6.50
7.16	8.16
33.9	35.7
40.9	44.6
62.0	68.5
62.8	67.1

Hg Measurements in Plants

New Method	Established Method
0.14	0.13
0.16	0.15
0.69	0.71
0.72	0.66
0.73	0.79
0.73	0.83
0.74	0.72
0.77	0.78
0.89	0.94
0.89	0.99
0.91	0.89
0.94	0.96
0.97	1.05
1.09	1.19
1.11	1.09
1.25	1.16
1.29	1.41
1.37	1.49
1.89	2.04
3.17	3.51
3.88	3.83
6.44	6.62
7.73	7.84

The researchers stated that the new and established methods were in agreement, and that the new method is reliable for measuring Hg in soils and plants. We'll confirm this via scatterplots and the sample correlations.

a) Make a scatterplot of the soil Hg measurements, with the established method on the x-axis and new method on the y-axis.

b) Compute the sample correlation between new method and the established method soil Hg measurements.

c) Make a scatterplot of the plant Hg measurements, with the established method on the x-axis and new method on the y-axis.

d) Compute the sample correlation between new method and the established method plant Hg measurements.

12.4 Ships are a significant contributor to worldwide air pollution emissions, largely from their main propulsion engines, but also from auxiliary engines that generate electrical power for ship services.

The table below shows representative emission values (g/kwh) of nitrogen oxide (NOx), sulphur dioxide (SO2), carbon dioxide (CO2), hydrocarbons (HC), and particulate matter (PM) for 14 ship types [12].

Emissions from Ships

Ship Type	NOx	SO_2	CO_2	HC	PM	SFC
Liquefied gas	8.8	12.4	816	0.31	1.03	257
Chemical	16.3	11.0	650	0.55	1.34	204
Oil	14.8	11.7	690	0.50	1.43	217
Other liquids	16.3	11.0	649	0.55	1.30	204
Bulk dry	17.7	10.6	627	0.59	1.61	197
General cargo	16.2	10.9	649	0.54	1.28	204
Container	17.3	10.8	635	0.57	1.56	200
Refrigerated cargo	17.1	10.8	636	0.57	1.47	200
Ro-Ro cargo	15.3	11.1	655	0.52	1.17	206
Passenger/Ro-Ro cargo	13.3	9.9	688	0.42	0.73	217
Passenger	13.2	11.8	697	0.46	0.81	219
Offshore supply	13.9	11.0	677	0.49	0.79	213
Research	14.1	11.5	675	0.48	0.85	212
Towing/Pushing	13.7	10.8	674	0.42	0.80	212

a) Make a scatterplot of PM (y-axis) versus NOx (x-axis).

b) Compute the sample correlation between PM and NOx.

c) Notice the outlier in the scatterplot of part a. Which ship type is the outlier?

d) Remove the outlier, recompute the correlation (with the outlier excluded), and compare its value to the one in part b.

12.5 Refer to the ship emissions study and data described in Problem 12.4.

a) Find the equation of the least squares regression line with PM as the response (Y) and NOx as the predictor (X).

b) Make a scatterplot of PM versus NOx with the regression line included in the plot.

c) Notice the outlier in the scatterplot. Remove the outlier, recompute the equation of the regression line (with the outlier excluded), and compare the slope b_1 to the one from part a.

12.6 A study was carried out to assess the effects of noise disturbance from aircraft on birds during the nesting season in the Colville River Delta, Alaska [9]. The table below shows the counts of landings or takeoffs for several types of aircraft at the region's airstrip for 45 days in 2001.

Airplane Landings and Takeoffs

Date	DC-6	CASA	Twin Otter/ Navajo/Beech	Small Planes
1 June	2	6	6	0
2 June	0	2	0	0
3 June	0	4	1	2
4 June	2	12	18	4
5 June	4	8	8	0
6 June	2	8	6	0
7 June	2	8	4	0
8 June	0	8	6	4
9 June	2	8	0	0
10 June	2	8	0	2
11 June	0	4	14	6
12 June	4	10	6	0
13 June	6	6	10	2
14 June	4	8	4	2
15 June	8	4	8	0
16 June	2	4	0	0
17 June	0	4	0	0
18 June	0	8	12	0
19 June	4	4	10	0
20 June	2	8	6	0
21 June	4	8	4	0
22 June	0	8	12	0
23 June	4	2	0	0
24 June	0	2	4	2
25 June	0	12	14	0
26 June	4	8	8	0
27 June	2	6	6	0
28 June	0	8	6	0
29 June	4	8	4	0
30 June	0	4	0	0
1 July	0	6	0	0
2 July	0	12	12	0
3 July	0	10	6	0
4 July	6	2	12	0
5 July	0	6	4	0
6 July	2	6	6	0
7 July	0	4	0	0
8 July	0	4	0	0
9 July	0	10	16	2
10 July	2	10	14	0
11 July	4	6	6	0
12 July	2	10	4	0
13 July	4	8	12	0
14 July	0	4	10	0
15 July	0	4	6	0

a) Calculate the correlation between DC-6 and Twin Otter/Navajo/Beech landings or takeoffs, and carry out a t test to decide if it's statistically significantly different from zero.

b) Calculate the correlation between CASA and Twin Otter/Navajo/Beech landings or takeoffs, and carry out a t test to decide if it's statistically significantly different from zero.

c) Calculate the correlation between Small Planes and Twin Otter/Navajo/Beech landings or takeoffs, and carry out a t test to decide if it's statistically significantly different from zero.

12.7 In the study to find out if concentrations of stable elements could be used to predict the concentrations of radioactive elements, described in Example 12.8, several elements were measured on eight plant species in the forest.

The data below are measurements of cesium (Cs), a proxy for ^{137}Cs, strontium (Sr), a proxy for radioactive ^{90}Sr, and the stable elements sodium (Na), potassium (K), and calcium (Ca) (all in mg/kg dry weight).

Radioactivity in Plants

Plant	Na	K	Cs	Ca	Sr
1	93	6900	0.024	2570	13.7
2	331	4250	0.017	6440	27.4
3	6040	15300	0.048	29800	108
4	713	8890	0.100	58300	151
5	1860	8830	0.037	25900	99.3
6	900	24500	0.112	28700	113
7	892	3960	0.022	5820	38.8
8	670	17900	0.052	6700	42.3

a) The authors of the study suggest that, in part because of a strong observed relationship between K and Cs, K could be used as an indicator of ^{137}Cs in plants. Calculate the correlation between Cs and K, and comment about whether it supports the authors' statement.

b) The authors of the study also suggest that, in part because of a strong observed relationship between Ca and Sr, Ca could be used as an indicator of radioactive ^{90}Sr in plants. Calculate the correlation between Ca and Sr, and comment about whether it supports the authors' statement.

c) The authors of the study state that there's very little relationship between Na and Cs. Calculate the correlation between Na and Cs, and comment about whether it supports the authors' statement.

12.8 Various numerical indices of ecological quality are used to assess the impact of anthropogenic environmental pressures such as pollution. In a comparative study of several such indices, samples of benthic (bottom dwelling) communities were collected using mesh sieves at 14 stations in the Aegean Sea on the coast of Greece, and the values the indices were determined for each sample [18].

The table below shows the values of six indices: The Bentix biotic index, the AMBI biotic index, Shannon's diversity index (H'), the species richness index (S), Pielou's evenness index (J), and the density of individuals per square meter (N/m^2). Also shown are the station depths.

Station	BENTIX	AMBI	H'	S	J	N/m^2
E3	3.46	1.81	4.95	42	0.83	1625
E5	3.76	2.01	5.69	86	0.82	3980
E8	4.04	1.63	5.61	91	0.80	4365
E10	3.95	1.68	5.83	68	0.87	2165
E11	5.13	1.37	6.06	79	0.88	3010
E13	4.61	1.87	6.19	97	0.87	3590
E16	3.98	2.15	5.75	83	0.84	4255
E20	3.60	2.38	5.32	67	0.81	3265
E24	4.10	2.12	5.69	65	0.87	2000
E25	4.13	1.79	5.85	67	0.88	2020
E28	3.36	1.95	5.84	87	0.83	4540
E29	3.97	1.95	5.62	55	0.88	1580
E30	4.08	1.86	6.21	77	0.91	2485
E31	4.21	1.98	4.39	21	0.88	535
DA3	4.22	2.13	4.75	95	0.72	8853
TP2	3.05	2.25	4.68	79	0.77	3873
TP13	3.17	2.59	5.85	138	0.86	5187
TP6	3.50	2.24	5.47	128	0.81	10247
TP7	3.01	3.30	4.68	81	0.77	3893
TP10	2.99	2.25	5.29	83	0.87	2933

The researchers were primarily interested in validating the newer Bentix index as an indicator of benthic ecological quality by showing that its value is related to those of the other, generally more established indices.

a) Make scatterplots of the Bentix index (y axis) versus each of the other indices (AMBI, H', S, J, and N/m^2) separately.

b) Based on the scatterplots of part *a*, with which of the other indices (AMBI, H', S, J, or N/m^2) is Bentix most strongly related (either positively or negatively)? With which does it appear to be least strongly related?

c) Compute the correlation between the Bentix index and each of the other indices (AMBI, H', S, J, and N/m^2) separately.

d) Based on the correlations of part *c*, with which of the other indices (AMBI, H', S, J, or N/m^2) is Bentix most strongly related (either positively or negatively)? With which is its relationship weakest?

12.9 Forests are potential sinks for atmospheric carbon dioxide, a greenhouse gas. Effective forest management practices can influence carbon sequestration rates, but they require estimating the carbon stocks in the forest from the biomass of its trees.

For live trees, biomass is approximated from stem diameter measurements [2]. For dead trees, with no leaves and with missing branches, the same approximation procedure can be used, but the amount of missing biomass in leaves and branches must be subtracted from the approximation. This requires knowing how much of a tree's total biomass is contained in its leaves and branches.

The table below shows the (estimated) percentage of above ground biomass that's contained in stems, branches, and leaves of hardwood and softwood trees for varying stem diameters (at breast height, DBH) [2], [8].

Hardwood				**Softwood**			
DBH	Stem	Branches	Leaves	DBH	Stem	Branches	Leaves
10	54	43	3	10	68	23	8
20	68	29	2	20	74	19	6
30	74	24	2	30	77	17	6
40	77	21	2	40	78	16	6
50	79	19	2	50	78	16	6
60	80	18	2	60	79	16	6
70	81	17	2	70	79	15	6
80	82	16	2	80	79	15	5
90	82	16	2	90	80	15	5
100	83	15	2	100	80	15	5

In this problem, we'll analyze the relationship between stem biomass and DBH in hardwood trees.

a) Make a scatterplot of stem biomass (y axis) versus diameter (x axis) for hardwood trees. Is the relationship linear or curved? Is it monotone?

b) Compute the Pearson correlation r and the Spearman rank correlation r_{sr}. Which is closer to 1.0? Why do you suppose this is the case?

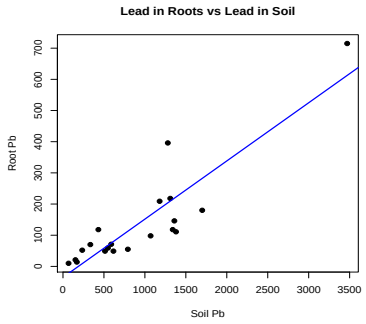
12.10 Lead is a toxic element that doesn't biodegrade or decay. Lead contamination in soil can be particularly high in the vicinity of houses whose exteriors have been painted with lead-containing paint. One concern is that this may contaminate garden vegetables.

The table below shows the lead (Pb) concentrations in garden soils (ppm) and roots of fruiting vegetables (μg/g) for a sample of 20 older (pre-1900) homes with gardens in Chicago, Illinois [4].

Lead in Soil and Roots of Fruiting Vegetables

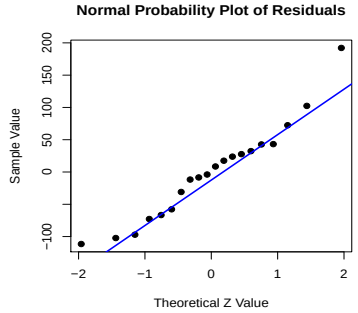
Home	Pb in Soil	Pb in Roots
1	1180	209
2	616	49
3	589	71
4	1360	146
5	1310	218
6	1340	118
7	549	59
8	1070	98
9	792	55
10	1280	396
11	68	10
12	152	21
13	1700	180
14	513	49
15	3470	715
16	1380	111
17	334	70
18	169	140
19	432	118
20	235	52

A plot of the data is below.

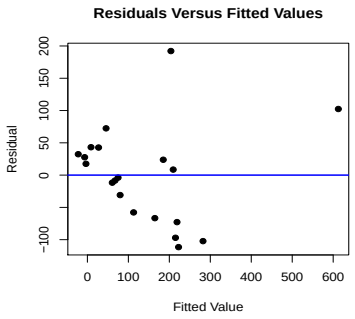


a) Carry out a linear regression analysis, with root lead concentration as the response and soil concentration as the predictor, to determine whether root concentrations are elevated when soil concentrations are higher. Give the equation of the fitted regression line and the results of the t test for the slope, and state the conclusion of the test.

b) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the t test, that the error term ϵ is normally distributed appear to be met?



c) A plot of the residuals versus the fitted values is below.

Based on the plot, does the assumption, required by the t test, that the standard deviation σ of the error distribution is the same for different soil lead concentrations appear to be met?

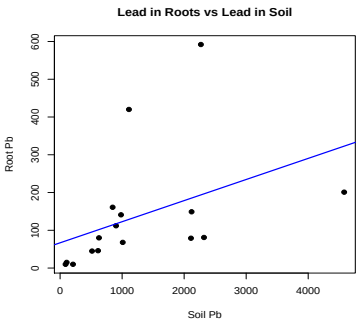
d) Assuming that σ is the same for the different soil lead concentrations, what's the estimated value of σ?

12.11 The table below shows the lead (Pb) concentrations in garden soils (ppm) and roots of leafy vegetables and herbs (μg/g) for a sample of 16 older (pre-1900) homes with gardens in Chicago, Illinois from the study described in Problem 12.10.

Lead in Soil and Roots of Leafy Vegetables

Home	Pb in Soil	Pb in Roots
1	612	46
2	208	10
3	515	45
4	2110	79
5	4580	201
6	982	141
7	1110	420
8	2120	149
9	847	161
10	2270	592
11	88	10
12	1010	68
13	2320	81
14	627	80
15	902	112
16	106	15

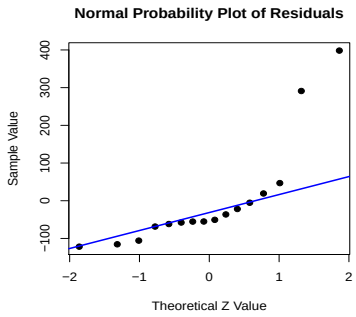
A plot of the data is below.



a) Carry out a linear regression analysis, with root lead concentration as the response and soil concentration as the predictor, to determine whether root concentrations are elevated when soil concentrations

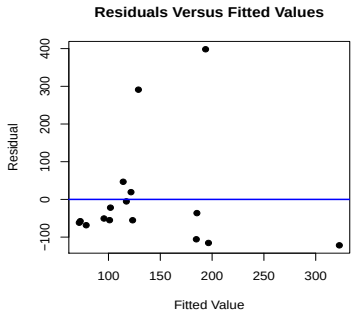
are higher. Give the equation of the fitted regression line and the results of the t test for the slope, and state the conclusion of the test.

b) A normal probability plot of the residuals is below.



Based on the plot, does the assumption, required by the t test, that the error term ϵ is normally distributed appear to be met?

c) A plot of the residuals versus the fitted values is below.



Based on the plot, does the assumption, required by the t test, that the standard deviation σ of the error distribution is the same for different soil lead concentrations appear to be met?

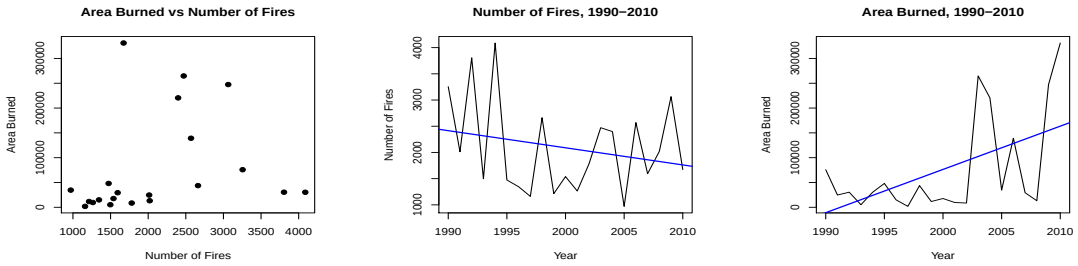
d) Assuming that σ is the same for the different soil lead concentrations, what's the estimated value of σ?

12.12 The Canadian Forest Service developed and maintains its National Forestry Database (NFD) to inform government agencies and the general public about forest management practices and forest resources. The NFD contains data on the total number of forest fires in Canada each year and the total area burned by those fires. The table below shows these data for the years 1990 - 2010 [17].

Forest Fires in Canada

Year	Number of Forest Fires	Area Burned (Thousands of Ha)
1990	3255	75.783
1991	2013	24.708
1992	3805	30.452
1993	1497	5.183
1994	4088	30.308
1995	1474	48.080
1996	1346	14.952
1997	1161	1.876
1998	2662	43.681
1999	1214	11.666
2000	1539	17.675
2001	1264	9.668
2002	1781	8.586
2003	2472	264.736
2004	2398	220.516
2005	971	34.664
2006	2569	139.201
2007	1594	29.416
2008	2020	13.211
2009	3064	247.419
2010	1673	331.108

Plots of the data are below.



a) Is there a correlation between the number of fires and the area burned? Calculate the correlation and carry out a t test to decide if the correlation is statistically significantly different from zero.

b) Carry out a linear regression analysis, with number of fires as the response and year as the predictor, to determine whether there's been a statistically significant trend in the number of fires over the years 1990-2010. Give the equation of the fitted regression line and the results of the t test for the slope, and state the conclusion of the test.

c) Carry out a linear regression analysis, with area burned as the response and year as the predictor, to determine whether there's been a statistically significant trend in the area burned over the years 1990-2010. Give the equation of the fitted regression line and the results of the t test for the slope, and state the conclusion of the test.

d) Compute the coefficient of determination R^2 between the area burned and year. What proportion of the variation in area burned can be explained by the time trend?

12.13 Yellowstone Lake is located in the southeastern part of Yellowstone National Park and covers an area about 136 mi^2 (352 km^2) depending on the level of water in the lake.

The water level in Yellowstone Lake varies from year to year in response to differences in the winter's snowpack accumulation, spring precipitation, and air temperatures. Restriction at the outlet of the lake retards the outflow, and water backs up in the lake during periods of high inflow. The U.S. Geological Survey started publishing Yellowstone Lake elevations in 1922 and outflows in 1926.

The table shows the maximum daily outflow (cubic feet per second) and maximum daily elevation (feet, as measured on Bridge Bay staff gage) for each of the years 1926 - 2001. [3].

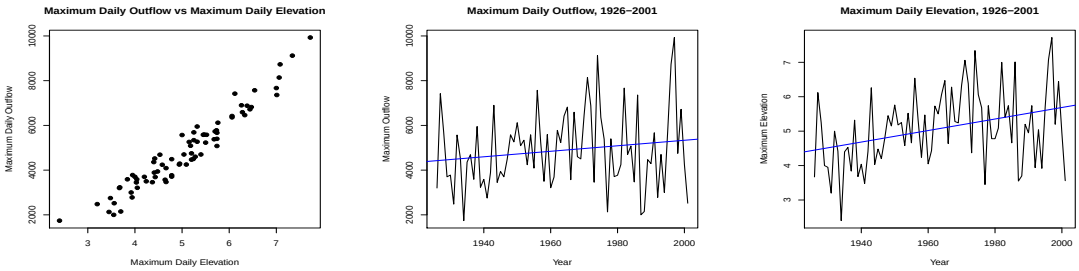
Yellowstone Lake, 1926-2001

Year	Maximum Daily Output	Maximum Daily Elevation
1926	3200	3.67
1927	7420	6.12
1928	5689	5.25
1929	3700	4.00
1930	3780	3.95
1931	2480	3.20
1932	5570	5.00
1933	4520	4.42
1934	1740	2.40
1935	4360	4.40
1936	4690	4.53
1937	3590	3.84
1938	5950	5.32
1939	3230	3.68
1940	3590	4.04
1941	2750	3.48
1942	3890	4.41
1943	6900	6.26
1944	3450	4.03
1945	3940	4.48
1946	3700	4.20
1947	4490	4.78
1948	5580	5.45
1949	5260	5.15
1950	6120	5.76
1951	5090	5.18
1952	5340	5.25
1953	4240	4.58
1954	5580	5.52
1955	4090	4.66
1956	7570	6.54
1957	5270	5.32
1958	3500	4.24
1959	5590	5.47
1960	3210	4.05
1961	3690	4.43
1962	5780	5.73
1963	5230	5.50
1964	6420	6.06
1965	6820	6.47
1966	3570	4.64
1967	6590	6.28
1968	4600	5.28
1969	4500	5.24
1970	6460	6.33
1971	8140	7.06
1972	6880	6.38
1973	3460	4.37
1974	9120	7.34
1975	6360	6.06
1976	5380	5.68
1977	2130	3.45
1978	5400	5.74
1979	3710	4.78
1980	3770	4.78
1981	4250	5.09
1982	7670	7.00
1983	4700	5.40
1984	5080	5.74
1985	3470	4.66
1986	7360	7.01
1987	2000	3.55
1988	2150	3.70
1989	4470	5.20
1990	4290	4.95
1991	5670	5.74
1992	2780	3.94
1993	4700	5.04
1994	3000	3.92
1995	5730	5.70
1996	8730	7.08
1997	9930	7.72
1998	4750	5.20
1999	6720	6.44
2000	4250	4.94
2001	2520	3.56

Plots of the data are below.

a) Is there a correlation between the maximum daily outflow and maximum daily elevation? Calculate the correlation and carry out a t test to decide if the correlation is statistically significantly different from zero.

b) Carry out a linear regression analysis, with maximum daily outflow as the response and year as the predictor, to determine whether there's been a statistically significant trend in the maximum daily outflow over the years 1926-2001. Give the equation of the fitted regression line and the results of the t test for the slope, and state the conclusion of the test.



c) Carry out a linear regression analysis, with maximum daily elevation as the response and year as the predictor, to determine whether there's been a statistically significant trend in the maximum daily elevation over the years 1926-2001. Give the equation of the fitted regression line and the results of the t test for the slope, and state the conclusion of the test.

d) Compute the coefficient of determination R^2 between the maximum daily elevation and year. What proportion of the variation in maximum daily elevation can be explained by the time trend?

12.14 In a study of the increase in solid waste resulting from rapid urban population growth in the Port Harcort metropolitan area of Nigeria, the size (number of residents) and annual refuse generation (metric tons per year) per household was determined for a sample of 46 households in the area [14]. A two-stage sampling scheme was used. In the first stage, 46 streets were randomly selected, and in the second, a single household was randomly selected from each of the 46 streets. The data are below.

Household Refuse in Nigeria

Address of House	Household Size	Annual Refuse
Ikwerre Road/465	12	5.6
Okpowu-Obasi/010	7	2.7
Worlu Street/013	5	2.0
Omachi Street/027	8	4.1
Eligbolo Road/266	11	5.3
Rumuagholu Road/159	9	4.5
Ovunwo Street/005	7	3.0
Nwachukwu Street/019	10	4.9
Worlu Eguma Street/015	8	3.0
David Ejekwu Street/020	6	2.8
Chinda Street/033	9	4.3
Rumuomoi/Orosi Road/115	11	5.1
Owhor Street/032	7	2.8
Obi Wali Road/108	6	2.0
Kesiolu Street/016	9	4.1
Mgbouba-Choba Road/305	7	2.5
Ehio Street/008	5	2.0
Ogbogoro Road/099	12	5.8
Kala Street/017	10	5.2
Ebara Street/049	9	5.0
Orazi Road/051	7	2.3
Eligbam Road/063	8	2.9
Rumuola Road/253	10	4.7
Mbarajah Street/021	12	5.5
PHC-Aba Express Road/378	9	4.0
Arochukwu Street/092	11	5.6
Uyo Street/036	8	2.7
Market Road/119	10	5.3
Bende Street/101	9	4.2
Geodetic Street/052	7	2.4
Wopara Street/014	6	2.1
Ekere Street/018	5	2.2
Enugu Street/003	8	2.5
Woji Road/367	10	5.1
Obadiah Street/030	9	3.9
Elitor Street/007	11	5.0
Ihunwo Street/012	8	3.8
Peace Crescent/009	6	2.6
Unity Avenue/001	5	2.1
Rumuibekwe Road/118	10	5.0
Old Aba Road/201	12	5.7
Okporo Rumuodara Road/204	9	4.0
Elelenwo Road/318	11	5.2
Rumuokwurusi-Igwuruta Road/476	10	4.8
Oroigwe Road/086	8	3.6
Eneka-Rukpokwu Road/011	6	2.4

a) Carry out a linear regression analysis, with annual refuse as the response and household size as the predictor, to estimate how much additional refuse is generated for each additional household resident. Give the equation of the fitted regression line and state the estimated amount by which the annual refuse increases per additional resident.

b) Compute a 95% confidence interval for the true (unknown) amount of additional refuse that's generated for each additional resident in a household.

12.15 Each solid waste disposal facility in California that's required to have a permit must pay a fee for each ton of nonhazardous solid waste landfilled at the facility. The table below shows amounts of waste (in thousands of tons) subject to the fee, as reported by the disposal facilities in several counties.

Landfilled Waste

Year	Inyo County	Los Angeles County	Monterey County	San Mateo County	Santa Barbara County
1990	15	12879	458	850	474
1991	22	11848	415	860	452
1992	20	12088	409	876	424
1993	9	11519	415	879	407
1994	12	12344	431	843	397
1995	12	11613	444	849	431
1996	8	10649	436	915	438
1997	17	9563	453	874	434
1998	15	10082	481	954	448
1999	12	10313	458	915	432
2000	16	10408	458	1002	414
2001	17	10134	439	953	392
2002	19	9227	439	857	338
2003	15	9441	445	789	415
2004	13	9361	518	773	422
2005	22	9852	568	397	429
2006	18	10045	545	748	390
2007	18	9162	525	696	377

Each county name refers to the county in which the waste facilities are located. The actual waste may have been generated elsewhere. In this problem, we'll decide if there's been a trend in Los Angeles County's landfilled waste.

a) Make a scatterplot of the waste landfilled in Los Angeles County (y) versus year (x).

b) Compute the least squares regression line. Give the equation of the line and graph it in the scatterplot of part *a*.

c) Based on the fitted regression line, by how much did the landfilled waste in Los Angeles County decrease in a typical year between 1990 to 2007?

d) Is the trend described in part *c* statistically significant? State the relevant hypotheses, give the observed value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$.

e) Check the assumption of normality of the error term ϵ in the linear regression model by making a normal probability plot and a histogram of the residuals. Does the normality assumption appear to be met?

12.16 Refer to the data on solid waste disposal in California counties given in Problem 12.15. In this problem we'll investigate the trend in landfilled waste in Monterey County.

a) Make a scatterplot of the waste landfilled in Monterey County versus year.

b) Compute the least squares regression line. Give the equation of the line and graph it in the scatterplot of part *a*.

c) Based on the regression line computed in part *b*, by how much did the landfilled waste in Monterey County increase in a typical year between 1990 to 2007?

12.17 Refer to the data on solid waste disposal in California counties given in Problem 12.15. In this problem we'll investigate the trend in landfilled waste in Inyo County.

a) Make a scatterplot of the waste landfilled in Inyo County versus year.

b) Compute the least squares regression line. Give the equation of the line and graph it in the scatterplot of part *a*.

c) Carry out a hypothesis test to decide if there was a statistically significant trend in the amount of landfilled waste in Inyo County between 1990 to 2007 (state the hypotheses, give the value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$).

d) Check the assumption that the error term ϵ in the linear regression model follows a normal distribution by making a histogram or normal probability plot of the residuals.

12.18 Refer to the data on solid waste disposal in California counties given in Problem 12.15. In this problem we'll investigate the trend in landfilled waste in San Mateo County.

a) Make a scatterplot of the solid waste landfilled in San Mateo County versus year.

b) Compute the least squares regression line. Give the equation of the line and graph it in the scatterplot of part *a*.

c) Based on the fitted regression line, by how much did the landfilled waste in San Mateo County decrease in a typical year between 1990 to 2007?

d) Is the trend described in part *c* statistically significant? State the relevant hypotheses, give the observed value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$.

e) Refer to the scatterplot of part *a*. Explain why the fitted regression line and results of the hypothesis test of part *d* don't adequately describe the trend in landfilled waste in this county.

12.19 Refer to the data on solid waste disposal in California counties given in Problem 12.15. In this problem we'll investigate the trend in landfilled waste in Santa Barbara County.

a) Make a scatterplot of the waste landfilled in Santa Barbara County versus year.

b) Compute the least squares regression line. Give the equation of the line and graph it in the scatterplot of part *a*.

c) Carry out a hypothesis test to decide if there was a statistically significant trend in the amount of landfilled waste in Santa Barbara County between 1990 to 2007 (state the hypotheses, give the value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$).

d) Check the assumption that the error term ϵ in the linear regression model follows a normal distribution by making a histogram or normal probability plot of the residuals.

12.20 The Columbia River and its drainage basin in the northwestern U.S. lie within one of the world's largest regions of economic mineral deposits. Industrial mining and ore processing to recover lead, zinc, and silver have been prominent in the region since the early 1900's, and have resulted in the release of effluent from smelters containing large amounts of heavy metals into the Columbia River.

Measures to reduce heavy metal releases into the river were implemented in the 1970's, and in the early 1980's improvements in wastewater treatment and termination of much of the mining and smelting operations led to a vast reduction in metals released into the river.

A study was conducted to examine the effects of these mitigation measures on metal concentrations in Columbia River sediments [10]. Because sediment accumulation in calm areas of the river can provide a chronological record of the changes in metal concentrations over time, sediment cores were taken from behind two dams, the Priest Rapids Dam and the McNary Dam. The table below shows the mean heavy metal concentrations (in ppm) in sediment cores from behind the Priest Rapids Dam, as well as the assigned dates based on depth in the cores.

Depth	Assigned Age/Date	Cu (ppm)	Zn (ppm)	As (ppm)	Ag (ppm)	Cd (ppm)	Pb (ppm)
1.2	1997	57	560	5.7	0.28	7.1	42
3.8	1994	60	620	4.4	0.34	7.9	49
6.4	1990	66	690	2.9	0.34	9.0	53
9.0	1986	59	780	<0.1	0.34	10.5	58
11.4	1983	59	870	<0.1	0.36	11.5	66
14.0	1980	48	1030	<0.1	0.20	13.0	58
16.6	1976	45	1020	<0.1	0.19	13.0	70
19.1	1972	45	1010	<0.1	0.21	12.1	77
21.7	1969	52	1200	<0.1	0.28	14.0	110
25.4	1966	44	840	<0.1	0.20	12.1	100

a) Make a scatterplot of the zinc (Zn) concentrations versus year.

b) Compute the least squares regression line and graph the line in the scatterplot of part *a*.

c) Write out the equation of the fitted regression line.

d) Based on the fitted regression line, by how much did the Zn concentration decrease per year, on average?

e) Carry out a hypothesis test to decide if there was a statistically significant trend in Zn concentration over time (state the hypotheses, give the value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$).

f) Calculate a 95% confidence interval for the slope β_1 of the true regression line.

g) Check the assumption that the error term ϵ in the linear regression model follows a normal distribution by making a histogram or normal probability plot of the residuals.

12.21 Refer to the study of heavy metals in sediments of the Columbia River described in Problem 12.20.

a) Make a scatterplot of the cadmium (Cd) concentrations versus year.

b) Compute the least squares regression line and graph the line in the scatterplot of part *a*.

c) Write out the equation of the fitted regression line.

d) Based on the fitted regression line, by how much did the Cd concentration decrease per year, on average?

e) Carry out a hypothesis test to decide if there was a statistically significant trend in Cd concentration over time (state the hypotheses, give the value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$).

f) Calculate a 95% confidence interval for the slope β_1 of the true regression line.

g) Check the assumption that the error term ϵ in the linear regression model follows a normal distribution by making a histogram or normal probability plot of the residuals.

12.22 Refer to the study of heavy metals in sediments of the Columbia River described in Problem 12.20.

a) Make a scatterplot of the lead (Pb) concentrations versus year.

b) Compute the least squares regression line and graph the line in the scatterplot of part *a*.

c) Write out the equation of the fitted regression line.

d) Based on the fitted regression line, by how much did the Pb concentration decrease per year, on average?

e) Carry out a hypothesis test to decide if there was a statistically significant trend in Pb concentration over time (state the hypotheses, give the value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$).

f) Calculate a 95% confidence interval for the slope β_1 of the true regression line.

g) Check the assumption that the error term ϵ in the linear regression model follows a normal distribution by making a histogram or normal probability plot of the residuals.

12.23 Refer to the study of heavy metals in sediments of the Columbia River described in Problem 12.20.

a) Make a scatterplot of the silver (Ag) concentrations versus year.

b) Compute the least squares regression line and graph the line in the scatterplot of part *a*.

c) Write out the equation of the fitted regression line.

d) Based on the fitted regression line, by how much did the Ag concentration increase per year, on average?

e) Carry out a hypothesis test to decide if there was a statistically significant trend in Ag concentration over time (state the hypotheses, give the value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$).

f) Calculate a 95% confidence interval for the slope β_1 of the true regression line.

g) Check the assumption that the error term ϵ in the linear regression model follows a normal distribution by making a histogram or normal probability plot of the residuals.

12.24 Refer to the study of heavy metals in sediments of the Columbia River described in Problem 12.20.

a) Make a scatterplot of the copper (Cu) concentrations versus year.

b) Compute the least squares regression line and graph the line in the scatterplot of part *a*.

c) Write out the equation of the fitted regression line.

d) Based on the fitted regression line, by how much did the Cu concentration increase per year, on average?

e) Carry out a hypothesis test to decide if there was a statistically significant trend in Cu concentration over time (state the hypotheses, give the value of the test statistic and the p-value, and state the conclusion using level of significance $\alpha = 0.05$).

f) Calculate a 95% confidence interval for the slope β_1 of the true regression line.

g) Check the assumption that the error term ϵ in the linear regression model follows a normal distribution by making a histogram or normal probability plot of the residuals.

12.25 Organic fluorochemical compounds such as perfluorooctane sulfonate (PFOS) are used in a variety of applications such as lubricants, fire retardants and pesticides. A field study was conducted to find out if a new method for measuring trace amounts of PFOS is effective in identifying sources of PFOS in the environment [6].

PFOS (ng/L) was measured every two miles along an 80 mile stretch of the Tennessee River near a fluorochemical manufacturing site in Decatur, AL, starting from mile marker 337 (furthest upstream) to mile marker 261. Discharge from the fluorochemical manufacturing facility enters the river at mile marker 301.

Mile marker	River depth (ft)	Conductance	PFOS
337	14	184	27.8
335	36	183	28.9
333	20	184	28.8
331	20	184	25.8
329	28	183	36.9
327	38	186	16.8
325	39	184	27.4
323	40	185	31.0
321	20	184	26.9
319	37	184	22.3
317	26	184	21.8
315	22	185	21.4
313	21	185	18.4
311	20	185	31.6
309	19	186	51.9
307	24	184	52.6
305	22	184	37.1
303	33	185	39.4
301	24	184	54.1
299	23	186	37.3
297	25	185	30.3
295	23	188	74.8
293	26	187	96.4
291	20	187	98.0
289	21	186	107
287	18	185	136
285	28	183	140
283	29	178	106
281	32	180	134
279	46	182	106
277	26	187	112
275	50	191	144
273	19	183	92.3
271	24	185	110
269	38	191	105
267	48	199	119
265	55	201	133
263	70	202	127
261	75	201	119

a) Make a time series plot of PFOS versus mile marker.

b) Describe how the plot provides evidence of a PFOS discharge from the fluorochemical manufacturing facility.

c) Suggest a two-sample hypothesis test for deciding if there is statistically significant evidence that the PFOS concentration is higher downstream of the fluorochemical facility than upstream.

12.26 Thermal spraying is an industrial process by which metals such as aluminum, nickel, and chromium are heated and sprayed along with other non-metallic materials onto surfaces such as automobile engines, aircraft bodies and engines, and bridges, to create a protective coating.

The data in the table below are measurements of aluminum, nickel, and chromium in the blood (μg/L) and urine (μg/g creatinine) of a 42 year old man employed as a thermal sprayer who was exposed to fumes for 6 hours when an exhauster malfunctioned. The metals were measured five times over a period of 1 year following the accident [16].

Time after exposure (days)	Al in blood	Al in urine	Ni in blood	Ni in urine	Cr in blood	Cr in urine
15	8.2	58.4	59.6	700	1.4	7.4
85	5.0	35.4	15.2	122	1.2	3.6
145	5.1	31.4	5.6	45	<0.5	1.7
272	2.6	15.1	3.4	23	0.9	1.6
365	1.8	10.5	2.2	21	<0.5	1.1

The measurements denoted "<0.5" are so-called *nondetects*, and refer to measurements whose exact values are known only to be less than the *detection limit* of 0.5 μg/L.

a) The authors of the cited study suggest that the Al in the blood exhibits a nonlinear exponential decay over time. Make a scatterplot of the Al in the blood versus time and comment on the pattern in the plot.

b) One way to "straighten out" (make more linear) the relationship between Al in the blood and time is to take the log of the Al measurements. Make a scatterplot of the log of the blood Al measurements versus time and comment on the pattern in the plot.

c) Fit the least squares regression line to the log of the Al blood measurements and graph the line in the scatterplot of part *b*.

d) Calculate the coefficient of determination R^2 between the log of the Al blood measurements and time.

e) Based on the value of R^2 calculated in part *d*, does the straight line model adequately describe the relationship between the Al level in blood and time? Explain your answer.

12.27 Refer to the study of metals in the blood and urine of the 42 year old man exposed to the metals in a thermal spraying accident described in Problem 12.26.

a) The authors of the cited study suggest that the Al in the urine exhibits a nonlinear exponential decay over time. Make a scatterplot of the Al in the urine versus time and comment on the pattern in the plot.

b) One way to "straighten out" (make more linear) the relationship between Al in the urine and time is to take the log of the Al measurements. Make a scatterplot of the log of the urine Al measurements versus time and comment on the pattern in the plot.

c) Fit the least squares regression line to the log of the Al urine measurements and graph the line in the scatterplot of part *b*.

d) Calculate the coefficient of determination R^2 between the log of the Al urine measurements and time.

e) Based on the value of R^2 calculated in part *d*, does the straight line model adequately describe the relationship between the Al level in urine and time? Explain your answer.

Bibliography

[1] Bruce Bauerle, David L. Spencer, and William Wheeler. The use of snakes as a pollution indicator species. *Copeia*, 1975(2):366 – 368, May 1975.

[2] S. Brown, D. Shoch, T. Pearson, and M. Delaney. Methods for measuring and monitoring forestry carbon projects in california. Technical report, Winrock International, for the California Energy Commission, PIER Energy-Related Environmental Research, 2004. 500-04-072F.

[3] Phillip E. Farnes. Natural variability in annual maximum water level and outflow of yellowstone lake. 6th Biennial Scientific Conference. Snowcap Hydrology, P.O. Box 691, Bozeman, Montana 59771-0691; farnes@montana.net.

[4] M. E. Finster, K. A. Gray, and H. J. Binns. Lead levels of edibles grown in contaminated residential soils: A field survey. *Science of the Total Environment*, 320:245 – 257, 2004.

[5] Fengxiang X. Han, W. Dean Patterson, Yunju Xia, B.B. Maruthi Sridhar, and Yi Su. Rapid determination of mercury in plant and soil samples using inductively coupled plasma atomic emission spectroscropy, a comparative study. *Water, Air, and Soil Pollution*, 170:161 – 171, 2006.

[6] K. J. Hansen et al. Quantitative characterization of trace levels of PFOS and PFOA in the Tennessee River. *Environmental Science and Technology*, 36(8):1681 – 1685, 2002.

[7] D.R. Helsel. *Nondetects and Data Analysis, Statistics for Censored Environmental Data*. John Wiley and Sons, Inc., 2005.

[8] J. C. Jenkins, D. C. Chojnacky, L. S. Heath, and R. A. Birdsey. National-scale biomass estimation for United States tree species. *Forest Science*, 49:12 – 35, 2003.

[9] Charles B. Johnson et al. Alpine aviation monitoring program, 2001, fourth annual and synthesis report. Technical report, ABR, Inc. - Environmental Research and Services, April 2003.

[10] Vernon G. Johnson, Robert E. Peterson, and Khris B. Olsen. Heavy metal transport and behavior in the lower Columbia River, USA. *Environmental Monitoring and Assessment*, 110:271 – 289, 2005.

[11] H. Kerndorff, S. Kuhn, T. Minden, D. Orlikowski, and T. Struppe. Effects of natural attenuation processes on groundwater contamination caused by abandoned waste sites in Berlin. *Environmental Geology*, 55(2):291 – 301, July 2008.

[12] Alper Kilic and Cengiz Deniz. Inventory of shipping emissions in Izmit Gulf, Turkey. *Environmental Progress and Sustainable Energy*, 29(2):221 – 232, July 2010.

[13] Ambe J. Njoh. Urbanization and development in sub-Saharan Africa. *Cities*, 20(3):167 – 174, 2003.

[14] D.N. Ogbonna, G.T. Amangabara, and T.O. Ekere. Urban solid waste generation in Port Harcourt Metropolis and its implications for waste management. *Management of Environmental Quality*, 18(1):71 – 88, 2007.

[15] William J. Ripple and Eric J. Larsen. Historic aspen recruitment, elk, and wolves in northern Yellowstone National Park, USA. *Biological Conservation*, 95:361 – 370, 2000.

[16] K. Schaller, G. Csanady, J. Filser, B. Jungert, and H. Drexler. Elimination kinetics of metals after an accidental exposure to welding fumes. *International Archives of Occupational and Environmental Health*, 80(7):635 – 641, 2007.

[17] Canadian Forest Service. National Forestry Database, http://nfdp.ccfm.org/index_e.php.

[18] N. Simboura and S. Reizopoulou. A comparative approach of assessing ecological status in two coastal areas of eastern Mediterranean. *Ecological Indicators*, 7:455 – 468, 2007.