12 Correlation and Linear Regression

MTH 3240 Environmental Statistics

Spring 2020

Notes

Objectives

Objectives:

- Obtain and interpret the correlation between two numerical variables.
- 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 the R^2 associated with a fitted regression model.
- Carry out a t test for the slope in a regression model.

Notes

Introduction to Correlation and Regression

- For *one-factor ANOVA*, the explanatory variable (or *factor*) was **categorical**.
- When the explanatory variable is **numerical**, we evaluate its **relationship** to the response variable using ***correlation*** and ***linear regression***.
  - The **correlation** summarizes the **strength** (and **direction**) of the relationship.
  - **Linear regression** gives the **equation** of the best line describing that relationship.

Notes

Example

In a study of the recent decline in the number of "overstory" aspen trees in Yellowstone National Park, the **ages** (yrs) and **diameters** (cm) at breast height of $n = 49$ aspen trees were recorded.

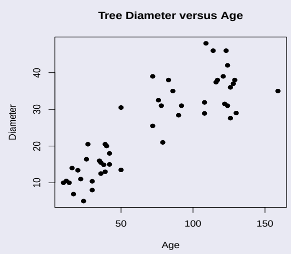
Notes

Ages and Diameters of Trees

Tree	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
⋮	⋮	⋮
47	129	38.0
48	124	42.0
49	123	46.0

Notes

Here's a ***scatterplot*** of the data.



Notes

- Data for which **two variables** are measured on each of n individuals are called ***bivariate data***.
- We'll denote the ***explanatory*** and **response** variables by X and Y, respectively, and store them in columns as below.

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

Notes

- Thus

$\boldsymbol{n}$ = The number of individuals upon which X and Y are measured, i.e. the 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.

Notes

Correlation

- When two variables exhibit (approximately) a **linear relationship**, we summarize that relationship by the ***sample correlation***, denoted r.

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

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

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.

Notes

- The following **properties** of the **correlation** r help us **interpret** its value:

1. The **value** of r will always lie between **-1.0** and **1.0**.
2. The **sign** of r tells us the **direction** of the relationship between X and Y:
   - Positive r values indicate a **positive** relationship.
   - Negative r values indicate a **negative** relationship.

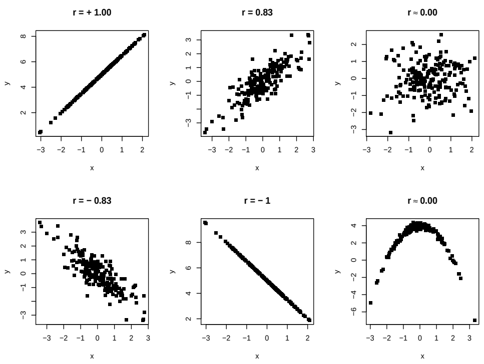
Notes

3. The **value** of r also tells us how **strong** the relationship between X and Y is:
   - r values near **zero** imply a very **weak** relationship or none at all.
   - r 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.

Notes

4. r only measures the strength of the **linear relationship** between X and Y. Curved relationships often have r near zero.
5. r is **not resistant** to outliers.

Notes



Notes

Example

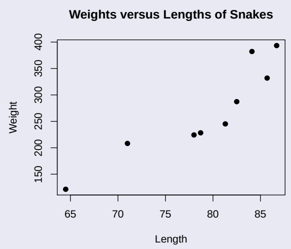
The data below are the **lengths** (cm) and **weights** (g) of $n = 9$ prairie rattlesnakes sampled from the Pawnee National Grassland in northeastern Colorado.

Lengths and Wts 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

The next slide shows a **scatterplot** of the data.

Notes



The **correlation** between **length** and **weight** (obtained using software) is $r = 0.90$, which summarizes the **strong**, **positive**, approximately **linear relationship** seen in the scatterplot.

Notes

Example

Data were collected for a study to determine if **urbanization** is associated with **development**.

Shown below, for each of $n = 40$ sub-Saharan countries, is the **urbanization rate** (percentage of the population living in cities) and **human development index** (**HDI**), which measures the country's health, education, and standard of living.

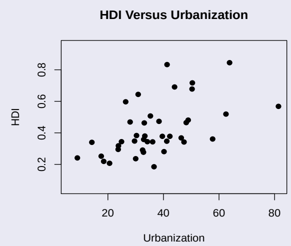
Notes

Urbanization and Development in Africa

Country	HDI	Urbanization
Angola	0.344	34.20
Benin	0.378	42.30
Botswana	0.678	50.30
BurkinaFaso	0.219	18.50
Burundi	0.241	9.01
Cameroon	0.481	48.90
CoteDIvoire	0.368	46.40
⋮	⋮	⋮
Zambia	0.378	39.60
Zimbabwe	0.507	35.30

A scatterplot of the **HDI** values versus **urbanization rates** is on the next slide.

Notes



The **correlation** between the **HDI** value and the degree of **urbanization** (obtained using software) is $r = 0.54$, which reflects the **moderate**, **positive relationship** seen in the plot.

Notes

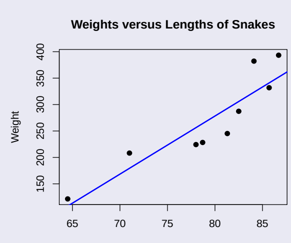
Introduction to Linear Regression

- ***Linear regression*** is a method for obtaining the **equation** of the **line** that best describes the relationship between two variables X and Y.
  1. The **slope** of the line quantifies the amount by which Y **changes** per **one-unit change** in X.
  2. The **equation** can be used to **predict** the value of Y from a given value of X (by plugging the X value into the equation).
  3. The line enhances the appearance of the scatterplot.

Notes

Example

Here's a scatterplot of the data on **lengths** and **weights** of snakes, with the so-called ***fitted regression line***.



Notes

The equation of the ***fitted regression line*** (obtained using software) is:

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

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

The "hat" over the Y indicates that it's the **fitted regression line**, not an observed snake's weight (which would be denoted Y_i).

The **slope**, **11.0**, says that on average, a snake's **weight increases** by about **11.0 g** for each additional **one-cm elongation**.

Notes

The **predicted weight** of a snake that's, say, **75 cm** long is obtained by **plugging** $\boldsymbol{X = 75}$ into the **equation**:

$$\hat{Y} = -601.1 + 11.0(75) = \mathbf{223.9}.$$

Thus we **predict** that a 75-centimeter-long snake will weigh **223.9 g**.

Notes

The Linear Regression Model (Optional for Spring 2020)

- We can describe **bivariate numerical** data using a **statistical model** called the ***linear regression model***.

  The model has a part representing a true (unknown) ***non-random*** **linear process**, which drives the straight line pattern in the data, and another representing ***random*** **deviations** away from that linear pattern.

Notes

(Optional for Spring 2020)

Simple Linear Regression Model: A statistical model for describing bivariate numerical data is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

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 explanatory variable for the ith individual.

Notes

(Optional for Spring 2020)

β_0 is the true (unknown) y***-intercept*** of the underlying ***true regression line***

β_1 is the true (unknown) ***slope*** of the ***true regression line***.

ϵ_i is a random error term following a $\mathbf{N}(0,\,\sigma)$ distribution (and the ϵ_i's are uncorrelated with each other.)

Notes

(Optional for Spring 2020)

- The (true) underlying linear process, $\beta_0 + \beta_1 X$, might represent a physical, chemical, or biological process, the exact nature of which **isn't known**, ...

  ... but the (unknown) intercept and slope coefficients, β_0 and β_1, can be **estimated** from the data.

- When we **estimate** the coefficients, we say that we've ***fitted*** the model to the data.

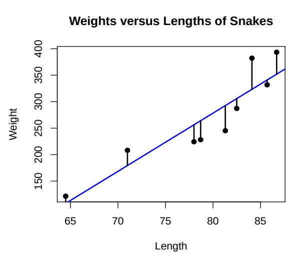
Notes

Estimation of Model Parameters

- We ***fit*** the ***model*** to the data using the ***method of least squares***, which says that the "best fitting" line is the one whose y-intercept b_0 and slope b_1 result in the smallest possible value for the sum of squared vertical deviations away from the line,

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

Notes



Notes

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

$$\hat{Y} = b_0 + b_1 X .$$

The y-intercept $\boldsymbol{b_0}$ and slope $\boldsymbol{b_1}$ (obtained using statistical software) are called the ***least squares estimates*** of the true (unknown) ***population*** (or ***model***) ***coefficients***, which are denoted $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$.

(For the snakes data, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ would be the y-intercept and slope of the line relating **weights** to **lengths** in the **population** of snakes.)

Notes

Example

For the data on **lengths** and **weights** of snakes, the **fitted regression line** given previously,

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

was obtained using statistical software, which reported the **estimated intercept** and **slope** as

$$\begin{aligned} \boldsymbol{b_0} &= \boldsymbol{-601.1} \\ \boldsymbol{b_1} &= \boldsymbol{11.0} \end{aligned}$$

Notes

Some Cautionary Notes

- Be aware:
  1. Linear regression should only be used if the data exhibit a **linear relationship**.
  2. ***Influential points*** are outliers that have a strong influence on the fitted regression line. Outliers in the horizontal (X) direction can be particularly influential.

Notes

Fitted Values and Residuals

- The ***fitted values***, denoted $\boldsymbol{\hat{Y}_i}$, are points on the fitted line that correspond to the **observed** $\boldsymbol{X_i}$**'s**.

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

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

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

The **fitted values** are $\boldsymbol{Y}$ values we'd **predict** for individuals **in** the **data set** that the line was fitted to, using that fitted line.

Notes

- A ***residual***, denoted e_i, is the difference between an **observed Y_i** value and that individual's **fitted value $\hat{Y}_i$**.

Residual: For the ith individual in the data set,

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

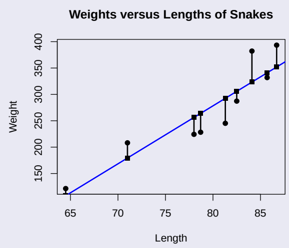
where Y_i is the observed response for that individual and $\hat{Y}_i$ is the fitted value.

The **residuals** are the **deviations** above or below the fitted line.

Notes

Example

For the snakes data, the **residuals** are the vertical deviations shown below, and the **fitted values** are the corresponding points on the fitted line.



Notes

- In a regression analysis, the **line** represents the (linear) **effect of X on Y**.

  The **residuals** represent the **net effect of all *other* variables *besides* X on Y**.

- **Example**: In the snakes regression analysis, a **residual** represents the effects of **all *other* variables *besides* length** on the snake's **weight** (e.g. it's bone density, girth, diet/caloric intake, metabolic rate, etc.).

Notes

R Squared

- The ***coefficient of determination***, denoted R^2 (usually just called **"R squared"**), measures **how well** the **fitted line fits** the **data**.
- One way to compute R^2 is to **square** the **correlation**:

Coefficient of Determination:

$$R^2 = r^2,$$

where r is the correlation.

(We'll see another way to compute it later.)

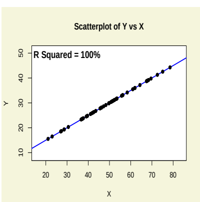
Notes

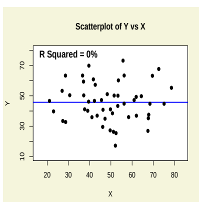
- R^2 is interpreted as the **proportion** of **variation in Y** that is **explained by X**:
  - An R^2 value **close to one** means most of the Y variation is explained by the X variable, and the **model fits** the data **well**.

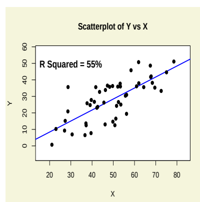
    This shows up as **small residuals**.
  - An R^2 value **close to zero** means very little or none of the Y variation is explained by X (but is explained by *other variables besides* X), and the **model doesn't fit** the data.

    This shows up as **large residuals**.

Notes







Notes

Example

For the data on lengths and weights of snakes

$$R^2 = 0.821.$$

(obtained using software). Thus **82.1%** of the variation in snakes' **weights** is attributable to differences in their **lengths**.

The other **17.9%** is due to the combined effects of **all other variables** (e.g. bone density, girth, diet/caloric intake, metabolic rate, etc.).

Notes

t Test for the Slope

- In the fitted regression line

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

  the **slope b_1** is the estimated average **change** in Y associated with a **one-unit increase** in X.
- **If b_1** was **zero**, there'd be **no change** in Y for any given change in X, i.e. **no relationship** between Y and X.
- A b_1 **different from zero** would mean there's a **relationship** between Y and X.

Notes

- But b_1 can differ from zero due to **sampling error** (because it's just an **estimate** based on data **sampled** from the population).
- We'll test the **null hypothesis** that there's **no relationship** between X and Y.

$$H_0 : \beta_1 = 0$$

where β_1 is the true (unknown) **population** (or **model**) slope coefficient.

Notes

- The alternative is that there's a **relationship** between X and Y.

$$H_a : \beta_1 \neq 0$$

Notes

t **Test Statistic for a Slope**:

$$t = \frac{b_1 - 0}{S_{b_1}},$$

where S_{b_1} is the (estimated) **standard error** of the estimated slope b_1.

- t indicates how many **standard errors** b_1 is **away from** 0, and in what direction (positive or negative).

Notes

- b_1 is an estimate of β_1, so ...
  - **If H_0 was true**, ...

    ... we'd expect b_1 to be close zero.
  - But **if H_a was true**, ...

    ... we'd expect b_1 to differ from zero in the direction specified by H_a.
- Thus ...
  1. t will be approximately **zero** (most likely) if H_0 is true.
  2. It will **differ from zero** (most likely) in the direction specified by H_a if H_a is true.

Notes

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 $H_a : \beta_1 \neq 0$.

Notes

- Now suppose the **residuals*** $e_1, e_2, ..., e_n$ are a sample from a **N**$(0,\ \sigma)$ distribution or that n is **large**.

  In this case, the **null distribution** is as follows.

* More formally, the ***errors*** $\epsilon_1, \epsilon_2, ..., \epsilon_n$ in the regression **model**.

Notes

Sampling Distribution of t Under H_0: If t is the test statistic in a t test for the slope, then when

$$H_0 : \beta_1 \ = \ 0$$

is true,

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

Notes

- **P-values** and **rejection regions** are obtained from the appropriate tail(s) of the $t(n-2)$ **distribution**.

Notes

- The t **test statistic** and **p-value** for the t ***test for the slope*** are reported in the output of statistical software.
- The software also reports results of a t ***test for the y-intercept***:

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

but this is usually of little interest.

Notes

- The software summarizes the results in a **regression table** of the form below.

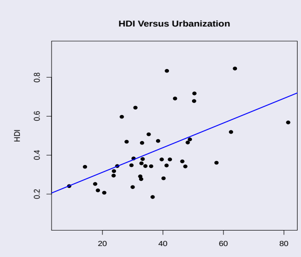
	Estimated Coefficent	Standard Error	t	P-value
Intercept	b_0	S_{b_0}	$t = b_0/S_{b_0}$	p
X	b_1	S_{b_1}	$t = b_1/S_{b_1}$	p

Notes

Example

A scatterplot of data on **urbanization rates** and **human development index** (**HDI**) values for $n = 40$ countries (from a previous example), with **fitted regression line**, is on the next slide.

Notes



Notes

The t **test** results (obtained using software) are below.

	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

Thus, the equation of the **fitted regression line** is

$$\boldsymbol{\hat{Y} \;=\; 0.1852 \,+\, 0.0063X}$$

Notes

For the t **test for the slope**, the hypotheses are

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

The observed **test statistic** value is $\boldsymbol{t = 3.979}$ and the **p-value** is **0.0003**.

Thus, using $\alpha = 0.05$, we **reject $\boldsymbol{H_0}$** and conclude that the observed linear **relationship** between **HDI** and **urbanization** is **statistically significant**.

The $\boldsymbol{R^2}$ value turns out to be **0.294**, so **29.4%** of the variation in **HDI** values can be attributed to differences in the countries' **urbanization** rates. The other **70.6%** is due to **other factors**.

Notes

Notes

Notes