Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

November 11, 2019

Notes

Topics

Notes

Objectives

Objectives:

- State the linear regression model.
- Obtain a fitted regression line, interpret its slope, and use it to predict values of the response variable.
- Interpret residuals and fitted values.
- Interpret sums of squares, degrees of freedom, and mean squares in a regression context.
- State the ANOVA-like partition of the total variation in the response variable.
- Compute and interpret the R-squared (coefficient of determination).

Notes

(**cont'd**)

- Carry out a t test for the slope.
- Use residuals to check normality and constant standard deviation assumptions.
- Compute and interpret a t confidence interval for the slope.
- Obtain and interpret a correlation.

Notes

Linear Regression

Notation

- We'll denote the observations in a ***bivariate*** numerical data set of size n by:

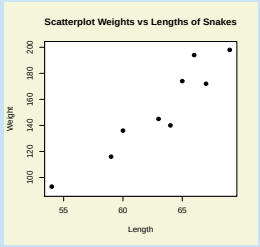
Observation	X variable	Y variable
1	X_1	Y_1
2	X_2	Y_2
3	X_3	Y_3
$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n

Example

The data and scatterplot below show the **lengths** (X) and **weights** (Y) of nine female snakes.

Lengths and Weights of Female Snakes

Snake	Length (cm)	Weight (g)
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145



Linear Regression Model

- A useful statistical model for bivariate data that exhibit a linear relationship is the ***linear regression model***.

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

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where ϵ is a $N(0, \sigma)$ random variable. In this model,

Y is the ***response*** variable.
X is the ***explanatory*** variable (or ***predictor*** variable).
β_0 is an ***intercept*** parameter.
β_1 is a ***slope*** parameter.
ϵ is a $N(0, \sigma)$ ***random error***.

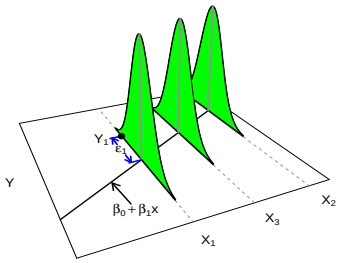
Notes

Notes

Notes

Notes

Linear Regression Model



Notes

- The model says the **expected value** of $\boldsymbol{Y}$ is a **linear function** of $\boldsymbol{X}$ (called the ***true regression line***):

$$E(Y) \;=\; \beta_0 + \beta_1 X$$

and that $\boldsymbol{Y}$ is **normally distributed**:

$$Y \;\sim\; N(\beta_0 + \beta_1 X,\;\; \sigma),$$

where σ **doesn't depend** on $\boldsymbol{X}$.

Notes

Estimating Model Parameters

- The ***least squares estimates*** of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, denoted $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$, are the values of β_0 and β_1 that minimize the sum of squared vertical deviations

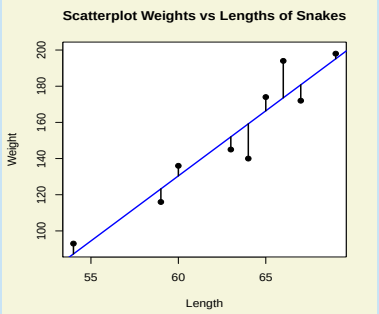
$$\sum_i (Y_i - (\beta_0 + \beta_1 X_i))^2$$

of the observed Y_i's away from the line.

Notes

Example

The **least squares regression line** minimizes the sum of squared deviations of the **weight**s ($\boldsymbol{Y}$) away from the line.



Notes

Proposition

It can be shown that the **least squares estimate** of the **slope** β_1 is given by

$$\hat{\beta}_1 \;=\; \frac{S_{xy}}{S_{xx}}\,,$$

where

$$\mathsf{S}_{xy} \;=\; \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) \qquad \text{and} \qquad \mathsf{S}_{xx} \;=\; \sum_i (X_i - \bar{X})^2\,,$$

and the **least squares estimate** of the **intercept** β_0 is

$$\hat{\beta}_0 \;=\; \bar{Y} - \hat{\beta}_1 \bar{X}\,.$$

Notes

- The resulting line,

$$\hat{Y} \;=\; \hat{\beta}_0 + \hat{\beta}_1 X\,,$$

  is sometimes called the ***fitted regression line***.

- We'll see later how to estimate σ.

Notes

- The **fitted regression line** is used to:
  - **Predict** a **new** Y value from a given X value (by plugging X into the equation of the line).
  - **Describe how** Y **changes** for a given change in X (by looking at the slope $\hat{\beta}_1$).

Notes

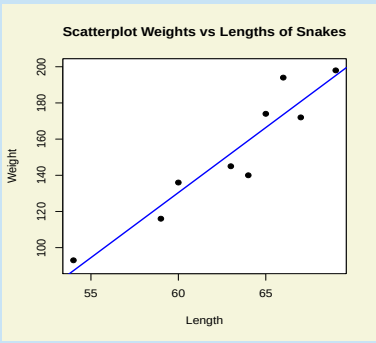
Example

The regression line fitted to the snake data is

$$\hat{Y} \;=\; -301.1 + 7.19X\,.$$

It's shown on the next slide.

Notes



The **predicted** weight of a **62** cm long snake is

$$\hat{Y} \;=\; -301.1 + 7.19(62) \;=\; \mathbf{144.7}$$

grams.

Also, we estimate that a snake's weight would **increase** by **7.19 grams** for each **1 cm** elongation.

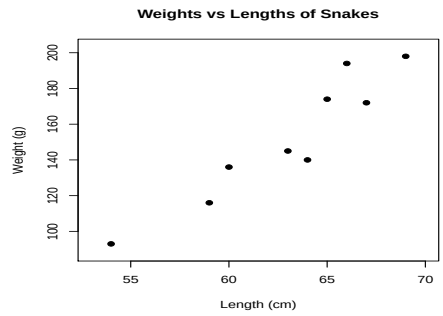
Proposition

The fitted regression line always passes through the "point of averages", $(\bar{X}, \bar{Y})$.

To see, try plugging $\bar{X}$ into

$$\hat{Y} \;=\; \hat{\beta}_0 + \hat{\beta}_1 X$$

and using the fact that $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.

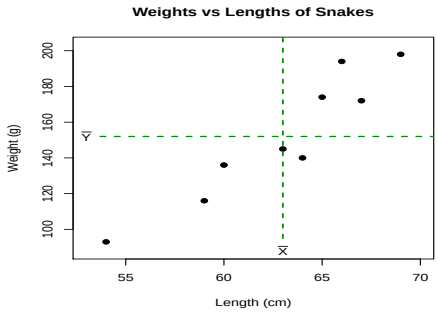


Notes

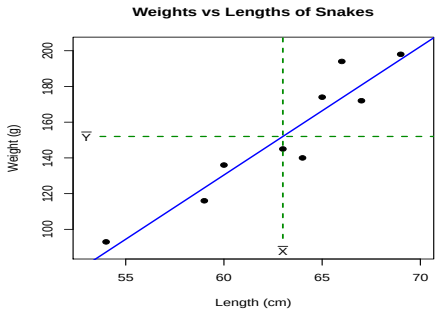
Notes

Notes

Notes



Notes



Notes

- **Some Cautionary Notes About Regression**
  1. **Beware of extrapolation** (predicting Y for values of X outside the range of the X_i's in the data set).
  2. **Beware of influential points** (outliers in the data set that have a strong influence on the fitted regression line). Outliers in the X direction can be particularly influential.

Notes

Fitted Values and Residuals

- The ***fitted value*** (or ***predicted value***) for the ith individual in the data set, $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$, is defined as:

$$\hat{Y}_i \;=\; \hat{\beta}_0 + \hat{\beta}_1 X_i \,,$$

  where $\boldsymbol{X_i}$ is the **observed** value of the explanatory variable for the $\boldsymbol{i}$**th individual**.

  The **fitted value** $\hat{Y}_i$ is the value we'd predict for the response of the ith individual in data set.

  The **fitted values** all lie **on** the **fitted regression line**.

Notes

- The ***residual*** for the ith individual in the data set, e_i, is defined as:

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

where Y_i is the **observed** response for the ith individual.

A **residual** is a **vertical deviation** of an observed response $\boldsymbol{Y_i}$ **away from** the **fitted line**.

A **residual** is interpreted as the **net effect** on $\boldsymbol{Y}$ of **all *other* factors *besides*** $\boldsymbol{X}$.

Notes

Proposition

The residuals sum to zero, i.e.

$$\sum_i e_i \;=\; 0\,.$$

Notes

- **Comment**: By the definitions of $\hat{Y}_i$ and e_i, we can write

$$Y_i \;=\; \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i\,,$$

The **residual** $\boldsymbol{e_i}$ corresponds to the **random error** term $\boldsymbol{\epsilon_i}$ in the model.

Notes

Sums of Squares and an ANOVA-Like Partition

- As in one-factor ANOVA, we can ***partition*** the **total variation** in the **response variable** into two parts:
  - Variation **due to** the **explanatory variable**.
  - Variation **due to random error**.

Notes

- The **partition** will involve the following ***sums of squares*** (shown with their **df**):
  - **SST** is the ***total sum of squares***, defined as

    $$\text{SST} = \sum_i (Y_i - \bar{Y})^2 \qquad \boldsymbol{df = n - 1}$$

    which measures the **total** variation in the Y_i's.

- (cont'd)
  - **SSR** is the ***regression sum of squares***, defined as

    $$\text{SSR} = \sum_i (\hat{Y}_i - \bar{Y})^2 \qquad \boldsymbol{df = 1}$$

    which measures variation in the Y_i's **due to variation** in the $\boldsymbol{X}$ **variable**.

- (cont'd)
  - **SSE** is the ***error sum of squares***, defined as

    $$\text{SSE} = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i e_i^2 \qquad \boldsymbol{df = n - 2}$$

    which measures variation in the Y_i's due to **random error**.

    Note that **SSE** is the **sum** of **squared residuals**.

Proposition

ANOVA-Like Partition of the Total Variation: It can be shown that

$$\text{SST} = \text{SSR} + \text{SSE}.$$

Notes

Notes

Notes

Notes

- **Comments**:
  - **SSR** is analogous to the treatment sum of squares SSTr in one-factor ANOVA.
  - The **df** for **SSE** is $\boldsymbol{n-2}$ since, it can be shown, only $n-2$ of the **residuals** are "free to vary" due to the fact that they satisfy **two constraints**:
    1. They sum to zero, i.e. $\sum_i e_i = 0$.
    2. The products $X_i e_i$ sum to zero, i.e. $\sum_i X_i e_i = 0$.
  - The **df** for **SSR** is only **1** because, it can be shown, any 1 of the deviations $\hat{Y}_i - \bar{Y}$ determines the values of the other $n-1$, thus only 1 of them is "free to vary".

Notes

Additive Property of Degrees of Freedom:

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

Notes

Mean Squares

- The ***mean squares*** are **sums of squares** divided by their **df**.

  The ***mean square for regression***, denoted **MSR**, is

$$\text{MSR} \;=\; \frac{\text{SSR}}{1},$$

  and the ***mean squared error***, denoted **MSE**, is

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

Notes

Estimating σ

- We **estimate** σ (of the $N(0, \sigma)$ **error** distribution) by $\sqrt{\text{MSE}}$ (which represents the **size** of a **typical residual**).

The Regression ANOVA Table

- The results are summarized in a ***regression ANOVA table***:

Source of Variation	df	Sum of Squares	Mean Square	f	P-value
Regression	1	SSR	MSR = SSR/1	MSR/MSE	p
Error	$n-2$	SSE	MSE $=$ SSE$/(n-2)$		
Total	$n-1$	SST			

(We'll discuss the F test statistic and p-value later.)

Notes

The Coefficient of Determination r^2

- The ***coefficient of determination***, denoted r^2, is

Coefficient of Determination:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST - SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}.$$

Notes

- r^2 lies **between 0** and **1**, and **measures how well** the **line fits** the data.
  - Values **close** to **1** imply a **good fit**.
  - Values **close** to **0** imply a **poor fit**.

Notes

- More precisely, r^2 represents the **proportion** of **variation** in the $Y_i's$ that's **due to variation** in the $X_i's$.

  To see why, consider the **two sources** of **variation** in Y:

  1. Variation in Y **due to** X.
  2. Variation in Y due to **all *other* factors *besides*** X (i.e. due to **random error**).

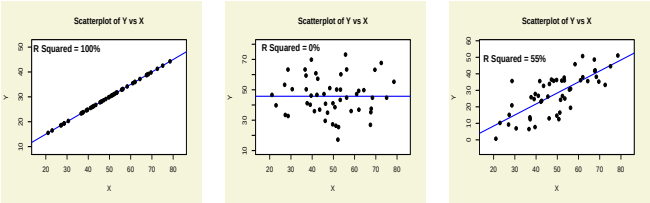
Notes

- Also note that,

$\frac{\text{SSE}}{\text{SST}}$ = The **proportion** of variation in the Y_i's due to **random error** (i.e. due to all ***other* factors *besides* X**),

so

$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$ = The **proportion** of variation in the Y_i's due to X.



Example

For the snakes data, recall that the **fitted regression line** is

$$\hat{Y} = -301.1 + 7.19X\,.$$

The **regression ANOVA table** (from R) is:

Source of Variation	df	Sum of Squares	Mean Square	f	P-value
Area	1	8896.3	8896.3	56.9	0.0001
Error	7	1093.7	156.2		
Total	8	9990.0			

Thus

$$\textbf{SST} = \mathbf{9990.0} \quad \text{and} \quad \textbf{SSE} = \mathbf{1093.7}$$

so

$$\boldsymbol{R^2} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{1093.7}{9990} \approx \mathbf{0.89}.$$

Thus **89%** of the **variation** in snakes' **weights** can be explained by variation in their **lengths**.

The other **11%** is due to all **other factors** (girth, bone density, metabolism, diet, physical activity, etc.).

Notes

Notes

Notes

Notes

t Test for the Slope β_1

- We'll want to test

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

The null hypothesis says that there's **no linear relationship** between the response variable Y and the explanatory variable X.

The alternative says **there's a relationship**.

- We'll need the following fact.

Notes

Proposition

When the error term ϵ in the linear regression model follows a $N(0,\, \sigma)$ distribution (and the responses $Y_1, Y_2, \ldots Y_n$ are independent of each other),

$$\hat{\beta}_1 \sim N\left(\beta_1,\ \sigma_{\hat{\beta}_1}\right),$$

where

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\mathsf{S}_{xx}}} \qquad \left(\text{with } \mathsf{S}_{xx} = \textstyle\sum_i (X_i - \bar{X})^2\right).$$

It follows that

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,\, 1).$$

Notes

- Thus the ***expected value*** of $\hat{\beta}_1$ is

$$E(\hat{\beta}_1) = \beta_1$$

and the ***standard error*** (standard deviation) of $\hat{\beta}_1$ is

$$\sigma_{\hat{\beta}_1} = SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\mathsf{S}_{xx}}}.$$

The standard error of $\hat{\beta}_1$ measures sample-to-sample **variation** in the **slope** of the **fitted line**.

It will be **small** when S_{xx} is large, i.e when the $\boldsymbol{X_i}$'s are **spread out**.

Notes

Estimated Standard Error of $\hat{\beta}_1$: Because $\sqrt{\mathsf{MSE}}$ estimates σ, we estimate $\sigma_{\hat{\beta}_1}$ by

$$S_{\hat{\beta}} = \frac{\sqrt{\mathsf{MSE}}}{\sqrt{\mathsf{S}_{xx}}}.$$

Notes

- To derive a hypotheses test procedure, we'll need the following.

Proposition

When the error term ϵ in the linear regression model follows a $N(0,\, \sigma)$ distribution (and the responses $Y_1, Y_2, \ldots Y_n$ are independent of each other),

$$T \;=\; \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}}} \;\sim\; t(n-2)\,.$$

Notes

t Test Statistic for the Slope β_1:

$$T \;=\; \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}}$$

Notes

1. **Large positive** values of t provide **evidence against H_0 in favor of $H_a: \beta_1 > 0$.**
2. **Large negative** values of t provide **evidence against H_0 in favor of $H_a: \beta_1 < 0$.**
3. **Large positive *and* large negative** values of t provide **evidence against H_0 in favor of $H_a: \beta_1 \neq 0$.**

Notes

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

$$H_0 : \beta_1 = 0$$

is true,

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

Notes

- The $t(n-2)$ curve gives us:
  - The ***rejection region*** as the **extreme 100α% of t values** (in the direction(s) specified by H_a).
  - The ***p-value*** as the **tail area(s) beyond the observed t value** (in the direction(s) specified by H_a).

- **Comment**: There's also a **test** for the **intercept** β_0, but we won't consider the details in these slides.

Summarizing the Results of a Regression Analysis

- The results are summarized in a ***regression table***:

	Estimated Coefficient	Standard Error	t	P-value
Intercept	$\hat{\beta}_0$	$S_{\hat{\beta}_0}$	$t = \hat{\beta}_0 / S_{\hat{\beta}_0}$	p
X Variable	$\hat{\beta}_1$	$S_{\hat{\beta}_1} = \sqrt{\text{MSE}}/\sqrt{\text{S}_{xx}}$	$t = \hat{\beta}_1 / S_{\hat{\beta}_1}$	p

Example

Yellowstone Lake in southeastern Yellowstone National Park covers an area of about 136 mi^2 (352 km^2). The lake's water level varies from year to year in response to differences in the winter's snow accumulation, spring precipitation, and air temperatures.

The U.S. Geological Survey started collecting data on the lake's elevation in 1922.

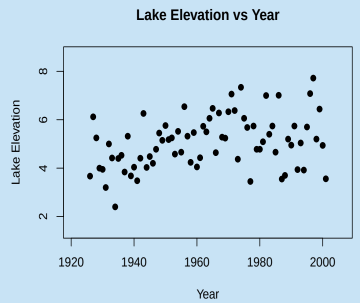
The maximum elevation (ft) is shown on the next slide for each of the $n = 76$ years 1926 - 2001.

Notes

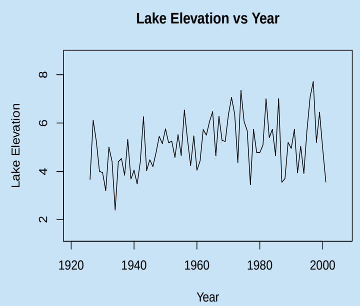
Notes

Notes

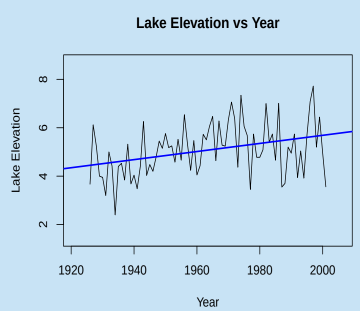
Notes



Notes



Notes



Notes

Results of a **regression analysis** (performed in R) is below.

	Estimated Coefficent	Standard Error	t	P-value
Intercept	-27.785	10.540	-2.636	0.0102
Year	0.017	0.005	3.118	0.0026

The **estimate** of the true (unknown) slope β_1 is $\hat{\beta}_1 = \mathbf{0.017}$ **ft/yr**, with a **standard error** $S_{\hat{\beta}_1} = \mathbf{0.005}$.

On average, the lake's elevation has been increasing by about **0.017 feet** (1/5 of an inch) **per year**, plus or minus about **0.005 feet** (1/10 of an inch).

Notes

For the test of

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

the t **test statistic** is $t = 0.005$, and the **p-value** is **0.0026** (from the $t(74)$ **distribution**).

Thus using $\alpha = 0.05$, we **reject** H_0 and conclude that the observed trend in the lake's elevation is statistically significant.

Notes

- **Comments**:
  - The F **test statistic** given in the **regression ANOVA table** is
    $$F = \frac{\text{MSR}}{\text{MSE}},$$
    and is for another test of
    $$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$
    The reported **p-value** is from the right tail of the $F(1, n-2)$ **distribution**.

Notes

- (cont'd)
  - If
    $$T = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}}$$
    is the t **test statistic** for the test of
    $$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$
    then
    $$F = T^2,$$
    and the **p-values** for the two tests will be **the same**.

Notes

Checking the Model Assumptions

- For the t **test** (and F **test**) for the **slope**, we assume the ϵ_i's are iid $N(0, \sigma)$.

  Note that σ is assumed to be **constant** (i.e. ***doesn't*** depend on X).
- **Checking the Normality Assumption**: Use a **histogram** or **normal probability plot** of the **residuals**.
- **Checking the Constant σ Assumption**: Plot either
  - The **residuals versus the X_i's**.
  - The **residuals versus the fitted values**.

  Usually, when σ *isn't* constant, it increases with the fitted value.

Notes

CI for the Slope β_1

- Using the fact that

$$\frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2)\,,$$

we can derive a **CI** for $\boldsymbol{\beta_1}$.

> **t CI for β_1**: For bivariate data following the linear regression model with true (unknown) slope β_1, a $\mathbf{100(1-\alpha)\%}$ ***t confidence interval for β_1*** is
>
> $$\hat{\beta}_1 \pm t_{\alpha/2,\,n-2} \cdot S_{\hat{\beta}_1}\,.$$

- The **CI** is valid if the **random errors ϵ_i** in the regression model follow a **normal** distribution (or n **is large**).

Example

Using the R output from the last example,

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = \mathbf{0.017} \qquad \text{and} \qquad \boldsymbol{S}_{\hat{\boldsymbol{\beta}}_{\mathbf{1}}} = \mathbf{0.005}\,,$$

so a **95% CI** for the true (unknown) slope $\boldsymbol{\beta_1}$ of the trend line describing Yellowstone Lake's elevation is

$$0.017 \pm 1.992(0.005) = \mathbf{(0.007,\ 0.027)}.$$

The fact that the interval doesn't contain zero is consistent with the results of the t test for β_1 in the previous example.

- There's also a **CI** for the **intercept $\boldsymbol{\beta_0}$**, but we won't consider the details here.

Notes

Notes

Notes

Notes

Correlation

Correlation

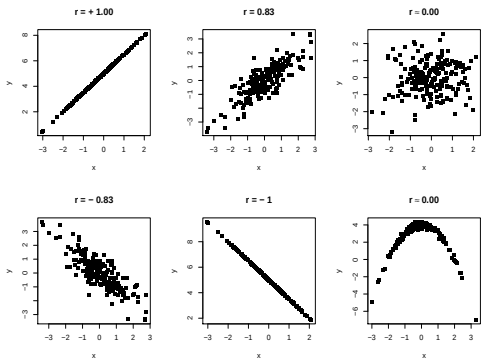
- We measure the **strength** and **direction** of the *relationship* between two variables using their ***correlation*** r.

Sample Correlation: The ***sample correlation***, denoted r, 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 $\bar{X}$, $\bar{Y}$, S_x, and S_y are the sample means and sample standard deviations of $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$.

Notes



Notes

- **Comment**: It's easy to show that an equivalent formula for r is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

where

$$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S_{yy} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

and

$$S_{xy} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Notes

- **Properties and Interpretation of r**:
  1. $-1 \leq r \leq 1$.
  2. $r > 0$ implies a **positive relationship** between X and Y, and $r < 0$ implies a **negative relationship**.
  3. $r \approx 0$ implies **little or no** (linear) **relationship** between X and Y, and of $r \approx \pm 1$ implies a **strong** (linear) **relationship**.
  4. $r = \pm 1$ only when there's a **perfect linear relationship** between X and Y.

Notes

- (cont'd)
  5. r only measures the strength and direction of the **linear relationship**. In particular, X and Y may have a strong curved relationship, but $r \approx 0$.
  6. r is **not resistant** to outliers.
  7. A correlation **doesn't imply** a **cause-and-effect relationship**. The relationship might be due instead to *confounding variables*.

Notes

Example

For the snakes data,

$$\bar{X} = 63 \qquad \bar{Y} = 152$$
$$S_x = 4.64 \qquad S_y = 35.34$$

So the **correlation** is

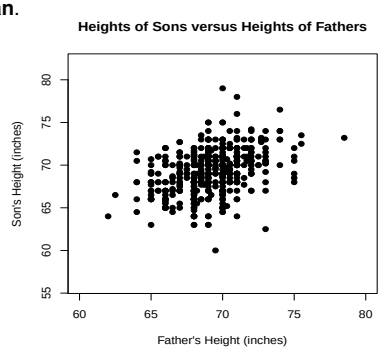
$$\begin{aligned} r &= \frac{1}{9-1}\left[\left(\frac{60-63}{4.64}\right)\left(\frac{136-152}{35.34}\right) + \left(\frac{69-63}{4.64}\right)\left(\frac{198-152}{35.34}\right) \right. \\ &\quad \left. + \cdots + \left(\frac{63-63}{4.64}\right)\left(\frac{145-152}{35.34}\right)\right] \\ &= \mathbf{0.943}\,. \end{aligned}$$

Notes

- **Comments**:
  - It can be shown that the **coefficient of determination** r^2 is the **square** of the **correlation** r.
  - An equivalent formula for the **slope** $\hat{\beta}_1$ of the **fitted regression line** is
    $$\hat{\beta}_1 = r \cdot \frac{S_y}{S_x}\,.$$
    Therefore:
    - $\hat{\beta}_1$ always has the **same sign** as r.
    - Each one-standard deviation increase in X leads to an increase of only r standard deviations in Y. This is called **"regression to the mean"**.

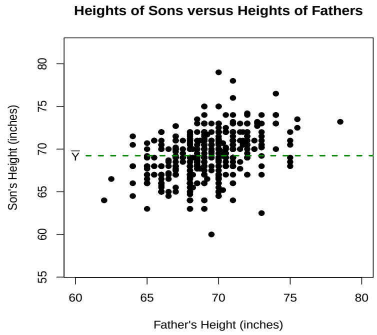
Notes

- Francis Galton's data on **heights** of **fathers** (x) and **heights** of their **sons** (y) illustrates **regression to the mean**.



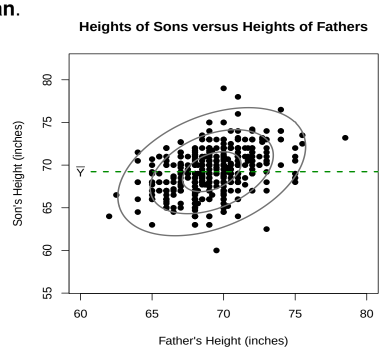
Notes

- Francis Galton's data on **heights** of **fathers** (x) and **heights** of their **sons** (y) illustrates **regression to the mean**.



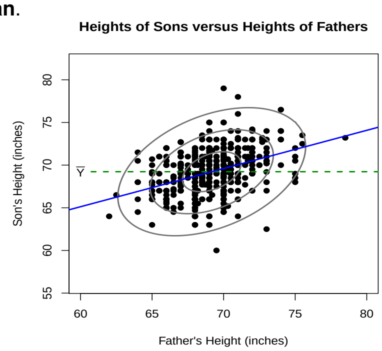
Notes

- Francis Galton's data on **heights** of **fathers** (x) and **heights** of their **sons** (y) illustrates **regression to the mean**.



Notes

- Francis Galton's data on **heights** of **fathers** (x) and **heights** of their **sons** (y) illustrates **regression to the mean**.



Notes

Notes