# STAT 212: Principles of Statistics II

## Lecture Notes: Chapter 1 (Part A)

## Simple Linear Regression

Sharmistha Guha

Dept. of Statistics

Texas A&M University

Fall 2022

# Simple Linear Regression and Correlation

Regression: *Methodology for studying the relationship between two sets of variables.*

Simplest case is where we have only two variables, say $x$ and $y$.

How does $y$ change when $x$ increases or decreases?

| $x$ | $y$ |
|---|---|
| Animal age | Animal weight |
| Height | Weight |
| Age | Blood pressure |
| Temperature | Chemical reaction time |

*Example 1*

Galileo discovered that gravity accelerates all objects at the same rate. Drop a bowling ball and a marble together from the top of a building and they will reach the ground at the same time.

$x =$ height $\quad y =$ time it takes object to reach ground when dropped from height $x$

Let $g$ be the gravitational constant. Isaac Newton argued that

$$y = \sqrt{\frac{2x}{g}}.$$

Under 'ideal' conditions, this equation is **exact**.

Suppose we didn't know about the equation and wanted to establish the relationship between $x$ and $y$ by experimentation.

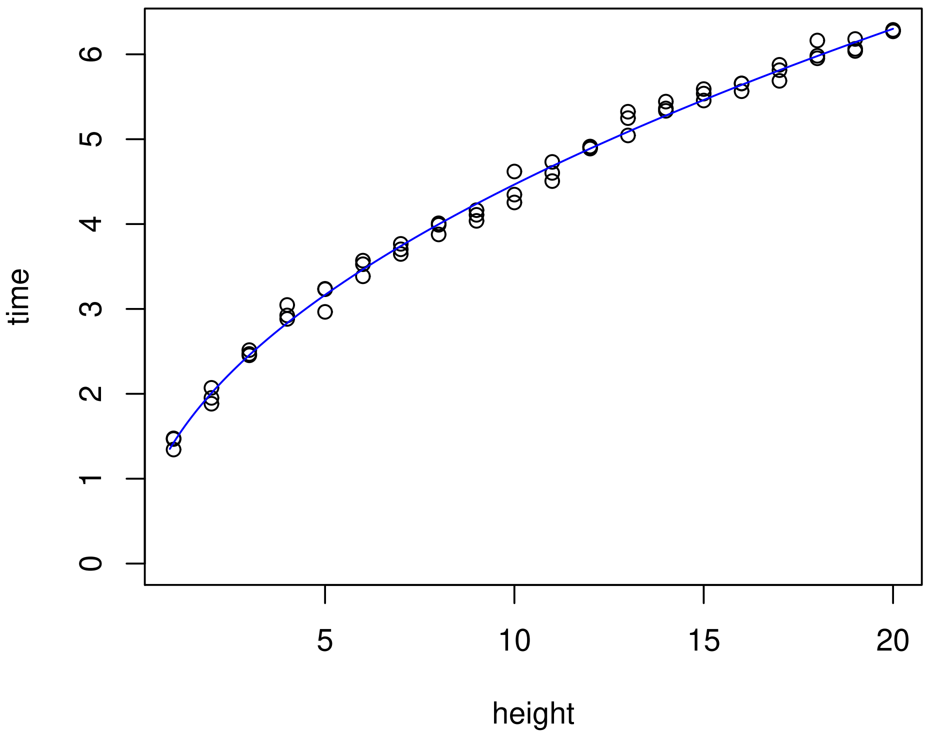
How the experiment might be done:

- Choose several values of $x$, say $n$ of them, and call them $x_1, \ldots, x_n$.

- For each height $x_i$, drop a ball three times from height $x_i$ and record the times it took the ball to reach the ground.

- Fit a curve through the $3n$ data points. This curve estimates the functional relationship between $x$ and $y$.

# *Scatterplot for gravity experiment*



The curve fitted through the points *estimates the true relationship* between height ($x$) and time ($y$).

*Note:* Differences between times at the same height are due completely to (random) errors (or ‘noise’) made in recording the times.

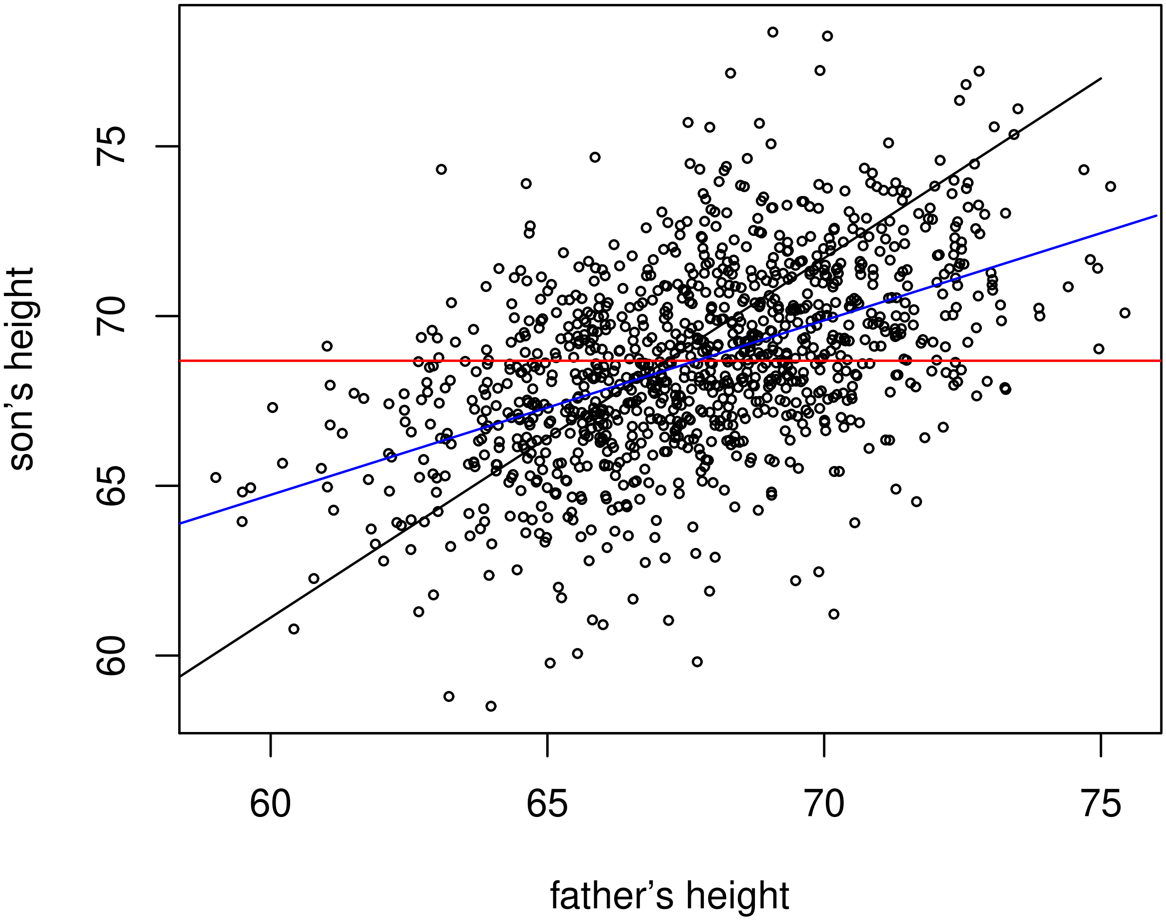
*Example 2*

Sir Francis Galton (1822-1911), cousin of Charles Darwin, was an English scientist who, among many other things, studied inheritance. (He coined the term "nature vs. nurture.")

Galton was an early pioneer in the field of statistics. He is credited with some of the earliest discoveries in the areas of *regression* and *correlation.*

Galton collected data on the heights of fathers and their sons. A famous data set of his on the heights of 1,078 pairs of fathers and sons is shown on the next page.

## *Galton's father-son height data*



The red line denotes the average height of all sons, blue line denotes the average son's height as a function of father's height, and the **black line** simply denotes the line $y = x + 1$.

Note: *The average height of all the sons is about one inch higher than that of the fathers.*

The line $y = x + 1$ is interesting because intuitively we might expect that a group of fathers all of whom have height $x$ would have sons whose heights average $x + 1$.

However, this isn't true. Sons' heights *regress towards the overall mean.* This is illustrated by the fact that the slope of the line of averages is less than 1.

Consider the heights of sons whose fathers are all the same height. Obviously the sons heights *won't* all be the same (so there is *still* 'noise').

In contrast to Example 1, *the differences between the heights of sons is not due solely to errors made in recording the sons' heights.*

Initially we study a fairly simple situation in which the relation between $y$ and $x$ is $\approx$ linear:

$$y \approx ax + b.$$

## Statistical model

$x_1, \ldots, x_n$ are values of a control or independent or explanatory variable (also called *covariate* or *predictor* or *regressor* or *feature* etc.)

These $x$ values are *fixed* by the experimenter. (Note: these may be random variables but are considered *"conditioned"*, i.e. fixed, here.)

$Y_1, \ldots, Y_n$ are corresponding values of a response or outcome or dependent variable. These are random variables (even though the respective $x_i$ is fixed!), as indicated by upper case $Y$'s.

We assume a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

The following assumptions are made about the preceding model:

(1) $\epsilon_1, \ldots, \epsilon_n$ are *unobserved* random variables.
Note: they are also called *errors* or *'noise'.*

(2) $\epsilon_1, \ldots, \epsilon_n$ are *independent* of each other.

(3) Each $\epsilon_i$ has a $N(0, \sigma^2)$ distribution. The variance $\sigma^2$ stays the same for all $\epsilon_i$'s. (Also called the *'homoskedasticity assumption'*).

(4) $\beta_0$, $\beta_1$ and $\sigma^2$ are unknown, but fixed, parameters. $\beta_0$ is called the *intercept*, $\beta_1$ the *slope* and $\sigma^2$ the *error (or 'noise') variance*.

**The statistical problem:**

Infer on the unknown parameters $\beta_0$, $\beta_1$ and $\sigma^2$ using the data $(x_1, y_1)$, $(x_2, y_2), \ldots, (x_n, y_n)$.

Note that, according to the model,

$$E(Y_i) = \beta_0 + \beta_1 x_i + E(\epsilon_i) = \beta_0 + \beta_1 x_i,$$

and $\text{Var}(Y_i) = \sigma^2$. Here $E(\cdot)$ and $\text{Var}(\cdot)$ respectively denote the expectation (i.e. average or mean) and variance of $Y$ given the predictor.

Interpretation: The "expected" (i.e. average) value of $Y$ in the sub-population where the predictor equals $x$ is a linear function of $x$.

The parameter $\sigma^2$ determines how tightly the data will cluster about the line $y = \beta_0 + \beta_1 x$. If $\sigma^2 = 0$, the relationship is deterministic.

An analysis of the relationship between $x$ and $y$ is referred to as a *regression analysis.*

Interpretation of slope parameter: $\beta_1$ represents the change in the *average* response when the independent variable increases by 1 unit.

Example: if $\beta_1 = -5$, this means the average response decreases by 5 when $x$ goes up 1 unit.

Given a set of data, how do we estimate the parameters $\beta_0$ and $\beta_1$? One method of doing so is based on the principle of *least squares.*

Given the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, find the values of $b_0$ and $b_1$ such that

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

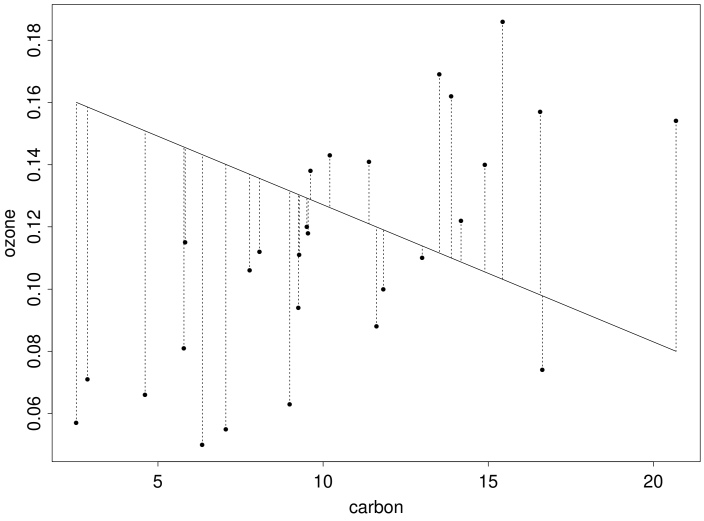
is minimized.

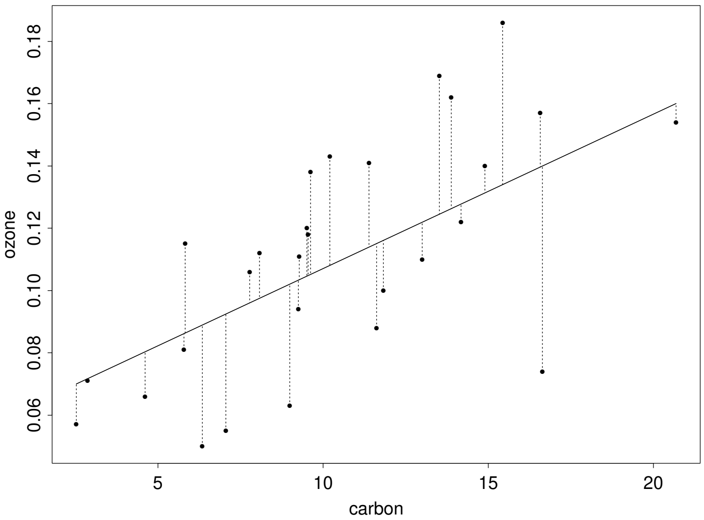
Carl Friedrich Gauss (1777-1855) is credited with discovering least squares.

- Take partial derivatives of $f$ wrt $b_0$ and $b_1$.

- Set derivatives to 0 and solve for $b_0$ and $b_1$.

# Illustration of the Least Squares Principle





*Two candidate lines and the vertical deviations of the data from each line*

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = 2\sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)](-1)$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = 2\sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)](-x_i).$$

Setting the first expression equal to 0 and solving for $b_0$ yields

$$b_0 = \bar{y} - b_1\bar{x}.$$

Substituting this value of $b_0$ into the second equation leads to

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \widehat{\beta}_1.$$

This implies that the best $b_0$ is

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}.$$

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are called the *least squares estimates* of $\beta_0$ and $\beta_1$.

*Example 3:* *Using R to find least squares line for Galton's data*

Suppose the response and independent variables are called `y` and `x`, respectively, in your R session.

The `R` command:

```
summary(lm(y~x))
```

gives the basic statistics for a linear regression fit, including the least squares line.

The estimated intercept and slope are:

$$\widehat{\beta}_0 = 33.887 \quad \text{and} \quad \widehat{\beta}_1 = 0.514.$$

The least squares line is:

$$y = 33.887 + 0.514x.$$

For the population of father-son pairs from which these data were collected, it is estimated that when father's height increases by 1 in., the son's height increases, on average, by 0.514 in.

## Estimation of $\sigma^2$

In our model $\sigma^2$ is the other unknown parameter (besides $\beta_0$ and $\beta_1$). Our model says that

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

where

$$\sigma^2 = \text{Var}(\epsilon_i).$$

Since

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i),$$

the definition of variance tells us that

$$\sigma^2 = E\left\{[Y_i - (\beta_0 + \beta_1 x_i)]^2\right\}.$$

It would thus make sense to estimate $\sigma^2$ by something like

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right]^2.$$

Define $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$, $i = 1, \ldots, n$. These are called the *predicted values*.

The so-called *residuals* are

$$e_i = y_i - \widehat{y}_i, \quad i = 1, \ldots, n.$$

The residuals serve as proxies for the unobservable error terms. Define the *error sum of squares*, or $SSE$, by

$$SSE = \sum_{i=1}^{n} e_i^2.$$

Our estimate of $\sigma^2$ will be:

$$\widehat{\sigma}^2 = \frac{SSE}{n-2}.$$

Dividing by $n-2$, rather than $n$, makes $\widehat{\sigma}^2$ an *unbiased estimator*. (Recall the case of sample variance where you need to divide by $n-1$.)

Here we subtract 2, rather than 1, from $n$ since we have had to estimate two parameters, $\beta_0$ and $\beta_1$, in order to construct the variance estimator $\widehat{\sigma}^2$.

## Use of residuals to check the model

The residuals are a main tool in checking to see whether or not our model assumptions are correct. If the model *is* correct, then $e_1, \ldots, e_n$ should behave very much like a random sample from a normal distribution.

Plotting the residuals in various ways allows us to check the assumptions.

- Plotting $e_i$ vs. $x_i$ checks on whether the *regression curve is more complicated than just a straight line*.

- Plotting $e_i$ vs. $\widehat{y}_i$ is a good way to check the *"equal variances" assumption*.

- A histogram, kernel density estimate or normal quantile plot of the residuals allow us to check the *normality assumption*.

In R, suppose we use the following commands:

```
fit=lm(y~x)
resid=fit$residuals
predict=fit$fitted.values
```

This puts the residuals into the vector called `resid` and the predicted values into the vector `predict`.

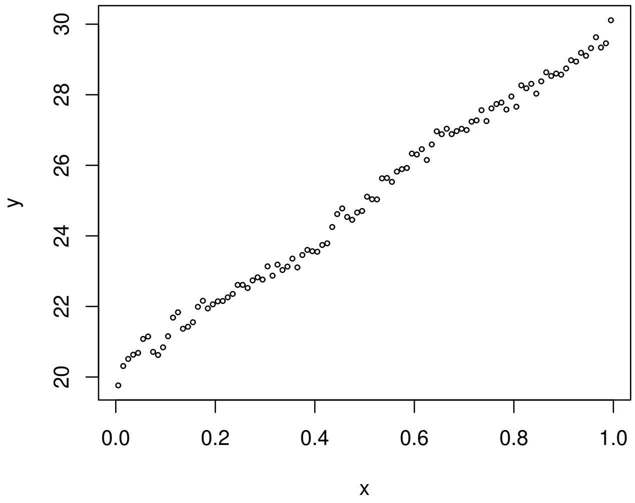
We can then produce the various plots on the previous page. The first two plots can be produced via the `plot ()` command, as follows.

```
plot(x, resid, xlab ="x", ylab ="Residuals")
plot(predict, resid, xlab ="yhat", ylab ="...")
```

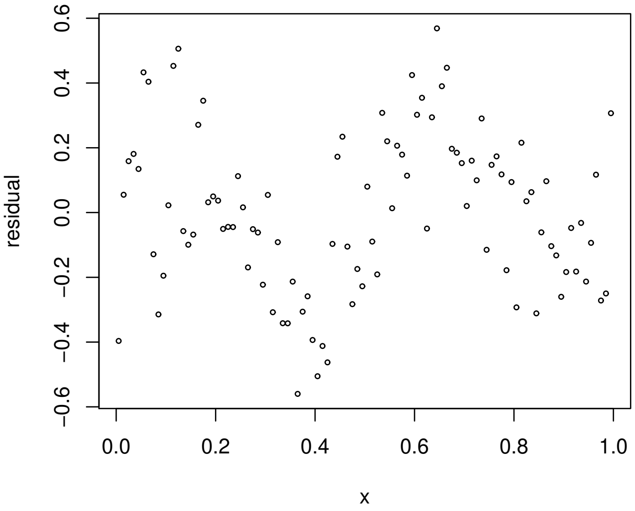
The third set of plots can be produced using: `hist(resid)` for histogram, `plot(density(resid))` for the density estimator and `qqnorm(resid)` for the normal quantile plot or the q-q plot which, ideally, should closely resemble the $y = x$ line.

Here are a few examples of the first two type of plots in the next two slides.

## *Scatterplot of $y$ versus $x$*

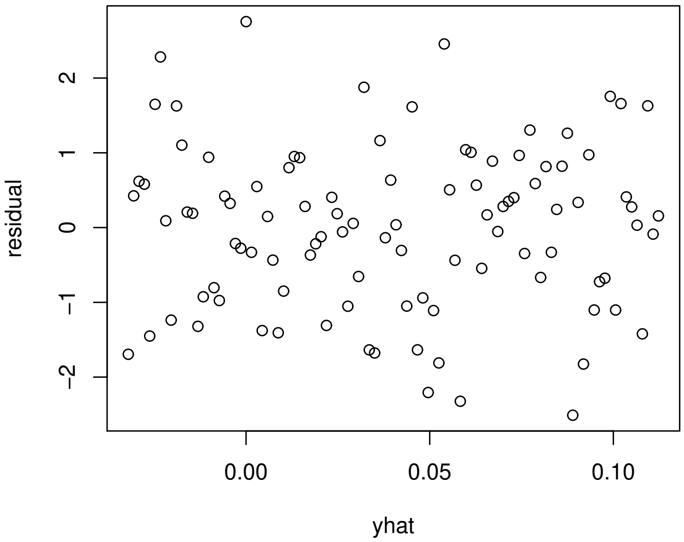


## *Plot of residuals from fitted straight line versus $x$*

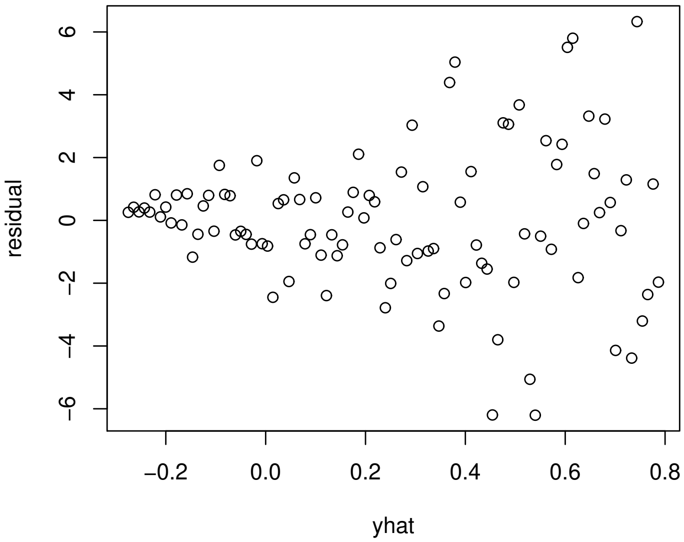


Note the pattern in the residuals.

## *A “good” plot of residuals versus $\widehat{y}$*



## *Plot of residuals versus $\widehat{y}$ in which the variance increases with $\widehat{y}$*



## Coefficient of determination

Define the *total sum of squares*, or $SST$, by

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

This measures *all* the variation in $y_i$s, from whatever source.

*ANOVA decomposition of SST*

$$\begin{aligned} SST &= \sum_{i=1}^{n} (y_i - \widehat{y}_i + \widehat{y}_i - \bar{y})^2 \\ &= SSE + \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \\ &\quad +2 \sum_{i=1}^{n} (y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}). \end{aligned}$$

Now we will argue that the very last term on the right-hand side above is 0.

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \bar{y} - \widehat{\beta}_1 \bar{x} + \widehat{\beta}_1 x_i \implies$$

$$\widehat{y}_i - \bar{y} = \widehat{\beta}_1 (x_i - \bar{x})$$

and

$$y_i - \widehat{y}_i = y_i - \bar{y} - \widehat{\beta}_1 (x_i - \bar{x})$$

Therefore

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}) =$$

$$\widehat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \widehat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 =$$

$$\widehat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - \widehat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = 0.$$

We've proven that

$$SST = SSE + \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2.$$

Define

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

which is called the *regression sum of squares*.

$SSR$: amount of variation in $y$ that is systematic and can be explained via the linear model, i.e. the linear relationship between $y$ and $x$.

$SSE$: amount of variation in $y$ that is unsystematic i.e. **not** explainable through $x$ and is purely due to random error or noise unrelated to $x$.

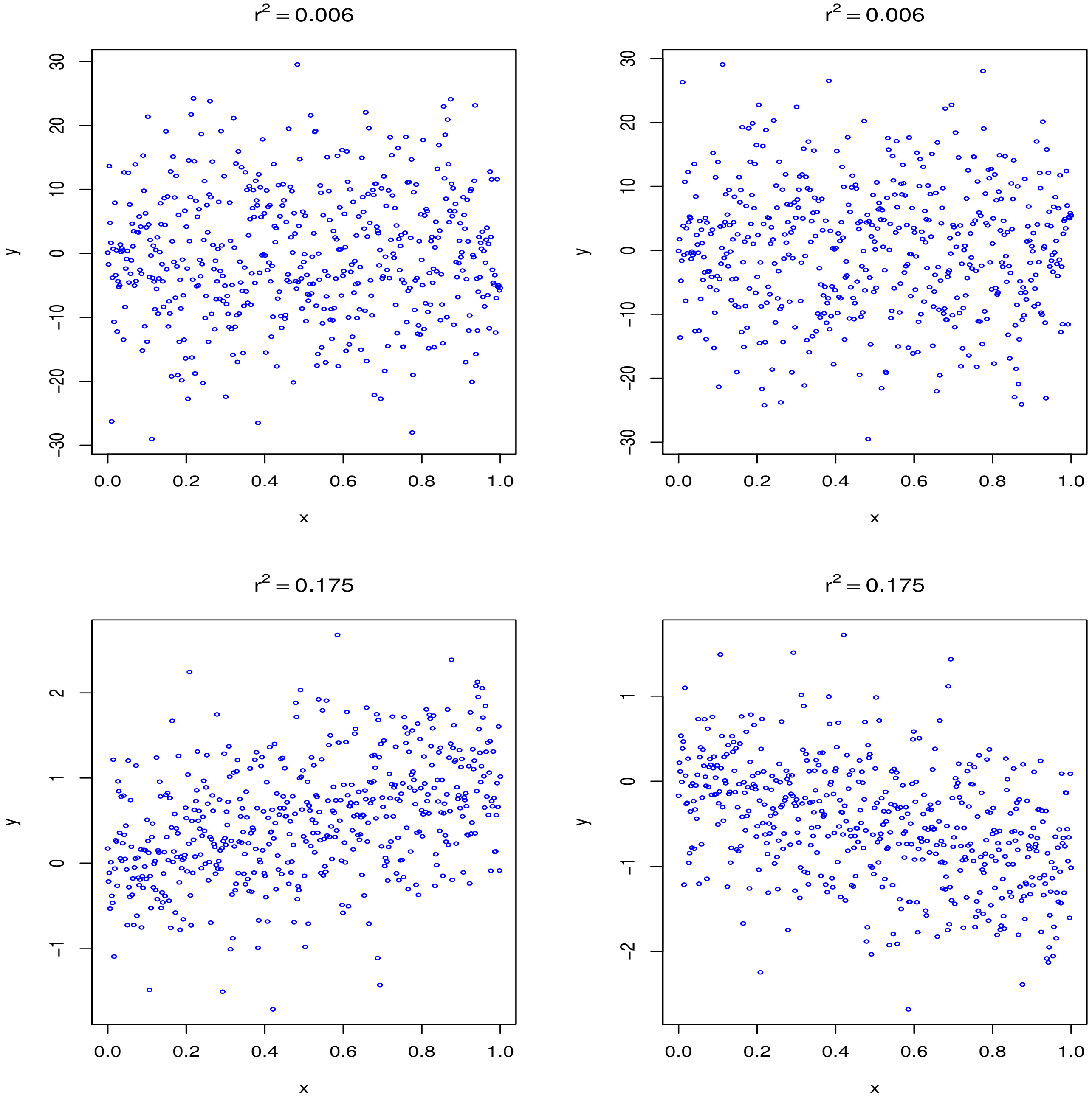
A nice summary measure is the coefficient of determination, given by the ratio:

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

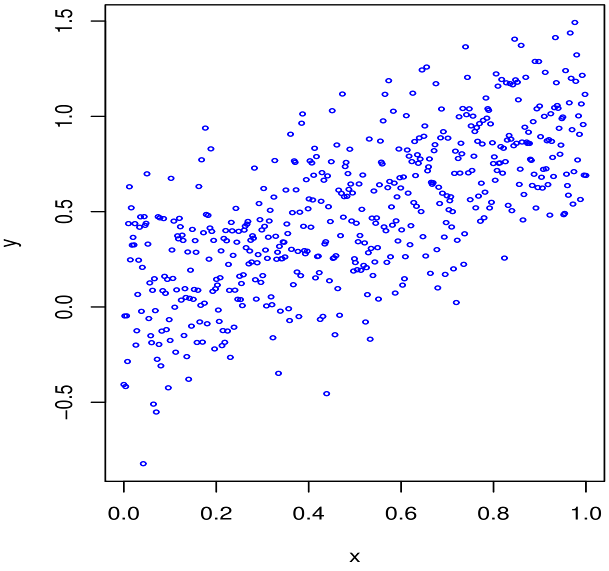
It's always true that $0 \leq R^2 \leq 1$.

$R^2$ is the *fraction of the total variation in $y$ that is due to (or explainable via) the straight line relationship between $y$ and $x$.*

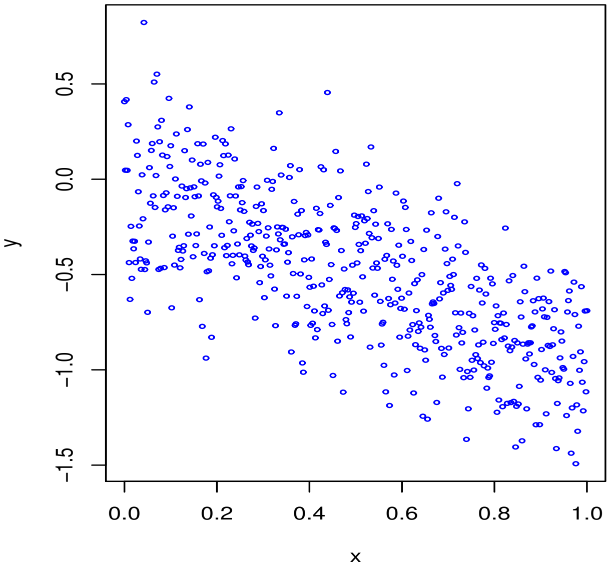
# *Scatterplots and associated values of $R^2$*



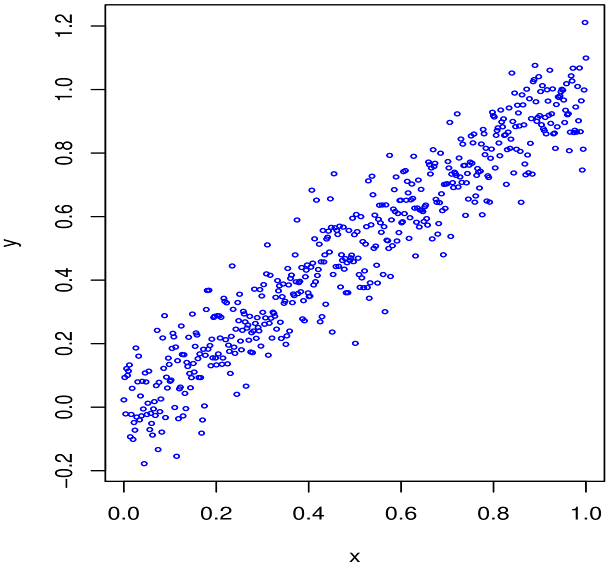
$r^2 = 0.491$



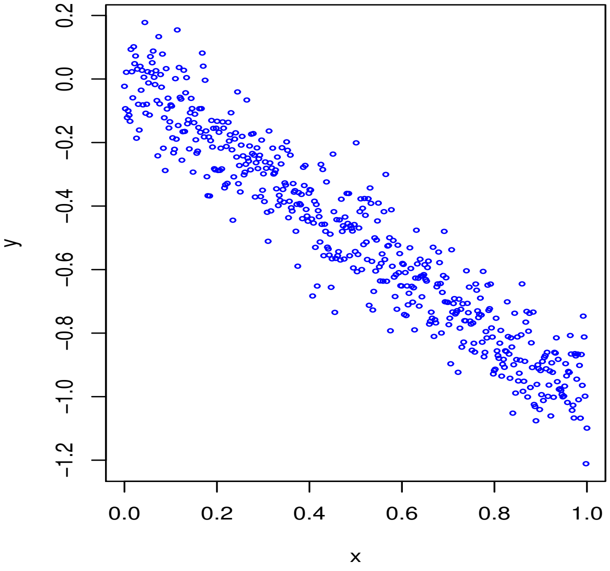
$r^2 = 0.491$

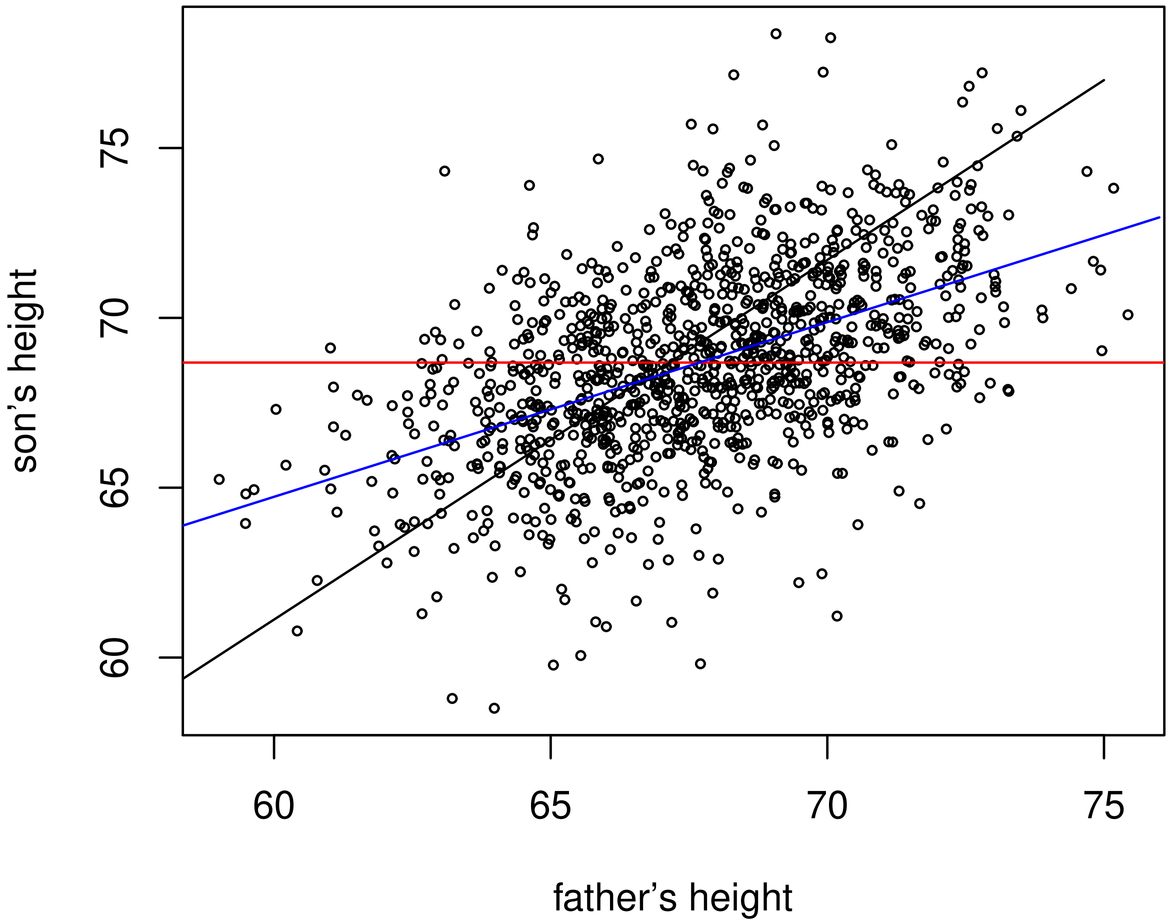


$r^2 = 0.908$



$r^2 = 0.908$





$R^2$ for Galton's data (Example 2) is 0.251. So, only about 25% of the total variation in sons' heights is explained by the linear relationship.

## Inference about $\beta_1$

Usually we're not satisfied with just having point estimates of parameters. We'd like to be able to say about how far off the point estimate could be. To this end we can construct a confidence interval and/or test a hypothesis.

First of all we consider the distribution of $\widehat{\beta}_1$.

Recall that: $\widehat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Properties of $\widehat{\beta}_1$:

1. $\widehat{\beta}_1$ is an unbiased estimator of $\beta_1$, i.e.

$$E(\widehat{\beta}_1) = \beta_1.$$

2. $\text{Var}(\widehat{\beta}_1) = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$.

3. $\widehat{\beta}_1$ is normally distributed.

## Proof of unbiasedness

The numerator of $\widehat{\beta}_1$ is

$$\sum_{i=1}^{n}(x_i - \bar{x})Y_i - \bar{Y}\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})Y_i.$$

Our model says the last quantity is

$$\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i) =$$

$$\beta_0 \textstyle\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) +$$
$$\textstyle\sum_{i=1}^{n} \epsilon_i(x_i - \bar{x}) =$$

$$\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n} \epsilon_i(x_i - \bar{x}).$$

Therefore,

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \epsilon_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

and so $E(\widehat{\beta}_1) = \beta_1$. *Why?*

Consider a standardized version of $\widehat{\beta}_1$:

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

The quantity

$$\widehat{\sigma}_{\widehat{\beta}_1} = \frac{\widehat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

estimates the standard error of $\widehat{\beta}_1$. *The probability distribution of $T$ is the $t$-distribution with $n-2$ degrees of freedom.*

The last fact leads to confidence intervals and tests for $\beta_1$.

A $(1-\alpha)100\%$ confidence interval for $\beta_1$ is

$$\widehat{\beta}_1 \pm t_{n-2;\alpha/2}\,\widehat{\sigma}_{\widehat{\beta}_1}.$$

Here $t_{n-2;\alpha/2}$ denotes $(1-\alpha/2)^{th}$ quantile of the $t_{n-2}$ distribution.

To test the hypothesis

$$H_0 : \beta_1 = \beta_{10},$$

use the test statistic

$$T = \frac{\widehat{\beta}_1 - \beta_{10}}{\widehat{\sigma}_{\widehat{\beta}_1}}.$$

The alternative could be any of $H_a : \beta_1 \neq \beta_{10}$, $H_a : \beta_1 > \beta_{10}$ or $H_a : \beta_1 < \beta_{10}$.

*The test is done exactly like the $t$-test for a population mean except using $n-2$ instead of $n-1$ degrees of freedom.* Rejection regions?

For two-sided $H_a$, reject $H_0$ at level $\alpha$ if $|T| > t_{n-2;\alpha/2}$. For a one-sided $H_a : \beta_1 > \beta_{10}$, or $H_a : \beta_1 < \beta_{10}$, reject $H_0$ at level $\alpha$ if $T > t_{n-2;\alpha}$, or if $T < -t_{n-2;\alpha}$, respectively.

Of particular interest is testing $H_0 : \beta_1 = 0$. *Why?*

## Prediction and inference for $E(Y)$

$x =$ father's height $\quad y =$ son's height

Two distinct problems:

- *A six foot tall father would like to predict the height of his unborn son Zach.*

- *Sir Francis Galton would like to estimate the average height of sons whose fathers are six feet tall.*

The first problem, called *prediction* is subject to more uncertainty than the second.

Suppose we *know* the average in the second problem. We would probably use this average as our prediction of Zach's height, but we wouldn't really expect this prediction to be right on the mark.

## *Inference for the mean of $Y$ at a given $x = x_0$*

$$E(Y) = \beta_0 + \beta_1 x_0$$

We estimate this by

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

Properties of the estimator:

1. $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is an unbiased estimator of $\beta_0 + \beta_1 x_0$.

2. Var$(\widehat{\beta}_0 + \widehat{\beta}_1 x_0)$ is
$$\sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

3. $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is normally distributed.

A $(1-\alpha)100\%$ confidence interval for $\beta_0+\beta_1 x_0$ is:

$$(\widehat{\beta}_0+\widehat{\beta}_1 x_0) \pm t_{n-2;\alpha/2} SE(x_0),$$

where

$$SE(x_0) = \hat{\sigma}\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]^{1/2}.$$

*Prediction intervals*

Predict the value of $Y$ for a subject whose $x$-value is $x_0$. Our model says

$$Y = \beta_0+\beta_1 x_0+\epsilon.$$

Suppose we guess $Y$ to be $\widehat{\beta}_0+\widehat{\beta}_1 x_0$. Then, the total error we committed in the process is:

$$\beta_0+\beta_1 x_0+\epsilon-(\widehat{\beta}_0+\widehat{\beta}_1 x_0) =$$

$$[(\beta_0+\beta_1 x_0)-(\widehat{\beta}_0+\widehat{\beta}_1 x_0)]+\epsilon.$$

$$[(\beta_0 + \beta_1 x_0) - (\widehat{\beta}_0 + \widehat{\beta}_1 x_0)] =$$

error due to estimating $\beta_0$ and $\beta_1$

$\epsilon =$deviation of $Y$ from $\beta_0 + \beta_1 x_0$

Suppose $\epsilon$ is independent of the data from which $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are computed. Then the variance of the prediction error is

$$\text{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) + \text{Var}(\epsilon) =$$

$$\sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) + \sigma^2 =$$

$$\sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right).$$

A $(1-\alpha)100\%$ prediction interval for $Y$ given $x = x_0$ is:

$$(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) \pm t_{n-2;\alpha/2} SE_{\text{pred}}(x_0),$$

where

$$SE_{\text{pred}}(x_0) = \hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^{1/2}.$$

*What is the correct interpretation of the prediction interval?*

Suppose, for example, that $\alpha = 0.05$. Then we are 95% confident that the future $Y$ value (corresponding to $x_0$) will be in the prediction interval.

Note: The prediction interval for $Y$ at $x$ will be wider than the corresponding confidence interval for the mean of $Y$ at $x$, for the same value of $x$. Makes sense (why?).

*Example 4:* *Using R to do inference for Galton's data*

The R command `summary(lm(y~x))` provides $\widehat{\beta}_1$ and its estimated standard error:

$$\widehat{\beta}_1 = 0.514 \quad \text{and} \quad \widehat{\sigma}_{\widehat{\beta}_1} = 0.027.$$

The $T$ for testing $H_0 : \beta_1 = 0$ is 19.01, and the associated $P$-value ($< 2 \cdot 10^{-16}$) is for the two-sided alternative.

So, there is strong evidence that the heights of fathers and their sons are positively related.

A 95% confidence interval for $\beta_1$ is:

$$0.514 \pm 1.96(0.027) = 0.514 \pm 0.05292,$$

or $(0.461, 0.567)$.

So we're 95% sure that the slope of the regression line is between 0.461 and 0.567. We can also get this interval using the R commands:

```
fit=lm(y~x)
confint(fit, level=0.95)
```

Now we do the following:

- *Find a 95% confidence interval for the average height of sons whose fathers are all 5 feet 8 inches tall*.

- *Predict Zach's height using an interval in which you have 98% confidence. (Zach's father is six feet tall.)*

In the first of these two problems $x_0 = 68$, and in the second $x_0 = 72$.

A description of how to solve both these problems in R is given on pg. 441-442 of your textbook. *This will be demonstrated in class.* Here are the R commands for getting these:

```
predict(lm(y~x), newdata=data.frame(x = 68),
interval="confidence", level=0.95)
```

The 95% confidence interval for the average is (68.70,68.99). So, we are 95% confident that the average height of sons whose fathers are 5’ 8” tall is between 68.70 and 68.99 inches.

```
predict(lm(y~x), newdata=data.frame(x = 72),
interval="prediction", level=0.98)
```

The 98% prediction interval for Zach’s height is (65.22,76.59). So, we’re 98% confident that Zach will be between 65.22 and 76.59 in. tall.

*It’s good to keep in mind that, at best, these inferences apply to a population of father-son pairs living in Galton’s time.*

## Correlation

In some cases, we want to know how *strongly* two variables are related without really identifying which of the two variables is the response and which is the predictor.

The (population) *correlation coefficient*, called $\rho$, is discussed in Section 4.5.2 of your text.

Given two random variables $X$ and $Y$ with some joint distribution and means $\mu_X$ and $\mu_Y$,

$$\rho \equiv \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}, \quad \text{where}$$

$$\sigma_X^2 = \mathrm{Var}(X), \quad \sigma_Y^2 = \mathrm{Var}(Y) \quad \text{and}$$

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Given data, we can estimate $\rho$. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed (i.i.d.) pairs of realizations of the random variables $(X, Y)$.

Real situation where this model holds:

> *Have a population of individuals. Each individual has an $x$ and a $y$ measurement. Take a random sample of individuals.*

How can we estimate $\rho = \text{Corr}(X_i, Y_i)$?

Estimate $\text{Cov}(X_i, Y_i)$ by

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}),$$

$\sigma_X^2$ by

$$\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

and $\sigma_Y^2$ by

$$\hat{\sigma}_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Now substitute these estimators for the parameters in the defn. of $\rho$ to get an estimator $\hat{\rho}$.

This yields

$$\widehat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)^{1/2}} = R.$$

This is called *Pearson's product moment correlation coefficient*, or simply the *sample correlation coefficient*.

The reason for calling $\widehat{\rho}$ also as $R$ is that $\widehat{\rho}^2$ also equals $R^2$, the coefficient of determination discussed on pg. 25N (i.e., pg. 25 of the notes).

$R$ (or equivalently, $\widehat{\rho}$) provides a point estimate of the $\rho$. The closer $R$ is to $\pm 1$, the stronger the relationship is between $X_i$ and $Y_i$. If $R \approx 0$, then it's safe to say that $\rho$ is close to 0.

Note also that the slope estimator $\widehat{\beta}_1 = R\frac{\widehat{\sigma}_Y}{\widehat{\sigma}_X}$.

*Unfortunately, $\rho = 0$ does not necessarily imply that $X_i$ and $Y_i$ are independent.* (Review Section 4.5.2 of textbook)

We may test the null hypothesis:

$$H_0 : \rho = 0$$

using the test statistic

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}.$$

When $H_0$ is true and $n$ is large, $T$ has an approximate standard normal distribution, i.e. $N(0,1)$. Tests are then done in the usual way.

The R command `cor.test(x,y)` provides the correlation coefficient between `x` and `y`, a 95% confidence interval for $\rho$ and a two-sided test of $H_0 : \rho = 0$ (at level 0.05).