Applied Probability II

Section 6: Linear Regression

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 6: Linear Regression

Section 6.1: Simple linear regression model

A model for the mean

In Section 4.1, we assumed that $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$.

If we collect a sample of data from this distribution, we can estimate μ using $\bar{y}$, the sample mean. We also described the sampling distribution of $\bar{Y}$, and used it to construct CIs and test hypotheses about μ.

We could also have referred to this as 'fitting a model for the mean' and written it as:

$$Y_i = \mu + \epsilon_i$$

where ϵ_i is a random variable which is normal with mean 0 and variance σ^2.

I.e., $E[\epsilon_i] = 0$, and $Var(\epsilon_i) = \sigma^2$, and $\epsilon_i \sim N(0, \sigma^2)$.

We can see that

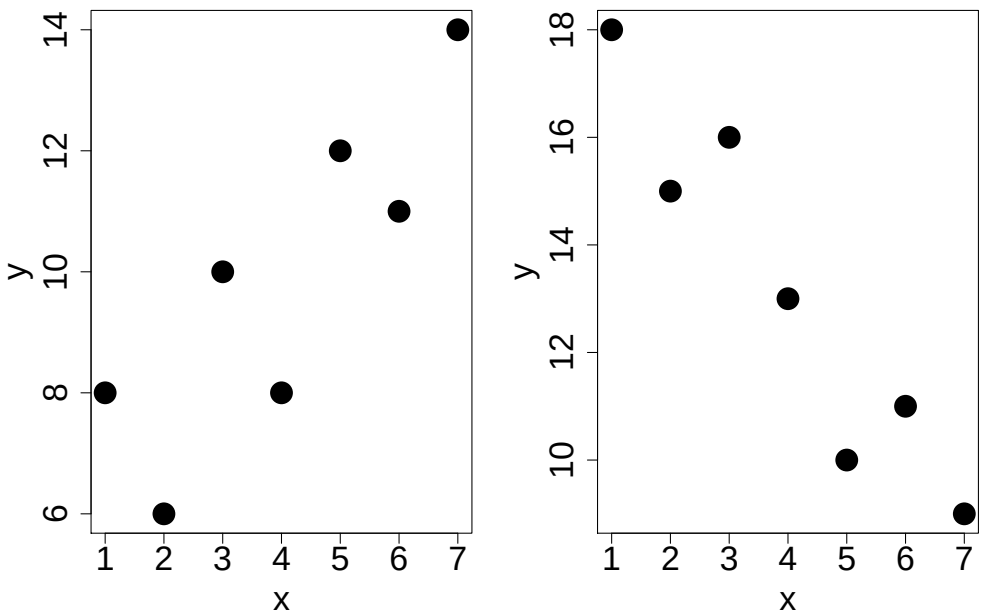
$$E[Y_i] = E[\mu + \epsilon_i] = \mu + E[\epsilon_i] = \mu$$
$$Var(Y_i) = Var(\mu + \epsilon_i) = Var(\epsilon_i) = \sigma^2$$

The ϵ_i terms are called the 'errors', i.e., how far away Y_i is from μ.

Simple linear regression motivation

Suppose now, in addition to observations $y_1, y_2, ..., y_n$, we have further data information $x_1, x_2, ...x_n$. We believe that knowing x can help us to predict y.

For example, a scatter plot could look like:



Simple linear regression equation

The simple linear regression model equation is

$$Y_i = \mu_{Y|x_i} + \epsilon_i$$

$$\mu_{Y|x_i} = E[Y|x_i] = \beta_0 + \beta_1 x_i$$

Or the more common way to express the simple linear regression model is:

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

Where

β_0: is the intercept parameter, the expected mean of Y when $x = 0$.

and

β_1: is the slope parameter, the change in the expected mean of Y for a one unit increase in x.

The simple linear regression model assumptions

For the model

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

We assume that

1. For a fixed x, Y is a normally distributed random variable with mean $\beta_0 + \beta_1 x$.
2. The variance of Y does not depend on x. I.e. $Var(Y|x) = Var(Y) = \sigma^2$.
3. The values of Y are independent (uncorrelated).
4. The model is linear: $E[Y|x] = \beta_0 + \beta_1 x$.

We can also express the assumptions in terms of the errors.

1. $E[\epsilon_i] = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ (and does not depend on i).
3. ϵ_i are independent.
4. $\epsilon_i \sim N(0, \sigma^2)$.

Predicting and residuals

To predict from a fitted simple linear regression model for any x value, we use:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{\beta}_0$ is the estimated intercept, and $\hat{\beta}_1$ is the estimated slope.

Side note on terminology: $\hat{\beta}_0$, β_0, $\hat{\beta}_1$, β_1. What's the difference?

To assess the model fit and assumptions, we can use residuals, where residuals are defined as

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

If we have fitted a simple linear regression model to n paired (x_i, y_i) data values, then we will have n observed values (y_i), n corresponding fitted or predicted values $(\hat{y}_i)$ and can find n corresponding residuals (observed minus predicted).

Simple linear regression example

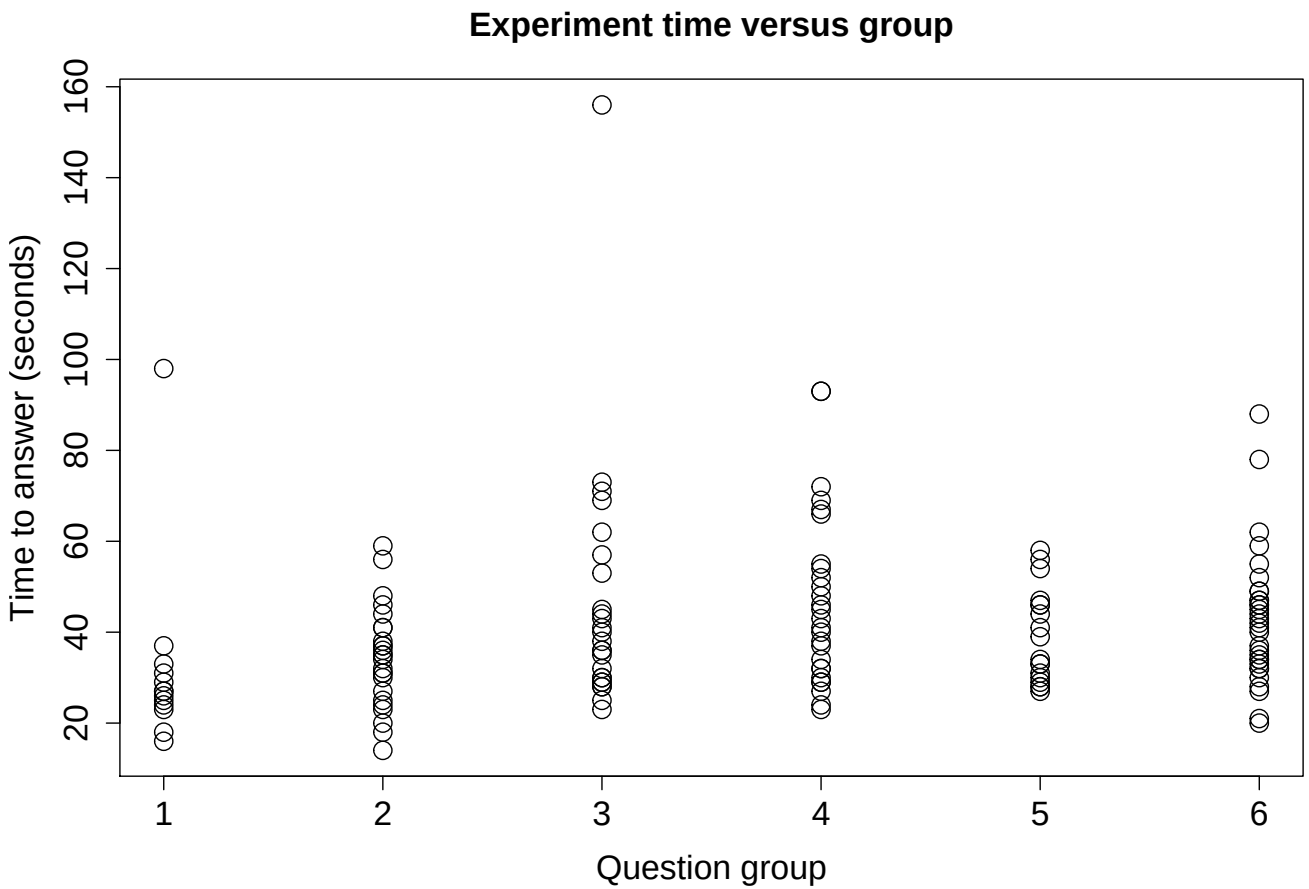
Remember the class experiment we did on the first day of term?

- Each person was asked just one question, but there were six different questions.
- Each question required you to put a list of words in alphabetical order; they were:
  - bouncing handle kitchen tracksuit university
  - **w**ashing **w**eird **w**hich **w**isdom **w**onderful
  - **mo**bile **mo**del **mo**ment **mo**untain **mo**vie
  - **sta**ck **sta**ke **sta**ndard **sta**pler **sta**tistics
  - **star**board **star**dom **star**fish **star**ry **star**tle
  - **cross**bar **cross**fire **cross**ing **cross**road **cross**word

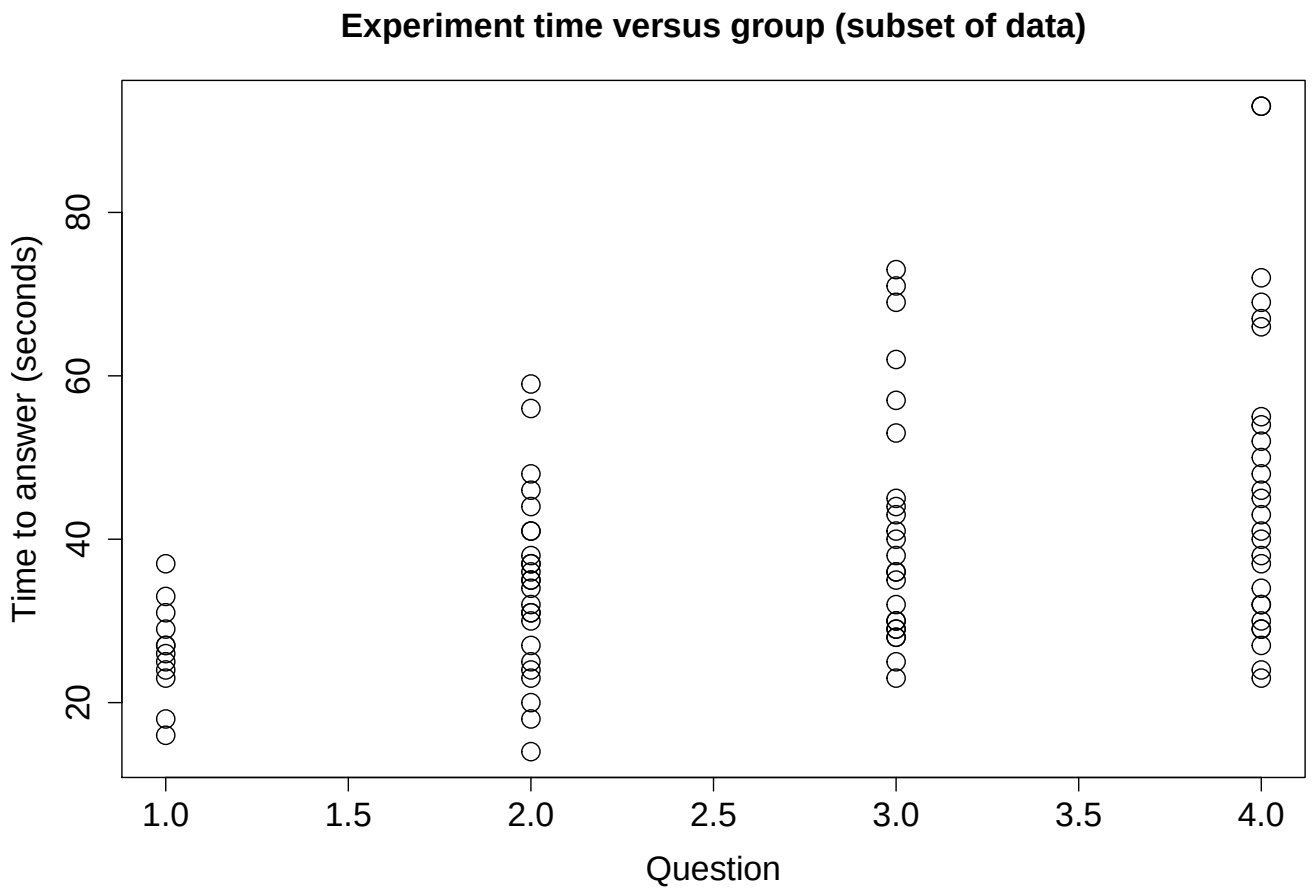
The first few rows of data:

```
##   Question Seconds Correct
## 1        1      16       1
## 2        1      18       1
## 3        1      23       1
## 4        1      24       1
## 5        1      25       1
## 6        1      26       1
```

A graph of the raw data



Let's focus in on the first four groups and omit outliers



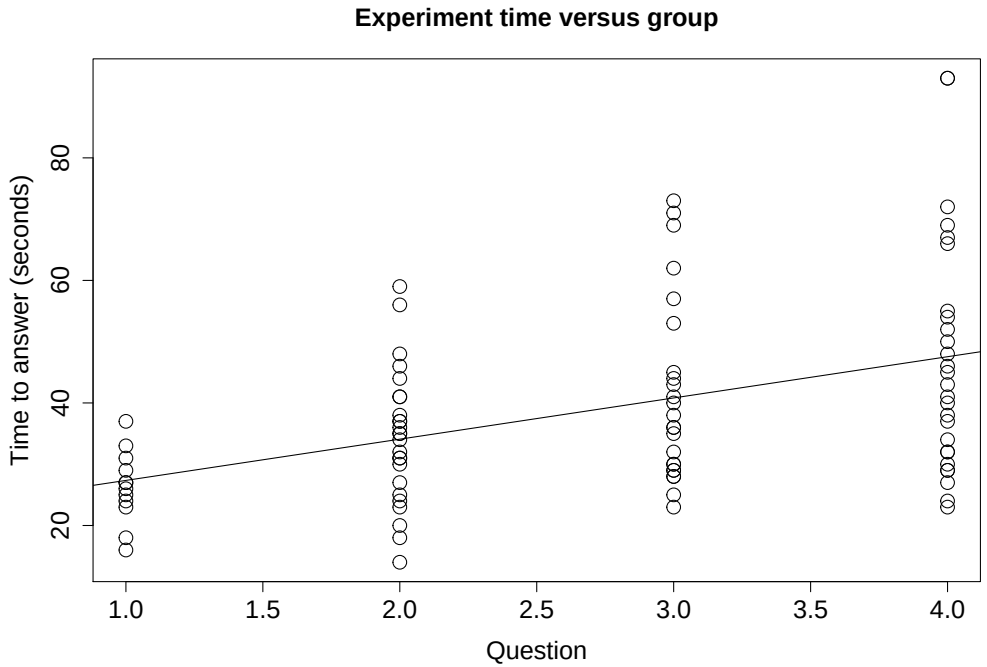
(We will come back to the full dataset during the lab session next week!)

Fit a simple linear regression model

```
lm1 <- lm(Seconds ~ Question, data = exp_subset)
summary(lm1)
```

```
##
## Call:
## lm(formula = Seconds ~ Question, data = exp_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.549 -10.617  -1.819   5.844  45.451
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   20.628      4.390   4.699 0.00000985 ***
## Question       6.730      1.494   4.505 0.00002075 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.52 on 86 degrees of freedom
## Multiple R-squared:  0.1909, Adjusted R-squared:  0.1815
## F-statistic:  20.3 on 1 and 86 DF,  p-value: 0.00002075
```

Fit a simple linear regression model

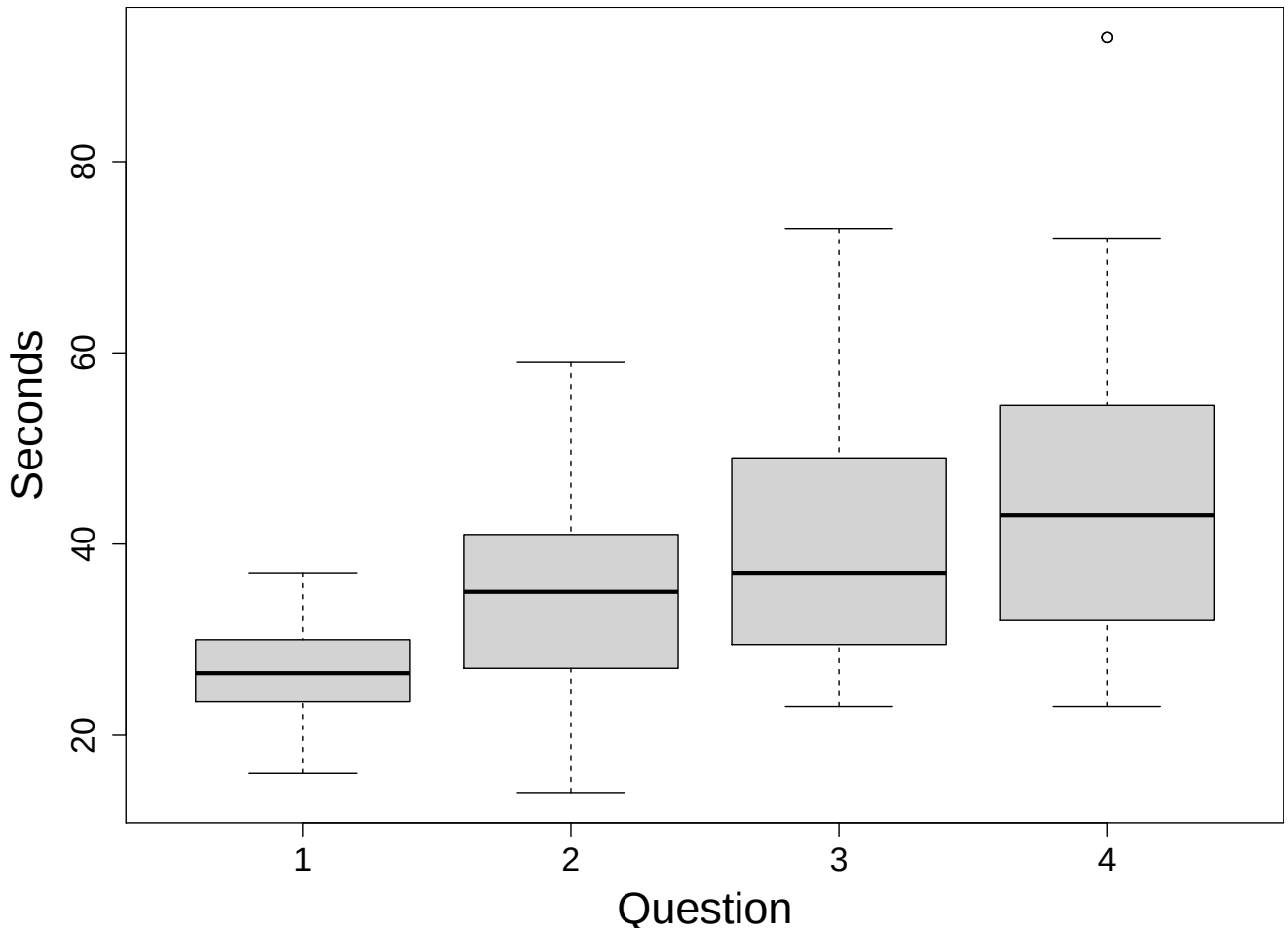


The equation of the line is: $\hat{y} = 20.63 + 6.73x$.

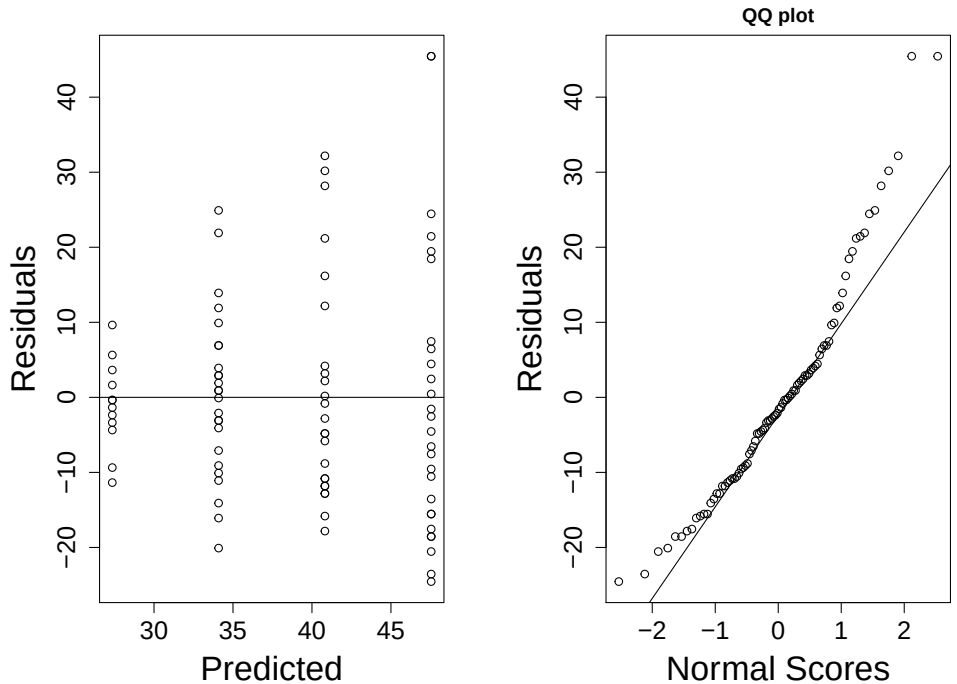
Intercept: the estimated average length of time to put the words in alphabetical order when question $= 0$ is 20.63 seconds.

Slope: the estimated average length of time to put the words in alphabetical order increases by 6.73 seconds for each unit increase in question.

Assess the model



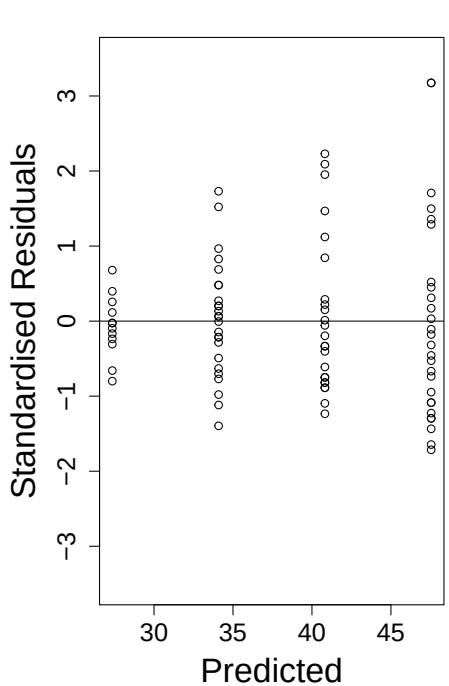
Assess the model (residuals)

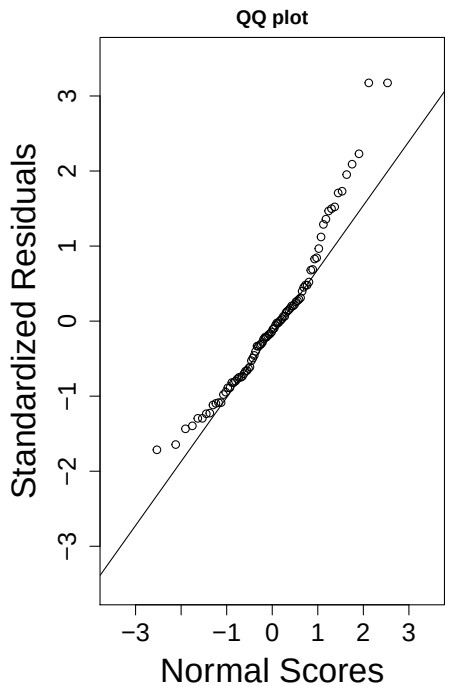


Assumptions:

1. $E[\epsilon_i] = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ (and does not depend on i).
3. ϵ_i are independent.
4. $\epsilon_i \sim N(0, \sigma^2)$.

Assess the model (standardised residuals)





Comparing models

It is often useful to think about the connections between different models.

The model for the mean is:

$$Y_i = \mu + \epsilon_i \qquad \text{with } \epsilon_i \sim N(0, \sigma^2) \text{ and IID.}$$

The simple linear regression model (SLR) is:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{with } \epsilon_i \sim N(0, \sigma^2) \text{ and IID.}$$

We can see that the mean model is a special case of the SLR model.

In fact, if we set $\beta_1 = 0$ in the SLR model, we get the mean model.

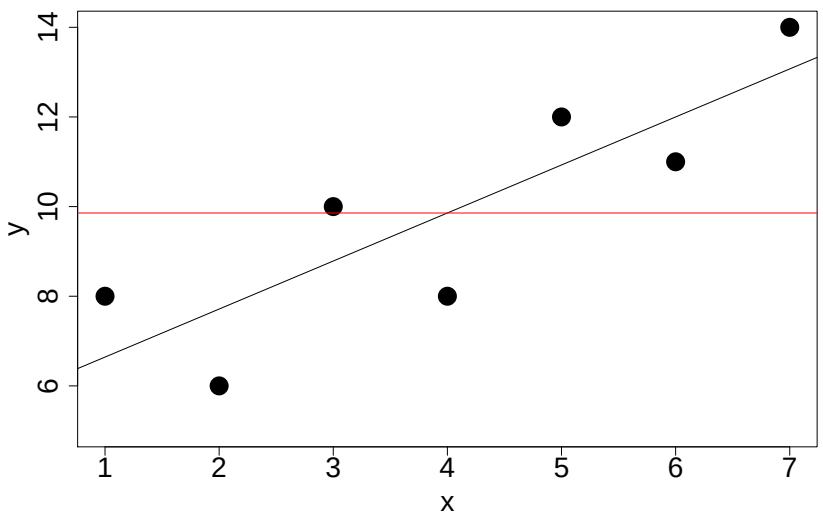
We will often be interested to test whether or not $\beta_1 = 0$. (Why?)

Comparing models visually

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

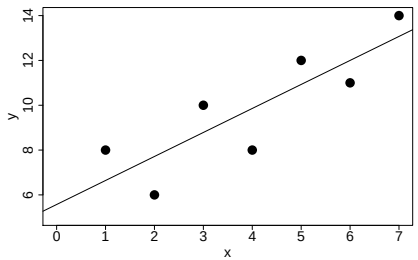
```
x <- c(1, 2, 3, 4, 5, 6, 7)
y <- c(8, 6, 10, 8, 12, 11, 14)
mean(y)
```

```
## [1] 9.857143
```



Note: $Y_i = \mu + \epsilon_i$ and $Y_i = \beta_0 + \epsilon_i$ are the same models, just different notation!

Simple linear regression in R




```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       1       2       3       4       5       6       7
##  1.3571 -1.7143  1.2143 -1.8571  1.0714 -1.0000  0.9286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5714     1.3477   4.134  0.00905 **
## x             1.0714     0.3014   3.555  0.01629 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.595 on 5 degrees of freedom
## Multiple R-squared:  0.7166, Adjusted R-squared:  0.6599
## F-statistic: 12.64 on 1 and 5 DF,  p-value: 0.01629
```

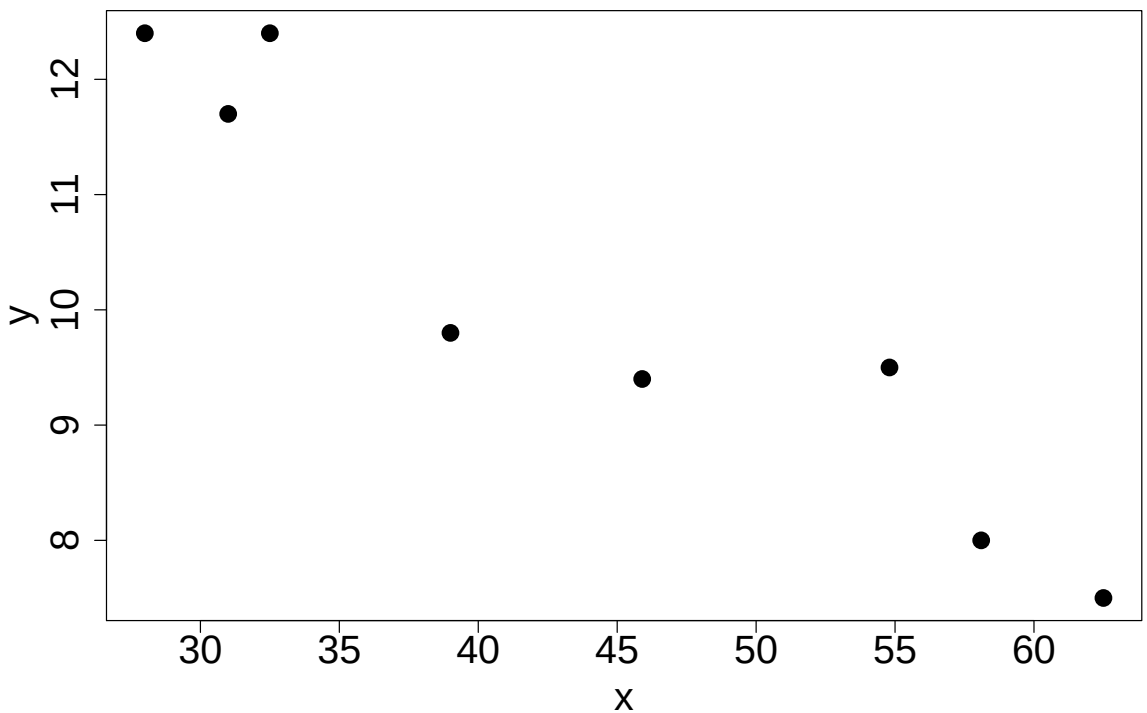
In the next sections, we will look at how to estimate the SLR parameters, and the underlying theory for conducting tests of hypothesis and condidence intervals for them.

Section 6.2: Least squares estimation

The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

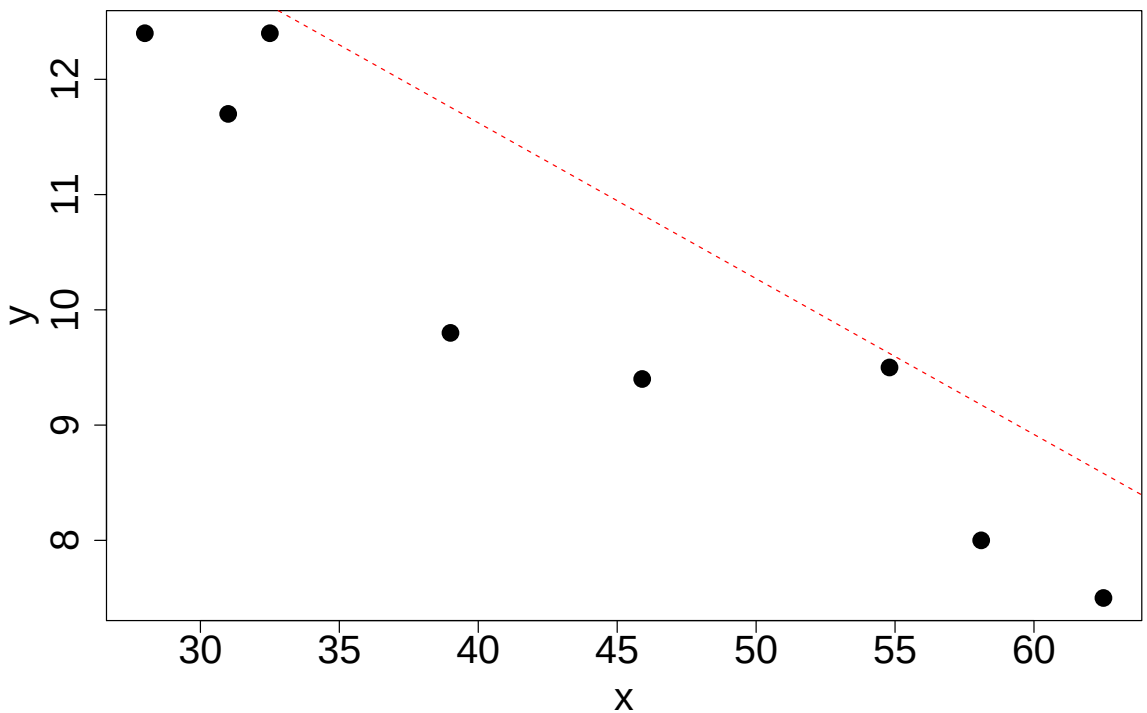
The method of ordinary least squares minimises the squared deviations from the line.



The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

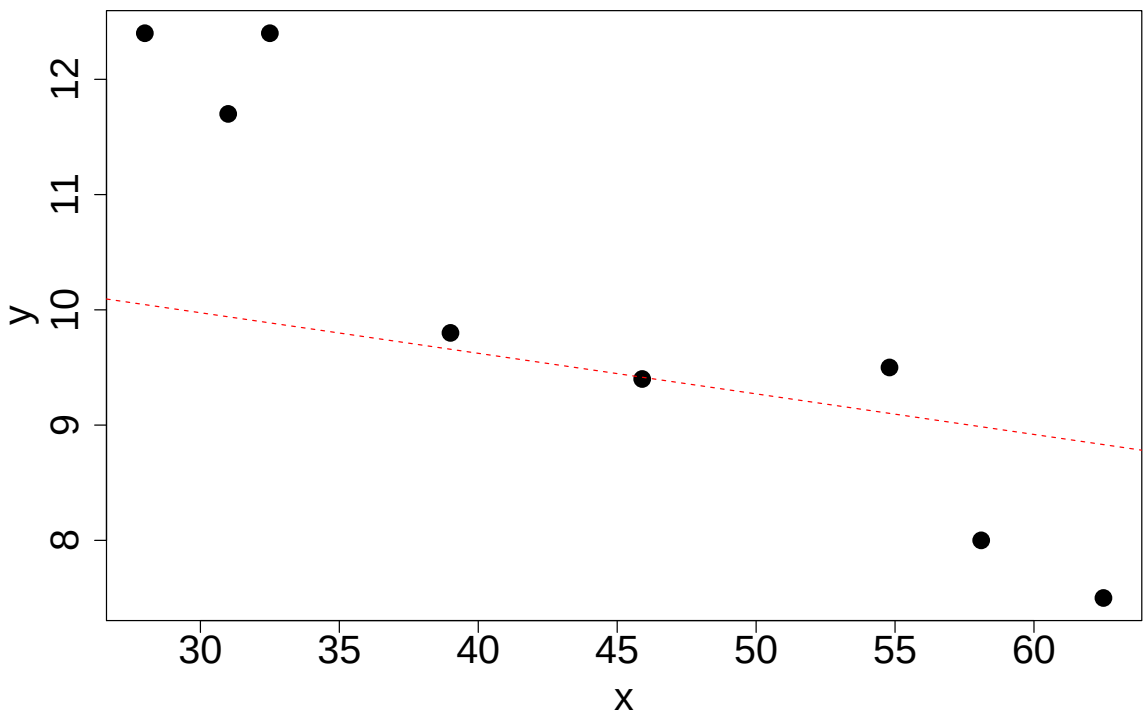
The method of ordinary least squares minimises the squared deviations from the line.



The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

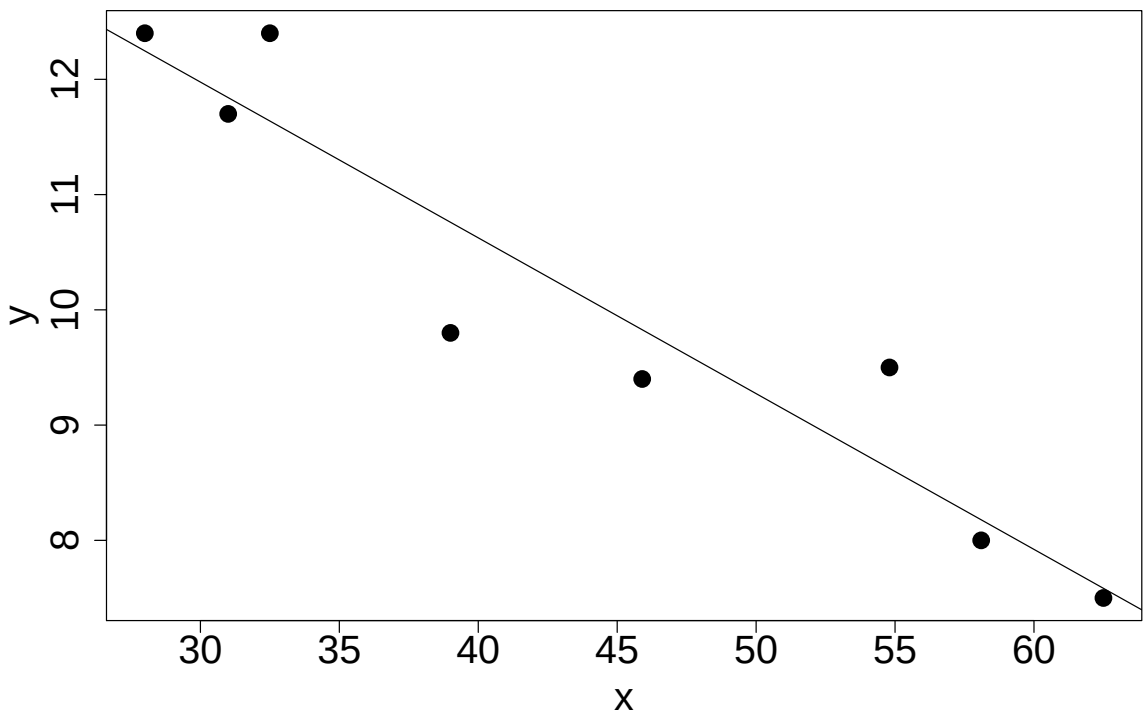
The method of ordinary least squares minimises the squared deviations from the line.



The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

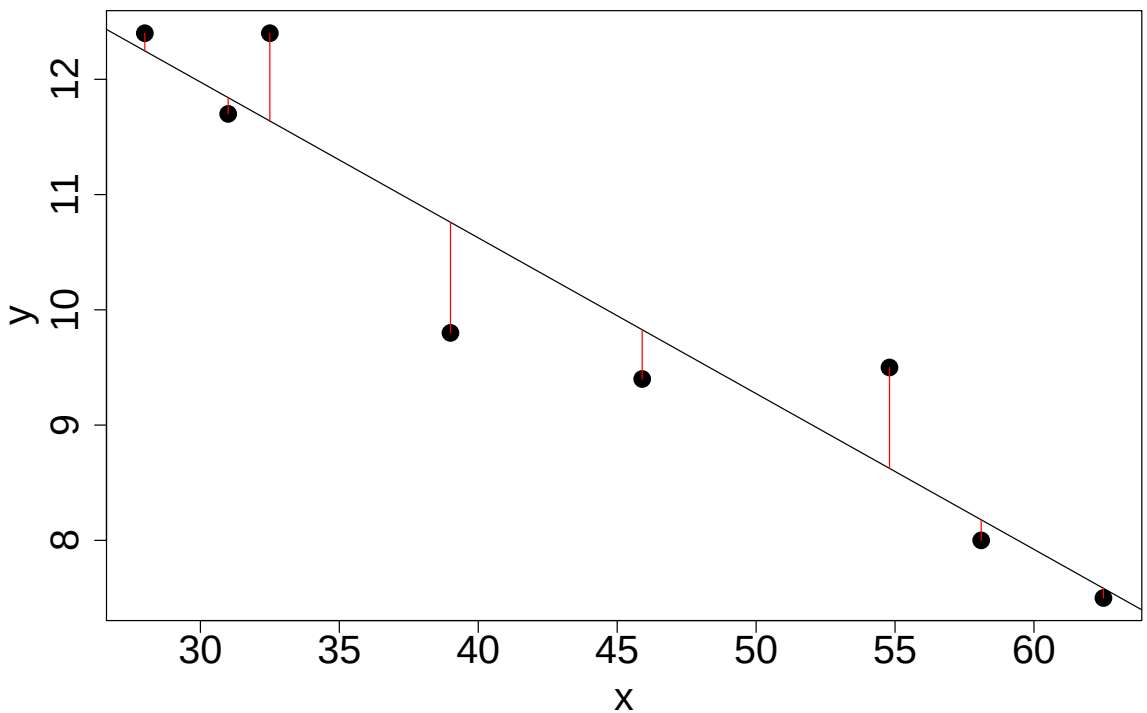
The method of ordinary least squares minimises the squared deviations from the line.



The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

The method of ordinary least squares minimises the squared deviations from the line.



Deriving the least squares estimates

Recall: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Define S, the sum of squared errors:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ must satisfy: $\frac{\delta S}{\delta \beta_0} = 0$ and $\frac{\delta S}{\delta \beta_1} = 0$.

$$\frac{\delta S}{\delta \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\delta S}{\delta \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Deriving the least squares estimates (contd)

Taking the derivative with respect to β_0:

$$\frac{\delta S}{\delta \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

and setting it equal to 0 at $\hat{\beta}_0$ gives:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

And we can get:

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

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

Deriving the least squares estimates (contd)

Taking the derivative with respect to β_1

$$\frac{\delta S}{\delta \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

and setting it equal to 0 at $\hat{\beta}_1$ gives:

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Substituting $\hat{\beta}_0$ in (where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$):

$$\sum_{i=1}^{n} x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = 0$$

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

Rearranging gives:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Some notation:

$$S_{xx} = \sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

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

$$S_{xy} = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

So, the equation of the ordinary least squares (OLS) fitted line is given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

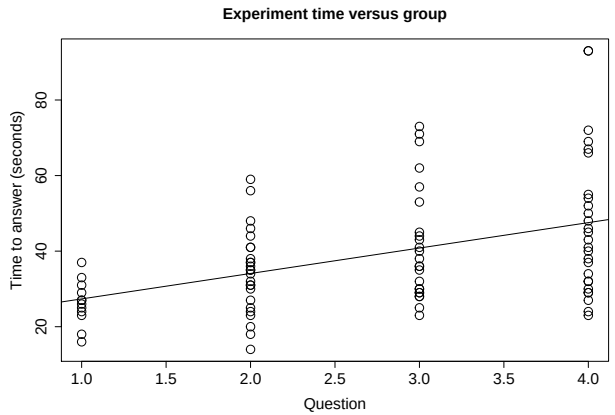
where

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

and

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

Back to our class experiment data example



The equation of the line is: $\hat{y} = 20.63 + 6.73x$. Let's confirm these values.

Some summary values: $n = 88$, $\sum x_i = 242$, $\sum y_i = 3444$, $\bar{x} = 2.75$, $\bar{y} = 39.1363636$, $\sum x_i^2 = 760$, $\sum y_i^2 = 157202$, $\sum x_i y_i = 10107$.

The slope

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{636}{94.5} = 6.73$$

The intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 39.1364 - 6.73 * 2.75 = 20.63$$

Section 6.3: Parameter inference

Sampling distributions of OLS estimators

When the simple linear regression model assumptions hold, then

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

are the sampling distributions of the least squares estimators.

We can think of $\hat{\beta}_1$ and $\hat{\beta}_0$ as being random variables in the context of any data collected from the population of interest.

In this section, we will prove these sampling distributions. This will allow us to perform tests of hypotheses and construct confidence intervals for the population parameters β_0 and β_1.

Sampling distribution of $\hat{\beta}_1$: Expected value

Recall that $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, and $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim$ IID $N(0, \sigma^2)$. Want: $E[\hat{\beta}_1] = \beta_1$.

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n a_i Y_i$$

where the a_i values depend only on x and are NOT random. By linearity of expectation:

$$E[\hat{\beta}_1] = \sum_{i=1}^n a_i E[Y_i] = \sum_{i=1}^n a_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n a_i + \beta_1 \sum_{i=1}^n a_i x_i = \beta_1$$

Since: $$\sum_{i=1}^n a_i = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0, \text{ and}$$

$$\sum_{i=1}^n a_i x_i = \frac{\sum_{i=1}^n x_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{S_{xx}} = \frac{S_{xx}}{S_{xx}} = 1, \text{ giving } E[\hat{\beta}_1] = \beta_1 \text{ as required.}$$

Sampling distribution of $\hat{\beta}_1$: Variance

We now want to show that $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$.

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 Var(Y_i) \text{ (since } Y_i\text{s are independent)} \\
&= \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 \\
&= \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \\
&= \sigma^2 \frac{S_{xx}}{(S_{xx})^2} = \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

Finally, the normality assumption follows as $\hat{\beta}_1$ is a linear combination of normal random variables (Y_is). So we have proven that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Sampling distribution of $\hat{\beta}_0$: Expected value

Recall that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, and $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim$ IID $N(0, \sigma^2)$, and $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$.

Want to show that $E[\hat{\beta}_0] = \beta_0$.

$$
\begin{aligned}
E[\hat{\beta}_0] &= E[\bar{Y} - \hat{\beta}_1 \bar{x}] = E[\bar{Y}] - \beta_1 \bar{x} \\
&= \frac{1}{n} \sum_{i=1}^{n} E[Y_i] - \beta_1 \bar{x} = \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} \\
&= \frac{1}{n} (n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i) - \beta_1 \bar{x} \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0
\end{aligned}
$$

Sampling distribution of $\hat{\beta}_0$: Variance

We now want to show that $\text{Var}(\hat{\beta}_1) = = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$.

$$\begin{aligned}\text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{Y} - \hat{\beta}_1 \bar{x}) \\ &= \text{Var}(\bar{Y}) + \bar{x}^2 \text{Var}(\hat{\beta}_1) - 2\bar{x}\text{Cov}(\bar{Y}, \hat{\beta}_1)\end{aligned}$$

Aside: $(\text{Cov}(aU + bV, cY + dZ) = ac\text{Cov}(U, Y) + bc\text{Cov}(V, Y) + ad\text{Cov}(U, Z) + bd\text{Cov}(V, Z))$.

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = \text{Cov}\left(\frac{1}{n} \sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} a_i Y_i \right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{n} a_i \text{Cov}(Y_i, Y_j) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \text{Cov}(Y_i, Y_j)$$

$$= \frac{1}{n} \sum_{i=1}^{n} a_i \text{Cov}(Y_i, Y_i) \text{ (since } Y_i \text{ are indep.)}$$

$$= \frac{\sigma^2}{n} \sum_{i=1}^{n} a_i = \frac{\sigma^2}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0$$

Sampling distribution of $\hat{\beta}_0$: Variance (contd)

Back to:

$$\begin{aligned}\text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{Y} - \hat{\beta}_1\bar{x}) \\ &= \text{Var}(\bar{Y}) + \bar{x}^2\text{Var}(\hat{\beta}_1) - 2\bar{x}\text{Cov}(\bar{Y}, \hat{\beta}_1)\end{aligned}$$

$$\begin{aligned}\text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{Y}) + \bar{x}^2\text{Var}(\hat{\beta}_1) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i) + \bar{x}^2\frac{\sigma^2}{S_{xx}} \\ &= \frac{1}{n^2}n\sigma^2 + \bar{x}^2\frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\end{aligned}$$

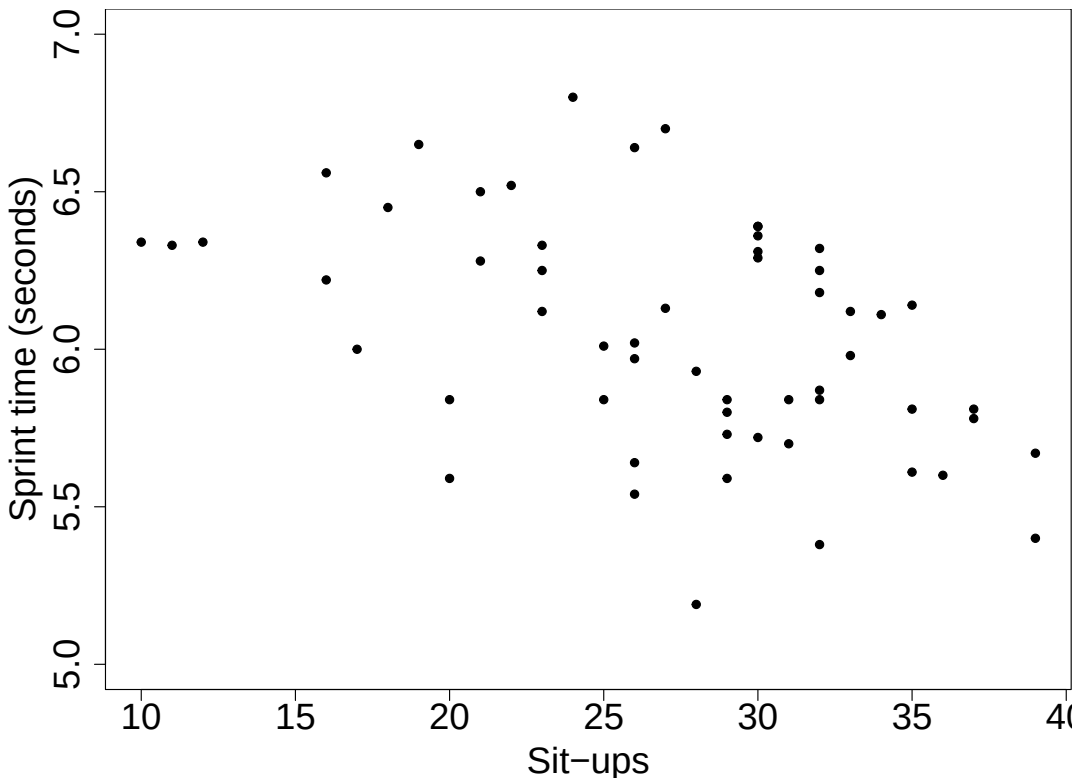
Finally, the normality assumption follows as $\hat{\beta}_0$ is a linear combination of normal random variables (Y_i values and $\hat{\beta}_1$). So we have proven that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

Athlete Example

Is there a relationship between how many sit-ups you can do and how fast you can sprint 40 yards? A study set up to examine this recorded details on 57 randomly selected female athletes.

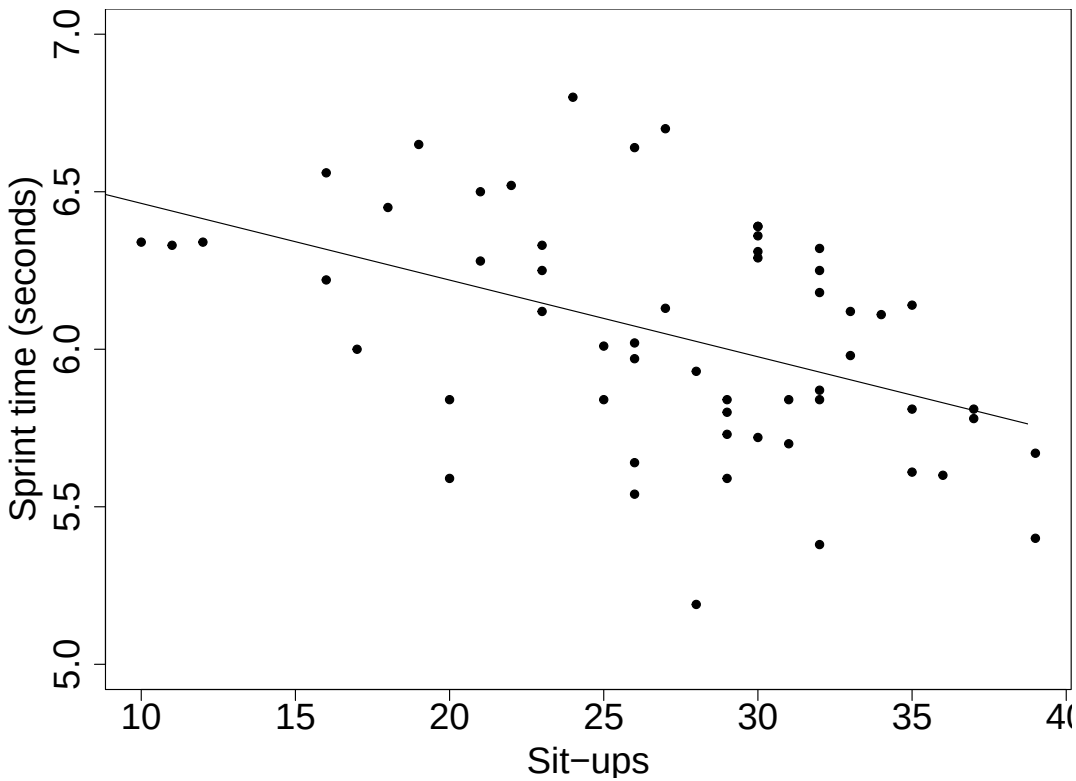
Here is a scatter plot of the data:



Athlete Example

Is there a relationship between how many sit-ups you can do and how fast you can sprint 40 yards? A study set up to examine this recorded details on 57 randomly selected female athletes.

Here is a scatter plot of the data with the linear regression line fitted:



Athlete Example

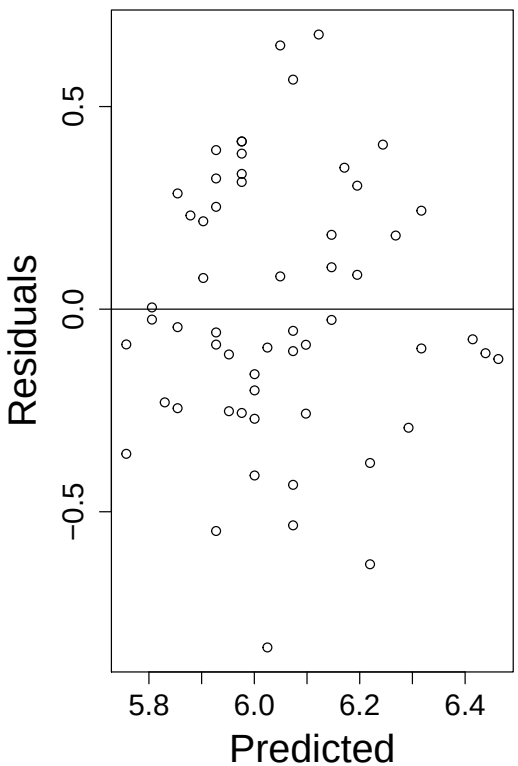
Let's fit the simple linear regression model and find confidence intervals for the parameters.

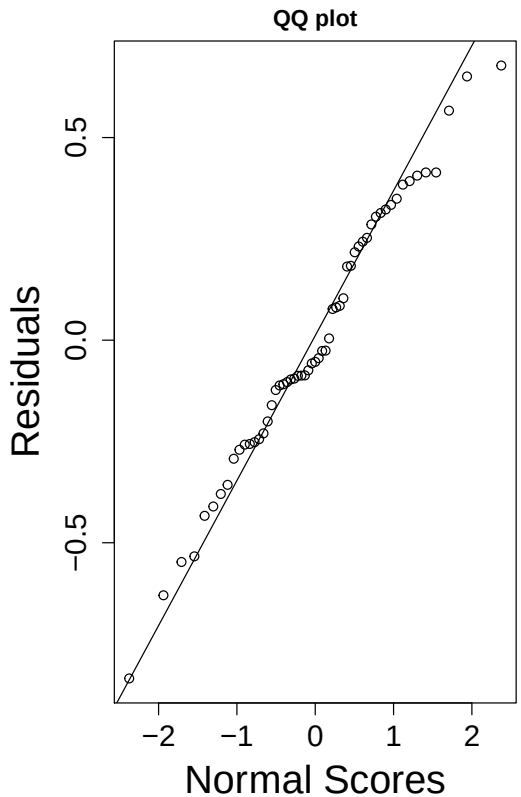
```
##
## Call:
## lm(formula = Sprint ~ Situp, data = SS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83484 -0.23007 -0.05353  0.25255  0.67778
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  6.706527   0.177891  37.700 < 0.0000000000000002 ***
## Situp       -0.024346   0.006349  -3.835             0.000326 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3272 on 55 degrees of freedom
## Multiple R-squared:  0.211,  Adjusted R-squared:  0.1966
## F-statistic: 14.71 on 1 and 55 DF,  p-value: 0.0003257
```

```
##               2.5 %  97.5 %
## (Intercept)  6.3500  7.0630
## Situp       -0.0371 -0.0116
```

Athlete Example

Testing the model assumptions:





Confidence interval for β_1

What theory allows us to construct a confidence interval for β_1?

We know that $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$, or equivalently $\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$.

And, when we replace σ by $\hat{\sigma}$ we have:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}.$$

The degrees of freedom (df) are $n-2$ because we have estimated two parameters (β_0 and β_1).

We estimate σ^2 by the mean squared error (MSE) by calculating:

$$\hat{\sigma}^2 = MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

A $(1-\alpha)\times 100\%$ confidence interval for β_1 is:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \times \sqrt{\frac{MSE}{S_{xx}}}$$

Hypothesis tests for β_1

We may wish to test

$H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$.

The null hypothesis here is that $E[Y] = \beta_0$, i.e., $E[Y]$ is not linearly related to x.

Under H_0 (i.e., if the null hypothesis is true):

$$t_{obs} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}.$$

Using a specified α level, we reject H_0 for extreme values of t_{obs}.

We could also test $H_0 : \beta_1 = b$ by computing:

$$\frac{\hat{\beta}_1 - b}{\text{S.E.}(\hat{\beta}_1)}.$$

where $b \neq 0$.

Back to the athlete example

Here are the model estimates and the confidence intervals:

```
##                Estimate  Std. Error
## (Intercept)  6.70652722 0.177891008
## Situp       -0.02434606 0.006348777

##                2.5 %  97.5 %
## (Intercept)  6.3500  7.0630
## Situp       -0.0371 -0.0116
```

The 95% confidence interval for β_1 is

$$\begin{aligned}\hat{\beta}_1 \pm t_{n-2,\alpha/2} \times \sqrt{\frac{MSE}{S_{xx}}} \\ = -0.02434606 \pm 2.0040448 \times 0.006348777 \\ = (-0.0371, -0.0116)\end{aligned}$$

The athlete example

Here are the model estimates and hypothesis tests:

```
##                Estimate  Std. Error   t value
## (Intercept)  6.70652722 0.177891008 37.700204
## Situp       -0.02434606 0.006348777 -3.834764
##                                                      Pr(>|t|)
## (Intercept) 0.00000000000000000000000000000000000000005563697
## Situp       0.00032571993194159000982457197181929586804471910
```

Now we'll look at the hypothesis test for β_1, using $\alpha = 0.05$.

$H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$.

$$t_{obs} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{-0.02434606}{0.006348777} = -3.835$$

If the H_0 is true, then, t_{obs} is a random draw from a $t(55)$ distribution.

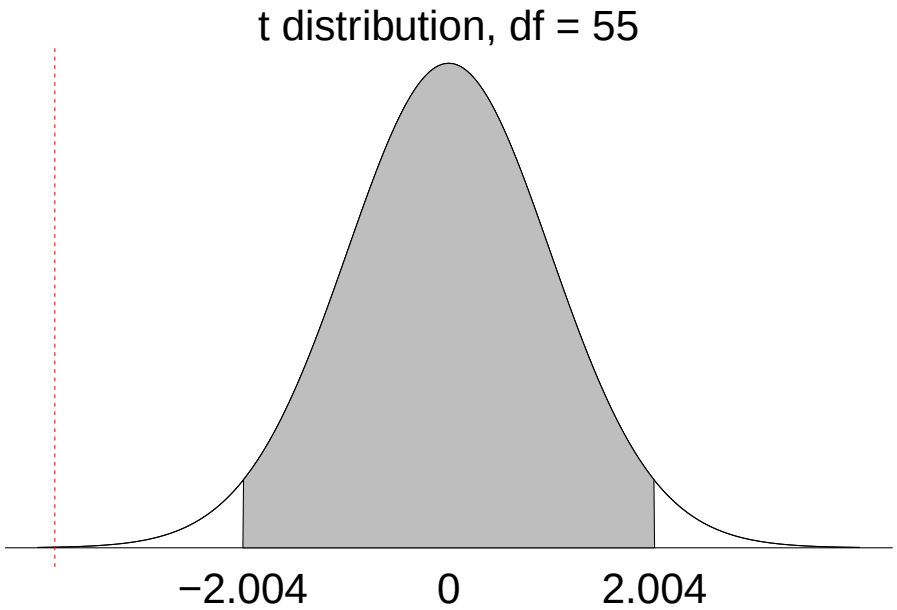
The critical values are -2.004 and 2.004. We reject H_0 if our observed test statistic is lower than -2.004, or greater than 2.004.

Let's take a look at this graphically...

The athlete example

Evaluate the hypothesis test using critical values.

The test statistic = -3.835, and is shown by the red line:

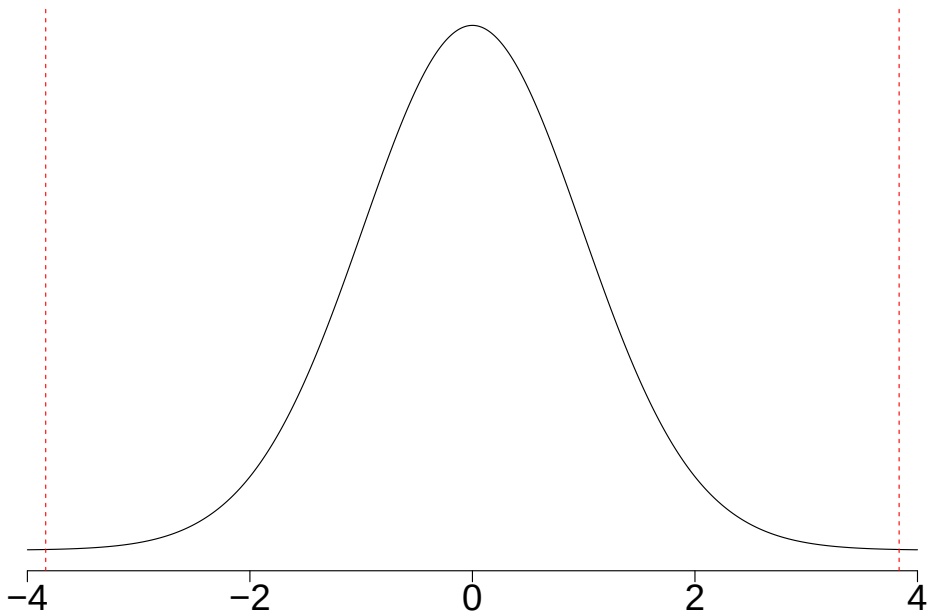


We reject the H_0, using $\alpha = 0.05$, and conclude that $\beta_1 \neq 0$. We have evidence that the true mean sprinting time for 40 yards is linearly related to the number of sit-ups that female athletes can do.

The athlete example

We can also evaluate the hypothesis test using the p-value. We find the probability of observing a test statistic as extreme, or more extreme than what we observed.

Here, p-value $= P(T(55) \geq |t_{obs}|) = 0.0003$.



As before, we reject the H_0, using $\alpha = 0.05$.

Confidence interval for β_0

What theory allows us to construct a confidence interval for β_0?

We know that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

We use this to construct a 95% C.I. for β_0 as:

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} \times \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

Hypothesis for β_0

If we wish to test for a particular value of β_0, we can perform a hypothesis test:

$H_0 : \beta_0 = 0$ vs. $H_A : \beta_0 \neq 0$

The null hypothesis here is that $E[y] = \beta_1 x$, i.e., the line passes through the origin.

Under the H_0, the test statistic

$$t_{obs} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t(n-2)$$

where $\hat{\sigma}^2$ is estimated by the MSE, as before.

Using a specified α level, we reject H_0 for extreme values of t_{obs}.

Back to the athlete example

Here are the model estimates, hypothesis tests and confidence intervals:

```
##                Estimate  Std. Error   t value
## (Intercept)  6.70652722 0.177891008 37.700204
## Situp       -0.02434606 0.006348777 -3.834764
##                                                        Pr(>|t|)
## (Intercept) 0.00000000000000000000000000000000000000005563697
## Situp       0.00032571993194159000982457197181929586804471910

##                2.5 %  97.5 %
## (Intercept)  6.3500  7.0630
## Situp       -0.0371 -0.0116
```

The 95% confidence interval for $\beta_0 = (6.350, 7.063)$.

For the hypothesis test, the test statistic $= 37.7$, with p-value < 0.0001. Using $\alpha = 0.05$, we reject the null hypothesis and conclude that $\beta_0 \neq 0$.

In practice, are we interested in the confidence interval and hypothesis test for the intercept in this example?

Section 6.4: Confidence intervals and prediction intervals

Using slr models: Confidence intervals and prediction intervals

Once we have estimated a simple linear regression model, there are two different ways to think about using it:

- estimating the mean of Y for a given x value,
- predicting Y for a new observation with a given x value.

These two scenarios are conceptually different; they yield the same point estimate or prediction, however, their standard errors are different. We can construct a confidence interval or a prediction interval using the respective standard errors.

For the athlete data with $x_0 = 30$, we get $\hat{y} = 5.976$, with:

Confidence interval

```
##        fit      lwr      upr
## 1 5.976145 5.882149 6.070141
```

Prediction interval

```
##        fit      lwr      upr
## 1 5.976145 5.313703 6.638588
```

Let's take a look at how to construct these.

Confidence intervals

Suppose we want to estimate $\mu = E[y]$ at a particular value of x.

At x_0 let:

$$\mu_0 = E[y_0] = \beta_0 + \beta_1 x_0$$

We can estimate μ_0 by:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

For the athlete example, let $x_0 = 30$, then

$$\hat{y}_0 = 6.7065 - 0.02435 \times 30 = 5.976$$

A 95% confidence interval for this estimate is:

$$\hat{y}_0 \pm t_{n-2,\alpha/2}\ \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$= 5.976 \pm 2.004 \times 0.3272 \times \sqrt{\frac{1}{57} + \frac{(30 - 27.17544)^2}{2656.246}}$$

$$= (5.88, 6.07)$$

Prediction intervals

Now, let's assume that we have a new female athlete that can do 30 sit-ups. What time would we predict for their sprint and how certain would we be of the prediction?

The point estimate prediction is:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 6.7065 - 0.02435 \times 30 = 5.976$$

This is the same as before.

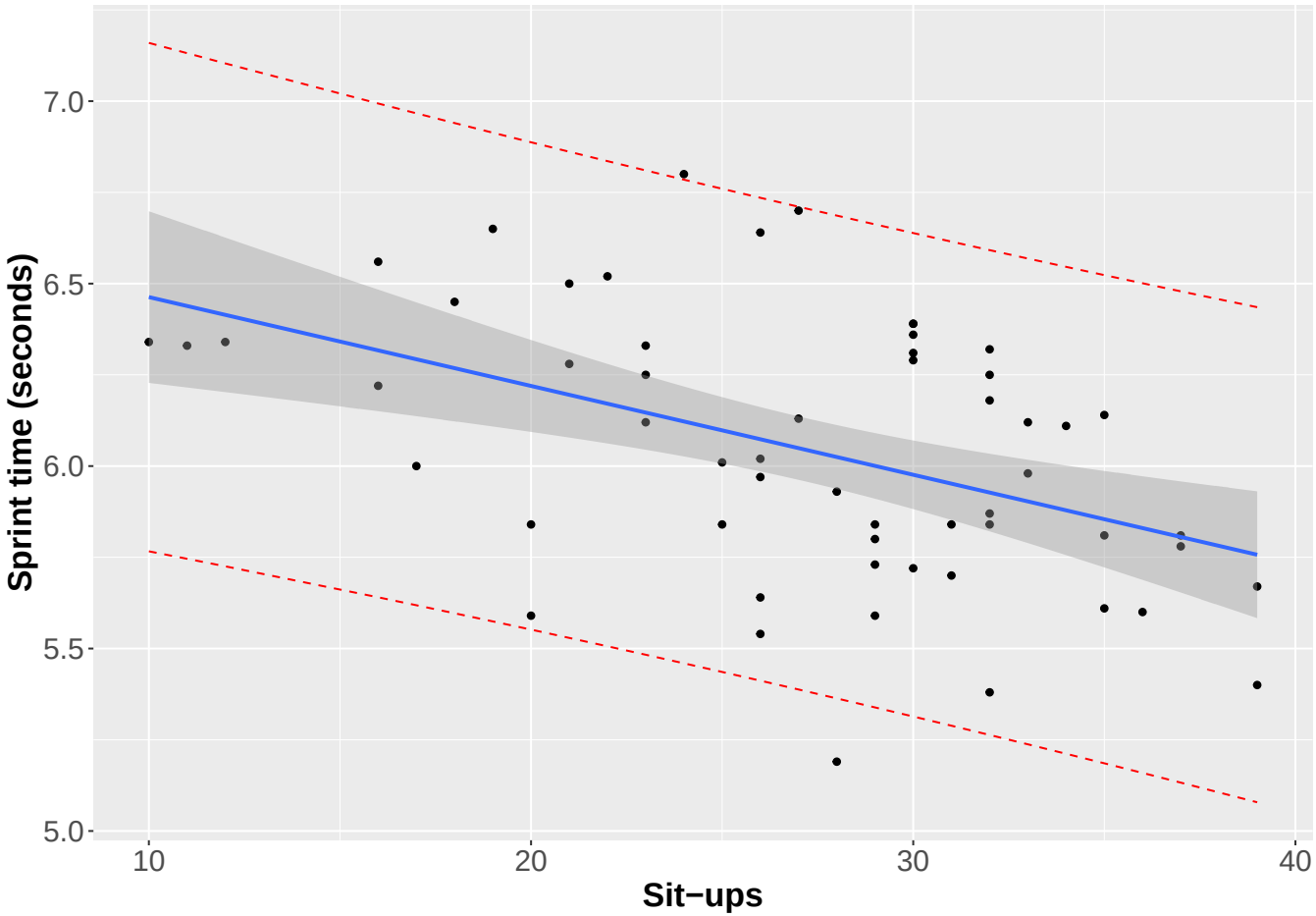
A 95% prediction interval is:

$$\hat{y}_0 \pm t_{n-2,\alpha/2}\ \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$= 5.976 \pm 2.004 \times 0.3272 \times \sqrt{1 + \frac{1}{57} + \frac{(30 - 27.17544)^2}{2656.246}}$$

$$= (5.31, 6.64)$$

Confidence intervals and prediction intervals graphically



The blue line shows the fitted regression line, the shaded grey region shows the confidence bounds and the dotted red lines show the prediction bounds.

Section 6.5: Matrix notation

Using matrix notation

We can express the simple linear regression model in matrix notation.

The regular form of the SLR model is

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

where $\epsilon_i \sim$ IID $N(0, \sigma^2)$.

This can be expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $\mathbf{Y}$ is the $n \times 1$ response vector,

$\mathbf{X}$ is an $n \times 2$ design matrix,

β is a 2×1 parameter vector,

and ϵ is the $n \times 1$ error vector

Using matrix notation contd.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

The variance of $\mathbf{Y}$ can be expressed as

$$\text{Var}(\mathbf{Y}) = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn}^2 \end{bmatrix}$$

Where the diagonal values are the variances, and off-diagonals are the covariances. But remember, for simple linear regression model we assume that the Y_i values are independent and have constant variance, giving

$$\text{Var}(\mathbf{Y}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \text{Var}(\epsilon)$$

Least squares estimation in matrix notation

The least squares estimate of the unknown parameter vector β is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

assuming that $\mathbf{X}^T\mathbf{X}$ is invertible.

The least squares estimate of β is unbiased, i.e., $E[\hat{\beta}] = \beta$, and the variance is given by $\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

The predicted $\mathbf{Y}$ values are given by

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is known as the hat matrix.

The corresponding vector of residuals is

$$\hat{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$$

The simple linear regression model in matrix notation

The $\mathbf{X}$ matrix is:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$\begin{aligned}(\mathbf{X}^T\mathbf{X})^{-1} &= \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \\ &= \frac{1}{n(\sum x_i^2 - n\bar{x}^2)}\begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \\ &= \frac{1}{S_{xx}}\begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\end{aligned}$$

The simple linear regression model in matrix notation contd.

Then to find the least squares estimates:

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum x_i y_i \end{bmatrix}$$

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \frac{1}{S_{xx}} \begin{bmatrix} \bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i \\ -n\bar{x}\bar{y} + \sum x_i y_i \end{bmatrix} \end{aligned}$$

With some algebra, this gives:

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

and

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

as before.

The simple linear regression model in matrix notation contd.

The variance of $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

$$= \frac{\sigma^2}{S_{xx}} \begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

which gives

$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{S_{xx}} \sum x_i^2/n = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$

$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, as before, and

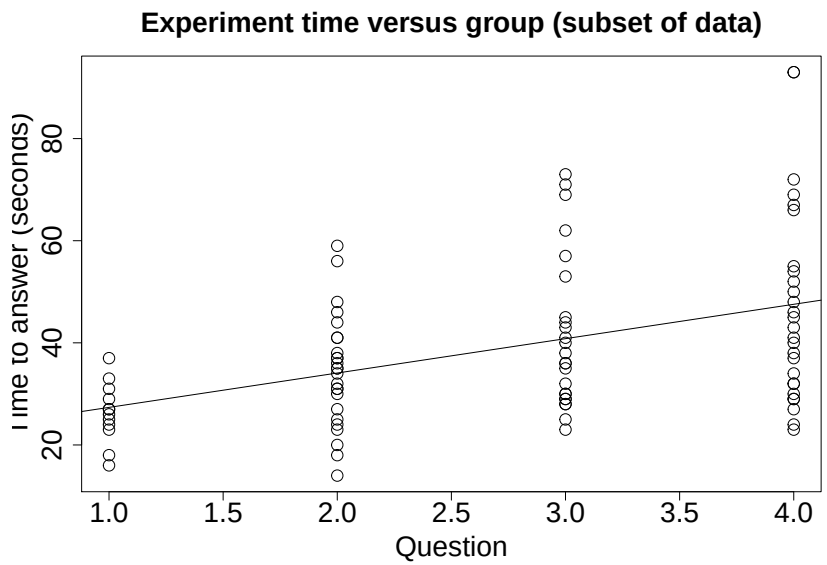
$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\frac{\sigma^2}{S_{xx}}$.

Section 6.6: Multiple regression

Multiple regression

Up until now, we have considered the simple linear regression model, with only one x predictor. We will now extend that so that the Y_i can depend on a number of possible independent variables $x_{i1}, x_{i2}, ..., x_{ik}$.

Let's think back to our in-class experiment. We looked at a simple linear regression model using question group as a predictor.



What other variables could we have recorded or considered?

The multiple regression model

A multiple regression model takes the form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \qquad \text{for } i = 1 \text{ to } n$$

and we can express this using matrix notation as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

with dimensions $(n \times 1), (n \times p), (p \times 1)$ and $(n \times 1)$, where $p = k + 1$ is the number of model parameters.

β is a vector of unknown parameters to be estimated from observed data. β_j is the change in the mean value of Y per unit change in x_j, assuming all other independent variables are held constant. Consequently, the β_j depend on which x's are included in the model.

Model assumptions

For the model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

we have

$$E[\mathbf{Y}] = \mathbf{X}\beta, \quad E[\epsilon] = 0, \quad \text{Var}(\mathbf{Y}) = \text{Var}(\epsilon) = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

The assumptions are:

1. Linearity: $E[\epsilon] = \mathbf{0}$, hence $E[\mathbf{Y}] = \mathbf{X}\beta$.
2. Constant variance and 0 covariances: $\text{Var}(\epsilon) = \sigma^2 I_n$ and $\text{Var}(\mathbf{Y}) = \sigma^2 I_n$.
3. Multivariate normal (MVN) distribution: $\epsilon \sim N_n(\mathbf{0}, \sigma^2 I_n)$ and $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 I_n)$

Sampling distribution of $\hat{\beta}$

Let $\hat{\beta}$ be the OLS estimator of β.

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

As before,

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is known as the hat matrix.

Also:

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

When the model assumptions hold:

$$\hat{\beta} \sim N_p(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

and

$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2)$$

where c_{jj} is the jj entry of $(\mathbf{X}^T\mathbf{X})^{-1}$ for $j = 0, ..., k$.

Franchise data example

- Data on 27 "All Greens" franchise branches
- There are multiple variables recorded on each branch
  - net sales (in thousands)
  - shop size, inventory (in thousands)
  - advertising revenue (in thousands)
  - size of sales district (thousands of families)
  - the number of competing stores in the sales district.
- Try to predict net sales
  - We could look at how net sales varies with each of the possible predictors.
  - Predictors could have an effect in tandem! For example advertising revenue and size of sales district could have different predictive power if considered together.
  - Which of the predictors are most important?

Franchise data matrix plot

Fitted model in R

```
##
## Call:
## lm(formula = net_sales ~ shop_size_sqft + inventory + advertising_rev +
##     size_sales_district + num_competing_stores, data = greens)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.338  -9.699  -4.496   4.040  41.139
##
## Coefficients:
##                       Estimate Std. Error t value    Pr(>|t|)
## (Intercept)          -18.85941   30.15023  -0.626    0.538372
## shop_size_sqft        16.20157    3.54444   4.571    0.000166 ***
## inventory              0.17464    0.05761   3.032    0.006347 **
## advertising_rev       11.52627    2.53210   4.552    0.000174 ***
## size_sales_district   13.58031    1.77046   7.671 0.000000161 ***
## num_competing_stores  -5.31097    1.70543  -3.114    0.005249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.65 on 21 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9916
## F-statistic: 611.6 on 5 and 21 DF,  p-value: < 0.00000000000000022
```

Model interpretation

- Shop site: $\hat{\beta}_1 = 16.202$
- Inventory: $\hat{\beta}_2 = 0.175$
- Advertising revenue: $\hat{\beta}_3 = 11.526$
- Size sales district: $\hat{\beta}_4 = 13.580$
- Number competing stores: $\hat{\beta}_5 = $ -5.311

How should we interpret these estimates? The intercept?

Advertising revenue:

For every extra \$1,000 spent on advertising, it is estimated that average sales increase by \$11,526, holding all other predictors constant.

Competing shop numbers:

For every extra competing shop in the district, it is estimated that the average sales decrease by \$5,311, keeping all other predictors constant.

Confidence interval for model parameters

To construct a 100(1 - α)% confidence interval for β_j, we use:

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\ \mathrm{SE}(\hat{\beta}_j)$$

$$= \hat{\beta}_j \pm t_{n-p,\alpha/2}\sqrt{\mathrm{MSE}\ c_{jj}}$$

where

$$\mathrm{MSE} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \sum_{j=i}^{k}\hat{\beta}_j x_{ij})^2$$

$$c_{jj} = \text{the } jj^{th} \text{ diagonal element of } (\mathbf{X}^T\mathbf{X})^{-1}$$

The degrees of freedom are equal to $n - p$ because we have estimated $p = k + 1$ parameters, where k is the number of predictors in the multiple regression model.

Confidence interval example for the franchise data

For the advertising revenue parameter for franchise data, the 95% confidence interval is

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\ \mathrm{SE}(\hat{\beta}_j)$$

$$= \hat{\beta}_3 \pm t_{27-6,0.05/2}\ \mathrm{SE}(\hat{\beta}_j)$$

$$= \hat{\beta}_3 \pm t_{21,0.025}\ \mathrm{SE}(\hat{\beta}_j)$$

$$= 11.52627 \pm 2.080 \times 2.53210 = (6.260, 16.793)$$

We are 95% confident that the true mean increase in average sales for a \$1000 increase in advertising expenditure lies between \$6,260 and \$16,793.

Hypothesis tests for model parameters

We can test the hypothesis:

H_0: $\beta_j = b$ vs H_0: $\beta_j \neq b$.

The test statistic is:

$$T_{obs} = \frac{\hat{\beta}_j - b}{\text{SE}(\hat{\beta}_j)}$$

Under the null hypothesis (that is, assuming that the H_0 is true), the test statistic is a random draw from a $t(n-p)$ distribution.

We evaluate the test by deciding if the observed test statistic is extreme, relative to the null hypothesis distribution.

Hypothesis test example for the franchise data

Advertising revenue. β_3 is the expected change in the mean sales (in thousands) for a \$1,000 increase in advertising revenue, holder other predictors constant.

H_0: $\beta_3 = 0$ vs H_0: $\beta_3 \neq 0$.

The test statistic is:

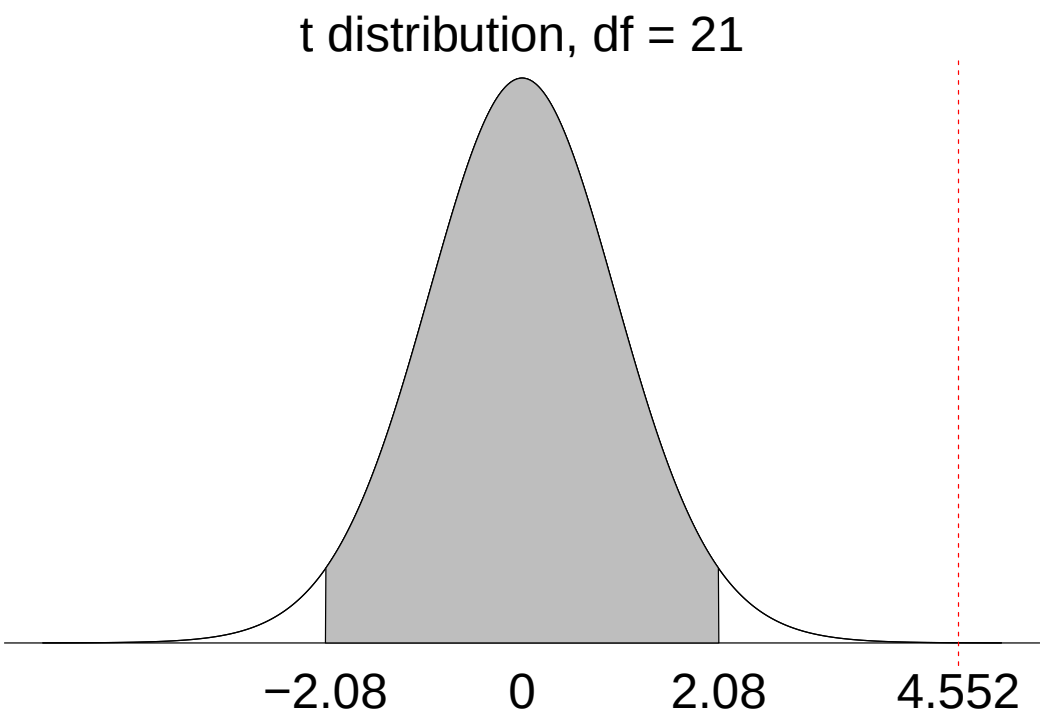
$$T_{obs} = \frac{\hat{\beta}_j - b}{\text{SE}(\hat{\beta}_j)} = \frac{11.52627 - 0}{2.53210} = 4.552$$

Using $\alpha = 0.05$, the critical values are $\pm t_{21,0.025} = \pm 2.080$. We reject H_0 if the observed test statistic is less than -2.080 or greater than 2.080.

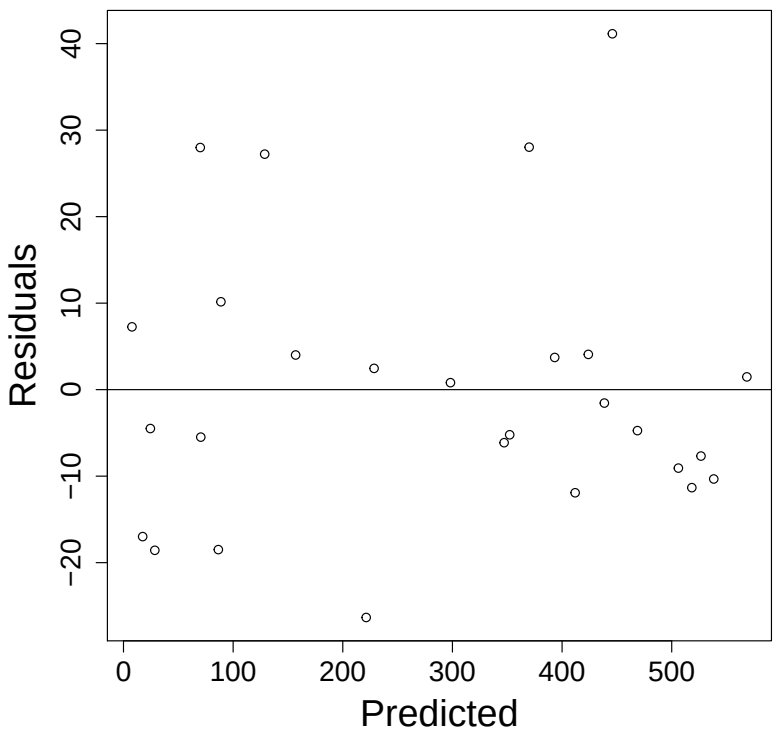
In this case, we reject the H_0 and conclude that $\beta_3 \neq 0$, since $4.552 > 2.080$.

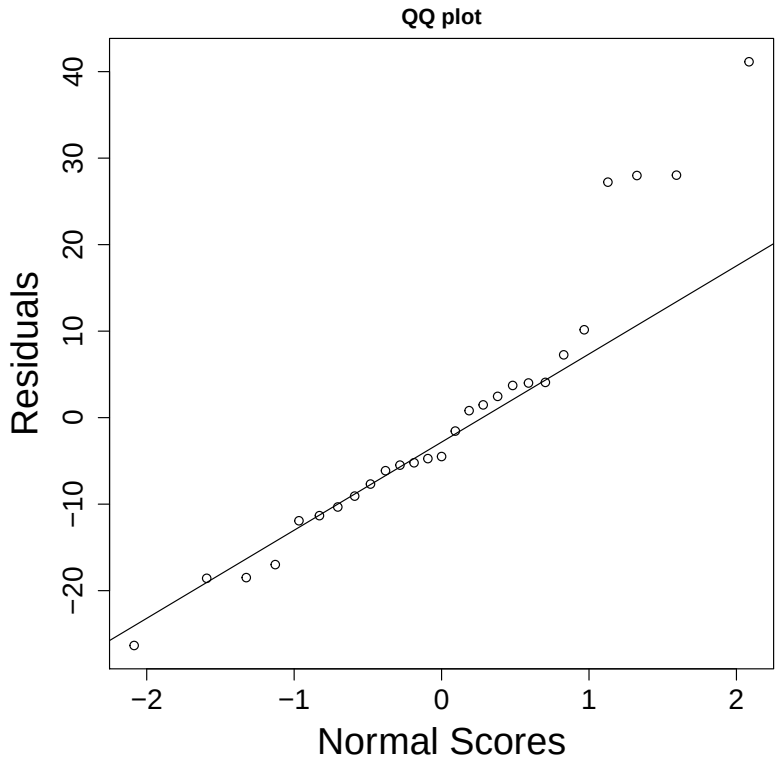
We could also evaluate the test from the p-value in the R output, the p-value $=$ 0.00174, since the p-value < 0.05, we reject the H_0 and conclude that $\beta_3 \neq 0$.

Assessing the hypothesis test visually



Model assumptions - residual plots for the franchise data





Further considerations

Next steps in the analysis of the franchise data

- Further diagnostic tests
- Multicollinearity (test using variance inflation factors - VIF)
- Interactions among the predictors?

Multiple regression in general

- Powerful and widely used tool.
- For valid inference:
  - must have an awareness of how the model should be correctly interpreted.
  - assumptions must be validated.
  - data source must be reliable and fit for purpose.