# STATS 2107
# Statistical Modelling and Inference II
# Solutions
# Workshop 3: Confidence intervals and hypothesis testing

Matt Ryan

Semester 2 2022

## Contents

# Confidence intervals for Simple Linear Regression Estimates

## The model

For data $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$, $x_i, Y_i \in \mathbb{R}$, consider the simple linear regression model

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

where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$

## The model estimates

Recall that the estimates for $\beta_0$ and $\beta_1$ are given by

$$\begin{aligned}\hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

where

$$\begin{aligned}S_{XY} &= \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) \\ S_{XX} &= \sum_{i=1}^{n} (x_i - \bar{x})^2\end{aligned}$$

## The distribution of $\hat{\beta}_1$

What is the distribution of $\hat{\beta}_1$?

1. Recall $\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$
2. Each $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
3. The sum of normals is normal

Hence $\hat{\beta}_1$ follows a normal distribution.

## The parameters

1. From last workshop, $\hat{\beta}_1$ is unbiased for $\beta_1$, so $\mathrm{E}\left[\hat{\beta}_1\right] = \beta_1$
2. The variance is given by:

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{XX}}$$

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

### A confidence interval for $\beta_1$

If $\sigma^2$ is known, a symmetric $(1-\alpha)\times 100\%$ confidence interval for $\beta_1$ is given by

$$\hat{\beta}_1 \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{S_{XX}}}\,.$$

**PROOF:**

## Your turn

### What to do

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

---

**Solutions:**
Recall from last workshop that $\hat{\beta}_0$ is the BLUE for $\beta_0$, that is, $\beta_0$ is linear and unbiased. You also showed that

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\,.$$

Hence, since the $Y_i$ are normal, $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\right)$.

---

2. Find a $(1-\alpha)\times 100\%$ confidence interval for $\beta_0$.

---

**Solutions:**
Using part 1, a $(1-\alpha)\times 100\%$ confidence interval for $\beta_0$ is

$$\hat{\beta}_0 \pm z_{\frac{\alpha}{2}}\, \sigma \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}$$

---

## Hypothesis testing for Simple Linear Regression Estimates

### What do we need

Recall we need the following 4 things for a hypothesis test:

1. A null hypothesis
2. An alternative hypothesis
3. A test statistic
4. A critical region.

### Some hypotheses

Let's suppose we are doing inference on $\beta_1$. Then, if we know $\sigma^2$, we can do a simple $Z$-test. Our hypothesis looks like:

$$H_0 : \beta_1 = \tilde{\beta}_1 \quad \text{vs} \quad H_a : \beta_1 \neq \tilde{\beta}_1$$

where $\tilde{\beta}_1$ is a fixed value.

### A test statistic

Our test statistic will be of the form:

$$\frac{\text{Best Guess} - \text{Null hypothesis}}{SE}$$

So

$$Z = \frac{\hat{\beta}_1 - \tilde{\beta}_1}{\sigma/\sqrt{S_{XX}}}$$

### Critical region

Since $Z \sim N(0,1)$ under the null hypothesis, our critical region at the $\alpha$-level of significance is

$$C_\alpha = \left\{ z : |z| > z_{\frac{\alpha}{2}} \right\}.$$

### Putting this together

We test the hypothesis

$$H_0 : \beta_1 = \tilde{\beta}_1 \quad \text{vs} \quad H_a : \beta_1 \neq \tilde{\beta}_1$$

with the test statistic

$$Z = \frac{\hat{\beta}_1 - \tilde{\beta}_1}{\sigma/\sqrt{S_{XX}}}$$

and critical region

$$C_\alpha = \left\{ z : |z| > z_{\frac{\alpha}{2}} \right\}.$$

## Your turn

### What to do

1. Describe an hypothesis test for $\beta_0$ assuming $\sigma^2$ is known.

---

**Solutions:**
We test the hypothesis

$$H_0 : \beta_0 = \tilde{\beta}_0 \quad \text{vs} \quad H_a : \beta_0 \neq \tilde{\beta}_0$$

with the test statistic

$$Z = \frac{\hat{\beta}_0 - \tilde{\beta}_0}{\sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}}$$

and critical region

$$C_\alpha = \left\{ z : |z| > z_{\frac{\alpha}{2}} \right\} .$$

---

# A practical example

## Some data

Let's look at the `FVC` dataset.

and the model

$$FVC_i = \beta_0 + \beta_1 Height_i + \varepsilon_i \, .$$

## Hypothesis to test

Let's test the hypothesis (at $\alpha = 0.05$) that

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 \neq 0 \, .$$

What do we need:

1. Our best guess $\hat{\beta}_1$.
2. Our SE, which involves
   a. $\sigma^2$
   b. $S_{XX}$.

## Getting $\hat{\beta}_1$

We can calculate this the hard way, or use R. To use R, let's fit the model (using `lm`) and view the output (using `summary`).

## Model output

```
fvc_lm <- lm(FVC ~ Height, data = fvc)
summary(fvc_lm)
```

```
##
## Call:
## lm(formula = FVC ~ Height, data = fvc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75507 -0.23898 -0.00411  0.21238  0.87589
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.064961   0.552593  -9.166 1.24e-15 ***
## Height       0.052194   0.003618  14.426  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3137 on 125 degrees of freedom
## Multiple R-squared:  0.6248, Adjusted R-squared:  0.6218
```

```
## F-statistic: 208.1 on 1 and 125 DF,  p-value: < 2.2e-16
```

### Getting $\hat{\beta}_1$

We can calculate this the hard way, or use R. To use R, let's fit the model (using `lm`) and view the output (using `summary`).

We find $\hat{\beta}_1 = 0.052194$.

### Getting $\sigma^2$ (doing a dodgy)

In truth, we don't know $\sigma^2$ and there are ways to deal with this. For today, we assuming that

$$\sigma^2 \approx s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \,.$$

**PLEASE NOTE: This is dodgy. We are doing this for illustrative purpose.**

### Model output

```r
fvc_lm <- lm(FVC ~ Height, data = fvc)
summary(fvc_lm)
```

```
##
## Call:
## lm(formula = FVC ~ Height, data = fvc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75507 -0.23898 -0.00411  0.21238  0.87589
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.064961   0.552593  -9.166 1.24e-15 ***
## Height       0.052194   0.003618  14.426  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3137 on 125 degrees of freedom
## Multiple R-squared:  0.6248, Adjusted R-squared:  0.6218
## F-statistic: 208.1 on 1 and 125 DF,  p-value: < 2.2e-16
```

### Getting $\sigma^2$ (doing a dodgy)

We take $\sigma^2 = 0.3137^2$.

### Getting $S_{XX}$

If you consider the data given by $x_1, x_2, \ldots, x_n$ (Height values), you see that the sample variance of this data (which we denote by $s_X^2$) is such that

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{S_{XX}}{n-1} \,.$$

### Getting $S_{XX}$

```
s2_x <- var(fvc$Height)
n <- length(fvc$Height)
(S_XX <- (n - 1) * s2_x)
```

```
## [1] 7517.512
```

### Getting $S_{XX}$

We take $S_{XX} = 7517.512$

### Putting it together

Using all this, we get the following Z-statistic:

$$z = \frac{0.052194 - 0}{0.3137/\sqrt{7517.512}} = 14.42591\,.$$

Since $z > 1.96$, we reject the null hypothesis that $\beta_1 = 0$.

## Your turn

### What to do

1. Construct a 95% confidence interval for the coefficient of Height in the model $FVC_i = \beta_0 + \beta_1 Height_i + \varepsilon_i$.

---

**Solutions:**
From above, we have that a 95% confidence interval will be given by

$$\hat{\beta}_1 \pm 1.96\frac{\sigma}{\sqrt{S_{XX}}}\,.$$

Thus the 95% confidence interval is given by

$$0.052194 \pm 1.96 \times \frac{0.3137}{\sqrt{7517.512}} \approx (0.0451, 0.0593)\,.$$

---

2. Test the hypothesis that $\beta_0 = -5$ at $\alpha = 0.05$.

---

**Solutions:**
Let's test the hypothesis (at $\alpha = 0.05$) that

$$H_0 : \beta_0 = -5\,.$$

What do we need:

1. Our best guess $\hat{\beta}_0$.
2. Our SE, which involves

a. $\sigma^2$,

b. $S_{XX}$,
c. $\bar{x}$.

The only thing we don't have from above is $\bar{x}$, which is found with

```
(xbar <- mean(fvc$Height))
```

```
## [1] 152.5433
```

**Solutions:**
Putting this together, we get that

$$z = \frac{-5.064961 + 5}{0.3137\sqrt{\left(\frac{1}{127} + \frac{152.5433^2}{7517.512}\right)}} = -0.1175521\,.$$

Since $|z| < 1.96$, there is insignificant evidence to reject the null hypothesis are the 5% level.