STAT241/251 Lecture Notes
Chapter 11 Part 2

Yew-Wei Lim

How to interpret $\hat{\beta}_0$ and $\hat{\beta}_1$ in our simple linear regression model?

1) $\hat{\beta}_0$ is the mean value of Y when x=0

2) The slope $\hat{\beta}_1$ represents the change in mean value of Y for 1 unit increase in x

Hypothesis test and confidence intervals concerning β_1

In simple linear regression, it is very common to test

$$\begin{aligned} & H_0: \quad \beta_1 = 0 \\ \text{vs} \quad & H_a: \quad \beta_1 \neq 0 \quad \text{(although } H_a: \quad \beta_1 > 0 \text{ or } H_a: \quad \beta_1 < 0 \text{ are also possibilities)} \end{aligned}$$

Why ? Because if H_0 is true, then it implies the mean value of Y for any value of x is the same, in which case it means that x is not useful in predicting Y.

Also, in some instances, we also test $H_0 : \beta_1 = a$ where a is a number.

Steps to hypothesis test concerning β_1

$$H_0: \quad \beta_1 = \text{number}$$
$$H_a: \quad \beta_1 < \text{number} \quad \underline{OR} \quad H_a: \beta_1 > \text{number} \quad \underline{OR} \quad H_a: \beta_1 \neq \text{number}$$

Under H_0, our test statistic is

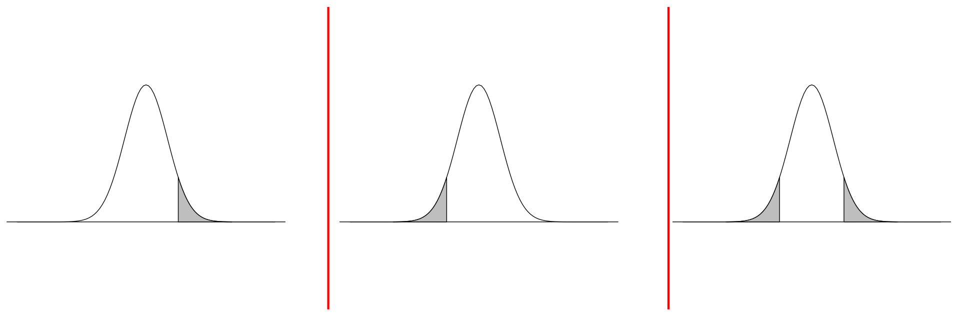
$$T = \frac{\hat{\beta}_1 - number}{s_{b_1}}$$

where $\underbrace{s_{b_1}^2}$ $= \dfrac{s^2}{\underbrace{\sum (x-\overline{x})^2}_{=\sum x^2 - \frac{(\sum x)^2}{n}}}$ careful! $s^2 = \frac{\sum(y-\hat{y})^2}{n-2}$

make sure
you know how to
get this value from R output

Critical region:

For $H_a: \beta_1 > \text{number}$ For $H_a: \beta_1 < \text{number}$ For $H_a: \beta_1 \neq \text{number}$



if $t_{obs} > t_\alpha$, reject H_0 if $t_{obs} < -t_\alpha$, reject H_0 if $|t_{obs}| > t_{\frac{\alpha}{2}}$, reject H_0

note: df = n-2

To find $100(1-\alpha)\%$ confidence interval for β_1 :

$$\hat{\beta}_1 \pm t_{\alpha/2, df=n-2} \cdot s_{b_1}$$

R command to fit simple linear regression model:

```
> x <- c(30, 300, 380, 275, 350, 190, 85)
> y <- c(957, 1125, 1202, 1028, 1134, 1124, 1062)
> fit1 <- lm(y~x)
> summary(fit1)
```

```
Call:
lm(formula = y ~ x)

Residuals:
      1      2      3       4       5      6      7
-37.391 1.151 39.793 -83.862 -13.823 52.893 41.238

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 980.0067 43.3783    22.592   3.16e-06 ***
x             0.4795  0.1662     2.885   0.0344 *
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 54.23 on 5 degrees of freedom
Multiple R-squared: 0.6247, Adjusted R-squared: 0.5496
F-statistic: 8.321 on 1 and 5 DF, p-value: 0.0344
```

Very important: $s_{b_1} = 0.1662$

Note: R is <u>always</u> testing
$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

Always done by R as $\neq$

t_{obs} =? (R says it is 2.885)
Compare t_{obs} against t-table using (n-2)df
Draw conclusion.

(Very useful "trick": can also use p-value $< \alpha$ to draw conclusion that we reject H_0. How?? p-value in R output is 0.0344. p-value < 0.05. Hence, reject H_0. It will be the same conclusion as if you had used t_{obs} against t-table.)

Example (on hypothesis testing and confidence interval involving β_1)

Consider the following data obtained in a simple linear regression study.

x	3.27	1.26	4.55	0.86	4.07	4.79	3.25
Y	16.67	19.93	14.65	17.48	18.18	13.58	15.70

(a) Find the estimated regression line

(b) Predict the mean value of Y when x=3

(c) Conduct the hypothesis test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ with significance level 0.01 (this is the same as saying $\alpha = 0.01$. Is there any evidence to suggest that x is useful in predicting y)

(d) Redo part c using $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 < 0$ with significance level 0.05.

(e) Redo part c using $H_0 : \beta_1 = -0.5$ versus $H_1 : \beta_1 \neq -0.5$ with significance level 0.05.

(f) Find a 95% confidence interval for β_1

Solution:

Instead of doing it by hand, let us use R to help get all the parameter estimates and standard error calculations.

R code for those who wish to try it themselves:

R Output:

```
> x <- c(3.27,1.26,4.55,0.86,4.07,4.79,3.25)
> y <- c(16.67, 19.93, 14.65, 17.48, 18.18, 13.58, 15.70)
> fit1 <- lm(y~x)
> summary(fit1)
```

```
Call:
lm(formula = y ~ x)

Residuals:
1 2 3 4 5 6 7
0.1938 1.4041 -0.5209 -1.4538 2.5196 -1.3462 -0.7966

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.8108     1.4841  13.349  4.22e-05 ***
x           -1.0197  0.4289     -2.377  0.0634 .
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.624 on 5 degrees of freedom
Multiple R-squared: 0.5306, Adjusted R-squared: 0.4367
F-statistic: 5.652 on 1 and 5 DF, p-value: 0.06337
```

see page 11

Solution:

(a) Estimated regression line:

$$\hat{y} = 19.8108 - 1.0197x$$

the ˆ means " estimated value of y "

(b) when $x = 3$

$$\hat{y} = 19.8108 - 1.0197(3) = 16.7517$$

(c) Two ways:

(1) We learn in class that R does the test

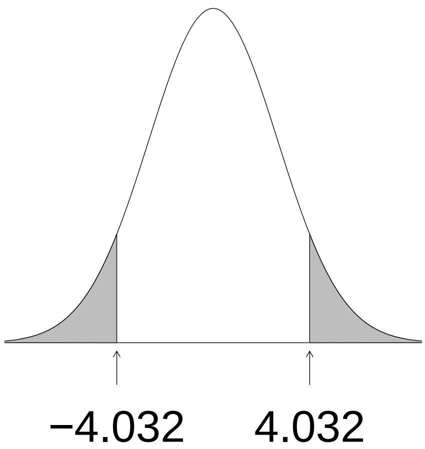
$$\begin{aligned} H_0 : &\beta_1 = 0 \quad vs \\ H_1 : &\beta_1 \neq 0 \quad \text{which matches the question} \end{aligned}$$

Therefore, we can read the answer directly from the R output.
P-value $= 0.0634 > \alpha$. Hence, we do not reject H_0.
(Rule : P-value $< \alpha$, we reject H_0)

(2) Do the test yourself.

$$t_{obs} = \frac{\hat{\beta}_1 - 0}{s_{b_1}} = \frac{-1.0197}{0.4289} = \underline{-2.377}$$

Note: agrees with R's output



$\alpha = 0.01$ in this question, so $\alpha/2 = 0.005$

Look up t-table with n-2=5 df

$t_{\alpha/2,df=5} = 4.032$

since t_{obs} does not lie within

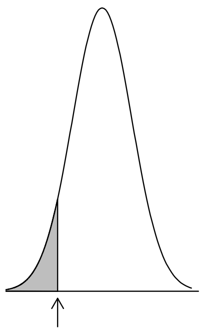
the critical region,

we do not reject H_0

(d)

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 < 0$$

We cannot use R's output since $H_a : \beta_1 < 0$ (R only produces output for $H_a : \beta_1 \neq 0$)



$t_{obs} = -2.377$ $\alpha = 0.05$ for this part

We look up $t_{0.05,df=5} = 2.015$ and

by symmetry, the critical region is $t < -2.015$.

Since $t_{obs} = -2.377$ lies in the critical

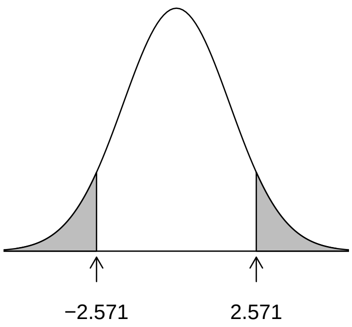
region, we reject H_0 (and conclude evidence supports $\beta_1 < 0$)

(e)

$$H_0 : \beta_1 = -0.5 \quad vs$$
$$H_1 : \beta_1 \neq -0.5$$

Under H_0, our test statistics is $T = \frac{\hat{\beta}_1 - (-0.5)}{s_{b_1}}$

$$
\begin{aligned}
t_{obs} =& \frac{\hat{\beta}_1 - (-0.5)}{s_{b_1}} \\
=& \frac{-1.0197 + 0.5}{0.4289} = -1.2117
\end{aligned}
$$



$\alpha = 0.05$ for this part, so $\alpha/2 = 0.025$

look up t-table using $t_{\alpha/2,df=5}$ and
since t_{obs} does not lies within the critical
region, we do not reject H_0
(and conclude there is no evidence to suggest $\beta_1 \neq -0.5$)

(f) 95% CI for β_1 :

$$
\hat{\beta}_1 \pm t_{\alpha/2,df=5} \cdot s_{b_1}
$$

$$
-1.0197 \pm (2.571)(0.4289)
$$

$$
(-2.122, 0.083)
$$

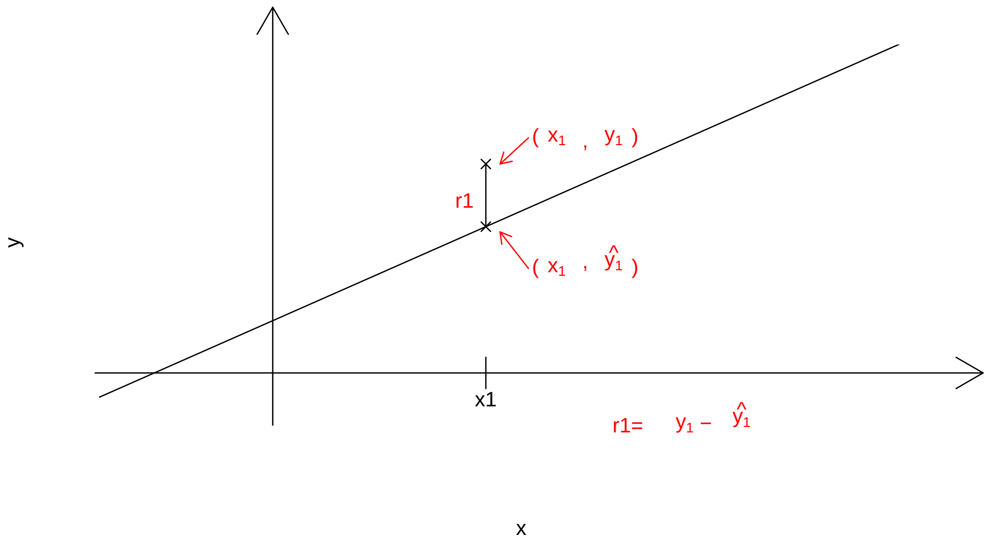
Multiple R^2

One other item given in the R output is worth noting. Pg 6

Look at R output and note that in our example, multiple $R^2 = 0.5306$

R^2 is a measure of the proportion of the variation in the data that is explained by the regression model. R^2 is a number between 0 and 1 (inclusive). The higher R^2, the better the model.

What are residuals?



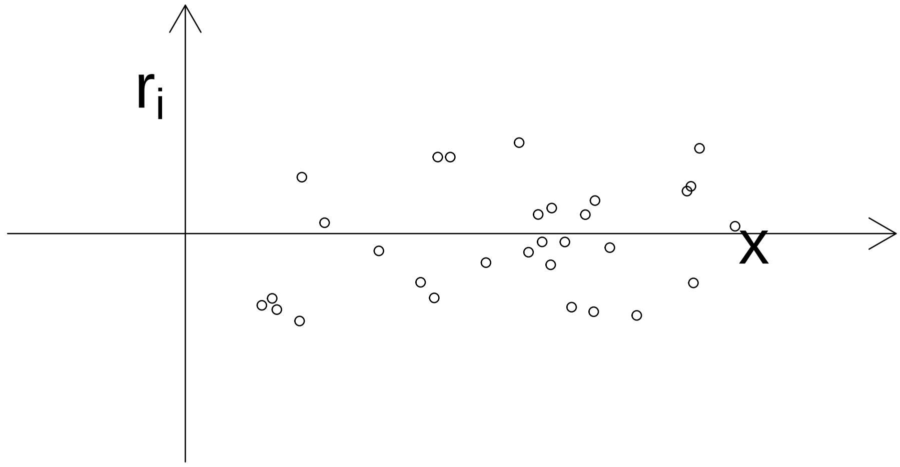
r_1 is the first residual
r_2 is the second residual
$\vdots$
there are n residuals

Recall that the assumptions in a simple linear regression model are stated in terms of ϵ_i, where $\epsilon_i \sim N(0, \sigma^2)$.

Since we do not know the value of ϵ_i, we use r_i to check assumptions violations.

(1) We use Normal probability plot of the residuals to check the Normality assumption.

(2) A scatterplot of the residuals vs the independent variable values is also used to check the simple linear regression assumptions.



If there are no violation in assumptions, the scatterplot should look like a horizontal band around zero with randomly distributed points and no discernible pattern.

Look at course notes pg 169.

Fig11.1(b) shows a curved residual plot. This suggests that a linear model is not appropriate.

Fig11.1(d) A residual plot with non-constant spread. This suggests that the variance is not the same for each value of x (hence, it violates the constant variance assumption.)