Homework 3 of Regression

1. (2.10) For each of the following questions, explain whether a confidence interval for a mean response or a prediction interval for a new observation is appropriate.

(a) Because we need to predict the humidity level tomorrow. We set the tomorrow temperature at 31 Celsius, so we need to prove the true temperature tomorrow with a prediction interval.

(b) The question has given a sample mean of disposable income. So we should focus on confidence intervals for a mean response for estimating their cost of meal out.

(c) Based on the present level of electricity consumption, the question will predict the next month electricity consumption in the Twin Cities service area. Thus, choosing the prediction interval for a new month's level is appropriate.

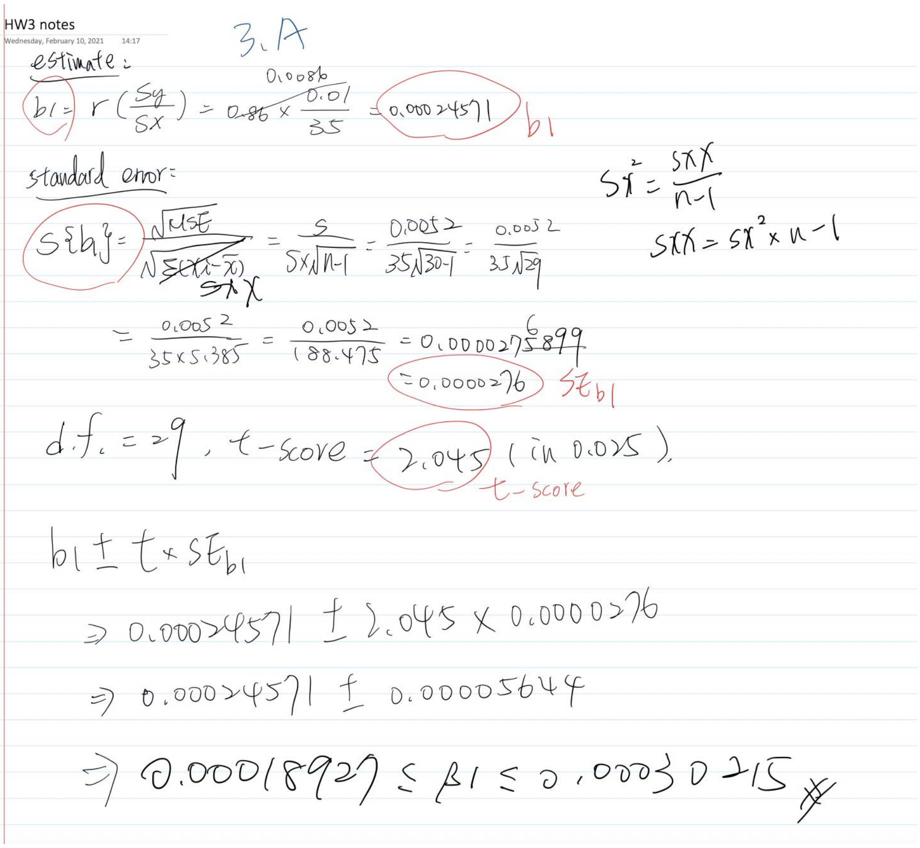
2. They are not the same. Firstly, we know X is an independent variable. The "mean response at X = x" indicates the population mean of a dependent variable when the X = x. On the other hand, the"mean of m new observations at X = x" points out m is just a sample mean when X = x.

3.

	X (files sizes)	Y (time)
samples	30	
Sample means	126	0.04
Standard deviation	35	0.01

correlation coefficient (r) = 0.86, s = 0.0052

a. 0.00018927 to 0.00030215 with 95% confidence interval.
   Reference: https://youtu.be/cql3ynkDId4



HW3 notes

Wednesday, February 10, 2021 14:17

3. A

estimate:

$$\boxed{b_1 =}\ r\left(\frac{S_y}{S_x}\right) = \cancel{0.86} \times \frac{\overset{0.0086}{\cancel{0.01}}}{35} = \boxed{0.00024571}\ b_1$$

standard error:

$$S_x^2 = \frac{S_{xx}}{n-1}$$

$$S_{xx} = S_x^2 \times n-1$$

$$\boxed{S\{b_1\} =}\ \frac{\sqrt{MSE}}{\sqrt{\underset{S_{xx}}{\cancel{\Sigma(x_i-\bar{x})}}}} = \frac{s}{S_x\sqrt{n-1}} = \frac{0.0052}{35\sqrt{30-1}} = \frac{0.0052}{35\sqrt{29}}$$

$$= \frac{0.0052}{35 \times 5.385} = \frac{0.0052}{188.475} = 0.000027\overset{6}{\cancel{5}}\cancel{899}$$

$$\boxed{= 0.0000276}\ SE_{b1}$$

$$d.f. = 29,\ t\text{-score} = \boxed{2.045}\ (\text{in } 0.025).$$

t-score

$$b_1 \pm t \times SE_{b1}$$

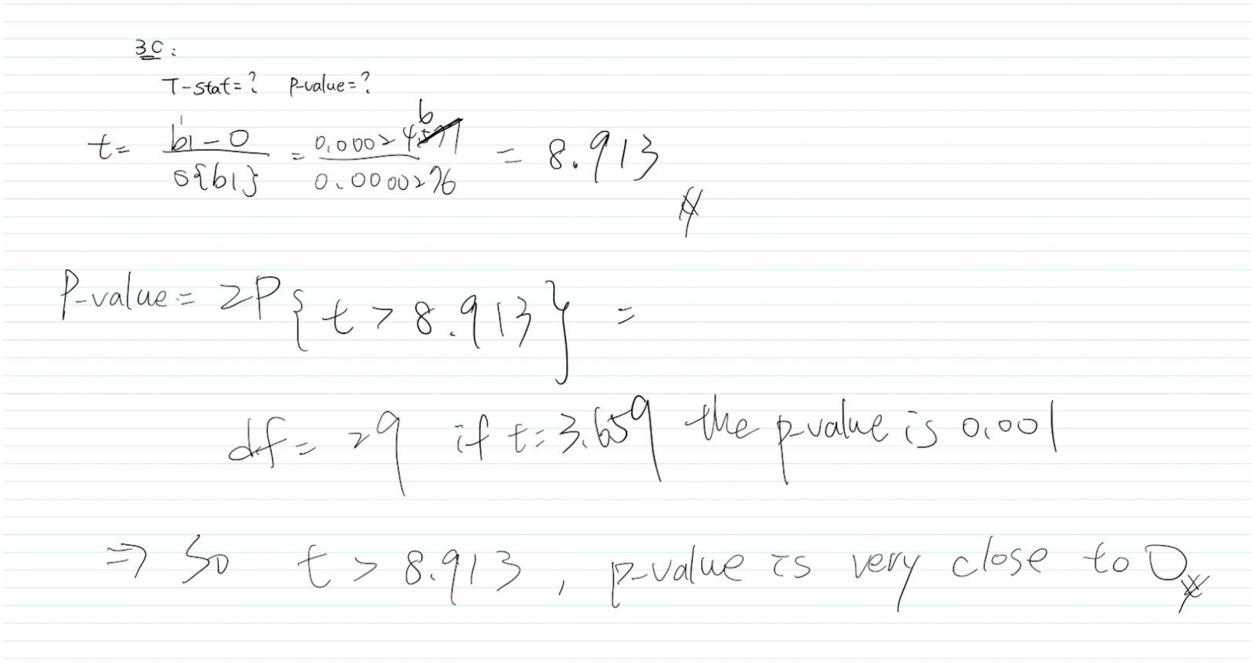
$$\Rightarrow 0.00024571 \pm 2.045 \times 0.0000276$$

$$\Rightarrow 0.00024571 \pm 0.00005644$$

$$\Rightarrow 0.00018927 \le \beta 1 \le 0.00030215\ \#$$

b. Because b1 is the slope of the regression line so we set H0: β1 = 0 v.s. Ha: b1 ≠ 0 with two-sided test.
   As we mentioned before, the b1 is 0.00024571 with 95 % confidence interval, so we have evidence to reject the null hypothesis. Thus, there is the slope at the 5 % level of significance.

c. H0: β1 = 0 v.s. Ha: β1 ≠ 0 from question 3b. We will be using b1 is not equal to zero.
   T statistic is 8.913
   P-value is very close to zero.



3c:

T-stat = ? p-value = ?

$$t = \frac{b_1 - 0}{s\{b_1\}} = \frac{0.00024\overset{6}{\cancel{577}}}{0.0000276} = 8.913$$

#

$$\text{P-value} = 2P\{t > 8.913\} =$$

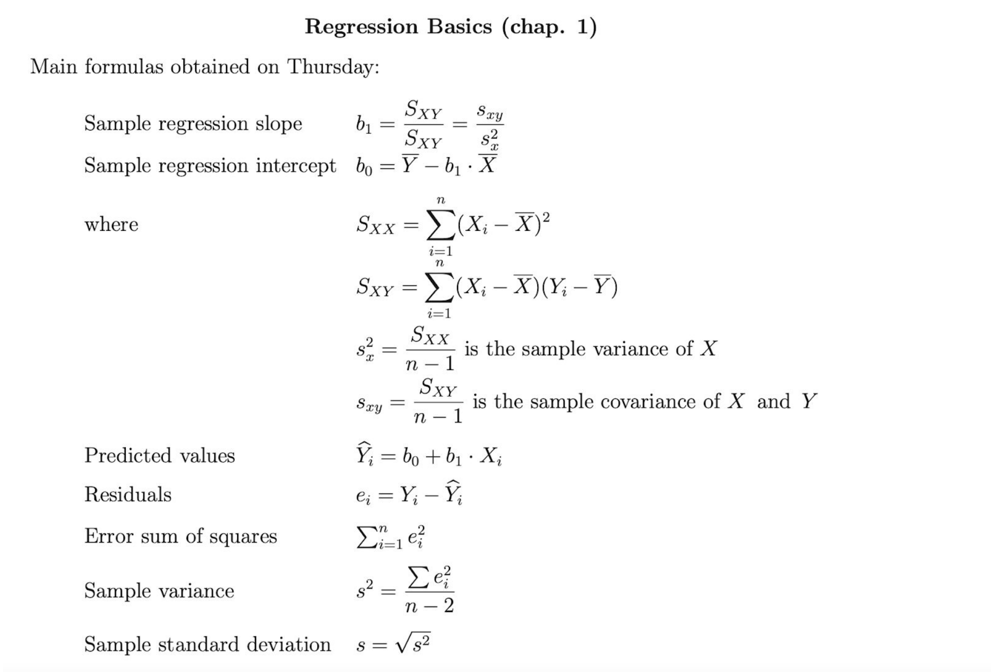
df = 29 if t = 3.659 the p-value is 0.001

⇒ So t > 8.913, p-value is very close to 0 #

T-table reference: https://www.tdistributiontable.com/

d. According to the question, consider a one-sided (right-tail alternative) is H0: β1 = 0 v.s. Ha: β1 > 0. Because β1 is the average change in Y for every one unit increase in X. In other words, β1 is a critical parameter to know the relationship between X and Y. If β1 > 0, one unit of X file size will increase β1 times. Conversely, if β1 < 0, it would take a negative size to transmit a file, which does not make sense. Compared to the left-tail and two-tail hypothesis of Ha, the right-tail alternative hypothesis is more reasonable.

4. Recall the hint table and HW2 solution



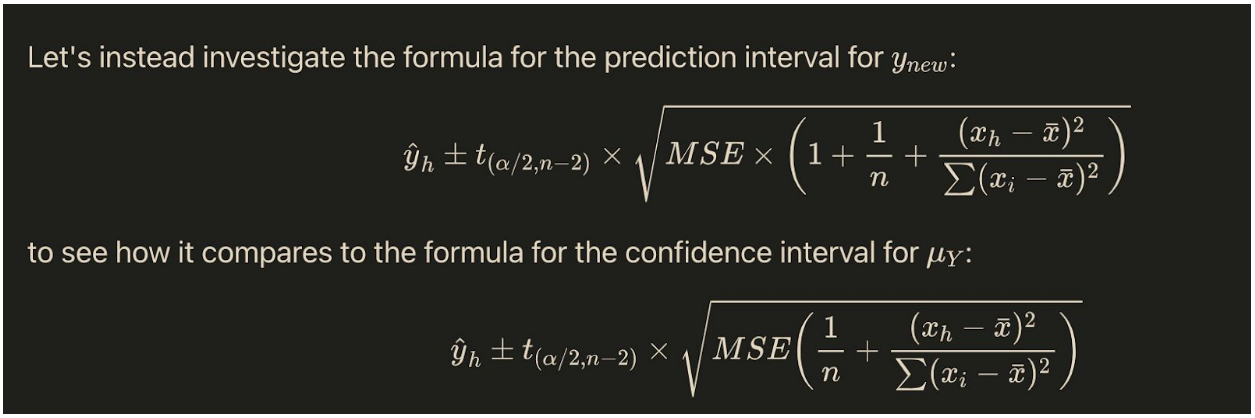
Regression Basics (chap. 1)

Main formulas obtained on Thursday:

Sample regression slope	$b_1 = \frac{S_{XY}}{S_{XY}} = \frac{s_{xy}}{s_x^2}$
Sample regression intercept	$b_0 = \overline{Y} - b_1 \cdot \overline{X}$
where	$S_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2$
	$S_{XY} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$
	$s_x^2 = \frac{S_{XX}}{n-1}$ is the sample variance of X
	$s_{xy} = \frac{S_{XY}}{n-1}$ is the sample covariance of X and Y
Predicted values	$\widehat{Y}_i = b_0 + b_1 \cdot X_i$
Residuals	$e_i = Y_i - \widehat{Y}_i$
Error sum of squares	$\sum_{i=1}^{n} e_i^2$
Sample variance	$s^2 = \frac{\sum e_i^2}{n-2}$
Sample standard deviation	$s = \sqrt{s^2}$

We know b1 = -0.0000395, b0 = 24.77, Yhat = 23.39

T-table: http://simulation-math.com/TDistTable.pdf



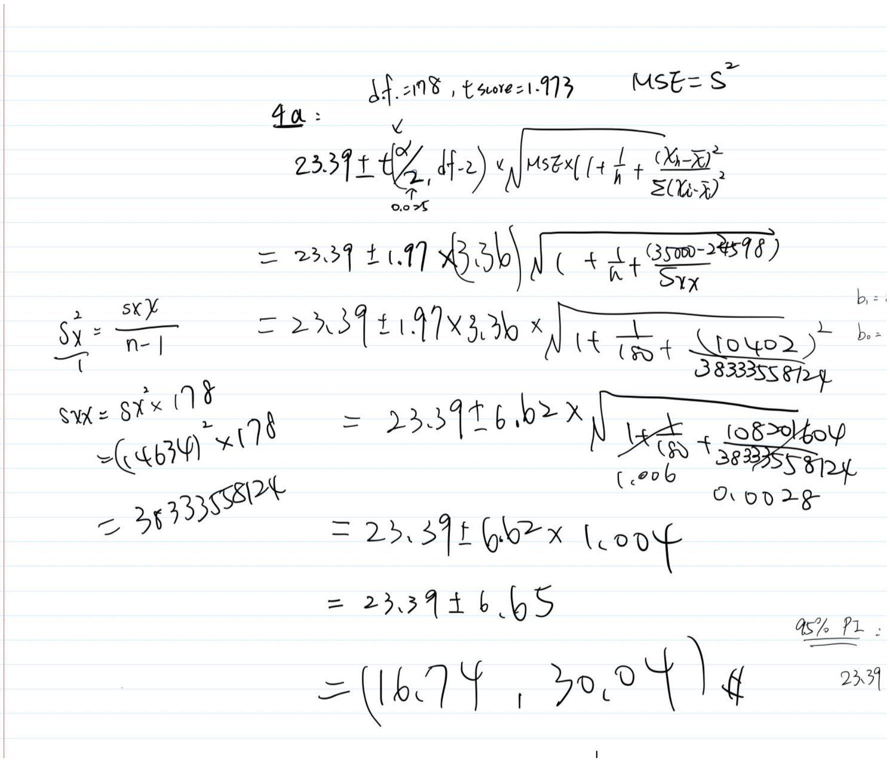
Let's instead investigate the formula for the prediction interval for y_{new}:

$$\hat{y}_h \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

to see how it compares to the formula for the confidence interval for μ_Y:

$$\hat{y}_h \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

4a. The 95% prediction interval for the miles per gallon of a used car with 35,000 miles is 16.74 to 30.04.



4a:

d.f.=178, t score=1.973 MSE = S^2

$$23.39 \pm t(\underset{0.025}{\alpha/2}, df-2) \times \sqrt{MSE\times(1+\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}}$$

$$= 23.39 \pm 1.97 \times (3.36) \sqrt{1+\frac{1}{n}+\frac{(35000-24598)}{Sxx}}$$

$$S_x^2 = \frac{Sxx}{n-1}$$

$$= 23.39 \pm 1.97 \times 3.36 \times \sqrt{1+\frac{1}{180}+\frac{(10402)^2}{38333558124}}$$

$$Sxx = Sx^2 \times 178 = (14634)^2 \times 178 = 38333558124$$

$$= 23.39 \pm 6.62 \times \sqrt{\underset{1.006}{\cancel{1+\frac{1}{180}}}+\underset{0.0028}{\frac{108201604}{38333558124}}}$$

$$= 23.39 \pm 6.62 \times 1.004$$

$$= 23.39 \pm 6.65$$

$$= (16.74, 30.04) \#$$

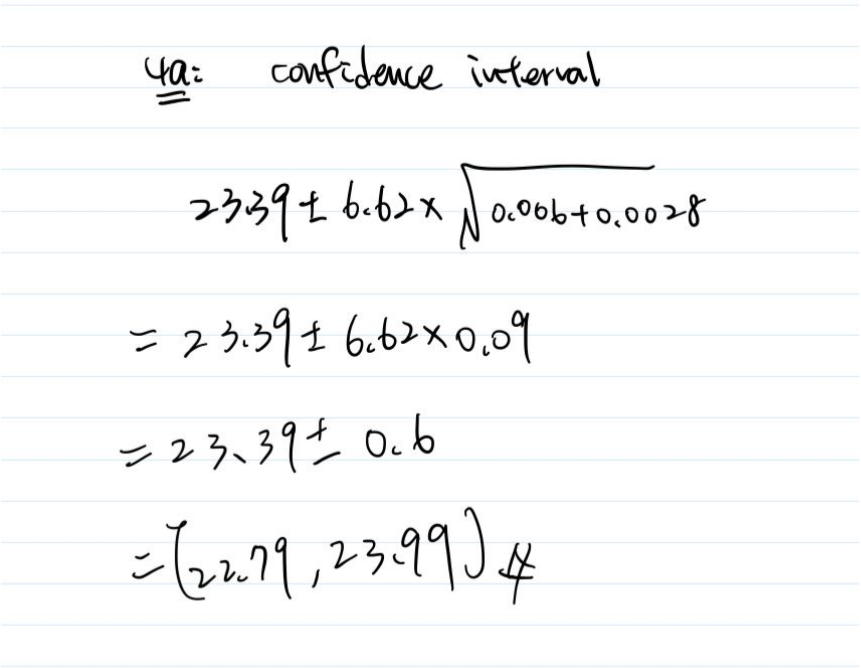
$b_1 =$

$b_0 =$

95% PI :

23.39

The 95% confidence interval for the miles per gallon of a used car with 35,000 miles is 22.79 to 23.99



4a: confidence interval

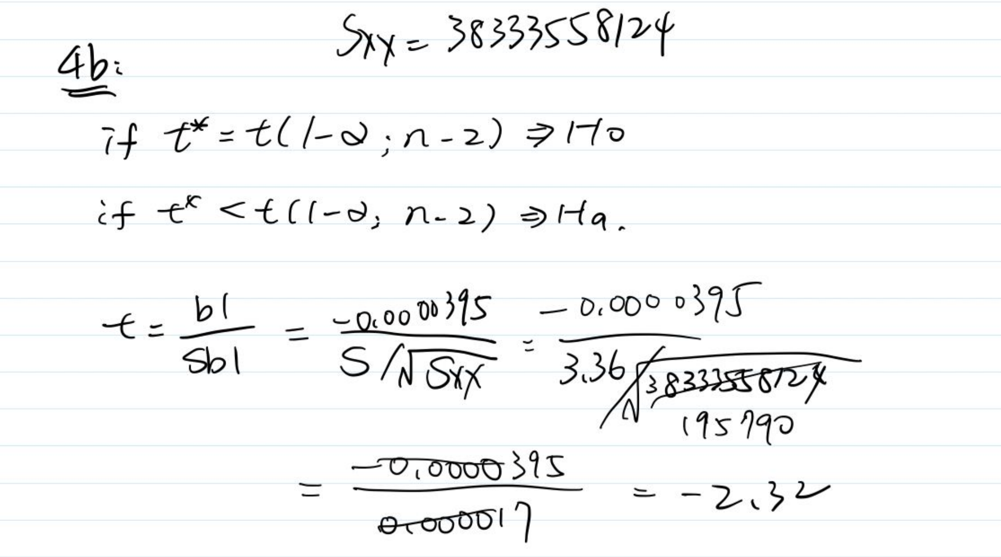
$$23.39 \pm 6.62 \times \sqrt{0.006+0.0028}$$

$$= 23.39 \pm 6.62 \times 0.09$$

$$= 23.39 \pm 0.6$$

$$= (22.79, 23.99) \#$$

4b. This is a one-sided t-test. According to the question description, we set H0: β1 = 0 v.s. Ha: β1 < 0. We calculate the t-score is -2.32.
With degrees of freedom is 179, the t-score is -2.347 with p-value 0.01, the t-score is -2.069 with p-value 0.02. The t-score -2.32 falls in between p-value 0.01 and 0.02 so we have evidence (one sided p-value < 0.05) to reject the null hypothesis in favor of the alternative hypothesis, meaning that we have a statistical result that cars with higher mileage are less economic.
Reference: https://youtu.be/tI6mdx3s0zk
https://www.resacorp.com/standerrorb(1).htm



$S_{xx} = 38333558124$

4b:

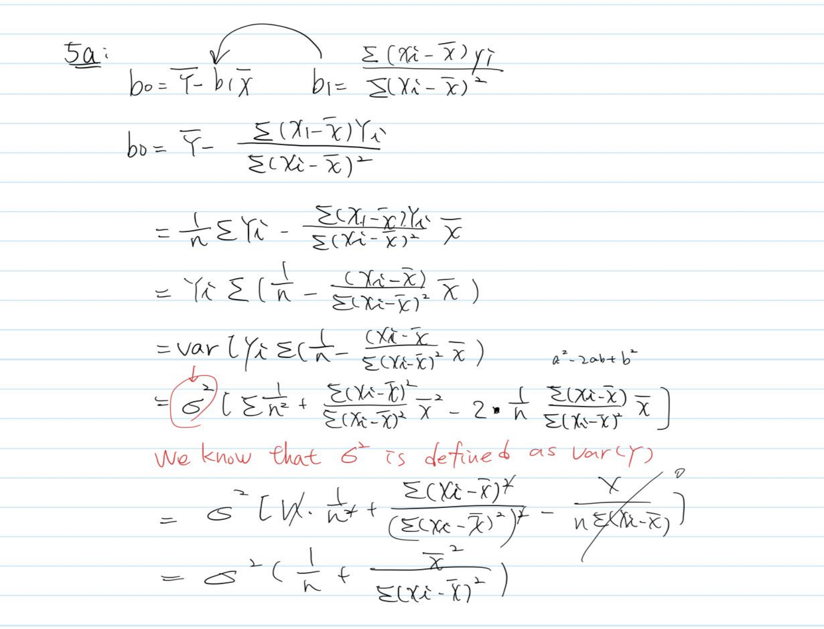
if $t^* = t(1-\alpha; n-2) \Rightarrow H_0$

if $t^* < t(1-\alpha; n-2) \Rightarrow H_a.$

$$t = \frac{b1}{Sb1} = \frac{-0.0000395}{S/\sqrt{S_{xx}}} = \frac{-0.0000395}{3.36/\sqrt{\frac{\cancel{38333558124}}{195790}}}$$

$$= \frac{-\cancel{0.0000}395}{\cancel{0.0000}17} = -2.32$$

5a.



5a:

$$b_0 = \bar{Y} - b_1 \bar{X} \qquad b_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{Y} - \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2}$$

$$= \frac{1}{n} \sum Y_i - \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2} \bar{x}$$

$$= Y_i \sum \left(\frac{1}{n} - \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \bar{x} \right)$$

$$= \mathrm{var}\left[Y_i \sum \left(\frac{1}{n} - \frac{(x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \bar{x} \right) \right. \qquad a^2 - 2ab + b^2$$

$$= \sigma^2 \left[\sum \frac{1}{n^2} + \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \bar{x}^2 - 2 \cdot \frac{1}{n} \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \bar{x} \right]$$

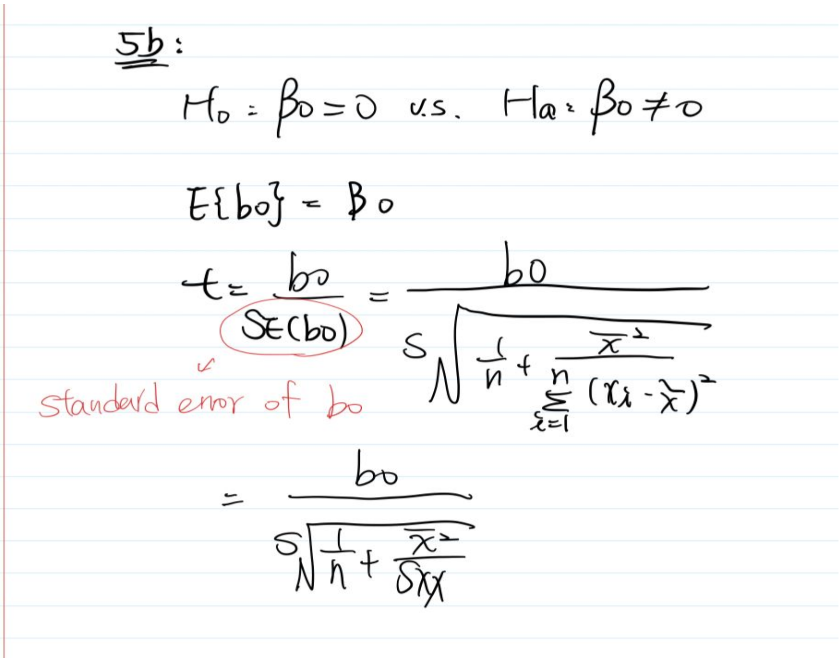
We know that σ^2 is defined as var(Y)

$$= \sigma^2 \left[\cancel{n} \cdot \frac{1}{\cancel{n^2}} + \frac{\sum (x_i - \bar{x})^{\cancel{2}}}{\left(\sum (x_i - \bar{x})^2\right)^{\cancel{2}}} - \cancel{\frac{\bar{x}}{n \sum (x_i - \bar{x})}}^{0} \right]$$

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

5b. Assumptions of Linear Regression:

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. Independence: The residuals are independent. ...
3. Homoscedasticity: The residuals have constant variance at every level of x.
4. Normality: The residuals of the model are normally distributed.
   Reference: https://www.xycoon.com/standarderrorb0.htm



5b:

$$H_0 : \beta_0 = 0 \quad \text{v.s.} \quad H_a : \beta_0 \neq 0$$

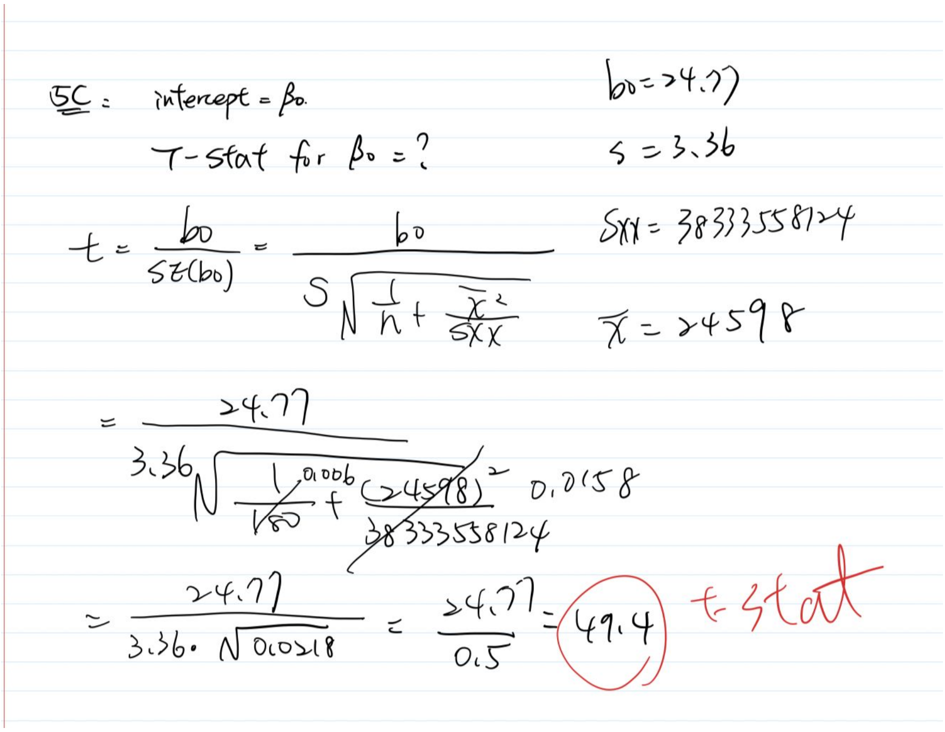
$$E\{b_0\} = \beta_0$$

$$t = \frac{b_0}{SE(b_0)} = \frac{b_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}$$

standard error of b_0

$$= \frac{b_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

5c. The t-stat of intercept is 49.4. Based on the t-table of df 179 again, if p-value = 0.0005 the t-score is 3.346 so we have strong evidence to conclude that the intercept is in the significant level.



5C: intercept = β_0

T-stat for β_0 = ?

$b_0 = 24.77$

$s = 3.36$

$Sxx = 38333558124$

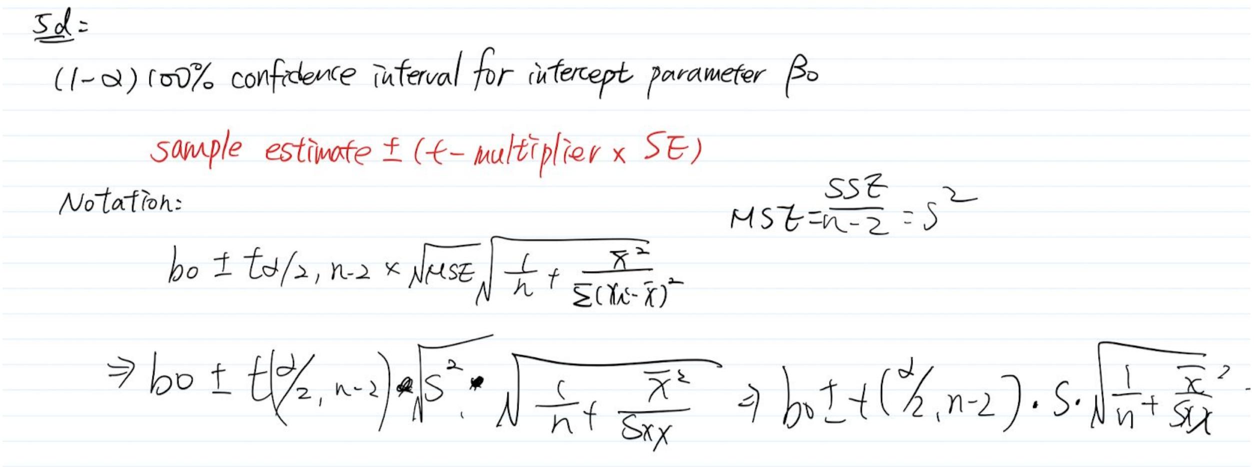
$\bar{x} = 24598$

$$t = \frac{b_0}{SE(b_0)} = \frac{b_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}}}$$

$$= \frac{24.77}{3.36\sqrt{\overset{0.006}{\frac{1}{180}} + \frac{(24598)^2}{38333558124}}} \quad 0.0158$$

$$= \frac{24.77}{3.36 \cdot \sqrt{0.0218}} = \frac{24.77}{0.5} = \boxed{49.4}$$ t-stat

5d. Reference : https://online.stat.psu.edu/stat501/lesson/2/2.1#paragraph--633



5d:

$(1-\alpha)100\%$ confidence interval for intercept parameter β_0

sample estimate ± (t- multiplier x SE)

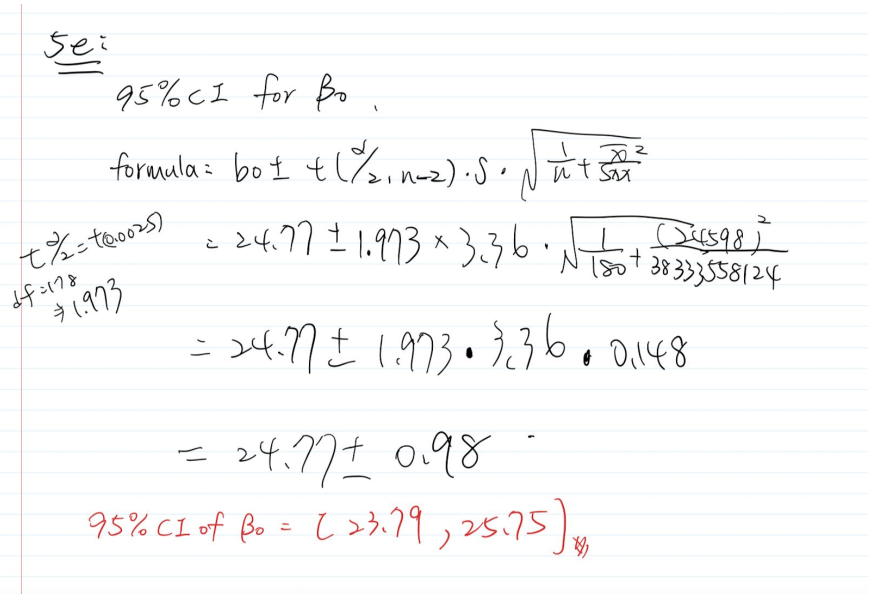
Notation:

$$MSE = \frac{SSE}{n-2} = S^2$$

$$b_0 \pm t_{\alpha/2, n-2} \times \sqrt{MSE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}}$$

$$\Rightarrow b_0 \pm t(\alpha/2, n-2) \cdot \sqrt{S^2} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}} \Rightarrow b_0 \pm t(\alpha/2, n-2) \cdot S \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}}$$

5e. 95% confidence interval for the intercept is 23.79 to 25.75.
T table: http://simulation-math.com/TDistTable.pdf



5e:

95% CI for β_0.

formula: $b_0 \pm t(\alpha/2, n-2) \cdot S \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

$t_{\alpha/2} = t(0.0025)$
$df = 178$
$\Rightarrow 1.973$

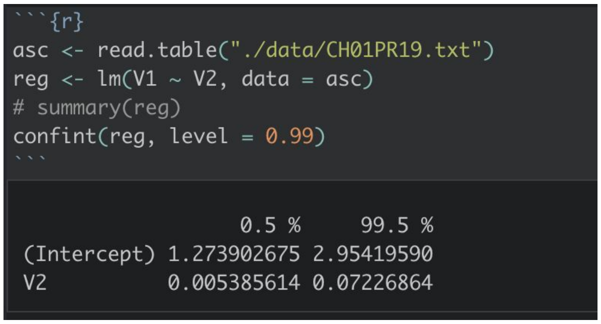
$= 24.77 \pm 1.973 \times 3.36 \cdot \sqrt{\frac{1}{180} + \frac{(24598)^2}{38333558124}}$

$= 24.77 \pm 1.973 \cdot 3.36 \cdot 0.148$

$= 24.77 \pm 0.98$

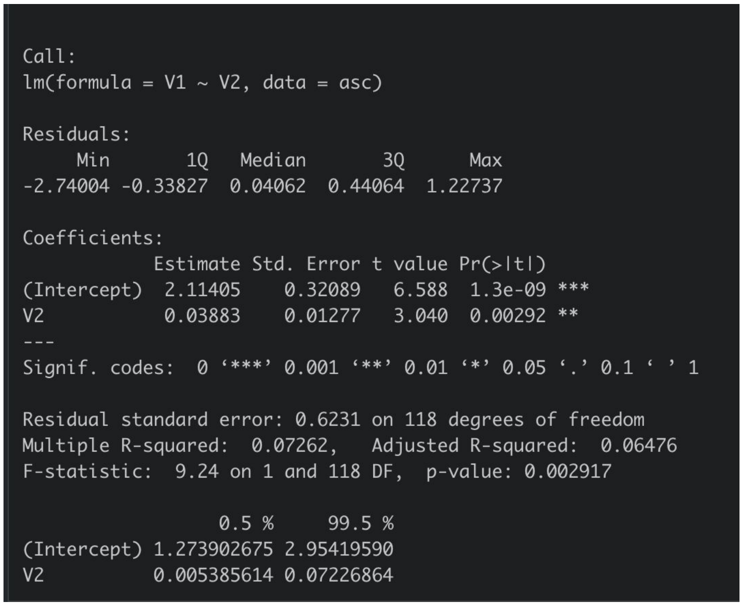
95% CI of β_0 = [23.79, 25.75]

6. (I have done before the in-class Lab 6 so I still keep it)
V1 = y = dependent variable = GPA
V2 = x = independent variable = ACT score



~~~
```{r}
asc <- read.table("./data/CH01PR19.txt")
reg <- lm(V1 ~ V2, data = asc)
# summary(reg)
confint(reg, level = 0.99)
```

                  0.5 %     99.5 %
(Intercept) 1.273902675 2.95419590
V2          0.005385614 0.07226864
~~~



```
Call:
lm(formula = V1 ~ V2, data = asc)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.74004 -0.33827  0.04062  0.44064  1.22737 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
V2           0.03883    0.01277   3.040  0.00292 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,	Adjusted R-squared:  0.06476 
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917

                  0.5 %     99.5 %
(Intercept) 1.273902675 2.95419590
V2          0.005385614 0.07226864
```

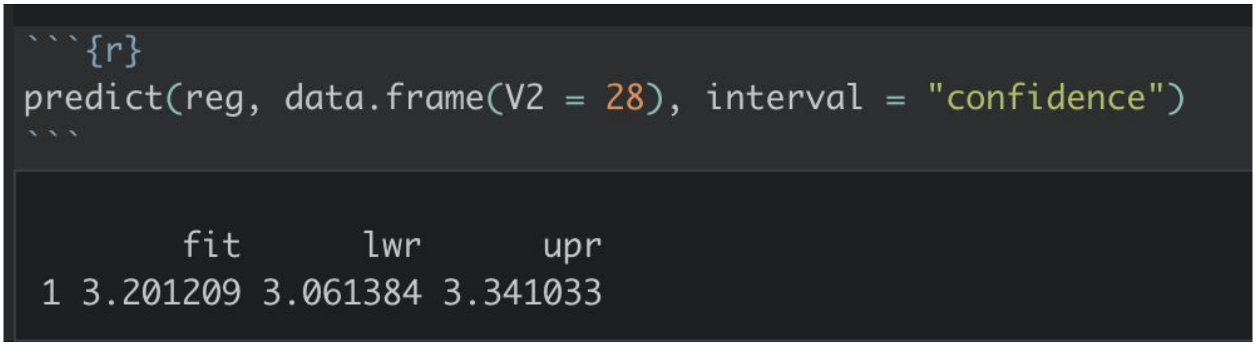
a. With a 99 % confidence interval of slope, the confidence interval is [1.274, 2.954] so there is not zero included. Because the β1 is at a significant level so the director of admissions can predict if the ACT score increases one unit, the GPA will be increased in 0.039 units, which means ACT variable is an important variable to derivative to GPA score.

a. Test whether or not a linear association we should focus on β1. So we set:
Ho: β1 = 0 (have NOT a linear association between ACT score (X) and
GPA at the end of the freshman year (Y)).
Ha: β1 ≠ 0 (have a linear association between ACT score (X) and
GPA at the end of the freshman year (Y)).

According to tables of a linear regression model above, we have evidence to reject the null hypothesis in favor of the alternative hypothesis with the 99 % confidence interval. Thus, there is a linear relationship between student’s ACT score (X) and
GPA at the end of the freshman year (Y).

b. We can see the p-value of β1 is 0.00292, which is less than the significance level of 0.01. We have the same conclusion that we reject the null hypothesis based on p-value provided.

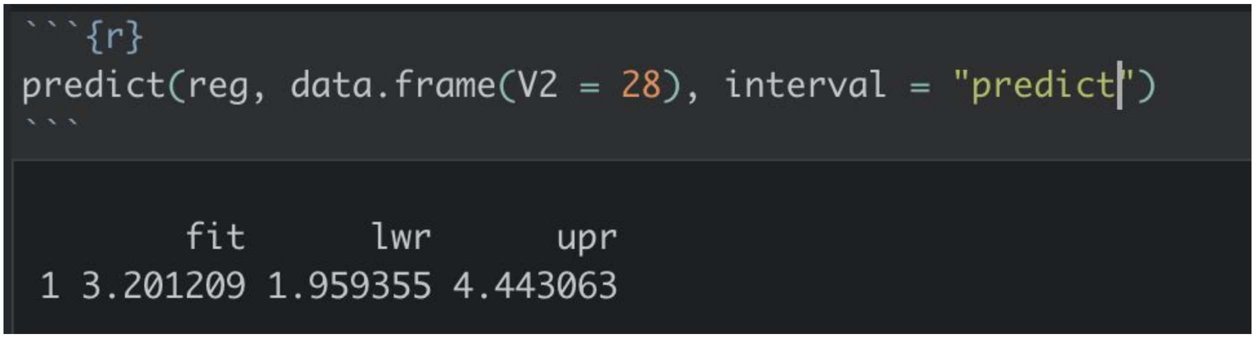
c. If ACT score is 28, we have 95% certainly contains the population mean of freshman GPA is 3.061384 to 3.341033.



~~~
```{r}
predict(reg, data.frame(V2 = 28), interval = "confidence")
```

       fit      lwr      upr
 1 3.201209 3.061384 3.341033
~~~

d. If Mary Jones obtained a 28 on the ACT, the 95 % prediction interval is 1.959355 to 4.443063 of his freshman GPA.



~~~
```{r}
predict(reg, data.frame(V2 = 28), interval = "predict")
```

       fit      lwr      upr
 1 3.201209 1.959355 4.443063
~~~

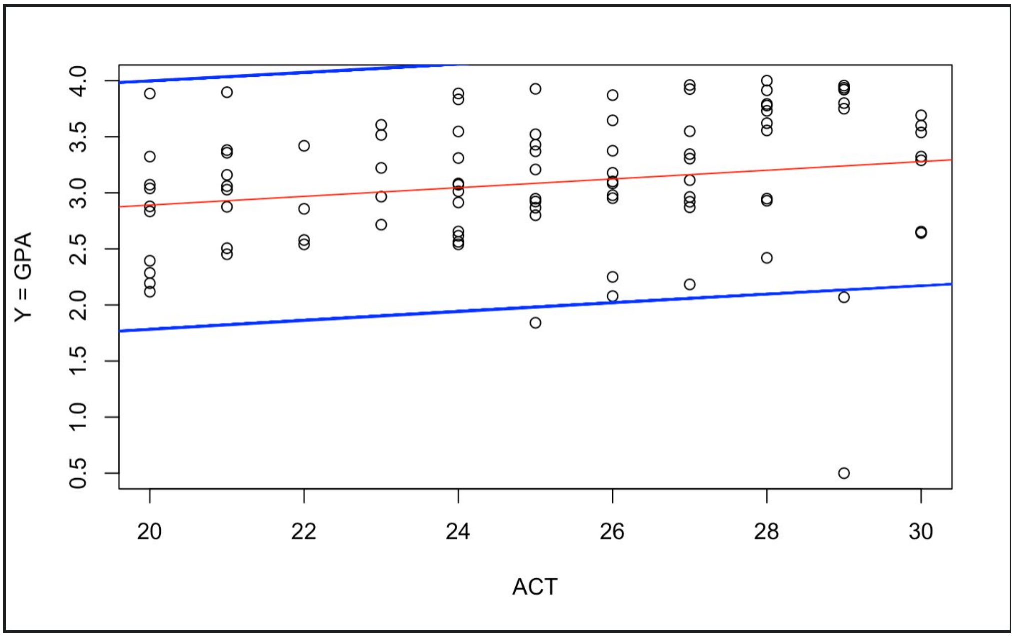
e. The majority of data points are in the range of upper band and lower band. We can conclude that the true regression relation has been precisely estimated.

Code:
n = length(X) #sample sizes
e = reg$residuals # residuals
s = sqrt(sum(e^2)/(n-2)) # estimated standard deviation = root MSE
s
W = sqrt(qf(0.95,2,n-2)) # quantity of F-distribution

```
W
Yhat = fitted.values(reg) # Yhat = b0 + b1x = predict(reg)
Sxx = (n-1)*var(X)

margin = W*s*sqrt(1/n + (X - mean(X))^2/Sxx)
upper.band = Yhat + W*s*sqrt(1 + 1/n + (X - mean(X))^2/Sxx)
lower.band = Yhat - W*s*sqrt(1 + 1/n + (X - mean(X))^2/Sxx)

plot(X,Y,xlab="ACT", ylab="Y = GPA", xlim = c(20,30))
abline(reg,col="red")
lines(X,upper.band,col="blue")
lines(X,lower.band,col="blue")
```


Y = GPA
ACT