Homework 1

Wenxiao Yang

Due Date: Tuesday 09/07 @ 11pm

Part II: HW Questions

1. SLR Reversed Consider the Simple Linear Regression model as defined in class:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{(Model I)}$$

where the ε random errors have mean zero, are homoscedastic and uncorrelated.

Now, assume that we interchange the response y_i with the predictor x_i and fit the following Simple Linear Regression model:

$$x_i = b_0 + b_1 y_i + \varepsilon_i^* \qquad \text{(Model II)}$$

where the ε^* random errors have mean zero, are homoscedastic and uncorrelated.

Derive (i.e. show your work step-by-step) the Least-Squares estimators for b_0 and b_1.

Let R_I^2 be the R^2 of model I and R_{II}^2 the R^2 of model II. Are they the same or not? Discuss.

1. $x_i = b_0 + b_1 y_i + \varepsilon_i^*$

(a) $(\hat{b}_0, \hat{b}_1) = \operatorname{argmin} \sum_i (x_i - (\hat{b}_0 + \hat{b}_1 y_i))^2$

$$\frac{\partial \sum_i (x_i - (\hat{b}_0 + \hat{b}_1 y_i))^2}{\partial \hat{b}_1} = -2\sum_i y_i (x_i - \hat{b}_0 - \hat{b}_1 y_i) = 0$$

$$\Rightarrow \sum_i y_i x_i = \hat{b}_0 \sum_i y_i + \hat{b}_1 \sum_i y_i^2 \quad (1)$$

$$\frac{\partial \sum_i (x_i - (\hat{b}_0 + \hat{b}_1 y_i))^2}{\partial \hat{b}_0} = -2\sum_i (x_i - \hat{b}_0 - \hat{b}_1 y_i) = 0$$

$$\Rightarrow \sum_i x_i = n\hat{b}_0 + \hat{b}_1 \sum_i y_i \quad (2)$$

According to (1), (2): $\begin{cases} \hat{b}_0 \sum_i y_i + \hat{b}_1 \sum_i y_i^2 = \sum_i y_i x_i \\ \hat{b}_0 n + \hat{b}_1 \sum_i y_i = \sum_i x_i \end{cases}$

we can infer that $\begin{cases} \hat{b}_1 = \dfrac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i y_i^2 - n\bar{y}^2} \\ \hat{b}_0 = \bar{x} - \bar{y}\hat{b}_1 = \bar{x} - \bar{y}\dfrac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i y_i^2 - n\bar{y}^2} \end{cases}$

(b) They are same.

As we know $R^2 = r_{xy}^2$. R^2 represents the degree of linear association between X and Y. which is same in the two models.

i.e. $R_I^2 = r_{xy}^2 = r_{yx}^2 = R_{II}^2 = \dfrac{(\sum_i (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}$

2. Stat Grades The `StatGrades.csv` data set contains 4 `Quiz Scores` and a `Final Exam` score from an Introductory Statistics course (actual course I taught several years ago!). Our goal in this example is to investigate if the average of the `Quizzes` can be used to explain the variation in the `Final Exam` scores by fitting a linear regression model of the `Final Exam`score vs. the `Average Quiz` score.

Compute the new variable `Quiz Average` and add it to the data frame.

```
Grades<-read.csv("StatGrades.csv",header=TRUE)
Grades$QuizAverage<-(Grades$quiz1+Grades$quiz2+Grades$quiz3+Grades$quiz4)/4
head(Grades)
```

```
##    quiz1  quiz2 quiz3  quiz4 final QuizAverage
## 1 100.00  87.50   100  98.33  93.5     96.4575
```

```
## 2  93.33 100.00   100  96.67  90.5     97.5000
## 3  93.33  86.25   100 100.00  95.0     94.8950
## 4  93.33  96.25   100 100.00  88.0     97.3950
## 5 100.00  78.75   100  91.67  91.0     92.6050
## 6 100.00 100.00   100  95.00  88.0     98.7500
```

Obtain the estimated regression line.

```
slr.fit<-lm(final~QuizAverage,data=Grades)
summary(slr.fit)
```

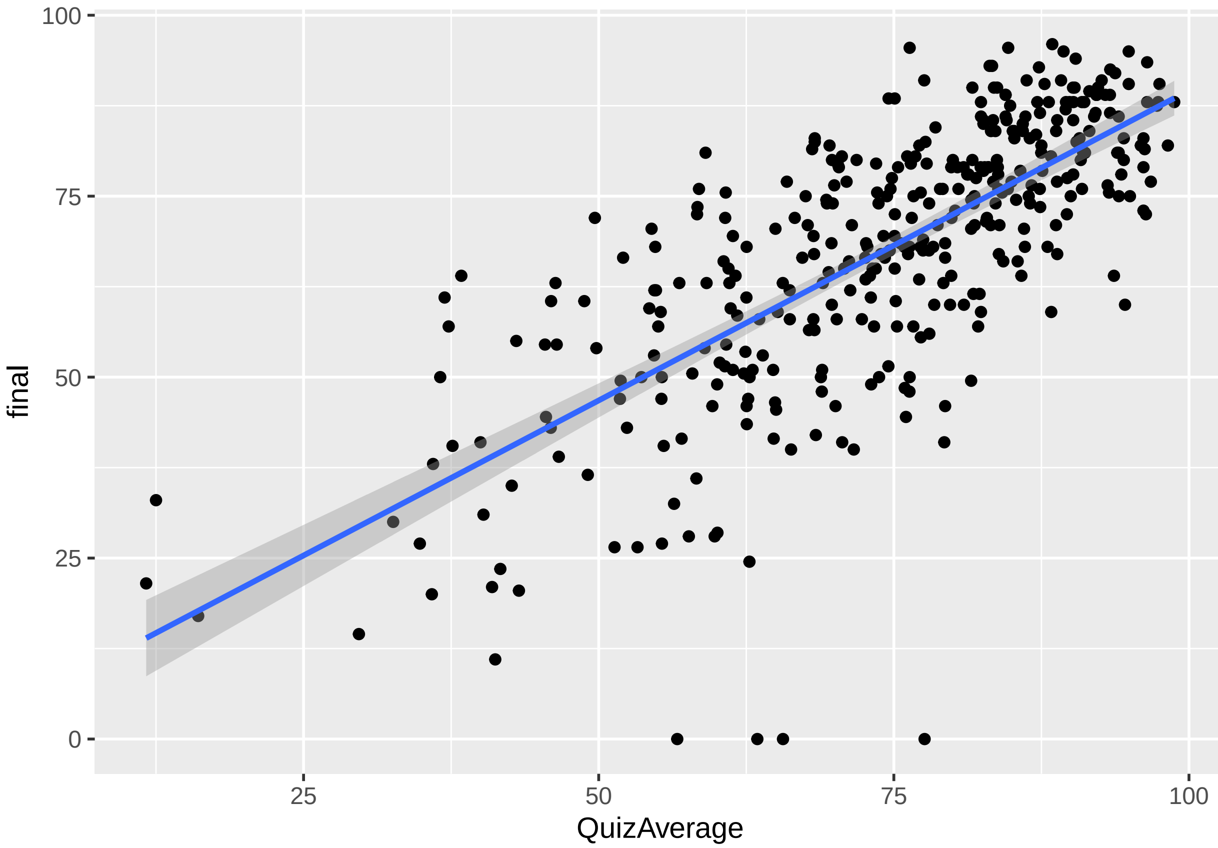
```
##
## Call:
## lm(formula = final ~ QuizAverage, data = Grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.429  -6.011   1.050   8.201  27.189
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.94717    3.15490   1.251    0.212
## QuizAverage  0.85667    0.04164  20.572   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.84 on 380 degrees of freedom
## Multiple R-squared:  0.5269, Adjusted R-squared:  0.5256
## F-statistic: 423.2 on 1 and 380 DF,  p-value: < 2.2e-16
```

$$\hat{final} = 3.94717 + 0.85667 QuizAverage$$

Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

```
library(ggplot2)
ggplot(Grades,aes(QuizAverage,final))+geom_point()+geom_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Not really good, there are still many variation of the data about the estimated regression line.

Interpret $\hat{\beta}_0$ in your estimated regression function. Does $\hat{\beta}_0$ provide any relevant information here? Explain.

```
grade.coef=summary(slr.fit)$coef
grade.coef[1,1]
```

```
## [1] 3.947172
```

$$\hat{\beta}_0 = 3.947172$$

No, it doesn't provide any information here because there are no students whose QuizAverage is 0.

Obtain a point estimate of the mean `Final Exam` score for a student with `Quiz Average` equal to 60.

```
predict(slr.fit, newdata = data.frame(QuizAverage=60))
```

```
##        1
## 55.34714
```

Estimate the difference in the mean `Final Exam` score for two students whose `Quiz Average` differs by 1. Use a 90% confidence interval.

```
confint(slr.fit, 'QuizAverage', level=0.9)
```

```
##                   5 %      95 %
## QuizAverage 0.7880018 0.9253306
```

Obtain a 90% confidence interval for the mean `Final Exam` score for students with `Quiz Average` equal to 85.

```r
predict(slr.fit,newdata = data.frame(QuizAverage=85), interval = 'confidence', level=0.9)
```

```
##       fit      lwr      upr
## 1 76.7638 75.44682 78.08077
```

Obtain a 90% prediction interval for the mean `Final Exam` score for a new student with `Quiz Average` equal to 85. Is your prediction interval wider than the corresponding confidence interval? Should it be?

```r
predict(slr.fit,newdata = data.frame(QuizAverage=85), interval = 'prediction', level=0.9)
```

```
##       fit      lwr      upr
## 1 76.7638 55.54486 97.98273
```

Prediction interval is wider. It should be wider.

Conduct an F test to determine whether or not there is a linear association between `Final Exam` score and `Quiz Average`. Use α equal to 0.1. State the alternatives, decision rule and conclusion.

$$\begin{cases} H_0 : \beta_1 = 0 \ (null) \\ H_\alpha : \beta_1 \neq 0 \ (alternative) \end{cases}$$

```r
grade.anova=anova(slr.fit)
grade.anova
```

```
## Analysis of Variance Table
##
## Response: final
##              Df Sum Sq Mean Sq F value    Pr(>F)
## QuizAverage   1  69812   69812  423.19 < 2.2e-16 ***
## Residuals   380  62687     165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's compute the p-value. If p-value$<\alpha$ we reject null.

```r
pf(grade.anova[1,4],df1=1,df2=380,lower.tail = FALSE)
```

```
## [1] 9.84919e-64
```

The p-value $9.84919e-64 < \alpha = 0.1$, so we can conclude that we reject the null: $\beta_1 = 0$, there is a linear association between `Final Exam` score and `Quiz Average`.

By how much relatively is the total variation in the `Final Exam` score reduced when the `Quiz Average` is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

```r
summary(slr.fit)$r.square
```

```
## [1] 0.5268859
```

$$R^2 = 0.5268859$$

0.5268859 of total variation in the `Final Exam` score reduced when the `Quiz Average` is introduced into the analysis. It is a relatively large reduction. The measure names Coefficient of Determination($R - square$).

Calculate r (the correlation coefficient) and attach the appropriate sign.

```r
sqrt(summary(slr.fit)$r.square)
```

```
## [1] 0.7258691
```

The sign is positive.

$$r = +|r| = +\sqrt{R^2} = +0.7258691$$