Valeria Duran

Math 6357

September 18, 2019

Homework 2

Problem 2.1:

a) Yes, the conclusion is warranted. The 95% confidence interval for the linear regression is (0.453, 1.06). Since this interval does not contain 0, the null hypothesis is rejected. The implied level of significance is that the slope is significantly different from 0 at a 0.05 level of significance.

b) Since X=0 is not within the scope of the model, it is not important if the interval contains negative numbers. If 0 is not within the scope of the interval, then the intercept coefficient has no interpretation.

Problem 2.4

a) The formula for the confidence interval for $\boldsymbol{\beta}_1$: $b_1 \pm t(1-\boldsymbol{\alpha}/2; n-2)s\{b_1\}$

Inserting the values found in HW 1: $0.03883 \pm t(0.995; 118) * (0.01277)$

$0.03883 \pm 2.6181 * (0.01277)$ $\therefore$ **99% CI: $0.00539 \leq \boldsymbol{\beta}_1 \leq 0.07227$.** We are 99% confidant that the interval (0.00539, 0.07227) covers $\boldsymbol{\beta}_1$

Using R to verify confidence interval:

```
> confint(grade.lm,level = 0.99)
                  0.5 %     99.5 %
(Intercept) 1.273902675 2.95419590
ACT         0.005385614 0.07226864
```

Since the confidence interval does not contain 0, the null hypothesis is rejected. The alternative hypothesis that $\boldsymbol{\beta}_1$ does not equal 0 will be accepted. If the confidence interval contains 0, then the effect $\boldsymbol{\beta}_1 = 0$ will not be significant.

b) H_0: $\boldsymbol{\beta}_1 = 0$; H_a: $\boldsymbol{\beta}_1 \neq 0$. Test statistic: $t^* = b_1 - \boldsymbol{\beta}_{10} / s\{b_1\}$ $\therefore$ $t^* = (0.03883 - 0)/0.01277$

∴ $t^* = 3.04072$ if $|t^* = 3.04072| \le 2.6181$, conclude H_0. Since t^* is larger, we conclude H_a.

Because we conclude the alternative hypothesis, we can conclude that there is a linear relationship between ACT and GPA.

c) The p-value is 0.0029. Since 0.0029 is less than the significance level of 0.01, we reject the null hypothesis (supports our decision in part b).

Problem 2.13

a) Solving for the confidence interval of a freshman with an ACT score of 28:

```
point.est.act <- grade.lm$coefficients[[1]] + grade.lm$coefficients[[2]]*28 #find point estimate
attach(grade.avg)
ACT.mean <- mean(grade.avg$ACT) #find the mean
sum.dev <- sum((grade.avg$ACT-ACT.mean)^2) #sum of squared deviations
var.grade <- MSE.grade*((1/120)+((28 - ACT.mean)^2)/sum.dev)
rm(sd.grade)
sd.grade <- sqrt(var.grade)
qt(0.975,118) #t value at 95%
point.est.act - 1.980272*sd.grade
point.est.act + 1.980272*sd.grade
```

The confidence interval at 95% is:

$\mathbf{3.061384 \le E\{Y_h\} \le 3.341033}$

Verifying that this is the CI:

```
act.new <- data.frame(ACT=28)
act.new.conf <- predict(grade.lm, act.new, interval = "confidence", level = 0.95, se.fit = T)
act.new.conf
$fit
      fit      lwr      upr
1 3.201209 3.061384 3.341033
```

According to this model, 95% of students with an ACT of 28 will have a B average freshman GPA between 3.2 and 3.34.

b) Solving for the 95% prediction interval for ACT=28:

```
#prediction interval
var.pred <- MSE.grade + var.grade #MSE + new variance of new value
```

sd.pred <- sqrt(var.pred)
#prediction interval at 95%:
point.est.act - 1.980272*sd.pred #lower bound
point.est.act + 1.980272*sd.pred #upper bound
95% prediction interval:
$1.959355 \leq E\{Y_{h(new)}\} \leq 4.443063$

Verifying PI:

act.new.pred <- predict(grade.lm, act.new, interval = "prediction", level = 0.95, se.firmt = T)

act.new.pred

fit lwr upr

1 3.201209 1.959355 4.443063

According to the 95% PI: Mary Jones will have a freshman GPA between 1.96 and 4.44.

c) The prediction interval in part c is wider than the confidence interval in part b since we are predicting the interval of a new point and not the interval of the mean of a parameter.

d) The 95% confidence band for the regression line at X= 28 is:

#confidence band at 95% for part d

W <- sqrt(2*qf(0.95,2,120-2))

act.new.conf$fit[,1]+W*act.new.conf$se.fit

act.new.conf$fit[,1]-W*act.new.conf$se.fit

> act.new.conf$fit[,1]+W*act.new.conf$se.fit

[1] 3.376258

> act.new.conf$fit[,1]-W*act.new.conf$se.fit

[1] 3.026159

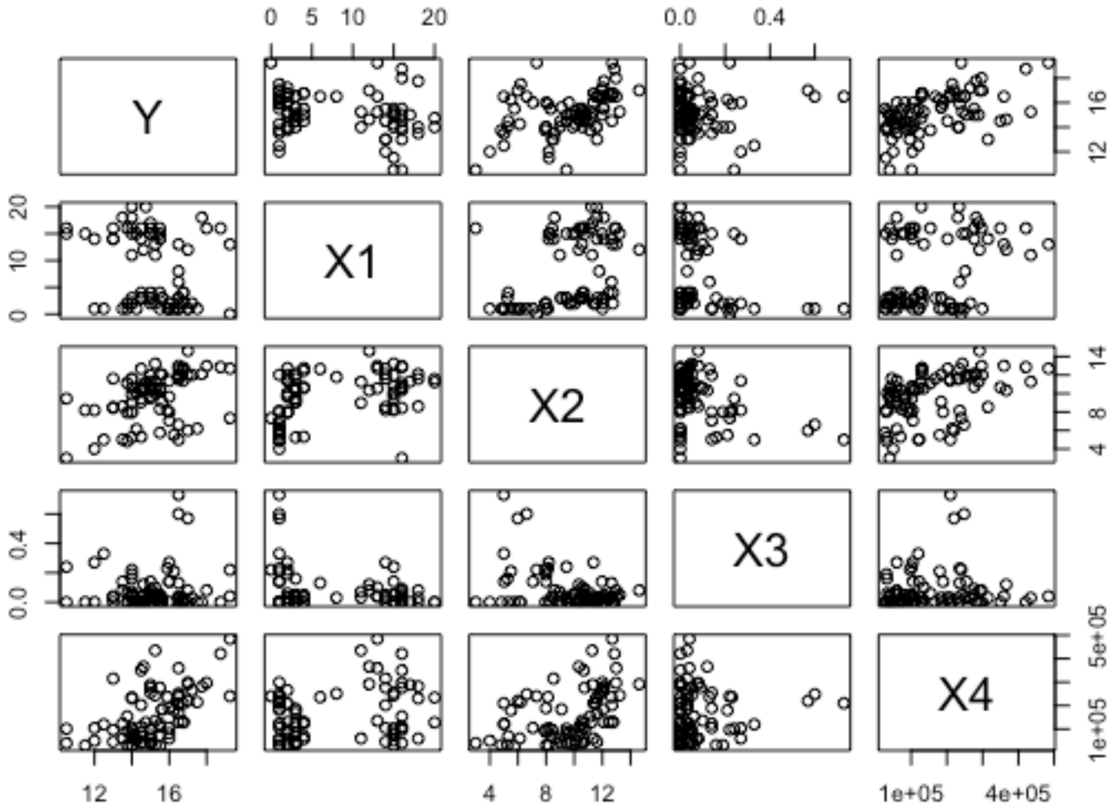
$3.026159 \leq \beta_0 + \beta_1 X_h \leq 3.376258$

The confidence band should be larger than the confidence interval in part a because it is representing the confidence interval for the entire regression line and not just the value X_h.

Problem 6.18:

a)
> property <- read.table("~/Downloads/CH06PR18.txt", quote="\"", comment.char="")
> View(property)
> names(property) <- c("Y","X1","X2","X3","X4")
> par(mfrow=c(3,2))

> pairs(property)



Based on the scatterplots, it seems that Y has a linear relationship with X2 (operating expenses) and X4 (square footage).

b) #multiple linear regression

```
property <- read.table("~/Downloads/CH06PR18.txt", quote="\"", comment.char="")
names(property) <- c("Y","X1","X2","X3","X4")
par(mfrow=c(3,2))
pairs(property)
attach(property)
prop.lm <- lm(Y ~ X1 + X2 + X3 + X4)
summary(prop.lm)
prop.lm1 <- lm(Y ~ X1 + X2 + X4) #drop parameter X3
summary(prop.lm1)
prop.lm2 <- lm(Y ~ X1 + X2) #drop parameter X4
```

```
summary(prop.lm2)
prop.lm3 <- lm(Y ~ X1 + X4) #drop parameter X2
summary(prop.lm2)
prop.lm4 <- lm(Y ~ X2 + X4) #drop parameter X1
summary(prop.lm2)
#Will leave parameters X1, X2, and X3 since it has the largest adjusted R^2
#and largest F statistic
> summary(prop.lm1)

Call:
lm(formula = Y ~ X1 + X2 + X4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0620 -0.6437 -0.1013  0.5672  2.9583

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.237e+01  4.928e-01  25.100  < 2e-16 ***
X1          -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
X2           2.672e-01  5.729e-02   4.663 1.29e-05 ***
X4           8.178e-06  1.305e-06   6.265 1.97e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.132 on 77 degrees of freedom
Multiple R-squared:  0.583,	Adjusted R-squared:  0.5667
F-statistic: 35.88 on 3 and 77 DF,  p-value: 1.295e-14
```

The estimated regression function is:

$Y = 1.237e+01 - 1.442e-01X_1 + 2.672e-01X_2 + 8.178e-06X_4$

Where $\boldsymbol{\beta_1} = -1.442e-01$, $\boldsymbol{\beta_2} = 2.672e-01$ and $\boldsymbol{\beta_4} = 8.178e-06$

The standard error values are:

2.092e-02 for X_1 , 5.729e-02 for X_2, and 1.305e-06 for X_4

The 95% CI for the parameters are:

```
> confint(prop.lm1, parm = c(2:4), level = 0.95)

          2.5 %       97.5 %

X1 -1.858219e-01 -1.025074e-01
```

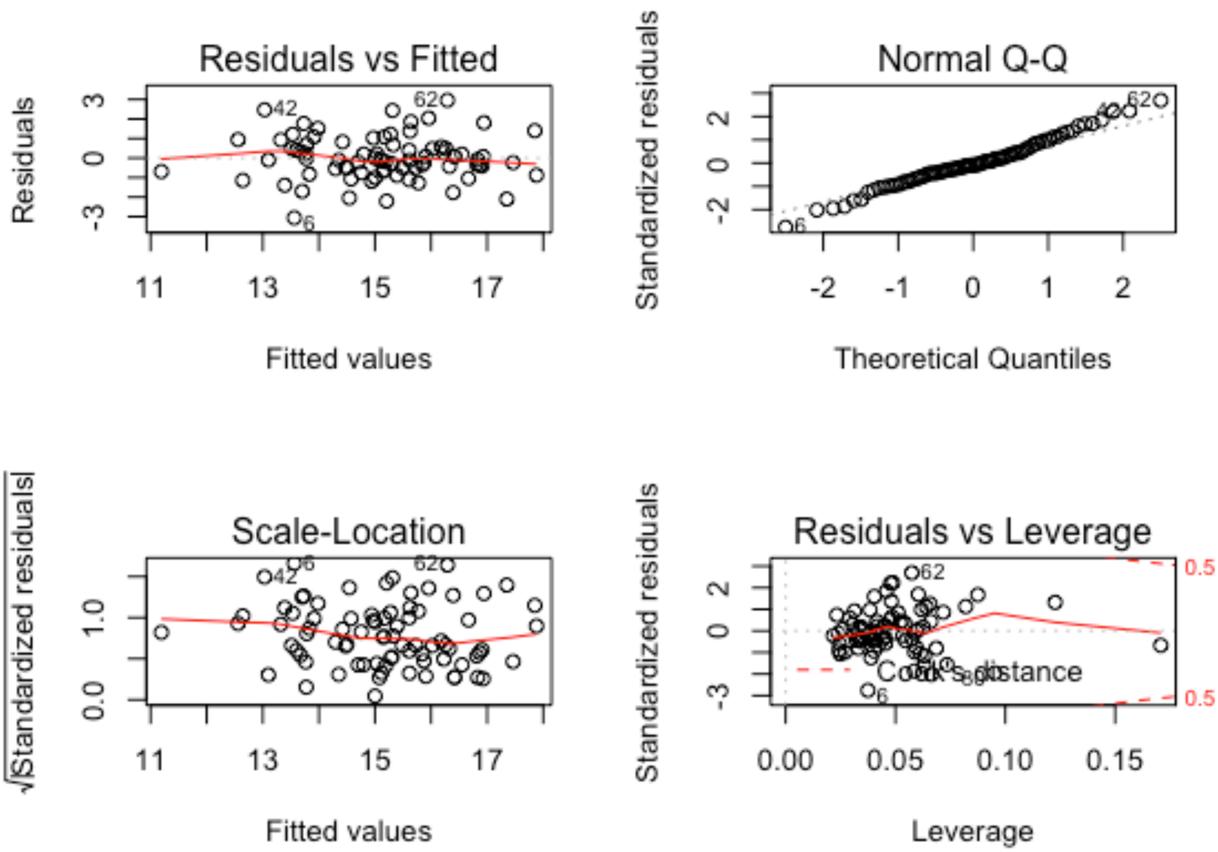
$\mathbf{-1.858219e-01 \leq \beta_1 \leq -1.025074e-01}$

X2 1.530784e-01 3.812557e-01 **$1.530784e-01 \leq \beta_1 \leq 3.812557e-01$**

X4 5.578873e-06 1.077755e-05 **$5.578873e-06 \leq \beta_1 \leq 1.077755e-05$**

c) par(mfrow=c(2,2))

plot(prop.lm1)



According to the Residual vs Fitted plot, linearity and homoscedasticity are satisfied. The QQ plot shows validates the normality of the regression. The parameters are also independent.

globaltest <- gvlma(prop.lm1)
summary(globaltest)
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance = 0.05

Call:

```
gvlma(x = prop.lm1)

                    Value p-value                Decision
Global Stat        1.33035  0.8562 Assumptions acceptable.
Skewness           0.31780  0.5729 Assumptions acceptable.
Kurtosis           0.51982  0.4709 Assumptions acceptable.
Link Function      0.05413  0.8160 Assumptions acceptable.
Heteroscedasticity 0.43860  0.5078 Assumptions acceptable.
```

The model passes the model diagnostic check.

Problem 6.21:

95% confidence intervals:

```
> #confidence intervals
> property.spec <- data.frame(X1 = c(4,6,12), X2 = c(10,11.5,12.5), X3= c(0.10,0,0.32),
+                              X4 = c(80000,120000,340000))
>
> property.conf <- predict(prop.lm, property.spec, interval = "confidence", level = 0.95, se.firmt = T)
> property.conf
       fit      lwr      upr
1 15.14850 14.76829 15.52870
2 15.54249 15.15366 15.93132
3 16.91384 16.18358 17.64410
```

$\mathbf{14.76829 \le E\{Y_h\} \le 15.52870}$

$\mathbf{15.15366 \le E\{Y_h\} \le 15.93132}$

$\mathbf{16.18358 \le E\{Y_h\} \le 17.64410}$

95% prediction intervals:

```
> #prediction interval
> property.pred <- predict(prop.lm, property.spec, interval = "prediction", level = 0.95, se.firmt = T)
> property.pred
       fit      lwr      upr
1 15.14850 12.85249 17.44450
2 15.54249 13.24504 17.83994
3 16.91384 14.53469 19.29299
```

$\mathbf{12.85249 \le E\{Y_{h(new)}\} \le 17.44450}$

$\mathbf{13.24504 \le E\{Y_{h(new)}\} \le 17.83994}$

$\mathbf{14.53469 \le E\{Y_{h(new)}\} \le 19.29299}$