512: Homework 3

Brody Vogel

2/8/2018

Preparation

```
set.seed(1234)
library(ISLR)
```

Problem 1 (#3)

a) [iii.] is correct. If the GPA is above 3.5, it looks like, the negative impact of the $\hat{\beta}_5$ coefficient on a female's salary outweighs the positive impact of the $\hat{\beta}_3$ coefficient.

b) $\hat{Y} = 50 + 20 \times (4) + .07 \times 110 + 35 \times (1) + .01 \times (4.0 \times 110) - 10 \times (4.0 \times 1) = \$137,100.$

c) False. The size of the coefficient doesn't tell us whether it is significant; the p-value associated with the interaction coefficient does, which is not provided here.

Problem 2 (#9e-f)

e) I first found the most highly-correlated predictor variables in the data set. These were: Cylinders, Horsepower, Displacement, and Weight. When I fit a model with all these predictors, though, there were no significant interaction effects, and the p-value for cylinders alone was .5. So I removed the cylinders predictor and ran things again. This time, the interaction between Horsepower and Displacement was highly significant (p = .000757).

```
data(Auto)
#cor(Auto[, -9])

fit.full <- lm(mpg~horsepower*displacement+horsepower*weight+displacement*weight, data = Auto)

summary(fit.full)
```

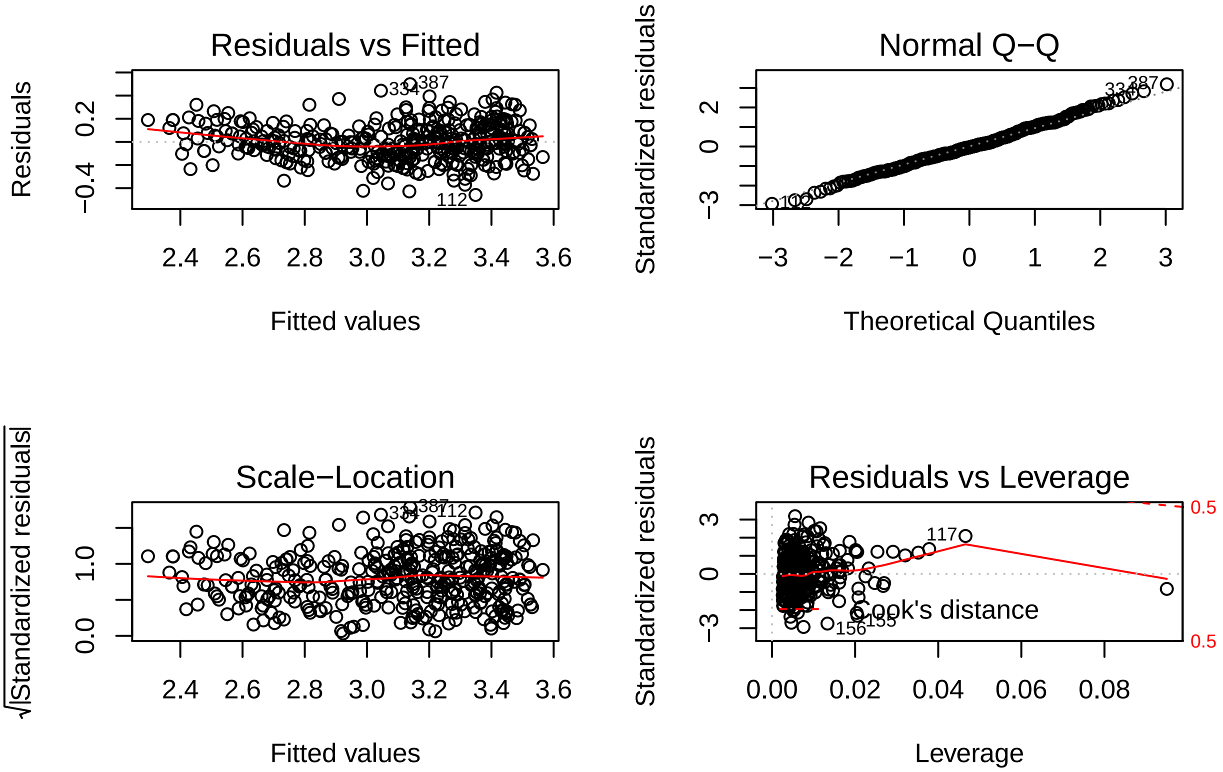
```
##
## Call:
## lm(formula = mpg ~ horsepower * displacement + horsepower * weight +
##     displacement * weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2588  -2.1593  -0.3806   1.8700  16.4104
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.665e+01  2.967e+00  19.093  < 2e-16 ***
## horsepower             -1.774e-01  5.271e-02  -3.366 0.000839 ***
## displacement           -7.609e-02  2.530e-02  -3.007 0.002811 **
## weight                  -4.612e-03  1.839e-03  -2.507 0.012578 *
## horsepower:displacement  3.779e-04  1.113e-04   3.395 0.000757 ***
## horsepower:weight        2.079e-06  1.873e-05   0.111 0.911649
## displacement:weight      6.225e-06  6.436e-06   0.967 0.334028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.875 on 385 degrees of freedom
## Multiple R-squared:  0.7574, Adjusted R-squared:  0.7536
## F-statistic: 200.3 on 6 and 385 DF,  p-value: < 2.2e-16
```

f) I first fit the model: $\hat{mpg} = log(horsepower) + \sqrt{displacement} + weight + weight^2$. Everything but the displacement term produced highly-significant p-values, but the residuals showed some heteroscedasticity, there were a few points with high leverage, and the Q-Q plot showed some non-normality among residuals. I then tried various transformations of the displacement variable - even leaving it out altogether - and still didn't get better results. I then moved the log transform over to the response variable, and this did produce better results, although the displacement term was still insignificant. Removing it again, I got my best model: $log(\hat{mpg}) = horsepower + weight$, which produced highly-significant p-values for both predictors and very good diagnostic plots.

```r
#full.fit1 <- lm(mpg~log(horsepower)+sqrt(displacement)+weight+I(weight^2), data = Auto)
full.fit1 <- lm(log(mpg)~horsepower+weight, data = Auto)
summary(full.fit1)
```

```
##
## Call:
## lm(formula = log(mpg) ~ horsepower + weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45865 -0.09660 -0.00648  0.10038  0.49885
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.1109472  0.0293674 139.983  < 2e-16 ***
## horsepower  -0.0025573  0.0004104  -6.231  1.2e-09 ***
## weight      -0.0002504  0.0000186 -13.462  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.157 on 389 degrees of freedom
## Multiple R-squared:  0.7879, Adjusted R-squared:  0.7869
## F-statistic: 722.7 on 2 and 389 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(full.fit1)
```



Problem 3 (#10)

a)

```
data(Carseats)
mod1 <- lm(Sales~Price+Urban+US, data = Carseats)
```

b)

Price: Based on the very low p-value, the model suggests the reponse, Sales, is negatively correlated with Price. As Price goes up, the model says Sales go down; for each unit increase in Price, it looks like the model predicts Sales to drop by ≈ 50.

Urban: Based on the high p-value, the model suggests there is no relationship between Sales and whether the store is in an Urban location.

US: Based on the very low p-value, the model suggests that Sales are positively correlated with a store being in the US. If a store is in the US, the model says it will have moe Sales; specifically, it looks like Sales will increase by ≈ 1,200 units.

```
summary(mod1)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936    
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335 
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

c) $\hat{Sales} = 13.04 - .054 \times Price - .022 \times Urban + 1.2 \times US$, assuming $Urban_{yes} = 1$ and $US_{yes} = 1$.

d) For Price and US, we can reject the null hypothesis $H_0 : B_j = 0$, because both have very low (and therefore significant) p-values, and because the model as a whole has a very low p-value corresponding to its F-statistic.

e) I just removed the Urban predictor, since it has an insignificant p-value.

```
mod2 <- lm(Sales~Price+US, data = Carseats)
```

f) Based on the R^2 values, neither of the models fit the data well. They appear to fit the data similarly - RSE of 2.472 for (a) and 2.469 for (e) - but neither capture anywhere close to the majority of the variance in the data. The R^2 from (a) is only .2393, which is identical to (e). So, essentially, the models perform equally poorly on the data, each missing over 3/4 of the variance in the data.

```
summary(mod2)
```

```
## 
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -6.9269 -1.6286 -0.0574  1.5766  7.0515 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354 
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

g)

$Intercept$: 11.7903 : 14.2712

$Price$: -.0648 : -.0442

US: .6915 : 1.7078

```
confint(mod2)
```

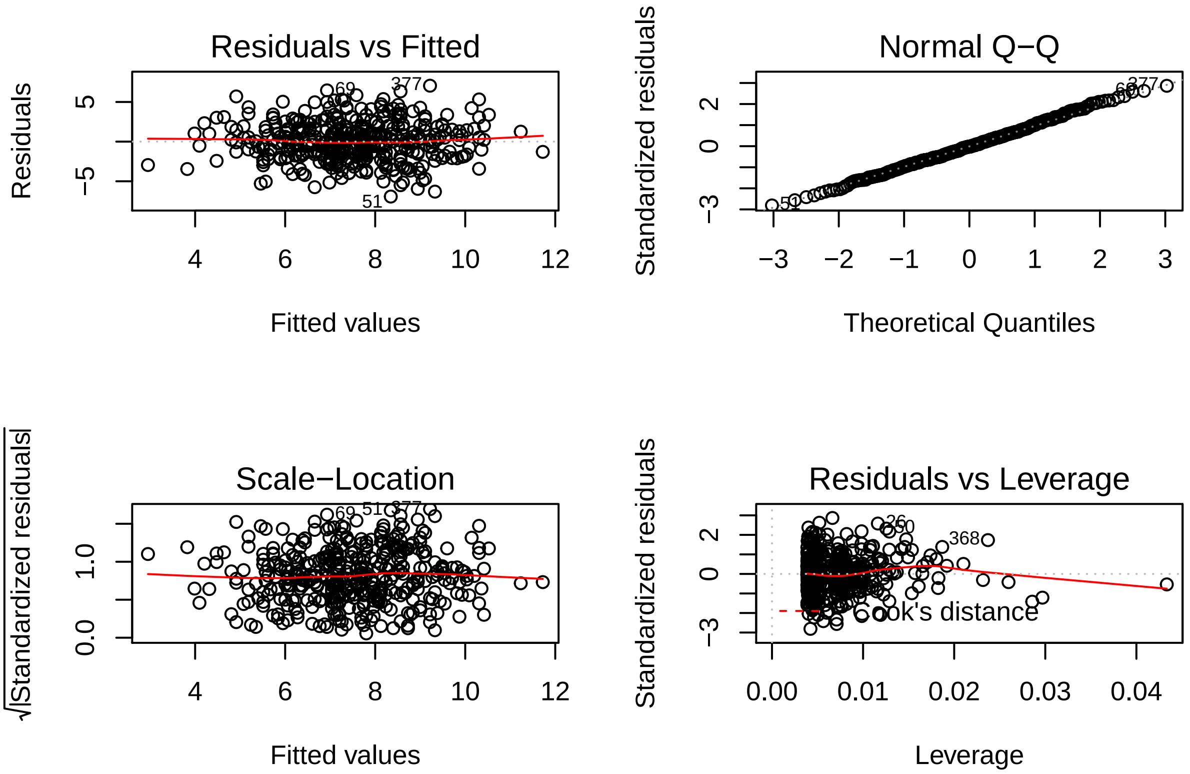
```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

h)

In the residuals plots, it looks like all the residuals stay within consistent bounds, and so there don't appear to be any outliers.

In the Residuals vs Leverage plot, though, there are a few points that stray very far from the far-left cluster, and so it would seem there are a couple points with very high leverage.

```
par(mfrow = c(2,2))
plot(mod2)
```



Problem 4 (#14a-c)

a) $Y = 2 + 2X_1 + .3X_2 + \epsilon$

$\beta_0 = 2; \beta_1 = 2; \beta_2 = .3$

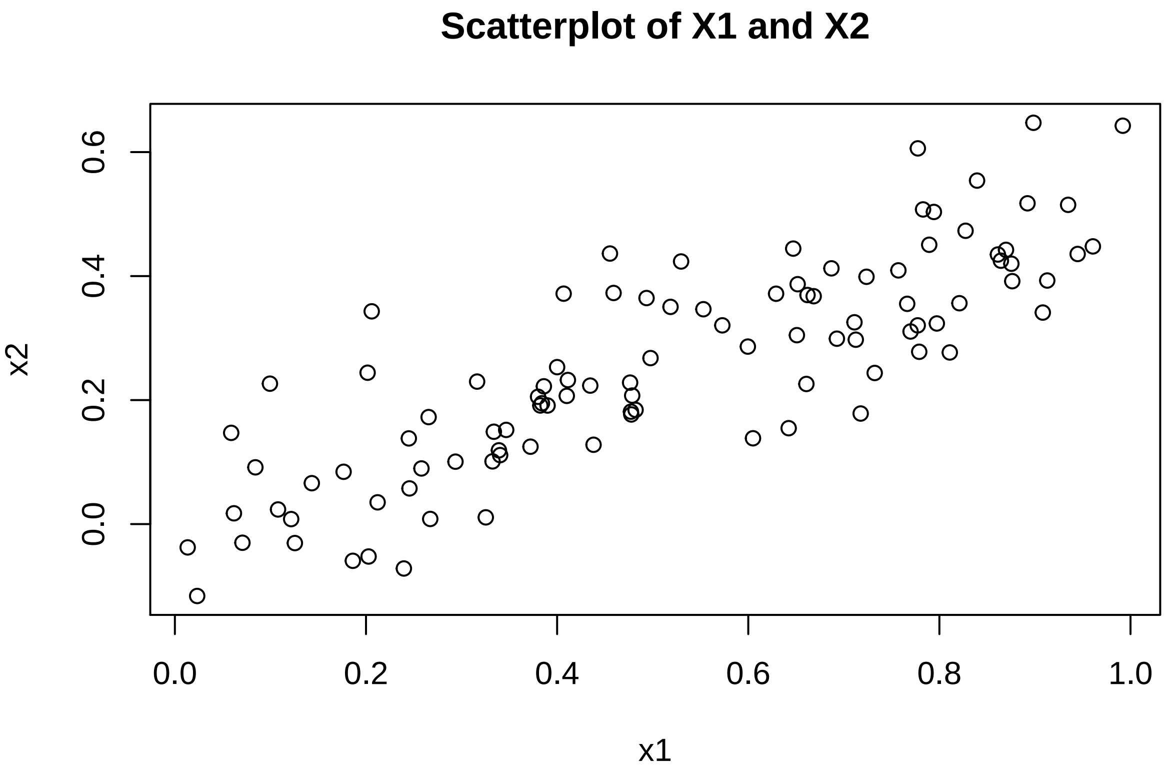
```
set.seed(1)
x1 <- runif(100)
x2 <- .5*x1 + rnorm(100)/10
y<- 2 + 2*x1 + .3*x2 + rnorm(100)
```

b) The correlation between x1 and x2 is $\approx .835$.

```r
cor(x1, x2)
```

```
## [1] 0.8351212
```

```r
plot(x1, x2, main = 'Scatterplot of X1 and X2')
```



c)

$\hat{\beta}_0 \approx 2.13; \hat{\beta}_1 \approx 1.44; \hat{\beta}_2 \approx 1.01$

Each estimate is somewhat close to the true coefficient, although the second and the third are comparatively far off. We can (barely) reject the null hypothesis $H_0 : \beta_1 = 0$, as it has a significantly low p-value of .0487; we cannot reject the null hypothesis $H_0 : \beta_2 = 0$, as it has a fairly high p-value of .3754.

```r
mod3 <- lm(y~x1+x2)
summary(mod3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```