HW5_BTindol

Blake Tindol

11/30/2020

Home Work 5

```
# Prep for questions

fit1 <- lm(y ~ x1 + x2 + x3, data = GroceryRetailer)
# Summary fitted values and residuals
summary(fit1); fit1$fitted.values[1:5]; fit1$residuals[1:5]
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = GroceryRetailer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -264.05 -110.73  -22.52   79.29  295.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.150e+03  1.956e+02  21.220  < 2e-16 ***
## x1           7.871e-04  3.646e-04   2.159   0.0359 *
## x2          -1.317e+01  2.309e+01  -0.570   0.5712
## x3           6.236e+02  6.264e+01   9.954 2.94e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.3 on 48 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6689
## F-statistic: 35.34 on 3 and 48 DF,  p-value: 3.316e-12
```

```
##        1        2        3        4        5
## 4296.063 4326.795 4338.825 4346.120 4869.066
```

```
##         1         2         3         4         5
## -32.06348 169.20509 -21.82543 -54.11955  75.93372
```

Question 1A)

Obtain the studentized deleted residuals and identify any outlying y observations. Use the case ID to identify them so we know which cases you have identified.

```r
s = summary(fit1)$sigma
residuals(fit1)[1:5]/s
```

```
##          1          2          3          4          5
## -0.2237672  1.1808621 -0.1523170 -0.3776939  0.5299324
```

```r
# completely studentized residuals
rstandard(fit1)[1:5] #first 5
```

```
##          1          2          3          4          5
## -0.2263377  1.2191338 -0.1723411 -0.3881136  0.5948397
```

```r
# Studentized Deleted Residuals
rstudent(fit1)[1:5] # first 5
```

```
##          1          2          3          4          5
## -0.2240872  1.2254901 -0.1705892 -0.3846535  0.5907924
```

```r
vec1 <- data.frame(y =GroceryRetailer$y, delresidual = rstudent(fit1))
vec1$absdelres <- abs(vec1$delresidual)
vec1[order(-vec1$absdelres),][1:5,]# no values over 3......
```

```
##       y delresidual absdelres
## 40 4555    2.178272  2.178272
## 38 4562    2.118786  2.118786
## 10 4560    2.036518  2.036518
## 32 3998   -1.997667  1.997667
## 34 4545    1.700417  1.700417
```

```r
# 1A) Answer!!!!!!!!!!!!!!!!!!!!
# Because there is no studentized deleted residuals over 3 there is nothing to look into but we will lo
# 40,38,10,32 34
```

Question 1B

Obtain the diagonal element of the hat matrix. Identify and outlying x observations using the rule of thumb discussed in class. Again use the case ID to identify them so we know which cases you have identified.

```r
lev = round(hatvalues(fit1),3)
vec1$lev <- lev;

# sum(hatvalues(fit1))

which(lev >= (2*3)/30)
```

```
##  3  5 16 22 43 44 48
##  3  5 16 22 43 44 48
```

```
vec1[order(-vec1$lev),][1:5,]
```

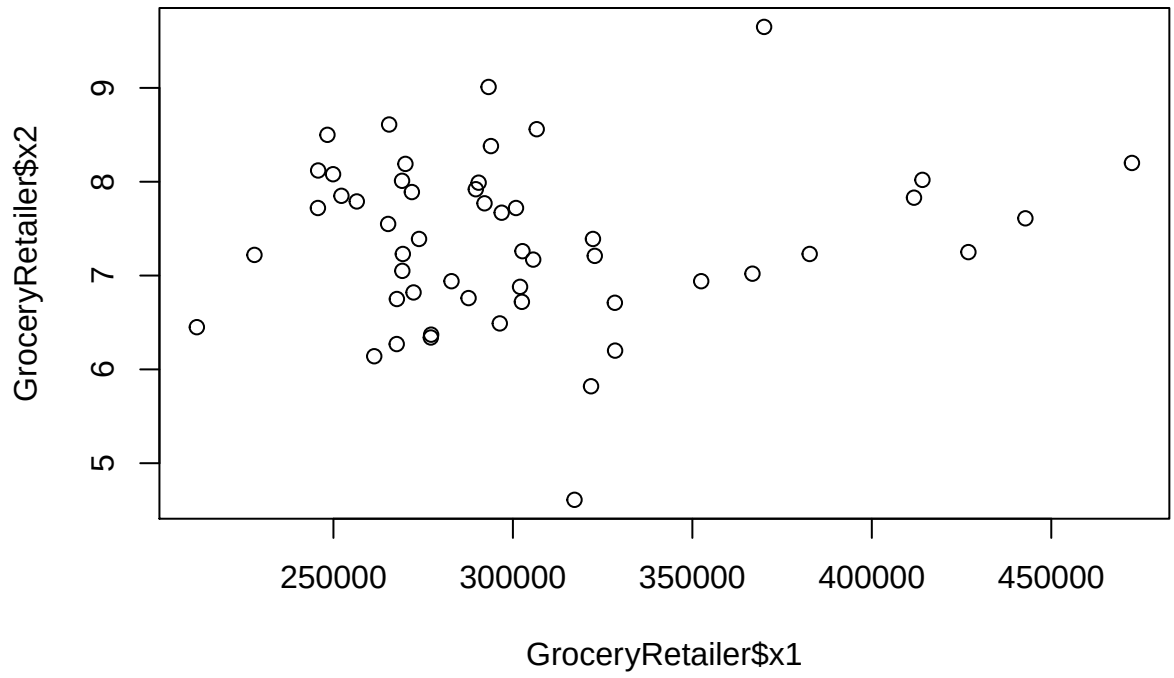
```
##       y delresidual  absdelres   lev
## 43 5045  0.88556630 0.88556630 0.287
## 48 4993 -0.23443689 0.23443689 0.282
## 22 4867  0.09348669 0.09348669 0.258
## 16 4833 -0.94585538 0.94585538 0.255
## 44 4469  0.43246959 0.43246959 0.220
```

```
# 1B) Answer: We have identified potential outlying x observations: (43,48,22,16,44) # they are the top
```

Question 1C

Suppose that a manger wishes to predict the total labor hours required to handle the next shipment containing Cases = 300,000 cases whose indirect costs of the total house is Costs = 7.2 and Holiday = 0, i.e. no holiday in week. Provide a scatter plot of Cases against Costs and determine visually whether this prediction involves an extrapolation beyond the range of the data.

```
plot(GroceryRetailer$x1,GroceryRetailer$x2)
```




```
new <- data.frame(x1=300000, x2=7.2,x3=0)
predict(fit1,new)
```

```
##        1
## 4291.216
```

```
#1C) Answer: I would say that there are visually this prediction fits into the model and does not invol
```

Question 1D

Obtain DFFITS and Cook's distance values for each of the outlying cases in part (a) and (b). What do these measures indicate about the influence of each of the cases?

```
## Calculate DFFITS
ddfits1 <- dffits(fit1)
vec1$ddfits <- ddfits1;
vec1[order(-vec1$absdelres),][1:5,]# order them in order of deleted studentized residuals in part a whe
```

```
##       y delresidual absdelres   lev     ddfits
## 40 4555    2.178272  2.178272 0.032  0.3967203
## 38 4562    2.118786  2.118786 0.032  0.3855177
## 10 4560    2.036518  2.036518 0.048  0.4586330
## 32 3998   -1.997667  1.997667 0.096 -0.6510771
## 34 4545    1.700417  1.700417 0.025  0.2732792
```

```
vec1[order(-vec1$lev),][1:5,] # for part b order them like this
```

```
##       y delresidual  absdelres   lev      ddfits
## 43 5045  0.88556630 0.88556630 0.287  0.56165186
## 48 4993 -0.23443689 0.23443689 0.282 -0.14684146
## 22 4867  0.09348669 0.09348669 0.258  0.05508583
## 16 4833 -0.94585538 0.94585538 0.255 -0.55399026
## 44 4469  0.43246959 0.43246959 0.220  0.22969393
```

```
## Calculate Cook's distance
cook1 <- cooks.distance(fit1)
vec1$cook <- cook1;
vec1[order(-vec1$absdelres),][1:5,] # for part a order them like this
```

```
##       y delresidual absdelres   lev     ddfits       cook
## 40 4555    2.178272  2.178272 0.032  0.3967203 0.03649915
## 38 4562    2.118786  2.118786 0.032  0.3855177 0.03463803
## 10 4560    2.036518  2.036518 0.048  0.4586330 0.04935012
## 32 3998   -1.997667  1.997667 0.096 -0.6510771 0.09975974
## 34 4545    1.700417  1.700417 0.025  0.2732792 0.01796257
```

```
vec1[order(-vec1$lev),][1:5,] # for part b order them like this
```

```
##       y delresidual  absdelres   lev      ddfits         cook
## 43 5045  0.88556630 0.88556630 0.287  0.56165186 0.0792193145
## 48 4993 -0.23443689 0.23443689 0.282 -0.14684146 0.0054988670
## 22 4867  0.09348669 0.09348669 0.258  0.05508583 0.0007746088
## 16 4833 -0.94585538 0.94585538 0.255 -0.55399026 0.0768950835
## 44 4469  0.43246959 0.43246959 0.220  0.22969393 0.0134170681
```

```
# 1D Answer: These observations could be potentially seen as influential because they may cause major c
```

1(e) Obtain Cook's distance for each case. Are any cases influential according to this measure?

```
## Calculate Cook's distance
## Using DFFITS to indentify influential observations
which(abs(dffits(fit1)) > 2) # no particular observation is significant
```

```
## named integer(0)
```

```
## Use Cook's distance to indentify influential observations
which(cooks.distance(fit1) > qf(.2, df1=4, df2=30-4))
```

```
## named integer(0)
```

```
# 1E) Answer: Using cooks distance for each case it does not look like there is any influential cases
```

2a) Fit the a regression model by using all predictor variables. Show the summary and a fitting regression line.

```
fit2 <- lm(height ~ exercise + male + dadht + momht, data = childHeight)
summary(fit2)
```

```
##
## Call:
## lm(formula = height ~ exercise + male + dadht + momht, data = childHeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6447  -1.4665   0.0499   1.5182   5.9399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.986458   4.688688   3.623 0.000392 ***
## exercise    -0.005263   0.041350  -0.127 0.898875
## male         5.311176   0.374290  14.190  < 2e-16 ***
## dadht        0.412572   0.051463   8.017 2.20e-13 ***
## momht        0.299162   0.069249   4.320 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 159 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6599, Adjusted R-squared:  0.6513
## F-statistic: 77.12 on 4 and 159 DF,  p-value: < 2.2e-16
```

```
#fit2$fitted.values
#plot(fit2)

# Answer 2A)equation is (height ~ 16.98645 + -0.0052*exercise + 5.311*male + 0.4127*dadht + 2.991*momht
```

2b) Based on the result in part (a), discuss why you decided to keep or drop variables.

```
# Answer2B): Based on summary exercise does not seem to be significant meaning no difference from 0 and
```

2 c) Conduct a F-test to determine whether the predictors in part (b) can be dropped from the regression model. Carefully state the null and alternative, test statistics and conclusion.

```
# Check each variable independently to see if they are a possible candidate to be dropped from model
mdl1 <- lm(height ~ exercise, data = childHeight) # exercise not significant
#mdl2 <- lm(height ~ dadht, data = childHeight)
#mdl3 <- lm(height ~ momht, data = childHeight)
#mdl4 <- lm(height ~ male, data = childHeight)
mdl1.sse =anova(mdl1)[2,2];mdl1.sse # Exercise is a potential candidate to be dropped
```

```
## [1] 2492.648
```

```
#mdl2.sse =anova(mdl2)[2,2];#mdl2.sse
#mdl3.sse =anova(mdl3)[2,2];#mdl3.sse
#mdl4.sse =anova(mdl4)[2,2];#mdl4.sse

# full model
full.mdl <- lm(height ~ exercise + male + dadht + momht, data = childHeight)
# reduced model
reduced.mdl1 <- lm(height ~ exercise, data =childHeight);summary(reduced.mdl1)# not significant this va
```

```
##
## Call:
## lm(formula = height ~ exercise, data = childHeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2015  -2.8695  -0.5105   2.6357   9.5535
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.2765     0.4436 149.402   <2e-16 ***
## exercise      0.1170     0.0683   1.713   0.0886 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.923 on 162 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.01779,    Adjusted R-squared:  0.01173
## F-statistic: 2.934 on 1 and 162 DF,  p-value: 0.08863
```

```
# Model Comparison with and without exercise
anova(reduced.mdl1, full.mdl)
```

```
## Analysis of Variance Table
##
## Model 1: height ~ exercise
## Model 2: height ~ exercise + male + dadht + momht
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    162 2492.65
## 2    159  863.15  3    1629.5 100.06 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# because of a significant p value we would reject the null hypothesis and keep exercise in the model?

#Answer2C):
# null hypotesis is that exercise is not significant than 0 meaning we do elimenate it from the model a
# the alternative hypotiseis is that exercsie is signiicant and we reject the null hypothesis and keep
```

2d) From part (b) and (c), fit a regression model by using predictors which you decided tokeep. Call this model Model D

```
model_d <- lm(height ~ exercise + male + dadht + momht, data = childHeight)
summary(model_d)
```

```
##
## Call:
## lm(formula = height ~ exercise + male + dadht + momht, data = childHeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6447  -1.4665   0.0499   1.5182   5.9399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.986458   4.688688   3.623 0.000392 ***
## exercise    -0.005263   0.041350  -0.127 0.898875
## male         5.311176   0.374290  14.190  < 2e-16 ***
## dadht        0.412572   0.051463   8.017 2.20e-13 ***
## momht        0.299162   0.069249   4.320 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 159 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6599, Adjusted R-squared:  0.6513
## F-statistic: 77.12 on 4 and 159 DF,  p-value: < 2.2e-16
```

2e) Obtain the studentized deleted residuals, the diagonal element of the hat matrix to identify any outlying observations. Choose an outlying case and explain why this case was flagged as unusual.

```
## Studentized deleted residuals
del <- rstudent(model_d)
err <- model_d$residuals
vec <- data.frame(del = del,err=err )
vec2<- vec[which(abs(err) > 3),]
vec2$absdel <- abs(vec2$del)
vec2[order(-vec2$absdel),]
```

```
##           del        err   absdel
## 130 -5.332221 -10.644729 5.332221
## 36   2.617004   5.939948 2.617004
## 16  -2.398970  -5.409392 2.398970
## 108 -2.330692  -5.259263 2.330692
## 104 -2.272546  -5.137287 2.272546
## 38   2.267616   5.152386 2.267616
## 23   2.241744   5.078257 2.241744
## 133  2.134069   4.858664 2.134069
## 45   1.982510   4.539389 1.982510
## 51   1.869352   4.273498 1.869352
## 18  -1.781972  -4.024548 1.781972
## 41   1.753830   3.987757 1.753830
## 19  -1.708727  -3.854956 1.708727
## 132 -1.682030  -3.862657 1.682030
## 3    1.638812   3.749303 1.638812
## 81   1.622119   3.606324 1.622119
## 159 -1.619614  -3.715235 1.619614
## 58  -1.583662  -3.634252 1.583662
## 14   1.476878   3.399662 1.476878
## 48   1.474273   3.374314 1.474273
## 73  -1.474140  -3.388825 1.474140
## 4   -1.472661  -3.360262 1.472661
## 124 -1.415554  -3.264333 1.415554
## 91  -1.414233  -3.245434 1.414233
## 141 -1.403095  -3.229204 1.403095
## 139 -1.380656  -3.182502 1.380656
## 156  1.367948   3.109979 1.367948
## 78   1.338270   3.082042 1.338270
## 103 -1.316627  -3.011942 1.316627
```

```
#Answer 2e) Part1: (Outlying y obersvations) Observation 130 is an outlier because the absolute value o
```

```
# Outling x obeservations!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!1
# Hii in matrix h
lev = round(hatvalues(model_d),3)
which(lev >= (2*5)/165) # these larger values indicate that the value is distant from the ith observati
```

```
##  20  21  77  81  86  99 122 130 137
##  20  21  77  81  86  99 122 130 136
```

```r
#observations (20,21,77,81,86,99,122,130,136,137)
outliers <- c('20','21','77','81','86','99','122','130','136','137')
vec[outliers,] # filter for the values that are outliers
```

```
##             del         err
## 20   0.80096435   1.7920757
## 21   0.68035211   1.5163654
## 77   0.33866084   0.7494227
## 81   1.62211909   3.6063239
## 86   0.12695394   0.2767036
## 99   0.99020686   2.0315766
## 122 -1.08411366  -2.4174905
## 130 -5.33222099 -10.6447285
## 136  0.05759748   0.1335946
## 137  0.98504208   2.1852927
```

```r
# Answer2e) Part2: Outlying x obersvations: (top 10) '20','21','77','81','86','99','122','130','136','1
```

2f) Obtain DFFITS and Cook's distance values for each of the outlying cases in part (e).#What do these measures indicate about the influence of each of the cases?

```r
## Calculate DFFITS
dff <- dffits(model_d)
## Calculate Cook's distance
cook <- cooks.distance(model_d)
vec3 <- data.frame(dff = dff,cook=cook);
vec3[outliers,] # THE COOKS AND DFITS FOR THE previous part e values
```

```
##              dff         cook
## 20   0.236091415 1.117302e-02
## 21   0.211374087 8.966089e-03
## 77   0.114736784 2.647648e-03
## 81   0.478885600 4.540049e-02
## 86   0.049147120 4.860958e-04
## 99   0.533072999 5.684033e-02
## 122 -0.326157777 2.125235e-02
## 130 -2.144496414 7.844332e-01
## 136  0.007154168 1.030099e-05
## 137  0.316465990 2.003389e-02
```

```r
## Using DFFITS to indentify influential observations
which(abs(dffits(model_d)) > 1)
```

```
## 130
## 130
```

```r
## Use Cook's distance to indentify influential observations
which(cooks.distance(fit2) > qf(.2, df1=3, df2=165-3))
```

```
## 130
## 130
```

```r
# Answer 2f): This cases might be influential in impacting the regression ID(130)
```

2g) Discuss whether any of the identified cases in part (e) and (f) should be removed from the analysis.

```r
# Looks like the case 130 should be removed from the analysis because it does affect the regression ana
```

2h) Find the Variance Inflation Factor (VIF) for each of the predictors in Model D. Based on the these VIF values, what do you suggest about what predictor variable(s) to include in your model?

```r
cor(childHeight[2:7],use = "complete.obs")
```

```
##             alcohol   exercise    height        male       dadht      momht
## alcohol   1.0000000 0.26337501 0.4351863  0.30555937  0.28953570 0.23658530
## exercise 0.2633750 1.00000000 0.1162157  0.20305276  0.03827115 0.01321576
## height    0.4351863 0.11621573 1.0000000  0.65179376  0.40524354 0.34010950
## male      0.3055594 0.20305276 0.6517938  1.00000000 -0.05971473 0.03019804
## dadht     0.2895357 0.03827115 0.4052435 -0.05971473  1.00000000 0.29798601
## momht     0.2365853 0.01321576 0.3401095  0.03019804  0.29798601 1.00000000
```

```r
vif(model_d)
```

```
## exercise     male    dadht    momht
## 1.038831 1.047991 1.125425 1.114388
```

```r
#Answer 2h): no vif is greater than 10 meaning there is no indicator of multicorilinearity
```