Multiple Regression Analysis

Andrew Wang

11/13/2021

Question 1 : Prediction from Multiple Regressions

Q1, part A

```
library(DataAnalytics)
data("multi")
mr = lm(Sales~p1+p2, data = multi)
lmSumm(mr)
```

```
## Multiple Regression Analysis:
##     3 regressors(including intercept) and 100 observations
##
## lm(formula = Sales ~ p1 + p2, data = multi)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)   115.70     8.548   13.54       0
## p1            -97.66     2.669  -36.60       0
## p2            108.80     1.409   77.20       0
## ---
## Standard Error of the Regression:  28.42
## Multiple R-squared:  0.987  Adjusted R-squared:  0.987
## Overall F stat: 3717.29 on 2 and 97 DF, pvalue= 0
```

Q1, part B

Setting p2=mean(p2) would be a bad choice, because that makes the assumption that p1 and p2 are uncorrelated with each other. However, based on what we can see from the simple linear regression below on p1, we can see that there is huge difference in the coefficients for p1 once we introduce p2 into the multiple regression (compare to part A). This is known as a confounding of effects, and so p1 and p2 must be correlated.

```
slr = lm(Sales~p1, data = multi)
lmSumm(slr)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 100 observations
##
## lm(formula = Sales ~ p1, data = multi)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)   211.20     66.49    3.18   0.002
## p1             63.71     13.04    4.89   0.000
```

```
## ---
## Standard Error of the Regression:  223.4
## Multiple R-squared:  0.196  Adjusted R-squared:  0.188
## Overall F stat: 23.87 on 1 and 98 DF, pvalue= 0
```

In addition, based on the summary of p1 data below we can see that setting p1=7.5 strays far away from its actual mean of about 4.8. Therefore, at p1=7.5 the value for p2 would most likely be quite different from mean(p2) since p1 and p2 have some correlation between them.

```
summary(multi$p1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4724  3.8902  4.8167  4.8023  5.7495  8.8071
```

Q1, part C

```
p2p1 = lm(p2~p1, data = multi)
lmSumm(p2p1)
```

```
## Multiple Regression Analysis:
##     2 regressors(including intercept) and 100 observations
##
## lm(formula = p2 ~ p1, data = multi)
##
## Coefficients:
##             Estimate Std Error t value p value
## (Intercept)   0.8773    0.6062    1.45   0.151
## p1            1.4830    0.1189   12.48   0.000
## ---
## Standard Error of the Regression:  2.037
## Multiple R-squared:  0.614  Adjusted R-squared:  0.61
## Overall F stat: 155.63 on 1 and 98 DF, pvalue= 0
```

Based on the regression above, we can predict p2 given p1=7.5: $\hat{Y} = 0.8773 + 1.4830(p1) = 0.8773 + 1.4830(7.5) = 12$. So p2 is approximately equal to \$12 when p1=7.5.

Q1, part D

Based on our above regression equation of Sales on p1 and p2, we can predict Sales as follows: $\hat{Y} = 115.70 + (-97.66)(p1) + (108.80)(p2) = 115.70 + (-97.66)(7.5) + (108.80)(12) = 688.85$. So, we predict Sales to be roughly 688.85 based on our estimated values for p2 and p1.

This is quite similar to our predicted value of Sales when regressing on just p1: $\hat{Y} = 211.20 + 63.71(p1) = 211.20 + 63.71(7.5) = 689.025$. This must be true because the simple linear regression of just p1 takes into account the co-movement of p1 along with all other variables (like p2). That's why the multiple regression and simple linear regression of Sales would predict the same value as each other.

Question 2: Interactions

```
library(DataAnalytics)
data("mvehicles")
cars = mvehicles[mvehicles$bodytype != "Truck",]
cars_mr = lm(log(emv)~luxury+sporty+luxury*sporty, data = cars)
lmSumm(cars_mr)
```

```
## Multiple Regression Analysis:
##     4 regressors(including intercept) and 1395 observations
```

```
##
## lm(formula = log(emv) ~ luxury + sporty + luxury * sporty, data = cars)
##
## Coefficients:
##               Estimate Std Error t value p value
## (Intercept)     9.7350   0.04385  222.02       0
## luxury          1.3220   0.10900   12.12       0
## sporty         -0.4096   0.11600   -3.53       0
## luxury:sporty   1.2930   0.22210    5.82       0
## ---
## Standard Error of the Regression:  0.3122
## Multiple R-squared:  0.588  Adjusted R-squared:  0.587
## Overall F stat: 662.49 on 3 and 1391 DF, pvalue= 0
```

Q2, part A

We first hold luxury constant at .3 units. Let's first assume sporty's value starts equals its mean of around 0.4 before we measure the change of increasing it by .1 units:

$log(emv) = 9.735 + 1.322(.3) + (-0.4096)(.4) + (1.2930)(.3)(.4) = 10.12292$

$emv = e^{10.12292} = 24907.39$

Now we compute the change in emv we would expect if sporty increased by .1 units:

$log(emv) = 9.735 + 1.322(.3) + (-0.4096)(.5) + (1.2930)(.3)(.5) = 10.12075$

$emv = e^{10.12075} = 24853.40$

Now we compute the change in emv:

$24853.40 - 24907.39 = -53.99$

Therefore, holding luxury constant .30 units and increasing sporty by .1 units, we would expect there to be a 53.99 decrease in emv.

Q2, part B

We first hold luxury constant at .7 units. Let's first assume sporty's value starts equals its mean of around 0.4 before we measure the change of increasing it by .1 units:

$log(emv) = 9.735 + 1.322(.7) + (-0.4096)(.4) + (1.2930)(.7)(.4) = 10.8586$

$emv = e^{10.8586} = 51979.26$

Now we compute the change in emv we would expect if sporty increased by .1 units:

$log(emv) = 9.735 + 1.322(.7) + (-0.4096)(.5) + (1.2930)(.7)(.5) = 10.90815$

$emv = e^{10.90815} = 54619.71$

Now we compute the change in emv:

$54619.71 - 51979.26 = 2640.45$

Therefore, holding luxury constant .70 units and increasing sporty by .1 units, we would expect there to be a 2640.45 increase in emv.

Q2, part C

The answers in part A and part B are different due to the effect of the interaction term which moderates the effect of luxury on the emv. In part A, the luxury constant was held at 0.3 while in part B, the luxury constant was held at 0.7 instead. A luxury value of 0.3 units was more limiting to influencing the effect

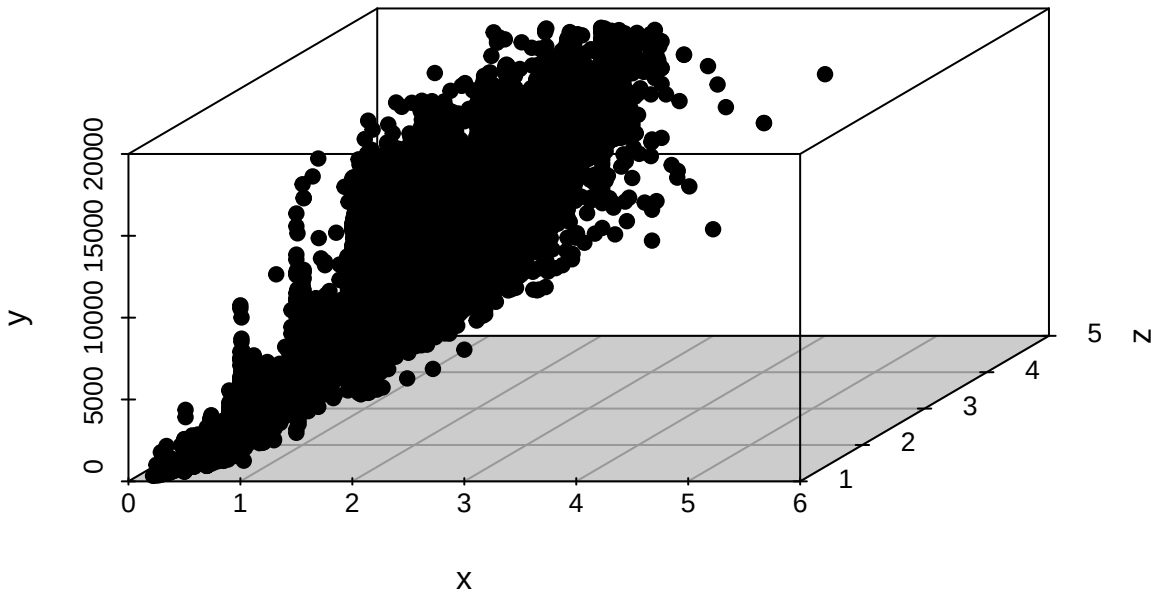
of changing the level of sporty. We can see this from taking the partial derivative of the conditional mean function as $\frac{\partial}{\partial X1} E[Y|X1, X2] = \beta_1 + \beta 3X2$, and so the interaction term makes intuitive sense to me.

Question 3: More on ggplot2 and regression planes

```
library(ggplot2)
data(diamonds)
cutf=as.character(diamonds$cut)
cutf=as.factor(cutf)
x = diamonds$carat
z = cutf
y = diamonds$price
```

1. Below, we regress price on both carat and cut and then begin to visualize the relationship between price and the two independent variables of carat and cut.

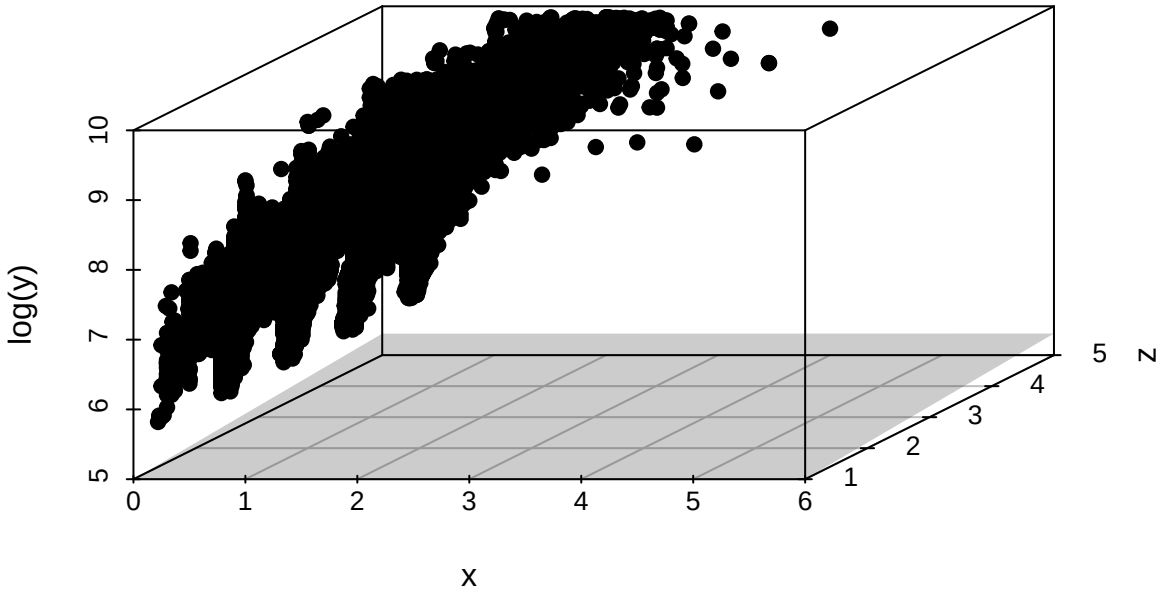
```
lm_diamonds = lm(y ~ x + z, data=diamonds)

library(scatterplot3d)
s3d <- scatterplot3d(x, z, y, pch = 19, type="p")

s3d$plane3d(lm_diamonds, draw_polygon = TRUE, draw_lines = FALSE)
```



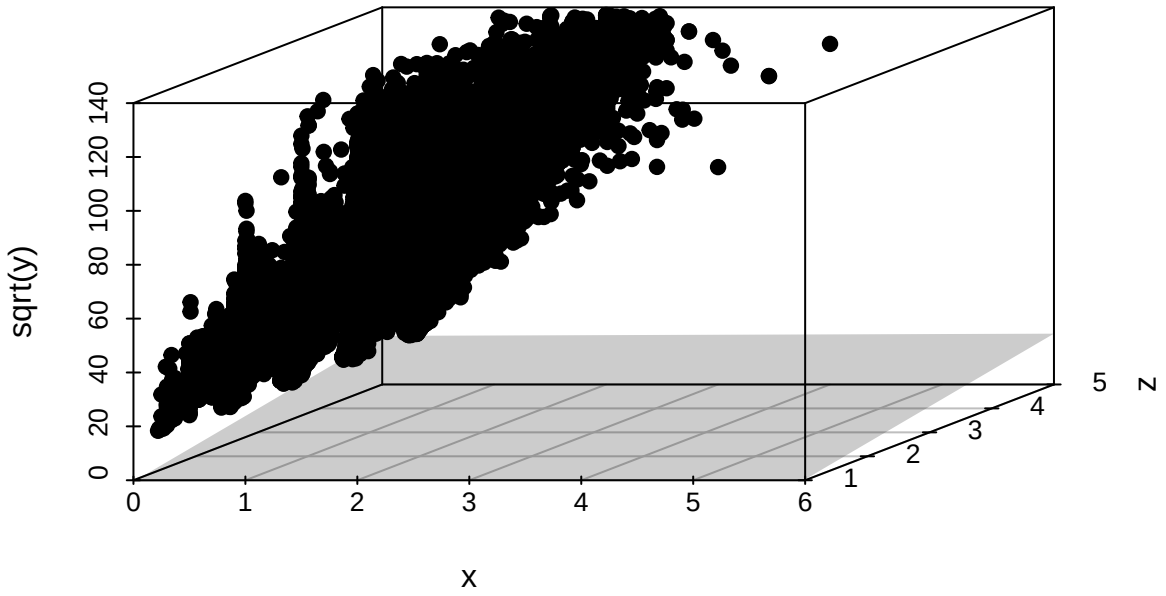
In addition, we also visualize the log() and sqrt() transformations below:

```
lm_diamonds_log = lm(log(price) ~ x + z, data=diamonds)

s3d_log <- scatterplot3d(x, z, log(y), pch = 19, type="p")
```

```
s3d$plane3d(lm_diamonds_log, draw_polygon = TRUE, draw_lines = FALSE)
```

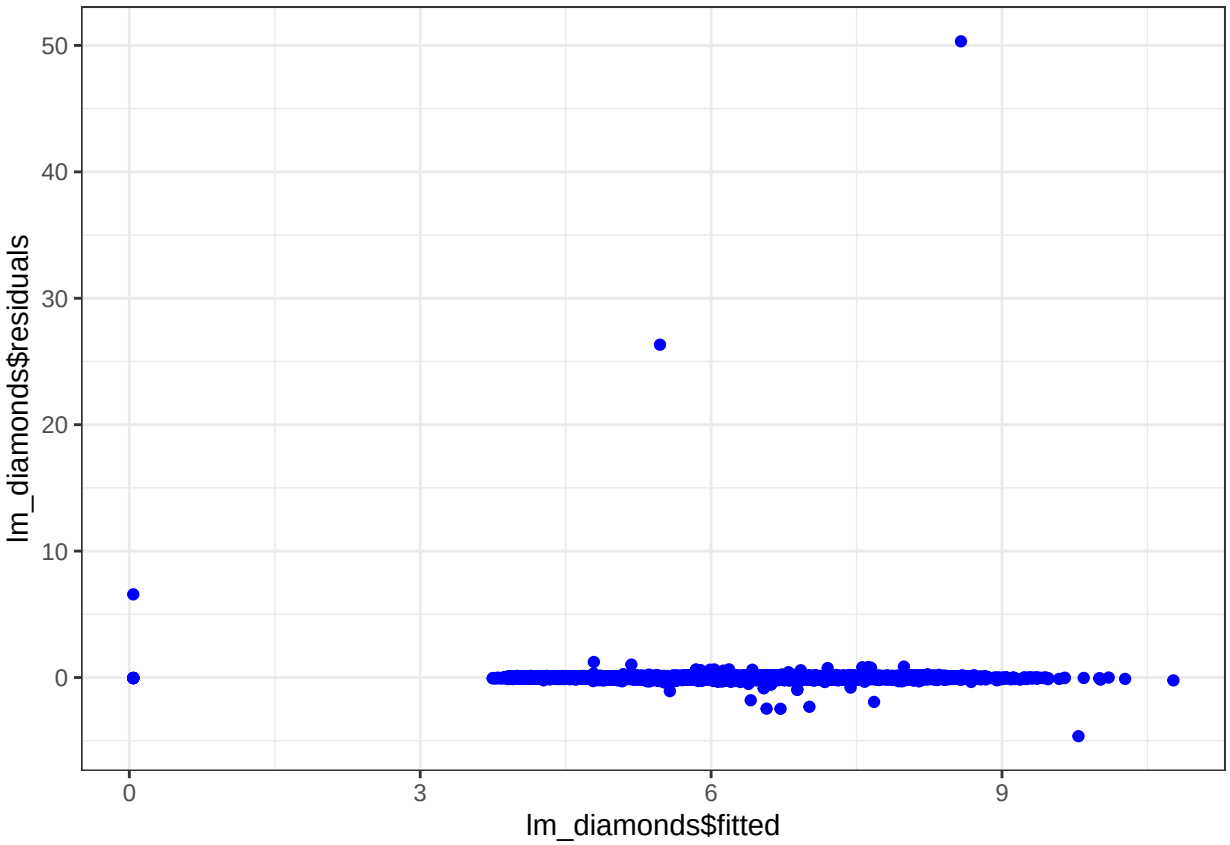



```
lm_diamonds_sqrt = lm(sqrt(price) ~ x + z, data=diamonds)

s3d_log <- scatterplot3d(x, z, sqrt(y), pch = 19, type="p")

s3d$plane3d(lm_diamonds_sqrt, draw_polygon = TRUE, draw_lines = FALSE)
```

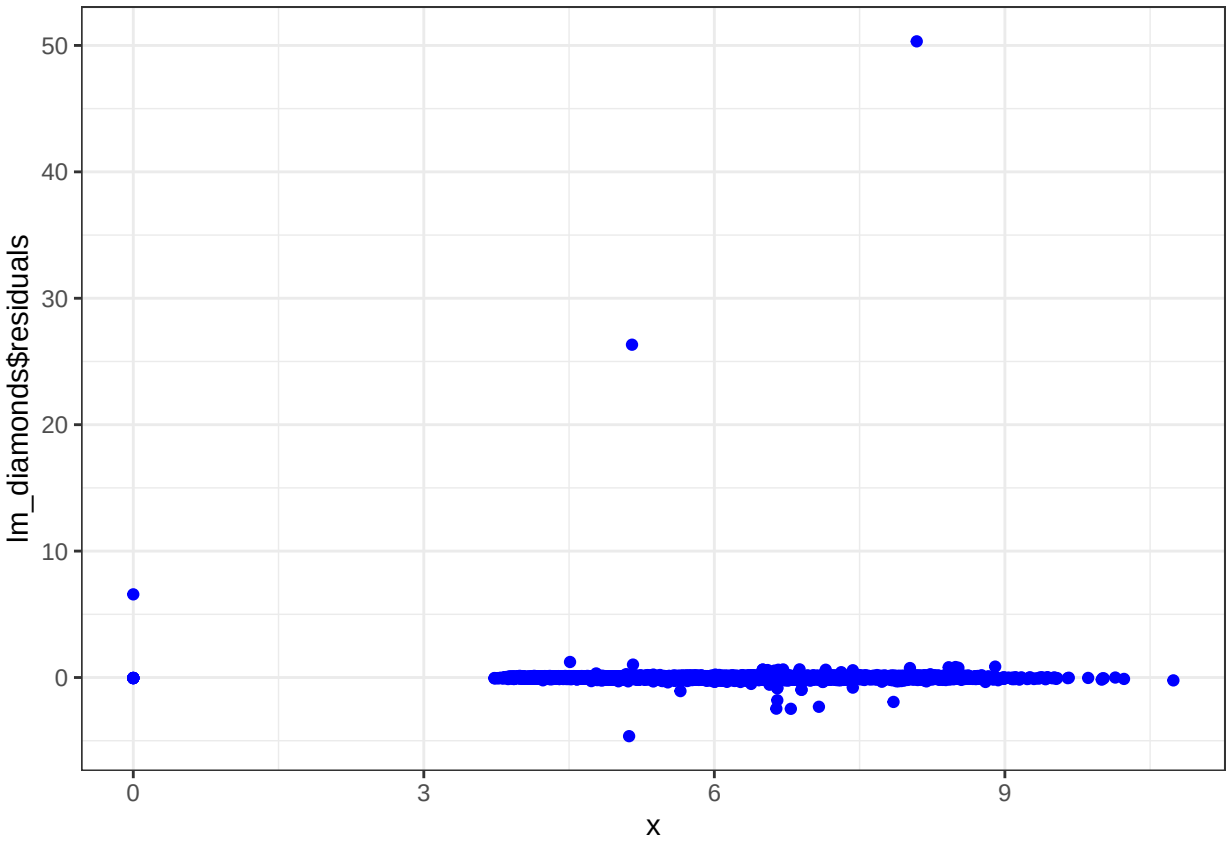


2. Below, we run the regular regression and perform residual diagnostics. Based on what we see below, we can conclude that there is not relationship between the residuals and the carat variable.

```
qplot(lm_diamonds$fitted,lm_diamonds$residuals,color=I("blue")) + theme_bw()
```




```
qplot(x,lm_diamonds$residuals,data=lm_diamonds,color=I("blue"))+theme_bw()
```




```
corr(data.frame(lm_diamonds$residuals, diamonds$carat))
```

```
## Full Correlation Matrix
##                        lm_diamonds.residuals diamonds.carat
## lm_diamonds.residuals                     1              0
## diamonds.carat                            0              1
##
## Correlation Matrix trimmed with cutoff = 0.25
##                         lm_diamonds.residuals  diamonds.carat
## lm_diamonds.residuals  1.00
## diamonds.carat                                1.00
```