STA 210: Lab 3

Jackson Hubbard

September 17, 2018

Question 1

There are 53,940 observations in the diamonds dataset

```r
glimpse(diamonds)
```

```
## Observations: 53,940
## Variables: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, ...
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very G...
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, ...
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI...
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, ...
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54...
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339,...
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, ...
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, ...
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, ...
```

Question 2

a

Because this ensures that the sample is random, which is a condition for independence which is necessary to do a proper experiment with no bias. The data could have been entered in order by a certain characteristic, such as diamond size, so selecting the first 1000 observations would not be representative of the poultation. Further, doing a simple random sample is very easy to do

b

```r
set.seed(123)
diamonds_samp <- sample_n(diamonds,1000,replace=F)
```

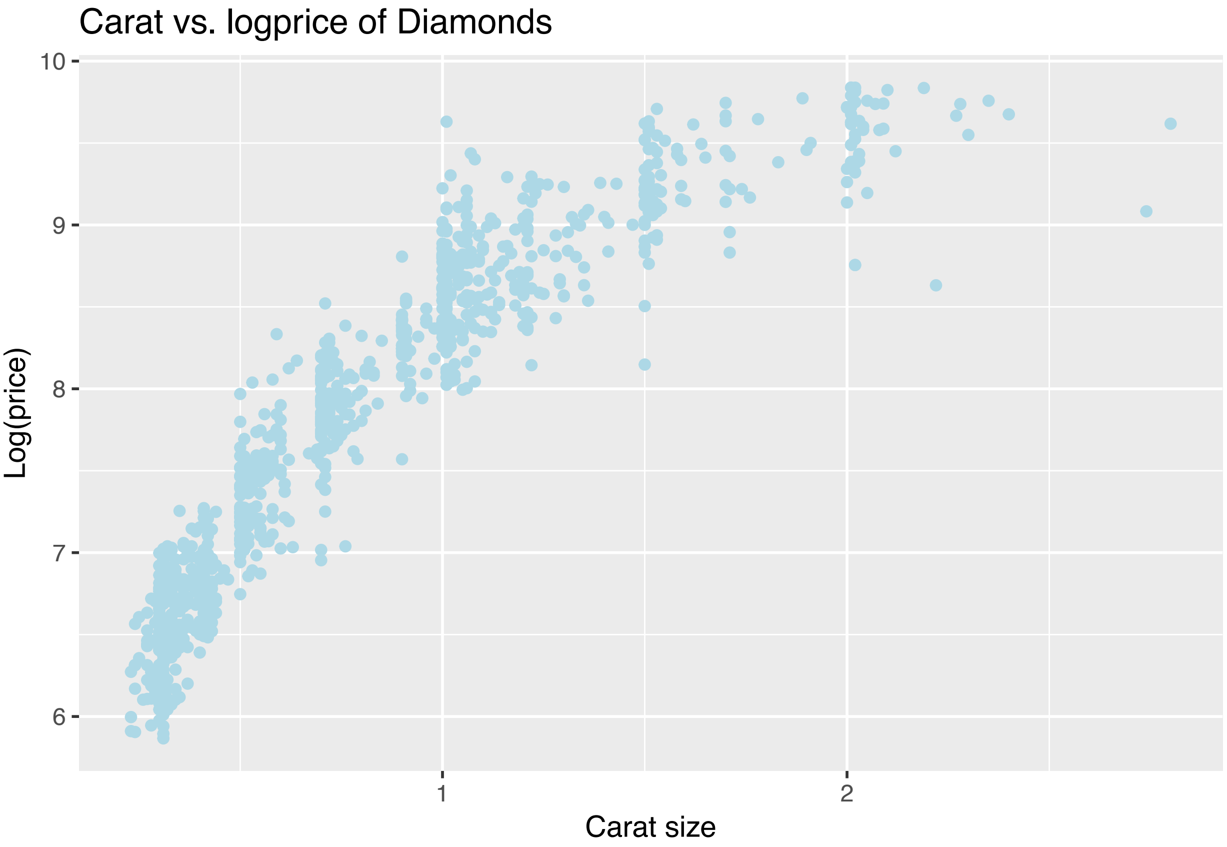
Question 3

```
diamonds_samp <- diamonds_samp %>% mutate(logprice = log(diamonds_samp$price))
```

Question 4

a

```
ggplot(diamonds_samp, aes(x = diamonds_samp$carat,
  y = diamonds_samp$logprice)) + geom_point(color = "light blue") +
  labs(title = "Carat vs. logprice of Diamonds", x= "Carat size",
       y = "Log(price)")
```



b

A linear model is probably not appropriate to model carat and logprice beacuse looking at the scatterplot there is a curve downward as the carat size increases. There is a linear relationship between carat and price, but the relationship seems to weaken as carat gets greater than 1. It is likely that transforming the carat variable will enable us to better fit the model.

Question 5

```
model1 <- lm(logprice ~ carat, data = diamonds_samp)
tidy(model1) #output model results
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)     6.25    0.0253     247.        0
## 2 carat           1.93    0.0268      71.9       0
```

a

$y = 1.931567x + 6.253803$
$\log(\text{price}) = 1.931567\ (\text{carat}) + 6.253803$

b

For every one unit of carat, there is a multiplicative change of e^(1.931567), which is 6.90031456711, in the median price of the diamond.

c

No it does not make sense to interpret the intercept in the context of this problem. This is because a diamond cannot have 0 carats.

Question 6

The R2 is 0.8383469. This means that 83.83469% of the variability in price is explained by the regression line.

```
glance(model1)
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
## *     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.838         0.838 0.405     5176.       0     2  -514. 1033. 1048.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

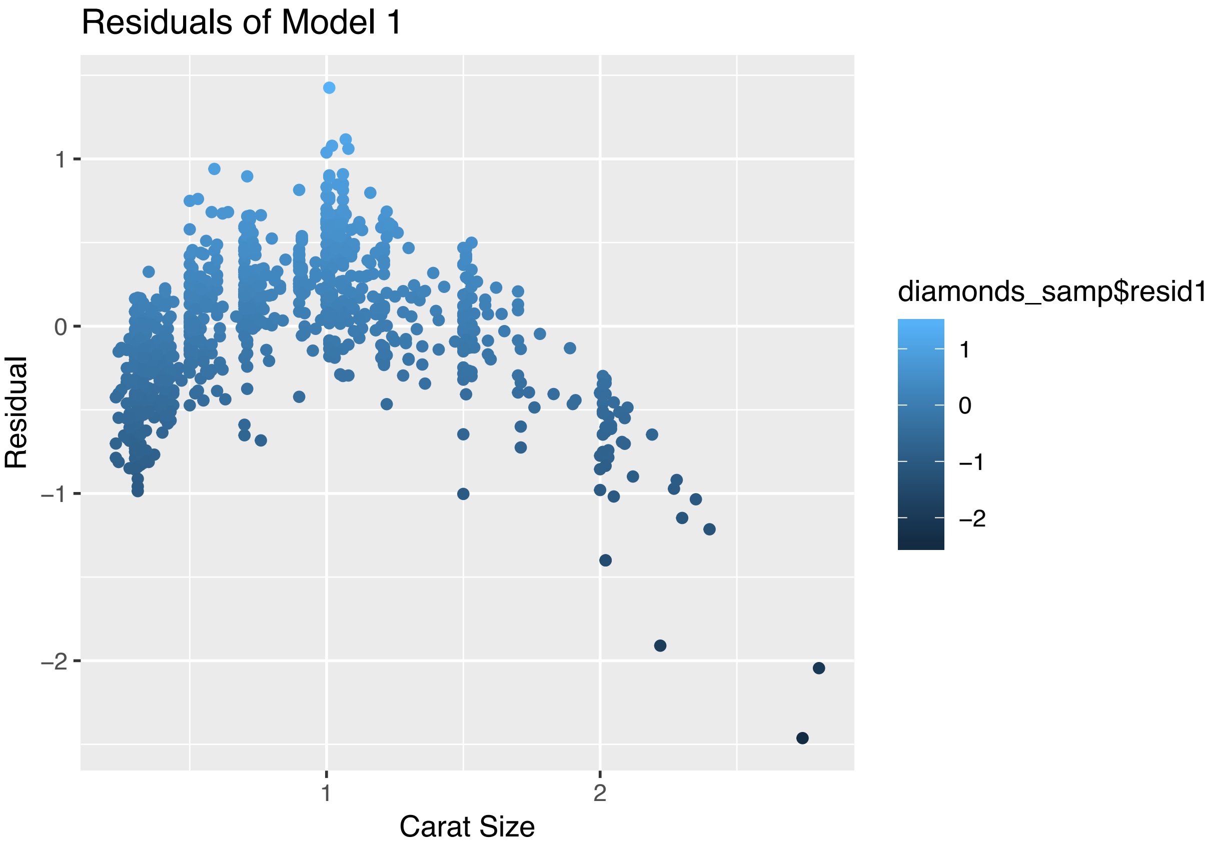
Question 7

a

```
resid1 <- resid(model1)
diamonds_samp <- diamonds_samp %>% mutate(resid1)
```

b

```
ggplot(diamonds_samp, aes(x = diamonds_samp$carat, y = diamonds_samp$resid1)) +
  geom_point(aes(color=diamonds_samp$resid1 )) +
  labs(title= "Residuals of Model 1", x = "Carat Size", y= "Residual" )
```



c

No the residual plot does not have constant variance. The value of the residual clearly depends on the value of the explanatory variable (carat). Hence, we will need to transform our model in order to accurately describe the relationship between carat and logprice.

Question 8

a

```
diamonds_samp <- diamonds_samp %>%
  mutate(carat2 = diamonds_samp$carat *diamonds_samp$carat)
```

b

```
model2 <- lm(logprice ~ carat + carat2, data=diamonds_samp)
tidy(model2)
```

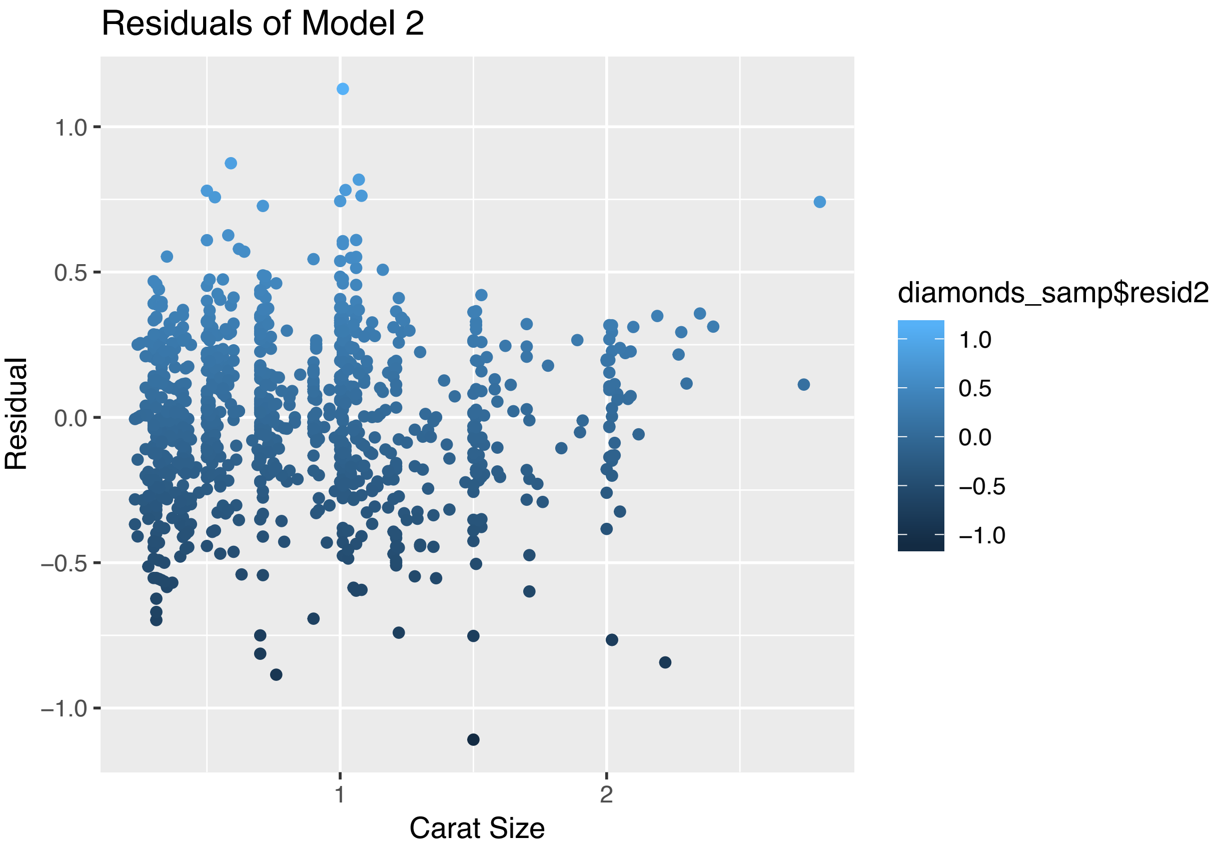
```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     5.39    0.0300     179.  0.
## 2 carat           4.12    0.0649      63.5 0.
## 3 carat2         -1.03    0.0293     -35.1 2.71e-176
```

c

```
resid2 <- resid(model2)
diamonds_samp <- diamonds_samp %>% mutate(resid2)
```

d

```
ggplot(diamonds_samp, aes(x = diamonds_samp$carat, y = diamonds_samp$resid2)) +
  geom_point(aes(color=diamonds_samp$resid2 )) +
  labs(title= "Residuals of Model 2", x = "Carat Size", y= "Residual" )
```



e

Yes model2 adequately describes the relationship between carat and logprice beacuse the variability of the residuals is now relatively constant. Just like the last model, model 2 still fits the assumptions of independence, linearity, and normality.

Question 9

We expect the price of a 0.75 carat diamond to be e^7.898352 which is $2692.84.

```r
newcarat <- 0.75
newdata=data.frame(carat=newcarat,carat2=newcarat^2)
#Note: the variable name(s) in the data frame must match the variable name(s) in the m
```

```r
predict.lm(model2,newdata)
```

```
##        1
## 7.898352
```