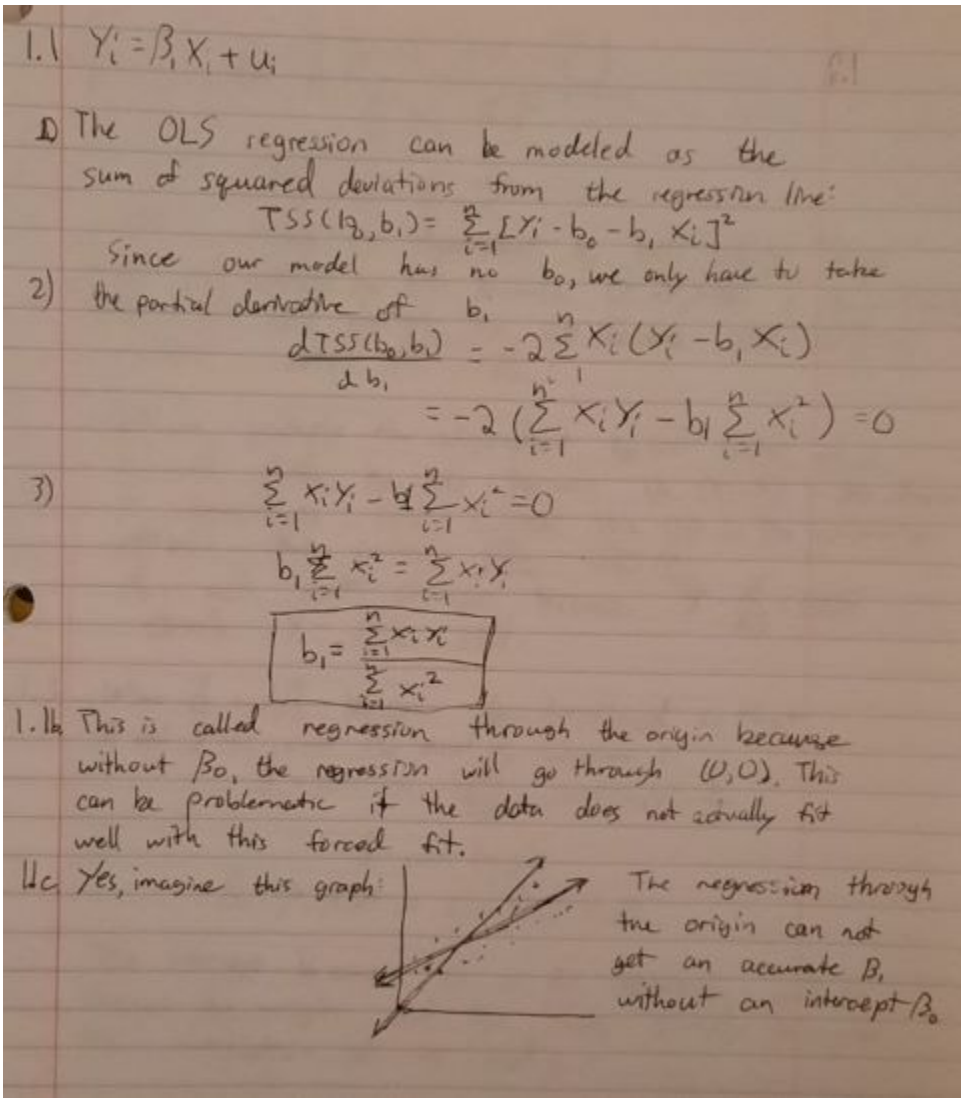
ECON-UB 251 Econometrics I Assignment 1

Kevin Song

Theory

1.1

1. Derive the OLS estimator of β_1



1.2

2. What happens to the OLS estimators if we multiply the independent or dependent variables by 100?

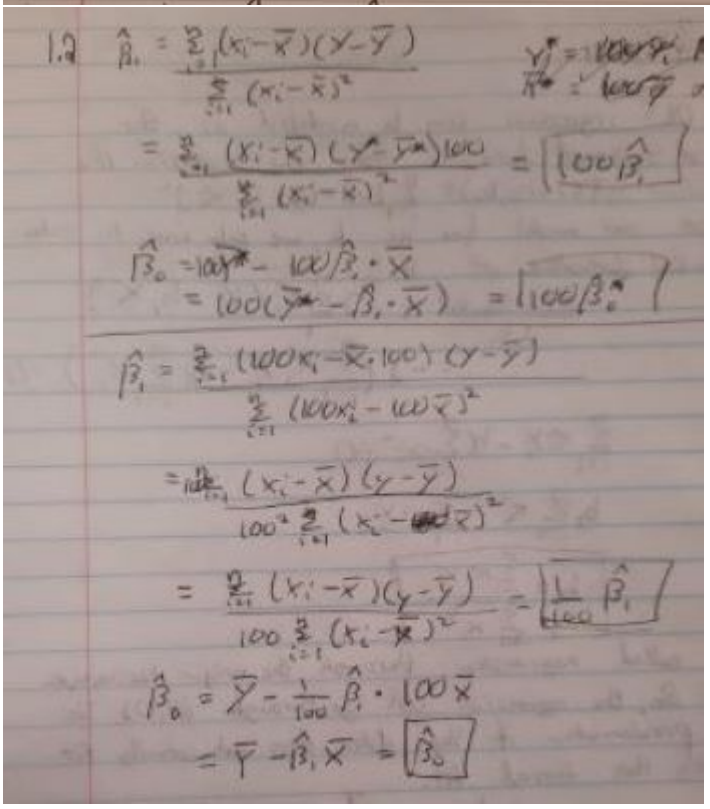
1.2

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

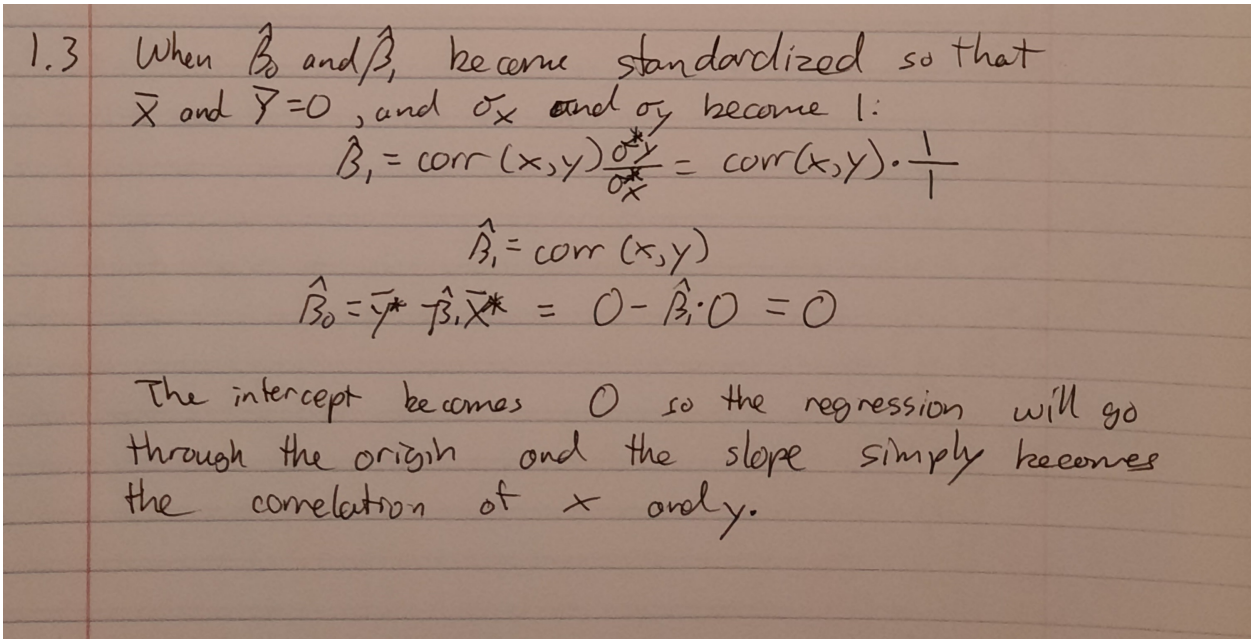
If we multiply the dependent by 100: $100Y_i$, then
$\hat{\beta}_1$ will multiply by 100 because $(Y_i - \bar{Y}) \cdot 100$ in the numerator.
$\hat{\beta}_0$ will also multiply by 100 because $100\bar{Y} - 100\hat{\beta}_1 \cdot \bar{X}$ equals $100(\bar{Y} - \hat{\beta}_1 \bar{X}) = \hat{\beta}_0$

If we multiply the independent by 100: $100X_i$, then
$\hat{\beta}_1$ will divide by 100 because the $(X_i - \bar{X})$ in the denom is squared, so $100^2 = 10000$. The 100 in the numerator offsets this partially: $\frac{100}{10000} = \frac{1}{100}\hat{\beta}_1$.
$\hat{\beta}_0$ will not change because $\bar{Y} - \frac{\hat{\beta}_1}{100} \cdot 100\bar{X}$ equals $\bar{Y} - \hat{\beta}_1 \bar{X}$ still.



1.3

3. What happens to the OLS estimators if we standardize the variables?



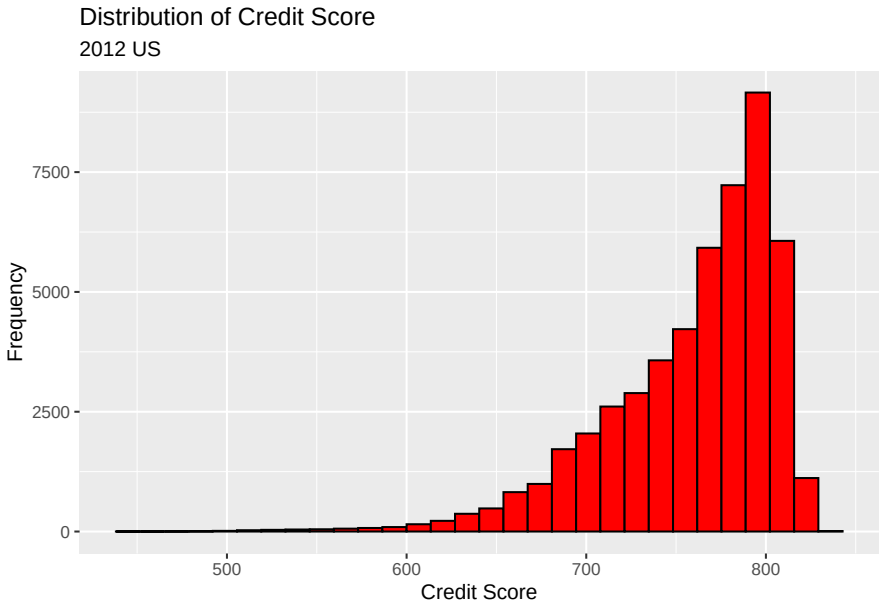
2.1

```
library(readr)
sample_orig_2012 <- read_delim("/cloud/project/sample_orig_2012.txt","|", escape_double = FALSE, col_nam



orig.names <- c("Credit_Score","First_Payment _ate","First_Time_Homebuyer", "Maturity_Date",
"MSA","Mortgage_Insurance_Percentage","Number_Units","Occupancy_Status","CLTV",
"DTI","UPB","LTV","Interest_Rate","Channel","Prepayment_Penalty",
"Amortization_Type","State","Property_Type","Postal_Code","Sequence_Number",
"Purpose","Loan_Term","Number_Borrowers","Seller_Name","Servicer_Name",
"Super_Conforming","Pre-HARP_Loan","Program_Indicator","HARP_Indicator",
"Valuation_Method","Interest_Only")
colnames(sample_orig_2012) <- orig.names
```

3.1

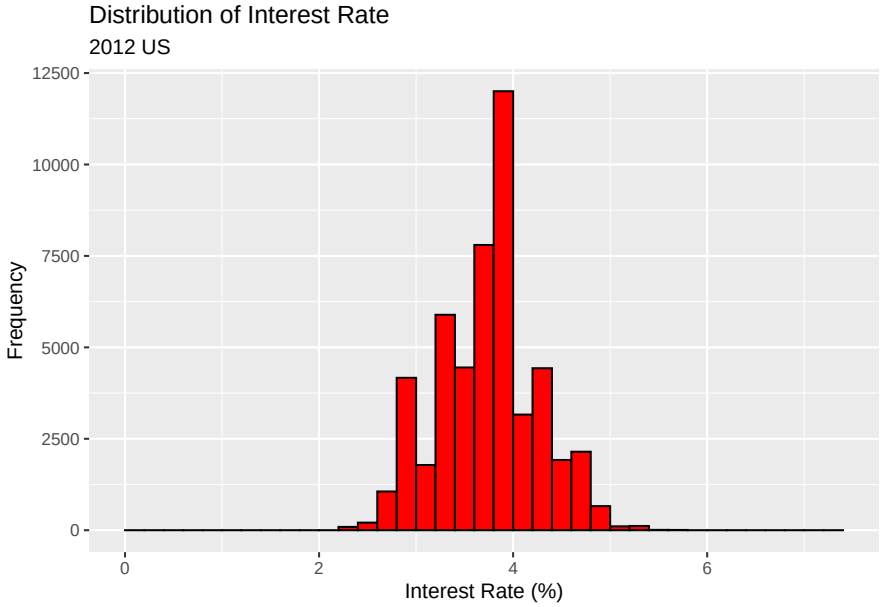
```
library(ggplot2)
library(dplyr)
Credit_Score <- sample_orig_2012$Credit_Score
Interest_Rate <- sample_orig_2012$Interest_Rate
sample_orig_2012 = filter(sample_orig_2012, sample_orig_2012$Credit_Score!= 9999)
ggplot(sample_orig_2012, aes(x = Credit_Score)) +
  geom_histogram(fill = "red", color = "black") +
  labs(title = "Distribution of Credit Score",
       subtitle = "2012 US",
```

```
      x = "Credit Score",
      y = "Frequency")
```



There are no outliers, nor missing values, but the data does skew to the left, with data points gradually going farther and farther left from the mode.

```
library(ggplot2)
ggplot(sample_orig_2012, aes(x = Interest_Rate)) +
  geom_histogram(breaks = seq(0, 7.5, 0.2), fill = "red", color = "black") +
  labs(title = "Distribution of Interest Rate",
       subtitle = "2012 US",
       x = "Interest Rate (%)",
       y = "Frequency")
```



There do not seem to be significant outliers for the interest rate. While the distribution is not smooth, it does seem to follow a relative bell shape with a mode near the median of the data.

3.2

```
# Calculate Mean
mean(Credit_Score)
```

```
[1] 759.1127
```

The mean is not within the modal bar. This is because of the left skew lowering the mean compared to the median and mode.

```
# Calculate Standard Deviation
sd(Credit_Score)
```

```
[1] 47.31753
```

Credit Score's standard deviation is about 47 units, which is quite a large spread. This number is large also in part due to the skew of the graph.

```
# Calculate Skewness
library(moments)
skewness(Credit_Score)
```

```
[1] -1.364729
```

The skewness is smaller than -1, meaning there is a high left skew, which affects the mean and std.

```
# Calculate Kurtosis
kurtosis(Credit_Score)
```

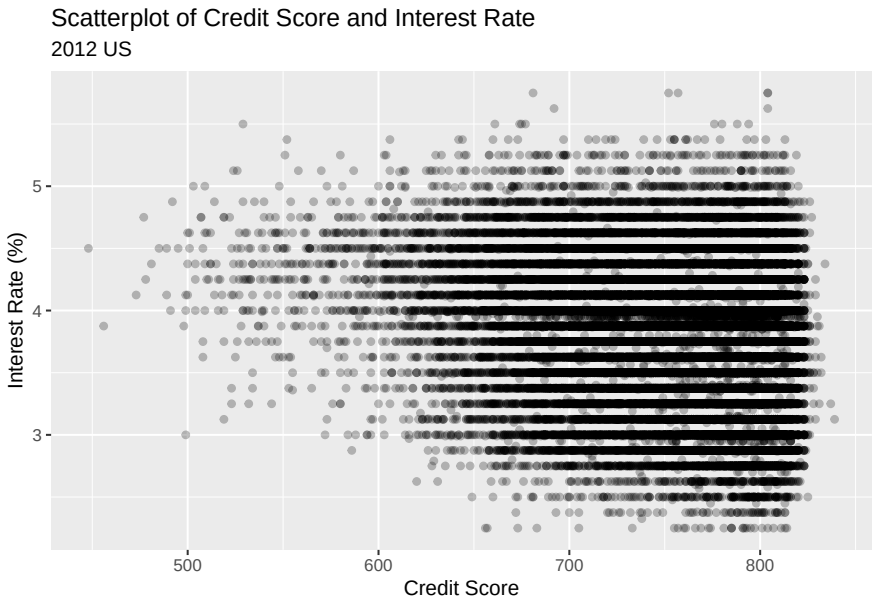
```
[1] 5.429348
```

Kurtosis is much greater than 3, meaning that there are high amounts of data in the tails. We can observe this in how there are data points below 500 in a graph with mean = 759 and std = 47. This is more than 5 standard deviations which in a normal distribution would be practically nonexistent.

3.3

Scatter plot of the credit score and interest rate variables:

```
library(ggplot2)
ggplot(sample_orig_2012, aes(x = sample_orig_2012$Credit_Score, y = sample_orig_2012$Interest_Rate)) +
  geom_point(color = "black", alpha = 0.25) +
  labs(title = "Scatterplot of Credit Score and Interest Rate",
       subtitle = "2012 US",
       x = "Credit Score",
       y = "Interest Rate (%)")
```



In the scatterplot, it does not seem like interest rate has much correlation with credit score. The regression appears to be slightly negative, but the two variables are probably independent from each other.

4.1

$$\hat{\beta}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

$\hat{\beta}_1$=-2.338e-03

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$\hat{\beta}_2$=5.532

4.2

Because $\hat{\beta}_1$=-2.338e-03, there is a negative linear relationship between Credit Scores and Interest rate; as Credit Score increases by 1 the Interest rate decreases by 2.338e-03. The sign is consistent, because higher credit score means that person is more consistently paying their card, meaning lower risk for the company and thus lower interest rates.

4.3

```
regression <- lm(Interest_Rate ~ Credit_Score)
summary(regression)
```

```
Call:
lm(formula = Interest_Rate ~ Credit_Score)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74798 -0.33196  0.04071  0.31536  2.09810
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.532e+00  3.587e-02  154.20   <2e-16 ***
Credit_Score -2.338e-03  4.717e-05  -49.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.499 on 49998 degrees of freedom
Multiple R-squared:  0.04685,   Adjusted R-squared:  0.04684
F-statistic:  2458 on 1 and 49998 DF,  p-value: < 2.2e-16
```

- R^2 is the correlation of determination, 0.04685 suggests very little of the variance in interest rate is explained by this credit score in this model. This supports the null hypothesis that interest rate is independent from credit score.
- SER is the average distance that the observed values fall from the regression. SER = 0.499 with a small 4.6% R^2 means that the error of the residuals are very large.