Homework 2

Andrew Wang

10/19/2021

Question 1

Q1, part A

We can show this by re-writing the equation to $R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$ and then substituting the top part changing it to $\frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{b_1^2 \sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2}$

Then, we can take the square root of this to get $R = \frac{b_1 \sum(X_i - \bar{X})}{\sum(Y_i - \bar{Y})} = \frac{b_1 s_X}{s_Y}$

Then, substitute in the formula for b_1: $r_{XY}(\frac{s_Y}{s_X})\frac{s_X}{s_Y} = r_{XY}$

Thus, we show that R^2 is the square of the sample correlation coefficient r_{XY}.

Q1, part B

We can substitute $corr(X, e) = 0$ as $cov(X, e) = 0 = \frac{1}{N-1}\sum(X_i - \bar{X})(e_i)$

Then we can substitute in for e_i as follows: $(X_i - \bar{X})(Y_i - b_0 - b_1 X_i) = \sum(X_i - \bar{X})(Y_i - [\bar{Y} - b_1\bar{X}] - b_1 X_i) = \sum(X_i - \bar{X})(Y_i - \bar{Y} - b_1(X_i - \bar{X})) = \sum[(X_i - \bar{X})(Y_i - \bar{Y}) - b_1(X_i - \bar{X})^2] = 0$

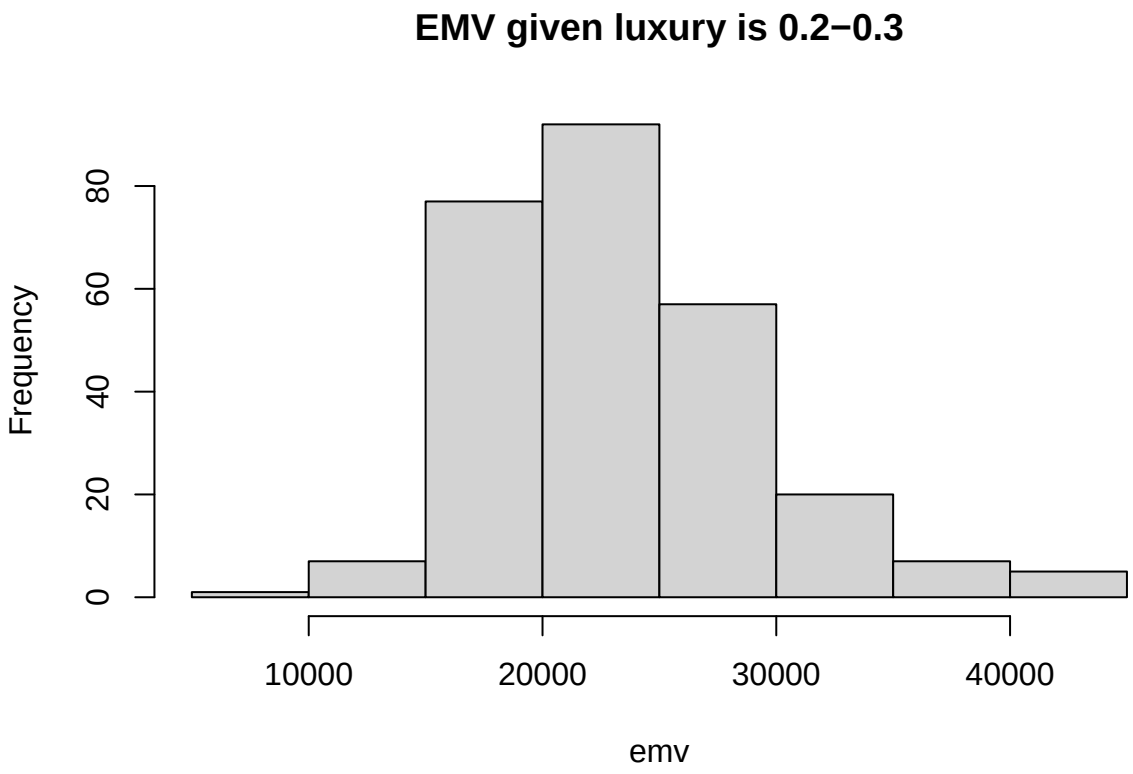
Next, we solve for b_1 which gives us: $b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$

Thus, we prove that $corr(X, e) = 0$ results directly from the formula for b_1.

Question 2 : More on Nearest Neighbor Approaches

Q2, part A

```
library(DataAnalytics)
data("mvehicles")
cars = mvehicles[mvehicles$bodytype != "Truck",]
cars2 = cars[cars$luxury >= 0.2,]
cars23 = cars2[cars2$luxury <= 0.3,]
emv <- cars23$emv
hist(emv, main = "EMV given luxury is 0.2-0.3")
```



Q2, part B

```
summary(emv)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9928   19079   22599   23377   26622   43822
```

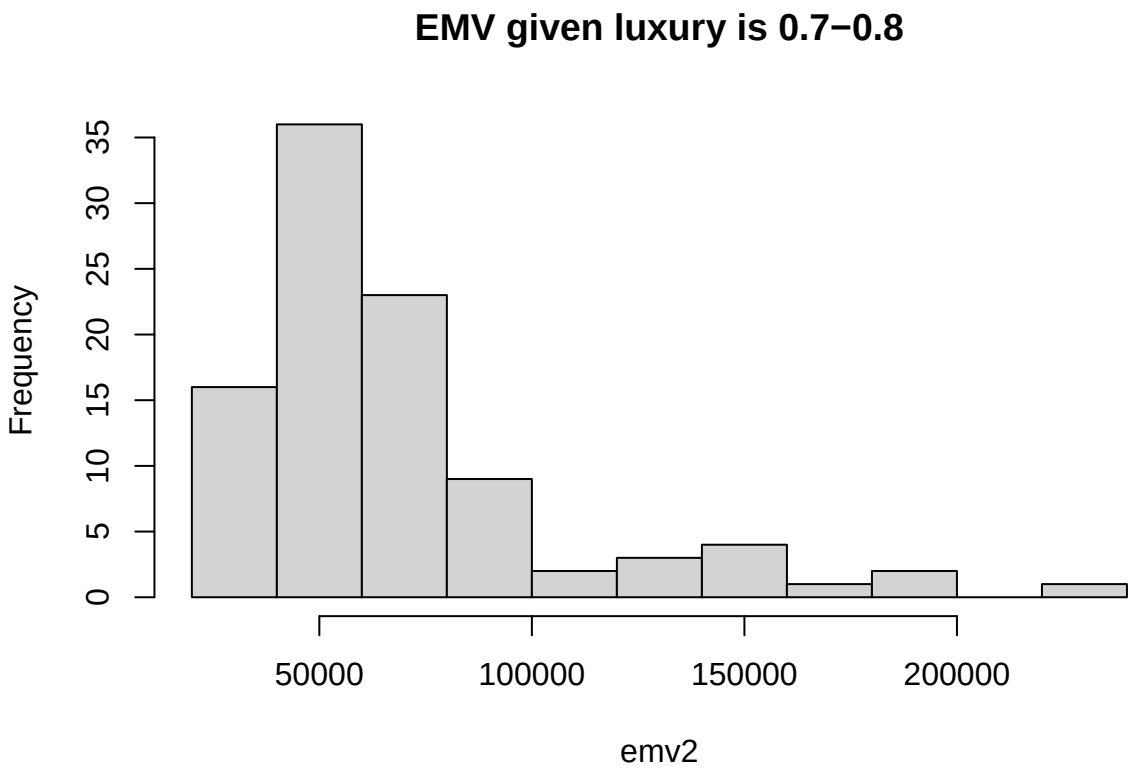
```
quantile(emv, probs = c(0.025, .975))
```

```
##     2.5%    97.5%
## 14699.09 38632.34
```

From the data above, we can see that our mean is 23377. And our 95% prediction interval is (14699.09, 38632.34).

Q2, part C

```
cars7 = cars[cars$luxury >= 0.7,]
cars78 = cars7[cars7$luxury <= 0.8,]
emv2 <- cars78$emv
hist(emv2, main = "EMV given luxury is 0.7-0.8")
```




```
summary(emv2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32232   43350   53520   68102   73119  225967
```

```
quantile(emv2, probs = c(0.025, .975))
```

```
##      2.5%     97.5%
##  34254.33 179179.22
```

From the data above, we can see that our mean is 68102. And our 95% prediction interval is (34254.33, 179179.22).

The difference between the two distributions is that the first one where luxury level is 0.2-0.3 more closely resembles a normal distribution while the distribution where luxury level is 0.7-0.8 is skewed right. The prediction interval for the second distribution also has a much higher prediction interval estimate.

Q2, part D

Luxury is not sufficiently informative to give accurate predictions of emv because it's only a single variable we're using to measure emv. When you only have one predictor variable, you're introducing a lot of potential for error in your prediction. In addition, by slicing the data at 0.7-0.8 luxury level we also saw that there was a lot of variation in the data which means we had to create a very large prediction interval which is not particularly useful for accurate predictions.

Question 3 : Optimal Pricing and Elasticities

Q3, part A

```
data("detergent")
log_det <- lm(log(q_tide128)~log(p_tide128), data = detergent)
summary(log_det)
```

```
##
## Call:
## lm(formula = log(q_tide128) ~ log(p_tide128), data = detergent)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7186 -0.4629 -0.0056  0.4339  2.9980
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13.29778    0.13689   97.14   <2e-16 ***
## log(p_tide128) -4.41205    0.06452  -68.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7358 on 14743 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2407
## F-statistic:  4676 on 1 and 14743 DF,  p-value: < 2.2e-16
```

From the data above, we can compute price elasticity of demand as -4.4 because in a log-log regression, the coefficient on log-price can be interpreted directly as an elasticity.

```
confint(log_det, level = 0.9)
```

```
##                       5 %      95 %
## (Intercept)    13.072601 13.522963
## log(p_tide128) -4.518186 -4.305912
```

A 90% CI for this elasticity is shown above as (-4.518186, -4.305912)

Q3, part B

If the retailer is earning a 25% gross margin, that means the elasticity consistent with that is equal to -4, meaning a 1 percent increase in price will reduce sales by 4 percent.

As we saw above, our 90% CI was (-4.518186, -4.305912) and since this -4 falls out of the range, we can say that we are not pricing optimally at the 90% CI.

Question 4

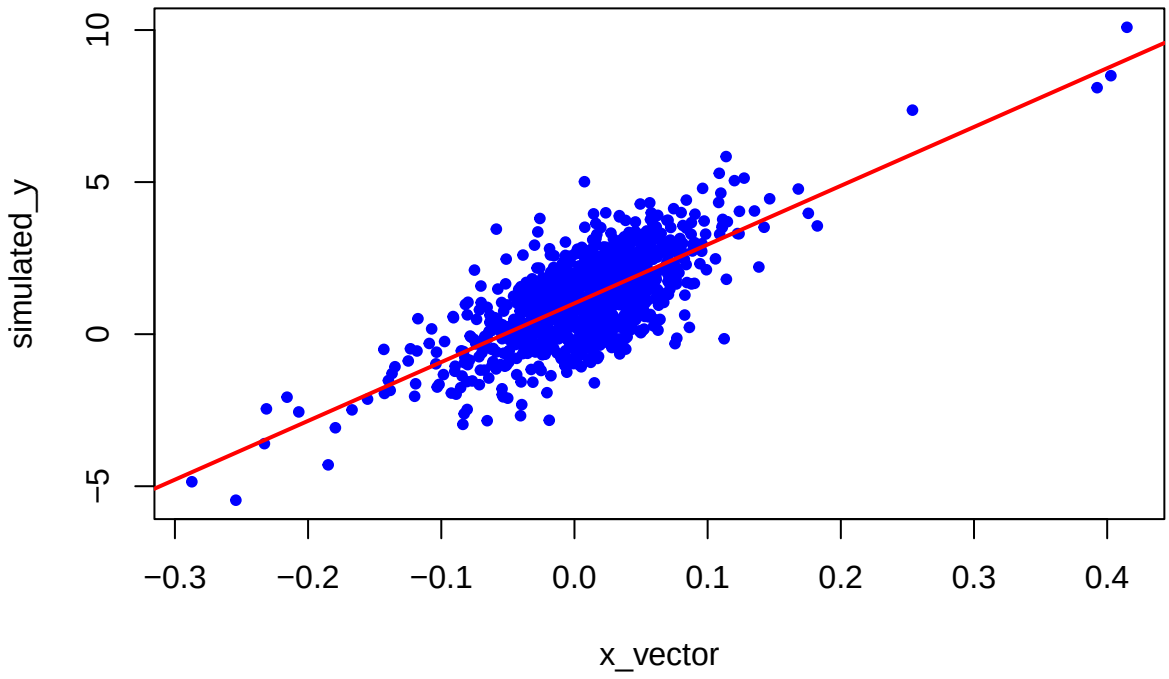
a.

```
y_values <- function(b_0, b_1, sigma, x){
  y <- b_0 + b_1*x + rnorm(length(x), sd=sigma)
}
```

b.

```
data("marketRf")
x_vector <- marketRf$vwretd
b_0 <- 1; b_1 <- 20; sigma <- 1
```

```r
simulated_y <- y_values(b_0, b_1, sigma, x_vector)
plot(x_vector, simulated_y, pch=20, col="blue")
outlm=lm(simulated_y~x_vector)
abline(outlm$coef,lwd=2,col="red")
```



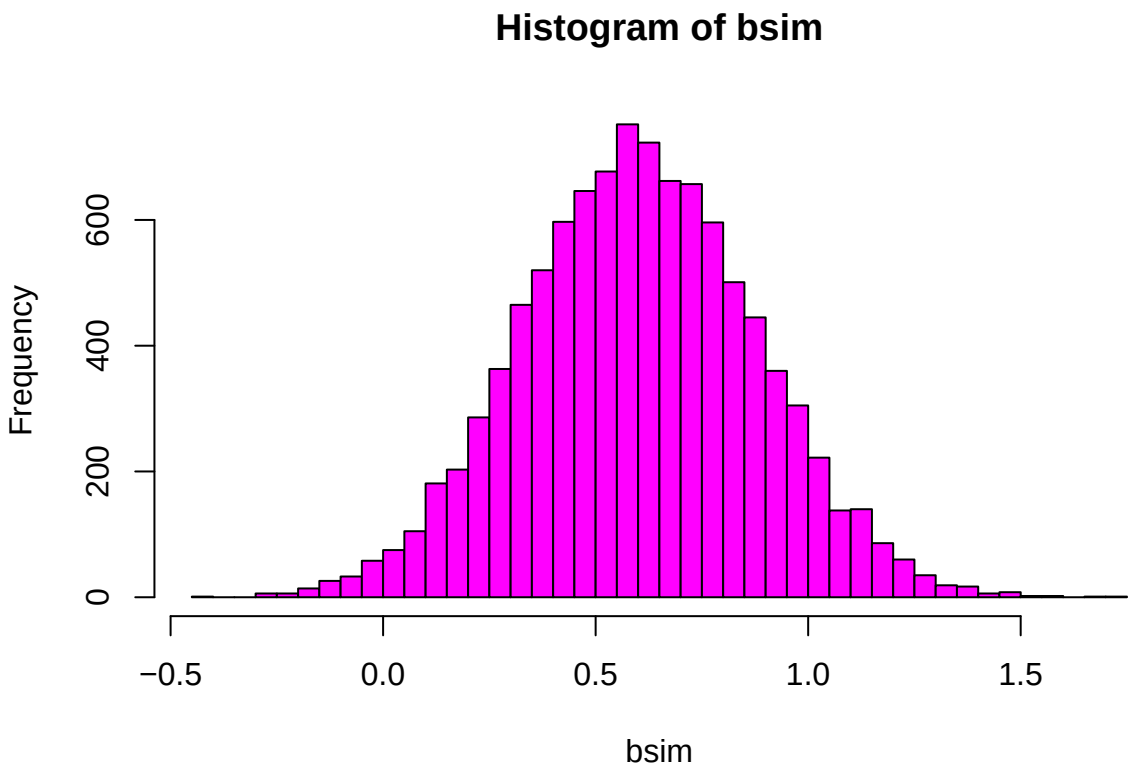
Question 5

a.

```r
x_unif = runif(n=300)
beta0 <- 2; beta1 <- 0.6; sigma2 <- sqrt(2)

nsample <- 10000
bsim <- double(nsample)
for(i in 1:nsample) {
  y <- y_values(beta0, beta1, sigma2, x_unif)
  bsim[i] <- lm(y~x_unif)$coef[2]
}

hist(bsim, breaks=40, col="magenta")
```



b.

```
mean(bsim)
```

```
## [1] 0.601855
```

The empirical value for $E[b_1]$ is equal to 0.6036787 which is very close to our theoretical value of 0.6.

c.

```
out = lm(y ~ x_unif)
var_x = var(x_unif)
N = length(x_unif)
s_sq = sum(out$residuals**2)/(N - 2)

sqrt(s_sq)
```

```
## [1] 1.403833
```

```
sqrt(2)
```

```
## [1] 1.414214
```

The empirical value for $Var(b_1)$ is 1.266362 while the theoretical value for $Var(b_1)$ is 1.414214. Again, these values are quite close to one another.

Question 6

Standard errors and p-values.

a. The standard error of a sampling statistic or an estimator $\hat{Y}$ is an estimate of its standard deviation. While standard error is a measure of variation across samples of a population, the standard deviation is a measure of variation within a sample.
b. Sampling error is the difference between an estimated population mean and a sample mean.The standard error helps capture sampling error by providing an estimate of how different the population mean is likely to be from the sample mean.
c. I would tell Steven that the standard error would tell us how accurate the mean of the sample is compared to the true population mean. And then based on the size of the parameter estimates and the standard error, we can determine how accurate and close we are to estimating the true population mean.
d. We can use the test statistic to help us determine what confidence interval we can construct to test against the null hypothesis. In addition, the p-value can help us describe how likely it is that our data could have occurred by random chance. Essentially, the lower the p-value, the stronger the evidence to suggest that we should reject the null hypothesis.