Homework 2
MSBA 400: Statistical Foundations for Data Analytics
Professor Rossi

Question 1

Q1, part A

In the class notes, we introduced the concept of R^2. Show that the formula $R^2 = \frac{SSR}{SST}$ implies that R^2 is the square of the sample correlation coefficient between X and Y, r_{XY}. Hint: recall from the notes how the fitted regression line can be expressed in terms of deviations from the mean.

Solution:

Proof. We begin with

$$R^2 = \frac{\sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}.$$

Working with the numerator, we have

$$\sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^N (b_0 + b_1 X_i - \bar{Y})^2 = \sum_{i=1}^N (\bar{Y} - b_1\bar{X} + b_1 X_i - \bar{Y})^2 = b_1^2 \sum_{i=1}^N (X_i - \bar{X})^2$$

Now,

$$b_1 = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

so we have

$$b_1^2 \sum_{i=1}^N (X_i - \bar{X})^2 = \frac{[\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})]^2 \sum_{i=1}^N (X_i - \bar{X})^2}{[\sum_{i=1}^N (X_i - \bar{X})^2]^2} = \frac{[\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^N (X_i - \bar{X})^2}.$$

Inserting the denominator, we get

$$\frac{[\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^N (X_i - \bar{X})^2 \sum_{i=1}^N (Y_i - \bar{Y})^2} = \left(\frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}} \right)^2 = r_{xy}^2.$$

Hence $R^2 = r_{xy}^2$. □

Q1, part B

In the class notes, we used intuition to argue the regression property, $corr(X, e) = 0$. Show this directly results from the formula for b_1. Hint: substitute, $e_i = Y_i - b_0 - b_1 X_i$ into $corr(X, e)$.

Solution:

Proof. We have

$$corr(X, e) = \frac{\text{Cov}(X, e)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(e)}}$$

and thus it suffices to show that $\text{Cov}(X, e) = 0$. Keeping in mind that $e = Y - b_0 - b_1 X$ and that $b_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$, we have

$$\begin{aligned}\text{Cov}(X, e) &= \text{Cov}(X, Y - b_0 - b_1 X) = \text{Cov}(X, Y - b_1 X) = \text{Cov}(X, Y) - b_1 \text{Cov}(X, X) \\ &= \text{Cov}(X, Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \text{Var}(X) = \text{Cov}(X, Y) - \text{Cov}(X, Y) = 0.\end{aligned}$$

Hence $corr(X, e) = 0$. □

Question 2 : More on Nearest Neighbor Approaches

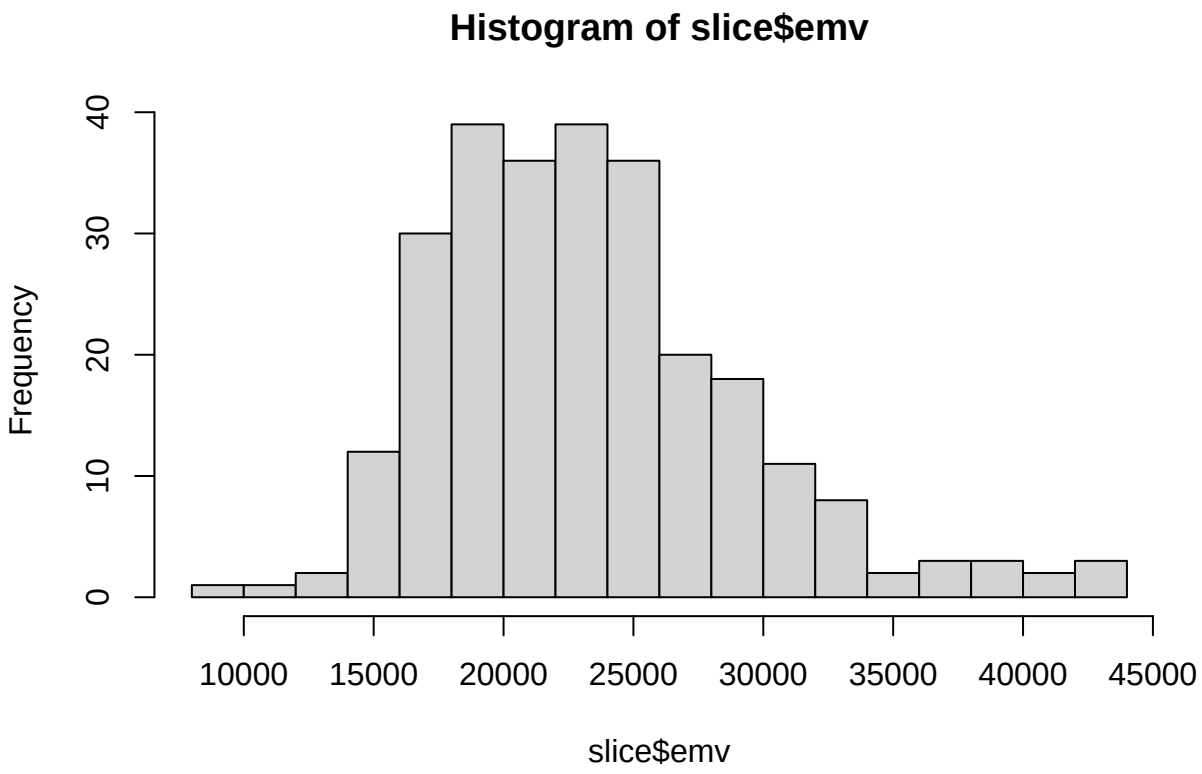
The simplest view of nearest neighbor methods is to "slice" into the data for only a small interval of X values. This distribution is called the conditional distribution of Y given X.

Q 2, part A

Display a histogram of the conditional distribution of emv given that luxury is in the interval (.2, .3) in the cars dataset (from problem set 1).

Solution:

```
# import package and data
library(DataAnalytics)
data(mvehicles)
# initialize cars dataframe
cars = mvehicles[mvehicles$bodytype != 'Truck',]
# slice data to contain only vehicles with luxury in (0.2, 0.3)
slice = cars[cars$luxury < 0.3 & cars$luxury > 0.2,]
# plot histogram of slice$emv
hist(slice$emv, breaks = 'FD')
```



Q 2, part B

Compute the mean of the conditional distribution in part A and compute a prediction interval that takes up 95% of the data (an interval that stretches from the .025 quantile (2.5 percentile) to the .975 (97.5 percentile)). Use the `quantile()` command.

Solution:

```
# reinitialize same slice
slice = cars[cars$luxury < 0.3 & cars$luxury > 0.2,]
# compute mean of the conditional distribution
mean(slice$emv)
```

```
## [1] 23376.5
```

```
# compute prediction interval that takes up 95% of the data
prediction_interval = quantile(slice$emv, probs = c(0.025, 0.975))
prediction_interval
```

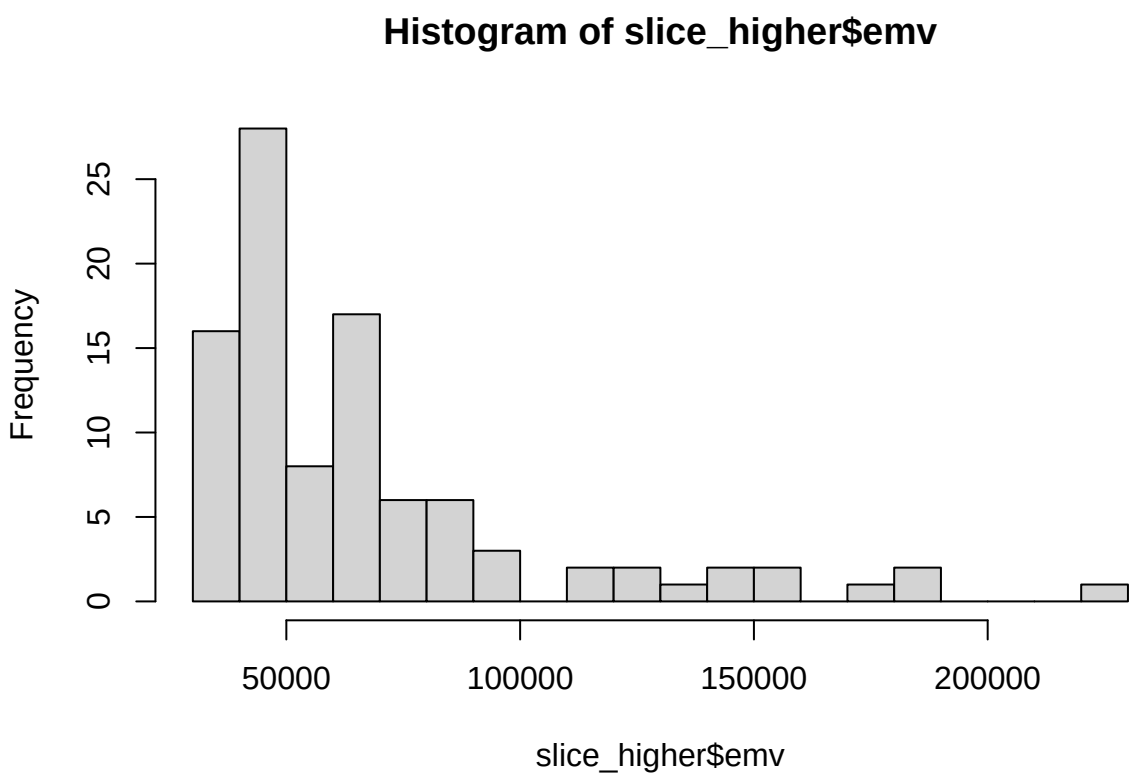
```
##     2.5%    97.5%
## 14699.09 38632.34
```

Q 2, part C

Repeat part A and B for a much higher level of luxury, namely the interval (.7, .8). Describe the difference between these two conditional distributions.

Solution:

```r
# slice data to contain only vehicles with luxury in (0.7, 0.8)
slice_higher = cars[cars$luxury < 0.8 & cars$luxury > 0.7,]
# plot histogram of slice_higher$emv
hist(slice_higher$emv, breaks = 'FD')
```




```r
# compute mean of the conditional distribution
mean(slice_higher$emv)
```

```
## [1] 68101.87
```

```r
# compute prediction interval that takes up 95% of the data
prediction_interval_higher = quantile(slice_higher$emv, probs = c(0.025, 0.975))
prediction_interval_higher
```

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

We can see that the conditional distribution of cars with luxury indices in the interval (0.2, 0.3) exhibited more of a normal distribution, while the cars with luxury indices in the interval (0.7, 0.8) were concentrated primarily at the lower end of the interval. This observation is supported by the means of the conditional distributions—the mean of the conditional distribution of cars in the (0.2, 0.3) interval is relatively close to the center of the interval while the mean of the conditional distribution of the cars in the (0.7, 0.8) interval is heavily skewed to the lower end of the interval.

Q 2, part D

Explain why the results of part B and C show that luxury is probably (by itself) not sufficiently informative to give highly accurate predictions of `emv`.

Solution: Luxury alone is likely not sufficient to give highly accurate predictions of `emv` simply due to the size of the 95% prediction intervals. For example, for cars in the conditional distribution of (0.2, 0.3), the prediction interval is (14699.09, 38632.34) and the mean is 23376.5, which implies a very high standard error and thus a poor predictive model. The same logic applies to the prediction interval and mean for cars in the conditional distribution of (0.7, 0.8), which has an even wider prediction interval.

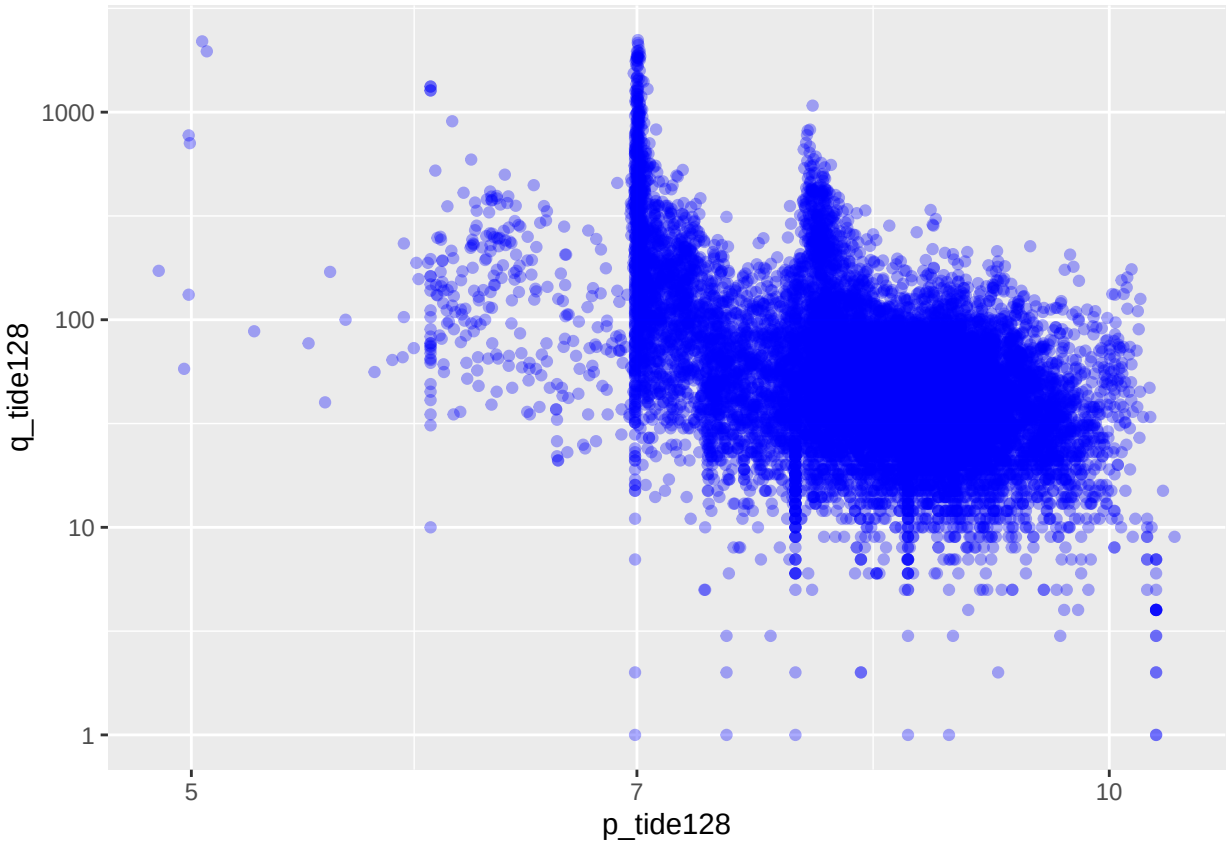
Question 3 : Optimal Pricing and Elasticities

Q 3, part A

Use the `detergent` dataset to determine the price elasticity of demand for 128 oz Tide. Compute the 90 percent confidence interval for this elasticity.

Solution:

```
# import detergent data
data(detergent)
# to compute price elasticity, run log-log regression and examine coefficient on log-price
# visualize log-transformed data
library(ggplot2)
qplot(p_tide128,q_tide128,log="xy",data=detergent,col=I("blue"),alpha=I(1/3))
```




```
# run log-log regression
lm_log = lm(log(q_tide128)~log(p_tide128), data = detergent)
summary(lm_log)
```

```
##
```

```
## 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
```

So we can see that the log-log price elasticity is -4.41.

```
# compute 90 percent confidence interval
confint(lm_log, level = 0.9)
```

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

Q 3, part B

One simple rule of pricing is the "inverse elasticity" rule that the optimal gross margin should be equal to the reciprocal of the absolute value of the price elasticity, i.e. Gross Margin $= \frac{1}{|elasticity|}$. For example, suppose we estimate that the price elasticity is -2 (a 1 per cent increase in price will reduce sales (in units) by 2 per cent. Then the optimal gross margin is 50 percent.

Suppose this retailer is earning a 25 per cent gross margin on 128 oz Tide. Perform appropriate hypothesis test to check if the retailer is pricing optimally at the 90 per cent confidence level.

Hints:
i. use the inverse elasticity rule to determine what elasticity is consistent with a 25 per cent gross margin.
ii. Use the confidence interval!

Solution: We know that the gross margin is 25 percent, so $0.25 = \frac{1}{|elasticity|}$ and thus the price elasticity, if this retailer is pricing optimally, should be -4. However, we can see that -4 is not in the 90 percent confidence interval (-4.518186 -4.305912) obtained in part A. The retailer thus does not appear to be pricing optimally.

Questions 4-5 explore the sampling properties of least squares

Question 4

a. Write your own function in `R` (using `function()`) to simulate from a simple regression model. This function should accept as inputs: b_0 (intercept), b_1 (slope), X (a vector of values), and σ (error standard deviation). You will need to use `rnorm()` to simulate errors from the normal distribution. The function should return a vector of Y values.

Solution:

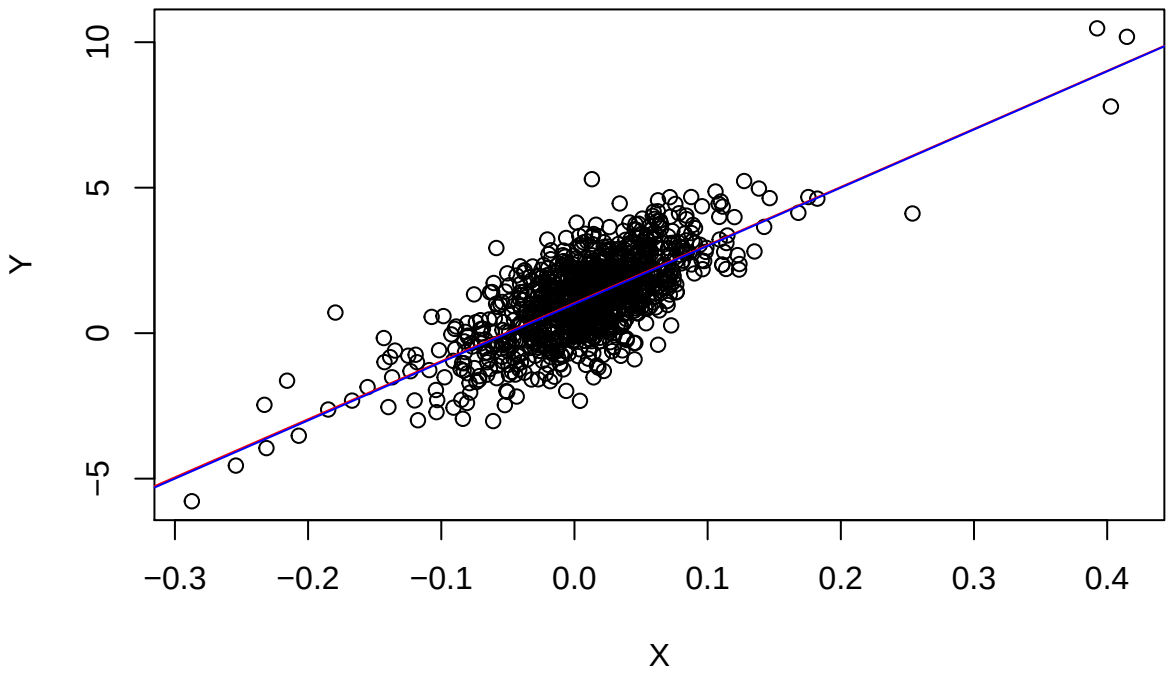
```
# simulate a regression
simulate_regression = function(b0, b1, X, sigma) {
  # initialize vector of error terms
  errors = rnorm(n = length(X), mean = 0, sd = sigma)
  # generate Y values
  Y = b0 + b1 * X + errors
  return(Y)
}
# test the simulated regression
X_test = c(1,2,3,4,5)
simulate_regression(0.1, 1.5, X_test, 0.2)
```

```
## [1] 1.492290 2.862192 4.404848 5.966939 7.736121
```

b. Simulate Y values from your function and make a scatterplot of X versus simulated Y. When simulating, use the `vwretd` data from the `marketRf` dataset as the X vector, and choose $b_0 = 1$, $b_1 = 20$, and $\sigma = 1$. Then add the fitted regression line to the plot as well as the true conditional mean line (the function `abline()` may be helpful).

Solution:

```
library(DataAnalytics)
data(marketRf)
# initialize X values
X = marketRf$vwretd
# generate simulated Y values
Y = simulate_regression(1, 20, X, 1)
plot(X, Y)
# overlay simulated regression line
reg = lm(Y~X)
abline(reg, col = 'red')
# overlay true conditional mean line
abline(1, 20, col = 'blue')
```



Question 5

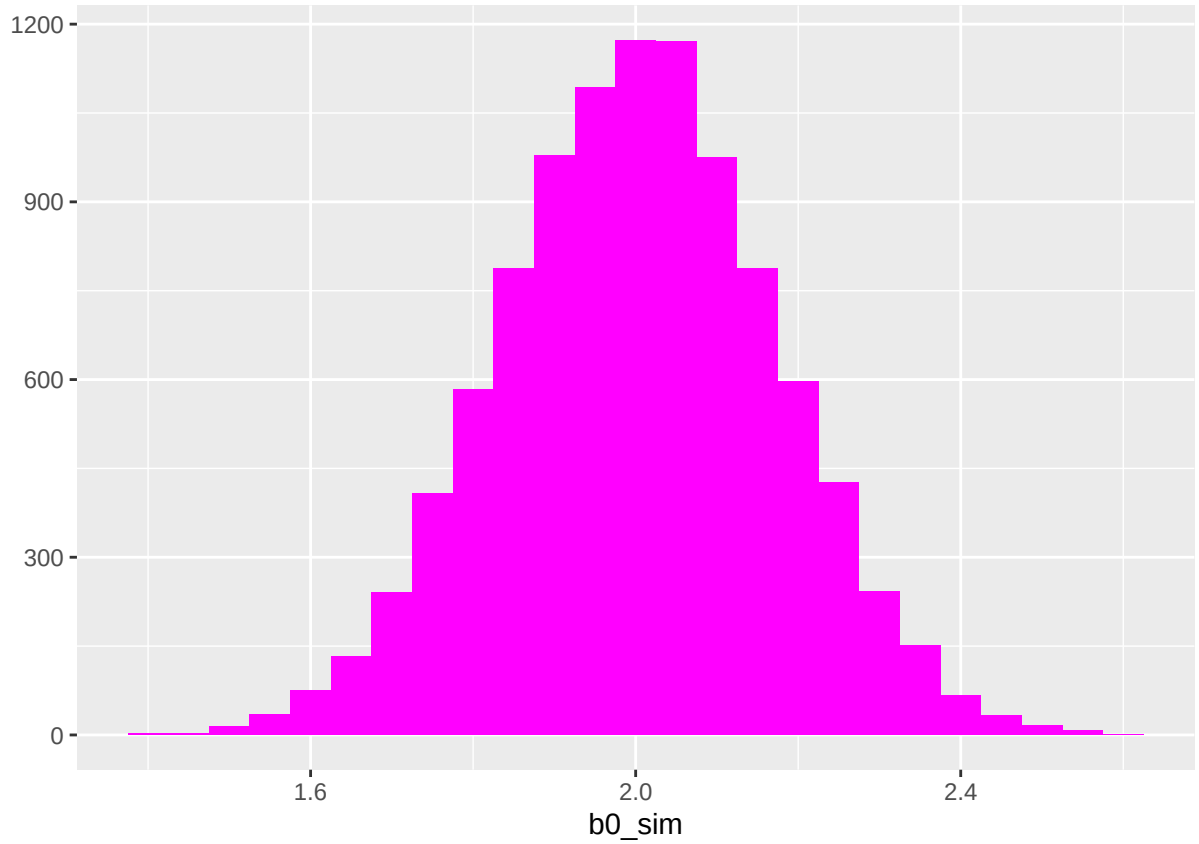
Assume $Y = \beta_0 + \beta_1 X + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Let $\beta_0 = 2$, $\beta_1 = 0.6$, and $\sigma^2 = 2$. You can make X whatever you like, for example you could simulate X from a uniform distribution.

a. Use your `R` function from question 4 to simulate the sampling distribution of the slope. Use a sample size of $N = 300$ and calculate b_0 & b_1 for 10,000 samples. Plot a histogram of the sampling distribution of b_0.

Solution:

```
library(ggplot2)
n = 300
# simulate X from a uniform distribution
X = runif(n)
# initialize known values
beta_0 = 2
beta_1 = 0.6
sigma = sqrt(2)
num_samples = 10000
# initialize vectors
b0_sim = double(num_samples)
b1_sim = double(num_samples)
# simulate and store values for b_0 and b_1
for (i in 1:num_samples){
  Y = simulate_regression(beta_0, beta_1, X, sigma)
```

```
  b0_sim_val = lm(Y~X)$coef[1]
  b1_sim_val = lm(Y~X)$coef[2]
  b0_sim[i] = b0_sim_val
  b1_sim[i] = b1_sim_val
}
# plot histogram of sampling distribution of b_0
qplot(b0_sim, fill = I('magenta'), binwidth = 0.05)
```



b. Calculate the empirical value for $\mathbb{E}[b_1]$ from your simulation and provide the theoretical value for $\mathbb{E}[b_1]$. Compare the simulated and theoretical values.

Solution:

The simulated value of $\mathbb{E}[b_1]$ is the mean of the sampling distribution, which we can calculate in R:

```
mean(b1_sim)
```

```
## [1] 0.5993079
```

We know that $\mathbb{E}[b_1] = \beta_1$ by construction, so the theoretical value is simply $\mathbb{E}[b_1] = \beta_1 = 0.6$ The simulated value differs slightly from the theoretical value, which is expected.

c. Calculate the empirical value for $\text{Var}(b_1)$ from your simulation and provide the theoretical value for $\text{Var}(b_1)$. Compare the simulated and theoretical values.

Solution:

The simulated value of $\text{Var}[b_1]$ is the variance of the sampling distribution, which we can calculate in R:

```
var(b1_sim)
```

```
## [1] 0.07960732
```

We know that the formula for $\text{Var}[b_1]$ is

$$\text{Var}[b_1] = \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2},$$

and we can use this to calculate the theoretical value:

```
# compute X bar
x_avg = mean(X)
# initialize vector
diffs_sq = double(length(X))
# compute all X_i - X bar and square them
for (i in 1:length(X)){
  diffs_sq[i] = (X[i] - x_avg)^2
}
# execute formula
var_b1 = sigma^2/sum(diffs_sq)
var_b1
```

```
## [1] 0.08015064
```

We can again see that the simulated value differs slightly from the theoretical value, as expected.

Question 6

Standard errors and p-values.

a. What is a standard error (of a sample statistic or an estimator)? How is a standard error different from a standard deviation?

   Solution: A standard error is an estimate of the extent to which sample means deviate from the population mean. That is, it measures the precision of estimates of sample mean. Standard deviation, however, measures the dispersion of the data in relation to the mean.

b. What is sampling error? How does the standard error capture sampling error?

   Solution: Sampling error is idea that a sample and the population it comes from are inherently different and thus the sample parameters often differ from population parameters. Standard error captures sampling error because it estimates the deviation of some estimated mean value from the true mean value, and this difference exists due to sampling error. When we give an estimate along with some standard error, we are able to get an idea of the sampling error present in that estimate.

c. Your friend Steven is working as a data analyst and comes to you with some output from some statistical method that you've never heard of. Steven tells you that the output has both parameter estimates and standard errors. He then asks, "how do I interpret and use the standard errors?" What do you say to Steven to help him even though you don't know what model is involved?

   Solution: I would tell him that the standard error (SE) can be used to calculate the margin of error, which can give him a measure of the dispersion of the estimated values from the true parameters values. More specifically, I would tell him that his parameter estimates are accurate roughly

up to within $\pm 2 \times SE$.

d. Your friend Xingua works with Steven. She also needs help with her statistical model. Her output reports a test statistic and the p-value. Xingua has a Null Hypothesis and a significance level in mind, but she asks "how do I interpret and use this output?" What do you say to Xingua to help her even though you don't know what model is involved?

Solution: I would tell her that she should first compute a rejection region

$$\left(-t^*_{N-2,\frac{\alpha}{2}},\ t^*_{N-2,\frac{\alpha}{2}} \right)$$

based on the significance level α. I would then tell her that, if her test statistic t fell within that rejection region, she should accept (or fail to reject) the Null Hypothesis and reject it otherwise. I would then tell her that the p-value could be used to determine the minimum significance level at which she should reject the Null Hypothesis.