Diana Liang

9/30/19

Stats 500 HW #2

**Part A: Effect of Education and Experience on US Wages**

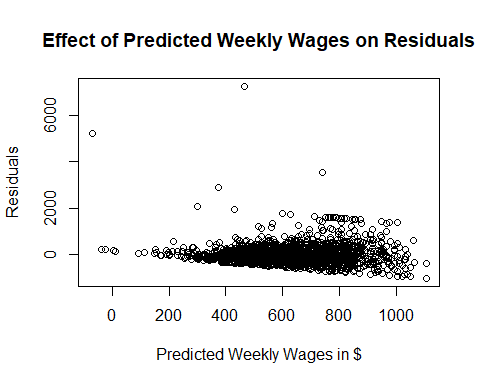
##   
## Call:  
## lm(formula = wage ~ educ + exper, data = uswages)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1014.7 -235.2 -52.1 150.1 7249.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -239.1146 50.7111 -4.715 2.58e-06 \*\*\*  
## educ 51.8654 3.3423 15.518 < 2e-16 \*\*\*  
## exper 9.3287 0.7602 12.271 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 426.8 on 1964 degrees of freedom  
## (33 observations deleted due to missingness)  
## Multiple R-squared: 0.1348, Adjusted R-squared: 0.1339   
## F-statistic: 153 on 2 and 1964 DF, p-value: < 2.2e-16

A regression model was fit on the US Wage data with the above output (Q1). The model provides a predicted value given inputted years of education and experience. So 2 people with the same education but only one year difference in experience would have a difference in $9.30 for predicted weekly wages (Q5). Although the p-values for the intercept and β’s confirm that the predictor variables are significant, the percentage of variation of R^2 = 0.135 (Q2) suggests that the predictors do not explain much of the variation in the response.

SUMMARY OF RESIDUALS

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1014.71 -235.22 -52.14 0.00 150.12 7249.17

The above summary of residuals provides a deeper insight into the response variation. The largest residual of 7249.2 for observation 1550 out of 2000, or participant 15387 (Q3). Meanwhile the mean of 0 is greater than the median of -52.1 (Q4), suggesting that most of the data points fall below in weekly wages compared to predicted weekly wages.



Correlation

## [1] 6.35678e-17

The correlation of 0 between the predicted weekly wages and the residuals confirms that there is no connection. Since the residuals are orthogonal to the predicted values by least squares regression, there is no way to project the residuals on the predicted values and, thus, no correlation.

**Part B: Y = Xβ + ϵ**

1. The β estimate based on formula is :

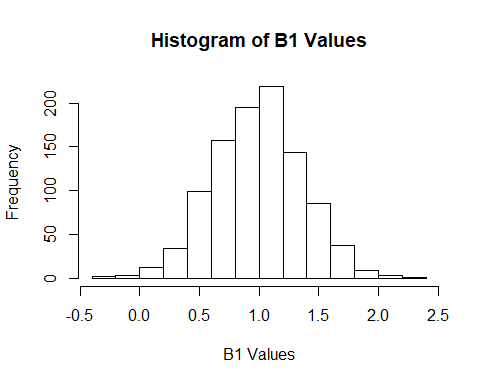
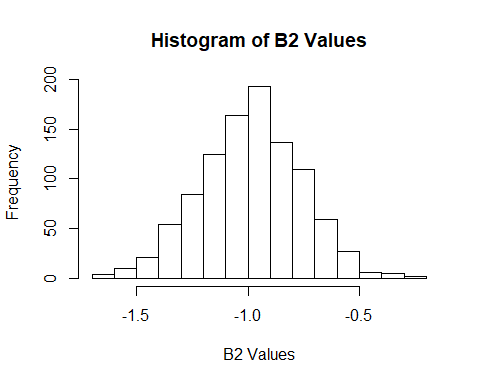
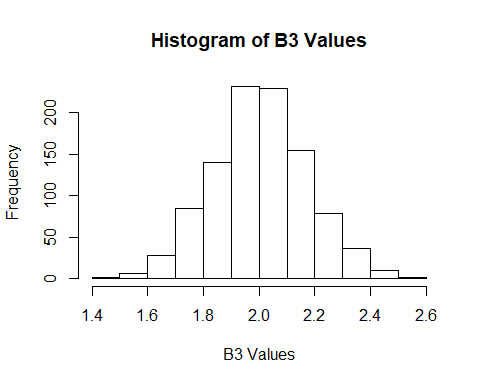
## [,1]  
## [1,] 1.0595019  
## [2,] -0.5545596  
## [3,] 1.8328164

1. The true variance of beta hat is the true variance multiplied by the inverse of X’X. Since we know that the true standard deviation is 1, the true variance is 1. So the true variance is:

## [,1] [,2] [,3]  
## [1,] 0.139180672 -0.042016807 -0.003413866  
## [2,] -0.042016807 0.050420168 -0.008403361  
## [3,] -0.003413866 -0.008403361 0.027442227

1. Variance = RSS/df so the estimate variance is:

## [1] 0.920632

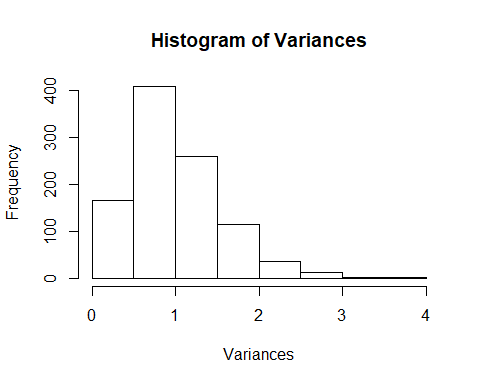
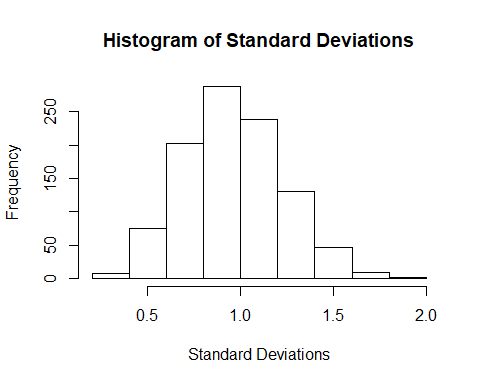
1.   

## Variance of B1: 0.1369088

## Variance of B2: 0.05087806

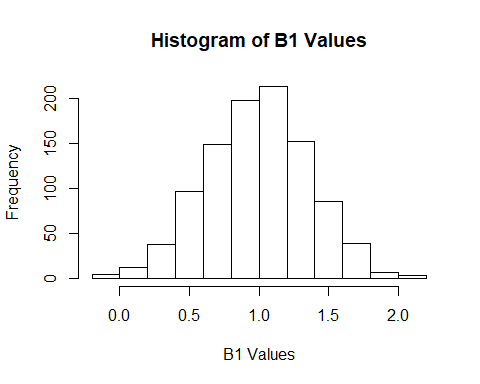
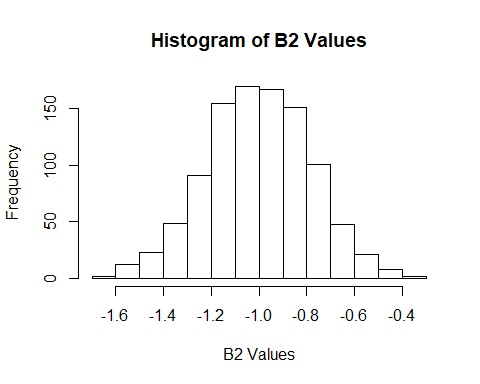
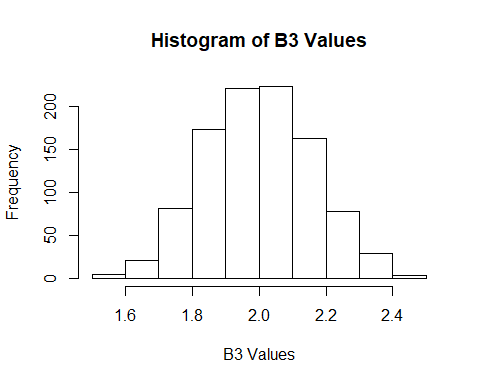
## Variance of B3: 0.02861535

The histograms of the β values show relatively symmetrical distributions around the true β values. And since the values of the variances are similar to those of the true variance diagonal, these variances match the true variance.

1.  

The histogram of the variance shows a skewed distribution about the true variance of 1, suggesting that these variances do not reliably estimate the true variance. On the other hand, the histogram of the standard deviation shows a mostly symmetrical distribution about the true standard deviation of 1, suggesting that the standard deviations are a reliable estimate of the true standard deviation and the variance by extension. Since the variance is a non-linear value, the skewed distribution is difficult to interpret, making the distribution of the standard deviation easier to interpret. The standard deviation is a reliable estimate of the true standard deviation, so the variance is a reliable estimate of the true variance.

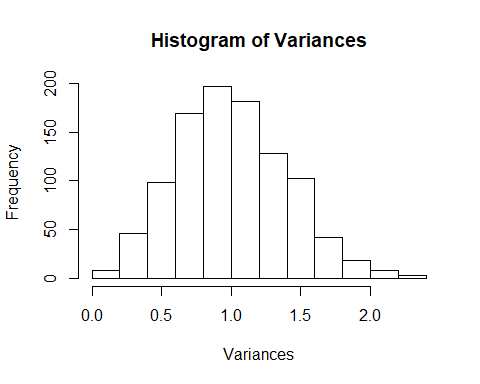
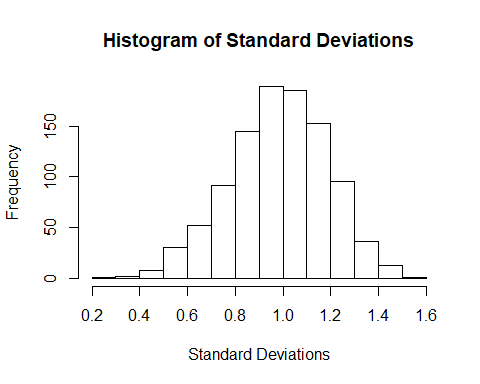
1. The same methodology was applied to a uniform distribution of mean = 0 and standard deviation = 1 to check if the produced β values and variances are reliable estimates of the true β and variance.

## Variance of B1: 0.1351695

## Variance of B2: 0.04794587

## Variance of B3: 0.02579371

The results are very similar to those produced using a standard normal distribution. This should be the case for every distribution, since sampling around known means and variances will provide values that are similar to these known values, especially if sampling a great number or a multitude of times.