Assignment 1

Milan Kuzmanovic, Mark McMahon
Martin Kotuliak, Jakub Polak

March 9, 2020

Task 1

A multiple linear regression model has been estimated to study the relation- ship between $Y =$ violent crime rate (per 100,000 people), $X_1 =$ poverty rate (percentage with income below the poverty line) and $X_2 =$ percentage living in urban area. Data are collected in 51 cities in the U.S.

The relevant equations that relate estimate, standard error, T-statistic, R-squared and Sum of Squares of residuals, regression and total are folloqwing.

$$\frac{\beta_i}{se(\beta_i)} = T_i \qquad R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}} \qquad \text{SST} = \text{SSreg} + \text{SSR}$$

Using these we just plug in the corresponding information that is already provided and compute the missing values.

```
a <- -498.683 / 140.988
b <- 4.885 * 6.677
c <- 9.112 / 6.900
d <- 1841257.15 / (1 - 0.5708)
e <- d - 1841257.15
```

The table below reports the output with filled in missing information.

	Est.	s.e.	t-value	p-value
Intercept	-498.683	140.988	[a] -3.537	0.009
X_1	[b] 32.617	6.677	4.885	0.001
X_2	9.112	[c] 1.321	6.900	0.001
R^2	0.5708			
SSreg	[e] 2448718			
SSR	1841257.15			
SSTotal	[d] 4289975			

The the coefficient of determination R^2 is 0.5708. This value measures the proportion of the variance in Y explained by the model. Hence, 57.08 % of the sample variability of Y can be explained by the linear combination of X_i'S given the sample data.

To compute the overall F-test we use the equation below and the statistic then follows an F distribution with corresponding degrees of freedom.

$$F = \frac{\text{SSreg} \,/\, p}{\text{SSR} \,/\, (n - (p+1))} \sim F_{p,n-(p+1)}$$

The data are collected in 51 cities, so $n = 51$ and we have 2 predictors, so $p = 2$. Other values we can easily obtain from the filled table above.

```r
(f = (e / 2) / (181257.15 / (51-(2+1)) ))

## [1] 324.2312
```

Hence, the F-statistic has a value of 324.2311914. To interpret this, the global F-test, tests a null hypothesis that all regression coefficients are simultaneously 0. In a mathematical notation $H_0 : \beta_1 = \beta_2 = 0$. To evaluate the test, we can compute its p-value. It is a quantile of the corresponding F-distribution for the given statistic or mass under the distribution.

```r
(p = pf(f,2,48,lower.tail = FALSE))

## [1] 1.319003e-28
```

The p-value is $1.3190029 \times 10^{-28}$, which is very low and therefore there is a significant evidence against the null hypothesis that all regression coefficients are simultaneously zero. Hence, the p-value of a given sample is small and it suggests that the model with all the covariates is better than the one with just intercept coefficient. This test therefore determines if the linear model is at least suitable to explain some of the variance in the outcome variable. Which in this case it does.

Task 2

Task 3

Task 4

Task Template

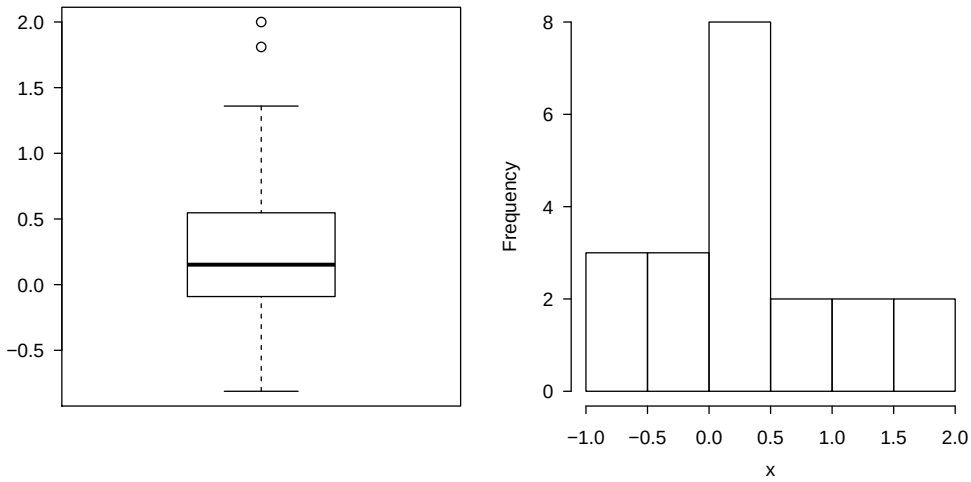
This is just an expample of how to use latex and r-code chunks in knitr.

You can test if **knitr** works with this minimal demo. OK, let's get started with some boring random numbers:

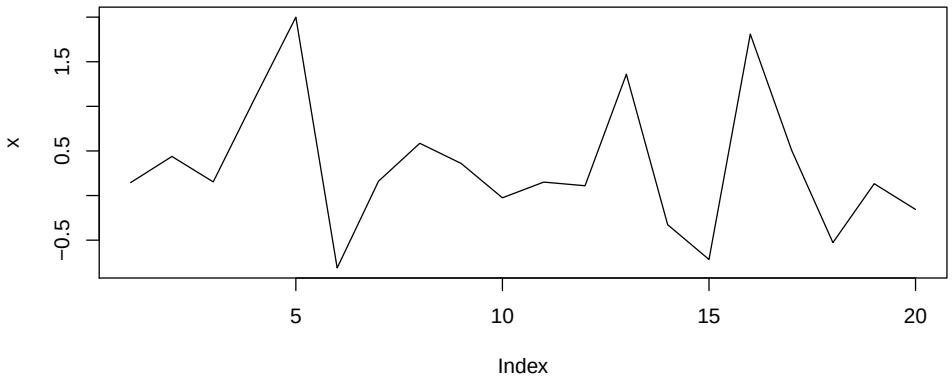
```r
set.seed(1121)
(x=rnorm(20))

##  [1]  0.1449583  0.4383221  0.1531912  1.0849426  1.9995449 -0.8118832  0.1602680
##  [8]  0.5858923  0.3600880 -0.0253084  0.1508809  0.1100824  1.3596812 -0.3269946
## [15] -0.7163819  1.8097690  0.5084011 -0.5274603  0.1327188 -0.1559430

mean(x);var(x)

## [1] 0.3217385
## [1] 0.5714534
```

The first element of `x` is 0.1449583. Boring boxplots and histograms recorded by the PDF device:

```r
par(mar=c(4,4,.1,.1),cex.lab=.95,cex.axis=.9,mgp=c(2,.7,0),tcl=-.3,las=1)
boxplot(x)
hist(x,main='')
```



Do the above chunks work? You should be able to compile the TEX The first element of `x` is 0.1449583. Boring boxplots and histograms recorded by the PDF device:



Do the above chunks work? You should be able to compile the TEX