Lucy Lin
Professor Finch
AMS 315
10 November 2020

AMS 315 Project 1 Part A

Introduction

The tasks for Part A involves merging and sorting the two CSV files for a randomized set of independent and dependent values. The next step is testing the null hypothesis that the slope of the regression line is zero, or that there is no correlation between the independent and dependent variables. These tasks were performed using R.

Methodology

The task of merging and sorting the data for the independent and dependent variables was done in R to find the number of missing values (with the help of the One Predictor Linear Regression Handout). 8 subject IDs did not have either value or 679 had at least one of the values. There were 162 subject IDs with at least one missing data value, 629 had an independent variable, 583 with a dependent variable, and 533 had both values. The pattern for missing data is included below.

For imputing the data values, the 8 empty values were deleted and “linear regression using bootstrap” was performed from R’s MICE package to generate appropriate estimates for the independent or dependent variables that were missing their dependent or independent values respectively.

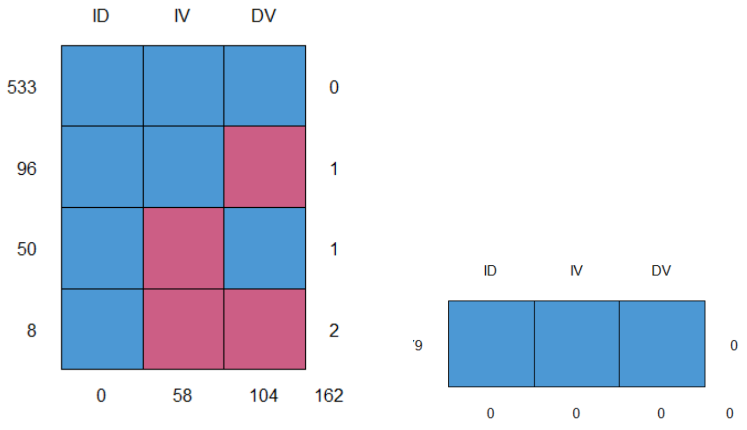
Results

The value for the linear regression line was R^2 =0.7021, which indicates that the fraction or percentage of about 70% of the variation in the dependent variable was explained. This value is shown in the ANOVA table attached. The fitted function of the linear regression is $(y) = 15.1887(x) + (-18.2517)$. The 95% confidence interval for the slope is (14.44214,15.93524), and the 99% confidence interval for the slope is (14.20654, 16.17084). For the null hypothesis that the slope is equal to 0, the p value was less than 2E-16, thus the null hypothesis is rejected.

Conclusion/Discussion

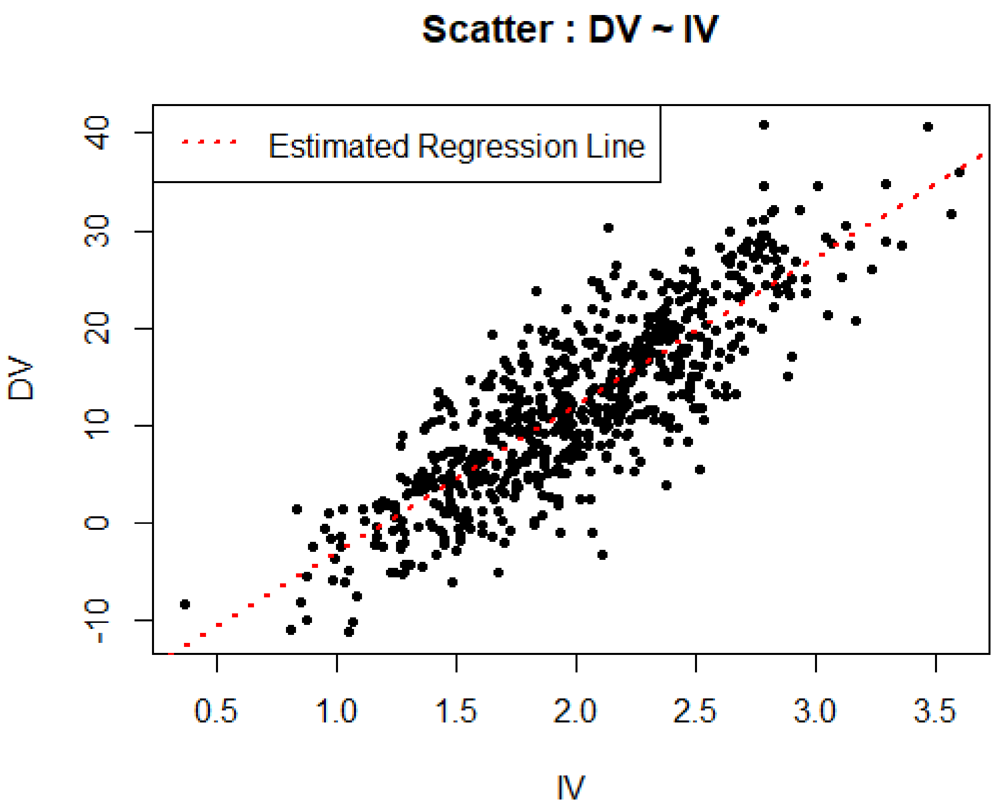
With the R^2value of 0.7021, it can be determined that there is a significant association between the independent and dependent variables. An R value of 0.7 shows that there is a correlation, but it is not particularly extreme. The fitted function was $y = 15.1887x - 18.2517$. Additionally, the p value indicates that the probability that the null hypothesis should be accepted is extremely low, further supporting the association between the variables.

Pattern of Missing Data (Before and After Imputation)



ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	38593.97	38593.97084	1595.777	0
Residuals	677	16373.29	24.18507	NA	NA



95% Confidence Interval of Slope		
	2.5%	97.5%
Intercept	-19.79255	-16.71091
IV	14.44214	15.93524

99% Confidence Interval		
	0.5%	99.5%
Intercept	-20.27880	-16.22466
IV	14.20654	16.17084

Summary():

```
Call:

lm(formula = DV ~ IV, data = PartA_complete)

Residuals:

Min           1Q            Median        3Q            Max

-17.0146      -3.2619       0.0325        3.1339        16.8096

Coefficients:

              Estimate      Std. Error    t value       Pr(>|t|)

(Intercept)   -18.2517      0.7847        -23.26        <2e-16 ***

IV            15.1887       0.3802        39.95         <2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.918 on 677 degrees of freedom

Multiple R-squared:  0.7021,    Adjusted R-squared:  0.7017

F-statistic:  1596 on 1 and 677 DF,  p-value: < 2.2e-16
```