First Computing Assignment

Vincent Forca (112620895)

April 6th, 2020

AMS 315
Data Analysis

Professor Stephen Finch

Stonybrook University

Spring 2020

AMS 315 data analysis computing assignment 1- part A

Vincent Forca (112620895)
Professor Stephen Finch

INTRODUCTION

The focus of Part A was to merge two data files containing dependent variables and the independent variables, deal with and impute missing data, and regenerate the regression equation used to obtain the value of the dependent variable based on the independent variable value.

METHODS AND MATERIALS

The statistical program used to achieve this outcome is called The R project for statistical Computing. First, the data files were separately read into the R program with code used to read the CSV file. Then, they were merged and sorted by ID number. Mice was utilized when imputing the missing data and the imputation method of choice was linear regression using bootstrap. Applying the 'Goodness of fit' test, the model was found to fit the data.

OUTCOMES

There were 713 total observations. 493 observations were complete data sets containing both an independent and dependent variable. An independent variable was missing in 69 data sets; there were 644 total cases with an independent variable. A dependent variable was missing in 171 data sets; there were 542 total cases with a dependent variable. There were 20 cases missing both an independent and dependent variable, leaving 693 data sets with at least 1 independent or 1 dependent variable. Using code in R, we were able to obtain a summary of the regression model:

Coefficients:		Estimate	Std. Error	t value	Pr(>|t|) (number of stars show significance)
	{Intercept	11.2612	0.9235	12.19	<2e-16 ***
	{IV	3.8135	0.1477	25.82	<2e-16 ***

With this data, we reconstructed the regression equation DV=B0+B1*IV: **Y(x) = 11.261+3.813x**.
The P-values shown are significant, so we can reject the null hypothesis that the slope is zero and accept the alternative hypothesis. Furthermore, we obtained the following 95% confidence intervals for the coefficients:
Intercept 95% CI: **[9.447057, 13.075549]**; and ***IV 95% CI:*** **[3.523504, 4.103448]**; Reject Ho.
The fraction of the variation of the dependent variables explained by the variation in the independent variable is given by our adjusted R squared value: 0.4903. (multiple R squared value: 0.4911)

The analysis of variance table:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
IV	1	41334.84	41334.8394	666.7297	0
Residuals	691	42839.51	61.9964	NA	NA

DISCUSSION

We rejected the null hypothesis and accepted the alternative hypothesis. Based on the P-values, it is safe to say that there is a relevant association between the two variables. The adjusted R-squared value is: 0.4903, which is a significant value, and shows a moderate association between the variables. The Multiple R-squared value is: 0.4911, the residual standard error value is: 7.874 on 691 DF, and the F-statistic value is relatively high: 666.7 on 1, 691 DF. Therefore, the model fits the data well.

AMS 315 data analysis computing assignment 1- part B

Vincent Forca (112620895)
Professor Stephen Finch

INTRODUCTION

The focus of part B was to recover the function using single predictor linear regression. Given a single .csv file containing a list of independent and dependent variables, we are expected to use a transformation to help the model fit better, bin near-data into one level, and apply an approximate lack of fit test.

METHODS AND MATERIALS

The data produced 431 observations. We achieved these outcomes using the R project for statistical Computing and attempted numerous transformations to help establish a better linear regression and obtain better results. We compared the results of the standard linear regression to the results of the Exponential model (DV=ln(y)), the Quadratic model (dv=sqrt(y)), the Reciprocal model (DV=1/y), the Logarithmic model (IV=ln(x)), and the Power model (DV=ln(y), IV= ln(x)). The most successful transformation was the Power model. We verified the fit of the transformation by binning near-data into groups separated by a value of 0.04 and applying an approximate lack of fit test.

OUTCOMES

The original data produced the following coefficient estimates: the coefficient Intercept estimate is 20.85130, the coefficient X estimate is 1.90056. This produced the following regression model DV = b0+b1*IV: Y=20.85130+1.90056x. The adjusted R-squared value was 0.5458. (multiple R squared was 0.5468). The original data's 95% CI for the intercept is (20.117205, 21.585397). The original data's 95% CI for X is (1.736375, 2.064741). The F statistic was 517.7 on 1, 429 DF and the residual standard error was 2.75 on 429 DF. After transforming the original data to the Power model, we obtained the coefficient Intercept estimate: 2.98552, and the coefficient X estimate: 0.26991.
This produced the following regression model: ln(DV)=BO + B1*ln(IV): **ln(y)=2.98552+0.26991*ln(x)**.
We obtained the following transformed 95% confidence intervals for the coefficients:
intercept 95% CI: **(2.9550169, 3.0160252)**; and ***X 95% CI:*** **(0.2483387, 0.2914775)**.
The Power model was the most successful transformation, resulting in the highest R-squared value out of all other transformation attempts: An adjusted R squared value of 0.5841. (multiple R squared was 0.5841). The improved F statistic is 604.9 on 1, 429 DF, and the improved residual standard error is 0.09544 on 429 DF. The p-value remained the same for both models and verified associated between variables: (P-value: < 2.2e-16). After binning near data, we calculated the following ANOVA table:

The Analysis of Variance Table:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	5.5115	5.5115	630.947	< 2e-16 ***
Residuals	429	3.9073	0.0091	-	-
Lack of fit	36	0.4743	0.0132	1.5082	0.03351 *
Pure Error	393	3.4330	0.0087	-	-

DISCUSSION

The Power model transformation was successful. The P-value on both the original, and transformed data is: (p-value: < 2.2e-16). This is significant, showing strong association between the variables, so we can reject the null hypothesis that the slope is zero and accept the alternative hypothesis. The R-squared values, F-values and residual error values all improved – and the lack of fit F value is 1.5082, showing that there is no significant lack of fit.

Part A Computer Data Appendix:

(some code and results from R)

```
ID IV  DV
493  1  1   1   0
151  1  1   0   1
49   1  0   1   1
20   1  0   0   2
   0 69 171 240
'ANOVA Table'
|          | Df|   Sum Sq|    Mean Sq|  F value| Pr(>F)|
|IV        |   1| 41334.84| 41334.8394| 666.7297|     0|
|Residuals | 691| 42839.51|    61.9964|       NA|    NA|
Coefficients:
(Intercept)          IV
   11.261       3.813
Residuals:
    Min      1Q   Median      3Q     Max
-22.2257  -5.0273   0.1224   5.0274  22.7346
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2612     0.9235   12.19   <2e-16 ***
IV            3.8135     0.1477   25.82   <2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.874 on 691 degrees of freedom
Multiple R-squared:  0.4911,    Adjusted R-squared:  0.4903
F-statistic: 666.7 on 1 and 691 DF,  p-value: < 2.2e-16
> confint(model)
                  2.5 %    97.5 %
(Intercept) 9.447957 13.074459
```

IV 3.523504 4.103448

> CI of each variable (LOWER BOUND)

1 2 3 4 5 6 7 8

31.098980 27.743309 36.324751 35.088917 35.163572 24.780029 33.703679 34.048954

9 10 11 12 13 14 15 16

17.326060 27.134040 37.486909 28.841262 31.286454 42.424078 41.838312 32.180875

……

END PART A APPENDIX

Part B Appendix:

(some code and results from R)

```
> data <- read.csv('P1B20895.csv', header = TRUE)
> str(data)
'data.frame':   431 obs. of  3 variables:
 $ ID: int  1 2 3 4 5 6 7 8 9 10 ...
 $ x : num  2.25 4.7 1.86 2.75 6.3 ...
 $ y : num  26.5 31.6 19.1 30.4 36.7 ...
> View(data)
> M <- lm(y ~ x, data = data)
```

ORIGINAL DATA

```
> confint(M)
               2.5 %    97.5 %
(Intercept) 20.117205 21.585397
x           1.736375  2.064741
lm(formula = y ~ x, data = data)
```

Residuals:

```
   Min     1Q  Median     3Q    Max
-6.9422 -1.7753  0.1061  1.7826  9.6538
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.85130    0.37349   55.83   <2e-16 ***
x            1.90056    0.08353   22.75   <2e-16 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.725 on 429 degrees of freedom
Multiple R-squared:  0.5468,    Adjusted R-squared:  0.5458
F-statistic: 517.7 on 1 and 429 DF,  p-value: < 2.2e-16
```

Transformed data:

ANOVA: (Tran)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	5.510751	5.5107509	604.9302	0
Residuals	429	3.908074	0.0091097	NA	NA

```
> tran <- transform(data, x = log(x), y = log(y))
> XD <- lm(y ~ x, data=tran)
> summary(XD)

Call:
lm(formula = y ~ x, data = tran)

Residuals:
     Min       1Q   Median       3Q      Max
-0.304018 -0.056840  0.007225  0.064941  0.297970
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.98552    0.01552  192.37   <2e-16 ***
x            0.26991    0.01097   24.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

summary(XD)
Call:
lm(formula = y ~ x, data = tran)
Residuals:
     Min       1Q   Median       3Q      Max
-0.304018 -0.056840  0.007225  0.064941  0.297970
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.98552    0.01552  192.37   <2e-16 ***
x            0.26991    0.01097   24.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.09544 on 429 degrees of freedom
Multiple R-squared:  0.5851,    Adjusted R-squared:  0.5841
F-statistic: 604.9 on 1 and 429 DF,  p-value: < 2.2e-16
> confint(XD)
                2.5 %    97.5 %
(Intercept) 2.9550169 3.0160252
x           0.2483387 0.2914775
> grou <- cut(tran$x,breaks=c(-Inf,seq(min(tran$x)+0.04, max(tran$x)-0.04,by=0.04),Inf))
> table(grou)
grou
```

```
 (-Inf,0.451] (0.451,0.491] (0.491,0.531] (0.531,0.571] (0.571,0.611]
            4            5            5            6           10
(0.611,0.651] (0.651,0.691] (0.691,0.731] (0.731,0.771] (0.771,0.811]
            5            6            4            8            6
(0.811,0.851] (0.851,0.891] (0.891,0.931] (0.931,0.971]  (0.971,1.01]
           11            6            9           11            8
 (1.01,1.05]   (1.05,1.09]   (1.09,1.13]   (1.13,1.17]   (1.17,1.21]
            6           11            5           14           14
 (1.21,1.25]   (1.25,1.29]   (1.29,1.33]   (1.33,1.37]   (1.37,1.41]
           14           11            9            9           10
 (1.41,1.45]   (1.45,1.49]   (1.49,1.53]   (1.53,1.57]   (1.57,1.61]
           10           16           21           16           11
 (1.61,1.65]   (1.65,1.69]   (1.69,1.73]   (1.73,1.77]   (1.77,1.81]
           12           20           18           16           23
 (1.81,1.85]   (1.85,1.89]   (1.89, Inf]
           18           21           22
> xxxx<-ave(tran$x, grou)
> data_binned <- data.frame(x=xxxx, y=tran$y)
> fitted_b <- lm(y ~ x, data = data_binned)
> pureErrorAnova(fitted_b)
Analysis of Variance Table
Response: y
              Df Sum Sq Mean Sq  F value  Pr(>F)
x              1 5.5115  5.5115 630.9478 < 2e-16 ***
Residuals    429 3.9073  0.0091
 Lack of fit  36 0.4743  0.0132   1.5082 0.03351 *
 Pure Error  393 3.4330  0.0087
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

End part B appendix

- End of Computing Assignment One -