Exam 1 Take Home Preparation Portion
QMB 3200: Advanced and Quantitative Methods
Fall 2019

Regression and log transformation model

Submitted to
Dr. Jim Dewey
Florida Polytechnic University

Submitted by
Luiz Gustavo Fagundes Malpele
Department of Data Science
Florida Polytechnic University

October 26th, 2019

Route Optimization System Analysis

1.1 Average Effect

Summary statistics of the variable undelivered given that the driver is using the system (1)

Variable	Obs	Mean	Std. Dev.	Min	Max
undelivered	**67**	**4.750038**	**2.082427**	**1.585087**	**14.70452**

The data from this table shows mean, standard deviation, minimum and maximum for the undelivered variable given that the driver is using the real time dynamic route optimization program. It noticeable that the mean value is lower than if the driver was not using the system (Table 3).

Summary statistics of the variable undelivered given that the driver is not using the system (2)

Variable	Obs	Mean	Std. Dev.	Min	Max
undelivered	**69**	**7.016306**	**5.886765**	**1.31344**	**21**

The data from this table shows mean, standard deviation, minimum and maximum for the undelivered variable given that the driver is not using the real time dynamic route optimization program. The mean difference between the two table are noticeable, if the driver is not using the system the mean undelivered packages is 7.016, while if the driver is using it is 4.75.

1.2 Confidence intervals

95% Confidence interval for undelivered given that the driver is using the system (3)

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
undelivered	**67**	**4.750038**	**.2544089**	**4.242094**	**5.257981**

From this confidence interval, it can be said that we are 95% confident that the mean undelivered packages for a driver that is using the real time dynamic route optimization program relies between 4.242 and 5.258.

95% Confidence interval for undelivered given that the driver is using the system (4)

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
undelivered	**69**	**7.016306**	**.7086832**	**5.60215**	**8.430461**

From this confidence interval, it can be said that we are 95% confident that the mean undelivered packages for a driver that is not using the real time dynamic route optimization program relies between 5.602 and 8.43. It does make a difference if homoscedasticity is assumed since the variances are different if the driver is using the system (Std. Err. = 0.254) or not (Std. Err. = 0.709).

1.3 T-test

T-test assuming heteroscedasticity to check if the system saves on average at least 1 package per truck per day (5)

Two-sample t test with unequal variances

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	69	7.0163	.7086874	5.8868	5.602136	8.430464
y	67	5.75	.2556273	2.0924	5.239624	6.260376
combined	136	6.392461	.38348	4.472107	5.634056	7.150866
diff		1.2663	.7533812		-.2315332	2.764133

diff = mean(x) - mean(y) t = 1.6808
Ho: diff = 0 Satterthwaite's degrees of freedom = 85.3571

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.9518 Pr(|T| > |t|) = 0.0965 Pr(T > t) = 0.0482

For this problem it was assumed the mean of undelivered packages for a driver that is using the real time dynamic route optimization program is 5.75 which is the same as 4.75 + 1, this assumption helped to set the Ho: diff = 1, Ha: diff > 1, and $\alpha = 0.05$. The null hypothesis is retained as the p-value is equal to 0.0482 and $\alpha > 0.0482$. We conclude that there is strong evidence to state that the system saves on average at least 1 undelivered package per truck per day.

T-test assuming homoscedasticity to check if the system saves on average at least 1 package per truck per day (6)

Two-sample t test with equal variances

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	69	7.0163	.7086874	5.8868	5.602136	8.430464
y	67	5.75	.2556273	2.0924	5.239624	6.260376
combined	136	6.392461	.38348	4.472107	5.634056	7.150866
diff		1.2663	.7620886		-.2409785	2.773579

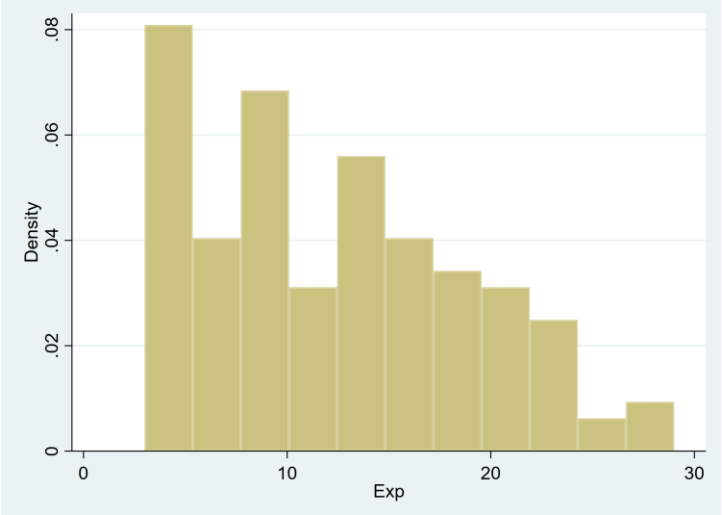
diff = mean(x) - mean(y) t = 1.6616
Ho: diff = 0 degrees of freedom = 134

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0
Pr(T < t) = 0.9505 Pr(|T| > |t|) = 0.0989 Pr(T > t) = 0.0495

Testing assuming homoscedasticity changed the p-value from 0.0482 to 0.0495, so it can be said there was a significance difference since p-value is much closer to $\alpha = 0.05$. If the null hypothesis is incorrectly rejected, a type 1 error.

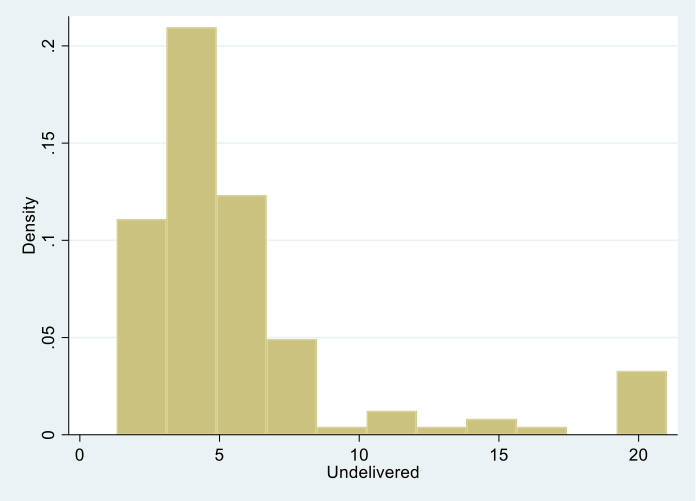
2.1 Frequency distribution and variables summary

Frequency distribution for experience (7)



The histogram seems slightly negatively skewed for drivers experience on similar jobs. It can be said that there is a significant number of drivers below 10 years of experience.

Frequency distribution for undelivered packages (8)



The histogram seems normal distributed with a mean around 4 and 6 undelivered packages. It is noticeable that there are outliers that are around 20 undelivered packages.

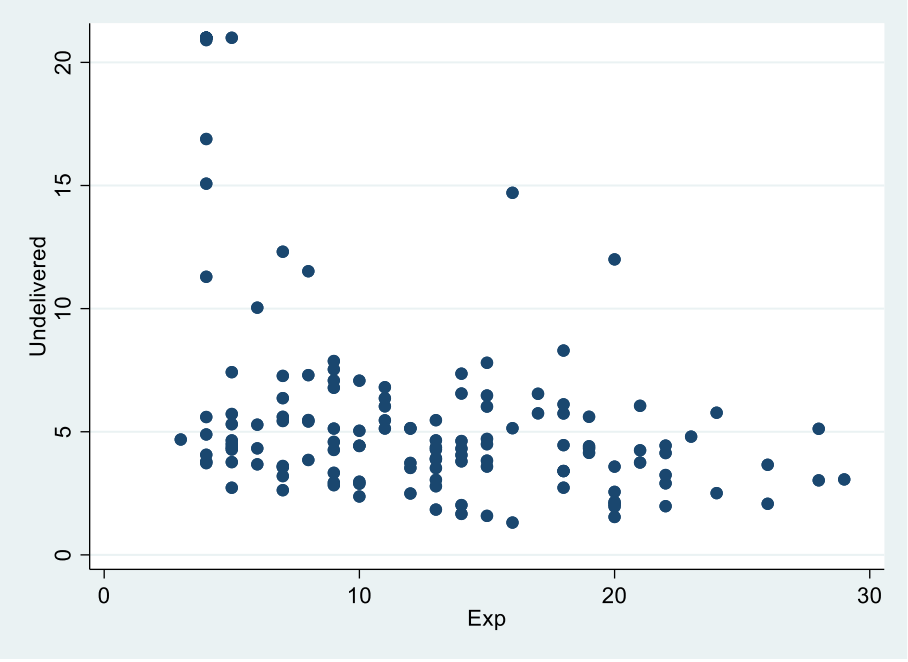
Summary statistics of the variables undelivered and exp (9)

Variable	Obs	Mean	Std. Dev.	Min	Max
exp	136	12.36765	6.404638	3	29
undelivered	136	5.899835	4.568221	1.31344	21

The data of the report is above, summary statistics, such as mean, standard deviation, minimum and maximum were used to describe the data for 2 different variables: exp (years the driver has been employed in similar jobs and undelivered (number of undelivered packages for the week). It noticeable that undelivered had a mean of 5.8998 undelivered packages a week and maximum of 21.

2.2 Scatterplot for data visualization

Scatterplot of undelivered packages against driver experience (10)



This scatterplot shows a trend that the more experienced the driver, the lower the number of packages he is likely to achieve. The max values related to undelivered are directly related to the less experienced drivers.

2.3 Regression data

Regression on Undelivered Packages against Driver Experience (11)

Source	SS	df	MS		
Model	481.515053	1	481.515053	Prob > F	0.0000
Residual	2335.7512	134	17.4309791	R-squared	0.1709
Total	2817.26625	135	20.8686389	Adj R-squared	0.1647
				Root MSE	4.175

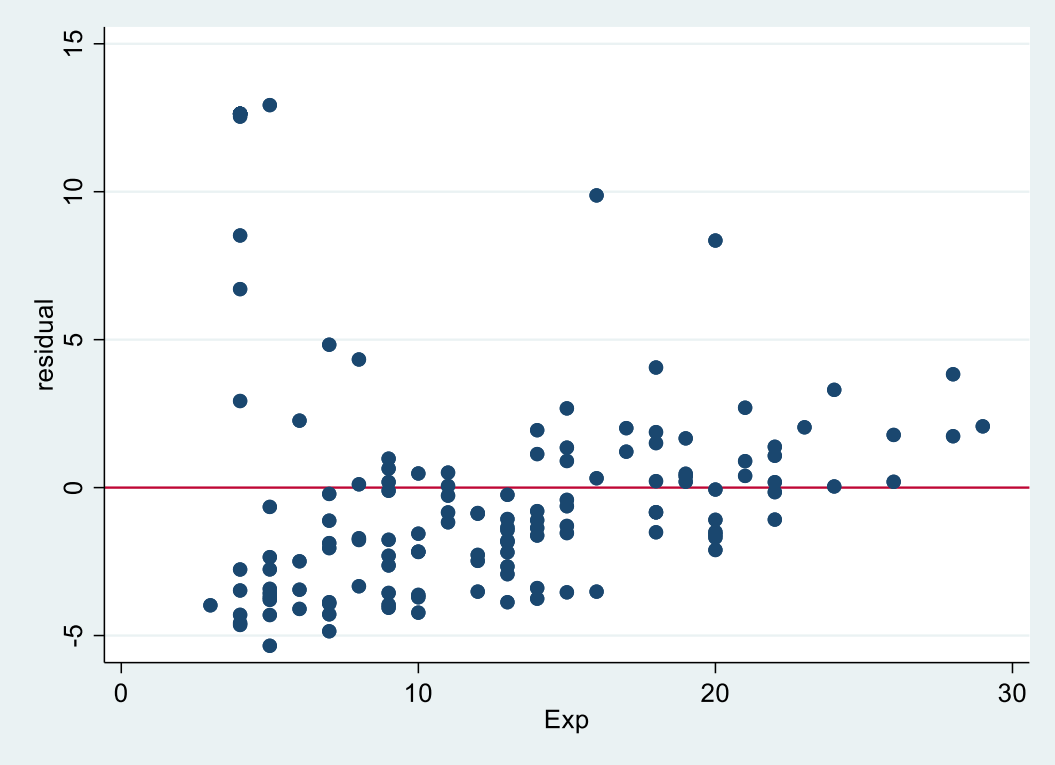
undelivered	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	-.2948787	.0561047	-5.26	0.000	-.4058441	-.1839134
_cons	9.546791	.7807965	12.23	0.000	8.002512	11.09107

The coefficients are $\beta\hat{}_0$ = 9.5468 which is the y-intercept and $\beta\hat{}_1$ = -0.2948 the slope of the regression line. While referring to the undelivered variable, the standard error of β_0 is 0.7808 and this standard error is used to calculate the 95% confidence interval. We are 95% that the coefficient relies between 8.003 and 11.0911. Lastly, the t-value = 12.23 and it can be said that there is 0.000% that Ho: β_0 = 0 and the null hypothesis is rejected.

The same concepts apply while referring to β_1 coefficient, the standard error is 0.0561. We are 95 % confident that this coefficient relies between -0.4058 and -0.1839. Lastly, the t-value is -5.26 and it can be said that there is 0.000% that Ho: β_0 = 0 and the null hypothesis is rejected.

2.4 Scatterplot of undelivered packages' residual and Driver Experience

Undelivered variable's residuals against driver experience (12)



The plot suggests heteroscedasticity, the variance of residuals on lower values for experience seems larger than on higher values of the Exp variable, this fact supports that the plot is heteroscedastic.

2.5 Robust Regression

Heteroscedastic Regression on Undelivered Packages against Driver Experience (13)

Linear regression

Number of obs	=	**136**	
F(1, 134)	=	**19.87**	
Prob > F	=	**0.0000**	
R-squared	=	**0.1709**	
Root MSE	=	**4.175**	

undelivered	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	**-.2948787**	**.0661532**	**-4.46**	**0.000**	**-.4257183**	**-.1640391**
_cons	**9.546791**	**1.102503**	**8.66**	**0.000**	**7.366232**	**11.72735**

Robust linear regression should be used since it provides a more accurate standard error for these two variables given that the variance of number of undelivered packages varies according years of experience. The variance of the number of undelivered packages for lower values of years of experience is much higher than for higher values of years of experience, so this fact implies heteroscedasticity and the use of robust standard errors.

2.6 Predicting the number of undelivered packages

Robust linear regression should be used since it provides a more accurate standard error for these two variables given that they are heteroscedastic and robust regression is the most appropriate method when variable have unequal variances.

$$\beta_0 + \beta_1 * X_i = 9.547 - 0.29488 * 12 = 6.008$$

A person on the 50th percentile is expected to have 6.008 undelivered packages according the linear regression predictions.

$$\beta_0 + \beta_1 * X_i = 9.547 - 0.29488 * 4 = 8.367$$

A person on the 10th percentile is expected to have 8.367 undelivered packages according the linear regression predictions.

2.7 Implication of additional 4 years of experience

Based on the regression line, we have that:

$$\beta_1 * X_i = -0.29488 * X_i$$

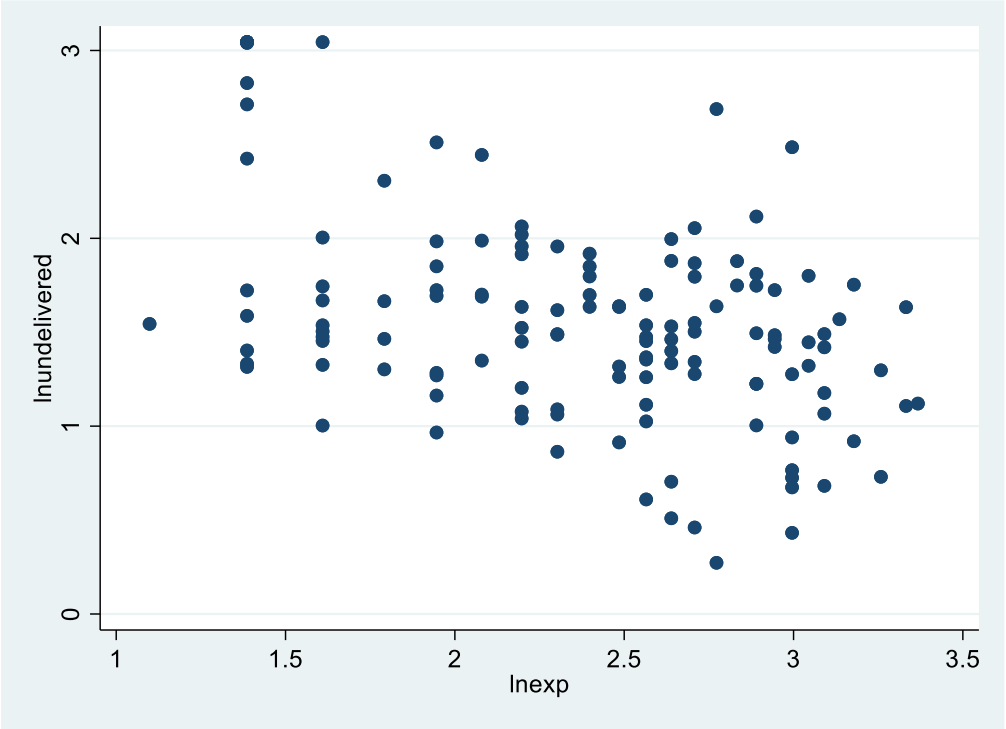
If $X_i = 4$, referring to the additional four years of experience a driver would have assuming the regression line is heteroscedastic:

$$\beta_1 * X_i = -0.29488 * 4 = -1.17952$$

Therefore, there is enough strong evidence to say that 4 years of additional experience is associated with a decrease in undelivered packages by at least 1 per truck per year.

2.8. Scatter of log transformed variables

Undelivered variable's residuals against driver experience (14)



The plot suggests heteroscedasticity, the variance of residuals on lower values for experience seems larger than on higher values of the lnexp variable, on [1,2] interval for lnundelivered values are around 1.5 and 3, while for the interval [2,3.5] of lnexperience the values for lnundelivered are more likely to range around 0.5 and 2, this fact supports that the plot is heteroscedastic.

2.9 Regression of log transformed variable

Regression on the natural log of undelivered against the natural log of experience (15)

Source	SS	df	MS		
				Number of obs	= 136
				F(1, 134)	= 40.35
Model	10.7056364	1	10.7056364	Prob > F	= 0.0000
Residual	35.5492373	134	.265292816	R-squared	= 0.2314
				Adj R-squared	= 0.2257
Total	46.2548737	135	.342628694	Root MSE	= .51507

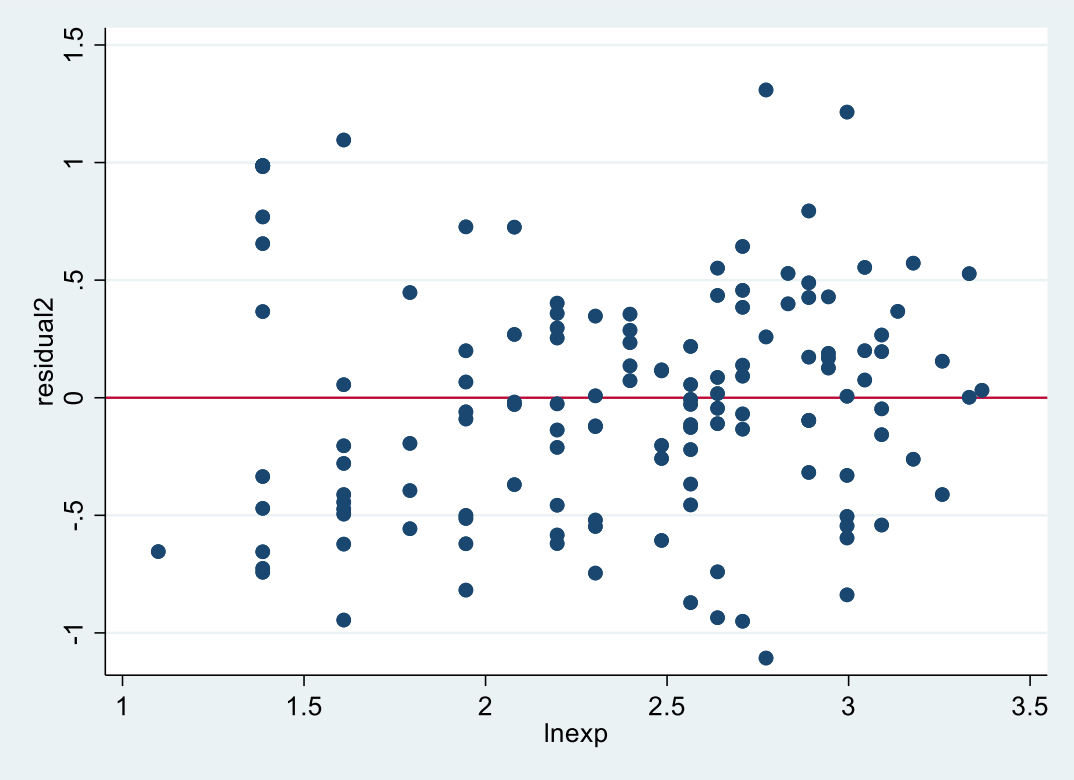
lnundelive~d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnexp	-.4893011	.0770252	-6.35	0.000	-.6416435	-.3369588
_cons	2.736098	.1874139	14.60	0.000	2.365426	3.10677

The coefficients are $\beta\hat{}_0$ = 2.7361 which is the y-intercept and $\beta\hat{}_1$ = -0.4893 the slope of the regression line. While referring to the undelivered variable, the standard error of β_0 is 0.1874 and this standard error is used to calculate the 95% confidence interval. We are 95% that the coefficient relies between 2.3654 and 3.1068. Lastly, the t-value = 14.60 and it can be said that there is 0.000% that Ho: β_0 = 0 and the null hypothesis is rejected.

The same concepts apply while referring to β_1 coefficient, the standard error is 0.077. We are 95 % confident that this coefficient relies between -0.6416 and -0.3370. Lastly, the t-value is -6.35 and it can be said that there is 0.000% that Ho: β_0 = 0 and the null hypothesis is rejected.

2.10 Scatterplot log transformed residual

Scatter plot of natural log of undelivered' s residual and the natural log of driver experience (16)



The plot suggests heteroscedasticity, the variance of residuals on lower values for experience seems larger than on higher values of the Exp variable, this fact supports that the plot is heteroscedastic.

2.11 Robust regression of log transformed variables

Heteroscedastic regression on the natural log of undelivered against the natural log of experience (17)

Linear regression

Number of obs	=	**136**
F(1, 134)	=	**28.44**
Prob > F	=	**0.0000**
R-squared	=	**0.2314**
Root MSE	=	**.51507**

lnundelive~d	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnexp	**-.4893011**	**.0917544**	**-5.33**	**0.000**	**-.6707755**	**-.3078268**
_cons	**2.736098**	**.2343507**	**11.68**	**0.000**	**2.272593**	**3.199603**

Robust linear regression should be used since it provides a more accurate standard error for these two variables given that the variance of number of undelivered packages varies according years of experience. The variance of the number of undelivered packages for lower values of years of experience is much higher than for higher values of years of experience, so this fact implies heteroscedasticity and the use of robust standard errors.

2.12 Predicting the number of undelivered packages

Robust linear regression should be used since it provides a more accurate standard error for these two variables given that they are heteroscedastic and robust regression is the most appropriate method when variable have unequal variances.

$$\beta_0 + \beta_1 * X_i = 2.7361 - 0.4893 * 2.4849 = 1.5202$$

$$y = e^{1.5202} = 4.5731$$

A person on the 50th percentile is expected to have 4.5731 undelivered packages according the linear regression predictions.

$$\beta_0 + \beta_1 * X_i = 2.7361 - 0.4893 * 1.3863 = 2.0578$$

$$y = e^{2.0578} = 4.5731$$

A person on the 10th percentile is expected to have 7.8287 undelivered packages according the linear regression predictions.

2.13 Implication of 1% additional experience

While referring to the additional one percent of experience a driver would have assuming the regression line is heteroscedastic and that the change in the natural logarithm is approximately the same as the percentage change, we have:

$$\beta_1 * X_i = -0.4893 * 0.01 * X_i \approx -0.49\% \; undelivered \; packages \; per \; truck \; per \; year$$

Therefore, there is enough strong evidence to say that 1% additional experience is associated with a decrease in at least 0.4% of undelivered packages by at least 1 per truck per year.

2.14 Models Comparison

While comparing the untransformed and the log transformed model, it can be concluded that the log transformed model is better to analyze the relation between experience and undelivered packages, this model provides a higher $R^2 \; value$ and a more precise description of the variable since the percentage change in the natural logarithm of a variable is approximately the same as the percentage change is a variable. Furthermore, the log model is useful to better visualize the relation heteroscedastic variable, which was the case of the association of experience and undelivered packages, while the person was less experienced the variance of undelivered packages was greater than if the person was experienced. Lastly, it was possible to transform the log variable into a regular variable by exponentiating the outcome by $e, such \; as \; e^x$.

2.15 Elements Analysis

SSM: $\sum(X_i - \bar{X})^2$	df_M= 1	MSM = $\frac{SSM}{df_M}$
SSR: $\sum(\varepsilon_i - \bar{\epsilon})^2$	df_R= N – 2	MSR = $\frac{SSR}{df_R}$
SST = SSM + SSR	df_t= $df_M + df_R$	MST = $\frac{SST}{df_T}$

$\widehat{\beta_0} = -0.4893$: Y-intercept of the regression line.

$\widehat{\beta_1} = 2.7360$: Coefficient of variable X_i, the slope of the regression line.

Std. Error: Standard error of the coefficients $\widehat{\beta_0} \; and \; \widehat{\beta_1}$ used to calculate CI and hypothesis testing.

t and P > $|t|$: used to test if $H_0 = 0$

Confidence Interval 95%: $\widehat{\beta_i} + 1.96 * SE_{\widehat{\beta_i}}$

$R^2 = \frac{SST}{SSM}$: Percentage of variability in the response variable.

Root MSE = $\sqrt{\frac{SSR}{df_R}} = \sqrt{\frac{\sum(\varepsilon_i - \bar{\epsilon})^2}{N-2}}$: is the standard deviation of residual, in other words.

Appendix A: Do-file-takehome

```
/* QMB 3200 Takehome exam */
clear
cd "C:\Users\luizg\Desktop\Stata"
log using "C:\Users\luizg\Desktop\Stata\Takehome.smcl", replace
import delimited "C:\Users\luizg\Desktop\Stata\undelivered.csv"

*Q1. summary statistics
sum undelivered
summarize undelivered if new
summarize undelivered if !new

*Q1.2 confidence intervals
ci means undelivered if new
ci means undelivered if !new

*Q1.3 t test with input values
//ttesti (obs1) (mean1) (std1) (obs2) (mean2) (std2)
ttesti 69 7.0163 5.8868 67 5.75 2.0924, unequal
ttesti 69 7.0163 5.8868 67 5.75 2.0924

*Q2.1 histogram and summary statistics
hist exp
hist undelivered
summarize exp undelivered

*Q2.2 scatter plot
scatter undelivered exp

*Q2.3 linear regression
regress undelivered exp

*Q2.4 generating residuals scatter
```

```
predict pundelivered
gen residual=undelivered-pundelivered
scatter residual exp, yline(0)

*Q2.5 robust regression
regress undelivered exp, robust

*Q2.6 generating the percentile values
egen p50 = pctile(exp), p(50)
egen p10 = pctile(exp), p(10)

*2.8 generating ln variables and scatter plot
gen lnexp=ln(exp)
gen lnundelivered = ln(undelivered)
scatter lnundelivered lnexp

*2.9 linear regression
regress lnundelivered lnexp

*2.10 scatter of ln residuals
predict plnundelivered
gen residual2=lnundelivered-plnundelivered
scatter residual2 lnexp, yline(0)

*2.11 robust regression of ln variables
regress lnundelivered lnexp, robust

*2.12 generating variables to check the percentile values
egen pe50 = pctile(lnexp), p(50)
egen pe10 = pctile(lnexp), p(10)

STOP
log close
```

Appendix B: Log-file-takehome

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   16.0   Copyright 1985-2019 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata license expires 16 Mar 2020:
       Serial number:  301609236389
         Licensed to:  Luiz Gustavo Fagundes Malpele
                       Florida Polytechnic University

Notes:
      1.  Unicode is supported; see help unicode_advice.

. doedit "C:\Users\luizg\Desktop\Stata\do-file-takehome.do"

. do "C:\Users\luizg\AppData\Local\Temp\STD4618_000000.tmp"

. /* QMB 3200 Takehome exam */
. clear

. cd "C:\Users\luizg\Desktop\Stata"
C:\Users\luizg\Desktop\Stata

. log using "C:\Users\luizg\Desktop\Stata\Takehome.smcl", replace
-------------------------------------------------------------------------------
-----------------------------------------
      name:  <unnamed>
```

```
      log:  C:\Users\luizg\Desktop\Stata\Takehome.smcl
  log type:  smcl
 opened on:  27 Oct 2019, 17:09:02

. import delimited "C:\Users\luizg\Desktop\Stata\undelivered.csv"
(3 vars, 136 obs)

.
. *Q1. summary statistics
. sum undelivered

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 undelivered |        136    5.899835    4.568221    1.31344         21

. summarize undelivered if new

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 undelivered |         67    4.750038    2.082427   1.585087   14.70452

. summarize undelivered if !new

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 undelivered |         69    7.016306    5.886765    1.31344         21

.
. *Q1.2 confidence intervals
. ci means undelivered if new

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
```

```
 undelivered |         67    4.750038    .2544089        4.242094    5.257981

. ci means undelivered if !new

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
 undelivered |         69    7.016306    .7086832         5.60215    8.430461

.
. *Q1.3 t test with input values
. //ttesti (obs1) (mean1) (std1) (obs2) (mean2) (std2)
. ttesti 69 7.0163 5.8868 67 5.75 2.0924, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      69      7.0163    .7086874      5.8868    5.602136    8.430464
       y |      67        5.75    .2556273      2.0924    5.239624    6.260376
---------+--------------------------------------------------------------------
combined |     136    6.392461      .38348    4.472107    5.634056    7.150866
---------+--------------------------------------------------------------------
    diff |              1.2663    .7533812               -.2315332    2.764133
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   1.6808
Ho: diff = 0                     Satterthwaite's degrees of freedom =  85.3571

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9518         Pr(|T| > |t|) = 0.0965          Pr(T > t) = 0.0482

. ttesti 69 7.0163 5.8868 67 5.75 2.0924

Two-sample t test with equal variances
```

```
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      69      7.0163    .7086874      5.8868    5.602136    8.430464
       y |      67        5.75    .2556273      2.0924    5.239624    6.260376
---------+--------------------------------------------------------------------
combined |     136    6.392461      .38348    4.472107    5.634056    7.150866
---------+--------------------------------------------------------------------
    diff |              1.2663    .7620886               -.2409785    2.773579
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   1.6616
Ho: diff = 0                                     degrees of freedom =      134

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9505         Pr(|T| > |t|) = 0.0989          Pr(T > t) = 0.0495

.
. *Q2.1 histogram and summary statistics
. hist exp
(bin=11, start=3, width=2.3636364)

. hist undelivered
(bin=11, start=1.3134398, width=1.7896873)

. summarize exp undelivered

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         exp |        136    12.36765    6.404638          3         29
 undelivered |        136    5.899835    4.568221    1.31344         21

.
. *Q2.2 scatter plot
```

```
. scatter undelivered exp

.
. *Q2.3 linear regression
. regress undelivered exp

      Source |       SS           df       MS      Number of obs   =       136
-------------+----------------------------------   F(1, 134)       =     27.62
       Model |  481.515053         1  481.515053   Prob > F        =    0.0000
    Residual |   2335.7512       134  17.4309791   R-squared       =    0.1709
-------------+----------------------------------   Adj R-squared   =    0.1647
       Total |  2817.26625       135  20.8686389   Root MSE        =     4.175

------------------------------------------------------------------------------
 undelivered |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |  -.2948787   .0561047    -5.26   0.000    -.4058441   -.1839134
       _cons |   9.546791   .7807965    12.23   0.000     8.002512    11.09107
------------------------------------------------------------------------------

.
. *Q2.4 generating residuals scatter
. predict pundelivered
(option xb assumed; fitted values)

. gen residual=undelivered-pundelivered

. scatter residual exp, yline(0)

.
. *Q2.5 robust regression
. regress undelivered exp, robust
```

```
Linear regression                               Number of obs     =        136
                                                F(1, 134)         =      19.87
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1709
                                                Root MSE          =      4.175

------------------------------------------------------------------------------
             |               Robust
 undelivered |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |  -.2948787   .0661532    -4.46   0.000    -.4257183   -.1640391
       _cons |   9.546791   1.102503     8.66   0.000     7.366232    11.72735
------------------------------------------------------------------------------

.
. *Q2.6 generating the percentile values
. egen p50 = pctile(exp), p(50)

. egen p10 = pctile(exp), p(10)

.
. *2.8 generating ln variables and scatter plot
. gen lnexp=ln(exp)

. gen lnundelivered = ln(undelivered)

. scatter lnundelivered lnexp

.
. *2.9 linear regression
. regress lnundelivered lnexp

      Source |       SS           df       MS      Number of obs   =       136
```

```
-------------+----------------------------------   F(1, 134)       =     40.35
       Model |  10.7056364         1  10.7056364   Prob > F        =    0.0000
    Residual |  35.5492373       134  .265292816   R-squared       =    0.2314
-------------+----------------------------------   Adj R-squared   =    0.2257
       Total |  46.2548737       135  .342628694   Root MSE        =    .51507

------------------------------------------------------------------------------
lnundelive~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnexp |  -.4893011   .0770252    -6.35   0.000    -.6416435   -.3369588
       _cons |   2.736098   .1874139    14.60   0.000     2.365426     3.10677
------------------------------------------------------------------------------

.
. *2.10 scatter of ln residuals
. predict plnundelivered
(option xb assumed; fitted values)

. gen residual2=lnundelivered-plnundelivered

. scatter residual2 lnexp, yline(0)

.
. *2.11 robust regression of ln variables
. regress lnundelivered lnexp, robust

Linear regression                               Number of obs     =        136
                                                F(1, 134)         =      28.44
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2314
                                                Root MSE          =     .51507

------------------------------------------------------------------------------
```

```
             |               Robust
lnundelive~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnexp |  -.4893011   .0917544    -5.33   0.000    -.6707755   -.3078268
       _cons |   2.736098   .2343507    11.68   0.000     2.272593    3.199603
------------------------------------------------------------------------------

.
. *2.12 generating variables to check the percentile values
. egen pe50 = pctile(lnexp), p(50)

. egen pe10 = pctile(lnexp), p(10)

.
end of do-file

. log close
      name:  <unnamed>
       log:  C:\Users\luizg\Desktop\Stata\Takehome.smcl
  log type:  smcl
 closed on:  27 Oct 2019, 17:09:35
-------------------------------------------------------------------------------------------
----------------------------------------
```