STAT 410 Project

University Students' Monthly Expenses

(Box Cox/Gamma Regressions)

Submitted to
Dr. Olga Korosteleva

Report Prepared By
Ethan Huang

November 30, 2022

Table of Contents

I. Introduction

For this project, I wanted to find a clean, workable dataset to conduct a regression analysis. As I looked through Kaggle, I found a dataset relevant to university students like myself. The dataset profiles university students, their habits, and their monthly expenses. The response variable, monthly expenses, is right-skewed. Therefore, we will be applying a Box-Cox Transformation and Gamma Regression Model to understand which factors influence students' monthly expenses.

II. The Dataset

The data, found through Kaggle, surveys 106 university students from an open-source survey. The explanatory variables are age, gender, year of study, living situation, transportation, scholarship, employment, drinking, games, and subscription. The response variable is total monthly expenses.

However, there are missing values in the original dataset, so I. replaced them. For numeric variables such as age and monthly expenses, I utilized the median value to substitute for the missing values. For categorical variables, I used the most frequent value for substitution. For example, if the dominant value of a variable is "yes", I would replace the missing value with "yes."

III. Box Cox Transformation

To find an optimal Lambda, I used SAS and R. Both software reported the optimal lambda to be -1.25. The transformed response is 4*(1-(1/y^0.25)).
The fitted model is as follows:

E(Transformed Monthly Expense)=2.7347 + 0.0073*(Male) + 0.0036*(Age) + 0.0192*(Home) + 0.0081*(Study Year) + 0.0071*(Scholarship) + 0.0034*(No Job) + 0.0768*(Car) + 0.0502*(Motorcycle) + 0.0151*(No Drinking) + 0.0369*(No Games) + 0.0400*(Monthly Subscription)

The model fits the data well. To prove this, we conduct the deviance test. Our test statistic is -2*(131.8037-154.3244)=50.787. 131.8037 is the log-likelihood for the null model, and 154.3244 is the log-likelihood for the fitted model. With 11 degrees of freedom, our p-value is .000004772, which is smaller than significance level=0.01. Therefore, the model is a good fit.

The significant predictors are Car, Motorcycle, No Games/Hobbies, and Monthly Subscription.

IV. Interpretation of Significant Box-Cox Coefficients

Monthly Subscription = 0.0400
For students with a monthly subscription, the estimated transformed mean monthly expense is 0.0400 units greater than that for students without a monthly subscription.

No Games/Hobbies = 0.0369
For students without games or hobbies, the estimated transformed mean monthly expense is 0.0369 units greater than that for students who have games/hobbies.

Transportation: Car = 0.0768
For students who transport by car, the estimated transformed mean monthly expense is 0.0768 units greater than that for students who don't commute.

Transportation: Motorcycle = 0.0502
For students who transport by motorcycle, the estimated transformed mean monthly expense is 0.0502 units greater than that for students who don't commute

V. Gamma Regression

Expected Average Monthly Expense=
exp(4.6081 + 0.0254(Male) + 0.0114(Age) + 0.0852(Gender=Male) +0.0393 (Study year)+ 0.0094(Scholarship) + 0.0219(Part Time job) + 0.2924(Transporting=Car) + 0.1811(Motorcycle)+ 0.0206(Does not Drink) +

0.1625(No Games/Hobbies) + 0.1484(monthly subscription = yes))

Alpha Hat = 1/22.8414 = 0.04378

The model fits the data well. To prove this, we apply the deviance test. Our test statistic is -2*(-570.6335-(-545.2400))=50.787. -570.6335 is the log-likelihood for the null model, and -545.2400 is the log-likelihood for the fitted model. With 11 degrees of freedom, our p-value is .000000452, which is smaller than 0.01 significance level. Therefore, the model is a good fit.

The significant predictors are car, motorcycle, no games/hobbies, and monthly subscription.

VI. Interpretation of Selected Gamma Coefficients

Monthly Subscription=0.1484
For students with a monthly subscription, the estimated mean monthly expense is exp(0.1484)*100%=115.998% of that for students without a monthly subscription.

No Games/Hobbies =0.1625
For students without games or hobbies, the estimated mean monthly expense is exp(0.1625)*100%=117.645% of that for students with games/hobbies.

Transportation Motorcycle = 0.1811
For students who transport by motorcycle, the estimated mean monthly expense is exp(0.1811)*100% = 119.854% of that for students who don't commute.

Transportation Car = 0.2924
For students who transport themselves by car, the estimated mean monthly expense is exp(0.2924)*100%=133.964% of that for students who don't commute.

VII. Predicted Probability

We predict the average monthly expense for a 22 year old 4th year student who lives at home, has a scholarship, does not work, gets to school by car, does not drink, has no hobbies, and has a monthly subscription.

Using the Box-Cox Model, we find that the estimated average monthly expense is $317.03, after converting the transformed response back to the actual response. Using the Gamma Model, we find that the estimated average monthly expense is $324.44.

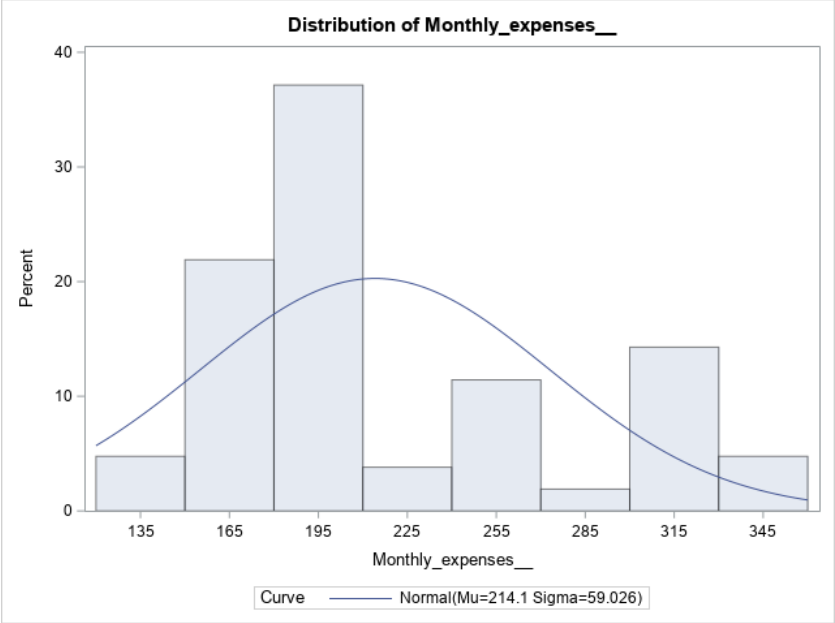
VIII. Conclusion

For both models, the significant predictors are car, motorcycle, no games or hobbies, and monthly subscription. People who commute and have high monthly expenses have greater monthly expenses. Surprisingly, people without games or hobbies have greater monthly expenses than people with games or hobbies. This contrasts popular belief as games/hobbies are thought to be more expensive. To reduce one's monthly expenses, one should reduce their commute, have games and hobbies, and cut their monthly subscriptions.

Appendix

SAS Code

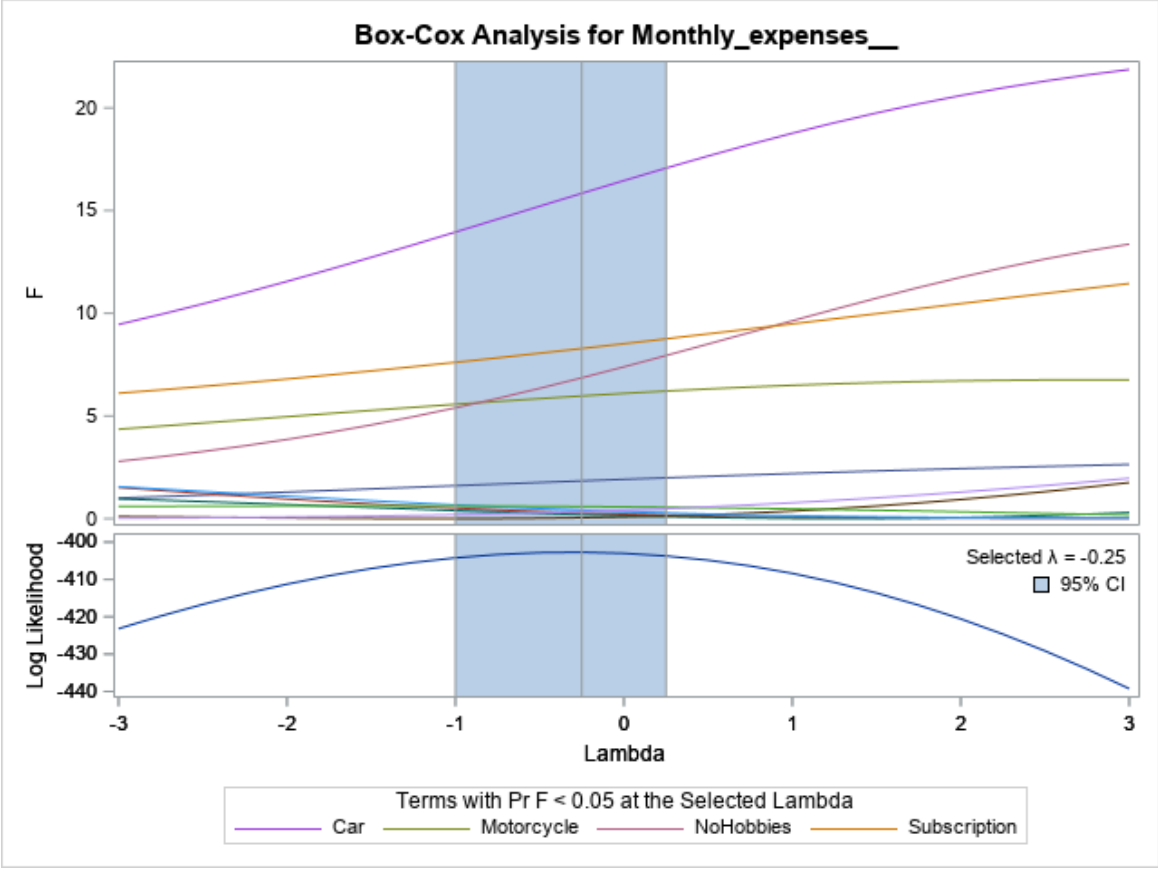
```
proc import
datafile = "C:\Users\colle\Downloads\University Students Monthly Expenses
Cleaned.csv"
out = expenses
dbms = csv replace
;

proc univariate;
var Monthly_expenses__;
histogram /normal;
run;
```



Box-Cox Transformation

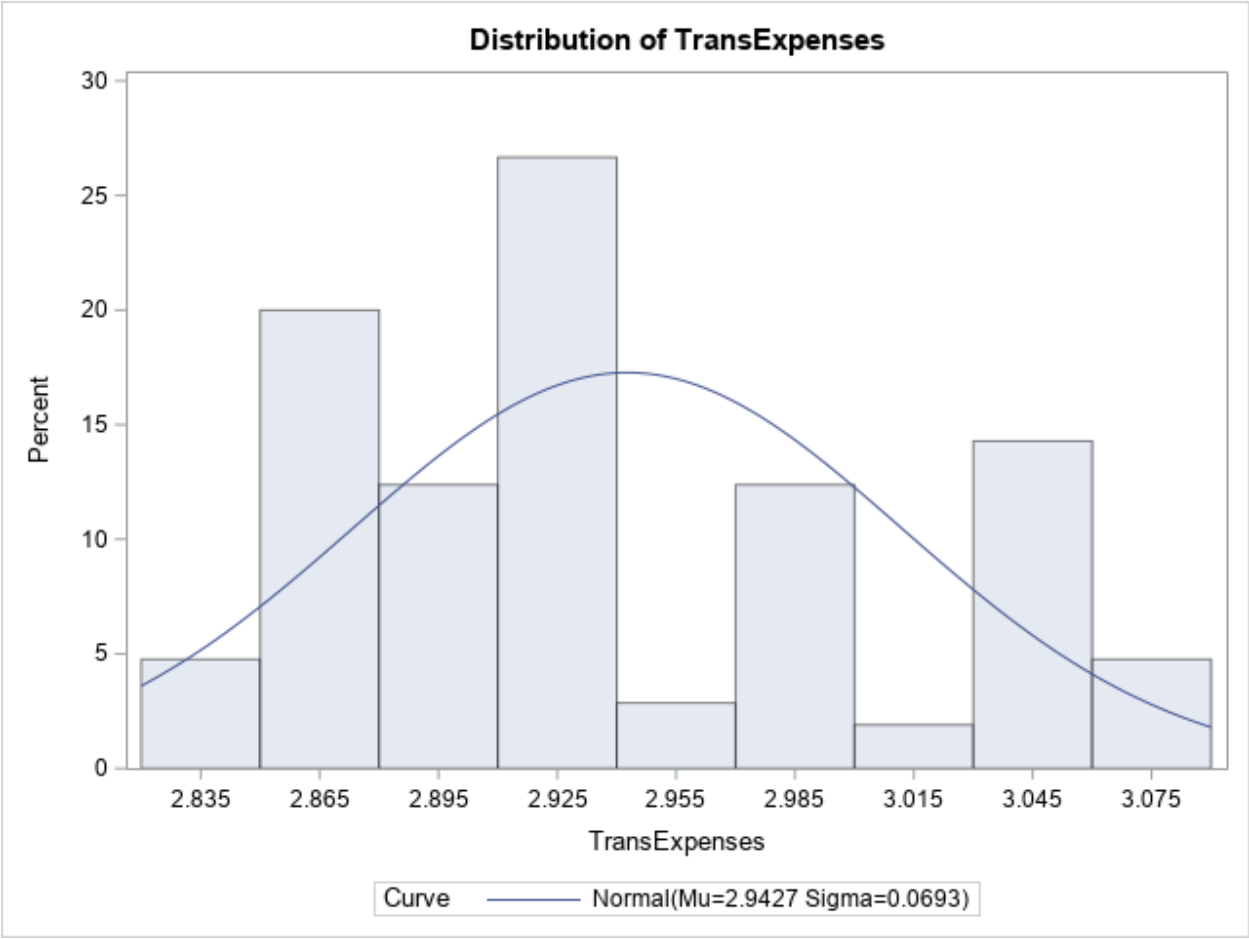
```
data expenses;
set expenses;
Home=(Living = 'Home');
Male=(Gender = 'Male');
IsScholarship=(Scholarship = 'Yes');
Unemployed=(Part_time_job = 'No');
Car=(Transporting='Car');
```

```
Motorcycle=(Transporting='Motorcycle');
NoAlcohol=(Drinks='No');
NoHobbies = (Games___Hobbies = 'No');
Subscription = (Monthly_Subscription = 'Yes')
;

proc transreg data=expenses;
model BoxCox(Monthly_expenses__)= identity(Home Male IsScholarship
Unemployed Car
Motorcycle NoAlcohol NoHobbies Subscription Study_year Age);
run;
```




```
data expenses;
set expenses;
TransExpenses = 4*(1-Monthly_expenses__**(-0.25));
run;

proc univariate;
var TransExpenses;
```

```
histogram /normal;
run;
```




```
proc genmod;
class Gender (ref="Female") Living Scholarship (ref="No")Part_time_job
Transporting Drinks Monthly_Subscription (ref="No") Games___Hobbies;
model TransExpenses = Gender Age Living Study_year
Scholarship Part_time_job Transporting Drinks Games___Hobbies
Monthly_Subscription
/ dist=normal link=identity;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	2.7347	0.1079	2.5233	2.9462	642.53	<.0001
Gender	Male	1	0.0073	0.0130	-0.0181	0.0327	0.31	0.5748
Gender	Female	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	0.0036	0.0058	-0.0077	0.0149	0.39	0.5328
Living	Home	1	0.0192	0.0133	-0.0069	0.0453	2.08	0.1497
Living	Hostel	0	0.0000	0.0000	0.0000	0.0000	.	.
Study_year		1	0.0081	0.0099	-0.0114	0.0275	0.66	0.4150
Scholarship	Yes	1	0.0071	0.0150	-0.0223	0.0365	0.22	0.6362
Scholarship	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Part_time_job	No	1	0.0034	0.0152	-0.0263	0.0332	0.05	0.8215
Part_time_job	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Transporting	Car	1	0.0768	0.0182	0.0412	0.1125	17.88	<.0001
Transporting	Motorcycle	1	0.0502	0.0193	0.0123	0.0881	6.75	0.0094
Transporting	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Drinks	No	1	0.0151	0.0222	-0.0284	0.0586	0.46	0.4964
Drinks	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Games___Hobbies	No	1	0.0369	0.0133	0.0109	0.0628	7.74	0.0054
Games___Hobbies	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Monthly_Subscription	Yes	1	0.0400	0.0131	0.0144	0.0656	9.35	0.0022
Monthly_Subscription	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.0556	0.0038	0.0486	0.0637		

```
proc genmod;
model TransExpenses = /dist = normal link=identity;
Run;
```

Log Likelihood =131.8037

```
data deviance_test;
deviance = -2*(131.8037-154.3244);
pvalue = 1-probchi(deviance,11);
run;
proc print data = deviance_test;
run;
```

Obs	deviance	pvalue
1	45.0414	.000004772

```
data prediction;
input Gender $ Living $ Scholarship $ Part_time_job $ Transporting $ Drinks$
Monthly_Subscription $ Games___Hobbies $
Age Study_year Smoking $ Cosmetics___Self_care $;
cards;
Male Home Yes No Car No Yes No 22 4 No No
;

data expenses;
set expenses prediction;
run;

proc genmod;
class Gender (ref="Female") Living Scholarship (ref="No")Part_time_job
Transporting Drinks Monthly_Subscription (ref="No") Games___Hobbies;
model TransExpenses = Gender Age Living Study_year
Scholarship Part_time_job Transporting Drinks Games___Hobbies
Monthly_Subscription
/ dist=normal link=identity;
output out=outdata p=pexpenses;
run;

data outdata;
set outdata;
pred_expenses=(1-0.25*pexpenses)**(-4);
run;

proc print data=outdata (firstobs=106 obs=106);
var pred_expenses;
run;
```

Obs	pred_expenses
106	317.033

Gamma Regression

```
proc genmod;
class Gender (ref="Female") Living Scholarship (ref="No")Part_time_job
Transporting Drinks Monthly_Subscription (ref="No") Games___Hobbies;
model Monthly_expenses__ = Gender Age Living Study_year
Scholarship Part_time_job Transporting Drinks Games___Hobbies
Monthly_Subscription
/ dist=gamma link=log;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	4.6081	0.4003	3.8236	5.3927	132.52	<.0001
Gender	Male	1	0.0254	0.0503	-0.0733	0.1240	0.25	0.6140
Gender	Female	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	0.0114	0.0220	-0.0316	0.0544	0.27	0.6032
Living	Home	1	0.0852	0.0506	-0.0141	0.1844	2.83	0.0925
Living	Hostel	0	0.0000	0.0000	0.0000	0.0000	.	.
Study_year		1	0.0393	0.0372	-0.0337	0.1122	1.11	0.2912
Scholarship	Yes	1	0.0094	0.0566	-0.1016	0.1204	0.03	0.8679
Scholarship	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Part_time_job	No	1	0.0219	0.0566	-0.0891	0.1328	0.15	0.6998
Part_time_job	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Transporting	Car	1	0.2924	0.0661	0.1628	0.4220	19.56	<.0001
Transporting	Motorcycle	1	0.1811	0.0718	0.0405	0.3218	6.37	0.0116
Transporting	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Drinks	No	1	0.0206	0.0838	-0.1436	0.1848	0.06	0.8057
Drinks	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Games___Hobbies	No	1	0.1625	0.0506	0.0633	0.2617	10.30	0.0013
Games___Hobbies	Yes	0	0.0000	0.0000	0.0000	0.0000	.	.
Monthly_Subscription	Yes	1	0.1484	0.0492	0.0520	0.2448	9.11	0.0025
Monthly_Subscription	No	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	22.8414	3.1297	17.4620	29.8780		

```
proc genmod;
model Monthly_expenses__ =/dist = gamma link=log;
run;

Log Likelihood= -570.6335

data deviance_test;
deviance = -2*(-570.6335-(-545.2400));
pvalue = 1-probchi(deviance,11);
run;
proc print data = deviance_test;
run;
```

Obs	deviance	pvalue
1	50.787	.000000452

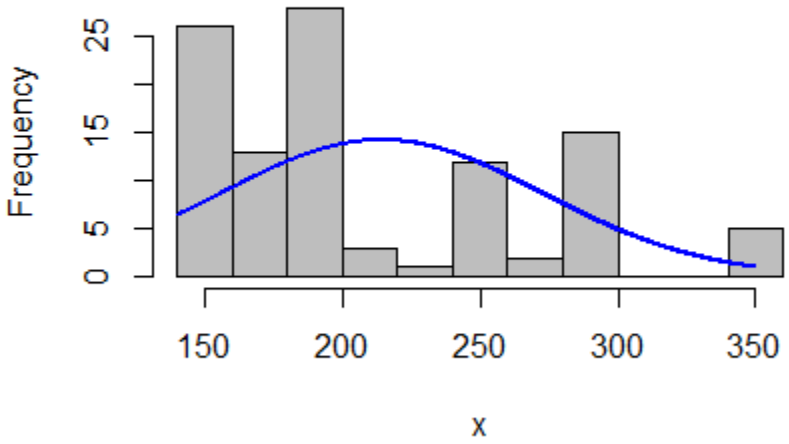
```
data prediction;
input Gender $ Living $ Scholarship $ Part_time_job $ Transporting $ Drinks$
Monthly_Subscription $ Games___Hobbies $
Age Study_year Smoking $ Cosmetics___Self_care $;
cards;
Male Home Yes No Car No Yes No 22 4 No No
;

data expenses;
set expenses prediction;
run;

proc genmod;
class Gender (ref="Female") Living Scholarship (ref="No")Part_time_job
Transporting Drinks Monthly_Subscription (ref="No") Games___Hobbies;
model Monthly_expenses__ = Gender Age Living Study_year
Scholarship Part_time_job Transporting Drinks Games___Hobbies
Monthly_Subscription
/ dist=gamma link=log;
output out=outdata p=pexpenses;
run;
```

```
data outdata;
set outdata;
run;

proc print data=outdata (firstobs=106 obs=106);
var pexpenses;
Run;
```

Obs	pexpenses
106	324.439

R Code and Output

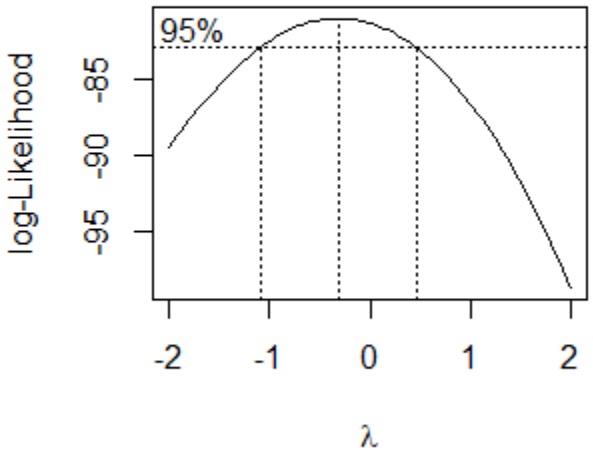
```
library(readr)
ProjectData <- read_csv("C://Users//colle//Downloads//University Students
Monthly Expenses Cleaned.csv")
shapiro.test(ProjectData$`Monthly_expenses_$`)
Shapiro-Wilk normality test

data:  ProjectData$`Monthly_expenses_$`
W = 0.89096, p-value = 3.199e-07

library(rcompanion)
plotNormalHistogram(ProjectData$`Monthly_expenses_$`)
```



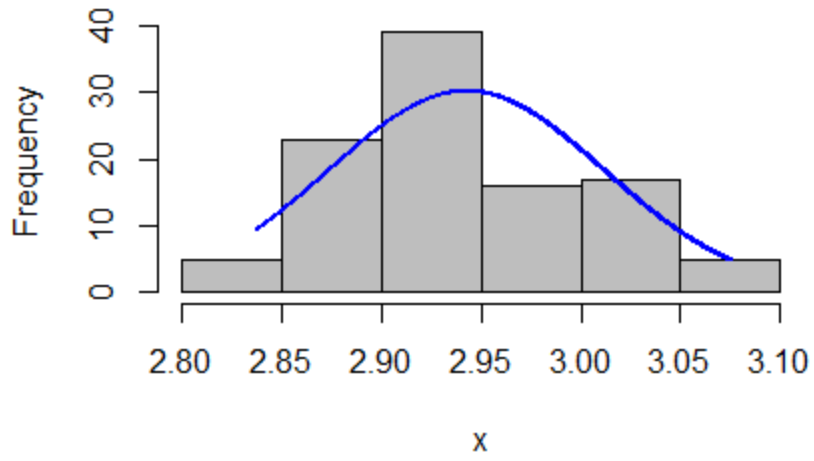
Box-Cox Transformation

```
GenderRef <- relevel(factor(ProjectData$Gender),ref="Female")
LivingRef <- relevel(factor(ProjectData$Living),ref="Hostel")
ScholarshipRef <- relevel(factor(ProjectData$Scholarship),ref="No")
EmploymentRef <- relevel(factor(ProjectData$Part_time_job),ref="Yes")
TransportationRef <- relevel(factor(ProjectData$Transporting),ref="No")
DrinksRef <- relevel(factor(ProjectData$Drinks),ref="Yes")
GamesRef <- relevel(factor(ProjectData$`Games_&_Hobbies`),ref="Yes")
SubscriptionRef <- relevel(factor(ProjectData$Monthly_Subscription),ref="No")
Age <- ProjectData$Age
```

```
Study_year <- ProjectData$Study_year

library(MASS)
BoxCox.fit <- boxcox(`Monthly_expenses_$` ~ GenderRef + LivingRef +
ScholarshipRef
+ EmploymentRef + TransportationRef + DrinksRef + GamesRef +
SubscriptionRef + Age + Study_year
, data=ProjectData)
BoxCox.data<- data.frame(BoxCox.fit$x, BoxCox.fit$y)
ordered.data<- BoxCox.data[with(BoxCox.data, order(-BoxCox.fit.y)),]
Ordered.data[1,]
```




```
 BoxCox.fit.x BoxCox.fit.y
43   -0.3030303    -80.94403

tr_expenses <- 4-4/((ProjectData$`Monthly_expenses_$`)^(1/4))
plotNormalHistogram(tr_expenses)
```



summary(fitted.model<- glm(tr_expenses ~ GenderRef + LivingRef + ScholarshipRef
+ EmploymentRef + TransportationRef + DrinksRef + GamesRef + SubscriptionRef + Age + Study_year
, data=ProjectData, family=gaussian(link=identity)))

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.734740   0.114636  23.856  < 2e-16
GenderRefMale               0.007269   0.013766   0.528 0.598732
LivingRefHome               0.019209   0.014169   1.356 0.178483
ScholarshipRefYes           0.007102   0.015953   0.445 0.657195
EmploymentRefNo             0.003424   0.016129   0.212 0.832331
TransportationRefCar        0.076839   0.019307   3.980 0.000137
TransportationRefMotorcycle 0.050223   0.020545   2.445 0.016389
DrinksRefNo                 0.015107   0.023600   0.640 0.523644
GamesRefNo                  0.036873   0.014082   2.618 0.010314
SubscriptionRefYes          0.039978   0.013890   2.878 0.004962
Age                         0.003599   0.006129   0.587 0.558454
Study_year                  0.008081   0.010535   0.767 0.444977
```

null.Model <-glm(tr_expenses2 ~ 1, family=gaussian(link=identity))
print(deviance <- -2*(logLik(null.Model)-logLik(fitted.model)))
'log Lik.' 45.04137 (df=2)
print(p.value <- pchisq(deviance, df=11,lower.tail=FALSE))
‘log Lik.' 4.771567e-06 (df=2)

```
TransPredExpense <- predict(fitted.model,type="response",
data.frame(GenderRef="Male",LivingRef="Home",ScholarshipRef="Yes",
EmploymentRef="No",TransportationRef="Car",DrinksRef="No",GamesRef="No
",SubscriptionRef="Yes",
Age=22,Study_year=4))

print(Expense <- (-0.25*TransPredExpense+1)^(-4))

  1
317.0332
```

Gamma Regression

```
GenderRef <- relevel(factor(ProjectData$Gender),ref="Female")
LivingRef <- relevel(factor(ProjectData$Living),ref="Hostel")
ScholarshipRef <- relevel(factor(ProjectData$Scholarship),ref="No")
EmploymentRef <- relevel(factor(ProjectData$Part_time_job),ref="Yes")
TransportationRef <- relevel(factor(ProjectData$Transporting),ref="No")
DrinksRef <- relevel(factor(ProjectData$Drinks),ref="Yes")
GamesRef <- relevel(factor(ProjectData$`Games_&_Hobbies`),ref="Yes")
SubscriptionRef <- relevel(factor(ProjectData$Monthly_Subscription),ref="No")
Age <- ProjectData$Age
Study_year <- ProjectData$Study_year

summary(GammaModel <- glm(ProjectData$`Monthly_expenses_$` ~ GenderRef
+ LivingRef + ScholarshipRef
+ EmploymentRef + TransportationRef + DrinksRef + GamesRef +
SubscriptionRef + Age +
Study_year, data=ProjectData, family=Gamma(link=log)))
```

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.608144   0.427975  10.767  < 2e-16 ***
GenderRefMale              0.025389   0.051393   0.494 0.622462
LivingRefHome              0.085158   0.052898   1.610 0.110820
ScholarshipRefYes          0.009424   0.059557   0.158 0.874617
EmploymentRefNo            0.021864   0.060215   0.363 0.717350
TransportationRefCar       0.292416   0.072081   4.057 0.000103 ***
TransportationRefMotorcycle 0.181144   0.076702   2.362 0.020279 *
DrinksRefNo                0.020606   0.088105   0.234 0.815589
GamesRefNo                 0.162490   0.052574   3.091 0.002634 **
SubscriptionRefYes         0.148393   0.051856   2.862 0.005206 **
Age                        0.011414   0.022882   0.499 0.619104
Study_year                 0.039279   0.039331   0.999 0.320535
```

> nullModel <- glm(ProjectData$`Monthly_expenses_$` ~ 1,
family=Gamma(link=log))
> print(deviance <- -2*(logLik(nullModel)-logLik(GammaModel)))
'log Lik.' 50.79139 (df=2)
> print(p.value <- pchisq(deviance,df=11,lower.tail=FALSE))
'log Lik.' 4.508581e-07 (df=2)

print(predict(GammaModel,type="response",
data.frame(GenderRef="Male",LivingRef="Home",ScholarshipRef="Yes",
EmploymentRef="No",TransportationRef="Car",DrinksRef="No",GamesRef="No
",SubscriptionRef="Yes",
Age=22,Study_year=4)))

1
324.4394

Works Cited

[1] “University Students Monthly Expenses”, Kaggle.com, https://www.kaggle.com/datasets/shariful07/nice-work-thanks-for-share, 2022