HEALTH INSURANCE PREMIUM CHARGES IN THE US

Box-Cox and Gamma Regression Analysis

STAT410 PROJECT

Submitted to Dr. Olga Korosteleva

Report Prepared by Nate Talampas

November 27, 2023

Contents

I. Introduction 1
II. Background 1
III. Data Description 2
IV. Results 2
V. Conclusion 3
VI. Appendix 4
A . SAS Code 1 4
B. R Code 1 6
C. SAS Code 2 7
D. R Code 2 7
E. SAS Code 3 8
F. R Code 3 9
G. SAS Code 4 11
H. SAS Code 4 Output 12
I. R Code 4 12
J. R Code 4 Output 13
K. Fitted Model Prediction Code and Output 13
VII. References 14

I. Introduction

I have developed a fascination towards the dynamic interplay between healthcare, insurance, and technology. My project is centered around understanding how Healthcare IT systems are utilized to streamline the claims processing workflow. The dataset I have employed comprises of 1338 rows of insured data, providing me with the opportunity to integrate both a Box-Cox and Gamma regression analysis to predict Insurance Premium Charges.

II. Background

An insurance premium is the amount of money an individual or business pays for an insurance policy. Insurance premiums are paid for policies that cover healthcare, auto, home, or life insurance [1]. When an individual signs up for an insurance policy, the insurer will charge a premium. The price of the premium depends on a variety of factors including: the type of coverage, age, area in which you live, any claims filed in the past, and moral hazard and adverse selection. With the emergence of artificial intelligence and sophisticated algorithms, the way insurance is priced and sold is being catered to these technological innovations. Insurers use the premiums paid to them by their customers and policyholders to cover liabilities associated with the policies they underwrite. Some insurers invest in the premium to generate higher returns, which helps maintain competitive prices within the market. In addition to an insurance plan's premium, total health care costs also include the deductible, copayment/coinsurance amounts, along with health and drug services [2]. A deductible is how much the individual spends for covered health services prior to the insurance company paying anything. Copayments and coinsurance are payments you make to your health care provider each time an individual gets care. Analytics is a vital component of the insurance industry because it allows for interpretation that drives business [3]. These methods of analytics often emphasize data-mining tools and statistical inference.

III. Data Description

The dataset was found through the website Kaggle, an online community platform for data scientists, where users can find and publish datasets. The original dataset included 1338 rows of data, with 7 variables including: age, sex (male/female), BMI, number of children, smoker (yes/no), region (southeast, southwest, northeast, northwest), and insurance charges. There are no missing or undefined values in the dataset. For this regression, I truncated the dataset to the first 100 observations. I set the insurance charges as the response variable and the rest of the variables as predictors. I then ran both a Box-Cox and a Gamma Regression to model the Insurance Premium Charges.

IV. Results

I was able to conclude that after I transformed the response variable for the Box-Cox Regression, it was not a successful fit. However, the Gamma Regression was a better fit to the data. After plotting the histogram, I could observe that the distribution of the charges was right skewed. I then calculated the optimal lambda for a Box-Cox transformation, which was equal to

0.25. This called for a square root transformation. After plotting a histogram to the transformed response and conducting a normality test, I noted that the p-value from its normality test was still large, and the transformed response was not normally distributed.

Alternatively, I fit a Gamma Regression model to the data. Significant predictors at the 5% level included age, BMI, indication of smoking, and region northwest. I noted that the log likelihood for the fitted model was -275.9746 and the log likelihood for the null model was -366.4021. After calculating the deviance and the p-value for the deviance test, an extremely small p-value was displayed, which SAS displayed as "0", indicating a good fit to the data. The fitted model can be written as: $\hat{E}(charges) = exp(-0.2226 + 0.0345 * age + 0.1195 * male + 0.0173*\mathrm{bmi} + 0.0520*\mathrm{children} + 1.6938*\mathrm{smoker} + 0.0093*\mathrm{northeast} + 0.2674*\mathrm{northwest} + 0.0782*\mathrm{southwest})$. The interpretation for the significant predictors is as follows: As age increases by one year, the estimated mean amount of premiums increases by 3.51%. As BMI increases by one point, the estimated mean amount of premiums increases by 1.745%. The estimated mean amount of premiums for smokers is 5.44% of that for nonsmokers. The estimated mean amount of premiums for people living in the northwest is 1.3066% of that for people living in the southeast.

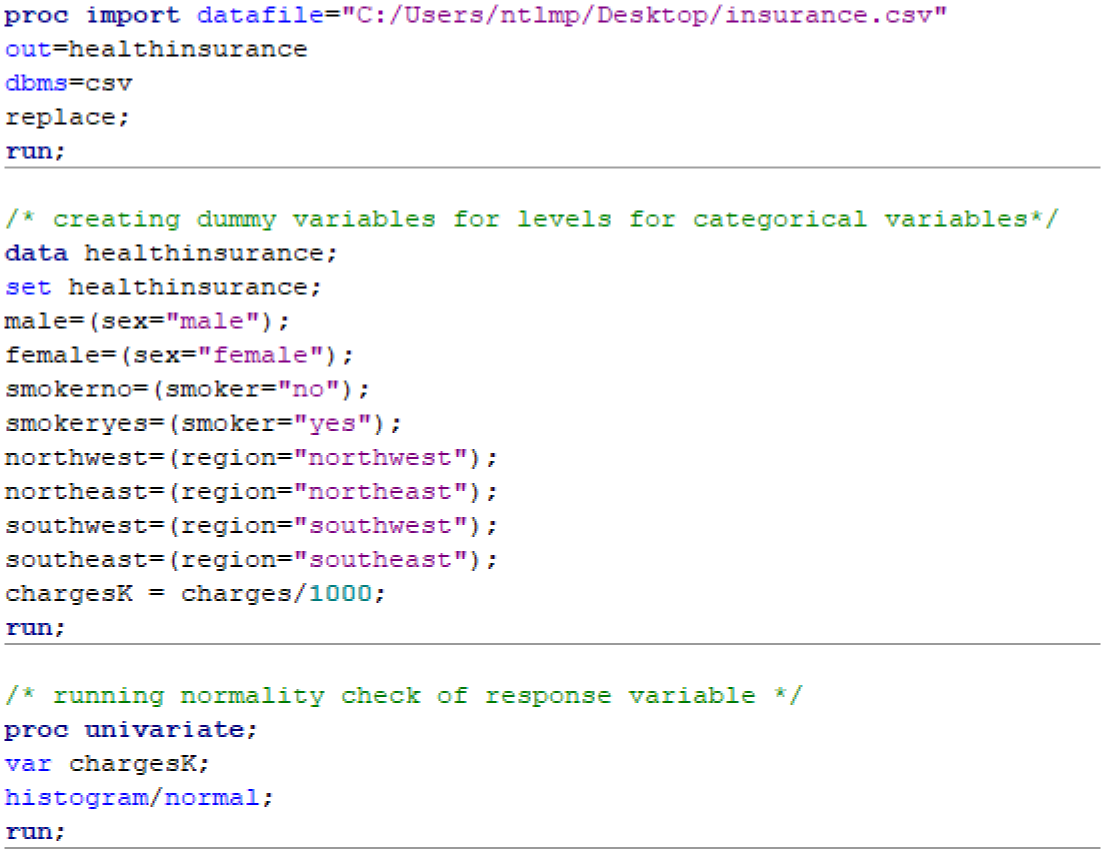
I then used the fitted model to predict the premium charge of a 22-year-old male with a BMI of 23.3, no children, nonsmoking, and living in the southwest. The value can be handwritten as: $\mathrm{exp}(-0.2226 + 0.0345(22) + 0.1195 + 0.0173(23.3) + 0.0782) * 1000 = 3117.994$. This value was very close to the output from SAS and R.

V. Conclusion

The Gamma Regression proved to be an effective and accurate way to model the health insurance premium charges. Given the inherent right-skewness of the insurance premium data, this regression method was well-suited for capturing and analyzing the asymmetry present in the dataset. In the future, I hope to apply my regression analysis skills to other types of financial metrics like claim amounts. This project allowed me to delve into understanding the risk underwriting in Health Insurance, the interplay of various attributes of the insured, and evaluate their effect on the insurance premium. I hope to eventually improve upon my project by applying machine learning and deep learning methods to big data in Healthcare IT, and benefit from the radically changing digital landscape of insurance.

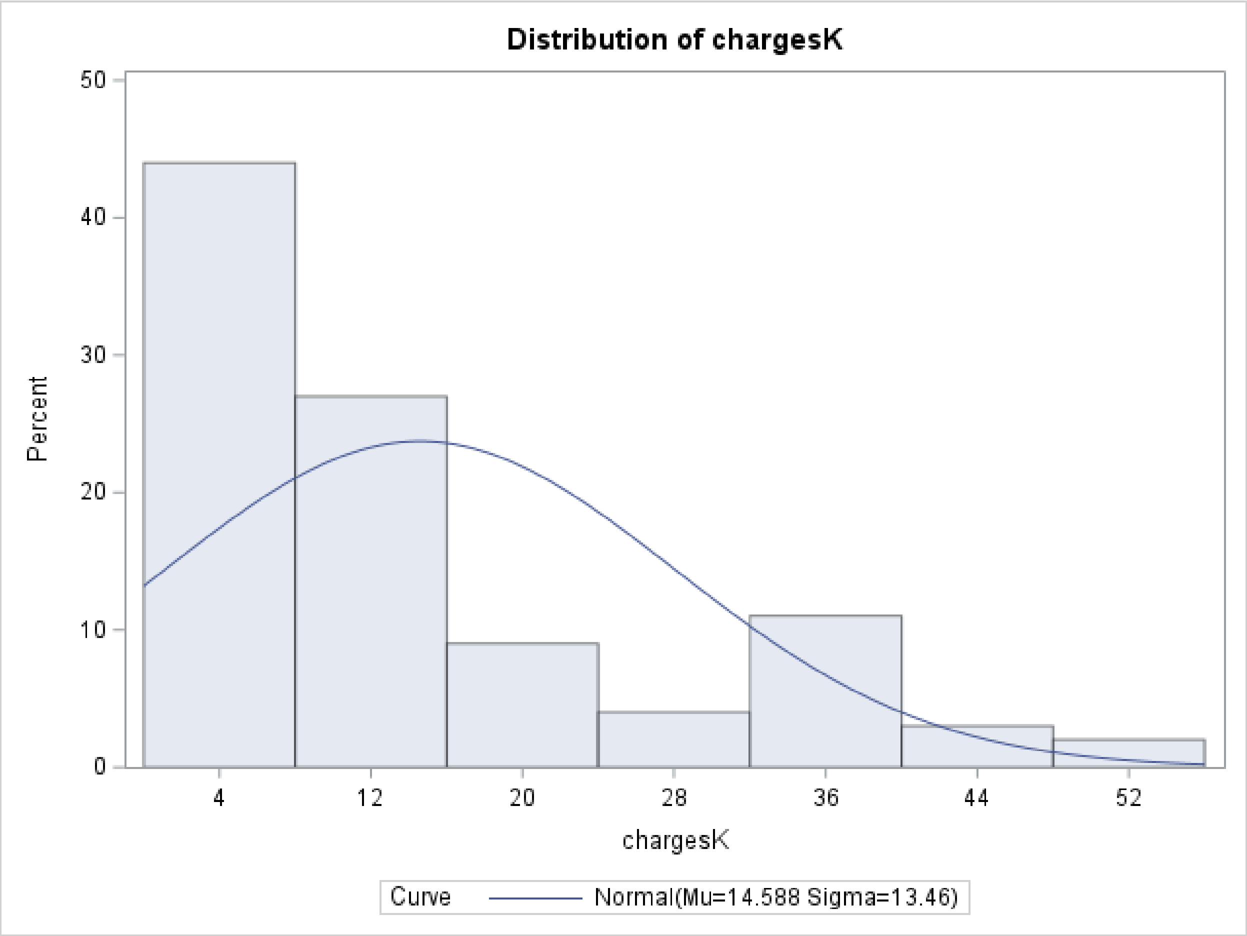
VI. Appendix

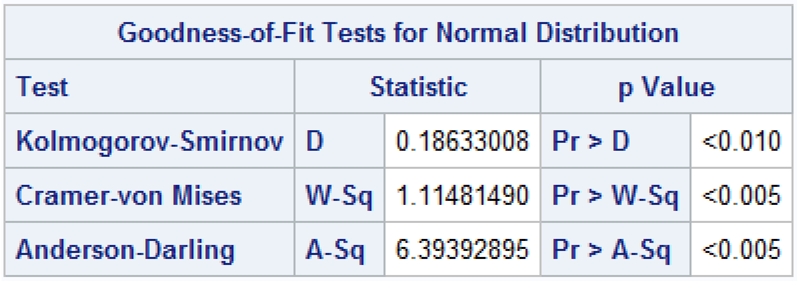
A . SAS Code 1



```
proc import datafile="C:/Users/ntlmp/Desktop/insurance.csv"
out=healthinsurance
dbms=csv
replace;
run;

/* creating dummy variables for levels for categorical variables*/
data healthinsurance;
set healthinsurance;
male=(sex="male");
female=(sex="female");
smokerno=(smoker="no");
smokeryes=(smoker="yes");
northwest=(region="northwest");
northeast=(region="northeast");
southwest=(region="southwest");
southeast=(region="southeast");
chargesK = charges/1000;
run;

/* running normality check of response variable */
proc univariate;
var chargesK;
histogram/normal;
run;
```

Figure 1 Running Normality Check on Response

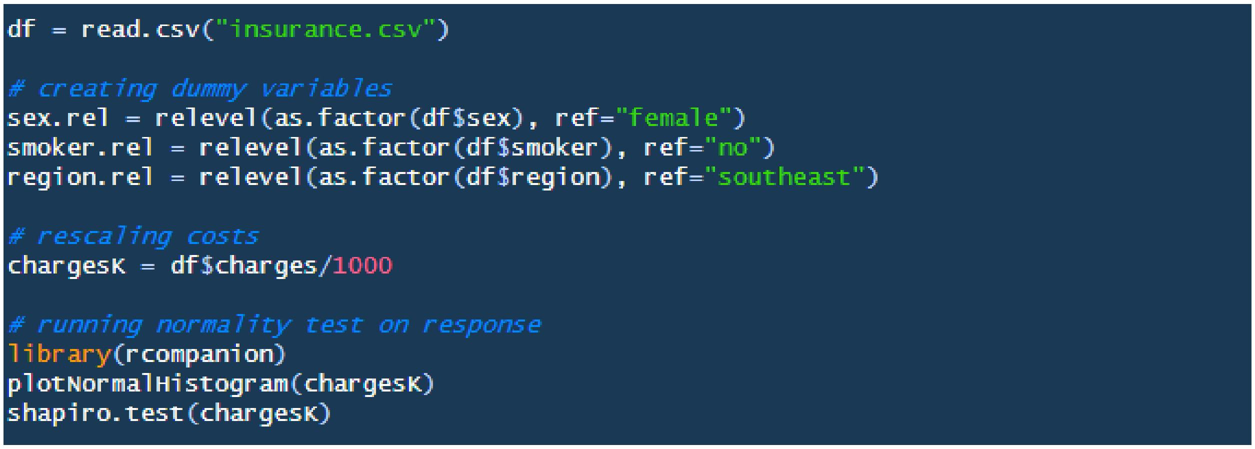




Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.18633008	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.11481490	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	6.39392895	Pr > A-Sq	<0.005

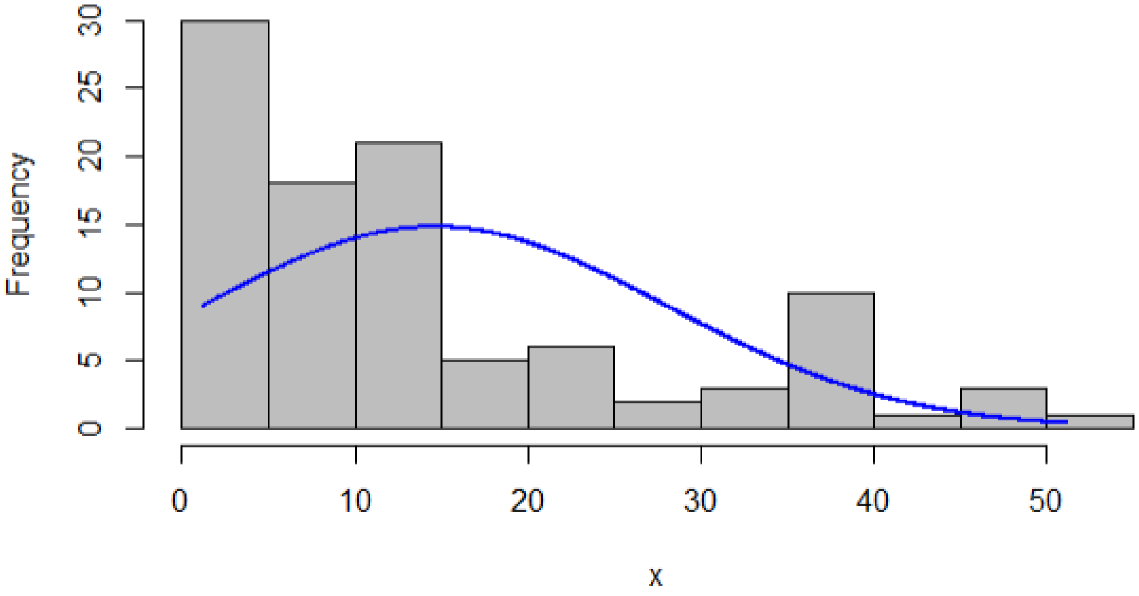
Figure 2 Histogram and Normality Tests

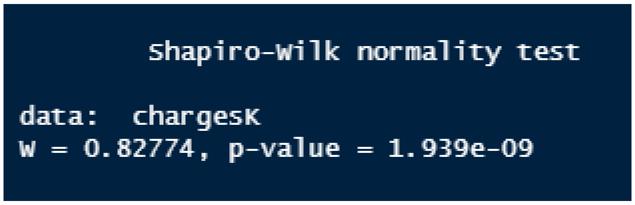
B. R Code 1



```r
df = read.csv("insurance.csv")

# creating dummy variables
sex.rel = relevel(as.factor(df$sex), ref="female")
smoker.rel = relevel(as.factor(df$smoker), ref="no")
region.rel = relevel(as.factor(df$region), ref="southeast")

# rescaling costs
chargesK = df$charges/1000

# running normality test on response
library(rcompanion)
plotNormalHistogram(chargesK)
shapiro.test(chargesK)
```

Figure 3 Running Normality Check on Response





```
        Shapiro-Wilk normality test

data:  chargesK
W = 0.82774, p-value = 1.939e-09
```

Figure 4 Histogram and Normality Test

C. SAS Code 2

```
/* finding optimal lambda for Box-Cox transformation*/
proc transreg;
model BoxCox(chargesK) =
    identity(age male female bmi children smokerno smokeryes northwest northeast southwest southeast);
run;
```

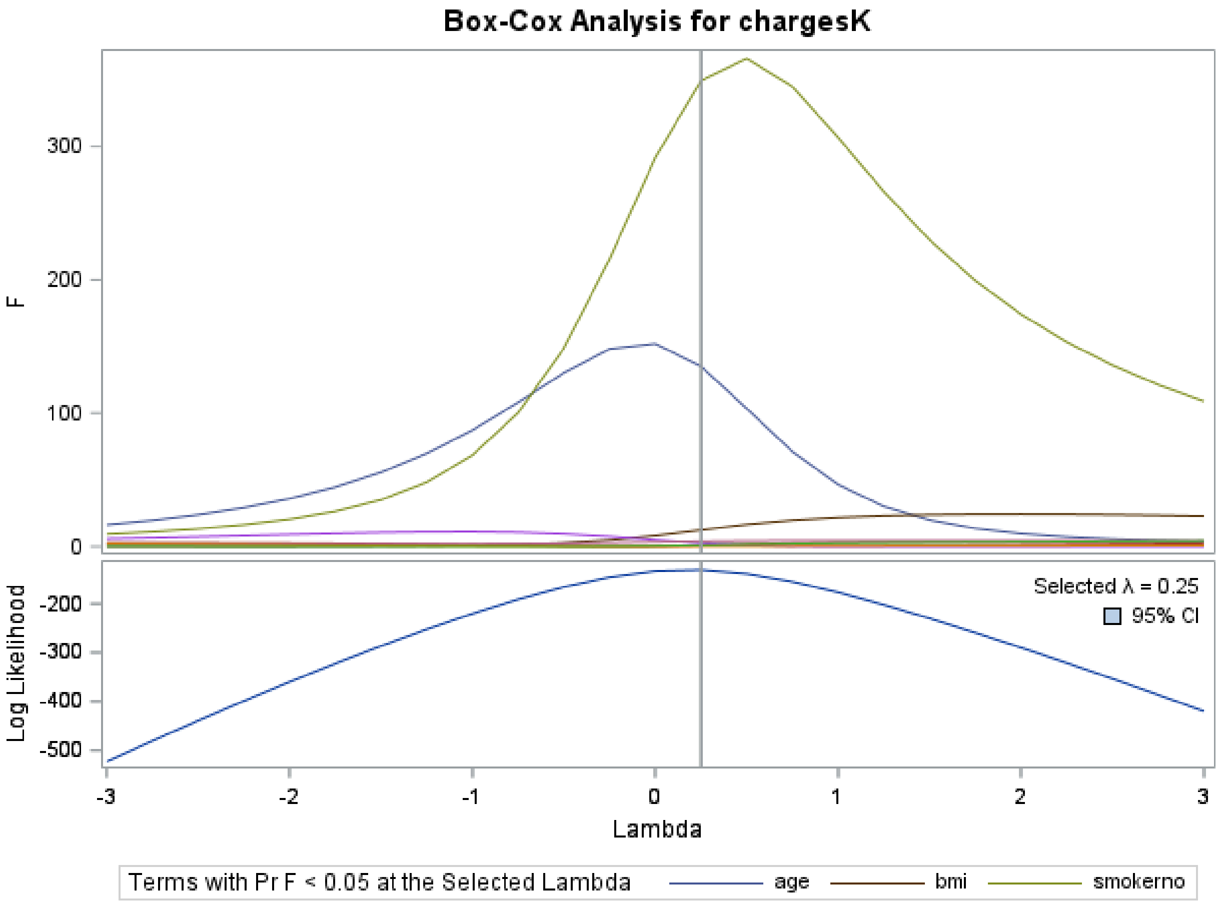
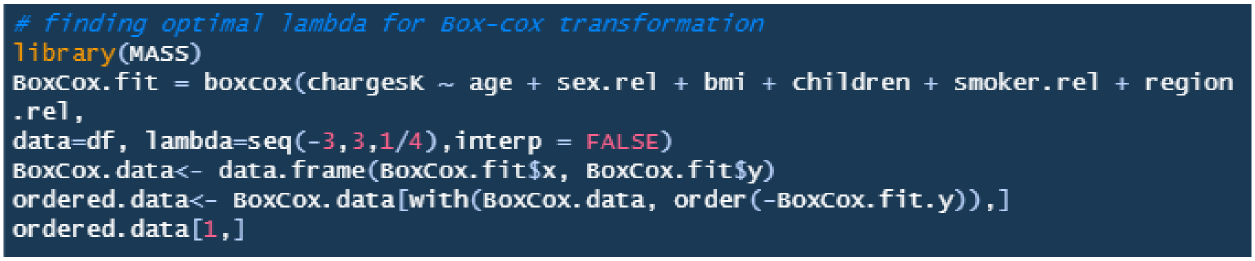
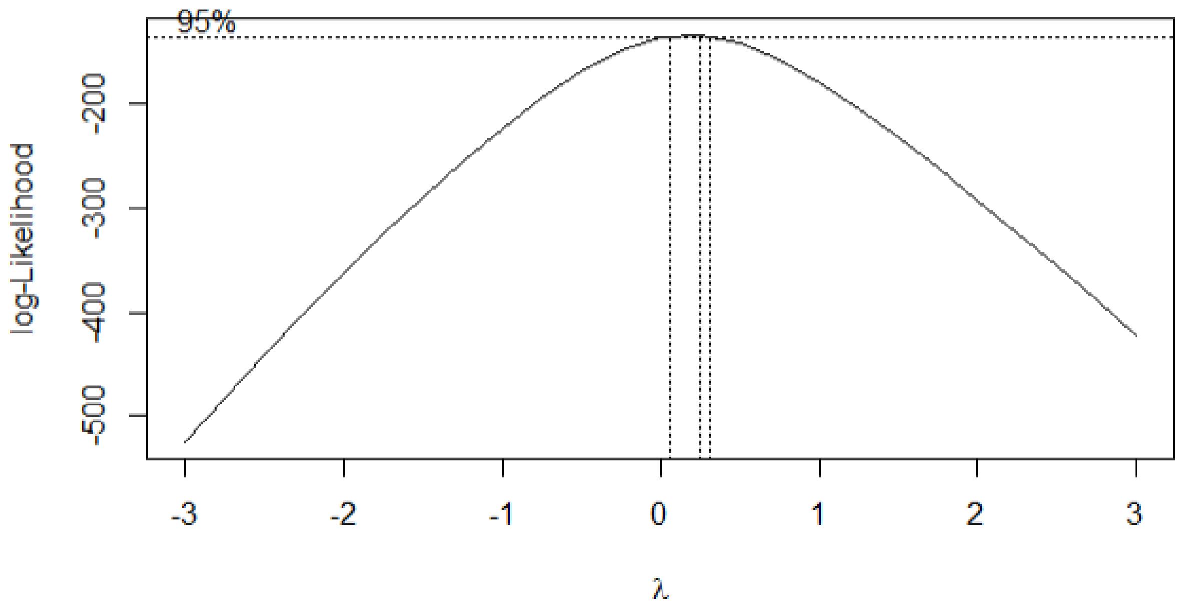


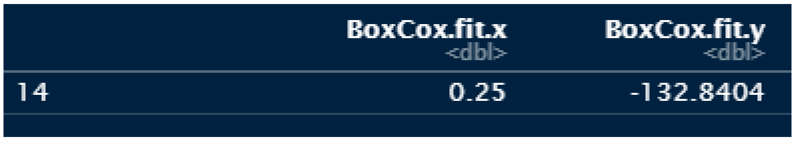
Figure 5 Finding optimal lambda for Box-Cox transformation

D. R Code 2



```
# finding optimal lambda for Box-cox transformation
library(MASS)
BoxCox.fit = boxcox(chargesK ~ age + sex.rel + bmi + children + smoker.rel + region
.rel,
data=df, lambda=seq(-3,3,1/4),interp = FALSE)
BoxCox.data<- data.frame(BoxCox.fit$x, BoxCox.fit$y)
ordered.data<- BoxCox.data[with(BoxCox.data, order(-BoxCox.fit.y)),]
ordered.data[1,]
```





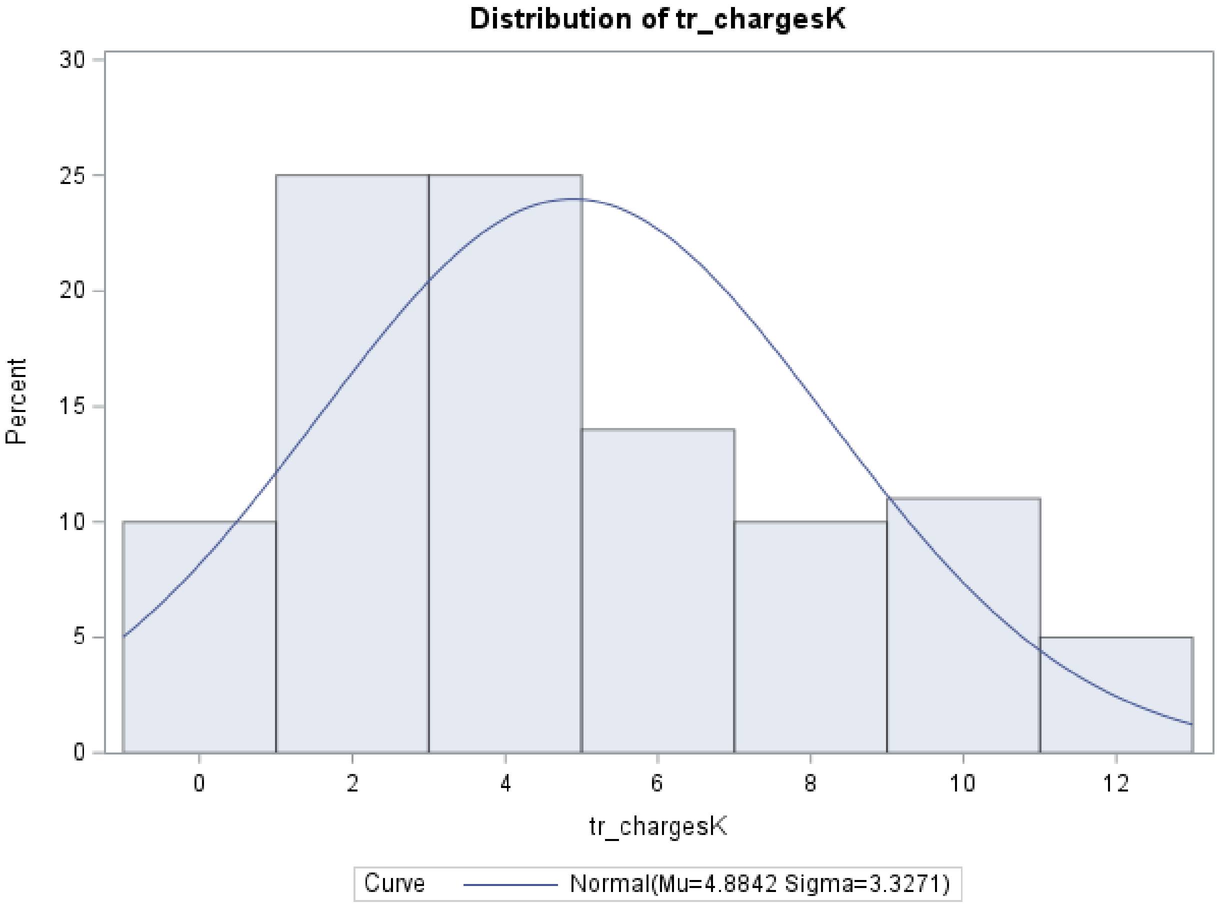
	BoxCox.fit.x <dbl>	BoxCox.fit.y <dbl>
14	0.25	-132.8404

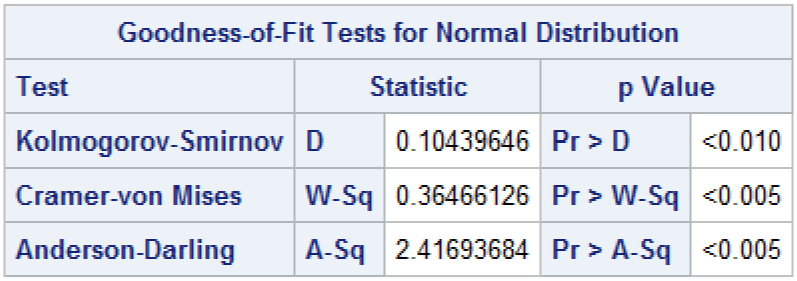
Figure 6 Finding optimal lambda for Box-Cox transformation

E. SAS Code 3

```
/* applying Box-Cox transformation with lambda=0.5*/
/* square root transformation */
data healthinsurance;
set healthinsurance;
tr_chargesK = 2 * (sqrt(chargesK) - 1);
run;
```

```
/* running normality check of transformed response*/
proc univariate;
var tr_chargesK;
histogram/normal;
run;
```

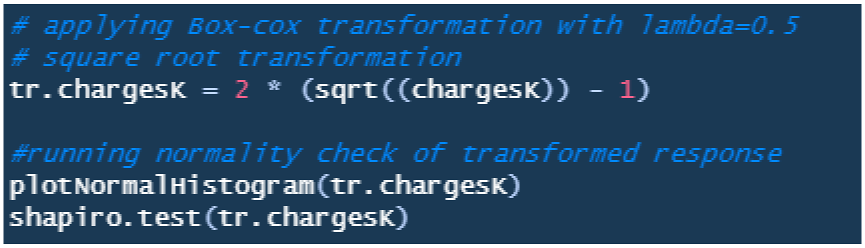




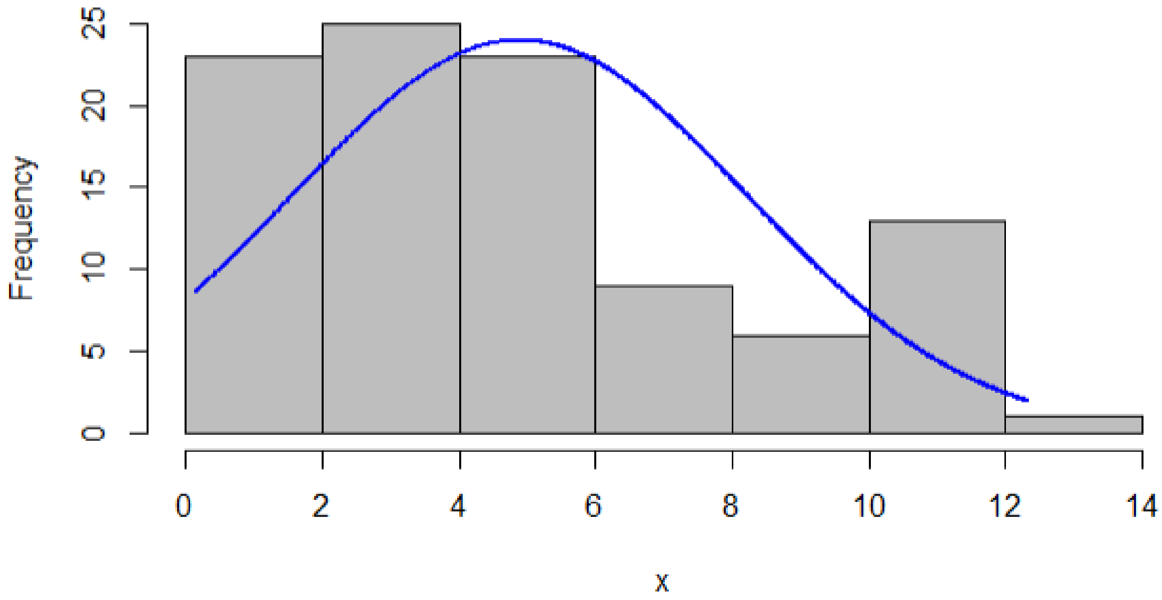
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.10439646	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.36466126	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2.41693684	Pr > A-Sq	<0.005

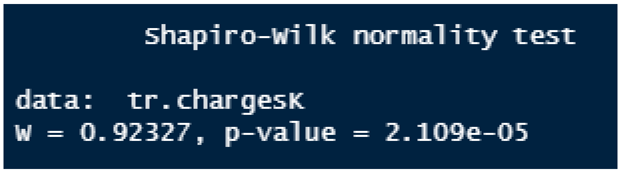
Figure 7 Running Normality Check on Transformed Response

F. R Code 3



```
# applying Box-cox transformation with lambda=0.5
# square root transformation
tr.chargesK = 2 * (sqrt((chargesK)) - 1)

#running normality check of transformed response
plotNormalHistogram(tr.chargesK)
shapiro.test(tr.chargesK)
```

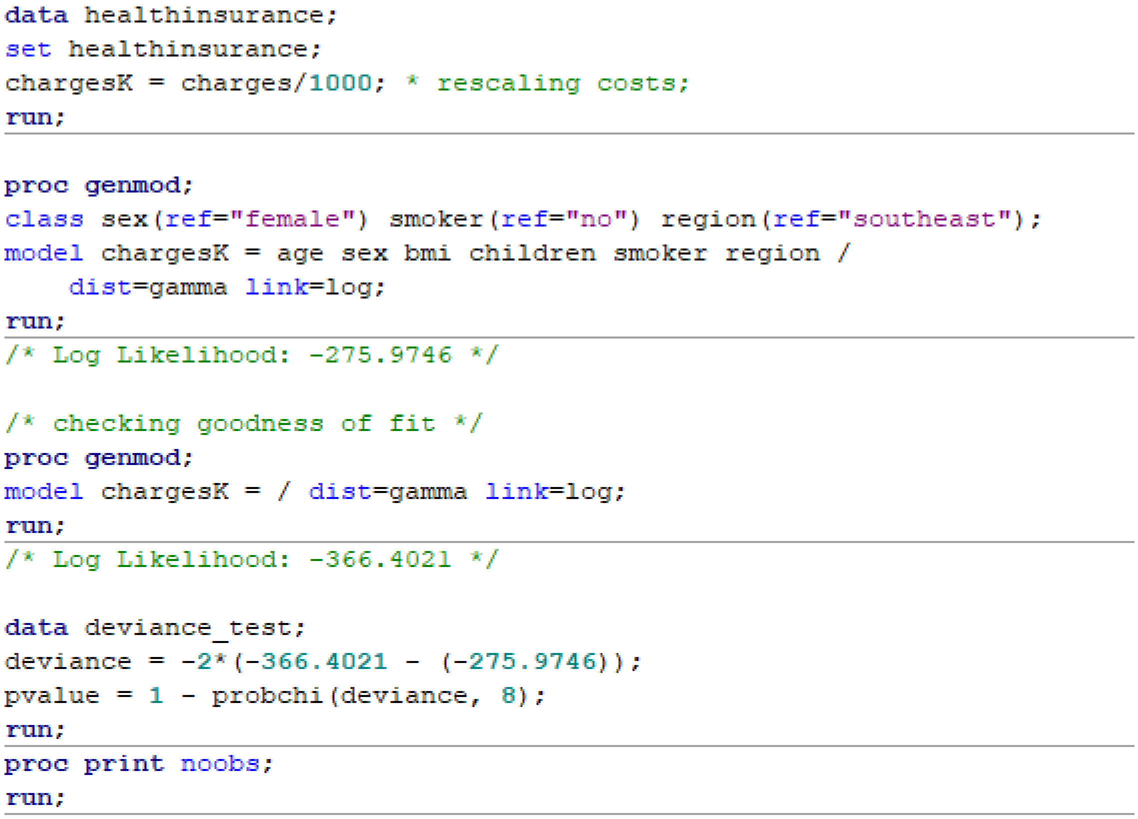




```
        Shapiro-Wilk normality test

data:  tr.chargesK
W = 0.92327, p-value = 2.109e-05
```

Figure 8 Running Normality Check on Transformed Response

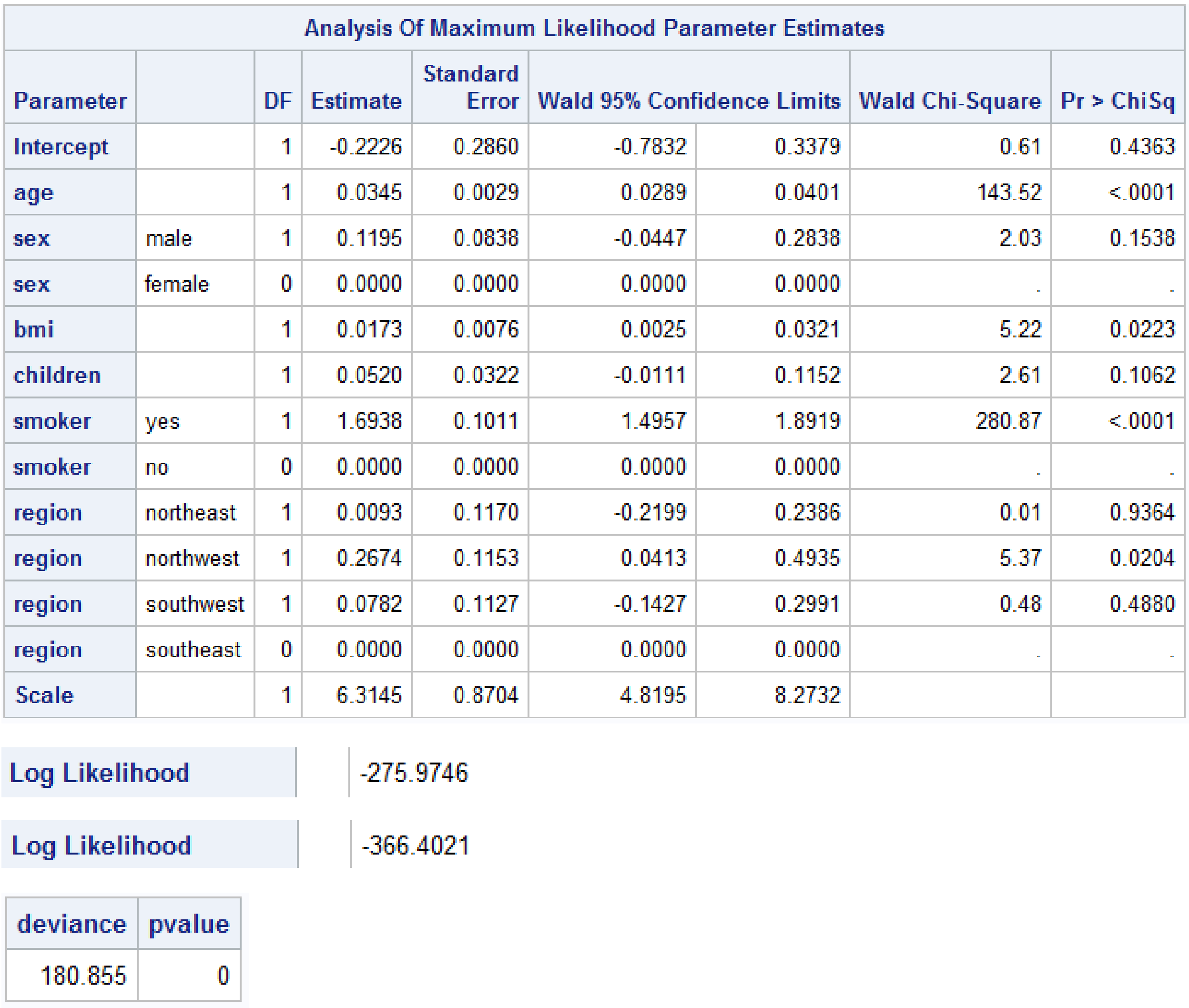
G. SAS Code 4



```
data healthinsurance;
set healthinsurance;
chargesK = charges/1000; * rescaling costs;
run;

proc genmod;
class sex(ref="female") smoker(ref="no") region(ref="southeast");
model chargesK = age sex bmi children smoker region /
    dist=gamma link=log;
run;
/* Log Likelihood: -275.9746 */

/* checking goodness of fit */
proc genmod;
model chargesK = / dist=gamma link=log;
run;
/* Log Likelihood: -366.4021 */

data deviance_test;
deviance = -2*(-366.4021 - (-275.9746));
pvalue = 1 - probchi(deviance, 8);
run;
proc print noobs;
run;
```

Figure 9 Fitting Gamma Regression

H. SAS Code 4 Output



Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2226	0.2860	-0.7832	0.3379	0.61	0.4363
age		1	0.0345	0.0029	0.0289	0.0401	143.52	<.0001
sex	male	1	0.1195	0.0838	-0.0447	0.2838	2.03	0.1538
sex	female	0	0.0000	0.0000	0.0000	0.0000	.	.
bmi		1	0.0173	0.0076	0.0025	0.0321	5.22	0.0223
children		1	0.0520	0.0322	-0.0111	0.1152	2.61	0.1062
smoker	yes	1	1.6938	0.1011	1.4957	1.8919	280.87	<.0001
smoker	no	0	0.0000	0.0000	0.0000	0.0000	.	.
region	northeast	1	0.0093	0.1170	-0.2199	0.2386	0.01	0.9364
region	northwest	1	0.2674	0.1153	0.0413	0.4935	5.37	0.0204
region	southwest	1	0.0782	0.1127	-0.1427	0.2991	0.48	0.4880
region	southeast	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	6.3145	0.8704	4.8195	8.2732		

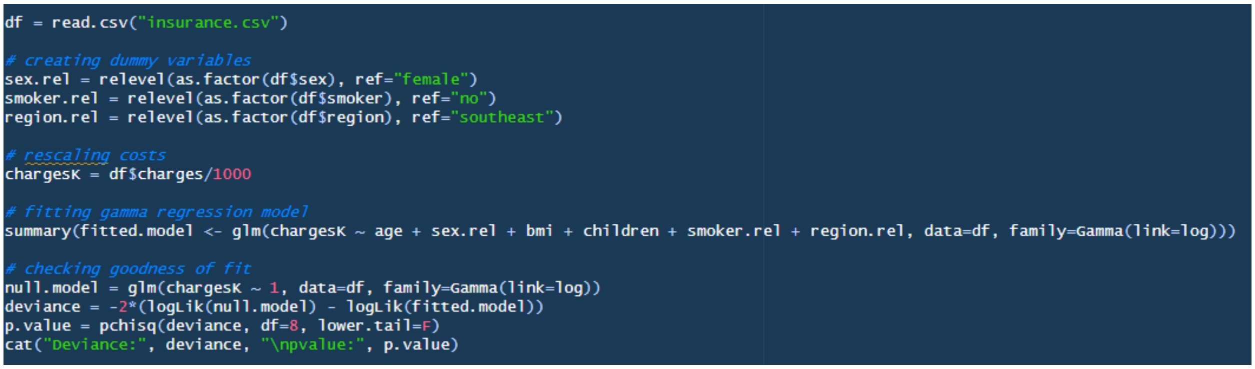
Log Likelihood	-275.9746

Log Likelihood	-366.4021

deviance	pvalue
180.855	0

Figure 10 Gamma Regression Output/ Checking Goodness-of-fit

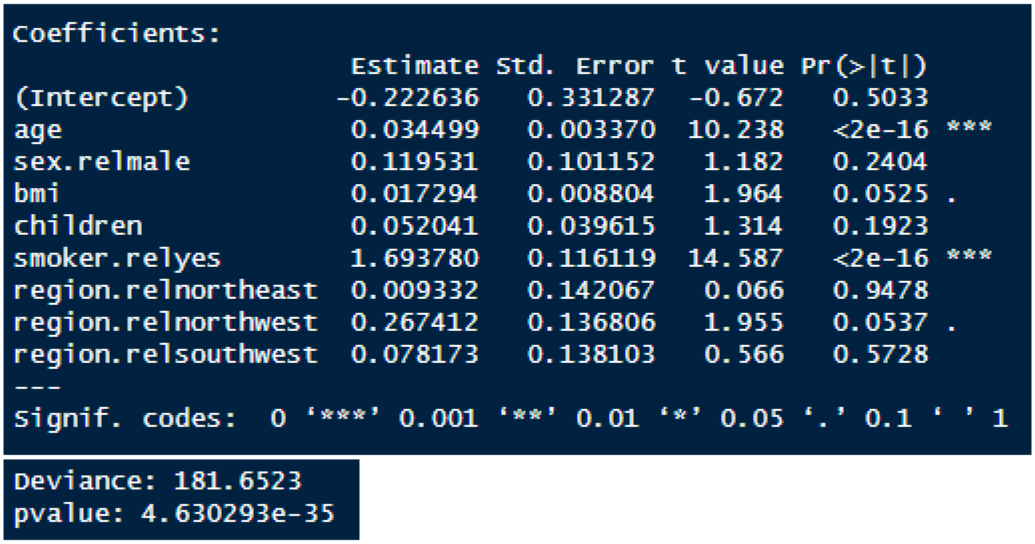
I. R Code 4



```
df = read.csv("insurance.csv")

# creating dummy variables
sex.rel = relevel(as.factor(df$sex), ref="female")
smoker.rel = relevel(as.factor(df$smoker), ref="no")
region.rel = relevel(as.factor(df$region), ref="southeast")

# rescaling costs
chargesK = df$charges/1000

# fitting gamma regression model
summary(fitted.model <- glm(chargesK ~ age + sex.rel + bmi + children + smoker.rel + region.rel, data=df, family=Gamma(link=log)))

# checking goodness of fit
null.model = glm(chargesK ~ 1, data=df, family=Gamma(link=log))
deviance = -2*(logLik(null.model) - logLik(fitted.model))
p.value = pchisq(deviance, df=8, lower.tail=F)
cat("Deviance:", deviance, "\npvalue:", p.value)
```

Figure 11 Fitting Gamma Regression

J. R Code 4 Output



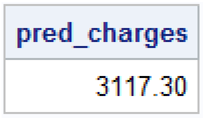
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -0.222636   0.331287  -0.672   0.5033    
age                   0.034499   0.003370  10.238   <2e-16 ***
sex.relmale           0.119531   0.101152   1.182   0.2404    
bmi                   0.017294   0.008804   1.964   0.0525 .  
children              0.052041   0.039615   1.314   0.1923    
smoker.relyes         1.693780   0.116119  14.587   <2e-16 ***
region.relnortheast   0.009332   0.142067   0.066   0.9478    
region.relnorthwest   0.267412   0.136806   1.955   0.0537 .  
region.relsouthwest   0.078173   0.138103   0.566   0.5728    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Deviance: 181.6523
pvalue: 4.630293e-35
```

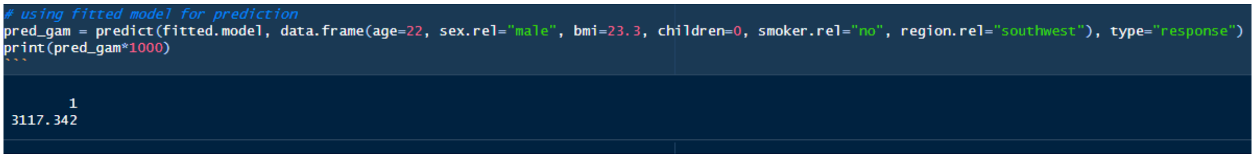
Figure 12 Gamma Regression Output/ Checking Goodness-of-fit

K. Fitted Model Prediction Code and Output

```
/* using fitted model for prediction */
data prediction;
input age sex$ 4-7 bmi children smoker$ 16-17 region$ 19-27;
cards;
22 male 23.3 0 no southwest
;

data healthinsurance;
set healthinsurance prediction;
run;

proc genmod;
class sex(ref="female") smoker(ref="no") region(ref="southeast");
model chargesK = age sex bmi children smoker region
/dist=gamma link=log;
output out=outdata p=pchargesK;
run;

data outdata;
set outdata;
pred_charges= 1000*pchargesK;
run;

proc print data=outdata(firstobs=101) noobs;
var pred_charges;
run;
```



pred_charges
3117.30



~~~
# using fitted model for prediction
pred_gam = predict(fitted.model, data.frame(age=22, sex.rel="male", bmi=23.3, children=0, smoker.rel="no", region.rel="southwest"), type="response")
print(pred_gam*1000)
```

        1
 3117.342
~~~

VII. References

Choi, Miri. “Medical Cost Personal Datasets.” *Kaggle*, 21 Feb. 2018, www.kaggle.com/datasets/mirichoi0218/insurance.

[1] “Your Total Costs for Health Care: Premium, Deductible, and out-of-Pocket Costs.” *Your Total Costs for Health Care: Premium, Deductible, and out-of-Pocket Costs | HealthCare.Gov*, www.healthcare.gov/choose-a-plan/your-total-costs/.

[2] Kagan, Julia. “Insurance Premium Defined, How It’s Calculated, and Types.” *Investopedia*, Investopedia, www.investopedia.com/terms/i/insurance-premium.asp/.

[3] Wisconsin School of Business. “How the Insurance Industry Uses Analytics to Make Decisions.” *Wisconsin School of Business*, 28 July 2021, business.wisc.edu/news/how-the-insurance-industry-uses-analytics-to-make-decisions/.