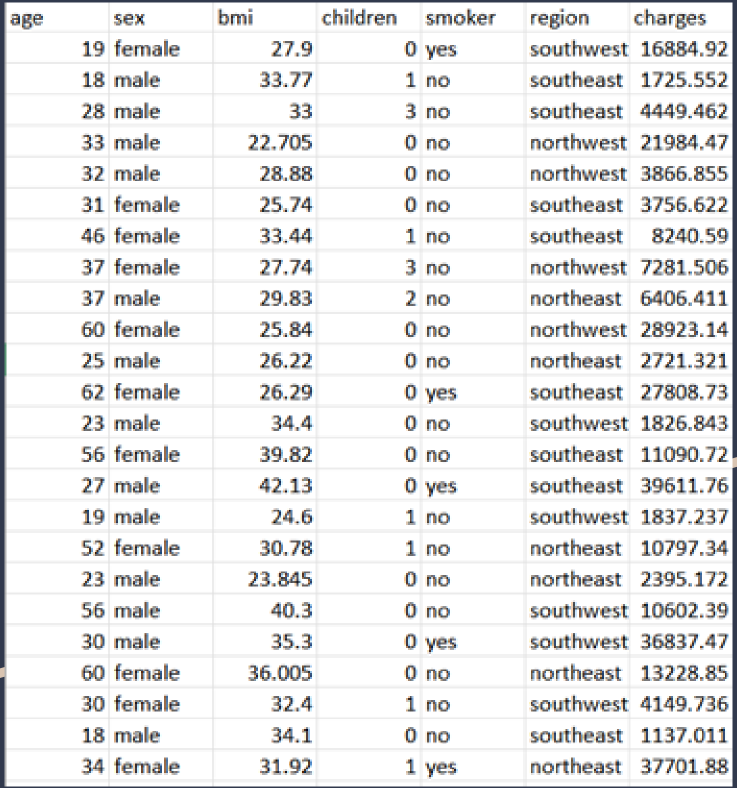
Health Insurance Premium Charges in US

By Nate Talampas
Model Used: Box-Cox and Gamma

Why health insurance?

I am fascinated by the intersection of health care, insurance, and technology. With technological innovations like AI and Blockchain making revolutionary changes in the industry, I wanted to base my project on understanding how insurance companies evaluate risks associated with patients.

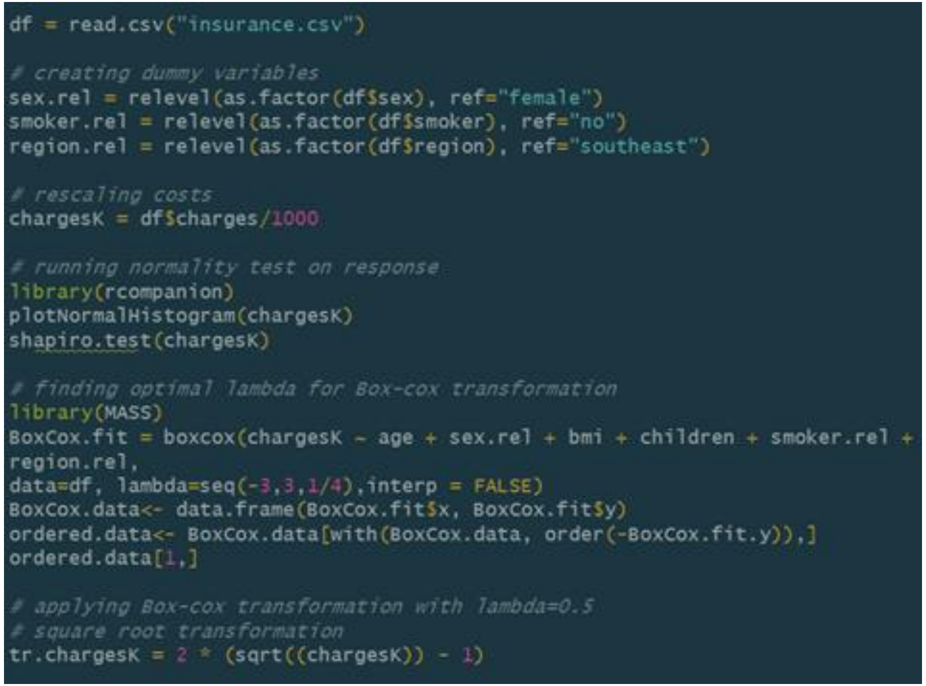
Data Overview



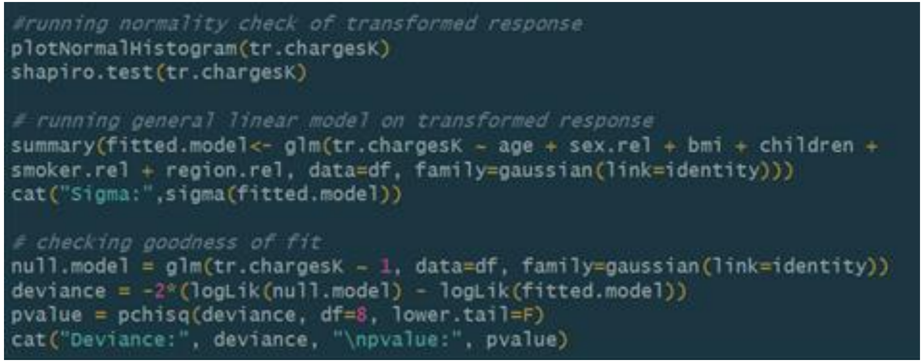
age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411
60	female	25.84	0	no	northwest	28923.14
25	male	26.22	0	no	northeast	2721.321
62	female	26.29	0	yes	southeast	27808.73
23	male	34.4	0	no	southwest	1826.843
56	female	39.82	0	no	southeast	11090.72
27	male	42.13	0	yes	southeast	39611.76
19	male	24.6	1	no	southwest	1837.237
52	female	30.78	1	no	northeast	10797.34
23	male	23.845	0	no	northeast	2395.172
56	male	40.3	0	no	southwest	10602.39
30	male	35.3	0	yes	southwest	36837.47
60	female	36.005	0	no	northeast	13228.85
30	female	32.4	1	no	southwest	4149.736
18	male	34.1	0	no	southeast	1137.011
34	female	31.92	1	yes	northeast	37701.88

The dataset contains 100 observations with the following attributes: Age, Sex, BMI, Number of Children, Smoker, Residential Region, and Individual Medical Costs Billed By Health Insurance.

Box-Cox Regression R Code

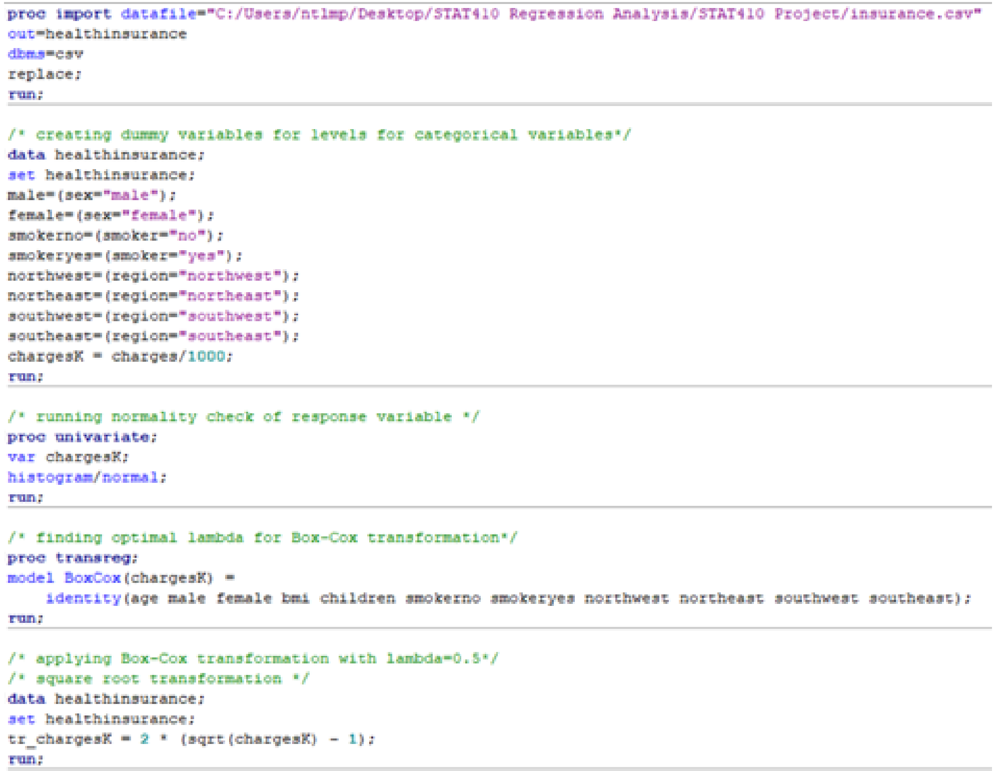


```r
df = read.csv("insurance.csv")

# 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)

# 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,]

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

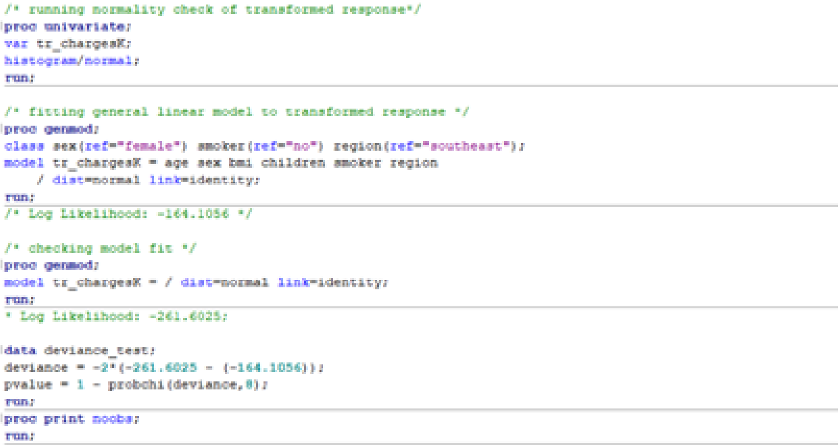


```r
#running normality check of transformed response
plotNormalHistogram(tr.chargesK)
shapiro.test(tr.chargesK)

# running general linear model on transformed response
summary(fitted.model<- glm(tr.chargesK ~ age + sex.rel + bmi + children +
smoker.rel + region.rel, data=df, family=gaussian(link=identity)))
cat("Sigma:",sigma(fitted.model))

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

Box-Cox Regression SAS Code

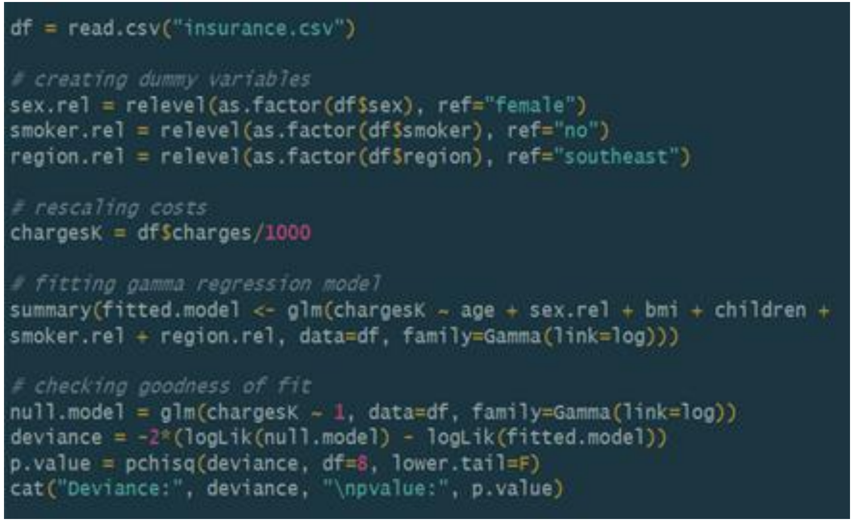


```
proc import datafile="C:/Users/ntlmp/Desktop/STAT410 Regression Analysis/STAT410 Project/insurance.csv"
out=healthinsurance
dbms=csv
replace;
run;

/* 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;

/* 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;

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

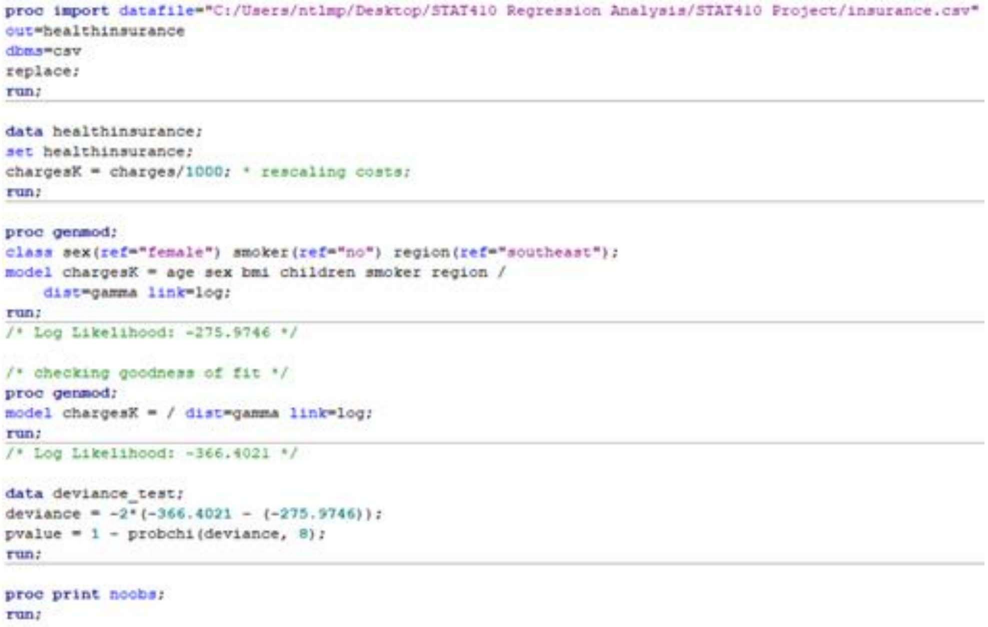


```
/* running normality check of transformed response*/
proc univariate;
var tr_chargesK;
histogram/normal;
run;

/* fitting general linear model to transformed response */
proc genmod;
class sex(ref="female") smoker(ref="no") region(ref="southeast");
model tr_chargesK = age sex bmi children smoker region
    / dist=normal link=identity;
run;
/* Log Likelihood: -164.1056 */

/* checking model fit */
proc genmod;
model tr_chargesK = / dist=normal link=identity;
run;
* Log Likelihood: -261.6025;

data deviance_test;
deviance = -2*(-261.6025 - (-164.1056));
pvalue = 1 - probchi(deviance,8);
run;
proc print noobs;
run;
```

Gamma Regression Code

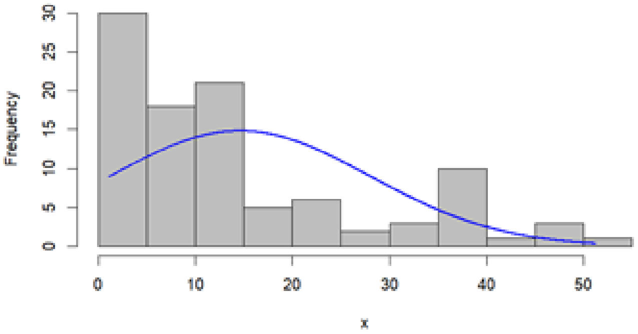


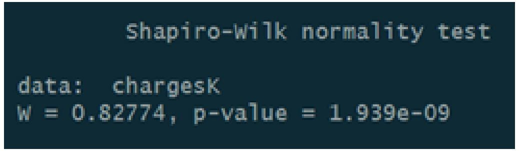
```r
df = read.csv("insurance.csv")

# 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

# 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)
```



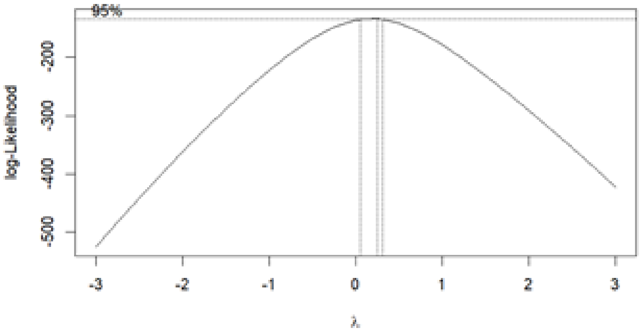
```sas
proc import datafile="C:/Users/ntlmp/Desktop/STAT410 Regression Analysis/STAT410 Project/insurance.csv"
out=healthinsurance
dbms=csv
replace;
run;

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;
```

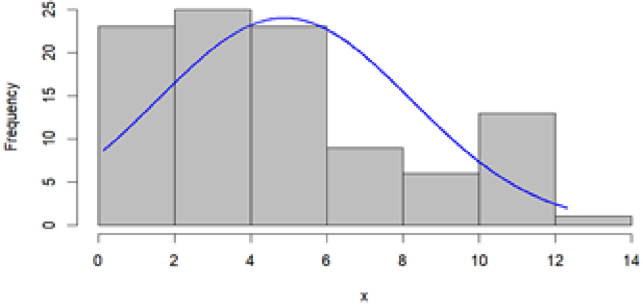
Box-Cox Regression R Output

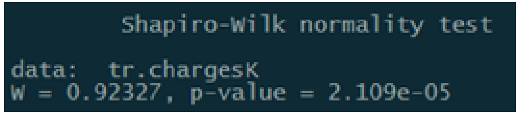




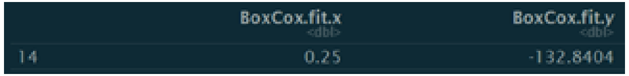
```
                 Shapiro-Wilk normality test

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





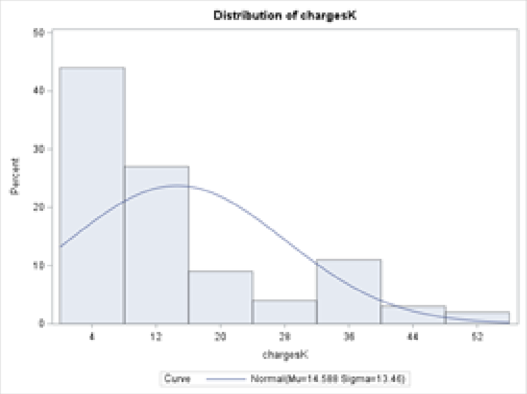


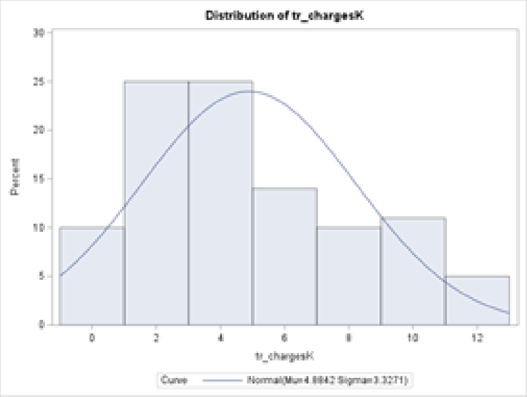
```
                 Shapiro-Wilk normality test

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

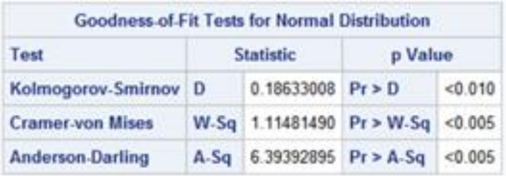


	BoxCox.fit.x <dbl>	BoxCox.fit.y <dbl>
14	0.25	-132.8404

Box-Cox Regression SAS Output

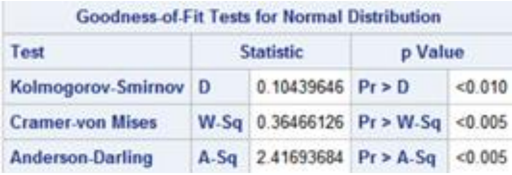






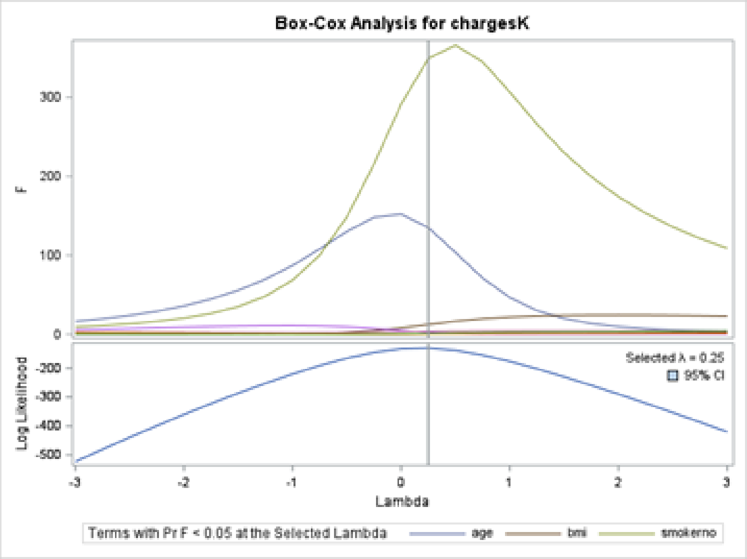
Goodness-of-Fit Tests for Normal Distribution

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

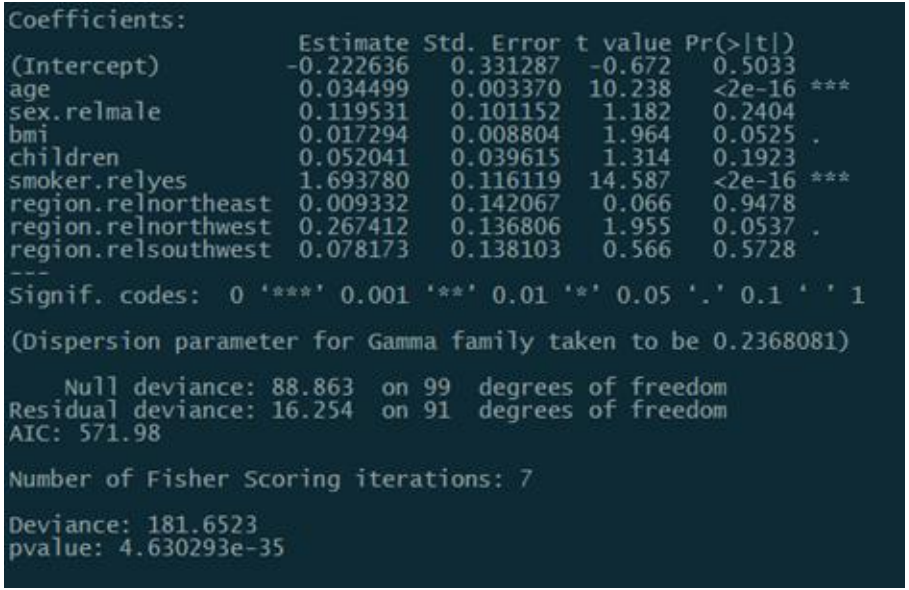


Goodness-of-Fit Tests for Normal Distribution

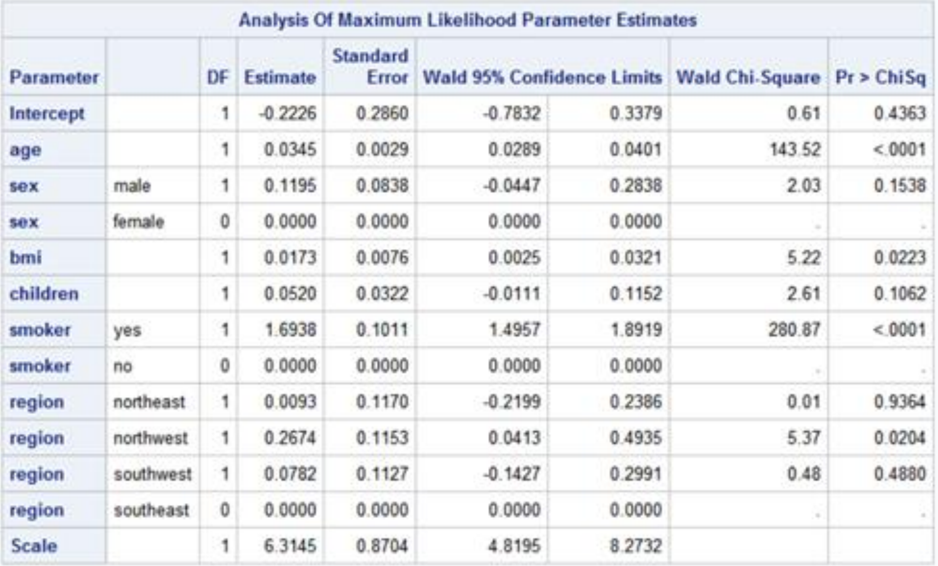
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.10439646	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.36466126	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2.41693684	Pr > A-Sq	<0.005



Gamma Regression Output

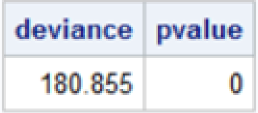


```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.222636   0.331287  -0.672   0.5033
age                   0.034499   0.003370  10.238   <2e-16 ***
sex.relmale           0.119531   0.101152   1.182   0.2404
bmi                   0.017294   0.008804   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

(Dispersion parameter for Gamma family taken to be 0.2368081)

    Null deviance: 88.863  on 99  degrees of freedom
Residual deviance: 16.254  on 91  degrees of freedom
AIC: 571.98

Number of Fisher Scoring iterations: 7

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



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		



deviance	pvalue
180.855	0

Fitted Model

The Gamma regression fitted model can be written as:

$$\hat{E}(charges) =$$

$$exp(-0.2226 + 0.0345 \cdot age + 0.1195 \cdot male + 0.0173 \cdot bmi + 0.0520 \cdot children + 1.6938 \cdot smoker + 0.0093 \cdot northeast + 0.2674 \cdot northwest + 0.0782 \cdot southwest)$$

Interpretation of Significant Predictors

Significant predictors at the 5% level include age, BMI, smoker, and region northwest.

- 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.

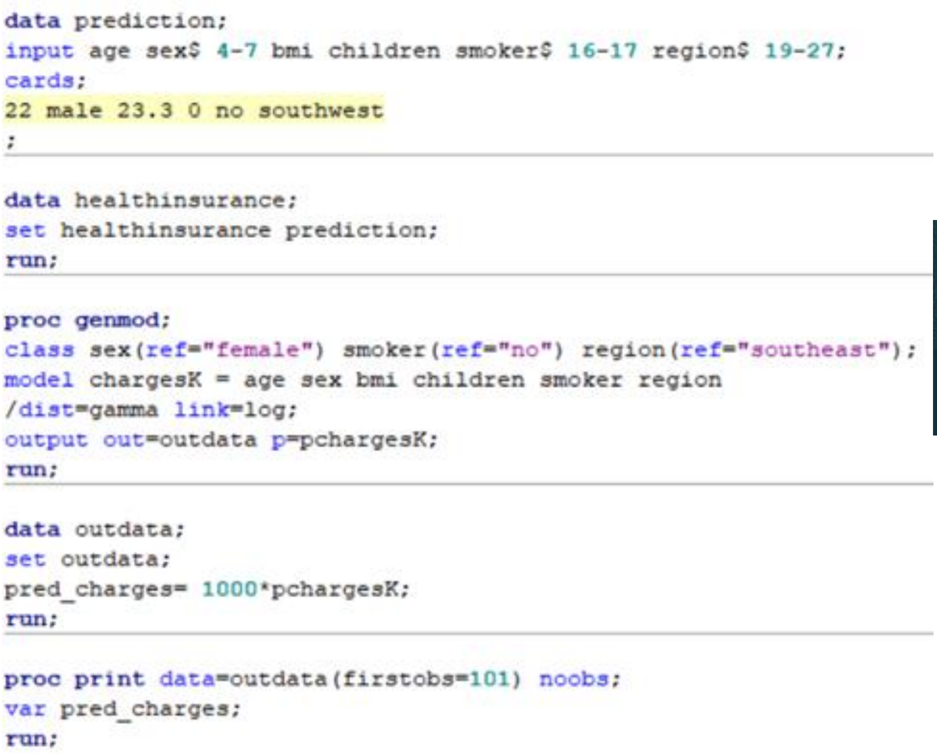
Fitted Model Prediction

I am a 22 year old male with a BMI of 23.3. I have no children, do not smoke, and I live in the southwest. What would my predicted health insurance costs be?

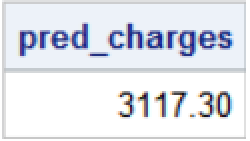
The Gamma fitted model value can be calculated as:

$$\exp\left(-0.2226 + 0.0345(22) + 0.1195 + 0.0173(23.3) + 0.0782\right) \cdot 1000 = 3117.994$$

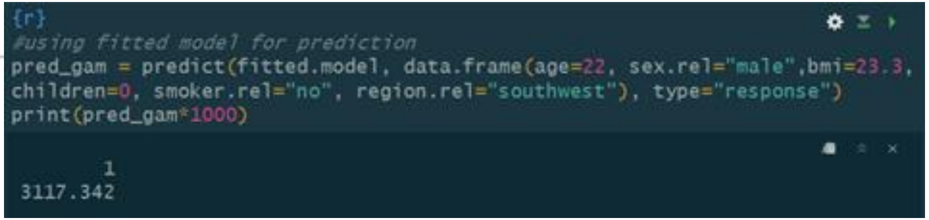
Gamma Fitted Model Code and Output



```
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



```
{r}
#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
```

Thank you, Dr. Olga and my fellow classmates!

I appreciate all of your time and attention. Good luck with the rest of your finals and have a fantastic break!