Health Insurance Costs

Predictive Analytics Project

Big Data and Business - ECON 386
Spring 2020

Carter Bradsky, Tom Hollerbach, Pedro Andrade, Taylor Danielson, Chase Lester, Alejandro Ortiz

Executive Summary

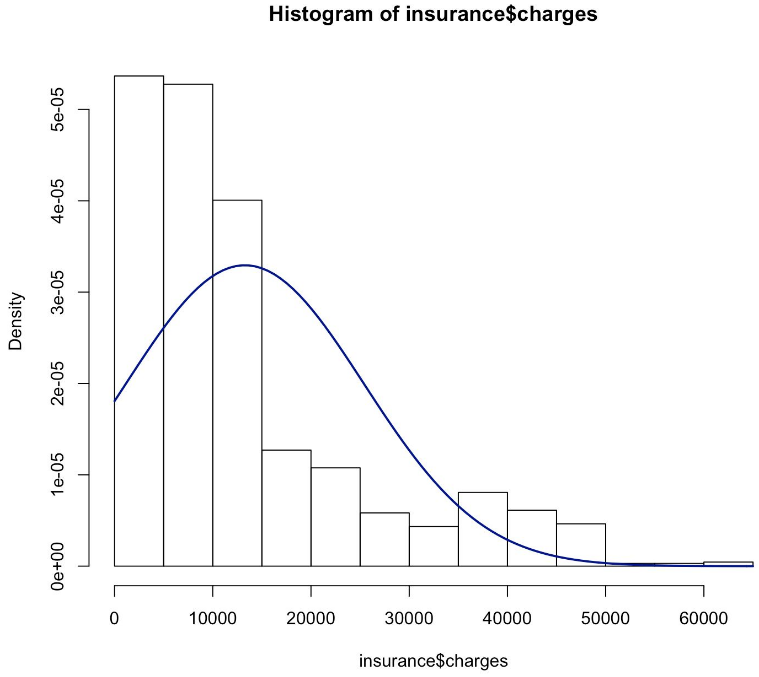
Section 1

The process of finding this health insurance data set was intensive and unexpected. We were looking for a business related data set that focused more on the stock market, but soon realized that a majority of the data sets that we found were time-series data sets, which would not run well with our project. So, we ended up looking to this insurance charges dataset that focused on seven main variables: Age, Sex, BMI, Number of Children, Smoker or not, location by region, and total health charges for that observation. We were interested to see what constituted a particular individual into being in the upper echelon of health charges, and, given the current pandemic, thought it would be interesting to see how things would be assessed during this global health crisis from an insurance company's perspective. This is a cross-sectional dataset with 1338 rows (observations) and 13 total columns (different variables) when factoring in the newly created dummy variables.

Exploratory Analysis

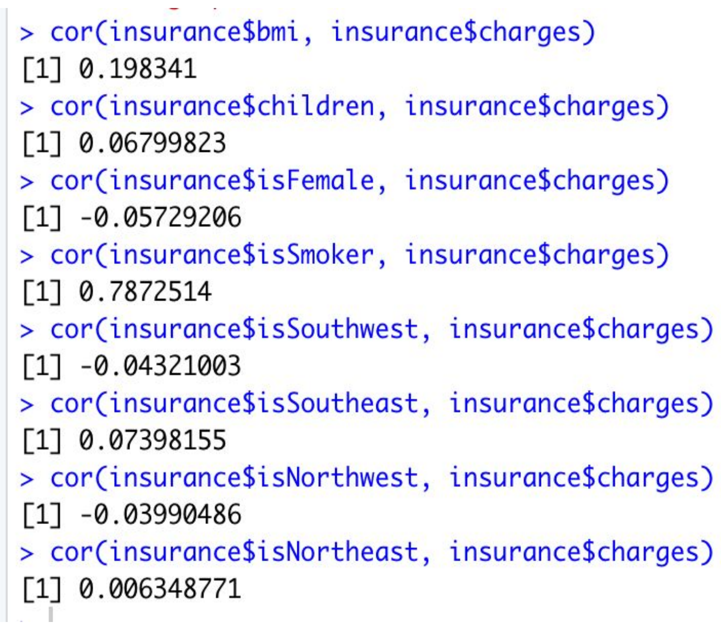
Section 2

The following is the descriptive analysis we ran:



Analysis:

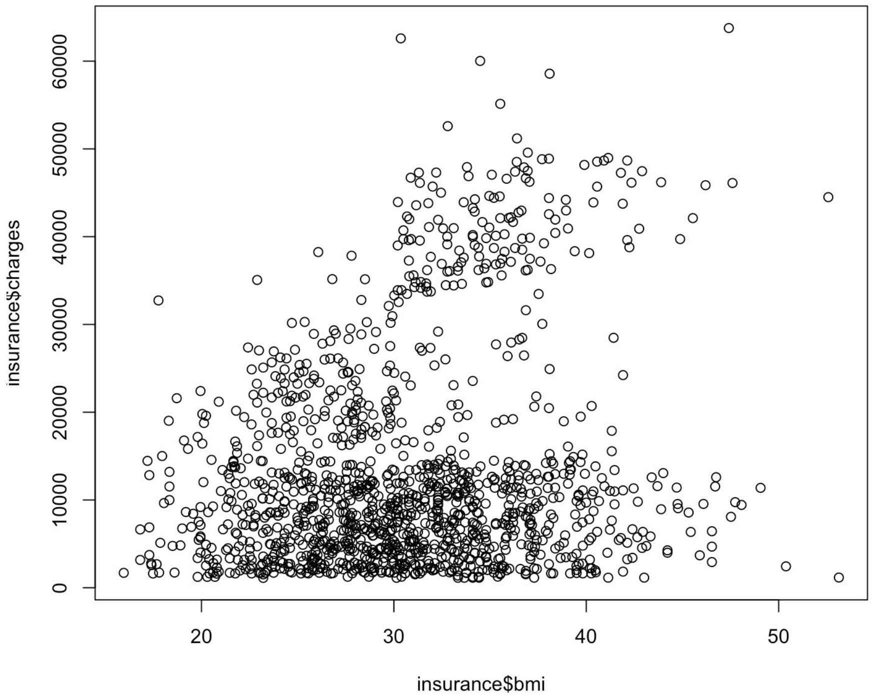
This histogram shows that the higher density of charges is more focused on the less costly amounts in the $100 to $10,000 range with a few skewed points reaching a max of $63,770 and a minimum amount of $1,122 (according to the summary statistics). The density curve shows that the max frequency is about $10,000.



```
> cor(insurance$bmi, insurance$charges)
[1] 0.198341
> cor(insurance$children, insurance$charges)
[1] 0.06799823
> cor(insurance$isFemale, insurance$charges)
[1] -0.05729206
> cor(insurance$isSmoker, insurance$charges)
[1] 0.7872514
> cor(insurance$isSouthwest, insurance$charges)
[1] -0.04321003
> cor(insurance$isSoutheast, insurance$charges)
[1] 0.07398155
> cor(insurance$isNorthwest, insurance$charges)
[1] -0.03990486
> cor(insurance$isNortheast, insurance$charges)
[1] 0.006348771
```

Analysis:

The correlations above display whether or not there is a significant relationship between two variables (positive or negative). The highest correlation resulted between the variables isSmoker and Charges at .787. This result makes sense given the considerable risk of illness that smoking is proven to make, thus insurance premiums should be higher to adjust for that risk. The lowest/most negative correlation resulted between isFemale (gender) and Charges at -.057. This tells us that there is close to no relationship between being male and female and increased insurance charges.

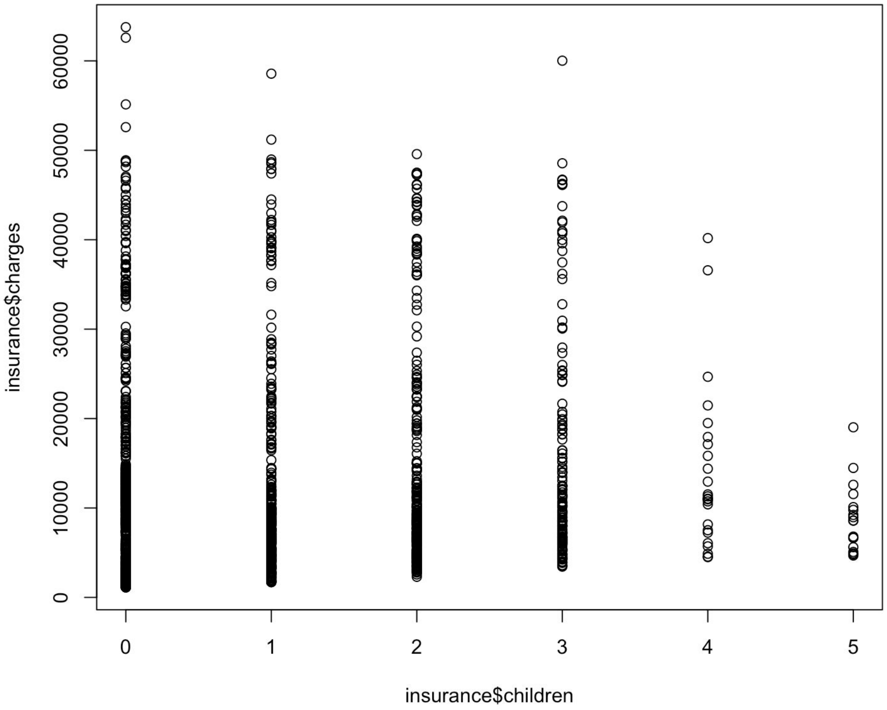


Analysis:

As we can see above, the relationship between Body Mass Index and Charges shows that the higher frequency, but lower insurance charges are centered around individuals with a BMI reading between 20-40, which, based on scientific research, is within the "over-weight" to "obese" classification. This result is expected given the fact that those with a higher BMI tend to be more prone to the risk of diseases such as heart disease and diabetes (to name a couple) both of these conditions pose significant financial burdens on their individuals, thus insurance companies adjust for that risk in their charges/use of funds.

Analysis:

When analyzing the relationship between Number of Children and Charges it can be seen that those with 0-3 children have higher and more frequent insurance charges. This proves unexpected in our a priori analysis because the assumption was that with increased dependents, there would be increased financial strain and potential for accidents to happen thus resulting in higher and more frequent insurance charges. However, it is proven that we are mistaken by the plot above.



Cleaning & Reprocessing

Section 2

Once we opened the dataset we followed these distinct steps:

1. Check for hard and soft data (fortunately this data set has hard only)
2. Searching for outliers/invalid/any missing data values in the dataset (of which there were none)
3. Assess whether there should be dummy variables created for categorical variables (some were created for region, gender, and smoker)
4. Prepare to the run the first regression tests

As can be seen, the data we gathered from this dataset was not very raw, and seemed to be pre-processed for analysis. As it made our cleaning process easier, we worry that this could lead to potential important data points being left out that could heavily influence our analysis but were not offered because the curators of this data did not think it was necessary.

Regression Task - Model Proposals and Diagnostics

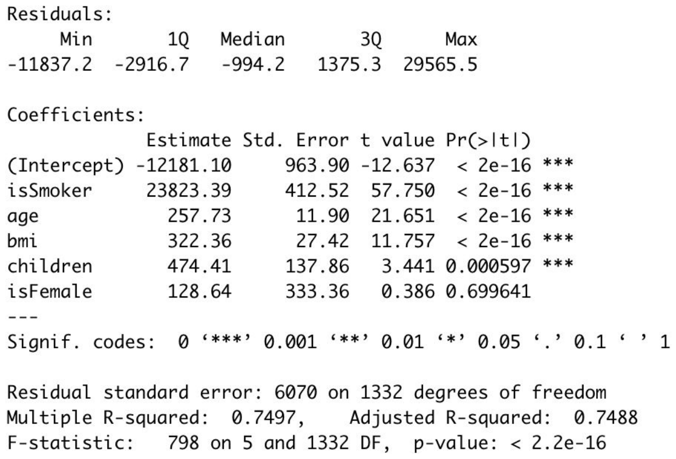
Section 3

Carter Bradsky

Variables used in regression: isSmoker, Age, BMI, Children, isFemale (Gender)

R-squared = .7497
isSmoker, Age, BMI, Children = Significant at the 100% level
isFemale = not statistically significant, will still include in model to prevent omitted variable bias
Equation: **Insurance Charges** = -12181.1 + **isSmoker** * 23,823.39 + **age** * 257.73 + **bmi** * 322.36 + **children** * 474.41 + **isFemale** * 128.64

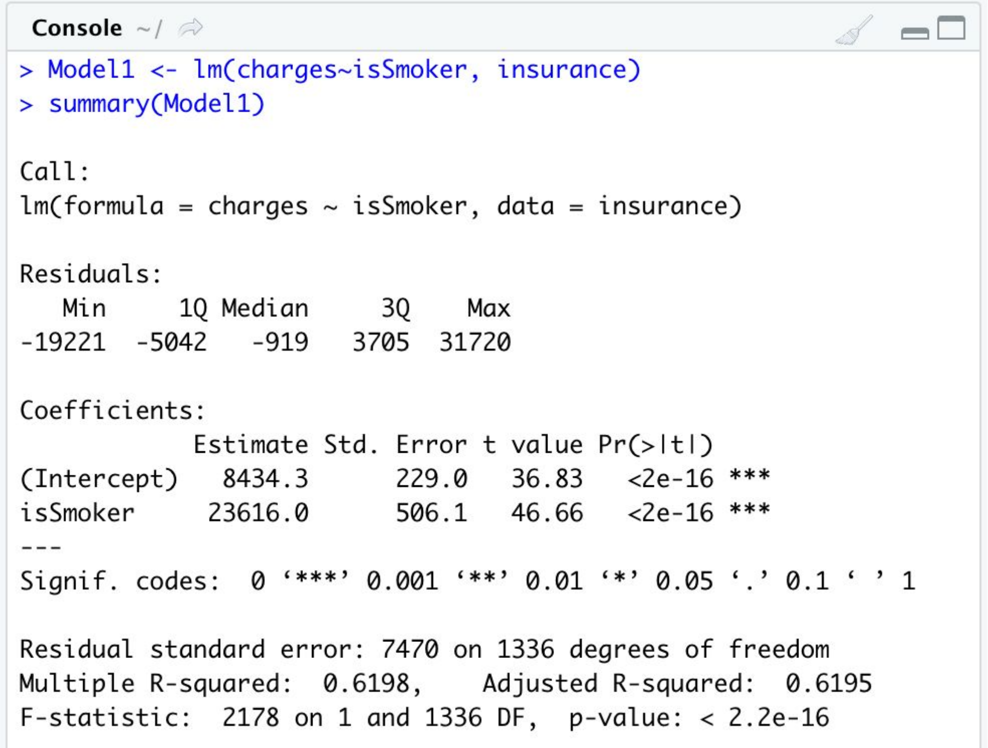
Summary: The model was selected out of 3 other regression models due to its high r-squared (74.97%), meaning the model explains a good amount of the variability around the data's mean, its high number of statistically significant variables, and an overall good model complexity to explain the data. The model as a whole is considered jointly significant with a p-value = 2.2e-16. No restraints such as regularization were used in this analysis. The out of sample error that i got for this model was the lowest of the variable combinations that I attempted and it came in at: 5,881 which signifies that the model may be too complex resulting in a high amount of variability being input given the higher number of independent variables taken into consideration. For this model, I split the data into a 70% training set and 30% testing set.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-11837.2  -2916.7   -994.2   1375.3  29565.5 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12181.10     963.90 -12.637  < 2e-16 ***
isSmoker     23823.39     412.52  57.750  < 2e-16 ***
age            257.73      11.90  21.651  < 2e-16 ***
bmi            322.36      27.42  11.757  < 2e-16 ***
children       474.41     137.86   3.441 0.000597 ***
isFemale       128.64     333.36   0.386 0.699641    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6070 on 1332 degrees of freedom
Multiple R-squared:  0.7497,	Adjusted R-squared:  0.7488 
F-statistic:   798 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Tom Hollerbach

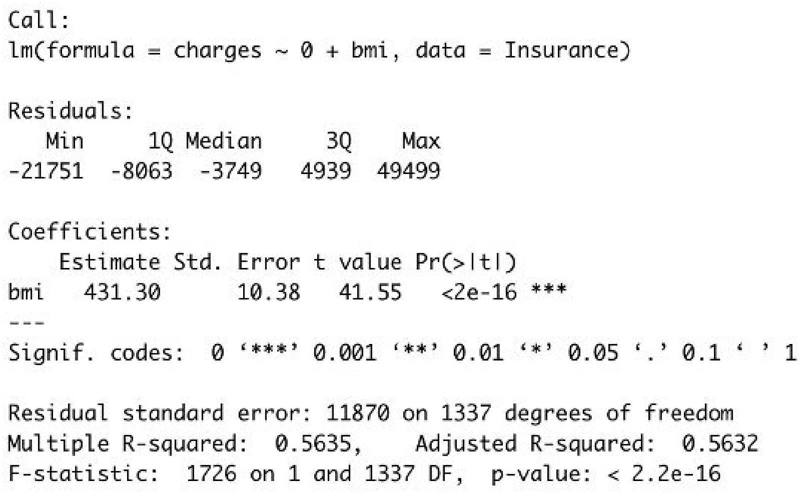
Variables used in linear regression:
Model1: isSmoker
Result (screenshot):



```
> Model1 <- lm(charges~isSmoker, insurance)
> summary(Model1)

Call:
lm(formula = charges ~ isSmoker, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max 
-19221  -5042   -919   3705  31720 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8434.3      229.0   36.83   <2e-16 ***
isSmoker     23616.0      506.1   46.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7470 on 1336 degrees of freedom
Multiple R-squared:  0.6198,	Adjusted R-squared:  0.6195 
F-statistic:  2178 on 1 and 1336 DF,  p-value: < 2.2e-16
```

This model has an R-squared of 0.6198 indicating that approximately 62% of the variation in charges can be attributed to whether or not someone is a smoker. The isSmoker variable had a beta parameter of 23,616.0 with a 95% confidence interval of (22623.175, 24608.752). The model overall was statistically significant with a p-value very close to 0 and all parameters were significant above the 99% level.

Pedro Andrade
Result (screenshot):
Linear Regression
Variables used in regression: BMI
Model1<-lm(charges~0+bmi, Insurance)



```
Call:
lm(formula = charges ~ 0 + bmi, data = Insurance)

Residuals:
   Min     1Q Median     3Q    Max 
-21751  -8063  -3749   4939  49499 

Coefficients:
    Estimate Std. Error t value Pr(>|t|)    
bmi   431.30      10.38   41.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11870 on 1337 degrees of freedom
Multiple R-squared:  0.5635,	Adjusted R-squared:  0.5632 
F-statistic:  1726 on 1 and 1337 DF,  p-value: < 2.2e-16
```

This model has an R-squared of 0.5635 indicating that approximately 56% of the variation in charges can be attributed to variations in the body mass index (BMI). The BMI variable had a beta parameter of 431.30 with a 95% confidence interval of (410.993, 451.6634). The model overall was statistically significant with a p-value very close to 0 and all parameters were significant above the 99% level. This model indicates that a 1 unit increase in BMI will increase charges by 431.30 dollars.

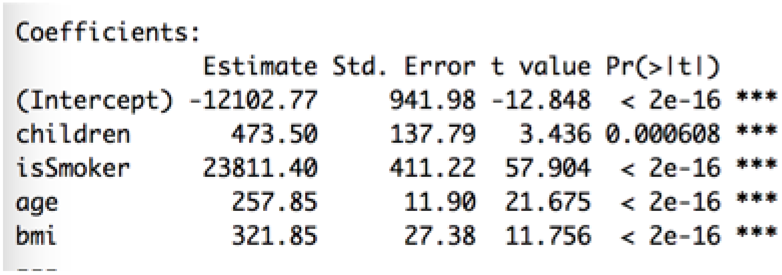
R-squared: 56.3%
Bmi significant at the 95% level (***)
Confidence Interval: (410.993, 451.6634)
RMSE= 11867.27

Taylor Danielson

Model:

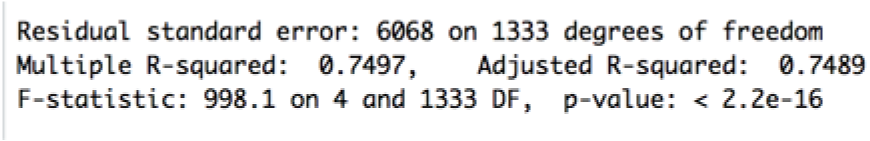
$$\textbf{charges} = -12102 + 473.50\ (\textbf{children}) + 23811.40\ (\textbf{isSmoker}) + 257.85\ (\textbf{age}) + 321.85\ (\textbf{bmi}) + e$$

Each variable in this model is statistically significant at the 0.1% level when controlling for children, smoking, age, and bmi. The beta values, standard error, and statistical significance of each variable is shown below.



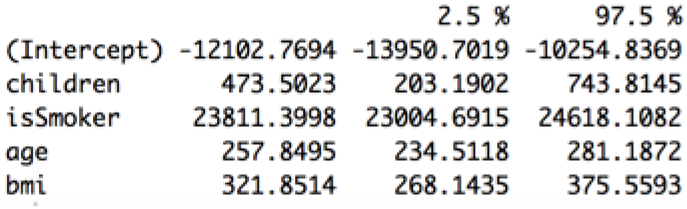
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
isSmoker     23811.40     411.22  57.904  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
---
```

The overall model has a high Adjusted R-squared value of 0.7489, demonstrating strong explanatory power, even when penalized for model complexity. This model shows that an individual's number of children, whether or not they're a smoker, their age, and their body mass index accounts for 74.89% of the variation in insurance charges. By running an F-test, we are able to determine that the model is jointly significant because the p-value is very close to zero, much smaller than 0.1%.



```
Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7489
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

The confidence intervals also give insight into the precision of the model. The confidence intervals show that the children variable has the largest confidence interval, showing that Beta prediction may be the least precise. However, none of the confidence intervals changed signs so there is strong evidence that each independent variable has a definitive positive relationship, on average, with insurance charges.

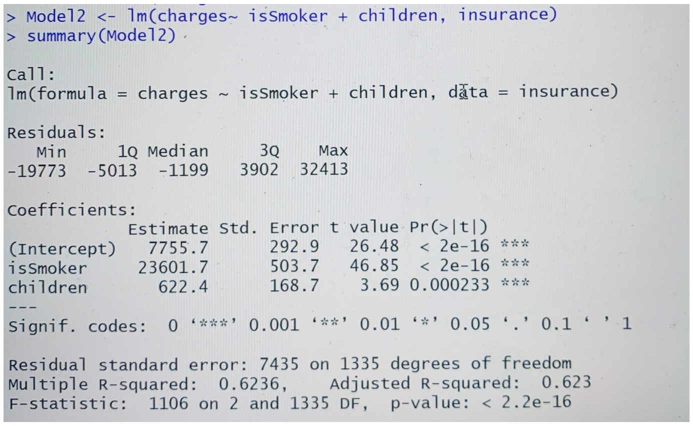


```
                    2.5 %      97.5 %
(Intercept) -12102.7694 -13950.7019 -10254.8369
children       473.5023    203.1902    743.8145
isSmoker     23811.3998  23004.6915  24618.1082
age            257.8495    234.5118    281.1872
bmi            321.8514    268.1435    375.5593
```

Chase Lester

Variables used in regression:
isSmoker + children

Result (screenshot):



```
> Model2 <- lm(charges~ isSmoker + children, insurance)
> summary(Model2)

Call:
lm(formula = charges ~ isSmoker + children, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max 
-19773  -5013  -1199   3902  32413 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7755.7      292.9   26.48  < 2e-16 ***
isSmoker     23601.7      503.7   46.85  < 2e-16 ***
children       622.4      168.7    3.69 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7435 on 1335 degrees of freedom
Multiple R-squared:  0.6236,	Adjusted R-squared:  0.623 
F-statistic:  1106 on 2 and 1335 DF,  p-value: < 2.2e-16
```

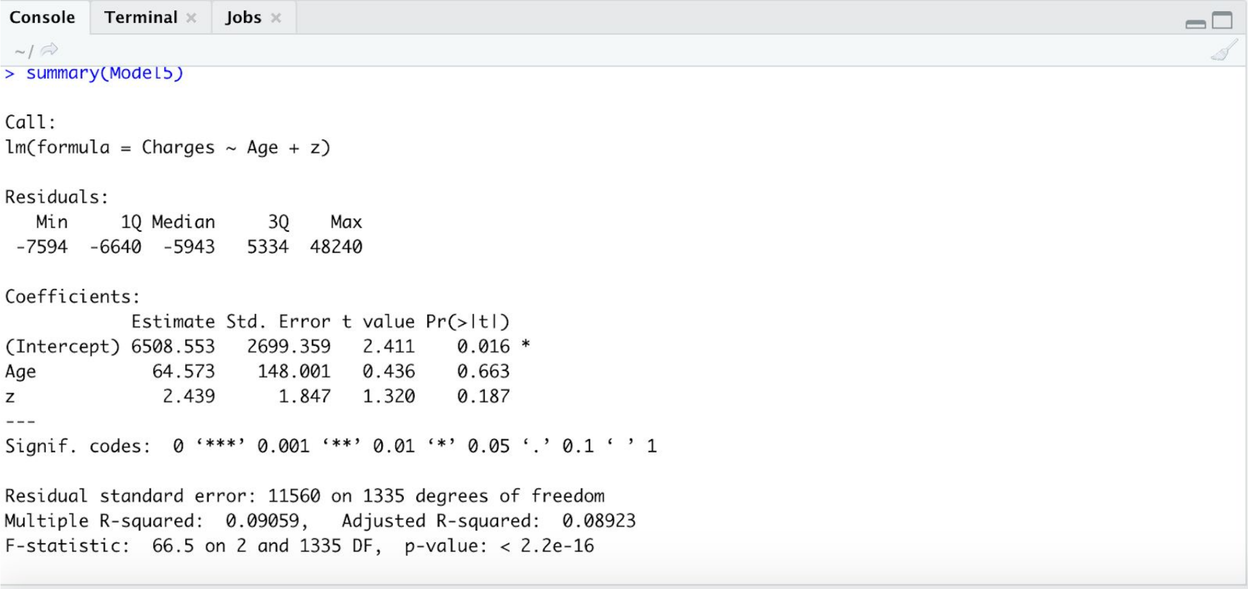
This model has an R-squared of .6236 which indicates that approximately 62% of the variation in charges can be attributed to the variations in isSmoker and children. The isSmoker variable had a beta parameter of 23,601 and the children had a beta parameter of 622. We can tell the model is jointly significant because the p-value of 2.2e-16 is close to zero.

R-squared: 62.36%
isSmoker and children significant at the 95% level
Confidence interval: isSmoker(22,594.3 - 24,609.1), children(285 - 959.8)

Alejandro Romo:

Variables used in regression:
Model5= Charges+Age

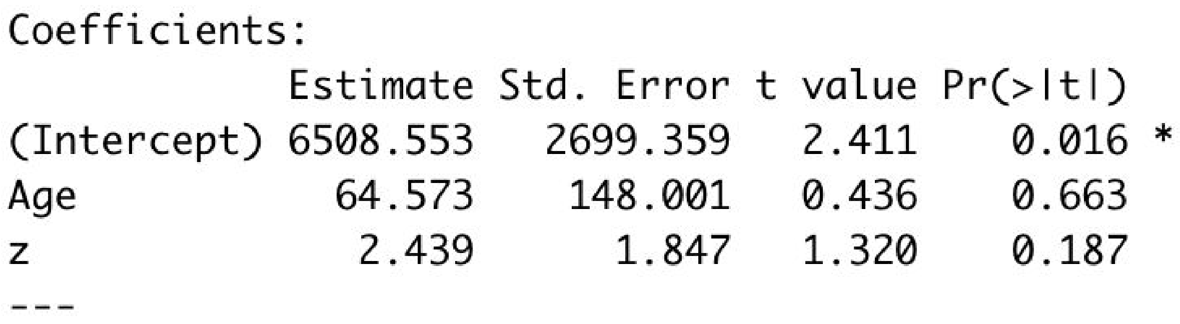
Result (screenshot):



```
> summary(Model5)

Call:
lm(formula = Charges ~ Age + z)

Residuals:
   Min     1Q Median     3Q    Max 
 -7594  -6640  -5943   5334  48240 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 6508.553   2699.359   2.411    0.016 *
Age           64.573    148.001   0.436    0.663  
z              2.439      1.847   1.320    0.187  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11560 on 1335 degrees of freedom
Multiple R-squared:  0.09059,	Adjusted R-squared:  0.08923 
F-statistic:  66.5 on 2 and 1335 DF,  p-value: < 2.2e-16
```

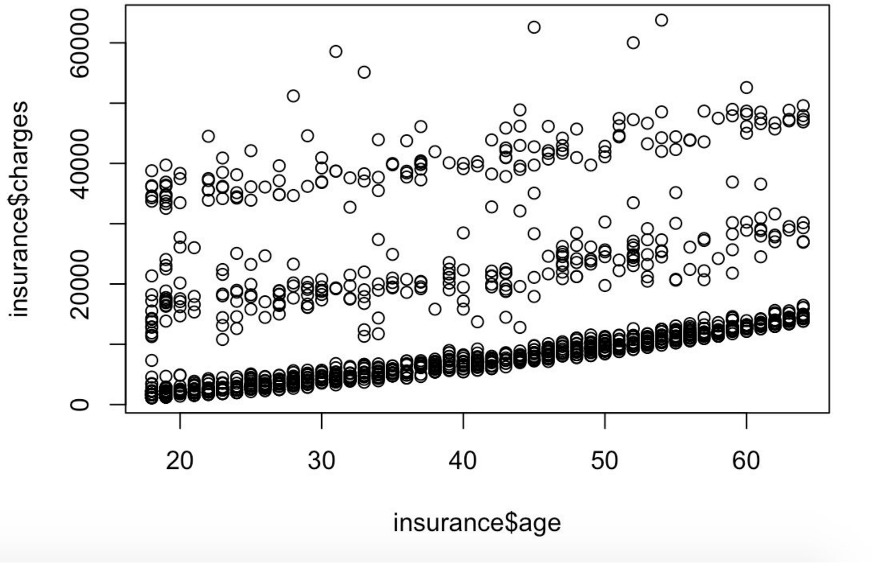
Equations:
Age <- c(0:64)
Charges <- 6508.553 + 64.573 * x + 2.439 * x^2

Each variable in this model is statistically significant at the 0.1% level when measuring the relation between the age and the charges of the applicants. The rest of the inputs including the standard error, beta values and statistical significance are shown below:



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 6508.553   2699.359   2.411    0.016 *
Age           64.573    148.001   0.436    0.663  
z              2.439      1.847   1.320    0.187  
---
```

The multiple R-Squared and the Adjusted R- Squared are not as high as we thought, as they show 0.0959 and 0.08923, however this doesn't is bad, because as show before in other regression models, the biggest correlations are found between variable charges and variable isSmoker as well as charges and bmi. So on Thomas's regression, the correlation found between the variables charges, age and issmoker, shows a higher R-Squared, meaning the goodness of fit is way higher. Another example of this, is Carter's regression, using more variables that have a bigger impact. Going back to my model, plotting the charges per age shown that in fact there is a high correlation but by levels:



The way I interpret this, is that age does play a factor in the amount of charge with every other relevant variable such as isSmoker, BMI and Children. The trends the graph shows that for each variable that is added to the regression, the charges do raise as well with the age. That is the reason why there are three main curves on the graph.

Regression Task: Validation and Model Selection

Section 4

Carter Bradsky

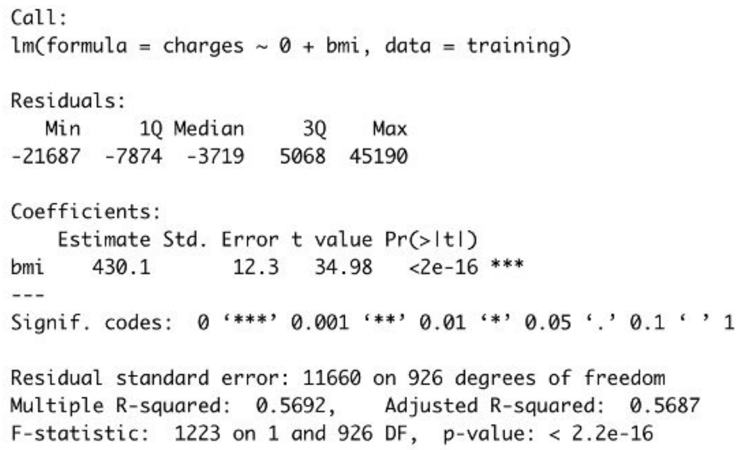
In order to validate this model, I randomly split the data into 70% training set and 30% test set, calculating the RMSE for the select number of variables in an effort to assess the out of sample error for this regression and see how well it performed relative to other models that were being tested. As was noted above, the out of sample error came in at around 5,881, which, compared to other models, was the lowest out of sample error I was able to calculate. However, it was still high given my peers' models so it can be interpreted as the model not being as accurate in its prediction when considering out of sample prediction.

Tom Hollerbach

In order to validate this model, I split the data randomly into 70% testing and 30% training. From this, I got an out of sample error of 0 which can be attributed to the simplicity of the model having only one independent, binary variable. No restraints, such as regularization, were used in this analysis.

Pedro Andrade

In order to validate my regression model, I split the data randomly into 70% training and 30% testing. The model I made with the training data gave me a beta coefficient of 430.1, R-squared of almost 57% and a residual standard error of 11,660.



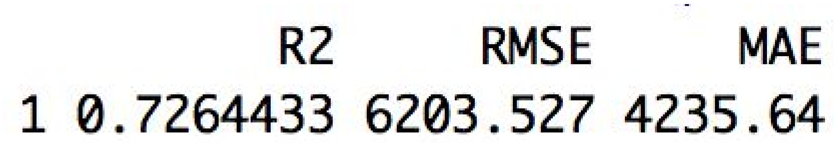
```
Call:
lm(formula = charges ~ 0 + bmi, data = training)

Residuals:
   Min     1Q Median     3Q    Max 
-21687  -7874  -3719   5068  45190 

Coefficients:
    Estimate Std. Error t value Pr(>|t|)    
bmi    430.1       12.3   34.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11660 on 926 degrees of freedom
Multiple R-squared:  0.5692,	Adjusted R-squared:  0.5687 
F-statistic:  1223 on 1 and 926 DF,  p-value: < 2.2e-16
```

Taylor Danielson

Model:

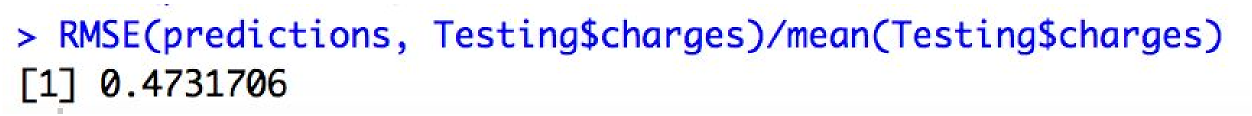
$$\textbf{charges} = -12102 + 473.50\ (\textbf{children}) + 23811.40\ (\textbf{isSmoker}) + 257.85\ (\textbf{age}) + 321.85\ (\textbf{bmi}) + e$$

To validate this model, I split my data randomly into a 70% train and 30% test partition. Then I calculated the value of the R2, RMSE, and MAE



```
         R2     RMSE     MAE
1 0.7264433 6203.527 4235.64
```

I noticed that the RMSE was a lot higher than I had expected it to be. So I decided to calculate the prediction error rate by dividing the RMSE by the average value of the outcome variable. Ideally, this value will be as close to zero as possible. However, the value I got for the error prediction rate was approximately 47%.



```
> RMSE(predictions, Testing$charges)/mean(Testing$charges)
[1] 0.4731706
```

Chase Lester

In order to validate this model, I randomly split the data into a 70% training set and 30% test partition. Then, I calculated the RMSE and it came back as zero. This

Alejandro Ortiz

Independent Variable(s): Age, Charges

Goodness of Fit: ~.10

Significance: ***Age^2

Confidence Interval: (-1.184622 , 6.06244)

In order to do the regression for this data, first we had to clean all the data and find correlations. Then after testing and trying with other variables, there was a problem with the data that found a different number of observations, so I had to ungroup the data and get that into vectors in order to be able to do the regressions. The RSME came back as zero.

Model Selected: Carter Bradsky

Insurance Charges = -12181.1 + **isSmoker** * 23,823.39 + **age** * 257.73 + **bmi** * 322.36 + **children** * 474.41 + **isFemale** * 128.64

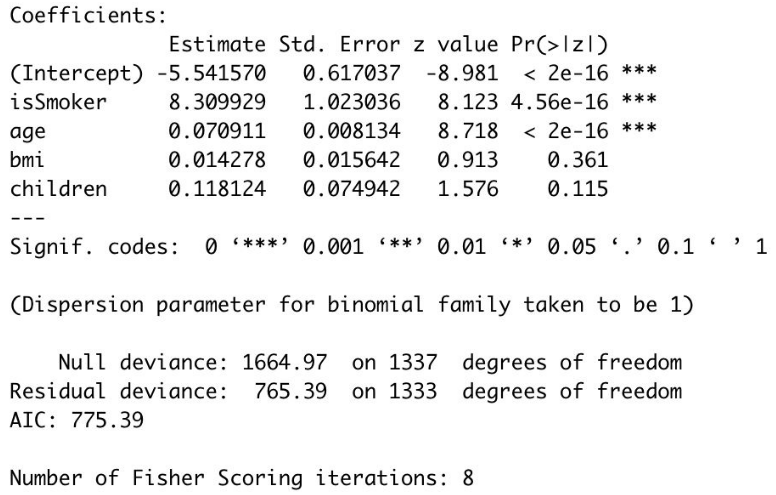
When selecting a final model we wanted to choose one that captured the effects that multiple variables had on charges. We concluded that Carter's models did this best with a high R-squared (74.97%) and joint model statistical significance at 95% (p-value = 2.2e-16).

Classification Task: Model Proposals and Diagnostics

Section 5

Carter Bradsky

Variables: isSmoker, age, bmi, children
Null deviance = 1664.97
Residual Deviance = 765.39
AIC = 775.39
Number of Iterations = 8 to reach optimum
Statistically Significant Variables (95% confidence interval) = isSmoker, age



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.541570   0.617037  -8.981  < 2e-16 ***
isSmoker     8.309929   1.023036   8.123 4.56e-16 ***
age          0.070911   0.008134   8.718  < 2e-16 ***
bmi          0.014278   0.015642   0.913    0.361    
children     0.118124   0.074942   1.576    0.115    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1664.97  on 1337  degrees of freedom
Residual deviance:  765.39  on 1333  degrees of freedom
AIC: 775.39

Number of Fisher Scoring iterations: 8
```

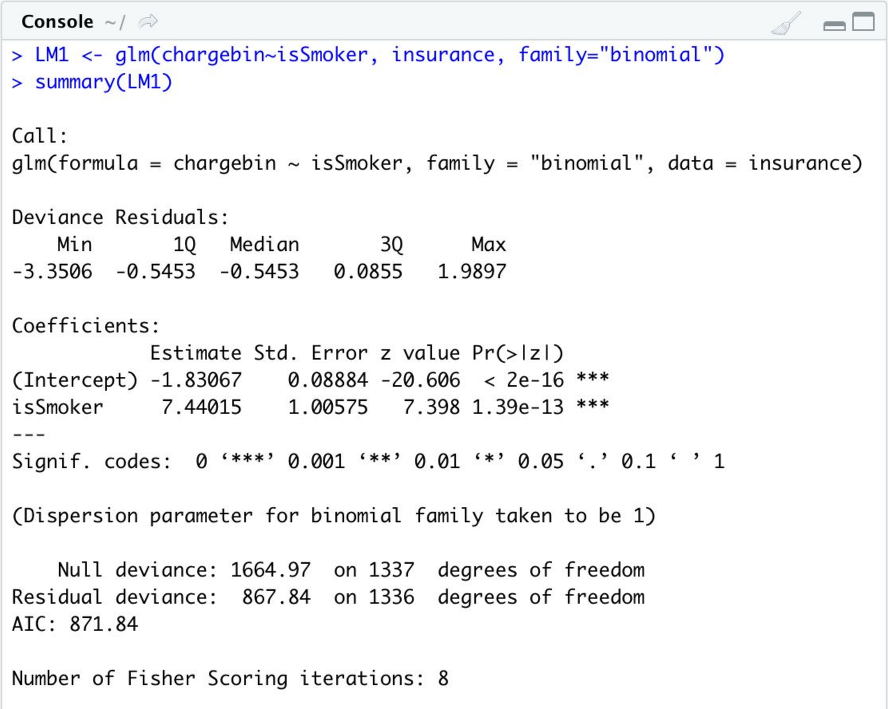
Analysis:

As can be seen above, the logistic regression that was enacted for the classification task gave interesting results. The only two independent variables that were significant at the 95% confidence interval were isSmoker and Age. The improved fit (Null Deviance - Residual Deviance) was significant at roughly 900. This tells us that with the addition of the independent variables taken into account with this model, we were able to better fit the data with the model significantly and thus have a better predictive feature. Unfortunately, I was unable to calculate the out of sample error for this model due to my use of multiple independent variables with a binary input. As you can see, the weights for the betas are significantly lower in this analysis versus the linear regression model and that is because we are aiming to place the input observation into an upper and a lower class using a binary dependent variable.

Tom Hollerbach

Variables used in logistic regression:

LM1: is Smoker



```
> LM1 <- glm(chargebin~isSmoker, insurance, family="binomial")
> summary(LM1)

Call:
glm(formula = chargebin ~ isSmoker, family = "binomial", data = insurance)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3506  -0.5453  -0.5453   0.0855   1.9897

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.83067    0.08884 -20.606  < 2e-16 ***
isSmoker     7.44015    1.00575   7.398 1.39e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1664.97  on 1337  degrees of freedom
Residual deviance:  867.84  on 1336  degrees of freedom
AIC: 871.84

Number of Fisher Scoring iterations: 8
```

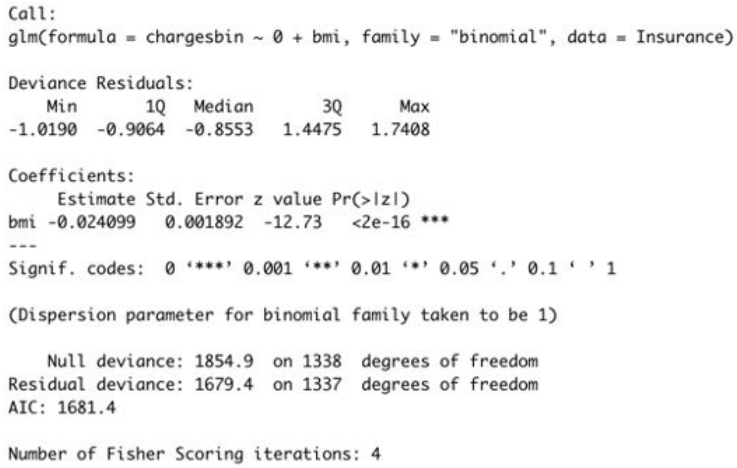
For the classification task, I decided to use logistic regression. Similar to the linear regression model, both parameters were significant above the 99% level and the overall improvement of the model by including the isSmoker variable was quite large, which can be seen by the null deviance being almost double the residual deviance. The beta value for is smoker was much smaller here because the output is binary ranking the data by charges into an upper or lower class.

Pedro Andrade

Model -> Logistic Regression
Variable: BMI

For the classification model, I used a logistic regression. In my classification model, the variable BMI appeared to be statistically significant with a p-value < 2e-16.However, the goodness of fit was quite low considering that there’s not a considerable difference between the null deviance and the residual deviance. Even though the variable is statistically significant, the effect of BMI on charges for this model is very low.
Bmi significant at the 95% confidence level (***)
Null deviance: 1854.9
Residual deviance: 1679.4

AIC: 1681.4

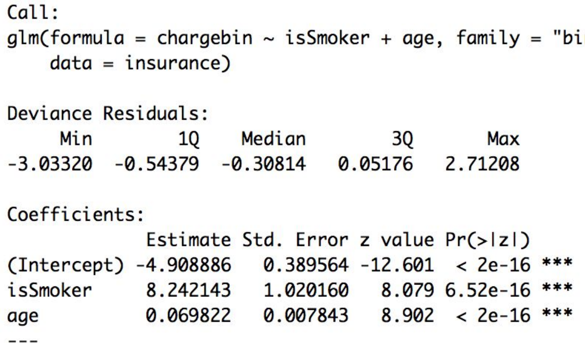


```
Call:
glm(formula = chargesbin ~ 0 + bmi, family = "binomial", data = Insurance)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0190  -0.9064  -0.8553   1.4475   1.7408

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
bmi -0.024099   0.001892  -12.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1854.9  on 1338  degrees of freedom
Residual deviance: 1679.4  on 1337  degrees of freedom
AIC: 1681.4

Number of Fisher Scoring iterations: 4
```

Taylor Danielson

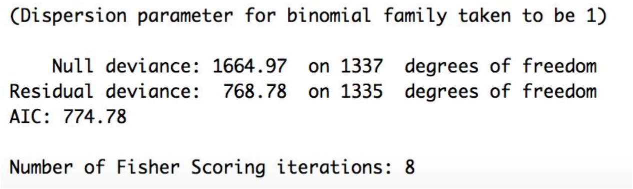
Model:
Logistic regression with variables isSmoker and age.

The variables I chose for this classification task are isSmoker and age. I chose these because I noticed that a lot of the variables that I had run previously in the regression task were not significant when I ran them through this classification model. So in order to decrease the out of sample error rates, I took out these variables, reducing the model complexity.



```
Call:
glm(formula = chargebin ~ isSmoker + age, family = "binomial",
    data = insurance)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.03320  -0.54379  -0.30814   0.05176   2.71208

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.908886   0.389564 -12.601  < 2e-16 ***
isSmoker     8.242143   1.020160   8.079 6.52e-16 ***
age          0.069822   0.007843   8.902  < 2e-16 ***
---
```

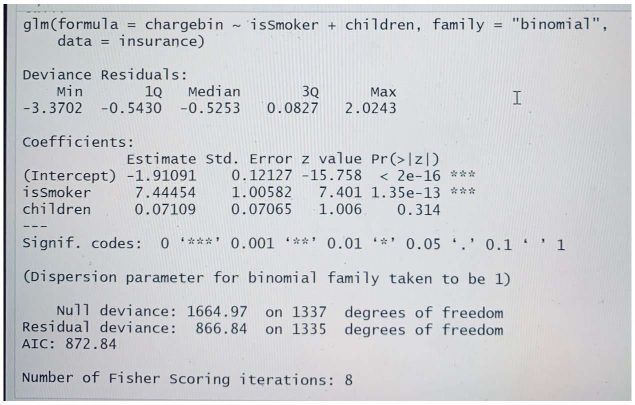
Now, all of the variables are significant at the 0.01% level. Additionally, I received a much higher Beta value for isSmoker, indicating that this variable has a larger impact on insurance charges on average.



```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1664.97  on 1337  degrees of freedom
Residual deviance:  768.78  on 1335  degrees of freedom
AIC: 774.78

Number of Fisher Scoring iterations: 8
```

The improved fit of the model was found by subtracting the residual deviance from the null deviance. This gave me an improved fit of 896.19. The AIC was 774.78 which is relatively high. This may indicate a poorer performance outside of the sample. Additionally, it took 8 iterations to reach the optima.

Chase Lester



```
glm(formula = chargebin ~ isSmoker + children, family = "binomial",
    data = insurance)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3702  -0.5430  -0.5253   0.0827   2.0243

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.91091    0.12127 -15.758  < 2e-16 ***
isSmoker     7.44454    1.00582   7.401 1.35e-13 ***
children     0.07109    0.07065   1.006    0.314
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1664.97  on 1337  degrees of freedom
Residual deviance:  866.84  on 1335  degrees of freedom
AIC: 872.84

Number of Fisher Scoring iterations: 8
```

For the classification task, I decided to use logistic regression. I kept the variables isSmoker and children the same. Unlike the linear regression model, children were not significant above the 99% level. However, isSmoker remained significant. The goodness of fit of my model was 797.03. The AIC was 872.84 which is very high. My model also took 8 fisher iterations to reach the optima.

Classification Task: Validation and Model Selection

Section 6

Carter Bradsky

Validating the logistic regression model was difficult to do because I was unable to calculate the out of sample error given the multiple independent variables with binary inputs into the model. Instead, focused on splitting the data into a 70% training dataset and 30% testing set, comparing the results with our initial regression and assessing the different Null and Residual deviances as well as the difference in AIC and Fisher iterations to reach the optima.

Tom Hollerbach

I was unable to calculate the out of sample error because the predict function cannot have a binary input and output.

Taylor Danielson

When validating this model, I was unable to partition the data into training and testing sets. And thus was unable to validate the classification model.

Chase Lester

I was not able to validate my data. I was unable to partition my data into training and testing sets. Therefore I was unable to validate the model.

Pedro Andrade

I was not able to partition the data into training and testing , and as result I was unable to validate my classification model.

Alejandro Romo

Independent Variable(s): Age, Charges
Goodness of Fit: 172.1
Significance: ***Age
Confidence Interval:(-1.6434, -1.1799)
The goodness of fit in this situation was the lowest as the r squared showed the minimal accuracy out of all the models we tested. The predict function wasn’t available for my model so I couldn't get a sample error.

Conclusion

Section 6

When selecting a final model we wanted to choose one that captured the effects that multiple variables had on charges. We concluded that Carter's models did this best with a high R-squared (74.97%) and joint model statistical significance at 95% (p-value = 2.2e-16). In our models, we included all regions as independent variables and never found any of them to be statistically significant, so we decided to exclude them from our final model.

Independent variables we used in this model include:

isSmoker (***), Age (***), BMI (***), Children (***), isFemale (not significant)

We used the same variables for the classification model which used logistic regression to classify individuals into an upper and lower class based on the amount they are charged for health insurance. In this model, we had similar findings except BMI was no longer statistically significant. Not to our surprise, whether one is a smoker or not and how many children an individual has had the greatest impact on health insurance costs and had the greatest impact on whether or not an individual would be classified into the upper or lower class.