Final Project

Feifei Li, Izzy Mika
Group 5
December 1st, 2025
STAT 311 Regression Analysis, Fall 2025

Modeling Health Insurance Charges: The Role of Smoking, BMI, and Demographic Factors

1. Introduction

1.1 Research Question

How do demographic and lifestyle factors influence annual medical insurance charges, and does the effect of BMI differ between smokers and non-smokers?

1.2 Significance

Understanding these drivers assists insurers, policymakers, and consumers.

1.3 Approach

Data screening, EDA, stepwise screening, model building (three models), diagnostics, comparison, and final model selection.

2. Data Description

The dataset “Health Insurance Charges” includes 1,338 observations sourced from Kaggle (link: https://www.kaggle.com/datasets/nalisha/health-insurance-charges-dataset), reduced to 1,335 after cleaning. Variables: age, sex, bmi, children, smoker, region, charges. Sex and region are not used for analysis, due to a low impact on a model’s ability to predict results (based on a stepwise model). We created an indicator variable for Smoker that indicates whether the insured individual smokes (baseline = “Yes”). Charges are simulated to reflect realistic U.S. healthcare cost patterns. We created a column for a log transformation of Charges (transformY) to account

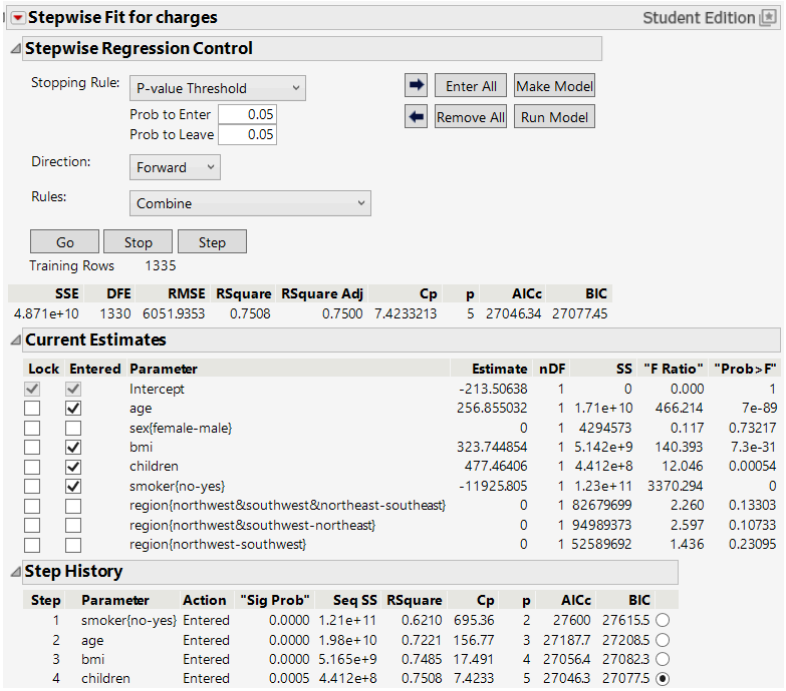
for outliers in the data and for the right-skewness of Charges' distribution.

3. Exploratory Data Analysis

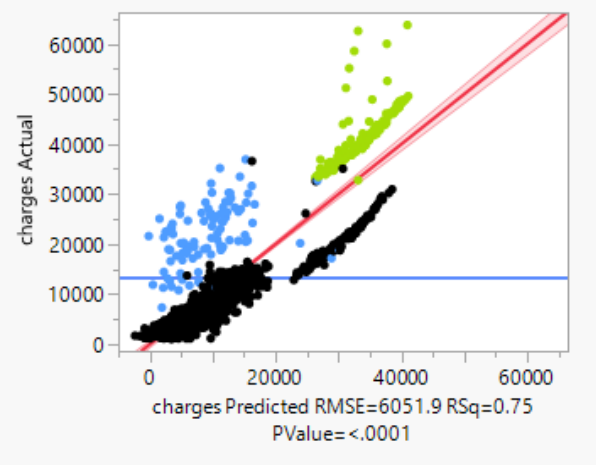
In our exploratory data analysis, we found that Charges (y) has a heavily right-skewed distribution, and that people who smoke see much higher costs, but folks with BMI have a less obvious jump unless you graph both factors in combination – in which case, there's a significant jump when an individual both smokes and has a BMI >= 30 (which is classified as obese). This jump also partly accounts for one of the groups of data points in Actual by Predicted plots (marked green) and its slight difference in behavior.

In a stepwise model, we found that Sex and Region are not significant in a First Order model, and thus removed it from our analysis.

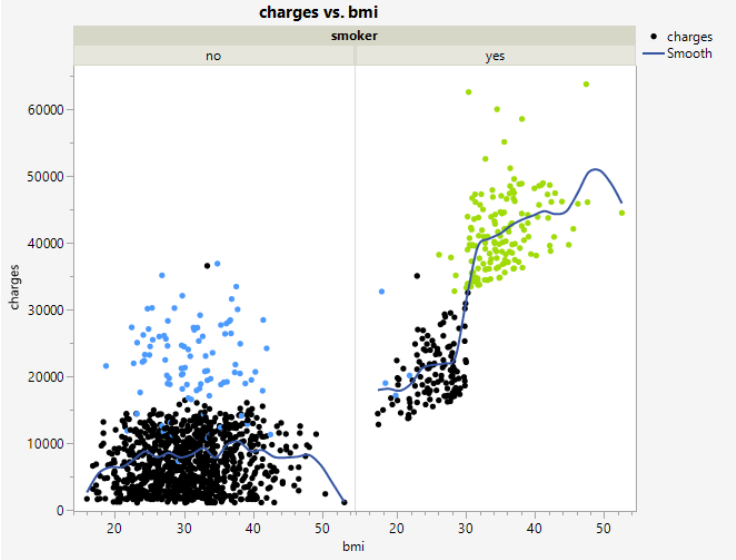
Stepwise Model – P-value Threshold to Enter & Leave = 0.05



When I was analyzing the data and models, I noticed there were different groups of data points – two of which consistently don't follow the fitted line as much as the main group of data points. I marked one blue and one green as a point of reference. Based on the scatterplot graphs, and the way that it otherwise



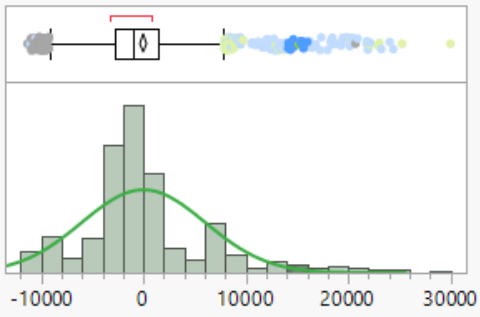
matches a similar trend as the black group of data points, I found that green is likely the result of health insurance treating the combination of BMI >= 30 and smoking as high-risk and increasing the costs in a less continuous way than usually expected. I tried to narrow down what was causing the blue group's behavior, but wasn't able to find something definitive. I included a few models that are grouped by color in my github repository, and a couple scatterplot graphs that combined the factors BMI, smoker, and number of children (which seems to create a more linear regression, but I'm not certain this is useful in this case) for reference.



The scatterplot of Charges vs BMI & Smoker shows that there's a jump when an individual both smokes and has a BMI >= 30. Suggests health insurance categorizes people who both smoke and have a BMI >= 30 as higher risk category, and auto-bumps their cost.

Distributions

Residual Charges



Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.8998445	<.0001*

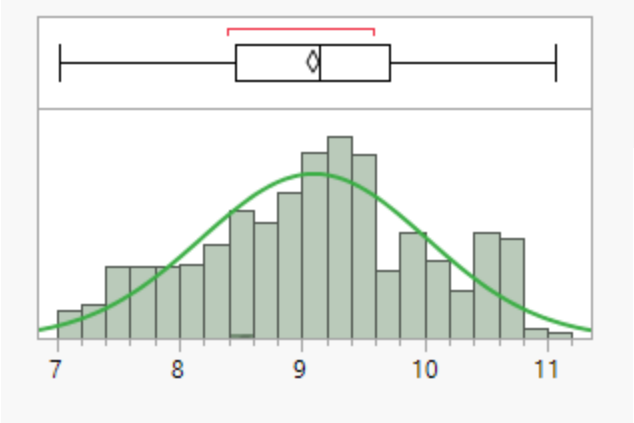
Quantiles

100.0%	maximum	29970.927
99.5%		22955.467
97.5%		16409.285
90.0%		7481.1648
75.0%	quartile	1413.7926
50.0%	median	-961.6472
25.0%	quartile	-2835.767
10.0%		-6592.646
2.5%		-10153.61
0.5%		-10905.69
0.0%	minimum	-11330.8

Summary Statistics

Mean	-8.77e-13
Std Dev	6028.2719
Std Err Mean	164.98794
Upper 95% Mean	323.66408
Lower 95% Mean	-323.6641
N	1335
N Missing	0

transformY



Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.983309	<.0001*

Shows an improvement in distribution / reduction in skewness or outliers, but Shapiro-Wilk test still shows that the data does not have a Normal Distribution.

4. Methods

Study Design: Multiple linear regression with model refinement.

Models:

Model 1: Main Effects

Model 1 is a First Order model that shows the main effects of an individual's demographics. It assumes there's no interaction between any of the variables, and that Region and Sex are not significant predictors to the accuracy of the model (based on a stepwise model). This model is used as the baseline for Models 2 and 3.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Where:

x_1 = Age

x_2 = BMI

x_3 = Children

x_4 = { 1, If Individual Does Not Smoke [no]

{ 0, if o/w [yes]

Model 2: Add BMI × Smoker Interaction

Model 2 adds to Model 1 by including the interaction between BMI and Smoker. This can demonstrate whether BMI's impact on healthcare costs (Charges) significantly changes based on whether the insured individual also smokes, or vice versa.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_2 x_4$$

Where:

x_1 = Age

x_2 = BMI

x_3 = Children

x_4 = { 1, If Individual Does Not Smoke [no]

{ 0, if o/w [yes]

Model 3: Transform Y - ln(Charges)

Model 3 has the same design as Model 2, but normalizes the dependent variable Charges with a log transformation (transformY), to account for the right-skewed distribution of Charges. Based on Model 1's Box-Cox Plot (Best λ=0.156 - Closest to 0), transformY = ln(Charges).

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_2 x_4$$

Where:

x_1 = Age

x_2 = BMI

x_3 = Children

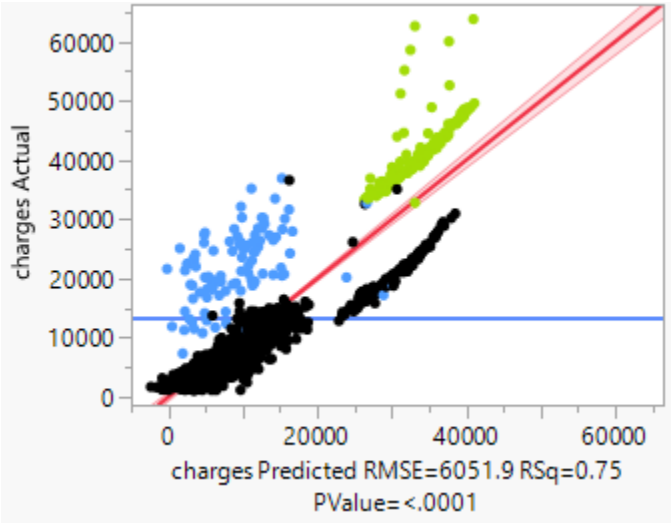
x_4 = { 1, If Individual Does Not Smoke [no]

{ 0, if o/w [yes]

Diagnostics included residual analysis, Box–Cox, global F-tests, partial F-tests, and validation metrics.

5. Results

Model 1: Main Effects - Statistically useful but fails diagnostics.



Summary of Fit	
RSquare	0.750763
RSquare Adj	0.750013
Root Mean Square Error	6051.935
Mean of Response	13255.38
Observations (or Sum Wgts)	1335

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	1.4673e+11	3.668e+10	1001.569
Error	1330	4.8712e+10	36625921	Prob > F
C. Total	1334	1.9545e+11		<.0001*

Global F-Test:

(p-value: See [Prob > F] under Analysis of Variance Table)

H_0: $\beta_1 = \beta_2 = \cdots = \beta_4 = 0$ *vs*

H_a: *At least one of the* $\beta_i \neq 0$

a = 0.05

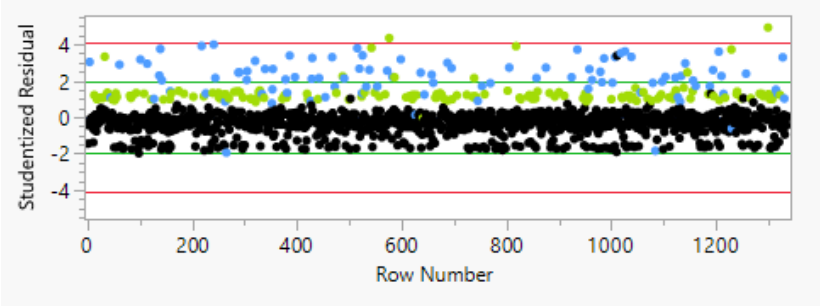
p < 0.0001

With a significance level of 5%, the Global F-Test shows strong evidence that at least one coefficient is not zero (F = 1001.569, p < 0.0001 < a = 0.05). This shows that the overall accuracy of the model is statistically significant in predicting insurance costs. However, R^2_{adj} shows that only 75% of the predicted results are accounted for by Model 1, which is low. Additionally, the Q-Q plot and Box-Cox Plot shows evidence that the assumption that ε is normally distributed and the assumption of constant variance is violated (see below). Since λ = 0.156 (closest to 0), we'll use the log transformation ln(Charges) for the Distribution of transformY and for Model 3.

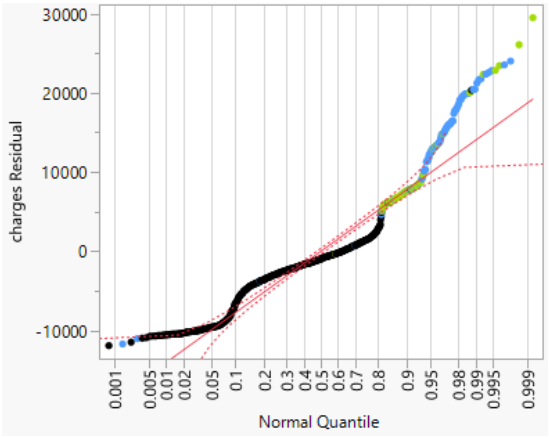
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-213.5064	942.4549	-0.23	0.8208
age	256.85503	11.89585	21.59	<.0001*
bmi	323.74485	27.32314	11.85	<.0001*
children	477.46406	137.5682	3.47	0.0005*
smoker[no]	-11925.8	205.4252	-58.05	<.0001*

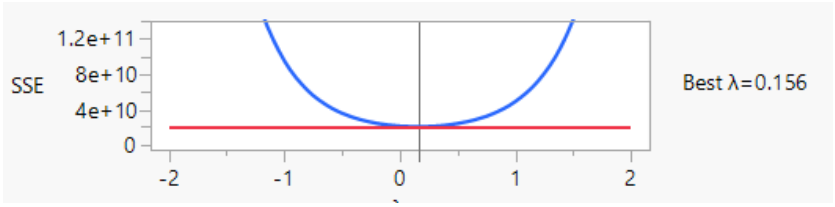
Fitted Model:

$E(y) = -213.5064 + 256.8550x_1 + 323.7449x_2 + 477.4641x_3 - 11925.80x_4$

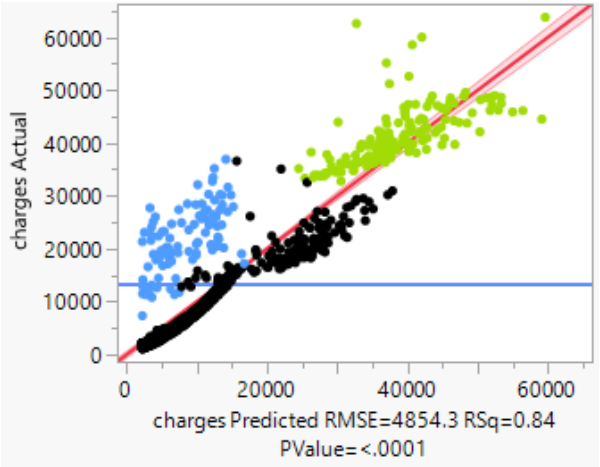


Studentized Residuals show that there are not many extreme outliers, but that there isn't constant variance (more data points above fitted line than below). Notably, most or all of the dots causing this behavior are from the two groups of data points (green and blue) (in Actual by Predicted Plot) that are farther from the fitted line and show a slightly different trend than the main group of data points.





Model 2: Interaction significant; improved structure.



Summary of Fit	
RSquare	0.839766
RSquare Adj	0.839164
Root Mean Square Error	4854.314
Mean of Response	13255.38
Observations (or Sum Wgts)	1335

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	1.6413e+11	3.283e+10	1393.028
Error	1329	3.1317e+10	23564363	Prob > F
C. Total	1334	1.9545e+11		<.0001*

Global F-Test:

(p-value: See [Prob > F] under Analysis of Variance Table)

H_0: $\beta_1 = \beta_2 = \cdots = \beta_5 = 0$ *vs*

H_a: *At least one of the* $\beta_i \neq 0$

a = 0.05

p < 0.0001

Partial F-Test:

H_0: $\beta_5 = 0$ *vs*

H_1: $\beta_5 \neq 0$

Custom Test	
Parameter	
Intercept	0
age	0
bmi	0
children	0
smoker[no]	0
bmi*smoker[no]	1
=	0
Value	-715.3352394
Std Error	26.328117836
t Ratio	-27.17001055
Prob>|t|	1.1816e-129
SS	17395436113

Sum of Squares	17395436113
Numerator DF	1
F Ratio	738.20947332
Prob > F	1.1816e-129

a = 0.05

p = 1.18E-129

F Ratio = 738.20947332

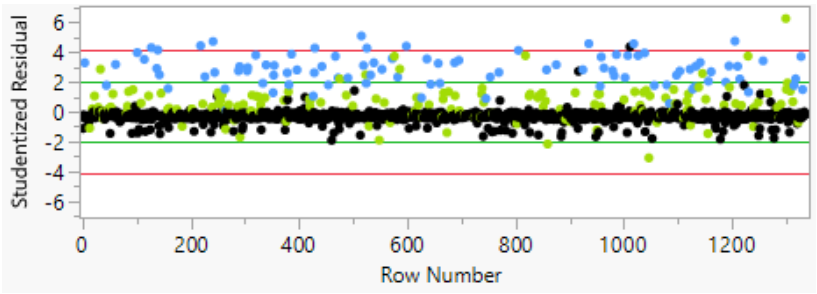
The Global F-Test shows strong evidence that at least one coefficient is not zero (F = 1393.028, p < 0.0001 < a = 0.05). The Partial F-Test also shows strong evidence that including the interaction of [BMIxSmoker] significantly improves the accuracy of the model. The interaction plots (see on following page) also show strong interaction between BMI and Smoker. RSquare Adj shows that 83.92% of results can be predicted by this model, which is an improvement from Model 1. The Root Mean Square Error (4854.314) is also an improvement. However, there is still differing behavior between the different groups of data points, and the Studentized Residual plot and Q-Q Plots show signs that the Assumption of Constant Variance is violated.

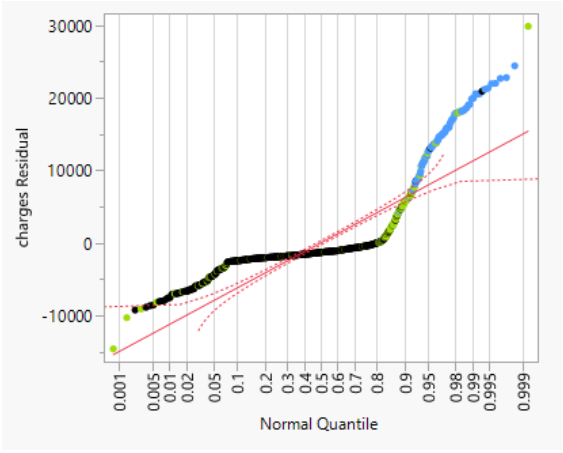
Parameter Estimates

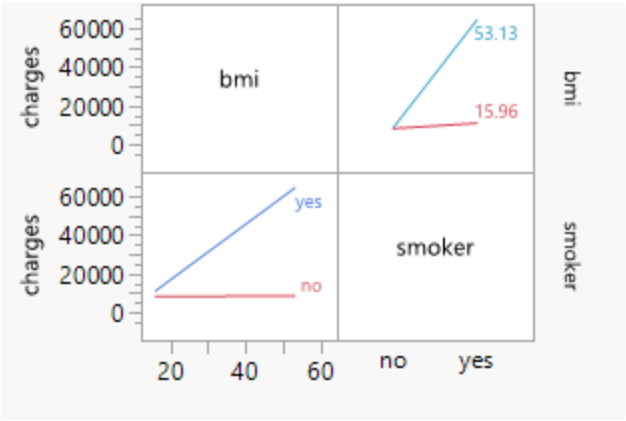
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-12815.19	886.8944	-14.45	<.0001*
age	264.00987	9.545404	27.66	<.0001*
bmi	723.35009	26.39376	27.41	<.0001*
children	513.60664	110.3528	4.65	<.0001*
smoker[no]	10032.646	824.8131	12.16	<.0001*
bmi*smoker[no]	-715.3352	26.32812	-27.17	<.0001*

Fitted Model:

E(y) = $-12815.19 + 264.00987x_1 + 723.35009x_2 + 513.60664x_3 + 10032.646x_4 - 715.3352x_2x_4$

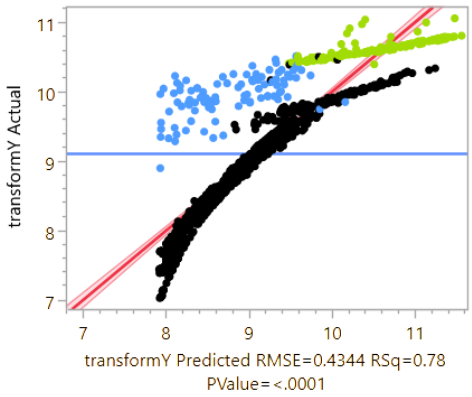






Model 3: Best overall

log transformation improves constant variance, improves normality, maintains interaction significance.



Summary of Fit	
RSquare	0.777158
RSquare Adj	0.77632
Root Mean Square Error	0.434356
Mean of Response	9.098084
Observations (or Sum Wgts)	1335

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	874.4412	174.888	926.9757
Error	1329	250.7363	0.189	Prob > F
C. Total	1334	1125.1776		<.0001*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	874.4412	174.888	926.9757
Error	1329	250.7363	0.189	**Prob > F**
C. Total	1334	1125.1776		<.0001*

Global F-Test:

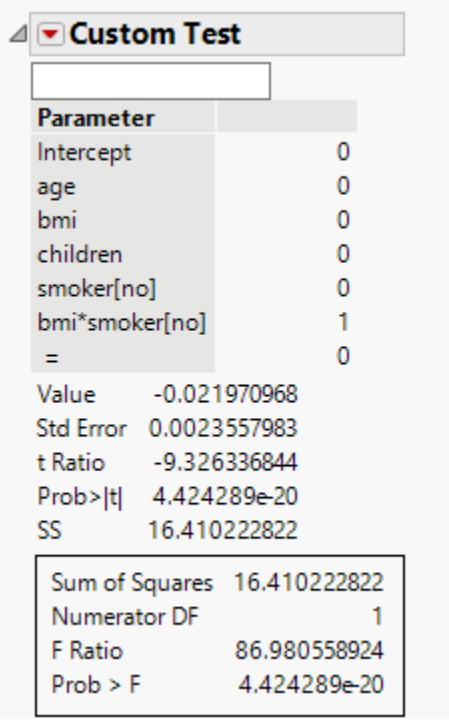
(p-value: See [Prob > F] under Analysis of Variance Table)

H_0: $\beta_1 = \beta_2 = \cdots = \beta_5 = 0$ *vs*

H_a: *At least one of the* $\beta_i \neq 0$

a = 0.05

p < 0.0001



Partial F-Test:

H_0: $\beta_5 = 0$ *vs*

H_1: $\beta_5 \neq 0$

a = 0.05

p = 4.424289e-20

The Global F-Test and Partial F-Test both show strong evidence that Model 3 is overall statistically significant in predicting the health insurance costs an individual has, and that it's an improvement from Model 2 due to the log transformation of Y (Charges).However, the RSquare Adjusted only covers 77.63% of predicted results, which is somewhat low and lower than Model 2, and there is a clear curvature in the Predicted by Actual Plot. I believe this is partly due to the fact that the groups of data

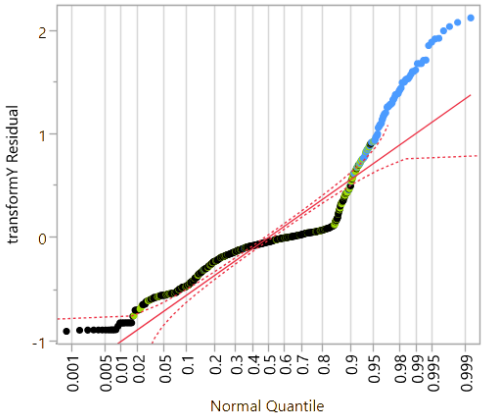
have slightly different behavior, and likely need different Y transformations and/or need to account for additional or different interactions. If you look past the end of this report, I've included some of the Predicted by Actual plots and data when you separate the Model results by the group of data points. The main group is much more accurate by itself (~97%) for Model 3, but the other 2 are actually worse than before the Y transformation (and are lower even without the transformation).

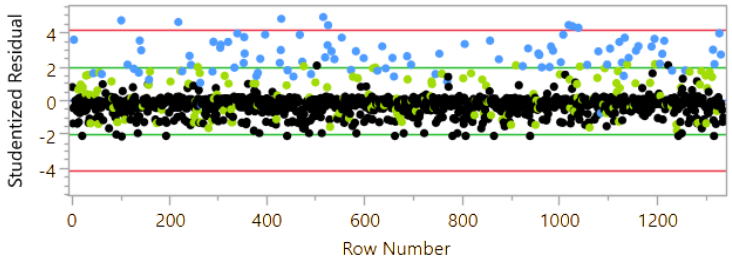
Parameter Estimates

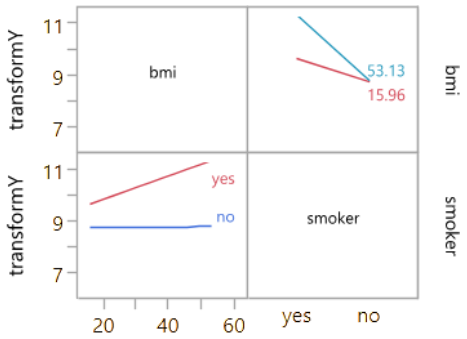
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.3669774	0.079358	92.83	<.0001*
age	0.0349422	0.000854	40.91	<.0001*
bmi	0.0230148	0.002362	9.75	<.0001*
children	0.1021047	0.009874	10.34	<.0001*
smoker[no]	-0.098531	0.073803	-1.34	0.1821
bmi*smoker[no]	-0.021971	0.002356	-9.33	<.0001*

Fitted Model:

$E(y) = 7.3669774 + 0.0349422x_1 + 0.0230148x_2 + 0.1021047x_3 - 0.098531x_4 - 0.021971x_2x_4$





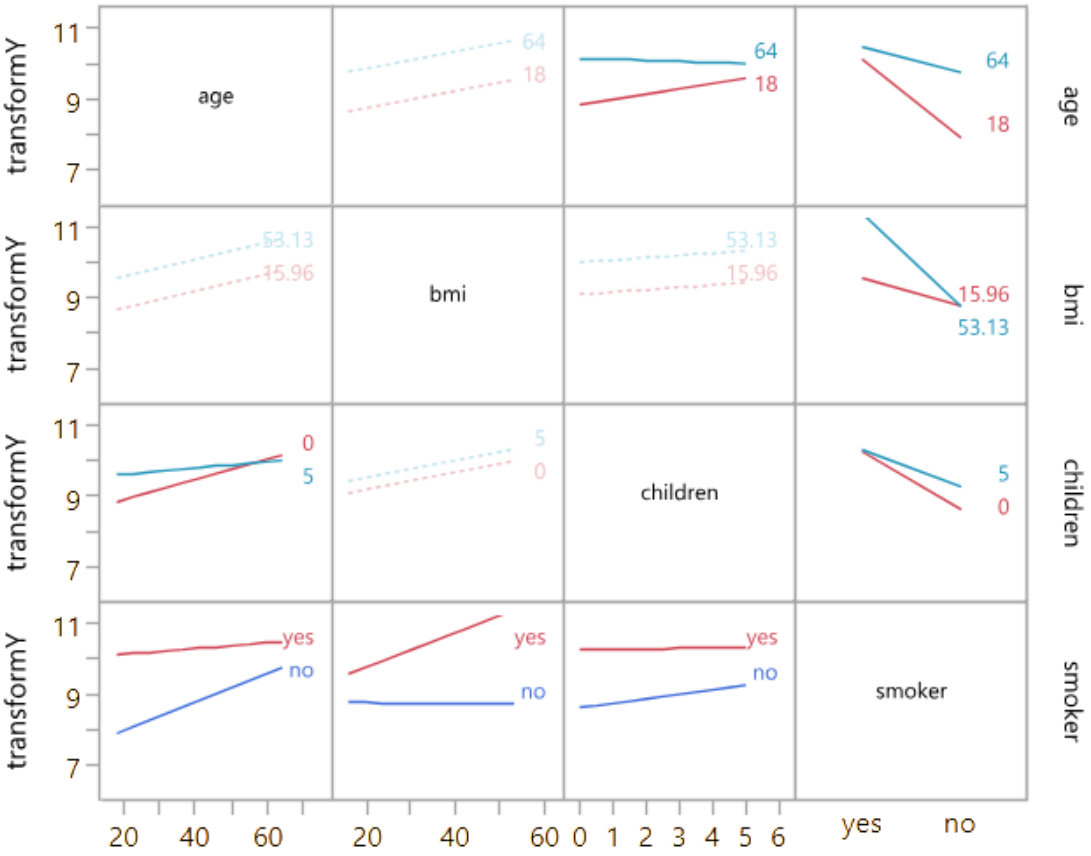


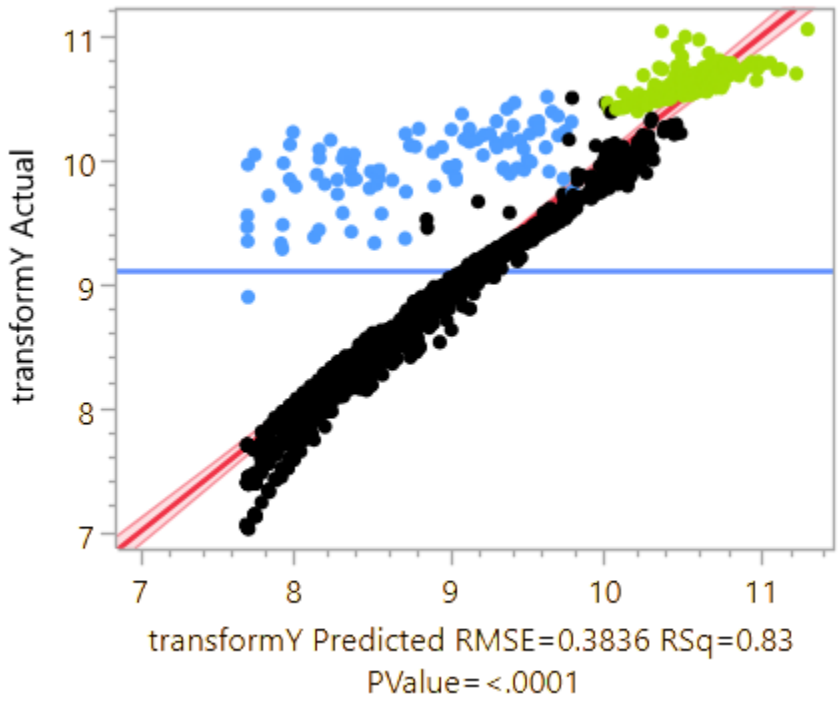
Model 3 chosen as final model for inference and prediction.

6. Conclusions and Discussion

Smoking drives the largest increases in healthcare charges. BMI strongly affects charges for smokers but minimally for non-smokers. Log transformation improves model adequacy. Final model meaningful for actuarial and policy use.

Overall Model 3 is the best of the options we initially planned, but I think if we were going to continue research, we would look into age x ___ interactions (see interaction plot below from a potential Model 4 I worked on - shows strong signs that Age impacts other factors), and see if splitting the groups into their own Models with their own Y transformations would improve results.


transformY
age
bmi
children
smoker
64
18
53.13
15.96
0
5
yes
no
20
40
60
0 1 2 3 4 5 6


transformY Actual
transformY Predicted RMSE=0.3836 RSq=0.83
PValue=<.0001
7
8
9
10
11

bmi*smoker[yes]

age*smoker[yes]

children*smoker[yes]

age*children

RSquare Adj	0.825497

7. Appendix

See github repository for exploratory data analysis and other plots.

https://github.com/izzyvmika/STAT311-50_Final-Project.git

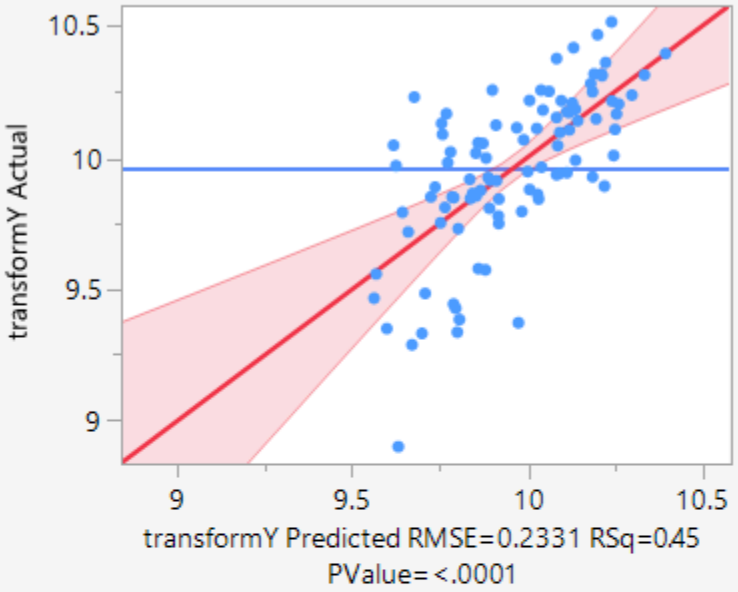
8. References

Metropolitan State University. (2025). STAT 311: Regression Analysis (Fall 2025).

Kaggle. (2025). Health Insurance Charges Dataset.

Color (Group)=37

Actual by Predicted Plot



Effect Summary

Summary of Fit

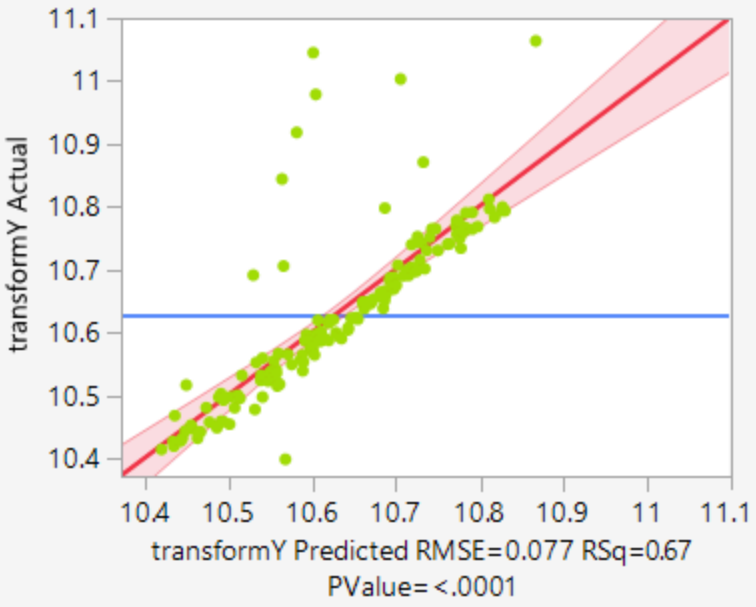
RSquare	0.452065
RSquare Adj	0.402253
Root Mean Square Error	0.233115
Mean of Response	9.960072
Observations (or Sum Wgts)	97

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	3.9454517	0.493181	9.0754
Error	88	4.7821651	0.054343	Prob > F
C. Total	96	8.7276169		<.0001*

Color (Group)=44

Actual by Predicted Plot



Effect Summary

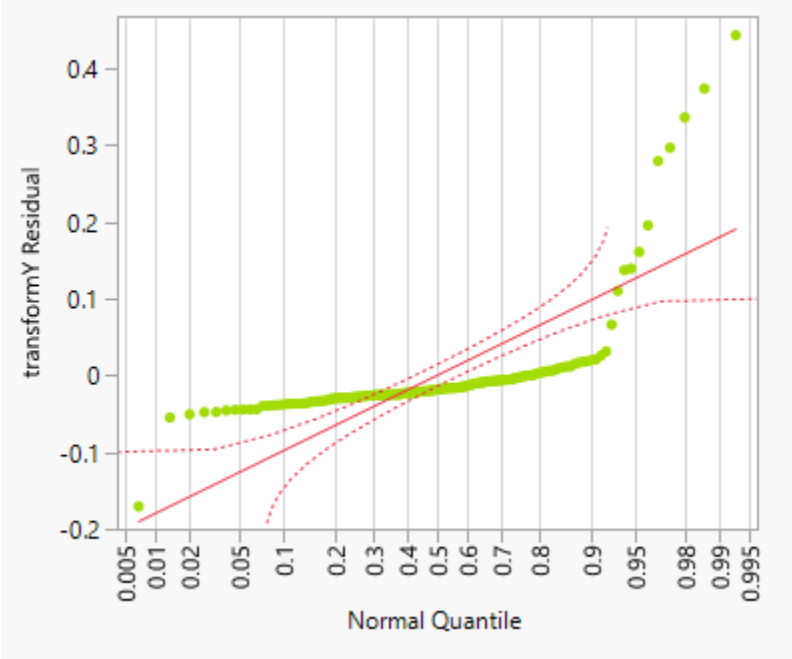
Summary of Fit

RSquare	0.674002
RSquare Adj	0.664819
Root Mean Square Error	0.077036
Mean of Response	10.62681
Observations (or Sum Wgts)	147

Analysis of Variance

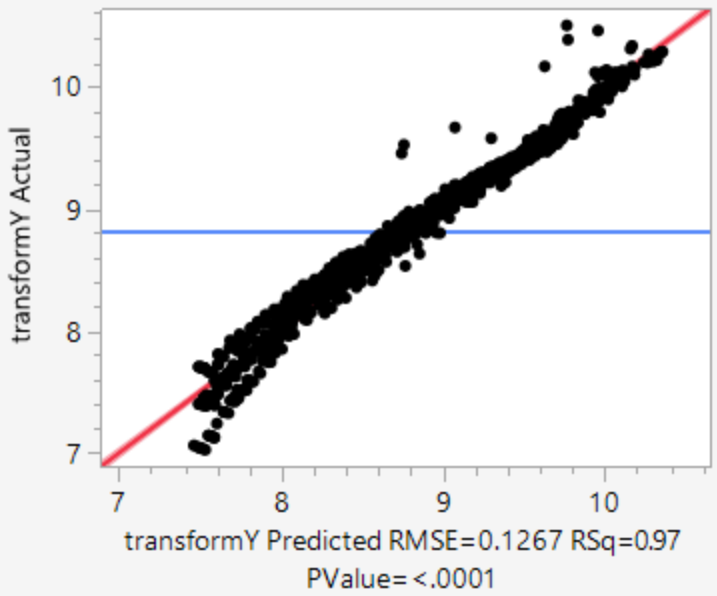
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	1.7422777	0.435569	73.3962
Error	142	0.8426979	0.005934	Prob > F
C. Total	146	2.5849756		<.0001*


Studentized Residual
5
0
-5
0
50
100
Row Number


transformY Residual
0.4
0.3
0.2
0.1
0
-0.1
-0.2
0.005
0.01
0.02
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.98
0.99
0.995
Normal Quantile

Color (Group)=0

Actual by Predicted Plot



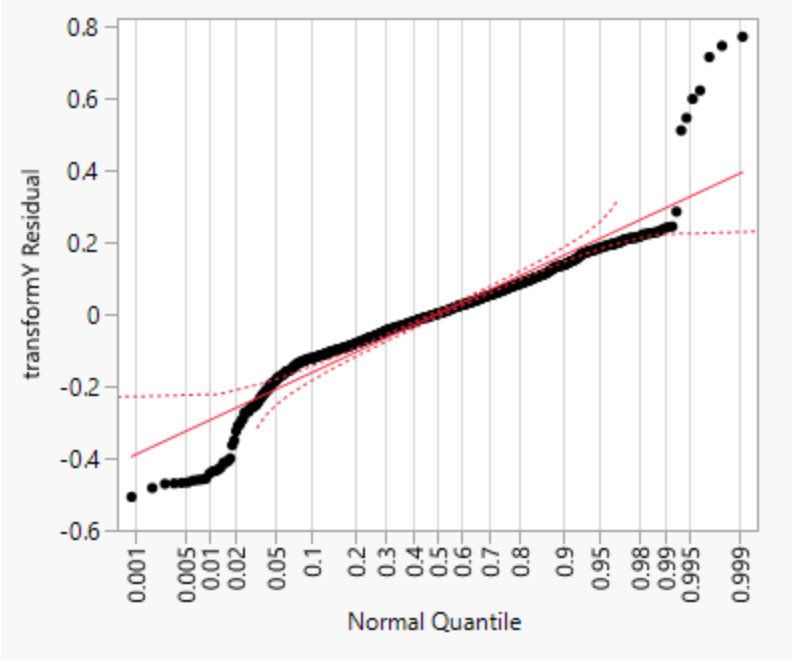
Effect Summary

Summary of Fit

RSquare	0.971575
RSquare Adj	0.971365
Root Mean Square Error	0.126705
Mean of Response	8.815467
Observations (or Sum Wgts)	1091

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	593.73981	74.2175	4622.921
Error	1082	17.37069	0.0161	Prob > F
C. Total	1090	611.11050		<.0001*


transformY Residual
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0.001
0.005
0.01
0.02
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.98
0.99
0.995
0.999
Normal Quantile