STAT 311 – Final Project Proposal

Title: Impact of Demographic and lifestyle Factors on health insurance charges

Team Members: Feifei Li & Izzy Mika

Course: STAT 311 Regression Analysis, Fall 2025

Proposal : November 15, 2025

1. Abstract

This project investigates how demographic and lifestyle variables influence annual U.S. health insurance charges. Using the Kaggle Health Insurance dataset (n = 1,338), we will evaluate how age, BMI, smoking status, number of children, sex, and geographic region contribute to medical cost variation. Our primary focus is whether the effect of BMI on health insurance charges differs between smokers and non-smokers, analyzed through a BMI × smoker interaction term.

We will build three regression models: (1) a main-effects base model, (2) a model including the interaction, and (3) a log-transformed charges model if diagnostics show heteroscedasticity. Each model will be evaluated using partial F-tests, adjusted R^2, and full residual diagnostics per Chapter 8. The final model will be selected based on statistical validity and interpretability.

2. Dataset and Source

We use the publicly available:

Kaggle: Health Insurance Charges dataset

https://www.kaggle.com/datasets/nalisha/health-insurance-charges-dataset

Dataset Summary

Observations: 1,338

Variables: 7 (mix of numeric and categorical)

Dependent Variable: `charges` — annual medical insurance cost

Variables Included

Variable	Type	Description
age	Numeric	Age of individual
sex	Categorical	Male, Female
bmi	Numeric	Body Mass Index
children	Numeric	Number of dependents
smoker	Categorical	Yes, No
region	Categorical	Northeast, Southeast, Southwest, Northwest
charges	Numeric	Annual insurance cost (response variable)

Potential Challenges

Only three values are missing (one each in age, BMI, and children), representing less than 0.3% of the dataset, so we will simply remove these rows to ensure a clean analysis.

Challenges include strong right-skewness in insurance charges, potential high-leverage points, and possible heteroscedasticity, all of which may require a log transformation and thorough diagnostics. Additional considerations include proper dummy coding of categorical variables and checking for multicollinearity after adding the BMI × smoker interaction term.

3. Research Questions

Primary Research Question

Does smoking modify the effect of BMI on health insurance charges?

→ Addressed through testing a **BMI × smoker** interaction.

Secondary Questions

1. What are the main effects of age, BMI, children, sex, smoker, and region?

2. Does adding an interaction significantly improve model fit (Partial F-test)?

3. Do we need a log(charges) transformation to satisfy linear model assumptions?

4. Which model provides the best fit?

4. Analytical Framework

We will fit three nested models that directly support hypothesis testing and model diagnostics.

Model 1 — Base Main-Effects Model

A foundational model including:

age, sex, BMI, children, smoker, and region dummy variables.

$$charges=\beta0 +\beta1\ (age)+\beta2\ (sex)+\beta3\ (bmi)+\beta4\ (children)+\beta5\ (smoker)+\beta6\ (NE)+\beta7\ (SE)+\beta8\ (SW)+\varepsilon$$

Purpose:

Establish core predictors

Provide baseline for model comparison

Evaluate coefficient significance with t-tests and overall usefulness via Global F-test

Model 2 — Interaction Model (Primary Hypothesis)

charges = Model 1+β(bmi×smoker)

Purpose:

Test if BMI’s impact differs for smoker's vs non-smokers

Use Partial F-test to compare Model 1 vs Model 2

Interpret interaction:

For nonsmokers: slope = β_bmi

For smokers: slope = β_bmi + β_interaction

Model 3 — Log-Transformed Response

Used only if residual diagnostics show clear violations.

Reason for log transformation:

- Charges are strongly right-skewed

- Heteroscedasticity (funnel pattern) is common

- Stabilizes variance and improves normality

- Coefficients become interpretable as % changes

5. Exploratory Data Analysis (EDA)

We will conduct a comprehensive EDA including:

Graphical EDA (Required by Rubric)

- **Histogram of charges (observe heavy skew)**
- **Boxplots:**

  charges vs smoker

  charges vs sex

  charges vs region
- **Scatterplots:**

  charges vs age

  charges vs bmi

  charges vs BMI, colored by smoker
- **Correlation matrix for quantitative predictors**

Summary statistics

- Mean, SD, quartiles, ranges

- Identify potential outliers or leverage points

6. Model Assumptions and Diagnostics

For each model, we will evaluate:

Residual Diagnostics

- **Residual vs fitted plot** (linearity + equal variance)
- **Residuals vs each predictor**
- **Normal probability plot (QQ plot)**
- **Shapiro–Wilk test** (supporting evidence only)

- **Influence statistics** (Cook's D, leverage, studentized residuals)

Model Comparison

- Adjusted R^2
- MSE / RMSE
- Partial F-test (Model 1 vs Model 2)
- Evaluate whether log transformation resolves heteroscedasticity

7. Data Preparation Plan

1. Remove three rows with missing values
2. Encode categorical variables using JMP indicator/dummy variables
3. Create interaction variable `bmi × smoker`
4. Identify outliers or high leverage points
5. Evaluate whether `charges` needs log transformation
6. No unnecessary new variables (keep model parsimonious)

8. Expected Outcomes

We expect:

Age, BMI, and smoking to be strong positive predictors

Smokers to have dramatically higher charges

BMI × smoker interaction to be statistically significant

Smokers will show a steeper BMI → charges slope

Log-transformed model to perform better in diagnostics

Final model likely to be log(charges) with interaction

References (APA Style)

Course Material

Metropolitan State University. (2025). *STAT 311: Regression Analysis (Fall 2025)*. Department of Mathematics & Statistics.

Dataset

Nalisha. (n.d.). *Health insurance charges dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/nalisha/health-insurance-charges-dataset

Data File Used

insurance.csv extracted from the Kaggle dataset listed above.

Workflow Chart

Date Range	Phase	Tasks	Responsibility
Nov 11–14	**Data Cleaning & EDA**	• Handle missing data • Histograms, boxplots, scatterplots • Summary statistics • Early outlier preview	**Both (Together)**
Nov 15–17	**Model 1: Main Effects**	• Both run Model 1 in JMP • Compare parameter estimates • Check VIF • Review residuals side-by-side	**Both**
Nov 17–19	**Model 2: BMI × Smoker Interaction**	• Both run Model 2 • Compare outputs to ensure identical results • Interpret interaction • Conduct Partial F-Test together	**Both**

Nov 20–22	**Diagnostics & Model 3**	• Check QQ plots, Cook's Distance, residuals • Identify any influential observations • Fit log(Y) model if needed (both run it) • Compare Models 1, 2, 3 together	**Both**
Nov 23–25	**Report Drafting**	**Feifei:** Intro, Interpretation, Discussion **Izzy:** Model building, diagnostics methods • Cross-review each other's sections	**Feifei + Izzy (split writing + joint review)**
Nov 26–27	**Final Report Polishing**	• Merge sections • Final edits and formatting • Fix tables & figure captions • Export PDF	**Both**
Nov 28–29	**Slide Deck Creation**	work together	**Both (different roles)**
Nov 30–Dec 1	**Presentation Practice**	• Practice #1 • Slide edits • Practice #2	**Both**
Dec 2	**Final Review**	• Light rehearsal • No new work added	**Both**