Regression Analysis: Predicting Medical Expenses

By Ambati Yaswanth & Rajeev Kumar

April 19, 2025

Introduction

- Understanding the factors that influence healthcare care costs is crucial for:
  - Insurance companies
  - Healthcare providers
  - Policymakers
  - Individuals
- Primary factors examined: smoking status, BMI, sex, and age.

Research Objectives

- Estimate future healthcare costs based on individual characteristics.
- Identify and quantify key risk factors that affect medical expenses.
- Provide insights for targeted cost-management interventions.

Data Description

Variable	Description
age	Age of the individual
sex	Gender (0=Female, 1=Male)
bmi	Body Mass Index
smoker	Smoking status (0=Non-smoker, 1=Smoker)
expenses	Medical expenses in USD (target variable)

age	sex	bmi	smoker	expenses
19	0	27.9	1	16884.92
18	1	33.8	0	1725.55
28	1	33.0	0	4449.46
33	1	22.7	0	21984.47
32	1	28.9	0	3866.86

Table: Sample of the dataset

Methodology

- **Data Preprocessing**
  - Encoding categorical variables
  - Data cleaning and outlier removal
  - Feature selection
- **Exploratory Data Analysis(EDA)**
  - Descriptive statistics
  - Correlation analysis
- **Statistical Modeling**
  - Multiple linear regression
  - Model evaluation metrics

Multiple Linear Regression Model

The general form of our regression model:

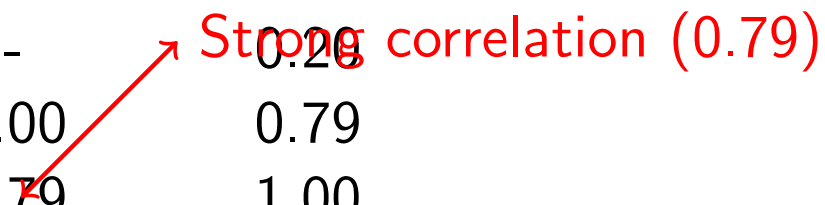
$$\text{Expenses} = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Sex} + \beta_3 \times \text{BMI} + \beta_4 \times \text{Smoker} + \varepsilon \quad (1)$$

- β_0 is the intercept (constant term)
- $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients for the predictors
- ε represents the error term

Model Evaluation Metrics	**Formula**
R^2 (Coefficient of Determination)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
MSE (Mean Squared Error)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Correlation Analysis

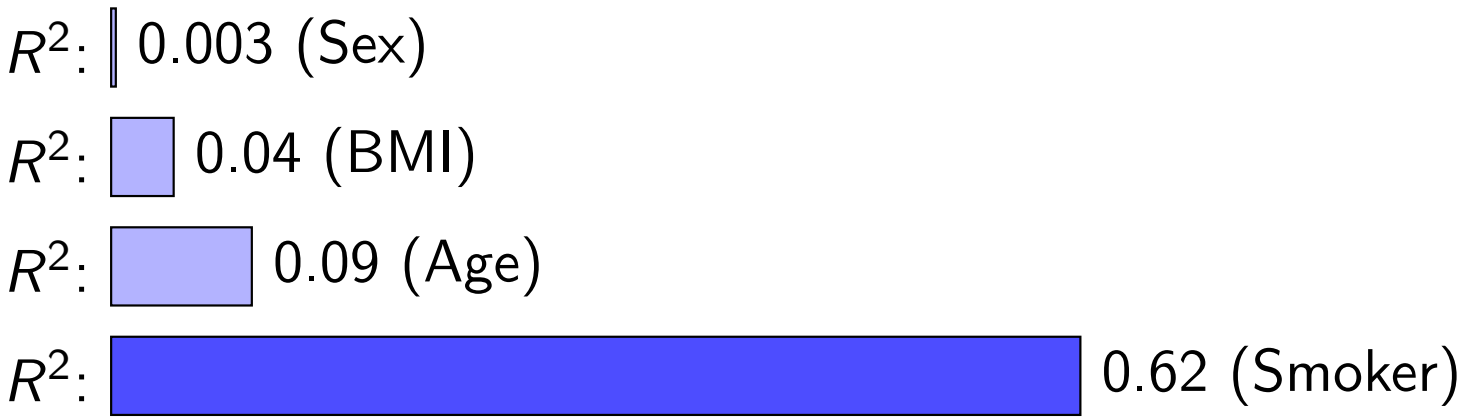
	Age	Sex	BMI	Smoker	Expenses
Age	1.00	-	-	-	0.30
Sex	-	1.00	-	-	0.06
BMI	-	-	1.00	-	0.20
Smoker	-	-	-	1.00	0.79
Expenses	0.30	0.06	0.20	0.79	1.00



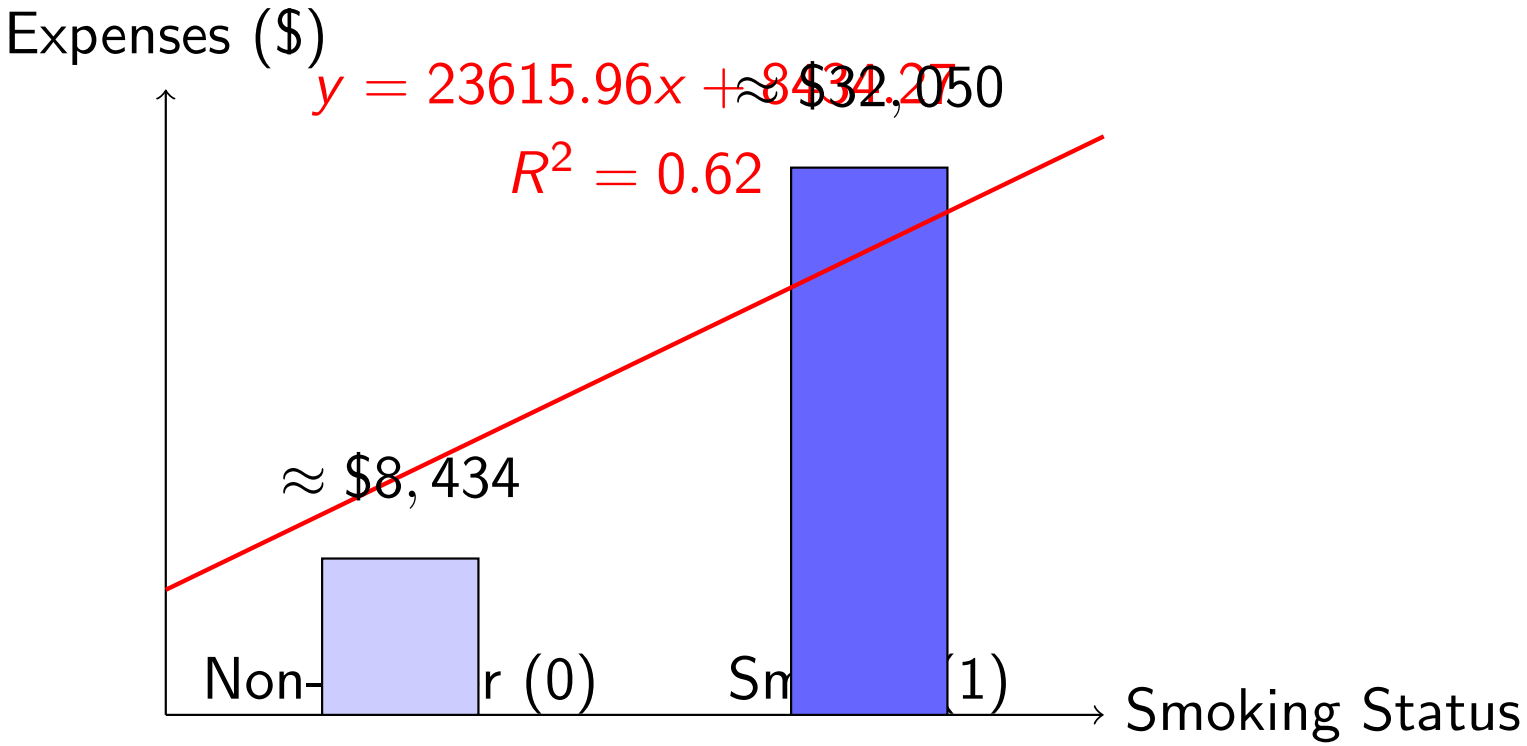
- **Key findings:**
  - **Smoking** has the strongest relationship with expenses (0.79)
  - **Age** shows moderate correlation (0.30)
  - **BMI** displays weak correlation (0.20)
  - **Sex** has minimal correlation (0.06)

Model Results: Individual Feature Contributions

Feature	R^2	**MSE**	**Unbiased** σ^2
Sex	0.0033	1.46×10^8	1.46×10^8
Age	0.0894	1.33×10^8	1.34×10^8
BMI	0.0394	1.41×10^8	1.41×10^8
Smoker	0.6198	5.57×10^7	5.58×10^7
Combined Model	**0.7475**	$\mathbf{3.70 \times 10^7}$	$\mathbf{3.71 \times 10^7}$

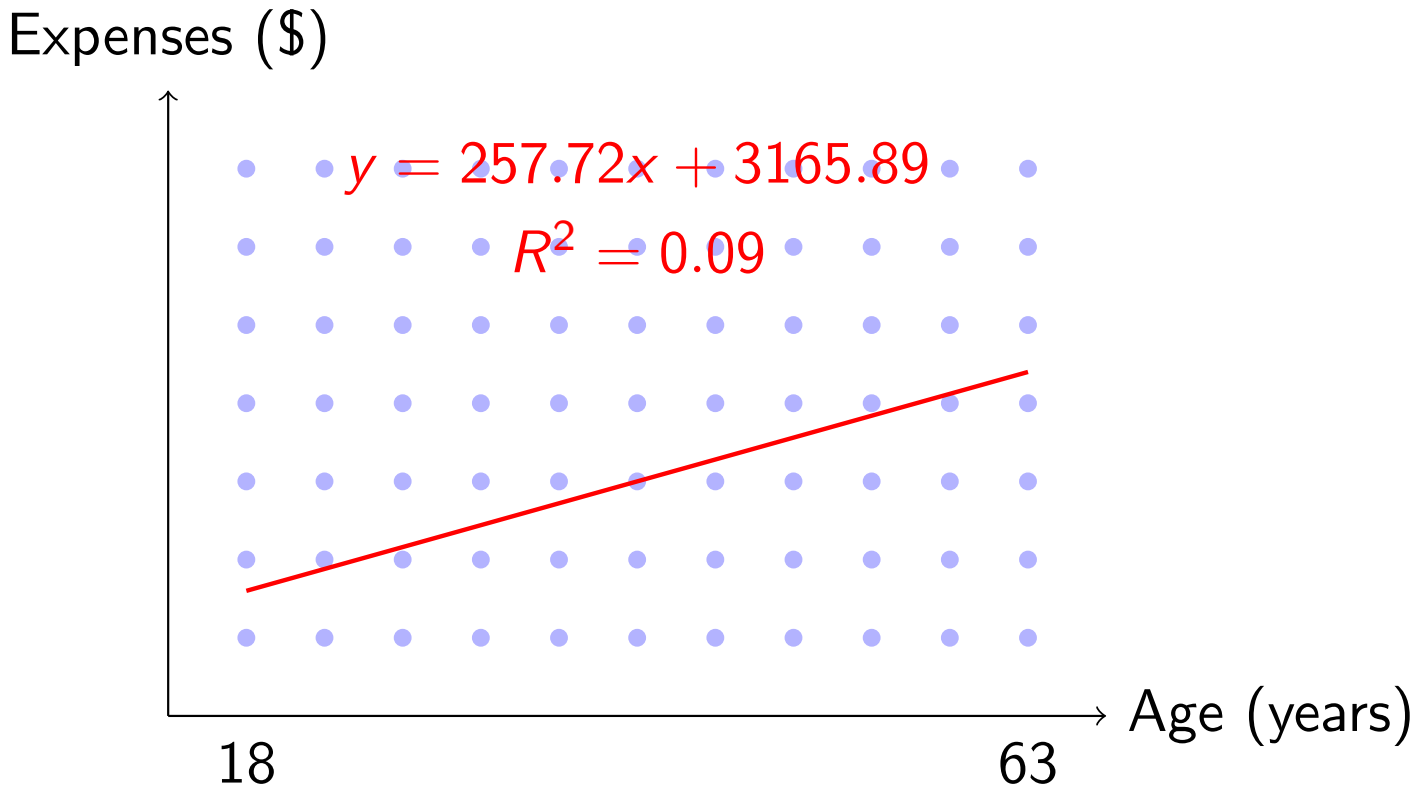


Expenses vs. Smoking Status



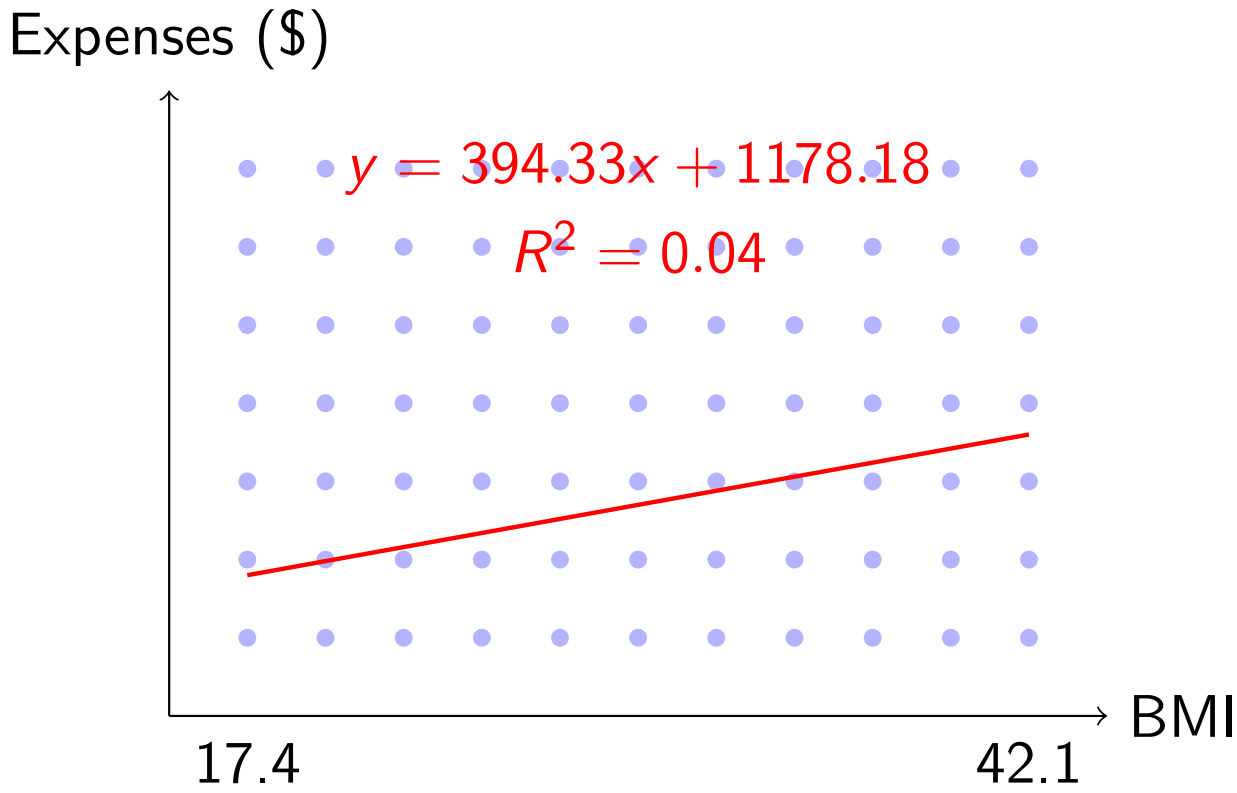
- Smokers have substantially higher medical expenses
- Difference of approximately $23,616 between smokers and non-smokers
- Smoking status alone explains 62% of variance in expenses

Expenses vs. Age



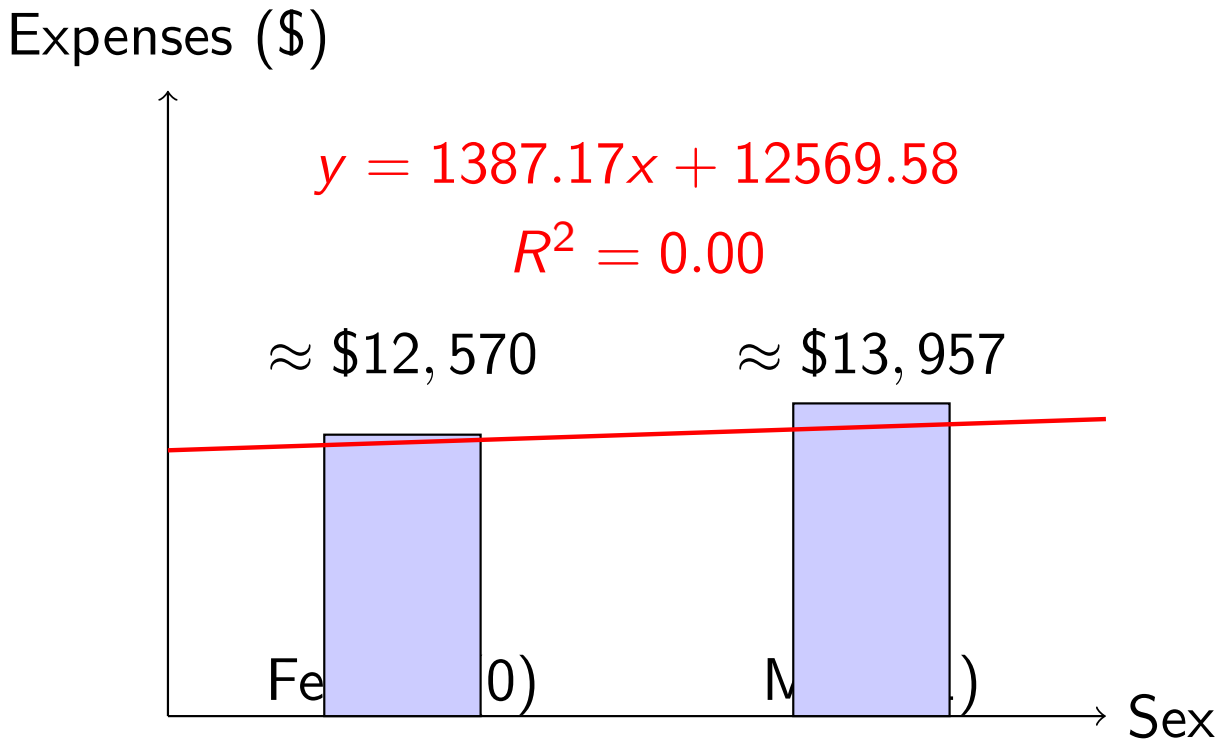
- Weak positive relationship between age and expenses
- Each additional year of age associated with $257.72 increase in expenses
- Age explains approximately 9% of variance in medical expenses

Expenses vs. BMI



- Weak positive relationship between BMI and expenses
- Each unit increase in BMI associated with $394.33 increase in expenses
- BMI explains only 4% of variance in medical expenses

Expenses vs. Sex



- Almost no relationship between sex and medical expenses
- Minimal difference in average expenses between males and females
- Sex alone explains less than 1% of variance in expenses

Combined Model

- Combined model explains 74.75% of variance in medical expenses
- Smoking status contributes most to the predictive power (62%)
- Age (9%) and BMI (4%) provide additional explanatory power
- Sex has minimal contribution (0.3%)
- About 25% of variance remains unexplained

Key Insights

- **Smoking Status:** Primary predictor of medical expenses
  - Smokers incur approximately $23,616 higher expenses than non-smokers
- **Age:** Secondary predictor with moderate influence
  - Each additional year associates with $257.72 increase in expenses
- **BMI:** Weak but measurable impact
  - Each unit increase associates with $394.33 higher expenses
- **Sex:** Minimal influence on medical expenses
  - Gender alone is not a meaningful predictor of healthcare care costs

Practical Applications

- **For Insurance Companies:**
  - Develop more accurate and equitable premium structures
  - Focus on smoking status as primary risk factor
  - Consider age and BMI as secondary factors
  - Avoid gender-based premium differentials
- **For Policymakers:**
  - Target smoking cessation programs for cost-effective interventions
  - Develop obesity prevention initiatives
- **For Individuals:**
  - Understand personal risk factors affecting healthcare costs
  - Make informed lifestyle decisions (particularly regarding smoking)

Conclusion

- Our regression analysis provides clear evidence on the factors that influence medical expenses:
  - Smoking is the dominant predictor (62% of variance)
  - Age and BMI provide additional predictive power
  - Sex has minimal impact on expenses
- The combined model explains approximately 75% of variance
- The findings can inform more equitable insurance pricing and targeted health interventions
- Future research should explore additional variables and causal mechanisms

Thank You!

By Ambati Yaswanth & Rajeev Kumar