



Linear Regression

Harshvardhan

University of Tennessee

Teaching Demo at UMass Amherst | December 2, 2024

Table of Contents

- What is Linear Regression?
- Mathematics of Linear Regression
- Interpreting Results and RSquared
- Assumptions and Model Diagnostics
- Conclusion



Case Study:
Health Insurance Premiums

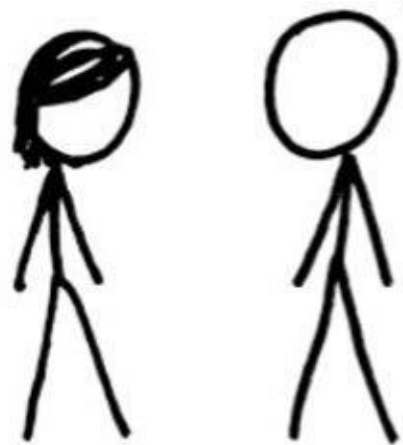
Linear Regression

Going from correlation to regression

What is Linear Regression?

- Linear regression is a *model* that estimates *linear relationship* between a dependent variable and one or more independent variable(s)
- It is an attempt to find the best fit line between independent and dependent variables
- **Strengths:**
 - Explainability and interpretability in understanding decisions
 - Strong statistical basis for usage and interpretation
- **Weaknesses:**
 - Assumes linear relationship – simple model is too simple
 - Sensitive to outliers
 - Assumptions – we will talk about them
 - **No causation implied, only a sophisticated form of correlation**

I USED TO THINK
CORRELATION IMPLIED
CAUSATION.



THEN I TOOK A
STATISTICS CLASS.
NOW I DON'T.



SOUNDS LIKE THE
CLASS HELPED.
WELL, MAYBE.



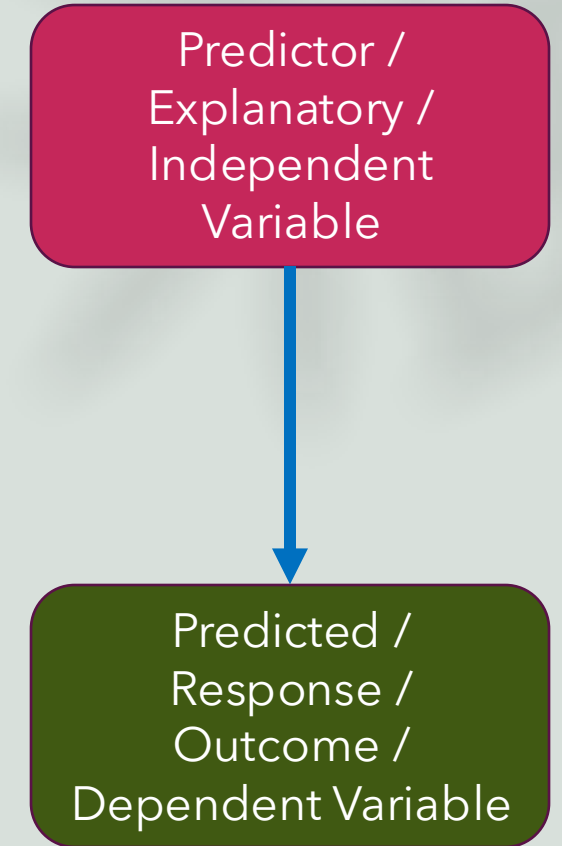
Key Components

- **Dependent Variable:**

- Variable to be predicted or explained
- Also known as **Response** or **Outcome** variables.
- Usually written as y_i
- Example: *Insurance Premium*

- **Independent Variable(s):**

- Variables used to predict or explain dependent variable
- Also known as **Predictor** or **Explanatory** variables
- Usually written as x_i
- Example: *BMI, Smoking Habits, # of Dependents, Age, etc.*

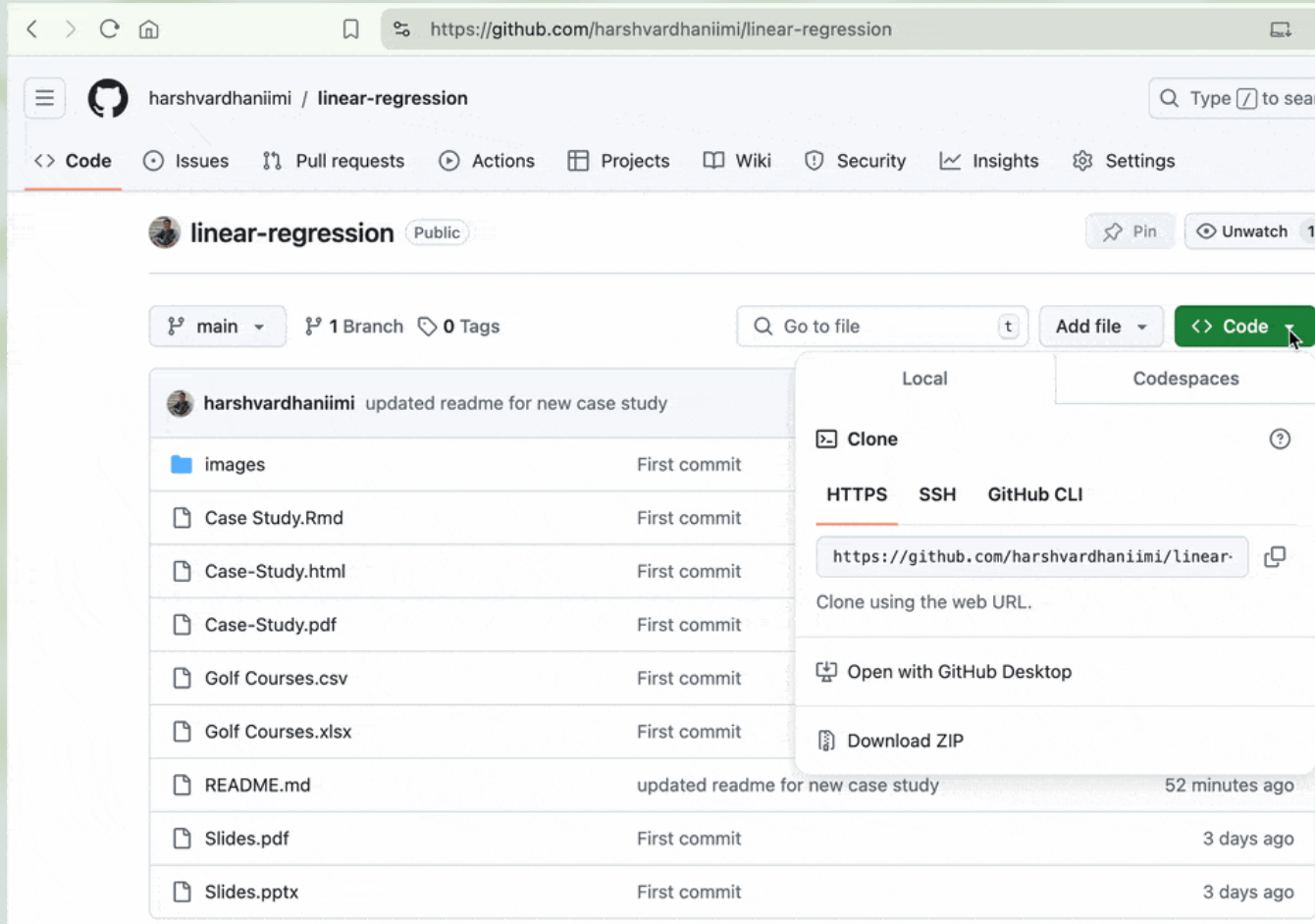


Case Study

What drives Health Insurance Premiums?



Kindly download slides,
data and R code.



Health Insurance Data

We will use data from Kaggle for a fictional case study.

Download files from Github:

github.com/harshvardhaniimi/linear-regression

Health Insurance Premiums in Massachusetts: What Drives Costs?

- In 2022, per capita health care spending was \$10,264, which was a 5.8% increase from 2021.
- The average cost of an individual health insurance plan in Massachusetts is \$721.19 per person.

What lifestyle factors drive insurance premium in general?

1.

2.

3.

Massachusetts Department of Public Health (MDPH)

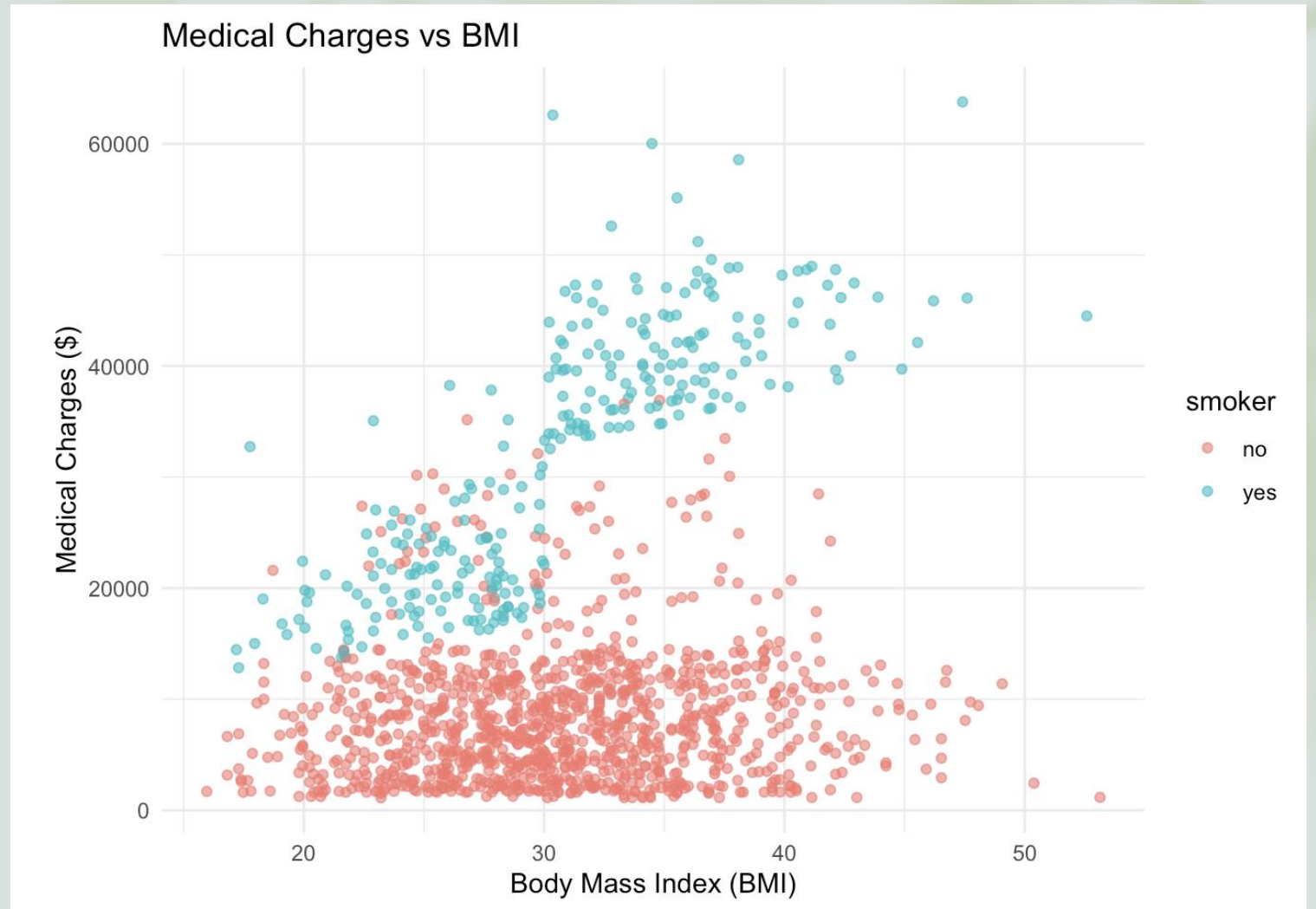
Identifying key factors driving health insurance charges.

- Goals
 1. Identify actionable Policy Insights for MDPH
 2. Empower insurers to structure premiums based on evidence rather than assumptions
- Variables
 1. Sex, Smoking Habits, Region
 2. Age, BMI, Dependents
 3. Insurance Charges

NOTE: Fictional Case Study

Scatterplot and Correlation

- Note for:
 - Direction
 - Strength
 - Outliers
- Linear regression is square of correlation between Y and X
- In multilinear regression, its squared correlation between \hat{Y} and Y (Why?)
- https://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html



Linear Regression Mathematics

Linear Model, Coefficients and Predicted Responses

Mathematics of Linear Regression

- Simple Linear Regression (single predictor)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Multiple Linear Regression (p predictors)

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_p x_i^p + \epsilon_i$$

$$Y = X\beta + E$$

Calculating Coefficients

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- $\beta_1 = \frac{Cov(x,y)}{Var(x)}$
$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$
- $\beta_0 = \bar{y} - \beta \bar{x}$
- β_1 measures how much y changes w.r.t. x , standardized by variability in x

Multiple Linear Regression

$$Y = X\beta + E$$

- $\beta = (X'X)^{-1}X'Y$
- where X is data matrix, Y is response vector

Predicted Response: \hat{Y}

- \hat{Y} is the estimated or predicted value of the dependent variable Y based on the estimated linear regression model
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ – Notice there is no residual term here. Why?
- **Interpretation:** \hat{y}_i is the best guess we have for y_i given information about (x_i, y_i)
- **Residual:** $\hat{\epsilon}_i = y_i - \hat{y}_i$
- **Example:** Predicted insurance premium based on lifestyle and other indicators

Interpretation, Assumptions and Diagnostics

How reliable are our conclusions?

Significance and Strength of Relationship

```
fit = lm(charges ~ bmi, data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20956  -8118  -3757    4722   49442
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1192.94    1664.80   0.717   0.474
## bmi           393.87     53.25   7.397 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862
## F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

- **P-value** measures significance of a relationship, typical threshold is 0.05
- **R-squared** measures proportion of variance in Y explained by X s
 - $R^2 = 0$ means no relationship
 - $R^2 = 1$ implies perfect relationship
 - Also called “goodness of fit”
- **Adjusted R-squared** accounts for number of variables
- **F-statistic** tests whether the regression model provides a better fit than a model with no predictors (i.e. simple mean)
 - Higher is better (and will have low p-value)
- **Residual Standard Error** measures average distance between \hat{Y} and Y

R-Squared: Goodness of Fit

- R^2 measures *the proportion of variance in dependent variable Y explained by independent variables X* in a linear regression model
 - $R^2 = 0.75$ means 75% of variance in Y is explained by X
- R lies between 0 (nothing can be explained) and 1 (everything explained)
- Higher R^2 generally implies better model
 - How high? Depends on field – economics or psychology (0.3 is okay), physics (even 0.9 may not be enough)
- Limitations:
 - R^2 doesn't indicate if the model is appropriate
 - Doesn't measure predictive performance for out of sample ("test") data
 - Doesn't imply causality
 - Will increase if # of independent variables increase
- Adjusted $RSquared$

Interpretation

```
fit = lm(charges ~ bmi, data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20956  -8118  -3757   4722  49442
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1192.94    1664.80   0.717   0.474
## bmi           393.87     53.25   7.397 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862
## F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

Based on the results:

1. Dependent variable (Y) = _____
2. Independent variable (X) = _____
3. Linear model, mathematically:

Interpretation

```
fit = lm(charges ~ bmi, data = df)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20956  -8118  -3757   4722  49442
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1192.94    1664.80   0.717   0.474
## bmi           393.87     53.25   7.397 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862
## F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

Is regression model significant?

- F-statistic =
- R^2 =
- Adj R^2 =

R^2 Interpretation:

Is the relationship between BMI and insurance premium significant?

- P-value =

Slope Interpretation:

Intercept Interpretation:

Multiple Linear Regression

- We can choose to include more than one variables in regression. Let's see an example.

```
Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
    region, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-11304.9	-2848.1	-982.1	1393.9	29992.8

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-11938.5	987.8	-12.086	< 2e-16 ***
age	256.9	11.9	21.587	< 2e-16 ***
sexmale	-131.3	332.9	-0.394	0.693348
bmi	339.2	28.6	11.860	< 2e-16 ***
children	475.5	137.8	3.451	0.000577 ***
smokeryes	23848.5	413.1	57.723	< 2e-16 ***
regionnorthwest	-353.0	476.3	-0.741	0.458769
regionsoutheast	-1035.0	478.7	-2.162	0.030782 *
regionsouthwest	-960.0	477.9	-2.009	0.044765 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom

Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494

F-statistic: 500.8 on 8 and 1329 DF, p-value: < 2.2e-16

Interpret R2:

Estimate effect of 'age':

Estimate effect of 'smoking':

Confidence Intervals of Estimated Coefficients

`confint(model)` function in R can give us 95% confidence intervals for all intercepts

```
confint(mlr, level = 0.95)
```

##	2.5 %	97.5 %
## (Intercept)	-13910.3546	-10070.18519
## age	233.6457	280.30145
## bmi	282.6391	394.69014
## children	204.3551	744.77779
## smokeryes	23028.3414	24644.25958
## regionnorthwest	-1286.2110	581.84680
## regionsoutheast	-1973.1303	-95.58998
## regionsouthwest	-1896.6557	-22.09365

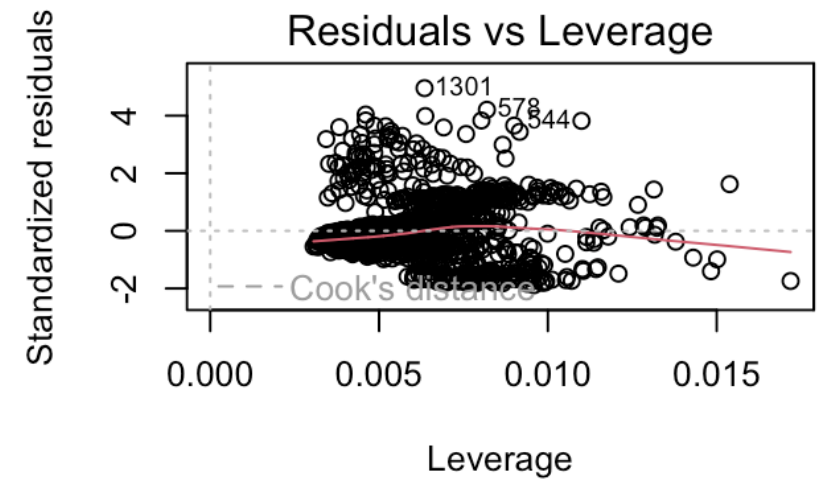
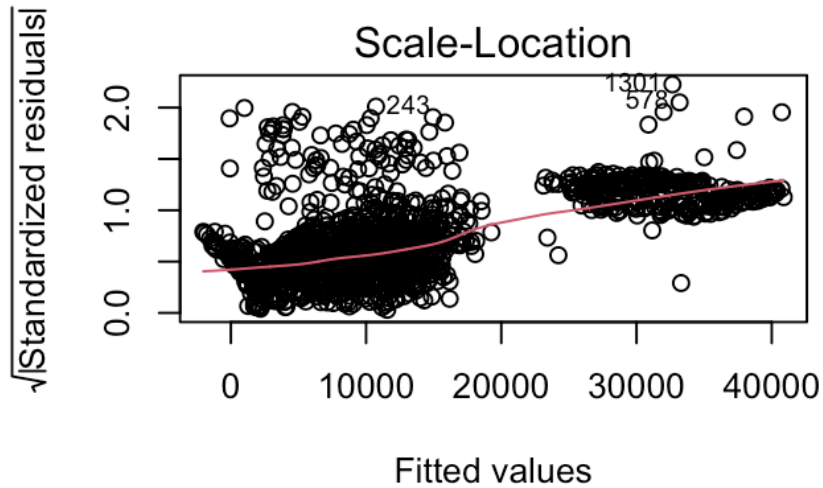
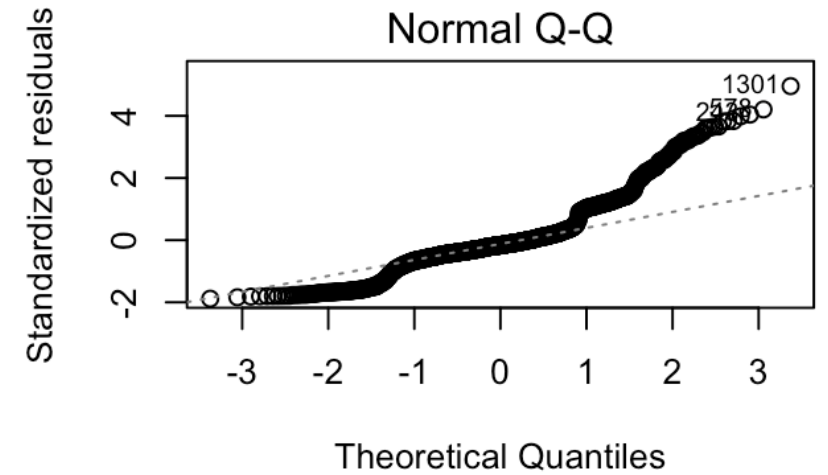
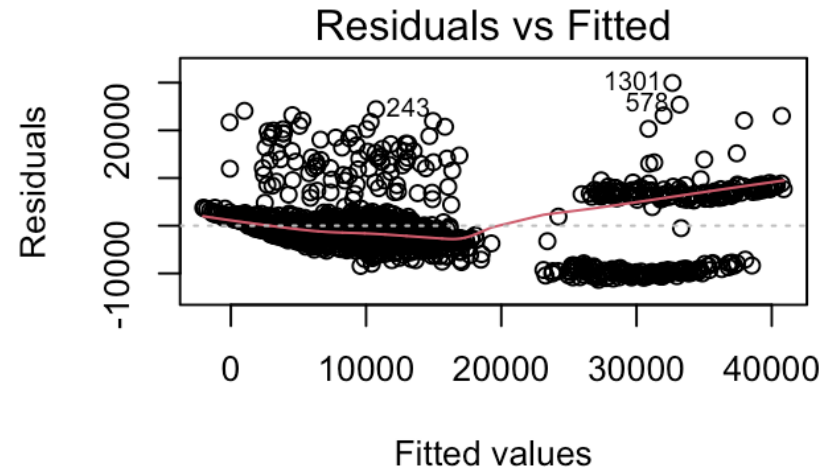
Assumptions in Linear Regression (LINE)

- Like all statistical models, Linear Regression works under certain assumptions:
 1. **L**inearity – assumed linear relationship between X and Y
 2. **I**ndependence – residuals (errors) are independent of each other
 3. **N**ormal distribution of residuals $N(0, \sigma^2)$
 4. **E**qual variance across values of X (homoskedasticity)

Model Diagnostics

- R provides four diagnostic plots:
 1. Residual vs Fitted plot (Linearity assumption)
 - Good if horizontal line shows with no distinct patterns
 - `plot(model, which = 1)`
 2. Normal Q-Q plot (Normality assumption)
 - Good if residuals follow diagonal dotted line
 - `plot(model, which = 2)`
 3. Scale-Location plot (Equal variance or Homoskedasticity assumption)
 - Good if horizontal line with equally spread points
 - `plot(model, which = 3)`
 4. Residuals vs Leverage (Detecting outliers)
 - Good if few points stand out
 - `plot(model, which = 4)`

Model Diagnostics Case



In-class Quiz

- Let's see how much we understood from today's class so far
- Visit <https://play.blooket.com/> and enter code XXXXXX

Business Insights

Real-World Implications and Actionable Insights

Impact of Smoking Habits, BMI and Others



Impact of Smoking on Charges

From our MLR results, we found that being a smoker is associated with an increase of approximately \$23,836 in annual medical charges, holding other variables constant.



Effect of BMI on Costs

For every one-unit increase in BMI, medical charges increase by around \$338.66, controlling for other factors like age and smoking.



Explaining Variability in Medical Charges

Smoking habits, BMI, age, number of dependents, and regional differences together explain 75% of the variability in health insurance premiums.

Recommended Policy Interventions



Expand the **Massachusetts Tobacco Cessation and Prevention Program (MTCP)** with targeted campaigns for high-risk populations – regions, incentives like tax benefits, or disincentives like higher taxes on cigarettes, etc.



Higher BMI leads to higher charges! Consider **subsidizing gym memberships, nutritional counseling, active transit like biking and walking** particularly for lower-income groups that face barriers to access



Extra premium for children's insurance causes higher childcare costs for parents. Consider subsidizing children's insurance!

Concluding Remarks

What we learned today?

- Linear regression helps us identify relationship and patterns between *independent* and *dependent* variables
- Smoking and BMI have high impact!
- Model coefficients tell us the impact of an individual variable
- RSquared tells us “goodness of fit” of a model
- Assumptions of LR should be verified before interpretation
- Residuals analysis can tell us more about the data than we think

What to remember?

- Regression’s strength is interpretability
- Regression coefficient being significant doesn’t imply causality
- Assumption of “linear model” might be too simplistic for pure prediction
- **LINE Assumptions:** Linearity, Independence, Normality, Equal Variance
- **There is more to regression! We will cover additional topics in coming classes**

Next Up...

- **Interaction between Variables**

- Identifying relationships that have combined/conditional effects, not additive effects
- Effect of exercise on weight might depend on diet type

- **Ranking Importance of Independent Variables**

- Scaling X by its mean and standard deviation

- **Linear regression for optimization**

- <https://blog.harsh17.in/using-linear-regression-to-find-optimal-value/>

- **Variable selection:** Lasso and ridge regressions



Questions?