SA2 Tutorials II

Manjusha Kancharla

3/26/2021

Agenda

1. Comparison of Two Populations (X)
2. Analysis of Variance (X)
3. Chi-square test (X)
4. Simple Linear regression

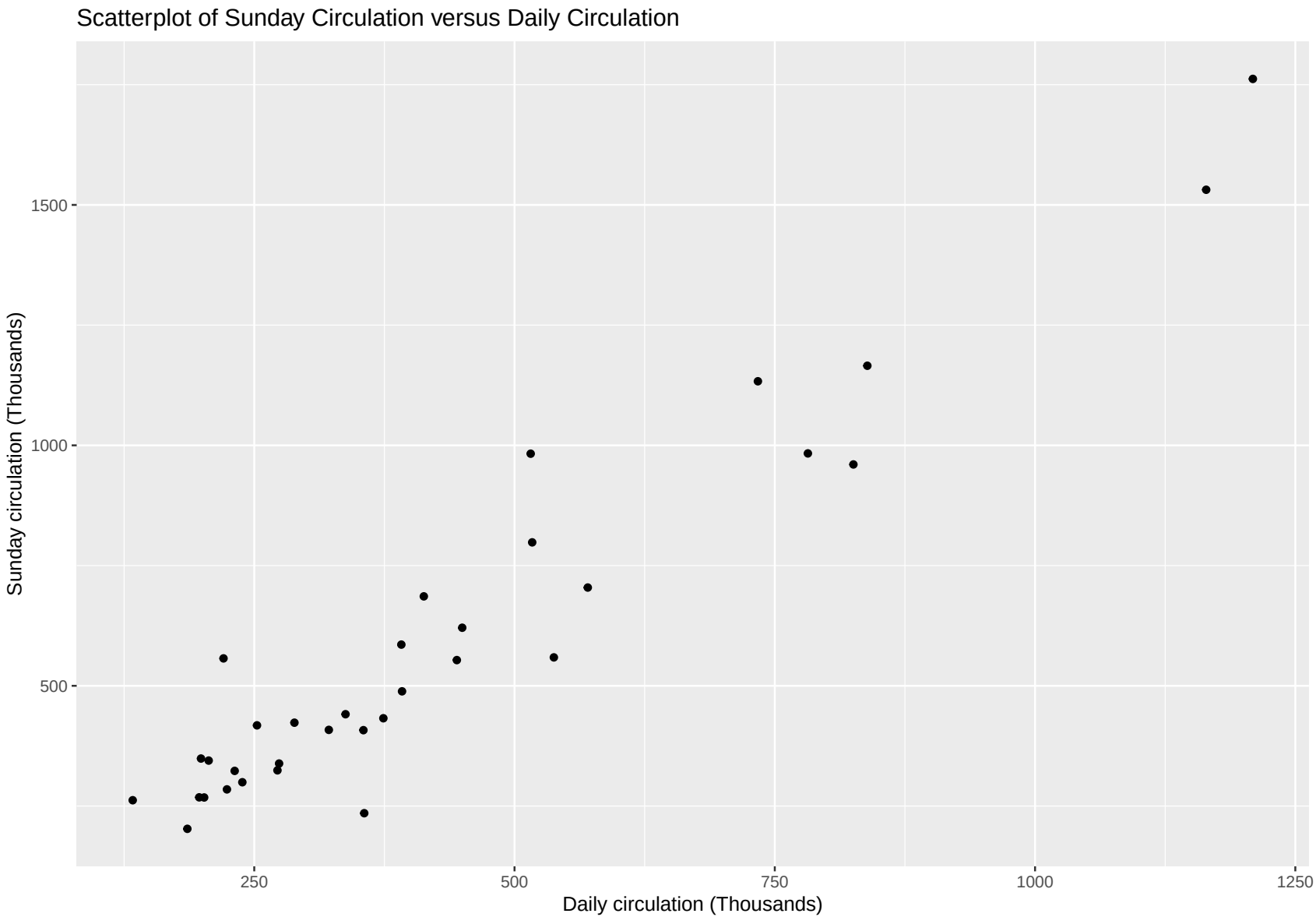
Practice Problem - 6

In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands) (Source: Gale Directory of Publications, 1994).

```
# A tibble: 6 x 3
  Newspaper           Daily Sunday
  <chr>               <dbl>  <dbl>
1 Baltimore Sun        392.   489.
2 Boston Globe         517.   798.
3 Boston Herald        356.   235.
4 Charlotte Observer   239.   299.
5 Chicago Sun Times    538.   559.
6 Chicago Tribune      734.  1133.
```

(a) Construct a scatter plot of Sunday circulation versus daily circulation. Does the plot suggest a linear relationship between daily and Sunday circulation? Do you think this is a plausible

Practice Problem - 6



Practice Problem - 6

(b) Fit a regression line predicting Sunday circulation from daily circulation.

Solution:

```
Call:
lm(formula = Sunday ~ Daily, data = Newspaper)

Coefficients:
(Intercept)        Daily
      13.84         1.34
```

Practice Problem - 6

The estimated best fit line:

$\hat{Sunday}_i = 13.84 + 1.34 \times Daily_i$

(c) Interpret the estimated intercept and slope.

Practice Problem - 6

The estimated best fit line:

$\hat{Sunday}_i = 13.84 + 1.34 \times Daily_i$

(c) Interpret the estimated slope.

Solution:

- For every one thousand increase in a newspaper's daily circulation, their Sunday circulation is expected to increase by 1.34 thousand.

Practice Problem - 6

(d) The particular newspaper that is considering a Sunday edition has a daily circulation of 500,000. What is the expected circulation of their Sunday edition (if they were to start a Sunday edition)?

Practice Problem - 6

(d) The particular newspaper that is considering a Sunday edition has a daily circulation of 500,000. What is the expected circulation of their Sunday edition (if they were to start a Sunday edition)?

Solution:

$\hat{Sunday_i} = 13.84 + 1.34 \times 500$

```
(Intercept)
    683.693
```

Practice Problem - 6

```
      1
683.693
```

So,the expected circulation of their Sunday edition is 683,693.

Practice Problem - 6

What proportion of the variability in Sunday circulation is accounted for by daily circulation?

```
Call:
lm(formula = Sunday ~ Daily, data = Newspaper)

Residuals:
    Min      1Q  Median      3Q     Max
-255.19  -55.57  -20.89   62.73  278.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.83563   35.80401   0.386    0.702
Daily        1.33971    0.07075  18.935   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.4 on 32 degrees of freedom
Multiple R-squared:  0.9181,    Adjusted R-squared:  0.9155
F-statistic: 358.5 on 1 and 32 DF,  p-value: < 2.2e-16
```

Practice Problem - 6

What proportion of the variability in Sunday circulation is accounted for by daily circulation?

Solution:

- $R^2 = 0.9181$
- 91.81% of the variability in Sunday circulation is accounted for by daily circulation

Practice Problem - 6

If daily and Sunday circulation were reversed in the above regression, what would you expect the R^2 to be?

Practice Problem - 6

If daily and Sunday circulation were reversed in the above regression, what would you expect the R^2 to be?

Solution:

- Wouldn't change
- Math: $R^2 = \widehat{Cor}(\boldsymbol{X}, \boldsymbol{Y})^2 = \widehat{Cor}(\boldsymbol{Y}, \boldsymbol{X})^2$
- Intuition: simple linear regression fits a line so as to minimize the distance between the observed points and the fitted line. That distance doesn't change if you flip the X and Y axes.

Practice Problem - 6

Obtain the 95% confidence intervals for $\beta_0 and \beta_1$.

```
                 2.5 %    97.5 %
(Intercept) -59.094743 86.766003
Daily         1.195594  1.483836
```

Practice Problem - 6

Suppose that daily circulation in a given newspaper is 650. What is the estimated Sunday circulation for the same newspaper? Construct the 95% confidence interval for the estimate.

Practice Problem - 6

Suppose that daily circulation in a given newspaper is 650. What is the estimated Sunday circulation for the same newspaper? Construct the 95% confidence interval (*Prediction Interval*) for the estimate.

```
       fit    lwr     upr
1 884.6502 656.32 1112.98
```

Practice Problem - 6

Is there a significant relationship between Sunday circulation and daily circulation? Justify your answer by a statistical test. Indicate what hypothesis you are testing and your conclusion.

Practice Problem - 6

Is there a significant relationship between Sunday circulation and daily circulation? Justify your answer by a statistical test. Indicate what hypothesis you are testing and your conclusion.

- Hypothesis

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$$

- Test Statistic is the t-statistic with $n - 2 = 34 - 2 = 32$ df.
- Observed t-value (from regression output) is 18.935 and the corresponding p-value is ≈ 0.
- Since the p-value of the slope is very small, we reject the null hypothesis, which means that there is a statistically significant (linear) relationship between Sunday and Daily Circulation.

Textbook data are unrealistic!! Real world is WAY messier.

- Goal: Can you accurately predict insurance costs?
- Data: 'insurance.csv'
- Source: Kaggle (https://www.kaggle.com/mirichoi0218/insurance)

Insurance Data

Columns/Variables in the dataset:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9 (**X**)
- children: Number of children covered by health insurance/Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance (**Y**)

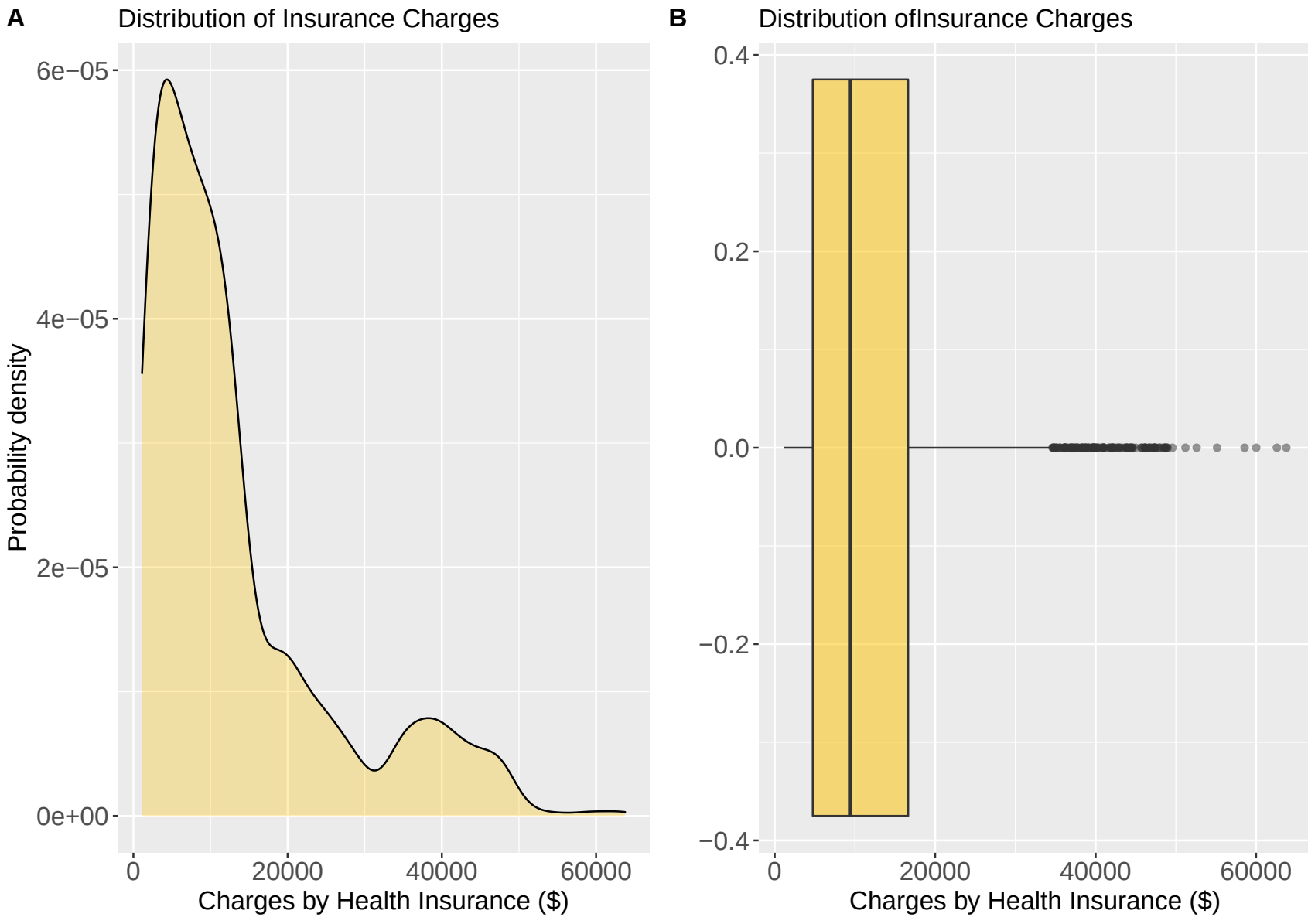
Insurance Data

Goal: Can you accurately predict insurance costs using BMI as the only predictor?

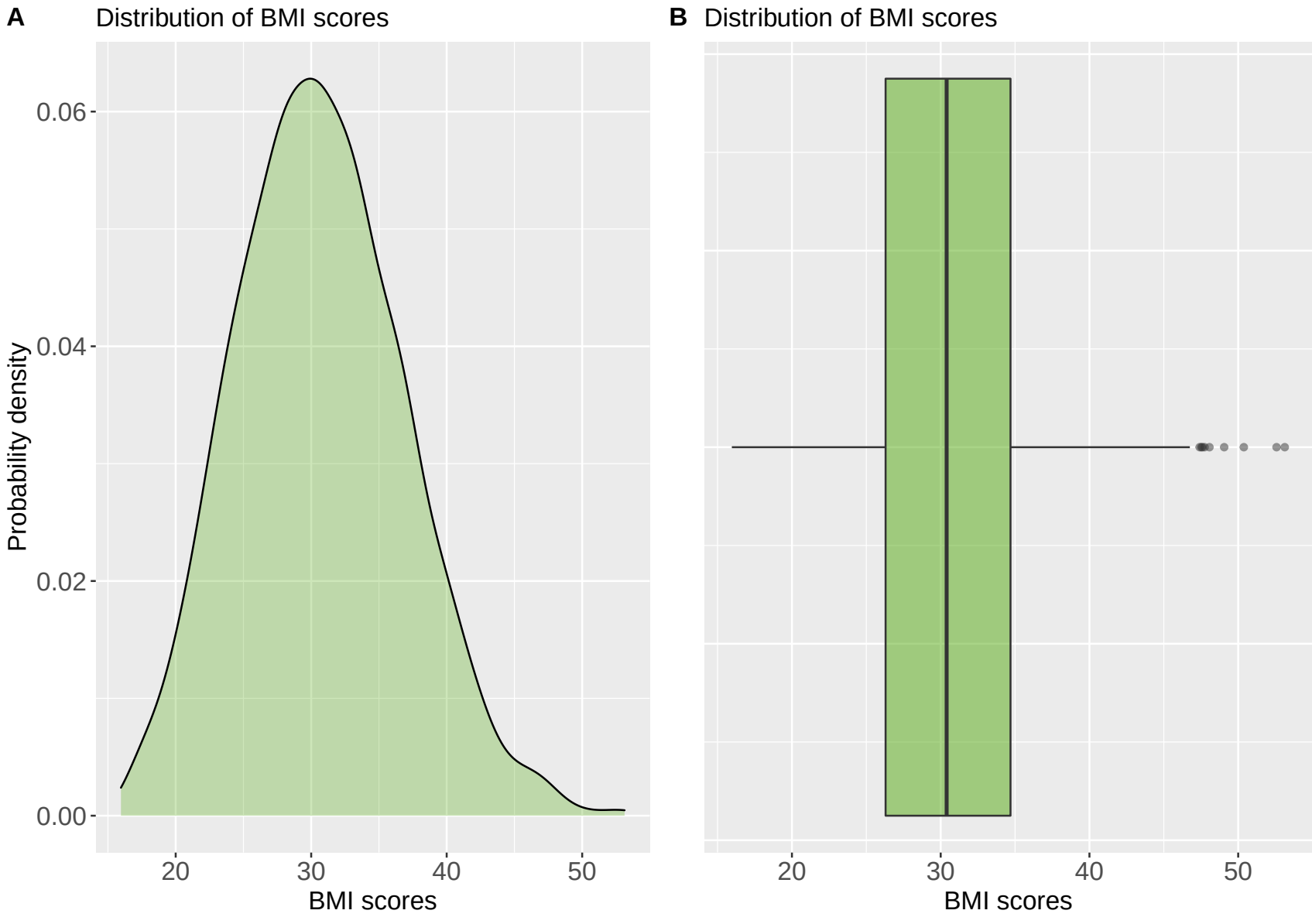
Insurance Data: EDA

```
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             1338
Number of columns          7
_______________________
Column type frequency:
  factor                   3
  numeric                  4
_______________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 sex                   0             1 FALSE          2 mal: 676, fem: 662
2 smoker                0             1 FALSE          2 no: 1064, yes: 274
3 region                0             1 FALSE          4 sou: 364, nor: 325, sou: 325, nor: 324

── Variable type: numeric ─────────────────────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate     mean       sd     p0    p25    p50     p75    p100 hist
1 age                   0             1    39.2     14.0    18     27     39      51      64   ▇▇▇▇
2 bmi                   0             1    30.7      6.10   16.0   26.3   30.4    34.7    53.1 ▃▇▇▂
3 children              0             1     1.09     1.21    0      0      1       2       5   ▇▂▁▁
4 charges               0             1 13270.   12110.   1122.  4740.  9382.  16640.  63770.  ▇▂▁▁
```

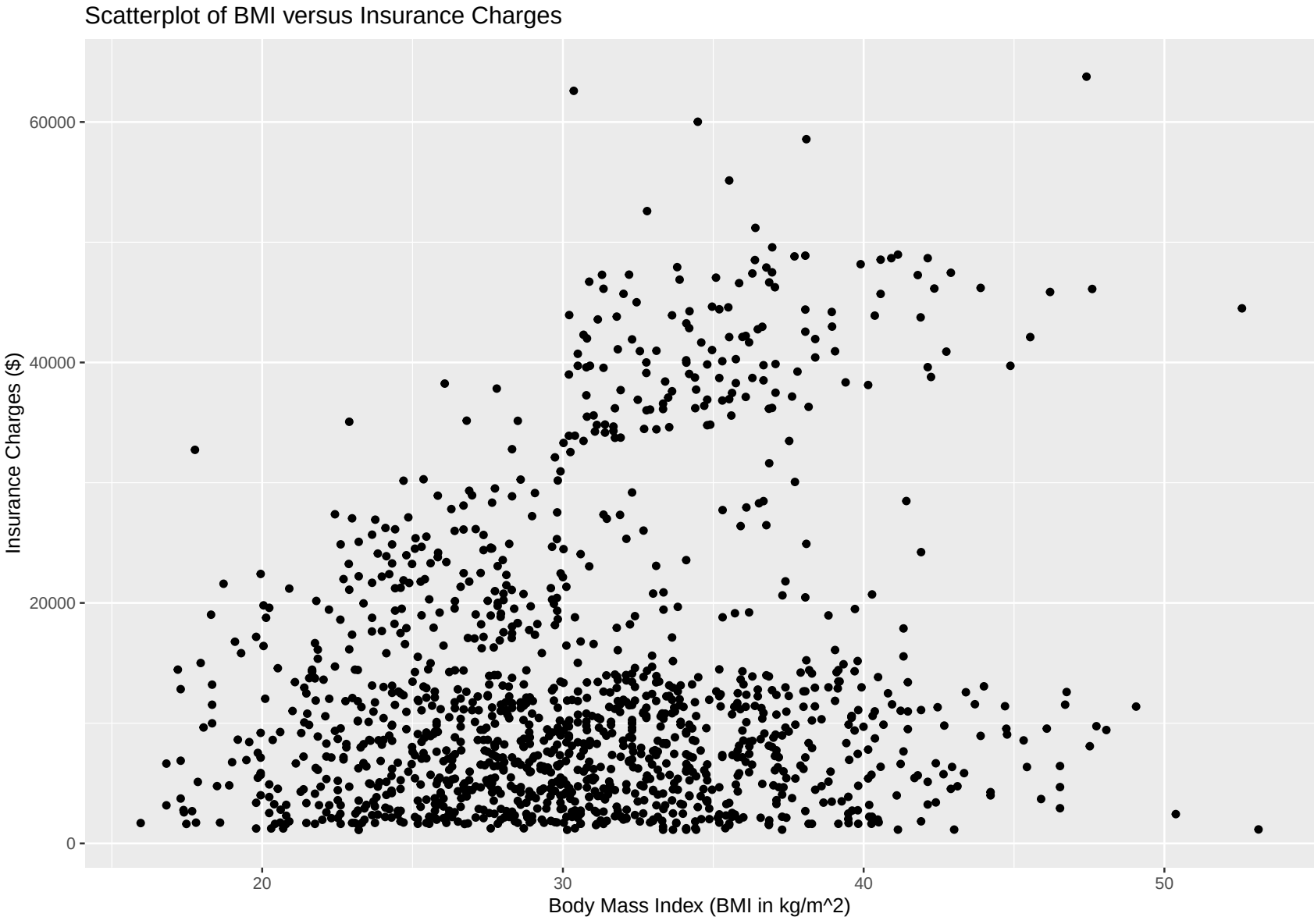
Insurance Data: EDA



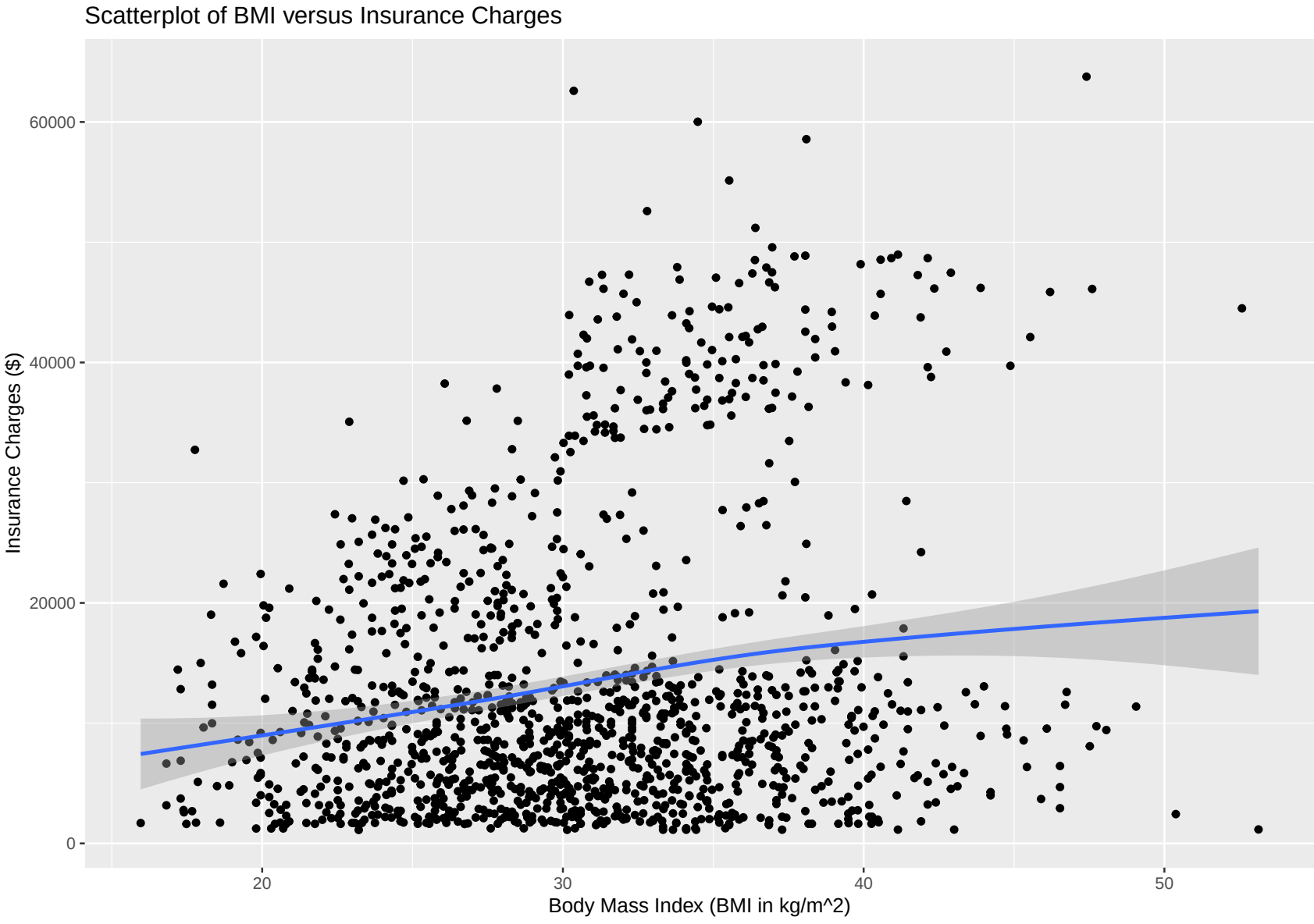
Insurance Data: EDA



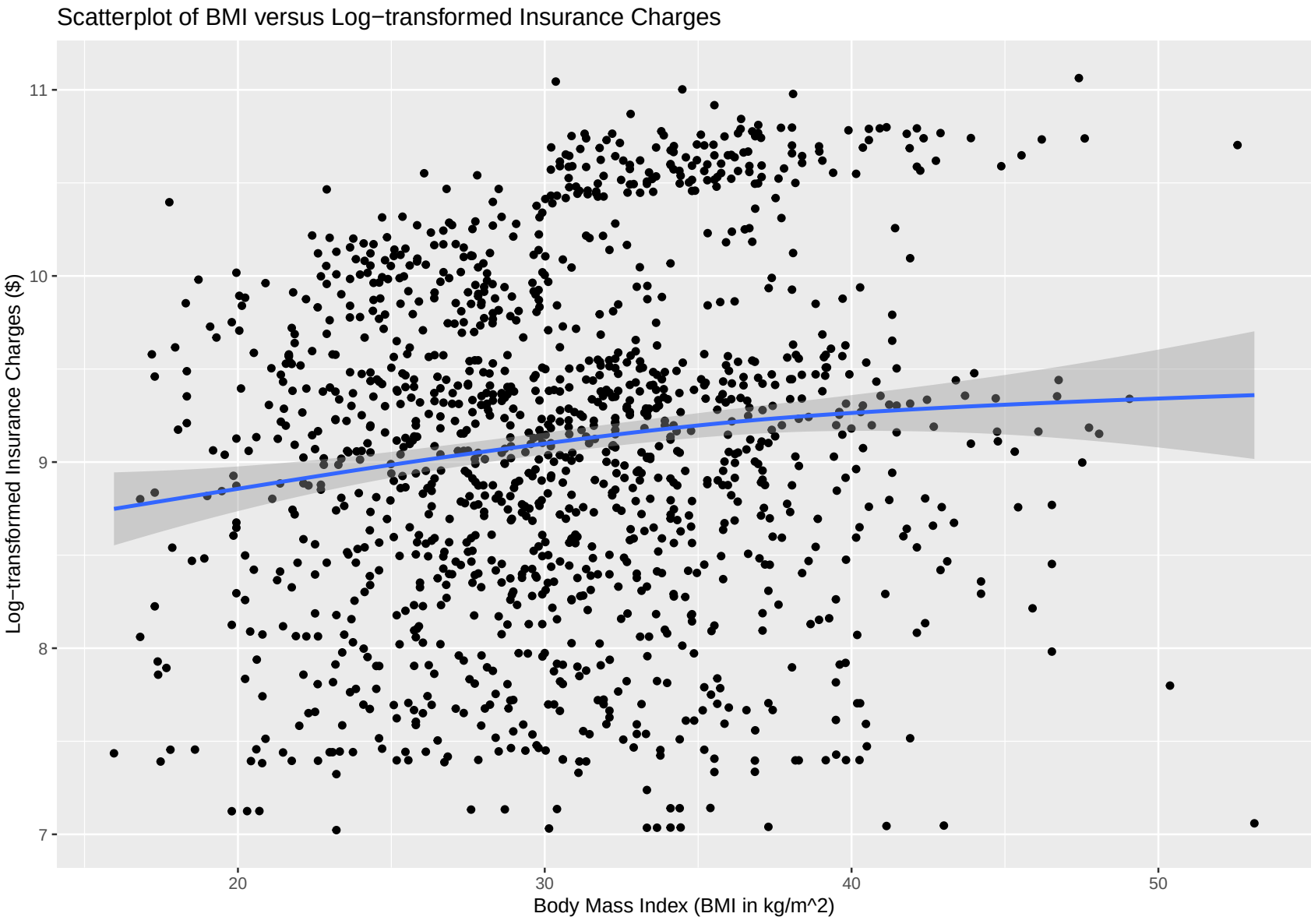
Insurance Data: EDA



Insurance Data: EDA



Insurance Data: EDA



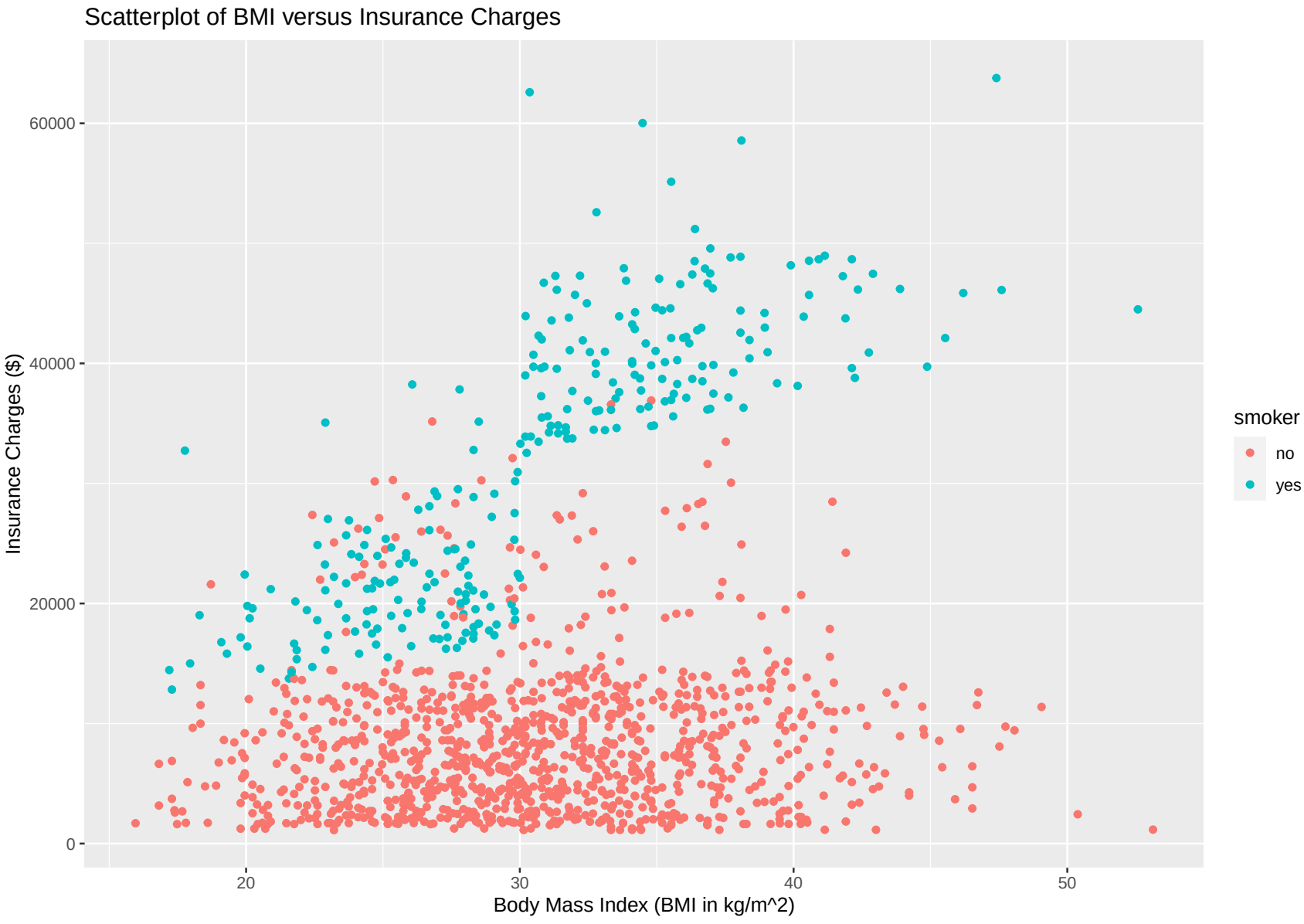
Insurance Data: Initial Model

```
Call:
lm(formula = charges_log ~ bmi, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.48894 -0.63536  0.03136  0.68007  1.95182

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.485243   0.127833  66.378  < 2e-16 ***
bmi         0.020005   0.004089   4.892 1.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9117 on 1336 degrees of freedom
Multiple R-squared:  0.0176,    Adjusted R-squared:  0.01687
F-statistic: 23.94 on 1 and 1336 DF,  p-value: 1.117e-06
```

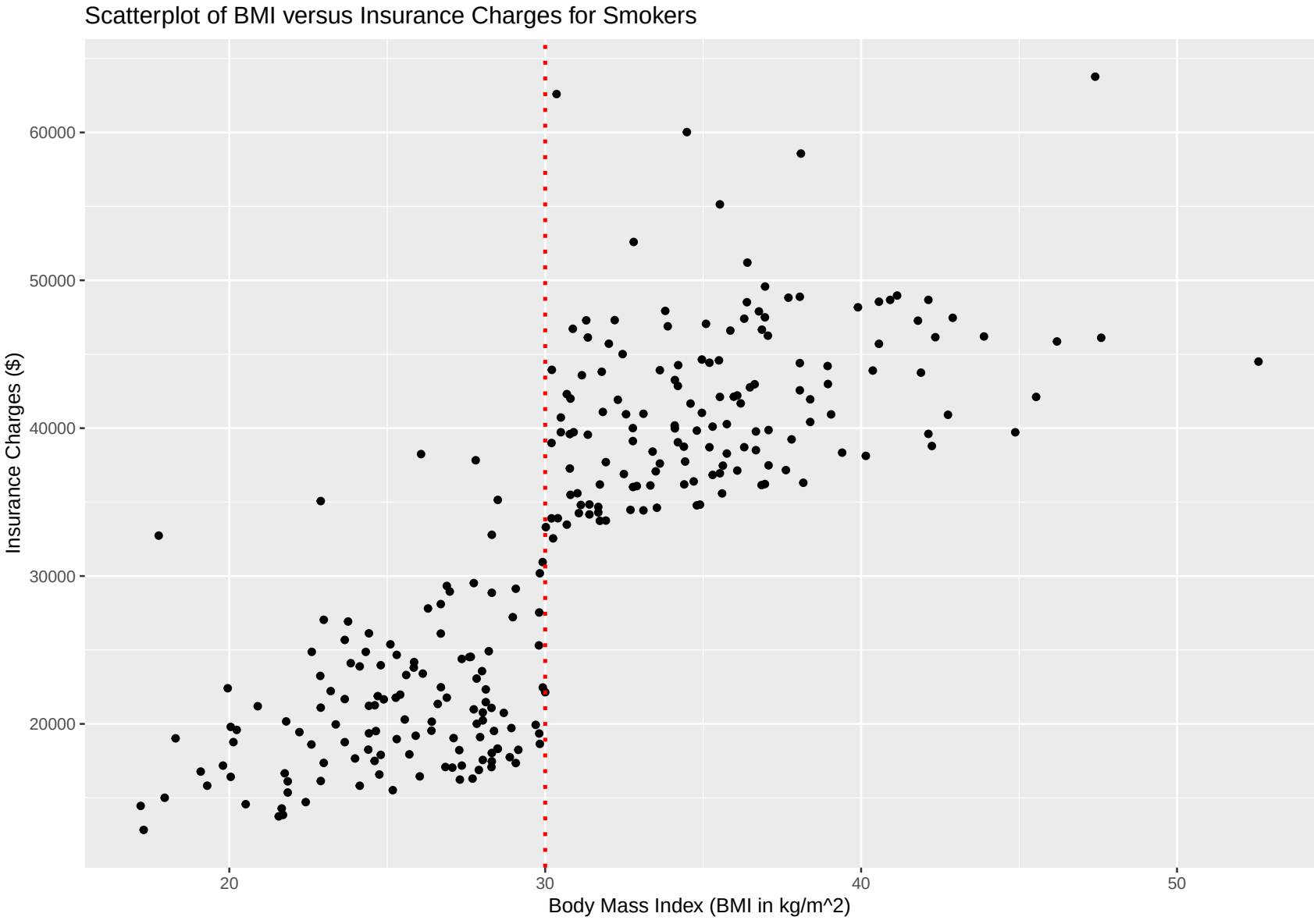
Insurance Data: Initial Model Interpretation

- If your BMI were to go up by one unit, you would expect a $100 * 0.02\% = 2\%$ increase in your insurance charges.

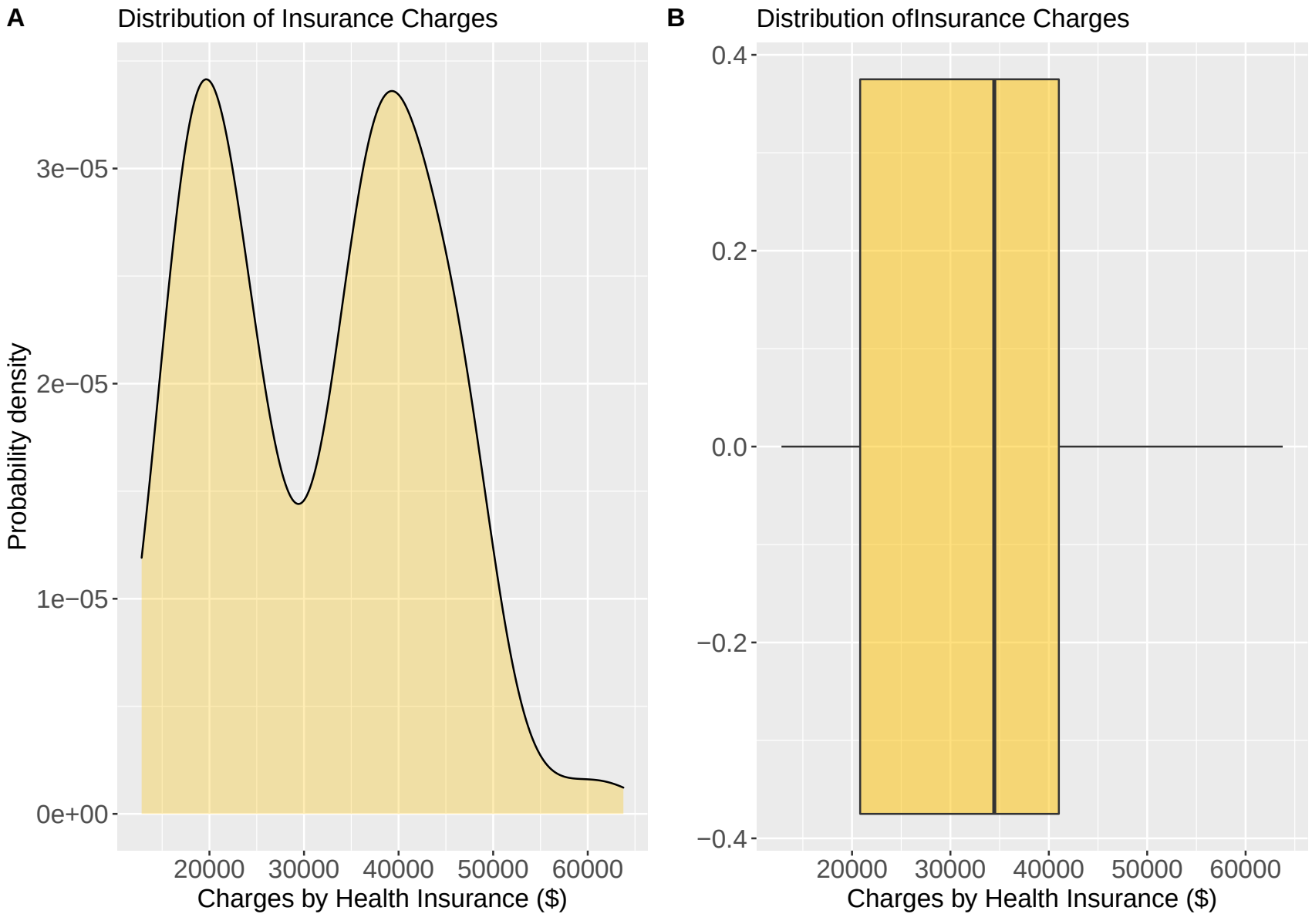
Insurance Data: DON'T GIVE UP ON EDA TOO SOON!



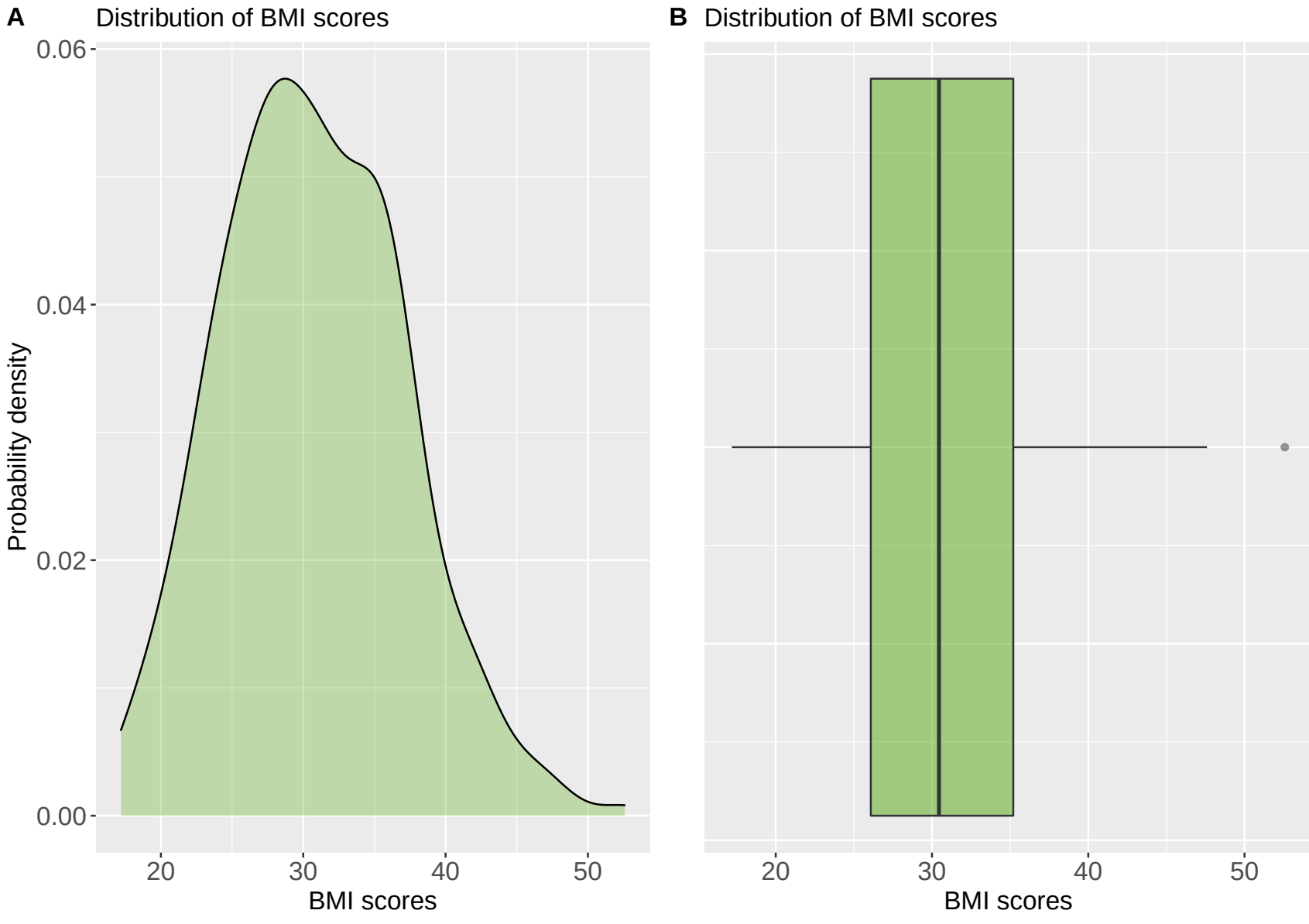
Insurance Data: Model for Smokers



Insurance Data: EDA



Insurance Data: EDA



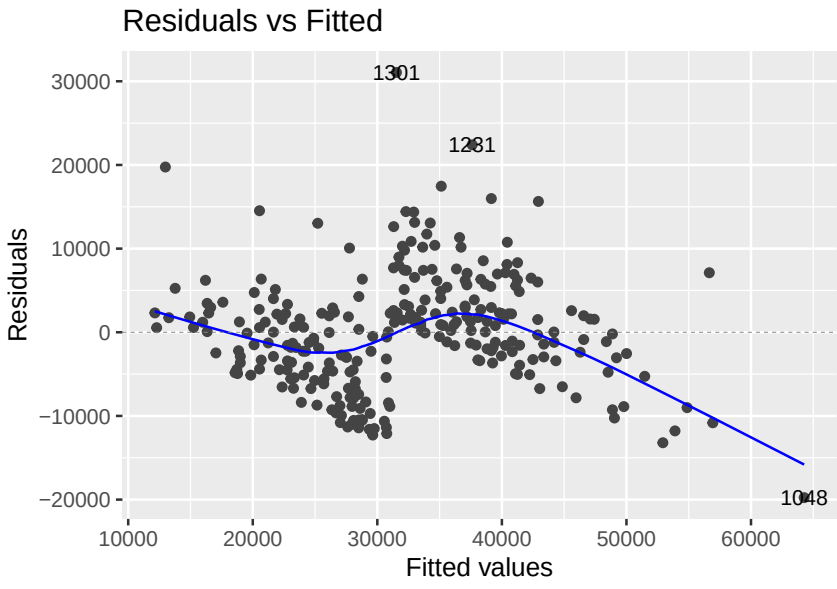
Insurance Data: Initial Model for Smokers

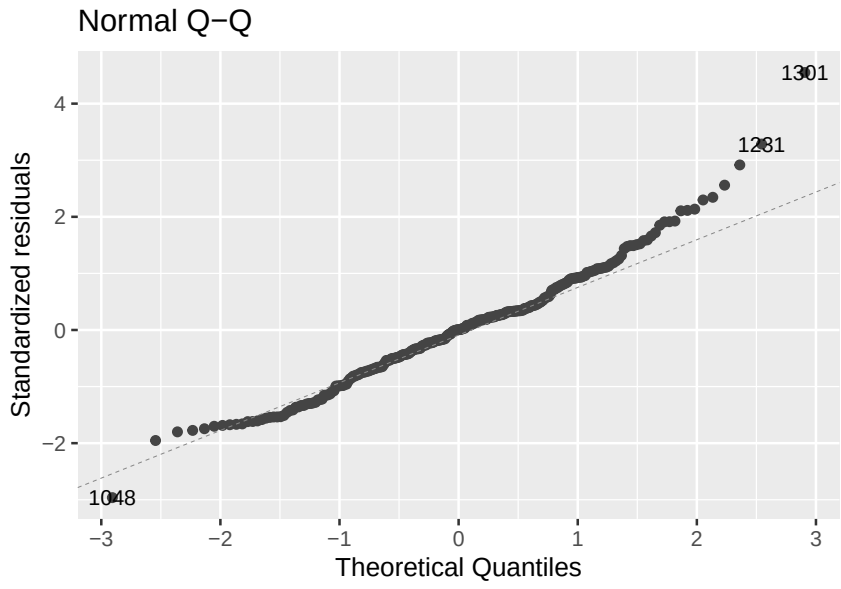
```
Call:
lm(formula = charges ~ bmi, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-19768.0  -4487.9     34.4   3263.9  31055.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13186.58    2052.88  -6.423 5.93e-10 ***
bmi           1473.11      65.48  22.496  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6837 on 272 degrees of freedom
Multiple R-squared:  0.6504,    Adjusted R-squared:  0.6491
F-statistic: 506.1 on 1 and 272 DF,  p-value: < 2.2e-16
```

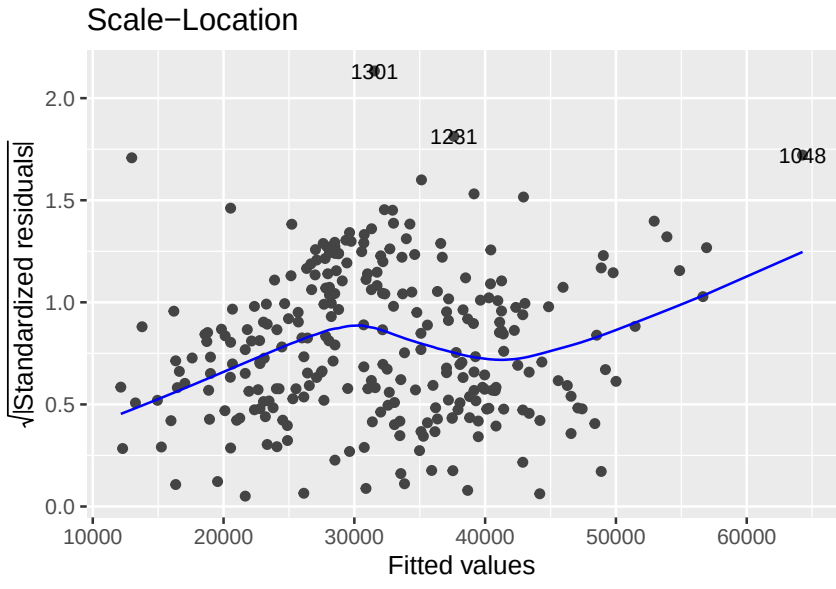
Insurance Data: Initial Model for Smokers Interpretation

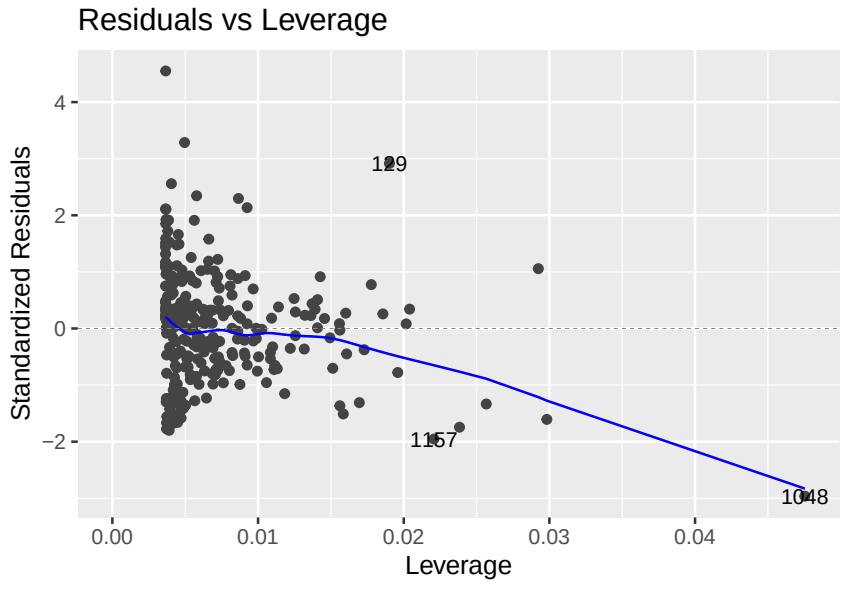
- If you are a smoker and your BMI were to go up by one unit, you would expect a $1473 increase in your insurance charges.
- Look at how the R^2 changed once we seperate the smokers out.

Insurance Data: Initial Model for Smokers Diagnostics

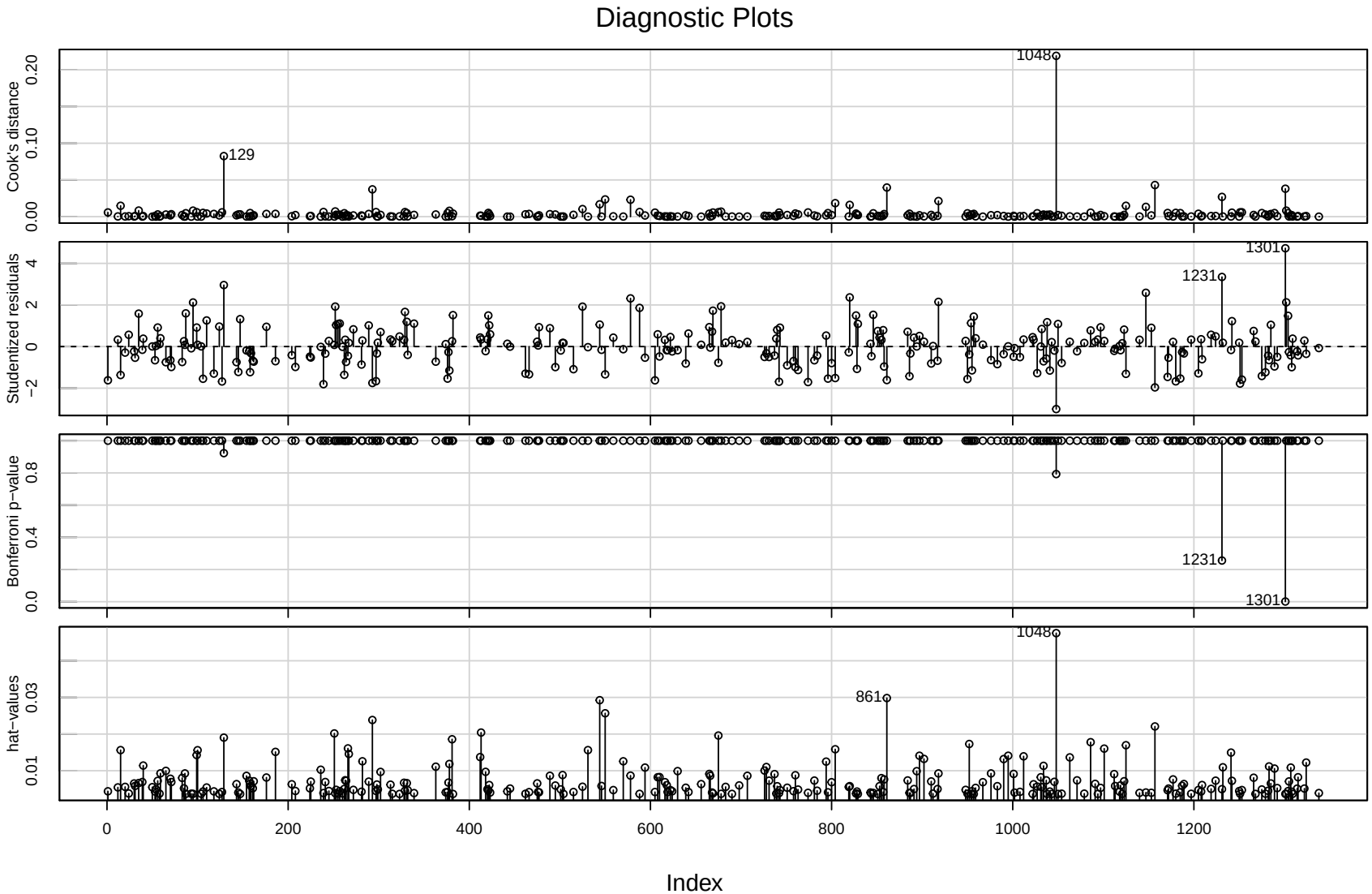








Insurance Data: Initial Model for Smokers Diagnostics



Insurance Data: Initial Model for Smokers Diagnostics

- Observation 1048 has the largest BMI (52)
- An influential point will change your model if it is included

Insurance Data: Initial Model for Smokers without Obs # 1048

```
Call:
lm(formula = charges ~ bmi, data = df[df$bmi < 52.58, ])

Residuals:
     Min       1Q   Median       3Q      Max
-13869.6  -4277.7     27.6   3333.5  30994.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14389.74    2062.42  -6.977 2.31e-11 ***
bmi            1514.75      66.01  22.948  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6738 on 271 degrees of freedom
Multiple R-squared:  0.6602,    Adjusted R-squared:  0.659
F-statistic: 526.6 on 1 and 271 DF,  p-value: < 2.2e-16
```

What happens when the assumption of Homosckedasticity fails?

- We are not able to properly estimate the variance-covariance matrix (error variance).
- One cannot compute any t-statistics and p-values and consequently hypothesis testing is not possible.
- Overall, under heteroscedasticity OLS loses its efficiency and is not BLUE (Best Linear Unbiased Estimator) anymore.
- Estimates are still unbiased.

Next steps

- There are methods to build models for bimodal dependent variables.
- Maybe look at a model for obese ($BMI > 30$) smokers only?

Thank you!