Project 1: Exploring Key Predictors of U.S. Health-Insurance Charges with Linear Regression

Muhammad Rafey Omer

09/28/2025

Problem Statement: What Drives Health-Insurance Costs?

Setup

```
library(dplyr)
library(ggplot2)
library(readr)
library(stringr)
library(tidyr)
library(knitr)
```

Data and Problem

Dataset: The dataset contains **1,338 health-insurance records**, each describing a single policyholder with demographics (`age`, `sex`, `region`), health indicators (`bmi`, `smoker`), family details (`children`), and the **annual medical charges** billed to the insurer (`charges`).

Goal: The aim of this analysis is to determine which personal and lifestyle characteristics have the greatest influence on medical insurance expenses, and to construct a simple regression model capable of predicting annual charges.

Import & Initial Checks

```
insurance <- read_csv("insurance.csv", show_col_types = FALSE)
glimpse(insurance)
```

```
## Rows: 1,338
## Columns: 7
## $ age      <dbl> 19, 18, 28, 33, 32, 31, 46, 37, 37, 60, 25, 62, 23, 56, 27, 1~
```

```
## $ sex      <chr> "female", "male", "male", "male", "male", "female", "female",~
## $ bmi      <dbl> 27.900, 33.770, 33.000, 22.705, 28.880, 25.740, 33.440, 27.74~
## $ children <dbl> 0, 1, 3, 0, 0, 0, 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0~
## $ smoker   <chr> "yes", "no", "no", "no", "no", "no", "no", "no", "no", "no", ~
## $ region   <chr> "southwest", "southeast", "southeast", "northwest", "northwes~
## $ charges  <dbl> 16884.924, 1725.552, 4449.462, 21984.471, 3866.855, 3756.622,~
```

```
# Missing values check
miss_summary <- sapply(insurance, function(x) sum(is.na(x)))
kable(data.frame(variable = names(miss_summary), n_missing = as.integer(miss_summary)),
      caption = "Missing values by column")
```

Table 1: Missing values by column

variable	n_missing
age	0
sex	0
bmi	0
children	0
smoker	0
region	0
charges	0

Wrangling

```
insurance <- insurance %>%
  mutate(
    sex    = as.factor(sex),
    smoker = as.factor(smoker),
    region = as.factor(region)
  )

summary(insurance)
```

```
##       age             sex            bmi           children      smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
##  Median :39.00                Median :30.40   Median :1.000
##  Mean   :39.21                Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##        region        charges
##  northeast:324   Min.   : 1122
```

```
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

Exploratory Data Analysis

Numeric summaries

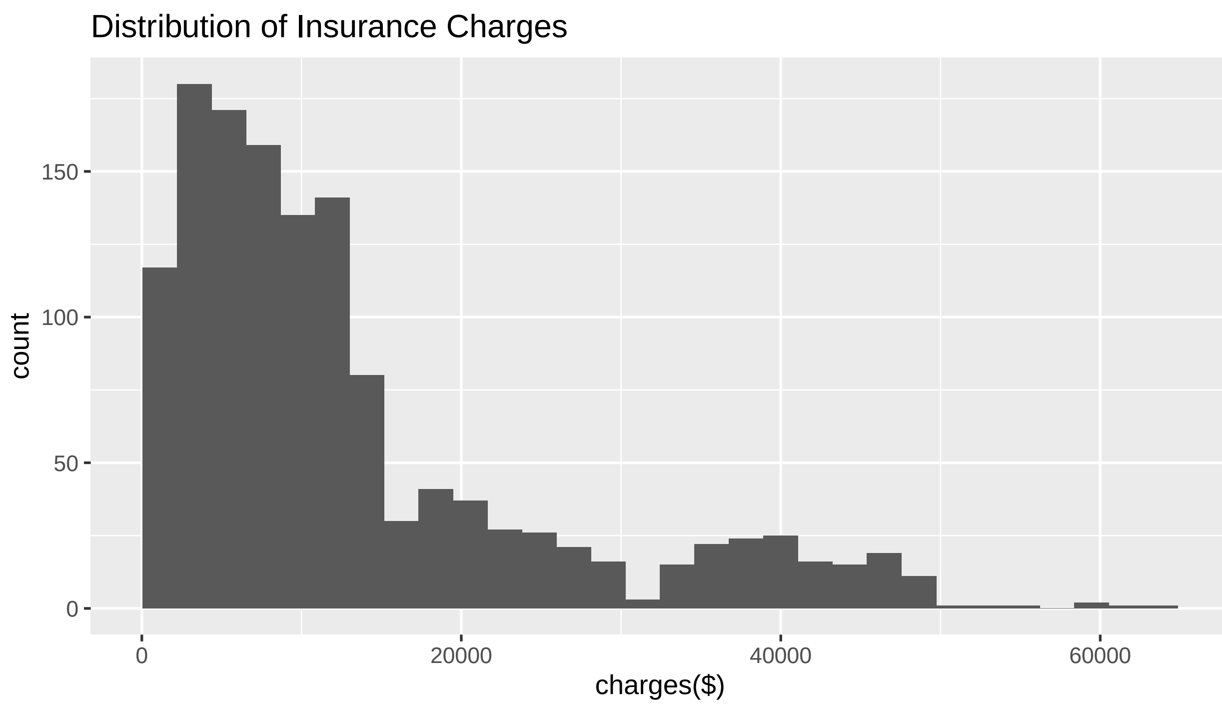
```r
num_cols <- c("age", "bmi", "children", "charges")
summ_tbl <- insurance %>%
  select(all_of(num_cols)) %>%
  summarise(across(everything(), list(min = min, q1 = ~quantile(.x, 0.25), median = median,
                                      mean = mean, q3 = ~quantile(.x, 0.75), max = max)))
summ_tidy <- summ_tbl %>%
  pivot_longer(everything(),
               names_to = c("variable", ".value"),
               names_sep = "_") %>%
  arrange(variable)

kable(summ_tidy, caption = "Summary statistics for numeric columns")
```

Table 2: Summary statistics for numeric columns

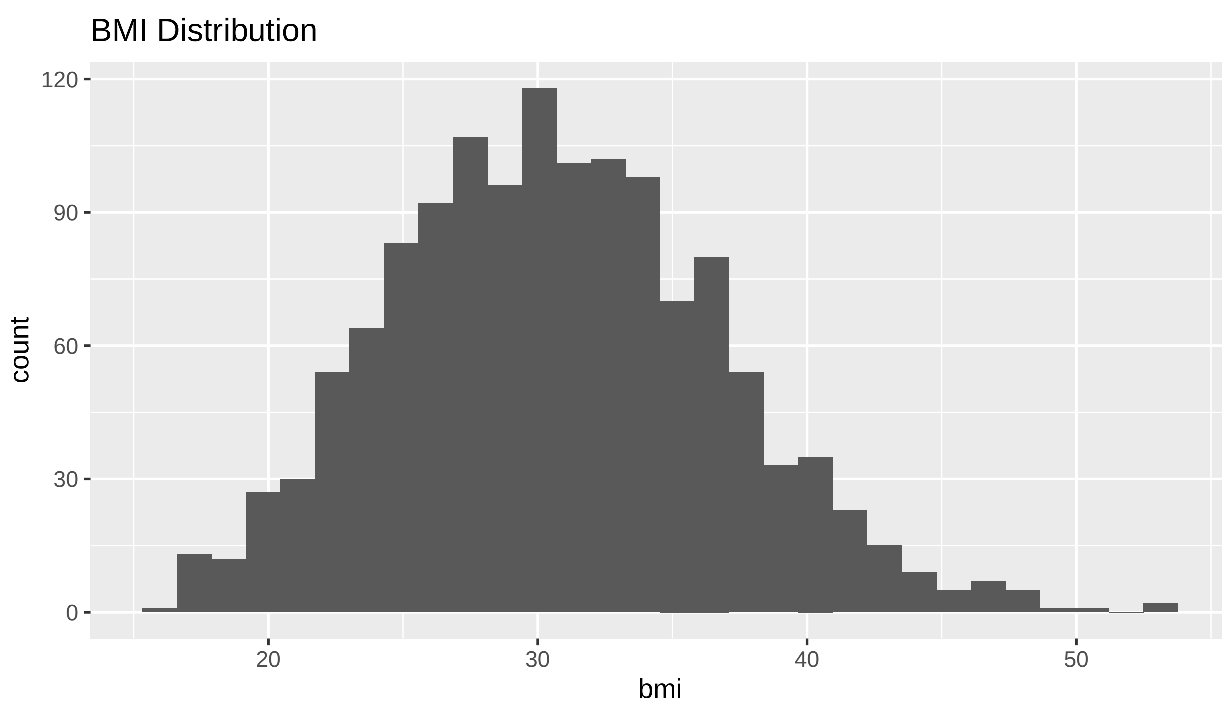
variable	min	q1	median	mean	q3	max
age	18.000	27.00000	39.000	39.207025	51.00000	64.00
bmi	15.960	26.29625	30.400	30.663397	34.69375	53.13
charges	1121.874	4740.28715	9382.033	13270.422265	16639.91251	63770.43
children	0.000	0.00000	1.000	1.094918	2.00000	5.00

Distributions

```r
ggplot(insurance, aes(charges)) +
  geom_histogram(bins = 30) +
  labs(title = "Distribution of Insurance Charges", x = "charges($)", y = "count")
```

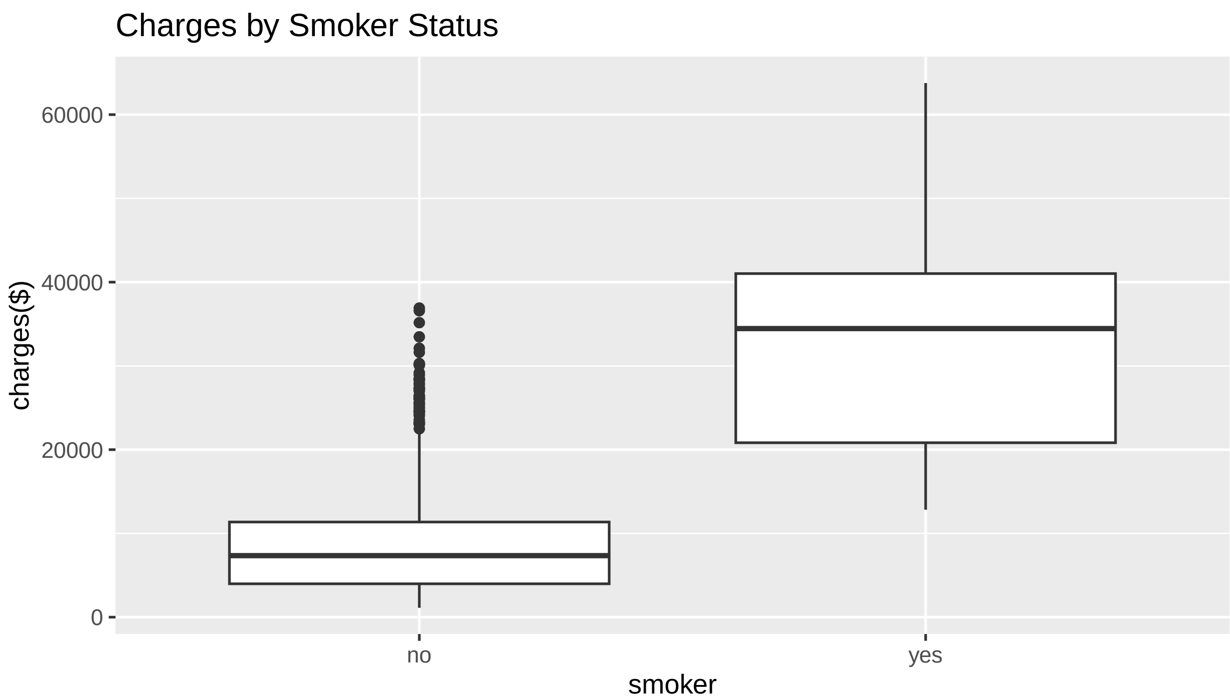



```r
ggplot(insurance, aes(bmi)) +
  geom_histogram(bins = 30) +
  labs(title = "BMI Distribution", x = "bmi", y = "count")
```

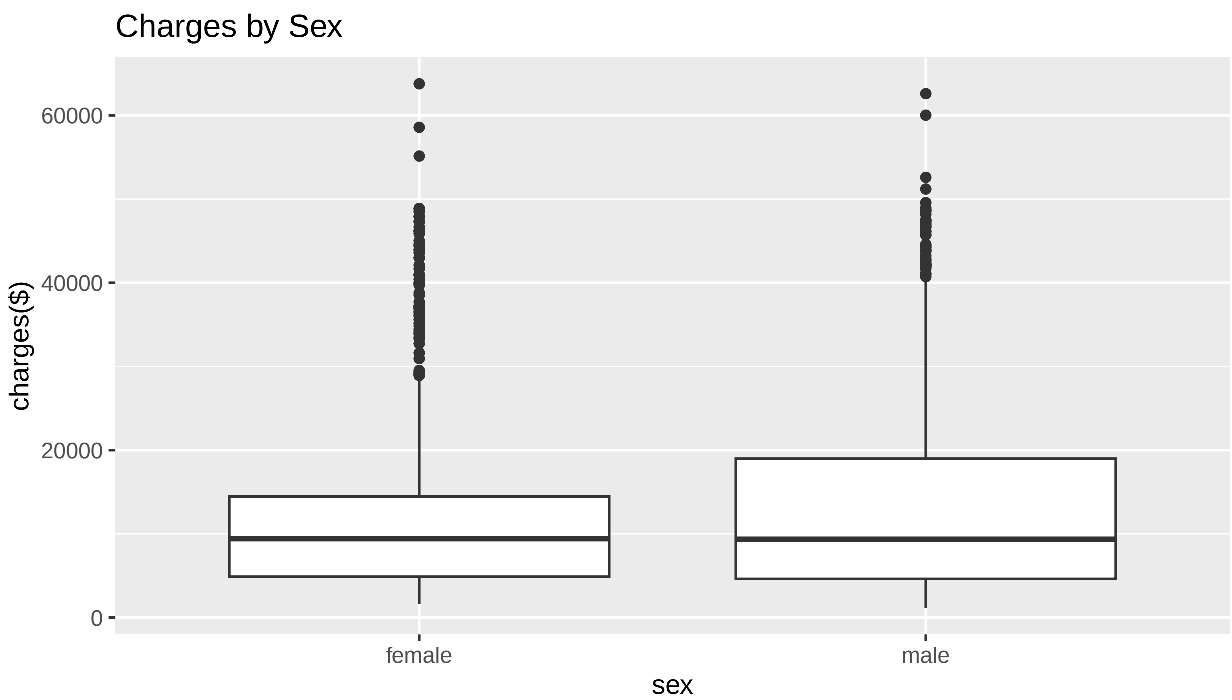


Charges by categories

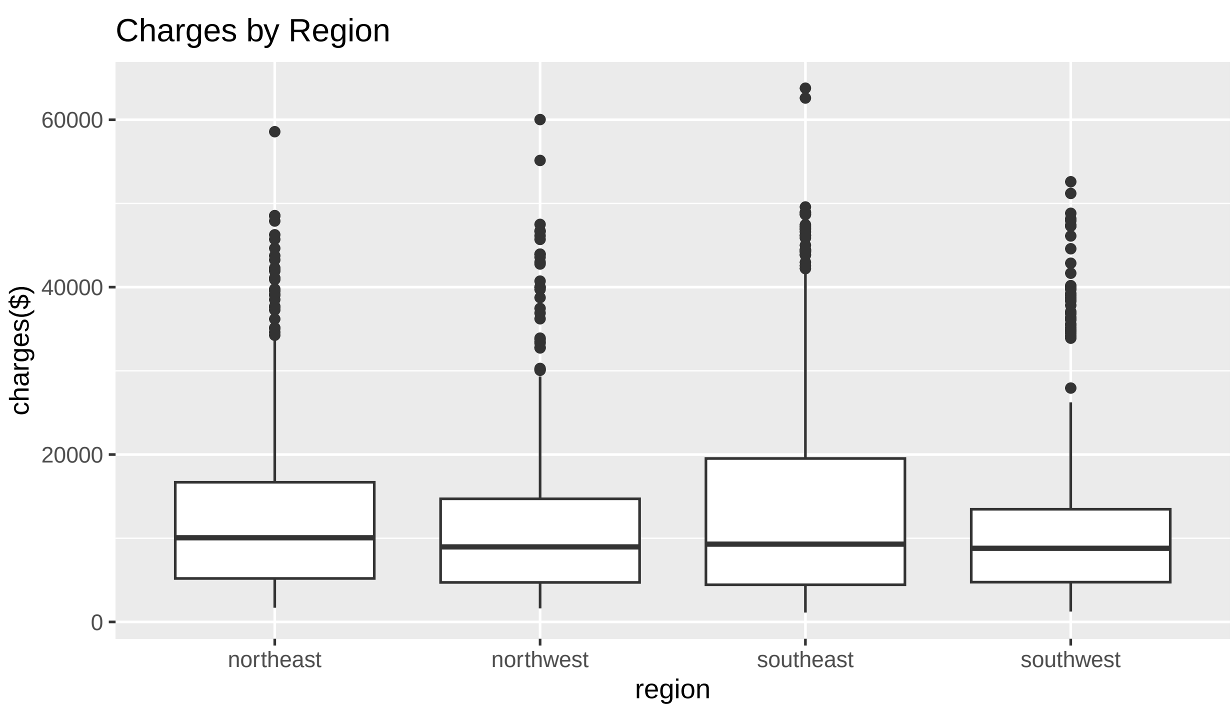
```
ggplot(insurance, aes(smoker, charges)) +
  geom_boxplot() +
  labs(title = "Charges by Smoker Status", x = "smoker", y = "charges($)")
```




```
ggplot(insurance, aes(sex, charges)) +
  geom_boxplot() +
  labs(title = "Charges by Sex", x = "sex", y = "charges($)")
```

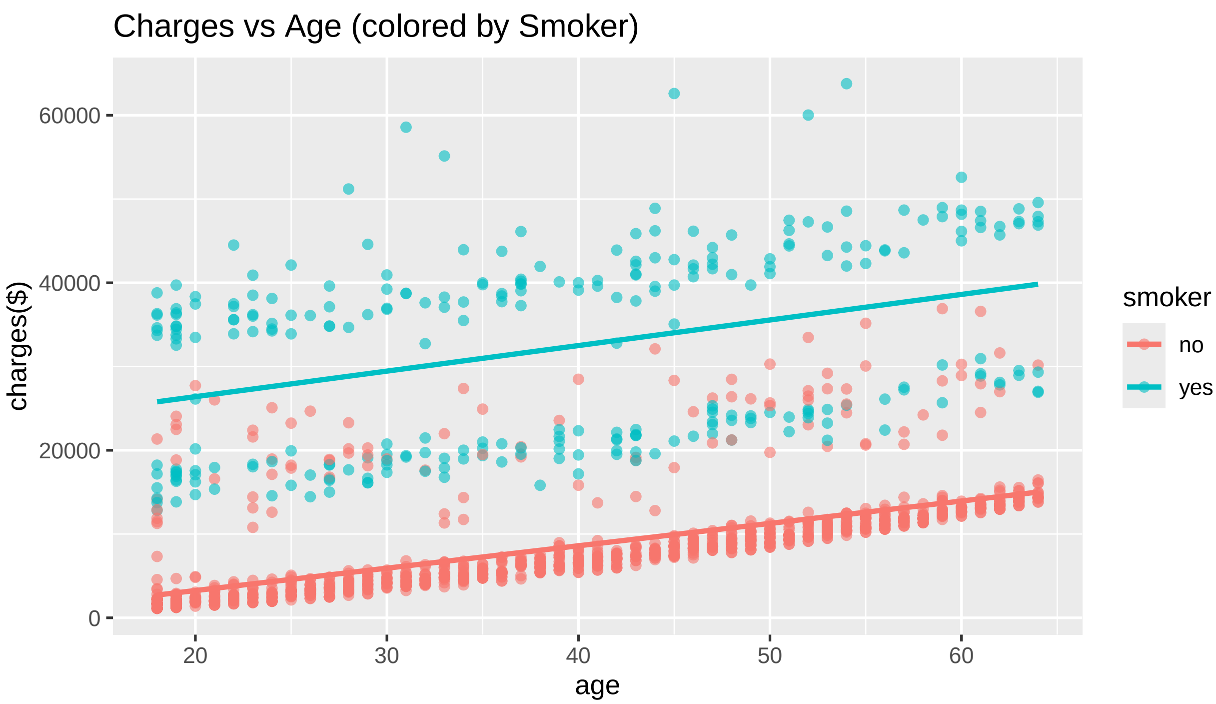



```r
ggplot(insurance, aes(region, charges)) +
  geom_boxplot() +
  labs(title = "Charges by Region", x = "region", y = "charges($)")
```

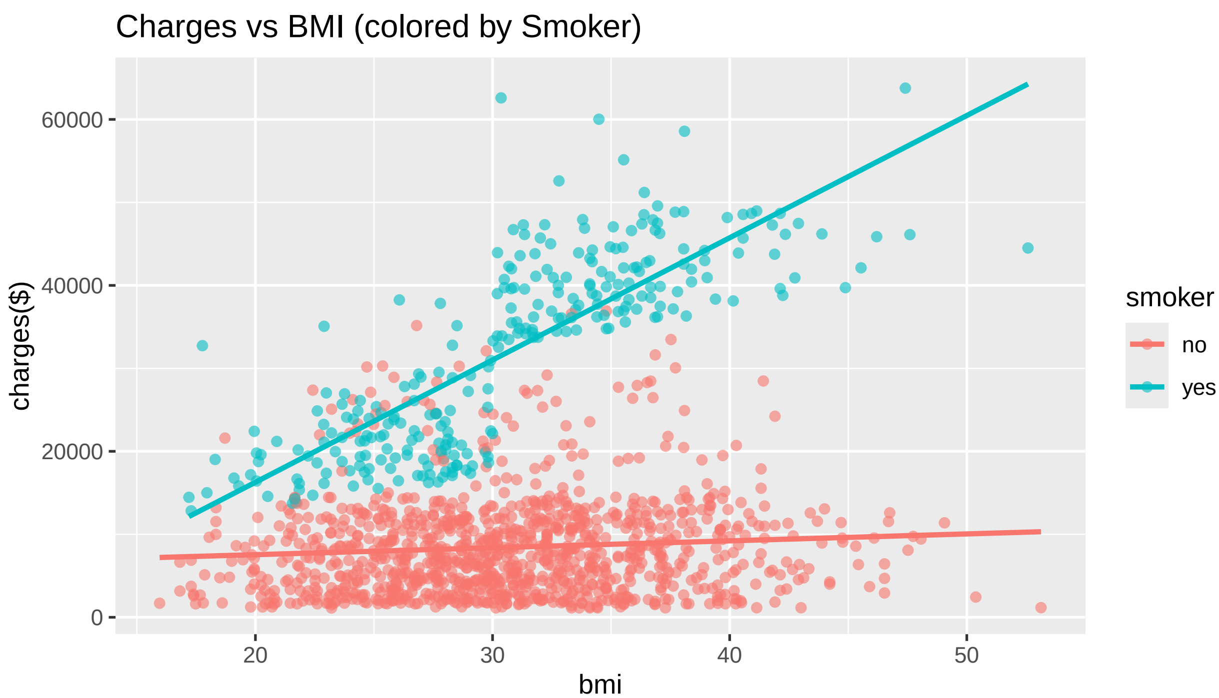


Relationships with age and BMI

```r
ggplot(insurance, aes(age, charges, color = smoker)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Charges vs Age (colored by Smoker)", x = "age", y = "charges($)")
```




```r
ggplot(insurance, aes(bmi, charges, color = smoker)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Charges vs BMI (colored by Smoker)", x = "bmi", y = "charges($)")
```



Modeling

We fit a **multiple linear regression** to explain `charges` as a function of demographics and habits. We include a simple interaction to let smoking modify the BMI effect.

```
set.seed(42)

n <- nrow(insurance)
idx <- sample.int(n, size = floor(0.8 * n))  # 80/20 split
train <- insurance[idx, ]
test  <- insurance[-idx, ]
```

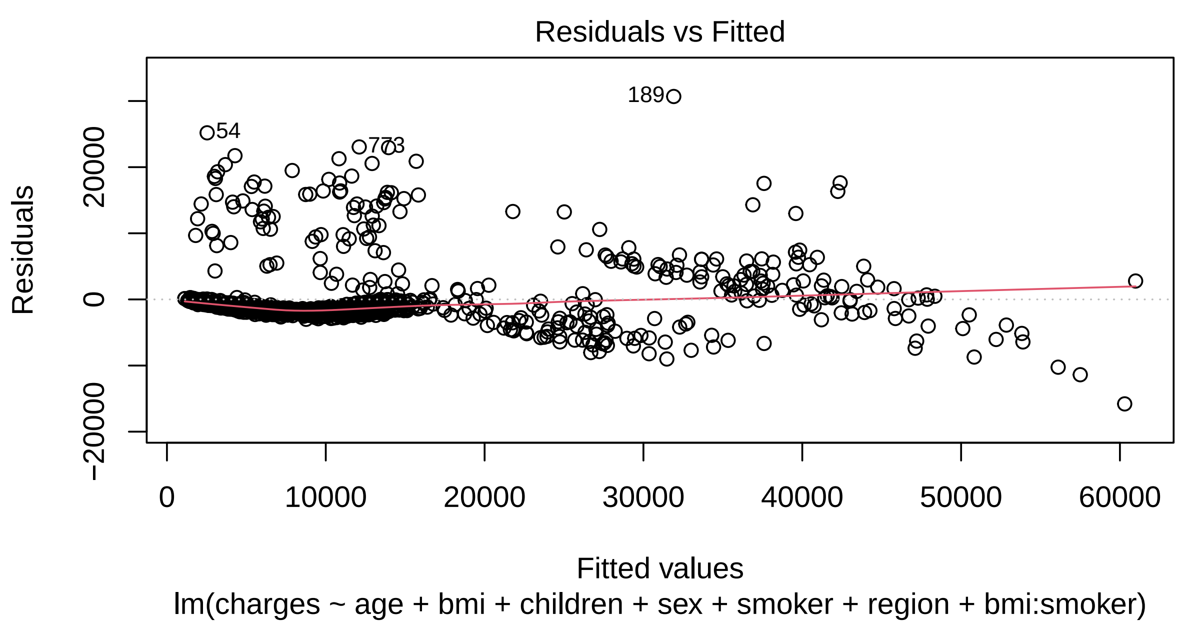
```
fit <- lm(charges ~ age + bmi + children + sex + smoker + region + bmi:smoker, data = train)
summary(fit)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + sex + smoker +
##     region + bmi:smoker, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15797.7  -1849.9  -1300.5   -386.4  30685.4
##
## Coefficients:
```

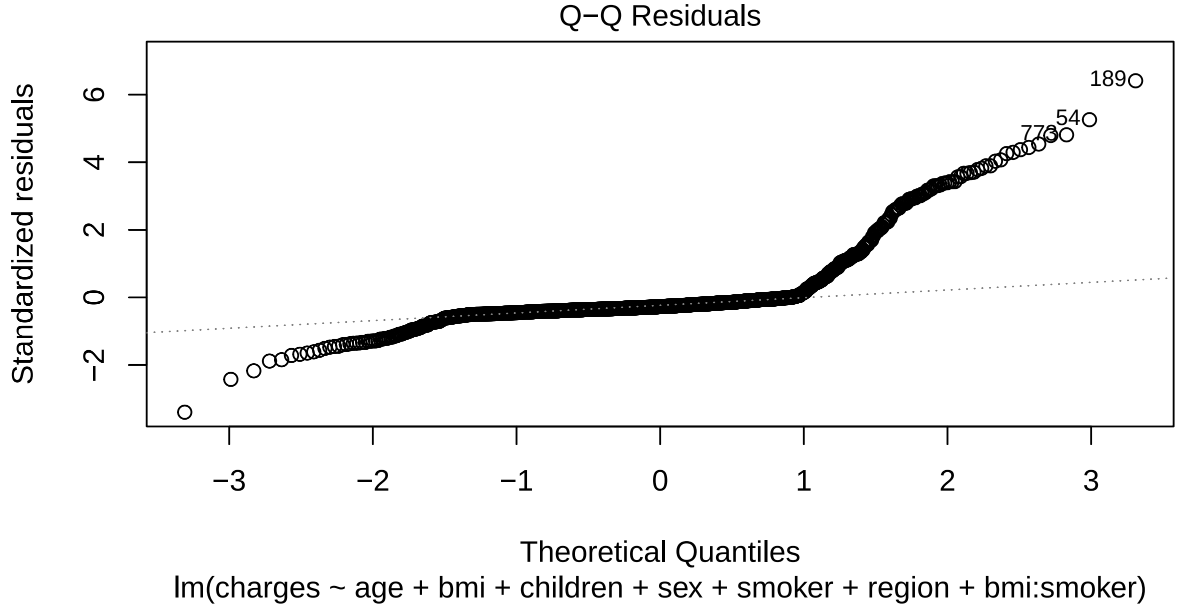
```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2719.00     962.57  -2.825 0.004821 **
## age                274.10      10.63  25.789  < 2e-16 ***
## bmi                 27.85      28.57   0.975 0.329823
## children           528.95     121.46   4.355 1.46e-05 ***
## sexmale           -395.20     296.16  -1.334 0.182350
## smokeryes       -22646.30    1859.62 -12.178  < 2e-16 ***
## regionnorthwest   -721.57     423.31  -1.705 0.088563 .
## regionsoutheast  -1349.73     426.68  -3.163 0.001604 **
## regionsouthwest  -1525.30     423.91  -3.598 0.000335 ***
## bmi:smokeryes     1509.81      59.30  25.461  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4807 on 1060 degrees of freedom
## Multiple R-squared:  0.8442, Adjusted R-squared:  0.8429
## F-statistic: 638.2 on 9 and 1060 DF,  p-value: < 2.2e-16
```

Model diagnostics (base plots)

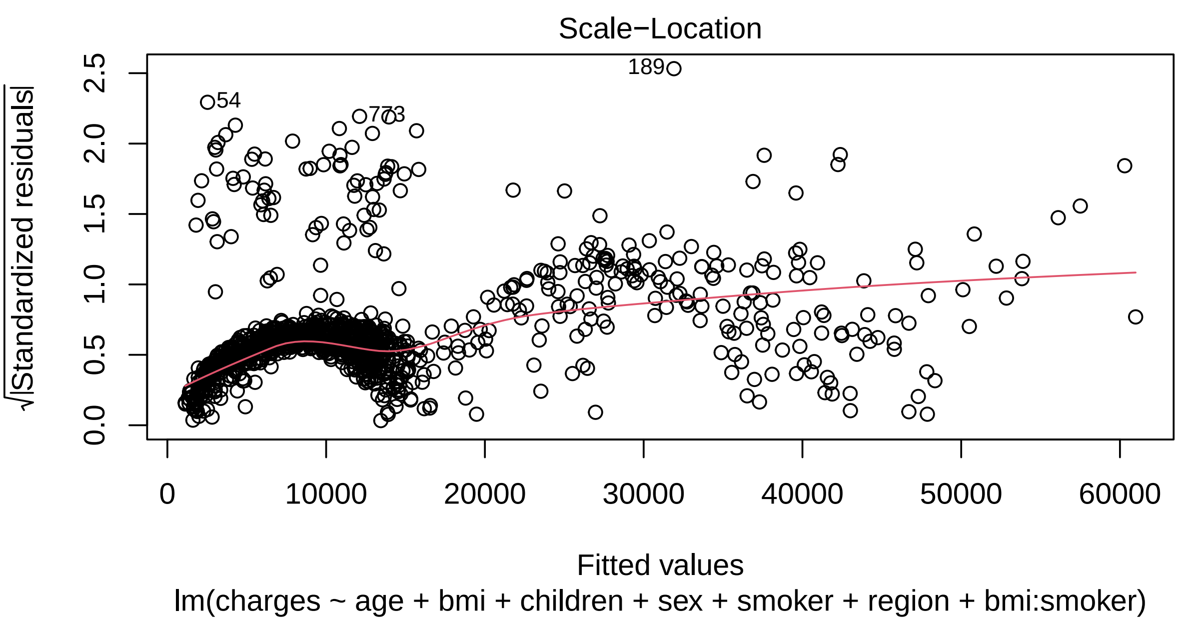
```r
plot(fit, which = 1)
```




```r
plot(fit, which = 2)
```




```r
plot(fit, which = 3)
```



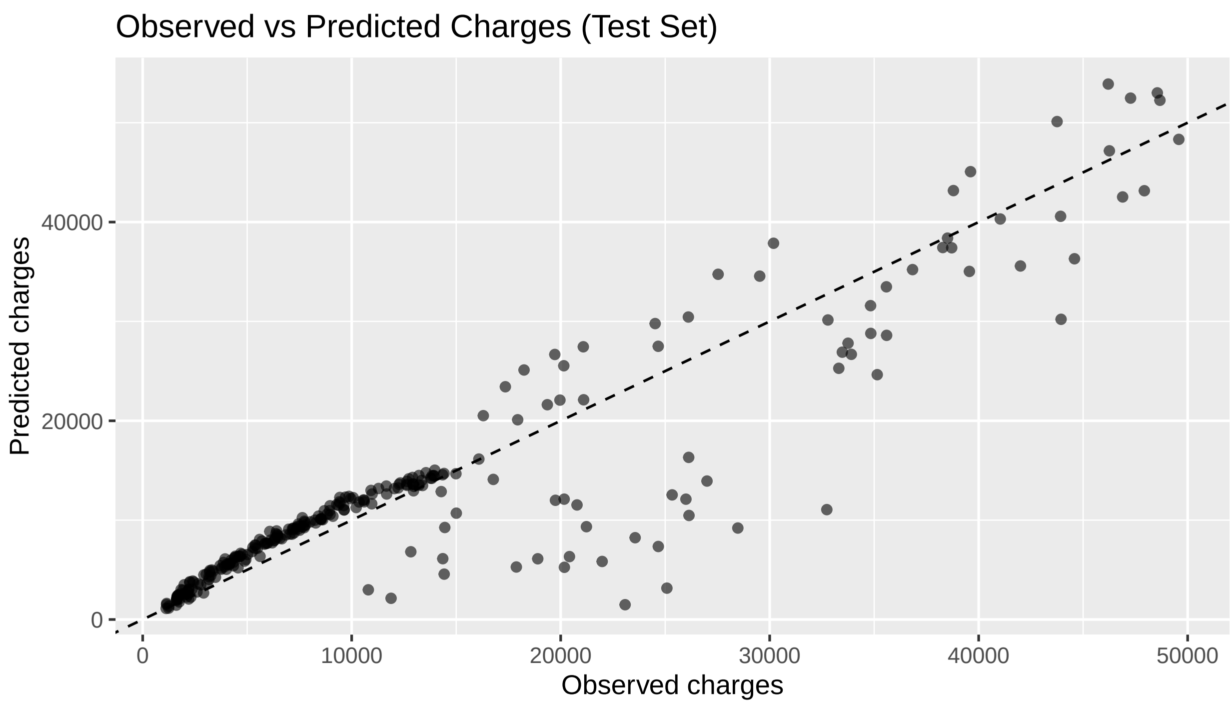
Test-set performance (manual RMSE and R^2)

```r
test$pred <- predict(fit, newdata = test)

rmse <- sqrt(mean((test$pred - test$charges)^2))

sse <- sum((test$charges - test$pred)^2)
sst <- sum((test$charges - mean(test$charges))^2)
rsq <- 1 - sse/sst

perf <- data.frame(
  .metric = c("rmse", "rsq"),
  .estimate = c(rmse, rsq)
)

kable(perf, digits = 4, caption = "Test-set performance (manual calculation)")
```

Table 3: Test-set performance (manual calculation)

.metric	.estimate
rmse	5042.3613
rsq	0.8245

```r
ggplot(test, aes(charges, pred)) +
  geom_point(alpha = 0.6) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed") +
  labs(title = "Observed vs Predicted Charges (Test Set)",
       x = "Observed charges", y = "Predicted charges")
```



Findings

The analysis indicates that **smoking status is the strongest driver of medical insurance charges**, showing a substantial positive association with costs. **Body mass index (BMI) and age** also exhibit positive relationships with charges, and an interaction between smoking and BMI suggests that the impact of smoking may vary depending on BMI levels. For modeling purposes, it is advisable to **log-transform the `charges` variable** if diagnostic plots reveal heteroscedastic residuals, as this transformation can help stabilize variance and improve model fit.

Conclusion

In conclusion, both demographic characteristics and lifestyle choices are key contributors to medical insurance costs. Among these, smoking stands out as the most influential factor, showing a strong positive association with higher charges. Age and body mass index (BMI) also play important roles, with older individuals and those with higher BMI generally incurring greater expenses. Together, these findings highlight the significant impact of personal health behaviors and demographic profiles on insurance charges.