Linear Regression

Application: Predicting health insurance expenses

Simon Lee, Jack Taiclet, Eric Li

4/23/2024

1. Overview

To set premiums for their beneficiaries, health insurers develop models to accurately forecast medical expenses, a challenge to estimate because the most costly conditions are rare and often occur randomly. Some conditions are more prevalent for certain segments of the population (e.g., lung cancer is more common among smokers than non-smokers).

The goal of this analysis is to use anonymized patient data to estimate average medical care costs based on individual characteristics. These estimates could be used by private insurers to set prices, or by government payers (e.g., Medicare) to predict costs.

2. Data Collection

We will use a simulated dataset containing hypothetical medical expenses for patients in the United States. This data was created using demographic statistics from the US Census Bureau, and thus, approximately reflect real-world conditions.

The insurance.csv file includes 1,338 beneficiaries currently enrolled in the insurance plan, with characteristics of the patient and total medical expenses charged to the plan for the calendar year. The variables are:

- **expenses:** Annual medical costs charged to the insurance plan.
- **age:** Age of the primary beneficiary (excluding those 65+, who are typically covered by Medicare).
- **sex:** Categorical variable for biological sex (male or female).
- **bmi:** Body mass index (BMI), which equals weight (in kg) divided by height (in meters) squared. An ideal BMI is between 18.5 and 24.9.
- **children:** Number of children/dependents covered by the insurance plan.
- **smoker:** Categorical variable for whether the individual regularly smokes tobacco (yes or no).
- **region:** Categorical variable for geographic residence in the US: northeast, southeast, southwest, or northwest.

It is important to think about how these variables may be related to medical expenses. For example, we might expect that older people and smokers tend to have higher medical expenses. In regression analysis, relationships among the variables are specified by the user rather than being detected automatically, as with machine learning.

3. Data Exploration

Let's first load some useful libraries. We add the option 'message = FALSE' to suppress printing the code output.

```
library(tidyverse)
library(formattable)
library(jtools)
library(stargazer)
```

Use read.csv() to load the insurance data for analysis. We add the option ‘stringsAsFactors = TRUE’ to convert the 3 categorical variables to factors:

```
setwd("/Users/simonlee/UCLA-Grad-Courses/MGMT298/p1/")
INSURANCE = read.csv("insurance.csv", stringsAsFactors = TRUE)
```

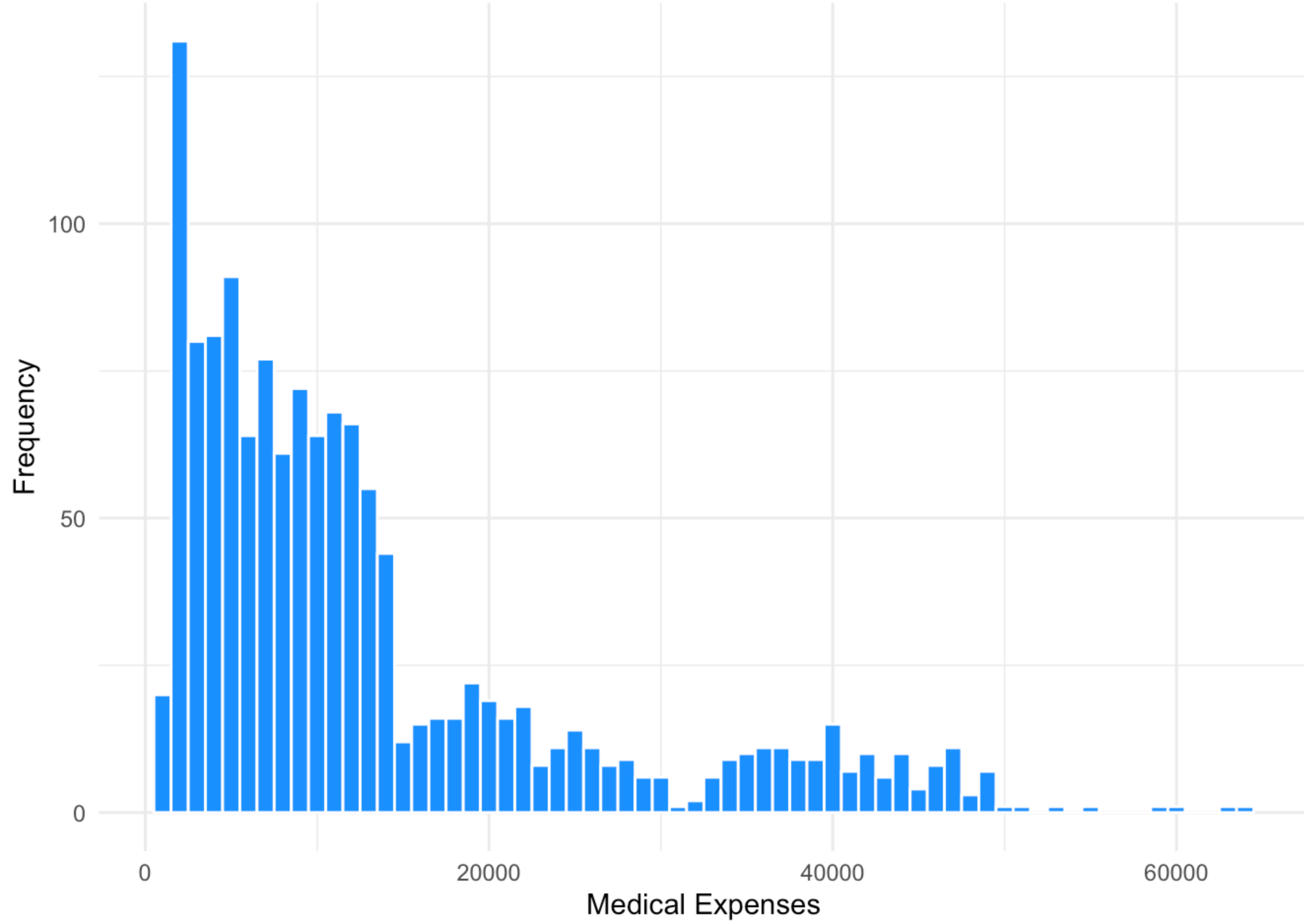
QUESTION 1: What are the mean and median medical expenses? What does this tell you about the distribution? Use ggplot to create a histogram of expenses (you can try different colors https://bookdown.org/hneth/ds4psy/D-3-apx-colors-basics.html (https://bookdown.org/hneth/ds4psy/D-3-apx-colors-basics.html))

```
summary(INSURANCE$expenses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4740    9382   13270   16640   63770
```

The mean of the medical expenses are $13270 and the median is $9382. The data is skewed due to the outliers. Median is a more robust statistic so it captures a better “middle” value (50% percentile).

```
ggplot(data = INSURANCE, aes(x = expenses)) +
  geom_histogram(fill = "dodgerblue", color = "white", binwidth = 1000) +
  labs(x = "Medical Expenses", y = "Frequency") +
  theme_minimal()
```



QUESTION 2: Create a correlation matrix for the 4 numeric variables. Which 2 variables are most strongly correlated with each other?

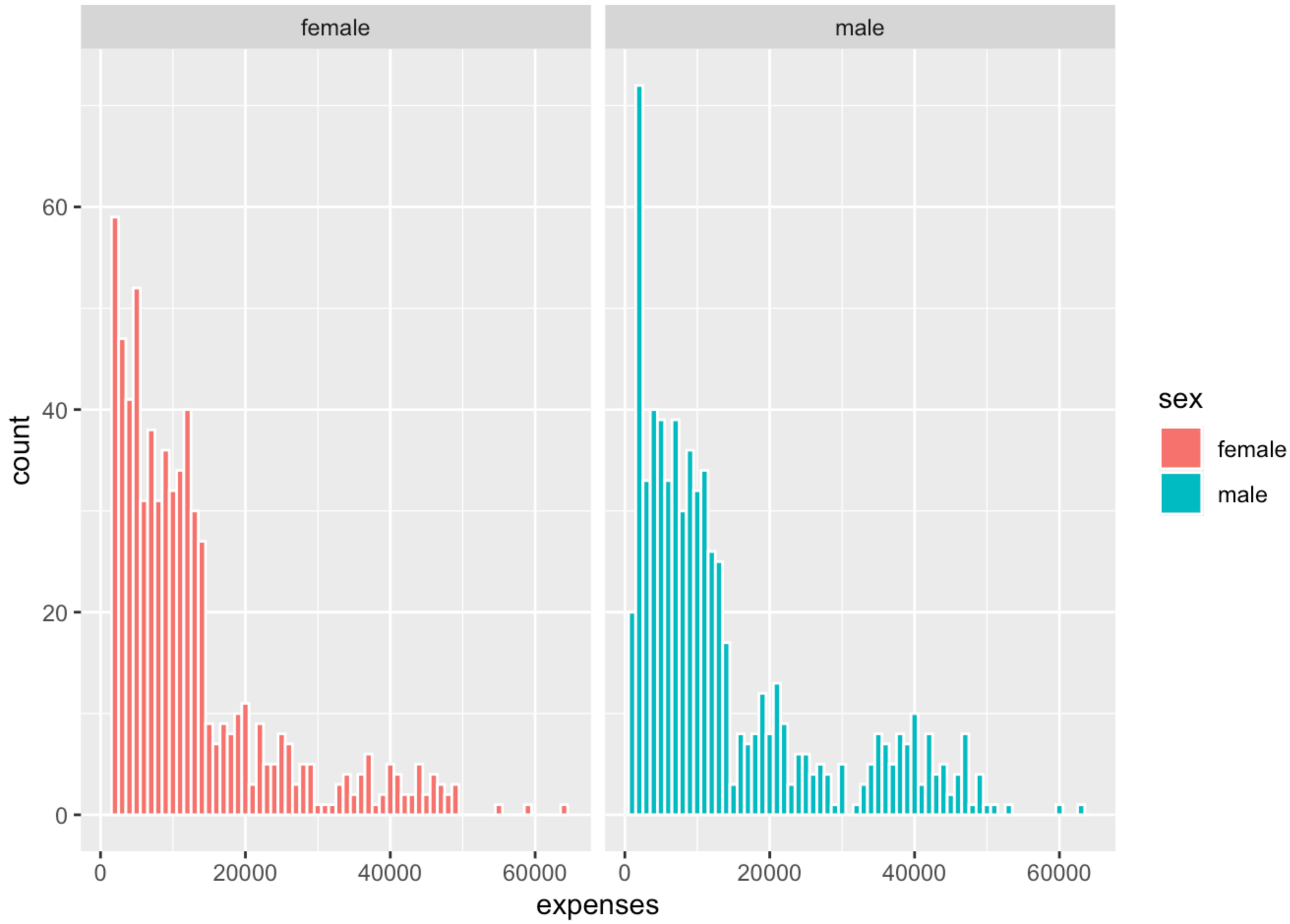
```
round(cor(INSURANCE[c("age", "bmi", "children", "expenses")]), 2)
```

```
##           age  bmi children expenses
## age      1.00 0.11     0.04     0.30
## bmi      0.11 1.00     0.01     0.20
## children 0.04 0.01     1.00     0.07
## expenses 0.30 0.20     0.07     1.00
```

The two most highly correlated variables are age and expenses.

QUESTION 3: Plot side-by-side histograms of expenses, by gender (using facet_wrap).

```
ggplot(data=INSURANCE, aes(x=expenses, fill=sex)) +
  geom_histogram(color="white", binwidth=1000) +
  facet_wrap(~sex)
```



QUESTION 4: Use a t-test to test whether mean expenses differ, on average, by sex.

```
t.test(expenses ~ sex, data=INSURANCE)
```

```
##
##  Welch Two Sample t-test
##
## data:  expenses by sex
## t = -2.1009, df = 1313.4, p-value = 0.03584
## alternative hypothesis: true difference in means between group female and group male
is not equal to 0
## 95 percent confidence interval:
##  -2682.48932   -91.85535
## sample estimates:
## mean in group female   mean in group male
##             12569.58             13956.75
```

QUESTION 5: Create a new dataframe called SMOKE with counts of beneficiaries by sex and smoking status. You can view this using the formattable package. What do you notice? Use a proportion-test to test whether smoking rates also differ by sex.

```
SMOKE = INSURANCE %>%
  group_by(sex) %>%
  summarize(total = n(), smokers = sum(smoker == "yes"))

SMOKE$proportion <- SMOKE$smokers / SMOKE$total

formattable(SMOKE)
```

sex	total	smokers	proportion
female	662	115	0.1737160
male	676	159	0.2352071

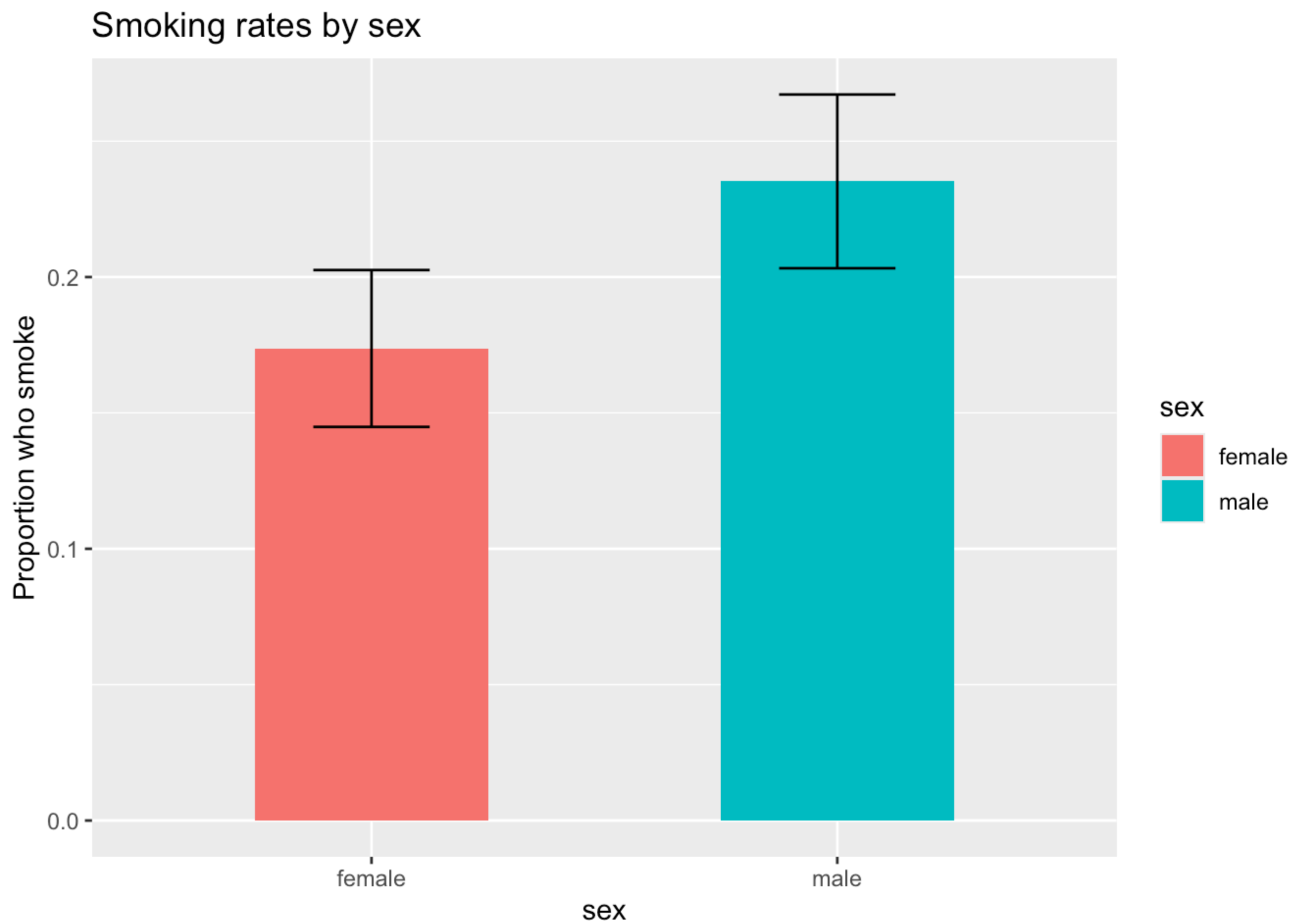
I notice that there is a slightly higher proportion of smokers in males than females.

```
prop.test(SMOKE$smokers, SMOKE$total)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  SMOKE$smokers out of SMOKE$total
## X-squared = 7.3929, df = 1, p-value = 0.006548
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.10605743 -0.01692475
## sample estimates:
##    prop 1    prop 2
## 0.1737160 0.2352071
```

QUESTION 6: Create a column chart with smoking rates, by sex, and add 95% confidence intervals.

```
ggplot(data = SMOKE, aes(x = sex, y = proportion, fill = sex)) +
  geom_col(width = 0.5) +
  geom_errorbar(aes(ymin = proportion - 1.96 * sqrt(proportion * (1 - proportion) / tota
l),
                    ymax = proportion + 1.96 * sqrt(proportion * (1 - proportion) / tota
l)),
                width = 0.25) +
  scale_y_continuous("Proportion who smoke") +
  ggtitle("Smoking rates by sex")
```



4. Linear Regression Analysis

We next use linear regression to examine the association between medical expenses and the independent variables.

QUESTION 7: Run a simple linear regression on just age (let's name this "regression1"). Use the summ() command from the jtools package to format the output. What is your interpretation of the coefficient on age? Is this statistically significant at the 5% level?

```
regression1 <- lm(expenses ~ age, data = INSURANCE)
summ(regression1, digits=3)
```

```
## MODEL INFO:
## Observations: 1338
## Dependent Variable: expenses
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,1336) = 131.174, p = 0.000
## R² = 0.089
## Adj. R² = 0.089
##
## Standard errors: OLS
## ------------------------------------------------------
##                         Est.      S.E.   t val.       p
## ----------------- ---------- --------- -------- -------
## (Intercept)         3165.885   937.149    3.378   0.001
## age                  257.723    22.502   11.453   0.000
## ------------------------------------------------------
```

The older you are the more it will cost you medically and this is pretty statistically significant given the p value. (Actually we don't necessarily know because of the rounding.)

QUESTION 8: Run a regression with with indepdendent variables age, children, BMI, sex, and smoking status (let's name this "regression2"). Do men have higher or lower expenses, holding all other variables constant? What about smokers? Is this consistent with your earlier t-test? What might explain this?

```
regression2 = lm(expenses ~ age + children + bmi + sex +smoker, data = INSURANCE)
summ(regression2, digits=3)
```

```
## MODEL INFO:
## Observations: 1338
## Dependent Variable: expenses
## Type: OLS linear regression
##
## MODEL FIT:
## F(5,1332) = 798.019, p = 0.000
## R² = 0.750
## Adj. R² = 0.749
##
## Standard errors: OLS
## ---------------------------------------------------------
##                           Est.      S.E.    t val.       p
## ----------------- ------------ --------- --------- -------
## (Intercept)         -12052.462   951.260   -12.670   0.000
## age                    257.735    11.904    21.651   0.000
## children               474.411   137.856     3.441   0.001
## bmi                    322.364    27.419    11.757   0.000
## sexmale               -128.640   333.361    -0.386   0.700
## smokeryes            23823.393   412.523    57.750   0.000
## ---------------------------------------------------------
```

Men have a lower expense. This is not consistent with the previous t test but this can be due to including all variables in the dataframe in the linear regression whereas the t test only computed from male to female.

QUESTION 9: Run another regression, adding region as an independent variable (let's name this "regression3"). Which geographic region has the highest medical expenses, controlling for the other variables?

```
regression3 = lm(expenses ~ age + children + bmi + sex + smoker + region, data = INSURAN
CE)
summ(regression3, digits=3)
```

```
## MODEL INFO:
## Observations: 1338
## Dependent Variable: expenses
## Type: OLS linear regression
##
## MODEL FIT:
## F(8,1329) = 500.811, p = 0.000
## R² = 0.751
## Adj. R² = 0.749
##
## Standard errors: OLS
## ---------------------------------------------------------------
##                              Est.      S.E.    t val.       p
## --------------------- ------------ --------- --------- -------
## (Intercept)            -11938.539   987.819   -12.086   0.000
## age                       256.856    11.899    21.587   0.000
## children                  475.501   137.804     3.451   0.001
## bmi                       339.193    28.599    11.860   0.000
## sexmale                  -131.314   332.945    -0.394   0.693
## smokeryes               23848.535   413.153    57.723   0.000
## regionnorthwest          -352.964   476.276    -0.741   0.459
## regionsoutheast         -1035.022   478.692    -2.162   0.031
## regionsouthwest          -960.051   477.933    -2.009   0.045
## ---------------------------------------------------------------
```

The region with the highest expense is northwest which spend roughly $-352 on average per every step in the x-axis based on the table. But the north east has 0 negative expenses so it is in theory spending the most.

QUESTION 10: Use the stargazer package to compare model performance. What fraction of the variation in medical expenses is explained by variation in these 6 variables in regression3?

```
stargazer(regression1, regression2, regression3, type ="text", digits = 2)
```

```
##
## ==========================================================================================
=========
##                                                      Dependent variable:
##                        -----------------------------------------------------------------
---------
##                                                            expenses
##                                  (1)                          (2)                         (3)
## ----------------------------------------------------------------------------------------
---------
## age                           257.72***                    257.73***                   256.86**
*
##                               (22.50)                      (11.90)                     (11.90)
##
## children                                                  474.41***                   475.50**
*
##                                                           (137.86)                    (137.8
0)
##
## bmi                                                       322.36***                   339.19**
*
##                                                           (27.42)                     (28.60)
##
## sexmale                                                   -128.64                     -131.31
##                                                           (333.36)                    (332.9
5)
##
## smokeryes                                               23,823.39***                  23,848.53
***
##                                                           (412.52)                    (413.1
5)
##
## regionnorthwest                                                                       -352.96
##                                                                                       (476.2
8)
##
## regionsoutheast                                                                      -1,035.02
**
##                                                                                       (478.6
9)
##
## regionsouthwest                                                                       -960.05*
*
##                                                                                       (477.9
3)
##
## Constant                     3,165.89***               -12,052.46***               -11,938.54
***
##                              (937.15)                     (951.26)                    (987.8
2)
##
## ----------------------------------------------------------------------------------------
```

```
---------
## Observations                    1,338                       1,338                       1,338
## R2                               0.09                        0.75                        0.75
## Adjusted R2                      0.09                        0.75                        0.75
## Residual Std. Error  11,560.31 (df = 1336)     6,069.73 (df = 1332)     6,062.10 (df
= 1329)
## F Statistic         131.17*** (df = 1; 1336) 798.02*** (df = 5; 1332) 500.81*** (df =
8; 1329)
## ===================================================================================
=========
## Note:                                                                *p<0.1; **p<0.05;
***p<0.01
```

In regression3, the R-squared value is 0.75. This means that approximately 75% of the variation in medical expenses can be explained by the variation in the six independent variables included in the model.

QUESTION 11: Let's go back to regression2, but add an interaction term for smoker and BMI. What is the interpretation of this coefficient? Let's also create a scatterplot for insurance expenses, showing lines for smokers and non-smokers.

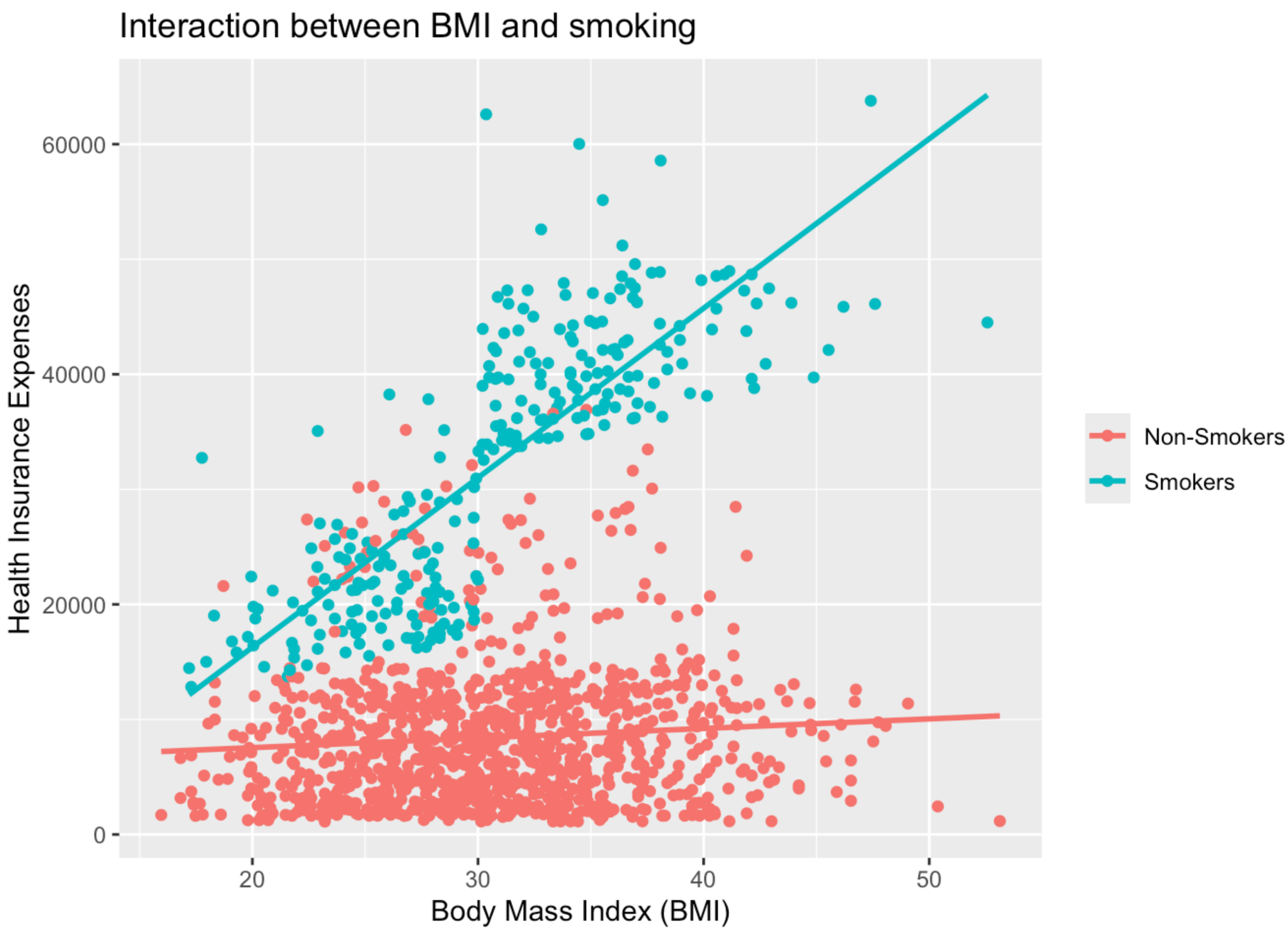
the coefficient for the interaction term captures how the relationship between BMI and medical expenses differs between smokers and non-smokers. It provides insight into whether the effect of BMI on medical expenses varies depending on smoking status.

```
regression4 = lm(expenses ~ age + children + bmi + sex + smoker + smoker * bmi, data = I
NSURANCE)
summ(regression4, digits=3)
```

```
## MODEL INFO:
## Observations: 1338
## Dependent Variable: expenses
## Type: OLS linear regression
##
## MODEL FIT:
## F(6,1331) = 1158.172, p = 0.000
## R² = 0.839
## Adj. R² = 0.839
##
## Standard errors: OLS
## ---------------------------------------------------------------
##                              Est.        S.E.    t val.       p
## ------------------- ------------ ---------- ---------- -------
## (Intercept)            -2503.041     839.433    -2.982   0.003
## age                      264.531       9.547    27.709   0.000
## children                 512.546     110.531     4.637   0.000
## bmi                        6.545      24.855     0.263   0.792
## sexmale                 -495.458     267.603    -1.851   0.064
## smokeryes             -20299.695    1653.971   -12.273   0.000
## bmi:smokeryes           1438.713      52.842    27.227   0.000
## ---------------------------------------------------------------
```

```
ggplot(data=INSURANCE, aes(x=bmi, y=expenses, color=smoker)) +
  geom_point() + geom_smooth(method="lm", se=FALSE) +
  scale_x_continuous("Body Mass Index (BMI)") +
  scale_y_continuous("Health Insurance Expenses") +
  ggtitle("Interaction between BMI and smoking") +
  scale_color_discrete(name = NULL, labels = c("Non-Smokers", "Smokers"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



QUESTION 12: Let's predict medical expenses using regression2 and regression4 for a 59-year old female with 2 children, BMI = 35, and smoker. What is a 95% prediction interval?

```
CUSTOMER = data.frame(age=59, sex="female", bmi=35, children=2, smoker="yes")
predict(regression2, CUSTOMER, interval="predict", level = 0.95)
```

```
##        fit      lwr      upr
## 1 39208.86 27260.98 51156.75
```

```
predict(regression4, CUSTOMER, interval="predict", level = 0.95)
```

```
##        fit      lwr      upr
## 1 44413.72 34827.47 53999.98
```

The 95% confidence intervals are displayed in the table above. To interpret is we can say that 95% of the prediction will fall into this range.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this: