STAT 308 – Chapter 11-12

Background Information

We have introduced the concept of multiple linear regression and how we can perform hypothesis tests on the significance of adding additional explanatory variables to the predictive performance of our overall model. Previously, we have only explored the inclusion of quantitative/numeric explanatory variables. In these lectures, we will introduce how to include qualitative/categorical explanatory variables to our linear regression models.

Motivating Example

Consider the `insurance.csv`, a dataset containing information on personal medical costs not covered by health insurance which includes

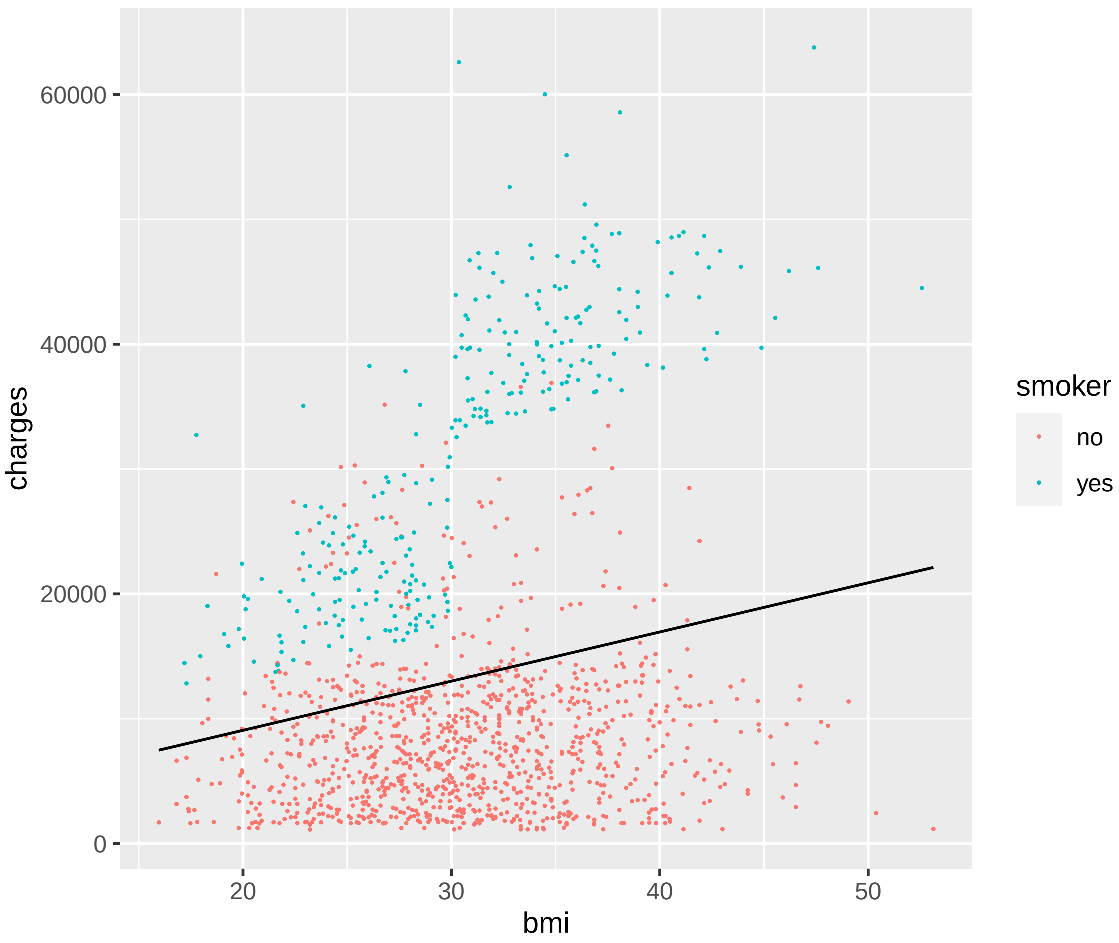
- `age`: Age of primary beneficiary
- `sex`: Gender of insurance contractor: male/female
- `bmi`: Body mass index (in kg/m^2), an understanding of body weights that are high or low relative to height
- `children`: Number of children covered by the health insurance contracts
- `smoker`: Yes/No on whether the primary beneficiary is a smoker
- `region`: The beneficiary's residential area in the US, northeast, southeast, southwest, northwest
- `charges`: Individual medical costs billed by health insurance

We wish to create a linear model for the individual's personal medical costs based on a subset of the additional variables in the model.

Suppose we already have knowledge that bmi is related linearly with individual medical charges. More specifically, we have

$$\hat{charges} = \beta_0 + \beta_1 bmi$$

```r
library(tidyverse)
insurance <- read.csv("../Data/insurance.csv")
mod <- lm(charges ~ bmi,insurance)
insurance %>%
  mutate(pred_charges = mod$fitted.values) %>%
  ggplot(aes(x=bmi,y=charges,colour=smoker)) +
  geom_point(size=0.1) +
  geom_line(aes(y=pred_charges),colour="black")
```



We wish to incorporate `smoker` into our linear model because of our belief that smokers, on average, spend more in medical bills on an annual basis. We have not discussed how to incorporate categorical variables into linear models.

Dummy/Indicator Variable: a variable that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

For a singular categorical variable with k different categories, $k-1$ different categorical variables must be added to our model

For our example where we include whether or not the primary beneficiary is a smoker, our new linear model is

$$\hat{charges} = \beta_0 + \beta_1 bmi + \beta_2 smoker_{yes}$$

where

$$smoker_{yes} = \begin{cases} 1 & \text{the primary benificiary is a smoker} \\ 0 & \text{the primary benificiary is not a smoker} \end{cases}$$

```
mod2 <- lm(charges ~ bmi + smoker,insurance)
head(model.matrix(mod2))
```

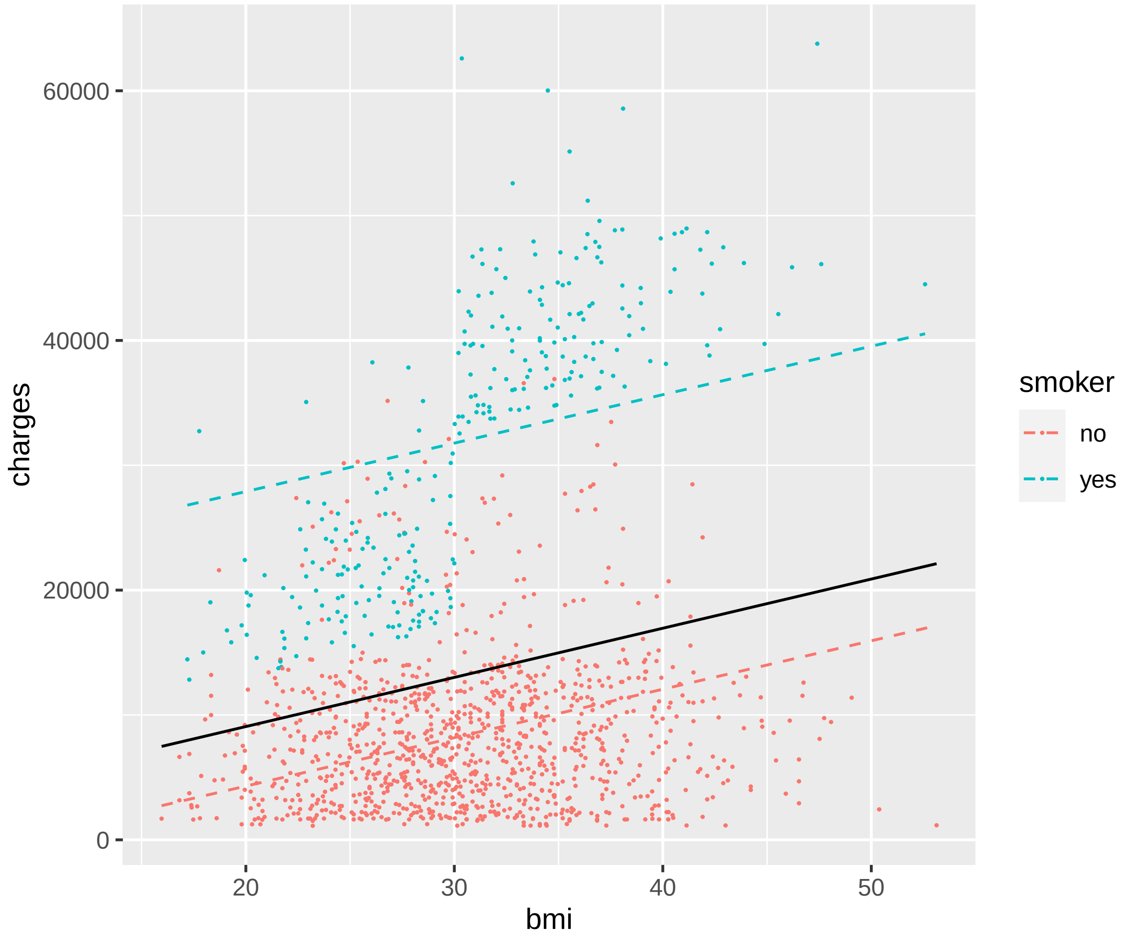
```
##   (Intercept)    bmi smokeryes
## 1           1 27.900        1
## 2           1 33.770        0
## 3           1 33.000        0
## 4           1 22.705        0
## 5           1 28.880        0
## 6           1 25.740        0
```

How would we perform a hypothesis test to determine if knowing whether or not the primary benficiary is a smoker improves the predictive ability of our linear model?

Because $smoker_{yes}$ is an indicator variable, we essentially have two different linear models, one for each level of smoker:

- For smokers: $\hat{charges} = (\beta_0 + \beta_2) + \beta_1 bmi$
- For non-smokers: $\hat{charges} = \beta_0 + \beta_1 bmi$

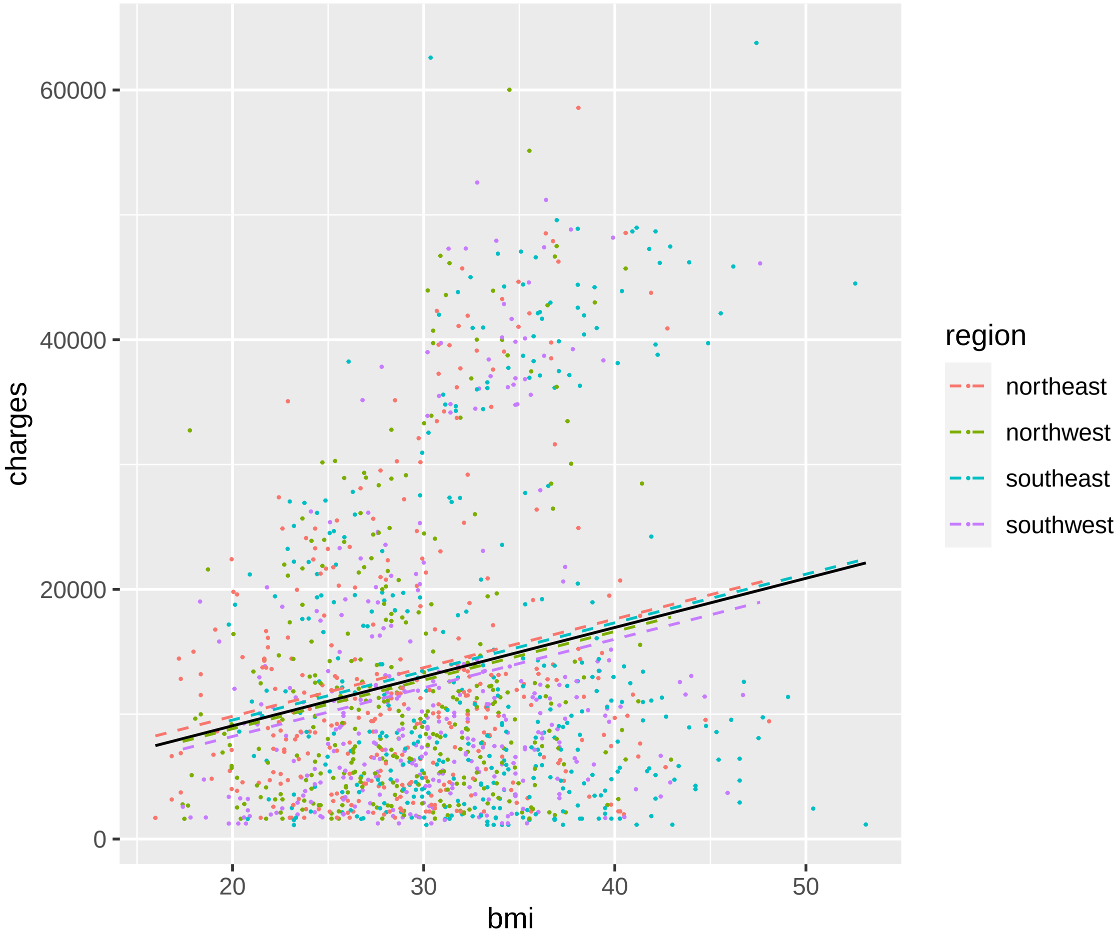
```
insurance %>%
  mutate(pred_charges = mod$fitted.values,
         pred_charges2 = mod2$fitted.values) %>%
  ggplot(aes(x=bmi,y=charges,colour=smoker)) +
  geom_point(size=0.1) +
  geom_line(aes(y=pred_charges),colour="black") +
  geom_line(aes(y=pred_charges2),lty=2)
```



Example

Suppose that for our linear model for `charges` that already includes `bmi`, perform a hypothesis test to determine if adding `region` to our model significantly improves the predictive ability?

```
mod2_region <-  lm(charges ~ bmi + region,insurance)
insurance %>%
  mutate(pred_charges = mod$fitted.values,
         pred_charges2 = mod2_region$fitted.values) %>%
  ggplot(aes(x=bmi,y=charges,colour=region)) +
  geom_point(size=0.1) +
  geom_line(aes(y=pred_charges),colour="black") +
  geom_line(aes(y=pred_charges2),lty=2)
```

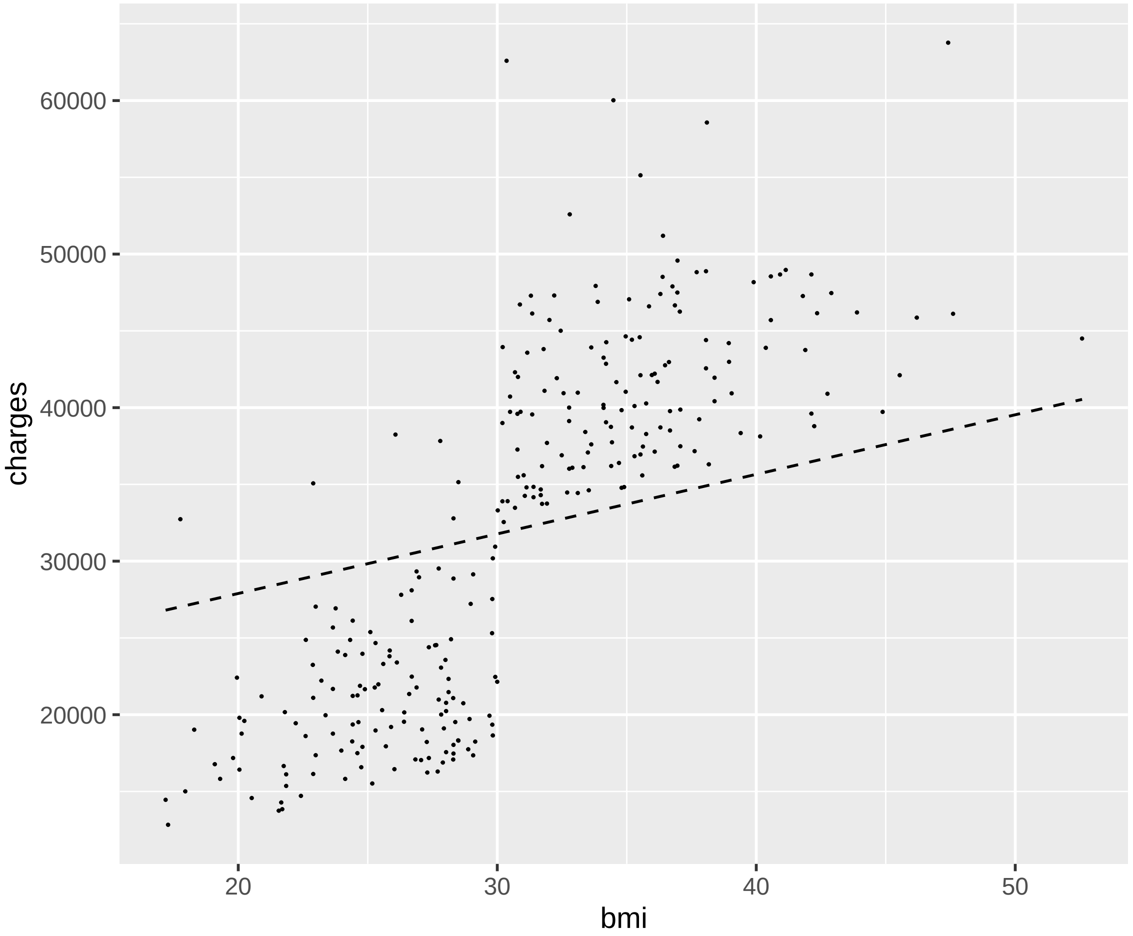


Motivating Example 2

We have determined that including whether or not the primary beneficiary is a smoker significantly improves our predictions for medical insurance charges.

However, let us observe the scatterplots and our least squares regression lines individually for smokers.

```
insurance %>%
  mutate(pred_charges2 = mod2$fitted.values) %>%
  filter(smoker=="yes") %>%
  ggplot(aes(x=bmi,y=charges)) +
  geom_point(size=0.1) +
  geom_line(aes(y=pred_charges2),lty=2)
```



Why does this regression line not seem to fit the data properly?

- The linear relationship between bmi and insurance charges may be **different** depending on whether or not the primary beneficiary is a smoker.

> **Interaction**: An effect in a linear model that quantifies the difference in a linear relationship between an one explanatory variable and the response variable when accounting for the value of another explanatory variable.

Recall for the insurance charges dataset, the model that we previously used was

$$\hat{charges} = \beta_0 + \beta_1 bmi + \beta_2 smoker_{yes}.$$

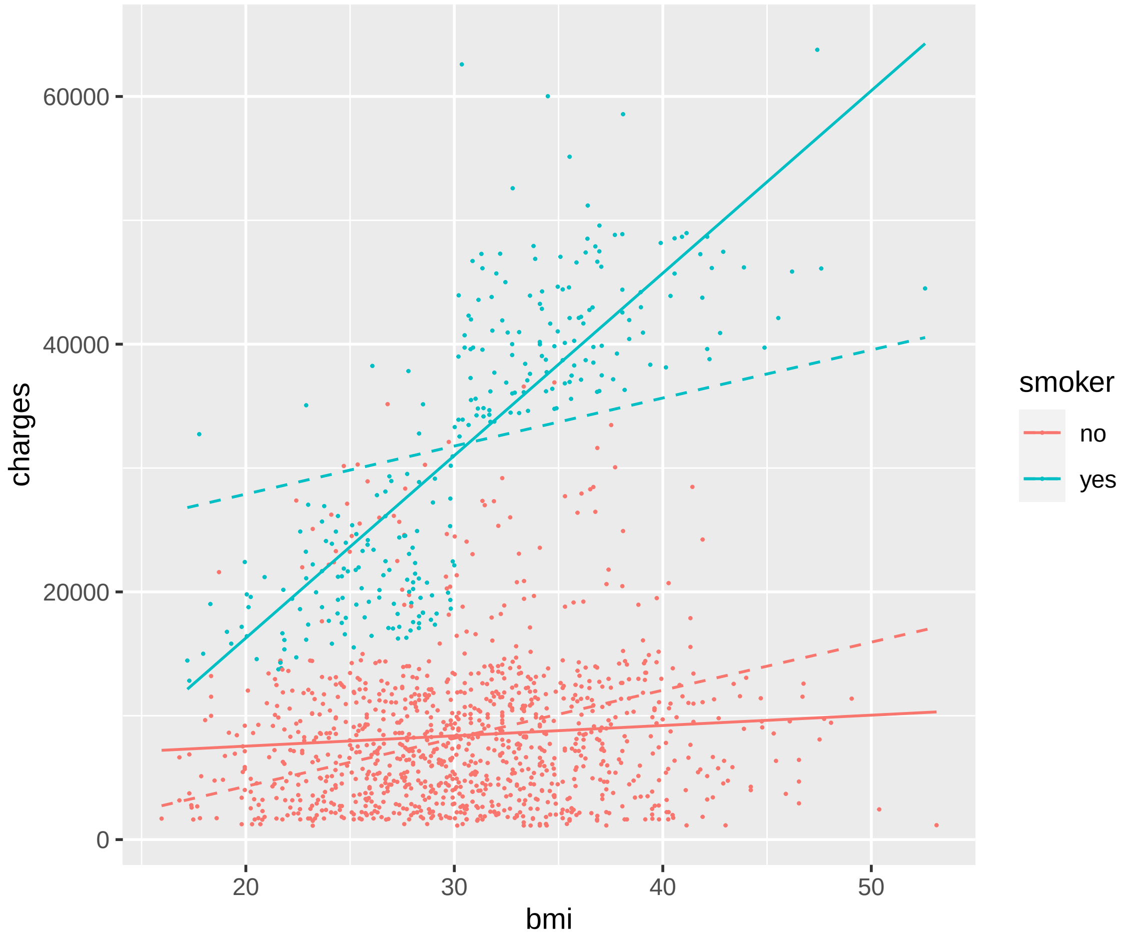
Adding an interaction term changes our linear model to be

$$\hat{charges} = \beta_0 + \beta_1 bmi + \beta_2 smoker_{yes} + \beta_3 (bmi \times smoker_{yes}).$$

We still have two different linear models for the two different levels of smoking:

- For smokers: $\hat{charges} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times bmi$
- For non-smokers: $\hat{charges} = \beta_0 + \beta_1 bmi$

```
mod2_inter <- lm(charges ~ bmi*smoker,insurance)
insurance %>%
  mutate(pred_charges = mod2$fitted.values,
         pred_charges2 = mod2_inter$fitted.values) %>%
  ggplot(aes(x=bmi,y=charges,colour=smoker)) +
  geom_point(size=0.1) +
  geom_line(aes(y=pred_charges),lty=2) +
  geom_line(aes(y=pred_charges2))
```



Example

Perform a hypothesis test to determine if the linear relationship between bmi and insurance charges is different for smokers vs. non-smokers.