PA 541 : Homework 2

Alexis Kwan

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
car_data <- read_csv("car_data.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   name = col_character(),
##   year = col_double(),
##   selling_price = col_double(),
##   km_driven = col_double(),
##   fuel = col_character(),
##   seller_type = col_character(),
##   transmission = col_character(),
##   owner = col_character()
## )
```

```r
insurance <- read_csv("insurance.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   age = col_double(),
##   sex = col_character(),
##   bmi = col_double(),
##   children = col_double(),
##   smoker = col_character(),
##   region = col_character(),
##   charges = col_double()
## )
```

```r
summary(car_data)
```

```
##      name                year      selling_price       km_driven
##  Length:4340        Min.   :1992   Min.   :  20000   Min.   :     1
##  Class :character   1st Qu.:2011   1st Qu.: 208750   1st Qu.: 35000
##  Mode  :character   Median :2014   Median : 350000   Median : 60000
##                     Mean   :2013   Mean   : 504127   Mean   : 66216
##                     3rd Qu.:2016   3rd Qu.: 600000   3rd Qu.: 90000
##                     Max.   :2020   Max.   :8900000   Max.   :806599
##      fuel           seller_type        transmission          owner
##  Length:4340        Length:4340        Length:4340        Length:4340
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

PART 1

Variable name (Description)

- name (Model of the car)
- year (Year of the car when it was bought)
- selling_price (Price at which the car is being sold in Indian Rupees)
- km_driven (Number of Kilometers the car is driven)
- fuel (Fuel type of car (petrol / diesel / CNG / LPG / electric))
- seller_type (Tells if a seller is Individual or a Dealer)
- transmission (Gear transmission of the car (Automatic/Manual))
- owner (Number of previous owners of the car.)

QUESTION 1 (10 pts)

a. What is the average selling price for automatic versus manual cars? (2 pts)
b. Of the automatic cars, which model was sold at the highest price? (2pts)
c. Plot the average selling price for each type of transmission. (3 pts)
d. Plot the relationship between selling price and year for the automatic and manual cars on the same plot. (3 pts)

```r
car_data %>%
  group_by(transmission) %>%
  summarise(avg_price = mean(selling_price))
```

```
## # A tibble: 2 x 2
##   transmission avg_price
##   <chr>            <dbl>
## 1 Automatic     1408154
## 2 Manual         400067.
```

The average selling price for automatic cars is about 1,000,000 more than manual cars.

```r
car_data %>%
  filter(transmission == "Automatic", selling_price == max(selling_price))
```

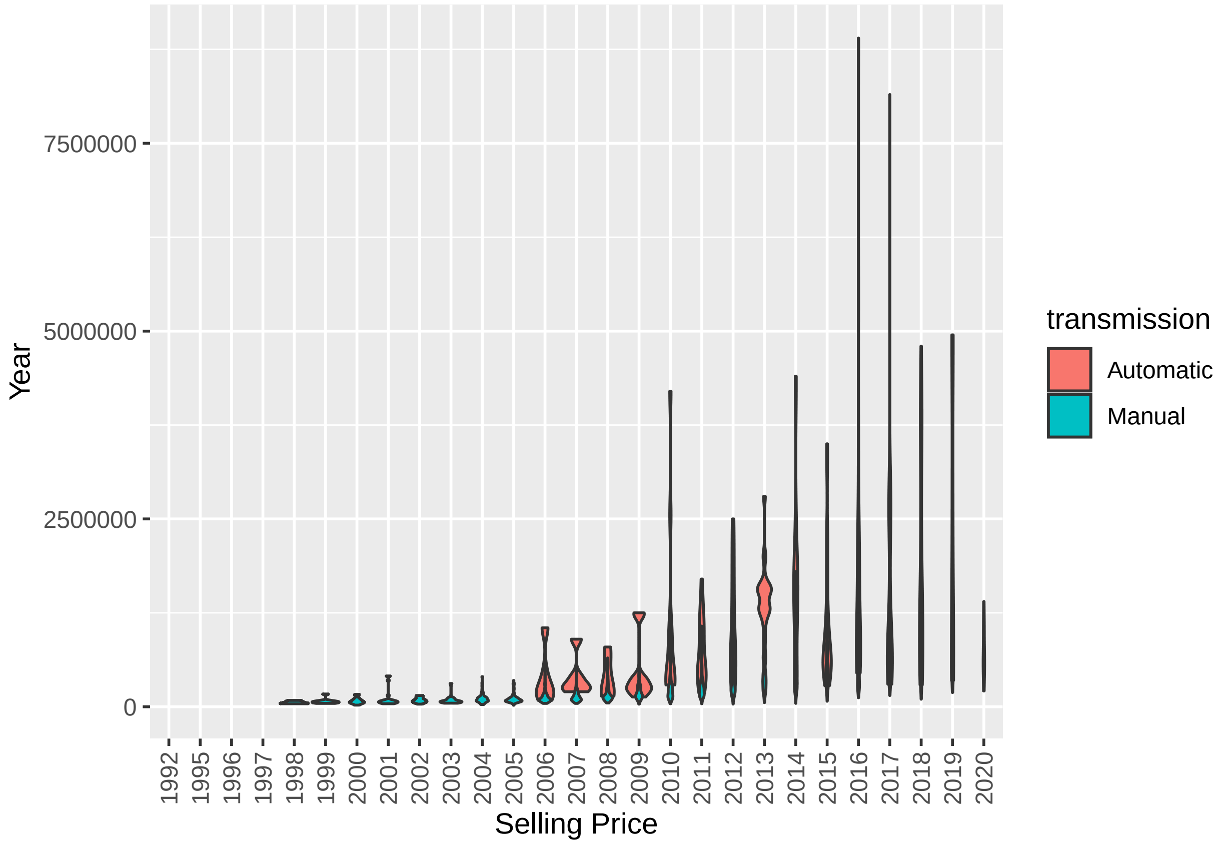
```
## # A tibble: 1 x 8
##   name        year selling_price km_driven fuel   seller_type transmission owner
##   <chr>      <dbl>         <dbl>     <dbl> <chr>  <chr>       <chr>        <chr>
## 1 Audi RS7 ~  2016       8900000     13000 Petrol Dealer      Automatic    Firs~
```

```r
car_data %>%
  group_by(transmission) %>%
  summarise(avg_price = mean(selling_price))
```

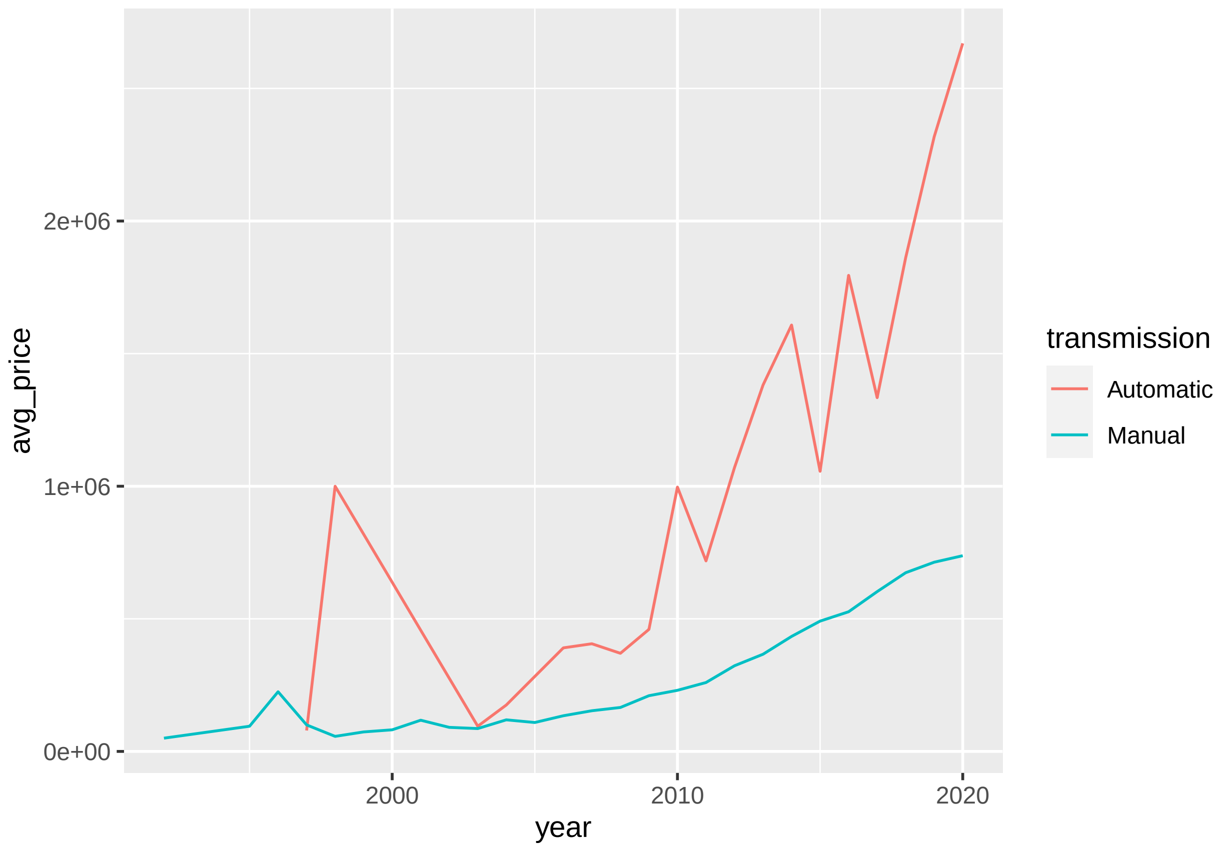
```
## # A tibble: 2 x 2
##   transmission avg_price
##   <chr>            <dbl>
## 1 Automatic     1408154
## 2 Manual         400067.
```

Plot the relationship between selling price and year for the automatic and manual cars.

```
ggplot(mapping = aes(x = reorder(year,selling_price), y = selling_price, fill=transmission)) +
  geom_violin(data = car_data %>% filter(transmission=="Automatic")) +
  geom_violin(data = car_data %>% filter(transmission=="Manual")) +
  xlab("Selling Price") +
  ylab("Year") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```
car_data %>%
  group_by(year, transmission) %>%
  summarise(avg_price = mean(selling_price)) %>%
  ggplot(mapping = aes(x = year, y = avg_price, color=transmission)) + geom_line()
```



QUESTION 2 (8 pts)

Estimate a model with selling price as the dependent variable and kilometers driven and transmission as the independent variables. (2pts)
Interpret the coefficients on all independent variables and the intercept. (6 pts)

```
price_mod1 <- lm(selling_price ~ km_driven + transmission, data = car_data)
summary(price_mod1)
```

```
##
## Call:
## lm(formula = selling_price ~ km_driven + transmission, data = car_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1296295  -217650   -70140   158559  7432492
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.489e+06  2.423e+04   61.43   <2e-16 ***
## km_driven           -1.618e+00  1.590e-01  -10.18   <2e-16 ***
## transmissionManual -9.783e+05  2.437e+04  -40.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 484900 on 4337 degrees of freedom
## Multiple R-squared:  0.2979, Adjusted R-squared:  0.2976
## F-statistic:   920 on 2 and 4337 DF,  p-value: < 2.2e-16
```

When the km driven is 0 and transmission is automatic, the base selling price, i.e. the intercept, is on average at a value of 1,489,000. It is also statistically significant. Per km driven, value drops by 1.618. This effect is also statistically significant. The average difference in value between manual and automatic transmission is -978,300, with automatic transmission being the higher valued one on average. This effect from transmission is also statistically significant. All coefficients significant at p-values much smaller than 0.05.

QUESTION 3 (6 points)

Now add year to the model. What happens to the coefficient on kilometers driven? Why?

```r
# year was recentered to make the intercept make more sense,
# since you cannot have negative base value
car_data$year_zeroed <- car_data$year - min(car_data$year)
price_mod2 <- lm(selling_price ~ km_driven + transmission + year_zeroed, data = car_data)
summary(price_mod2)
```

```
##
## Call:
## lm(formula = selling_price ~ km_driven + transmission + year_zeroed,
##     data = car_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1262774  -164383   -31830   103566  7443587
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.018e+05  4.964e+04   6.079 1.31e-09 ***
## km_driven            1.543e-01  1.614e-01   0.956    0.339
## transmissionManual -9.153e+05  2.269e+04 -40.329  < 2e-16 ***
## year_zeroed          4.803e+04  1.792e+03  26.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 449200 on 4336 degrees of freedom
## Multiple R-squared:  0.3977, Adjusted R-squared:  0.3973
## F-statistic: 954.3 on 3 and 4336 DF,  p-value: < 2.2e-16
```

The coefficient of the km_driven variable switched while the magnitude of the coefficient went down by an order of 10 and also became statistically insignificant. The effect from year is much great than from km driven and more statistically significant.

```r
cor.test(car_data$year, car_data$km_driven)
```

```
##
##  Pearson's product-moment correlation
##
## data:  car_data$year and car_data$km_driven
## t = -30.454, df = 4338, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4438979 -0.3948662
## sample estimates:
##        cor
## -0.4196881
```

Given that there is a statistically signficant negative correlation between the variables, and the effect from year is positive, the bias is negative. Since the new coefficient is greater than the old, the effect from km_driven was undercompensating for the hidden missing effect from year on price resulting in a downwards bias in the original model.

QUESTION 4 (6 points)

Now add the categorical variable owner to the previous model (the one that included km_driven, transmission, and year). Make "first owner" the reference group for the owner variable (hint: you would need to tranform the variable "owner" into a factor before determining the reference group). (2 pts) Interpret the coefficients of owner. (4 pts)

```r
# change reference group to First Owner
car_data$owner_cat <- factor(car_data$owner)
levels(car_data$owner_cat)
```

```
## [1] "First Owner"          "Fourth & Above Owner" "Second Owner"
## [4] "Test Drive Car"       "Third Owner"
```

```r
price_mod3 <- lm(selling_price ~ km_driven + transmission + year_zeroed + owner_cat, data = car_data)
summary(price_mod3)
```

```
##
## Call:
## lm(formula = selling_price ~ km_driven + transmission + year_zeroed +
##     owner_cat, data = car_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1263128  -161480   -31446   103607  7438553
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.640e+05  5.332e+04   6.827 9.88e-12 ***
## km_driven                     2.485e-01  1.636e-01   1.519  0.12886
## transmissionManual           -9.152e+05  2.268e+04 -40.345  < 2e-16 ***
## year_zeroed                   4.559e+04  1.948e+03  23.404  < 2e-16 ***
## owner_catFourth & Above Owner -2.404e+04  5.227e+04  -0.460  0.64553
## owner_catSecond Owner        -5.282e+04  1.726e+04  -3.059  0.00223 **
## owner_catTest Drive Car       1.955e+05  1.097e+05   1.782  0.07484 .
## owner_catThird Owner         -5.775e+04  2.893e+04  -1.996  0.04596 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 448600 on 4332 degrees of freedom
## Multiple R-squared:  0.3996, Adjusted R-squared:  0.3986
## F-statistic: 411.9 on 7 and 4332 DF,  p-value: < 2.2e-16
```

For owner, since the base level is the "first owner", it would be included in the intercept. This means that when the kilometers driven is 0, transmission is automatic, the year of the car is 1992 AND when the car is on its first owner, then the value of the is 364,000. When the car is on the second owner, 52,820 in value is lost on average compared to having just one owner in the past. This is statistically significant at a p-value of 0.05 and below, so we can reject the null hypothesis that the coefficient is 0 or that there is no difference in price between a car having 2nd owner and a 1st owner. When the car is on its third owner, 57,750 is lost in value on average from the base or first owner price point. This effect is also statistically significant at a p-value of 0.05, so we can again reject the null hypothesis that there is no difference. The average effect from the car being on fourth owner status is at -24,040 on the price (compared to single owner) but is not remotely statistically significant. The test drive owner status does statistically significant effect but only at a significance level 0.10. The average difference in price for test drive ownership status is 195,500.

QUESTION 5 (4 points)

What would be the predicted selling price of an automatic 2012 car with 100,000 kilometers and whose owner category is first owner?

```
test_df <- data.frame(km_driven=c(100000),
                      transmission=c("Manual"),
                      year_zeroed=c(2012-min(car_data$year)),
                      owner_cat="First Owner")
predict(price_mod3, newdata = test_df)
```

```
##        1
## 385508.6
```

```
# did it two different ways to confirm
test <- c(1,100000,1,2012-min(car_data$year),0,0,0,0)
crossprod(test, price_mod3$coefficients)
```

```
##          [,1]
## [1,] 385508.6
```

The predicted price would be 385508.6

QUESTION 6 (6 points)

The model above implicitly assumes the effect of year is the same regardless of the kilometers driven. Test whether this assumption is true and briefly discuss your results (i.e., tell me whether the assumption is true or not).

```
price_mod3 <- lm(selling_price ~ km_driven + transmission + year_zeroed + owner_cat + year_zeroed*km_dri
summary(price_mod3)
```

```
##
## Call:
## lm(formula = selling_price ~ km_driven + transmission + year_zeroed +
##     owner_cat + year_zeroed * km_driven, data = car_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1278323  -159226   -30478   102707  7431606
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.329e+05  7.439e+04   3.131  0.00175 **
## km_driven                     2.209e+00  7.936e-01   2.784  0.00540 **
## transmissionManual           -9.113e+05  2.272e+04 -40.106  < 2e-16 ***
## year_zeroed                   5.154e+04  3.057e+03  16.860  < 2e-16 ***
## owner_catFourth & Above Owner -3.282e+04  5.235e+04  -0.627  0.53068
## owner_catSecond Owner        -5.108e+04  1.727e+04  -2.958  0.00311 **
## owner_catTest Drive Car       1.615e+05  1.105e+05   1.462  0.14379
## owner_catThird Owner         -6.198e+04  2.896e+04  -2.140  0.03240 *
## km_driven:year_zeroed        -9.712e-02  3.846e-02  -2.525  0.01161 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 448400 on 4331 degrees of freedom
## Multiple R-squared:  0.4005, Adjusted R-squared:  0.3994
## F-statistic: 361.7 on 8 and 4331 DF,  p-value: < 2.2e-16
```

Given that the interaction coefficient is statistically significant, the assumption that there was no interaction between kilometers driven and year may be wrong. The year of the car has a significant moderation effect on the km driven and we can reject the null hypothesis that there is no effect. For every additional year on the car, the effect of km driven decreases by 0.09712 per km on the selling price.

PART 2

Load the data file called 'insurance.csv'. This data contains medical information and costs billed by health insurance companies. The data contains the following variables:

- age (age of primary beneficiary)
- gender (insurance contractor gender, female, male)
- bmi (Body mass index)
- children (Number of children covered by health insurance / Number of dependents)
- smoker (Fuel type of car (petrol / diesel / CNG / LPG / electric))
- region (the beneficiary's residential area in the US)
- charges (Individual medical costs billed by health insurance)

QUESTION 7 (8 points)

Write out a model (in notation similar to that which we use in class or the Wooldridge text; in other words write out the regression model) that predicts the charges based on age, sex, bmi and smoker. You can use the Microsoft word equation editor or simply enter the model using regular text in word. (2 pts)
Given the model and how the variables are defined in the dataset, what is the base group? (2 pts)
Write out the condition expectation for a female smoker. (2 pts)
Write out the conditional expectation for a male nonsmoker. (2 pts)

$$\widehat{charges} = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 sex + \hat{\beta}_3 bmi + \hat{\beta}_4 smoker$$

Assuming that the categorical variable base levels are chosen based on alphabetic order or set by us, the base group should be where age is 0, the gender is female, BMI is at 0 and they are non-smokers.

Write out the condition expectation for a female smoker.

$$E(y_i|age, sex = 0, bmi, smoker = 1) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_3 bmi + \hat{\beta}_4$$

Write out the conditional expectation for a male nonsmoker.

$$E(y_i|age, sex = 1, bmi, smoker = 0) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 + \hat{\beta}_3 bmi$$

QUESTION 8 (8 points)

Run the model discussed in question 7. (2pts)
Interpret the coefficients on sex and smoker (4 pts).
Look at standard errors on coefficients for sex and smoker. Why are they different? (2 pts)
[Hint: look at the formula for how we calculate the variance of our coefficient estimates]

```r
charges_mod <- lm(charges ~ age + sex + bmi + smoker, data = insurance)
summary(charges_mod)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12364.7  -2972.2   -983.2   1475.8  29018.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11633.49     947.27 -12.281   <2e-16 ***
## age            259.45      11.94  21.727   <2e-16 ***
## sexmale       -109.04     334.66  -0.326    0.745
## bmi            323.05      27.53  11.735   <2e-16 ***
## smokeryes    23833.87     414.19  57.544   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6094 on 1333 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7467
## F-statistic: 986.5 on 4 and 1333 DF,  p-value: < 2.2e-16
```

The intercept, which represents the previously described base group, is -11633.49, is statistically significant under level of 0.05 (and even below 0.001). Age increases the charges by 259.45 per additional year and is also very statistically significant. The effect from gender, though not statistically significant even at a high significance level, shows a difference of -109.04 in charges on average if the group is male. Since the variance of the coefficients is the ratio of deviation of residuals over sample size times the variable value times the variance of the variable times 1 minus the goodness of fit, the coefficients of sex and smoker will be different because they have different variances, and values. The variances of the variables are different because, even though they are both binary variables in this case, there are different numbers and different distribution "peak width" (as described by standard deviation) of males compared to smokers.

```
var(ifelse(insurance$sex == 'male', 1, 0))
```

```
## [1] 0.2501596
```

```
var(ifelse(insurance$smoker == 'yes', 1, 0))
```

```
## [1] 0.1629689
```

As we can see above, the values of the variation of the sex and smoker variables is different, resulting in the calculation of coefficient standard error being different.

QUESTION 9 (12 points)

The model above implicitly assumes the effect of bmi is the same for both smokers and nonsmokers. Test whether this assumption is true and briefly discuss your results (i.e., tell me whether the assumption is true or not). (4 pts)
Interpret the simple main effect of bmi and smoker as well as the interaction. (4 pts)
What are the estimated charges for a 38 years old non smoker man with 25 bmi? (2 pts)
What are the estimated charges for a 25 years old smoker woman with 30 bmi? (2 pts)

```
charges_mod2 <- lm(charges ~ age + sex + bmi + smoker + smoker*bmi, data = insurance)
summary(charges_mod2)
```

```
##
## Call:
```

```
## lm(formula = charges ~ age + sex + bmi + smoker + smoker * bmi,
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14524.3  -1967.9  -1337.7   -396.7  29516.9
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2071.077    840.644  -2.464   0.0139 *
## age             266.372      9.612  27.713   <2e-16 ***
## sexmale        -473.495    269.612  -1.756   0.0793 .
## bmi               7.969     25.044   0.318   0.7504
## smokeryes    -20193.152   1666.491 -12.117   <2e-16 ***
## bmi:smokeryes  1435.608     53.242  26.964   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4904 on 1332 degrees of freedom
## Multiple R-squared:  0.8367, Adjusted R-squared:  0.836
## F-statistic:  1365 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Given that moderation effect of smoking status on bmi (or the other way around) is statistically significant it would seem the assumption that the effect of bmi is not the same for both smokers and non-smokers.

At the base level where there is no interaction since it is a group of aged 0 female non-smokers at bmi 0, charges average at -2071.07. Every additional year in age adds 266.37 to the charges and this effect is statistically significant. On average males pay 473.50 on average less than females in charges. This effect from gender is statistically significant but only at a relatively high significance level of 0.1. Every one unit increase in BMI results in a simple main effect of 7.969 increase in insurance charges conditioned on the person being a non-smoker. It should be noted that this effect is statistically insignificant. It also contributes relatively little to the charges compared to the other factors based on magnitude. If the person is a smoker, there is an average decrease of 20,193 in charges compared to a non-smoker, given that BMI is 0. Finally, for a group of smokers, moderation of smoking on bmi towards charges additionally increases charges by 1435.61 for every unit of increase in bmi, an incremental change in slope. In other words, for smokers every 1 bmi increase charges increase by approximately 1443.57, holding all else constant.

```r
# 38 years old non smoker man with 25 bmi
test_df1 <- data.frame(age=38,
                       sex="male",
                       bmi=25,
                       smoker="no")
predict(charges_mod2, newdata = test_df1)
```

```
##        1
## 7776.793
```

38 years old non smoker man with 25 bmi would get charged 23473.84.

```r
# 25 years old smoker woman with 30 bmi
test_df2 <- data.frame(age=25,
                       sex="female",
                       bmi=30,
                       smoker="yes")
predict(charges_mod2, newdata = test_df2)
```

```
##        1
## 27702.38
```

25 years old smoker woman with 30 bmi would get charged 27702.38.

QUESTION 10 (4 points)

Do you trust the coefficients in the model above? In other words, do you consider these to be reasonable causal estimates of the effects of the different variables? Why or why not?

Given that bmi's coefficient is far from being statistically significant and the gender coefficient is barely above a signficance level of 0.05, I would not trust these estimates. I might trust these coefficients if literature has previously shown that they are causally linked to charges. However, since the R-squared of the new model is higher by about 0.1, it would seem that the new model is at least more trustable than the older model without the interaction.