# PA 541 Homework 2 (72 points)

Michael D. Siciliano

March 3rd, 2021

## PART ONE

Load the data file called 'car_data.csv'. This data contains information about cars and motorcycles listed on CarDekho.com. The data contains the following variables:

**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 One (tidyverse/data wrangling work – 10 points)

*a. What is the average selling price for automatic versus manual cars? (2 pts)*

```
car_data1 = car_data %>%
  group_by(transmission) %>%
  summarise(mean.price = mean(selling_price))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
car_data1
```

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

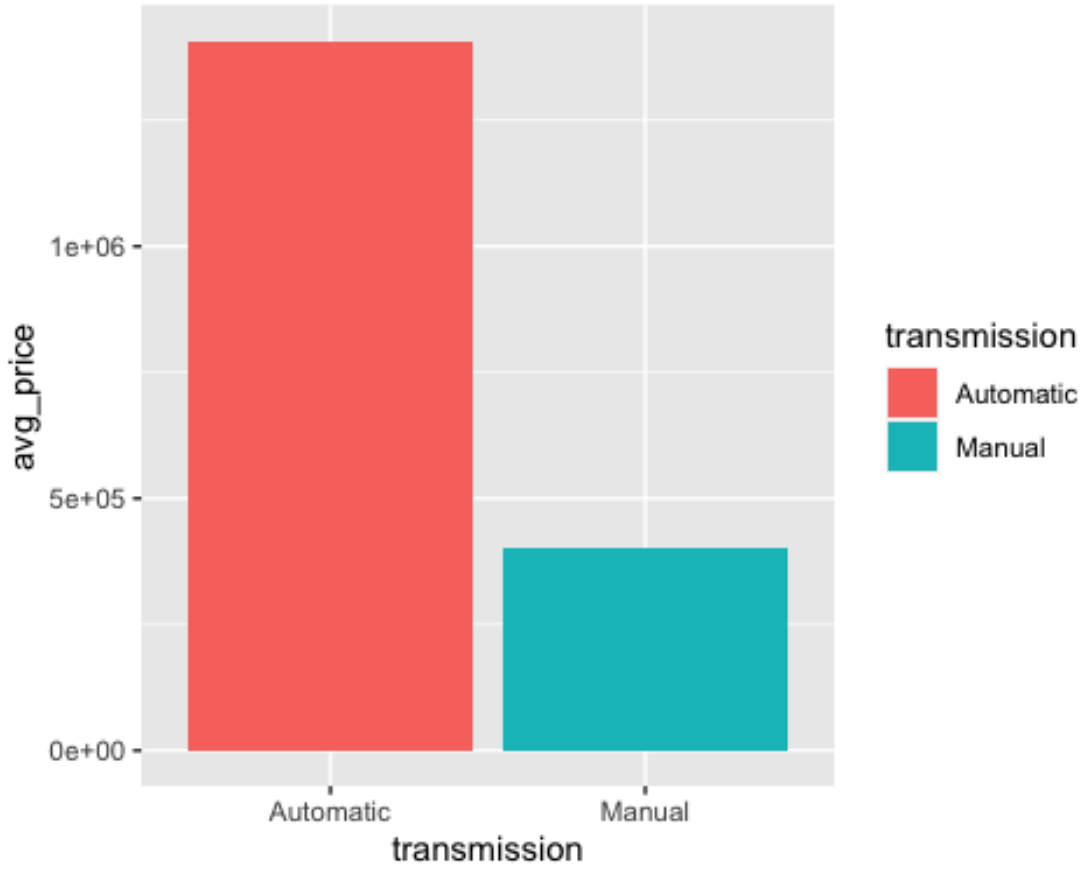
*b. Of the automatic cars, which model was sold at the highest price? (2pts)*

```
car_data2 = car_data %>%
  filter(transmission == "Automatic") %>%
  select(selling_price, name) %>%
  arrange(desc(selling_price))
car_data2
```

```
## # A tibble: 448 x 2
##    selling_price name
##            <dbl> <chr>
##  1       8900000 Audi RS7 2015-2019 Sportback Performance
##  2       8150000 Mercedes-Benz S-Class S 350d Connoisseurs Edition
##  3       5500000 Mercedes-Benz GLS 2016-2020 350d 4MATIC
##  4       4950000 BMW X5 xDrive 30d xLine
##  5       4950000 BMW X5 xDrive 30d xLine
##  6       4950000 BMW X5 xDrive 30d xLine
##  7       4950000 BMW X5 xDrive 30d xLine
##  8       4950000 BMW X5 xDrive 30d xLine
##  9       4950000 BMW X5 xDrive 30d xLine
## 10       4950000 BMW X5 xDrive 30d xLine
## # … with 438 more rows
```

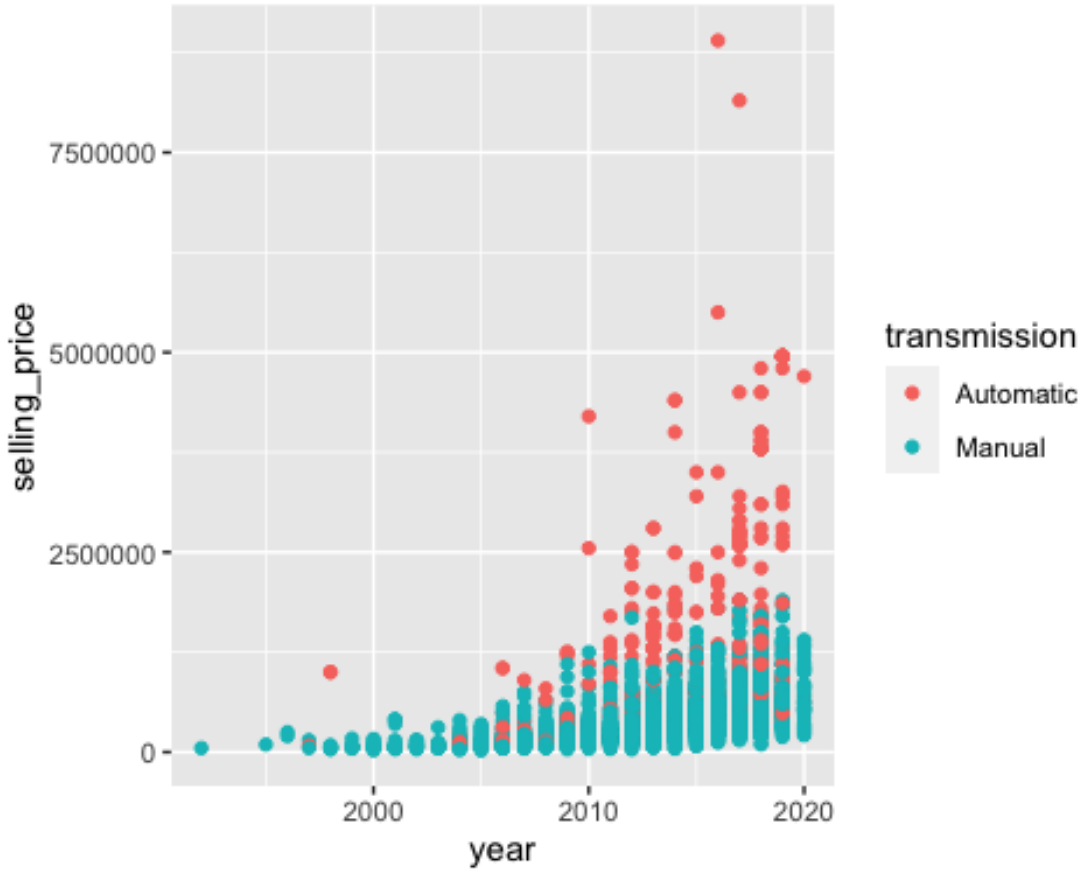
*c. Plot the average selling price for each type of transmission. (3 pts)*

```
avg.price = car_data %>%
  group_by(transmission) %>%
  summarise(avg_price = mean(selling_price))

## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(avg.price, aes (x = transmission, y = avg_price, fill = transmission))
+ geom_col()
```

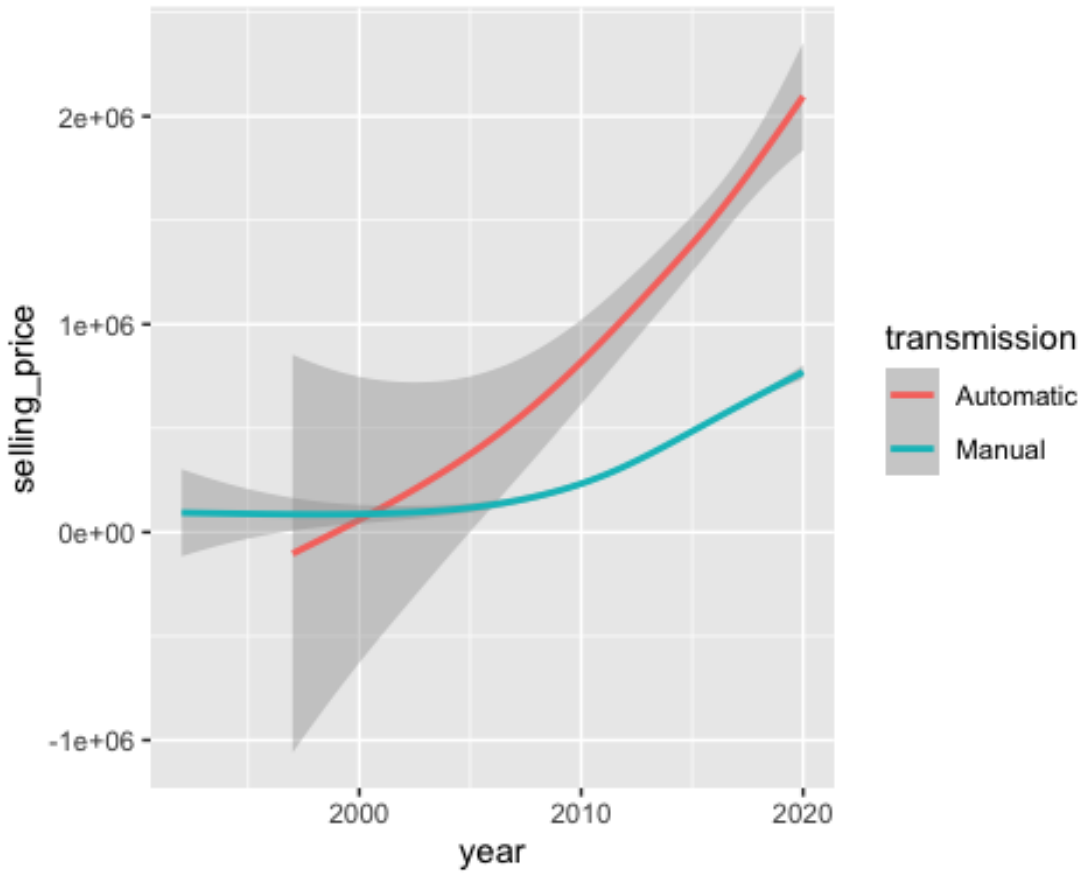



*d. Plot the relationship between selling price and year for the automatic and manual cars on the same plot. (3 pts)*

```
ggplot(car_data, aes (x = year, y = selling_price, color = transmission)) +
geom_point()
```




```
#we didn't cover this directly in class...but a better approach here is:

ggplot(car_data, aes (x = year, y = selling_price, color = transmission)) +
geom_smooth()

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




## 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)*

```
mod1 <- lm(selling_price ~ km_driven + transmission, data = car_data)
summary(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
```

*##Intercept is 1489000, which is the expected selling price of an automatic car with 0 kilometers driven.*

*##The slope of km_driven indicates that for each additional kilometer driven, the selling price of the car is expected to decrease by 1.618 Indian Rupees. The effect is significant at the .001 level.*

*##The reference group is automatic cars. On average, manual cars are sold for 978,300 Indian Rupees less than the automatic cars. The effect is significant at the .001 level.*

## QUESTION 3 (6 points)

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

```
mod2 <- lm(selling_price ~ km_driven + transmission + year, data = car_data)
summary(mod2)
```

```
##
## Call:
## lm(formula = selling_price ~ km_driven + transmission + year,
##     data = car_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1262774  -164383   -31830   103566  7443587
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -9.537e+07  3.614e+06 -26.391   <2e-16 ***
## km_driven            1.543e-01  1.614e-01   0.956    0.339
## transmissionManual -9.153e+05  2.269e+04 -40.329   <2e-16 ***
## year                 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
```

*##Year was in the error term of the above model and, based on the results of the current model, it looks like it exerted a substantial effect on the selling price and was correlated with the kilometers driven.  If you do a correlation test, you will find an expected negative correlation between year and km driven.  As year gets larger, i.e. new cars, the kms driven goes down.*

*Couple that negative correlation with a positive effect effect of year on selling price, and you have a negative bias. Thus we see the coefficient on km_driven go from a .15 in the model with year to a -1.618 when year is excluded. Thus, including the year the car was sold caused the coefficient on kilometer driven to decrease and go from statistically significant to statistically insignificant. The explanation is that cars that new cars (i.e. larger value on year) tend to have fewer kilometers driven. When including only the kilometers in the model and not year, it picked up part of the effect of year.*

## 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)*

```
car_data$owner <- as.factor(car_data$owner)
car_data$owner <- relevel(car_data$owner, ref = "First Owner")
mod3 <- lm(selling_price ~ km_driven + transmission + year + owner, data =
car_data)
summary(mod3)
```

```
##
## Call:
## lm(formula = selling_price ~ km_driven + transmission + year +
##     owner, data = car_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1263128  -161480   -31446   103607  7438553
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -9.045e+07  3.929e+06 -23.023  < 2e-16 ***
## km_driven                   2.485e-01  1.636e-01   1.519  0.12886
## transmissionManual         -9.152e+05  2.268e+04 -40.345  < 2e-16 ***
## year                        4.559e+04  1.948e+03  23.404  < 2e-16 ***
## ownerFourth & Above Owner -2.404e+04  5.227e+04  -0.460  0.64553
## ownerSecond Owner          -5.282e+04  1.726e+04  -3.059  0.00223 **
## ownerTest Drive Car         1.955e+05  1.097e+05   1.782  0.07484 .
## ownerThird 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
```

*##The coefficient for second owner is -52820, which means on average the selling price of cars that have two previous owners is 52820 Indian Rupees*

*less compared to cars that have one previous owner. The relationship is significant at the .01 level.*

*##The coefficient for third owner is -57750, which means on average the selling price of cars that have three previous owners is 57750 Indian Rupees less compared to cars that have one previous owner. The relationship is significant at the .05 level.*

*##The coefficient for fourth and above owner is -24040, which means on average the selling price of cars that have four or more previous owners is 24040 Indian Rupees less compared to cars that have one previous owner. The relationship is not significant.*

*##The coefficient for test drive car is 195500, which means on average the selling price of test drive cars is 195500 Indian Rupees more compared to cars that have one previous owner. This is not statistically significant at the two-tailed alpha = 0.05 level (but is at the .10 level).*

## 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?*

```
y = -90450000 + (0.2485*100000) + (45590 * 2012)
y
```

```
## [1] 1301930
```

## 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).*

```
mod4 <- lm(selling_price ~ km_driven*year + transmission + owner, data =
car_data)
summary(mod4)
```

```
##
## Call:
## lm(formula = selling_price ~ km_driven * year + transmission +
##     owner, data = car_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1278323  -159226   -30478   102707  7431606
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.024e+08  6.161e+06 -16.629  < 2e-16 ***
## km_driven                   1.957e+02  7.740e+01   2.528  0.01151 *
## year                        5.154e+04  3.057e+03  16.860  < 2e-16 ***
## transmissionManual         -9.113e+05  2.272e+04 -40.106  < 2e-16 ***
## ownerFourth & Above Owner -3.282e+04  5.235e+04  -0.627  0.53068
## ownerSecond Owner          -5.108e+04  1.727e+04  -2.958  0.00311 **
## ownerTest Drive Car         1.615e+05  1.105e+05   1.462  0.14379
## ownerThird Owner           -6.198e+04  2.896e+04  -2.140  0.03240 *
## km_driven:year             -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
```

*##The results show a significant interaction between year and kilometers driven. Thus, the implicit assumption that the effect of year is the same regardless of the kilometers driven does not hold. It is clear that kilometers driven moderate the effect of year on the selling price. Year has a higher impact on selling price for cars less kilometers.*

## PART TWO

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:

**Variable name (Description)**

- *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* (whether the individual is a smoker or not (yes/no))
- *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)*

$$charges_i = \beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 bmi_i + \beta_4 smoker_i + \epsilon_i$$

```
##The base category here is female, nonsmoker

##Conditional expectation for a Female smoker:
```

$$E(charges_i | Age_i, gender_i = female, bmi_i, smoker = yes)$$
$$= \beta_0 + \beta_1 age_i + \beta_2 bmi_i + \beta_3 smoker_i$$

```
##Conditional expectation for a male nonsmoker:
```

$$E(charges_i | Age_i, gender = male, bmi_i, smoker = no)$$
$$= \alpha_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 bmi_i$$

## 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]*

```
mod5 <- lm(charges ~ age + sex + bmi + smoker, data = insurance)
summary(mod5)

##
## 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 coefficient for sex is -109.04, which means that males on average were charged 109.54 dollars less compared to women. The relationship is not statistically significant. The coefficient for smoker is 23833.87 which means that smokers on average were charged 23833.87 more than nonsmokers. The relationship is significant at .001 level.*

*##From the equation for the variance of Bj we know that the factors that influence it are the fit of the model (sigma squared), sample size (n), and variance of the jth variable (var(Xj). See the formula below. The (1 - R2j) is the term in the equation for the variance in a multivariate model that deals with the amount of correlation between the jth variable and the other predictors in the model.*

*##Since the fit of the model and the sample size are the same for all three of these variables, the difference in the standard errors must come driven mostly by the variance of the independent variable*

$$var(\hat{\beta}_j) = \frac{\sigma^2}{N * var(X_j) * (1 - R_j^2)}$$

```
##So let's look at the variance for the dummy variables:

insurance$sex2 <- ifelse(insurance$sex == "female", 0, 1)
var(insurance$sex2)

## [1] 0.2501596

insurance$smoker2 <- ifelse(insurance$smoker == "no", 0, 1)
var(insurance$smoker2)

## [1] 0.1629689

##So the variance for sex is larger compared to smoker. This is why the standard error for sex is lower compared to smoker.
```

## 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)*

```
mod6 <- lm(charges ~ age + sex + bmi * smoker, data = insurance)
summary(mod6)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi * smoker, 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
```

*##Base group: female nonsmoker*

*##The results show a significant interaction between smoker and bmi.  So the effect of bmi is not the same for smokers and nonsmokers. Whether one is a smoker or not moderates the effect of bmi on charges.*

*##The coefficient for smoker is -20193. Because smoker is part of an interaction, we have to interpret it as simple main effect. Smoker with 0 bmi will be charged 20193 dollars less compared to nonsmoker with 0 bmi. The relationship is significant  (p<.001)*

*##bmi is part of an interaction, so we have to interpret it as simple main effect. The coefficient for bmi shows the effect of bmi only on nonsmokers. The coefficient is 8 which means that nonsmokers pay 8 dollars more for each additional unit of bmi. The relationship is significant (p<.001).*

*##The coefficient for bmi:smokeryes is 1436. This value is the difference in effect of bmi between smokers and nonsmokers. For smokers, the effect of bmi is 1436 dollars more for each additional unit of bmi compared to nonsmokers. So, nonsmokers are charged  8 dollars more for each unit increase in bmi and smokers are charged 1444 dollars (1436 + 8) more for each unit increase in bmi.*

*##a 38 years old non smoker man with 25 bmi:*

```
y1 = -2071.1 + (266.4*38) - 473.5 + 8*25
y1
```

```
## [1] 7778.6

##a 25 years old smoker woman with 30 bmi?
y2 = -2071.1 + (266.4*25)  + 8*30 - 20193  + (1436*30)
y2

## [1] 27715.9

#an alternative way to do this, as we saw in the class scripts:
p1 = c(1,38,1,25,0,0)
crossprod(p1,coef(mod6))

##          [,1]
## [1,] 7776.793

p2 = c(1,25,0,30,1,30)
crossprod(p2, coef(mod6))

##          [,1]
## [1,] 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?*

```
##discuss about omitted variable bias
```