PA 541 Homework 2 (100 pts)

**DUE March 1st at 3:00pm**

NAME \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 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 1 (5 pts)

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

##### Of the automatic cars, which model was sold at the highest price? (3 pts)

## QUESTION 2 (10 pts)

##### Estimate a model with selling price as the dependent variable and kilometers driven and transmission as the independent variables. (4 pts)

##### Interpret the coefficients on all independent variables and the intercept. (6 pts)

## QUESTION 3 (10 points)

##### Add year to the model. What happens to the coefficient on kilometers driven? Why?

## QUESTION 4 (10 points)

##### 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. (8 pts)

## QUESTION 5 (5 points)

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

## QUESTION 6 (10 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).

# 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* (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 (10 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. (4 pts)

##### Given the model and how the variables are defined in the dataset, what is the base/reference category? (2 pts)

##### Write out the equation for a female smoker. (2 pts)

##### Write out the equation for a male nonsmoker. (2 pts)

## QUESTION 8 (10 points)

##### Run the model discussed in question 7 (4 pts)

##### Interpret the coefficients on sex and smoker (4 pts).

##### B. 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 equality of our coefficient estimates]

## QUESTION 9 (20 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). (5 pts)

##### Interpret the simple main effect of bmi and smoker as well as the interaction. (5 pts)

##### What are the estimated charges for a 38 years old non smoker man with 25 bmi? (5 pts)

##### What are the estimated charges for a 25 years old smoker woman with 30 bmi? (5 pts)

## QUESTION 10 (5 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?

## QUESTION 11 (5 points)

Discuss the differences between statistical and practical significance. Give an example of when a variable may be statistically significant, but has little practical significance. (Hint: Think like a data analyst who is sharing results with a policymaker).