STAT 363 Homework 1

Merve Ogretmek

November 15, 2020

1 True/False

i. FALSE: In order to say that a model is nonlinear that model has to be a nonlinear function of a parameter instead of being a nonlinear function of the x term. To check if a model is nonlinear, first calculate the E(y) and take the derivative with respect to parameters as following

$$E(y) = E(\beta_0 + \beta_1 e^x + \epsilon) = \beta_0 + \beta_1 E(e^x) \tag{1}$$

$$\frac{dE(y)}{d\beta_0} = 1 \tag{2}$$

Derivative of the expected value of y with respect to the parameter β_0 does not contain the parameter β_0. The model is linear in terms of parameter β_0 and

$$\frac{dE(y)}{d\beta_1} = E(e^x) \tag{3}$$

again does not contain the parameter β_1. Therefore, model is linear in terms of parameters as a result we can say that this model is a "linear" model.

ii. TRUE: Let's check if this model is nonlinear or linear in the same way in previous question

$$E(y) = E(\beta_0 + \frac{1}{\beta_1} x + \epsilon) = \beta_0 + \frac{1}{\beta_1} E(x) \tag{4}$$

$$\frac{dE(y)}{d\beta_0} = 1 \tag{5}$$

Derivative of the expected value of y with respect to the parameter β_0 does not contain the parameter β_0. The model is linear in terms of parameter β_0 and

$$\frac{dE(y)}{d\beta_1} = \frac{-E(x)}{\beta_1^2} \tag{6}$$

above equation contains β_1. Therefore, model is a nonlinear function of β_1. We can conclude that the model is a "nonlinear" model.

iii. FALSE: $\hat{\beta}_1$ is known, it's an estimator for the unknown β_1 so we can't apply a test on $\hat{\beta}_1$ because we can find its exact value by fitting the observed sample. We can apply test on unknown parameters such as β_1. If the null hypothesis was H_0: $\beta_1 = 0$ and if we cannot reject the null hypothesis, we could conclude that there is no statistically significant linear association between x and y.

iv. TRUE: β_0, β_1 and σ^2 are unknown parameters in a model and usually our aim is to estimate them with some methods. ϵ_i's are not observable because we can only estimate the unknown parameters and fit them in to a model. Therefore, we can only calculate the residuals $\hat{\epsilon_i}$'s after fitting the model.

v. FALSE: For this question, I have made some research on the word "line" because I was confused about if the word "line" by definition means it is straight. I have found that Euclid used the word "line" when also referring to a "curve". For the straight line he explicitly used the word "straight line".[1] Therefore, I will assume that regression line by definition does not mean it is straight.
As a result, a regression line may not be straight if we are appyling a logistic regression and due to the same reason our dependent variable (y) does not have to be quantitative. Logistic regression is usually used for the categorized meaning qualitative data.

vi. FALSE: In the model, only response variable y and the error term ϵ is random. Independent variables are the data we are using in the model. Response variable is random because the error term is random.Also, there is no assumption on the distribution of independent variables.
For the error term, we might assume there is normality because it is justifiable in many cases. However, even if we don't assume that error is distributed normally OLS will still provide the BLUE estimators. We have to assume a form of distribution for the error term when we are applying tests and confidence intervals.
For the response value, if error term is normally distributed then the response value will be distributed normally. While applying the OLS method, we only have to assume that error is iid with mean 0 and constant variance σ^2 so that we don't have to deal with the heterocedasticity.

vii. TRUE: Yes, estimated slope and intercept terms have sampling distributions(they are statistics) meaning they have a probability distribution. If the normality assumption of the error holds the sampling distribution of the estimated slope and intercept is normal distributed, too. For the parameters, they are constants and unknowns we try to estimate them with the estimators I just mentioned.

viii. FALSE: True version is we need to assume that the **variance** of the response variable is constant for each value of an independent variable (x), and this is called homoscedasticity. Mean of the response variable is not constant for each value in most cases.

ix. FALSE: Spread of our data in the direction of response variable refers to the variance of the response variable. If we are appyling OLS method, it has to be constant for each value of independent variable because one of the assumptions of OLS is homocedasticity. If this assumption does not hold we would have to use methods like Generalized Least Squares.

[1] Euclid, Simson, R., amp; Todhunter, I. (1979). Euclid's Elements: Books I-VI, XI and XII. London: Dent.

x. TRUE: We can show that average(expected value) of residuals will be equal to 0 in a simple way. The i^th residual is the difference between the observed value Y_i and the fitted value $\hat{Y}_i$. Under the OLS method we try to minimize the following,

$$SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X)^2 \tag{7}$$

to minimize it we take the derivates with respect to parameters,

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum Y_i - \hat{Y}_i = \sum e_i = 0 \tag{8}$$

Normal equation proves that the sum of residuals are equal to zero. Since the sums are zero, 0/number of observations meaning the average will be equal to zero.

xi. TRUE: According to the Gauss-Markov Theorem, the OLS method yields the best linear unbiased estimators. Therefore, the estimators of β_0, $\beta 1$ and σ^2 are unbiased.

xii. FALSE: Showing a positive relation in terms of a regression between two variables does not necessarily imply causality. For the case given in question, it could be said that both number of television per person and the life expectancy affected by the wealth of a country. If a country is wealthy, we might expect people to afford many technological devices and in wealthy countries, standard of living is expected to be high. Also, health system is better mostly in wealthy countries. Therefore, there could be another variable that affecting the both number of television and the life expectancy in a country in a positive manner.

2 Open Ended Questions

a. Underlying assumptions of a simple linear regression analysis as following:
- Model is a linear function of parameters but it doesn't have to be linear in terms of the x terms. This is compulsory for the model to be a linear model.
- Error term has to be iid with N(0,σ^2). If it is does not have zero mean, and constant variance for all observations, the OLS estimators wouldn't be BLUE according to the Gauss-Markov Theorem. For the normality case, to make use of hypothesis testing and and confidence intervals so that we can make an inference, we have to use the normality assumption of the error term.
- No autocorrelation between the error terms meaning the error terms are all statistically independent. Again, if this does not hold OLS estimators wouldn't be BLUE.

b. Yes, it would be risky to use a sample linear regression equation to predict or to estimate outside the range of values of the independent variable. This is an issue about the scope of the model. When we are creating a regression model, we have to choose a scope meaning an interval of values for the independent variable. The scope of the model is usually limited by the range of data. Getting outside the data would be risky because we wouldn't have any proof to support our estimates. This act is called the "extrapolation". [2]

[2]Kutner, M. H., Nachtsheim, C. J., Neter, J., amp; Li, W. (2013). Applied linear statistical models. New Delhi: McGraw-Hill Education.

c. A regression equation could be used for the following two purposes: (I would like to answer this question with my favourite two subjects related to the Statistics: Machine Learning and Econometrics.)

- **Prediction:** Many machine learning applications uses regression models to predict some y value in the scope of the model. Their aim is to foresee the result y when the x values are observed out of the sample. They would train the model with some data set and fit a regression equation. When an out of sample x value observed, a prediction of y will be made for that x value. For example, I would like to start a new company and I want to predict my profits(y) by looking at the other companies' features(x) and profits.
- **Coefficient Estimation:** Econometrics is usually interested with the coefficient estimation($\hat{\beta}$ values), it is not interested with the prediction of response value (y). This due to fact that econometrics tries to understand how X is affecting the Y so that they can make policy decisions like how taxes imposed on imports(x) affects the international trade(y) so that they can choose to optimal tax amount.
In conclusion, regression models can be used for different purposes by different disciplines. Therefore, it is crucial to make the best use out of the regression models depending on the problem.

d. Dependent variable is considered as a random variable because we set the model as following,

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{9}$$

On the right hand-side of the equation, the error term(ϵ) is a random variable. As a result, dependent variable (Y) becomes also a random variable.

e. We say that it is "simple" because it has only one explanatory variable (X-term). If it has more than one explanatory variable then we would say that model is a "multiple" linear regression model.

f. The given data in the question is a time-series data. In time-series data every observation is affected by the previous observations. In this case, one of the assumptions of the OLS method wouldn't hold which is no autocorrelation between error terms. As a result, the estimates provided by the OLS wouldn't be BLUE. We have to get rid of this serial-correlation by applying some time-series methods like using a lagged variable.

3 Regression Model

a. Since the slope term(1.7) is positive, I would say that these two variables have positive linear relationship.

b. For the simple regression model, 1.7 corresponds for the slope term. Interpretation is that one unit increase(decrease) in the rides would increase(decrease) the mean of the distribution of overall satisfaction.

c. Regression model is,

$$\hat{Y} = -94.96 + 1.7X \tag{10}$$

X = 86 given, by fitting it into the equation above we predict $\hat{Y} = 51.24$. Since the observed value Y = 43, the residual would be equal to Y-$\hat{Y}$ = 43 - 51.24 = -8.24.

d. It wouldn't be convenient to make a prediction about the response value when the observed X value(50) is not in our data range. Maybe, the characteristic of the regression changes for the X values smaller than 60. Therefore, it would be an extrapolation again and without any evidence it wouldn't be right to expect something for the response only by looking at the regression model in the question.