Study Questions For Predict 410

Topic: Ordinary Least Squares Regression

Our learning format requires that you complete the assigned readings efficiently and “intelligently”. In order to help you focus your attention on important concepts in the course reading, we have constructed a list of study questions for each topic covered in PREDICT 410. You should preview each list of study questions before you begin your reading, and then answer the questions in a notebook while you are performing you reading. If you cannot answer a question, then you should look up the answer. If you cannot find the answer, then you should post a question in your Blackboard course shell.

(1) When we refer to a 'simple linear regression', to what type of model are we referring? How does a 'simple linear regression' differ from a 'multiple regression'?

(2) In statistics, and in this course, we use the term 'regression' as a general term. What do we mean by the term 'regression'? What is the objective of a 'regression model'?

(3) What do we mean by 'linear regression'? Which equations represent a linear regression?

(a) Y = b0 + b1*X1
(b) Y = b0 + b1*X1 + b2*(X2^2)
(c) Y = b0 + exp(b1)*X1

(4) Before building statistical models, it is a common and preferred practice to perform an Exploratory Data Analysis (EDA). What constitutes an EDA for a simple linear regression model? Is this EDA satisfactory for a multiple regression model, or do we need to change or extend the EDA? As we move forward in this course we will also learn about logistic regression models and survival regression models, will these methods need their own EDA or is EDA general to all statistical models?

(5) In the simple linear regression model what is the relationship between R-squared and the correlation coefficient rho?

(6) How do we interpret a regression coefficient in OLS regression?

(7) Frequently, as a form of EDA for OLS regression we make a scatterplot between the response variable Y and a predictor variable X. As an assumption of OLS, the response variable Y must be continuous. However, the predictor variable X could be continuous or discrete. When the predictor variable is discrete, does a scatterplot still make sense? If not, what type of visual EDA does make sense? Does the appropriateness of the scatterplot make sense if the discrete variable takes on many discrete values (such as the set of integers, think of dollar amounts rounded to the nearest dollar) versus only a few discrete values(such as a coded categorical variable which only takes the values 1, 2, or 3)?

(8) The simple linear regression model is a special case of 'Multiple Regression' or 'Ordinary Least Squares'(OLS) regression. (We will typically use the term OLS regression.) What are the assumptions of OLS regression? In the final step of a regression analysis we perform a 'check of model adequacy'. What model diagnostics do we use to validate our fitted model against the model assumptions of OLS regression?

(9) How are the parameters, i.e. the model coefficients, estimated in OLS regression? How does this relate to maximum likelihood estimation? How do you show the relationship between OLS regression and maximum likelihood estimation?

(10) What is the overall F-test? What is the null hypothesis and what is the alternate hypothesis? The overall F-test is also called the 'test for a regression effect'. Why is it called this?

(11) What is the difference between R-squared and adjusted R-squared? How is each measure computed, and which measure should we prefer? How does the interpretation of R-squared change as we move from the simple linear regression model to the multiple regression model?

(12) The simple linear regression model Y = b0 + b1*X1 has three parameters. Two of the parameters are b0 and b1. What is the third parameter?

(13) What is a sampling distribution? What theoretical distribution do the parameter estimates have in OLS regression? What distribution do we use in practice? Why do we use a different distribution in practice?

(14) The final step of a regression analysis is a 'check of model adequacy'. This 'check of model adequacy' or 'goodness-of-fit' is a very important step in regression analysis. Why? Which quantities in the regression output are affected when the fitted model deviates from the underlying assumptions of OLS regression?

(15) Nested Models: Given two regression models M1 and M2, what does it mean when we say that 'M2 nests M1'?

(16) What is the Analysis of Variance Table for a regression model? How do we interpret it and what statistical tests and quantities can be computed from it?

(17) When the intercept is excluded in a regression model, how does the computation and the interpretation of R-squared change? Fit a no intercept model in SAS and check the SAS output for any noted differences.

(18) How do we interpret the diagnostic plots output by the PLOTS(ONLY)=(DIAGNOSTICS) option in PROC REG in SAS?

(19) Why do we plot each predictor variable against the residual as a model diagnostic?

(20) Why do we perform transformations in the construction of regression models? Name at least two reasons.

(21) What is multicollinearity and how does it affect the parameter estimates in OLS regression? How do we diagnose multicollinearity?

(22) What is a Variance Inflation Factor (VIF) and how does it relate to multicollinearity?

(23) Given a set of predictor variables X1,..., Xn, which are determined to show a high degree of multicollinearity between some of the variables, how should we choose a subset of these predictor variables to reduce the degree of multicollinearity and improve our OLS regression performance?

(24) Variable Selection: How does forward variable selection work? How does backward variable selection work? How does stepwise variable selection work?