Exam 2

Applied Stats/Quant Methods 1

Due: December 08, 2021

Instructions

- Instructions Please read carefully: You have until 23:59 Wednesday December 8 to complete the exam. Please export your answers as a single PDF file and include all code you produce in a supporting R file, which you will upload to Blackboard. The exam is open book; you can consult any materials you like. You must not collborate with or seek help from other students. In case of questions or technical difficulties, you can contact me via email. You should write-up your answers in R and LaTeX as you would for a problem set. Please make sure to concisely number your answers so that they can be matched with the corresponding questions.

Question 1: Stock Market DONE

(a) What concerns might we have about using the value of our company 'as is' in a model that regresses 'Total Stock Value' on 'Months After Purchase'? Skewness and Kurtosis.

This appears to be an example of data that are drawn from a distribution that is right-skewed or positive skewed (in this case it appears to be the exponential distribution).

Another concern for this data may be excess kurtosis, either positive or negative. Underdispersed data has negative excess kurtosis and therefore a reduced number of outliers. Overdispersed data has a positive excess kurtosis and an increased number of outliers.

(b) How could we address these concerns? Skew: There are apparently numerous ways to address skew in data.One way is the data can be transformed, using log and root for exponential distributions and Box Cox for skewed distributions. Outliers can also be removed manually, identifying them using z scores and interquartile range.

Kurtosis: Transforming the data. The Box-Cox transformation is a useful technique for trying to normalise a data set. Another technique available is the probability plot correlation coefficient plot and the probability plot to investigate a good distributional model for the data.

Question 2: Lambs DONE

(a) Write out the fitted model for a ewe lamb using the estimated coefficients. Write the prediction equation based on the result.

$\hat{Y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i D_i$

This is the equation or fitted regression line that applied in the last data set. However in this equation there is a second dummy variable

$\hat{Y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i D_1 i + \beta_3 x_i D_2 i$

Predicted value of Fatness = -18.137 + 2.298*Weight -8.362*Wether:NotWether -4.072*Ram:NotRam

DONE

(b) What is the predicted Fatness index of a wether lamb that weighs 14kg?

```
Predicted value of Fatness = -18.137 + 2.298*14 -8.362*1 -4.072*0

Predicted value of Fatness = -18.137 + 32.172 -8.362

 Predicted Fatness index = 5.673

 DONE
```

(c) Which lamb group has the highest Fatness index for every weight?

Perhaps need to perform a test in Rstudio for this one?

Predicted Fatness Index of Wether group = -18.137 -8.362*1 -4.072*0 Predicted Fatness Index of Wether group = -18.137 -8.362 Predicted Fatness Index of Wether group = -26.499

Predicted Fatness Index of Ram group = -18.137 -8.362*0 -4.072*1 Predicted Fatness Index of Ram group = -18.137 -4.072*1 Predicted Fatness Index of Ram group = -18.137 -4.072 Predicted Fatness Index of Ram group = -22.209

Predicted Fatness Index of Ewe group = -18.137 -8.362*0 -4.072*0 Predicted Fatness Index of Ewe group = -18.137

The negative values perhaps make sense in the context of the overall question which is not part of this exam, so if I ignore the negative values, then the Wether group has the highest Fatness index for every weight.

DONE

Question 3: Arsenic

(a) So, we successfully estimated an additive model with well depth and distance to the nearest factory as the two predictors of a household's arsenic level. The estimated coefficients are found in the first column of Table 1. Interpret the estimated coefficients for the intercept and each predictor.

An additive model does no contain an interaction and so Model 1 of Table 1 is of interest for this interpretation.

For Model 1, distance in kilometers/100 to the closet known commercial factor was the only statistically significant predictor ($p < 0.001$).

See slide about 47 of multireg2 printable

Write out prediction equation $\hat{Y}_i = 1.02 - 0.62 * well_depth - 4.33 * dist100$

Those with deep wells are predicted to have .62 less units of hundreds of micrograms per liter of arsenic in household's drinking water. For every kilometer/1000 farther from the closest known commercial factory, 4.33 less units of hundreds of micrograms per liter of arsenic in household's drinking wate is predicted. The predicted value of arsenic levels when wells are not deep and the distance in kilometers/ 100 to the closest know commercial factory is 0 is 1.02 units of hundreds of micrograms per liter according to this first prediction equation.

(b) Does the coefficient estimate for the closest known factory vary based on whether or not a house has a deep well? If so, change your interpretation of the estimated coefficients in part (a) to conform with the interactive model in column 2 of Table 1. Provide the appropriate test to determine whether we should model the relationship between distance, well depth, and arsenic levels using an additive or interactive model.

(c) Compute the average difference in arsenic levels between two households that have a deep well (=1), but one is closer to a factory (dist100 = 0.4) than the other (dist100 = 2.14).

Question 4:Multiple Choice DONE

(a) For explanatory variables with multi-collinearity, the corresponding estimated slopes have larger standard errors. VIFDONE .

(b) The coefficients in an ordinary least squares regression model are generalized additive estimates. DONE

(c) We can calculate our standard errors by taking the square root of the off-diagonal elements in our variance-covariance matrix. FALSE diagonal elements DONE

(d) Which of the following plots is used to check for normality in the assumptions of linear regression? The QQ plot of residuals. DONE

Question 5: Climate Action HALF OF B AND ALL OF C LEFT

(a) Interpret the coefficients for Age and Education Here those who are 70 and above have 8.413 less feeling thermometer ratings units.

For every year more the respondents indicated they attended school for Education, 7.987 less feeling thermometer rating units for support for climate action were applicable to them.

DONE

(b)) The author claims that she 'cannot reject the null hypothesis that Age has no effect on support for climate action ($H\Theta : \beta Age = \Theta$) . Using the coefficient estimate and the standard error for Age construct a 95support for climate action. Based on the confidence interval, do you agree with the author? Explain your answer.

The 95 percent confidence interval for the slope is the estimated coefficient (-8.413) $\pm$ two standard errors (4.539*2).

-8.413-9.078 -8.413+9.078

$-17.491 \geq x \leq 0.665$

```
Using R:
confint(fit, 'Age', level= 0.95)
```

BASED ON YOUR ANSWER, DO YOU AGREE WITH THE AUTHOR?

DONE

(c) Calculate the first difference in support for climate action between low and high values of Education for young respondents holding Party constant at its sample mean. Use 3.93 as the mean of Party and use +/- one standard deviation around the mean of Education (from 10.99 to 12.99) for low and high values of Education respectively

Predicted support for climate action =

Question 6: Define importance of terms DONE

Define and describe why the following four (4) terms are important to hypothesis testing and/or regression. You can earn full credit with just two or three sentences, but please be specific and thorough.

(a) Partial F-test Used to determine whether there is a statistically significant difference between a regression model and a subset of variables/ nest version of the overall regression model. Used to test the usefulness of a group of specific predictors in the overall model i.e. improve the fit of the model. The null hypothesis is that all coefficients removed from the overall model are zero. The alternative is that at least one of these removed coefficients in the nest model is not zero. DONE

(b) Categorical data/dummy variables It does not make sense to assign values of 1,2,3, to categorical data. Decomposing categorical variables into dummy variables in this case is important, assigning values of instead 1 or 0. This allows us to use Categorical data to predict in our regression. Dummy variable models apparently are also important provide correct results for un/imbalanced data. DONE

(c) Constituent term, alse referred to as constitutive terms, is used in relation to the interaction effect. Example of constituent or constitutive terms is the/a coefficient. The product of constitutive terms is interaction terms. DONE

(d) The Test statistic is important as it let's us know whether we can reject or fail to reject the null hypothesis test. If the standardised test statistic is more extreme than the critical value, reject. If not more extreme, fail to reject. For example, a two sample t-test tests if the means of two populations are equal. DONE