202021 - Statistics for Enterprise Data Analysis - S2

Dashboard / My courses / 2020-2021 / 2º Ciclo / Pós-Graduações / Primavera / SfEDA-400089-202021-S2 / Week #13 - May 17th 2021 / Lab 4

Lab 4

exclude this student from the analysis.

Hint: Look up this student's studentized residual and leverage to see why his/her Cook's distance is high and if you should investigate further.

Exercise 6

The **BEVERAGE** data file (kindly provided by Dr. Wolfgang Jank at the University of Maryland) contains the following quarterly data on sales of a popular soft drink for the period 1986-1998:

- $Sales$ = quarterly sales (U.S $ million)
- $Time$ = time period (consecutive numbers from 1 to 52)
- D_1 = indicator variable for quarter 1
- D_2 = indicator variable for quarter 2
- D_3 = indicator variable for quarter 3

The reference level for the indicator variables is quarter 4.

1. Draw a scatterplot of $Sales$ on the vertical axis versus $Time$ on the horizontal axis and connect the points on the plot with lines. Based on the plot, why would it be inappropriate to fit a simple linear regression model with $Sales$ as the response variable and Time as the predictor variable?
2. One way to take into account the seasonality of these data is to fit a multiple linear regression model with Sales as the response variable and $(D_1, D_2, D_3, Time)$ as the predictor variables. Fit this model and draw a scatterplot of the studentized residuals from the model on the vertical axis versus $Time$ on the horizontal axis. Based on the plot, why is this model also inappropriate?
3. One possible remedy for models with excessive autocorrelation is to try adding the lag-1 response variable as a predictor variable. Create the lag-1 response variable, $LagSales$. Next remove the first observation of each variable (since there is no value of $LagSales$ when $Time = 1$). Then fit a multiple linear regression model with sales as the response variable and $(D_1, D_2, D_3$ and $LagSales)$ as the predictor variables. Draw a scatterplot of the studentized residuals from this model on the vertical axis versus $Time$ on the horizontal axis. Comparing this plot with the residual plot from part 2. , does including appear to correct any autocorrelation problems?
4. Use a simple linear regression model with $Sales$ as the response variable and $Time$ as the predictor variable to predict for the four quarters in 1999. Calculate the prediction errors if $Sales$ for the four quarters in 1999 were actually 4,428, 5,379, 5,195, and 4,803, respectively. Also calculate the prediction errors for the models you fit in parts 2. and 3. Which model provides the best predictions overall?

◀ Linear regression - Part II - Handouts

Jump to...

NY ▶

Useful Links

Support

Serviços de Informática / IT Services

si@novaims.unl.pt

NOVA IMS' Accreditations and Certifications

Cofinanced by