STAT 425 Assignment 4

Due Tuesday, March 22, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Most relevant class notes: 4.1.Collinearity, 4.2.GLS, 4.3.TestFit, 5.1.Polynomial

Problem 1

A study was conducted to see if a certain food dye (Acid Red 118) affected mutation rates in salmonella bacteria. The data are in the included file, "reddye.csv." The variables are concentration of red dye on the plate (`conc`), and number of mutation colonies developed on the plate (`colonies`).

a) Read the data into an R data frame, and make a scatterplot with `colonies` as the response and `conc` as the predictor. Include the least squares line on the graph. What does the slope of the LS line suggest about the relationship between `conc` and `colonies`? Does the line seem to fit the data?

Answer:

b) Test the fit of the linear model in part a) versus the more general alternative model where conc is treated as a factor variable. Recall from class: if `conc` is treated as a factor variable, then we assume only that y has a different mean for each value of `conc` and nothing more. What do you conclude from the results?

Answer:

c) Since the concentrations range over several orders of magnitude, a logarithmic transformation of `conc` might help. There is a problem, though, because the zero concentration would transform to $-\infty$. Instead, consider the constructed variable `lgconc=log10(1+conc)`. Make a scatterplot of `colonies` versus `lgconc`, including the least squares line for the corresponding linear regression model. What does the slope of the LS line suggest about the relationship between `lgconc` and `colonies`? Does the line seem to fit the data?

Answer:

d) Test the fit of the model in c) by comparing it to the more general model that treats `lgconc` as a factor variable.

Answer:

e) Obtain the R-square values and model F test p-values for three models:

$$\text{colonies} = \beta_0 + \beta_1 \texttt{conc} + \texttt{error}$$
$$\text{colonies} = \beta_0 + \beta_1 \texttt{lgconc} + \texttt{error}$$
$$\text{colonies} = \beta_0 + \beta_1 \texttt{lgconc} + \beta_2 \texttt{lgconc}^2 + \texttt{error}$$

Based on the results, which of these three models seems most reasonable and why? **Answer:**

Problem 2

The `aatemp` data in the `faraway` library comes from the U.S. Historical Climatological Network. The data report annual mean temperatures in Ann Arbor Michigan for roughly 150 years.

a) Fit a linear trend model to temperature as a function of year and display the model summary. Does there appear to be a trend?

Answer:

b) For the model in a), plot residuals versus year, connecting the dots in the plot (`type='o'`), and adding a horizontal reference line at 0. Is there any evidence of serial correlation in the graph?

Answer:

c) Based on the model in a), test for serial correlation between successive years using the Durbin-Watson test, and state your conclusion.

Answer:

d) Using the `gls` function from the `nlme` library, fit a linear trend model using the AR1 form of correlation between years. Display the model summary. Does the trend line change much? How much correlation is there, based on the estimated AR1 correlation parameter?

Answer:

e) Again using the `gls` function, fit a cubic model (third order polynomial) for temperature as a function of year, with the AR1 form of correlation. Display the model summary. Make a scatter plot of `temp` versus `year`, and add the fitted curves from the linear and 3rd order polynomial models to the graph. One way to do this is with `lines` command after creating the plot, e.g. `lines(mod$fitted~year, data=aatemp)`. Which model seems to track the data better, based on what you see?

Answer:

Problem 3:

We delve into the theory for added variable plots and variance inflation factors. Consider a model of the form

$$\mathbf{y} = \mathbf{X}_0\beta + \mathbf{z}\gamma + \mathbf{e}$$

where $\mathbf{X_0}$ is $n \times p$ and full rank, $\mathbf{z}$ is $n \times 1$ and linearly independent of the columns of $\mathbf{X_0}$, $E(\mathbf{e}) = \mathbf{0}$, and $Cov(\mathbf{e}) = \sigma^2\mathbf{I}$.

The partial regression plot for $\mathbf{z}$ shows its association with the response $\mathbf{y}$ after adjusting for the variables in $\mathbf{X_0}$. This plot is obtained by plotting residuals $\mathbf{r}_y$ from the LS regression of $\mathbf{y}$ on $\mathbf{X}_0$ versus residuals $\mathbf{r}_z$ from the LS regression of $\mathbf{z}$ on $\mathbf{X}_0$.

a) Show that $\mathbf{r}_y = (\mathbf{I} - \mathbf{H}_0)\mathbf{y}$ and $\mathbf{r_z} = (\mathbf{I} - \mathbf{H}_0)\mathbf{z}$, where $\mathbf{H}_0 = \mathbf{X_0}(\mathbf{X_0^T X_0})^{-1}\mathbf{X_0^T}$.

Answer:

b) Under the model assumptions given above, show that $E(\mathbf{r}_y) = \mathbf{r}_z\gamma$, so the expected slope of the line is the same as the coefficient of $\mathbf{z}$ in the full model.

Answer:

c) The conditional expectation model in b) has the form of simple linear regression through the origin (no intercept). Show that fitting this "model" by LS regression gives estimated slope:

$$\hat{\gamma} = \frac{\mathbf{z^T}(\mathbf{I} - \mathbf{H_0})\mathbf{y}}{\mathbf{z^T}(\mathbf{I} - \mathbf{H_0})\mathbf{z}} = \frac{\sum_{i=1}^n (z_i - \hat{z}_i)y_i}{\sum_{i=1}^n (z_i - \hat{z}_i)^2}$$

where $\hat{\mathbf{z}} = \mathbf{H}_0\mathbf{z}$ and y_i, z_i and $\hat{z}_i$ are the ith components of $\mathbf{y}$, $\mathbf{z}$ and $\hat{\mathbf{z}}$ respectively.

Answer:

d) Using c) and the model assumptions, show $E(\hat{\gamma}) = \gamma$ and

$$var(\hat{\gamma}) = \frac{\sigma^2}{\sum_{i=1}^n (z_i - \hat{z}_i)^2}$$

Answer:

e) Let R_z^2 denote the multiple R-square statistic for the regression of $\mathbf{z}$ on the variables in $\mathbf{X}_0$. It can be shown that

$$R_z^2 = 1 - \frac{\sum_{i=1}^n (z_i - \hat{z}_i)^2}{\sum_{i=1}^n (z_i - \bar{z}_i)^2}$$

where $\bar{z} = n^{-1}\sum_{i=1}^n z_i$. Using this fact, show that

$$var(\hat{\gamma}) = VIF_z * \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z}_i)^2}$$

where VIF_z is the "variance inflation factor" given by

$$VIF_z = \frac{1}{1 - R_z^2}$$

Answer: