Linear Regression Models
Statistics GR5205/GU4205 — Fall 2020

Homework 4

The following problems are due on Monday, Nov 9th, 11:59pm.

1. (Project 8.38 in KNN) Return to the *SENIC Project* data from earlier assignments. Here we consider the regression relating number of nurses (Y) to available facilities and services (X).

   (a) Make a scatterplot of the data, and overlay a lowess smoother. Does a linear mean function seem plausible for these data?

   (b) Fit the second order mean function

   $$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x + \beta_{11} x^2$$

   assuming a constant variance

   $$\mathrm{Var}[Y|X = x] = \sigma^2$$

   Overlay the estimated mean function on your scatterplot from part (a). How closely does your fitted model agree with the lowess smoother?

   (c) Assume normality and test whether the quadratic term can be dropped from the model. Clearly state your null and alternative hypotheses, find and interpret the P-value, and clearly state your conclusion.

   (d) Obtain separate 95% prediction intervals for the number of nurses at two hospitals, one with an AFS percentage of 30 and one with an AFS percentage of 60. Interpret the resulting intervals. What is your simultaneous confidence level in the correctness of *both* intervals?

   (e) Prepare residuals plots: (i) residuals versus $\hat{Y}$; (ii) residuals versus X; and (iii) normal probability plot of residuals. Do the model assumptions – quadratic mean function, constant variance, normality – appear to be satisfied for these data? Comment on the impact this may have on your significance test from part (c), and prediction intervals from part (d).

2. (Project 8.40 in KNN) Continue with the *SENIC Project* data, but this time we consider regressing infection risk `Risk` against average length of stay `Stay`, average age of patients `Age`, routine chest X-ray ratio `Xray` (three continuous predictors), and medical school affiliation `MS`, which takes the value 1 if Yes and 2 if No.

   (a) Change the affiliation variable `MS` that takes the value 1 if Yes and 0 if No.

   (b) Prepare a scatterplot matrix of the response and three continuous predictor variables, where data points corresponding to hospitals with a medical school affiliation are indicated by a different plotting symbol. Describe the relationships among the variables.

   (c) Letting Y denote the response and X_1, X_2, X_3 the continuous predictors and X_4 the indicator variable for medical school affiliation, fit the mean functions

   $$\mathrm{E}[Y|\mathbf{X}=\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

   and

   $$\mathrm{E}[Y|\mathbf{X}=\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{14} x_1 x_4 + \beta_{24} x_2 x_4 + \beta_{34} x_3 x_4$$

   assuming constant variance and normality in both cases.

      i. Explain in plain English what each of these models means exactly. (You don't have to include this in your answer, but you should know the interpretation of *every single parameter* in both mean functions.)

      ii. Conduct an F-test of the reduced model versus the full model, that is, a test of the null hypothesis

      $$H_0 : \beta_{14} = \beta_{24} = \beta_{34} = 0 \qquad v.s. \qquad H_1 : \text{other wise.}$$

      What is your conclusion?

   (d) Working with the reduced model, estimate the effect of medical school affiliation on infection risk using a 95% confidence interval. Interpret your interval estimate.

3. (Project 8.41 in KNN) Continuing with the *SENIC Project* data, but here we will regress average length of stay `Stay` on age `Age`, routine culturing ratio `Cult`, average daily census `Cen`, and available facilities and services `AFS`; we also consider geographical region `Reg`, a categorical variable taking the values

$$\texttt{Reg} = \begin{cases} 1 & \text{Northeast} \\ 2 & \text{Midwest} \\ 3 & \text{South} \\ 4 & \text{West} \end{cases}$$

   (a) Prepare a scatterplot matrix of the continuous variables (don't worry about separate markings for geographic region). Briefly describe the relationships among the variables, and any other interesting features of the data.

   (b) Fit the first-order regression model with separate intercepts for the four regions. Report your estimated mean function.

   (c) Carefully interpret the estimated coefficient of routine culturing ratio `Cult`. Obtain and interpret a 99% confidence interval for the *true* regression coefficient of `Cult`.

   (d) Test the null hypothesis that average length of stay does not vary by geographic region. Clearly state your null and alternative hypotheses, obtain and interpret a P-value, and clearly state your conclusion.

4. (Ridge Regression)

   (a) Derive the expression for the estimator $\widehat{\beta}_\lambda$ for the Ridge regression criterion:

$$\min_\beta Q(\beta) := \|Y - x\beta\|^2 + \lambda\|\beta\|^2.$$

   (b) We can easily see that $\widehat{\beta}_\lambda = (x^\top x)^{-1}x^\top Y$ when $\lambda = 0$. Show that $\widehat{\beta}_\lambda \to 0$ as $\lambda \to \infty$.

   (c) Consider the following modified ridge estimator:

$$\widetilde{\beta}_\lambda = \lambda\left(x^\top x + \lambda I_p\right)^{-1} x^\top Y.$$

   What does this estimator converge to as $\lambda \to \infty$?