Homework 9

Here we present solutions to the problems posted in the ninth (and last) homework assignment. The solutions and related commentary are put in italics. Remember that problems can have several different but still correct ways of solving them.

Multiple choice questions

1. Which one of the following is correct?

   ***A** the residuals form a basis for assessing the standard deviation of the error term in a well fitted multiple regression;

   B the residuals are defined as the fitted values less the mean observed value;

   C the residuals are only important for diagnostic purposes and are not important for prediction purposes;

   D the residuals are always within 2 standard deviations of the predicted values;

   E none of statements A, B, C, D above is correct.

   Solution: *Residuals play an important role both in the diagnostic stage of model fitting and in estimation of standard deviation of the error term once the model is well fitted. They are defined as the observed values less the predicted values.*

2. t-ratios in multiple regression

   A are expected to vary between +2 and -2;

   B test the significance of differences between regression coefficients;

   ***C** test the statistical significance of individual regression coefficients;

   D exceed 2 only if the corresponding X variables are not contributing to variability of the response variable;

   E may be plotted against the fitted values to assess the goodness of fit.

   Solution: *t-ratios in a well fitted multiple regression follow Student-t distribution and are used to test for significance of the corresponding coefficients.*

3. Which one of the following is correct?

   A the predicted values are the values of the X variables multiplied by the corresponding coefficients;

   B the predicted values can be defined only for the X-values that have been used to fit the regression model;

   ***C** in order to evaluate residuals the predicted values have to be evaluated at the X-values used to fit the regression model;

   D the average of predicted values is always zero;

   E the predicted values are not affected by deleting outliers from the data.

Solution: *The residuals are defined as the observed values minus the predicted (fitted) values, thus the latter have to be evaluated first.*

4. Which of the following is *not* a standard graphical tool in the diagnostic of multivariate regression models?

   A a plot of deleted residuals against fitted values;

   ***B** a plot of the fitted values against the observed values;

   C a Normal plot of the deleted residuals;

   D dotplots of the response and explanatory variables;

   E a time series plot of the deleted residuals.

   Solution: *A preliminary diagnostic is to examine both the explanatory and response variable for extremal values which the best is done by looking at their dotplots. In a further search for exceptional deviation of the data from the regression model, deleted residuals are typically examined through: a plot of deleted residuals of against the fitted (predicted) values, a Normal plot of the deleted residuals, a time series plot of the deleted residuals. Plotting the fitted (predicted) values against the observed values typically does not provide with an insight into deviation from the regression model.*

5. Given a set of explanatory variables, the parameters of a multiple regression model

   A emerge from the initial diagnostic stage of the model fitting;

   ***B** include the regression coefficients which determine the prediction formula and the error standard deviation which is the key parameter in determining prediction error;

   C determine the observed values of the response variable;

   D determine the exceptional values of the response variable.

   E should be first tested for significance before any diagnostic of the model is performed.

   Solution: *The parameters in the multiple regression model include the constant term, the coefficients standing by each explanatory variable (X-variables) and the standard deviation of the error term. All except the last one are used to obtain predicted values while the standard deviation is used to determine the prediction error.*

Problems

1. In the textbook and lecture, the dependence of 'Jobtime' variable on the explanatory variables: 'Units', 'Operations per Unit', 'Total Operations' and 'Rushed' has been studied through the multivariate regression analysis. After the diagnostic process, three values have been excluded corresponding to the order numbers: 9, 11, and 16. Here is the computer printout representing the multivariate regression fit to all the data

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  77.24147   44.75814   1.726    0.105
   Units     -0.15074    0.11215  -1.344    0.199
   OperU      7.15154    4.30466   1.661    0.117
   TotOper    0.11460    0.01322   8.668  <0.00001 ***
   Rush     -24.94413   19.11023  -1.305    0.211
---

Residual standard error: 37.46 on 15 degrees of freedom
Multiple R-Squared: 0.9729
```

and the one for the data from which 9, 11, and 16 orders have been excluded

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.216498   9.079992   4.870 0.000385 ***
    Units    -0.069307   0.028526  -2.430 0.031758 *
    OperU     9.828585   0.887341  11.076 <0.00001 ***
    TotOper   0.107795   0.004114  26.202 <0.00001 ***
    Rush    -37.960103   3.857205  -9.841 <0.00001 ***
---

Residual standard error: 7.413 on 12 degrees of freedom
Multiple R-Squared: 0.9978
```

- Write down the models corresponding to these two fits involving the two sigma prediction band.
- Give the arguments based on the computer printouts for deletion of 9, 11, and 16 orders when fitting the model.
- Evaluate the residuals to the 9th, 11th, and 16th orders for both the fits. Compare them with two sigma prediction band and comment.
- According to the second fit how much time is saved when the requested job is rushed?
- Suppose that the company has received an order for a job that has 85 units and 14 operations per unit. The client would like to know the rushed and non-rushed jobtimes that will be required for completion of the order. Provide with the appropriate information based on the fitted regression model.
- Suppose that the client opted for the rushed order and it took 272 hours to complete the job. Is it in line with the prediction given to the client prior to taking the order?
-

Solution:

- *The model fit on all the data:*

  $$Jobtime = 77.24 - 0.15 * Units + 7.15 * OperU + 0.115 * TotOper - 24.9 * Rushed \pm 2 * 37.46$$

The model fit on the data with the 9th, 11th, and 16th orders excluded:

$$Jobtime = 44.22 - 0.069 * Units + 9.82 * OperU + 0.108 * TotOper - 37.9 * Rushed \pm 2 * 7.41$$

- The main difference between the two models is the reduction in the prediction standard error from 37.46 in the first model to 7.41 in the second model. This favors the second model as being more precise in the predictions. Additionally we observe that the second model has improved R^2 coefficient (0.9978 vs. 0.9729) although in both the cases it is very high. Finally, in the second model all coefficients in the regression are significant while in the first one only one is significant (TotOper).
- The residuals for the 9th, 11th, and 16th observations for the first model are

  $$\begin{aligned} Res_9 &= 260 - (77.24 - 0.15 * 21 + 7.15 * 9 + 0.115 * 189 - 24.9) = 124.73 \\ Res_{11} &= 835 - (77.24 - 0.15 * 426 + 7.15 * 14 + 0.115 * 5964) = 35.7 \\ Res_{16} &= 775 - (77.24 - 0.15 * 520 + 7.15 * 12 + 0.115 * 6240) = -27.64 \end{aligned}$$

  and the corresponding ones for the second model are

  $$\begin{aligned} Res_9 &= 260 - (44.22 - 0.069 * 21 + 9.82 * 9 + 0.108 * 189 - 37.9) = 146.34 \\ Res_{11} &= 835 - (44.22 - 0.069 * 426 + 9.82 * 14 + 0.108 * 5964) = 38.58 \\ Res_{16} &= 775 - (44.22 - 0.069 * 520 + 9.82 * 12 + 0.108 * 6240) = -25.1. \end{aligned}$$

  We observe that the residual for the 9th order is not in line with the two sigma band for either of the models. While the residuals for the 11th and 16th observations are within the two sigma band for the first model. We conclude that although the second model is more accurate for 'typical' data it fails to provide accurate predictions for the extremal data.
- We observe that the coefficient for 'Rushed' variable is -37.9 which represents the reduction in jobtime if 'Rushed' is changed from 0 to 1.
- We will use the second fit and we obtain

  $$Jobtime1 = 44.22 - 0.069 * 85 + 9.82 * 14 + 0.108 * 85 * 14 \pm 2 * 7.41 = 304.4 \pm 14.8$$

  for the non-rushed job and

  $$Jobtime2 = 304.3 - 37.9 \pm 14.8 = 266.4 \pm 14.8$$

  for the rushed job.
- The two sigma prediction band for the non-rushed job time is [251.6,281.2] and thus the result is in line with the initial prediction.