Statistics 206

Homework 9

Not Due

Problems 1 and 2. Model validation and model diagnostic case study in R. *Diabetes data (Cont'd from homework 8). This data consist of* 19 *variables on* 403 *subjects from* 1046 *subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. We will consider building regression models with* `glyhb` *as the response variable as Glycosolated Hemoglobin* > 70 *is often taken as a positive diagnostics of diabetes. The data set and description are under Files/Homework. Please attach your R codes and plots.*

1. **Model validation.** We now consider validation of the two models fs1 and fs2 selected by the forward stepwise procedure.

   (a) **Internal validation of Models fs1 and fs2**. For this purpose, we need to compute C_p and $Press_p$ for these models. For C_p, we need an unbiased estimator of the error variance σ^2. The largest model we have considered so far is Model 2. However, this model has a very large number of regression coefficients (relative to the sample size), making its parameter estimation unreliable due to large sampling variability. Therefore, we decided to use a smaller model consisting of all predictors identified by Model fs1 (the forward stepwise selected first-order model), as well all the 2-way interaction terms among these predictors. Denote this model by Model 3. Note that, Model fs2 is also a sub-model of Model 3. How many regression coefficients are there in Model 3 ? What is MSE from Model 3? Calculate SSE_p, MSE_p, C_p and $Press_p$ for Models fs1 and fs2 and briefly comment on the results, e.g., does it appear to be substantial model bias in these two models? Should overfitting be a concern?

   (b) **External validation using the validation set**. We now fit Models fs1 and fs2 on the validation data set. Compare the fitted regression coefficients from the training data and those from the validation data. Are the two sets of estimated regression coefficients having the same sign? Are their values similar? How about the two sets of standard errors? Does it appear that Models fs1 and fs2 have consistent estimates on the training data and validation data? Calculate the mean squared prediction error (MSPE) using the validation data for each of the two models. How do these $MSPE_v$ compare with the respective $Press_p/n$ and SSE_p/n ()Note here n is the sample size of the training data, i.e., 183)? Which model among the two has a smaller $MSPE_v$?

   (c) Based on both internal and external validation, which model you would choose as the final model? Fit the final model using the entire data set (training and validation combined). Write down the fitted regression function and report the R summary() and anova() output.

2. **Model diagnostic: Outlying and influential cases.** Conduct model diagnostic for the final model from the previous problem.

   (a) Draw residual vs. fitted value plot and residual Q-Q plot and comment on these plots.

   (b) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure at $\alpha = 0.1$.

   (c) Obtain the leverage and identify any outlying X observations. Draw residual vs. leverage plot.

   (d) Draw an influence index plot using Cook's distance. Are there any influential cases according to this measure?

   (e) Calculate the average absolute percent difference in the fitted values with and without the most influential case identified from the previous question. What does this measure indicate the influence of this case?

3. **(Optional Problem). Studentized deleted residuals.** In the following, no assumption is made on the data or the model unless it is explicitly stated.

   (a) Assume the observed response vector $\mathbf{Y} \in \mathbb{R}^n$ has $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$. Show that, the ith deleted residual $d_i = Y_i - \widehat{Y}_{i(i)}$ has

$$Var(d_i) = \frac{\sigma^2}{1 - h_{ii}}.$$

   (b) Let

$$SSE_{(i)} = \sum_{j:j \neq i} (Y_j - \widehat{Y}_{j(i)})^2, \quad MSE_{(i)} = \frac{SSE_{(i)}}{n - p - 1},$$

   i.e., $SSE_{(i)}$ and $MSE_{(i)}$ are the SSE and MSE of the regression fit excluding case i, respectively. Show that

$$SSE_{(i)} = SSE - \frac{e_i^2}{1 - h_{ii}}.$$

   Hints: Recall that

$$SSE_{(i)} = \tilde{\mathbf{Y}}^T (\mathbf{I} - \mathbf{H}) \tilde{\mathbf{Y}},$$

   where

$$\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{d}_{(i)}, \quad \textit{where}, \quad \mathbf{d}_{(i)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ d_i \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

   i.e., $\tilde{\mathbf{Y}}$ *is the same as* $\mathbf{Y}$ *except for the* i*th element, where it is* $\widehat{Y}_{i(i)}$.

(c) Show that the studentized deleted residual

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{d_i}{\sqrt{MSE_{(i)}/(1-h_{ii})}}$$

can be computed by:

$$t_i = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}.$$

(d) Under the Normality assumption, i.e., $\mathbf{Y}$ is an n-dimensional Normal random vector with $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, show that $SSE_{(i)}$ is independent with Y_i and $\widehat{Y}_{i(i)}$. Therefore, $SSE_{(i)}$ is independent with d_i. If we further assume that the model is correct, then the deleted residual d_i has mean zero and the studentized deleted residual t_i follows a $t_{(n-p-1)}$ distribution.

4. **(Optional Problem). Cook's distance.** Show that the Cook's distance

$$D_i := \frac{\sum_{j=1}^n (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \times MSE}, \quad i = 1, \cdots, n$$

can be computed by:

$$D_i = \frac{e_i^2}{p \times MSE} \frac{h_{ii}}{(1-h_{ii})^2}.$$

Hints: Note that

$$\sum_{j=1}^n (\widehat{Y}_j - \widehat{Y}_{j(i)})^2 = (\mathbf{Y} - \tilde{\mathbf{Y}})^T \mathbf{H} (\mathbf{Y} - \tilde{\mathbf{Y}}).$$

5. **Regression formulations of one-way ANOVA model**.

(a) Consider the *cell means formulation* discussed in class:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \cdots, I, j = 1, \cdots, n_i.$$

Express this model as a linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Specify $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. What is $\mathbf{X}^T\mathbf{X}$ and what is $\mathbf{X}^T\mathbf{Y}$? What is the LS estimator of $\boldsymbol{\beta}$? Derive the fitted values and residuals.

(b) Consider the alternative formulation used by `R` function `lm`:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, I, j = 1, \cdots, n_i, \quad \alpha_1 = 0.$$

Express this model as a linear regression model by specifying $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. Compare it with the linear regression model with $I-1$ indicator variables for factor levels (with level 1 as the reference class). What do you find?

6. **One-way ANOVA case study in R.**

A company uses six filling machines of the same make and model to place detergent into cartons that show a label weight of 32 ounces. The production manger has complained that the six machines do not place the same amount of fill into the cartons. A consultant requested that 20 filled cartons be selected randomly from each of the six machines and the content of each carton weighted. The observations were recorded in terms of the deviations of weights from 32 ounces. The data is under Files/Homework/filling.txt: The first column is the observation, the second column is the index for the filling machine and the third column is the index for the carton. Consider fitting the one-way ANOVA model to this data.

(a) What is the response variable? What is the factor? How many levels are there for this factor? Name the design of this study.

(b) Draw side-by-side box plots of the response for the factor levels. Do the factor level means appear to differ? Does the variability of the observations within each factor level appear to be approximately the same for all factor levels?

(c) Fit the one-way ANOVA model. Draw residual versus fitted value plot and residual Q-Q plot. Comment on model assumptions. Are remedial measures needed? *(Hint: When using the* `lm` *function, remember to declare the factor as* `factor`*.)*

(d) Obtain the estimated factor levels means and obtain the ANOVA table.

(e) Test whether or not the mean fill differs among the six machines at level 0.05. State the null and alternative hypotheses, the decision rule and the conclusion.

(f) Construct a 99% confidence interval for μ_2. If we are interested in all factor level means, what multiple comparison procedure shall we use? What would be the corresponding 99% confidence interval for μ_2?

(g) How many pairwise comparisons of factor levels means are there? If we are interested in all these pairwise comparisons, what multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? Construct the corresponding 95% confidence intervals for all pairwise comparisons. At familywise significance level 0.05, which pairs of factor level means should be declared as being different?

(h) What if we are only interested in 6 **pre-specified** pairwise comparisons, which multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? What if we are only interested in the 6 pairwise comparisons that show the most differences in the data (i.e corresponding to the six largest $|\widehat{D}|$), which procedure shall we use and what are the corresponding C.Is?

(i) If we are interested in all possible contrasts, which multiple comparison procedure shall we use? Construct the corresponding 95% confidence interval for the contrast:

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}.$$

Test whether or not $L = 0$ with family-wise-error-rate controlled at 0.05. What if we are interested in 20 pre-specified contrasts (including L), which multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$?

Construct the corresponding 95% confidence interval for L (assuming L is in this prespecified set of contrasts). What do you find?