Statistics 206

Homework 7

Due : November 23, 2015, In Class

1. Tell true or false of the following statements.

   (a) With many potential X variables, we can first fit a model with all these variables, and then drop those having non-significant regression coefficients by t-tests.

   (b) A correct model must be a good model.

   (c) With too many nuisance X variables, the model tends to have a large model bias.

   (d) With the R_p^2 criterion, we aim to select the model with the largest R_p^2.

   (e) For models of the same size, their $Press_p$ values are monotonically decreasing with the decreasing of SSE_p.

   (f) For models of the same size, their C_p, AIC_p, BIC_p values are monotonically decreasing with the decreasing of SSE_p.

   (g) For a given model, its SSE_p is always no greater than its $Press_p$.

   (h) Compared with AIC_p, BIC_p criterion tends to select smaller models because it puts more penalty on model size.

   (i) The stepwise procedures are guaranteed to find the best model according to a given criterion.

Problems 2 to 6. Model Selection, Validation and Diagnostic. *Diabetes data. 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 Resources/Homework. Please attach your R codes and plots.*

2. **Processing of the data.**

   (a) Read the data into R. Replace the missing values in the variable `frame` (indicated by an empty string ‘’) by ’NA’ and drop the old class ‘’.

   (b) Drop `id, bp.2s, bp.2d` from the data. The column `id` are patient IDs and thus is not a meaningful predictor. The variables `bp.2s, bp.2d` have many missing values. You may use the code:

   ```
   > drops=c("id","bp.2s", "bp.2d")
   > data=diabetes[,!(names(diabetes)%in%drops)]
   ```

(c) Which of the (remaining) variables are quantitative variables and which are qualitative variables? Draw histogram for `glyhb` and comment on its distribution. Draw histograms for the rest quantitative variables and draw pie charts for qualitative variables.

(d) It turns out that the distribution of `glyhb` is severely right-skewed. Thus, you want to consider some transformations. Draw histogram for $\log(glyhb)$, $\sqrt{glyhb}$ and $\frac{1}{glyhb}$, respectively. Which distribution appears to be the most Normal like among the three? Denote it by `glyhb*`.

(e) Replace the column `glyhb` in `data` by `glyhb*` and refer to `glyhb*` as `glyhb` hereafter and use it as the response variable.

(f) Drop all the cases having missing value. You may use the code:

```
> index.na=apply(is.na(data), 1, any)
                                 ## identify cases with missing value.
> data.s=data[index.na==FALSE,]  ##drop cases with missing value.
> any(is.na(data.s)) ## this should return FALSE -- no NA in data.s
> dim(data.s)  ##this should return 366 16: 366 cases, 16 variables.
> table(data.s$frame)  ## this should show three classes.
```

(g) Draw scatterplot matrix and obtain the pairwise correlation matrix for all quantitative variables. Do you observe nonlinearity?

(h) Draw side-by-side box plots to show how `glyhb` is distributed in male and female, and how it is distributed in the three `frame` classes.

(i) Randomly split data into two equal halves: a training data set and a validation data set. You may use the code:

```
> set.seed(10) ## set seed for random number generator
               ##so everyone gets the same split of the data.
> n.s=nrow(data.s) ## number of cases in data.s (366)
> index.s=sample(1: n.s, size=366/2, replace=FALSE)
               ## randomly sample 183 cases to form the training data.
> data.c=data.s[index.s,]  ## get the training data set.
> data.v=data.s[-index.s,]
               ## the remaining 183 cases form the validation set.
```

(j) Examine whether the training data and validation data look alike. Draw side-by-side boxplots for `glyhb`, `stab.glu`, `ratio`, `age`, `bp.1s` and `waist`, in training data and validation data, respectively. Are these variables having similar distributions in these two sets?

3. **Selection of first-order effects.** We now consider subsets selection from the pool of all first-order effects of the 15 predictors.

(a) Fit a model with all first-order effects (Model 1). How many regression coefficients are there in this model? What is the MSE from this model? Apply box-cox procedure on this model. Does it appear that any transformation of the response variable is still needed?

(b) Consider best subsets selection using the R function `regsubsets()` from the `leaps` library with Model 1 as the full model. Return the top 1 best subset of all subset sizes up to 16. Get SSE_p,R_p^2, $R_{a,p}^2$, C_p, AIC_p,BIC_p for each of these models. Identify the best model according to each criterion. For the best model according to C_p criterion, what do you observe about its C_p value? Do you have a possible explanation?

(c) We now explore stepwise procedures. Apply the `forward stepwise procedure` using R function `stepAIC()`, starting from the null-model and using the AIC_p criterion. What is the model being selected? Denote this model by Model fs1. Is it the "best" model according to AIC_p criterion identified in the previous question? If not, how its AIC value compare with AIC of the "best" model?

(d) Comment on the residual vs. fitted value plot and the residual Q-Q plot of Model fs1. Does this model appear to be adequate?

4. **Selection of first- and second- order effects.** We now consider subsets selection from the pool of first-order effects as well as 2-way interaction effects of the 15 predictors.

   (a) Fit a model with all first-order and 2-way interaction effects (Model 2). How many regression coefficients are there in this model? What is the MSE from this model? Do you have any concern about the fitting of this model and why?

   (b) Apply the `forward stepwise procedure` using R function `stepAIC()`, starting from the null-model and using the AIC_p criterion. What is the model being selected? Denote this model by Model fs2. Compare its AIC value with that of Model fs1. What do you find?

   (c) Comment on the residual vs. fitted value plot and the residual Q-Q plot of Model fs2. Does this model appear to be adequate?

   (d) Apply the `forward selection procedure`. What model do you end up with?

   Notes: You could try best subsets selection using the R function `regsubsets()` from the `leaps` library, e.g. return the top 1 best subset of all subset sizes up to 16 with the full model being Model 2. However, be careful, you may have to stop the R session due to the slowness of this procedure! So save all that you want to save before you try this.

5. **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.

6. **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?

7. **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 ith 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) **(Optional Problem).** 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.

8. **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

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