ASSIGNMENT 2: Multiple Linear Regression

Model Selection

Problem 1: The amount of water used by the production facilities of a plant varies. Observations on water usage and other,possibility related,variables were collected for 249 months. The data are given in **water.csv file** The explanatory variables are

TEMP= average monthly temperature(degree celsius)

PROD=amount of production(10cubic)

DAYS=number of operationing day in the month

HOUR=number of hours shut down for maintenance

The response variable is USAGE=monthly water usage (gallons/minute)

From Exercise 1 and 2, assume that the best fitted model is

$$\widehat{USAGE} = \hat{\beta}_0 PROD + \hat{\beta}_1 TEMP + \hat{\beta}_2 HOUR + \hat{\beta}_3 PROD * TEMP + \hat{\beta}_4 PROD * HOUR$$

a) Many researchers avoid the problems of multicollinearity by always omitting all but one of the ‘’redundant’’ variables from the model. By checking all pairwise combinations of predictors in scatterplots and using the VIF function, do you detect any high correlation (r>0.8) between predictors? Does there appear to be any problem with multicollinearity assumption?

b) Conduct a test for heteroscedasticity (non constant varaince) and plot a residual plot. Does there appear to be any problem with homoscedasticity assumption?

c) Provide a histogram for residuals, a normal Q-Q plot, and the Shapiro -Wilk test. Does there appear to be any problem with normality assumption?

d) Plot the residuals vs predicted value $\hat{Y}$ plot, do you detect any patterns? Does there appear to be any problem with linearity assumption?

e) Do you detect any outliers by using Cook’s distance measure (using cooks.distance()>1) and Residual vs Leverage plot?

f) From part a-e, determine whether your model meets the assumptions of the analysis.

Suggestion!

For situations in which the errors are homoscedastic but nonnormal, normalizing transformations are available. This family of transformations on the dependent variable includes $\sqrt{y}$ and log(y) ,as well as such simple transformations as y^2,$1/\sqrt{y}$, and $1/y$. Box and Cox have developed a procedure for selecting the appropriate transformation to use.

Problem 2. **Collusive bidding in road construction**. Road construction contracts in the state of Florida are awarded on the basis of competitive, sealed bids; the contractor who submits the lowest bid price wins the contract. During the 1980s, the Office of the Florida Attorney General (FLAG) suspected numerous contractors of practicing bid collusion (i.e., setting the winning bid price above the fair, or competitive, price in order to increase proect margin). By comparing the bid prices (and other important bid variables) of the fixed (or rigged) contracts to the competitively bid contracts, FLAG was able to establish invaluable benchmarks for detecting future bid-rigging. FLAG collected data for 279 road construction contracts. For each contract, the following variables shown below were measured and are only considered for this problem.

1. Price of contract ($) bid by lowest bidder, LOWBID.
2. Department of Transportation (DOT) engineer's estimate of fair contract price ($), DOTEST.
3. Status of contract (1 if fixed, 0 if competitive), STATUS
4. District (1, 2, 3, 4, or 5) in which construction project is located, DISTRICT.
5. Number of bidders on contract, NUMIDS.
6. Estimated number of days to complete work, DAYSEST.
7. Length of road project (miles), RDLNGTH.
8. Percentage of costs allocated to liquid asphalt, PCTASPH.
9. Percentage of costs allocated to base material, PCTBASE.
10. Percentage of costs allocated to excavation, PCTEXCAV.
11. Percentage of costs allocated to mobilization, PCTMOBIL.
12. Percentage of costs allocated to structures, PCTSTRUC.
13. Percentage of costs allocated to trafic control, PCTTRAF.

The data are saved in the file named **FLAG2.txt**

(a) Consider building a model for the low-bid price (Y). Apply **Stepwise Regression Procedure with pent=0.05 and prem=0.1** to the data to find the independent variables most suitable for modeling Y.

(b) Consider building a model for the low-bid price (Y). Apply **Forward Regression Procedure with pent=0.05** :*ols_step_forward_p(fullmodel,pent=0.05)* to the data to find the independen t variables most suitable for modeling Y.

(c) Consider building a model for the low-bid price (Y). Apply **Backward Regression Procedure with prem=0.05** :*ols_step_backward_p(fullmodel,prem=0.05)* to the data to find the independent variables most suitable for modeling Y.

(d) Using the full model with all predictors, test the individual t-test at $\alpha = 0.05$. what predictors should be added to the model.

(e) Compare the results, parts a-d. Which independent variables consistently are selected as the ''best'' predictors for the first order model? Write the first order model for predicting Y.

(f) Interpret the regression coefficients for each β_i

(g) Apply **All Possible Regressions Selection Procedure** to confirm that the independent variables in part (d) are suitable for modeling Y. Provide all three criteria value (Cp,AIC,R^2_{adj}) for the model selected.

(h) Build a complete first order model with interaction term. Would you suggest this model for predicting Y? Explain.

(i) Compare the RMSE from the first order model in part (d) with the interation model in part (h). Interpret the result.

(j) Find the R^2_{adj} and interpret the result from part (h)

Problem 3: An author studied family caregiving in Korea of older adults with dementia. The outcome variable, caregiver burden (BURDEN), was measured by the Korean Burden Inventory (KBI) where scores ranged from 28 to 140 with higher scores indicating higher burden. The following independent variables were reported by the researchers:

1. CGAGE: caregiver age (years)
2. CGINCOME: caregiver income (Won-Korean currency)
3. CGDUR: caregiver-duration of caregiving (month)
4. ADL: total activities of daily living where low scores indicate the elderly perform activities independently.
5. MEM: memory and behavioral problems with higher scores indicating more problems.
6. COG: cognitive impairment with lower scores indicating a greater degree of cognitive impairment.
7. SOCIALSU: total score of perceived social support (25-175, higher values indicating more support). The reported data are in file **KBI.csv**.

Answer the following questions

a) Use stepwise regression (with stepwise selection) to find the ''best'' set of predictors of caregiver burden. [Hint: Use pent =0.1 and prem=0.3].

b) Use backward elimination regression to find the ''best'' set of predictors of caregiver burden. [Hint: Use prem=0.1]

c) Use forward elimination regression to find the ''best'' set of predictors of caregiver burden. [Hint: Use pent=0.1]

d) Use all-possible-regressions-selection to find the ''best'' predictors of caregiver burden (Cp, AIC, Adjuster R^2, R^2)

e) Compare the results, parts a-c. Which independent variables consistently are selected as the ''best'' predictors? Comment the value of the adjusted R^2.

f) Explain how you would use the results, parts a-c, to develop a model for caregiver burden. Check for interactions, normality and linearity assumptions.

g) Do you detect any outliers by using leverage values greater that $\frac{2p}{n}$? Remove those outliers and fit again the model with variables selected in question a.

h) Do you detect any outliers by using leverage values greater that $\frac{3p}{n}$? Remove those outliers and fit again the model with variables selected in question a.

i) Do you notice any difference in the results with the model from part a and the best fit model between part g and h? Comment.