Code ▾

Multiple Linear Regression

ASSIGNMENT 2

First-order Model with Interaction Terms (Quantitative and Qualitative Variable and Model Selection

Deadline: Nov. 25 , 2022, by 11:59 pm. Submit to Gradescope.ca

© Thuntida Ngamkham 2022

Problem 1. The file **tires.csv** provides the results of an experiment on tread wear per 160 km and the driving speed in km/hour. The researchers looked at 2 types of tires and tested 20 random sample tires. The response variable is the tread wear per 160 km in the percentage of tread thickness, and the quantitative predictor is the average speed in km/hour.

Hide

```
tires=read.csv("c:/Users/thunt/OneDrive - University of Calgary/dataset603/tires.CSV", header =
TRUE)
str(tires)
```

```
'data.frame':   140 obs. of  3 variables:
 $ type: chr  "A" "A" "A" "A" ...
 $ wear: num  0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.4 0.4 0.4 ...
 $ ave : int  80 80 80 80 80 80 80 88 88 88 ...
```

Answer the following questions

a. Use the individual T-test to evaluate the significant predictors from the full model at $\alpha = 0.05$ and write the estimated best fit model.
b. Based on the output in (a), define the dummy variable that explains the two types of tires.
c. From the best fit model in part (a), interpret all possible regression coefficient estimates,$\hat{\beta}_i$.
d. From the best fit model in part (a), you can improve this model by adding an interaction term(s). Evaluate whether the interaction term(s) is(are) significant to be added in the model at $\alpha = 0.05$. Summarize which model would you suggest using for predicting y.
e. From the model in part (d), report the adjusted-R2 value from the model selected and interpret its value.
f. Predict the average tread wear per 160 km in the percentage of tread thickness for a car with type A with the average speed of 100 km/hour from the model selected in part (d).

Problem 2. A team of mental health researchers wishes to compare three methods (A, B, and C) of treating severe depression. They would also like to study the relationship between age and treatment effectiveness as well as the interaction (if any) between age and treatment.

Each member of a simple random sample of 36 patients, comparable with respect to diagnosis and severity of depression, was randomly assigned to receive treatment A, B, or C. The data are given in **MentalHealth.csv**.

Answer the following questions

a. Which is the dependent variable (the response variable)?

b. What are the independent variables (the predictors)?

c. Draw a scatter diagram of the sample data with EFFECT on the y-axis and AGE on the x-axis using different symbols/colors for each of the three treatments. Briefly summarize the visualization. [Hint: Check MLR part II under Interaction Effect in MLR with both Quantitative and Qualitative Variable models].

d. Is there any interaction between age and treatment? Test the hypothesis at $\alpha = 0.05$.

e. From part (d), write the final model with sub-models for predicting the treatment effectiveness. Please ensure you substitute all regression coefficients to the models.

f. Interpret the effect of treatment from sub-models in part (e).

g. Plot the three regression lines on the scatter diagram obtained in part (c). May one have the same conclusion as in part (f)?

Problem 3. **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 the 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 traffic control, PCTTRAF.

The data are saved in the file named **FLAG2.txt**. Answer the following questions:

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 independent 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. Test the individual t-test at $\alpha = 0.05$ to evaluate the variables in the model. What predictors should be kept in the model?

e. Compare the results, parts (a)-(d). Which independent variables consistently are selected as the ''best'' predictors for the model? Write all possible additive model(s) for predicting Y. Note! Proposing more than one model is acceptable.

f. Assume that your model selected in part (e) contains the following predictors: DOTEST,STATUS,NUMBIDS, and DISTRICT. Calculate the absolute difference in average contact bid price (by the lowest bidder) between District 1 and 4,when other predictors are held as a constant.

g. Assume that your model selected in part (e) contains the following predictors: DOTEST,STATUS,NUMBIDS, and DISTRICT. Calculate the difference in average contact bid price (by the lowest bidder) between District 2 and 5,when other predictors are held as a constant.

h. Assume that your model selected in part (e) contains the following predictors: DOTEST,STATUS,NUMBIDS, and DISTRICT. Build the first order model with interaction terms. Write the best fit model for predicting Y.

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

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

Problem 4: 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 the file **KBI.csv**.

Answer the following questions:

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

b. Use all-possible-regressions-selection to find the ‘’best’’ predictors of caregiver burden (Cp, AIC, RMSE, Adjusted R^2). Report all significant predictors.

c. Compare the results, parts a-b. Which independent variables consistently are selected as the ‘’best’’ predictors? Build the first order model with interaction terms, evaluate which interation terms are significant to be added in the model, and conclude the the final model for the prediction.