ASSIGNMENT 1: Multiple Linear Regression

First order Model with Interaction Term (Quantitative and Qualitative Variable)

Deadline: Nov. 1st, 2019, by 9pm. Submit to Dropbox via D2L.

Problem 1. (From Exercise 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 250 months. The data are given in **water.csv file**. The explanatory variables are

TEMP= average monthly temperature (degree celsius)

PROD=amount of production (in hundreds of cubic)

DAYS=number of operationing day in the month (days)

HOUR=number of hours shut down for maintenance (hours)

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

a. Fit the model containing all four independent variables. What is the estimated multiple regression equation?

b. Test the hypothesis for the full model i.e the test of overall significance. Use significance level 0.05.

c. Would you suggest the model in part b for predictive purposes? Which model or set of models would you suggest for predictive purposes? Hint: Use Individual Coefficients Test (t-test) to find the best model.

d. Use Partial F test to confirm that the independent variable (removed from part c) should be out of the model at significance level 0.05.

e. Obtain a 95% confidence interval of regression coefficient for TEMP from the model in part c. Give an interpretation.

f. Use the method of Model Fit to calculate R^2_{adj} and RMSE to compare the full model and the model in part c. Which model or set of models would you suggest for predictive purpose? For the final model, give an interpretation of R^2_{adj} and RMSE.

g. (From Exercise 2) Build an interaction model to fit the multiple regression model from the model in part f. From the output, which model would you recommend for predictive purposes?

Problem 2. A collector of antique grandfather clocks sold at auction believes that the price received for the clocks depends on both the age of the clocks and the number of bidders at the auction. Thus, (s)he hypothesizes the first-order model

$$\begin{aligned} y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\ & \textit{where} \\ y &= \text{Auction price (dollars)} \\ X_1 &= \text{Age of clock (years)} \\ X_2 &= \text{Number of bidders} \end{aligned}$$

A sample of 32 auction prices of grandfather clocks, along with their age and the number of bidders, is given in data file **GFCLOCKS.CSV**

a. Use the method of least squares to estimate the unknown parameters β_0 , β_1 , β_2 of the model.

b. Find the value of SSE that is minimized by the least squares method.

c. Estimate s, the standard deviation of the model, and interpret the result.

d. Find and interpret the adjusted coefficient of determination, R^2_{Adj}.

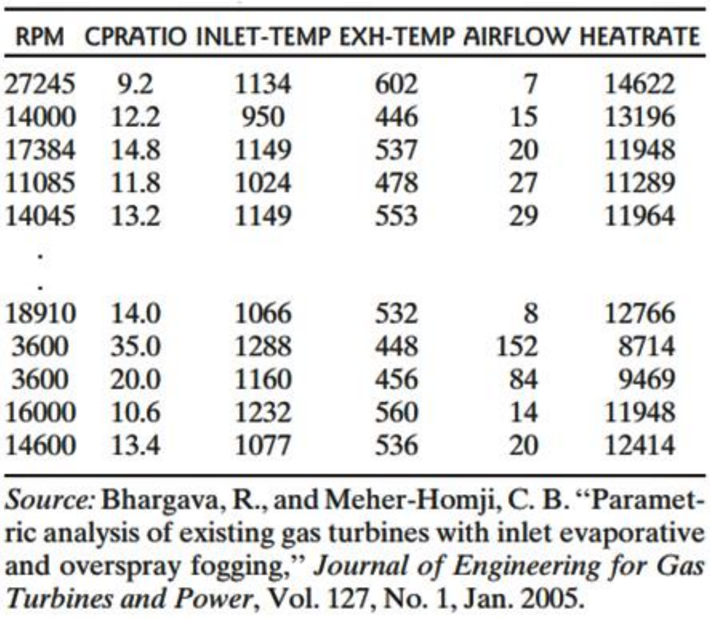
e. Construct the Anova table for the model and test the global F-test of the model at the α = 0.05 level of significance.

f. Test the hypothesis that the mean auction price of a clock increases as the number of bidders increases when age is held constant (i.e., when $\beta_2 \neq 0$). (Use α = 0.05)

g. Find a 95% confidence interval for β_1 and interpret the result.

h. Test the interaction term between the 2 variables at $\alpha = .05$. What model would you suggest to use for predicting y? Explain.

Problem 3. **Cooling method for gas turbines.** Refer to the Journal of Engineering for Gas Turbines and Power (January 2005) study of a high pressure inlet fogging method for a gas turbine engine. The heat rate (kilojoules per kilowatt per hour) was measured for each in a sample of 67 gas turbines augmented with high pressure inlet fogging. In addition, several other variables were measured, including cycle speed (revolutions per minute), inlet temperature (degree celsius), exhaust gas temperature (degree Celsius), cycle pressure ratio, and air mass flow rate (kilograms persecond). The data are saved in the **TURBINE.CSV** file.



RPM	CPRATIO	INLET-TEMP	EXH-TEMP	AIRFLOW	HEATRATE
27245	9.2	1134	602	7	14622
14000	12.2	950	446	15	13196
17384	14.8	1149	537	20	11948
11085	11.8	1024	478	27	11289
14045	13.2	1149	553	29	11964
.					
.					
18910	14.0	1066	532	8	12766
3600	35.0	1288	448	152	8714
3600	20.0	1160	456	84	9469
16000	10.6	1232	560	14	11948
14600	13.4	1077	536	20	12414

Source: Bhargava, R., and Meher-Homji, C. B. "Parametric analysis of existing gas turbines with inlet evaporative and overspray fogging," *Journal of Engineering for Gas Turbines and Power*, Vol. 127, No. 1, Jan. 2005.

The first and last five observations are listed in the table.

(a) Write a first-order model for heat rate (y) as a function of speed, inlet temperature, exhaust temperature, cycle pressure ratio, and air flow rate.

(b) Test the overall significance of the model using $\alpha = 0.01$

(c) Fit the model to the data using the method of least squares. (Suggestion! check both models with and without a predictor that has p-value close to 0.05, and propose the best model.)

(d) Test all possible interaction terms for the best model in part (c) at $\alpha = .05$. What is the final model would you suggest to use for predicting y? Explain.

(e) Give practical interpretations of the β_i estimates.

(f) Find RMSE, s from the model in part (d)

(g) Find the adjusted-R2 value from the model in part (d) and interpret it.

(h) Predict a heat rate (y) when a cycle of speed = 273,145 revolutions per minute, inlet temperature= 1240 degree celsius, exhaust temperature=920 degree celsius, cycle pressure ratio=10 kilograms persecond, and air flow rate=25 kilograms persecond.

Problem 4. 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 percentage of tread thickness and the quantitative predictor is average speed in km/hour.

(a) Define the dummy variable that explains the two types of tires.

(b) Test the additive model at $\alpha = 0.05$ and write a first-order model for the tread wear per 160 km as a function of average speed and type of tires.

(c) Interpret all possible regression coefficient estimates.

(d) Test the interaction term between the 2 variables at $\alpha = .05$. What model would you suggest to use for predicting y? Explain.

(e) From the model in part (d) Find the adjusted-R2 value and interpret it.

(f) Predict the tread wear per 160 km in percentage of tread thickness for a car that has type A with an average speed 100 km/hour.

Problem 5. 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**.

a. Which is the dependent variable?

b. What are the independent variables?

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. Comment.

d. Is there any interaction between age and treatment? [Hint: Use dummy variable coding, the least square method and $\alpha = 0.05$.]

e. How would you interpret the effect of treatment?

f. Plot the three regression lines on the scatter diagram obtained in c. May one have the same conclusion as in question d.?

Problem 6 [Optional]. Erecting boiler drums In a production facility, an accurate estimate of hours needed to complete a task is crucial to management in making such decisions as the proper number of workers to hire, an accurate deadline to quote a client, or cost-analysis decisions regarding budgets. A manufacturer of boiler drums wants to use regression to predict the number of hours needed to erect the drums in future projects. To accomplish this, data for 35 boilers were collected. In addition to hours (y), the variables measured were boiler capacity (x_1 =lb/hr), boiler design pressure (x_2 =pounds per square inch, or psi), boiler type (x_3 =1 if industry field erected, 0 if utility field erected), and drum type (x4 =1 if steam, 0 if mud).The data are saved in the **BOILERS.csv** file.

(a) Write the first order model for hours.

(b) Construct the Anova table for the first order model (the additive model).

(c) Use the Anova table from part b to conduct a test for the full model (Use α = .01).

(d) Would you drop any predictors out of the full model? Explain.

(e) Test individually the interaction terms at $\alpha = .05$. What model would you suggest to use for predicting y? Explain.

(f) Write all possible submodels for two categorical variables (do not have to substitute values of β_i)