Sean Anselmo Assignment 1

First order Model with Interaction Term (Quantitative Variables)

Deadline: Mar. 8, 2024, by 11:59 pm. Submit to Gradescope.ca

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. Ensure that you include the hypotheses, p-value, and conclusion.

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. Make sure you provide your hypotheses, p-value(s), and final 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. Ensure that you include the hypotheses, p-value, and conclusion.

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 RSE to compare the full model in part a 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 RSE.

g. 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? Be sure to explain your process.

```
water <- read.csv("~/603/Ass1/water.csv")
Q1a = lm(USAGE~PROD+TEMP+HOUR+DAYS, data=water)


Q1c =  lm(USAGE~PROD+TEMP+HOUR, data=water)
Q1d = anova(Q1a, Q1c)
```

```
Q1e = confint(Q1c, parm = 'TEMP')

fullRsquared = summary(Q1a)$adj.r.squared
reducedRsquared = summary(Q1c)$adj.r.squared

Q1f = lm(USAGE~(PROD+TEMP+HOUR+DAYS)^2,data=water)
Q1fReduced = lm(USAGE~PROD+TEMP+HOUR+PROD*TEMP+PROD*HOUR, data=water)
```

Answer To Question 1a

$$y = 5.89 + 0.04X_1 + 0.17X_2 - 0.07X_3 - 0.02X_4 + \epsilon$$

where

$$y = \text{Usage}$$
$$X_1 = \text{Prod}$$
$$X_2 = \text{Temperature}$$
$$X_3 = \text{Hour}$$
$$X_4 = \text{Days}$$

Answer To Question 1b The null hypothesis is that Temperature, production, days and hours all do not have an effect on the outcome, Usage.

$$H_0 : PROD/DAYS/HOUR/TEMP = 0$$

. This means the model does not explain any variance in the outcome, Usage. The alternate hypothesis is

$$H_1 : PROD/DAYS/HOUR/TEMP \neq 0$$

However, after looking at the p values of the independent variables it can be concluded that all variables excluding Days have a significant relationship with Usage (p values are below 0.05). The p value of the model is 2x10^-16, which means at least one predictor significantly affects the outcome variable, Usage.

Prod = p value is 3x10^-8 **Temp** = p value is 2x10^-16 **Hour** = p value is 4.1x10^-5 **Days** = p value is 0.502

This means we are able to reject our null hypothesis in favour of our alternate, because there are predictors below our threshold of significance at 0.05. The predictors below this threshold are determined to be significantly affecting the outcome variable in our model.

Answer to Question 1c I would not suggest the model in part B for predictive purposes, because it includes an independent variable that does not significantly affect the outcome variable. The p value for Days is 0.502, which is not below our significance threshold of 0.05. I would use the model that omits Days, and uses Temperature, Hours, and production as predictors. The p values for these predictors are all below 0.05.

Prod = p value is 2x10^-16 **Temp** = p value is 2x10^-16 **Hour** = p value is 4.23x10^-5 **Days** = p value is 0.502

The final model would look as this:

$$Usage = \beta_0 + \beta_1 * Prod + \beta_2 * Temp + \beta_3 * Hour$$

Answer to Question 1d The anova test shows a test stat of 0.4514, and a p value of 0.5023. This means that we do reject the null hypothesis in favour of our alternate.

H_0 = Reduced model and Full model have no significant difference H_1 = Reduced model has significantly better fit than the full model

Therefore we should drop the Days variable as an indicator.

Answer to Question 1e We are 95% confident the Temperature variable has an impact between 0.153, and 0.185. This means that Temperature usually has a positive relationship with the outcome variable.

Answer to Question 1f The $R^2_{adj} = 0.8867$ for our full model, and the RSE is 1.768. Our reduced model has an $R^2_{adj} = 0.8869$ and an RSE of 1.766. The higher adjusted R squared indicates the model's inputs more accurately describes the variance. The RSE being slightly lower indicates the reduced model may be more reliable than the full model.

Answer to Question 1g I would recommend the interaction model that has interactions between PROD and HOUR, and PROD and TEMP. The $R^2_{adj} = 0.9651$, which is noticeawbly higher than 0.8869 of the reduced additive model.

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

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where

y = Auction price (dollars)

X_1 = Age of clock (years)

X_2 = Number of bidders

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 $\alpha = 0.05$ level of significance. Be sure to state your hypotheses, p-values, and conclusion.

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

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.

```
GFCLOCKS <- read.csv("~/603/Ass1/GFCLOCKS.csv")

Q2a = lm(PRICE~AGE+NUMBIDS, data = GFCLOCKS)
Q2b = sum(Q2a$residuals^2)

#TO find s we need to solve for sqrt(MSE), where MSE = SSE/df
Q2c = sqrt(Q2b/29)
Q2d = summary(Q2a)$adj.r.squared
```

```
Q2e = anova(Q2a)
Q2f = summary(Q2a)
Q2g = confint(Q2a, "AGE")
Q2h = lm(PRICE~AGE+NUMBIDS+AGE*NUMBIDS, data = GFCLOCKS)
```

Answer to Question 2a $\beta_0 = -1338.95, \beta_1 = 12.74, \beta_2 = 85.95$. The equation for this model would be:

$$y = -1338.95 + 12.74X_1 + 85.95X_2 + \epsilon$$

Answer to Question 2b 516726.5

Answer to Question 2c 133.4846678 This standard deviation means the actual values deivate from our prediction by about 133.4847 kilojoules per kilowatt-hour.

Answer to Question 2d Our $R^2_{Adj} = 0.8849$. This means our model is explaining the variance in prices well. This means there is a strong relationship between the predictors and the outcome variable, Price.

Answer to Question 2e

Source of Variation	df	Sum of Squares	Mean Square	F Value	Pr(>F)
AGE	1	2555224	2555224	143.406	9.527x10^-13
NUMBIDS	2	1727838	1727838		9.345x10^-11
Residuals	29	516727	17818		

Our null hypothesis is: H0 = None of the predictors have an impact on the outcome variable. This means the model does not explain any of the variance. Our alternate hypothesis is H1 = One or more of the predictors have an impact on the outcome variable. This means the model is able to explain some of the variance seen.

Age = p value is 9.53x10^-13 **NUMBIDS** = p value is 9.34x10^-11.

This means both Age and NUMBIDS have a significant impact on the price of the clock. The conclusion is that we are able to reject our null hypothesis in favour of the alternate hypothesis.

Answer to Question 2f The p value for NUMBIDS is 9.34x10^-11, which is below the significance threshold of 0.05. Therefore we can reject the null hypothesis and conclude that NUMBIDS has a significant positive impact on the price of the clock, while age is held constant.

Answer to Question 2g The lower limit for our confidence interval for Age is 10.89, and the upper limit is 14.59. This means we are 95% confident the coefficient for Age is between these values. In clearer terms, we can say that age has a positive correlation with price, increasing it between 10.89 and 14.59.

Answer to Question 2h I would use the interactive prediction model. This is because the adjusted R squared is higher compared to the additive model, (0.9489 versus 0.8849). The RSE is also lower for the interactive model (88.91 versus 133.5). The p value for the interaction term is below 0.05, meaning there is a significant interaction between Age and NUMBIDS affecting Price.

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.

[The first and last five observations for the turbine data 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 $\alpha = 0.01$ Be sure to state your hypotheses, p-values, and conclusion.

(c) Fit the best additive model to the data using the method of least squares. Test significance of predictors at $\alpha = 0.06$.Be sure to state your hypotheses, p-values, and conclusion.

(d) Test all possible interaction terms for the best model in part (c) at $\alpha = .06$. 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 RSE, 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.

```
TURBINE <- read.csv("~/603/Ass1/TURBINE.csv")
Q3a = lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP+CPRATIO+AIRFLOW, data=TURBINE)
#B is CPRATIO and AIRFLOW not sig
Q3c = lm(HEATRATE~RPM+INLET.TEMP+EXH.TEMP, data=TURBINE)

Q3d = lm(HEATRATE~(RPM+INLET.TEMP+EXH.TEMP)^2, data=TURBINE)#INTERACTIONS NOT SIG, PARENT NOT SIG
Q3d2 = lm(HEATRATE~(RPM+INLET.TEMP+EXH.TEMP+RPM*INLET.TEMP+RPM*EXH.TEMP), data=TURBINE)#INTERACTIONS NO
Q3d2.5 = lm(HEATRATE~(RPM+INLET.TEMP+EXH.TEMP+RPM*INLET.TEMP), data=TURBINE)#INTERACTION NOT SIG
Q3d2.6 = lm(HEATRATE~(RPM+INLET.TEMP+EXH.TEMP+RPM*EXH.TEMP), data=TURBINE)#INTERACTION NOT SIG
Q3d3 = lm(HEATRATE~(RPM+INLET.TEMP+EXH.TEMP+EXH.TEMP*INLET.TEMP), data=TURBINE) #NO GOOD
Q3d3.5 = lm(HEATRATE~(RPM+INLET.TEMP+EXH.TEMP+EXH.TEMP*INLET.TEMP+RPM*EXH.TEMP), data=TURBINE) #NO GOOD




newdata = data.frame(RPM=273145, INLET.TEMP=1240, EXH.TEMP=920)
predict(Q3c,newdata,interval="predict",level = .95)
```

```
##        fit      lwr      upr
## 1 43075.61 37718.03 48433.19
```

Answer to Q3a

$$y = 13,610 + 0.088X_1 - 9.2X_2 + 14.39X_3 + 0.35X_4 - 0.85X_5 + \epsilon$$

Where X1 is RPM, X2 is Inlet Temp, X3 is Exhaust Temp, X4 is CPRatio, and X5 is Airflow.

Answer to Q3b The null hypothesis is

$$H_0 : RPM/InletTemp/ExhuastTemp/CPRatio/Airflow = 0$$

. This means the predictors have no influence on the outcome variables. The alternate hypothesis is

$$H_0 : RPM/InletTemp/ExhuastTemp/CPRatio/Airflow \neq 0$$

. This means one or more predictor has an influence on the outcome, Heatrate. The p values of all variables except for CPRatio and Airflow are below the threshold of 0.01.

RPM = p value of 2.64x10ˆ-8

InletTemp= p value of 6.86x10ˆ-8

ExhaustTemp = p value of 0.000102

CPRatio = p value of 0.9905

Airflow= p value of 0.0598

These p values mean that we are able to say Airflow and CPRatio do not significantly contribute to the model, and should not be used as predictors.

This means that CPRatio and Airflow do not contribute to the model and do not serve predictive purposes. We are also able to reject our null hypothesis in favor of our alternate, since one or more predictors had a significant impact on the outcome, Heatrate.

Answer to Q3c Our null hypothesis is

$$H_0 : RPM/InletTemp/ExhuastTemp = 0$$

Our alternate hypothesis is

$$H_0 : RPM/InletTemp/ExhuastTemp \neq 0$$

. The equation of the best additive model is:

$$Heatrate = 14,360 + RPMX_1 + InletTempX_2 - ExhaustTempX_3 + \epsilon$$

The p values for the best additive model were:

RPM = 2.55x10ˆ-14

InletTemperature = 2x10ˆ-16

ExhaustTemperature = 1.06x10ˆ-7

This means that all of these terms have a significant impact on the Heatrate. We can therefore reject our null hypothesis in favor of our alternate hypothesis.

Answer to Q3d The final interactive model equation is:

$$Heatrate = \beta_0 + \beta_1 * RPM + \beta_2 * InletTemp + \beta_3 * Exhaust + \beta_1 * \beta_3 * RPM * ExhaustTemp$$

This model has an $R^2_{Adj} = 0.9145$, and an RSE of 468.4. This model has the best adjusted r squared and lowest RSE of the interaction models. With that being said, it is a worse predictor model than the additive model. This is because many of the predictors are no longer significant, and the interactions are not significant.

Therefore, the best predictive model is the additive model in Q3c.

Answer to Q3e The beta estimates show how much our outcome variable heatrate is expected to change for each increase in the respective predictor variable, while other variables are held constant.

Answer to Q3f The RSE of the model is 465kJ per kW per hour.

Answer to Q3g This model has an $R^2_{Adj} = 0.915$, meaning that the model is excellent at explaining the variance in responses about the mean.

Answer to Q3h The prediction did not include Airflow and pressure, as it was determined to not have a significant impact on Heatrate. The prediction is: 43075.61, with an upper limit of 48433.19 and a lower limit of 37718.03. We are 95% sure the value rests between these limits.