# MET CS 555 Assignment 5 – 20 points

SUBMISSION REQUIREMENTS: Please submit a single document (word or PDF) for submission. Your submission should contain a summary of your results (and answers to questions asked on the homework) as well as your R code used to generate your results (please append to the end of your submission). Please use R for the calculations whenever possible. You will lose points if you are not utilizing R. You will also lose 10 points per day for late submissions unless prior arrangements are made with your teaching team.

1. [Total 16 points: 14 points + 2 Extra Credit points] A producer of various feed additives for cattle conducts a study of the number of days of feedlot time required to bring beef cattle to market weight. Eighteen steers of essentially identical age and weight are purchased and brought to a feedlot. Each steer is fed a diet with a specific combination of protein content, antibiotic concentration, and percentage of feed supplement. The data are as follows:

| STEER | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PROTEIN | 10 | 10 | 10 | 10 | 10 | 10 | 15 | 15 | 15 | 15 | 15 | 15 | 20 | 20 | 20 | 20 | 20 | 20 |
| ANTIBIO | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 2 |
| SUPPLEM | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 | 3 | 5 | 7 |
| TIME | 88 | 82 | 81 | 82 | 83 | 75 | 80 | 80 | 75 | 77 | 76 | 72 | 79 | 74 | 75 | 74 | 70 | 69 |

(1) Obtain the regression equation relating feedlot time to the three diet variables. (2 points)
(2) Find the value of Residual Standard Deviation $S$. (1 point)
(3) Find the $R^2$ value. (1 point)
(4) How much of a collinearity problem is there with these data (using vif function in car package in R)? [Variance Inflation Factor (VIF) measures how much the standard error of a regression coefficient ($b_i$) is increased due to collinearity. If the value of VIF is very large, such as 10 or more, collinearity is a serious problem.] (2 Extra Credit points)
(5) Predict the feedlot time required for a steer fed 15% protein, 1.5% antibiotic concentration, and 5% supplement. (1 point)
(6) Do these values of the independent variables represent a major extrapolation from the data? (1 point)
(7) Give a 95% confidence interval for the mean time predicted in part (5). (3 points)
(8) Analyze the data using a regression model with only protein content as an independent variable.
(8a) Obtain the regression equation. (1 point)
(8b) Find the $R^2$ value. (1 point)
(8c) Test the null hypothesis that the coefficients of ANTIBIO and SUPPLEM are zero at $\alpha = .05$. (3 points)

**[Please refer to Supplemental Materials for further topics in Multiple Regression]**

2. [Total 8 points: 6 points + 2 Extra Credit points] Following the in-person lecture notes, load the data set http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data ,

```
wine <- read.table("http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data",sep=",")
wine_data <- wine[,2:14]
```

utilizing the following R packages:

```
install.packages("corrr")
install.packages("ggcorrplot")
install.packages("FactoMineR")
install.packages("factoextra")
install.packages("car")
library('corrr')
library("ggcorrplot")
library("FactoMineR")
library("factoextra")
library("car")
```

and perform PCA analysis to obtain:

(1) A Matrix Scatterplot of all 13 variables of chemical concentration. (1 point)
(2) Summary Statistics (mean, standard deviation) of all 13 variables of chemical concentration using sapply() function. (1 point)
(3) After standardizing the data, use the cor() function from the corrr package to calculate the correlation matrix, and uset he ggcorrplot() to generate a visualization. (1 point)
(4) Use princomp() to obtain PCA results, use summary table and scree plot to decide how many Principal Components to retain, and explain why. (2 points)
(5) Make a biplot combined with cos2 (attributes importance), and combined with (4) to discuss the contributions by important variables of chemical contributions in the new space spanned by the retained Principal Components. (1 point + 2 Extra Credit points)

---

# Supplemental Materials: **Extrapolation in Multiple Regression**

The notion of **extrapolation** is more subtle **in multiple regression** than in simple linear regression. In simple regression, extrapolation occurred when we tried to predict *y* using an *x*-value that was well beyond the range of the data. In multiple regression, we must be concerned not only about the range of each individual predictor but also about the set of values of several predictors together. It might well be reasonable to use multiple regression to predict the salary of a 30-year-old middle manager or the salary of a middle manager with 25 years of experience, but it would *not* be reasonable to use regression to predict the salary of a 30-year-old middle manager with 25 years of experience! Extrapolation depends not only on the range of each separate *xj* predictor used to develop the regression equation but also on the correlations among the $x_j$ values. In the salary prediction example, obviously age and experience will be positively correlated, so the combination of a low age and high amount of experience wouldn't occur in the data. When making forecasts using multiple regression, we must consider not only whether each independent variable value is reasonable by itself but also whether the chosen combination of predictor values is reasonable.

**Example:** A state fisheries commission wants to estimate the number of bass caught in a given lake during a season in order to restock the lake with the appropriate number of young fish. The commission could get a fairly accurate assessment of the seasonal catch by extensive "netting sweeps" of the lake before and after a season, but this technique is much too expensive to be done routinely. Therefore, the commission samples a number of lakes and records $y$, the seasonal catch (thousands of bass per square mile of lake area); $x_1$, the number of lakeshore residences per square mile of lake area; $x_2$, the size of the lake in square miles; $x_3$ = 1 if the lake has public access, 0 if not; and $x_4$, a structure index. (Structures are weed beds, sunken trees, drop-offs, and other living places for bass.) The data are shown in the table below.

| Lake | Catch | Residence | Size | Access | Structure |
|---|---|---|---|---|---|
| 1 | 3.6 | 92.2 | 0.21 | 0 | 81 |
| 2 | 0.8 | 86.7 | 0.3 | 0 | 26 |
| 3 | 2.5 | 80.2 | 0.31 | 0 | 52 |
| 4 | 2.9 | 87.2 | 0.4 | 0 | 64 |
| 5 | 1.4 | 64.9 | 0.44 | 0 | 40 |
| 6 | 0.9 | 90.1 | 0.56 | 0 | 22 |
| 7 | 3.2 | 60.7 | 0.78 | 0 | 80 |
| 8 | 2.7 | 50.9 | 1.21 | 0 | 60 |
| 9 | 2.2 | 86.1 | 0.34 | 1 | 30 |
| 10 | 5.9 | 90 | 0.4 | 1 | 90 |
| 11 | 3.3 | 80.4 | 0.52 | 1 | 74 |
| 12 | 2.9 | 75 | 0.66 | 1 | 50 |
| 13 | 3.6 | 70 | 0.78 | 1 | 61 |
| 14 | 2.4 | 64.6 | 0.91 | 1 | 40 |
| 15 | 0.9 | 50 | 1.1 | 1 | 22 |
| 16 | 2 | 50 | 1.24 | 1 | 50 |
| 17 | 1.9 | 51.2 | 1.47 | 1 | 37 |
| 18 | 3.1 | 40.1 | 2.21 | 1 | 61 |
| 19 | 2.6 | 45 | 2.46 | 1 | 39 |
| 20 | 3.4 | 50 | 2.8 | 1 | 53 |

The commission is convinced that residences and size are important variables in predicting catch because they both reflect how intensively the lake has been fished. However, the commission is uncertain whether access and structure are useful as additional predictor variables. Therefore, two regression models (with all four predictor variables entered linearly) are fitted to the data, the first model with all four variables and the second model without access and structure.

(A) Write the complete and reduced models.
(B) Write the null hypothesis for testing that the omitted variables have no (incremental) predictive value.
(C) Perform an F test for this null hypothesis.

**Answer:**

(A) The complete and reduced models are, respectively,

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

and

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The corresponding multiple regression least-squares equations based on the sample data are
Complete:

$$\hat{y} = -2.7840 + 0.0268 x_1 + 0.5035 x_2 + 0.7429 x_3 + 0.0511 x_4$$

Reduced:

$$\hat{y} = -0.8709 + 0.0394 x_1 + 0.8280 x_2$$

(B) The appropriate null hypothesis of no predictive power for $x_3$ and $x_4$ is $H_0$: $\beta_3 = \beta_4 = 0$.

(C) The test statistic for the $H_0$ of part (B) makes use of $SS(Regression, complete)$ = 0.585673 + 2.326834 + 1.921788 + 19.228068 = 24.0624, $SS(Regression, reduced)$ = 0.585673 + 2.326834 = 2.913, SS(Residual, complete) = 2.2756, df(Regression,complete) = 4, df(Regression, reduced) = 2, and n = 20:

$$F = \frac{(SS(Regression, complete) - SS(Regression, reduced))/(4-2)}{SS(Residual, complete)/(20-5)}$$
$$= \frac{(24.0624 - 2.913)\ /2}{2.2756/15} = 69.705$$

Therefore, $P < 0.0001$, we reject $H_0$. There is convincing evidence that the Access and Structure variables add predictive value.

The R scripts and results are the following:

```
bass_catch <- read.csv("Bass_catch.csv")
attach(bass_catch)
fm1 <- lm(Catch~Residence+Size+Access+Structure)
summary(fm1)

Call:
lm(formula = Catch ~ Residence + Size + Access + Structure)

Residuals:
    Min      1Q  Median      3Q     Max
-0.85859 -0.14400 -0.04054  0.21234  0.72653

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.784001   0.815706  -3.413  0.00385 **
Residence    0.026794   0.009141   2.931  0.01032 *
Size         0.503508   0.220767   2.281  0.03760 *
Access       0.742933   0.202128   3.676  0.00225 **
Structure    0.051129   0.004542  11.258 1.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3895 on 15 degrees of freedom
Multiple R-squared:  0.9136,      Adjusted R-squared:  0.8906
F-statistic: 39.65 on 4 and 15 DF,  p-value: 8.296e-08

fm2 <- lm(Catch~Residence+Size)
summary(fm2)

Call:
lm(formula = Catch ~ Residence + Size)
```

Residuals:
```
    Min      1Q  Median      3Q     Max
-2.24338 -0.49886 -0.02312  0.56536  2.89304
```

Coefficients:
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.87086    2.40884  -0.362    0.722
Residence    0.03941    0.02733   1.442    0.168
Size         0.82801    0.63720   1.299    0.211
```

Residual standard error: 1.174 on 17 degrees of freedom
Multiple R-squared:  0.1106,	Adjusted R-squared:  0.005945
F-statistic: 1.057 on 2 and 17 DF,  p-value: 0.3693

**aov(fm1)**

Call:
```
  aov(formula = fm1)
```

Terms:

| | Residence | Size | Access | Structure | **Residuals** |
|---|---|---|---|---|---|
| Sum of Squares | 0.585673 | 2.326834 | 1.921788 | 19.228068 | **2.275636** |
| Deg. of Freedom | 1 | 1 | 1 | 1 | **15** |

Residual standard error: 0.3894985
Estimated effects may be unbalanced

**aov(fm2)**

Call:
```
  aov(formula = fm2)
```

Terms:

| | Residence | Size | **Residuals** |
|---|---|---|---|
| Sum of Squares | 0.585673 | 2.326834 | **23.425492** |
| Deg. of Freedom | 1 | 1 | **17** |

Residual standard error: 1.17387
Estimated effects may be unbalanced

**anova(fm1,fm2)**

Analysis of Variance Table

Model 1: Catch ~ Residence + Size + Access + Structure
Model 2: Catch ~ Residence + Size

```
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     15  2.2756
2     17 23.4255 -2    -21.15 69.705 2.545e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The state fisheries commission hoped to use the Bass catch data to predict the catch at a lake with 8 residences per square mile, a size of .7 square mile, 1 public access, and a structure index of 55 and also for another lake with 48 residences per square mile, a size of 1.0 square mile, 1 public access, and a structure index of 40. The R scripts and results are the following:

```
bass_catch_newdata <- read.csv("Bass_catch_newdata.csv")
predict.lm(fm1, bass_catch_newdata, se.fit=TRUE, interval="prediction")
```

```
Obs  Residence Size Access Structure
1        8          0.7     1       55
2       48          1.0     1       40

$fit
        fit       lwr          upr
1 1.337851 -0.2081428 2.883844
2 1.793742  0.8476458 2.739839

$se.fit
       1       2
0.6118719 0.2128750

$df
[1] 15

$residual.scale
[1] 0.3894985
```

Locate the 95% prediction intervals for the two new lakes. Why is the first interval so much wider than the second?

The prediction intervals are given by the respective 95% PI values, (-0..2081, 2.8838) for the first lake and (0.8476, 2.7398) for the second lake. The first interval carries a warning: a point that is an extreme outlier in the predictors. A check of the data for the original 20 lakes reveals no lake had even close to eight residences per square mile. Thus, the prediction for this set of values of the predictors would be an extrapolation well beyond the data used to fit the model. For this case, the problem is with the value for just one of the explanatory variables, residence; the values for the remaining predictor variables are well within the range of the data.