# Exam 2

MATH 4387/5387: Applied Regression Analysis

Spring 2020

## Problem 1 (20 points)

In your own words precisely explain (also keep it *concise* ) followings:

a. Bias-variance trade off
b. Overfitting
c. Why we expect test error to be higher than training error
d. Cross validation
e. Permutation tests
f. Bootstrap confidence interval
g. Prediction vs confidence interval
h. Bonferroni correction
i. Outlier and influencial observation
j. In interpretation why we need to say "controlling for other variables"?

Though you are free to explore books, internet or any other sources, don't state definitions directly from any source other than yourself.

## Problem 2 (20 points)

Download `amazon-books.csv` from canvas. Data description:

The books data frame includes data harvested from Amazon.com related to 325 books. The data include:

- Amazon.Price - the amazon price,
- List.Price - list (publisher) price
- Hard_Paper - a categorical variable indicating whether the variable is a hardback (H) or paperback (P)
- NumPages - the number of pages
- Pub.year - the year the book was published
- Height - the height of the book (in)
- Width - the width of the book (in)
- Thick - the thickness of the book (in)
- Weight_oz - the weight of the book (oz).

Answer followings based on your analysis.

a. Run a multiple linear regression model with Amazon price as the response variable. As regressor use `List.Price`, `Hard_Paper`, `NumPages`, `Pub.year`, `Height`, `Width`, `Thick`, `Weight_oz`. Lets call the model `lmod`. Do you think all of the variables included in the full model (lmod) are relevant for predicting the Amazon price? Why?

b. Define `rmod` as `Amazon.Price ~ List.Price + Hard_Paper + NumPages + Pub.year`. Perform an $F$ test comparing `rmod` to `lmod`. Clearly state $H_0$, $H_a$, manually compute the test statistic (you can use RSS and df from the model summary), state the p-value, and interpret your results in the context of the problem.

c. Define `cmod` as `Amazon.Price ~ List.Price + NumPages + Pub.year`. Can a permutation test be used to compare rmod to cmod? Why? If so, a permutation test. Interpret your result from the test.

d. Find the best model for predicting Amazon price.

e. Peform an $F$ test comparing rmod and lmod.

f. Consider the model regressing `Amazon.Price` on `List.Price`, `NumPages`, and `Pub.year`. Is there enough evidence to conclude that the coefficient for `List.Price` is less than 1 and the `NumPages` coefficien is less than 0? Why?

g. Consider the `rmod` fitted model. What kind of causal conclusions can we make from our results?

h. Does the `rmod` fitted model who evidence of structural problems? Why? If so, discuss the problem.

## Problem 3 (20 points)

a. Download `simu1_exam2.txt` data from canvas. Run a linear regression model with `Y` as the response and all other variables as predictors.
b. Perform model selection.
c. Is there any evidence of structural problems? Why?
d. Answer this part if you answered *Yes* on (c). What remedial measure do you suggest. Perform model selection after you take step(s) to fix structural problems.

## Problem 4 (20 points)

(a) Fit a multiple linear regression model on the data you selected for the final project five predictors (If your data do not have five predictors use all that are available).
(b) Report $R^2$ and $F$-test from the model fit. What do you think about predictive ability of your model? Why?
(c) Take three new observations as
- $x_1$: mean values of the predictors
- $x_2$: medians of the predictors
- $x_3$ : 98th percentile of the predictors

(d) Calculate the prediction and confidence interval for $x_1$, $x_2$, $x_3$. Interpret the intervals according to the context of your data.
(e) Compare the following intervals. Discuss the differences and the reasons behind then.
- Compare prediction intervals at $x_1$ and $x_3$
- Compare prediction and confidence interval at $x_2$
- Compare confidence intervals at $x_1$ and $x_3$

## Problem 5 (20 points)

A large, national grocery retailer tracks productivity and costs of its facilities closely. Data below were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped ($X_1$), the indirect costs of the total labor hours as a percentage ($X_2$), a quantitative predictor called holiday ($X_3$) that is coded 1 if the week has a holiday and 0 otherwise, and the total labor hours ($Y$). Download the data from the canvas.

(a) Regress the total labor hours (response variable) on the other variables in the data. State the estimated regression function. How are the coefficients interpreted here?
(b) Check normality assumption.
(c) Test for outlier with $\alpha = 0.05$. For the hypothesis test state the decision rule.
(d) Check for leverage points?
(e) Management wishes to predict the total labor hours required to handle the next shipment containing 300,000 cases whose indirect costs of the total hours is $X_2 = 7.2$ and $X_3 = 0$ (no holiday in week).

Construct a scatter plot of $X_2$ against $X_1$ and determine visually whether this prediction involves an extrapolation beyond the range of the data.

(f) To spot hidden extrapolation in high dimensional predictor space, we can utilize the direct leverage calculation for new set of $X$ values for which inverences are to be made:

$$h_{\text{new, new}} = X_{\text{new}}^T (X^T X)^{-1} X_{\text{new}}$$

where $X_{\text{new}}$ is the vector containing the $X$ values for which an inference about a mean response or a new observation is to be made, and the $X$ matrix is the one based on the data set used for fitting the regression model. If $h_{\text{new, new}}$ is well within the range of leverage values $h_{ii}$ for the cases in the data set, no extrapolation is involved. On the other hand, if $h_{\text{new, new}}$ is much larger than the leverage values for the cases in the data set, an extrapolation is indicated.

Use this numerical technique to determine whether an extrapolation is involved. Do your calculations from the two methods agree?

(g) Identify influential observations if any.