# STOR 664: FALL 2022
# Final Exam, December 3–4, 2022

Time limit: 6 hours. The exam will be posted on gradescope and available from 6:00 am Saturday, December 3 through 7:00 pm Sunday, December 4, but once you download the exam, you have 6 hours in which to complete it and upload your solutions.

This is an open-book, take-home exam. Access to course materials and standard computational tools (in particular, R) is allowed; communication with other students or with anybody via the internet, other than the instructor, is not. The university Honor Code is in effect at all times. Answers may be given in any of the following formats (including combinations of more than one): R Markdown, Word, Latex or handwritten pages scanned or photographed; if handwritten, it is recommended you use blue or black ink on plain sheets of white paper. They should then be uploaded in gradescope. Each question will be graded out of 100 points and then averaged to give a total score of 100 for the whole exam. An error in one part-question will not prevent you gaining full credit in other parts of the same question even if you carry over the error.

It is strongly recommended that you do not spend more than 2 hours on any one question. If you find yourself getting near that limit, leave it and go to the next question. If you still have time left over, you can return to the question you left. Answers out of sequence will not be penalized so long as you make clear which part of your answer refers to which part of the question sheet.

The exam refers to two datasets as .csv files that are not preloaded in R. These will be placed on the course sakai page, in the "Final Exam" directory under "Resources". [Note subsequent to exam: they have since been placed on the web, see "http://rls.sites.oasis.unc.edu/Viscosity.csv" and "http://rls.sites.oasis.unc.edu/Quadratic.csv".]

1. The viscosity of an elastomer is affected by the level of oil and also by the level of any fillers. An experiment was conducted using three filler types, six levels of the filler used, and four levels of oil, with the following results (numbers in columns 3–8 are the measured viscosities):

| | | Filler Level | | | | | |
|---|---|---|---|---|---|---|---|
| Oil Level | Filler Type | 0 | 12 | 24 | 36 | 48 | 60 |
| 0 | N990 | 26 | 28 | 30 | 32 | 34 | 37 |
| 0 | SilA | 26 | 38 | 50 | 76 | 108 | 157 |
| 0 | SilB | 25 | 30 | 35 | 40 | 50 | 60 |
| 10 | N990 | 18 | 19 | 20 | 21 | 24 | 24 |
| 10 | SilA | 17 | 26 | 37 | 53 | 83 | 124 |
| 10 | SilB | 18 | 21 | 24 | 28 | 33 | 41 |
| 20 | N990 | 12 | 14 | 14 | 16 | 17 | 17 |
| 20 | SilA | 13 | 20 | 27 | 37 | 57 | 87 |
| 20 | SilB | 13 | 15 | 17 | 20 | 24 | 29 |
| 30 | N990 | — | 12 | 12 | 13 | 14 | 14 |
| 30 | SilA | — | 15 | 22 | 27 | 41 | 63 |
| 30 | SilB | 11 | 14 | 15 | 17 | 18 | 25 |

The data are presented in a form suitable for R analysis in the file Viscosity.csv (the two values with dashes are missing data and are represented NA in the csv file.)

The objective is to find a model for predicting viscosity as a function of oil level, filler type and filler level. As a guideline to what models to consider, the response function could be linear or quadratic in the two continuous variables (oil level and filler level) and could also include a cross-product. However, the model would obviously be easier to interpret if it was linear in the oil level and filler level. I don't recommend including any higher-order terms beyond quadratic. The coefficients could be independent of filler type or they could be completely dependent (i.e. a separate set of coefficients for each type). You should also consider whether it's appropriate to use any transformation of the response variable (viscosity).

(a) Propose a model for predicting viscosity as a function of filler type, filler level and oil level, explaining in detail the steps you take to select such a model, but also remembering that a simple model is easier to understand and to implement than a complicated one. Your final result should include a table of the selected model (including standard errors, t statistics, etc.) and supported by whatever graphics you consider appropriate.

(b) For the model selected in part (a), carry out appropriate diagnostic tests and comment on any features that may indicate the model is not a good fit. You should use plots and/or numerical calculations as appropriate.

(c) A new test is to be conducted with any of the filler types, filler level 40, and oil level 15. Find a 95% prediction interval for the resulting viscosity (three intervals, one for each type).

(d) How would you use your result to determine the combination of filler type, filler level and oil level that maximizes the viscosity?

2. An experiment with one $y$ and one $x$ variable produced the following data (see file Quadratic.csv):

| x | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| y | 3.12 | 4.80 | 6.34 | 5.48 | 4.74 | 4.26 |

It is believed that an appropriate model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \tag{1}$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent.

(a) Fit the linear model (1) in R and hence estimate the value $x^*$ which maximizes $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. Find an approximate 95% confidence interval for $x^*$ using the delta method.

(b) Repeat the calculations of (a) using Fieller's method — in other words, find an exact 95% confidence interval for $x^*$.

(c) Now you learn that the true model that generated the data was not (1), but

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i, \tag{2}$$

again with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent.

Suppose you want to predict a new response at the value $x = 2$. Show that,

i. If you use model (1), the prediction error variance will be $1.3714\sigma^2$.

ii. If you use model (2), the prediction error variance will be $1.4603\sigma^2$.

iii. However if you use model (1), the prediction will also have a bias of $-2.4\beta_3$.

iv. Hence, using mean squared error as a means of comparing predictions, you should still prefer to use model (1) if

$$\left|\frac{\beta_3}{\sigma}\right| < 0.124.$$

[**Note.** Part (c) is more of a theoretical exercise; it uses the $x$ values but not the $y$s. However, you will almost certainly need to use R or some other progamming language to derive the numerical answers. For this question, a clear explanation of the method is more important than an accurate numerical calculation.]

3. Consider the following dataset that you can download in R: `data(gasoline,package='pls')`. The dataset consists of 60 gasoline samples for which the octane level (response) is measured (`gasoline$octane`) as well as the near infrared spectrum at 401 frequences (`gasoline$NIR`). Remove every tenth observation from the data to use as a test sample. Use the remaining data as a training sample to build the following models:

   (a) Linear regression with all predictors;

   (b) Linear regression with variables selected by AIC;

   (c) Principal component regression;

   (d) PLS regression;

   (e) LASSO.

   Use the models you find to predict the response in the test sample. Make a report comparing the performance of the different models.