# Homework 5

36-401, Fall 2015

Due at the start of class on 22 October 2015

1. The file gpa.txt contains a bivariate dataset for a sample of 120 students at a small college. The variables are, in order,

   **GPA** = grade point average at end of first year

   **ACT** = score on the ACT standardized test for admissions

   Suppose all the assumptions of the Gaussian-noise linear regression model are satisfied.

   (a) (5) Estimate the regression function for predicting GPA from ACT. What is the predicted value of GPA when ACT=30?

   (b) (8) What is the log-likelihood of the data under the simple linear regression model? What is the log-likelihood under the intercept-only model? What is the log likelihood ratio? What is the $p$-value for testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ using the likelihood ratio test?

   (c) (5) Calculate the $t$ statistic for testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$. Clearly explain where you get each component of the test. What is the $p$-value?

   (d) (4) Give a 50% prediction interval for the GPA of students with the median ACT score. About how many students should be in this interval? How many are in this interval?

2. Refer back to the regression-through-the-origin model from previous homeworks,

   $$Y_i = \beta X_i + \epsilon_i,$$

   and suppose $n = 4$.

   (a) (3) Write the regression model using properly defined matrices $\mathbf{x}$, $\mathbf{y}$ and $\epsilon$.

   (b) (5) Derive the matrix version of the following estimating equation:

   $$\hat{\beta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

   Solve for $\hat{\beta}$.

(c) (4) Using matrix notation, show that $\hat{\beta}$ is an unbiased estimator.

(d) (6) Find the variance-covariance matrix of the vector of fitted values $\hat{m}$, in terms of the hat matrix $\mathbf{H} = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$.

3. (a) (5) Suppose there are two predictor variables, $X_1$ and $X_2$, so

$$\mathbf{x} = \left[ \begin{array}{ccc} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} \end{array} \right]$$

Express $n^{-1}\mathbf{x}^T\mathbf{x}$ in terms of $\overline{x_1}$, $\overline{x_2}$, $\overline{x_1^2}$, $\overline{x_2^2}$, and $\overline{x_1 x_2}$.

For the remaining parts of this problem, suppose the data looks as follows:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $X_1$ | 4 | 1 | 2 | 3 | 3 | 5 |
| $X_2$ | 4 | 3 | 10 | 9 | 5 | 8 |
| $Y$ | 16 | 5 | 10 | 15 | 13 | 22 |

(b) (4) Use matrix multiplication (in R) to calculate $\mathbf{x}^T\mathbf{x}$ and $(\mathbf{x}^T\mathbf{x})^{-1}$.

(c) (4) Check that the value you got for $\mathbf{x}^T\mathbf{x}$ matches what you would get using your answer from 3a.

(d) (4) What is $\hat{\beta}$? (Calculate this using matrices, not by calling `lm`.)

(e) (4) Using matrix multiplication, give estimates of $\mathrm{Var}[\hat{\beta}_0]$ and $\mathrm{Cov}[\hat{\beta}_1, \hat{\beta}_2]$. (*Hint:* You will have to estimate $\sigma^2$ first.)

(f) (4) Use `lm` to get values of $\hat{\beta}$ and $\mathrm{Var}[\hat{\beta}_0]$, and verify your answers from the previous parts.

(g) (5) Find the hat matrix $\mathbf{H}$ by matrix multiplication; include a printout (to a *reasonable* number of decimal places). Use $\mathbf{H}$ to find a fitted value for each data point. Compare your answer to using `fitted` and `lm`.

(h) (Extra credit, 1): Use `lm`, and at most one other function, to get $\mathrm{Cov}[\hat{\beta}_1, \hat{\beta}_2]$.

4. (5) Suppose that we have $p$ predictor variables in a multiple linear regression. Show that $\mathrm{tr}\,\mathbf{H}$, the trace of the hat matrix, is exactly $p + 1$.

*Hint:* Use the "cyclic rule" for traces: for any three matrices $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$, where $\mathbf{abc}$ is a square matrix, $\mathrm{tr}\,(\mathbf{abc}) = \mathrm{tr}\,(\mathbf{bca}) = \mathrm{tr}\,(\mathbf{cab})$.

5. A commercial real estate company wants to predict market rental rates in a particular metropolitan area, as a service to clients. It aims to do this by fitting a linear model, with rental rates as the response, and various features of rental properties as the predictors. The file `commercial.txt` contains information for 81 suburban commercial properties that are newest,

best located, most attractive, and expensive for five specific geographic areas. The response variable of interest is rental rate (`rent`), and the rest variables are predictors, including age (`age`), operating expenses and taxes (`expense`), vacancy rates (`vacancy`), and total square footage (`space`).

(a) (5) *Univariate EDA*: Use graphs and numerical summaries to describe the four predictor variables and the response variable. Describe the variables' distributions; do you have any possible outliers?

(b) (5) *Bivariate EDA*: Create the pairs plot with correlations. Interpret/describe the relationships you see between all the pairs of variables.

(c) (5) Estimate the regression coefficients $\hat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$. State the estimated regression function. Interpret $\hat{\beta}_0$ and $\hat{\beta}_{\text{age}}$.

Rubric: The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate R file. In the former case, please only submit the knitted file. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.