Problem Set 01
Linear Models -- Fall 2015
Due date: 2015-09-16

Problems sets are due at the start of class on the due date. Please hand write or type up and print the solutions; we will not accept e-mail solution sets except in exceptional circumstances. You may discuss problem sets with others, but must write up your own solutions. This means that you should have no need to look at other's final written solutions. Many of these problems come from a variety of textbooks, which are referenced in the problems. These are for citation purposes and not because you will need to consult the text itself (though you may feel free to do so).

1. [Sheather 2009, pg 39] A story by James R. Hagerty entitled *With Buyers Sidelined, Home Prices Slide* published in the Thursday October 25, 2007 edition of the *Wall Street Journal* contained data on so-called fundamental housing indicators in major real estate markets across the US. The author argues that... *prices are generally falling and overdue loan payments are piling up.* Thus, we shall consider data presented in the article on:

> Y = Percentage change in average price from July 2006 to July 2007 (based on the S&P/Case-Shiller national housing index); and
> x = Percentage of mortgage loans 30 days or more overdue in latest quarter (based on data from Equifax and Moody's).

The data are available at `euler.stat.yale.edu/~tba3/psets/pset01/data/indicators.csv`. Fit a simple linear regression model with an intercept to the data.

(a) Find a 95% confidence interval for the slope of the regression model, β. On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.

(b) Use the fitted regression model to estimate $\mathbb{E}(Y|X = 4)$. Find a 95% confidence interval for $\mathbb{E}(Y|X = 4)$. Is 0% a feasible value for $\mathbb{E}(Y|X = 4)$? Give a reason to support your answer.

2. The Cramér–Rao bound gives a lower bound on the variance of any unbiased estimator for a given deterministic parameter. Specifically, if $\widehat{\theta}$ is an unbiased estimator of θ, then the following inequality holds:

$$\mathbb{V}\widehat{\theta} \geq \frac{1}{I(\theta)} \tag{1}$$

Where $I(\theta)$ is the Fisher information. The Fisher information is defined in terms of the log-likelihood ℓ:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] \tag{2}$$

Calculate the Fisher information for β in the simple linear regression model with no intercept and normal, i.i.d. errors. Use this to establish that the MLE estimator for this model achieves the Cramér–Rao bound (i.e., it has the lowest allowable variance amongst the class of unbiased estimators).

3. For a given sample size n, consider observing y_i from a simple linear model without an intercept, with normal i.i.d. errors and $x_i = \frac{i}{n}$. For $\widehat{\beta}_{MLE}$ show that:

(a) Without calculating an analytic form of the variance, argue that $\widehat{\beta}_{MLE}$ is a consistent estimator of β.

(b) Find an analytic expression for the variance of $\widehat{\beta}_{MLE}$. Hint: $\sum_{i=1}^{k} i^2 = \frac{k^3}{3} + \frac{k^2}{2} + \frac{k}{6}$.

(c) Assume that σ is equal to 2 and known. Find the smallest n such that the z-test for the null hypothesis $H_0 : \beta = 0$ will yield a p-value less than 0.05 when the true β is greater than 1.

4. [Sheather 2009, pg 38] The web site www.playbill.com provides weekly reports on the box office ticket sales for plays on Broadway in New York. We shall consider the data for the week October 11–17, 2004 (referred to below as the current week). The data are in the form of the gross box office results for the current week and the gross box office results for the previous week (i.e., October 3–10, 2004). The data are available at `euler.stat.yale.edu/~tba3/psets/pset01/data/playbill.csv`

Fit the following model to the data: $Y = \alpha + \beta x + e$ where Y is the gross box office results for the current week (in $) and x is the gross box office results for the previous week (in $).

(a) Find a 95% confidence interval for the slope of the regression model, b1. Is 1 a plausible value for b1? Give a reason to support your answer.

(b) Test the null hypothesis $H_0 : \beta = 10000$ against a two-sided alternative. Interpret your result.

(c) Use the fitted regression model to estimate the gross box office results for the current week (in $) for a production with $400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office.

5. [Sheather 2009, pg 106] We consider here a real example involving the management at a Canadian port on the Great Lakes who wish to estimate the relationship between the volume of a ship's cargo and the time required to load and unload this cargo. It is envisaged that this relationship will be used for planning purposes as well as for making comparisons with the productivity of other ports. Records of the tonnage loaded and unloaded as well as the time spent in port by 31 liquid-carrying vessels that used the port over the most recent summer are available. The data are available at: `euler.stat.yale.edu/~tba3/psets/pset01/data/glakes.csv`.

We consider two models fit to the data. The first being:

$$\text{Time} = \beta_0 + \beta_1 \cdot \text{Tonnage} + e \tag{3}$$

And the second model:

$$\log(\text{Time}) = \beta_0 + \beta_1 \cdot \text{Tonnage}^{0.25} + e \tag{4}$$

Note that the second can be fit in **R** by using the lm function and the formula:

```
> lm(log(Time) ~ I(Tonnage^0.25), data=glakes)
```

(a) Does the first regression model seem to fit the data well? If not, list any weaknesses apparent in model.

(b) Suppose that model (3.8) was used to calculate a prediction interval for Time when Tonnage = 10,000. Would the interval be too short, too long or about right (i.e., valid)? Give a reason to support your answer.

(c) Is the second model an improvement over the first in terms of predicting Time? If so, please describe all the ways in which it is an improvement.

(d) List any weaknesses apparent in the second model.