STAT 306 Finding Relationships in Data

Lab 4 - Properties of Least Squares Estimator

David Lee

University of British Columbia
Department of Statistics

February 3/5, 2015

The regression model

- Recall our regression model for a single explanatory variable x_i (assumed deterministic) and response Y_i, $i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{1}$$

The parameter estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the solutions to the minimization

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

where y_i's are the realizations of the random variables Y_i's.

Properties of the estimators

- If we assume that $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for some constant σ^2, we obtain the following distributional results:
  - The regression parameter estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, i.e. $\mathbb{E}\left(\hat{\beta}_m\right) = \beta_m$, $m = 0, 1$.
  - The slope estimator $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$ where $\bar{x}$ is the sample mean of the x_i's.
  - When σ^2 is unknown and is estimated by $s^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-2)$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ are the fitted values, we have

$$\frac{\hat{\beta}_1 - \beta_1}{s \Big/ \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2},$$

    i.e. when scaled by its mean and standard error, $\hat{\beta}_1$ has a t distribution with $n-2$ degrees of freedom.

Properties of the estimators (cont.)

- The estimated variance (of the error term) S^2 is unbiased, i.e. $\mathbb{E}(S^2) = \sigma^2$. Here S^2 is the random variable from which s^2 is realized.
- When scaled appropriately, the estimated variance has a chi-squared distribution with $n-2$ degrees of freedom:

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}.$$

- The regression parameters are uncorrelated with the estimated variance:

$$\text{Cor}(\hat{\beta}_0, S^2) = \text{Cor}(\hat{\beta}_1, S^2) = 0.$$

  They are, in fact, independent. This fact is used in constructing the t-test for the parameters. (Recall how a normal r.v. and a chi-squared r.v., independent of each other, can be used to form a t-distributed r.v.)

- Some of the results above do not require the normality assumption of the ϵ_i's. Can you point out which?

Model simulation

- The R file for this lab involves a simulation. Simulation studies are often used in statistics to explore the properties of estimators (assuming the model is true) which are otherwise hard to obtain analytically.
- The purpose of this simulation is to demonstrate the above properties of a simple linear regression. The steps involved are:
  - Simulate from a model with known parameters (i.e. the regression model (1)). We generate the error terms and these are used to perturb the observations around the true regression line.
  - This simulated data set is fitted by the regression model. We obtain one set of estimated parameters and estimated variance of the error, and thus one value for the t statistic (say, for the test of $\beta_1 = 0$).
  - The above two steps are repeated M times (`nsim` in the code).

Model simulation (cont.)

- At the end, we obtain M sets of parameter estimates, and M values for the test statistic. These can be plotted using histograms, illustrating their sampling distributions. Such distribution will resemble the true distribution when $M \to \infty$.
- Of course, in this case the true distributions are known; the simulation confirms that they are valid and gives you an idea of how the estimators should perform.

Lab question

- See WeBWorK for this week's lab question. Question 1 requires you to write some elaborate R code to obtain the result. The easiest way to do this is to write something that resembles the following:

```
[Initialize a 'total' variable]
for (i in 1:upperbound){
    if ([Divisibility condition is met]){
     [Increment the 'total' variable
    by the square root of i]
    }
}
```

Phrases in blue are pseudo-codes; you need to write it in the language recognizable by R. Refer to the lab R file (or use search engine!) to see how loops and conditionals should be written in R.

Lab question (cont.)

- Here are some hints:
  - There are a few lines of code. I suggest you write your code in a separate text editor, and then paste them into R.
  - The modulo operator in R is %%.
  - The logical OR is given by the | or || operators. There is a difference between the two, although it should not matter for the purpose of this exercise. Type `help("|")` for the relevant help document.
- The other questions are related to the simulation we've done today. Submit your response by Friday 10 pm.