BST 140.752
Problem Set 1
Due: November 7, 2017

1 Inference and estimation in linear models

1. Consider the random variables $Y_1, \ldots, Y_n$ defined as $Y_i = U + Z_i$, where $U \sim N(\xi, \tau^2)$, Z_i are i.i.d $N(\mu, \sigma^2)$, and U and Z_i are independent. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

   (a) What is the distribution Y_i?

   (b) Find $cov(Y_i, Y_j)$ for $i \neq j$.

   (c) What is the distribution of $\mathbf{Y}$?

   (d) Consider the estimator $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Is $\bar{Y}$ an unbiased estimator for $E[Y_1]$?

   (e) Consider the estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$. Show that $(n-1)S^2 = \mathbf{Y}'\mathbf{P}\mathbf{Y}$ for some projection matrix $\mathbf{P}$.

   (f) What is the distribution of S^2?

   (g) Is S^2 an unbiased estimator for $var(Y_1)$?

   (h) Let $\mathbf{V} = var(\mathbf{Y})$. Find the inverse $\mathbf{V}^{-1}$ and the determinant $det(\mathbf{V})$.

2. Let $Y_{ij} = \mu_i + \epsilon_{ij}$ for $i = 1, \ldots I$ and $j = 1, \ldots, J_i$ where the $\epsilon_{ij} \sim N(0, \sigma^2)$ are iid.

   (a) For $I = 2$ show that the unbiased estimate of σ^2 is the so-called pooled variance estimate, $S_p^2 = \frac{1}{J_1+J_2-2}\{(J_1-1)S_1^2 + (J_2-1)S_2^2\}$ where S_i^2 is the standard variance estimate within group i. Derive a T confidence interval for $\mu_1 - \mu_2$ and test of $\mu_1 = \mu_2$.

   (b) For a general value of I derive an overall F test for the hypothesis that $H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$ versus the alternative that at least two are unequal. Argue that this F-test compares the variation between the groups to that within the groups.

3. Consider the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Argue that the variance of $\hat{\beta}_1$ is minimized with the variance in the observed X_i is maximized. For what pattern in the X_i is the lowest variance estimate is obtained?

2 Coding and data analysis exercises

1. Randomly generate $1,000 \times 20$ normals with mean 5 and variance 2. Place these results in a matrix with $1,000$ rows. Using two `apply` statements on the matrix, create two vectors, one of the sample mean from each row and one of the sample standard deviation from each row. From these $1,000$ means and standard deviations, create $1,000$ t statistics. Now use R's `rt` function to directly generate $1,000$ t random variables with 19 df. Use R's `qqplot` function to plot the quantiles of the constructed t random variables versus R's t random variables. Do the quantiles agree? Describe why they should.

2. Simulate $1,000$ sample variances of 20 observations from a normal distribution with mean 5 and variance 2. Convert these to sample variances which should be chi-squared random variables with 19 degrees of freedom. Now simulate $1,000$ random chi-squared variables with 19 degrees of freedom using R's `rchisq` function. Use R's `qqplot` function to plot the quantiles of the constructed chi-squared random variables versus those of R's random chi-squared variables. Do the quantiles agree? Describe why they should.

3. Extend the `R` function `mylm()` you created in a previous homework to return a list with a T table (estimate, standard error, t statistics, P-value), as well as the results of a test of overall regression. Find a dataset to try out your function (you can simulate one if you like), and compare the results to the one from the `lm()` function.