EEB 581, Problem Set One: Solutions

1 : Data was measured on 50 individuals for arm size (x) and brain size (y), with the following results:

$$\overline{x} = 10, \quad \overline{y} = 50, \quad \sum_{i=1}^{50} (x_i - \overline{x})^2 = 100, \quad \sum_{i=1}^{50} (y_i - \overline{y})^2 = 400, \quad \sum_{i=1}^{50} (x_i - \overline{x})(y_i - \overline{y}) = 175$$

(a) Compute the variances of x and y, their covariance, and their correlation.

$$\textbf{Var}(x) = \frac{100}{49} = 2.04, \quad \textbf{Var}(y) = \frac{400}{49} = 8.16, \quad \textbf{Cov}(x,y) = \frac{175}{49} = 3.57$$

$$\textbf{Corr}(x,y) = \frac{3.57}{\sqrt{2.04} \cdot \sqrt{8.16}} = 0.88$$

(b) What the best linear regression of arm size (x) on brain size (y)?

$$b_{x\,|\,y} = \frac{3.57}{8.16} = 0.44, \qquad a = \overline{x} - b_{x\,|\,y}\,\overline{y} = 10 - 0.44 \cdot 50 = -11.88$$

Hence, the regression is (Arm size) = -11.88+ 0.44(Brain size)

(c) What the best linear regression of brain size (y) on arm size (x)?

$$b_{y\,|\,x} = \frac{3.57}{2.04} = 1.75, \qquad a = \overline{y} - b_{y\,|\,x}\,\overline{x} = 50 - 1.75 \cdot 10 = 32.58$$

Hence, the regression is (Brain size) = 32.50 + 1.75(Arm size)

(d) What fraction of the total variance in brain size does the regression account for?

Fraction of the total variance explained by the regression is the squared correlation, or $0.88^2 = 0.766$

2 : Use the properties of covariances to show that $E[\,(x - \mu_x)^2\,] = E[\,x^2\,] - \mu_x^2$.

$$E[\,(x-\mu_x)^2\,] = E[\,x^2 - 2\mu_x x + \mu_x^2\,] = E[\,x^2\,] - 2\mu_x E[\,x\,] + \mu_x^2 = E[\,x^2\,] - 2\mu_x^2 + \mu_x^2 = E[\,x^2\,] - \mu_x^2$$

3 : What is the covariance between a particular data point (x_i) and the sample mean $\overline{x}$?

$$\sigma\left(x_i, \frac{1}{n}\sum_{j=1}^{n} x_j\right) = \frac{1}{n}\sum_{j=1}^{n} \sigma\,(x_i, x_j) = \frac{\sigma^2(x_i)}{n} + \frac{1}{n}\sum_{j \neq i}^{n} \sigma\,(x_i, x_j)$$

4: Assuming the appropriate normality assumptions, compute the 95% confidence intervals for σ_x^2 and σ_y^2 using the data in (1). (Hint: Use R to obtain the appropriate χ^2 values).

Since $\sum^n (x_i - \overline{x})^2 \simeq \sigma_x^2 \chi_{n-1}^2$**, it follows that**

$$\sum^n (x_i - \overline{x})^2 / \sigma_x^2 \sim \chi_{n-1}^2$$

Define $\chi_n^2(\alpha/2)$ **as satisfying**

$$\Pr(\chi_n^2 < \chi_n^2(\alpha/2)) = \alpha/2 \qquad \textbf{so that} \qquad \Pr(\chi_n^2 > \chi_n^2(1-\alpha/2)) = \alpha/2$$

Thus the upper cutoff for σ_x^2 in an $(1-\alpha)$ confidence interval of σ_x^2 satisfies

$$\Pr\left(\frac{\sum^n (x_i - \overline{x}\,)^2}{\sigma_x^2} \leq \chi_n^2(\alpha/2)\right) \quad \textbf{or} \quad \sigma_x^2 \leq \frac{\sum^n (x_i - \overline{x}\,)^2}{\chi_n^2(\alpha/2)}$$

The lower cuttoff follows similiarly, giving the $(1-\alpha)$ confidence interval of σ_x^2 as

$$\frac{\sum^n (x_i - \overline{x}\,)^2}{\chi_{n-1}^2(1-\alpha/2)} \leq \sigma_x^2 \leq \frac{\sum^n (x_i - \overline{x}\,)^2}{\chi_{n-1}^2(\alpha/2)}$$

For a 95% confidence interval, $\alpha = 0.05$. Hence we seek the value X such that $\Pr(\chi_{49}^2 \leq X) = 0.0975$ From R, `qchisq(0.975,49 )` returns `70.22` and hence $\chi_{49}^2(0.975) = 70.22$.

Likewise `qchisq(0.025,49)` returns `= 31.56` and hence $\chi_{49}^2(0.025) = 31.56$. Thus the 95% confidence interval for σ_x^2 is $1.42 \leq \sigma_x^2 \leq 3.17$ (note this confidence interval is not symmetric around the sample variance). Likewise, the interval for σ_y^2 is $5.68 \leq \sigma_y^2 \leq 12.68$.