EMTH211 2014 Exam Solutions for Statistics

6 (a) The correlation between two data vectors $\mathbf{x}$, $\mathbf{y}$ is defined as

$$cor(\mathbf{x}, \mathbf{y}) = \frac{\tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}}}{\|\tilde{\mathbf{x}}\| \|\tilde{\mathbf{y}}\|}$$

(b) The correlation between the number of sheets of fly paper and number of flies caught is positive. If the number of sheets of fly paper increases the number of flies caught increases as well. The values of the observations tend to go into the same direction.

(c) The correlation between F and r is negative. If we increase the distance r, the gravitational force decreases. If the distance is decreased, F will increase. The values of the observations tend to go into opposite directions.

(d) The correlation would still be -0.21 since the correlation is scale invariant

$$cor(100\mathbf{x}, \mathbf{y}) = sign(100) cor(\mathbf{x}, \mathbf{y}) = cor(\mathbf{x}, \mathbf{y}).$$

(e) In this experiment the outcome of the second draw does not depend on the outcome of the first draw because all cards are back in the hat. Hence, there is no linear relation between the two experiments and the correlation will tend to be 0.

(e) In this experiment the outcome of the second draw does depend on the outcome of the first draw because the first card can not be drawn again. If the first card had a small number, e.g. between 1 and 5, the chance to get a card with a large value, e.g. between 6 and 13 are a bit higher. If the first card had a larger value, e.g. between 7 and 13, the chance to get a small number in the second draw becomes larger. This means that the observations tend to go into opposite directions. The correlation will be negative.

7 (a) (i) We have the explicit formula for $\widehat{b}_1$ in the simple regression model

$$\widehat{b}_1 = \frac{\tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}}}{\|\tilde{\mathbf{x}}\|^2}.$$

$\overline{\mathbf{x}} = 2.5$ and $\overline{\mathbf{y}} = 3.75$
$\tilde{\mathbf{x}} = (-1.5, -0.5, 0.5, 1.5)$ and $\tilde{\mathbf{y}} = (-2.75, -1.75, 0.25, 4.25)$.
$\tilde{\mathbf{x}} \cdot \tilde{\mathbf{y}} = 4.125 + 0.875 + 0.125 + 6.375 = 11.5$
$\|\tilde{\mathbf{x}}\|^2 = 2.25 + 0.25 + 0.25 + 2.25 = 5$
Hence, $\widehat{b}_1 = 11.5/5 = 2.3$.

(ii) The intercept $\widehat{b}_0$ is given by the formula

$$\widehat{b}_0 = \overline{\mathbf{y}} - \widehat{b}_1 \overline{\mathbf{x}}.$$

Hence, the intercept is $\widehat{b}_0 = 3.75 - 2.3 \times 2.5 = -2$.

(iii) The $100(1-\alpha)\%$ confidence interval is given by $b_1 = \widehat{b}_1 \pm t_{n-2}(1-\alpha/2)se(\widehat{b}_1)$ with

$$se(\widehat{b}_1) = \sqrt{\frac{1}{n-2}\frac{SSR}{\|\tilde{\mathbf{x}}\|^2}}.$$

$$\begin{aligned} SSR &= (1 - 2.3 + 2)^2 + (2 - 2.3 \times 2 + 2)^2 + (4 - 2.3 \times 3 + 2)^2 \\ &\quad + (8 - 2.3 \times 4 + 2)^2 \\ &= 0.7^2 + 0.6^2 + 0.9^2 + 0.8^2 \\ &= 2.3 \end{aligned}$$

$$se(\widehat{b}_1) = \sqrt{\frac{1}{n-2}\frac{SSR}{\|\tilde{\mathbf{x}}\|^2}} = \sqrt{\frac{1}{2}\frac{2.3}{5}} \approx 0.48$$

Hence, $b_1 = \widehat{b}_1 \pm t_2(0.975)se(\widehat{b}_1) \approx 2.3 \pm 4.3 \times 0.48 = 2.3 \pm 2.06$
Or equivalently, $b_1 \in [0.24, 4.36]$.

(b) (i) The predicted value is $\widehat{y} = 2000 + 500 \times 2 + 500 \times 2 + 2 = 4002$

(ii) We can expect a strong correlation between length, height and width of an elephant. This will most likely lead to multicollinearity in this regression which can cause large estimation errors in $\widehat{b}_1$, $\widehat{b}_2$, and $\widehat{b}_3$.

(iii) Qualitative information can be incorporated in the regression as a dummy variable. We can, for example, assign the value $X_4 = 0$ to male and $X_4 = 1$ to female elephants.