STAC67: Regression Analysis

Assignment 1
(Total: 90 points)

Please submit R Markdown file for Q. 1 and Q. 6 along with your submission of the assignment.

Q. 1 (10 pts) This question is to practice R to sample from a Normal distribution. Obtain random samples from a Normal with mean $\mu = 100$, $\sigma = 20$ of size n = 100, 1000, 10,000, 100,000.

When you generate a random number, use R code, **set.seed(your student number)** before the R codes of generating a random number, so that we can replicate the result.

(a) (5 pts) On a single page (2 rows, 2 columns) give the histograms on the same set of bins, with a normal density superimposed on each. Comment on the approximation accuracy.

(b) (5 pts) For each sample size, give the mean, standard deviation, and the following percentiles (2.5, 25, 50, 75, 97.5). Compare these with the theoretical values.

Q. 2 (14 pts) (a) (4 pts) Prove the following equalities.

(i) $S_{XX} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$

(ii) $S_{XY} = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$

(b) Suppose that $(Y_1, X_1), \ldots, (Y_n, X_n)$ is a data set to which we fit a simple linear regression. Let $\hat{\beta}_1$ be the least squares estimate of the slope with Y and let r be the sample correlation coefficient.

(i) (5 pts) Show that

$$\hat{\beta}_1 = r\frac{s_Y}{s_X}$$

where s_Y and s_X are the sample standard deviations of $Y_1 \ldots Y_n$ and $X_1 \ldots X_n$ respectively.

(ii) (5 pts) Show that

$$\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Q. 3 (16 pts) Anastrozole is a drug often used to treat breast cancer patients. One study attempted to see if the effect of Anastrozole is associated with the age of patients. The response variable Y is the change the levels of cortisol-binding globulin (CBG). and the covariate x is age. The following summary statistics were reported.

$$n = 26 \qquad \sum X_i = 1613 \qquad \sum Y_i = 281.9$$
$$S_{XX} = 3756.96 \qquad S_{YY} = 465.34 \qquad S_{XY} = -757.64$$

(a) Find the least squares estimates of the intercept and slope.

(b) Give the standard errors for your estimates in (a).

(c) Construct 95% confidence intervals for the true intercept and true slope.

(d) What conclusions would you draw from your results?

Q. 4 (24 pts) We fit the linear regression model without the intercept, $Y_i = \beta_1 X_i + \epsilon_i, i = 1, \ldots n$,

(a) (5 pts) Find the least square estimator of β_1.

(b) (5 pts) Denote the estimator by $\hat{\beta}_1$ then the estimated model is $\hat{Y}_i = \hat{\beta}_1 X_i$. Let $e_i = Y_i - \hat{Y}_i$. Can you conclude $\sum_{i=1}^{n} e_i = 0$?

(c) (4 pts) Assume that the error term are independent and identically distributed, $N(0, \sigma^2)$ with σ^2 unknown. Find the Standard Error for the estimator of β_1.

(d) (5 pts) Design a procedure to test

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

(e) (5 pts) The data is collected for six observations.

i :	1	2	3	4	5	6
X_i :	7	12	4	14	25	30
Y_i :	128	213	75	250	446	540

Find the maximum likelihood estimator of β_1 and evaluate its value.

Q. 5 (10 pts) Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ for } \mathrm{i} = 1, 2, \cdots, \mathrm{n},$$

where $\epsilon_i \sim N(0, \sigma^2)$ and

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

is the least squares estimator of β_1. Imagine $\hat{\beta}_1^* = \sum_{i=1}^{n} c_i Y_i$ is any other unbiased estimator of β_1 with c_i being arbitrary constant. Prove that $Var(\hat{\beta}_1) \leq Var(\hat{\beta}_1^*)$. That is, prove that the least squares estimator of β_1 has the minimum variance among all other linear unbiased estimators of β_1.

Q. 6 (16 pts) (4 pts each) For this question, use R Markdown file. The data set, "MiceWeightGain.xls" is posted at Quercus. The data consists as different levels of a nutrient (X), and the weight gain of mice (Y) which is the response variable. There are 30 sampled mice of a particular breed and assigns them randomly to one of 6 levels of the nutrient (0, 20, 40, 60, 80, 100).

(a) Obtain a scatter plot of weight change versus nutrient level

(b) Fit a simple linear regression, relating weight change to nutrient level

(c) Test whether there is a positive association between weight change and nutrient level

(d) Give a 95% confidence interval for the mean change in weight as nutrient level is increased by 1 unit.