Tutorial Problems

Question 1

Suppose the linear regression model is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1 \ldots, n \geq 2$. Assume all these assumptions are satisfied. In this question, we will fill out some mathematical details and derive some useful properties for the least square estimators.

(a) In the lecture, we derive the normal equations for the OLS estimate to be

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i = 0, \quad \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0.$$

Solve this system of equation in detail to obtain the OLS estimate as written in the lecture. Some equalities that may be useful include $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y}) = 0$.

(b) Prove that $R^2 = r_{xy}^2$, where $R^2 = SSR/SST$ is the coefficient of determination, and r_{xy} is the sample correlation.

(c) Prove

$$E(SSR) = \sigma^2 + \beta_1^2 S_{xx}.$$

Question 2

Show that the F-test statistic for testing $H_0 : \beta_1 = 0$,

$$F = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\sigma}^2},$$

can be written as

$$F = \frac{r^2(n-2)}{1-r^2},$$

where r is the sample coefficient of correlation between x and y. (Assumed knowledge)

Computer Problems

For the following questions, use the `olympic.txt` dataset that consists of the winning heights or distances (in inches) for the High Jump, Discus and Long Jump events at the Olympics up to 1996.

Question 1

(a) Store the `olympic.txt` dataset in R as the data frame `olympic`. **Hint**: the dataset is tab delimited (`sep="\t"`).

(b) Describe, and where possible explain, any unusual features about `olympic`.

(c) Create a new data frame `olympicMetric` that has measurements in metres, by using the conversion 1 m = 39.3701 inches, and the full year (e.g. 1900 rather than 0). Show the first 6 rows of the `olympicMetric`. **Hint**: The Olympics were held every 4 years except for 1916, 1940, and 1944 due to war.

You should use `olympicMetric` for the next question.

Copyright © The University of Sydney

Question 2

(a) Plot the first 20 values of LongJump (x_i) against the first 20 values of HighJump (y_i). Briefly comment on the pattern.

(b) Fit the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for $i = 1, \ldots, 20$ using

```
olympicLm <- lm(HighJump ~ LongJump, data = olympicMetric)
```

(c) Find the least square estimates for the parameters (β_0, β_1, σ^2) using a `summary` output of `olympicLm`.

(d) Construct 95% confidence intervals for β_1. Manually compute it using the `summary` output and verify the results with

```
confint(olympiclm)
```

(e) Use the `anova` function to produce the ANOVA table for the fitted linear model. Verify the relationship between the F-test and the t-test for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

(f) Formally test the hypotheses $H_0 : \beta_1 = 0.25$ versus $H_1 : \beta_1 > 0.25$. Include the value of the test statistic, p-value and the conclusion.

Copyright © The University of Sydney