Homework to Week 3

Statistics: Principle, Methods and R (II)

Gao Fengnan

Week 4, 20 March 2017

The homework is due on Monday, 27 March 2017. Please hand in the solutions to the teaching assistant He Siyuan at the beginning of the lecture.

1. Prove the following statement. Let $Y \sim N_n(\mu, \Sigma)$ with Σ positive definite, C be an $m \times n$ matrix of rank m, and d be an $m \times 1$ vector, then $CY + d \sim N_m(C\mu + d, C\Sigma C^T)$. Hint: The moment generating function of $Y \sim N(\mu, \Sigma)$ is

$$\mathbb{E}[\exp(s^T Y)] = \exp(s^T \mu + s^T \Sigma s/2).$$

   Calculate the moment generating function of $Z = CY + d$

$$\mathbb{E}[\exp(u^T Z)] = \mathbb{E}[\exp(u^T CY + u^T d)],$$

   and try to rewrite it into the following form

$$\mathbb{E}[\exp(u^T Z)] = \exp(u^T (C\mu + d) + u^T C\Sigma C^T u/2).$$

2. Find the proper rejection threshold c_α such that under the null hypothesis $H_0 : \beta_j = 0$,

$$\mathbb{P}(|z_j| > c_\alpha | H_0) = \alpha,$$

   where the z-score is defined as

$$z_j = \hat{\beta}_j / (\hat{\sigma} \sqrt{v_j}). \tag{1}$$

   The above statement is the same as that the probability for the test to make the type-I error is α. For simplicity, find only c_α for $\alpha = 0.05$ and $n - p - 1 = 10$. You may need the fact that under the null hypothesis $\beta_j = 0$, $z_j \sim t_{n-p-1}$.

3. Find the proper rejection threshold c_α such that under the null hypothesis $H_0 : \beta_{p_0+1} = \beta_{p_0+2} = \cdots = \beta_{p_1} = 0$

$$\mathbb{P}(F > c_\alpha | H_0) = \alpha,$$

   where the test statistics F is defined in lecture as

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(n - p_1 - 1)}. \tag{2}$$

You may use the fact that under the null hypothesis (i.e., the smaller model is correct), the F statistic has a $F_{p_1-p_0,n-p_1-1}$ distribution. For simplicity, find only c_α for $\alpha = 0.05$ with $p_1 = 5$ and $p_0 = 2$ and $n = 20$.

4. Show the F statistic defined above in (2) for dropping a single coefficient from a model is equal to the square of the corresponding z-score defined above in (1).

5. In this exercise you will create some simulated data with `R` and will fit simple linear regression models to it. Make sure to use `set.seed(100)` prior to starting part 5a) to ensure consistent results. **Please attach your code in your solution.**

   a) Using the `rnorm()` function, create a vector `x`, containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, X.

   b) Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a $N(0, 0.25)$ distribution.

   c) Using `x` and `eps`, generate a vector `y` according to the model

   $$Y = -1 + 0.5X + \varepsilon. \tag{3}$$

   What is the length of the vector `y`? What are the values of β_0 and β_1 in this *simple linear model*?

   d) Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.

   e) Fit a least squares linear model to predict `y` using `x`. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

   f) Display the least square line on the scatterplot obtained in 5d). Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.