Statistics: Homework 3

Due on Aug 11, 2014

Instructor: Rados Radoicic 6:00 pm

Weiyi Chen

Problem 1

Problem T1: Prediction with confidence

1.

Mean response is an estimate of the mean of the y population associated with x, that is $E(y|x) = \hat{y}$. The variance of the mean response is given by

$$\text{Var}\left(\hat{\alpha} + \hat{\beta}x\right) = \text{Var}\left(\hat{\alpha}\right) + \left(\text{Var}\hat{\beta}\right)x^2 + 2x\text{Cov}\left(\hat{\alpha}, \hat{\beta}\right). \tag{1}$$

This expression can be simplified to

$$\text{Var}\left(\hat{\alpha} + \hat{\beta}x\right) = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right). \tag{2}$$

The $100(1-\alpha)\%$ confidence intervals are computed as $y \pm t_{\frac{\alpha}{2},n-2}\sqrt{\text{Var}}$, since σ is unknown, we estimate it by $s = \sum(x_i - \overline{x})^2/(n-1)$, therefore the confidence interval is

$$\beta' x \pm t_{\frac{\alpha}{2},n-2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)} \tag{3}$$

2.

The predicted response distribution is the predicted distribution of the residuals at the given point x. So the variance is given by

$$\text{Var}\left(y - \left[\hat{\alpha} + \hat{\beta}x\right]\right) = \text{Var}\left(y\right) + \text{Var}\left(\hat{\alpha} + \hat{\beta}x\right). \tag{4}$$

The second part of this expression was already calculated for the mean response. Since $\text{Var}\left(y\right) = \sigma^2$ (a fixed but unknown parameter that can be estimated), the variance of the predicted response is given by

$$\text{Var}\left(y - \left[\hat{\alpha} + \hat{\beta}x\right]\right) = \sigma^2 + \sigma^2\left(\frac{1}{m} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) = \sigma^2\left(1 + \frac{1}{m} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right). \tag{5}$$

In the same way I did in part 1, the $100(1-\alpha)\%$ confidence interval is

$$\beta' x \pm t_{\frac{\alpha}{2},n-2}\sqrt{s^2\left(1 + \frac{1}{m} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)} \tag{6}$$

Problem 2

Problem T2: Linear hypothesis

Answer

According to the null hypothesis, which can be written as $H_0 : R\beta = r$, where

$$R = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \end{pmatrix} \text{ and } r = \begin{pmatrix} 4 \\ 0.5 \end{pmatrix} \tag{7}$$

Calculate the F-ratio as:

$$F = \frac{(Rb - r)'(R(X'X)^{-1}R')^{-1}(Rb - r)/m}{s^2} = 0.0614 \tag{8}$$

where $m = rank(R) = 2$.
The critical value is:

$$F_\alpha(m, n-k) = F_{0.05}(2, 20-3) = 3.591 > 0.0614 \tag{9}$$

Therefore accept H_0.

Problem 3

Problem T6: Hide and Seek

Answer

For the column df, just remember the rule "minus one":

$$\text{We have 3 different factors} \Rightarrow df_{Regression} = 2 \tag{10}$$

$$df_{Regression} + df_{error} = df_{total} \Rightarrow df_{error} = 14 - 2 = 12 \tag{11}$$

For the column MS just remember the rule $MS = SS/df$, then:

$$MS_{error} = 3.250/12 = 0.2708 \tag{12}$$

Given the p-value as 0.05 and df above, the F-statistics is

$$F_{2,12}(0.05) = 3.8853 \tag{13}$$

The F-value is also given by:

$$F = MS_{Regression}/MS_{error} \Rightarrow MS_{Regression} = F \times MS_{error} = 1.0523 \tag{14}$$

$$MS = SS/df \Rightarrow SS_{Regression} = MS_{Regression} \times df_{Regression} = 2.1045 \tag{15}$$

The total SS is always equal to the sum of the other SS:

$$SS_{total} = SS_{Regression} + SS_{error} = 5.3545 \tag{16}$$

Therefore the ANOVA table is:

Source	df	SS	MS	F	p-value
Regression	2	2.1045	1.0523	3.8853	0.05
error	12	3.2500	0.2708		
total	14	5.3545			

R^2 and adjusted R^2 are:

$$R^2 = 1 - \frac{SS_{error}}{SS_{total}} = 0.3930 \tag{17}$$

$$\overline{R}^2 = 1 - \frac{MS_{error}}{MS_{total}} = 0.2047 \tag{18}$$

Problem 4

Problem T7: Matrix algebra of fitted value and residuals

(A)

Suppose b is a "candidate" value for the parameter β. The quantity $y_i x_i' b$ is called the residual for the i-th observation, it measures the vertical distance between the data point (x_i, y_i) and the hyperplane $y = x'b$, and thus assesses the degree of fit between the actual data and the model. The sum of squared residuals (SSR) is a measure of the overall model fit:

$$S(b) = \sum_{i=1}^{n} (y_i - x_i' b)^2 = (y - Xb)^T (y - Xb), \tag{19}$$

The value of b which minimizes this sum is called the OLS estimator for β. The function $S(b)$ is quadratic in b with positive-definite Hessian, and therefore this function possesses a unique global minimum at $b = \hat{\beta}$, which can be given by the explicit formula:

$$\hat{\beta} = \arg\min_b S(b) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i y_i \tag{20}$$

or equivalently in matrix form,

$$\hat{\beta} = (X^T X)^{-1} X^T y \ . \tag{21}$$

After we have estimated β, the fitted values (or predicted values) from the regression will be

$$\hat{y} = X\hat{\beta} = Py, \tag{22}$$

where $P = X(X^T X)^1 X^T$ is the projection matrix. The annihilator matrix $M = I_n P$ is a projection matrix onto the space orthogonal to X. Both matrices P and M are symmetric and idempotent (meaning that $P^2 = P$), and relate to the data matrix X via identities $PX = X$ and $MX = 0$. Matrix M creates the residuals from the regression:

$$e = y - X\hat{\beta} = My = M\varepsilon. \tag{23}$$

(B)

Using these residuals we can estimate the value of σ^2:

$$s^2 = \frac{e'e}{n-p} = \frac{y'My}{n-p} = \frac{S(\hat{\beta})}{n-p} \tag{24}$$

where we can find that

$$SSR = S(\hat{\beta}) = e'e = (M\varepsilon)'(M\varepsilon) = \varepsilon' M \varepsilon \tag{25}$$

using the conclusion of part(A) as well as the symmetric and idempotent properties of M.

Problem 5

Problem T8: No Covariance

Answer

By definition,

$$Cov(b, e|X) = E[(b - E[b|X])(e - E[e|X])'|X] \tag{26}$$

Since $E[b|X] = \beta$, we have

$$b - E[b|X] = A\epsilon \tag{27}$$

where $A = (X'X)^{-1}X'$.
Use (A) from the previous problem,

$$e - E[e|X] = M\epsilon - 0 = M\epsilon \tag{28}$$

Therefore,

$$Cov(b, e|X) = E[A\epsilon(M\epsilon)'|X] = AE[\epsilon\epsilon']M \tag{29}$$

since both A and M are functions of X. Further,

$$AE[\epsilon\epsilon']M = E[\epsilon\epsilon']AM = E[\epsilon\epsilon'](X'X)^{-1}X'M = E[\epsilon\epsilon'](X'X)^{-1}(MX)' = 0 \tag{30}$$

Therefore using $MX = 0$,

$$Cov(b, e|X) = 0 \tag{31}$$

Problem 6

Problem T9: Variance of s^2

Answer

The estimator $\hat{\beta}$ is normally distributed, with mean and variance as given in the lecture note:

$$\hat{\beta} \sim \mathcal{N}\big(\beta,\ \sigma^2(X'X)^{-1}\big) \tag{32}$$

Consider the z-statistic:

$$z = \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2(X'X)^{-1}}} \tag{33}$$

then $z|X \sim N(0, 1)$.
σ^2 in z is replaced by s^2 in t, consider the t-statistic:

$$t = \frac{z}{s^2/\sigma^2} = \frac{z}{\sqrt{e'e/[(n-k)\sigma^2]}} = \frac{z}{\sqrt{q/(n-k)}} \tag{34}$$

where $q = ee'/\sigma^2$.
According to the conclusion in problem T7,

$$q = \frac{e'e}{\sigma^2} = \frac{\epsilon' M\epsilon}{\sigma^2} = \frac{\epsilon'}{\sigma}M\frac{\epsilon}{\sigma} \tag{35}$$

Using the fact that if $a = \frac{\epsilon'}{\sigma} \sim N(0, I)$, and M is an idempotent matrix, then "the quadratic form" $a'Aa$ has a χ^2-distribution with # of degrees of freedom = rank(M), therefore the estimator s^2 will be proportional to the chi-squared distribution:

$$s^2 \ \sim \ \frac{\sigma^2}{n-k} \cdot \chi^2_{n-k} \tag{36}$$

According to the hint, since the mean of s^2 is $\mu(s^2|X) = \frac{\sigma^2}{n-k}$, then

$$Var(s^2|X) = 2\mu(s^2|X) = \frac{2\sigma^2}{n-k} \tag{37}$$