31092924

Lu Wang

MATH 644: Regression Analysis Methods

MID-TERM EXAM
Fall, 2012

(Time allowed: TWO Hours)

INSTRUCTIONS TO STUDENTS:

1. This test contains **FIVE** questions and comprises **SEVEN** printed pages.
2. Answer **ALL** questions for a total of 100 marks.
3. This is a **closed-book** test; only a one-page formula sheet and non-programmable calculators are allowed.
4. Write your name on the front of your answer booklet and on any additional sheets you write on.

1. A regression analysis relating test scores (Y) to training hours (X) produced the following fitted equation: $\hat{y} = 10 + 0.56x$.

(a) What is the fitted value of the response variable corresponding to $x = 7$?

(b) What is the residual corresponding to the data point with $x = 7$ and $y = 17$?

(c) If the number of training hours is increased by 10, how is the expected test score affected?

(d) Consider the data point in part (b). An additional test score is to be obtained for a new observation at $x = 7$. Would the test score for the new observation necessarily be 17? Explain.

(e) The sums of squares error (SSE) for this model was found to be 11. If there were $n = 18$ observations, provide the best estimate for σ^2.

(f) Rewrite the regression equation in terms of x^*, where x^* is training time measured in minutes.

(a) $X=7, \quad \hat{Y} = 10+0.56(7) = 13.92$

(b) $e = Y - \hat{Y} = 17 - 13.92 = 3.08$

(c) $X' = X + 10 = 17 \quad \hat{Y}' = 10 + 0.56(17) = 19.52, \quad \hat{Y}' - \hat{Y} = 5.6$, the test score increased by 5.6

(d) No, since $Y_i = 10 + 0.56X + \varepsilon_i$ is a random varible, the new observation is not necessarily 17

(e) $SSE = 11 \quad MSE = \frac{SSE}{n-2} = \frac{11}{18-2} = 0.6875$

The best estimate for σ^2 is $MSE = 0.6875$

(f) $X^* = 60 \cdot X, \quad b_1^* = \frac{\Sigma(X_i^* - \bar{X}^*)(Y_i - \bar{Y})}{\Sigma(X_i^* - \bar{X}^*)^2} = \frac{60\,\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{60^2 \cdot \Sigma(X_i - \bar{X})^2} = \frac{1}{60} b_1 = 0.0093$

$b_0^* = \bar{Y} - b_1^* \bar{X}^* = \bar{Y} - \frac{1}{60} b_1 \cdot (60\bar{X}) = \bar{Y} - b_1 \bar{X} = 10$

Thus $\hat{Y}^* = 10 + 0.0093 \cdot X^*$

2. A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79. The results follow; X is age, and Y is a measure of muscle mass. Assume that the simple linear regression model is appropriate. The following is the R output of regressing Y with respect to X.

```
Call: lm(formula = Y ~ X)

Residuals:
     Min         1Q    Median         3Q       Max
-16.1368   -6.1968   -0.5969    6.7607   23.4731

Coefficients:
            Estimate    Std. Error    t value    Pr(>|t|)
(Intercept) 156.3466    5.5123        28.36      <2e-16 ***   \\
X           -1.1900     0.0902        -13.19     <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.173 on 58 degrees of freedom Multiple
R-squared: 0.7501,      Adjusted R-squared: 0.7458  \\
F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

Given $\bar{X} = 60$, $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 8210$, based on the R output given as above, obtain the following:

(a) Conduct a test using t-statistic to decide whether or not there is a linear association between amount of muscle mass and age. Control the risk of Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

(b) Estimate with a 95 percent confidence interval the difference in expected muscle mass for women whose ages differ by five year.

(c) Obtain a 95 percent confidence interval for the mean muscle mass for women of age 60. Interpret your confidence interval.

(d) Predict the muscle mass for a woman of age 50 using a 95% prediction interval. Interpret your prediction interval.

(a) $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

Since $\frac{b_1 - \beta_1}{s(b_1)}$ follows a t distribution with degree of freedom $n-2$, we test $t^* = \frac{b_1}{s(b_1)}$

If $|t^*| > t(0.975, 58)$ reject H_0 ; If $|t^*| < t(0.975, 58)$, accept H_0

$|t^*| = \left|\frac{-1.19}{0.0902}\right| = |-13.19| = 13.19 > t(0.975, 58) = 2.002$ P-value $< 2 \times 10^{-16} < 0.05$

Thus we reject H_0, and conclude that there is linear relationship between muscle mass and age

(b) Suppose $X_2 = X_1 + 5$ $E(Y_2) - E(Y_1) = (\beta_0 + \beta_1 X_2) - (\beta_0 + \beta_1 X_1) = \beta_1 (X_2 - X_1) = 5\beta_1$

$b_1 = -1.19$ $s(b_1) = 0.0902$

The 95% CI of β_1 is: $b_1 \pm t(n-2, 1-\alpha/2) s(b_1) = -1.19 \pm 2.002(0.0902) = -1.19 \pm 0.18$ or $(-1.37, -1.01)$

The 95% CI of $5\beta_1$ is $5[b_1 \pm t(n-2, 1-\alpha/2) s(b_1)] = -5.95 \pm 0.90$ or $(-6.85, -5.05)$

(c) $X = 60$ $\hat{Y} = 156.3466 - 1.19(60) = 84.9466$

$Var(\hat{Y}) = MSE\left[\frac{1}{n} + \frac{(X-\bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right] = (8.173)^2 \left[\frac{1}{60} + \frac{(60-60)^2}{8210}\right] = 1.11$

$s(\hat{Y}) = \sqrt{1.11} = 1.06$

The 95% CI of $E(\hat{Y})$ is: $\hat{Y} \pm t(1-\alpha/2, n-2) s(\hat{Y}) = 84.9466 \pm 2.002(1.06) = 84.9466 \pm 2.1221$ or $(82.8245, 87.0687)$

With 95% confidence interval, the mean muscle mass at age 60 is between 82.8245 and 87.0687

(d) $X = 50$, $\hat{Y} = 156.3466 - 1.19(50) = 96.8466$

$Var(pred) = MSE\left[1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right] = (8.173)^2 \left[1 + \frac{1}{60} + \frac{(50-60)^2}{8210}\right] = 68.72$

$s(pred) = \sqrt{68.72} = 8.29$

The 95% prediction interval is: $\hat{Y} \pm t(1-\alpha/2, n-2) s(pred) = 96.8466 \pm 2.002(8.29) = 96.8466 \pm 16.5966$ or $(80.25, 113.44$

With 95% prediction interval, the muscle mass at age 50 is between 80.25 and 113.4432

3. Based on the R output given in Problem 2, do the following:

(a) Set up the ANOVA table.

(b) Test whether or not $\beta_1 = 0$ using an F test with $\alpha = 0.05$. State the alternatives, decision rule, and conclusion.

(c) What proportion of the total variation in muscle mass remains "unexplained" when age is introduced into the analysis? Is this proportion relatively small or large?

(d) Obtain R^2 and r.

(a) $MSE = (8.173)^2 = 66.7979$ $\quad SSE = MSE \cdot (n-2) = 66.7979(58) = 3874.2799$

$F^* = \frac{MSR}{MSE} = 174.1$, then $MSR = F^* \cdot MSE = 174.1(66.7979) = 11629.5194$

$SSR = MSR \cdot 1 = 11629.5194$ $\quad SST = SSE + SSR = 15503.7993$

ANOVA:

Source	df	SS	MS	F	P-value
Regression	1	11629.5194	11629.5194	174.1	$< 2.2 \times 10^{-16}$
Error	58	3874.2799	66.7979		
Total	59	15503.7993			

(b) $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

Since $\frac{MSR}{MSE}$ follows a F distribution with degree of freedom 1, 58, ~~we test~~ $F^* = \frac{MSR}{MSE}$

If $F^* > F(1-\alpha, 1, 58)$ reject H_0 ; If $F^* < F(1-\alpha, 1, 58)$, accept H_0

$F^* = 174.1 > F(0.95, 1, 58) = 4.007$

Thus we reject $\beta_1 = 0$ and conclude that there is linear relationship between muscle mass and age

(c) $R^2 = 0.7501$ $\quad 1-R^2 = 0.2499$

Thus 24.99% of the total variation in muscle mass remains "unexplained" when age is introduced.

(d) $R^2 = \frac{SSR}{SST} = 0.7501$

Since muscle mass and age have negative relationship,

$r = -\sqrt{R^2} = -0.866$

4. For model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$, we have $\sum_{i=1}^n Y_i = 0, SSR = 15, SSE = 5$. Let e_i be the fitted residuals of the least squares estimation. Find $\sum_{i=1}^n (Y_i + 5e_i)^2$.

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, then $\hat{Y}_i = b_0 + b_1 X_i$, where $b_0 = \bar{Y} - b_1 \bar{X}$, $b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$

Thus $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i = 0$, $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = 0$

Since $SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum_{i=1}^n \hat{Y}_i^2 - n\bar{\hat{Y}}^2 = \sum_{i=1}^n \hat{Y}_i^2 - n\bar{Y}^2 = \sum_{i=1}^n \hat{Y}_i^2$,

$\sum_{i=1}^n \hat{Y}_i^2 = 15$.

Since $e_i = Y_i - \hat{Y}_i$, $Y_i = e_i + \hat{Y}_i$

Thus $\sum_{i=1}^n (Y_i + 5e_i)^2 = \sum_{i=1}^n (\hat{Y}_i + e_i + 5e_i)^2 = \sum_{i=1}^n (\hat{Y}_i + 6e_i)^2 = \sum_{i=1}^n \hat{Y}_i^2 + 12 \sum_{i=1}^n \hat{Y}_i e_i + 36 \sum_{i=1}^n e_i^2$

Since $\sum_{i=1}^n \hat{Y}_i e_i = 0$, $SSE = \sum_{i=1}^n e_i^2 = 5$,

$\sum_{i=1}^n (Y_i + 5e_i)^2 = 15 + 36(5) = 195$

5. An analyst wanted to fit the regression model

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, i = 1, \ldots, n$$

by the least squares estimation when it is known that $\beta_2 = 4$. State the least square criterion and derive the least squares normal equations.

$Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_1 X_{i1} - 4X_{i2} - \beta_3 X_{i3})^2$, we want to find estimates b_1 and b_3 which minimize Q

Let $\dfrac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n} -2X_{i1}(Y_i - \beta_1 X_{i1} - 4X_{i2} - \beta_3 X_{i3})^2 = 0$

$\dfrac{\partial Q}{\partial \beta_3} = \sum_{i=1}^{n} -2X_{i3}(Y_i - \beta_1 X_{i1} - 4X_{i2} - \beta_3 X_{i3})^2 = 0$

Then $\sum_{i=1}^{n} b_1 X_{i1}^2 + 4\sum_{i=1}^{n} X_{i1}X_{i2} + \sum_{i=1}^{n} b_3 X_{i1}X_{i3} = \sum_{i=1}^{n} X_{i1}Y_i$ or $\sum_{i=1}^{n} b_1 X_{i1}^2 + \sum_{i=1}^{n} b_3 X_{i1}X_{i3} = \sum_{i=1}^{n} X_{i1}Y_i - 4\sum_{i=1}^{n} X_{i1}X_{i2}$

$\sum_{i=1}^{n} b_1 X_{i1}X_{i3} + 4\sum_{i=1}^{n} X_{i2}X_{i3} + \sum_{i=1}^{n} b_3 X_{i3}^2 = \sum_{i=1}^{n} X_{i3}Y_i$ or $\sum_{i=1}^{n} b_1 X_{i1}X_{i3} + \sum_{i=1}^{n} b_3 X_{i3}^2 = \sum_{i=1}^{n} X_{i3}Y_i - 4\sum_{i=1}^{n} X_{i2}X_{i3}$

Then we can find unbiased estimates for β_1 and β_3. Besides, b_1 and b_3 have the minimum variance among all the unbiased estimates