Statistics for Financial Engineering - Refresher Seminar, Homework 3

Radoš Radoičić

July 31, 2017

Homework has two types of problems: theoretical (T) and practical (P).

You need to show all your work. It is due on 08/23/17.

We will be using the notation from the lecture notes, as well as the Assumptions 1-5.

Problem T1: "Prediction with confidence":

1. Our regression model can be written as $y_i = \beta' x_i + \epsilon_i$, $1 \leq i \leq n$. Find the $100(1-\alpha)\%$ confidence interval for the *mean response* $\beta' x$ at a given vector x of regressor observations.
2. Now suppose a future observation y is taken at a regressor x. Find a $100(1-\alpha)\%$ prediction interval for y.

Problem T2: "Linear hypothesis": Suppose that in the multiple linear regression model $Y = X\beta + \epsilon$ with 20 observations, we have $s^2 = 1.5$, $b = (1, 2, 1)'$, and

$$(X'X)^{-1} = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Test the hypothesis

$$H_0 : \beta_1 = 4 - 2\beta_3, \text{ and } \beta_2 = 0.5 + \beta_3,$$

at the level of significance $\alpha = 0.05$.

Problem T3: "Quadratic polynomial regression" Suppose that you are given the following (X, Y) data:

$$(1.9, 0.7), (0.8, -1.0), (1.1, -0.2), (0.1, -1.2), (-0.1, -0.1),$$

$$(4.4, 3.4), (4.6, 0.0), (1.6, 0.8), (5.5, 3.7), (3.4, 2.0).$$

We will fit quadratic regression $y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon$, using the multiple linear regression model.

1. Determine the values of b_1, b_2, b_3, and s^2.
2. Determine the values of $\text{Var}(b_1)$, $\text{Var}(b_2)$, $\text{Var}(b_3)$, $\text{Cov}(b_1, b_2)$, $\text{Cov}(b_1, b_3)$, and $\text{Cov}(b_2, b_3)$.
3. Carry out a test of the following hypothesis $H_0 : \beta_3 = 0$.
4. Determine the value of (uncentered) R^2 and adjusted R^2.
5. Carry out a test of the following hypothesis $H_0 : \beta_2 = \beta_3 = 0$.
6. Construct a 95% confidence interval for y at a given x.

Problem T4: "Centering" Problem 12.4 from the textbook.

Problem T5: "Regression through the origin" Problem 12.8 from the textbook.

Problem T6: "Hide and Seek" Complete the following ANOVA table relative to the model $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$.

Source	df	SS	MS	F	p-value
Regression	?	?	?	?	0.05
error	?	3.250	?		
total	14	?			

What is R^2? What is adjusted R^2?

Problem T7: "Matrix algebra of fitted values and residuals" Show the following:

(A) $\hat{y} = Py$, $e = My = M\epsilon$.

(B) $SSR = \epsilon' M \epsilon$.

(We are using the notation in Lecture Notes, Part 2, see page 6.)

Problem T8: "No Covariance" Prove (under the Assumptions $1--4$ in Lecture Notes, Part 2) that $\text{Cov}(b, e|X) = 0$ (where $e = y - Xb$; see page 11. in the Lecture Notes, Part 2).

Hint: By definition, $\text{Cov}(b, e|X) = E[(b - E[b|X])(e - E[e|X])'|X]$. Since $E[b|X] = \beta$, we have $b - E[b|X] = A\epsilon$, where $A = (X'X)^{-1}X'$. Use (A) from the previous problem to show that $e - E[e|X] = M\epsilon$. $E[A\epsilon\epsilon' M|X] = AE[\epsilon\epsilon'|X]M$ since both A and M are functions of X. Finally, use $MX = 0$.

Problem T9: "Variance of s^2" Prove (under the Assumptions $1 - -5$ of Lecture Notes, Part 2) that

$$\text{Var}(s^2|X) = \frac{2\sigma^4}{n - K}.$$

Hint: If a random variable is distributed as $\chi^2(m)$, then its mean is m and variance $2m$.

Problem P1: "NASDAQ" The file **d_nasdaq_82stocks.txt** contains the daily log returns of the NASDAQ Composite Index and 82 stocks from January 3, 1990 to December 29, 2006. We want to track the returns of NASDAQ by using only a small number of stocks from the given 82 stocks.

(A) Construct a full regression model.

(B) Use partial F-statistics and backward elimination to select variables from the full regression model in (A). Write down the selected model.

(C) Compare the full and selected models. Summarize your comparison in an ANOVA table.

(D) For the selected regression model in (B), perform residual diagnostics.

(E) If you can only use at most five stocks to track the daily NASDAQ log returns, describe your model selection procedure and your constructed model.

Problem P2: Chapter 12 R-lab.

Problem P3: Problem 13.1 from textbook.

Problem P4: Problem 13.5 from textbook.

CHALLENGE: "Wald vs. Likelihood Ratio" In the restricted least squares, the sum of squared residuals is minimized subject to the constraint implied by the null hypothesis $R\beta = r$. Form the Lagrangian as

$$L = \frac{1}{2}(y - X\tilde{\beta})'(y - X\tilde{\beta}) + \lambda'(R\tilde{\beta} - r),$$

where λ is the m-dimensional vector of Lagrange multipliers (recall: R is $m \times k$, $\tilde{\beta}$ is $k \times 1$, and r is $m \times 1$). Let $\hat{\beta}$ be the restricted least squares estimator of β. It is the solution to the constrained minimization problem.

(A) As usual, let b be the unrestricted OLS estimator. Show

$$\hat{\beta} = b - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(Rb - r), \text{ and } \lambda = [R(X'X)^{-1}R']^{-1}(Rb - r).$$

Hint: The first-order conditions are $X'y - (X'X)\hat{\beta} = R'\lambda$ or $X'(y - X\hat{\beta}) = R'\lambda$. Combine this with the constraint $R\hat{\beta} = r$ to solve for λ and $\hat{\beta}$.

(B) Let $\hat{\epsilon} = y - X\hat{\beta}$ be the residuals from the restricted regression. Show

$$SSR_R - SSR_U = (b - \hat{\beta})'(X'X)(b - \hat{\beta}) = (Rb - r)'[R(X'X)^{-1}R']^{-1}(Rb - r),$$

thereby,

$$SSR_R - SSR_U = \lambda'R(X'X)^{-1}R'\lambda = \hat{\epsilon}'P\hat{\epsilon},$$

where P is the projection matrix.
Hint: For the first equality, use the add-and-subtract strategy: $SSR_R = (y - X\hat{\beta})'(y - X\hat{\beta}) = [(y - Xb) + X(b - \hat{\beta})]'[(y - Xb) + X(b - \hat{\beta})]$, and then use the normal equations $X'(y - Xb) = 0$. For the second and third equality use (A). For the fourth equality, use the first-order condition mentioned in (A) that $R'\lambda = X'\hat{\epsilon}$.

(C) Verify that the F-ratio derived by Wald principle in Lecture Notes (on page 19.) is the same as the F-ratio stated later in the notes by the likelihood-ratio principle (on page 21.)

CHALLENGE: "Decomposition" Take the unrestricted model to be a regression where one of the regressors is a constant, and the restricted model to be a regression where the only regressor is a constant.

(A) Show that (B) in the previous problem implies the decomposition of the total sum of squares into the regression sum of squares and the residual error sum of squares

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2,$$

stated in the Lecture Notes (on page 9). Hint: What is $\hat{\beta}$ in this case? Show $SSR_R = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $(b - \hat{\beta})'(X'X)(b - \hat{\beta}) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

(B) For a regression where one of the regressors is a constant, prove that the F-ratio satisfies $F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$.

CHALLENGE: "Jackknife" In this exercise, you will be proving the relationship between the studentized and standardized residuals given in the Lecture Notes.

(A) Let A be a $p \times p$ nonsingular matrix and U and V be $p \times 1$ vectors. Use the following matrix inversion trick from linear algebra:

$$(A + UV')^{-1} = A^{-1} - A^{-1}U(I + V'A^{-1}U)^{-1}V'A^{-1},$$

to prove

$$(X'_{(-i)}X_{(-i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1 - h_{ii}},$$

and, hence,

$$x_i'(X'_{(-i)}X_{(-i)})^{-1}x_i = \frac{h_{ii}}{1 - h_{ii}}.$$

(B) Show that, after deletion of the observation (x_i, y_i), we have

$$\hat{\beta}_{(-i)} = \hat{\beta} - \frac{(X'X)^{-1}x_i}{1 - h_{ii}} e_i,$$

where $e_i = y_i - x_i'\hat{\beta}$. Letting $\hat{e}_{(-i)} = y_i - x_i'\hat{\beta}_{(-i)}$ and $e_i' = \frac{e_i}{s\sqrt{1-h_{ii}}}$, derive the representations:

$$\hat{e}_{(-i)} = \frac{e_i}{1 - h_{ii}} = \frac{e_i'}{\sqrt{1 - h_{ii}}} s, \quad s^2_{(-i)} = \frac{n - p - (e_i')^2}{n - p - 1} s^2.$$

(C) Show that $\widehat{\mathrm{Var}}(\hat{e}_{(-i)}) = s^2_{(-i)}/(1 - h_{ii})$, and hence prove that

$$\hat{e}_{(-i)} = \frac{e_i'}{\sqrt{\frac{n-p-(e_i')^2}{n-p-1}}}.$$