Ma 439 Test — Linear Statistical Models — Fall 2010
Model Solutions

Prof. Sawyer — Washington Univ. — Test date October 26, 2010

1. (Let X be a d-dimensional random vector that is normally distributed with parameters μ and Σ (that is, $X \approx N(\mu, \Sigma)$).

Show that $(X-\mu)'\Sigma^{-1}(X-\mu)$ has a chi-square distribution with r degrees of freedom for some r, and find r.)

Solution: If $X \approx N(\mu, \Sigma)$, then $X = \mu + \Sigma^{1/2}N$ where $N \approx N(0, I_d)$. Thus $X - \mu = \Sigma^{1/2}N$ and $Q = (X-\mu)'\Sigma^{-1}(X-\mu) = (\Sigma^{1/2}N)'\Sigma^{-1}\Sigma^{1/2}N = N'\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}N = N'N = \sum_{i=1}^{d} N_i^2 \approx \chi^2_d$ where $N_1, \ldots, N_d$ are independent $N(0,1)$, so that $Q \approx \chi^2_r$ with $r = d$.

2. Let $X = (X_1, \ldots, X_r)'$ and $Y = (Y_1, \ldots, Y_s)'$ be r- and s-dimensional random vectors, respectively. Let $C = \mathrm{Cov}(X, Y)$ be the $r \times s$ matrix with entries

$$C_{ij} = \mathrm{Cov}(X_i, Y_i), \qquad 1 \le i \le r,\ 1 \le j \le s$$

Let A be an $a \times r$ matrix and B a $b \times s$ matrix.

(i) Prove that

$$\mathrm{Cov}(AX,\, BY) \;=\; A\,\mathrm{Cov}(X, Y)B' \tag{1}$$

(ii) Prove that (1) is dimensionally correct: That is, the matrix dimensions are such that the right-hand side of (1) is defined, and the two sides of (1) have the same matrix dimensions.

Solutions: (i) **(First Proof)** Since $\mathrm{Cov}(X_i, Y_j) = E(X_iY_j) - E(X_i)E(Y_j)$,

$$\mathrm{Cov}\big((AX),(BY)\big)_{pq} \;=\; \mathrm{Cov}\big((AX)_p, (BY)_q\big) \;=\; \mathrm{Cov}\left(\sum_{i=1}^{r} A_{pi}X_i,\, \sum_{j=1}^{s} B_{qj}Y_j\right)$$

$$= \sum_{i=1}^{r} A_{pi}\,\mathrm{Cov}\left(X_i,\, \sum_{j=1}^{s} B_{qj}Y_j\right) \;=\; \sum_{i=1}^{r} A_{pi} \sum_{j=1}^{s} B_{qj}\,\mathrm{Cov}(X_i, Y_j)$$

$$= \sum_{i=1}^{r} A_{pi} \sum_{j=1}^{s} \mathrm{Cov}(X_i, Y_j)(B')_{jq} \;=\; (A\,\mathrm{Cov}(X, Y)B')_{pq}$$

This implies $\mathrm{Cov}(AX), BY) = A\,\mathrm{Cov}(X, Y)B'$, since the matrix components (or matrix entries) are the same.

(i) **(Second Proof)** Since $\mathrm{Cov}(X_i, Y_j) = E(X_iY_j) - E(X_i)E(Y_j)$, we have $\mathrm{Cov}(X, Y) = E(XY') - E(X)E(Y)'$ and

$$\begin{aligned}
\mathrm{Cov}(AX, BY) \;&=\; E\big((AX)(BY)'\big) - E(AX)E(BY)' \\
&=\; E\big(AXY'B'\big) - AE(X)\big(BE(Y)\big)' \\
&=\; AE\big(XY'\big)B' - AE(X)E(Y)'B' \\
&=\; A\big(E(XY') - E(X)E(Y)'\big)B' \;=\; A\,\mathrm{Cov}(X, Y)B'
\end{aligned}$$

(ii) The left-hand side of (1) is $a \times b$. The right-hand side (in terms of the dimensions) is $(a \times r)(r \times s)(s \times b) = a \times b$.

3. (Suppose that we have a univariate regression

$$Y_i \;=\; \beta_1 \;+\; \beta_2 X_{i2} \;+\; \beta_3 X_{i3} \;+\; \beta_4 X_{i4} \;+\; \beta_5 X_{i5} \;+\; e_i \tag{3}$$

for $1 \le i \le n$, where e_i are independent $N(0, \sigma^2)$. We can write (3) in matrix form as $Y = X\beta + e$ where X is $n \times 5$ and β is 5×1. Let $\widehat{\beta}$ be the least-squares estimator of the vector β, and let $\text{SSE} = \sum_{i=1}^n \Big(Y_i - (X\widehat{\beta})_i\Big)^2$ be the usual error sum of squares.

Assume that we want to test the hypothesis

$$H_0 : \beta_2 = \beta_4 \quad \text{and} \quad \beta_3 = \beta_5 \tag{4}$$

which we can write in the form $H_0 : A\beta = 0$ where A is the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix}$$

Recall that $\widehat{\beta}$ is independent of SSE for the regression (3), and hence $A\widehat{\beta}$ is also independent of SSE. Assuming standard results about the distribution of $\widehat{\beta}$ and SSE,

(i) State the distribution of SSE in terms of X, β, σ, and n.

(ii) Find the distribution of the random vector $A\widehat{\beta}$.

(iii) Find $G = \text{Cov}(A\widehat{\beta})/\sigma^2$, show that G does not depend on β or σ, and find the distribution of $Q = (A\widehat{\beta} - A\beta)'G^{-1}(A\widehat{\beta} - A\beta)/\sigma^2$. (*Hint:* You can use the results of an earlier problem.)

(iv) What is the dimension of the matrix A?

(v) Use part (iii) to write down a test statistic for $H_0 : A\beta = 0$ that (a) is similar to Q, (b) whose distribution does not depend on β or σ, and (c) that has an F-distribution F_{df_1, df_2} given H_0. Find the degrees of freedom df_1 and df_2.)

Solutions: (i) Since the model (3) has $p = 5$ linear parameters other than σ, $\text{SSE} \approx \sigma^2 \chi^2_{n-p} = \sigma^2 \chi^2_{n-5}$.

(ii) A is 2×5 and $\widehat{\beta} \approx N(\beta,\, \sigma^2 (X'X)^{-1})$ is 5×1, so that $A\widehat{\beta}$ is 2×1 and joint normal with parameters

$$A\widehat{\beta} \;\approx\; N\big(A\beta,\, A\sigma^2 (X'X)^{-1} A'\big) \;\approx\; N\big(A\beta,\, \sigma^2 A (X'X)^{-1} A'\big)$$

(iii) $G = \text{Cov}(A\widehat{\beta})/\sigma^2 = A(X'X)^{-1}A'$ is 2×2 and does not depend on β or σ^2. Thus $(A\widehat{\beta} - A\beta)/\sigma \approx N(0, G)$ so that $\big((A\widehat{\beta} - A\beta)/\sigma\big)' G^{-1} (A\widehat{\beta} - A\beta)/\sigma \approx \chi^2_2$ by Problem 1.

(iv) A is a 2×5 matrix.

(v) Let $W = (1/2)(A\widehat{\beta})'\big(A(X'X)^{-1}A'\big)^{-1}(A\widehat{\beta})$. If $H_0 : A\beta = 0$ is true, then $Q = (A\widehat{\beta})' G^{-1} (A\widehat{\beta})/\sigma^2 = 2W/\sigma^2 \approx \chi^2_2$ and $W \approx (1/2)\sigma^2 \chi^2_2$.

Since $\widehat{\beta}$ and SSE are independent, W is independent of $\text{MSE} = \text{SSE}/(n-5) \approx (1/(n-5))\sigma^2 \chi^2_{n-5}$ by part (i). Thus, given $H_0 : A\beta = 0$,

$$F \;=\; \frac{W}{\text{MSE}} \;=\; \frac{W/\sigma^2}{\text{MSE}/\sigma^2} \;\approx\; F(2, n-5)$$

and F can be used for an F-test of $H_0 : A\beta = 0$. In part (v) above, $df_1 = 2$ and $df_2 = n - 5$.

4. (An experimenter does a regression of for 35 values of an observed variable `YY` along with four covariates that he calls `AA1`, `BB1`, `CC1`, and `DD1`. Part of the SAS output for this regression is in Table 1 below.

Note that, in addition to the Parameter Estimate table, the output has two additional tables with P-values for the four covariates. Call these the "Type I" and "Type III" tables from the "Type I SS" and "Type III SS" in the table headings. Both tables have columns with F statistics and corresponding P-values, but they are not the same.

(i) What is the difference between the two sets of P-values, in the Type I and Type III tables? What are the hypotheses H_0 that are tested in each case?

(ii) Give or describe formulas for the Type I and the Type III SS values.

(iii) Each table has an "SS" column, which is the numerator of the F-statistic involved, but does not explicitly list the denominator. What is the denominator of the F statistics for these tests?

(iv) F-statistics have two degrees of freedom, one for the numerator and one for the denominator. The degrees of freedom for the numerator are listed in the table. What is the number of degrees of freedom for the denominator?

(v) Note that the F-statistics and the P-values for the last entry, `DD1`, are the same in both tables. Is this a coincidence? Does this happen in general? If so, why?)

Solutions: Write the experimenter's regression as $Y = X\beta + e$ where Y and e are $n \times 1$ and X is $n \times p$ for $n = 35$ and $p = 5$. If $\widetilde{X}$ is an arbitrary $n \times s$ matrix, let $\mathrm{SSE}(\widetilde{X})$ be the error sum of squares $\sum_{i=1}^{n}(Y_i - (\widetilde{X}\widetilde{\beta}))^2$ for $\widetilde{\beta} = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}Y$.

For $1 \le i \le p = 5$, let X_i be the $n \times i$ matrix composed of the first i columns of X, INCLUDING the intercept column. (Thus X_1 corresponds to the intercept and X_2 to the intercept plus the first non-intercept covariate.) The Type I SS for the i^{th} covariate for $2 \le i \le p$ (ignoring the intercept) is $\mathrm{SSI}_i = \mathrm{SSE}(X_{i-1}) - \mathrm{SSE}(X_i)$. The Type III SS for the i^{th}covariate is $\mathrm{SSIII}_i = \mathrm{SSE}(X_{(i)}) - \mathrm{SSE}(X)$, where $X_{(i)}$ is the $n \times (p-1)$ matrix obtained by dropping the i^{th} column of X corresponding to the i^{th} covariate.

(i) The null hypothesis in both cases is $H_0 : \beta_i = 0$: that is, that the i^{th} covariate does not contribute to the regression. (This counts the intercept as the first covariate, so $2 \le i \le p$ for the non-intercept covariates.)

Type I P-values are based on the statistics $T_i = \mathrm{SSI}_i/\mathrm{MSE}$ where $\mathrm{MSE} = \mathrm{SSE}(X)/(n-p)$ for the full regression. Given $H_0 : \beta_i = 0$, $T_i \approx F(1, n-p) = F(1, 30)$ in this case. Type III P-values are based on $T_i = \mathrm{SSIII}_i/\mathrm{MSE}$, which have the same distribution given H_0. The P-values are different because the two sets of SS statistics are in general different.

Type I P-values correspond to testing whether the i^{th} covariate provides a significantly better fit to the dependent variables Y_i than do the first $i-1$ covariates (including the intercept). Type III P-values correspond to testing whether the i^{th} covariate provides a significantly better fit to the Y_i than do the $p-1$ covariates (including the intercept) other than the i^{th}covariate.

(ii) See above.

(iii) The denominator is $\mathrm{MSE} = \mathrm{SSE}(X)/(n-p)$, where $\mathrm{SSE}(X)$ is the error sum of squares for the full regression.

(iv) The error degrees of freedom is $n - p = 35 - 5 = 30$ in this case.

(v) Since $\text{SSI}_i = \text{SSIII}_i$ for the last covariate by definition, the Type I and Type III F-statistics and P-values for the last covariate are the same. This happens in general and is not a coincidence.

5. (Let $A = \text{Corr}(X)$ the the correlation matrix of a random vector $X \in R^n$. Let $A = RDR'$ be the spectral decomposition of A, where R is an orthogonal matrix and D is diagonal with diagonal entries $\lambda_1, \ldots, \lambda_n$. Set $g_{ij} = R_{ij}\sqrt{\lambda_j}$ where R_{ij} are the matrix entries of R.

(i) Show that $\sum_{i=1}^n g_{ij}^2 = \lambda_j$ for $1 \le j \le n$ and $\sum_{i=1}^n g_{ij_1} g_{ij_2} = 0$ for $j_1 \ne j_2$, $1 \le j_1, j_2 \le n$.

(ii) Show that $\sum_{j=1}^n g_{ij}^2 = 1$ for each i, $1 \le i \le n$. (*Hint:* Be careful!))

Solutions: Note $G = RD^{1/2}$, so that $G'G = (RD^{1/2})'RD^{1/2} = D^{1/2}R'RD^{1/2} = D$ and $GG' = RD^{1/2}(RD^{1/2})' = RD^{1/2}D^{1/2}R' = RDR' = A$. Thus

(i) $\sum_{i=1}^n g_{ij_1} g_{ij_2} = (G'G)_{j_1 j_2} = D_{j_1 j_2} = 0$ if $j_1 \ne j_2$ and equals λ_j if $j_1 = j_2 = j$.

(ii) $\sum_{j=1}^n g_{i_1 j} g_{i_2 j} = (GG')_{j_1 j_2} = A_{j_1 j_2}$, so that $\sum_{j=1}^n g_{ij}^2 = A_{jj} = 1$ for $1 \le j \le n$, since A is a correlation matrix.

6. (Let $X_1, \ldots, X_n$ be d blood-protein measurements for n individuals who have the rare disease D. Let $Y_1, \ldots, Y_n$ be the same blood-protein measurements for n unaffected individuals who were chosen so that the i^{th} individual is randomly chosen from unaffected individuals with the same age, gender, income, and place of birth as the i^{th} affected individual. (This is called a "Case-Control" design, where X_i are the "cases" and Y_i are the "controls". Note that X_i and Y_i are not independent, since they were chosen to have the same values of age, gender, etc.)

Assume that the pairs (X_i, Y_i) are independent and normal, and that we want to test H_0 : $E(X) = E(Y)$. If $Z_i = X_i - Y_i$, then $Z_1, \ldots, Z_n$ are independent normal $N(\mu_Z, \Sigma_Z)$ and H_0 is equivalent to $H_0 : E(Z) = 0$.

Suppose that the person who gathered the data uses SAS to carry out the appropriate test of $H_0 : E(Z) = 0$, and that an important part of her output is in Table 2 below.

(i) Does the person accept of reject H_0? What is the P-value?

(ii) The test is equivalent to an F-test, with degrees of freedom listed in the output. How did SAS calculate the two degrees of freedom? In terms of these two numbers, what is the dimension d and the sample size n?)

Solutions: (i) Reject H_0 at level of significance $\alpha = 0.05$. All four multivariate tests have $P = 0.0216$, when is taken as the P-value for the test.

(ii) The F-statistics are distributed as $F(d, n-d)$ given H_0, or more generally $F(d, n-p-d+1)$ where p is the number of free covariate parameters, which is $p = 1$ here for the intercept. Thus $d = 3$ (from `Num DF`) and $n - d = 34$ from `Den DF`, so that the sample size is $n = 34 + d = 37$.

7. (Let $X1, X_2, \ldots X_m \in R^d$ be an independent vector-valued sample with distribution $N(\mu_X, \Sigma)$, and let $Y_1, Y_2, \ldots Y_n \in R^d$ be an independent sample with distribution $N(\mu_Y, \Sigma)$. Let H_0 be the hypothesis $H_0 : \mu_X = \mu_Y$. Write down the (Hotelling-T^2) statistic that is used for the Hotelling T^2 test of H_0.)

Solution:

$$T^2 = \frac{mn}{m+n}(\overline{Y} - \overline{X})' S^{-1} (\overline{Y} - \overline{X}) \quad \text{where}$$

$$S = \frac{1}{m+n-2}\left(\sum_{i=1}^m (X_i - \overline{X})(X_i - \overline{X})' + \sum_{j=1}^n (Y_i - \overline{Y})(Y_i - \overline{Y})' \right)$$

for $\overline{X} = (1/m)\sum_{i=1}^m X_i$ and $\overline{Y} = (1/n)\sum_{j=1}^n Y_j$.