STATISTICS 174: COMP EXAM 2000

AUGUST 18 2000

Part I

Assume data are generated by a model $y = X_1\beta_1 + X_2\beta_2 + \epsilon$ where y is $n \times 1$, X_1 is a $n \times p_1$ matrix of covariates, X_2 is another $n \times p_2$ matrix of covariates, β_1 and β_2 are respectively $p_1 \times 1$ and $p_2 \times 1$ parameter vectors, and ϵ is a $n \times 1$ vector of independent normally distributed random errors with mean 0 and variance σ^2. Suppose the statistician ignores or is unaware of the X_2 covariates and fits the model $y = X_1\beta_1 + \epsilon$, calculating the standard least squares estimator $\hat{\beta}_1$ under this assumption.

(a) Calculate the mean and variance of $\hat{\beta}_1$. Is the estimator biased or unbiased? If biased, write down the bias.

(b) The statistician forms a predictor vector $\hat{y} = X_1\hat{\beta}_1$ and calculates the residual sum of squares, $R = (y - \hat{y})^T(y - \hat{y})$. Show that the expected value of R is of the form $\beta_2^T C\beta_2 + (n - p_1)\sigma^2$, and give an explicit expression for the matrix C.

(c) What is the distribution of R? (Just write down the answer if you know it — no derivation is required for this part.)

(d) Suppose a second sample $y^* = X_1\beta_1 + X_2\beta_2 + \epsilon^*$ is to be taken, where ϵ^* is independent of ϵ but has the same distribution. Note that we are assuming that the covariate matrices X_1 and X_2 are the same for both samples. Again, the statistician uses $\hat{y}$ (as in part (b)) as a predictor. Calculate the expected sum of squared prediction errors, $E\{(y^* - \hat{y})^T(y^* - \hat{y})\}$, under this scenario.

(e) Now let us compare this with what would have happened if the statistician had used the correct model from the beginning, i.e. including X_2. Show that the mean squared prediction error in (d) is smaller than the corresponding mean squared prediction error when the statistician uses the correct model if and only if

$$\beta_2^T C\beta_2 < p_2\sigma^2.$$

Part II

Table 1 shows a set of data originally given by Longley (1967). The objective is to predict the variable y, total derived employment, as a function of six other variables $x1, ..., x6$. All the regression models include an intercept.

(a) Table 2 shows the residual sum of squares (RSS) for all possible models containing linear combinations of $x1, ..., x6$. Based on these, which model would you choose?

(You may use any method of variable selection you prefer, but be sure to indicate the rationale behind your selection.)

(b) Now consider the model dropping $x5$ but including all the other variables. (*Note:* There is no reason why this should be the same model as you selected in (a).) Table 3 shows the parameter values, standard errors, t statistics and p-values. Fig. 1 is a plot which shows the R-student (or externally studentized) residuals against (a) time, and (b) fitted values, for this model.

Based on Table 3, Fig. 1 and any other features of the data that occur to you, write a brief summary report of your conclusions. Your summary should include statistical conclusions, such as whether the model appears to fit the data well, but should also explain the implications of the analysis that might be of interest to an economist. If you had the opportunity to perform further analyses, what would you try?

x1	x2	x3	x4	x5	x6	y
83.0	234289	2356	1590	107608	1947	60323
88.5	259426	2325	1456	108632	1948	61122
88.2	258054	3682	1616	109773	1949	60171
89.5	284599	3351	1650	110929	1950	61187
96.2	328975	2099	3099	112075	1951	63221
98.1	346999	1932	3594	113270	1952	63639
99.0	365385	1870	3547	115094	1953	64989
100.0	363112	3578	3350	116219	1954	63761
101.2	397469	2904	3048	117388	1955	66019
104.6	419180	2822	2857	118734	1956	67857
108.4	442769	2936	2798	120445	1957	68169
110.8	444546	4681	2637	121950	1958	66513
112.6	482704	3813	2552	123366	1959	68655
114.2	502601	3931	2514	125368	1960	69564
115.7	518173	4806	2572	127852	1961	69331
116.9	554894	4007	2827	130081	1962	70551

Table 1. Longley's data. Variables are:
x1: Gross National Product implicit price deflator (1954=100)
x2: Gross National Product
x3: Unemployment
x4: Size of armed forces
x5: Non-institutional population 14 years of age and over
x6: Year
y: Total derived employment

Variables	RSS
x1 x2 x3 x4 x5 x6;	836424.05549
x1 x2 x3 x4 x5;	2335237.5051
x1 x2 x3 x4 x6;	841173.00360
x1 x2 x3 x5 x6;	2997329.5373
x1 x2 x4 x5 x6;	2426562.0273
x1 x3 x4 x5 x6;	942730.31445
x2 x3 x4 x5 x6;	839348.03187
x1 x2 x3 x4;	2683826.9047
x1 x2 x3 x5;	3246013.6349
x1 x2 x3 x6;	3121919.5214
x1 x2 x4 x5;	2533302.2294
x1 x2 x4 x6;	4898726.1696
x1 x2 x5 x6;	3197698.061
x1 x3 x4 x5;	3537492.3408
x1 x3 x4 x6;	1322077.3641
x1 x3 x5 x6;	3165993.016
x1 x4 x5 x6;	9519274.966
x2 x3 x4 x5;	2366597.2129
x2 x3 x4 x6;	858680.40583
x2 x3 x5 x6;	3236865.8501
x2 x4 x5 x6;	3029239.821
x3 x4 x5 x6;	985719.64799
x1 x2 x3;	3560224.0666
x1 x2 x4;	5686283.534
x1 x2 x5;	3259976.3912
x1 x2 x6;	4899207.5743
x1 x3 x4;	5510107.7078
x1 x3 x5;	4573506.7168
x1 x3 x6;	3165993.7244
x1 x4 x5;	9948802.5874
x1 x4 x6;	9537605.0942
x1 x5 x6;	9659766.8401

Table 2 (part 1). Various models and associated residual sums of squares (RSS).

Variables	RSS
x2 x3 x4;	2756711.6889
x2 x3 x5;	3482242.1172
x2 x3 x6;	3239267.6143
x2 x4 x5;	3050739.013
x2 x4 x6;	4907747.2763
x2 x5 x6;	3811970.1926
x3 x4 x5;	5619322.2252
x3 x4 x6;	1323360.7427
x3 x5 x6;	3260101.5123
x4 x5 x6;	9776262.2717
x1 x2;	5824195.1764
x1 x3;	7597740.5494
x1 x4;	10602630.171
x1 x5;	10187326.108
x1 x6;	9756466.2106
x2 x3;	3579064.9691
x2 x4;	5959487.7837
x2 x5;	3874361.4669
x2 x6;	4910943.9004
x3 x4;	81250446.738
x3 x5;	5755028.5301
x3 x6;	3272124.7031
x4 x5;	11908512.561
x4 x6;	9850233.958
x5 x6;	10062884.494
x1;	10611376.221
x2;	6036140.1661
x3;	138293297.4
x4;	146317919.47
x5;	14365926.087
x6;	10456528.953
None	185008830

Table 2 (part 2). Various models and associated residual sums of squares (RSS).

Variable	DF	Estimate	Stan. Err.	t statistic	p-value
INTERCEP	1	–3564922	772385.59420	–4.615	0.0010
X1	1	27.714878	60.74979084	0.456	0.6580
X2	1	–0.042127	0.01761875	–2.391	0.0379
X3	1	–2.103944	0.30293168	–6.945	0.0001
X4	1	–1.042377	0.20018388	–5.207	0.0004
X6	1	1869.116966	399.35328119	4.680	0.0009

Table 3. Table of parameter estimates for model containing variables x1, x2, x3, x4, x6.

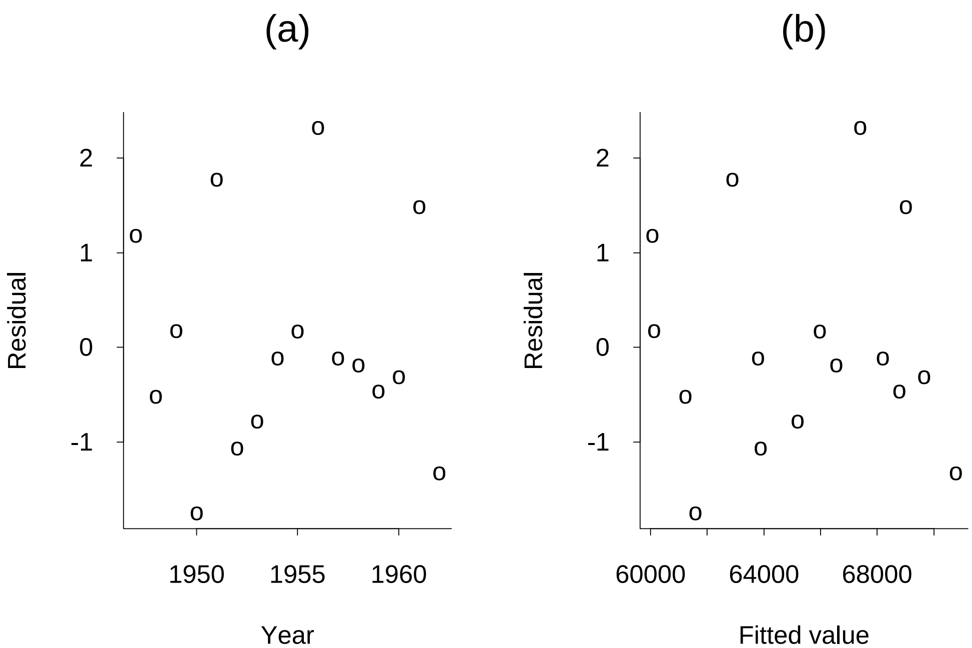

Fig. 1. Plot of R-studentized residuals against (a) year, (b) fitted values, for the model of Table 3.

Solution

Part I

(a) $\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y$ which has variance $(X_1^T X_1)^{-1} \sigma^2$ (same proof as in standard case) and mean $\beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$. If $\beta_2 \neq 0$, this is a biased estimator with bias $(X_1^T X_1)^{-1} X_1^T X_2 \beta_2$.

(b) $\hat{y} = H_1 y$ where $H_1 = X_1 (X_1^T X_1)^{-1} X_1^T$. Thus $y - \hat{y} = (I - H_1) X_2 \beta_2 + (I - H_1)\epsilon$, and

$$(y - \hat{y})^T (y - \hat{y}) = \beta_2^T X_2^T (I - H_1) X_2 \beta_2 + 2\beta_2^T X_2^T (I - H_1)\epsilon + \epsilon^T (I - H_1)\epsilon$$

(recall that the matrix H_1 is symmetric, idempotent). Taking expectations, the middle term vanishes and the last term has expectation $(n - p_1)\sigma^2$ as in standard least squares theory, so the answer is of the form given in the question, with $C = X_2^T (I - H_1) X_2$.

(c) The distribution of R is $\chi'^2_{n-p;\delta}$ (in words: the noncentral chi-squared distribution with $n - p$ degrees of freedom and noncentrality parameter δ), where $\delta = \sqrt{\beta_2^T C \beta_2}$. (For this problem the precise specification of δ may be rather difficult, but just "non-central chi-squared" is sufficient for at least partial credit, and any extra detail that is provided will earn more.)

(d) $y^* - \hat{y} = (I - H_1) X_2 \beta_2 + \epsilon^* - H_1 \epsilon$. By similar reasoning to part (b), $E\{(y^* - \hat{y})^T (y^* - \hat{y})\} = \beta_2^T C \beta_2 + (n + p_1)\sigma^2$.

(e) If we repeat the calculation of part (d) for the case when the statistician uses the full model including X_2, the mean squared prediction error is $(n + p_1 + p_2)\sigma^2$. The answer comes from comparing the two mean squared errors.

Part II

(a) If we perform backward selection then we start with the model containg all six variables (plus an intercept) and successively drop variables $x1, x5, x2, x4$. The corresponding RSS values are 836424 (9 degrees of freedom for error), 839348 (10 DF), 858680 (11 DF), 1323361 (12 DF), 3272125 (13 DF), with successive F statistics (for each model in turn as the null against its immediate predecessor as the alternative) of .031, .230, 5.95, 17.67. For example, for testing the fourth model in the sequence against the third, the F statistic is $\frac{1323361 - 858680}{1} \cdot \frac{11}{858680} = 5.95$, which is statistically significant, whereas .031 and .230 are not significant. Therefore, backward selection leads to the model containing the variables $x2, x3, x4, x6$. Other forms of model selection will be accepted provided they are backed up with appropriate details.

(b) There are actually a lot of possibilities here so what follows is meant just to indicate some of them. Credit will be given for any reasonably well-argued points. Table 3 includes $x1$ but this is not statistically significant — therefore we should presumably ignore that variable but all the rest are significant, so this is additional confirmation of the model selected in (a). For interpretation to an economist, it appears that both unemployment and enrollment in the armed services have a negative impact on total employment. The explanation is presumably that army service takes people away from regular employment (especially at times of high military activity, as during the Korean war), while we would expect unemployment to be low when employment is high and vice versa. The model suggests that GNP has a negative influence which seems contradictory, but there is also a positive time trend (the $x6$ variable) so it may be that there is collinarity between GNP and the time trend. Also, the information presented in $x1$ suggests that maybe we should actually be using GNP adjusted for inflation (i.e. the variable $x2/x1$), and if we did this we might find that the dependence on GNP is stronger (and with the right sign) than the dependence on time. From the point of view of statistical interpretation, apart from pointing out that the variable $x1$ is not statistically significant in a linear regression, it also looks from the plots in Fig. 1 that the time trend is non-linear. For possible further analyses, the above remarks suggest (i) try using $x2/x1$ in place of $x2$ to see if this gives a more satisfactory linear trend without $x6$, (ii) if there is still evidence of a non-linear time trend, try using a quadratic trend in either $x2/x1$ or $x6$.