KTH dep. of mathematics
Examiner:
Tatjana Pavlenko

Regression Analysis, sf2930[1]
Examination Mars 18, 2016

Aids: Pocket calculator

Teacher:
Harald Lang
073 9868 996

Skriv gärna lösningarna på svenska!

1. You have run a regression with eight covariates and 50 observations. The $R^2 = 0.710$. You feel that perhaps you should exclude the last two covariates from the regression, and run the equation with only the six remaining covariates. For this regression the $R^2 = 0.700$. Which is your decision? (You must of course motivate.)

2. You want to test the two grape varieties Cabernet Sauvignon and Shiraz (Syrah) for red wine. You have 24 brands of wine from Cabernet Sauvignon and 19 from Shiraz. You have a wine expert who (in random order) test all 43 brands, and give each one of the quality ratings A, B or C, where A is the excellent quality, B good quality and C mediocre quality. The outcome is as follows

	C	*B*	*A*
Cabernet Sauvignon	5	12	7
Shiraz	2	5	12

   (So five of the Cabernet Sauvignon wines get the rating C, and so on.) A homogeneity test (independent test, contingency table) results in a *p*-value 0.0833 for the null "no difference". Suggest and perform another analysis to determine if one of the grape varieties dominate the other, and give also an estimate of the difference. Use a significance level 5%. (N.B.: These data are of course complete fake!)

3. A friend was interested to see if wages are "socially inherited", in the sense that parents' wages influence the children's wages, *ceteris paribus* ("all else held equal".) She had a very large amount of observations on individual wages, education working experience and parents' wages. She ran a regression of log(wage) on parent's wages, dummies for highest university or college degree (bachelor's degree, etc.; "no degree" was benchmark), working experience and years of study at college or university.

   The coefficients came out with the sign she had expected, except for the coefficient for years of study, which to her surprise came out slightly *negative* (indicating that college studies are detrimental to wage opportunities.) To test, she computed a confidence interval for that coefficient, but also this interval was all negative. *What is your explanation for this?*

4. You run three regressions:

$$y = \hat{\beta}_0 + x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \hat{e}$$

$$x_2 = \hat{\alpha}_0 + x_1\hat{\alpha}_1 + \hat{\varepsilon}$$

$$y = \hat{\gamma}_0 + x_1\hat{\gamma}_1 + \hat{u}$$

   You know that $\hat{\beta}_2 = 1.4$, $\hat{\alpha}_1 = 1.1$, $\hat{\gamma}_1 = 2.04$, $|\hat{\varepsilon}|^2 = 2.1$ and $|\hat{u}|^2 = 5.0$. Compute $|\hat{e}|^2$.

[1] This is also a re-examination for sf2950

5. You want to run a regression where you consider one covariate, x_4, endogeneous:

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + e$$

You have two covariates, z_1 and z_2 which are canditates for instruments. You will now test each of them, and the two together, for "weak instruments".

a) What does it mean that instruments are "weak", and what is the harm if they are?

b) How do you perform the test, and how do you choose between the three combinations (z_1 alone, z_2 alone and z_1 plus z_2 together)?

6. You have run a regression

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + e$$

Now you want to calculate a confidence interval for $\beta_2 - \beta_3 - \beta_4$. How do you do that? (You don't need to describe how to obtain standard errors for individual coefficient estimates, nor the estimated covariance matrix.)

Short answers

1. The smaller model. *F*-value for the two coefficients be zero is less than one, Akaike for the smaller is 2.3 less than for the larger (for BIC the difference is 6.13.)

2. Wilcoxon's rank sum test gives *p*-value 0.0383, so there is a difference.
Estimated Pr(Shiraz better than CS) – Pr(CS better than Shiraz) = 0.34, so Shiraz is better than CS.

3. This is a *ceteris paribus* confusion: *for a given academic degree*, a longer time spent on studying is not merit.

4. Employ the *normal equations* a couple of times, and finally "Pythagoras' Theorem".

$$|\hat{e}|^2 = |\hat{u}|^2 - |\hat{\varepsilon}|^2 \hat{\beta}_2^2 = 0.884.$$

5. The instruments are not sufficiently correlated with the endogeneous variable; the bias is unacceptably large. See "Elements …" p.29–30.

6. The easiest way is to run the regression as

$$y = \beta_0 + x_1\beta_0 + (x_2 + x_3)\beta_3 + (x_2 + x_4)\beta_4 + x_2\gamma + e$$

and compute the confidence interval for γ.