KTH dep. of mathematics
Examiner:
Tatjana Pavlenko

Regression Analysis, sf2930*
Examination June 9, 2016

Aids: Pocket calculator

Teacher:
Harald Lang
073 9868 996

Skriv gärna lösningarna på svenska!

1. You want to predict the annual income for a 26 year old, white male with 16 years of schooling, and you plan to employ an OLS regression equation of log(wage) on age, age^2, race, gender and years.of.schooling for the purpose.

   A colleague points out that "years of schooling" is endogeneous – the choice to study may well depend on your wage opportunities other than those captured by the other covariates, i.e., it may depend on the size of the residual.

   What do you do?

2. You use the data set "auto.csv" to compute a prediction interval for the mileage (miles per gallon) of an American car with four cylinders, 100 cubic inches displacement, 80 horsepower and 2200 pounds weight. You run a (homoskedastic) OLS regression and get the following result:

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           28.2031290  0.3873962  72.802  < 2e-16
Japanese               2.3771648  0.6036143   3.938 9.74e-05
I(cylinders - 4)      -0.5309896  0.4036015  -1.316    0.189
I(displacement - 100)  0.0065680  0.0089950   0.730    0.466
I(horsepower - 80)    -0.0538805  0.0129440  -4.163 3.88e-05
I(weight - 2200)      -0.0048308  0.0007126  -6.779 4.55e-11

Residual standard error: 4.165 on 386 degrees of freedom
Multiple R-squared:  0.7189,    Adjusted R-squared:  0.7153
F-statistic: 197.5 on 5 and 386 DF,  p-value: < 2.2e-16
```

   Here "`I(cylinders - 4)`" is the generated variable "no. of cylinders – 4" etc, "`Japanese`" is a dummy for "made in Japan"; there are also American and European cars in the data.

   Compute a prediction interval with confidence level 95%.

3. We run an OLS regression $y_i = x_i\beta + e_i, \quad i = 1 \dots n,$ and get the estimated coefficients $\hat{\beta}.$ The predicted *y*-values are $\hat{y}_i = x_i\hat{\beta}.$ Prove that $\sum_1^n \hat{y}_i^2 \le \sum_1^n y_i^2.$

4. We use the data in auto.csv again. Now we want to estimate the probability that a car is of American origin, given the car's weight. We run a Logit of American on weight and get

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.6547433  0.7214296  -9.224   <2e-16
weight       0.0026186  0.0002761   9.484   <2e-16
```

   so a *high* weight predicts that the car is of American origin (*p*-value almost = 0.)

   Now we test displacement as a predictor:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.546779   0.664840  -8.343  < 2e-16
displacement  0.042142   0.005324   7.915 2.47e-15
```

   so *large* displacement predicts that the car is of American origin (*p*-value almost = 0.)

---

* This is also a re-examination for sf2950

Now we employ both weight and displacement as predictors:

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9135068  1.0312016  -1.856   0.0635 .
weight       -0.0037925  0.0008543  -4.439 9.03e-06
displacement  0.0886811  0.0133117   6.662 2.70e-11
```

But now the coefficient for weight is *negative*, indicating that a *low* weight predicts that the car is of American origin! What went wrong? Explain!

5. In *Wine Economics and Policy* 3 (2014), Snipes and Taylor regressed the rating of wine on its price and eight dummies for various grapes (so there were in all nine varieties of grapes). The Akaike value for this regression was AIC = 1109.10.

   They also ran a regression where they added in dummies for eight regions (so there were in all nine regions: California, France, Italy …) The Akaike value for this regression was AIC = 1112.51. There were 197 observations in these regressions.

   Compute the effect size η^2 of "region" on rating (when grape is constant.) (In some literature η^2 is labelled "partial-η^2.")

6. Sociologists often conduct experiments to investigate the relationship between socioeconomic status and college performance. Socionomic status is generally partitioned into three groups: lower class, middle class, and upper class. Consider the problem of comparing the mean grade point average of those college freshmen associated with lower class, those associated with middle class, and those associated with upper class. The grade point averages (GPAs) for random samples of seven college freshmen associated with each of the three socioeconomic classes were selected from a university's files at the end of the academic year. The data are recorded below.

Grade point averages		
Lower class	middle class	upper class
2.9	3.2	2.3
2.2	3.5	3.1
3.1	2.8	2.4
2.5	3.8	2.5
1.8	3.0	3.3
3.0	3.5	2.8
2.2	3.0	1.4

   Do the data provide sufficient evidence to indicate a difference among the freshmen GPAs for the three socioeconomic classes?

Short answers:

1. You should go on with the regression as planned, and tell your colleague that endogeneity is not an issue when the equation is used for prediction. The problem with endogeneity arises when the equation is given a structural interpretation.

2. (cf "more exercises", 10) The prediction is the coefficient for the intercept. The standard error of the prediction is
   $\sqrt{0.3874^2 + 4.165^2} \approx 4.18$. $F_{0.05}(1, 386) = 3.866$, so the prediction interval is
   $28.2 \pm 4.18\sqrt{3.866} = 28.2 \pm 8.2.$

3. We have $y = \hat{y} + \hat{e},$ and $\hat{y} \perp \hat{e}$ (the "normal equations".) Hence, by Pythagoras' theorem,
   $|y|^2 = |\hat{y}|^2 + |\hat{e}|^2 \geq |\hat{y}|^2.$

4. (cf "more exercises", 15) Nothing went wrong. American cars are heavier and have more displacement that European and Japanese cars, on average. However, if we compare cars of equal displacement, American cars are in fact lighter. Indeed, the ratio pounds/cub.inch is 14.9 for American cars, and 22.2 for non-American cars.

5. Manipulating the formulae for AIC and η^2 gives
   $$\text{AIC} - \text{AIC}_* = n\ln(1-\eta^2) + 2r,$$
   where AIC is for the larger model, and AIC_* for the smaller model; *n* is the number of observations, and *r* the number of restrictions (coefficients set to zero.) We conclude that $\eta^2 \approx 0.062.$

6. (cf "more exercises", 17) This is a one-way ANOVA with no interactions. The easiest way to perform the calculations is as is outlined on p.36 in the booklet. The residuals when the responses are regressed on dummies for the social classes are the response minus the average of the responses of the groups. We get (with obvious notation) $|\hat{e}|^2 = 4.53.$ The residuals when the response is regressed on an intercept is the response minus the grand mean. We get $|\hat{e}_*|^2 = 6.96.$ Hence
   $$F = \frac{R^2}{1-R^2}\frac{18}{2} = \frac{6.96 - 4.53}{4.53}\frac{18}{2} = 4.83.$$
   Under the null, this is an observation of an F(2, 18), variable, so the *p*-value for the null is 0.021. It seems reasonable to reject the null, and accept that there are differences.