Lecture notes VI

Multiple regression continued...

SW ch. 6-7.

Multicollinearity:

The example at the end of the previous lecture notes shows that multicollinearity implies that one variable can be eliminated.

Instead of eliminating "Man" we could have eliminated "Woman":

$Woman_i = 1 - Man_i$ ⇒

$$Y_i = \beta_0 + \beta_1 Man + \beta_2 (1 - Man_i) + u_i$$

$$= \beta_0 + \beta_2 + (\beta_1 - \beta_2) Man_i + u_i$$

Or we could have eliminated the constant term:

$$Y_i = \beta_0 (Man_i + Woman_i) + \beta_1 Man_i + \beta_2 Woman_i + u_i$$

$$= (\beta_0 + \beta_1) Man_i + (\beta_0 + \beta_2) Woman_i + u_i$$

If we exclude "Man" (or "Woman"), the category "Man" ("Woman") can be seen as the basis or reference category. The parameter of the included category "Woman" ("Man") can then be interpreted as the effect of being a "Woman" ("Man") relative to the basis category.

Generally, let there be M mutually exclusive categories for some variables (e.g. "Style" $\Rightarrow M = 19$).

Let Z_{ji} be one if unit i is in category j, $j = 1, \ldots, M$

$$Z_{ji} = \begin{cases} 1 & \text{if } i \text{ is in category } j \\ 0 & \text{else} \end{cases}$$

with $j = 1, \ldots, M$.

Then $Z_{1i} + Z_{2i} + \cdots + Z_{Mi} = 1$, since i can only be in one of the M categories.

Assume the regression model is

$$Y_i = \beta_0 + \sum_{j=1}^{M} \beta_j Z_{ji} + \sum_{k=1}^{r} \lambda_k X_{ki} + u_i$$

where $X_{1i}, \ldots, X_{ri}$ are other regressors.

Since $\sum_{j=1}^{M} Z_{ji} = 1$ we can exclude

e.g. $j = 1$, and use this as basis

category: $Z_{1i} = 1 - \sum_{j=2}^{M} Z_{ji}$ ⇒

$$Y_i = \beta_0 + \beta_1 \left(1 - \sum_{j=2}^{M} Z_{ji}\right) + \sum_{j=2}^{M} \beta_j Z_{ji} + \sum_{k=1}^{r} \lambda_k X_{ki} + u_i$$

$$= \beta_0 + \beta_1 + \sum_{j=2}^{M} (\beta_j - \beta_1) Z_{ji} + \sum_{k=1}^{r} \lambda_k X_{ki} + u_i$$

Thus: the "effect" of category j is $\beta_j - \beta_1$, which is relative to the excluded variable Z_{1i}.

Perfect multicollinearity is not restricted to categorical variables. E.g. from the national accounts:

$$\underbrace{GDP}_{\text{gross domestic product}} = \underbrace{C}_{=\text{consumption}} + \underbrace{J}_{=\text{gross investments}} + \underbrace{A}_{=\text{exports}} - \underbrace{B}_{=\text{imports}}$$

Thus cannot include all variables GDP, C, J, A and B in a linear regression.

In general, if

$$Z_{1i} = \sum_{j=2}^{m} a_j Z_{ji}, \quad \text{with } a_j \neq 0,\ j = 2, \dots, m$$

then Z_{1i} can be excluded from the regression:

$$Y_i = \beta_0 + \beta_1 \left(\sum_{j=2}^{m} a_j Z_{ji} \right) + \sum_{j=2}^{m} \beta_j Z_{ji} + \sum_{k=1}^{r} \beta_k X_{ki} + u_i$$

$$= \beta_0 + \sum_{j=2}^{m} \left(\beta_1 a_j + \beta_j \right) Z_{ji} + \sum_{k=1}^{r} \lambda_k X_{ki} + u_i$$

Cannot estimate β_j — only $\beta_1 a_j + \beta_j$

V

Imperfect multicollinearity means that some regressors are highly correlated, but not perfectly linear in their relationship.

That is, there exist some variable Z_{1i} and some linear combination $\sum_{j=2}^{m} a_j Z_{ji}$ — for constants $a_j \neq 0$ — such that $Corr(Z_{1i}, \sum_{j=2}^{m} a_j Z_{ji}) \approx 1$

Measures of fit in Multiple Regression:

Assume homoscedasticity:

$$Var(u_i | X_{1i}, \ldots, X_{ki}) = \sigma_u^2,$$

i.e. the variance of the error term, u_i, is independent of the regressors.

Then an unbiased and consistent estimator of σ_u^2 is:

$$\hat{\sigma}_u^2 = \frac{SSR}{n-k-1}, \text{ where } SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

$\hat{\sigma}_u$ is called the standard error of regression (SER).

The R^2:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

Same definition as in univariate regression, since it still holds that:

$$TSS = ESS + SSR$$

$$\Downarrow$$

$$\underbrace{\sum_i (Y_i - \bar{Y})^2}_{TSS} = \underbrace{\sum_i (\hat{Y}_i - \bar{Y})^2}_{ESS} + \underbrace{\sum_i \hat{u}_i^2}_{SSR}$$

But R^2 increases whenever k increases, - unless the estimated regression coeff. $\hat{\beta}_k = 0$

$\Rightarrow$ cannot use R^2 to compare the fit of different models directly

The adjusted R^2 : $\bar{R}^2$

$$\bar{R}^2 = 1 - \underbrace{\frac{n-1}{n-k-1}}_{(>1)} \frac{SSR}{TSS} \quad \left(< R^2 \right)$$

When k increases, the term $\frac{n-1}{n-k-1}$ increases, while SSR decreases. Thus $\bar{R}^2$ may decrease even if k increases.

The term $\frac{n-1}{n-k-1}$ is a penalty term, which penalizes using too many regressors.

$\bar{R}^2$ is often used as a criterion to choose between different models (i.e. different sets of regressors).

Still, there is no guarantee that:

- non-significant variables are included
- significant ones are excluded, leading maybe to omitted variable bias
- the estimated parameters do not reflect true causality.

Hypothesis tests and confidence intervals (SW ch. 7)

In large samples the OLS estimator $\hat{\beta}_j$, $j = 0, 1, \ldots, k$ will be approximately $N(\beta_j, SE(\hat{\beta}_j)^2)$ - distributed

Thus

$$t = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim N(0,1)$$

when n is large.

95% Confidence intervals: $\left[\hat{\beta}_j \pm 1.96\, SE(\hat{\beta}_j)\right]$

90% Confidence intervals: $\left[\hat{\beta}_j \pm 1.64\, SE(\hat{\beta}_j)\right]$

Two-sided hypothesis of a single coefficient:

$$H_0: \beta_j = \beta_j^0 \quad \text{vs} \quad H_1: \beta_j \neq \beta_j^0$$

$$t = \frac{\hat{\beta}_j - \beta_j^0}{SE(\hat{\beta}_j)}$$

$$\text{p-value} = \Pr(|t| \geq |t^{obs}|)$$

$$= 2\Phi(-|t^{obs}|)$$

Reject if p-value < level of significance, e.g. 0.05

Joint hypothesis

Restrictions involving more than one parameter

Generally, let q denote the number of restrictions

Example 1:

$H_0: \beta_1 = \beta_2$ vs $H_1: \beta_1 \neq \beta_2$

Here $q = 1$ (but H_0 involves 2 parameters)

Example 2:

$H_0: \beta_1 = 0$ and $\beta_2 = 0$ vs $\beta_1 \neq 0$ or $\beta_2 \neq 0$

Here $q = 2$ (two restrictions and two param.)

Example 3:

$H_0: \beta_1 = \beta_1^0, \beta_2 = \beta_2^0, \ldots, \beta_m = \beta_m^0$

H_0: At least one of the restrictions in H_0 does not hold

Here $q = m$.

Example 4:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_m$

H_1: At least one of the equalities in H_0 do not hold

Here $q = m - 1$

To test a joint hypothesis, we can use the F-statistic:

Define:

$SSR_{restricted}$ = SSR under H_0

$SSR_{unrestricted}$ = SSR under H_1

Assuming homoscedastic errors:

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{\hat{\sigma}^2}$$

where $\hat{\sigma}^2 = \frac{SSR_{unrestricted}}{n - k_{unrestricted} - 1}$

When $n \to \infty$, then F is distributed as $F_{q,\infty}$, which is tabulated in the Appendix.

p-value $= Pr(F_{q,\infty} \geq F^{obs})$ where F^{obs} is the observed value of the F-statistic in the data.

UNIVERSITY OF OSLO

DEPARTMENT OF ECONOMICS

Exam: ECON4135 - Applied statistics and econometrics, fall 2006

Date of exam: 29 November 2006

Time for exam: 2.30 p.m. – 5.30 p.m.

The problem set covers 7 pages (including STATA printouts)

Resources allowed:
- All written and printed resources, as well as calculators, are allowed

Grades given: A (best), B, C, D, E and F, with E as the weakest passing grade.

SkatteFUNN is a system for tax deduction of Research and Development expenses (R&D). It became law in 2003 and applies to any Norwegian enterprise. If a tax deduction is granted to a firm, it gives the right to a 20 percent tax deduction of R&D, limited upwards to 800,000 NOK for one year. If the firm is not in a tax position, i.e., its payable tax is zero, the R&D subsidy is paid out in cash.

To obtain the R&D subsidy through SkatteFUNN, certain formal requirements must be fulfilled: The project must have the character of R&D, and specifically have "a goal to acquire new knowledge, information or experience which is assumed to be useful for the firm with regard to developing new products, services or production processes".[1] However, there is no substantive evaluation of the project or its chances of success. The application process takes just two weeks. In 2004, 70% of the applications were accepted: 6073 projects with total tax deductions of 1.6 billions.

SkatteFUNN is popular among businesses because of its simple application and reporting procedures. However, critiques of the scheme claim that this is a waste of the taxpayer's money. While SkatteFUNN is meant to generate more R&D and change in R&D behaviour in enterprises, it may not be effective, either because the firms would undertake the R&D projects anyway, or because the system gives an incentive for creative bookkeeping, whereby ordinary operating expenses are classified as R&D.

[1] The Norwegian Research Council's SkatteFUNN report, 2002

We will examine the following problem: What characterizes the firms that obtain R&D subsidies through SkatteFUNN? In this problem set we will look at data for 2004. Our sample consists of (almost all) joint stock companies (i.e., they are not personally owned) in the manufacturing sector. The sample consists of 3716 firms, of which 583 obtained the R&D subsidy in 2004. We will analyse the following variables:

Table 1. Variable list

Variable name:	Description:
RDsubsidy	R&D subsidy from SkatteFUNN, in 1000 NOK
y	Dummy (binary): =1 if *RDsubsidy* > 0 (and =0 if *RDsubsidy*=0)
taxposition	Dummy: =1 if the firm has positive payable tax (prior to any tax deduction)
share_high	Share of the firm's workers with at least 17 years of schooling
*VA_empl**	Value added per employee in 1000 NOK
firmage_10y	Dummy: =1 if the firm's age is 10 years or less
emply	Number of employees
empl1	Dummy: =1 if $emply \leq 5$
empl2	Dummy: =1 if $5 < emply \leq 10$
empl3	Dummy: =1 if $10 < emply \leq 50$
empl4	Dummy: =1 if $50 < emply \leq 100$
empl5	Dummy: =1 if $emply > 100$

* Value added is defined as sales less intermediate input, a measure of the productive contribution of labor and capital in the firm. It is measured here as the sum of operating profits (the reward to capital owners) and wage costs (the reward to the workers). Value added per employee is often termed labor productivity.

R1 below presents some summary statistics for these variables. Summary statistics for the sub-sample consisting of the 583 firms that obtained the R&D subsidy, i.e., with $y = 1$, is shown in **R2**. A detailed summary for the variable *RDsubsidy* is shown in **R3**.

1. It was decided to estimate the following linear regression model using Ordinary Least Squares (OLS):

$$RDsubsidy = \beta_0 + \beta_1 taxposition + \beta_2 share\_high + \beta_3 VA\_empl + \beta_4 firmage\_10y + \beta_5 emply + u,$$

where u is the error term. What assumptions must be satisfied to obtain estimators which are (i) unbiased and (ii) asymptotically normally distributed when N, the number of firms, tend to infinity? The results from the estimation are reported in **R4**. To try to improve the fit of the model, it was decided to replace the continuous variable *emply* with the dummy variables *empl2*,....., *empl5*. The results are reported in **R5**. Which of the models, **R4** or **R5**, appears to be the better? How would you describe the relation between employment and R&D subsidy? What would happen if the dummy variable *empl1* is also included in the regression?

2. Give 99% confidence intervals for the parameters of the variables *VA_empl*, *share_high* and *firmage_10y*, based on the results in **R5**. What does it mean that the degree of confidence is 99%? Which of these variables are significant determinants of the dependent variable *RDsubsidy* at the 1% level of significance? Give a brief interpretation of your findings with regard to these three variables.

3. Let us now turn to the variable *taxposition*, which takes the value one if the firm is in tax position in 2004, i.e., if it pays tax, and zero if not. An opponent of SkatteFUNN has the following claim:

   "SkatteFUNN does not attract firms with a potential for R&D, but rather (a disproportionate number of) firms with cash flow problems. Hence the system is a failure."

   Can you formulate a relevant statistical hypothesis to investigate this claim (or at least a part of it)? Set up a test statistic and calculate the p-value of the test. What does the p-value say? It appears that there is a very significant, negative relation between tax position and R&D subsidy. Can *you* suggest an explanation of this finding?

4. Let us now, for a moment, turn away from our empirical model and consider the general regression model

   $$Y = \alpha + \beta X + u\,.$$

   What does it mean that X is exogenous in this model? Assume now that X is *endogenous*. State formally the requirements for a variable Z to be a valid instrumental variable (IV) for X.

5. The R&D subsidy has a direct effect on a firm's operating profits, because (in the accounts) the subsidy is considered as a reduction in operating costs, typically wage costs. Is it then plausible to assume that *VA_empl* is an exogenous regressor? What type of bias do you think endogeneity will lead to in this case (positive or negative)? Discuss briefly whether there could be endogeneity problems also with regard to some of the other regressors.

6. Because of concern about potential bias due to endogeneity of *VA_empl*, it was proposed to use instrumental variable estimation (2SLS), with value added per employee in 2003 (termed *VA_empl_2003*) as an instrument. Do you think this is a reasonable choice of instrument? The results are shown in **R6**. Is *VA_empl* a significant determinant of *RDsubsidy* according to the 2SLS estimator?

7. A detailed summary of the sample distribution of the dependent variable *RDsubsidy* is shown in **R3**. Does the distribution of *RDsubsidy* look like a normal distribution? Do you think a linear regression model is appropriate, in view **R3**? An alternative model would be a binary regression model with the binary variable y as the dependent variable. The results from a logit model with the same regressors as in the previous models, except that *VA_empl* was replaced with *VA_empl_2003*, are shown in **R7**. Do you have to modify any of the main conclusions about the relation between SkatteFUNN and the explanatory variables in view of the new results?

8. The estimated probability that $y = 1$ is equal to 0.52 for a firm with the following values of the regressors: *taxposition=0, share_high=.03, VA_empl_2003=507, firmage_10y=0, empl4=1.* What is, *ceteris paribus*, the change in the probability that $y = 1$ when the variable *taxposition* changes from 0 to 1 ?

R1: Summary statistics

```
. summarize  RDsubsidy y  taxposition VA_empl firmage_10y  emply empl2 empl3
empl4 empl5
```

Variable	Obs	Mean	Std. Dev.	Min	Max
RDsubsidy	3716	60.93126	175.3727	0	800
y	3716	.1568891	.3637452	0	1
taxposition	3716	.6324004	.4822164	0	1
share_high	3716	.0249045	.1049061	0	1
VA_empl	3716	453.8682	1010.88	-24250	31211
firmage_10y	3716	.4477933	.4973339	0	1
emply	3716	32.07374	170.4802	1	3451
empl2	3716	.2341227	.4235062	0	1
empl3	3716	.3167384	.4652671	0	1
empl4	3716	.0497847	.2175291	0	1
empl5	3716	.0468245	.2112913	0	1

R2: Summary statistics for the firms which obtained R&D subsidy

```
. summarize  RDsubsidy y  taxposition VA_empl firmage_10y  emply empl2 empl3
empl4 empl5
```

Variable	Obs	Mean	Std. Dev.	Min	Max
RDsubsidy	583	388.3715	262.547	1.068	800
y	583	1	0	1	1
taxposition	583	.5231561	.4998924	0	1
share_high	583	.0552841	.1257722	0	1
VA_empl	583	580.1343	1842.422	-2178	31211
firmage_10y	583	.4511149	.4980318	0	1
emply	583	60.93139	153.2501	1	1960
empl2	583	.1423671	.3497263	0	1
empl3	583	.4768439	.4998924	0	1
empl4	583	.1492281	.3566192	0	1
empl5	583	.1114923	.3150111	0	1

R3. Summary of *RDsubsidy*

```
. summarize  RDsubsidy,detail
```

RDsubsidy

Percentiles		Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	3716
25%	0	0	Sum of Wgt.	3716
50%	0		Mean	59.98739
		Largest	Std. Dev.	174.078
75%	0	800		
90%	221.192	800	Variance	30303.16
95%	534.539	800	Skewness	3.156037
99%	800	800	Kurtosis	12.06535

R4. Linear regression 1

```
. regress    RDsubsidy  taxposition  share_high VA_empl firmage_10y emply,
robust

Linear regression                                      Number of obs =    3716
                                                       F(  5,  3710) =   13.93
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0451
                                                       Root MSE      =  171.49

------------------------------------------------------------------------------
             |               Robust
   RDsubsidy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 taxposition |  -35.05432    6.25826    -5.60   0.000    -47.32429   -22.78435
  share_high |   235.9143   46.35583     5.09   0.000     145.0289    326.7997
     VA_empl |   .0131114   .0046915     2.79   0.005     .0039133    .0223096
 firmage_10y |   5.864701   5.695922     1.03   0.303    -5.302745    17.03215
       emply |   .0859267   .0345811     2.48   0.013     .0181269    .1537265
       _cons |   65.89127   5.792417    11.38   0.000     54.53464    77.24791
------------------------------------------------------------------------------
```

R5. Linear regression 2

```
. regress    RDsubsidy  taxposition  share_high VA_empl firmage_10y  empl2 empl3
> empl4 empl5, robust

Linear regression                                      Number of obs =    3716
                                                       F(  8,  3707) =   39.93
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1390
                                                       Root MSE      =   162.9

------------------------------------------------------------------------------
             |               Robust
   RDsubsidy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 taxposition |  -31.00485   6.011071    -5.16   0.000    -42.79018   -19.21952
  share_high |   246.7327   43.00345     5.74   0.000     162.4199    331.0454
     VA_empl |   .0121639   .0050599     2.40   0.016     .0022434    .0220844
 firmage_10y |   21.85446   5.492125     3.98   0.000     11.08658    32.62235
       empl2 |   20.70427   4.735327     4.37   0.000     11.42017    29.98837
       empl3 |   76.01872   6.347054    11.98   0.000     63.57466    88.46278
       empl4 |    204.878   21.73546     9.43   0.000     162.2634    247.4926
       empl5 |   172.2839   22.57573     7.63   0.000     128.0218    216.5459
       _cons |   11.89461   5.312467     2.24   0.025     1.478964    22.31025
------------------------------------------------------------------------------
```

R6. Instrumental variable estimation (2SLS)

```
. ivreg    RDsubsidy  taxposition (VA_empl=VA_empl_2003) firmage_10y  empl2
empl3 empl4 empl5, robust

Instrumental variables (2SLS) regression          Number of obs =    3713
                                                  F(  7,  3705) =   38.80
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.1090
                                                  Root MSE      =  165.76

------------------------------------------------------------------------------
             |               Robust
   RDsubsidy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     VA_empl |  -.0035754   .0134786    -0.27   0.791    -.0300017    .0228508
 taxposition |  -32.93748   6.226047    -5.29   0.000     -45.1443   -20.73067
 firmage_10y |    23.0984   5.607941     4.12   0.000     12.10345    34.09336
       empl2 |   15.50027   5.087173     3.05   0.002     5.526332     25.4742
       empl3 |   71.64897   6.440869    11.12   0.000     59.02098    84.27697
       empl4 |   204.4647   22.07005     9.26   0.000     161.1941    247.7353
       empl5 |   175.5888   22.82496     7.69   0.000      130.838    220.3395
       _cons |   28.35262   7.793745     3.64   0.000     13.07217    43.63307
------------------------------------------------------------------------------
Instrumented:  VA_empl
Instruments:   taxposition firmage_10y empl2 empl3 empl4 empl5 VA_empl_2003
------------------------------------------------------------------------------
```

R.7 Logistic regression (logit)

```
Logistic regression                               Number of obs   =       3713
                                                  Wald chi2(8)    =     393.67
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -1379.2434                 Pseudo R2       =     0.1454

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 taxposition |  -.5005766   .0998658    -5.01   0.000      -.69631   -.3048433
  share_high |   3.071107     .37473     8.20   0.000     2.336649    3.805564
VA_empl_2003 |   .0000123   .0001394     0.09   0.930    -.0002608    .0002855
 firmage_10y |   .3497588   .0988605     3.54   0.000     .1559958    .5435218
       empl2 |   .8730689   .1825514     4.78   0.000     .5152747    1.230863
       empl3 |   2.009296   .1575086    12.76   0.000     1.700585    2.318008
       empl4 |   3.064406   .2071933    14.79   0.000     2.658315    3.470497
       empl5 |   2.536805    .217765    11.65   0.000     2.109994    2.963617
       _cons |  -3.073021   .1679754   -18.29   0.000    -3.402247   -2.743795
------------------------------------------------------------------------------
```