Lecture notes to Stock and Watson chapter 6

Multiple linear regression

Tore Schweder

September 2009

Example: desired number of children by age and civil status

Does desired number of children increase by age when controlling for civil status? Data from the class:

Not married

	22	23	24	25	26	28	29	37	38
0	0	2	1	0	1	0	0	0	0
1	0	0	0	1	1	1	0	0	0
2	2	1	4	2	2	0	0	0	1
3	0	2	1	1	0	1	1	0	0
4	1	0	0	0	1	0	0	0	0
5	0	0	0	0	0	0	0	0	0

Married

	22	23	24	25	26	28	29	37	38
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0	0

Figure: Number of students by number of desired children Y (row), Age X_1 (column) and civil status X_2 (panel), 2009.

Data in 2008

non-married

	19	20	21	22	23	24	25	26	28	34
0	1	0	0	0	1	1	2	0	1	0
1	0	0	0	1	1	0	1	0	0	0
2	0	2	0	7	1	2	3	1	0	0
3	0	0	2	0	2	1	0	2	1	0
5	0	0	0	0	0	0	0	0	0	0

married

	19	20	21	22	23	24	25	26	28	34
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	1	1
3	0	0	0	0	0	0	0	1	0	0
5	0	0	0	0	1	0	0	0	0	0

Example cont.

Structural model:

$$\begin{aligned} E\left[Children|Age = x_1, Status = x_2\right] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ E\left[Y|X_1 = x_1, X_2 = x_2\right] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \end{aligned}$$

X_2 is a "dummy" variable, $X_2 = 1$ for married students, $X_2 = 0$ for unmarried students.
Parallel regression:

$$E\left[Children|Age = x_1, Status = x_2\right] = \begin{array}{ll} \beta_0 + \beta_1 x_1 & x_2 = 0 \\ (\beta_0 + \beta_2) + \beta_1 x_1 & x_2 = 1 \end{array}$$

OVB when Married is excluded?

. gen Age25=Age-25
. reg Children Age25, robust

```
Linear regression                               Number of obs  =      30
                                                F(  1,    28)  =    0.27
                                                Prob > F       =  0.6085
                                                R-squared      =  0.0039
                                                Root MSE       =  1.2435

                    Robust
Children       Coef.  Std. Err.      t      P>t    [95% Conf.     Interval]

Age25     .0205963     .0397526      0.52   0.608   -.0608332      .1020258
_cons     2.124408     .2344238      9.06   0.000   1.644213       2.604604
```

. reg Children Age25 Married, robust

```
Linear regression                               Number of obs  =      30
                                                F(  2,    27)  =    4.20
                                                Prob > F       =  0.0258
                                                R-squared      =  0.1829
                                                Root MSE       =  1.1469

                    Robust
Children       Coef.  Std. Err.      t      P>t    [95% Conf.     Interval]

Age25          -.0175595   .0401834   -0.44   0.666   -.1000091      .0648901
Married        1.760284    .6273679    2.81   0.009   .4730319       3.047537
_cons          1.964914    .2257845    8.70   0.000   1.501643       2.428186
```

Example cont.

Regressing desired number of children on 5 regressors

. reg Children Age25 Married Siblings Gender Norwegian, robust

```
Linear regression                               Number of obs  =        30
                                                F(  5,    24)  =      3.67
                                                Prob > F       =    0.0131
                                                R-squared      =    0.4194
                                                Root MSE       =    1.0254

                             Robust
Children        Coef.   Std. Err.      t    P>t     [95% Conf. Interval]

Age25        -.017243   .0418212    -0.41   0.684    -.1035577   .0690718
Married      1.771837   .7071885     2.51   0.019     .3122716   3.231402
Siblings     .3907208   .1299175     3.01   0.006     .1225843   .6588573
Gender       .0860467   .4255109     0.20   0.841    -.7921646   .9642579
Norwegian    -.4128417  .4371262    -0.94   0.354    -1.315026   .4893424
_cons        1.410037   .4794367     2.94   0.007     .4205281   2.399545
```

Multiple linear regression model

- Data: Observed values on response variable Y and explanatory variables $X_1, \cdots, X_k$ on n sampling units
- Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i$
  1. $E[u_i | X_{1i}, \cdots, X_{ki}] = 0 \Leftrightarrow E[Y_i | X_{1i}, \cdots, X_{ki}] = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
  2. $(Y_i, X_{1i}, \cdots, X_{ki}) \quad i = 1, \cdots, n$ is a random sample from a population
  3. Large outliers are unlikely: all variables have finite fourth moments
  4. No multicollinearity: $R^2 < 1$ when X_j is regressed on the other explanatory variables $j = 1, \cdots, k$
- Comments (relative to key concept 6.4)
  1. is the structural regression model
  2. $(Y_i, X_{1i}, \cdots, X_{ki}) \quad i = 1, \cdots, n$ are independent and identically distributed (iid)
  3. (when the fourth moment of X_j is zero, there is no variability in X_j!)
  4. SW require there is no perfect colinearity - to ensure identifiability of all regression coefficients

OLS regression

$$SS\left(b_0, b_1, \cdots, b_k\right) = \sum_{i=1}^{n} \left(Y_i - \left(b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}\right)\right)^2$$

is the sum of squared differences between observed responses and corresponding values on the candidate regression plane
$y = b_0 + b_1 x_1 + \cdots + b_k x_k$

- $\left(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k\right) = \arg\min SS\left(b_0, b_1, \cdots, b_k\right)$
- $\left(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k\right)$ is approximately multinormally distributed around $\left(\beta_0, \beta_1, \cdots, \beta_k\right)$
- Linear parameters $\theta = \sum_{j=0}^{k} a_j \beta_j$ are estimated by $\widehat{\theta} = \sum_{j=0}^{k} a_j \widehat{\beta}_j$ which are approximately $N\left(\theta, SE\left(\widehat{\theta}\right)^2\right)$
- $SE\left(\widehat{\theta}\right)$ depends on the a-coefficients and the estimated variances and covariances of the regression estimates. It is estimated by Stata under homoskedacticity, or as heteroskedasticity-robust

Measures of fit

Fitted values: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \cdots + \widehat{\beta}_k X_{ki} \quad i = 1, \cdots, n$
(Empirical) residuals: $\widehat{u}_i = Y_i - \widehat{Y}_i$

- $SER = s_{\widehat{u}}, \quad s^2_{\widehat{u}} = \frac{1}{n-k-1} \sum_{i=1}^n \widehat{u}_i^2 = \frac{1}{n-k-1} SS\left(\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_k\right)$
  - Homoscedasticity, $var\left[Y_i | X_{1i}, \cdots, X_{ki}\right] = \sigma^2 \Rightarrow E s^2_{\widehat{u}} = \sigma^2$
- $R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$ is the ratio of explained to total sums of squares
  - $TSS = \sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^n \left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^n \left(\widehat{Y}_i - \overline{Y}\right)^2 = SSR + ESS$
  - R^2 cannot decrease when a regressor is added.
- Adjusted $R^2 = \overline{R}^2 = 1 - \frac{n-1}{n-k-1}\left(1 - R^2\right) = 1 - \frac{s^2_{\widehat{u}}}{s^2_Y}$
  - Fluctuates around 0 when there is no regression effect, $E\left[Y_i | X_{1i}, \cdots, X_{ki}\right] = \beta_0$
  - Tend to Increase when a new regressor X_j with $|\beta_j| > 0$ is added, and decrease when $\beta_j = 0$.

Potential problems

- Omitted variable bias
  - One regressor X_1, but correct structural model $E[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - $E[X_2|X_1] = \gamma_0 + \gamma_1 X_1 \Rightarrow$

$$\begin{aligned} E[Y|X_1] &= E[E[Y|X_1, X_2]|X_1] \\ &= E[\beta_0 + \beta_1 X_1 + \beta_2 X_2|X_1] = \beta_0 + \beta_1 X_1 + \beta_2 (\gamma_0 + \gamma_1 X_1 \\ &= (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) X_1 \\ &= \beta_0' + \beta_1' X_1 \end{aligned}$$

  - When X_2 is omitted one estimates $\left(\beta_0', \beta_1'\right)$, not (β_0, β_1)!
- Multicollinearity
  - avoid the dummy variable trap for nominal (categorical) covariates
  - solve by excluding regressors linearly dependent on other regressors
- Near multicollinearity $\Rightarrow$ some/all $\widehat{\beta}_j$ are highly correlated and have inflated SE

Example cont.

adding a correlated explanatory variable to the regression

```
. gen S2=Siblings^2
. reg  Children Age25  Married  Siblings Gender Norwegian  S2, robust

Linear regression                               Number of obs =      30
                                                F(  6,    23) =    3.29
                                                Prob > F      =  0.0173
                                                R-squared     =  0.4223
                                                Root MSE      =  1.0449

                             Robust
Children        Coef.   Std. Err.      t    P>t     [95% Conf. Interval]

Age25          -.01448   .0431835    -0.34   0.740    -.1038119    .0748519
Married        1.767208   .7180732     2.46   0.022     .2817601    3.252655
Siblings       .4956215   .3271087     1.52   0.143    -.1810544    1.172297
Gender         .1065104   .4359644     0.24   0.809    -.7953508    1.008372
Norwegian      -.4210967  .4423494    -0.95   0.351    -1.336166    .4939727
S2             -.0224536  .0497008    -0.45   0.656    -.1252675    .0803604
_cons          1.323851   .5754465     2.30   0.031     .1334489    2.514253
```

Example cont.

. corr Age25 Gender Married Norwegian Siblings S2

(obs=30)

	Age25	Gender	Married	Norweg~n	Siblings	S2
Age25	1.0000					
Gender	-0.1746	1.0000				
Married	0.2643	- 0.2075	1.0000			
Norwegian	-0.1153	-0.0847	-0.0454	1.0000		
Siblings	-0.0359	0.1442	-0.0086	-0.1787	1.0000	
S2	0.0237	0.1857	-0.0115	-0.1984	0.9026	1.0000

Problems to be done in class

- SW: 6.6
- SW: 6.11