6. Regression with Qualitative Information

Qualitative information: "family owns a car or not", "a person smokes or not", "firm is in bankruptcy or not", "industry of a firm", "gender of a person", etc

Some of the above examples can be both in a role of background (independent) variable or dependent variable.

Technically these kinds of variables are coded by a binary values (1 = "yes") (0 = "no"). In Econometrics these variables are called generally called dummy variables.

Remark 6.1 Usual practice is to denote the dummy variable by the name of one of the categories. For example, instead of using gender one can define the variable e.g. as female, which equals 1 if the gender is female and 0 if male.

6.1 Single Dummy Independent Variable

Dummy variables can be incorporated into a regression model as any other variables.

Consider the simple regression

$$y = \beta_0 + \delta_0 D + \beta_1 x + u, \tag{1}$$

where $D = 1$ if individual has the property and $D = 0$ otherwise, and $\mathbb{E}[u|D, x] = 0$. Parameter δ_0 indicated the difference with respect to the reference group ($D = 0$), for which the parameter is β_0.

Then

$$\delta_0 = \mathbb{E}[y|D = 1, x] - \mathbb{E}[y|D = 0, x]. \tag{2}$$

The value of x is same in both expectations, thus the difference is only due to the property D.

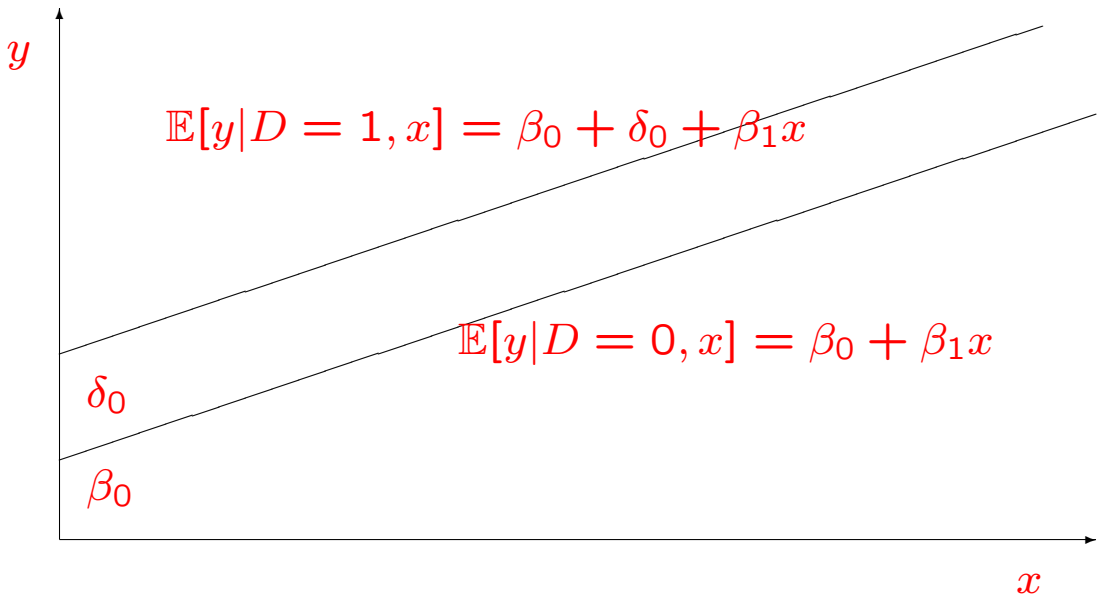

Figure 6.1: $\mathbb{E}[y|D,x] = \beta_0 + \delta_0 D + x + u$, $\delta_0 > 0$.

The category with $D = 0$ makes the reference category, and δ_0 indicates the change in the intercept with respect to the reference group.

From interpretation point of view it may also be beneficial to associate the categories directly to the regression coefficients.

Consider the wage example, where

(3)
$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + u.$$

Suppose we are interested about the difference in wage levels between men an women. Then we can model β_0 as a function of gender as

(4)
$$\beta_0 = \beta_{\text{m}} + \delta_{\text{f}} \text{female},$$

where subscripts m and f refer to male and female, respectively. Model (3) can be written as

(5)
$$\begin{aligned} \text{wage} \;=\; & \beta_{\text{m}} + \delta_{\text{f}} \text{female} + \beta_1 \text{educ} \\ & + \beta_2 \text{exper} + \beta_3 \text{tenure} + u. \end{aligned}$$

In the model the female dummy is zero for men. All other factors remain the same. Thus the expected difference between wages in terms of the model is according to (2) equal to δ_{f}

We can also run a regression of wage on the female dummy alone, without any additional controls. This is a convenient form of running the independent sample t-test known from the introductory statistics course. The intercept β_{m} equals then the average wage of men and δ_{f} the average difference between mens and womens wages.

If we use logarithmic wages $log(w)$ instead of levels on the left hand side, then $100\,\delta_{\mathsf{f}}$ approximates the relative difference in percentages. The log approximation may be inaccurate if the percentage difference is large.

A more accurate approximation is obtained by using the fact that

$$\log(w_f) - \log(w_m) = \log(w_f/w_m) = \delta_{\mathsf{f}},$$

where w_f and w_m refer to a woman's and a man's wage, respectively, thus

$$100\left(\frac{w_f - w_m}{w_m}\right)\% = 100\,(\exp(\delta_{\mathsf{f}}) - 1)\,\%. \tag{6}$$

Example 6.1: Augment the wage example with squared exper and squared tenure to account for possible reducing incremental effect of experience and tenure and account for the possible wage difference with the female dummy. Thus the model is

$$\begin{aligned} \log(w) &= \beta_m + \delta_{\mathrm{f}}\mathrm{female} + \beta_1\mathrm{educ} \\ &\quad + \beta_2\mathrm{exper} + \beta_3\mathrm{tenure} \\ &\quad + \beta_4(\mathrm{exper})^2 + \beta_5(\mathrm{tenure})^2 + u. \end{aligned} \tag{7}$$

```
EViews Estimation Results:
Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: 1 526
Included observations: 526
=========================================================
Variable  Coefficient   Std. Error  t-Statistic  Prob.
---------------------------------------------------------
C            0.416691     0.098928     4.212066 0.0000
FEMALE      -0.296511     0.035805    -8.281169 0.0000
EDUC         0.080197     0.006757     11.86823 0.0000
EXPER        0.029432     0.004975     5.915866 0.0000
TENURE       0.031714     0.006845     4.633036 0.0000
EXPER^2     -0.000583     0.000107    -5.430528 0.0000
TENURE^2    -0.000585     0.000235    -2.493365 0.0130
=========================================================
R-squared          0.440769 Mean dependent var     1.623268
Adjusted R-squared 0.434304 S.D. dependent var     0.531538
S.E. of regression 0.399785 Akaike info criterion  1.017438
Sum squared resid  82.95065 Schwarz criterion      1.074200
Log likelihood    -260.5861 F-statistic            68.17659
Durbin-Watson      1.795726 Prob(F-statistic)      0.000000
=========================================================
```

Using (6), with $\hat{\delta}_{\mathrm{f}} = -0.296511$

$$100\frac{\widehat{w}_f - \widehat{w}_m}{\widehat{w}_m} = 100[\exp(\hat{\delta}_{\mathrm{f}}) - 1] \approx -25.7\%, \tag{8}$$

which suggests that, given the other factors, women's wages (w_f) are on average 25.7 percent lower than men's wages (w_m).

It is notable that exper and tenure squared have statistically significant negative coefficient estimates, which supports the idea of diminishing marginal increase due to these factors.

6.2 Multiple categories

Additional dummy variables can be included to the regression model as well. In the wage example if *married* (married = 1, if married, and 0, otherwise) is included we have the following possibilities

female	married	characteization
1	0	single woman
1	1	married woman
0	1	married man
0	0	single man

and the intercept parameter refines to

$$\beta_0 = \beta_{sm} + \delta_f \text{female} + \delta_{ma} \text{marr}. \quad (9)$$

Coefficient δ_{ma} is the wage "marriage premium".

Example 6.2: Including "married" dummy into the wage model yields

```
Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: 1 526
Included observations: 526
```

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.41778	0.09887	4.226	0.0000
FEMALE	-0.29018	0.03611	-8.036	0.0000
MARRIED	0.05292	0.04076	1.299	0.1947
EDUC	0.07915	0.00680	11.640	0.0000
EXPER	0.02695	0.00533	5.061	0.0000
TENURE	0.03130	0.00685	4.570	0.0000
EXPER^2	-0.00054	0.00011	-4.813	0.0000
TENURE^2	-0.00057	0.00023	-2.448	0.0147

R-squared	0.443	Mean dependent var	1.623
Adjusted R-squared	0.435	S.D. dependent var	0.532
S.E. of regression	0.400	Akaike info crit.	1.018
Sum squared resid	82.682	Schwarz criterion	1.083
Log likelihood	-259.731	F-statistic	58.755
Durbin-Watson stat	1.798	Prob(F-statistic)	0.000

The estimate of the "marriage premium" is about 5.3%, but it is not statistically significant.

A major limitation of this model is that it assumes that the marriage premium is the same for men and women.

We relax this next.

Generating dummies: "singfem", "marrfem", and "marrmale" we can investigate the "marriage premiums" for men women.

The intercept term becomes

$$\begin{aligned}\beta_0 &= \beta_{sm} + \delta_{mm}\text{marrmale} \\ &\quad +\delta_{mf}\text{marrfem} + \delta_{sf}\text{singfem}.\end{aligned} \tag{10}$$

The needed dummy-variables can be generated as cross-products form the "female" and "married" dummies.

For example, the "singfem" dummy is

$$\text{singfem} = (1 - \text{married}) \times \text{female}.$$

Example 6.3: Estimating the model with the intercept modeled as (10) gives

Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.3214	0.1000	3.213	0.0014
MARRMALE	0.2127	0.0554	3.842	0.0001
MARRFEM	-0.1983	0.0578	-3.428	0.0007
SINGFEM	-0.1104	0.0557	-1.980	0.0483
EDUC	0.0789	0.0067	11.787	0.0000
EXPER	0.0268	0.0052	5.112	0.0000
TENURE	0.0291	0.0068	4.302	0.0000
EXPER^2	-0.0005	0.0001	-4.847	0.0000
TENURE^2	-0.0005	0.0002	-2.306	0.0215

R-squared	0.461	Mean dependent var	1.623
Adjusted R-squared	0.453	S.D. dependent var	0.532
S.E. of regression	0.393	Akaike info crit.	0.988
Sum squared resid	79.968	Schwarz criterion	1.061
Log likelihood	-250.955	F-statistic	55.246
Durbin-Watson stat	1.785	Prob(F-statistic)	0.000

The reference group is single men.

Single women and men estimated wage difference: -11.0%, just borderline statistically significant in the two sided test.

"Marriage premium" for men: 21.3% (more accurately, using (7), 23.7%).

Married women are estimated to earn 19.8% less than single men.

"Marriage premium" for women:

$$\hat{\delta}_{\text{mf}} - \hat{\delta}_{\text{sf}} = -0.198 - (-0.110) \approx -0.088$$

or -8.8%.

The statistical significance of this can be tested either by redefining the dummies such that the single women become the base.

Another option is to use advanced econometric software. EVievs Wald test produces for

(11) $$H_0 : \delta_{\text{mf}} = \delta_{sf}$$

$F = 2.821$ (df 1 and 517) with p-value 0.0937, which is not statistically significant and there is not empirical evidence for wage difference between married and single women.

In the same manner testing for $H_0 : \delta_{mm} = \delta_{mf}$ produces $F = 80.61$ with p-value 0.0000, i.e., highly statistically significant. Thus, there is strong empirical evidence of 'marriage premium' for men.

Remark 6.1: If there are q then $q-1$ dummy variables are needed. The category which does not have a dummy variable becomes the base category or benchmark.

Remark 6.2: "Dummy variable trap". If the model includes the intercept term, defining q dummies for q categories leads to an exact linear dependence, because $1 = D_1 + \cdots + D_q$. Note also that $D^2 = D$, which again leads to an exact linear dependency if a "dummy squared" is added to the model. All these cases which lead to the exact linear dependency with dummy-variables are called the "dummy variable trap".

Ordinal Information

If the categories include ordinal information (e.g. 1 = "good", 2 = "better", 3 = "best"), sometimes people these variables as such in regressions. However, interpretation may be a problem, because "one unit change" implies a constant partial effect. That is the difference between "better" and "good" is as big as "best" and "better".

The usual alternative to use dummy-variables. In the above example two dummies are needed. $D_1 = 1$ is "better", and 0 otherwise, $D_2 = 1$ for "best" and 0 otherwise. As a consequence, the reference group is "good".

The constant partial effect can be tested by testing the restricted model

(12) $$y = \beta_0 + \delta(D_1 + 2D_2) + x + u$$

against the unrestricted alternative

(13) $$y_i = \beta_0 + \delta_1 D_1 + \delta_2 D_2 + x + u.$$

Example 6.4: Effects of law school ranking on starting salaries. Dummy variables top10, $r11\_25$, $r26\_40$, $r41\_60$, and $r61\_100$. The reference group is the schools ranked below 100.

Below are estimation results with some additional covariates (Wooldridge, Example 7.8).

Dependent Variable: LOG(SALARY)
Method: Least Squares
Sample (adjusted): 1 155
Included observations: 136 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.1653	0.4114	22.277	0.0000
TOP10	0.6996	0.0535	13.078	0.0000
R11_25	0.5935	0.0394	15.049	0.0000
R26_40	0.3751	0.0341	11.005	0.0000
R41_60	0.2628	0.0280	9.399	0.0000
R61_100	0.1316	0.0210	6.254	0.0000
LSAT	0.0057	0.0031	1.858	0.0655
GPA	0.0137	0.0742	0.185	0.8535
LOG(LIBVOL)	0.0364	0.0260	1.398	0.1647
LOG(COST)	0.0008	0.0251	0.033	0.9734

R-squared	0.911	Mean dependent var	10.541
Adjusted R-squared	0.905	S.D. dependent var	0.277
S.E. of regression	0.086	Akaike info crit.	-2.007
Sum squared resid	0.924	Schwarz criterion	-1.792
Log likelihood	146.452	F-statistic	143.199
Durbin-Watson stat	1.829	Prob(F-statistic)	0.000

The estimation results indicate that the ranking has a big influence on the staring salary. The estimated median salary at a law school ranked between 61 and 100 is about 13% higher than in those ranked below 100. The coefficient estimate for the top 10 is 0.6996, using (7) we get $100\,[\exp(0.6996) - 1] \approx 101.4\%$, that is median starting salaries in top 10 schools tend to be double to those ranked below 100.

Example 6.5: Although not fully relevant, let us for just illustration purposes test constant partial effect hypothesis. I.e., whether

$$H_0 : \begin{cases} \delta_{\text{top10}} = 5\delta_{61\_100}, \\ \delta_{11\_25} = 4\delta_{61\_100}, \\ \delta_{26\_40} = 3\delta_{61\_100}, \\ \delta_{41\_60} = 2\delta_{61\_100}. \end{cases} \tag{14}$$

Using Wald test for coefficient restrictions in EViews gives $F = 1.456$ with $df_1 = 4$ and $df_2 = 126$ and p-value 0.2196. This indicates that the there is not much empirical evidence against the constant partial effect for the starting salary increment. The estimated constant partial coefficient is 0.139782, i.e., at each ranking class starting median salary is estimated to increase approximately by 14%.

6.3 Different Slopes

Consider

(15) $$y = \beta_0 + \beta_1 x_1 + u.$$

If the slope depends also on the group, we get in addition to

(16) $$\beta_0 = \beta_{00} + \delta_0 D,$$

for the slope coefficient similarly

(17) $$\beta_1 = \beta_{11} + \delta_1 D.$$

The regression equation is then

(18) $$y = \beta_{00} + \delta_0 D + \beta_{11} x_1 + \delta_1 D x_1 + u.$$

Example 6.6: Wage example. Test whether return of eduction differs between women and men. This can be tested by defining

(19) $$\beta_{\text{educ}} = \beta_{\text{meduc}} + \delta_{\text{feduc}}\text{female}.$$

The null hypothesis is $H_0 : \delta_{\text{feduc}} = 0$.

```
Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: 1 526
Included observations: 526
```

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.31066	0.11831	2.626	0.0089
MARRMALE	0.21228	0.05546	3.828	0.0001
MARRFEM	-0.17093	0.17100	-1.000	0.3180
SINGFEM	-0.08340	0.16815	-0.496	0.6201
FEMALE*EDUC	-0.00219	0.01288	-0.170	0.8652
EDUC	0.07976	0.00838	9.521	0.0000
EXPER	0.02676	0.00525	5.095	0.0000
TENURE	0.02916	0.00678	4.299	0.0000
EXPER^2	-0.00053	0.00011	-4.829	0.0000
TENURE^2	-0.00054	0.00023	-2.309	0.0213

R-squared	0.461	Mean dependent var	1.623
Adjusted R-squared	0.452	S.D. dependent var	0.532
S.E. of regression	0.394	Akaike info crit.	0.992
Sum squared resid	79.964	Schwarz criterion	1.073
Log likelihood	-250.940	F-statistic	49.018
Durbin-Watson stat	1.785	Prob(F-statistic)	0.000

$\widehat{\delta}_{\text{feduc}} = -0.00219$ with p-value 0.8652. Thus there is no empirical evidence that the return of education would differ between men and women.

Chow Test

Suppose there are two populations (e.g. men and women) and we want to test whether the same regression function applies to both groups.

All this can be handled by introducing a dummy variable, D with $D = 1$ for group 1 and zero for group 2.

If the regression in group g ($g = 1, 2$) is

(20) $$y_{g,i} = \beta_{g,0} + \beta_{g,1} x_{g,i,1} + \cdots + \beta_{g,k} x_{g,i,k} + u_{g,i},$$

$i = 1, \ldots, n_g$, where n_g is the number of observations from group g.

Using the group dummy, we can write

(21) $$\beta_{g,j} = \beta_j + \delta_j D,$$

$j = 0, 1, \ldots, k$. An important assumption is that in both groups $\mathbb{V}\text{ar}[u_{g,i}] = \sigma_u^2$.

The null hypothesis is

(22) $$H_0 : \delta_0 = \delta_1 = \cdots = \delta_k = 0.$$

The null hypothesis (22) can be tested with the F-test, given in (4.20).

In the first step the unrestricted model is estimated over the pooled sample with coefficients of the form in equation (20) (thus $2(k+1)$-coefficients).

Next the restricted model, with all δ-coefficients set to zero, is estimated again over the pooled sample.

Using the SSRs from restricted and unrestricted models, test statistic (4.20) becomes

$$F = \frac{(\text{SSR}_r - \text{SSR}_{ur})/(k+1)}{\text{SSR}_{ur}/[n-2(k+1)]}, \tag{23}$$

which has the F-distribution under the null hypothesis with $k+1$ and $n-2(k+1)$ degrees of freedom.

Exactly the same result is obtained if one estimates the regression equations separately from each group and sums up the SSRs. That is

$$\text{SSR}_{ur} = \text{SSR}_1 + \text{SSR}_2, \tag{24}$$

Where SSR_g is from the regression estimated from group g, $g = 1, 2$.

Thus, statistic (23) can be written alternatively as

$$F = \frac{[\text{SSR}_r - (\text{SSR}_1 + \text{SSR}_2)]/(k+1)}{(\text{SSR}_1 + \text{SSR}_2)/[n - 2(k+1)]}, \tag{25}$$

which is known as Chow statistic (or Chow test).