Chapter 6
Analysis with Qualitative Information: Dummy Variables

In previous chapters, the dependent and the independent variables in our regression equations had a *quantitative* meaning. That is, the magnitude of the variable had a useful information, for example, years of education, years of experience, unemployment rate, or wage. In this chapter we will analyze how to introduce *qualitative* information into a regression equation. Example of qualitative information includes marital status, gender, race, industry (manufacturing, retail, etc.) or geographical region (south, north, west, etc.).

6.1 Describing qualitative information

Qualitative factors often come in the form of binary information: a person is female of male; a person does or does not own a computer; a person is married or not. In all these cases the relevant information can be captured by a binary variable, also called a dummy variable or zero-one variable. In defining a dummy variable we must decide which event is assigned a value of one and which a value of zero. Table 6.1 shows how two dummy variables (`female` and `married`) look in the data set.

Table 6.1 A partial Listing of the Data in Wage.xls

`person`	`wage`	`educ`	`exper`	`female`	`married`
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

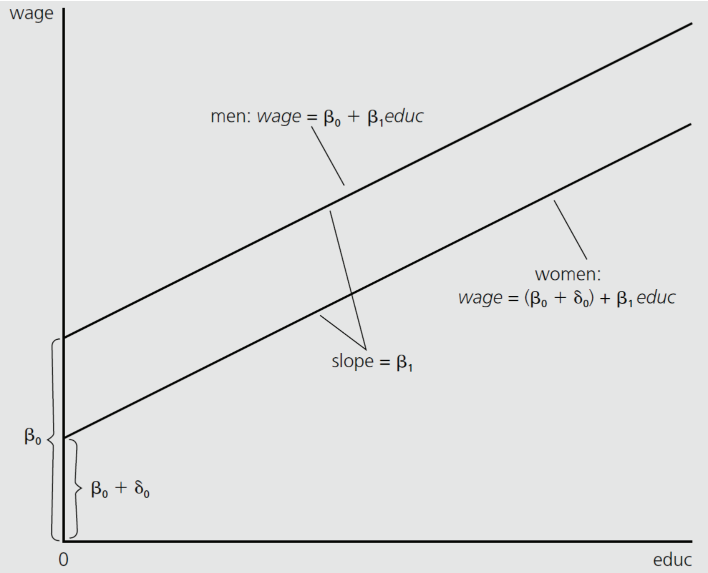

Fig. 6.1 Graph of $\texttt{wage} = \beta_0 + \delta_0 \texttt{female} + \beta_1 \texttt{educ}$ for $\delta_0 < 0$.

6.2 A single dummy independent variable

The simplest case is when we have a single dummy independent variable. Let's consider the following model:

$$\texttt{wage} = \beta_0 + \delta_0 \texttt{female} + \beta_1 \texttt{educ} + \varepsilon \tag{6.1}$$

We use the parameter δ_0 to emphasize the fact that `female` corresponds to a dummy variable. If the person is a female we have `female` = 1, and if the person is a male, we have `female` = 0. The parameter δ_0 has the following interpretation: δ_0 is the difference in hourly wage between females and males, given the same amount of education (and the error term ε). Thus, the coefficient δ_0 determines whether there is discrimination against women: if $\delta_0 < 0$, it means that on average, women earn less than men.

The interpretation of δ_0 (when $\delta < 0$) can be depicted graphically in Figure 6.1 as an *intercept shift* between males an females.

Let's estimate the following more interesting model:

$$\texttt{wage} = \beta_0 + \delta_0 \texttt{female} + \beta_1 \texttt{educ} + \beta_2 \texttt{exper} + \beta_1 \texttt{tenure} + \varepsilon \tag{6.2}$$

The regression output in Gretl is:

```
Model 2: OLS, using observations 1-526
Dependent variable: wage
```

```
             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -1.56794      0.724551     -2.164    0.0309    **
  female     -1.81085      0.264825     -6.838    2.26e-011 ***
  educ        0.571505     0.0493373    11.58     9.09e-028 ***
  exper       0.0253959    0.0115694     2.195    0.0286    **
  tenure      0.141005     0.0211617     6.663    6.83e-011 ***

Mean dependent var   5.896103   S.D. dependent var   3.693086
Sum squared resid    4557.308   S.E. of regression   2.957572
R-squared            0.363541   Adjusted R-squared   0.358655
F(4, 521)            74.39801   P-value(F)           7.30e-50
Log-likelihood      -1314.228   Akaike criterion     2638.455
Schwarz criterion    2659.782   Hannan-Quinn         2646.805
```

$$\widehat{\texttt{wage}} = \underset{(0.72455)}{-1.56794} - \underset{(0.26483)}{1.81085}\,\texttt{female} + \underset{(0.049337)}{0.571505}\,\texttt{educ} + \underset{(0.011569)}{0.0253959}\,\texttt{exper}$$

$$+ \underset{(0.021162)}{0.141005}\,\texttt{tenure}$$

$$N = 526 \quad \bar{R}^2 = 0.3587 \quad F(4,521) = 74.398 \quad \hat{\sigma} = 2.9576$$

(standard errors in parentheses)

Where it is easy to see that $\delta_0 = -1.81$. If we want to test the null hypothesis that there is no difference between men and women, $H_0 : \delta_0 = 0$. The alternative hypothesis is that there is discrimination against women, $H_1 : \delta_0 < 0$. Based on the p-value we reject the null and conclude that there is discrimination, females make two dollars and twenty seven cents less per hour than males. This is after controlling for differences in education, experience and tenure.

It is illustrative to additionally estimate the following equation:

$$\texttt{wage} = \beta_0 + \delta_0\texttt{female} + \varepsilon \tag{6.3}$$

where we do not control for education, experience or tenure. The regression output is:

```
Model 3: OLS, using observations 1-526
Dependent variable: wage

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const        7.09949      0.210008    33.81     8.97e-134 ***
  female      -2.51183      0.303409    -8.279    1.04e-015 ***

Mean dependent var   5.896103   S.D. dependent var   3.693086
Sum squared resid    6332.194   S.E. of regression   3.476254
R-squared            0.115667   Adjusted R-squared   0.113979
F(1, 524)            68.53668   P-value(F)           1.04e-15
Log-likelihood      -1400.732   Akaike criterion     2805.464
Schwarz criterion    2813.995   Hannan-Quinn         2808.804
```

$$\widehat{\texttt{wage}} = \underset{(0.21001)}{7.09949} - \underset{(0.30341)}{2.51183}\,\texttt{female}$$

$$N = 526 \quad \bar{R}^2 = 0.1140 \quad F(1,524) = 68.537 \quad \hat{\sigma} = 3.4763$$

(standard errors in parentheses)

The expected (predicted) wage for females is $\widehat{\texttt{wage}} = 7.099 - 2.512\texttt{1} = 4.587$, while the expected wage for males is $\widehat{\texttt{wage}} = 7.099 - 2.512\texttt{0} = 7.099$. This is not controlling for differences in education, experience or tenure. Once we control for those differences, the wage gap between these two groups is smaller and equal to $\delta_0 = -1.81$.

What is the interpretation of the coefficient on a dummy variable if the dependent variable is in logs? Here the coefficient has a *percentage* interpretation. Let's say we want to estimate the following equation:

$$\log \texttt{wage} = \beta_0 + \delta_0 \texttt{female} + \beta_1 \texttt{educ} + \beta_2 \texttt{exper} + \beta_3 \texttt{tenure} + \varepsilon \qquad (6.4)$$

that has the following Gretl estimation output:

$$\widehat{\texttt{logwage}} = \underset{(0.10190)}{0.501348} - \underset{(0.037246)}{0.301146}\,\texttt{female} + \underset{(0.0069389)}{0.0874623}\,\texttt{educ} + \underset{(0.0016271)}{0.00462938}\,\texttt{exper}$$

$$+ \underset{(0.0029762)}{0.0173670}\,\texttt{tenure}$$

$$N = 526 \quad \bar{R}^2 = 0.3876 \quad F(4,521) = 84.072 \quad \hat{\sigma} = 0.41596$$

(standard errors in parentheses)

The coefficient on `female`, δ_0, implies that for the same levels of education, experience, and tenure, women earn approximately 100(0.301) = 30.1% less than men.

6.3 Dummy variables for multiple categories

One can use several dummy variables in the same equation. For example, we can add the dummy variable `married` to Equation 6.3 to obtain:

$$\texttt{wage} = \beta_0 + \delta_0 \texttt{female} + \delta_1 \texttt{married} + \varepsilon \qquad (6.5)$$

In Gretl we have,

$$\widehat{\texttt{wage}} = \underset{(0.29634)}{6.18043} - \underset{(0.30261)}{2.29440}\,\texttt{female} + \underset{(0.30971)}{1.33948}\,\texttt{married}$$

$$N = 526 \quad \bar{R}^2 = 0.1429 \quad F(2,523) = 44.779 \quad \hat{\sigma} = 3.4190$$

(standard errors in parentheses)

The coefficient on `married` gives the (approximate) difference in wages between married and non married individuals. Based on these results, married individuals have higher hourly wages. On important restriction in Equation 6.5 is that it restricts the effect of marital status on wages is the same whether you are male of female. If we are interested in this difference we can estimate an alternative model with additional categories. In particular we need four categories: (1) married men, (2) married women (3) single men, and (4) single woman. We must select a base group (for example, single men) and create the dummy variables for the other three groups.

$$\begin{aligned}\texttt{marrmale} &= \texttt{married} \times (1 - \texttt{female})\\ \texttt{marrfem} &= \texttt{married} \times \texttt{female}\\ \texttt{singfem} &= (1 - \texttt{married}) \times \texttt{female}\end{aligned}$$

The equation we want to estimate is:

$$\log \texttt{wage} = \beta_0 + \delta_0 \texttt{marrmale} + \delta_1 \texttt{marrfem} + \delta_2 \texttt{singfem} + \varepsilon \tag{6.6}$$

and the estimation output is:

$$\widehat{\texttt{logwage}} = \underset{(0.050987)}{1.5201} + \underset{(0.061554)}{0.4267}\ \texttt{marrmale} - \underset{(0.065524)}{0.0797}\ \texttt{marrfem} - \underset{(0.066804)}{0.1316}\ \texttt{singfem}$$

$$N = 526 \quad \bar{R}^2 = 0.2087 \quad F(3,522) = 47.149 \quad \hat{\sigma} = 0.47284$$

(standard errors in parentheses)

The interpretation of each of the δ coefficients is with respect to the base group. For example $\delta_2 = 0.1316$ means that single females earn approximately 13.16% lower hourly wages than single men (the base group).

6.4 Incorporating ordinal information

Suppose we want to estimate the effect of city credit ratings on the municipal bond interest rate (MBR). The credit rating (CR) is an ordinal variable and suppose it goes from zero (worst credit) to four (best credit). Under these consideration, a potential candidate for our model is:

$$\texttt{MBR} = \beta_0 + \beta_1 \texttt{CR} + other factors + \varepsilon \tag{6.7}$$

where $other factors$ are just other variables in the model. On concern with this specification is that it is hard to interpret one unit increase in *CR*. It is easy to talk about an additional year of education or an additional year of income, but credit ratings usually have only an ordinal meaning. Moreover, it is restrictive to assume that each additional unit increase in CR has the same effect on MBR. An alternative approach is to create separate dummy variables for each of the values of CR, that is,

$$\begin{aligned}
\texttt{CR}_1 &= 1 \text{ if } \texttt{CR} = 1 \\
&= 0 \text{ otherwise.} \\
\texttt{CR}_2 &= 1 \text{ if } \texttt{CR} = 2 \\
&= 0 \text{ otherwise.} \\
\texttt{CR}_3 &= 1 \text{ if } \texttt{CR} = 3 \\
&= 0 \text{ otherwise.} \\
\texttt{CR}_4 &= 1 \text{ if } \texttt{CR} = 4 \\
&= 0 \text{ otherwise.}
\end{aligned}$$

Then we can focus on estimating the following model:

$$\texttt{MBR} = \beta_0 + \delta_1 \texttt{CR}_1 + \delta_2 \texttt{CR}_2 + \delta_3 \texttt{CR}_3 + other factors + \varepsilon \tag{6.8}$$

Again, we omit one category ($\texttt{CR}_4$) and the interpretation of the dummy coefficients is relative to the omitted category. For example, δ_2 represents the difference in municipal bond interest rate between ratings $\texttt{CR}_2$ and $\texttt{CR}_4$.

6.5 Interactions involving dummy variables

Just as quantitative variables can have interactions, so can dummy variables. Actually, we revisit the estimation of Equation 6.6 to see that the same results can be achieved by including the interaction term between `female` and `married`. The model we want to estimate is:

$$\log \texttt{wage} = \beta_0 + \delta_0 \texttt{female} + \delta_1 \texttt{married} + \delta_2 (\texttt{female} \times \texttt{married}) + \varepsilon \tag{6.9}$$

Estimating in Gretl we have:

$$\begin{aligned}
\widehat{\texttt{logwage}} = &\underset{(0.050987)}{1.5201} - \underset{(0.066804)}{0.1316}\ \texttt{female} + \underset{(0.061554)}{0.4267}\ \texttt{married} \\
&- \underset{(0.085708)}{0.3748}\ \texttt{female} \times \texttt{married}
\end{aligned}$$

$$N = 526 \quad \bar{R}^2 = 0.2087 \quad F(3,522) = 47.149 \quad \hat{\sigma} = 0.47284$$

(standard errors in parentheses)

Notice that this regression output is equivalent as the one obtained from Equation 6.6.

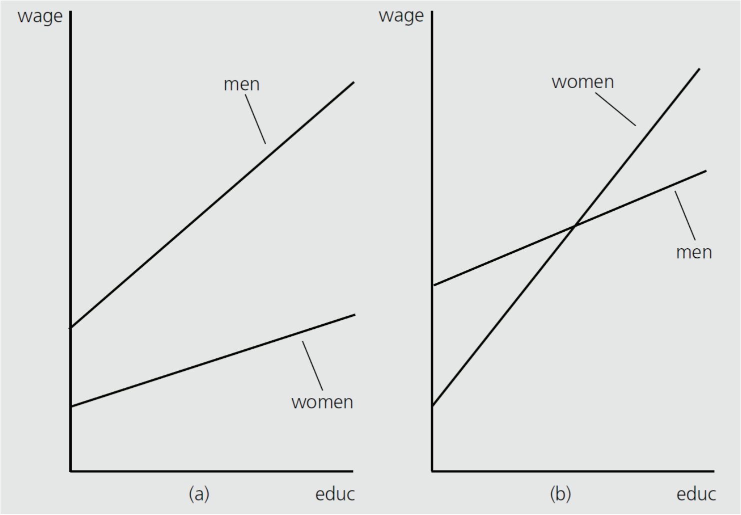

Fig. 6.2 Graph of $\texttt{wage} = \beta_0 + \delta_0 \texttt{female} + \beta_1 \texttt{educ} + \delta_1 \texttt{educ} \times \texttt{female}$.

6.5.1 Allowing for different slopes

Consider the case where we want to estimate the effect of education on hourly wage and in addition, we want for the marginal effect to change based on your gender. This can be done by interacting the `educ` with `female` and estimating the following model:

$$\texttt{wage} = \beta_0 + \delta_0 \texttt{female} + \beta_1 \texttt{educ} + \delta_1 (\texttt{female} \times \texttt{educ}) + \varepsilon \tag{6.10}$$

A graphical approach to this problem in presented in Figure 6.2. The output in Gretl is

$$\widehat{\texttt{wage}} = \underset{(0.84356)}{0.200496} - \underset{(1.3250)}{1.19852}\,\texttt{female} + \underset{(0.064223)}{0.539476}\,\texttt{educ}$$
$$-\underset{(0.10364)}{0.0859990}\,\texttt{female} \times \texttt{educ}$$
$$T = 526 \quad \bar{R}^2 = 0.2555 \quad F(3,522) = 61.070 \quad \hat{\sigma} = 3.1865$$

(standard errors in parentheses)

6.5.2 Testing for differences in regression functions across groups

So far we saw that interacting a dummy variable with other independent variables is a powerful tool. Now, we can use this tool to test the null hypothesis that two groups follow the same regression function, against the alternative that one or more of the slopes differs across the two groups. Suppose we want to test whether the same regression model describe college GPA for males and for females. The model is

$$\texttt{cumgpa} = \beta_0 + \beta_1 \texttt{sat} + \beta_2 \texttt{hsperc} + \beta_3 \texttt{tothrs} + \varepsilon, \qquad (6.11)$$

where `cumgpa` is cumulative college GPA, `sat` is the SAT score, `hsperc` is the high school rank percentile, and `tothrs` is the total hours of college courses. The regression results in Gretl are

$$\widehat{\texttt{cumgpa}} = \underset{(0.22855)}{0.929111} + \underset{(0.000208)}{0.0009028}\,\texttt{sat} - \underset{(0.00157)}{0.006379}\,\texttt{hsperc} + \underset{(0.000931)}{0.01198}\,\texttt{tothrs}$$

$$N = 732 \quad \bar{R}^2 = 0.2323 \quad F(3,728) = 74.717 \quad \hat{\sigma} = 0.86711$$

(standard errors in parentheses)

To allow for a difference in the intercept we just need to include the dummy variable `female`. Then, to allow differences in the slope parameters we need to include interaction terms for each of the variables and `female`. That is

$$\begin{aligned} \texttt{cumgpa} = {} & \beta_0 + \delta_0 \texttt{female} + \beta_1 \texttt{sat} + \delta_1 \texttt{sat} \cdot \texttt{female} \\ & + \beta_2 \texttt{hsperc} + \delta_2 \texttt{hsperc} \cdot \texttt{female} \\ & + \beta_3 \texttt{tothrs} + \delta_3 \texttt{tothrs} \cdot \texttt{female} + \varepsilon \end{aligned} \qquad (6.12)$$

The parameter δ_0 is the difference in the intercepts between females and males, δ_1 is the slope difference with respect to `sat` between females and males, and so on. The null hypothesis that `cumgpa` follows the same model for females and males is

$$H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0 \qquad (6.13)$$

If at least one of the δ_j is different from zero, then the model is different for men and women. After creating the interaction terms, the estimated model in Gretl is

```
Model 2: OLS, using observations 1-732
Dependent variable: cumgpa

                 coefficient    std. error    t-ratio     p-value
  --------------------------------------------------------------
  const           1.21398       0.264828       4.584      5.37e-06  ***
  sat             0.000611312   0.000235026    2.601      0.0095    ***
  hsperc         -0.00596745    0.00177646    -3.359      0.0008    ***
  tothrs          0.0103004     0.00109284     9.425      5.65e-020 ***
  female         -1.11364       0.528539      -2.107      0.0355    **
  satfemale       0.00111674    0.000500034    2.233      0.0258    **
  hspercfemale    5.07597e-05   0.00410253     0.01237    0.9901
  tothrsfemale    0.00555989    0.00206958     2.686      0.0074    ***
Mean dependent var   2.080861   S.D. dependent var   0.989617
Sum squared resid    534.3092   S.E. of regression   0.859067
R-squared            0.253652   Adjusted R-squared   0.246436
F(7, 724)            35.15106   P-value(F)           2.54e-42
Log-likelihood      -923.4440   Akaike criterion     1862.888
Schwarz criterion    1899.654   Hannan-Quinn         1877.071
```

$$
\begin{aligned}
\widehat{\texttt{cumgpa}} = {} & \underset{(0.26483)}{1.21398} + \underset{(0.00023503)}{0.000611312}\,\texttt{sat} - \underset{(0.0017765)}{0.00596745}\,\texttt{hsperc} + \underset{(0.0010928)}{0.0103004}\,\texttt{tothrs} \\
& - \underset{(0.52854)}{1.11364}\,\texttt{female} + \underset{(0.00050003)}{0.00111674}\,\texttt{satfemale} + \underset{(0.0041025)}{5.07597\text{e–}005}\,\texttt{hspercfemale} \\
& + \underset{(0.0020696)}{0.00555989}\,\texttt{tothrsfemale}
\end{aligned}
$$

$$
N = 732 \quad \bar{R}^2 = 0.2464 \quad F(7,724) = 35.151 \quad \hat{\sigma} = 0.85907
$$

(standard errors in parentheses)

Now, to test the null hypothesis presented in Equation 6.13 from the window that shows the regression output, we need to go to `Tests` → `Omit variables` and a new window will open. We then have to select the variables to omit. There are `female`, `satfemale`, `hspercfemale`, and `tothrsfemale`. This will estimate the restricted model and the comparison between the restricted model (Equation 6.11) and the full model (Equation 6.12),

```
  Model 3: OLS, using observations 1-732
  Dependent variable: cumgpa

               coefficient    std. error    t-ratio    p-value
    ----------------------------------------------------------
    const       0.929111      0.228552       4.065    5.32e-05  ***
    sat         0.000902834   0.000207870    4.343    1.60e-05  ***
    hsperc     -0.00637913    0.00156785    -4.069    5.24e-05  ***
    tothrs      0.0119779     0.000931383   12.86     2.96e-034 ***

  Mean dependent var   2.080861   S.D. dependent var   0.989617
  Sum squared resid    547.3649   S.E. of regression   0.867107
  R-squared            0.235416   Adjusted R-squared   0.232265
  F(3, 728)            74.71707   P-value(F)           3.87e-42
  Log-likelihood      -932.2797   Akaike criterion     1872.559
  Schwarz criterion    1890.942   Hannan-Quinn         1879.651

  Comparison of Model 2 and Model 3:

    Null hypothesis: the regression parameters are zero for the variables
      female, satfemale, hspercfemale, tothrsfemale

    Test statistic: F(4, 724) = 4.4227, with p-value = 0.00154347
    Of the 3 model selection statistics, 1 has improved.
```

The *F* statistics that Gretl is reporting comes from

$$F = \frac{RSS - RSS_{\text{UR}}}{RSS_{\text{UR}}} \cdot \frac{n-2k}{q}, \tag{6.14}$$

where RSS is the residual sum of squares of the model estimates in Equation 6.11 and RSS_{UR} is the unrestricted model in Equation 6.12. n is the sample size, k is the number of parameters we are estimating, and q is the number of restrictions when comparing the model in Equation 6.11 and in Equation 6.12. Substituting the values we obtain,

$$F = \frac{547.3649 - 534.3092}{534.3092} \cdot \frac{732 - 2 \cdot 4}{4} = 0.024434 \cdot 181 = 4.4227, \tag{6.15}$$

An alternative way to calculate this F statistic is to follow the formula,

$$F = \frac{RSS - (RSS_1 + RSS_2)}{RSS_1 + RSS_2} \cdot \frac{n-2k}{k}, \tag{6.16}$$

where RSS is the residual sum of squares of the model estimates in Equation 6.11. RSS_1 and RSS_2 are the residual sum of squares of the model estimated in Equation 6.11 using only the females in the sample (RSS_1) and using only the males in the sample (RSS_2). As before, n is the sample size and k is the number of parameters we are estimating. The estimation of Equation 6.11 with just females is:

```
Model 5: OLS, using observations 1-180
Dependent variable: cumgpa

             coefficient   std. error    t-ratio    p-value
  ---------------------------------------------------------
  const       0.100346     0.481095       0.2086   0.8350
  sat         0.00172805   0.000464216    3.723    0.0003    ***
  hsperc     -0.00591669   0.00388949    -1.521    0.1300
  tothrs      0.0158603    0.00184854     8.580    4.82e-015 ***

Mean dependent var   2.268611   S.D. dependent var   1.126549
Sum squared resid    143.6897   S.E. of regression   0.903559
R-squared            0.367483   Adjusted R-squared   0.356702
F(3, 176)            34.08447   P-value(F)           2.03e-17
Log-likelihood      -235.1319   Akaike criterion     478.2638
Schwarz criterion    491.0356   Hannan-Quinn         483.4422
```

and with just males is:

```
Model 6: OLS, using observations 1-552
Dependent variable: cumgpa

             coefficient    std. error    t-ratio    p-value
  ----------------------------------------------------------
  const       1.21398       0.260270       4.664    3.90e-06  ***
  sat         0.000611312   0.000230981    2.647    0.0084    ***
  hsperc     -0.00596745    0.00174588    -3.418    0.0007    ***
  tothrs      0.0103004     0.00107403     9.590    3.06e-020 ***

Mean dependent var   2.019638   S.D. dependent var   0.933655
```

```
Sum squared resid    390.6194   S.E. of regression   0.844280
R-squared            0.186740   Adjusted R-squared   0.182288
F(3, 548)            41.94377   P-value(F)           2.06e-24
Log-likelihood      -687.8093   Akaike criterion     1383.619
Schwarz criterion    1400.873   Hannan-Quinn         1390.360
```

Using the formula in Equation 6.17,

$$F = \frac{547.3649-(143.6897+390.6194)}{143.6897+390.6194} \cdot \frac{732-2\cdot 4}{4} = 0.024434 \cdot 181 = 4.4227, \tag{6.17}$$

which is the same result as in Equation 6.15. This version of the F test is know also as the *Cho* test. A large F statistic is evidence against the null hypothesis. In our example the F statistic of 4.4227 has an associated p-value of 0.0015, below the usual 0.05 (or 5%). Hence, we reject the null hypothesis that there is no difference between the equation for females and the equation for males. This means that there is difference and we are better off estimating Equation 6.12 instead of Equation 6.11.

The key to estimate Equation 6.11 with just the female portion of the data change the sample. To do this go to `Sample` → `Restrict, based on criterion...`, then after a new window shows up, select the “use dummy variable” and then `female`. Once the sample is restricted, just estimate the model using Ordinary Least Squares again.

6.6 The dummy variable trap

The dummy variable trap occurs when there is an exact linear relationship among the variables in the regression model. That is the reason why we do not include `female` and `male` in the same regression equation because `female` + `male` = 1. The same occurs when we have more than one category and we should always omit one of the categories (base group). Than is why `singmen` does not appear in Equation 6.6 (`marrmale` + `singmale` + `marrfem` + `singfem` = 1).