Lecture 13. Dummy variables

Types of variables

- Continuous (income, height, weight, etc.)
- Discrete (gender, season, points scored etc.)

Continuous variables have

- Origin, i.e. value is 0
- Unit of measurement

Often obvious, e.g. price in US$.

In regression both origin and unit of measurement can be changed.
Discrete variables: three types

- Counts, e.g. number of runs scored
- Ordinal, e.g. agree/neural/disagree
- Nominal/categorical, e.g. gender

With counts there is obvious origin and also unit of measurement is obvious

Continuous variables and counts together are called quantitative variables
With ordinal variables there is no origin and no unit of measurement, but there is an order.

With nominal variables there is no unit of measurement and no origin and even no order.

Ordinal and nominal variables are called qualitative variables.
Discrete variables can be

- Dependent variable
- Independent variable

If dependent variable is discrete various problems, e.g. in

\[ Y = \alpha + \beta X + u \]

random error \( u \) cannot be continuous variable and hence cannot have a normal distribution.

In this lecture we consider qualitative variables as independent variables in linear regression models.
To use a qualitative variable as an independent variables in a linear regression

\[ Y = \alpha + \beta X + u \]

we must first attach numerical values to the categories.

For this dummy/indicator variables are very useful. A dummy/indicator variable \( D \) is a variable that has two values: 0 and 1.
Consider gender with categories female and male. We could choose

\[ D_i = 0 \quad \text{if } i \text{ is female} \]

(1)

\[ D_i = 1 \quad \text{if } i \text{ is male} \]

or

\[ D_i^* = 0 \quad \text{if } i \text{ is male} \]

(2)

\[ D_i^* = 1 \quad \text{if } i \text{ is female} \]

Because the labels are arbitrary this should not make a difference.

Note the 0 is not the origin and 1 is not the unit of measurement. They are just labels and we could have used –2 and 99 instead (but that is not a convenient choice).
The category with label 0 is called the control or reference category (I prefer reference category)

Now consider the regression model

\[ Y = \alpha + \beta D + u \]

with \( D \) as in (1) and with \( Y \) is monthly salary.

What is the interpretation of \( \alpha, \beta \)?
If assumption 2 of the CLR model holds, then

\[ E(u \mid D = 0) = E(u \mid D = 1) = 0 \]

and hence

\[ E(Y \mid D = 0) = \alpha + E(u \mid D = 0) = \alpha \]

\[ E(Y \mid D = 1) = \alpha + \beta + E(u \mid D = 1) = \alpha + \beta \]

with

\[ E(Y \mid D = 0) \text{ is average monthly salary female employees (reference category)} \]

\[ E(Y \mid D = 1) \text{ is average monthly salary male employees} \]
This suggests for OLS estimators $\hat{\alpha}, \hat{\beta}$

$$\hat{\alpha} = \bar{Y}_{female}$$

$$\hat{\alpha} + \hat{\beta} = \bar{Y}_{male}$$

and hence

$$\hat{\beta} = \bar{Y}_{male} - \bar{Y}_{female}$$

Intercept is average for reference category
Example: Sample of 49 employees

\[ n_{\text{male}} = 26, \quad n_{\text{female}} = 23 \]

\[ \bar{Y}_{\text{male}} = 2086.93, \quad \bar{Y}_{\text{female}} = 1518.70 \]

Compare with regression results:

\[ \hat{\alpha} = 1518.70, \quad \hat{\beta} = 568.23 \]

Advantage of regression: direct confidence interval of/test for salary difference between male and female employees
Dependent Variable: WAGE
Method: Least Squares
Date: 10/30/01   Time: 06:58
Sample: 1 49
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1518.696</td>
<td>122.5373</td>
<td>12.39374</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>568.2274</td>
<td>168.2208</td>
<td>3.377868</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

R-squared       | 0.195343    | Mean dependent var | 1820.204 |
Adjusted R-squared | 0.178223    | S.D. dependent var | 648.2687 |
S.E. of regression | 587.6681    | Akaike info criterion | 15.63016 |
Sum squared resid  | 16231629    | Schwarz criterion  | 15.70738 |
Log likelihood    | -380.9390   | F-statistic       | 11.40999 |
Durbin-Watson stat | 1.664603    | Prob(F-statistic) | 0.001476 |
If we replace $D$ by $D^*$, i.e. now 0 indicates male and 1 female we have the regression model

$$Y = \alpha^* + \beta^* D^* + u$$

and

$$E(Y \mid D^* = 0) = \alpha^*$$

$$E(Y \mid D^* = 1) = \alpha^* + \beta^*$$

and hence

$$\hat{\alpha}^* = \bar{Y}_{male} \quad \quad \hat{\beta}^* = \bar{Y}_{female} - \bar{Y}_{male}$$
Dependent Variable: WAGE
Method: Least Squares
Date: 10/30/01   Time: 07:15
Sample: 149   Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2086.923</td>
<td>115.2512</td>
<td>18.10760</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDERC</td>
<td>-568.2774</td>
<td>168.2208</td>
<td>-3.377868</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

R-squared 0.195343  Mean dependent var 1820.204
Adjusted R-squared 0.178223  S.D. dependent var 648.2687
S.E. of regression 587.6681  Akaike info criterion 15.63016
Sum squared resid 16231629  Schwarz criterion 15.70738
Log likelihood -380.9390  F-statistic 11.40999
Durbin-Watson stat 1.664603  Prob(F-statistic) 0.001476
For the OLS estimates we find

\[ \hat{\alpha}^* = 2086.92 \quad \hat{\beta}^* = -568.23 \]

Note \( \hat{\beta}^* = -\hat{\beta} \) and standard error is identical: tests/confidence intervals give same conclusion.

Is the result a proof of gender discrimination? Why (not)?
Now consider two dummy variables

\[ D_{i1} = 0 \quad \text{if } i \text{ is female} \]

\[ D_{i1} = 1 \quad \text{if } i \text{ is male} \]

and

\[ D_{i2} = 0 \quad \text{if } i \text{ is nonwhite} \]

\[ D_{i2} = 1 \quad \text{if } i \text{ is white} \]
We consider the following models

(1) \[ Y = \beta_1 + \beta_2 D_1 + \beta_3 D_3 + u \]

(2) \[ Y = \beta_1 + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_1 D_2 + u \]

We consider the salary difference between men and women by ethnicity.
In model (1)

\[ E(Y \mid D_1 = 1, D_2 = 0) - E(Y \mid D_1 = 0, D_2 = 0) = \beta_2 = \]

\[ = E(Y \mid D_1 = 1, D_2 = 1) - E(Y \mid D_1 = 0, D_2 = 1) \]

Restriction: Salary difference the same for whites and nonwhites

In model (2)

\[ E(Y \mid D_1 = 1, D_2 = 0) - E(Y \mid D_1 = 0, D_2 = 0) = \beta_2 \]

and

\[ E(Y \mid D_1 = 1, D_2 = 1) - E(Y \mid D_1 = 0, D_2 = 1) = \beta_2 + \beta_4 \]
Estimation results: Salary difference only for whites.

Also: Race difference only for men.

Model (2) has an interaction term $D_1D_2$. 
Dependent Variable: WAGE
Method: Least Squares
Date: 10/30/01   Time: 07:42
Sample: 1 49
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1305.974</td>
<td>161.7846</td>
<td>8.072299</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>592.7722</td>
<td>163.9563</td>
<td>3.615427</td>
<td>0.0007</td>
</tr>
<tr>
<td>RACE</td>
<td>326.1732</td>
<td>167.9351</td>
<td>1.942258</td>
<td>0.0582</td>
</tr>
</tbody>
</table>

R-squared      | 0.256330    | Mean dependent var | 1820.204
Adjusted R-squared | 0.223997 | S.D. dependent var | 648.2687
S.E. of regression | 571.0668 | Akaike info criterion | 15.59216
Sum squared resid | 15001395 | Schwarz criterion | 15.70799
Log likelihood  | -379.0079   | F-statistic | 7.927696
Durbin-Watson stat | 1.822353 | Prob(F-statistic) | 0.001101
Dependent Variable: WAGE
Method: Least Squares
Date: 10/30/01   Time: 07:43
Sample: 1
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1635.000</td>
<td>186.9467</td>
<td>8.745809</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>24.45455</td>
<td>245.6961</td>
<td>0.099532</td>
<td>0.9212</td>
</tr>
<tr>
<td>RACE</td>
<td>-178.3333</td>
<td>231.4920</td>
<td>-0.770365</td>
<td>0.4451</td>
</tr>
<tr>
<td>RACEGENDER</td>
<td>919.2788</td>
<td>312.4829</td>
<td>2.941853</td>
<td>0.0051</td>
</tr>
</tbody>
</table>

R-squared          | 0.376284    | Mean dependent var | 1820.204 |
Adjusted R-squared | 0.334703    | S.D. dependent var | 648.2687 |
S.E. of regression | 528.7651    | Akaike info criterion | 15.45707 |
Sum squared resid  | 12581662    | Schwarz criterion | 15.61151 |
Log likelihood     | -374.6983    | F-statistic | 9.049421 |
Durbin-Watson stat | 1.780860    | Prob(F-statistic) | 0.000084 |
Next, we consider qualitative variable with more than 2 categories

Examples: State of residence, level of education, income category (grouped continuous variable)

\[ S = 0 \text{ if no high school diploma} \]
\[ S = 1 \text{ if high school diploma, but no college degree} \]
\[ S = 2 \text{ if college degree} \]

Using \( S \) in this way is bad idea (why?)
Instead we introduce two dummy variables

\[ S_1 = 1 \quad \text{if high school diploma, but no college degree} \]

\[ S_1 = 0 \quad \text{otherwise} \]

and

\[ S_2 = 1 \quad \text{if college degree} \]

\[ S_2 = 0 \quad \text{otherwise} \]

Note: reference group has not a high school diploma
Regression model

\[ Y = \beta_1 + \beta_2 S_1 + \beta_3 S_2 + u \]

Now

\( \beta_1 \) is average of \( Y \) for reference group (no high school diploma)

\( \beta_1 + \beta_2 \) is average of \( Y \) for group with high school diploma, but no college degree

\( \beta_1 + \beta_3 \) is average of \( Y \) for group with college degree
How do you test

- Education has no impact on income
- The return (in income) to having a college degree is 0

Give $H_0$ and indicate which test you want to use.

Define

$$S_3 = 1 \quad \text{if no high school diploma}$$

$$S_3 = 0 \quad \text{otherwise}$$

Consider the regression model

$$Y = \beta_1 + \beta_2 S_1 + \beta_3 S_2 + \beta_4 S_3 + u$$

Why can the coefficients of this model not be estimated?
This is called the dummy variable trap

Example: Monthly salary and type of work

Maint=maintenance work

Crafts=works in crafts

Clerical=clerical work

Reference category is professional

Interpret the constant and the other coefficients.
Dependent Variable: WAGE
Method: Least Squares
Date: 10/30/01  Time: 08:20
Sample: 1 49
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2670.222</td>
<td>152.5465</td>
<td>-17.50431</td>
<td>0.0000</td>
</tr>
<tr>
<td>MAINT</td>
<td>-1052.040</td>
<td>205.6937</td>
<td>-5.114597</td>
<td>0.0000</td>
</tr>
<tr>
<td>CRAFTS</td>
<td>-641.1313</td>
<td>205.6937</td>
<td>-3.116922</td>
<td>0.0032</td>
</tr>
<tr>
<td>CLERICAL</td>
<td>-1279.222</td>
<td>186.8306</td>
<td>-6.846964</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R-squared         0.532794  Mean dependent var 1820.204
Adjusted R-squared 0.501647  S.D. dependent var 648.2687
S.E. of regression 457.6396  Akaike info criterion 15.16815
Sum squared resid   9424530  Schwarz criterion 15.32258
Log likelihood      -367.6196  F-statistic 17.10576
Durbin-Watson stat  1.780224  Prob(F-statistic) 0.000000
Combining quantitative and qualitative independent variables

Consider the model

\[ Y = \beta_1 + \beta_2 D + \beta_3 X + u \]

with \( Y \) is log of monthly salary, \( D \) is gender and \( X \) is education (in years of schooling)

In relation between \( Y \) and \( X \) the intercept is \( \beta_1 \) for women and \( \beta_1 + \beta_2 \) for men (see figure)

Estimation results (what is interpretation of coefficient of gender?)

Note that gender difference is not due to difference in level of education.
Consider two other models

(3) \[ Y = \beta_1 + \beta_2 X + \beta_3 DX + u \]

In this model intercept is the same but slope is different for men and women (see figure)

For women slope is \( \beta_2 \)

For men slope is \( \beta_2 + \beta_3 \)

(4) \[ Y = \beta_1 + \beta_2 D + \beta_3 X + \beta_4 DX + u \]

In this model both slope and intercept are different
FIGURE 7.1  An Example of an Intercept Shift Using a Dummy Variable

\[ \hat{\alpha}_1 + \hat{\alpha}_2 + \beta X \]

\[ \hat{\alpha}_2 + \beta X \]
### Regression Results

**Dependent Variable:** LNWAGE  
**Method:** Least Squares  
**Date:** 10/30/01  
**Time:** 08:35  
**Sample:** 1 49  
**Included observations:** 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.305416</td>
<td>0.058705</td>
<td>124.4423</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.281818</td>
<td>0.080591</td>
<td>3.496883</td>
<td>0.0010</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>Note</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.206459</td>
<td>Mean dependent var</td>
<td>7.454952</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.189575</td>
<td>S.D. dependent var</td>
<td>0.312741</td>
</tr>
<tr>
<td>S.E. of regression</td>
<td>0.281541</td>
<td>Akaike info criterion</td>
<td>0.342879</td>
</tr>
<tr>
<td>Sum squared resid</td>
<td>3.725458</td>
<td>Schwarz criterion</td>
<td>0.420096</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-6.400539</td>
<td>F-statistic</td>
<td>12.22819</td>
</tr>
<tr>
<td>Durbin-Watson stat</td>
<td>1.622408</td>
<td>Prob(F-statistic)</td>
<td>0.001039</td>
</tr>
</tbody>
</table>
Dependent Variable: LNWAGE
Method: Least Squares
Date: 10/30/01   Time: 08:36
Sample: 1 49
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.024929</td>
<td>0.111997</td>
<td>62.72427</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.273708</td>
<td>0.075079</td>
<td>3.645624</td>
<td>0.0007</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.045753</td>
<td>0.015946</td>
<td>2.869199</td>
<td>0.0062</td>
</tr>
</tbody>
</table>

R-squared 0.326916       Mean dependent var 7.454952
Adjusted R-squared 0.297652       S.D. dependent var 0.312741
S.E. of regression 0.262096       Akaike info criterion 0.219060
Sum squared resid 3.159944       Schwarz criterion 0.334886
Log likelihood -2.366973       F-statistic 11.17108
Durbin-Watson stat 1.809135       Prob(F-statistic) 0.000111
Model for women

\[ Y = \beta_1 + \beta_3 X + u \]

and for men

\[ Y = \beta_1 + \beta_2 + (\beta_3 + \beta_4) X + u \]

This amounts to splitting the sample and estimating two separate regressions

OLS estimates

Advantage dummy approach: Tests
FIGURE 7.2  An Example of a Slope Shift Using a Dummy Variable

\[ \hat{\alpha} + (\hat{\beta}_1 + \hat{\beta}_2)X \]

\[ \hat{\alpha} + \hat{\beta}_1X \]
Figure 7.3: An Example of a Shift in the Intercept and Slope

\[ (\hat{\alpha}_1 + \hat{\alpha}_2) + (\hat{\beta}_1 + \hat{\beta}_2)x \]

\[ \hat{\alpha}_1 + \hat{\beta}_1x \]
Dependent Variable: LNWAGE
Method: Least Squares
Date: 10/30/01   Time: 08.47
Sample (adjusted): 248 IF GENDER=1
Included observations: 26 after adjusting endpoints

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.232265</td>
<td>0.145793</td>
<td>49.60624</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.056276</td>
<td>0.021111</td>
<td>2.665669</td>
<td>0.0135</td>
</tr>
</tbody>
</table>

R-squared          | 0.228439    | Mean dependent var | 7.587234 |
Adjusted R-squared | 0.196291    | S.D. dependent var | 0.337607 |
S.E. of regression | 0.302664    | Akaike info criterion | 0.521415 |
Sum squared resid  | 2.198528    | Schwarz criterion  | 0.618192 |
Log likelihood     | -4.778394   | F-statistic       | 7.105789 |
Durbin-Watson stat | 1.926152    | Prob(F-statistic) | 0.013528 |
Dependent Variable: LNWAGE
Method: Least Squares
Date: 10/30/01   Time: 08:49
Sample: 1 49 IF GENDER=0
Included observations: 23

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.230143</td>
<td>0.160588</td>
<td>45.02294</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.012278</td>
<td>0.025268</td>
<td>0.485930</td>
<td>0.6320</td>
</tr>
</tbody>
</table>

R-squared   0.011119   Mean dependent var 7.305416
Adjusted R-squared -0.035970   S.D. dependent var 0.199545
S.E. of regression 0.203102   Akaike info criterion -0.267274
Sum squared resid 0.866261   Schwarz criterion -0.168535
Log likelihood 5.073647   F-statistic 0.236128
Durbin-Watson stat 1.591244   Prob(F-statistic) 0.632048
### OLS Regression Results

**Dependent Variable:** LNWAGE  
**Method:** Least Squares  
**Date:** 10/30/01  
**Time:** 08:50  
**Sample:** 1-49  
**Included observations:** 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.231691</td>
<td>0.106183</td>
<td>68.10581</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.012044</td>
<td>0.018206</td>
<td>0.661531</td>
<td>0.5116</td>
</tr>
<tr>
<td>EDUCGENDER</td>
<td>0.044308</td>
<td>0.011391</td>
<td>3.889899</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stats</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.347184</td>
<td>Mean dependent var</td>
<td>7.454952</td>
<td></td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.318800</td>
<td>S.D. dependent var</td>
<td>0.312741</td>
<td></td>
</tr>
<tr>
<td>S.E. of regression</td>
<td>0.258120</td>
<td>Akaike info criterion</td>
<td>0.188486</td>
<td></td>
</tr>
<tr>
<td>Sum squared resid</td>
<td>3.064794</td>
<td>Schwarz criterion</td>
<td>0.304312</td>
<td></td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-1.617911</td>
<td>F-statistic</td>
<td>12.23196</td>
<td></td>
</tr>
<tr>
<td>Durbin-Watson stat</td>
<td>1.832642</td>
<td>Prob(F-statistic)</td>
<td>0.000055</td>
<td></td>
</tr>
</tbody>
</table>
Dependent Variable: LNWAGE
Method: Least Squares
Date: 10/30/01  Time: 08:51
Sample: 1 49
Included observations: 49

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>7.230143</td>
<td>0.206344</td>
<td>35.03922</td>
<td>0.0000</td>
</tr>
<tr>
<td>GENDER</td>
<td>0.002122</td>
<td>0.241622</td>
<td>0.008780</td>
<td>0.9930</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.012278</td>
<td>0.032467</td>
<td>0.378176</td>
<td>0.7071</td>
</tr>
<tr>
<td>EDUCGENDER</td>
<td>0.043997</td>
<td>0.037222</td>
<td>1.182013</td>
<td>0.2434</td>
</tr>
</tbody>
</table>

R-squared          | 0.347185    | Mean dependent var | 7.454952 |
Adjusted R-squared | 0.303664    | S.D. dependent var | 0.312741 |
S.E. of regression  | 0.260972    | Akaike info criterion | 0.229301 |
Sum squared resid   | 3.064789    | Schwarz criterion   | 0.383735 |
Log likelihood      | -1.617869   | F-statistic         | 7.977402 |
Durbin-Watson stat  | 1.832670    | Prob(F-statistic)   | 0.000227 |