Dummy Variables

- A dummy variable (binary variable) D is a variable that takes on the value 0 or 1.

- Examples: EU member ($D = 1$ if EU member, 0 otherwise), brand ($D = 1$ if product has a particular brand, 0 otherwise), gender ($D = 1$ if male, 0 otherwise)

- Note that the labelling is not unique, a dummy variable could be labelled in two ways, i.e. for variable `gender`:

  – $D = 1$ if male, $D = 0$ if female;
  – $D = 1$ if female, $D = 0$ if male.

Regression Models with Dummy Variables

Consider a regression model with one continuous variable X and one dummy variable D:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + u.$$

If $D = 0$, then:

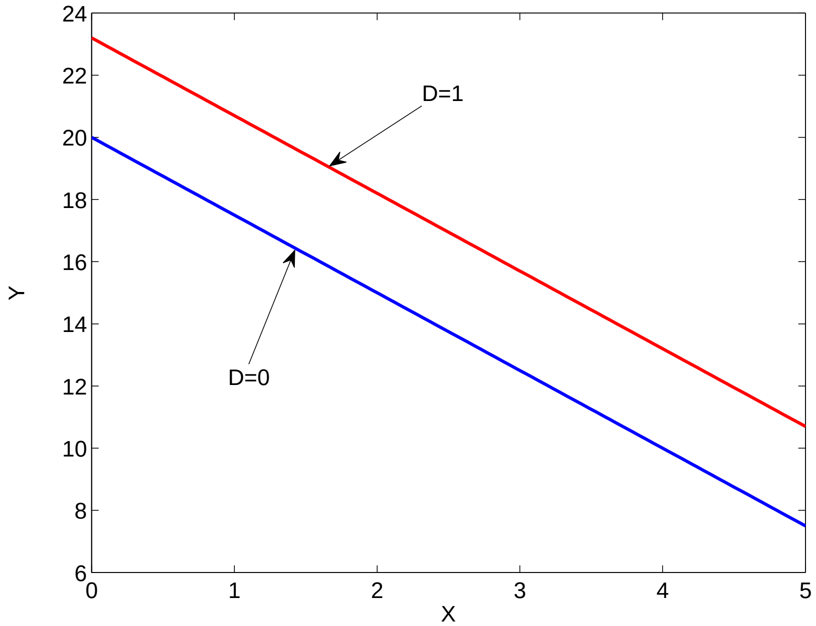
$$Y = \beta_0 + \beta_2 X + u.$$

If $D = 1$, then:

$$Y = \beta_0 + \beta_1 + \beta_2 X + u.$$

Regression Models with Dummy Variables

Example: $Y = 20 + 3.2 \cdot D - 2.5 \cdot X$



Regression Models with Dummy Variables

Interpretation:

- The observed units are split into 2 groups according to D (e.g. into men and women).

- The group with $D = 0$ is called the baseline (e.g. men).

- The regressin coefficient β_1 of D quantifies the expected effect of considering the other group (e.g. women) on the dependent variable Y, while holding all other variables (e.g. X) fixed.

- The null hypothesis $\beta_1 = 0$ corresponds to the assumption that the average value of Y is the same for both groups.

Regression Models with Dummy Variables

Consider model $Y = 20 + 3.2 \cdot D - 2.5 \cdot X + u$, where $D = 1$, if female. Assume that $X = 4$:

- expected value for Y for a man: $\mathrm{E}(Y|X = 4) = 20 - 2.5 \cdot 4 = 10$;
- expected value for Y for a woman: $\mathrm{E}(Y|X = 4) = 20 + 3.2 - 2.5 \cdot 4 = 13.2$;
- expected difference, if we consider a woman: $\beta_1 = 3.2$;
- expected difference between women and men is equal to $\beta_1 = 3.2$, even if we change X.

Combining more than one dummy variable

Estimate a model where D_1 is the gender (1: female, 0: male), D_2 is the brand (1: specific brand, 0: no-name), and P is the price:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 P + u,$$

- β_0 corresponds to the baseline (male, no-name product)
- β_1: difference in the expected rating between male and female consumers (same product).
- β_2: difference in the expected rating between the specific brand and a no-name product (same person, same price).

Categorical Variables

We can use dummy variables to control for characteristics with multiple categories (K categories, $K-1$ dummies).

Suppose one of the predictors is the highest level of education. Such variables are often coded in the following way:

edu	
1	high school dropout
2	high school degree
3	college degree

What is the effect of education on a variable Y, e.g. hourly wages?

Categorical Variables

Including `edu` directly into a linear regression model would mean that the effect of a high school degree compared to a drop out is the same as the effect of a college degree compared to a high school degree.

To include the highest level of education as predictor in a regression model, define 2 dummy variables D_1 and D_2:

	`edu`	D_1	D_2
1	high school dropout	0	0
2	high school degree	1	0
3	college degree	0	1

Categorical Variables

- Baseline (all dummies 0): high school dropout
- $D_1 = 1$, if highest degree from high school, 0 otherwise;
- $D_2 = 1$, if college degree, 0 otherwise.

Include D_1 and D_2 as dummy predictors in a regression model:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X + u.$$

The intercept β_0 corresponds to the baseline $(D_1 = 0, D_2 = 0)$.

Categorical Variables

- β_1 is the effect of a high school degree compared to a drop out.
- β_2 is the effect of a college degree compared to a drop out.

Testing hypothesis:

- Is the effect of a high school degree compared to a drop out the same as the effect of a college degree compared to a high school degree?
- Test, if $2\beta_1 = \beta_2$, or equivalently, test the linear hypothesis $2\beta_1 - \beta_2 = 0$.

Case Study Marketing

There are 5 different brands of mineral water (KR,RO,VO,JU,WA):

- Select one mineral water as baseline, e.g. KR.
- Introduce 4 dummy variables $D_1, \ldots, D_4$, and assign each of them to the remaining brands, e.g. $D_1 = 1$, if brand is equal to RO and $D_1 = 0$, otherwise; $D_2 = 1$, if brand is equal to VO and $D_2 = 0$, otherwise; etc.

The model reads:

$$Y = \beta_0 + \beta_1 D_1 + \ldots + \beta_4 D_4 + \beta_5 P + u. \tag{66}$$

Case Study Marketing

- the expected rating for the brand corresponding to the baseline is given by $\beta_0 + \beta_5 P$;

- the expected rating for the brand corresponding to D_j is given by $\beta_0 + \beta_j + \beta_5 P$;

- the coefficient β_j measures the effect of the brand D_j in comparison to the brand corresponding to the baseline;

- the difference in the expected average rating between two arbitrary brands D_j and D_k is equal to $\beta_j - \beta_k$. Is the rating different for the brands D_j and D_k? Test $\beta_j - \beta_k = 0$.

Case Study Marketing

Including an additional dummy variable D_5, where $D_5 = 1$, if brand equal to KR, i.e.

$$Y = \beta_0 + \beta_1 D_1 + \ldots + \beta_5 D_5 + \beta_6 P + u,$$

leads to a model which is not identified, because:

$$D_1 + D_2 + \ldots + D_5 = 1.$$

Hence, the set of regressors $D_1, \ldots, D_5$ is perfectly correlated with the regressor '1' corresponding to the intercept. (EViews produces an error message indicating difficulties with estimating the model.)

Case Study Marketing

It is possible to include all 5 regressors, if no constant is included in the model, with a slightly different interpretation of the coefficients:

$$Y = \beta_1 D_1 + \ldots + \beta_5 D_5 + \beta_6 P + u.$$

- β_j is a brand specific intercept of the regression model for the brand corresponding to D_j.
- For a given price level P, the expected rating for the brand corresponding to D_j is given by: $\beta_j + \beta_6 P$.
- The difference in the expected average rating between two arbitrary brands D_j and D_k is still equal to $\beta_j - \beta_k$.

II.8 Model Comparison Using R^2 and AIC/BIC

- Model evaluation using the coefficient of determination R^2
- Problems with R^2: R^2 increases with increasing number of variables, because SSR decreases $\Rightarrow$ may lead to overfitting
- Model comparison using AIC and SC (BIC): Penalize the ever decreasing SSR by including the number of parameters

Coefficient of Determination R^2

Coefficient of determination R^2 (SST is the squared sum of residuals of the simple model without predictor):

$$\mathrm{R}^2 = \frac{\mathrm{SST} - \mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} \tag{67}$$

- Close to 1, if $\mathrm{SSR} << \mathrm{SST}$; close to 0, if $\mathrm{SSR} \approx \mathrm{SST}$.

SSR is always smaller than SST. If SSR is much smaller than SST, then the regression model $\mathcal{M}_1$ is much better than the simple model $\mathcal{M}_0$.

EVIEWS Exercise II.8.1

Discuss in EVIEWS, where to find SSR and R^2; discuss by including an increasing number of predictors, how SSR and R^2 change when increasing the number of predictors

- Case Study profit, workfile `profit`;
- Case Study Chicken, workfile `chicken`;
- Case Study Marketing, workfile `marketing`;

Case Study Chicken

Predictor included	SSR	R^2
pchick	0.273487	0.647001
income	0.041986	0.945807
income, pchick	0.015437	0.980074
income, pchick, pport	0.014326	0.981509
income, pchick, pport, pbeef	0.013703	0.982313

Problems with R^2

- Choosing the model with the smallest SSR (largest R^2) leads to overfitting: R^2 increases with increasing number of variables as SSR decreases.

- R^2 is 1 for $K = N - 1$, because $\mathrm{SSR} = 0$, if we include as many predictors as observations, even if the predictors are useless.

- The increase of adding a useless predictor, however, is small $\Rightarrow$ penalize the ever decreasing SSR by including the number of parameters used for estimation which is an increasing function of the number of parameters.

Model choice criteria

Definition of model choice criteria

$$\log(\mathrm{SSR}) + m \cdot \text{Number of parameters} \tag{68}$$

- $m = 2$ ... AIC (Akaike Information Criterion)
- $m = \log(\text{Number of observations})$ ... SC (Schwarz Criterion), also called BIC

Choose the model that minimize a particular criterion

EVIEWS Exercise II.7.2

Discuss in EVIEWS, where to find AIC and Schwarz criterion; discuss how to choose predictors based on these model choice criteria

- Case Study profit, workfile `profit`;
- Case Study Chicken, workfile `chicken`; estimate log-linear model
- Case Study Marketing, workfile `marketing`;

Case Study Chicken (log-linear model)

Predictor included	SSR	R^2	AIC	SC
pchick	0.273487	0.647001	-1.420206	-1.321468
income	0.041986	0.945807	-3.294124	-3.195386
income, pchick	0.015437	0.980074	**-4.207711**	**-4.059603**
income, pchick, pporK	0.014326	0.981509	-4.195488	-3.998011
income, pchick, ppork, pbeef	0.013703	0.982313	-4.152987	-3.906140

Comparing linear and log-linear models

The residual sum of squares SSR depends on the scale of y_i, therefore AIC and SC are scale dependent

AIC and SC could not be used directly to compare a linear and a log-linear model.

AIC and SC of the log-linear model could be matched back to the original scale by adding 2 times the mean of the logarithmic values of y_i.

Comparing linear and log-linear models

Correction formula:

$$AIC = AIC^{\star} + 2\frac{1}{N}\sum_{i=1}^{N}\log(y_i) \tag{69}$$

$$SC = SC^{\star} + 2\frac{1}{N}\sum_{i=1}^{N}\log(y_i) \tag{70}$$

$AIC^{\star}$ and $SC^{\star}$ are the model choice criteria for the log-linear model

EVIEWS Exercise II.7.3 - Case Study Chicken

Predictor included	SSR	R^2	AIC	SC
income, pchick (log-linear)	0.015437	0.980074	-4.207711	-4.059603
income, pchick (linear)	106.65	0.9108	4.633	4.781

Transform AIC and SC of the log-linear model:

AIC$=$ -4.207711 $+ 2 \cdot$ 3.663887 $=$ 3.1201

SC$=$ -4.059603 $+ 2 \cdot$ 3.663887 $=$ 3.2682

log-linear model preferred