ECON3150/4150 Spring 2015

Lecture 7&8, February 9
Multiple regression model

Siv-Elisabeth Skjelbred

University of Oslo

Last updated: February 16, 2015

Outline

- Omitted variable bias
- Multiple linear regression model
    - Estimation
    - Properties
    - Measures of fit
- Data scaling
- Dummy variables in MLRM

The zero conditional mean assumption

- In the last lecture you saw that $E(u|X) = 0$ is important in order for the OLS estimator to be unbiased.
- This assumption is violated if we omit a variable from the regression that belongs in the model.
- The bias that arise from such an omission is called omitted variable bias.
- Comparing to the IRC experiment an omitted variable means that there is systematic difference between the "treatment" group and the "control group".

Omitted variable bias

Omitted variable bias

The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called omitted variable bias. For omitted variable bias to occur, the omitted variable "Z" must satisfy two conditions:

- The omitted variable is correlated with the included regressor (i.e. $corr(Z, X) \neq 0$)
- The omitted variable is a determinant of the dependent variable (i.e. Z is part of u)

Example: $Corr(Z, X) \neq 0$

The omitted variable (Z) is correlated with X, example

$$\text{wages} = \beta_0 + \beta_1 \text{educ} + \underbrace{u_i}_{\delta_1 pinc + v_i}$$

- Parents income is likely to be correlated with education, college is expensive and the alternative funding is loan or scholarship which is harder to acquire.

Example: Z is a determinant of Y

The omitted variable is a determinant of the dependent variable,

$$\text{wages} = \beta_0 + \beta_1 \text{educ} + \underbrace{u_i}_{\delta_2 MS + v_i}$$

- Market situation is likely to determine wages, workers in firms that are doing well are likely to have higher wages.

Example: Omitted variable bias

The omitted variable is both determinant of the dependent variable, i.e. $corr(X_2, Y) \neq 0$ and correlated with the included regressor

$$\text{wages} = \beta_0 + \beta_1 \text{educ} + \underbrace{u_i}_{\delta_3 ability + v_i}$$

- Ability - the higher your ability the "easier" education is for you and the more likely you are to have high education.
- Ability - the higher your ability the better you are at your job and the higher wages you get.

Omitted variable bias

The direction of bias is illustrated in the the following formula:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X} \tag{1}$$

where $\rho_{Xu} = corr(X_i, u_i)$. The formula indicates that:

- Omitted variable bias exist even when n is large.
- The larger the correlation between X and the error term the larger the bias.
- The direction of the bias depends on whether X and u are negatively or positively correlated.

How to overcome omitted variable bias

1. Run a ideal randomized controlled experiment
2. Do cross tabulation
3. Include the omitted variable in the regression

Cross tabulation

One can address omitted variable bias by splitting the data into subgroups. For example:

	College graduates	High school graduates
High family income	$\bar{Y}_{HFI,C}$	$\bar{Y}_{HFI,H}$
Medium family income	$\bar{Y}_{MFI,C}$	$\bar{Y}_{MFI,H}$
Low family income	$\bar{Y}_{LFI,C}$	$\bar{Y}_{LFI,H}$

Cross tabulation

- Cross tabulation only provides a difference of means analysis, but it does not provide a useful estimate of the ceteris paribus effect.
- To quantify the partial effect on Y_i on the change in one variable (X_{1i}) holding the other independent variables constant we need to include the variables we want to hold constant in the model.
- When dealing with multiple independent variables we need the multiple linear regression model.

Multiple linear regression model

Multiple linear regression model

- Have used only one dependent variable for simplicity.
- However, you may want to add more than one independent variable to the model.
  - You are interested in the ceteris paribus effect of multiple parameters.
  - Y is a quadratic function of X (more in chapter 8)
  - You fear violation omitted variable bias.
- When you are having more than one independent variable you have a multiple linear regression model.

Y	X	Other variables
Wages	Education	Experience, Ability
Crop Yield	Fertilizer	Soil quality, location (sun etc)
Test score	Expenditure per student	Average family income

Multiple linear regression model

The general multiple linear regression model for the population can be written in the as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

- Where the subscript i indicates the i^{th} of the n observations in the sample.
- The first subscript, 1,2,...,k, denotes the independent variable number.
- The intercept β_0 is the expected value of Y when all the X's equal zero.
- The intercept can be thought of as the coefficient on a regressor, X_{0i}, that equals zero for all i.
- The coefficient β_1 is the coefficient of X_{1i}, β_2 the coefficient on X_{2i} etc.

Multiple linear regression model

The average relationship between the k independent variables and the dependent variable is given by:

$$E(Y_i|X_{1i} = x_1, X2i = x_2, ..., X_{ki} = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

- β_1 is thus the effect on Y of a unit change in X_1 holding all other independent variables constant.
- The error term includes all other factors than the X's that influence Y.

Example

To make it more tractable consider a model with two independent variables. Then the population model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u$$

Example:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exp_i + u_i$$

$$wage_i = \beta_0 + \beta_1 exp_i + \beta_2 exp_i^2 + u_i$$

Interpretation of the coefficient

In the two variable case the predicted value is given by:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Thus the predicted change in y given the changes in X_1 and X_2 are given by:

$$\Delta \hat{Y} = \hat{\beta}_1 \Delta X_1 + \hat{\beta}_2 \Delta X_2$$

Thus if x_2 is held fixed then:

$$\Delta \hat{Y} = \hat{\beta}_1 \Delta X_1$$

$\hat{\beta}_1$ measures the partial effect of X_1 on Y holding the other independent variables (here X_2) fixed.

Interpretation of the coefficient

Using data on 526 observations on wage, education and experience the following output was obtained:

```
1 . reg wage educ exper

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  2,   523) =   75.99
       Model |  1612.2545     2  806.127251           Prob > F      =  0.0000
    Residual |  5548.15979   523  10.6083361           R-squared     =  0.2252
-------------+------------------------------           Adj R-squared =  0.2222
       Total |  7160.41429   525  13.6388844           Root MSE      =   3.257

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6442721   .0538061    11.97   0.000     .5385695    .7499747
       exper |   .0700954   .0109776     6.39   0.000     .0485297    .0916611
       _cons |  -3.390539   .7665661    -4.42   0.000    -4.896466   -1.884613
------------------------------------------------------------------------------
```

Holding experience fixed another year of education is predicted to increase your wage by 0.64 dollars.

Interpretation of the coefficient

If we want to change more than one independent variable we simply add the two effects.

Example:

$$\hat{wage} = -3.39 + 0.64educ + 0.07exp$$

If you increase education by one year and decrease experience by one year the predicted increase in wage is 0.57 dollars. (0.64-0.07)

Example: Smoking and birthweight

Using the data set birthweight_smoking.dta you can estimate the following regression:

$$\widehat{birthweight} = 3432.06 - 253.2\text{Smoker}$$

If we include the number of prenatal visits:

$$\widehat{birthweight} = 3050.5 - 218.8\text{Smoker} + 34.1\text{nprevist}$$

Example education

The relationship between years of education of male workers and the years of education of the parents.

```
8 . reg educ meduc feduc, robust

Linear regression                               Number of obs =       1129
                                                   F(  2,  1126) =     159.83
                                                   Prob > F      =     0.0000
                                                   R-squared     =     0.2689
                                                   Root MSE      =     2.2595

------------------------------------------------------------------------------
             |               Robust
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       meduc |   .1844065   .0223369     8.26   0.000     .1405798    .2282332
       feduc |   .2208784   .0259207     8.52   0.000     .1700201    .2717368
       _cons |   8.860898   .2352065    37.67   0.000     8.399405     9.32239
------------------------------------------------------------------------------
```

- Interpret the coefficient on mother's education.
- What is the predicted difference in education for a person where both parents have 12 years of education and a person where both parents have 16 years of education?

Example education and siblings

From stata:

```
. display _cons+_b[meduc]*12+_b[feduc]*12
5.8634189

. display _cons+_b[meduc]*16+_b[feduc]*16
7.4845585

.
. display 7.484-5.863
1.621

.
. *or
.
. display _b[meduc]*4+_b[feduc]*4
1.6211396
```

Or by hand:

$$0.1844 * (16 - 12) + 0.2209 * (16 - 12) = 1.6212$$

Multiple linear regression model

Advantages of the MLRM over the SLRM:

- By adding more independent variables (control variables) we can explicitly control for other factors affecting y.
- More likely that the zero conditional mean assumption holds and thus more likely that we are able to infer causality.
- By controlling for more factors, we can explain more of the variation in y, thus better predictions.
- Can incorporate more general functional forms.

Comparing estimates from simple and multiple regression

What is the return to education? Simple regression:

```
1 . reg wage educ, robust

  Linear regression                                     Number of obs =        935
                                                           F(  1,   933) =     95.65
                                                           Prob > F      =    0.0000
                                                           R-squared     =    0.1070
                                                           Root MSE      =    382.32

  --------------------------------------------------------------------------------
               |               Robust
          wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
  -------------+------------------------------------------------------------------
          educ |   60.21428   6.156956     9.78   0.000      48.1312    72.29737
         _cons |   146.9524   80.26953     1.83   0.067    -10.57731    304.4822
  --------------------------------------------------------------------------------
```

Can we give this regression a causal interpretation? What happens if we include IQ in the regression?

▸ forth

Comparing estimates from simple and multiple regression

Call the simple regression of Y on X_1 (think of regressing wage on education)

$$\tilde{Y} = \tilde{\beta_0} + \tilde{\beta_1} X_1$$

while the true population model is:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$$

The relationship between $\tilde{\beta_1}$ and β_1 is:

$$\tilde{\beta_1} = \beta_1 + \beta_2 \tilde{\delta}_1$$

where $\tilde{\delta}_1$ comes from the regression $\hat{X}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 X_1$

Comparing estimates from simple and multiple regression

Thus the bias that arise from the omitted variable (in the model with two independent variables) is given by $\beta_2 \tilde{\delta}_1$ and the direction of the bias can be summarized by the following table:

	$corr(x_1, x_2) > 0$	$corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Comparing estimates from simple and multiple regression

- Deriving the sign of omitted variable bias when there are more than two independent variables in the model is more difficult.
- Note that correlation between a single explanatory variable and the error generally results in all OLS estimators being biased.
- Suppose the true population model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

- But we estimate

$$\tilde{Y} = \tilde{\beta_0} + \tilde{\beta_1} X_1 + \tilde{\beta_2} X_2$$

- If $Corr(X_1, X_3) \neq 0$ while $Corr(X_2, X_3) = 0$ $\tilde{\beta_2}$ will also be biased unless $corr(X_1, X_2) = 0$.

Comparing estimates from simple and multiple regression

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

- People with higher ability tend to have higher education
- People with higher education tend to have less experience
- Even if we assume that ability and experience are uncorrelated β_2 is biased.
- We cannot conclude the direction of bias without further assumptions

Comparing estimates from simple and multiple regression

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	42.05762	6.810074	6.18	0.000	28.69276	55.42247
IQ	5.137958	.9266458	5.54	0.000	3.319404	6.956512
_cons	-128.8899	93.09396	-1.38	0.167	-311.5879	53.80818

IQ	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	3.533829	.1839282	19.21	0.000	3.172868	3.89479
_cons	53.68715	2.545285	21.09	0.000	48.69201	58.6823

$$\tilde{\beta}_1 = 60.214 \approx 42.047 + 3.533 * 5.137$$

back

Causation

- Regression analysis can refute a causal relationship, since correlation is necessary for causation.
- But cannot confirm or discover a causal relationship by statistical analysis alone.
- The true population parameter measures the ceteris paribus effect which holds all other (relevant) factors equal.
- However, it is rarely possible to literally hold all else equal, but one way is to take advantage of "natural experiments" or "quasi-experiments".
- One way to deal with unobserved factors is to use an instrument.

Estimation of MLRM

Assumptions of the MLRM

1. Random sampling
2. Large outliers are unlikely
3. Zero conditional mean, i.e the error u has an expected value of zero given any value of the independent variables

$$E(u|X_1, x_2, ...X_k) = 0$$

4. (There is sampling variation in X) **and there are no exact linear relationships among the independent variables.**
5. (The model is linear in parameters)

Under these assumptions the OLS estimators are unbiased estimators of the population parameters. In addition there is the homoskedasticity assumption which is necessary for OLS to be BLUE.

No exact linear relationships

Perfect collinearity

A situation in which one of the regressors is an exact linear function of the other regressors.

- This is required to be able to compute the estimators.
- The variables can be correlated, but not perfectly correlated.
- Typically perfect collinearity arise because of specification mistakes.
  - Mistakenly put in the same variable measured in different units
  - The dummy variable trap: Including the intercept plus a binary variable for each group.
  - Sample size is to small compared to parameters (need at least k+1 observations to estimate k+1 parameters)

No perfect collinearity

Solving the two 1oc for the model with two independent variables gives:

$$\hat{\beta}_1 = \frac{\hat{\sigma}^2_{X_2}\hat{\sigma}_{Y,X_1} - \hat{\sigma}_{Y,X_2}\hat{\sigma}_{X_1,X_2}}{\hat{\sigma}^2_{X_1}\hat{\sigma}^2_{X_2} - \hat{\sigma}_{X_1,X_2}}$$

where $\hat{\sigma}^2_{X_j}$ $(j = 1, 2)$, $\hat{\sigma}^2_{Y,X_j}$ and $\hat{\sigma}^2_{X_1,X_2}$ are empirical variances and covariances. Thus we require that:

$$\hat{\sigma}^2_{X_1}\hat{\sigma}^2_{X_2} - \hat{\sigma}_{X_1,X_2} = \hat{\sigma}^2_{X_1}\hat{\sigma}^2_{X_2}(1 - r^2_{X_1,X_2}) \neq 0$$

Thus must have that $\hat{\sigma}^2_{X_1} > 0$, $\hat{\sigma}^2_{X_2} > 0$ and $r^2_{X_1,X_2} < 1$. Thus the sample correlation coefficient between X_1 and X_2 cannot be one or minus one.

OLS estimation of MLRM

The procedure for obtaining the estimates is the same as with one regressor. Choose the estimate that minimize the sum of squared errors. If k=2 then minimize

$$S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2$$

- The estimates $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ are chosen simultaneously to make the squared error as small as possible.
- The i subscript is for the observation number, the second subscript is for the variable number.
- β_j would thus be the coefficient on variable number j.
- For even moderately sized n and k solving the first order conditions by hand is tedious.
- Computer software can do the calculation as long as we assume the FOCs can be solved uniquely for the $\hat{\beta}_j$'s.

OLS estimation of MLRM

The solution to the FOCs give you:

- The ordinary least square estimators $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ of the true population coefficients $(\beta_0, \beta_1, \beta_2)$.
- The predicted value $\hat{Y}$ of Y_i given X_{1i} and X_{2i}.
- The OLS residuals $\hat{u}_i = Y_i - \hat{Y}_i$.

OLS estimation of MLRM

The OLS fitted values and residuals have the same important properties as in the simple linear regression:

- The sample average of the residuals is zero and so $\bar{Y} = \bar{\hat{Y}}$
- The sample covariance between each independent variable and the OLS residuals is zero. Consequently, the sample covariance between the OLS fitted values and the OLS residuals is zero.
- The point $(\bar{X}_1, \bar{X}_2, ..., \bar{X}_k, \bar{Y})$ is always on the OLS regression line.

Properties of the MLRM OLS estimator

- Under the OLS assumptions the OLS estimators of MLRM are unbiased and consistent estimators of the unknown population coefficients.

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, 2, ...k$$

- The homoskedasticity only variance is:

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2(1 - R_j^2)}, j = 0, 1, 2, ..., k,$$

- Where R_j^2 is the R-squared from regressing x_j on all other independent variables.
- In large samples the joint samling distribution of $\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_k$ is well approximated by a multivariate normal distribution.

Properties of the MLRM OLS estimator

- Under the OLS assumptions, including homoskedasticity, the OLS estimators $\hat{\beta}_j$ are the best linear unbiased estimators of the population parameter β_j.
- Thus when the standard set of assumptions holds and we are presented with another estimator that are both linear and unbiased then we know that the variance of this estimator is at least as large as the OLS variance.
- Under heteroskedasticity the OLS estimators are not necessarily the one with the smallest variance.

Variance of the OLS estimator

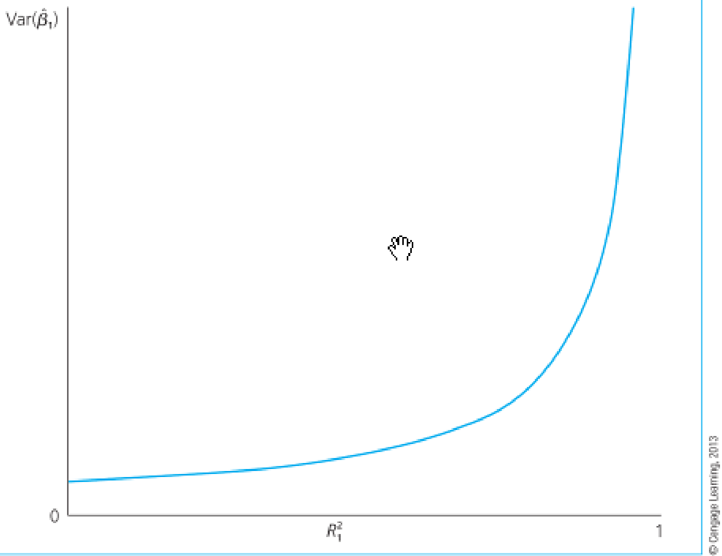
Variance:

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2(1 - R_j^2)}, j = 0, 1, 2, ..., k,$$

- As in the SLRM the OLS variance of $\hat{\beta}_1$ depend on the variance of the error term and the sample variance in the independent variable.
- In addition it depends on the linear relationship among the independent variables R_j^2

Variance of the OLS estimator

FIGURE 3.1 $Var(\hat{\beta}_1)$ as a function of R_1^2.



Imperfect collinearity

- Occurs when two or more of the regressors are highly correlated (but not perfectly correlated).
- High correlation makes it hard to estimate the effect of the one variable holding the other constant.
- For the model with two independent variables and homoskedastic errors:

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\left(\frac{1}{1-\rho^2_{X_1,X_2}}\right)\frac{\sigma^2_u}{\sigma^2_{X_1}}$$

- The two variable case illustrates that the higher the correlation between X_1 and X_2 the higher the variance of $\hat{\beta}_1$.
- Thus, when multiple regressors are imperfectly collinear, the coefficients on one or more of these regressors will be imprecisely estimated.

Overspecification

- The OVB problem may lead you to think that you should include all variables you have in your regression.
- If an explanatory variable in a regresion model has a zero population parameter in estimating an equation by OLS we call that variable irrelevant.
- An **irrelevant variable** has no partial effect on y.
- A model that includes irrelevant variables is called an overspecified model.
- An overspecified model gives unbiased estimates, but it can have undesirable effects on the variances of the OLS.
- Omitted variable bias occurs from excluding a **relevant variable**, thus the model can be said to be underspecified.

Controlling for too many factors

- In a similar way we can over control for factors.
- In some cases, it makes no sense to hold some factors fixed, precisely because they should be allowed to change.
- If you are interested in the effect of beer taxes on traffic fatalities it makes no sense to estimate:

$$fatalities = \beta_0 + \beta_1 tax + \beta_2 beercons + ...$$

- As you will measure the effect of tax holding beer consumption fixed, which is not particularly interesting unless you want to test for some indirect effect of beer taxes.

Consistency

Clive W. J. Granger (Nobel Prize-winner) once said:

> *If you can't get it right as n goes to infinity you shouldn't be in this business.*

- Which indicate that if your estimator of a particular population parameter is not consistent then you are wasting your time.
- Consistency involves a thought experiment about what would happen as the sample size gets large. If obtaining more and more data does not generally get us cloesr to the parameter of interest, then we are using a poor estimation procedure.
- The OLS estimators are inconsistent if the error is correlated with any of the independent variables.

Goodness of fit

- SST, SSE and SSR is defined exactly as in the simple regression case.
- Which means that the R^2 is defined the same as in the regression with one regressor.
- However R^2 never decrease and typically increase when you add another regressor as you explain at least as much as with one regressor.
- This means that an increased R^2 not necessarily means that the added variable improves the fit of the model.

The adjusted R-squared

- The adjusted R-squared is introduced in MLRM to compensate for the increasing R-squared.
- The adjusted R-squared includes a "penalty" for including another regressor thus $\bar{R}^2$ does not necessarily increase when you add another regressor.

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1} \right) \frac{SSR}{TSS} \tag{2}$$

Properties of $\bar{R}^2$

- Since $\frac{n-1}{n-k-1} > 1$ $R^2 > \bar{R}^2$
- Adding a variable may decrease or increase $\bar{R}$ depending on whether the increase in explanation is large enough to make up for the penalty
- $\bar{R}^2$ can be negative.

Note on caution about $R^2/\bar{R}^2$

- The goal of regression is not to maximize $\bar{R}^2$ (or R^2) but to estimate the causal effect.
- R^2 is simply an estimate of how much variation in y is explained by the independent variables in the population.
- Although a low R^2 means that we have not accounted for several factors that affect Y, this does not mean that these factors in u are correlated with the independent variables.
- Whether to include a variable should thus be based on whether it improves the estimate rather than whether it increase the fraction of variance we can explain.
- A low R^2 does imply that the error variance is large relative to the variance of Y, which means we may have a hard time precisely estimating the β_j.
- A large error variance can be offset by a large sample size, with enough data one can precisely estimate the partial effects even when there are many unobserved factors.

The standard error of the regression

Remember that the standard error of the regression (SER) estimates the standard deviation of the error term u_i:

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2} \text{ where } s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1} \qquad (3)$$

The only difference from the SLRM is that the number of regressors k is included in the formula.

Heteroskedasticity and OVB

- Pure heteroskedasticity is caused by the error term of a correctly specified equation.
- Heteroskedasticity is likely to occur in data sets in which there is a wide disparity between the largest and smallest observed values.
- Impure heteroskedasticity is heteroskedasticity caused by an error in specification, such as an omitted variable.

Effects of data scaling on OLS

Consider an example

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc$$

where:

- bwght = child birth weights, in ounces.
- cigs = number og cigarettes smoked by the mother while pregnant, per day
- faminc = annual family income, in thousands of dollars

Effects of data scaling on OLS

```
1 . reg bwght cigs faminc

      Source |       SS           df       MS      Number of obs   =      1388
-------------+----------------------------------   F(  2,  1385)   =     21.27
       Model |  17126.2088         2  8563.10442   Prob > F        =    0.0000
    Residual |  557485.511      1385  402.516614   R-squared       =    0.0298
-------------+----------------------------------   Adj R-squared   =    0.0284
       Total |   574611.72      1387  414.283864   Root MSE        =    20.063

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.4634075   .0915768    -5.06   0.000    -.6430518   -.2837633
      faminc |   .0927647   .0291879     3.18   0.002     .0355075    .1500219
       _cons |   116.9741   1.048984   111.51   0.000     114.9164    119.0319
------------------------------------------------------------------------------
```

Effects of data scaling on OLS

Alternatively you can specify the model in pounds so that $bwghtlbs = bwght/16$ Then:

$$bw\hat{ght}/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16) * cigs + (\hat{\beta}_1/16) faminc$$

- So it follows from previous lectures that each new coefficient will be the corresponding old coefficient divided by 16.
- Once the effects are transformed into the same units we get exactly the same answer, regardless of how the dependent variable is measured.
- It has no effect on the statistical significance. The t-statistic is independent, but the standard errors are scaled with the coefficient.

Effects of data scaling on OLS

Alternatively one could measure cigs in cigarette packs instead. Then:

$$\hat{bwght} = \hat{\beta}_0 + 20\hat{\beta}_1 (cigs/20) + \hat{\beta}_2 faminc \hat{bwght} = \hat{\beta}_0 + 20\hat{\beta}_1 (packs) + \hat{\beta}_2 faminc$$

The only effect is that the coefficient on packs is 20 times higher than the coefficient on cigarettes, and so will the standard error be.

Effects of data scaling on OLS

The below figure show the three regressions including the goodness of fit measures.

TABLE 6.1 Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Independent Variables			
cigs	−.4634 (.0916)	−.0289 (.0057)	—
packs	—	—	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	.0298	.0298	.0298
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

- The R^2 from the three regressions are the same (as they should be)
- The SSR and SER are different in the second specification.
- Actually SSR is 256 (16^2) larger in one and three than two.
- And SER is 16 times smaller in two than in one and three.
- Because SSR is measured in squared units of the dependent variable, while SER is measured in units of the dependent variable.
- Thus we have not reduced the error by chaning the units.

Measuring effects in standard deviations

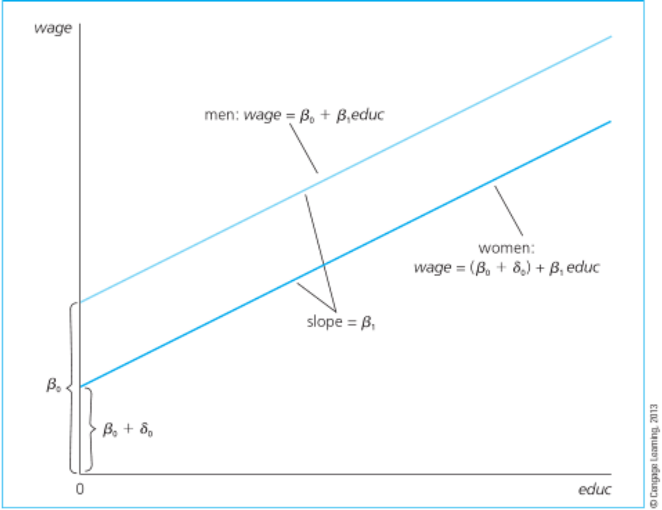
- Sometimes a key variable is measured on a scale that is difficult to interpret.
- An example is test score in labor economists wage equations which can be arbitrarily scored and hard to interpret.
- Then it can make sense to ask what happens if test score is one standard deviation higher.
- A variable is standardized by subtracting off its mean and dividing by the standard deviation.
- You can make a regression where the scale of htm regressors are irrelevant by standardizing all the variables in the regression.

Dummy variables in MLRM

- The multiple regression model allows for using several dummy independent variables in the same equation.
- In the multiple regression model a dummy variable gives an intercept shift between the groups.
- If the regression model is to have different intercepts for, say, g groups or categories, we need to include g-1 dummy variables in the model along with an intercept.
- The intercept for the base group is the overall intercept in the model
- The dummy variable coefficient for a particular group represents the estimated difference in intercepts between that group and the base group.
- An alternative is to suppress the intercept, but it makes it more cumbersome to test for differences relative to a base group.

Dummy variables in MLRM

FIGURE 7.1 Graph of $wage = \beta_0 + \delta_0\, female + \beta_1\, educ$ for $\delta_0 < 0$.



Dummy variables in MLRM

- Variables with are ordinal can either be entered to the equation in its form or you can create a dummy variable for each of the values.
- Creating a dummy variable for each value allow the movement between each level to be different so it is more flexible than simply putting the variable in the model.
- F.ex you can have a credit rate ranking between 0 and 4. Then you can include 4 dummy variables in your regression.