ECON3150/4150 Spring 2016

Lecture 6
Multiple regression model

Siv-Elisabeth Skjelbred

University of Oslo

February 5th
Last updated: February 3, 2016

Outline

- Multiple linear regression model and OLS
  - Estimation
  - Properties
  - Measures of fit
- Data scaling
- Dummy variables in MLRM

Estimation of MLRM

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.

$$min \sum_{i=1}^{n} \hat{u}_i^2$$

- 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 is the coefficient on variable number j.
- In the general form we have k independent variables, thus k+1 first order conditions.

OLS estimation of MLRM

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 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.

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

- 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.
- Under the OLS assumptions, including homoskedasticity, the OLS estimators $\hat{\beta}_j$ are the best linear unbiased estimators of the population parameter β_j.
- Under heteroskedasticity the OLS estimators are not necessarily the one with the smallest variance.

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.*

- 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 closer 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.

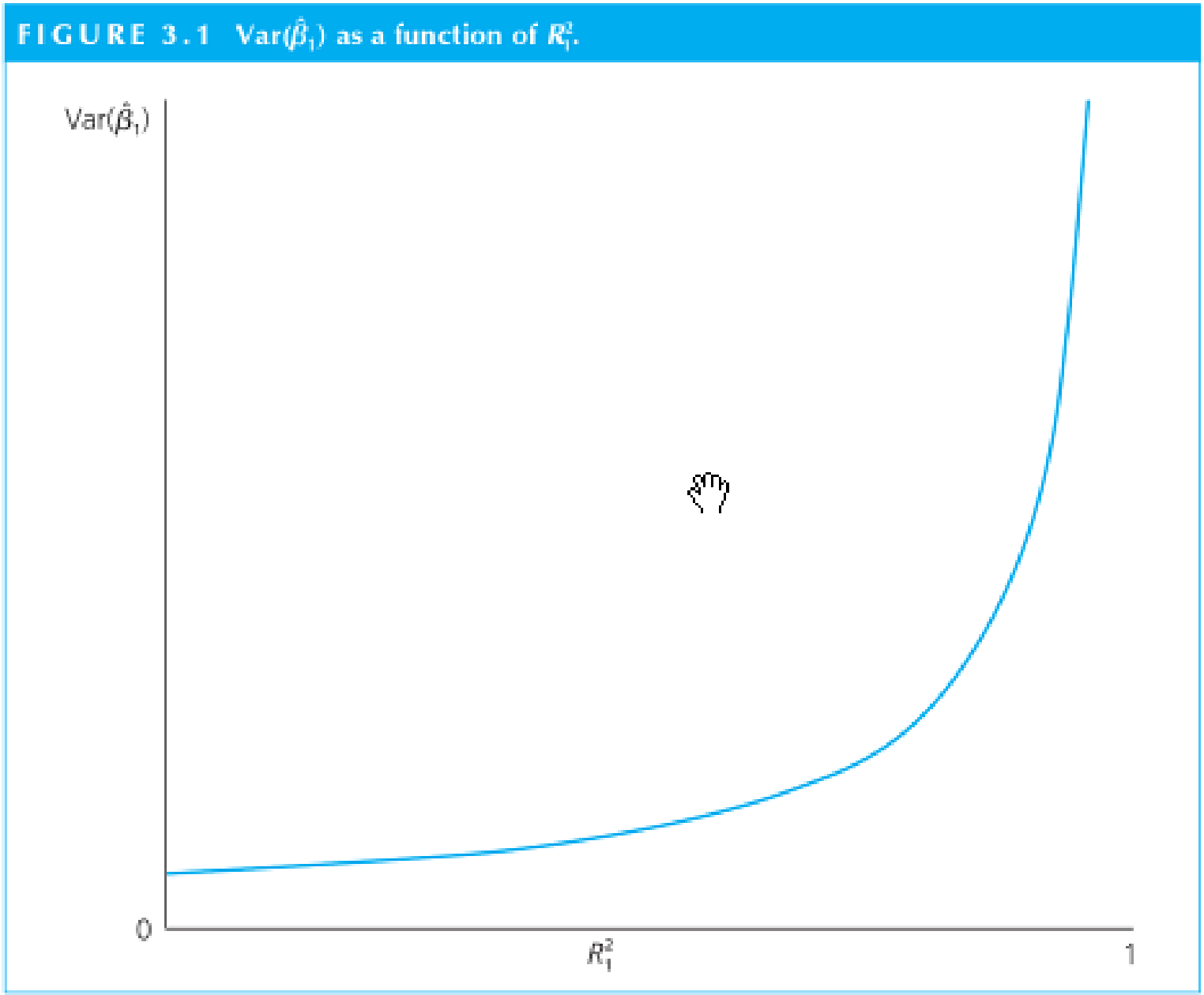
Variance of the OLS estimator

Under the OLS assumptions, conditional on the sample values of the independent variables:

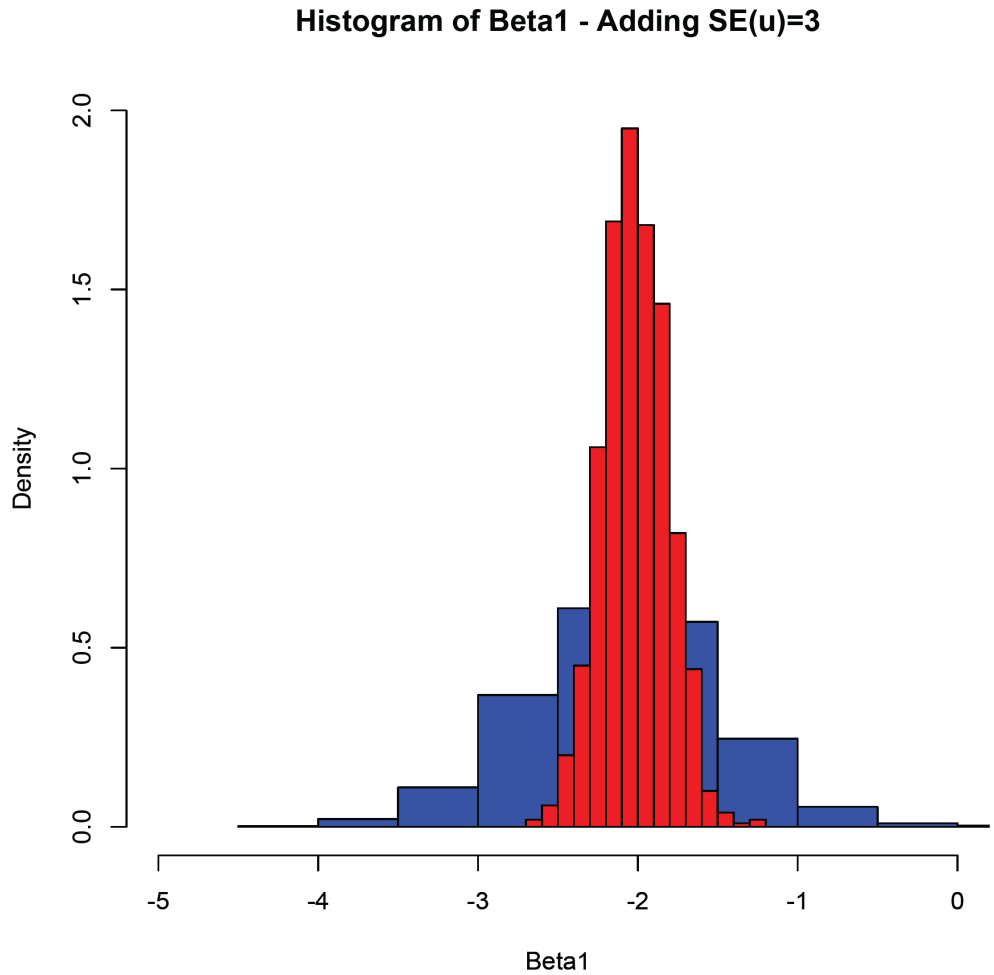
$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2(1 - R_j^2)}, j = 0, 1, 2, ..., k,$$

- Where R_j^2 is the R-squared from regressing x_j on all other independent variables.
- 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



Variance of the OLS estimator



Blue: error distributed normal with mean 0 and standard deviation 10.
Red: Error distributed normal with mean 0 and standard deviation 3.

Estimate variance

If σ^2 is not known we need to estimate it:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2$$

- Higher variance gives higher standard errors, lower precision due to more noise.

Added OLS assumption

Normality assumption

The population error u is independent of the explanatory variables and is normally distributed with zero mean and variance σ^2 : $u \sim Normal(0, \sigma^2)$

If the other OLS assumptions plus this one holds the OLS estimator has an exact normal sampling distribution and the homoskedasticity only t-statistic has an exact Student t distribution.

- The OLS estimators are jointly normally distributed.
- Each $\hat{\beta}_j$ is distributed $N(\beta_j, \sigma^2_{\hat{\beta}_j})$.

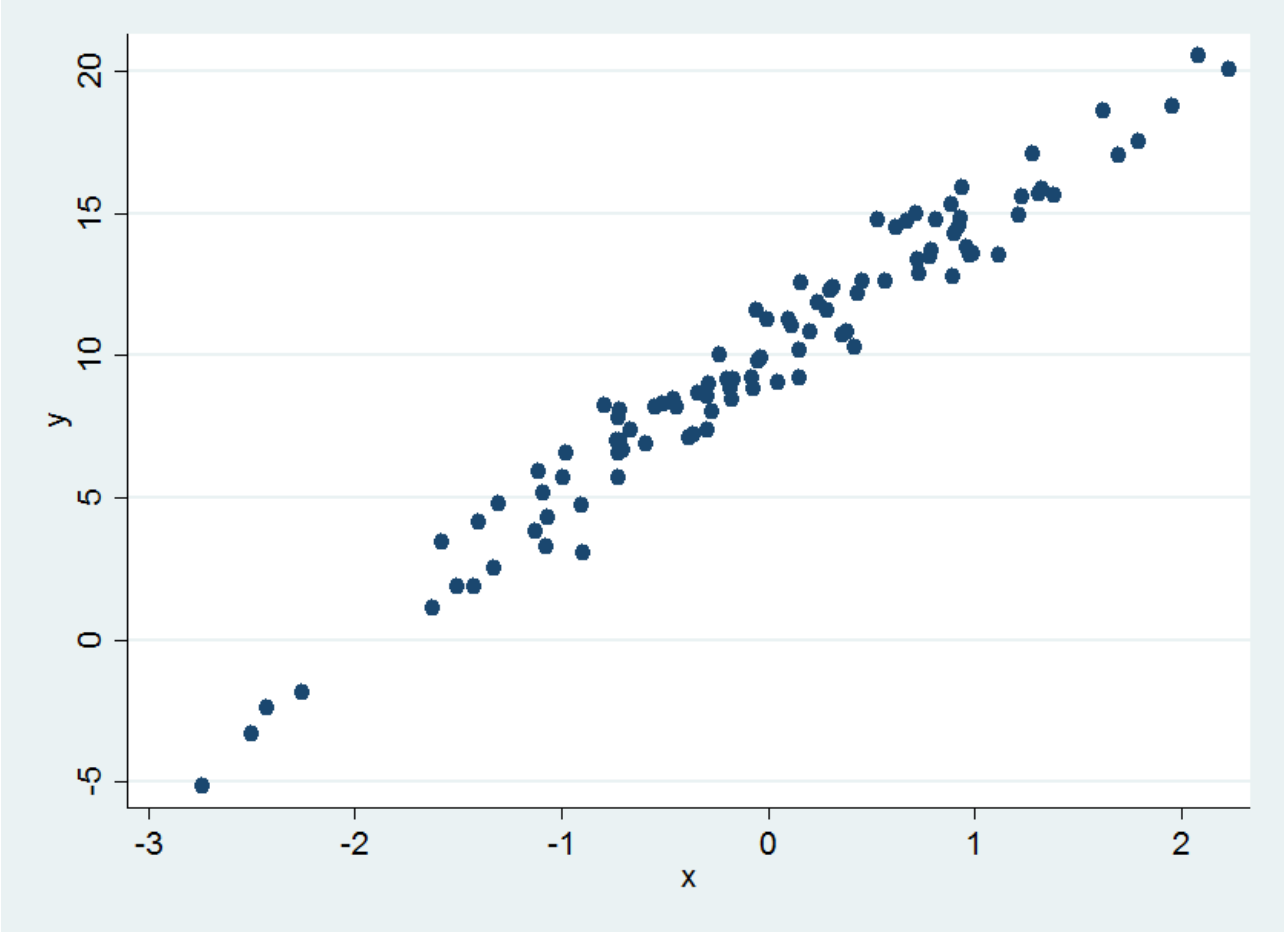
Simulation

A simulation is a fictitious computer representation of reality. Steps in simulation:

1. Choose the sample size n
2. Choose the parameter values and functional form of the population regression function.
3. Generate n values of x randomly in Stata
4. Choose probability distribution of the error term and generate n values of u
5. Estimate the model
6. Repeat step 1 through 5 multiple times and look at the summary statistics over the repetitions.

Monte Carlo simulation

Example: A random realization of X and u for 100 observations with the true population function: $Y = 10 + 5x + u$.



How does OLS perform in estimating the underlying population function?

Simulation

```
1 . reg y x
```

Source	SS	df	MS
Model	**2438.45884**	**1**	**2438.45884**
Residual	**100.53965**	**98**	**1.02591479**
Total	**2538.99849**	**99**	**25.6464494**

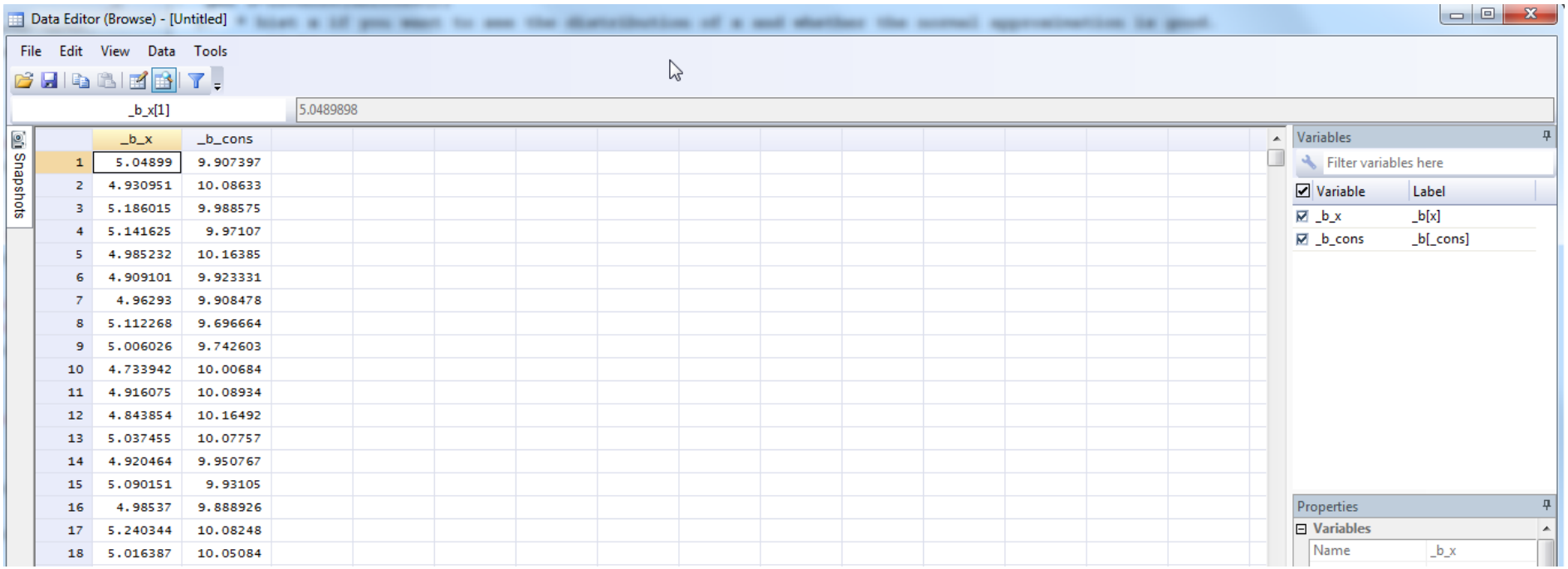
Number of obs =	**100**
F(1, 98) =	**2376.86**
Prob > F =	**0.0000**
R-squared =	**0.9604**
Adj R-squared =	**0.9600**
Root MSE =	**1.0129**

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	**4.86004**	**.0996868**	**48.75**	**0.000**	**4.662214**	**5.057865**
_cons	**9.951091**	**.1013099**	**98.22**	**0.000**	**9.750045**	**10.15214**

The coefficients are close to the true population coefficients.

Simulation

So running one simulation got us close to the estimate, how if we simulate 1000 times? Gives us 1000 estimates for β_0 and β_1



	_b_x	_b_cons
1	5.04899	9.907397
2	4.930951	10.08633
3	5.186015	9.988575
4	5.141625	9.97107
5	4.985232	10.16385
6	4.909101	9.923331
7	4.96293	9.908478
8	5.112268	9.696664
9	5.006026	9.742603
10	4.733942	10.00684
11	4.916075	10.08934
12	4.843854	10.16492
13	5.037455	10.07757
14	4.920464	9.950767
15	5.090151	9.93105
16	4.98537	9.888926
17	5.240344	10.08248
18	5.016387	10.05084

Simulation

```
1 . sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	**1000**	**5.000788**	**.1048841**	**4.679994**	**5.318002**
_b_cons	**1000**	**9.997947**	**.0994027**	**9.696664**	**10.33181**

The estimated OLS coefficients approximate to the true population coefficient. Thus OLS gives an unbiased estimate for the slope coefficient and the constant term.

Simulation: Bias

Another example. Suppose that the population is characterized by:

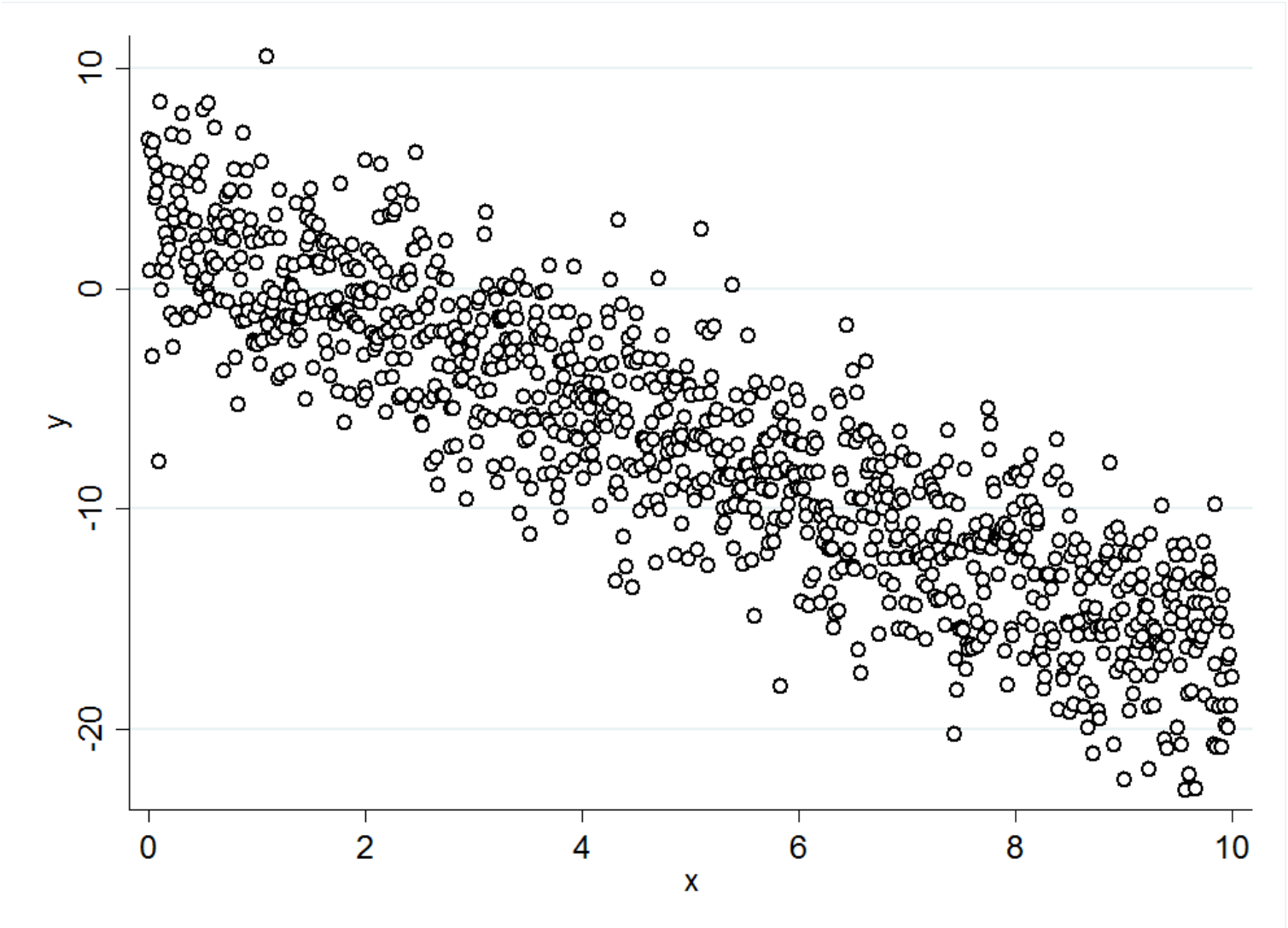
$$y = 3 - 2x_1 + u$$

- $\beta_0 = 3$
- $\beta_1 = -2$
- u is distributed normal, mean 0 and standard deviation 3.
- x's are between 0.01 and 10 spaced evenly
- n= 1000

Estimate using $y = \beta_0 + \beta_1 x_1 + u$ and plot y on X.

Simulation: Bias

Scatterplot of the simulated data:



Simulation: Bias

How to make the scatter plot in Stata:

```
.  *drop any observations that are currently in memory.
.  set obs 1000
obs was 0, now 1000

.  *In the runs of the regression we want there to be .. observations.
.  gen u=rnormal(0,3)

.  *generates distributions of x
. range x 0 10 1000

. *Generate Y:
. gen y=3-2*x+u


. twoway (scatter y x, mfcolor(white) mlcolor(black)), graphregion(color(white))
```

Simulation: Bias

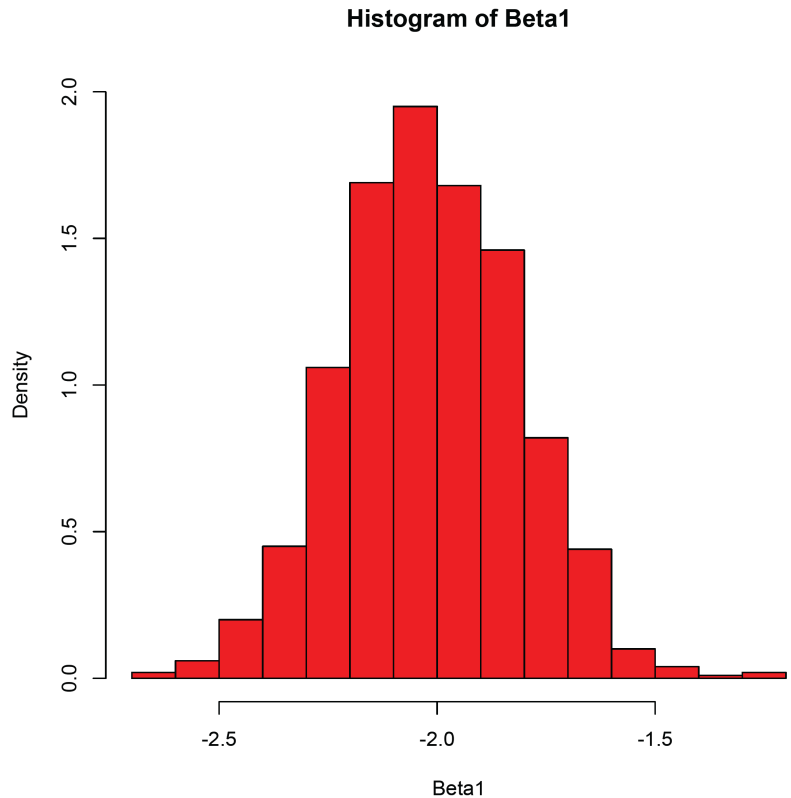
Suppose that we sample 30 people from the population and estimate β_1 via OLS

- First sample $\hat{\beta}_1 = -1.951$
- Second sample: $\hat{\beta}_1 = -1.890$
- Second sample: $\hat{\beta}_1 = -1.559$

None of them equals the population parameter of 2. Is this a problem?

Simulation: Bias

Keep sampling! If we sample 1000 times we get the following histogram of the estimates of β_1.



Variance observations

10 observations repeated on 1000 samples with model from example 2:

```
2 .  sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	**1000**	**4.997962**	**.3770747**	**3.270008**	**6.512727**
_b_cons	**1000**	**10.01276**	**.3351356**	**8.923193**	**11.20773**

100 observations repeated on 1000 samples.

```
1 . sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	**1000**	**5.000788**	**.1048841**	**4.679994**	**5.318002**
_b_cons	**1000**	**9.997947**	**.0994027**	**9.696664**	**10.33181**

Simulation

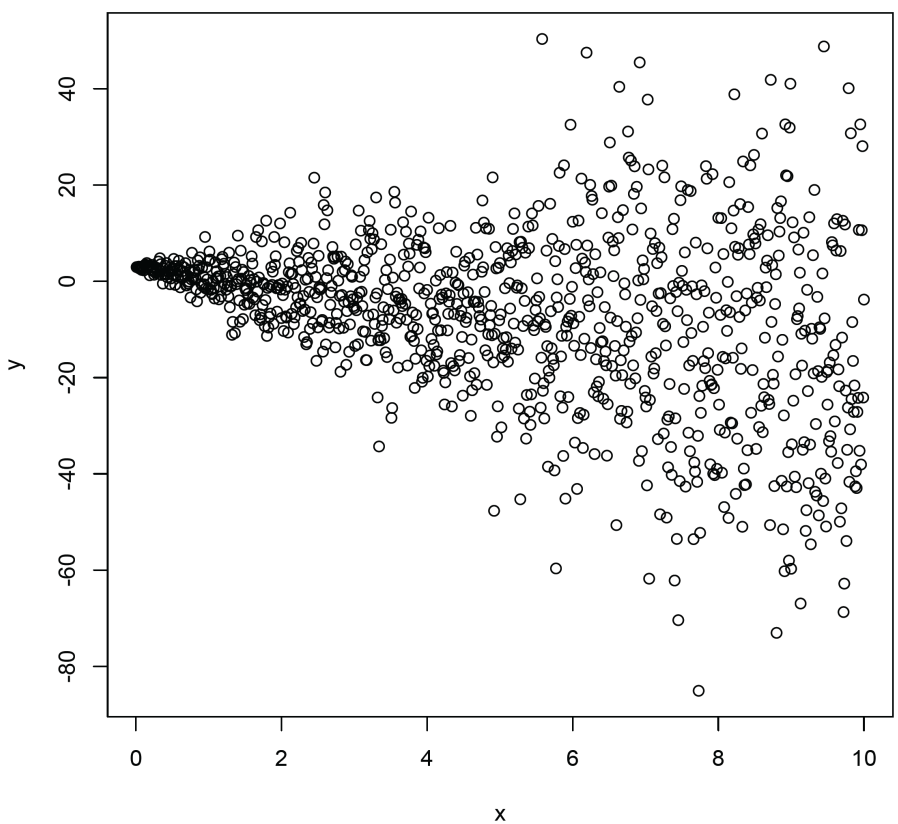
The errors are not normally distributed with mean 0, but with mean 3. So $u \sim N(3, 1)$

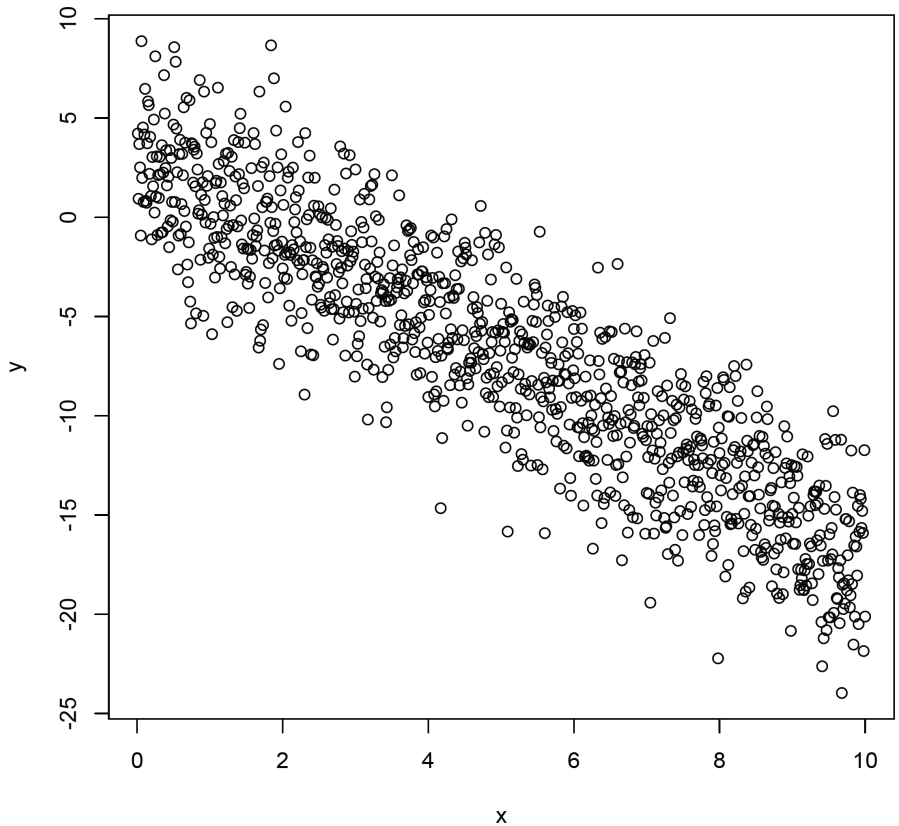
Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	**1000**	**4.998194**	**.0975124**	**4.67178**	**5.298087**
_b_cons	**1000**	**13.0041**	**.1030487**	**12.65324**	**13.33964**

- As long as X and u are uncorrelated $\hat{\beta}_1$ is unbiased.
- The constant term and the error term is correlated in this situation so β_0 is biased.

Simulation: Heteroskedasticity

By assumption the variance of errors is common across x. (homoskedasticity assumption).





Which graph illustrates a violation of this assumption?

Measures of fit

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.

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

Properties of $\bar{R}^2$

- Since $\frac{n-1}{n-k-1} < 1 \rightarrow R^2 > \bar{R}^2$
- 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:

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

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

Heteroskedasticity and OVB

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

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.

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.

Measurement of variable

Effects of data scaling on OLS

Consider an example

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

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

using bwght.dta

Effects of data scaling on OLS

```
1 . reg bwght cigs faminc
```

Source	SS	df	MS
Model	**17126.2088**	**2**	**8563.10442**
Residual	**557485.511**	**1385**	**402.516614**
Total	**574611.72**	**1387**	**414.283864**

Number of obs =	**1388**
F(2, 1385) =	**21.27**
Prob > F =	**0.0000**
R-squared =	**0.0298**
Adj R-squared =	**0.0284**
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 grams so that $bwghtgram = bwght28.35$ Then:

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

- So it follows from previous lectures that each new coefficient will be the corresponding old coefficient divided by 16.
- If you wanted grams each coefficient would be multiplied by 28.349
- Once the effects are transformed into the same units we get exactly the same answer, regardless of how the dependent variable is measured.

Effects of data scaling on OLS

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

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

$$\widehat{bwght} = \hat{\beta}_0 + 20\hat{\beta}_1(packs) + \hat{\beta}_2 faminc \quad (3)$$

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 differ in the second specification from the two others.
- Remember SSR is measured in squared units of the dependent variable, while SER is measured in units of the dependent variable.

Standardizing variables

- Sometimes a key variable is measured on a scale that is difficult to interpret.
- An example is test scores - tests can be arbitrarily scored
- 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 the regressors are irrelevant by standardizing all the variables in the regression.

Standardizing variables

Suppose we start with the model:

$$Y = \beta_0 + \beta_1 X + u \tag{4}$$

Take the mean of the entire equation:

$$\hat{\mu}_u = \beta_0 + \beta_1 \hat{\mu}_x \tag{5}$$

Subtract 1 from 2 and

$$(Y - \hat{\mu}_y) = \beta_1 (x - \hat{\mu}_x) + u$$

Divide both sides by $\hat{\sigma}_y$ and multiply β_1 by $\hat{\sigma}_x / \hat{\sigma}_x$ and manipulate such that:

$$\frac{(y - \hat{u}_y)}{\hat{\sigma}_y} = \frac{\hat{\sigma}_x}{\hat{\sigma}_y} \beta_1 \frac{(x - \hat{\mu}_x)}{\hat{\sigma}_x} + \frac{1}{\hat{\sigma}_y} u$$

$$y^s = \tilde{\beta}_1 x^s + \tilde{u}$$

Standardizing variables

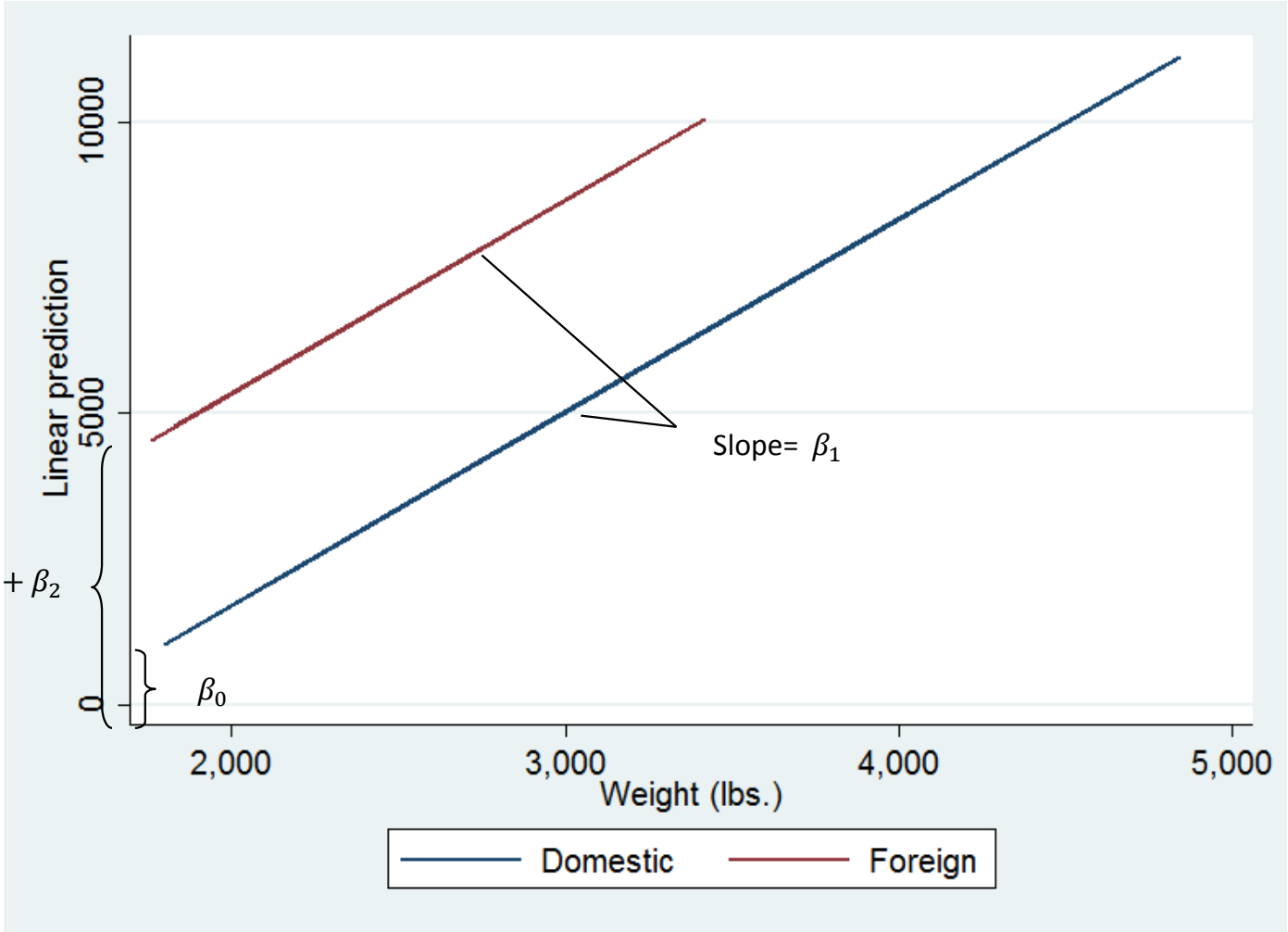
- Note: the coefficients and standard errors change
- The t-statistics, p-values and R^2 do not change
- Standardizing variables allows a comparison of size of coefficients across variables.

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

$$price = \beta_0 + \beta_1 weight + \beta_2 Foreign$$



Dummy variables in MLRM

- Variables which 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.

Alternative specification OLS

The OLS minimization problem:

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

can alternatively be written as:

$$S(\alpha, \beta_1) = \sum_{i=1}^{n} (Y_i - \alpha - \beta_1 (X_i - \bar{X}))^2$$

where the intercept parameter is redefined to: $\alpha = \beta_0 + \beta_1 \bar{X}$

Alternative specification OLS

$$\frac{\partial S(\alpha, \beta_1)}{\partial \alpha} = -2 \sum_{i=1}^{n} [Y_i - \alpha - \beta_1 (X_i - \bar{X})]$$

$$\frac{\partial S(\alpha, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} [Y_i - \alpha - \beta_1 (X_i - \bar{X})] * (X_i - \bar{X})$$

$\hat{\alpha}$ and $\hat{\beta}_1$ are the values of α and β_1 for which the FOC is equal to zero.
Solution:

$$\hat{\alpha} = \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

and β_1 as before.