Topic 2: Revision of Simple Regression

Quick Revision For Simple Regression

There are causal relationships which influence the interaction between actors in society and in the economy. These relationships are not exact. They are not predictable in the way that, say, the force of gravity is predictable.

One reason that they are not exact, is that we observe only a limited selection of the influences that affect behaviour.

The investigator is only able to observe a sample from the population. He/she uses the sample to make inferences about the underlying population.

Simple Linear Model

Population Model

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad i = 1.....n$$

The index *i* indicates the relationship for a representative individual. The sample has size *n*. The observable variables are *y* and *x* and the unknown population parameters are β_1 and β_2. β_1 is known as the constant and β_2 is known as the slope. The unobserved disturbance term is ε_i

Estimated Relationship

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + e_i \qquad i = 1.....n$$

Simple Linear Model: Terminology

Unknown population parameters

Constant Slope Disturbance

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad i = 1..n$$

Dependent variable

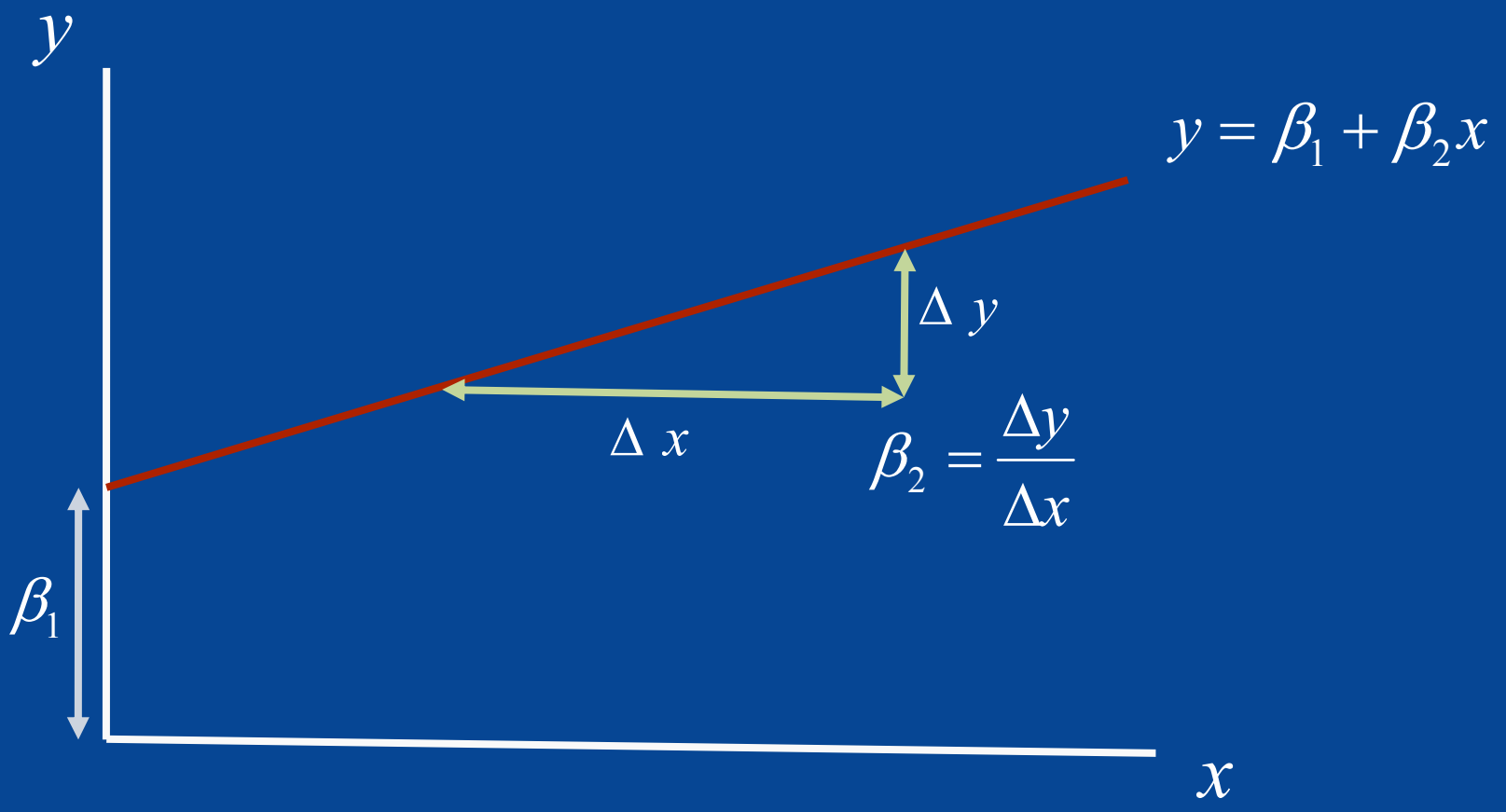
Independent variable

Sample size

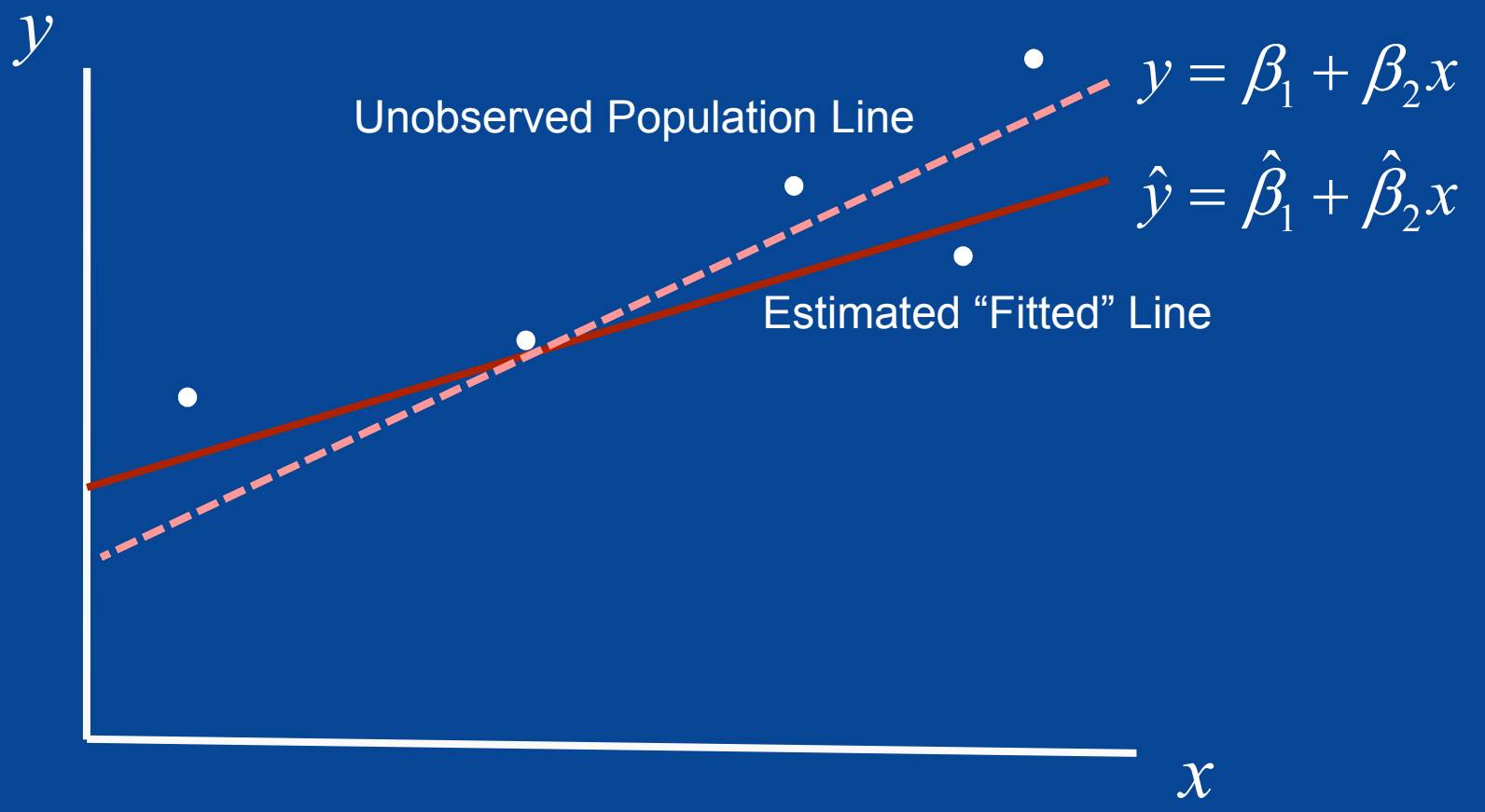
Disturbance *i* assumed to be distributed normally with constant variance (homoscedastic) and uncorrelated with other disturbances (non-autocorrelated) – written as:

$$\varepsilon_i \sim N(0, \sigma^2)$$

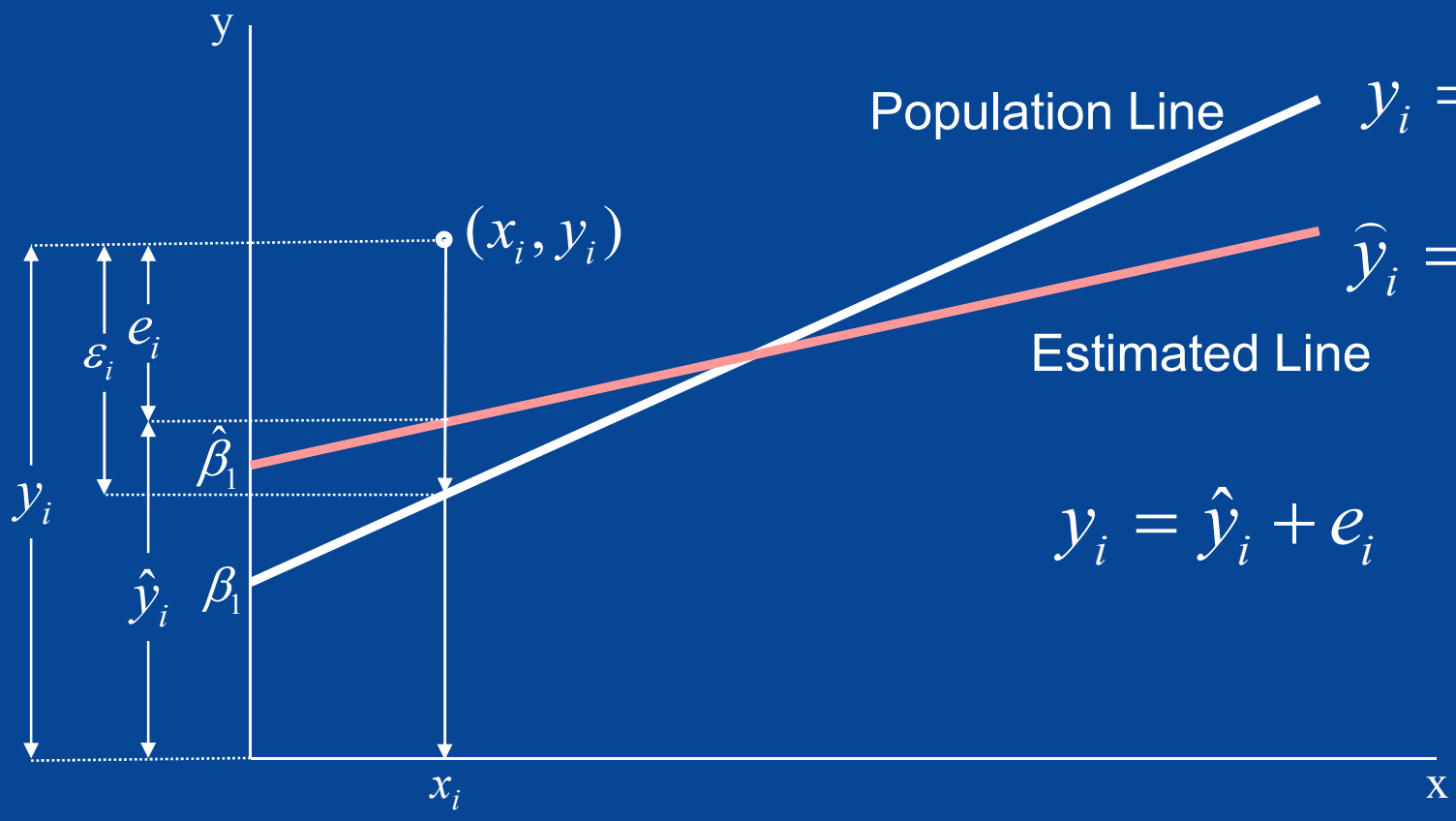
Graphical Representation



Graphical Representation



Graphical Representation



Five assumptions about the disturbance - ε_i

1. Mean is zero $E(\varepsilon_i)=0$
2. Variance is constant $Var(\varepsilon_i)=\sigma_\varepsilon^2$
3. Mutually independent $Cov(\varepsilon_i \varepsilon_j)=0$
4. Independent of x variables $Cov(x_i \varepsilon u_i)=0$

To be able to conduct hypothesis tests, we add:

5. Assuming the Central Limit Theorem holds, then the disturbances are normally distributed $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

Estimation Methods

- Least Squares (Minimise the Sum of Squared Errors)
- Maximum Likelihood (What parameters are most likely to have generated this sample, given the assumed distribution of the disturbances?)
- General Method of Moments (Equate sample moments with unobservable population moments and then solve resulting equations)

OLS is BLUE

If these five assumptions hold, then the least squares estimator is BLUE (Best Linear Unbiased Estimator)

This result is also known as the Gauss-Markov Theorem

Unbiasedness

$$E(\hat{\beta}_2) = E\left(\beta_2 + \frac{\text{cov}(xu)}{\text{var}(x)}\right)$$

$$= \beta_2$$

Efficiency

- The least squares estimate is “best” in the sense that it has the smallest possible variance in the class of linear estimators

Consistency

- As the sample size increases, the estimates of the population parameters converge on their true values.

Dummy Variables

Sometimes we need to take account of *qualitative* factors in a regression (things that have categories rather than numbers associated with them e.g. day of the week, gender, ethnicity, level of qualification)

Dummy Variables

You could run different regressions for each *state* or category of the qualitative variable.

May not have enough observations to do this

Alternatively assume that the slope is the same in each *state*, but there is a step change in the *constant* each time the state changes

Dummy Variables

How do we do this?

$$y_i = \beta_1 + \beta_2 x + \beta_3 \text{Female} + \varepsilon_i$$

$$y_i = \beta_1 + \beta_2 x + \varepsilon_i$$

We estimate the shift δ by introducing a *dummy* variable, *Female*, which takes the value 0 if the sampled individual is male and the value 1 if female.

Thus we actually estimate: $y_i = \beta_1 + \beta_2 x + \beta_3 \text{Female} + \varepsilon_i$

The result might come out like: $y = 2 - 3x + 5\text{Female}$

Dummy Variables

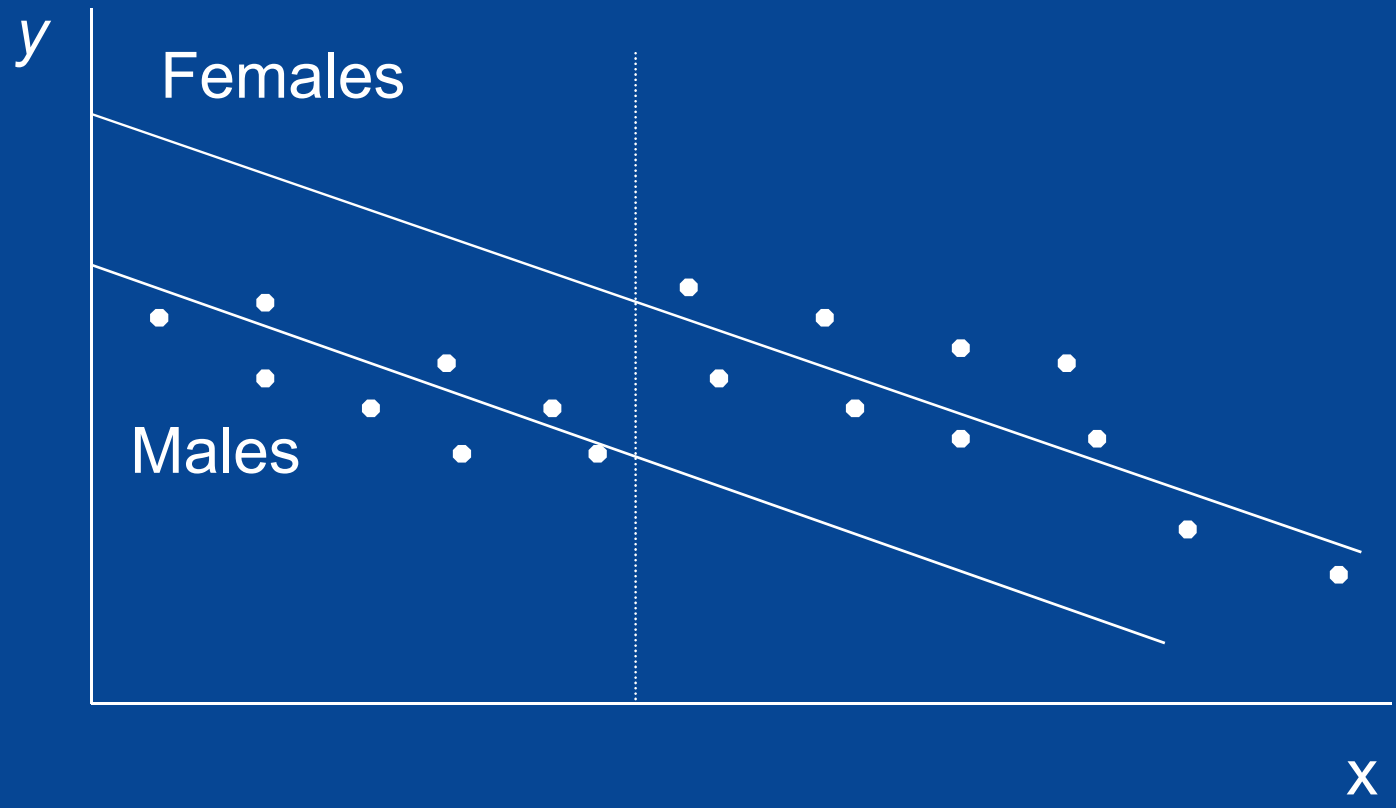
Can say there is a significant difference between males and females? The usual hypothesis testing procedure applies:

So we would test

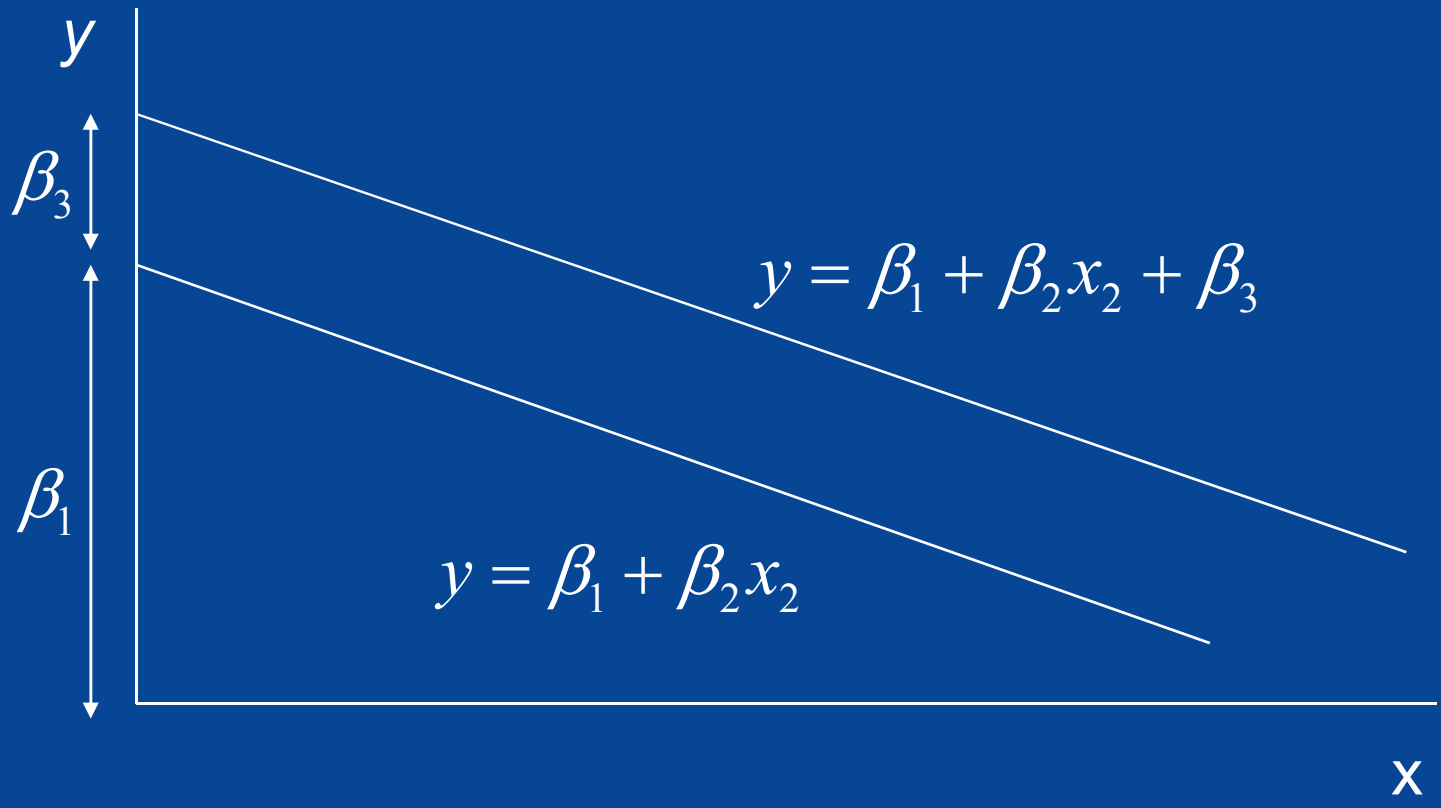
$H_0: \beta_3 = 0$ against

$H_1: \beta_3 \neq 0$.

Dummy Variables



Dummy Variables



Dummy Variables

The previous example has two **categories**

We may have many more - e.g. days of the week, firms, industries, countries.
For these we need a **set** of dummy variables.

We usually include a constant and have one fewer dummy than we have categories e.g. suppose we are considering the **five** working days in the week. We would require **four** dummies.

We can arrive at the same results by **excluding the constant** and having as many dummies as there are categories.

Dummy Variables

Suppose you observe earnings for a sample comprising equal numbers of men and women. You have no explanatory variables and estimate a model trying to explain earnings on the basis of gender alone. What would the least squares estimator for the gender variable in this model be?

$$y_i = \beta_1 + \beta_2 F_i + \varepsilon_i \quad i = 1 \ldots n$$

$$\hat{\beta}_2 = \frac{\sum (y_i - \bar{y})(F_i - \bar{F})}{\sum (F_i - \bar{F})^2} = \frac{(1/2)\left(\sum_{FEM} (y_i - \bar{y}) - \sum_{MALE} (y_i - \bar{y}) \right)}{n/4}$$

$$= \frac{2*(n/2)(\bar{y}_{FEM} - \bar{y}_{MALE})}{n} = (\bar{y}_{FEM} - \bar{y}_{MALE})$$

$$\hat{\beta}_1 = \bar{y} - (\bar{y}_{FEM} - \bar{y}_{MALE})/2 = (\bar{y}_{FEM} + \bar{y}_{MALE})/2 - (\bar{y}_{FEM} - \bar{y}_{MALE})/2$$

$$= \bar{y}_{MALE}$$

Omitted Variables

Population relationship: $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$

Estimated relationship: $y_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + e_i$

We have wrongly omitted the explanatory variable x_3 from the relationship we have estimated.

Does this matter?

Yes - it can cause **omitted variables bias**.

Omitted Variable Bias

The omitted variable x_3 will have an *indirect* effect through x_2 on the dependent variable. The size of this effect depends on the relationship between x_2 and x_3 as shown above.

If they are unrelated, then omission of this variable will have no impact on the estimate of the coefficient on x_2 .

If they are related, you will get a biased estimate of this coefficient - because x_2 is now doing two tasks:

- having its own independent effect on *y*
- proxying the effect of x_3 on *y*

Omitted Variable Bias

Suppose $x_{3i} = g + hx_{2i} + v_i$

then $\hat{h} = \dfrac{\text{cov}(x_2 x_3)}{Var(x_2)}$

$$and\ y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 (g + hx_{2i} + v_i) + u_i$$

$$= \beta_1^* + (\beta_2 + \beta_3 h) x_{2i} + u_i^*$$

and the coefficient on x_2 estimates $\beta_2 + \beta_3 h$

and will be given by $\hat{\beta}_2 + \hat{\beta}_3 \dfrac{\text{cov}(x_2 x_3)}{\text{var}(x_2)}$