ECN 102, Summer 2020

Week 3 Recap
Correlation and Regression

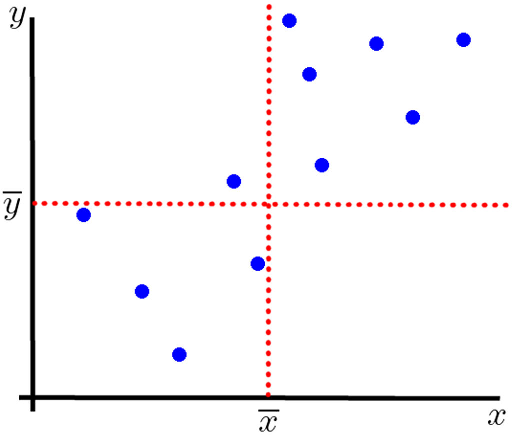
Correlation

- Sample correlation between x and y is given by

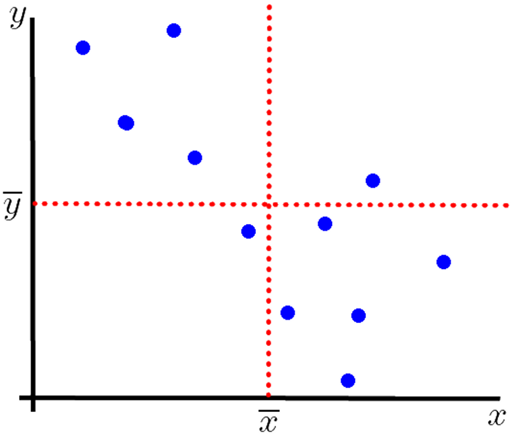
$$r_{xy} \equiv \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- Positive if X and Y both tend to be above their means simultaneously; and if both tend to be below their means simultaneously
- Negative if X tends to be above its mean when Y is below its mean; and vice versa
- $r_{xy} = 1$ is a perfect positive correlation, $r_{xy} = -1$ is perfect negative correlation
- ρ_{xy} is the population correlation

Positive Correlation



Negative Correlation



Correlation Test

- Two-tailed example: $H_0 : \rho_{xy} = 0$ against $H_1 : \rho_{xy} \neq 0$
- Therefore rejecting null hypothesis means we conclude non-zero correlation
- Use test-statistic

$$t \equiv \frac{r_{xy} - 0}{\sqrt{\frac{1-r_{xy}^2}{n-2}}} \sim T(n-2)$$

- Then interpret t-statistic in the same way as before: find a critical value or a p-value
- Can also use `cor.test()` function in R

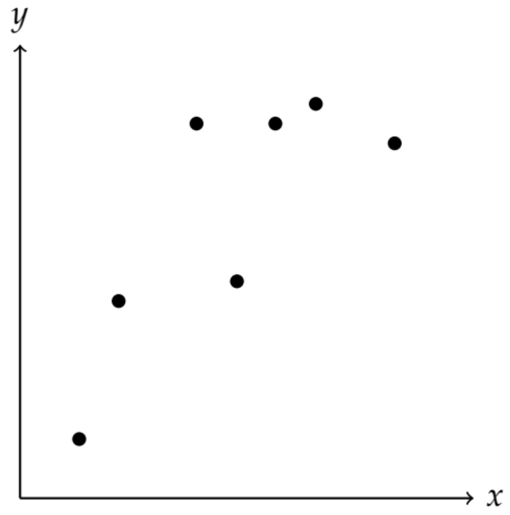
Population Regression

- Suppose we have population data for both X and Y
- The regression line is the line of best fit
- It's a line, so it can expressed in terms of an intercept and a slope

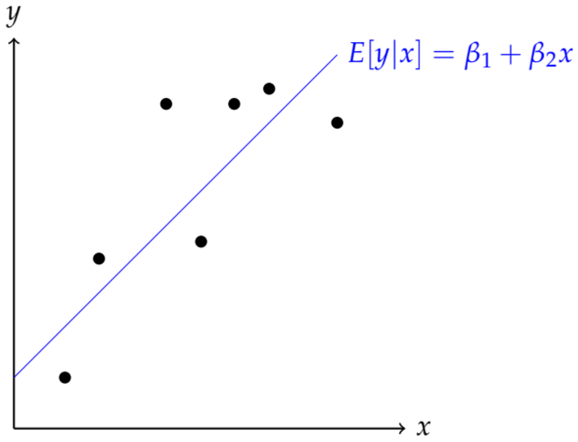
$$y = \beta_1 + \beta_2 x$$

- But it's only a line of *best* fit, usually not a line of *perfect* fit
- The difference between the actual data and the population regression line are called *disturbances*, denoted ϵ

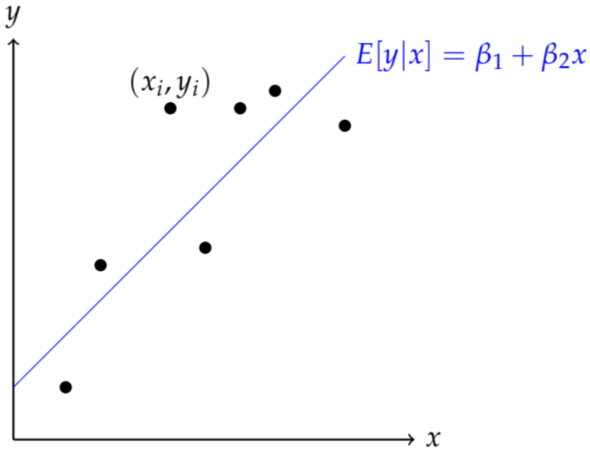
Population Regression Illustrated



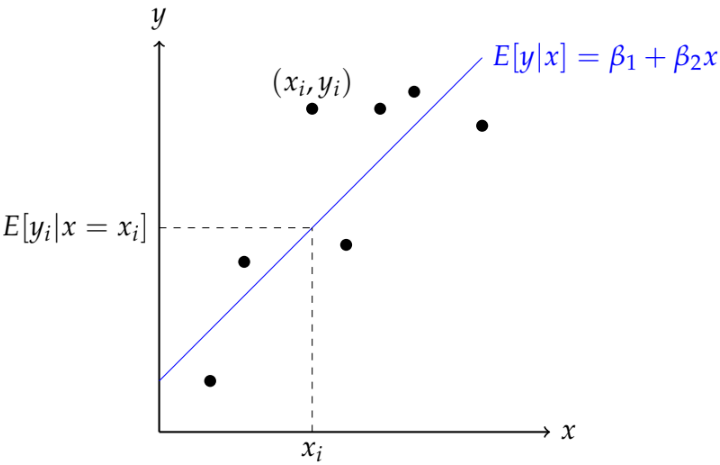
Population Regression Illustrated



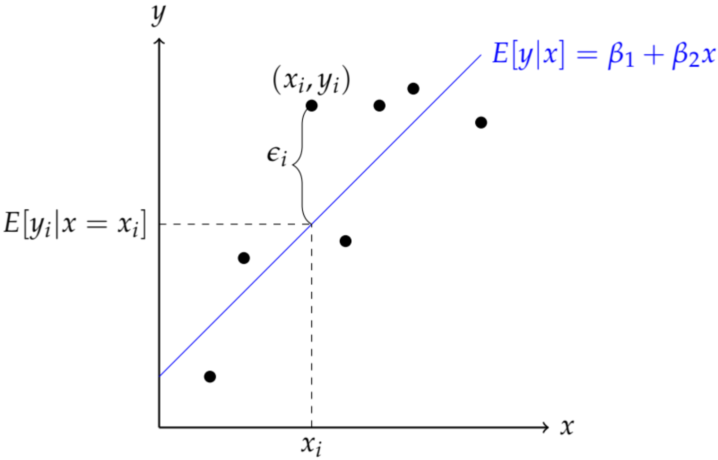
Population Regression Illustrated



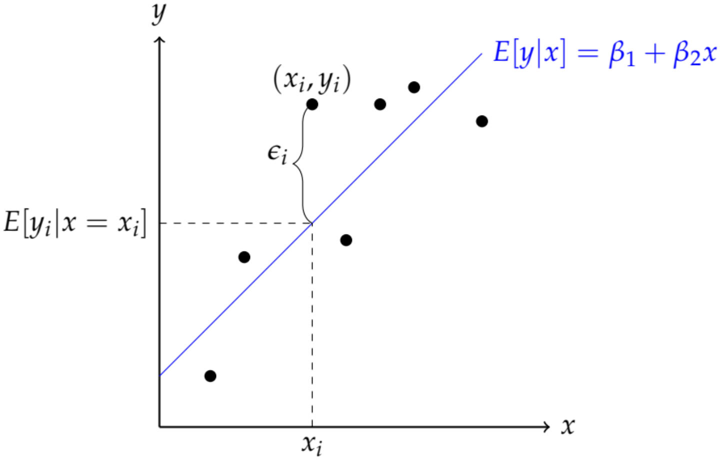
Population Regression Illustrated



Population Regression Illustrated



Population Regression Illustrated



Therefore can write $y_i = \beta_1 + \beta_2 x_i + \epsilon_i$ as exact relationship

Sample Regression

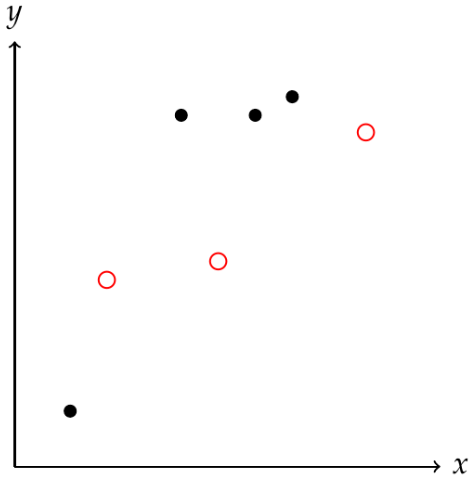
- We almost never have population data, we just have a sample
- So we do a regression line through the sample instead
- It's a line, so it can expressed in terms of an intercept and a slope

$$y = b_1 + b_2 x$$

- So b_1 is an estimate of β_1 and b_2 is an estimate of β_2
- But it's only a line of *best* fit, usually not a line of *perfect* fit
- The difference between the actual data and the sample regression line are called *residuals*, denoted e

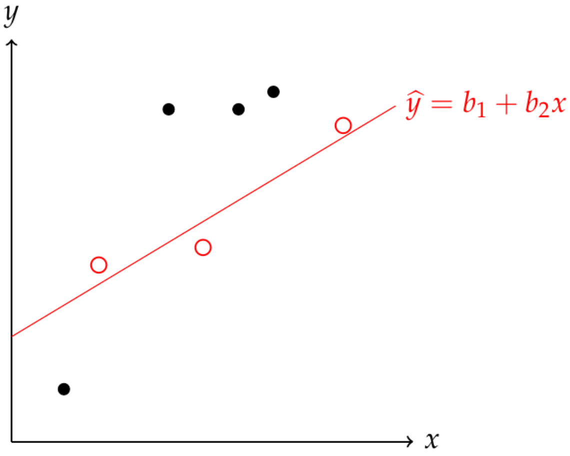
Sample Regression Illustrated

Only have sample of red dots, so regression line is a little different



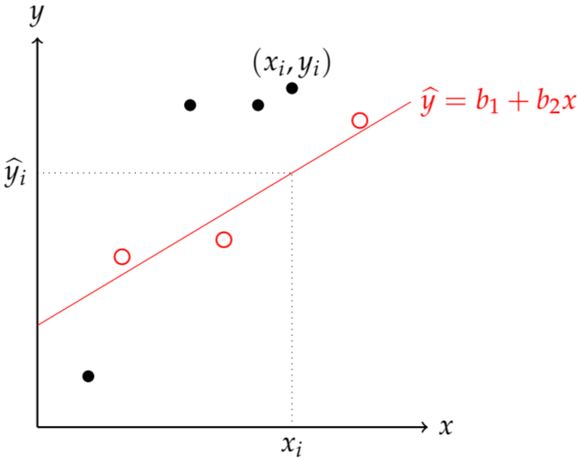
Sample Regression Illustrated

Only have sample of red dots, so regression line is a little different



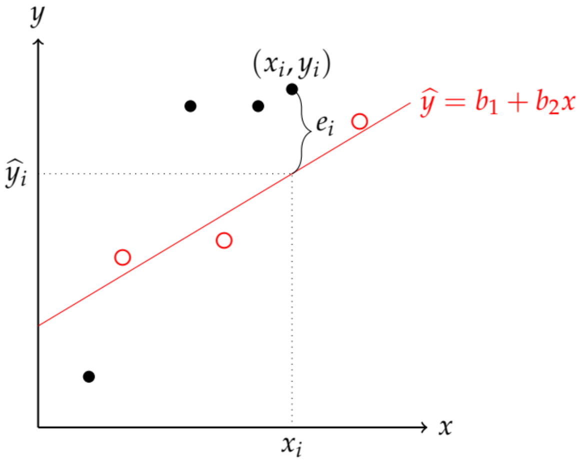
Sample Regression Illustrated

Only have sample of red dots, so regression line is a little different

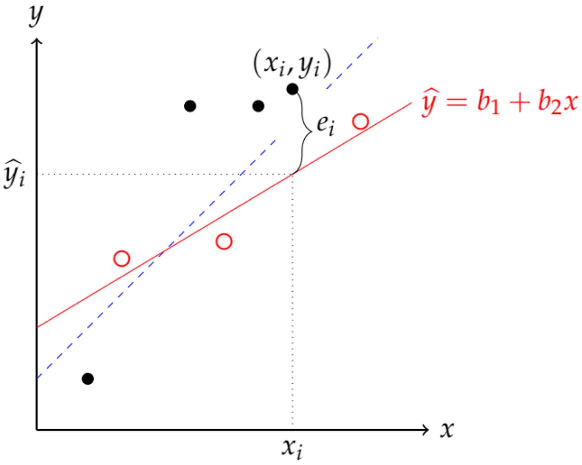


Sample Regression Illustrated

Only have sample of red dots, so regression line is a little different

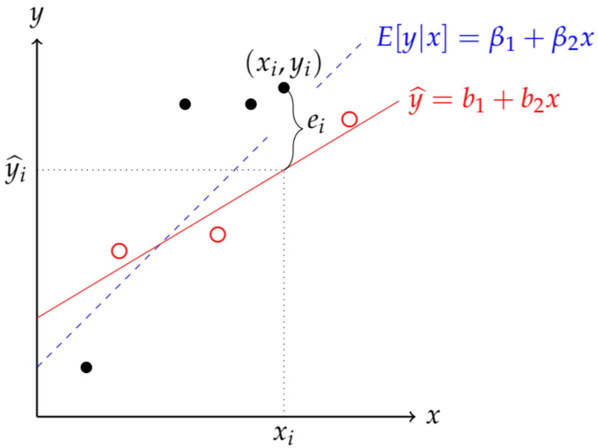


Only have sample of red dots, so regression line is a little different



Sample Regression Illustrated

Only have sample of red dots, so regression line is a little different



Therefore can write $y_i = b_1 + b_2 x_i + e_i$ as exact relationship

Finding Line of Best Fit

- The residuals capture how far the data are from the line
- Intuition: maximizing rightness is equivalent to minimizing wrongness
- The residuals capture how "wrong" the line is relative data
- So let's minimize an overall measure of the residuals
- We minimize the **residual sum of squares (RSS)**,

$$\text{RSS} \equiv \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

- This process gives b_1 and b_2, and the technique is called **ordinary least squares (OLS)**

Formulas for Line of Best Fit

Don't have to know how to minimize RSS yourself (ECN 140 stuff), so here are the results.

$$b_2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = r_{xy} \times \frac{s_y}{s_x}$$

$$b_1 = \overline{Y} - b_2\overline{X}$$

Coefficient of Determination

- Total variation in y is given by

$$\text{Total Sum of Squares (TSS)} \equiv \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Variation in y explained by x is given by

$$\text{Explained Sum of Squares (ESS)} \equiv \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

- Variation in y not explained by x is given by

$$\text{Residual Sum of Squares (RSS)} \equiv \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

- R^2 captures proportion of variation in y explained by x

$$R^2 \equiv \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{TSS} = r_{xy}^2$$

Testing Regression

- The **standard error of the regression** (sometimes called the residual standard error or root mean square error) is really just the standard deviation of the residuals, given by

$$s_e \equiv \sqrt{\frac{\text{RSS}}{n-2}}$$

- The standard error of b_2 is given by

$$\text{se}(b_2) = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

- For two-sided hypothesis $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$, we use t-statistic

$$t \equiv \frac{b_2 - 0}{\text{se}(b_2)} \sim T(n-2)$$

- R automatically tests this

Testing R-squared

- Test $H_0 : R^2 = 0$ against $H_1 : R^2 > 0$
- One-sided test, so only reject null if test statistic is sufficiently larger than 0 in the positive direction
- Use the F-statistic

$$F \equiv \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(k-1, n-k),$$

  where $k = 2$ is the number of coefficients estimated (i.e. β_1 and β_2)
- R automatically tests this as well