Simple Linear Regression

CIVL 7012/8012

Regression Analysis Recap

- We used method of moments to estimate the coefficients of a simple linear regression

So the OLS estimated slope is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$\text{provided that} \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 > 0$$

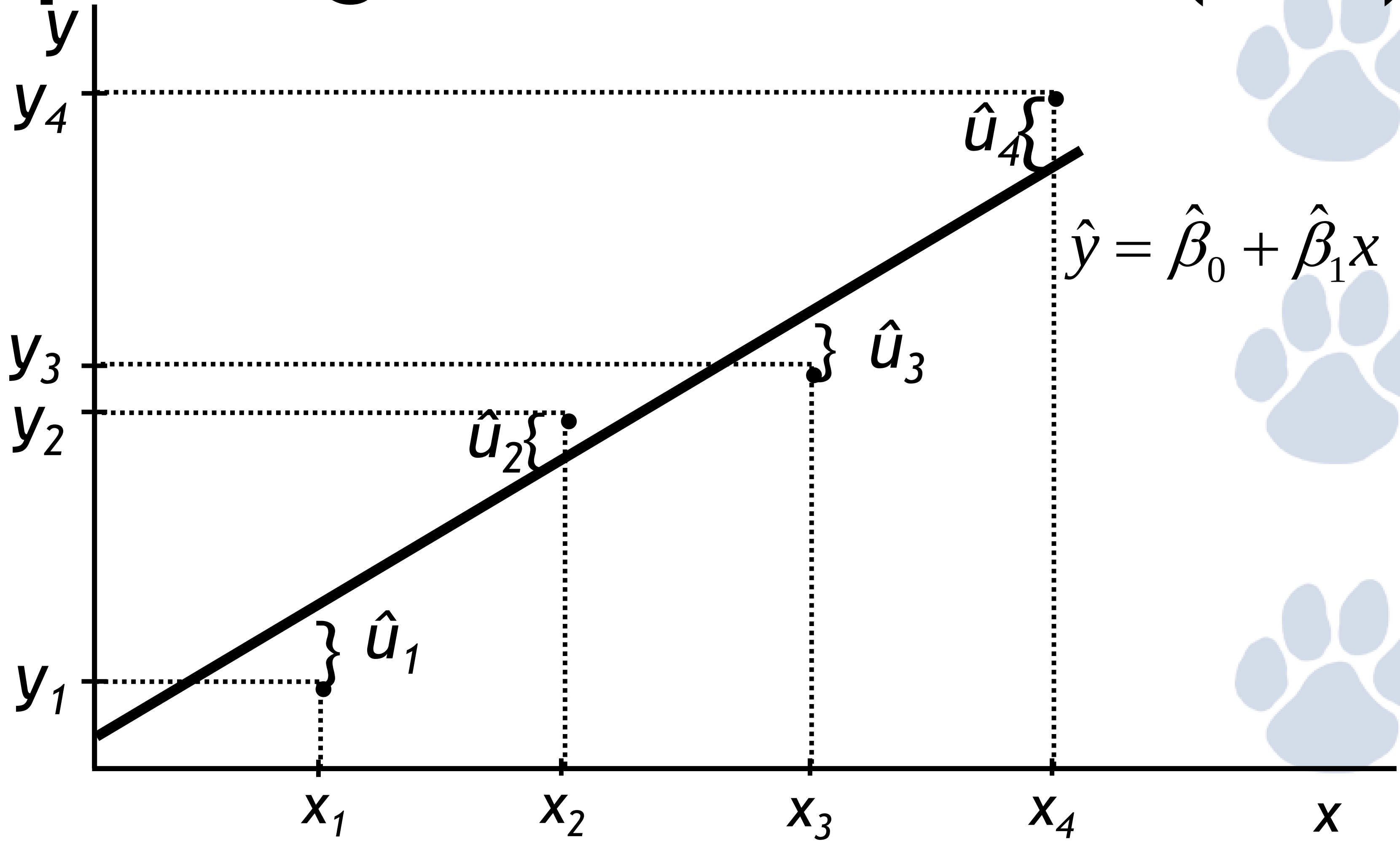
Summary of OLS slope estimate

- The slope estimate is the sample covariance between *x* and *y* divided by the sample variance of *x*
- If *x* and *y* are positively correlated, the slope will be positive
- If *x* and *y* are negatively correlated, the slope will be negative
- Only need *x* to vary in our sample

More OLS

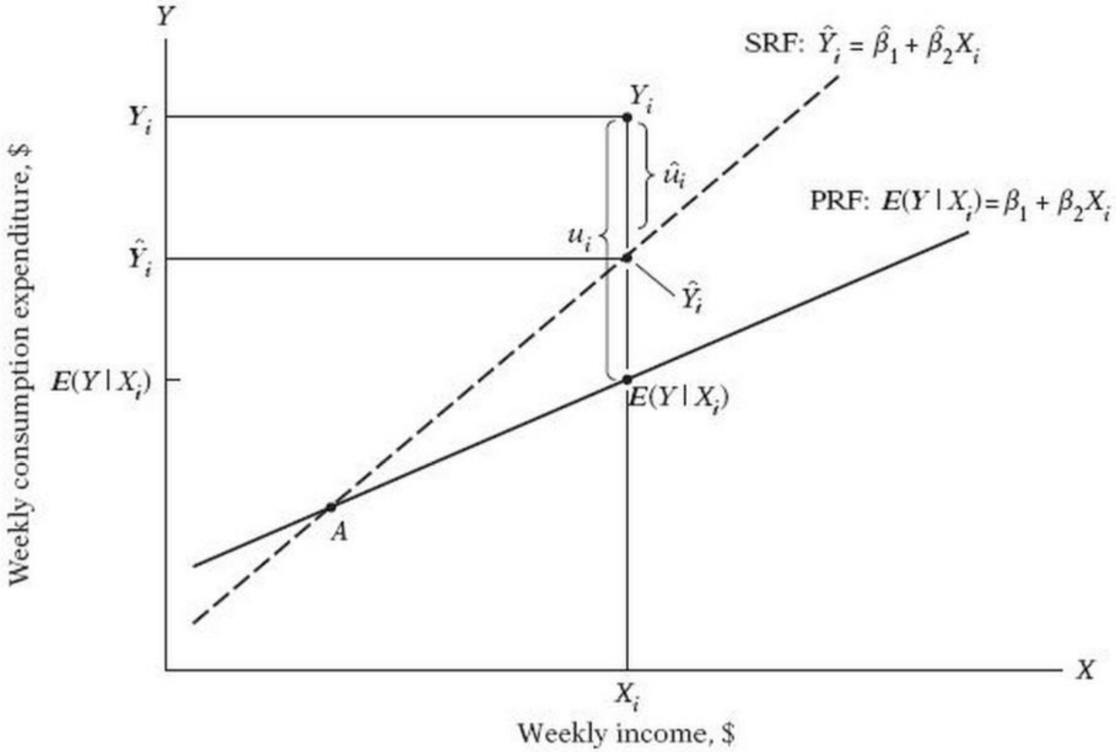
- Intuitively, OLS is fitting a line through the sample points such that the sum of squared residuals is as small as possible, hence the term least squares
- The residual, $\hat{u}$, is an estimate of the error term, u, and is the difference between the fitted line (sample regression function) and the sample point

Simple Regression Function (SRF)



PRF and SRF Relationship

- Observe the relationship between error term (u), and the residuals ($\hat{u}$)


Y
SRF: $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
Y_i
Y_i
$\hat{u}_i$
PRF: $E(Y \mid X_i) = \beta_1 + \beta_2 X_i$
u_i
$\hat{Y}_i$
$\hat{Y}_i$
$E(Y \mid X_i)$
$E(Y \mid X_i)$
A
Weekly consumption expenditure, $
X
X_i
Weekly income, $

Algebraic Properties of OLS

- The sum of the OLS residuals is zero
- Thus, the sample average of the OLS residuals is zero as well
- The sample covariance between the regressors and the OLS residuals is zero
- The OLS regression line always goes through the mean of the sample

Algebraic Properties (precise)

$$\sum_{i=1}^{n} \hat{u}_i = 0 \text{ and thus, } \frac{\sum_{i=1}^{n} \hat{u}_i}{n} = 0$$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

More terminology

We can think of each observation as being made up of an explained part, and an unexplained part, $y_i = \hat{y}_i + \hat{u}_i$ We then define the following:

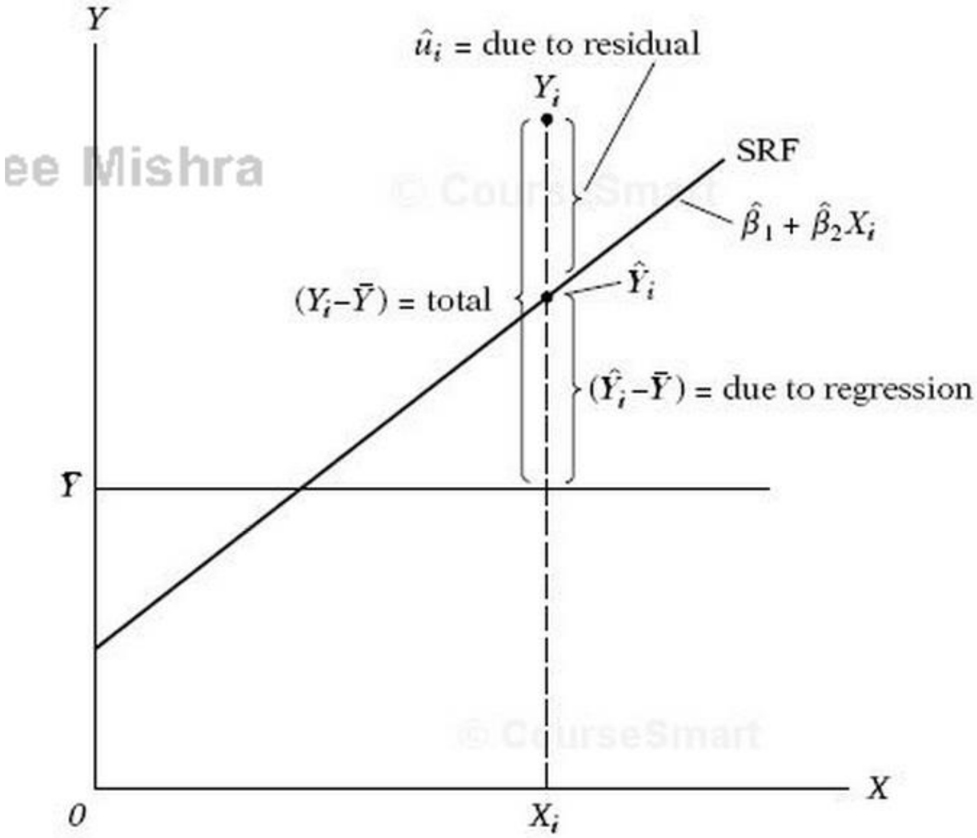
$\sum (y_i - \bar{y})^2$ is the total sum of squares (SST)

$\sum (\hat{y}_i - \bar{y})^2$ is the explained sum of squares (SSE)

$\sum \hat{u}_i^2$ is the residual sum of squares (SSR)

Then $\text{SST} = \text{SSE} + \text{SSR}$

- SST = SSE + SSR


Y
$\hat{u}_i$ = due to residual
Y_i
SRF
$\hat{\beta}_1 + \hat{\beta}_2 X_i$
$\hat{Y}_i$
$(Y_i - \bar{Y})$ = total
$(\hat{Y}_i - \bar{Y})$ = due to regression
$\bar{Y}$
0
X_i
X

Proof that SST = SSE + SSR

$$\sum\left(y_i-\bar{y}\right)^2=\sum\left[\left(y_i-\hat{y}_i\right)+\left(\hat{y}_i-\bar{y}\right)\right]^2$$

$$=\sum\left[\hat{u}_i+\left(\hat{y}_i-\bar{y}\right)\right]^2$$

$$=\sum\hat{u}_i^2+2\sum\hat{u}_i\left(\hat{y}_i-\bar{y}\right)+\sum\left(\hat{y}_i-\bar{y}\right)^2$$

$$=\mathrm{SSR}\ +2\sum\hat{u}_i\left(\hat{y}_i-\bar{y}\right)+\mathrm{SSE}$$

$$\text{and we know that }\sum\hat{u}_i\left(\hat{y}_i-\bar{y}\right)=0$$

Goodness-of-Fit

- How do we think about how well our sample regression line fits our sample data?

- Can compute the fraction of the total sum of squares (SST) that is explained by the model, call this the R-squared of regression

- $R^2 = SSE/SST = 1 – SSR/SST$

Unbiasedness of OLS

- Assume the population model is linear in parameters as $y = \beta_0 + \beta_1 x + u$
- Assume we can use a random sample of size n, $\{(x_i, y_i): i=1, 2, \ldots, n\}$, from the population model. Thus we can write the sample model $y_i = \beta_0 + \beta_1 x_i + u_i$
- Assume $\mathrm{E}(u|x) = 0$ and thus $\mathrm{E}(u_i|x_i) = 0$
- Assume there is variation in the x_i

Unbiasedness of OLS (2)

- In order to think about unbiasedness, we need to rewrite our estimator in terms of the population parameter
- Start with a simple rewrite of the formula as

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{s_x^2}, \text{ where}$$

$$s_x^2 \equiv \sum (x_i - \bar{x})^2$$

Unbiasedness of OLS (cont)

$$\sum\left(x_i - \bar{x}\right)y_i = \sum\left(x_i - \bar{x}\right)\left(\beta_0 + \beta_1 x_i + u_i\right) =$$

$$\sum\left(x_i - \bar{x}\right)\beta_0 + \sum\left(x_i - \bar{x}\right)\beta_1 x_i$$

$$+ \sum\left(x_i - \bar{x}\right)u_i =$$

$$\beta_0 \sum\left(x_i - \bar{x}\right) + \beta_1 \sum\left(x_i - \bar{x}\right)x_i$$

$$+ \sum\left(x_i - \bar{x}\right)u_i$$

Unbiasedness of OLS (cont)

$$\sum(x_i - \bar{x}) = 0,$$

$$\sum(x_i - \bar{x})x_i = \sum(x_i - \bar{x})^2$$

so, the numerator can be rewritten as

$$\beta_1 s_x^2 + \sum(x_i - \bar{x})u_i, \text{ and thus}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})u_i}{s_x^2}$$

Unbiasedness of OLS (cont)

$$\text{let } d_i = (x_i - \bar{x}), \text{ so that}$$

$$\hat{\beta}_i = \beta_1 + \left(1 / s_x^2 \right) \sum d_i u_i, \text{ then}$$

$$E\left(\hat{\beta}_1\right) = \beta_1 + \left(1 / s_x^2 \right) \sum d_i E(u_i) = \beta_1$$

Unbiasedness Summary

- The OLS estimates of β_1 and β_0 are unbiased
- Proof of unbiasedness depends on our 4 assumptions – if any assumption fails, then OLS is not necessarily unbiased
- Remember unbiasedness is a description of the estimator – in a given sample we may be "near" or "far" from the true parameter

Variance of the OLS Estimators

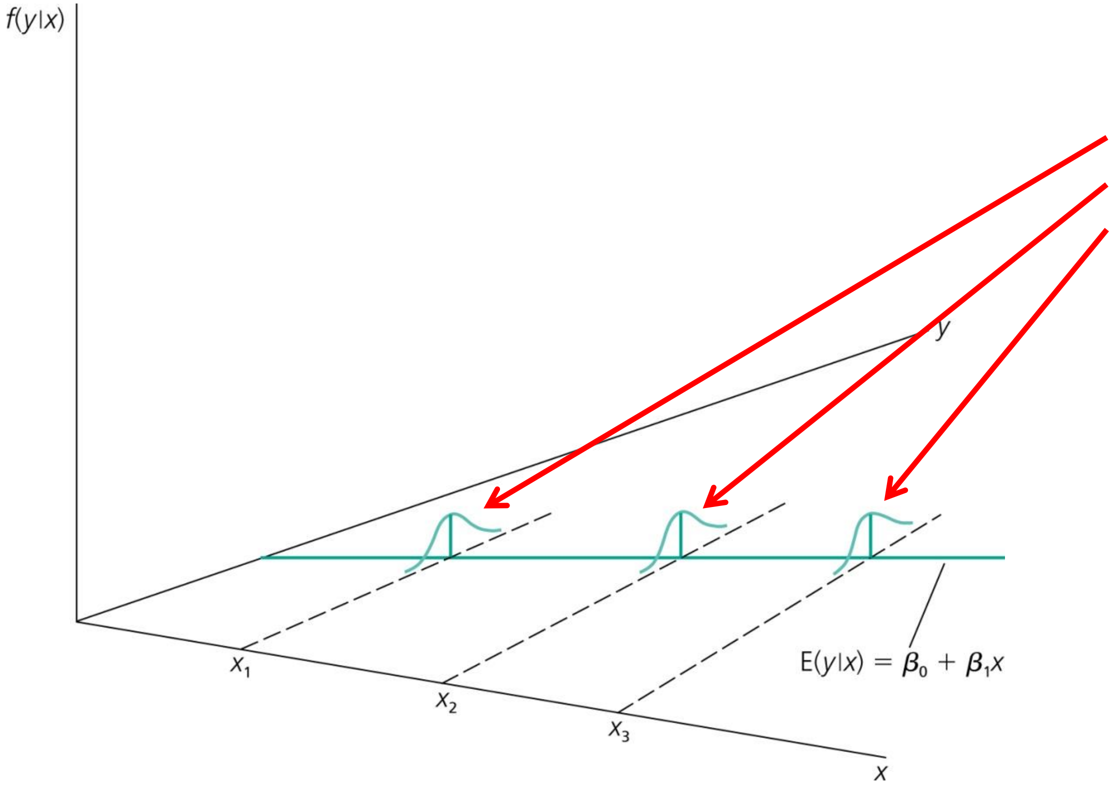
- Now we know that the sampling distribution of our estimate is centered around the true parameter
- Want to think about how spread out this distribution is
- Much easier to think about this variance under an additional assumption, so
- Assume Var($u|x$) = σ^2 (Homoskedasticity)

Variance of OLS (cont)

- $\mathrm{Var}(u|x) = E(u^2|x) - [E(u|x)]^2$
- $E(u|x) = 0$, so $\sigma^2 = E(u^2|x) = E(u^2) = \mathrm{Var}(u)$
- Thus σ^2 is also the unconditional variance, called the error variance
- σ, the square root of the error variance is called the standard deviation of the error
- Can say: $E(y|x)=\beta_0 + \beta_1 x$ and $\mathrm{Var}(y|x) = \sigma^2$

Homoskedasticity

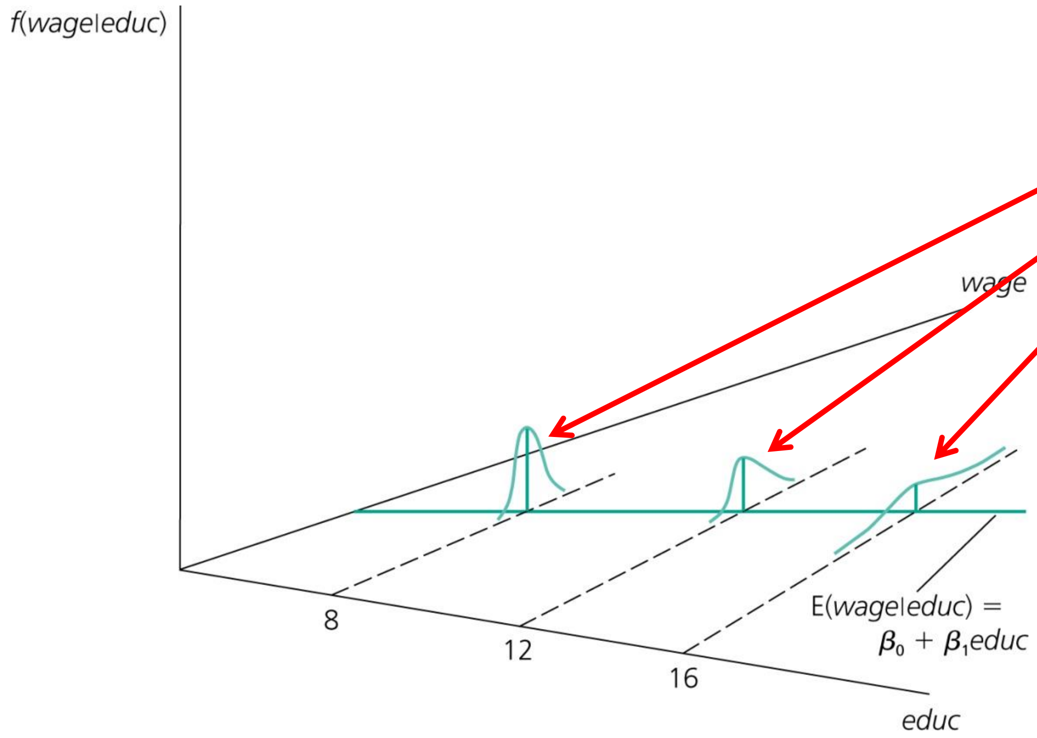
- Graphical illustration of homoskedasticity



The variability of the unobserved influences does not dependent on the value of the explanatory variable

Heteroskedasticity

- **An example for heteroskedasticity: Wage and education**



The variance of the unobserved determinants of wages increases with the level of education

Variance of OLS (cont)

$$Var\left(\hat{\beta}_1\right) = Var\left(\beta_1 + \left(1/s_x^2\right)\sum d_i u_i\right) =$$

$$\left(1/s_x^2\right)^2 Var\left(\sum d_i u_i\right) = \left(1/s_x^2\right)^2 \sum d_i^2 Var(u_i)$$

$$= \left(1/s_x^2\right)^2 \sum d_i^2 \sigma^2 = \sigma^2 \left(1/s_x^2\right)^2 \sum d_i^2 =$$

$$\sigma^2 \left(1/s_x^2\right)^2 s_x^2 = \sigma^2/s_x^2 = Var\left(\hat{\beta}_1\right)$$

Variance of OLS Summary

- The larger the error variance, σ^2, the larger the variance of the slope estimate
- The larger the variability in the x_i, the smaller the variance of the slope estimate
- As a result, a larger sample size should decrease the variance of the slope estimate
- Problem that the error variance is unknown

Estimating the Error Variance

- We don't know what the error variance, σ^2, is, because we don't observe the errors, u_i
- What we observe are the residuals, $\hat{u}_i$
- We can use the residuals to form an estimate of the error variance

Error Variance Estimate (cont)

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$= (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)$$

Then, an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum \hat{u}_i^2 = SSR/(n-2)$$

Error Variance Estimate (cont)

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \text{Standard error of the regression}$$

$$\text{recall that } \operatorname{sd}\left(\hat{\beta}\right) = \sigma / s_x$$

$$\text{if we substitute } \hat{\sigma} \text{ for } \sigma \text{ then we have}$$

$$\text{the standard error of } \hat{\beta}_1 \text{ ,}$$

$$\operatorname{se}\left(\hat{\beta}_1\right) = \hat{\sigma} / \left(\sum (x_i - \bar{x})^2\right)^{1/2}$$

Gauss-Markov Assumptions (1)

- <u>Standard assumptions for the linear regression model</u>
- **Assumption SLR.1 (Linear in parameters)**

$$y = \beta_0 + \beta_1 x + u$$

In the population, the relationship between y and x is linear

- **Assumption SLR.2 (Random sampling)**

$$\{(x_i, y_i) : \; i = 1, \ldots n\}$$

The data is a random sample drawn from the population

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Each data point therefore follows the population equation

Gauss-Markov Assumptions (2)

- Assumptions for the linear regression model (cont.)
- Assumption SLR.3 (Sample variation in explanatory variable)

$$\sum_{i=1}^{n}(x_i-\bar{x})^2>0$$

The values of the explanatory variables are not all the same (otherwise it would be impossible to study how different values of the explanatory variable lead to different values of the dependent variable)

- Assumption SLR.4 (Zero conditional mean)

$$E(u_i|x_i)=0$$

The value of the explanatory variable must contain no information about the mean of the unobserved factors

- Assumption SLR.5 (Homoskedasticity)

$$Var(u_i|x_i)=\sigma^2$$

The value of the explanatory variable must contain no information about the variability of the unobserved factors

Simple Regression Estimation

Term	Formulae
Coefficient of x	$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
Intercept coefficient	$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$
Coefficient of Determination (r^2)	$$r^2 = \frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2}$$
Standard error (β_1)	$\mathrm{se}(\hat{\beta}_1) = \hat{\sigma} / \left(\sum(x_i - \bar{x})^2\right)^{1/2}$ where, $\hat{\sigma}^2 = \frac{\sum_i \widehat{u_i}^2}{n-2}$
Standard error (β_0)	se(β_0)=sqrt $\left(\frac{\hat{\sigma}^2 * \sum_i x_i^2}{n\sum_i(x_i - \bar{x})^2}\right)$
t-stat	Coefficient/standard error