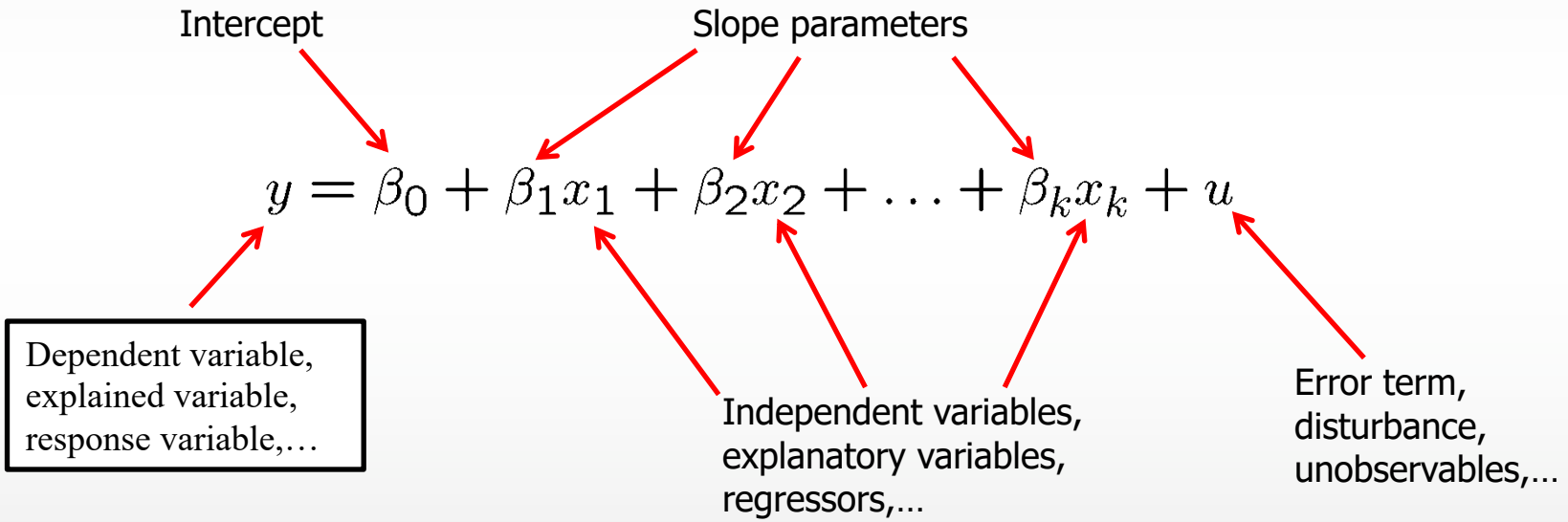
Multiple Regression

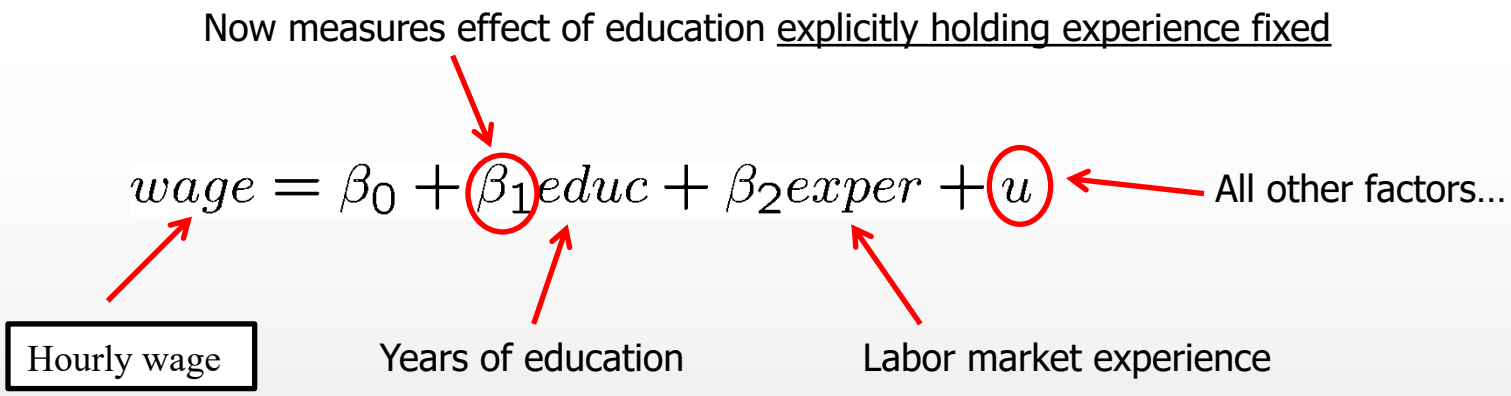
- **Definition of the multiple linear regression model**

Explains variable y in terms of variables $x_1, x_2, \ldots, x_k$



Multiple Regression

- **Motivation for multiple regression**
  - Incorporate more explanatory factors into the model
  - Explicitly hold fixed other factors that otherwise would be in u
  - Causal inference and *ceteris paribus*.
- **Example: Wage equation**



Multiple Regression

- **OLS Estimation of the multiple regression model**
- **Random sample**

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \ i = 1, \ldots n\}$$

- **Regression residuals**

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}$$

- **Minimize sum of squared residuals**

$$\min \sum_{i=1}^{n} \hat{u}_i^2 \quad \rightarrow \quad \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$$

Minimization will be carried out by computer

Multiple Regression

- Solutions (for a model with 2 independent variables):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

$$\hat{\beta}_1 = \frac{\sum (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \sum (x_{2i} - \bar{x}_2)^2 - \sum (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2)^2 - \left[\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]^2} =$$

$$\frac{Cov(x_1, y) * Var(x_2) - Cov(x_2, y) * Cov(x_1, x_2)}{Var(x_1) * Var(x_2) - [Cov(x_1, x_2)]^2} =$$

$$\left(\frac{r_{y,x_1} - r_{x1,x2} * r_{y,x2}}{1 - (r_{x1,x2})^2} \right) * \left(\frac{s_y}{s_{x_1}} \right)$$

$\hat{\beta}_2$: Interchange x_1 and x_2 in the numerator of $\hat{\beta}_1$. Denominator is the same. In correlation equation, interchange all.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

- No solution to $\hat{\beta}_1$ if: (1) x_1 and x_2 are perfectly correlated or (2) x_1 or x_2 does not vary.

Multiple Regression: Assumptions

- **Standard assumptions for the multiple regression model**
- **Assumption MLR.1 (Linear in parameters)**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

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

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : \; i = 1, \ldots n\}$$

The data is a random sample drawn from the population

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$

Each data point therefore follows the population equation

Multiple Regression: Assumptions

- **Standard assumptions for the multiple regression model (cont.)**
- **Assumption MLR.3 (No perfect collinearity)**

  In the sample (and therefore in the population), none of the independent variables is constant and there are no exact relationships among the independent variables

- **Remarks on MLR.3**
  - The assumption only rules out perfect collinearity/correlation between explanatory variables; correlations < 1 are allowed.
  - If an explanatory variable is a perfect linear combination of other explanatory variables it is superfluous and will be eliminated
  - Constant variables are also ruled out (collinear with intercept)

Multiple Regression: Assumptions

- **Standard assumptions for the multiple regression model (cont.)**
- **Assumption MLR.4 (Zero conditional mean)**

$$E(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

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

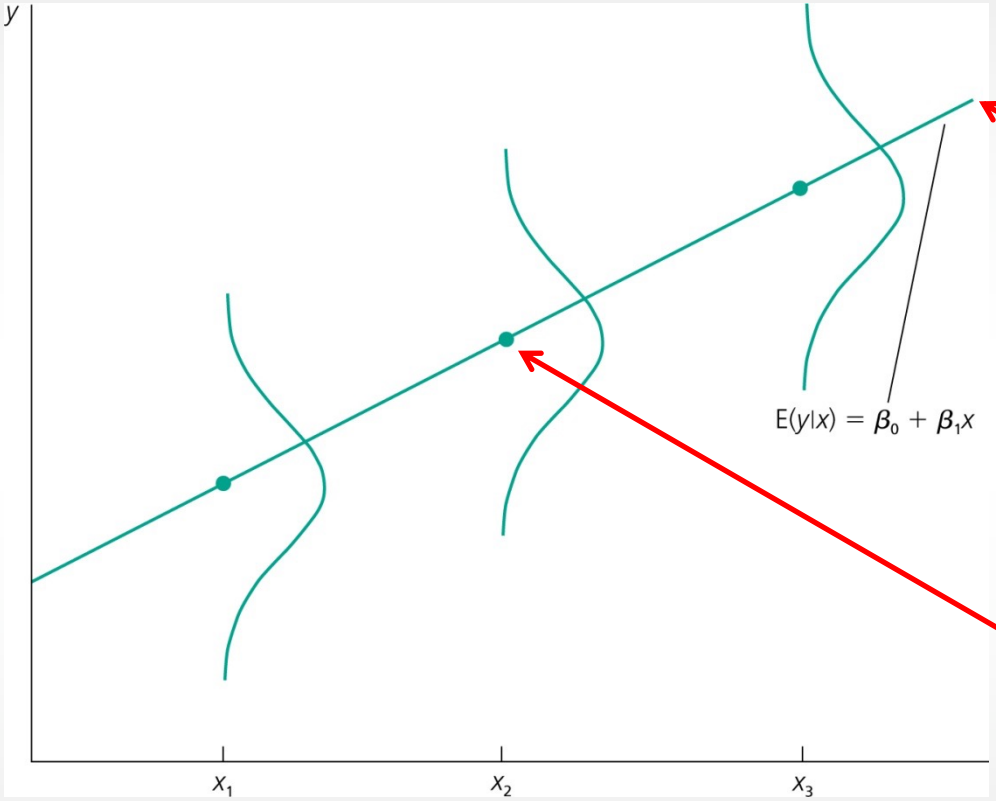
Multiple Regression

- **Discussion of the zero conditional mean assumption**
  - Explanatory variables that are correlated with the error term are called endogenous; endogeneity is a violation of assumption MLR.4
  - Explanatory variables that are uncorrelated with the error term are called exogenous; MLR.4 holds if all explanatory variables are exogenous
  - Exogeneity is the key assumption for a causal interpretation of the regression, and for unbiasedness of the OLS estimators
  - Omited variable bias can induce a violation of this assumption
  - Wrong functional form can induce a violation, too.
- **Theorem 3.1 (Unbiasedness of OLS)**

$$MLR.1-MLR.4 \quad \Rightarrow \quad E(\hat{\beta}_j) = \beta_j, \quad j = 0, 1, \ldots, k$$

  - Unbiasedness is an average property in repeated samples; in a given sample, the estimates may still be far away from the true values

Population Regression



Population regression function

For individuals with $x = x_2$, the average value of y is $\beta_0 + \beta_1 x$ and average value of u = 0.

Multiple Regression: Assumptions

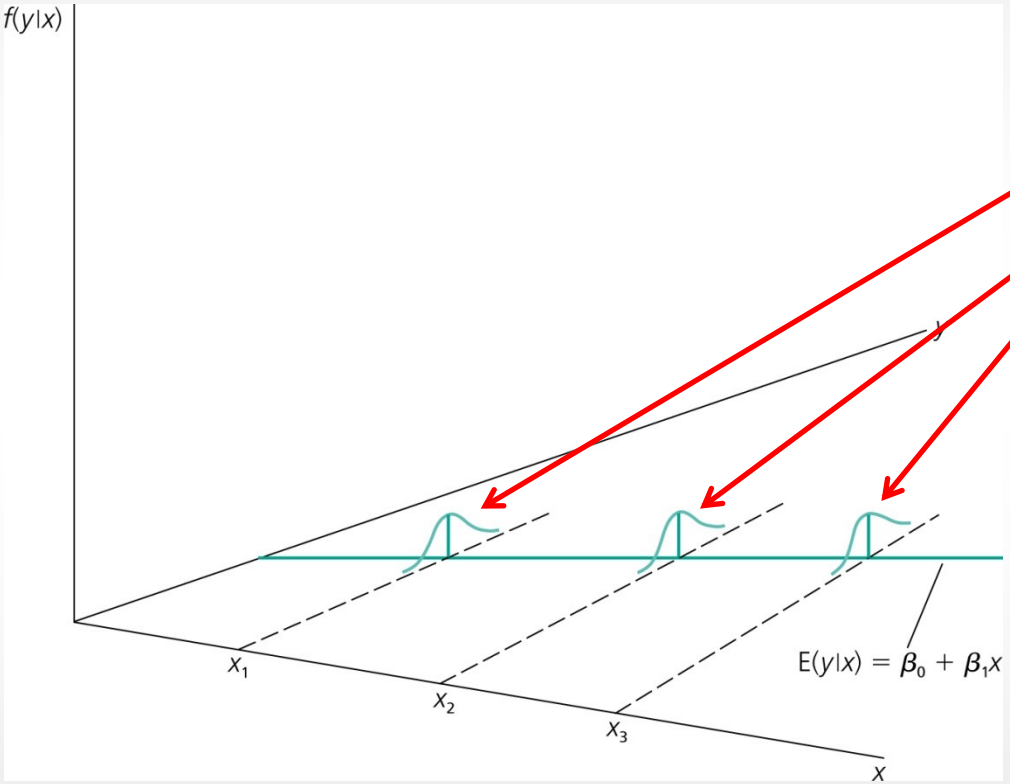
- **Standard assumptions for the multiple regression model (cont.)**
- **Assumption MLR.5 (Homoscedasticity)**

$$Var(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2$$

Error variance is assumed to the same for all observations; it is not conditional on independent variables.

Homoskedasticity

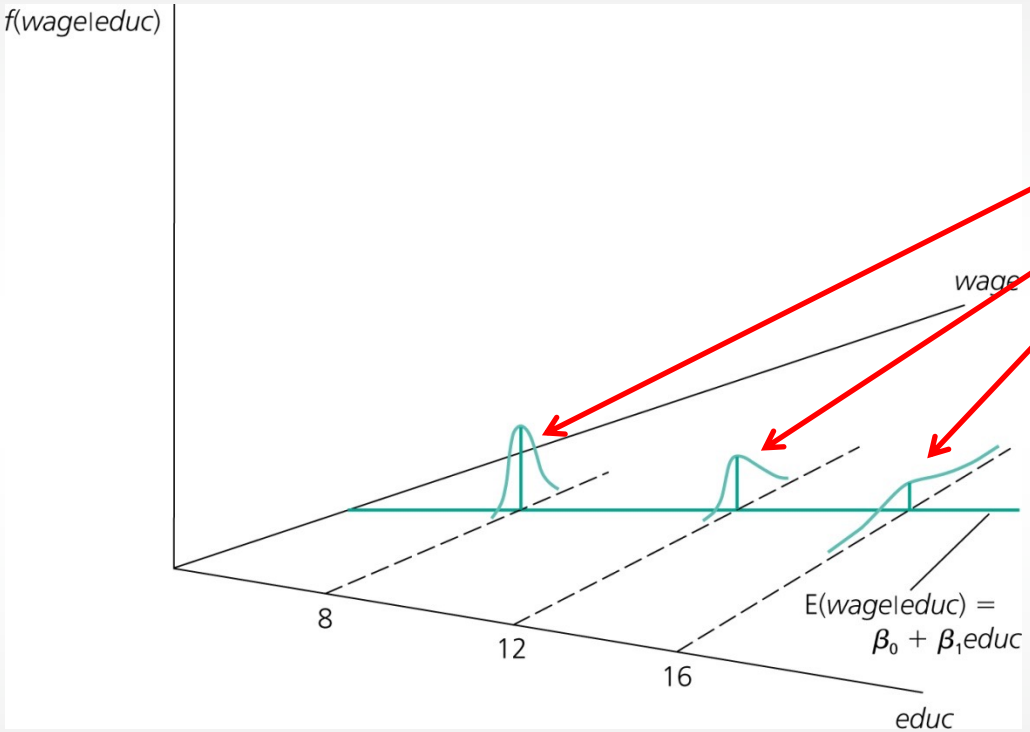
- **Graphical illustration of homoskedasticity**



The variability of the unobserved influences does not depend on the value of the explanatory variable; error variance is constant across the range of x.

Heteroskedasticity

- **An example for heteroskedasticity**

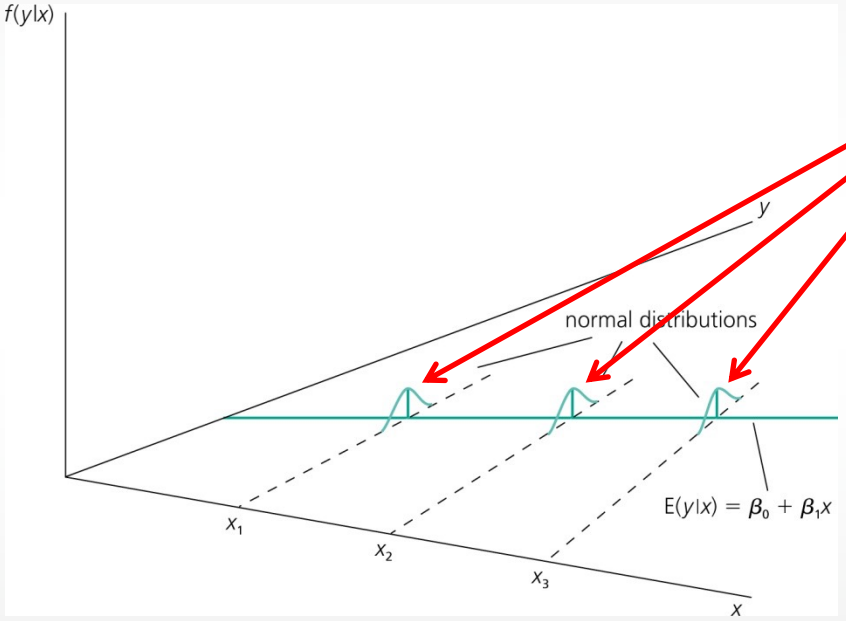


The variance of the unobserved determinants of wages increases with the level of education; non-constant error variance, related to x.

Multiple Regression: Assumptions

- **Assumption MLR.6 (Normality of error terms)**

$u_i \sim N(0, \sigma^2)$ independently of $x_{i1}, x_{i2}, \ldots, x_{ik}$



It is assumed that the unobserved factors are normally distributed around the population regression function.

The form and the variance of the distribution does not depend on any of the explanatory variables.

It follows that:

$$y|\mathbf{x} \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$$

- MLR.6 needed for statistical inference; hypothesis testing.

Error Variance in Multiple Regression

- **Estimating the error variance in multiple regression**

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2 \right) / [n - k - 1]$$

An unbiased estimate of the error variance can be obtained by substracting the number of estimated regression coefficients from the number of observations. n – k – 1 = degrees of freedom. Note k=number of independent variables (or number of partial slopes estimated; the 1 is for the intercept.

- **Theorem 3.3 (Unbiased estimator of the error variance)**

$$MLR.1 - MLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

Standard Error of the Regression (SER)

- Simply the square root of the *error variance*.
- Calculated as (simple regression):

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

- Using rules of standard deviation and the normal distribution: about 2/3 of the residuals are $\pm$ 1 *SER* from zero. 95% are $\pm$ 2 *SER*'s from zero. [Note: the mean of $\hat{u}_i$ is zero.]
- Aka, "root mean squared error"; "Root MSE" [Stata]; "standard error of the estimate"

Multiple Regression: Interpretation

- In multiple regression, we have *partial slopes* (aka, *partial coefficients* or *partial effects*)
  - Effect of independent variable *controlling for*, or *holding constant*, the other independent variables.
  - Connection to causality and control….

- *Marginal effects* for a linear function
  - Partial derivative of function (effect of x1 while holding x2 constant)

Multiple Regression: Interpretation

- **Interpretation of the multiple regression model**

$$\beta_j = \frac{\partial y}{\partial x_j}$$

By how much does the dependent variable change if the j-th independent variable is increased by one unit, holding all other independent variables and the error term constant

  - The multiple linear regression model manages to hold the values of other explanatory variables fixed even if, in reality, they are correlated with the explanatory variable under consideration;
  - Design v. statistics...
  - *For a one-unit increase in x_1, y changes by $\hat{\beta}_1$ units, holding x_2 constant.* Or…
  - *Controlling for x_2, a one-unit increase in x_1 produces a $\hat{\beta}_1$ unit change in y.*
  - Interpreting $\hat{\beta}_0$: $\hat{y}$ when both x_1 ***and*** x_2 are zero.
  - Goodness of fit: same as in simple regression.

Multiple Regression: Interpretation

- **Example: Determinants of college GPA**

$$\widehat{colGPA} = 1.29 + .453hsGPA + .0094ACT$$

Grade point average at college

High school grade point average

Achievement test score

- **Interpretation**
  - Holding ACT fixed, another point on high school grade point average is associated with another .453 points college grade point average
  - Or: If we compare two students with the same ACT, but the hsGPA of student A is one point higher, we predict student A to have a colGPA that is .453 higher than that of student B
  - Holding high school grade point average fixed, another 10 points on ACT are associated with less 1/10 of a point on college GPA

Multiple Regression: Interpretation

- **Partialling out interpretation of multiple regression**
- **One can show that the estimated coefficient of an explanatory variable in a multiple regression can be obtained in two steps:**
  - 1) Regress the explanatory variable on all other explanatory variables
  - 2) Regress *y* on the residuals from this regression
- **Why does this procedure work?**
  - The residuals from the first regression is the part of the explanatory variable that is uncorrelated with the other explanatory variables
  - The slope coefficient of the second regression therefore represents the isolated effect of the explanatory variable on the dep. variable
  - *Netting out* interpretation

Standardized Coefficients

- Pros and cons

- Norms in political science: *very rarely* used.

- Fundamental issue: *Comparing* effects of independent variables when their measurement units are different. For instance:
  - Education may be measured in the number of years (units of measurement).
  - Income is measured in dollars.
  - How do we compare slopes for these, since a "one unit increase" in both is heavily dependent on the units of measurement each variable takes on?
  - Bigger issue: How do we compare effects (slopes) of independent variables in multiple regression?

Standardized Coefficients

- Standardized coefficients: "Standardize" units of measurement across different IVs into "standard deviation units."

  Standardized variable: $x^* = (x_i - \bar{x}) / s_x$

  Mean of x^*=0; *s.d.*=1.

- Now, slopes from a regression on fully standardized data (standardize y and x's) are interpreted as: *For a one standard deviation increase in x, y changes by* $\hat{\beta}$ *standard deviation units, controlling for the other x's.*

- In simple regression, a standardized $\hat{\beta}_1 = r_{Y,X}$; ranges from -1 to 1.

- In multiple regression, standardized partial slopes are *partial correlations* (again, range from -1 to 1).

- Problems?

Goodness of Fit: R^2

- **Measures of Variation**

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Total sum of squares, represents total variation in dependent variable

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Explained sum of squares, represents variation explained by regression

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

Residual sum of squares, represents variation not explained by regression

R^2

- **Decomposition of total variation**

$$SST = SSE + SSR$$

Total variation — Explained part — Unexplained part

- **Goodness-of-fit measure (R-squared)**

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

R-squared measures the fraction of the total variation that is explained by the regression

R^2

- **Goodness-of-Fit**

- **R-squared**

Notice that R-squared can only increase if another explanatory variable is added to the regression

$$R^2 = SSE/SST = 1 - SSR/SST$$

Adjusted R^2

- Since R^2 can only increase as you add independent variables, we can correct for this by incorporating information about degrees of freedom in the SSR part.

$$Adj.R^2 = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}}$$

- k = number of independent variables; (n-k-1)=degrees of freedom (df)
- In multiple regression, the effect of this correction will be to reduce the regular R^2.

Sampling Variances and SEs of the Parameters in Simple Regression

Sampling Variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ in simple regression:

$$Var(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

Standard errors (SE) of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$SE(\hat{\beta}_0) = \sqrt{\frac{\hat{\sigma}^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}} = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Standard Errors

- What makes the standard error of the slope large or small (in simple regression)?
  1. Standard error increases (worsens) as **error variance** increases.
  2. Standard error decreases (improves) as **variance in x** increases.
  3. Standard error decreases (improves) as **sample size** (*n*) increases (*n* enters in through the error variance).
- Exercise: Calculating SEs

Calculations, Simple Regression

ID	y_i	x_i						
1	10	2						
2	4	3						
3	5	0						
4	12	10						
5	20	8						

Sampling Variances and SEs in Multiple Regression

Sampling variances for slopes in multiple regression:

$$Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\left(\sum (x_{ij} - \bar{x}_j)^2\right)\left(1 - R_j^2\right)}$$

$$SE(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{\left(\sum (x_{ij} - \bar{x}_j)^2\right)\left(1 - R_j^2\right)}} = \frac{\hat{\sigma}}{\sqrt{\left(\sum (x_{ij} - \bar{x}_j)^2\right)\left(1 - R_j^2\right)}}$$

R-squared from a regression of explanatory variable x_j on all other independent variables (including a constant)

Standard error (SE) of $\hat{\beta}_1$ when there are **two independent variables**:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\left(\sum (x_{1i} - \bar{x}_1)^2\right)\left(1 - R_{x_1, x_2}^2\right)}}$$

In a model with just two independent variables, this is simply squared value of the correlation between the two x variables.

Note: We'll skip estimating the SE of the intercept for multiple regression.

Sampling Variances and SEs in Multiple Regression

Standard error (SE) of $\hat{\beta}_1$ when there are two independent variables:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\left(\sum (x_{1i} - \bar{x}_1)^2\right)\left(1 - R^2_{x_1, x_2}\right)}}$$

In a model with just two independent variables, this is simply squared value of the correlation between the two x variables.

- Calculating $R^2_{x1,x2}$

$$r = \frac{Cov(x_1, x_2)}{s_{x_1} s_{x_2}} = \frac{\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2)^2}}$$

- After calculating *r* between x1 and x2, don't forget to **square it** before putting into the equation for the standard error.

Standard Errors

- What makes the standard error of the partial slopes large or small in multiple regression?
  1. Standard error increases (worsens) as **error variance** increases.
  2. Standard error decreases (improves) as **variance in x** increases.
  3. Standard error decreases (improves) as **sample size** (n) increases (n enters in through the error variance).
  4. Standard error for a particular $\hat{\beta}$ increases as its corresponding x's **correlation with the other *x* variables** increases (**multicollinearity**).
- Consequences for inference…

Efficiency of OLS

- **Efficiency of OLS: The Gauss-Markov Theorem**
  - Under assumptions MLR.1 - MLR.5, OLS is unbiased
  - However, under these assumptions there may be many other estimators that are unbiased
  - Which one is the unbiased estimator with the smallest variance?
  - In order to answer this question one usually limits oneself to linear estimators, i.e. estimators linear in the dependent variable

Gauss-Markov Theorem

- **Theorem 3.4 (Gauss-Markov Theorem)**
  - Under assumptions MLR.1 - MLR.5, the OLS estimators are the **best linear unbiased estimators** (BLUEs) of the regression coefficients, i.e.

$$Var(\hat{\beta}_j) \leq Var(\tilde{\beta}_j) \quad j = 0, 1, \ldots, k$$

for all $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$ for which $E(\tilde{\beta}_j) = \beta_j, j = 0, \ldots, k.$

- **OLS is only the best estimator if MLR.1 – MLR.5 hold; if there is heteroscedasticity for example, there are better estimators.**

Calculations, Simple Regression

ID	y_i	x_i	$\bar{y}$	$\bar{x}$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^* (y_i - \bar{y})$
1	10	2						
2	4	3						
3	5	0						
4	12	10						
5	20	8						

Calculations, Simple Regression

ID	y_i	x_i	$\bar{y}$	$\bar{x}$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^* (y_i - \bar{y})$
1	10	2	*10.2*	*4.6*	*-2.6*	*-0.2*	*6.76*	*0.52*
2	4	3	*10.2*	*4.6*	*-1.6*	*-6.2*	*2.56*	*9.92*
3	5	0	*10.2*	*4.6*	*-4.6*	*-5.2*	*21.16*	*23.92*
4	12	10	*10.2*	*4.6*	*5.4*	*1.8*	*29.16*	*9.72*
5	20	8	*10.2*	*4.6*	*3.4*	*9.8*	*11.56*	*33.32*
						$\sum_{i=1}^{n}$	***71.2***	***77.4***

$$\hat{\beta}_1 = 77.4/71.2 = 1.087$$

$$\hat{\beta}_0 = 10.2 - 1.087*4.6 = 5.20$$

Statistical Inference

- Use standard errors in conjunction with slope estimates to conclude whether there is a statistically significant impact of *x* on *y*.

- What does **statistical significance** mean?
  - Confidence intervals in classical inference.
  - Why do we choose such wide confidence bands (e.g., 95%) and low alpha (α) levels?
  - Type 1 and Type 2 errors
    - Scientific norms → α =.05
  - *Is the effect statistically different from zero?*

Statistical Inference

- **Hypothesis testing for slopes – 5 steps** (Problem Set 1)

1. Set up null and alternative hypotheses

   - 1-tailed or 2-tailed test?

2. Choose alpha level – i.e., a confidence level.
3. Calculate test statistic

   - t-distribution

4. Decision rule for rejecting or failing to reject the null hypothesis.
5. Assess the evidence – make decision.

Omnibus F-Test

- Goal: Test whether the effect of multiple X's in multiple regression are simultaneously equal to zero.
- **Omnibus Test**:
  - Comparing model to a model with no independent variables.
  - Test whether *all* X's are simultaneously equal to zero.
  - H_0: $\beta_1=\beta_2=\ldots\beta_k = 0$

F-Test: $$F = \frac{SSE/k}{SSR/(n-k-1)} = \frac{n-k-1}{k} * \frac{R^2}{1-R^2}$$

- Note: k = the number of *independent variables*
- Retrieve p-value from the F-distribution (right-skewed distribution; ratio of two chi-squared distributions).
  - df=(k, n-k-1) [note: df associated with numerator given first]
  - To find p-value, use Stata: di Ftail(dfnum, dfdenom, F)

"Subset" F-Test

- Let's say we had 6 independent variables, and we wanted to test simultaneously whether the effects of X5 and X6 were simultaneously equal to zero.

  H_0: $\beta_5=\beta_6=0$

- Why would we want to do this?
- You're comparing a "full model" to a "reduced model," just like we did in the omnibus test. Difference is with respect to the reduced model.

F-test: $$F = \frac{(SSE_1 - SSE_0)/q}{SSR_1/(n-k-1)} = \frac{n-k-1}{q} * \frac{R_1^2 - R_0^2}{1-R_1^2}$$

- Note: subscript 1 indicates "full model"; subscript 0 indicates "reduced model."
- q = the number of independent variables excluded from full model; or the number of parameters listed in H_0.

Omnibus and Subset F-Tests in Stata

- Omnibus: R output provides this
- Subset:
  - run the two different models, get the various estimates, and estimate F