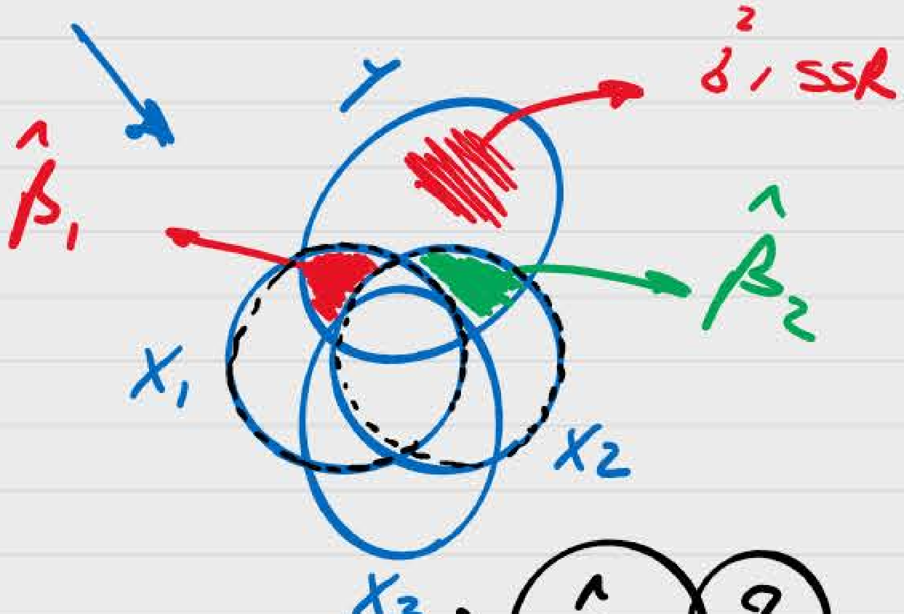
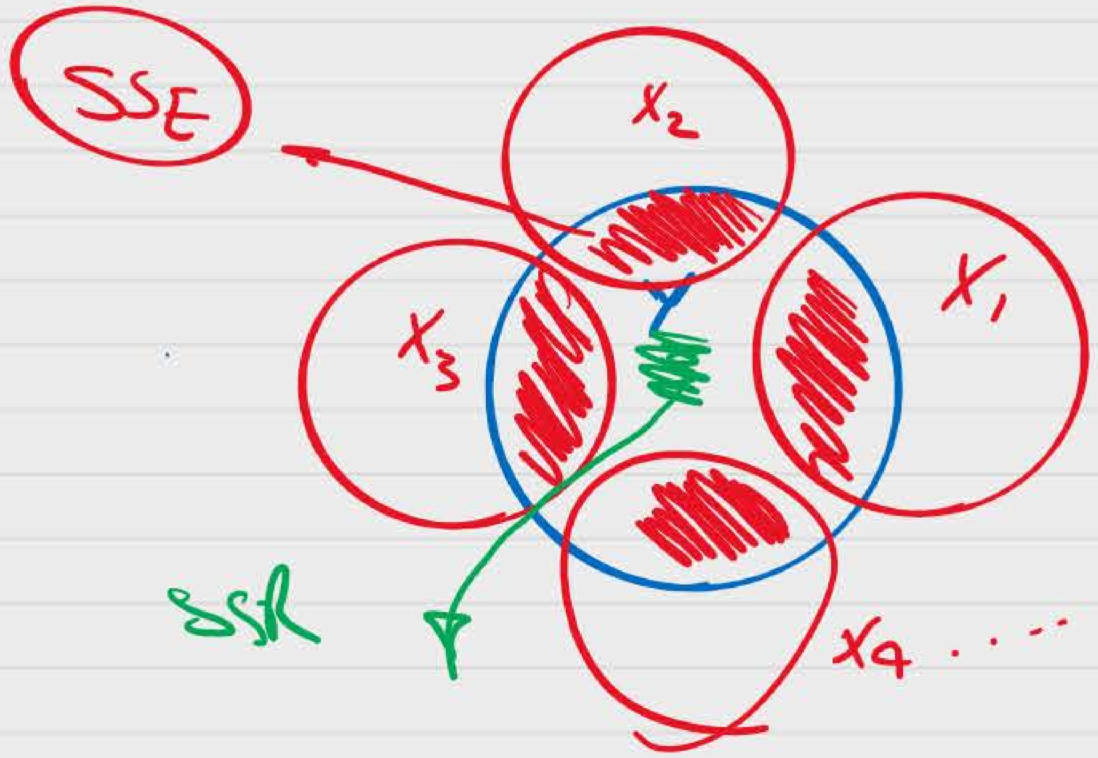
Venn Diagrams

$$\underset{}{wage} \sim \overset{x_1}{educ} + \overset{x_2}{exp} + \overset{x_3}{age}$$

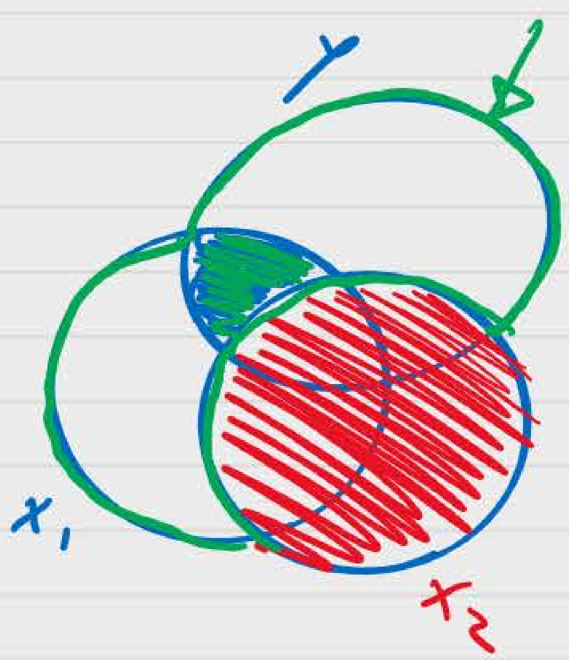


Real world

$\hat{\beta}_3$? $\quad Var(\hat{\beta}_3) \uparrow$



ideal!



OLS Assumptions :

SRM

1. linear in β_j
2. Random Sampling

3. $Var(X) > 0$
4. $E(u|X) = 0$
5. $Var(u|X) = \sigma^2$

MRM

1. linear in β_j
2. Random Sampling

3. $Var(X) > 0$ + No Perfect Collinearity
4. $E(u|X) = 0$
5. $Var(u|X) = \sigma^2$

*

Class 12 – Multiple Regression Model Estimation (Part II)

Pedram Jahangiry

JON M. HUNTSMAN
SCHOOL OF BUSINESS
UtahStateUniversity

Introducing Venn Diagrams

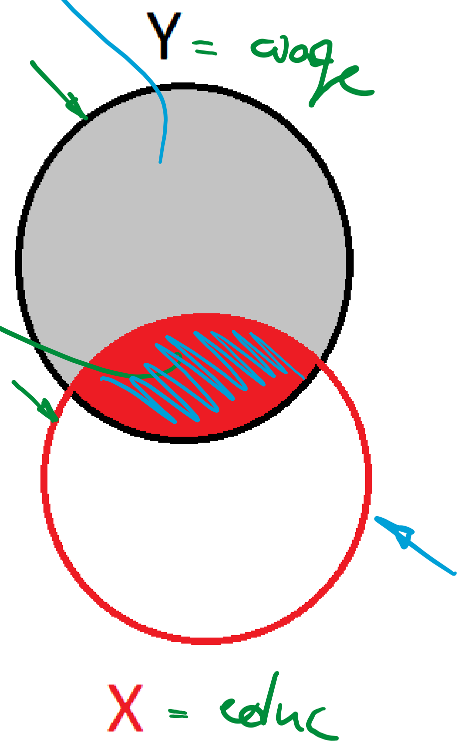
In a Simple Regression Model : $Y = \beta_o + \beta_1 X + u$,

Var(u)

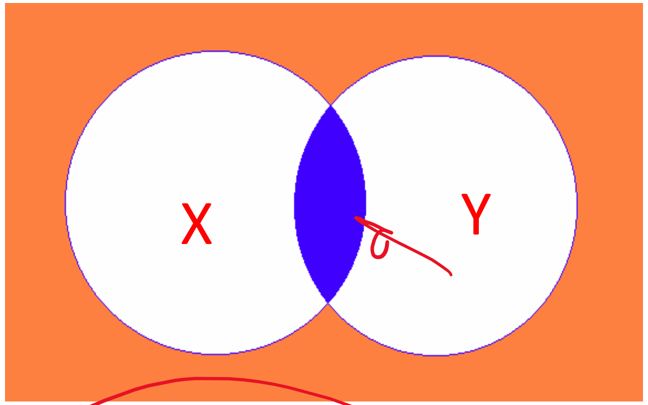
SSR
σ^2
$\hat{\sigma}^2$

- Black circle *represents* the **variations in Y**
- Red circle *represents* the **variations in X**
- Red shaded area *represents*
  - variation in Y **explained** by X **(SSE)**
  - Correlation between Y and X
  - $\boldsymbol{R^2}$
  - $\boldsymbol{\hat{\beta}_1}$
- Gray shaded area *represents*
  - **unexplained** variations in Y **(SSR)**
  - $\boldsymbol{\sigma^2}$
  - Variations of residuals

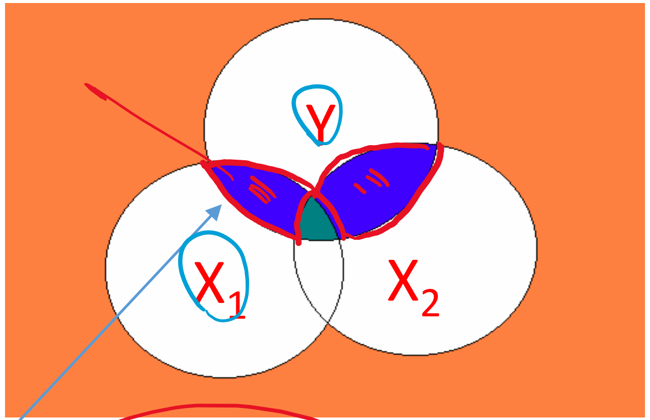
$\leq$ SSE
$\neq$ Corr(x,y)
$\neq$ R^2
$\neq$ $\hat{\beta}_1$



Correlation vs. Partial correlation



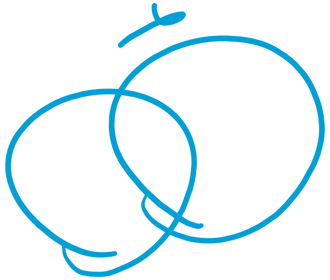
Correlation
Simple Regression Model



Partial Correlation
Multiple Regression Model

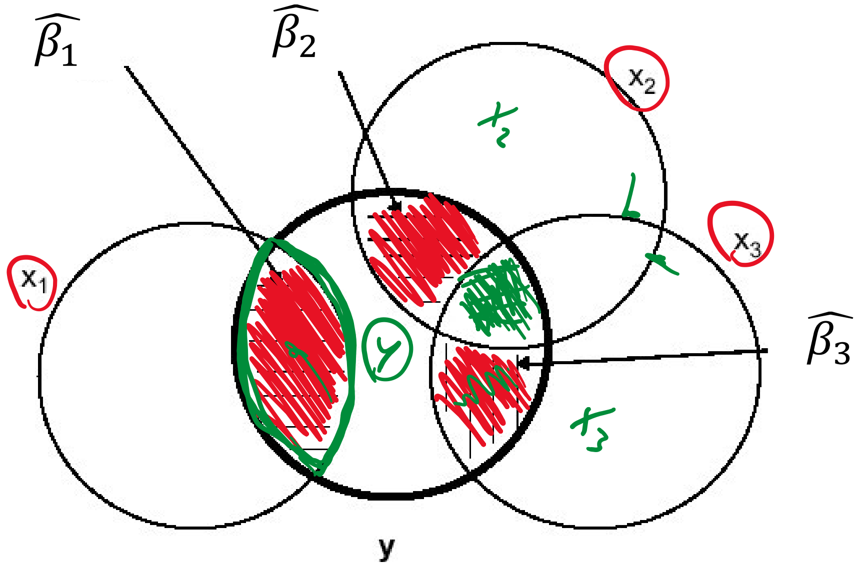
- Partial correlation between Y and X1 is defined as the correlation between Y and X1 while controlling (netting out) the effect of X2

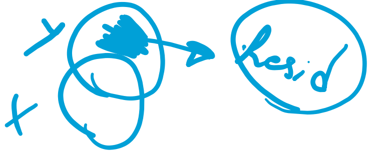
Venn Diagram Depiction of MRM coefficients

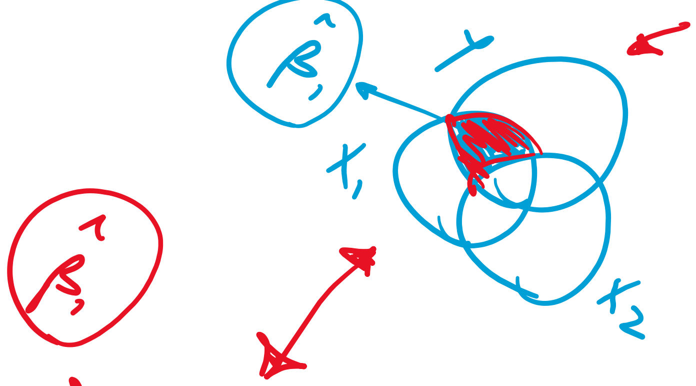


Multiple Regression Model (Ceteris paribus interprepation)

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$





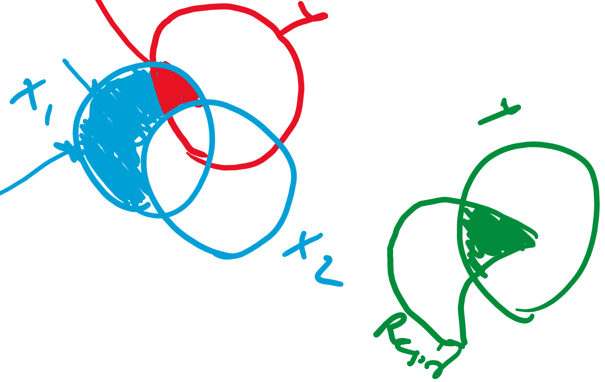


"Partialling out" interpretation of multiple regression

One can show that the estimated coefficient of an explanatory variable in a multiple regression can be obtained in **two steps**:

X_1 X_2

1) Regress the explanatory variable on all other explanatory variables
2) Regress y on the residuals from this regression



* Why does this procedure work?

X_1

- The residuals from the first regression is the part of the explanatory variable that is uncorrelated with the other explanatory variables

X_2

- The slope coefficient of the second regression therefore represents the isolated effect of the explanatory variable on the dependant variable

X_1 Y

R^2

Goodness-of-Fit measure for a given model

$\sum(\hat{y}_i - \bar{y})^2$

$\sum(y - \bar{y})^2$ $\sum(y_i - \hat{y}_i)^2$

- Decomposition of total variation

$$\boxed{SST = SSE + SSR}$$

- R-squared

$$R^2 \equiv \underline{SSE}/SST = 1 - \underline{SSR}/SST$$

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

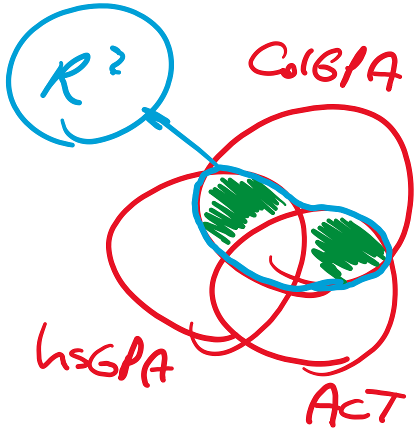
- Alternative expression for R-squared

$$R^2 = \underline{corr(y, \hat{y})^2} = \left[\frac{cov(y, \hat{y})}{sd(y)\ sd(\hat{y})}\right]^2$$

$\neq Cor(y, x)^2$

?

R-squared is equal to the squared correlation coefficient between the actual and the predicted value of the dependent variable



EXAMPLE 3.4 Determinants of College GPA

From the grade point average regression that we did earlier, the equation with R^2 is

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

$$n = 141,\ R^2 = \boxed{.176}.$$

This means that *hsGPA* and *ACT* together explain about 17.6% of the variation in college GPA for this sample of students. This may not seem like a high percentage, but we must remember that there are many other factors—including family background, personality, quality of high school education, affinity for college—that contribute to a student's college performance. If *hsGPA* and *ACT* explained almost all of the variation in *colGPA*, then performance in college would be preordained by high school performance!

Standard assumptions for the multiple regression model

Assumption MLR.1 **Linear in Parameters**

The model in the population can be written as

10

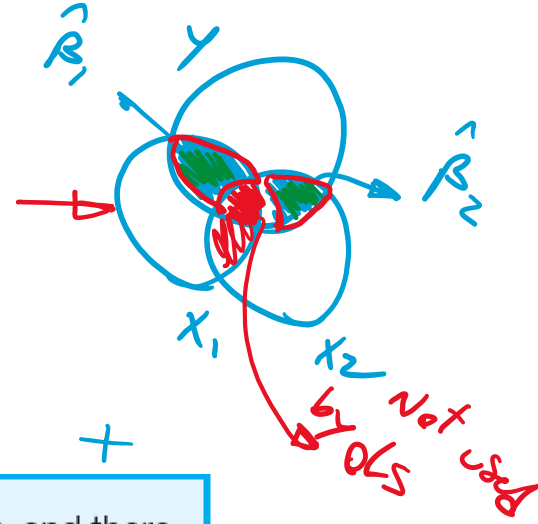
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u, \qquad [3.31]$$

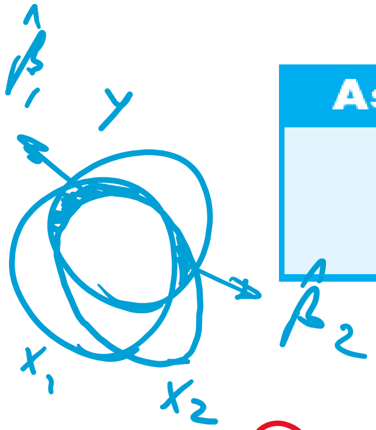
where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

Assumption MLR.2 **Random Sampling**

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

Standard assumptions for the multiple regression model

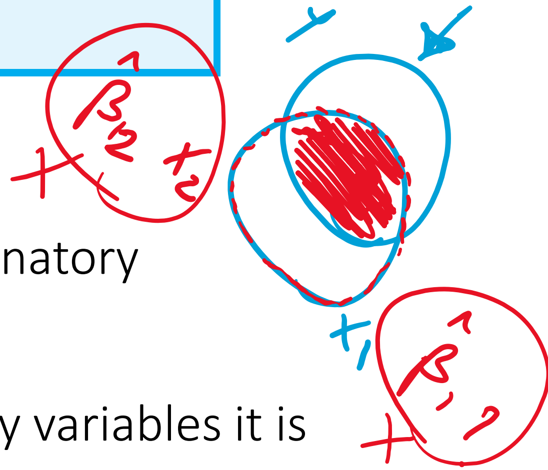




$Var(X) > 0$

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 linear* relationships among the independent variables.



1. The assumption only rules out perfect collinearity/correlation between explanatory variables; imperfect correlation is allowed
2. If an explanatory variable is a perfect linear combination of other explanatory variables it is superfluous and may be eliminated
3. MLR.3 fails if $n < k + 1$. Intuitively, this makes sense: to estimate $k + 1$ parameters, we need at least $k + 1$ observations.

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

$n = 3$

Examples for perfect collinearity

$$\boxed{X_1 + X_2 = X_3}$$

$$\boxed{X_1 + X_2 - X_3 = 0}$$

$$voteA = \beta_o + \beta_1 expendA + \beta_2 expendB + \beta_3\ totalexpend\ + u$$

Either expendA or expendB or totalexpend will have to be dropped from the regression because there is an exact linear relationship between them: expendA + expendB = totalexpend

- Let $\boldsymbol{vote\ A}$ be the percentage of the vote for Candidate A
- let $expendA$ be campaign expenditures by Candidate A, let $expendB$ be campaign expenditures by Candidate B, and let $totexpend$ be total campaign expenditures

Standard assumptions for the multiple regression model (cont.)

Assumption MLR.4 **Zero Conditional Mean**

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0. \implies Cov(u, x_j) = 0 \quad [3.36]$$

- The value of the explanatory variables x_j must contain no information about the mean of the unobserved factors
- In a multiple regression model, the zero conditional mean assumption is much more likely to hold because fewer things end up in the error.
- Example: Average test scores

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 \cancel{avginc} + u$$

SRM

$Cov(u, exp) \neq 0$

avginc

MRM

$Cov(u, exp) = 0$

If avginc was not included in the regression, it would end up in the error term; it would then be hard to defend that expend is uncorrelated with the error

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$

$Corr(X_1, u) \neq 0 \Rightarrow X_1$ is endog

Endogeneity vs. Exogeniety

- Explanatory variables that are **correlated** with the error term are called **endogenous**;
  endogeneity is a violation of assumption MLR.4

$$\boxed{Corr(x, u) \neq 0}$$

- Explanatory variables that are **uncorrelated** with the error term are called **exogenous**;
  MLR.4 holds if all explanatory variables are exogenous

$$\boxed{Corr(x, u) = 0}$$

$Corr(X_2, u) = 0$

$Corr(X_3, u) = 0$

$X_2, X_3 \rightarrow$ exog

- Exogeneity is the key assumption for a **causal interpretation** of the regression, and for **unbiasedness of the OLS estimators**

Theorem 3.1 (Unbiasedness of OLS)

THEOREM 3.1

UNBIASEDNESS OF OLS

Under Assumptions MLR.1 through MLR.4,

$$E(\hat{\beta}_j) = \beta_j, \; j = 0, 1, \ldots, k, \qquad [3.37]$$

for any values of the population parameter β_j. In other words, the OLS estimators are unbiased estimators of the population parameters.

$E(\hat{\beta}_1) = \beta_1$

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

Standard assumptions for the multiple regression model (cont.)

Assumption MLR.5 **Homoskedasticity**

The error u has the same variance given any value of the explanatory variables. In other words, $\mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2$.

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

- Example: Wage equation

$$Var(u_i|educ_i, exper_i, tenure_i) = \sigma^2$$

This assumption may also be hard to justify in many cases

Theorem 3.2 (Sampling variances of the OLS slope estimators)

SRM $Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_X}$

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

THEOREM 3.2

SAMPLING VARIANCES OF THE OLS SLOPE ESTIMATORS

Under Assumptions MLR.1 through MLR.5, conditional on the sample values of the independent variables,

MLR

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \quad [3.51]$$

x_1, x_2, x_3

SST_j

for $j = 1, 2, \ldots, k$, where $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all other independent variables (and including an intercept).

R_j^2 => Captures the relationship btw x_j and x_{-j}

wage ~ educ + age + exper

x_1 = educ x_{-1} : age, expe

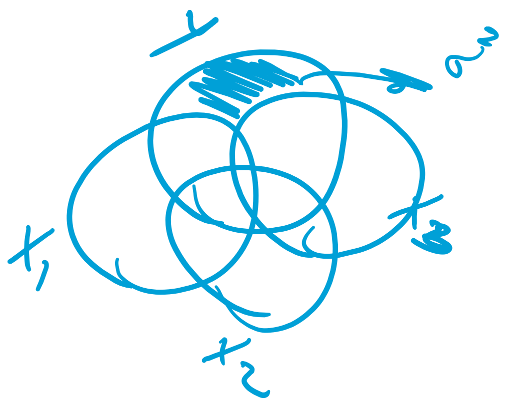
j explan
i obs

$x_j = x_3$ = tenure

$SST_j = \sum(\text{tenure}_i - \overline{\text{tenure}})^2$

Components of OLS Variances:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{SST}_j(1 - R_j^2)},$$

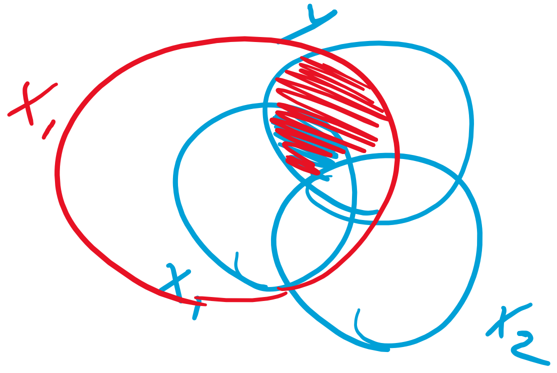


1) The error variance

- A high **error variance** increases the sampling variance because there is more "noise"
- The error variance **does not decrease with sample size**. Remember, σ^2 is a feauter of the population, it has nothing to do with the sample size.
- there is really only **one way to reduce the error variance**, and that is to **add more explanatory variables** to the equation (Not always possible to find good candidates though!)

Components of OLS Variances:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$



2) The total sample variation in the explanatory variable

SST_j

- More sample variation in explanatory variable j leads to more precise estimates $\boxed{\text{(lower varicane of } \beta_j\text{)}}$
- Total sample variation automatically increases with the sample size
- Increasing the sample size is thus a way to get more precise estimates

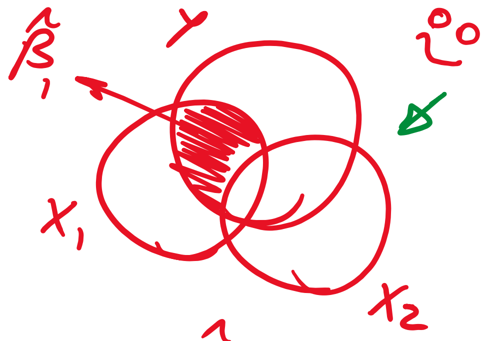
$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \uparrow$$

$n \uparrow \quad \mathrm{Var}(\hat{\beta}_j) \downarrow$

New

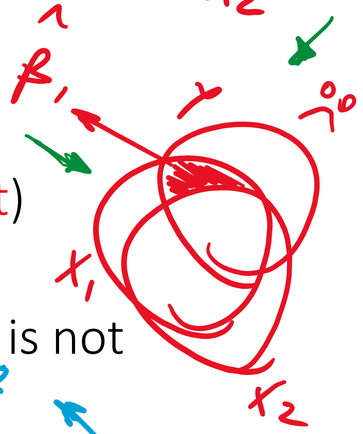
Components of OLS Variances (cont'd)

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$



3) Linear relationships among the independent variables

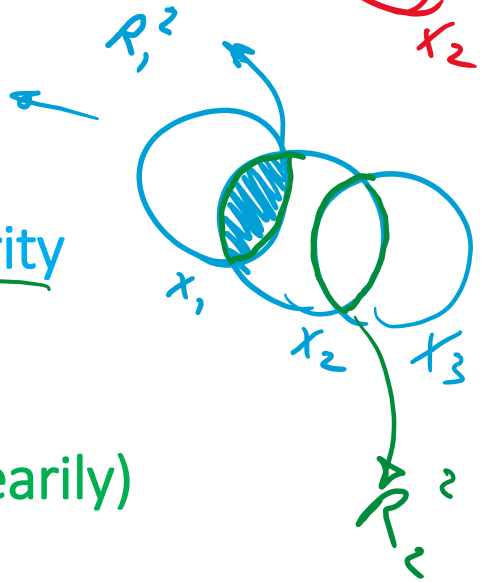
R_j^2 comes from **regressing x_j on x_{-j}** : all other independent variables including a constant)



Higher R_j^2 means that x_j can be **better** explained by the other independent variables. This is not a good thing!

* R_1^2 : X_1 on X_2, X_3

* R_2^2 = Reg X_2 on X_1, X_3



The problem of almost linearly dependent explanatory variables is called multicollinearity

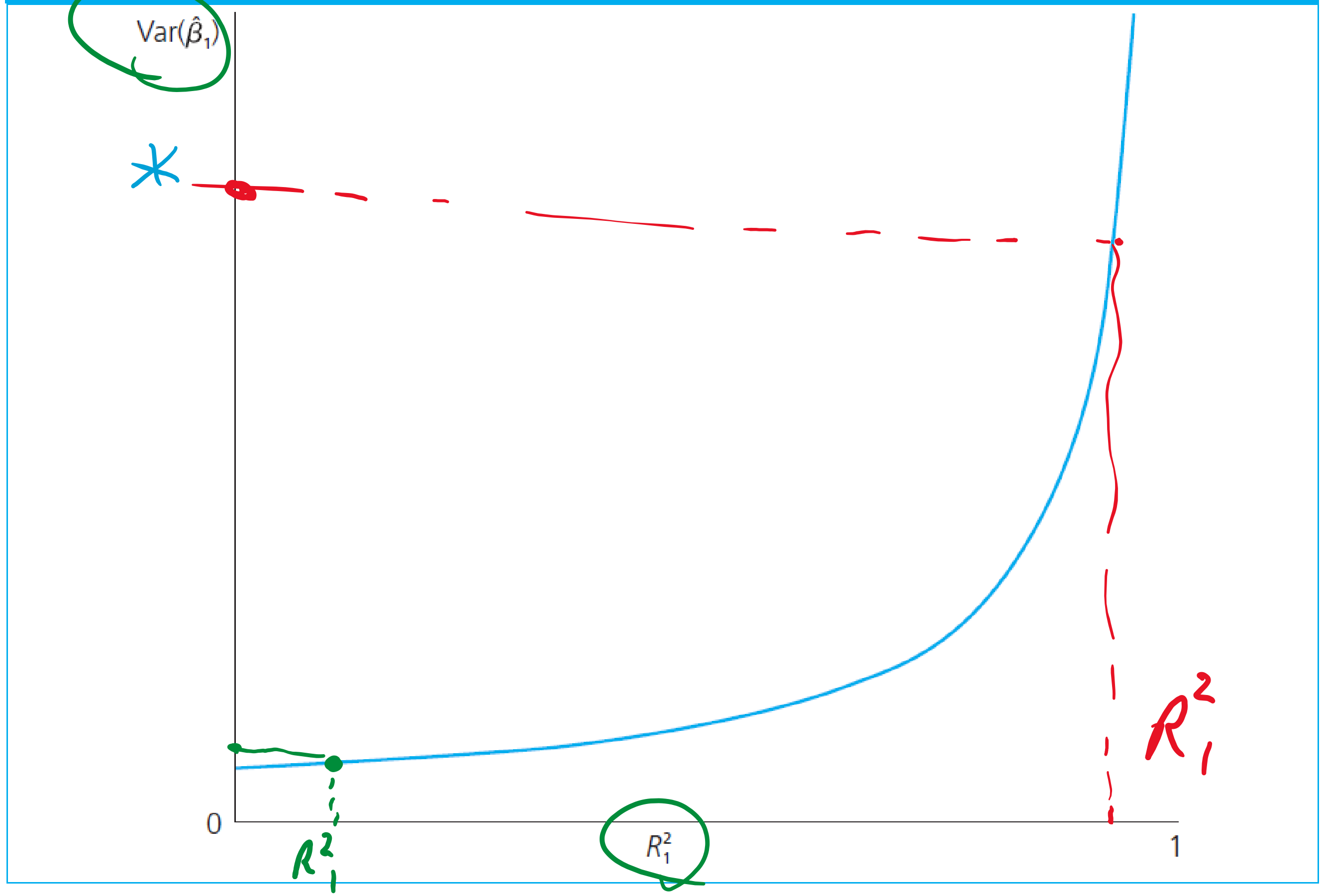
$$R_j^2 \to 1$$

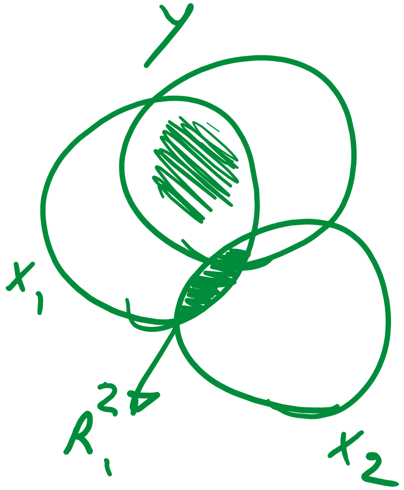
Multicollinearity is NOT a violation of MLR3 (No perfect collinearily)

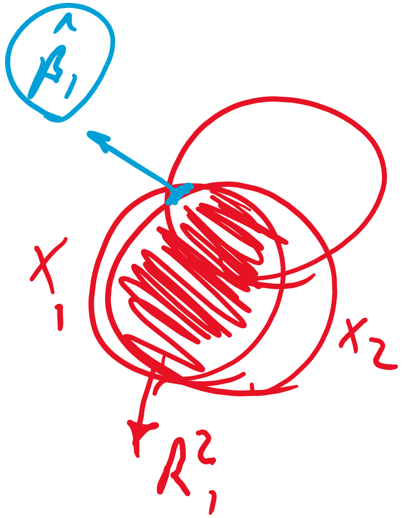
Components of OLS Variances (cont'd)

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{SST}_j(1 - R_j^2)},$$

FIGURE 3.1 $\mathrm{Var}(\hat{\beta}_1)$ as a function of R_1^2.







Next class?

1. Including **irrelevant variable** (overspecification)
2. **Ommitting relevant** variables (omitted variable bias)
3. How to deal with **multicollinearity**?
4. Estimation of sampling **variances** of the OLS estimators
5. **Efficiency** of OLS