Economics 403A

Multiple Regression Concepts

Dr. Randall R. Rojas

Today's Class

- Introductory Concepts
  - Projection
  - Frisch-Waugh Theorem
  - Partial Correlation
  - Adjusted R^2

Residuals vs. Disturbances

ε = Disturbances (Population)

e = Residuals (Sample)

$$\text{In the population}: E[\mathbf{X}'\boldsymbol{\varepsilon}] = \mathbf{0}$$

$$\text{In the sample}: \quad \frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i} e_i = \mathbf{0}$$

Residuals vs. Disturbances

Disturbances (population) $y_i - \mathbf{x}_i'\beta = \varepsilon_i$

Partitioning **y**: $\mathbf{y} = E[\mathbf{y}|\mathbf{X}] + \boldsymbol{\varepsilon}$

= conditional mean + disturbance

Residuals (sample) $y_i - \mathbf{x}_i'\mathbf{b} = e_i$

Partitioning **y** : $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$

= projection + residual

(**Note : Projection into the column space of X**, i.e., the set of linear combinations of the columns of **X** - **Xb** is one of these.)

Projection

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i$$ (Stochastic Relation)

$$E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\beta$$ (Population Regression)

The estimate of $E[y_i|\mathbf{x}_i]$ is denoted by $\widehat{y_i} \rightarrow \widehat{y_i} = \mathbf{x}_i'\mathbf{b}$

$$\varepsilon_i = y_i - \mathbf{x}_i'\beta$$ (Disturbance associated with point *i*)

$$e_i = y_i - \mathbf{x}_i'b$$ (Residual = estimate of ε_i)

$$\boxed{y_i = \mathbf{x}_i'\beta + \varepsilon_i = \mathbf{x}_i'\mathbf{b} + e_i}$$

Projection

In general for the multiple regression case:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \mathbf{X}\mathbf{b} \\ &= \mathbf{M}\mathbf{y} \end{aligned}$$

M = Residual Maker because **MX** = **0**

$$\boxed{\mathbf{e} = \mathbf{M}\mathbf{y}}$$

Projection

Recall that: $\mathbf{y} = \widehat{\mathbf{y}} + \mathbf{e}$, where **e** = $\mathbf{My}$

$$
\begin{aligned}
\widehat{\mathbf{y}} &= \mathbf{y} - \mathbf{e} \\
&= \mathbf{y} - \mathbf{My} \\
&= (\mathbf{I} - \mathbf{M})\mathbf{y} \\
&= \mathbf{Py}
\end{aligned}
$$

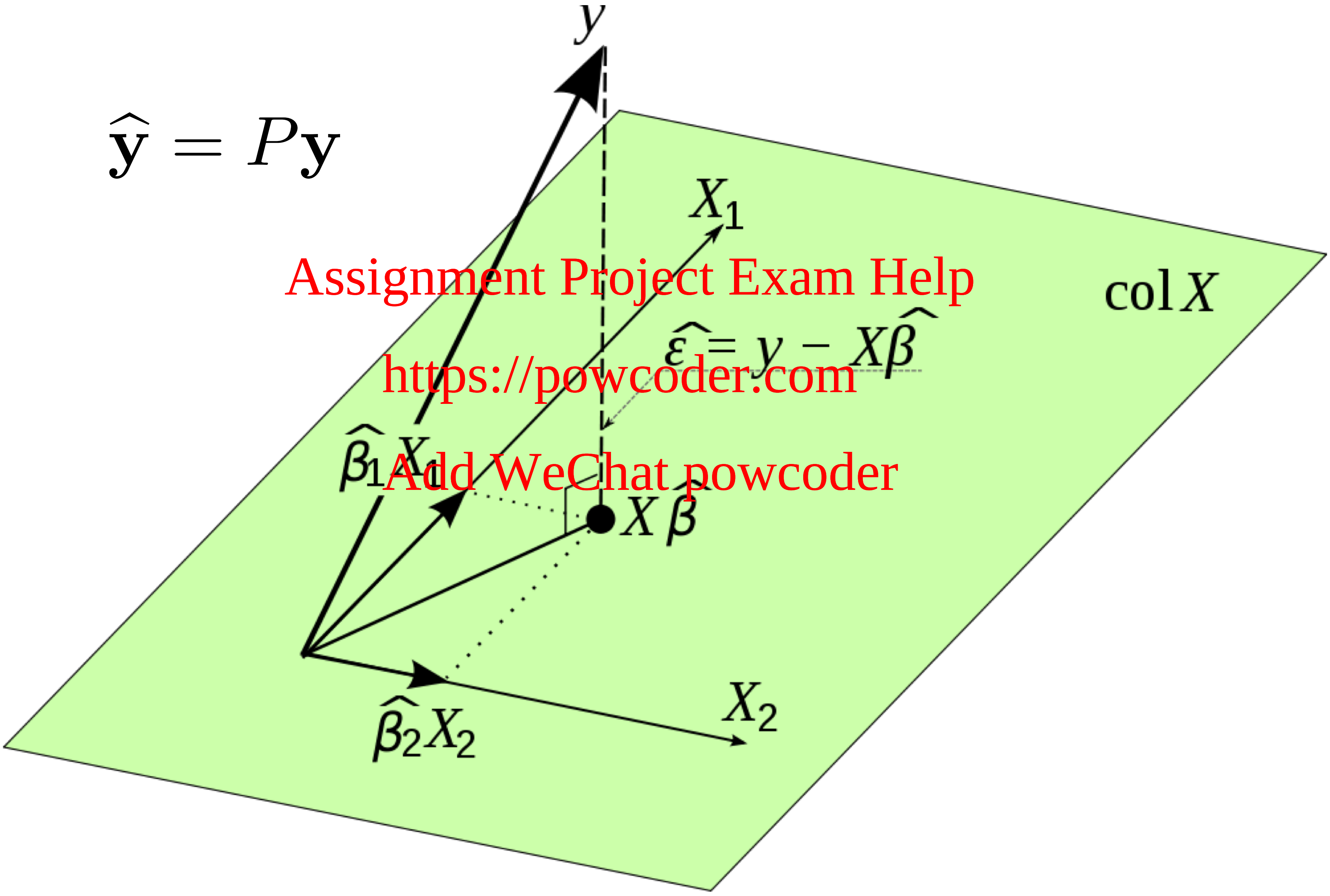
(**P** = Projection Matrix)

$$
\boxed{\mathbf{y} = \mathbf{Py} + \mathbf{My}}
$$

(Projection + Residual)

Projection of **y** into the column space of **X**

$$\widehat{\mathbf{y}} = P\mathbf{y}$$



Frisch-Waugh (1933) Theorem

- Consider the Model: $\mathbf{y \sim \beta_1 X_1 + \beta_2 X_2}$

- The Frisch-Waugh theorem says that the multiple regression coefficient of any single variable can also be obtained by first netting out (*partialing out*) the effect of other variable(s) in the regression model from both the dependent variable and the independent variable.

Frisch-Waugh (1933) Theorem
Algorithm

- **Model:** log(wages) ~ $\beta_1$educ + $\beta_2$exper
  - X_1 = educ, X_2 = exper, y =log(wages)
- **Goal:** Estimate β_1
- **Step 1:** regress y ~ X_2 → save the residuals (= res1)
- **Step 2:** regress X_1 ~ X_2 → save the residuals. (= res2)
- **Step 3:** regress res1 ~ res2 → β_1

Frisch-Waugh (1933) Theorem
R Example

```
library(AER)
library(np)
data(wage1)
```


Estimate the model to compare results:

Step 0: m0 = lm(lwage ~ educ + exper, data = wage1)

Step 1: m1 = lm(lwage ~ exper, data = wage1) → res1 = m1$res

Step 2: m2 = lm(educ ~ exper ,data = wage1) → res2 = m2$res

Step 3: m3 = lm(res1 ~ res2) → β_1 (coefficient for educ)

Frisch-Waugh Result

We can show after some algebraic manipulation that:

$$\mathbf{b}_2 \ = \ [\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2]^{-1}[\mathbf{X}_2'\mathbf{M}_1\mathbf{y}].$$

- This is Frisch and Waugh's famous result - the "double residual regression."

- How do we interpret this? A regression of residuals on residuals.

- "We get the same result whether we (1) detrend the other variables by using the residuals from a regression of them on a constant and a time trend and use the detrended data in the regression or (2) just include a constant and a time trend in the regression and not detrend the data"

- "Detrend the data" means compute the residuals from the regressions of the variables on a constant and a time trend.

Partial Correlation

- Let X and Y be the two variables we have found to be correlated.
- Let r(X,Y) = simple correlation
- Introduce a third variable Z which may or may not mediate the relationship between X and Y.
- Let r((X,Y)|Z) = partial correlation of X and Y, controlling for Z.
- Result: If r(X,Y) is relatively large, but r((X,Y)|Z) is much smaller, we can conclude that Z is a mediating variable. Z may explain, at least in part, the observed relationship between X and Y.

Partial Correlation

- Study on language skills (=x) and children's toe size (=y).
  - **Finding**: strong correlation between language skills and size of the big toe.
- Changes in X may be a cause of changes in Y (or vice versa).
- How about a third variable, Z (e.g., age)?
- Could Z be producing changes in both X and Y?

Partial Correlation

- Study results:
- X = measure of language skills
- Y = size of big toe
- Z = child's age
- r(X,Y) = 0.40
- r(X,Z) = 0.55, r(Y,Z) = 0.65
- r((X,Y)|z) = 0.07 (much smaller than 0.4)

→ Age explains the relationship between language and big toe size

Partial Correlation

- Consider the Model:

  $\log(\text{wages}) \sim \beta_1 \text{educ} + \beta_2 \text{exper}$

Q: How much of the measured correlation between wages and educ reflects a direct relation between them, instead of the fact that wages and education level tend to rise with age?

A: *Partial Correlation Coefficient*

Partial Correlation

Algorithm

- **Step 1:** y_* = the residuals in a regression of wages on a constant and age

- **Step 2:** z_* = the residuals in a regression of education on a constant and age

- **Step 3:** r^*_{yz} = partial correlation of wages and education controlling for age.
- *Note:* In R, you can use the "ppcor" package

$$\boxed{r^{*2}_{yz} = \frac{(\mathbf{z}'_*\mathbf{y}_*)^2}{(\mathbf{z}'_*\mathbf{z}_*)(\mathbf{y}'_*\mathbf{y}_*)}}$$

Partial Correlation

R Example

```
library(AER)
library(np)
library(ppcor)

data("PSID1976", package = "AER")
PSID1976$kids <- with(PSID1976, factor((youngkids + oldkids) > 0,
  levels = c(FALSE, TRUE), labels = c("no", "yes")))
PSID1976$nwincome <- with(PSID1976, (fincome - hours * wage)/1000)
wage = PSID1976$wage
age = PSID1976$age
education = PSID1976$education
y.data <- data.frame(wage, age, education)
# Compute/Interpret 'simple correlation'
cor(y.data)
# Compute/Interpret 'partial correlation'
pcor(y.data)
# Partial correlation between "wage" and "educ" given "age"
pcor.test(y.data$wage,y.data$education,y.data[,c("age")])
```

Measure of Fit
Theorem

Change in R^2 When a Variable is Added to the Regression:

- Let $R^2_{\mathbf{X}z}$ = R^2 from the regression of y on **X** and *additional variable* z.
- Let $R^2_{\mathbf{X}}$ = R^2 from the regression of y on **X**.
- Let r^*_{yz} = partial correlation between y and z controlling for X.

$$\boxed{R^2_{\mathbf{X}z} = R^2_{\mathbf{X}} + (1 - R^2_{\mathbf{X}}) r^{*\,2}_{yz}}$$

Measure of Fit
Adjusted R^2

- Let n = number of observations and K = number of parameters estimated. We define the adjusted R^2 as:

$$\overline{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2)$$

- Theorem: *Change in adjusted R^2 When a Variable is Added to the Regression:*

In a multiple regression, adjusted R^2 will fall (rise) when the variable x is deleted from the regression if the square of the t-ratio associated with this variable is greater(less) than 1.

Adjusted R^2
R Example

```
library(AER)
library(np)
library(ppcor)
library(AER)
data("PSID1976", package = "AER")
```


Model 1:

```
fit1 = lm(wage ~ experience+education, data =  PSID1976)
summary(fit1)
```

Q: Look at R^2 and adjusted R^2. What should happen to them if we add another variable to model?

Model 2:

```
fit2 = lm(wage ~ experience+education+youngkids, data =  PSID1976)
summary(fit2)
```