Probability & Statistics Review, Linear Regression Models, and OLS Inference

Review Session

CLABE 2025/2026

Marco Rosso

15 December 2025

Review Goals

- Refresh large-n inference for means: test statistic, p-value, confidence intervals (normal approximation)
- Revisit the Simple Linear Regression Model (SLRM): assumptions, OLS estimator, sampling variability
- Revisit Multiple Linear Regression Models (MLRM): ceteris paribus interpretation and omitted variable bias intuition
- Practice OLS inference: (asymptotic) t/z tests, confidence intervals, and goodness-of-fit (R^2, $\bar{R}^2$)
- Joint hypotheses / exclusion restrictions: restricted vs unrestricted models and the F-statistic
- Translate each step into (i) built-in Stata commands and (ii) "by hand" computations

Foundations: Random Variables and Distributions

Population Parameter (unknown):

$$\mu_Y = E(Y) = \int y \cdot f(y)\, dy \quad \text{(population mean)}$$

Population Variance (measures spread):

$$\sigma_Y^2 = \text{Var}(Y) = E[(Y - \mu_Y)^2]$$

Covariance (linear relationship):

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Correlation (scaled covariance):

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \quad -1 \leq \text{Corr}(X, Y) \leq 1$$

Random Sampling: $Y_1, Y_2, \ldots, Y_n$ are **i.i.d.** (independently and identically distributed)

Sample Mean: Estimation and Properties (1)

Sample Mean (unbiased estimator of μ_Y):

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Key Properties:

- **Unbiasedness:** $E(\overline{Y}) = \mu_Y$
- **Variance of $\overline{Y}$:** $\text{Var}(\overline{Y}) = \frac{\sigma_Y^2}{n}$
- **Standard Error (estimated):** $\text{SE}(\overline{Y}) = \frac{s_Y}{\sqrt{n}}$

where

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

is the sample variance

Sample Mean: Estimation and Properties (2)

Large-n inference (CLT / normal approximation):

$$t \equiv \frac{\overline{Y} - \mu_0}{s_Y/\sqrt{n}} \quad \xrightarrow[H_0]{a} \quad N(0,1)$$

95% CI (large n):

$$\overline{Y} \pm 1.96 \cdot \frac{s_Y}{\sqrt{n}}$$

Sample Mean Inference in Stata (1)

Estimate mean and standard error:

```
// Load data
use firm_data.dta, clear

// Descriptive statistics
summarize sales_employee fwp

// Standard error of mean (manual)
display "SE(mean) = " %6.4f r(sd)/sqrt(r(N))
```

Hypothesis test (Stata built-in): $H_0 : \mu_Y = \mu_0$

```
ttest sales_employee = 230
```

Sample Mean Inference in Stata (2)

Confidence interval (manual large-n):

```
summarize sales_employee
scalar ybar = r(mean)
scalar sd   = r(sd)
scalar n    = r(N)

display "95% CI: [" %6.2f (ybar - 1.96*sd/sqrt(n)) ///
    ", " %6.2f (ybar + 1.96*sd/sqrt(n)) "]"
```

Manual: Difference-in-Means (Two Groups) (1)

Statistic:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \qquad \text{(large } n\text{: } t \approx N(0,1)\text{)}$$

```
// Plug-in example numbers (replace with your output)
scalar n1 = 238
scalar n2 = 238
scalar mean1 = 238.49
scalar mean2 = 223.96
scalar sd1 = 149.09
scalar sd2 = 101.63
```

Manual: Difference-in-Means (Two Groups) (2)

```
scalar se_diff = sqrt((sd1^2)/n1 + (sd2^2)/n2)
scalar t_stat = (mean1 - mean2) / se_diff

// Large-n two-sided p-value
scalar pval = 2 * (1 - normal(abs(t_stat)))

display "t = " %6.3f t_stat
display "p-value = " %6.4f pval
```

Manual: Confidence Interval for Difference in Means

95% CI (large n):

$$(\overline{Y}_1 - \overline{Y}_2) \pm 1.96 \cdot SE(\overline{Y}_1 - \overline{Y}_2)$$

```
scalar diff = mean1 - mean2
scalar ci_l = diff - 1.96*se_diff
scalar ci_u = diff + 1.96*se_diff

display "Diff = " %6.2f diff
display "95% CI: [" %6.2f ci_l ", " %6.2f ci_u "]"
```

Key rule: CI and two-sided tests are equivalent.

Simple Linear Regression Model (SLRM)

Population Regression Function:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Key assumption (mean independence):

$$E(\varepsilon_i \mid X_i) = 0 \quad \Rightarrow \quad E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$

Other common assumptions:

- Random sampling: (Y_i, X_i) are i.i.d.
- No perfect collinearity: X varies in the sample
- Homoskedasticity (when used): $\text{Var}(\varepsilon_i \mid X_i) = \sigma^2$

Interpretation: β_1 is the change in $E(Y \mid X)$ for a one-unit increase in X.

OLS Estimators

OLS minimizes the Sum of Squared Residuals (SSR):

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

OLS slope and intercept:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Key properties (under standard assumptions):

- Unbiasedness: $E(\hat{\beta}_1) = \beta_1$
- Consistency: $\hat{\beta}_1 \xrightarrow{p} \beta_1$

Simple Regression: Stata

Estimate SLRM:

```
regress sales_employee fwp
```

Test on slope:

```
test fwp = 0

// Access stored results
display "b = "  _b[fwp]
display "se = " _se[fwp]
display "z/t = " _b[fwp]/_se[fwp]
```

Fitted values and residuals:

```
predict yhat, xb
predict ehat, resid
summarize yhat ehat
```

Manual Computation: OLS Slope from Summary Stats (1)

Using correlation identity (simple regression):

$$\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \rho_{XY} \cdot \frac{s_Y}{s_X}$$

```
// Get sd(y), sd(x), corr(x,y)
summarize sales_employee
scalar sd_y = r(sd)
summarize fwp
scalar sd_x = r(sd)

correlate sales_employee fwp
matrix C = r(C)
scalar corr_xy = C[1,2]
```

Manual Computation: OLS Slope from Summary Stats (2)

```
// beta1 = corr * sd_y / sd_x
scalar beta1 = corr_xy * (sd_y/sd_x)

// intercept: beta0 = ybar - beta1*xbar
summarize sales_employee, meanonly
scalar ybar = r(mean)
summarize fwp, meanonly
scalar xbar = r(mean)

scalar beta0 = ybar - beta1*xbar

display "beta1 (manual) = " %8.4f beta1
display "beta0 (manual) = " %8.4f beta0
```

Manual Computation: $\widehat{\sigma}^2$ and SE($\hat{\beta}_1$) (SLRM) (1)

```
// After: regress y x
// (here y = sales_employee, x = fwp)
predict yhat, xb
gen ehat = sales_employee - yhat

// SSR = sum ehat^2
gen ehat2 = ehat^2
quietly summarize ehat2, meanonly
scalar SSR = r(sum)

// n and k (SLRM has k=2: constant + x)
scalar n = e(N)
scalar k = 2
```

Manual Computation: $\widehat{\sigma}^2$ and SE($\hat{\beta}_1$) (SLRM) (2)

```
// sigma^2 hat = SSR/(n-k) = SSR/(n-2)
scalar sig2hat = SSR/(n-k)

// SSX = sum (x - xbar)^2
quietly summarize fwp, meanonly
scalar xbar = r(mean)
gen xdev2 = (fwp - xbar)^2
quietly summarize xdev2, meanonly
scalar SSX = r(sum)

// Var(b1) = sigma^2 / SSX ; SE = sqrt(Var)
scalar var_b1 = sig2hat/SSX
scalar se_b1  = sqrt(var_b1)

display "SSR = " SSR
display "sig2hat = " sig2hat
display "Manual SE(b1) = " se_b1
```

Manual: Test and Confidence Interval on Slope (Large n) (1)

Test statistic (usual t-stat; large n normal approximation):

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} \approx N(0,1)$$

```
// After regress
scalar b1  = _b[fwp]
scalar se1 = _se[fwp]
scalar b10 = 0

scalar tstat = (b1 - b10)/se1
scalar pval  = 2*(1 - normal(abs(tstat)))
```

Manual: Test and Confidence Interval on Slope (Large n) (2)

```
display "t/z = " %6.3f tstat
display "p-value = " %6.4f pval

// 95% CI (large-n)
scalar ci_l = b1 - 1.96*se1
scalar ci_u = b1 + 1.96*se1
display "95% CI: [" %6.3f ci_l ", " %6.3f ci_u "]"
```

Goodness of Fit: R^2

Decomposition:

$$\text{TSS} = \text{ESS} + \text{SSR}$$

- $\text{TSS} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$
- $\text{SSR} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$
- $\text{ESS} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = \text{TSS} - \text{SSR}$

R^2:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}$$

In simple regression: $R^2 = \text{Corr}(X, Y)^2$.

Manual Computation: R^2 via *TSS* and *SSR* (1)

```
// After: regress y x
predict yhat, xb
gen ehat = sales_employee - yhat

// SSR
gen ehat2 = ehat^2
quietly summarize ehat2, meanonly
scalar SSR = r(sum)
```

Manual Computation: R^2 via *TSS* and *SSR* (2)

```
// TSS = sum (y - ybar)^2
quietly summarize sales_employee, meanonly
scalar ybar = r(mean)
gen ydev2 = (sales_employee - ybar)^2
quietly summarize ydev2, meanonly
scalar TSS = r(sum)

// ESS = TSS - SSR
scalar ESS = TSS - SSR

scalar R2  = ESS/TSS
display "R2 (manual) = " R2
```

Multiple Linear Regression Model (MLRM)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Key assumption:

$$E(\varepsilon_i \mid X_{1i}, \ldots, X_{ki}) = 0$$

Interpretation: β_j is the **ceteris paribus** effect of X_j on $E(Y \mid X)$.

Omitted Variable Bias (intuition, 2 regressors):

$$\text{Bias}(\hat{\beta}_1) = \beta_2 \cdot \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$

Multiple Regression: Stata (1)

Estimate MLRM:

```
regress sales_employee fwp mgmt us_location
```

Single coefficient test:

```
test fwp = 0
```

Joint hypothesis test:

```
test mgmt us_location
```

Multiple Regression: Stata (2)

LR test (if shown in class notes for OLS):

```
// Store models
regress sales_employee fwp
estimates store r

regress sales_employee fwp mgmt us_location
estimates store ur

// LR test compares Gaussian log-likelihoods
lrtest ur r
```

Note: In linear regression, LR tests correspond to comparing restricted vs unrestricted fits under a Gaussian likelihood.

F-statistic for Joint Restrictions (1)

Null: q exclusion restrictions.

Restricted vs Unrestricted:

- Restricted (under H_0): SSR_0
- Unrestricted (under H_1): SSR_1
- q = number of restrictions
- k = number of regressors **including the constant**

Note: *In some textbooks k denotes the number of slope regressors only, and the denominator is written as $\mathrm{SSR}_1/(n-k-1)$. The two formulas are* **numerically identical***; only the notation differs.*

F-statistic for Joint Restrictions (2)

Statistic:

$$F = \frac{(\mathrm{SSR}_0 - \mathrm{SSR}_1)/q}{\mathrm{SSR}_1/(n-k)}$$

Asymptotic distribution under H_0 (large n):

$$F \xrightarrow[H_0]{a} \frac{\chi^2_q}{q}$$

Intuition: how much the SSR increases when moving from the unrestricted to the restricted model, relative to the unexplained variance in the unrestricted model.

Manual Computation: F from SSR_0 and SSR_1

```
// Example: test mgmt=0 and us_location=0 jointly

// Restricted
regress sales_employee fwp
scalar SSR0 = e(rss)

// Unrestricted
regress sales_employee fwp mgmt us_location
scalar SSR1 = e(rss)
scalar n    = e(N)

// k includes the constant
scalar k = 1 + 3   // constant + (fwp mgmt us_location)
scalar q = 2       // restrictions: mgmt=0 and us_location=0

scalar Fstat = ((SSR0 - SSR1)/q) / (SSR1/(n - k))
display "F = " %8.4f Fstat
```

Dummy Variables

Binary regressor: $US = 1$ if firm is in US, else 0.

```
regress sales_employee fwp us_location
```

Interpretation:

- For $US = 0$: $E(Y \mid X, US = 0) = \beta_0 + \beta_1 X$
- For $US = 1$: $E(Y \mid X, US = 1) = (\beta_0 + \beta_2) + \beta_1 X$
- β_2 is the ceteris paribus mean difference (US vs non-US)

Multiple categories: omit one category (reference) to avoid the dummy trap.

Stata Command Summary

Task	Stata Command
Summary statistics	`summarize y x1 x2`
Correlation	`correlate y x1 x2`
Difference-in-means	`ttest y, by(group)`
Simple regression	`regress y x`
Multiple regression	`regress y x1 x2 x3`
Test single coef	`test x1 = 0`
Joint test (F)	`test x2 x3`
Fitted values	`predict yhat, xb`
Residuals	`predict ehat, resid`
Model comparison (if shown)	`lrtest ur r`

Key Formulas

- **Sample mean:** $\overline{Y} = \frac{1}{n}\sum Y_i, \quad SE(\overline{Y}) = \frac{s_Y}{\sqrt{n}}$
- **Mean test statistic:** $t = \frac{\overline{Y} - \mu_0}{s_Y/\sqrt{n}} \approx N(0,1)$ (large n)
- **95% CI (large n):** $\overline{Y} \pm 1.96 \cdot SE(\overline{Y})$
- **OLS slope (SLRM):** $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
- **Error variance (SLRM):** $\hat{\sigma}^2 = \frac{\mathrm{SSR}}{n-2}$ (more generally $\mathrm{SSR}/(n-k)$)
- **Regression test statistic:** $t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)} \approx N(0,1)$ (large n)
- **R^2:** $R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{TSS}}$
- **F-statistic:**

$$F = \frac{(\mathrm{SSR}_0 - \mathrm{SSR}_1)/q}{\mathrm{SSR}_1/(n-k)} \xrightarrow{a} \frac{\chi^2_q}{q}$$

Common Mistakes & Sanity Checks

Common mistakes:

1. $\mathrm{SSR} > \mathrm{TSS}$: impossible (check computations)
2. $R^2 > 1$ or $R^2 < 0$: computational error
3. All dummies included: perfect multicollinearity
4. Correlation $\neq$ causation
5. Forgetting df: use $(n - k)$ with k incl. constant
6. Mixing finite-sample vs large-n critical values without stating it

Sanity checks:

- Do signs make economic sense?
- Are magnitudes reasonable?
- Are SEs smaller when n is larger (all else equal)?
- Does adding controls change $\hat{\beta}_1$ in the direction predicted by OVB?

Tips

1. Use **large-n critical values**: 1.96 for 5% two-sided
2. **Confidence intervals** and **two-sided tests** are equivalent
3. In MLRM, coefficients are **ceteris paribus** effects
4. Joint hypotheses: use **restricted vs unrestricted** (F test)
5. Dummy variables: always define a **reference** group
6. OVB sign depends on β_2 and $\text{Cov}(X_1, X_2)$
7. R^2 is fit, not causality; compare specifications carefully
8. Always check degrees of freedom: $n - k$ with k incl. constant

Worked Example: Complete Analysis (1)

Question: Effect of family-friendly policies on productivity, with controls?

```
// 1. Explore
summarize sales_employee fwp mgmt us_location
correlate sales_employee fwp mgmt us_location

// 2. Simple regression
regress sales_employee fwp
estimates store r

// 3. Multiple regression
regress sales_employee fwp mgmt us_location
estimates store ur
```

Worked Example: Complete Analysis (2)

```
// 4. Joint test (built-in)
test mgmt us_location

// 5. LR test (if shown in notes for OLS)
lrtest ur r

// 6. "By hand" sanity check (t/z)
scalar tstat = _b[fwp]/_se[fwp]
display "t/z = " tstat
```