Wrap up

Final Lecture: Part I

Nayel Bettache

Linear Model: Assumptions and Interpretations

Model:

$$Y = X\beta + \varepsilon \quad \text{where } \varepsilon \text{ represents random errors.}$$

Key Assumptions:

- **Linearity:** The expected value of Y is a linear function of X.
- **Independence:** The errors ε_i are independent across observations.
- **Homoscedasticity:** The errors have constant variance: $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all i.
- **Normality:** The errors ε_i are normally distributed (required mainly for valid *inference*, not for point estimates).
- **Low Multicollinearity:** Predictors should not be nearly perfectly correlated (to ensure stable estimation of β).

Notes:

- Normality is *not required* for estimating β, but important for hypothesis tests.
- Multicollinearity does not bias estimates, but increases their variance.

Consequences of Assumption Violations

Reminder: Under the Gauss-Markov assumptions, OLS provides the **Best Linear Unbiased Estimator (BLUE)**.

If assumptions are violated:

- **Linearity violation:**
  $\rightarrow$ Model is misspecified $\Rightarrow$ Biased estimates, poor predictions.
- **Independence violation:**
  $\rightarrow$ Standard errors are underestimated $\Rightarrow$ Inference becomes invalid (e.g., misleading p-values).
- **Homoscedasticity violation:**
  $\rightarrow$ Estimates remain unbiased but are inefficient $\Rightarrow$ Larger standard errors, invalid usual inference (need robust methods).
- **Normality violation:**
  $\rightarrow$ OLS estimates are still unbiased and consistent, but *small-sample inference* (e.g., t-tests, F-tests) may be invalid.
- **Multicollinearity:**
  $\rightarrow$ Estimates remain unbiased, but coefficients are highly sensitive to data perturbations $\Rightarrow$ Inflated standard errors, unstable predictions.

Checking Model Assumptions

Visual Diagnostics:

- **Residual plots**:
  Check for patterns $\Rightarrow$ Linearity and homoscedasticity.
- **Q-Q plots (Quantile-Quantile plots)**:
  Check if residuals align with a normal distribution.

Statistical Tests:

- **Durbin-Watson test**:
  Detect autocorrelation in residuals (independence violation).
- **Variance Inflation Factor (VIF):**
  Quantify multicollinearity among predictors.

Note: Always combine visual inspection and statistical tests for robust diagnostics.

Multiple Linear Regression: Fundamentals

Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

where ε_i are random errors with mean 0 and constant variance σ^2.

Matrix Form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Least Squares Estimation:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

(provided that $\mathbf{X}^T\mathbf{X}$ is invertible)

Fitted Values:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Objective: Minimize the residual sum of squares (RSS): $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

Multiple Linear Regression: Interpretation

- **Coefficient β_j:**
  Expected change in Y for a one-unit increase in X_j, *holding all other predictors constant*.
- **Coefficient of Determination (R^2):**
  Proportion of variance in Y explained by the model.
  $$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$
- **Adjusted R^2:**
  Adjusts R^2 for the number of predictors.
  Penalizes overfitting.
- **F-test:**
  Tests the joint hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$.
- **t-tests:**
  Test if individual β_j significantly differs from zero.

Categorical Variables and Interactions

Handling Categorical Variables:

- **Dummy variables:**

$$X_{ij} = \begin{cases} 1 & \text{if observation } i \text{ belongs to category } j \\ 0 & \text{otherwise} \end{cases}$$

- **Reference category:**
  One category is omitted to avoid the "dummy variable trap" (perfect multicollinearity).

Modeling Interactions:

- Include *products* of predictors to allow non-additive effects.
- **Interaction model:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon$$

- **Interpretation of β_3:**
  β_3 measures how the effect of X_1 on Y changes depending on the level of X_2.

Sampling Distributions: Key Concepts

Definition:

- The **sampling distribution** of a statistic is the probability distribution of that statistic across all possible random samples from the population.

Examples:

- The sample mean $\bar{X}$ has a normal distribution (by the Central Limit Theorem) when n is large.
- The t-statistic follows a t-distribution under the null hypothesis (for small samples).

Importance:

- Fundamental for building **confidence intervals** and conducting **hypothesis tests**.
- Explains the variability of estimators from sample to sample.

Key Idea:

- Even though a parameter (like μ) is fixed, the statistic (like $\bar{X}$) is random across samples.

Central Limit Theorem (CLT)

Statement:

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$.
- Then, as $n \to \infty$:

$$\sqrt{n}\left(\bar{X} - \mu\right) \xrightarrow{d} N(0, \sigma^2)$$

- Equivalently:

$$\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Implications:

- The sampling distribution of $\bar{X}$ becomes approximately normal, regardless of the original distribution.
- Justifies normal-based inference (confidence intervals, hypothesis tests) for large n.

Note:

- Convergence can be slower for heavy-tailed or skewed distributions.

Important Sampling Distributions

When population variance σ^2 is known (or n large):

- **Sample mean:**

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- **Sample proportion:**

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

(valid for large n by the Central Limit Theorem)

- **Difference of two sample means:**

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Important Sampling Distributions II

When population variance σ^2 is unknown:

- **Student's t-distribution:**

  If σ is unknown and replaced by the sample standard deviation s, then

  $$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

  where t_{n-1} is a t-distribution with $n-1$ degrees of freedom.

For variance-related statistics:

- **Chi-square distribution:**

  $$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

  (used in confidence intervals and tests for variance)

- **F-distribution:**

  $$\frac{\text{Variance}_1/\sigma_1^2}{\text{Variance}_2/\sigma_2^2} \sim F_{d_1,d_2}$$

  (used for comparing two variances)

Confidence Intervals: Fundamentals

General form:

$$\text{Point estimate} \pm (\text{Critical value} \times \text{Standard error})$$

Correct Interpretation:

- If we repeated the sampling procedure infinitely many times, approximately 95% of the constructed confidence intervals would contain the true population parameter.

Precision vs Confidence:

- **Wider intervals:** Higher confidence, but lower precision.
- **Narrower intervals:** Higher precision, but lower confidence.
- Increasing the sample size n leads to narrower intervals without sacrificing confidence.

Common Misinterpretation (WRONG):

- It is **incorrect** to say that there is a 95% probability that the true parameter lies inside a *realized* interval.
- Once the interval is calculated, the parameter is either inside it or not.

Common Confidence Intervals

Population Mean (Known Variance σ^2):

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

(when σ known, $z_{\alpha/2}$ from standard normal distribution)

Population Mean (Unknown Variance):

$$\bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$$

(when σ unknown, s sample standard deviation, t-distribution with $n-1$ degrees of freedom)

Population Proportion:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(for large n, by Central Limit Theorem)

Confidence Intervals for Regression Coefficients

Confidence interval for each regression coefficient β_j:

$$\hat{\beta}_j \pm t_{\alpha/2,\, n-p-1} \times SE(\hat{\beta}_j)$$

where:

- $SE(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \, [(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$
- $\hat{\sigma}^2 = \dfrac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ (residual variance estimate)
- Degrees of freedom: $n - p - 1$ (n: observations, p: predictors)

Notes:

- t-distribution used because σ^2 is unknown.
- $\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{jj}$ is the j-th diagonal element of the inverse Gram matrix, capturing the variability of $\hat{\beta}_j$.

Hypothesis Testing Framework

Key Elements:

- **Null Hypothesis (H_0):**
  Statement of no effect, no difference, or status quo. (Assumed true at the start.)
- **Alternative Hypothesis (H_1 or H_A):** Statement representing an effect, difference, or deviation from H_0.
- **Test Statistic:** Quantity computed from sample data, with a known distribution under H_0.
- ***p*-value:** Probability, assuming H_0 is true, of observing a test statistic as extreme or more extreme than the one observed.
- **Significance Level (α):** Pre-specified threshold (commonly 0.05) for deciding whether to reject H_0.

Decision Rule:

- Reject H_0 if p-value $\leq \alpha$.
- Fail to reject H_0 if p-value $> \alpha$.

Types of Errors and Power

	H_0 **True**	H_0 **False**
Reject H_0	Type I error (α)	**Correct decision** (Power)
Fail to reject H_0	**Correct decision**	Type II error (β)

Key Concepts:

- **Type I error** (α): Rejecting H_0 when it is actually true (false positive).
- **Type II error** (β): Failing to reject H_0 when it is actually false (false negative).
- **Power** ($1 - \beta$): Probability of correctly rejecting a false H_0.

Factors that Increase Power:

- Larger sample size (n).
- Larger effect size (true difference from H_0).
- Higher significance level (α).
- Lower variability (smaller σ^2).

Common Hypothesis Tests

One-Sample t-Test:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad \sim \quad t_{n-1}$$

Testing $H_0 : \mu = \mu_0$. Requires approximate normality of the population.

Two-Sample t-Test (Independent Samples):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \sim \quad t_{df}$$

Testing $H_0 : \mu_1 = \mu_2$. Assumes independence between samples. Equal variance assumption may or may not be made (Welch's test if variances unequal).

Common Hypothesis Tests II

Paired t-Test:

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \quad \sim \quad t_{n-1}$$

Testing $H_0 : \mu_d = 0$ for the mean difference between paired observations. Assumes differences are approximately normally distributed.

F-Test (One-Way ANOVA):

$$F = \frac{MS_{between}}{MS_{within}} \quad \sim \quad F_{k-1,n-k}$$

Testing H_0: All group means are equal.

- $k =$ number of groups
- $MS_{between} =$ mean square between groups
- $MS_{within} =$ mean square within groups (residuals)

Hypothesis Tests in Regression: t-test

t-Test for Individual Coefficients:

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{SE(\hat{\beta}_j)} \quad \sim \quad t_{n-p-1}$$

Testing:

$$H_0 : \beta_j = \beta_{j0} \quad \text{vs.} \quad H_1 : \beta_j \neq \beta_{j0}$$

(Typically, $\beta_{j0} = 0$ to test if a predictor is useful.)

Hypothesis Tests in Regression: F-test

F-Test for Overall Model Significance:

$$F = \frac{MSR}{MSE} = \frac{\text{Regression Sum of Squares}/p}{\text{Residual Sum of Squares}/(n-p-1)}$$

Testing:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{At least one } \beta_j \neq 0$$

- MSR: Mean Square Regression
- MSE: Mean Square Error (Residual)
- p: Number of predictors

Partial F-Test:

- Compares two nested models (a simpler model inside a more complex one).
- Tests if a subset of predictors improves the model significantly.

The Multiple Testing Problem

Key Issue:

- When performing multiple hypothesis tests (m tests), the probability of making at least one Type I error increases.

Definitions:

- **Family-Wise Error Rate (FWER):** Probability of making at least one Type I error across all tests.
- **False Discovery Rate (FDR):** Expected proportion of Type I errors among all rejections.

Without Correction (Independent Tests):

$$\text{FWER} = 1 - (1 - \alpha)^m, \quad \text{where:}$$

- $\alpha =$ significance level for each individual test
- $m =$ number of independent tests

The Multiple Testing Problem II

Example:

- $m = 20$ tests, $\alpha = 0.05$ per test
- FWER $\approx 1 - (1 - 0.05)^{20} \approx 0.64$

Solutions (briefly):

- **FWER control**: Bonferroni correction, Holm's procedure (very conservative).
- **FDR control**: Benjamini-Hochberg procedure (more powerful, used in large-scale testing).

Multiple Testing Correction Methods

Bonferroni Correction (Controls FWER):

$$\alpha_{\text{corrected}} = \frac{\alpha}{m}$$

- Test each hypothesis at level α/m.
- Very simple, but conservative (especially when m is large).

Holm-Bonferroni Method (Controls FWER, Less Conservative):

1. Order p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. For each i, compare $p_{(i)}$ to $\dfrac{\alpha}{m-i+1}$:
   - If $p_{(i)}$ is significant, reject $H_{0(i)}$ and continue.
   - Stop at the first non-significant p-value.

- Less conservative than Bonferroni while still controlling FWER.

Multiple Testing Correction Methods II

Benjamini-Hochberg Procedure (Controls FDR):

1. Order p-values: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
2. Find the largest k such that:

$$p_{(k)} \leq \frac{k}{m}\alpha$$

3. Reject all $H_{0(i)}$ for $i = 1, 2, \ldots, k$.

- Controls the expected proportion of false discoveries among all rejections (FDR).
- Widely used in large-scale testing (e.g., genomics, machine learning).

Practical Considerations in Multiple Testing

Key Trade-Off:

- Stricter corrections (e.g., Bonferroni) reduce false positives (Type I errors) but increase false negatives (Type II errors).
- Balancing error types depends on research priorities.

Independence Assumption:

- Many multiple testing corrections assume independent or weakly dependent tests.
- Violations (strong correlations) may affect FWER/FDR control.

Context Matters:

- **FWER control**: Prefer when false positives are very costly (e.g., clinical trials, regulatory decisions).
- **FDR control**: Prefer when discovery is more important than strict certainty (e.g., genomics, exploratory research).

Reporting Standards:

- Always report:
  - The correction method used (e.g., Bonferroni, BH procedure).
  - The type of error rate controlled (FWER or FDR).

Understanding Heteroskedasticity

Definitions:

- **Homoskedasticity:**

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \text{for all } i$$

  The variance of the errors is constant across observations.

- **Heteroskedasticity:**

$$\text{Var}(\varepsilon_i) = \sigma_i^2 \quad \text{varies with } i$$

  The variance of the errors differs across observations.

Consequences of Heteroskedasticity:

- OLS estimators remain **unbiased** but are **no longer efficient**.
- Estimated standard errors are biased $\rightarrow$ Hypothesis tests and confidence intervals are invalid.
- Predictions have suboptimal (non-minimal) variance.
- Violates a Gauss-Markov assumption: OLS is no longer the **Best Linear Unbiased Estimator (BLUE)**.

Detecting Heteroskedasticity

Visual Methods:

- **Residual vs. Fitted Values Plot:** Look for patterns — a funnel shape suggests heteroskedasticity.
- **Residual vs. Predictor Plots:** Examine key predictors individually for changing spread.
- **Scale-Location Plot:** Plot $\sqrt{|e_i|}$ versus $\hat{y}_i$ to better detect non-constant variance.

Statistical Tests:

- **Breusch-Pagan Test:**
  - Regress squared residuals on predictors.
  - H_0: Homoskedasticity (constant variance).
- **White Test:**
  - General test allowing for nonlinear forms of heteroskedasticity.
  - H_0: Homoskedasticity.
- **Goldfeld-Quandt Test:**
  - Test for monotonic changes in variance across ordered data.
  - Compare variance across different subsamples.

Addressing Heteroskedasticity

Transformation Approaches:

- Apply a transformation to the response variable to stabilize variance (e.g., log, square root).
- **Box-Cox Transformation:**

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases}$$

where λ is estimated from the data.

Weighted Least Squares (WLS):

- Minimize a weighted sum of squared residuals:

$$\sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2$$

- Use weights $w_i = 1/\hat{\sigma}_i^2$, giving more importance to observations with smaller variance.

Addressing Heteroskedasticity II

Robust Standard Errors:

- Use heteroskedasticity-consistent (HC) standard errors (e.g., White's estimator).
- Coefficient estimates stay the same; only the standard errors are adjusted for valid inference.

Summary:

- If you care about improving model fit $\rightarrow$ use transformations or WLS.
- If you mainly care about valid inference $\rightarrow$ use robust standard errors.

Bootstrapping: Fundamentals

Basic Principle:

- Approximate the sampling distribution of a statistic by resampling from the observed data.
- Does not require assumptions about the underlying population distribution.

Bootstrap Process:

- Draw a random sample **with replacement** of size n from the observed dataset.
- Compute the statistic of interest (e.g., mean, median, regression coefficient) from the resample.
- Repeat the process a large number of times (e.g., 1000 or more).
- Use the distribution of bootstrap replicates to estimate:
  - Standard error
  - Confidence intervals
  - Bias (if needed)

Bootstrapping: Advantages

Advantages of Bootstrapping:

- **Distribution-free**: No assumptions about the form of the population.
- Handles complex statistics for which theoretical formulas are difficult or unknown.
- Easy to implement with modern computing.

Types of Bootstrap Methods

Nonparametric Bootstrap:

- Resample directly with replacement from the observed data.
- No assumptions about the underlying population.

Block Bootstrap:

- Used when data are dependent (e.g., time series).
- Resample blocks of consecutive observations to preserve local dependence structure.

Residual Bootstrap (for Regression Models):

- Assumes the model form is correct but allows for random errors.
- Process:
  1. Fit the regression model and calculate residuals.
  2. Resample residuals and generate new response values: $\hat{y}_i + e_i^*$.
  3. Refit the model using these new responses.

Bootstrap Confidence Intervals

Idea:

- Use the empirical distribution of bootstrap estimates to construct confidence intervals.
- No assumptions about normality or standard errors needed.

Percentile Method:

- Take the $\alpha/2$ and $1-\alpha/2$ quantiles from the bootstrap distribution.
- Confidence interval:

$$\left[\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}\right]$$

Basic Bootstrap Method:

- Symmetric adjustment around the original estimate $\hat{\theta}$.
- Confidence interval:

$$\left[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)},\ 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\right]$$