STAT 5200 Handout #6.a Checking Assumptions (Ch. 6)

Before trusting any inference (significance tests, confidence intervals), we must first verify that model assumptions are satisfied. Recall effects model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$i = 1, \ldots, g$ factor levels
$j = 1, \ldots, n$ exp. unit
$N = g \cdot n$ ← overall sample size

increasing importance ↑

Three key assumptions (that allow inference)

I. ϵ_{ij}'s are independent
II. ϵ_{ij}'s have constant variance (σ^2)
III. ϵ_{ij}'s are normally distributed

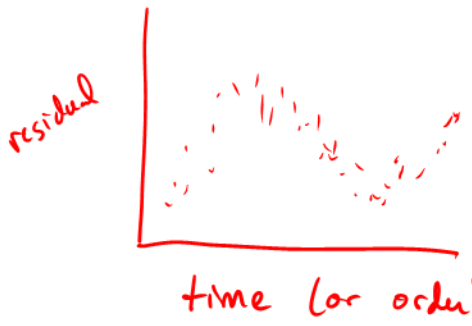
Shorthand: ϵ_{ij}'s are i.i.d. $N(0, \sigma^2)$ → independent and identically distributed

(*) We can assess these using the residuals, relying on graphical checks. (don't over-emphasize numerical checks)

residuals → $e_{ij} = \hat{\epsilon}_{ij} = y_{ij} - \hat{y}_{ij}$ = observed minus predicted

I. Independence of error terms

- critical (but not focus of this class)
- sometimes violated when data are collected sequentially or spatially
- sequential dependence: check with time series model (STAT 5100)
  - book introduces Durbin-Watson test
  - can plot residuals vs. time to look for any rough trend → likely have some dependence



- spatial dependence: check with spatial (or spatio-temporal) model (STAT 5410)
  - book introduces variogram

- If have multiple obs. per subject, like at time points → Repeated Measures design w/ dependence within subject (more later)

II. Constant error variance

- our significance tests rely on estimate of σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{ij}\left(Y_{ij} - \hat{Y}_{ij}\right)^2}{N-g} = MSE$$

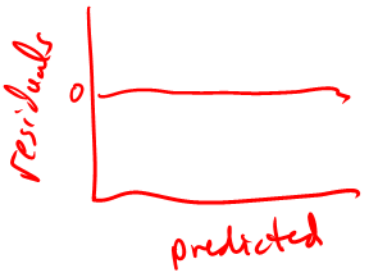
→ denominator in F-statistic

Note that when $g = 2$, this is the pooled variance estimate in the t-test:

$$s_p^2 = \frac{\sum_{j=1}^{n_1}\left(Y_{1j} - \bar{Y}_{1\cdot}\right)^2 + \sum_{j=1}^{n_2}\left(Y_{2j} - \bar{Y}_{2\cdot}\right)^2}{n_1 + n_2 - 2} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

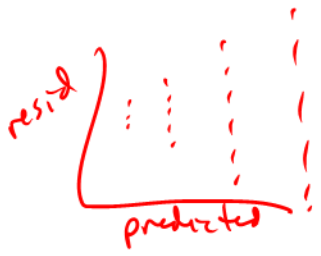
$\left(\hat{Y}_{ij} = \bar{Y}_{i\cdot}\right)$

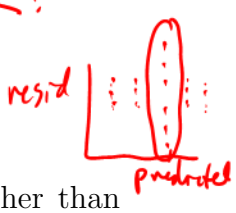
- It can sometimes happen that σ^2 depends on i
  - So variance is larger or smaller for some factor levels i)
    - this is called heteroscedasticity
  - Then the estimate $\hat{\sigma}^2 = MSE$ is wrong
- Assess by plotting residuals vs. predicted:



If no trend, assumption is okay.

A common trend is megaphone





- Plotting residuals vs. factor levels can reveal which factor levels are "problematic"
- A "test" for constant variance is most appropriate when changes in variance (rather than means) are of primary interest
  - Example: a quality control situation where you want a supplier who provides the lowest variance for some characteristic
  - Modified Levene test (HOVtest option in PROC GLM):

Let σ_i^2 be variance for ε_{ij}'s in factor level i

$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_g^2$

Fit ANOVA F-test to "response" $d_{ij} = |Y_{ij} - \tilde{Y}_{i\cdot}|$ → MAD = median absolute deviation

$$d_{ij} = \mu^{(d)} + \alpha_i^{(d)} + \varepsilon_{ij}^{(d)}, \quad \varepsilon_{ij}^{(d)} \text{ iid } N(0, \sigma^2_{(d)})$$

$\tilde{Y}_{i\cdot}$ = sample median for factor level i

III. Normality of error terms

central limit theorem

- Critical for small N, but unnecessary for large N (due to CLT). How large does N need to be?
  – see Lumley *et al.* 2002, *Annual Review of Public Health* "The Importance of the Normality Assumption in Large Public Health Data Sets"
  - depends on degree of non-normality
  - safest bet if $N \geq 500$
- Recall normal probability plot to assess [on residuals]; numerical checks are available (but with much lesser emphasis)
- An extreme form of non-normality: outliers (glaringly different Y values)
  Formal tests exist for detection (STAT 5100), but here we'll rely on visual checks
  – be extremely reluctant to throw out valid data
  – rough visual check: RStudent values for outside ref. lines

What to do if assumptions are violated?

I. independence → identify & account for dependence structure (example later on HO # 23, repeated measures design)

II. constant variance

III. normality

} → first consider transforming y

What transformations to consider? Usually something from the "power" family:

$$\ldots, -y^{-2}, -y^{-1.5}, -y^{-1}, -y^{-.5}, \log y, y^{.5}, y^{1}, y^{1.5}, y^{2}, \ldots$$

A useful transformation-finding tool: Box-Cox

- Let $\dot{Y}$ be geometric mean:

$$\dot{Y} = \left(\prod_{ij} Y_{ij} \right)^{\frac{1}{N}}$$

- Define

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda \dot{Y}^{\lambda-1}} & \text{for } \lambda \neq 0 \\ \dot{Y} \log Y & \text{for } \lambda = 0 \end{cases}$$

In practice

$\rightarrow \text{sign}(\lambda) \cdot y^{\lambda}$

$\rightarrow \log y$

- Fit ANOVA model for several λ values, and choose λ that minimizes SS_E

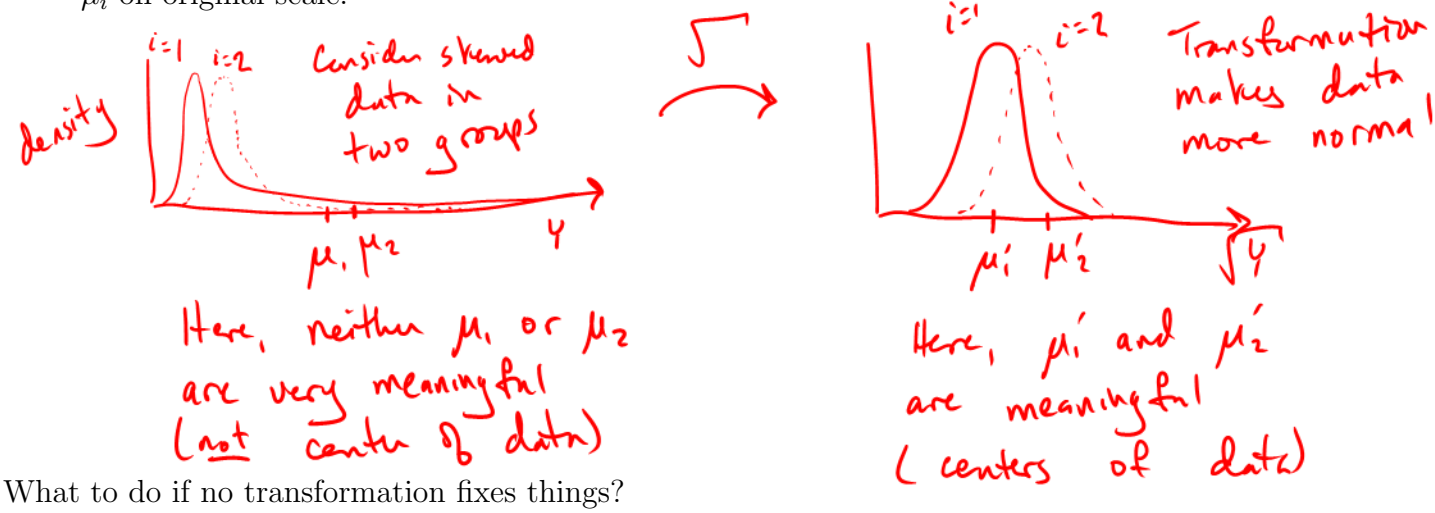
  – will also make residuals most normal and variance most constant

- Implemented in PROC TRANSREG
  → look for "convenient" and apply that
  inside shaded CI
  λ (*** interpretability ***)
  power transformation

After making transformation, re-fit model and re-check assumptions

Why isn't making a transformation "cheating"?

- Just re-scale data
- Parameters μ_i' on new scale should more accurately represent center of treatment effects than μ_i on original scale:



"prime" = not derivative, just a new parameter on a new scale

What to do if no transformation fixes things?

- Balanced designs help mitigate (but do not fully resolve) effects of assumption violations
- For non-constant variance, consider weighted least squares or Brown-Forsythe
- For non-normality, ANOVA F-test is fairly robust, but could also go to a non-parametric test (Kruskal-Wallis)
  - Recall two factor levels in Wilcoxon Rank-Sum
  - Kruskal-Wallis generalizes rank-based test to g factor levels
  - See documentation and examples in PROC NPAR1WAY
- If distribution is non-normal for a reason (counts, for example), can fit a generalized linear mixed model (PROC GLIMMIX) – will return to this at end of semester if time permits

NOTE: If your assessment of normality or constant variance hinges on a single point (cover it up and your assessment changes, e.g.), then you probably don't have a gross or systematic violation of model assumptions.