Recap: Violations of Assumptions

1) Bias
2) Non-Normality
   - Outliers

$y \sim N(Z\beta, \sigma^2 V)$

3) Now: Heteroscedusticity
   (non constant Variance)

$V \neq I$

$\boxed{\text{Heteroscadasticity}}$

Suppose $y \sim N(Z\beta, \sigma^2 V)$
V is full rank, not the identity

- Example 1: $V = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, as is the case where different amounts of data goes into each measurement

- Example: AR(1):

$$\begin{pmatrix} 1 & \rho & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} p & 1 & p \\ 0 & p & 1 \end{pmatrix} \qquad n=3$$

- If we know V, we can use generalized LS:

$$V = P^T \Lambda P \qquad P^T P = P P^T = I$$

$$\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

let: $D := \Lambda^{-\frac{1}{2}} P$

- we have: $\underbrace{Dy}_{\tilde{y}} = D(Z\beta + \varepsilon) = \underbrace{DZ}_{\tilde{Z}}\beta + \underbrace{D\varepsilon}_{\tilde{\varepsilon}}$

$$\boxed{\tilde{y} = \tilde{Z}\beta + \tilde{\varepsilon}}$$

Where: $\operatorname{Var}(\tilde{\varepsilon}) = D \underbrace{\operatorname{var}(\varepsilon)}_{V} D^T$

$$= D P^T \Lambda P D$$

$$= \Lambda^{-\frac{1}{2}} \underbrace{P P^T}_{I} \Lambda \underbrace{P P^T}_{I} \Lambda^{-\frac{1}{2}} = I$$

- we take:

$$\hat{\beta}_{GLS} = (\tilde{Z}^T \tilde{Z})^{-1} \tilde{Z}^T \tilde{y}$$

so that $\hat{\beta}_{GLS}$ minimizes

$$\| Dy - DZ\beta \|^2$$

- If $V = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2)$

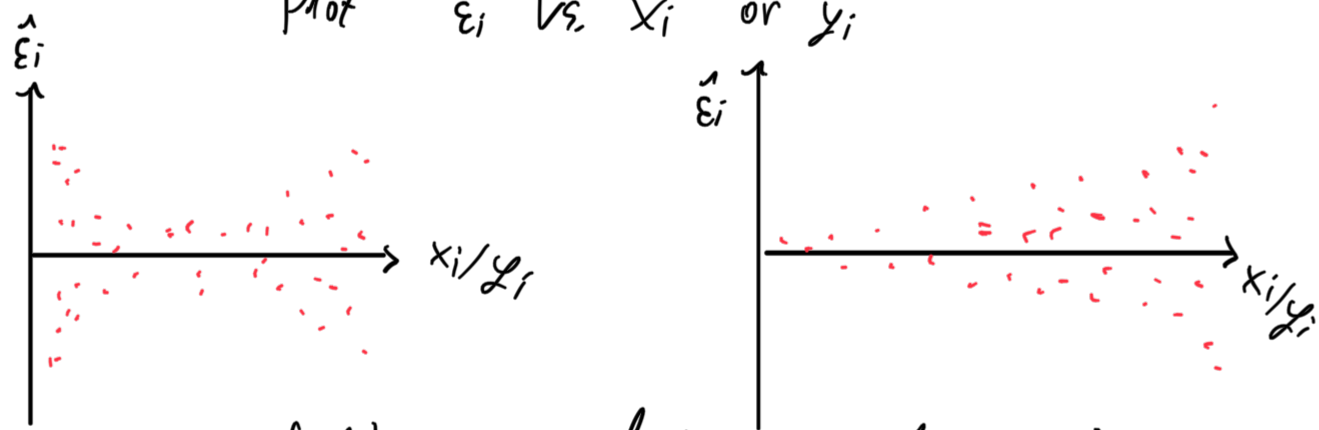
then we can use

$$\hat{\sigma}_i^2 = \hat{u}_i \quad , \quad \hat{u}_i = y_i - z_i^T \hat{\beta}_{OLS}$$

Huber-White residuals $\hat{\beta}_{OLS} = (Z'Z)Z'y$

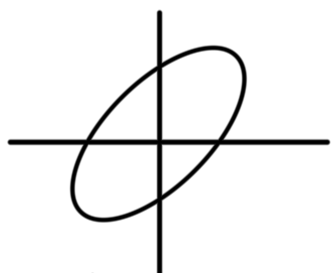
Detection

plot $\hat{\varepsilon}_i$ vs. x_i or y_i



(things we don't want to see)

- Another approach is to plot $\hat{\varepsilon}_i$ vs $\hat{\varepsilon}_{i+1}$ a pattern like



indicates dependence in errors (like in AR(1) say)

- This issue needs to be addressed using time-series methods

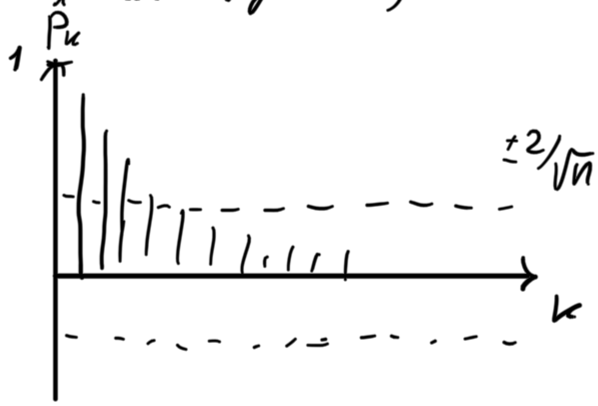
- While $\hat{\beta}$ is still unbiased, we get erroneous variance so we cannot do proper statistical tests.

- We can also compute the correlation between the residuals:

$$\hat{\rho}_k = \frac{1}{n}\sum_{i=k+1}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-k}$$

$$\frac{\cdots}{\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

(autocorrelation at lag k)



- under the null hypothesis when there is no correlation between the residuals, $\rho_k = 0$, $k \geq 1$

- Tests for non-zero correlation: Durbin-Watson & Ljung-Box ...

- See class on time-series analysis Stanford STATS 207

$\boxed{\text{Course Summary}}$

Additional Topics

- ANOVA with random effects:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad , \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

$$i = 1, \dots, k \qquad \alpha_i \overset{iid}{\sim} N(0, \sigma_A^2)$$

$$j = 1, \dots, n_i$$

- Two-Factor ANOVA:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \quad , k=1,\dots,n_{ij} \; , \; \begin{matrix} i=1,\dots,I \\ j=1,\dots,J \end{matrix}$$

$$\mu_{ij} = \alpha_i + \beta_j + \gamma_{ij} \quad 0 = \sum_i \alpha_i = \sum \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij}$$

Also: Fixed × Fixed , Fixed × Random
Random × Random

- Causality
- Bootstrapped Regression:
  - Pairs: resample n pairs from $(x_i, y_i)_{i=1}^n$ with repetition
  - Residuals: resample n times from $\{\hat{\varepsilon}_i\}_{i=1}^n$
    use $y_i^* = z_i^T \hat{\beta} + \hat{\varepsilon}_i^*$

Course Overview

- Applied Stats & Linear Model
- We have (x_i, y_i) pairs & we want to model how y depends on x
- For example: we want to predict y_{n+1} from x_{n+1}
- Also: how things generalizes.

What can go wrong?

$\boxed{\text{The Pipeline}}$

(1) Data. $\{(x_i, y_i)\}_{i=1}^{n}$, $y_i \in \mathbb{R}$. But x_i can be 1 group, k groups, kxr groups, or $\mathbb{R}^d$

(2) Features. Z, e.g. $z_i^T = (1, x_{i1}, \dots, x_{id})$

- Can also add-in non-linear features like x_{i3}^2 or $1(x_{i2} \leq 7)$

- The inclusion/exclusion of features can be based on intuition, experience, science, or machine learning.

(3) Model. $y \sim N(Z\beta, \sigma^2 I)$

(4) Fitting. Estimate $\hat{\beta}$ and $\hat{\sigma}$ (based on linear algebra)

(5) Inference. Derive confidence intervals, p-values, hypothesis testing, power calculations (based on normal dist. theory)

(6) Interpretation. Association between predictors and respose. Concerns about causality and the meaning of a "true" β_j, because it depends on whether the k-th predictor is in the model or not.

(7) Model Selection. Using methods like AIC, BIC, CV, lasso, ridge regression, optimized for prediction, but not necessarily for understanding of causal mechanisims.

(8) Problems & Fixes

- Non-normality: CLT fixes most cases
- Non-constant variance: Bootsrapping or Auber-White residuals in GLS
- Bias. No good solutions, context and experience on topic are useful. Lack of fit sum of squares may help identify bias but can do little to solve it.
- Outliers: some methods exist, but these are tough to overcome in high dimensions. If both y_i & z_i are outliers, it is a good indication that this sample should be removed.
- Correlations within residuals. Time series methods

$\boxed{\text{Followups}}$

- Generalized Linear Models (GLM):

$$E[y_i|z_i] = \varphi(z_i^T\beta) \qquad \varphi: \mathbb{R} \to \mathbb{R}$$

- Look at $y_i \in \{0,1\}$, $y_i \in \{0,1,2,\dots,k\}$

$$y_i \in \mathbb{R}^p$$

- More computationaly sophisticated algorithms than LS:

$$E[y_i|z_i] = \varphi_\beta(z_i^T\beta)$$

- Causal conclusions from observations

- Time series analysis (dependency between (x_i, y_i) over i)

$\boxed{\text{GLM}}$ $\varphi : \mathbb{R} \to \mathbb{R}$

$$E[y_i | z_i] = \varphi(z_i^T \beta)$$

φ is called the "link function"

Examples:

$\varphi(x) = x$

$\varphi(x) = \frac{1}{1+e^{-x}}$ logit model

$\varphi(x) = \Phi(x)$ probit model

$\varphi(x) = e^{x}$ Poisson Reression

Logit Regression: $\varphi(x) = \frac{1}{1+e^{-x}}$

- We have $y_i \in \{0,1\}$

- We continuous predictors $x_1, \ldots, x_n$ $x_i \in \mathbb{R}^p$

- We propose that the log-likelihood ratio between the calsses is of the form:

$$\left(\log\left(\frac{\overbrace{Pr(Y_i = 1 | x_i)}^{p}}{\quad} \right) = x_i^T \beta \leftarrow \text{logit}\right.$$

$\quad \ell\left(\underbrace{Pr(Y_i=0 \mid x_i)}_{1-p}\right) \quad$ space

So that $p = p(\beta, x_i) = E[Y_i \mid x_i] = Pr(Y_i = 1 \mid x_i)$

$$= \frac{1}{1+e^{-x_i^T\beta}}$$

- The log-likelihood function:

$$\ell(\beta; \{y_i, x_i\}) = \sum_{i: y_i=1} \log(p(\beta; x_i)) + \sum_{i: y_i=0} \log(1-p(\beta; x_i))$$

$-\ell(\beta; \{y_i, x_i\})$ is convex, has a unique minimum