① Classical linear model $Y = X\beta + \varepsilon$ (X full rank)

- **Key Assumptions**:
  1. Linearity of cov. effects → transform
  2. Additivity of errors → can apply kg. transformation for multiplic. ε
  3. Error mean zero expectation: $E[\varepsilon] = 0$
  4. Homoscedastic error var: $Var[\varepsilon_i] = \sigma^2 \ \forall \varepsilon_i$
  5. Uncorrelated errors: $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$
  6. Gaussian errors: $\varepsilon \sim N(0, \sigma^2 I)$

- **Estimation**: $\min RSS \ \ \varepsilon^T\varepsilon$ (A)
  - $\|\varepsilon\|_2 = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta$ (same equations via MLE)
  - Rules for diff: $\frac{\partial}{\partial\beta}\beta^TA\beta = 2A\beta$ (A)

- **Pred. values and Residuals**: $\hat{Y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{H}Y = HY$

- **Successive Orthogonalization** (Gram-Schmidt Orthogonalization)
  1. Initialize $z^0 = x^0 = 1$
  2. Regress x_j on $z^0, z_1, \dots$ the prev. ie (for $j = 1..k$) to produce
  $$\hat{\gamma}_{ej} = \frac{z^{e^T} \cdot x_j}{z^{eT} z_j}, \quad e = 0 \dots, j-1 \text{ and } z_j = x_j - \sum_{e=0}^{j-1} \hat{\gamma}_{ej} z^e$$
  3. Regress y on residuals z^k to give $\hat{\beta}_k = \frac{z^{kT} \cdot y}{z^{kT} z^k}$

→ QR decomp: Write output of step 2 as
$$X = Z D^{-1} D \Gamma = QR \rightarrow \hat{\beta} = R^{-1}Q^TY$$

- **Error Variance**: $\hat{\sigma}^2 = \frac{1}{n-p}\hat{\varepsilon}^T\varepsilon \rightarrow E[\hat{\sigma}^2] = \sigma^2$

- **Distribution of Parameter** (without distr. / normality assumptions)
  - $E[\hat{\beta}] = \beta$ → still unbiased
  - $Cov[\hat{\beta}] = \sigma^2(X^TX)^{-1}$ → $\widehat{Cov}(\hat{\beta}) = \hat{\sigma}^2(X^TX)^{-1}$ with $\hat{\sigma}^2 = \frac{1}{n-p}\hat{\varepsilon}^T\hat{\varepsilon}$
  - $Var[\hat{\beta}_j] = \sigma^2 / (z^{jT} z^j)$ → high corr → z^j ↓ → Var ↑

- **Gauss-Markov**: LSE is BLUE (without normality, others still hold)
  $$Var(\hat{\beta}_i) \leq Var(\beta_j^{Lin}) \rightarrow \text{smallest var of all lin. unbiased}$$

- **MLE**: Add ass. + normality of error
  → LSE has min. var among all unbiased (not only linear) estimators
  → attains the Cramér Rao lower bound
  → $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

- **Asymptotics**: Consistency and asympt. normality of LSE → 2 assumpt.
  1. $\lim_{n\to\infty}(X_n^TX_n)^{-1} = \lim_{n\to\infty}(\sum x_i x_i^T)^{-1} = 0$ "covariate info ↑ as n↑"
  2. $\lim_{n\to\infty} \max_{i=1,\dots,n} x_i^T(X_n^TX_n)^{-1}x_i = 0$ "influence of each x_i as negligible relative to the entire info $x_i^TX_n$"

- **Properties of residuals**
  - $\hat{\varepsilon} = Y - \hat{Y} = (I-H)Y$ → residuals are orthogonal to X: $X^T\hat{\varepsilon} = 0$
  - $E[\hat{\varepsilon}] = 0$, $Cov[\hat{\varepsilon}] = \sigma^2(I-H)$, $Var[\hat{\varepsilon}_i] = \sigma^2(1-h_{ii})$
  - $\hat{Y}^T\hat{\varepsilon} = 0$ (residuals and pred. values are orthogonal + independent)

---

② 1. Hypothesis Testing

- **General linear hypothesis**: $C\beta = d$ vs. $C\beta \neq d$
  ($r \times p$ matrix, $rk(C) = r \leq p$ (r = # constraints))

- **F-stat**: $\frac{n-p}{r}\frac{\Delta SSE}{SSE} \sim F_{r,n-p}$ (derivation based on the normality of errors)

↳ **Wald Test**: $F = [(C\hat{\beta}-d)^T \widehat{Cov}(C\hat{\beta})^{-1}(C\hat{\beta}-d)]/r \rightarrow W = rF$

ⓐ **Test of sign. (t-test)**: $t_j = \frac{\hat{\beta}_j}{se_j} \sim t_{n-p}$ as $F = \frac{\hat{\beta}_j^2}{\widehat{Var}(\hat{\beta}_j)} \sim F_{1,n-p}$

ⓑ **Complete test of subvector (Partial F-test)**: $\beta_2 = (\beta_{s_1}, \dots \beta_r)^T$

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

$\rightarrow F = \frac{1}{r}\hat{\beta}_1^T \widehat{Cov}(\hat{\beta}_1)^{-1}\hat{\beta}_1 \sim F_{r,n-p}$ where $\widehat{Cov}(\hat{\beta}_1)$ could be corr. elements of $\hat{\sigma}^2(X^TX)^{-1}$

ⓒ **Test for sign. of regression (global F-test)**

$H_0: \beta_i = 0 \ \forall i = 1..k$, H_1: at least one > 0 → $\frac{n-p}{k}\frac{R^2}{1-R^2} \sim F_{k,n-p}$

2. Confidence and Prediction Intervals

ⓐ CI based on t_j: $P(\hat{\beta}_j - t_{n-p(1-\frac{\alpha}{2})} \cdot se_j < \beta_j < \hat{\beta}_j + t_{n-p,1-\frac{\alpha}{2}} se_j] = \alpha$ → can extract $\hat{\beta}_j$

ⓑ For multiple parameters (subvector β_1) → Confidence Ellipsoid

$CE: [\beta_1 : \frac{1}{r}(\hat{\beta}_1 - \beta_1)^T \widehat{Cov}(\hat{\beta}_1)^{-1}(\hat{\beta}_1 - \beta_1) \leq F_{r,n-p}(1-\alpha)]$

- **CI for $\mu_0 = E[y_0]$** (for future obs. y_0 at location x_0)

$[x_0^T\hat{\beta} \pm t_{n-p}(1-\frac{\alpha}{2})\hat{\sigma}(x_0^T(X^TX)^{-1}x_0)^{1/2}]$

- **Prediction Intervals** (for prediction $\hat{y}_0 = x_0^T\hat{\beta}$ for new/future obs. x_0)

$[x_0^T\hat{\beta} \pm t_{n-p}(1-\frac{\alpha}{2})\hat{\sigma}(1 + x_0^T(X^TX)^{-1}x_0)^{1/2}]$ contains future obs. y_0 with Pr $1-\alpha$

3. Multiple Testing

- **P value**: $= P_{H_0}(|T| \geq |t_{obs}|) \sim U(0,1)$ (under H_0)

- **Errors**:

Decision \ Reality	H_0	H_1
H_0	Type I = α	correct
H_1	correct	Type II = β

- discovery = rejecting H_0
- false discovery: type I error
- power: $1-$ Type II error $(1-\beta)$

→ as n↑ or effect size↑ → β↓ → $(1-\beta)$↑ but α same

- **FWER**: $P(V \geq 1) = 1-(1-\alpha)^m$ (m = total # of hypotheses)
  - for independent test = α
  - for pos. dependent tests: FWER $< \alpha$ (cons.); for neg. dependence → FWER $> \alpha$ (lib.)

- **Bonferroni**: reject $H_0^{(i)}$ when $p_i \leq \frac{\alpha}{m}$; can also be used to construct CI
  ↳ $\alpha^* = \alpha/m$, OR $p^* = p_i \cdot m$

- **Bonferroni-Holms (BH)**: sequentially reject when $P_i \leq \frac{\alpha}{m-i+1}$
  ↳ also needs no assumption on dependence structure of $H^{(i)}$ → more powerful than BF

- **Permutation test**: permute test statistic $T(S_x, S_y) \rightarrow P_H(|T| \geq T(S_x, S_y))$
  → calc p-values after permutation: $p\text{-value} = \frac{\sum_{i=1}^{B} \mathbb{1}(|T^{*(i)}| \geq |T(S_x, S_y)|) + 1}{B+1}$ (B # permutations)
  ⊕ non-parametric approach that works with finite sample
  ⊖ comp. expensive + only a global test

- **FDR**: weaker error control than FWER (FWER also controls FDR)

$FDR = E[FDP] = E[\underbrace{V/R}_{FDP}]$ (exp. false disc. / all disc.)

- **BH$_\alpha$**:
  - Order p-values (same as for BH)
  - R be the largest k s.t. $P_k \leq \frac{\alpha \cdot k}{m}$ → R will be #
  - Reject all null hypotheses for which $P_j \leq P_{k/R}$

↳ only works well for independent and pos. dependent (PRDS) p-values
↳ under arbitrary dependence: FDR $\approx \alpha \cdot \log(m)$ and FDP high variance!

4. Model Specification

Truth \ Fitted	Reduced model ($\hat{\beta}_1$)	Full model ($\hat{\beta}$)
Reduced model ($\tilde{\beta}_1$)	$E[\tilde{\beta}_1] = \beta_1$ (both unbiased) $Cov(\tilde{\beta}_1) = \sigma^2(X_1^TX_1)^{-1}$	① Irrelevant Variables - $\hat{\beta}_1$ unbiased - $Cov(\hat{\beta}_1) - Cov(\tilde{\beta}_1) \geq 0$
Full model ($\tilde{\beta}_1$)	② Missing variables - $\tilde{\beta}_1$ is biased - $Cov(\tilde{\beta}_1) - Cov(\hat{\beta}_1) \geq 0$ ↳ smaller Var / underestimate	$E[\hat{\beta}] = \beta$ $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

- for Var (⊜ if X_1, X_2 (partitions) are orthogonal
- Precision decreases (Var ↑) if model contains irrelevant variables

③ Prediction Quality: → Consider SPSE for assessment

$$SPSE = \sum_{i=1}^{n}(y_{new,i} - \hat{y}_{i,in})^2 \quad \hat{y}_{i,in} \text{ is pred. based on a model}$$
$$= \underbrace{n \cdot \sigma^2}_{\text{I}} + \underbrace{|M|(\sigma^2 + \sum_{i=1}^{n} E[(\hat{y}_{i,in} - \mu_i)^2])}_{\text{II + III} = MSE}$$

I) Irreducible prediction error → just depends on σ^2
II) Variance: depends on $|M| = \#$ variables included in model
III) Squared bias: smaller as more variables are included (if they have expl. power)

5. Model Choice Criteria: need to use estimated SPSE:

a) Estimate using independent data → split into training + validation
b) Use all existing data to estimate $\hat{\beta}_M$ → correct training error → C_p

a) **Mallows's C_p**: $C_p = \frac{\sum_{i=1}^{n}(y_{i,i} - \hat{y}_{i,in})^2}{\hat{\sigma}^2 \text{ = est. on full model}} - n + 2|M| \rightarrow \text{Min.}$

b) **AIC**: $AIC = -2\ell(\hat{\beta}_M, \hat{\sigma}^2_{ML}) + 2(|M|+1) \rightarrow \text{Min.}$
$= n \cdot \log(\hat{\sigma}^2_{ML}) + 2(|M|+1)$ for lin. model with $\varepsilon \sim N(\mu, \sigma^2)$

c) **BIC**: $BIC = -2\ell(\hat{\beta}_M, \hat{\sigma}^2_{ML}) + \log(n)(|M|+1)$
$= n \cdot \log(\hat{\sigma}^2_M) + \log(n)(|M|+1)$ for lin. model with $\varepsilon \sim N(\mu, \sigma^2)$

d) **Adjusted Coef. of Determination (R^2)**: always smaller than mult. R^2

$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum(y_i - \bar{y})^2} \rightarrow \bar{R}^2 = 1 - \frac{(n-1)}{(n-p)}(1-R^2)$

e) **Cross Validation**: can address overfitting / selection bias by splitting up data
- Partition data set in r sets of similar size
- take set i as validation set, $r-1$ sets left for estimation
→ get SPSE for all r subsets
→ CV = $\sum$ SPSE$^{(i)}$ → use model with smallest sum of SPSE

↳ **LOOCV**: use all observations except one data point
- $CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i,in}^{-i})^2 \ \left(= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_{i,in}}{1-h_{ii,in}}\right)^2 \text{ for LSE}\right)$

↳ select model using the one-SE rule

6. Model Selection Procedures: (at no pre-selection exists)

a) **Complete model selection** → all start with H_0 (null model)
1. go through every model for $k = 1, 2, \dots, p$ → $\binom{p}{k}$ possible models
2. Pre-Select from models with same p every RSS (e.g. diff. specification)
3. Select single best model based on model choice criteria
→ can apply leaps & bounds algo.

b) **Forward Selection**
- Consider all $p-k$ models that add an add. predictor to M_k
- as above → variable which offers greatest reduction in the pre-selected model choice criteria is chosen
- also applicable if $p > n$ (unlike back ward selection)

c) **Backward Selection**: as above but starts with full model

d) **Stepwise Selection**: combination of b) & c), every step inclusion or deletion of a variable is considered

- Model selection based on sign. not useful as t-test would not $\sim t$
→ as n↑, any variable ($\neq 0$) would become sign.

7. Model Diagnostics: assessing validity of the model

- look at residuals (but can be heteroscedastic & correlated
→ a) Standardized residuals (not useful as unknown dist).
b) Studentized residuals (based on LOO-estimator)

$r^* = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}(1 + x_i^T(X_{(i)}^TX_{(i)})^{-1}x_i)} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \sim t_{n-p-1}$

A+B) Mean zero & Homoscedasticity

1) Tukey-Anscombe Plot: raw residuals $\hat{\varepsilon}_i$ vs. $\hat{y}_i$ plot
2) Scale-Location plot: $(|\text{standardized residuals}|)^{1/2}$ vs. $\hat{y}_i$ plot

→ if heteroscedastic errors do seem to occur: Transformation OR WLS

i) fix non-linearity → pol. transform x and y
ii) stabilize the var → transform Y

→ often use particular transformation family:

- **Power family**: x^λ for $\lambda \neq 0$ and $\log(x)$ (for $\lambda = 0$)
- Transforming only predictor: **Scaled power**

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

→ preserves direction of assoc. of x and y (not as power family)

- Transforming response: **Box-Cox Procedure**
  - need to have $y > 0$ (otherwise add constant)
  - automatically identifies a transformation (estimates λ) from the (scaled) power family
- Variance-Stabilizing Transformations: most common $\log(y)$, also $\sqrt{y}$

C) Autocorrelated errors

$\varepsilon \sim \mathcal{N}(0, \Sigma) \rightarrow \hat{\beta} \sim \mathcal{N}(\beta, \underbrace{(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}}_{\text{size of corr. depends on } X \text{ & } \Sigma})$

- **AR1 process**: $\varepsilon_i = \rho\varepsilon_{i-1} + u_i$, $-1 < \rho < 1$
  ↳ then $E[\varepsilon] = 0$, $\text{Cov}(\varepsilon) = \frac{\sigma_u^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & & \ddots & & \vdots \\ \rho^{n-1} & \dots & & & 1 \end{pmatrix} = \sigma^2 W^{-1}$

→ decay of correlation between ε_i and ε_{i-j} as $\text{Lag } j \uparrow$
  - for $\rho > 0$ → geometric decay
  - for $\rho < 0$ → corr. decreases with alternating signs

- **ACF** $(\rho_j) = \frac{\text{Cov}(\varepsilon_{i}, \varepsilon_{i-j})}{V \quad \varepsilon} = \rho^j$ → for AR1: $j = 1$ (can both be estimated)
- **PACF** (ϕ_j) is jth reg. coef. in $\varepsilon_i = \alpha_1\varepsilon_{i-1} + \dots + \alpha_j\varepsilon_{i-j} + u_i$

→ diagnose by looking at residual plots over time

D) Normality of errors

: use QQ / Normal Plot

E) Linear model → multicollinearity issue

- $\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{z_j^T z_j} = \frac{\sigma^2}{(1-R_j^2)\sum(x_{ij}-\bar{x}_j)^2}$
  - $z_j \perp$ with corr. X
  - $R_j^2 \uparrow$ with higher dependence (where R_j^2 is R^2 from reg. on x_j)
- **VIF** $= \frac{1}{1-R_j^2}$, problem if > 10
- **Scaled Residuals**: $\tilde{r} = \frac{y - X\hat{\beta}}{\|y - X\beta\|_2} = \frac{(I-H)\varepsilon/\sigma}{\|(I-H)\varepsilon/\sigma\|_2}$ (scaled by L2 norm)

and $\tilde{r} \sim \mathcal{N}(0,1)$ → simulate under H_0 and compare with the observed scaled residuals

- **Residual Prediction Test**: H_0: $y = x^T\beta + \varepsilon$ vs H_A: $y = f(x) + \varepsilon$
  → Compare scaled residuals for both models
  → can get p-values as we know that under H_0 $\tilde{r} \sim \mathcal{N}(0,1)$

8. Outlier Detection

when $E[y_i | x_i] = x_i^T\beta + d$

- may affect both parameter estimates (part. if not outer) and σ^2
- look for Large residuals (→ see Studentized residuals)

→ observations with Large r^* are pot. outliers (can also test)

- **Leverage**: outliers in x direction (extreme values of x)
  diagonal elements of H: $\frac{1}{n} \leq h_{ii} \leq 1$ ($h_{ii} > \frac{2p}{n}$ examine!)
  - Large leverage → $\text{Var}(\varepsilon_i)$ small
  - for only one covariate x_i: $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$

→ lev ↑ as x_i further away from $\bar{x}$
→ obs. with large lev → large influence on analysis

- **Cooks Distance**: compare $\hat{y}$ vs $\hat{y}_{(i)}$ → look at euclid. distance

$$D_i = \frac{1}{p}\frac{\hat{\varepsilon}_i^2}{\hat{\sigma}^2(1-h_{ii})} \cdot \frac{h_{ii}}{1-h_{ii}}$$ (depends on both lev and residuals)

$$= \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T(\hat{\beta}_{(i)} - \hat{\beta})}{p \cdot \hat{\sigma}^2} = \frac{(\hat{y}_{(i)} - \hat{y})^T(\hat{y}_{(i)} - \hat{y})}{p \cdot \hat{\sigma}^2}$$

$D_i > 0.5$ attention, > 1 always examine

9. General Linear Model (with heteroscedastic and/or corr. errors)

- Now: $\text{Cov}(\varepsilon) = \sigma^2 W^{-1}$ ($W \neq \mathbb{1}$), for hetero var but no correlation: $\text{Cov}(\varepsilon_i) = \frac{\sigma^2}{w_i}$
- **Aitken/WLS Estimator**: $\hat{\beta}_{WLS} = (X^TWX)^{-1}X^TWy$ (= MLE for $\varepsilon \sim \mathcal{N}$)
  - from minimizing: $WLS(\beta): (y - X\beta)^TW(y - X\beta)$
    for hetero + uncorr. $WLS(\beta): \sum_{i=1}^n w_i(y_i - x_i^T\beta)^2$
  - Properties: $E[\hat{\beta}_{WLS}] = \beta$, $\text{Cov}(\beta_{WLS}) = \sigma^2(X^TWX)^{-1}$, GM holds
  - with $\varepsilon \sim \mathcal{N}(0, \sigma^2W^{-1})$: $\hat{\sigma}^2_{ML} = \frac{1}{n}\hat{\varepsilon}^TW\hat{\varepsilon}$ → biased estimator (WLS residual)
  → REML estimator: $\frac{1}{n-p}\hat{\varepsilon}^TW\hat{\varepsilon}$
- **Two-stage Least Squares** (simultaneous estimation of β, σ^2, w_i)
  → want to estimate the var through $\varepsilon_i^2 = \sigma_z^2 + \nu$ → $\sigma_i^2 = z_i^T\alpha$ (z^T often $= x^T$)
  1. Obtain $\hat{\beta}$ through OLS → get residuals $\hat{\varepsilon}_i$
  2. Obtain $\hat{\alpha}$ from unweighted reg (OLS) between $\hat{\varepsilon}_i^2 \sim z_i$
  → for GLM with $\hat{w}_i = \frac{1}{z_i^T\hat{\alpha}}$ → $= \frac{1}{\exp(z_i^T\hat{\alpha})}$ to ensure $z_i^T\alpha > 0$
- **White Estimator**: $\widehat{\text{Cov}}(\hat{\beta}) = (X^TX)^{-1}X^T\,\text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2)X(X^TX)^{-1}$
- **Two-stage Least Squares for AR(1)**: as above but in step 2:

$$\hat{\rho} = \sum_{i=2}^n \hat{\varepsilon}_i\hat{\varepsilon}_{i-1} \Big/ \left(\sqrt{\textstyle\sum\hat{\varepsilon}_i^2}\sqrt{\textstyle\sum\hat{\varepsilon}_{i-1}^2}\right)$$

3. Insert $\hat{\rho}$ into weight matrix $W \rightarrow \hat{W} \rightarrow$ apply WLS

10. Robust Regression

(when we have obs. with high influence and/or leverage)

- **Influence function**: local concept → measure sensitivity of $\hat{\beta}$ w.r.t. one outlier
  - Version of empirical IF for $n \to \infty$: Consider cont. dist of $\hat{\beta}$
    $F = (1-\varepsilon)F + \varepsilon\delta_x$ (as $\varepsilon \to 0$) (measures sensitivity of β to an. small change ε)
  - Shape of IF for LSE: $\psi_{LS}(\varepsilon) = 2\varepsilon$ (prop. to ε)
- **Breakdown Point**: global concept → how much contamination data until useless
  - Finite sample BP: Fraction of data that can be given arbitrary values without making estimator arb. bad
    ↳ cannot be $> \frac{1}{2}$
  - asymptotic BP: Limit of the finite sample BP as $n \to \infty$
    e.g.) for OLS (with mean) the BP $= 1/n$; median could be $1/2$
- **M-estimator**: method of estimation; pays less attention to obs. with large ε
  1) min. obj function $\sum\rho(\varepsilon_i) = \sum\rho(y_i - x_i^T\beta)$ ($\rho(\cdot)$ is obs. function)
  2) ↳ min. means: $0 = \frac{\partial}{\partial\beta}\sum\rho(\varepsilon_i) = \sum\psi(y_i - x_i^T\beta)\cdot x_i^T$ (↑ influence curve!)
  3) Set $w(\varepsilon) = \psi(\varepsilon)/\varepsilon$ → $\sum w_i(y_i - x_i^T\beta)x_i^T = 0$
  4) → like WLS → solve via IRLS
- **Objective functions**: → the $\rho(\cdot)$ for M-Estimation
  a) LS estimator: $\rho_{LS}(\varepsilon) = \varepsilon^2$, $w_{LS}(\varepsilon) = 1$
  b) LAV estimator: $\rho_{LAV}(\varepsilon) = \sum|y_i - x_i|$, $w_{LAV}(\varepsilon) = \frac{1}{|\varepsilon|}$
  c) Huber loss: combines LS and LAV for large $|\varepsilon|$: $\rho_H(\varepsilon) = \begin{cases} \frac{1}{2}\varepsilon^2 & \text{if } |\varepsilon| \leq c \\ c(|\varepsilon| - c/2) & \text{if } |\varepsilon| > c \end{cases}$
     → tuning parameter $c \approx 2 \cdot MAD$ where $MAD \equiv \text{median}|y_i - \hat{\mu}|$ (use median as estimator)
  d) Bisquare loss: for large $|\varepsilon|$ → weighting becomes constant / levels off

$$\rho_{BS}(\varepsilon) = \begin{cases} \frac{c^2}{6}\left\{1 - \left[1 - \left(\frac{\varepsilon}{c}\right)^2\right]^3\right\} & \text{if } |\varepsilon| \leq c \\ c^2/6 & \text{if } |\varepsilon| > c \end{cases} \quad \text{with } c \approx 7 \cdot MAD$$

11. Generalized Linear Models (GLMs) - Binary ($y \sim \text{Bin}(\pi)$)

- **Logit Model**:
  - Logistic response: $\pi = h(\eta) = \exp(\eta)/(1+\exp(\eta))$
  - Logit link: $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{P(y_i=1)}{P(y_i=0)}\right) = \eta = x^T\beta$

→ Log effect on odds

- **Probit Model**: · Probit response: $\pi = \Phi(\eta) = \Phi(x^T\beta)$
  · Var: $\sigma^2 = \pi^2/3$ → to compare Logit and probit ($\sigma^2 = 1$): $\hat{\beta}_L = \frac{\pi}{\sqrt{3}}\hat{\beta}_P$
- **Latent linear models**: $y_i = \begin{cases} 0 \text{ if } \varepsilon_i \leq 0 \\ 1 \text{ if } \varepsilon_i > 0 \end{cases}$ Assume $\varepsilon_i = x_i^T\beta - \varepsilon_i$
  $\pi_i = P(y_i = 1) = P(\varepsilon_i < x_i^T\beta)$:
  - for $\varepsilon_i \sim \mathcal{N}(0,1)$ → Probit (given a often normal → Probit calculated)
    ↳ only identifiable to $\eta = \beta/\sigma$ due to std. → look at β_R/β_L
  - for $\varepsilon_i \sim$ Logistic → Logit
- **For Grouped Data**: rel. freq. (per group): $\bar{y}_i \sim \text{B}(n_i, \pi_i)/n_i$
  · $E[\bar{y}_i] = \pi_i$, $\text{Var}[\bar{y}_i] = [\pi_i(1-\pi_i)]/n_i$
  $= [\bar{y}_i(1-\bar{y}_i)]/n_i$ (estimate)
- **Overdispersion**:
  · correct with $\phi > 1$: $\text{Var}[\bar{y}_i] = \phi\frac{\pi_i(1-\pi_i)}{n_i}$
  · estimate: $\hat{\phi}_D = \frac{1}{n-p}\cdot D$
- **Maximum Likelihood Estimation**
  · when there is no closed form solution to $S(\beta) = 0$
  a) Newton-Raphson method
  1. At $\theta^{(0)}$ score is approx. through 1st order Taylor expansion: $S(\theta^*) \approx S(\theta^{(0)}) + S'(\theta^{(0)})(\theta - \theta^{(0)})$
  2. Improve approx. iteratively: $\theta^{(t+1)} = \theta^{(t)} + I(\theta^{(t)})^{-1}\cdot S(\theta^{(t)})$
  b) Fisher Scoring algorithm
  2. use $J(\theta^{(t)})$ instead obs. FI: $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + J(\theta^{(t)})^{-1}\cdot S(\hat{\theta}^{(t)})$
  with exp. FI $J(\theta) = E[-S'(\theta)] = E[S(\theta)\cdot S^T(\theta)]$
  → often estimating $\hat{\beta}$ in logit: $\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (X^TV^{(t)}X)^{-1}X^T(y - \pi^{(t)})$
  3. Stop when convergence criterion is met ↳ need X full rank for converge
- **Testing linear hypotheses**
  · same setup as in classical linear model with diff. statistics:

a) Likelihood Ratio Test: $LR = 2(\ell(\hat{\beta}) - \ell(\tilde{\beta})) \overset{a}{\sim} \chi^2_r$
b) Wald test: $W = (C\hat{\beta} - d)^T(C\,J(\hat{\beta})^{-1}C^T)(C\hat{\beta} - d) \overset{a}{\sim} \chi^2_r$
→ reject if LR, W are "large"

- **Deviance**: measures goodness-of-fit for GLMs
  · compares perfect fit / saturated model to training model

a) For ind. data: $D(\hat{\pi}) = -2\ell(\hat{\pi})$
b) For grouped data: $D(\hat{\pi}) = 2\cdot(\ell(\tilde{\pi}) - \ell(\hat{\pi}))$ with $\tilde{\pi}_i = \frac{y_i}{n_i} = \bar{y}_i$
↳ if models are nested = LR

- **Count Data Regression**: often based on Poisson distribution
  a) Log-linear Poisson model: $\lambda_i = \exp(\eta_i)$ with $\eta = x_i^T\beta$
  b) Linear Poisson model: $\lambda_i = x_i^T\beta$ (if x have a linear effect on λ)
- **Unified framework for GLM**: $\mu = E[y] = h(\eta)$
  · part of univariate exp. family: $f(y_i|\theta) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi,\omega)\right\}$
  θ: canonical parameter (of interest), ϕ: dispersion parameter (nuisance), ω: c

12. Penalized Regression

(when LSE unstable (collinearity or $p > n$)

- Penalized LS objective: $PLS(\beta) = (y - X\beta)^T(y - X\beta) + \lambda\text{pen}(\beta)$

a) **Ridge Regression**: $\text{pen}(\beta) = \|\beta\|_2^2 = \beta^T\beta$ → $\hat{\beta}_R = (X^TX + \lambda I)^{-1}X^Ty$
  · $E[\hat{\beta}_R] = (X^TX + \lambda I)^{-1}X^TX\beta$ (biased)
  · $\text{Cov}[\varepsilon] = \sigma^2 I$; $\text{Cov}(\beta_R) = \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$
  ↳ smaller var than β_{OLS} for $\lambda > 0$
  · For orthogonal covariates: $\beta_R = n/(n+\lambda)\beta_{OLS}$ and $\text{Cov}(\beta_R) = \sigma^2(n+\lambda)^{-2}I_p$
  · high Var / corr. variables are penalized less

b) **LASSO Regression** $\text{pen}(\beta) = \|\beta\|_1 = |\beta| \Rightarrow \hat{\beta}^{LASSO} = \arg\min_\beta (y - X\beta)^T(y - X\beta) + \lambda\sum_{j=1}^p |\beta_j|$
  · incl. model selection → can shrink $\beta_j = 0$
  · also biased but with $\text{Var} \leq \text{Var}(\hat{\beta}_{OLS})$ (# non-zero parameters)

c) **Best subset selection**: $\arg\min_\beta (y - X\beta)^T(y - X\beta) + \lambda\sum_{j=1}^p \mathbb{1}(\beta_j \neq 0)$