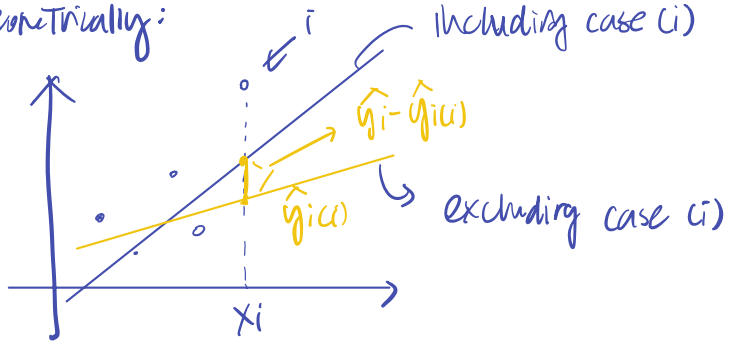
Lecture 16 2023년 10월 25일 화요일.

1) LOOCV DFFITS, DFBETAS, Cook's distance.
2) Multicollinearity

LOOCV: leave-one-out cross validation
idea: to access how including/excluding case i affects $\hat{y}_i$, or bk, $\hat{y}$

▶ $(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$
differences

Geometrically:



$Var(\hat{y}) = Var(H \cdot y) = H \underbrace{Var(y)}_{\sigma^2 I_n} H^T = \sigma^2 HH^T = \sigma^2 H$

$Var(\hat{y}_i) = \sigma^2 h_{ii}$

$\widehat{Var}(\hat{y}_i) = MSE_{(i)} h_{ii}$ (estimated). ~~$(MSE\, h_{ii})$~~ $MSE_{(i)}$ might be less bias than MSE.
exclude case i

Fact: $DFFITS_i = e_i \left(\frac{(n-p-1)}{SSE(1-h_{ii}) - e_i^2}\right)^{\frac{1}{2}} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}}$

↳ $= t_i$ (extremely studentized residual, studentized di).

Robust linear Regression might be a useful model to use in real world if there are many outliers in the dataset. (Not covered in the class).

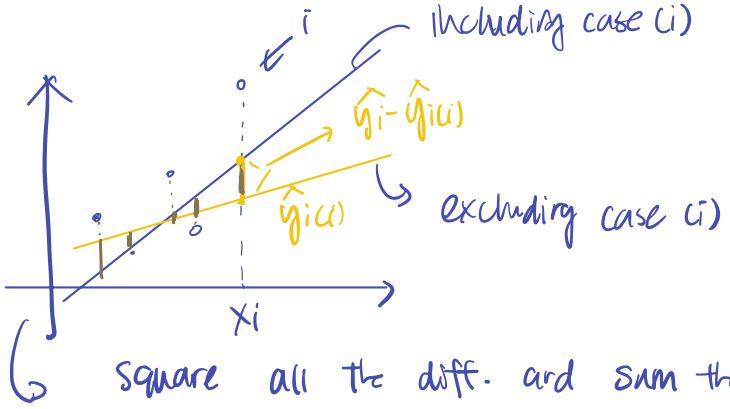
Rule of thumb: influential on $\hat{y}_i$
if $|DFFITS_i| > 1$ for "small"/"medium" datasets. Should be # of b.
$|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ for "large" datasets. p = # of the predictors.

$DFFITS_i$ = influence on ith fitted value by the ith case X_i.

Cook's distance := influence on All fitted values by the ith case X_i.

$D_i = \sum_{j=1} (\hat{y}_j - \hat{y}_{j(i)})^2 / (p \cdot MSE) = \|\hat{y} - \hat{y}_{(i)}\|^2 / p \cdot MSE$

Cook's Distance.



square all the diff. and sum them up.

Exercise: $\|\hat{y} - \hat{y}_{(i)}\|^2 = (b - b_{(i)})^T (X^T X)(b - b_{(i)})$ ← Check that this is χ^2_p. (*)

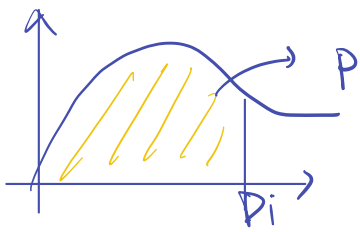
(since $b \sim N(\beta, (X^T X)^{-1}\sigma^2)$

$\|\hat{y} - \hat{y}_{(i)}\| / \sigma^2 \sim \chi^2_p$ $\quad SSE/\sigma^2 \sim \chi^2_{n-p}$ $\quad$ So: $(\chi^2_p / p) / (\chi^2_{n-p}/n-p)$

$\Rightarrow$ $\boxed{D_i \sim F(p, n-p)}$ (*) and that numerator / denominator are $\perp$. (independent)

square all the diff. and sum them up.
and then normalized by pMSE to obtain D_i.

Standardize : make the data with mean zero, variance 1.
normalize : make the data in between 0 & 1.

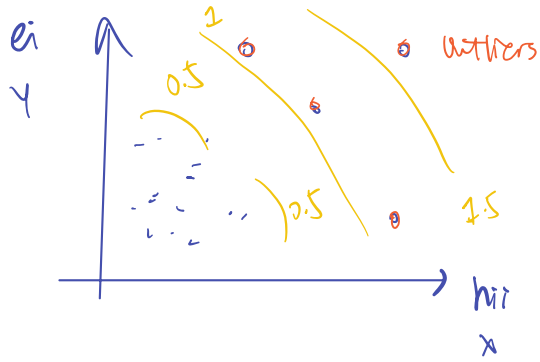


Rule of thumb: $P \in [0.1, 0.2]$ ← no influence
$P \in (0.5, 1]$ ← major influence.

Fact: $D_i = (e_i^2 / pMSE) \cdot (h_{ii}/(1-h_{ii})^2) = P(e_i, h_{ii})$ ↳ leverage.
↳ residual

Influential case: if 1) e_i is large, h_{ii} is moderate $\quad 0 < h_{ii} \leq 1$
2) e_i is med., h_{ii} large
3) both are large.

In R: plot(lm(...)).



outliers. $D_i(e_i, h_i)$ — y, x.

→ Influence on a regression coeff. (of a single case (i)).

$$DFBETAS_{k(i)} = \text{(difference)} = (b_k - b_{k(i)}) \Big/ \sqrt{MSE_{(i)}\, C_{kk}}$$

where $C_{kk} = [(X^TX)^{-1}]_{k+1,k+1}$, $k = 0 \dots p-1$.

= a change in the kth estimated b_k when the ith obsn. X_i is removed, studentized.

$Var(b) = \sigma^2(X^TX)^{-1} \Rightarrow Var(b_k) = \sigma^2 \underbrace{[(X^TX)^{-1}]_{k+1,k+1}}_{C_{kk}}$

Rule of thumb.: $|DFBETAS_{k(i)}| > 1$, "small" or "medium" dataset
$|DFBETAS_{k(i)}| > 2/\sqrt{n}$, "large" dataset

X_i is influential on b_k

Note: D_i (Cook's distance) $\approx \sum_{k=0}^{p-1} (DFBETAS)^2_{k(i)}$

▼ Multicollinearity (M.C.)

$Corr(X_i, \sum_{j \neq i} a_j X_j) \Rightarrow \det(X^TX) \approx 0.$

↑ Cols X are almost linearly dependent

Causes multicollinearity.

→ Informal diagnostics of M.C.

① b_k change a lot when include / exclude predictor X_k.
② β_k is not significant for important predictors
③ $b_k > 0$ but should be < 0 (or vice versa)
④ large Γ_{XX}
⑤ wide CI's for β_k

→ Variance inflation factor (VIF) $\leftarrow$ Var of b_k's

$Var(\underline{b}) = \sigma^2(X^TX)^{-1}$
$Var(\underline{b}^*) = \sigma^{2*}(\Gamma_{XX})^{-1}$ (standardized model).

Def: $VIF_k = [\Gamma_{XX}^{-1}]_{kk}$.

Fact: $VIF_k = \frac{1}{1-R_k^2}$, $k=1,2,\cdots,p-1$.

R_k^2 = coeff. of multiple determination when X_k is regressed on $X_1,\cdots,X_{k-1},X_{k+1},\cdots,X_{p-1}$

$p-2$ predictors (b/c excluding X_k).

If $R_k^2=0$ i.e. X_k is not linearly related with other pred. $\Rightarrow VIF_k=1$.
If $R_k^2\to 1 \Rightarrow VIF_k\to\infty \Rightarrow Var(b_k^*)\to\infty$.

Rule of thumb.: $VIF_k > 10 \Rightarrow$ bad, M.C. is present
↑
$R_k^2 \approx 0.9$.

ISR Internal studentized residual

$\frac{e_i}{\sqrt{MSE(1-h_{ii})}}$

internal.

ESR external stu. res.

$\frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$

excluded, external.

★ Lecture 18 2023年11月6日.

★★ weighted LS

$y = X\beta+\varepsilon$, $\varepsilon \sim N(0,\sigma^2 I)$ $\quad E(\varepsilon_i)=0$

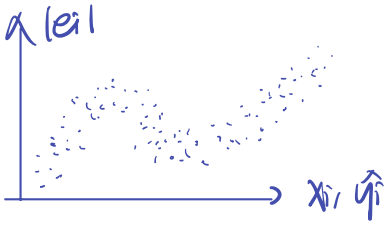
$cov(\varepsilon_i,\varepsilon_j)=0$, $i\neq j$

$Var(\varepsilon_i)=\sigma_i^2$ ← ~~$Var(\varepsilon_i)=\sigma^2$~~

heteroskedasticity

Before: How to fix nonconstant variance? Transform y via Box-cox transform.
But you may destroy the relation btw. X & y. var. stabilizing.

OR:



messy!
transforms fail to stabilized variance.

Answer: WLS (weighted Least Square)
as opposed to OLS. (Ordinary Least Square).

Case 1: Assume σ_i^2's are given

$L(\beta) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma_i^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i-\beta_0-\beta_1 x_{i1}-\cdots-\beta_{p-1}x_{i,p-1})^2}{\sigma_i^2}\right)$

↓ likelihood

$w_i := \frac{1}{\sigma_i^2}$, $L(\beta) = \prod_{i=1}^{n}\left(\frac{w_i}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} w_i(y_i-\beta_0-\beta_1 x_{i1}-\cdots-\beta_{p-1}x_{i,p-1})^2\right)$

$= Q_W(\beta)$

Maximizing $L(\beta)$ wrt β $\Leftrightarrow$ minimizing $Q_W(\beta)$ wrt β

$$Q_W(\beta) = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i,p-1})^2$$
$$= (y - X\beta)^T W (y - X\beta)$$

where $W = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$

observations with high σ_i^2 have low w_i
observations with low σ_i^2 have high w_i

$\min_{\beta} Q_W(\beta) \Rightarrow$ normal equations for WLS.

$$(X^T W X) b_W = X^T W y \Rightarrow \boxed{b_W = (X^T W X)^{-1} X^T W y}$$

estimated β using WLS.

If $\text{var}(\varepsilon_i) = \sigma^2 \Rightarrow$ you can show that $b_W = b^{OLS}$

Properties: b_W = unbiased
= min variance
= consistent (i.e. $n \to \infty \Rightarrow b_W \to \beta$)

Case 2: Variances are known up to a proportionality constant, i.e. we know $\frac{\sigma_i^2}{\sigma_j^2}$

Define $w_i = \frac{k}{\sigma_i^2}$ → unknown
→ this is all we know.

Example: happens when response variable is an average of groups of sizes n_i:

$\bar{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$ constant variance σ^2

$\bar{y}_i$ ↓ response variable

$\text{Var}(\bar{y}_i) = \frac{\sigma^2}{n_i}$ → estimate
→ known

$\Rightarrow \text{Var}(\varepsilon_i) = \sigma_i^2, \quad \sigma_i^2/\sigma_j^2 = n_j/n_i$

In this case, $\text{var}(b_W) = k(X^T W X)^{-1}$

$S^2(b_W) = MSE_W (X^T W X)^{-1}$

MSE_W = estimator of $k = \frac{1}{n-p} \sum_{i=1}^{n} w_i e_i^2 = \hat{k}_{MLE}$.

Before: $\text{Var}(b) = \sigma^2 (X^T X)^{-1}$, $S^2(b) = MSE (X^T X)^{-1}$

$\frac{e^T W e}{n-p}$ [ ]

Case 3: Variances σ_i^2 are unknown

$\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) - (\underbrace{E\varepsilon_i}_{0})^2 = E(\varepsilon_i^2) = \sigma_i^2 \approx e_i^2$

Plot $|e_i|$ or e_i^2 vs. X_{ik} or vs. $\hat{y}_i$.

If $|e_i|$ or e_i^2 plot has shape like  , then regress $|e_i|$ or e_i^2 onto X_k wrt X_{ik}

or $X_1 + \cdots + X_{p-1}$ wrt $\hat{y}_i$

- $|e_i| \sim X_{ik} \Rightarrow$ fitted values $|e_i| := \hat{s}_i$ estimate of sd σ_i
- $e_i^2 \sim X_{i1} + \cdots + X_{p-1} \Rightarrow$ fitted values of $e_i^2 := \hat{v}_i$ estimate of variance σ_i^2

choose $w_i = \frac{1}{\hat{s}_i^2}$ or $\frac{1}{\hat{v}_i} \Rightarrow b_w = 1 \cdot (X^T W X)^{-1} X^T W y$ the WLS estimate

sanity check $MSE_w \approx 1$ since $MSE_w = \hat{K}$

Procedure: IWLS / IRLS

↓ iteratively weighted LS ↓ iteratively reweighted LS

- Step 1: Fit regression via OLS $\Rightarrow e_i$
- Step 2: Regress $|e_i|$ or e_i^2 onto X_{ik} or $\hat{y}_i$
- Step 3: get $w_i = 1/\hat{s}_i^2$ or $w_i = 1/\hat{v}_i$
- Step 4: fit regression using WLS

  If b_w is significantly different than b

  ↳ means residual estimates were not good enough.

  $\Rightarrow$ do steps (2)–(4) again

  (new e_i's in (2) are from (4)) $e_i = y_i - X b_w$

Usually 2-3 iterations is enough.

What if we use OLS in case $\sigma_i^2 \neq$ constant?

$b = (X^T X)^{-1} X^T y \rightarrow$ unbiased, consistent, not min. variance

$Var(b) \neq \sigma^2 (X^T X)^{-1}$ ↘ $\neq$

$= (X^T X)^{-1} (X^T \underbrace{Var(y)}_{\text{unknown}} X)(X^T X)^{-1}$

$$\begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \sigma_2^2 & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$S^2(b) = (X^T X)^{-1} (X^T S_e X)(X^T X)^{-1}$ where $S_e = \begin{bmatrix} e_1^2 & & 0 \\ & \ddots & \\ 0 & & e_n^2 \end{bmatrix}$

White's estimator / robust covariance matrix of b.

Inference in WLS case:

$\boxed{S^2(b_w) = MSE_w (X^T W X)^{-1}}$ $\rightarrow$ CI's, PI's etc.

$\Updownarrow$

$Var(b_w) = 1 \cdot (X^T W X)^{-1}$

$S^2(b) = MSE\,(X^T X)^{-1}$

g g$ sigma

g (g)$ sigma

Summary

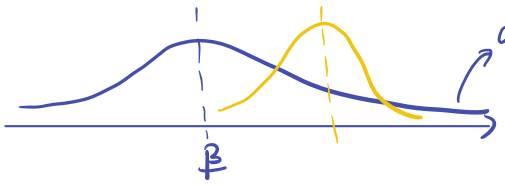
II. Ridge Regression (11.2)

remedial measure for Multicollinearity.

Signs of M.C.: sd(bi) are very large ⇒ OR
bi's for important pred's not significant.

Idea: reduce variance, sacrifice unbiasedness.

i.e.: finds $\underline{b}^R$ s.t. "$Var(b^k)$ is small", but $E(\underline{b}^R) \neq \beta$



distribution of $\underline{b}$ (OLS) is wider.

distribution of $\underline{b}^R$ (biased)
closer to β due to smaller variance.

X_{ik}, y_i → correlation transform
⇒ $\underline{b} = \underline{r}_{xx}^{-1} \underline{r}_{yx}$
⇒ Ridge regression: $\underline{b}^R = (\underline{r}_{xx} + c \cdot \underline{I})^{-1} \cdot \underline{r}_{yx}$

improving conditioning of the matrix to −1
appropriately chosen constant.

★ Lecture 19. 2023年11月8日

(I) Ridge Regression
↳ Remedial measure for M.C.
y_i, X_{ik} → cor'n transform (entries btw −1, 1)
y_i^*, X_{ik}^*

OLS: $\underline{b} = \underline{r}_{xx}^{-1} \underline{r}_{yx}$ Ridge regression: $\underline{b}^R = (\underline{r}_{xx} + c \cdot \underline{I})^{-1} \underline{r}_{yx}$

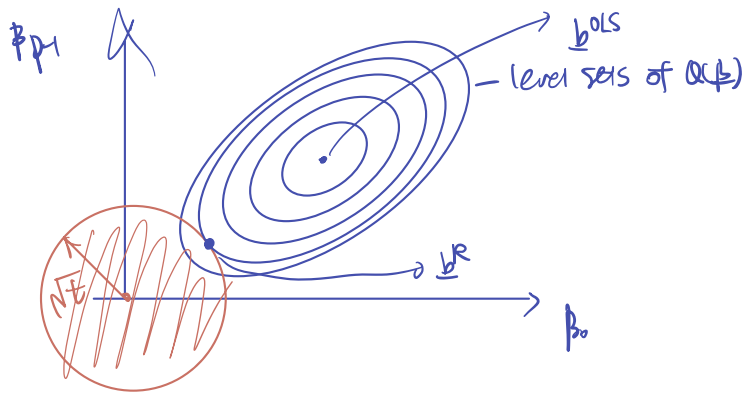
↑ appropriately chosen constant

★★ Point of view (I): minimization with regularization.

$$\underline{b}^R = \underset{\beta}{\operatorname{argmin}} \underbrace{\|\underline{y} - \underline{X}\beta\|^2}_{OLS} + \underbrace{c\|\beta\|^2}_{\text{regularization. (penalty term)}}$$

Objective function.
$Q(\beta)$.

; $\mathbf{X}^T\mathbf{X} = \mathbf{r}_{XX}$ and $\mathbf{X}^T\mathbf{Y} = \mathbf{r}_{YX}$,
cient of simple correlation betwee



✳✳ Point of view (II): Minimization with Constraints ⇔ POV. I.

$$\underline{b}^R = \underset{\underline{\beta}}{\operatorname{argmin}}\ Q(\underline{\beta}) = \underset{\underline{\beta}}{\operatorname{argmin}} \|\underline{y} - X\underline{\beta}\|^2 \qquad \text{subject to: } \|\underline{\beta}\|^2 \le t.$$

✳✳ simultaneously minimizing residuals AND minimizing $\underline{\beta}$'s.

If $c=0$: $\underline{b}^R = \underline{b}^{OLS}$
If $c\to\infty$: $\underline{b}^R = \underline{0}$
$\Rightarrow$ As $c\uparrow$, bias $\uparrow$.

POV I ≡ POV II under KKT conditions.

✳✳ POV (III): Linear Algebra.

M.C. $\Rightarrow \det(X^TX) \approx 0.$

$= \eta_1 \cdot \eta_2 \cdot \eta_3 \cdots \eta_K \approx 0.$

↳ at least one eigenvalue $\eta_i \approx 0$.

eigenvalues $(X^TX + c\cdot I) \Rightarrow$ eigenvalues $\eta_1 + c$;
$\eta_2 + c$;
$\vdots$
$\eta_K + c$;

$\det(X^TX + cI) = (\eta_1+c)(\eta_2+c)\cdots(\eta_K+c) \neq 0.$

✳✳ POV (IV): Geometric Interpretation.

Say $X^TX = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ $\quad c_1 = c_3$

$\underline{c}_1$ $\underline{c}_2$ $\underline{c}_3$

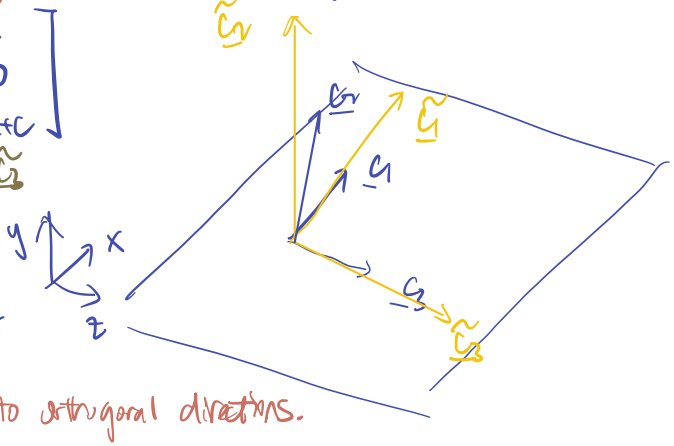
$X^TX + c\cdot I = \begin{bmatrix} 1+c & 0 & 1 \\ 0 & 1+c & 0 \\ 1 & 0 & 1+c \end{bmatrix}$ (x, y, z)

$\tilde{\underline{c}}_1$ $\tilde{\underline{c}}_2$ $\tilde{\underline{c}}_3$

Add the ridge correction term makes these columns less linearly dependent.

being pushed onto orthogonal directions.

Consider a general case where the $\underline{c}_k$'s are linearly dependent.



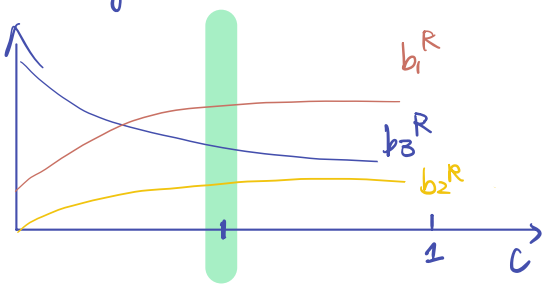
→ How to choose c?

Plot b_k^R for different values of $c \in [0,1]$

Plot VIF$_k$ for different values of $c \in [0,1]$

$VIF_k = \frac{1}{1-R_k^2}$ → regressing X_k onto the other predictors. (↑ if correlated).

Plot b_k against c.



Choose this c.

b/c b_k^R are "stable".

Downsides:
- choice of c is subjective.
- No distributional results (CIs, PI's)

↳ Bootstrapping

(II) Robust Regression (Ch. 11.3).

↳ useful when we have clusters of outliers.

OLS is sensitive to influential obsn's

Idea: create regression dampens influential obsn's

i) Robust regression I: LAR / LAD / L^1

least absolute residuals

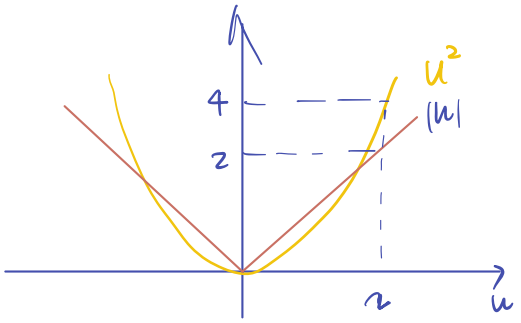
least absolute deviations

L^1 norm

obj. $\min_{\beta} \sum_{i=1}^{n} |y_i - (\beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_{p-1} X_{i,p-1})|$

L^2 norm: $\|u\|_2^2 = \sum_{i=1}^{n} u_i^2$ ← OLS $= \|u\|_{L^2}^2$

L^1 norm: $\|u\|_1 = \sum_{i=1}^{n} |u_i|$ ← LAR / LAD / L^1 $= \|u\|_{L^1}$

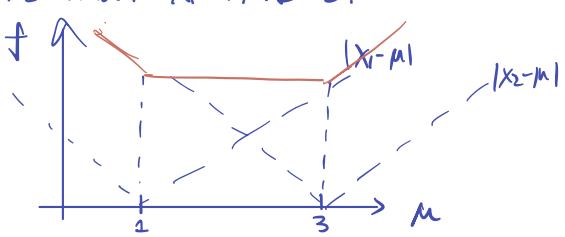


Downside:

not unique solutions.

Compare $f = \sum_i |x_i - \mu|$ & $f = \sum_i (x_i - \mu)^2$

Let's consider $x_1 = 1$, $x_2 = 3$.



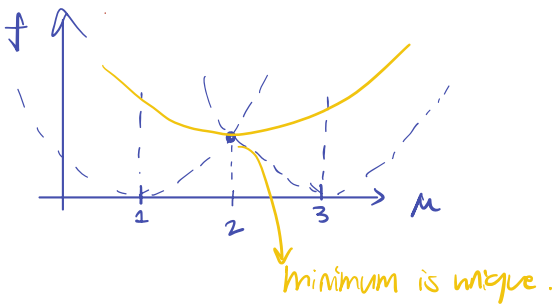
•) Robust Regression II: LMS - Least median of squares

$$\min_{\beta} \text{median } [y_i - (\beta_0 - \beta_1 x_1 - \cdots - \beta_{p-1}(x_{i,p-1})]^2$$

$$\min \text{median}(e_i^2)$$

	mean	median
1 2 3	2	2
1 2 10	≈4	2

robust.



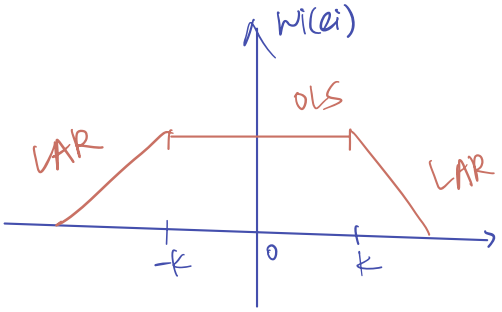
Downside: Hard to optimize. median not differentiable.

•) Robust regression III: IRLS. iteratively reweighted least squares.
similar to WLS, but influential obsn's are downweighted while others treated the same.

$$\min \sum_i w(e_i) e_i^2$$

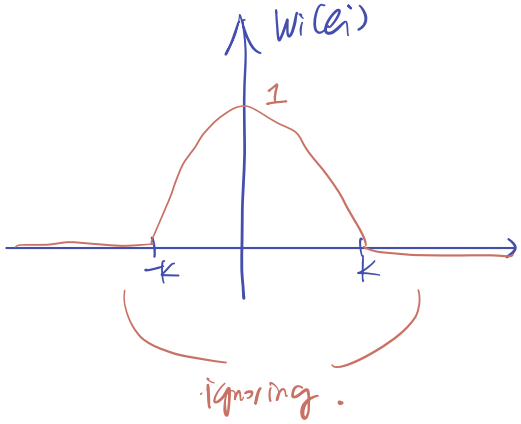
Huber

$$w(e_i) = \begin{cases} 1, & |e_i| < k \\ \frac{k}{|e_i|}, & |e_i| > k \end{cases}$$



Bisquare

$$w(e_i) = \begin{cases} \left[1 - \left(\frac{e_i}{k}\right)^2\right]^2, & |e_i| < k \\ 0, & |e_i| > k. \end{cases}$$



※ Lecture 20. 2023年11月13日(一).

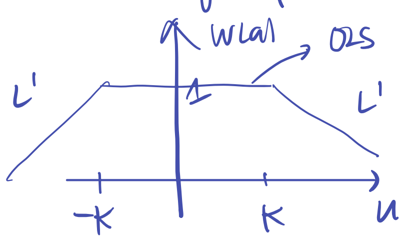
(I) Robust Regression — remedial measure for outliers (clusters)

LAD / LAR / L^1

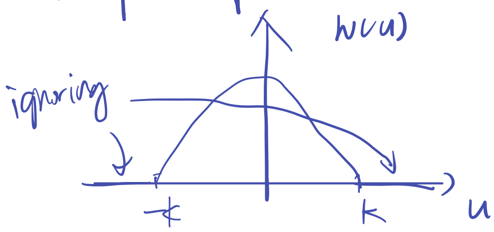
LMS (median)

IRLS

Huber weight func.



Bisquare function.



$u_i = e_i / MAD$

MAD = robust estimator of σ $= \frac{\sum_{i=1}^{n} |x_i - \text{median}|}{n}$

= median absolute deviation.

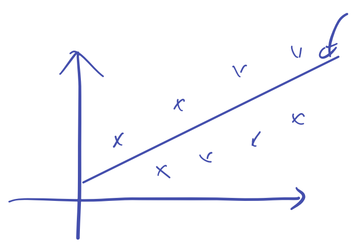
* If $e_i \sim N(0, \sigma^2)$, $E(MAD) = \sigma$

Do IRLS iteratively: – obtain initial weights with OLS

– iterate until b_k^{RR} stabilize OR look at $w(u_i)$, e_i / u_i, $\hat{y}_i$

Usually start w/ Huber, switch to bisquare later.

※※ Properties



- Requires knowledge of reg. func.
- If reg. func. unknown / difficult to establish
  ⇒ use nonparametric reg. methods.
  (lowess, regression trees)
- Can identify multiple outliers.
- Rob. regression is used to confirm OLS: if $b_k^{RR} \approx b_k^{OLS}$ ⇒ no influential obsn's
  $\hat{y} \neq H \cdot y$

<u>Downside</u>: can't obtain CI's, PI's, etc. Solution: bootstrap

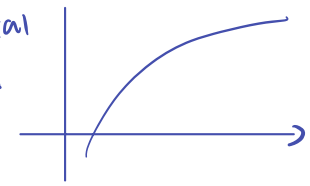
we eliminated influence of y outlying obsn's

For X outlying obsn's: bounded influence regression

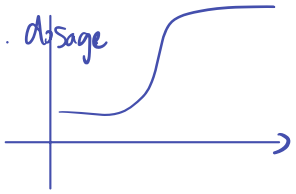
(II) Nonlinear regression. (Ch 13)

Lots y_i's are nonlinear func. of predictors.

biological growth



drug vs. dosage



$E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

Nonlinear reg:

$$y_i = f(\underset{q\times 1}{x_i}, \underset{p\times 1}{\gamma}) + \varepsilon_i$$

p & q not necessarily the same.

Examples: Eg. 1: $f(x_i, \gamma) = x_i^T \gamma$ ← MLR

Eg. 2: exponential reg. model: (A) $y_i = \gamma_0 \exp(\gamma_1 x_i) + \varepsilon_i$ $\quad q=1, p=2$

(B) $y_i = \gamma_0 + \gamma_1 \exp(\gamma_2 x_i) + \varepsilon_i$ $\quad q=1, p=3$.

Eg. 3: Logistic Reg.: $y_i = \dfrac{\gamma_0}{1+\gamma_1 \exp(\gamma_2 x_i)} + \varepsilon_i$ $\quad q=1, p=3$.

- Linear (MLR)
- Intrinsically linear (A). $f(x_i, \gamma) = \gamma_0 \exp(\gamma_1 x_i)$

$$\log f = \underbrace{\log \gamma_0}_{\tilde{\gamma}_0} + \gamma_1 x_i = \tilde{\gamma}_0 + \gamma_1 x_i$$

How to obtain γ? use OLS, MLE.

OLS: $Q(\gamma) = \sum_{i=1}^n (y_i - f(x_i, \gamma))^2$

$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}}\, Q(\gamma)$ $\quad \dfrac{\partial Q}{\partial \gamma_k} = 2\sum (y_i - f(x_i, \gamma)) \dfrac{\partial f}{\partial \gamma_k}(x_i, \gamma) = 0.$ (Normal Eqn.)

MLE: $L(\gamma, \sigma^2) = \dfrac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\dfrac{1}{2\sigma^2} \sum (y_i - f(x_i, \gamma))^2\right)$

For MLR, e.g. $x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{i,p-1} \end{bmatrix}$

(III) Probit / Logit Regression (Ch. 14)

Nonlinear Regression when y is categorical / quantitative with 2 outcomes.

e.g. y_i = High BP / low BP.
= voted yes / no.
= bought stocks / didn't
= took a drug / didn't

$$y_i = \begin{cases} 1 & \text{w/ prob. } \pi_i \\ 0 & \text{w/ prob. } 1-\pi_i \end{cases}$$

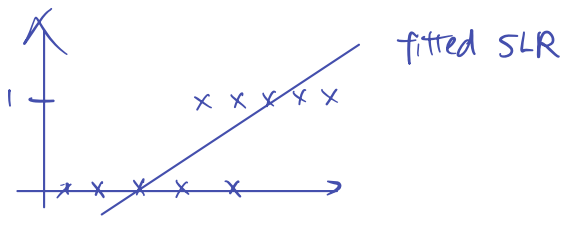
$y_i \sim \text{Bernoulli}(\pi_i)$

If use SLR: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$E y_i = \beta_0 + \beta_1 x_i = 1(\pi_i) + 0(1-\pi_i) = \pi_i$

Has to be in $[0, 1]$

For all x_i depending on x_i, this does not happen.



E.g. X = alcohol use

y^c = duration of pregnancy (continuous resp. variable)

SLR: $y_i^c = \beta_0^c + \beta_1^c x_i + \varepsilon_i^c$, $\varepsilon_i^c \sim N(0, \sigma_c^2)$

Dichotomization: $y_i = \begin{cases} 1, & y_i^c \le 38 \text{ weeks} \quad \text{preterm} \\ 0, & y_i^c > 38 \text{ weeks} \quad \text{full term.} \end{cases}$

$$\pi_i = P(y_i = 1) = P(y_i^c \le 38) = P(\beta_0^c + \beta_1^c x_i + \varepsilon_i^c \le 38)$$
$$= P\left(\frac{\varepsilon_i^c}{\sigma_c} \le \underbrace{\frac{38-\beta_0^c}{\sigma_c}}_{\beta_0^*} - \underbrace{\frac{\beta_1^c}{\sigma_c}}_{\beta_1^*} x_i\right)$$

$\varepsilon_i^c \sim N(0, \sigma_c^2)$

$z = \frac{\varepsilon_i^c}{\sigma_i} \sim N(0,1)$

$$= P(z \le \beta_0^* + \beta_1^* x_i)$$
$$= \Phi(\beta_0^* + \beta_1^* x_i)$$

$$E(y_i) = \pi_i = \underbrace{\Phi(\beta_0^* + \beta_1^* x_i)}$$

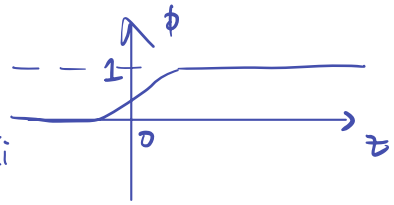
Nonlinear response func.: $f(x_i, \beta) = \Phi(\beta_0^* + \beta_1^* x_i)$

Probit mean func.

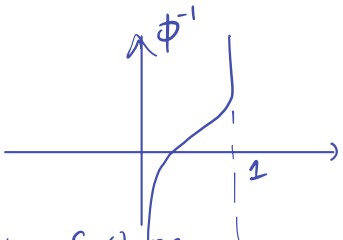
$\pi_i = \Phi(\beta_0^* + \beta_1^* x_i)$

$\pi_i' = \Phi^{-1}(\pi_i) = \beta_0^* + \beta_1^* x_i$

probit — probit transform



$\Phi^{-1}: [0,1] \to (-\infty, \infty)$



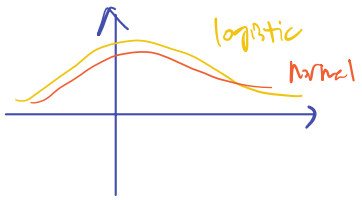
In general, $\Phi(\beta_0^* + \beta_1^* x_i)$ has a S shape.

If $\beta_1^* \uparrow$, If $\beta_1^* > 0$, $\beta_1^* < 0$,

$y_i' = 1 - y_i \Rightarrow \beta_0' = -\beta_0^*$, $\beta_1' = -\beta_1^*$

Logit Mean Response

Logistic R.V.: $f_L(\varepsilon) = \frac{\exp(\varepsilon)}{[1+\exp(\varepsilon)]^2}$ $\quad E(\varepsilon) = 0$, $\mathrm{var}(\varepsilon) = \pi/\sqrt{3}$



modify error term: $\varepsilon_i \sim N(0, \sigma^2)$

$\varepsilon_{L,i} \sim \text{logistic}(\text{mean} = 0, \text{var} = \sigma_L^2)$

$$E(y_i) = \pi_i = P(y_i = 1) = P(y_i^c \le 38)$$
$$= P(\varepsilon_i^L \le 38 - \beta_0 - \beta_1 x_i)$$
$$= P\left(\frac{\pi \varepsilon_i^L}{\sqrt{3}\sigma_c} \le \beta_0^* + \beta_1^* x_i\right)$$
$$= P(\varepsilon \le \beta_0^* + \beta_1^* x_i)$$
$$= F_L(\beta_0^* + \beta_1^* x_i)$$ CDF of logistic R.V.

$\pi_i = F_L(\beta_0^* + \beta_1^* x_i)$

$\pi_i' = F_L^{-1}(\pi_i) = \beta_0^* + \beta_1^* x_i$.

$$F_L(\varepsilon) = \frac{1}{1+\exp(-\varepsilon)} = \exp(\varepsilon)/1+\exp(\varepsilon)$$ (also S-shaped)

☆ Lecture 21 2023年11月15日(三)。

Logistic Regression

probit mean response func.

$\mathbb{E}y_i = \pi_i = \Phi(\beta_0 + \beta_1 X_i)$

↳ squashes $\beta_0+\beta_1 X_i$ into $[0,1]$

$\pi_i' = \Phi^{-1}(\pi_i) = \beta_0 + \beta_1 X_i$

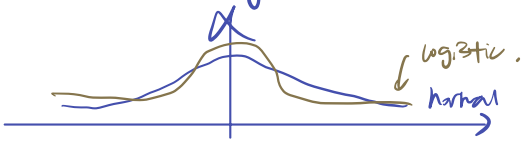
↳ stretches $(0,1)$ into $[-\infty, +\infty]$

$y_i \sim \text{Bernoulli}(\pi_i)$

$y_i = \begin{cases} 1 & \text{w.p. } \pi_i \\ 0 & \text{w.p. } (1-\pi_i) \end{cases}$

Logit mean response func.

errors ~ logistic R.V.'s



$\mathbb{E}y_i = \pi_i = F_L(\beta_0 + \beta_1 X_i)$

$\pi_i' = F_L^{-1}(\pi_i) = \beta_0 + \beta_1 X_i$

$F_L(y) = \frac{1}{1+\exp(-y)}$ = cdf of logistic. R.V.

$\pi_i' = F_L^{-1}(\pi_i) = \log \frac{\pi_i}{1-\pi_i}$ (log odds)

$\text{odds} = \frac{P(S)}{P(F)} = \frac{\pi_i}{1-\pi_i}$

If $\pi_i = \frac{1}{2} \Rightarrow \text{odds} = 1$

If $\pi_i = \frac{1}{10} \Rightarrow \text{odds} = \frac{\frac{1}{10}}{1-\frac{1}{10}} = \frac{1}{9}$

another option for reference: error terms ~ gumbel R.V.

$\pi_i = \log(-\log(1-\pi_i)) = \beta_0 + \beta_1 X_i$ (3rd most common)

⇒ most common: Logistic regression.

→ Simple logistic regression.

$\mathbb{E}y_i = \pi_i = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)} \Leftrightarrow \pi_i' = \beta_0 + \beta_1 X_i$

Coefficients are estimated via MLE:

$P(y_i = 1) = \pi_i$; $P(y_i = 0) = 1-\pi_i$

$P(y_i = y) = \pi_i^{y}(1-\pi_i)^{1-y}$

$L(y_1, \dots, y_n \mid \beta_0, \beta_1) = \text{joint pmf of } y_1, \dots, y_n = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

($y_1, \dots, y_n$: given obsns. ; β_0, β_1: unknowns)

↓ take the log

$\ell(y_1, \dots, y_n \mid \beta_0, \beta_1) = \log L = \sum y_i \log \pi_i + (1-y_i)\log(1-\pi_i)$

$= \sum_{i=1}^{n} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)$

① ②

① log odds $= \pi_i' = \beta_0 + \beta_1 X_i$

$= \sum_{i=1}^{n} \{ y_i (\beta_0 + \beta_1 X_i) - \log(1 + \exp(\beta_0 + \beta_1 X_i)) \}$

→ Solve numerically to find maximizers. $b_0 = \hat{\beta}_0^{MLE}$, $b_1 = \hat{\beta}_1^{MLE}$

$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)}$ fitted value for the ith obsn

✡✡ Interpretation of b_1

$\pi_i' = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

$\hat{\pi}_i'(X_i) = b_0 + b_1 X_i$

$\hat{\pi}_i'(X_i+1) = b_0 + b_1 X_i + b_1$

$\Rightarrow$ $b_1 = \hat{\pi}_i'(X_i+1) - \hat{\pi}_i'(X_i)$

$b_1 = \hat{\pi}_i'(2) - \hat{\pi}_i'(1)$

$= \log(\text{odds}_2) - \log(\text{odds}_1)$

$= \log \frac{\text{odds}_2}{\text{odds}_1}$

$\text{odds} = \frac{\pi_i}{1-\pi_i}$

$\text{odds ratio} = \widehat{OR} = \frac{\text{odds}(X_i+1)}{\text{odds}(X_i)} = e^{b_1}$

Increase of predictor value by 1 leads to increase of odds by $\exp(b_1)$.

$b_1 = 0.1615 \Rightarrow \exp(b_1) = 1.175$

As X increases by 1, odds of y are increased by 17.5%.

✡ Multiple logistic Regression.

$E y_i = \pi_i = \frac{\exp(X^T\beta)}{1+\exp(X^T\beta)} = (1+\exp(-X^T\beta))^{-1}$

$\pi' = X^T\beta$ → vector of predictors w/ intercept.

↓ n×p, p+1 predictors

✡ Inference (chp. 14.5)

we only have large sample results for $s^2(b_1)$ via theory of MLE.

$s^2(b) \approx (-g_{ij})_{\beta=b}^{-1}$ where $g_{ij} = \frac{\partial^2 \log L}{\partial\beta_i \partial\beta_j}$ Hessian of log-likelihood

✡ Wald-Test (t-test)

$H_0: \beta_k = 0$ vs. $H_a: \beta_k \neq 0$.

$z^* = \frac{b_k}{s(b_k)}$

CI for odds ratio: $OR = e^{\beta_k}$

$\exp(\underbrace{b_k \pm t(1-\frac{\alpha}{2})\, s(b_k)}_{\text{CI for } \beta_k})$

$|z|^* = \left|\frac{b_k}{s(b_k)}\right| > t(1-\frac{\alpha}{2})$, rej. H_0

$\leq t(1-\frac{\alpha}{2})$, accept H_0.

Mean response X_h

$\hat{\pi}_h' \pm t(1-\frac{\alpha}{2}) \sqrt{X_h^T S^2(b) X_h} = [L, U]$

↘ sd of mean response at X_h.

For π_h : $(L^*, U^*]$ where $\boxed{L^* = (1+e^{-L})^{-1}, \quad U^* = (1+e^{-U})^{-1}}$ $\quad \pi_h = (1+e^{-\pi_h'})^{-1}$

$X_h = [1 \;\; X_{h1} \;\; X_{h2} \;\; \cdots \;\; X_{h,p-1}]$

★ Polychotomous Logistic Regression (Ch. 14.11, p608–611) Multinomial

As opposed to dichotomous (binomial)

Response variable y_i = qualitative / categorical with > 2 categories

Eg. : y_i = product A / OR / product B / OR / product C) nominal

nominal

ordinal - there is a natural ordering to classes.

y_i = mild illness / moderate illness / severe illness) ordinal.

$y_{i1} = \begin{cases} 1 & \text{if cat 1} \\ 0 & \text{o.w.} \end{cases}$

$y_{i2} = \begin{cases} 1 & \text{if cat 2} \\ 0 & \text{o.w.} \end{cases}$

$y_{i3} = \begin{cases} 1 & \text{if cat 3.} \\ 0 & \text{o.w.} \end{cases}$

$\sum_{j=1}^{3} y_{ij} = y_{i1} + y_{i2} + y_{i3} = 1$

one hot encoding.

! Not dummy coding .!

$\pi_{ij} = P(y_{ij} = 1)$ = " prob. that cat j is selected in ith response"

One can show that :

$\hat{\pi}_{ij} = \dfrac{\exp(X_i^T b_j)}{1 + \sum_{k=1}^{J-1} \exp(X_i^T b_k)}$ where J = # of categories y_i can assume.

we skipped category J.

Note: If $J=2 \Rightarrow$ regular logistic regression.

☆ Lecture 22 2023年11月27日(一)

☆ Poisson Regression (Ch. 14.13, pp. 618–620)

y_i = count of "rare" events

y_i = # of trips family takes to mall in 1 month (1)

y_i = # of hospitalizations in a year (2)

X_i = # of members, income, distance, ... (1)

X_i = preexisting conditions, age, more (2)

$y \sim \text{Poisson}(\mu)$

p.m.f. $f(y) = \frac{\mu^y \exp(-\mu)}{y!}$, $y = 0, 1, 2, 3, \ldots$ $\quad \mathbb{E}y = \mu$, $\text{var}(y) = \mu$.

$\text{Poisson}(\mu) \approx \text{Binomial}(n, p)$ as $n \to \infty$, $p \to 0$.

$\mu = np$. "rare events"

☆ Poisson regression

y_i are independent Poisson r.v.s with parameter μ_i (= expected value μ_i)

$\mu_i = \mu(X_i, \beta)$ — X_i: vector of predictors; β: vector of params.

(☆) 3 options for $\mu(X_i, \beta)$:

$\mu(X_i, \beta) = X_i^T \beta$

$\mu(X_i, \beta) = \exp(X_i^T \beta)$

$\mu(X_i, \beta) = \log(X_i^T \beta)$

β is estimated via MLE:

$L(\beta) = \prod_{i=1}^n \frac{\mu(X_i, \beta)^{y_i} \exp(-\mu(X_i, \beta))}{y_i!}$

$\ell(\beta) = \log L(\beta) = \sum_i y_i \log(\mu(X_i, \beta)) - \sum_i \mu(X_i, \beta) - \sum_i \log(y_i!)$

↓

maximize to find b!

Fitted values $\hat{\mu}_i = \mu(X_i, b)$

☆ General Linear Models (GLiM) Ch. 14.15.

(bern., normal, poisson, exp, ...)

1) $y_1, \ldots, y_n$ are resp. var's coming from exponential family of dist'n's

2) A linear predictor $X_i^T \beta = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$

3) A link function g → monotonic & differentiable.

$X_i^T \beta = g(\mu_i)$ mean of y_i, $\mathbb{E}(y_i) = \mu_i$

$\text{var}(y_i) = \sigma_i^2$ (nonconstant) but $\sigma_i^2 = f(\mu_i) = f(g^{-1}(X_i^T \beta))$

E.g. ·) Normal error model $g(\mu_i) = \mu_i$

$X_i^T \beta = g(\mu_i) = \mu_i \Rightarrow X_i^T \beta = \mu_i \to \mathbb{E}y_i = X_i^T \beta$

·) Logistic Regression: $g = F_L^{-1}(\mu_i)$ → cdf of logistic R.V.

$X_i^T \beta = g(\mu_i) \Rightarrow X_i^T \beta = F_L^{-1}(\mu_i) = F_L^{-1}(\pi_i)$

$$\Rightarrow X_i^T\beta = \pi_i' = \frac{\log \pi_i}{1-\pi_i}$$

· Poisson regression: $g(\mu_i) = \log(\mu_i)$

$X_i^T\beta = \log(\mu_i) \Rightarrow \mu_i = \exp(X_i^T\beta)$

☆ One factor ANOVA (chp. 16.)

Factor: predictor variable = X qualitative / categorical

levels of factor = treatments = values of X. $X_1, X_2, \dots, X_r$

E.g. factor = dosage of drug.

① treatments = low, medium, high.

y = effectiveness of 3 dosages of drug to reduce blood pressure.

② factor := brand

treatments = B_1, B_2, B_3, B_4.

y = absorbency of paper towel.

y_{ij} = response of the jth replicate of ith treatment

treat 1	treat 2	treat 3
$y_{11}\ y_{12} \cdots y_{1n_1}$	$y_{21}\ y_{22} \cdots y_{2n_2}$	$y_{31}\ y_{32} \cdots y_{3n_3}$

n_i = # of replicates in treatment i

total # of observations = $n_T = \sum_i n_i$

☆ Cell means model for ANOVA (CMM)

$y_{ij} = \mu_i + \varepsilon_{ij}$ → response variable in jth replicate / trial for ith factor level or treatment.

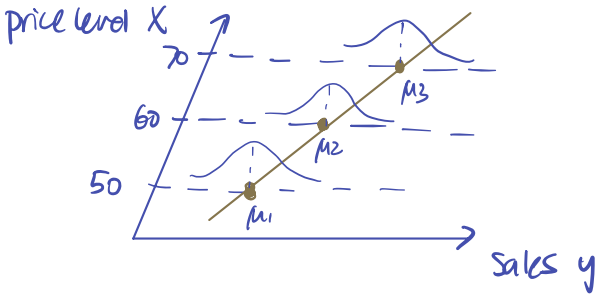
μ_i = parameters

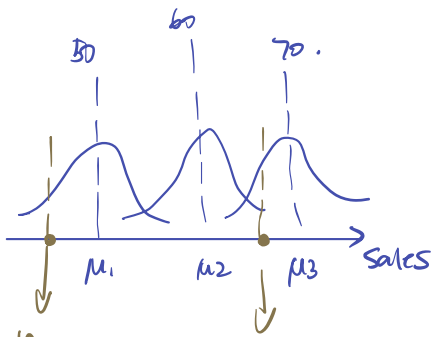
$\varepsilon_{ij} \sim$ indep. $N(0, \sigma^2)$

$i = 1, \dots, r$ # of factor levels / treatments

$j = 1, \dots, n_i$ # of observations in treatment i

$\sum_i n_i = n_T$ total # observations.





$y_{11} = \mu_1 + \varepsilon_{11}$ $y_{31} = \mu_3 + \varepsilon_{31}$.

$y_{ij} \sim N(\mu_i, \sigma^2)$, independent.

CELL Means Model

ANOVA as a linear model

treatments $r=3$, $n_1 = n_2 = n_3 = 2$

$$\underline{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} \quad \underline{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \quad \Rightarrow y_{ij} = \mu_i + \epsilon_{ij} \Leftrightarrow \underline{y} = \underline{X}\underline{\beta} + \underline{\epsilon}.$$

$$\underline{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \underline{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix}$$

$$\mathbb{E}\underline{y} = \underline{X}\underline{\beta} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_3 \\ \mu_3 \end{bmatrix}$$

μ_i = population mean for treatment i

σ^2 = "variability" of responses in population.

$\text{Var}\,\underline{y} = \text{Var}\,\underline{\epsilon} = \sigma^2 \underline{I}_6$ ↪ 6×6 identity matrix.

What if we just did linear regression w/ dummy coding?

$$X_{ij1} \begin{cases} 1, & \text{if } T_1 \\ 0, & \text{o.w.} \end{cases} \qquad X_{ij2} \begin{cases} 1, & \text{if } T_2 \\ 0, & \text{o.w.} \end{cases}$$

defining dummy coding

$$y_{ij} = \beta_0 + \beta_1 X_{i,j1} + \beta_2 X_{i,j2} + \epsilon_{ij}$$

$\mathbb{E}y_{j1} = \beta_0 + \beta_1$ (if treatment 1) measures effect of treatment 1.

$\mathbb{E}y_{j2} = \beta_0 + \beta_2$

Regression approach to ANOVA or effects model.

CMM → cell means μ_1, μ_2, μ_3.

We study only fixed effects ANOVA: factor levels are fixed / predetermined.

We do NOT study random effects ANOVA.

factor levels → chosen randomly from some population.

Notation:

$y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$ = sum of observations in ith treatment

$\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n_i} y_{i\cdot}$ = avg. of obsns in ith treatment.

$y_{\cdot\cdot} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} y_{ij}$

$\bar{y}_{\cdot\cdot} = \frac{1}{n_T} y_{\cdot\cdot} = \sum_{i=1}^{r} \frac{n_i}{n_T} \bar{y}_{i\cdot}$ ↪ weighted sum of treatment averages.

Overall mean

→ Estimate of cell means (LS, MLE). ★Lecture 13.

1) LS.

$Q(\mu_1,\dots,\mu_r) = \sum_i \sum_j (y_{ij}-\mu_i)^2$ → objective function.

$= \sum_j (y_{1j}-\mu_1)^2 + \sum_j (y_{2j}-\mu_2)^2 + \cdots + \sum_j (y_{rj}-\mu_r)^2$

minimize wrt cell means μ_i

$\hat{\mu}_i^{LS} = \bar{y}_{i\cdot}$ fitted values $\hat{y}_{ij} = \bar{y}_{i\cdot}$

In vector notation, $\hat{y} = X b$, where $b = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{bmatrix}$

2) MLE

$L(\mu_1,\dots,\mu_r,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2} \cdot \sum_{i,j} (y_{ij}-\mu_i)^2\right)$

maximize $L \rightarrow \hat{\mu}_i^{MLE} = \hat{\mu}_i^{LS} = \bar{y}_{i\cdot}$ for $i = 1,\dots,r$.

Residuals: $e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot}$ $\sum_{j=1}^{n_i} e_{ij} = 0$. for $i = 1,\dots,r$

In vector notation: $e = y - Xb$.

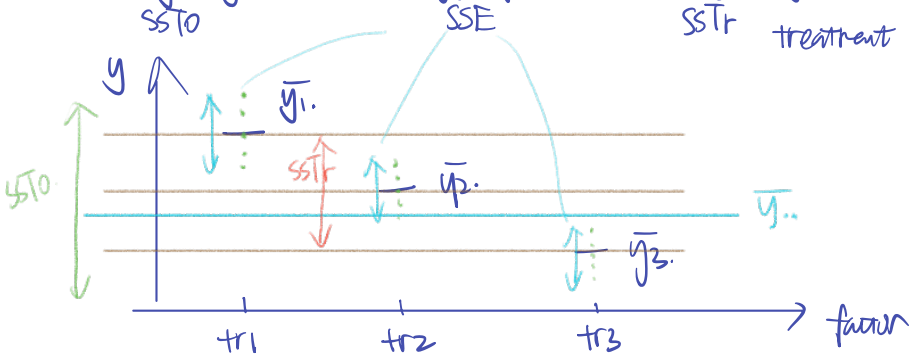
ANOVA identity: $y_{ij} - \bar{y}_{\cdot\cdot} = (y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$

total deviation ↓ deviation of estimated treatment means around overall mean (btw Deviation).

deviation around estimated treatment mean (within deviation)

$\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_i n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$

SSTO SSE SSTr treatment



ANOVA Table:

SS	df	MS	E(MS)
SSE	$n_T - r$	$MSE = SSE/(n_T - r)$	σ^2 → unbiased estimator of variance.
SSTr	$r-1$	$MSTr = SSTr/(r-1)$	$\sigma^2 + \sum_i \frac{n_i(\mu_i-\mu_0)^2}{(r-1)}$ → where $\mu_0 = \sum_{i=1}^r \frac{n_i}{n_T}\mu_i$ weighted mean.
SSTO	$n_T - 1$		

Scenario 1): $SSTr \gg SSE$ → treatment works. i.e. μ_i are different.

Scenario 2): $SSTr \ll SSE$ → treatment prob. doesn't work, i.e. μ_i are "same"

$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$, i.e. treatments don't work.
H_a: at least one μ_i is different, so treatment works.
under H_0, $SSE/\sigma^2 \sim \chi^2_{n_T - r}$
$SSTr/\sigma^2 \sim \chi^2_{r-1}$

$$\frac{SSTr/\sigma^2/(r-1)}{SSE/\sigma^2/(n_T - r)} = \frac{MSTr}{MSE} \sim F(r-1, n_T - r) \text{ under } H_0. \qquad F^* = \frac{MSTr}{MSE}.$$

if $F^* > F(1-\alpha, r-1, n_T - r)$ conclude H_a
between group diff. ($SSTr$) >> within group difference (SSE).
if $F^* \le F(1-\alpha, r-1, n_T - r)$ conclude H_0.

Intuition: $E(MSE) = \sigma^2$
$E(MSTr) = \sigma^2 + \frac{\sum n_i(\mu_i - \mu_0)^2}{r}$ ≈ 0 on average when we conclude H_0.

under H_0, $MSE \approx MSTr$ on average
under H_a, $MSE < MSTr$ on average
in M.L.R, F^* was MSR/MSE

★ Factor effects model (Ch. 16.7)
Recall, in cell means model $y_{ij} = \mu_i + \varepsilon_{ij}$ Factor effect model: (ε_{ij} iid. $N(0, \sigma^2)$)

$$\mu_i = \mu_0 + (\mu_i - \mu_0) = \mu_0 + \tau_i$$ → ith factor level effect / ith treatment effect.
(μ_0: constant component of all obsn's)

So $y_{ij} = \mu_0 + \tau_i + \varepsilon_{ij}$, $i = 1, \ldots, r$, $j = 1, \ldots, n_i$

★ Definition of μ_0 = unweighted mean.
$\mu_0 = \frac{1}{r}\sum_{i=1}^{r} \mu_i \Rightarrow \sum_{i=1}^{r} \tau_i = 0$

e.g. ($r = 3$): $\frac{1}{3}(\underset{\mu_1}{70} + \underset{\mu_2}{60} + \underset{\mu_3}{53}) = 61$

$\tau_1 = 9$, $\tau_2 = -1$, $\tau_3 = -8$
→ $\tau_1 + \tau_2 + \tau_3 = 0$.
→ treatment 1 increases the mean of obsn's from that group by 9 units.

★ weighted mean $\mu_0 = \sum_{i=1}^{r} w_i \mu_i$, $\sum_i w_i = 1$
$\Rightarrow \sum_i w_i \tau_i = 0$.
Choose w_i: according to sample size: $w_i = n_i/n_T \rightarrow \sum_i n_i/n_T = \frac{1}{n_T}\sum_i n_i = n_T/n_T = 1$.
OR
according to the importance of the treatments to the experimenter.

In CRM,

H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$

H_a: not all μ_i are equal

In factor effects model,

H_0: $\tau_1 = \tau_2 = \cdots = \tau_r = 0$.

H_a: not all $\tau_i = 0$.

vs.

"treatments have no effect on response variable."

★ Lecture 24 2023年12月4日(一).

Use factor ANOVA

regression approach to ANOVA

multiple inferences for factor level means.

★ Regression Approach to ANOVA (Ch. 16.8).

3 models: (A) Factor effects with unweighted mean.

(B) " " " weighted "

(C) Cell means model.

$y = X\beta + \varepsilon$ ← (A), (B), (C)

? ?

(A) Factor effects w/ unweighted mean

$y_{ij} = \mu_0 + \tau_i + \varepsilon_{ij}$, $i = 1, \ldots, r$

(μ_0: common; τ_i: effect of treat i)

Need to estimate $(r+1)$ params. μ_0 & τ_i, $i = 1, \ldots, r$

only r are independent: $\sum_i \tau_i = 0$

$\Rightarrow \tau_r = -\tau_1 - \tau_2 - \cdots - \tau_{r-1}$

e.g. $r = 3$; $n_1 = n_2 = n_3 = 2$

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} \quad \beta = \begin{bmatrix} \mu_0 \\ \tau_1 \\ \tau_2 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

$\mu_0 - \tau_1 - \tau_2 = \mu_0 + \tau_3$

$y = X\beta + \varepsilon \Leftrightarrow y_{ij} = \mu_0 + \tau_i + \varepsilon_{ij}$

$$E\{y\} = X\beta = \begin{bmatrix} \mu_0 + \tau_1 \\ \mu_0 + \tau_1 \\ \mu_0 + \tau_2 \\ \mu_0 + \tau_2 \\ \mu_0 - \tau_1 - \tau_2 \\ \mu_0 - \tau_1 - \tau_2 \end{bmatrix}$$

$$y_{ij} = \mu_0 + \tau_1 X_{ij1} + \tau_2 X_{ij2} + \cdots + \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij}$$

μ_0: intercept; $\tau_1, \tau_2, \ldots, \tau_{r-1}$: slope cells.

$y_{rj} = \mu_0 - \tau_1 - \tau_2 \cdots - \tau_{r-1} + \varepsilon_{rj} = \mu_0 + \tau_r + \varepsilon_{rj}$.

$X_{ij1} = \begin{cases} 1 & \text{if treat } 1 \\ -1 & \text{if treat } r \cdots \\ 0 & \text{o.w.} \end{cases}$ $X_{ij(r-1)} = \begin{cases} 1 & \text{if treat } r-1 \\ -1 & \text{if treat } r \\ 0 & \text{o.w.} \end{cases}$

"effect" coding

L.S. estimates: $\hat{\mu}_0 = \frac{1}{r}\sum_{i=1}^{r} \bar{y}_{i\cdot}$, $\hat{\tau}_i = \bar{y}_{i\cdot} - \hat{\mu}_0$

β_1 β_2 β_{r-1}

$H_0: \tau_1 = \tau_2 = \cdots = \tau_{r-1} = 0$ → Can test this using the overall model F test.
H_a: at least one $\tau_i \neq 0$.

$F^* = MSR/MSE$

(B) Factor effects with weighted mean.

$\mu_0 = \sum_i w_i \mu_i = \sum_i \frac{n_i}{n_T} \mu_i$, $n_T = \sum_i n_i$

↳ weighted by sample size.

$\sum_{i=1}^{r} \frac{n_i}{n_T} \tau_i = 0 \Rightarrow \tau_r = -\frac{n_1}{n_r}\tau_1 - \frac{n_2}{n_r}\tau_2 - \cdots - \frac{n_{r-1}}{n_r}\tau_{r-1}$

$y_{ij} = \mu_0 + \tau_1 X_{ij1} + \tau_2 X_{ij2} + \cdots + \tau_{r-1} X_{ij(r-1)}$

$X_{ij1} = \begin{cases} 1 & \text{if treat } 1 \\ -n_1/n_r & \text{if treat } r \\ 0 & \text{o.w.} \end{cases}$ $\cdots$ $X_{ij(r-1)} = \begin{cases} 1 & \text{if treat } (r-1) \\ -n_{r-1}/n_r & \text{if treat } r \\ 0 & \text{o.w.} \end{cases}$

Note: if $n_1 = n_2 = \cdots = n_r$, then we get (A).

e.g. if $r=4$, $n_1 = n_2 = n_4 = 5$, $n_3 = 4$.

$X_{ij1} = \begin{cases} 1 & \text{if treat } 1 \\ -5/5 & \text{if treat } 4 \\ 0 & \text{o.w.} \end{cases}$, $X_{ij2} = \begin{cases} 1 & \text{if treat } 2 \\ -5/5 & \text{if treat } 4 \\ 0 & \text{o.w.} \end{cases}$, $X_{ij3} = \begin{cases} 1 & \text{if treat } 3. \\ -4/5 & \text{if treat } 4. \\ 0 & \text{o.w.} \end{cases}$

$H_0: \tau_1 = \tau_2 = \cdots = \tau_{r-1} = 0.$
H_a: at least one $\tau_i \neq 0$. $F^* = MSR/MSE.$

(C) Cell means model.

$y_{ij} = \mu_i + \epsilon_{ij} \Leftrightarrow \underline{y} = X\beta + \underline{\epsilon}$

$\beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{bmatrix}$, $X = \begin{bmatrix} \mathbb{1}_{n_1} & \underline{0}_{n_1} & \cdots & \underline{0}_{n_1} \\ \underline{0}_{n_2} & \mathbb{1}_{n_2} & \underline{0}_{n_2} \cdots & \underline{0}_{n_2} \\ \vdots & & & \vdots \\ \underline{0}_{n_r} & \cdots & & \underline{0}_{n_r}\ \mathbb{1}_{n_r} \end{bmatrix}$ (★)

where $\mathbb{1}_{n_1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n_1 \times 1}$ $\underline{0}_{n_1} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{n_1 \times 0}$

Eg.: Consider $r=3$ & n_1, n_2, n_3

$b = (X^TX)^{-1}X^Ty = \begin{bmatrix}\bar{y}_{1\cdot}\\ \bar{y}_{2\cdot}\\ \bar{y}_{3\cdot}\end{bmatrix}$ (☆)

$X_1 = \begin{cases}1 & \text{if treat 1}\\ 0 & \text{o.w.}\end{cases}$ ⋯ $X_r = \begin{cases}1 & \text{if treat } r\\ 0 & \text{o.w.}\end{cases}$

no intercept. !

$y_{ij} = \mu_1 X_{ij1} + \mu_2 X_{ij2} + \cdots + \mu_r X_{ijr} + \varepsilon_{ij}$ (☆☆)

$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$

$H_a:$ at least one μ_i is $\neq$ than the others.

NOT the same as the overall model fit. test.

use general linear test.

H_0: reduced model $y_{ij} = \mu_c + \varepsilon_{ij}$.

H_a: Full model (☆☆)

Reduced: $y = X\beta + \varepsilon$, $X = \begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}$, $\beta = [\mu_c]$

$F^* = \dfrac{SSE(R) - SSE(F)}{SSE(F)} \cdot \dfrac{df(F)}{df(R) - df(F)}$ ← $df(F) = n_T - r$

$n_T - 1 - (n_T - r) = r - 1$

$= MSTr / MSE$

☆ Inference for factor level means

CMM: $y_{ij} = \mu_i + \varepsilon_{ij}$

Before, we could conduct: $H_0: \mu_1 = \mu_2 = \cdots = \mu_r$

vs.

$H_a:$ not all μ_i are $=$.

Which $\mu_i \neq \mu_j$?

OR What's the CI for μ_i?

OR What's the CI for $(\mu_i - \mu_j)$?

☆ Estimation / testing for single factor level mean μ_i

L.S.: $\hat{\mu}_i = \bar{y}_{i\cdot}$ $\quad y_{ij} \sim N(\mu_i, \sigma^2)$, indep.

$\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \Rightarrow E(\bar{y}_{i\cdot}) = \mu_i$

$Var(\bar{y}_{i\cdot}) = \sigma^2/n_i$ $\quad s^2(\bar{y}_{i\cdot}) = MSE/n_i$

$\bar{y}_{i\cdot} \sim N(\mu_i, \sigma^2/n_i)$ $\quad (\bar{y}_{i\cdot} - \mu_i)/s(\bar{y}_{i\cdot}) \sim t(n_T - r)$

dfs $= n_T - r$

because $SSE/\sigma^2 \sim \chi^2_{n_T - r}$

$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$ ↳ r

(!) We use full sample n_T to estimate σ^2.

$\star\star$ $CI_{(1-\alpha)}$ For $\mu_i = \bar{y}_{i\cdot} \pm t(1-\alpha/2, n_T - r)\, s(\bar{y}_{i\cdot})$, $i = 1, \dots, r$

Hypothesis Test: $H_0: \mu_i = c$, vs. $H_a: \mu_i \neq c$.

$$t^* = \frac{\bar{y}_{i\cdot} - c}{s(\bar{y}_{i\cdot})} \sim t(n_T - r)$$

If $|t^*| \leq t(1-\alpha/2; n_T - r) \Rightarrow H_0$

" $>$ " $\Rightarrow H_a$.

$\star$ Estimation / Testing for difference of 2 factor level means.

Define $D = \mu_i - \mu_j \Rightarrow \hat{D} = \hat{\mu}_i - \hat{\mu}_j = \bar{y}_{i\cdot} - \bar{y}_{j\cdot}$

$$E(\hat{D}) = E(\bar{y}_{i\cdot}) - E(\bar{y}_{j\cdot}) = \mu_i - \mu_j = D$$

$$Var(\hat{D}) = Var(\bar{y}_{i\cdot}) + Var(\bar{y}_{j\cdot}) - 2cov(\bar{y}_{i\cdot}, \bar{y}_{j\cdot}) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)$$

$= 0$ because obsn's in ith cell $\perp\!\!\!\perp$ obsn's in jth cell.

$$s^2(\hat{D}) = MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)$$

$$\Rightarrow \frac{\hat{D} - D}{s(\hat{D})} \sim t(n_T - r)$$

$CI_{(1-\alpha)}$ For $D = \hat{D} \pm t(1-\alpha/2, n_T - r)\, s(\hat{D})$

$\hat{D} = (\bar{y}_{i\cdot} - \bar{y}_{j\cdot})$

$s(\hat{D}) = \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

Hypothesis Test: $H_0: D = 0$ $(\mu_i = \mu_j)$ vs. $H_a: D \neq 0$ $(\mu_i \neq \mu_j)$

$$t^* = \hat{D} / s(\hat{D}) \sim t(n_T - r)$$

Lecture 25 12/06 (F)

Multiple comparisons

Say $r=3$. Test 1: $H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 \neq \mu_2$.
Test 2: $H_0: \mu_1 = \mu_3$ vs. $H_a: \mu_1 \neq \mu_3$.
Test 3: $H_0: \mu_2 = \mu_3$ vs. $H_a: \mu_2 \neq \mu_3$.

Each test conducted at $\alpha = 0.05$. FTR = "failing to reject"
$P(\text{FTR } H_0 \text{ in } T_1 \mid H_0 \text{ is true}) = 1-\alpha = 0.95$
$P(\text{FTR } H_0 \text{ in } T_1 \& T_2 \& T_3 \mid H_0 \text{ is true}) \cong 0.95^3 = 0.857$
$P(\text{rejecting at least once in } T_i\text{'s} \mid H_0) = 1 - 0.857 = 0.143 = 14\%$

Type I error for 3 tests = "family-wise" error (FWE).
$\Rightarrow$ FWE $\leq \alpha$ (this is what we want)

→ Tukey multiple comparison procedure. (= "Honestly Significant Differences") = HSV in R.

Defn: Studentized Range Distribution (SRD).

Consider i.i.d. sample $y_1, \ldots, y_r \sim N(\mu, \sigma^2)$
$W = \max_i(y_i) - \min_i(y_i)$ = range.
Estimate σ^2 by s^2 with $\nu \to \infty$
Define new r.v.: $q(r, \nu) = W/S$ Where r = # obsn's in samples
ν = # dfs of s^2
$q(r, \nu)$ = studentized range.

Percentiles of $q(r, \nu)$ are in Table B.9 (KNNL)
e.g. $P(\frac{W}{S} = q(5, 10) \leq 4.65) = 0.95$
↳ in R. $q(0.95, 5, 10)$
$1-\alpha$ r ν

Simultaneous CI for all possible differences.

$\hat{D} \pm T \cdot S(\hat{D})$, $\hat{D} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot}$, $S(\hat{D}) = \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$, $T = \frac{1}{\sqrt{2}} q(1-\alpha, r, n_T - r)$

↳ $\frac{r(r-1)}{2}$ CIs if we have r factor levels.

$\frac{\hat{D}}{S(\hat{D})} = \frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{S(\hat{D})} \leq \frac{\max_i(\bar{y}_{i\cdot}) - \min_i(\bar{y}_{i\cdot})}{S(\hat{D})}$ worst case scenario.

H.T.: (Simultaneous): $H_0: \mu_i = \mu_j$ vs. $H_a: \mu_i \neq \mu_j$ for all pairs.

$q^* = \frac{\sqrt{2}\hat{D}}{S(\hat{D})}$

$|q^*| \leq q(1-\alpha, r, n_T - r)$, H_0.
overall type I error $\leq \alpha$.

Bonferroni $CI_{1-\alpha}$ (simultaneous)

$\hat{D} \pm B \cdot S(\hat{D})$, $B = t(1 - \alpha/2g, n_T - r)$, g = # comparisons.
if all pairwise comparisons: $g = \frac{r(r-1)}{2}$
↳ Tukey is better, less conservative.

If $g < \frac{k(k-1)}{2}$, then Bonf. could be better than Tukey.

☆☆ Two factor ANOVA ("Two way ANOVA")

Case where y depends on 2 categorical variables.

- y = sales
- pred. var 1 = factor A, e.g. price level ($50, $60, $70) # levels = a.
- pred. var 2 = factor B, e.g. marketing (IG, FB) campaign # levels = b.

total number of treatments: ab.

We shouldn't do one-factor ANOVA: we fail to access the joint effects of both factors A & B.

factor A \ factor B	$j=1$	$j=2$
$i=1$	$y_{111}, y_{112}, \dots$	$y_{121}, y_{122}, \dots$
$i=2$	$y_{211}, y_{212}, \dots$	$y_{221}, y_{222}, \dots$
$i=3$	$y_{311}, y_{312}, \dots$	$y_{321}, y_{322}, \dots$

CMM: $y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$

Effects: $y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$.

- α_i: main effect of factor A at level i
- β_j: main effect of factor B at level j
- $(\alpha\beta)_{ij}$: new term c (not the multiplication) interaction of A at level i & B at level j