#3 - F2017

Review Linear Regression

Outline

- Linear regression with normality assumption
- Estimation (LSE and MLE); BLUE
- Inference on parameter estimators
- Diagnostics
- Many others (Please review them on your own).

Additional texts:

1. McCullagh and Nelder (1990), Generalized Linear Models, 2nd edt.
2. Hosmer and Lemeshow (2005), Applied Logistic Regression, 2nd Edt
3. Agresti (2002), Categorical Data Analysis, 2nd Edt

Introduction

- Y_i: response variable (dependent variable, outcome) for the ith subject. **RANDOM!**
- $\mathbf{X}_i^T = (X_{i1}, X_{i2}, \quad, X_{ip})$: known values of p predictor variables (independent variable, covariate) for the ith subject (Very often consider $X_{i1} \equiv 1$, intercept) for all i.
- i: index of subject
- n: total number of subjects in the data
- Independent

* Note: Very often $\mathbf{X}_i$ is assumed deterministic (not random). In general, it can be assumed random too. Later we will see it does not make any difference in parameter estimation if there is no missing data.

Model

- Linearly relate predictors to the mean response (assume X is deterministic).
  - "Linear" in parameter β.
- For ith subject,

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \quad + \beta_p X_{ip} + \epsilon = \mathbf{X}_i^T \beta + \epsilon_i,$$

where $X_{i1} \equiv 1$ (intercept), $\beta^T = (\beta_1, \quad, \beta_p)$ and $\epsilon \sim N(0, \sigma^2)$.
In matrix form with all n subjects

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \ \ \epsilon \sim MVN_n(\mathbf{0}, \sigma^2 I).$$

i.e.

$$\mathbf{Y} \sim MVN_n(\mathbf{X}\beta, \sigma^2 I).$$

with

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\right\}$$

Interpretation

- Basic idea (in simple linear regression)

$$Y_i = \beta_1 + \beta_2 X_{i2} + \epsilon_i,$$

  - β_2: change in mean response per unit increase of X_2.

$$\beta_2 = \frac{E[Y_i|X_{i2} = x+1] - E[Y_i|X_{i2} = x]}{(x+1) - x} \equiv \frac{\partial E[Y_i|X_{i2}]}{\partial X_{i2}}.$$

- Multiple linear regression (no interaction term): need to adjust other covariates (or holding them constant).
- MLR with interaction terms: look at which covariates are involved in interaction terms.

Estimation

- Need to estimate β.
- Least Square Estimation (LSE): minimizing the sum of squared errors (nothing to do with distributional assumption)

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta).$$

- Maximum Likelihood Estimation (MLE): maximizing the likelihood (or log likelihood) (it needs distributional assumption)

$$\max_{\beta} L(\beta; \mathbf{Y}) \quad \text{or} \quad \max_{\beta} \log L(\beta; \mathbf{Y}).$$

- Here, the MLE and LSE of β are the same when the error distribution is $N(0, \sigma^2)$.

Estimation (continued)

2 Then we show $E(\hat\sigma^2)=\sigma^2$, $\hat\sigma^2$ is an unbiased estimator of σ^2

- The OLS estimator of β is

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

where $\mathbf{X}$ is of full rank.

  - $\hat{\beta}$ is unbiased for β.
  - The variance of $\hat{\beta}$ is $V(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

- An unbiased estimator of σ^2 is,

$$\hat{\sigma}^2 = \frac{1}{n-p}SSE = \frac{1}{n-p}\mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

- Residual: estimated error

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta}.$$

1. First show $SSE = Y^T(I-H)Y$, where $H = X(X^TX)^{-1}X^T$

we see $H^2 = H$ $H = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$

Then $SSE = \sum_{i=1}^n (y_i - x_i^T\beta)^2 = (Y - X^T\hat\beta)^T(Y - X^T\hat\beta) = e^Te$

Recall $\hat\beta = (X^TX)^{-1}X^TY$, we have $e = Y - X^T\hat\beta = Y - X(X^TX)^{-1}X^TY = (I - X(X^TX)^{-1}X^T)Y$,

so $e^Te = Y^T(I - X(X^TX)^{-1}X^T)^T(I - X(X^TX)X^T)Y$

$= Y^T(I - X(X^TX)^{-1}X^T)^T(I - X(X^TX)^{-1}X^T)Y$

$= Y^T(I-H)^T(I-H)Y$

$= Y^T(I-H)Y$. #

Since $Y = X\beta + \varepsilon$,

$Y^T(I-H)Y = (\beta^TX^T + \varepsilon^T)(I-H)(X\beta+\varepsilon)$

$= \beta^TX^T(I-H)X\beta + \varepsilon^T(I-H)\varepsilon + \boxed{2\beta^TX^T(I-H)}\varepsilon$,

By $E[(2\beta^TX^T)(I-H)\varepsilon] = 0$,

$E(\varepsilon^T(I-H)\varepsilon) = [\mathrm{trace}(I-H)]\sigma^2 = (n - \mathrm{trace}\,H)\sigma^2 = (n-p)\sigma^2$,

$X^T(I-H)X = X^TX - X^THX = X^TX - X^TX = 0$,

it is seen that

$E[Y^T(I-H)Y] = (n-p)\sigma^2$. Hence,

$E(\hat\sigma^2) = \frac{1}{n-p}E[Y^T(I-H)Y] = \sigma^2$ #

 Note: $H^2 = H$, $(I-H)^2 = I-H$, H is idempotent.

- Analysis of Variance
  - SST: total variation of $\mathbf{Y}$ around mean
  - SSR: variation of $\mathbf{Y}$ explained by regression
  - SSE: variation of $\mathbf{Y}$ unexplained by regression

$$SST = SSR + SSE.$$

$SST = \sum_{i=1}^n (y_i - \bar y)^2$,

$SSR = \sum_{i=1}^n (\hat y_i - \bar y)^2$,

$SSE = \sum_{i=1}^n (y_i - \hat y_i)^2$

$R^2 = \frac{SSR}{SST}$,

R^2 is called the multiple coefficient of determination

Equivalent to

To show $\sum_{i=1}^n (Y_i - \bar Y)^2 = \sum_{i=1}^n (\hat Y_i - \bar Y)^2 + \sum_{i=1}^n (y_i - \hat y_i)^2$

Since $\sum_{i=1}^n (y_i - \bar y)^2 = \sum_{i=1}^n (\hat y_i - \bar y + y_i - \hat y_i)^2$

$= \sum_i (\hat y_i - \bar y)^2 + 2\sum_i (\hat y_i - \bar y)(y_i - \hat y_i) + \sum_i (y_i - \hat y_i)^2$ (*)

$\hat y_i = x_i^T\hat\beta$, $\hat Y = X\hat\beta = HY$, $Y - \hat Y = (I-H)Y$,

Let $L = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$, then

$\sum_i (\hat y_i - \bar y)(y_i - \hat y_i) = (\hat Y - \frac{1}{n}LL^TY)^T(Y - \hat Y)$

$= [\hat Y^T - \frac{1}{n}Y^TLL^T](Y - \hat Y)$

$= [Y^TH - \frac{1}{n}Y^TLL^T](I-H)Y$

$= Y^T\underbrace{H(I-H)}_{0}Y - \frac{1}{n}Y^TL\underbrace{L^T(I-H)}_{0}Y$ (See proofs on right)

$= Y^T0Y - \frac{1}{n}Y^T0Y = 0$, therefore, by (*), $\sum_{i=1}^n (y_i - \bar y)^2 = \sum_{i=1}^n (\hat y_i - \bar y)^2 + \sum_{i=1}^n (y_i - \hat y_i)^2$.

Proofs of $H(I-H) = 0$ and $LL^T(I-H) = 0$

Since $H^2 = H$, then $H(I-H) = H - H^2 = 0$.

Since $X^T(I-H) = X^T - X^TH = X^T - X^T = 0$, when there is an intercept, the first row of X^T is L^T, which implies $L^T(I-H) = 0$.

Note: When there is an intercept, $\sum_{i=1}^n \hat\varepsilon_i = L^T(I-H)Y = 0$.

GLSE and WLSE

- Consider $\mathbf{Y} \sim MVN_n(\mathbf{X}\beta, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ arbitrary positive definite symmetric var-cov matrix.
  - There exists a non-singular matrix $\mathbf{\Psi}$ s.t. $\mathbf{\Sigma} = \mathbf{\Psi}\mathbf{\Psi}^\mathbf{T}$
  - Then, $\mathbf{\Psi}^{-1}\mathbf{Y} = \mathbf{\Psi}^{-1}\mathbf{X}\beta + \mathbf{\Psi}^{-1}\epsilon$.
  - Set $\mathbf{Z} = \mathbf{\Psi}^{-1}\mathbf{Y}$ and $\mathbf{W} = \mathbf{\Psi}^{-1}\mathbf{X}$,

$$\mathbf{Z} = \mathbf{W}\beta + \epsilon^*, \;\; \epsilon^* \sim MVN(\mathbf{0}, \mathbf{I}).$$

  - Then, the OLS estimator of β based on $\mathbf{Z}$ and $\mathbf{W}$ is

$$\hat{\beta}_G \equiv (\mathbf{W}^\mathbf{T}\mathbf{W})^{-1}\mathbf{W}^\mathbf{T}\mathbf{Z} = (\mathbf{X}^\mathbf{T}\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{\Sigma}^{-1}\mathbf{Y},$$

  which is called the Generalized LSE (GLSE) of β.
- GLS can obviously handle correlations among Y_i's.
- It also can handle independent Y_is with unequal variances (WLSE).

we have

$W^T Z = (\Psi^{-1} Z)^T \Psi^{-1} Y$

$= Z^T (\Psi \Psi^T)^{-1} Y$

$= Z^T (\Sigma)^{-1} Y$

$(W^T W)^{-1}$

$= \left((\Psi^{-1} X)^T (\Psi^{-1} X) \right)^{-1}$

$= (X^T (\Psi \Psi^T)^{-1} X)^{-1}$

$= (X^T \Sigma^{-1} X)^{-1}$

BLUE: Gauss-Markov Theorem

- **Gauss-Markov Theorem**: Consider a r.v. $\mathbf{Y}$ with $V(\mathbf{Y}) = \sigma^2\mathbf{I}$. Let $\mathbf{T} = \mathbf{A}\mathbf{Y}$ be an unbiased estimator of β. Then the OLS estimator $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of β (i.e. $V(\hat{\beta}) \leq V(\mathbf{T})$).
  - Generalized LSE and Weighted LSE are also BLUE.
- Note: The BLUE property of the LSE (OLS, GLS and WLS) is not dependent upon any particular distribution for the Y_i's. However if we want to make inference we need some distribution assumption.

Review Matrix Calculus

- Derivative of an inner product $x^T a = a_1 x_1 + \cdots + a_n x_n$, $= a^T x$, then $\frac{\partial}{\partial x}(x^T a) = \frac{\partial}{\partial x}(a^T x) = a$
- Let A be a $q \times c$ matrix, then $\frac{\partial}{\partial x}(x^T A) = A$.
- Let B be a $r \times q$ matrix, then $\frac{\partial}{\partial x}(Bx) = B^T$
- Derivative of a quadratic form, then $\frac{\partial}{\partial x}(x^T A x) = Ax + A^T x$. If A is symmetric, $\frac{\partial}{\partial x}(x^T A x) = 2Ax$.

Maximum Likelihood Estimation (MLE)

- Suppose Y_i's are jointly normal and independent, i.e., $Y_i \sim N(X_i^T\beta, \sigma_i^2)$, for $i = 1, \quad, n$. i.e.

$$\mathbf{Y} \sim MVN_n(\mathbf{X}\beta, \mathbf{V}),$$

where $\mathbf{V} = \text{diag}(\sigma_1^2, \quad, \sigma_n^2)$. The likelihood function for β is

$$L(\beta|\mathbf{Y}) = \frac{1}{(2\pi)^{n/2}|\mathbf{V}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta)\right\}$$

determinant

then the MLE is $\hat{\beta}_{MLE} = (X^T V X)^{-1} X^T V^{-1} Y = B_V Y$

$$\ell(\beta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V| - \frac{1}{2}(Y - X\beta)^T V^{-1}(Y - X\beta)$$

$$\frac{\partial \ell(\beta)}{\partial (Y - X\beta)} = -V^{-1}(Y - X\beta) \qquad \frac{\partial \ell(\beta)}{\partial \beta} = \frac{\partial (Y - X\beta)}{\partial \beta}\frac{\partial \ell(\beta)}{\partial (Y - X\beta)} = -X(-V^{-1})(Y - X\beta)$$

$$\frac{\partial \ell(\beta)}{\partial \beta} = X^T V^{-1}(Y - X\beta) = 0$$

Therefore, $\hat{\beta}_{MLE} = (X^T V X)^{-1} X^T V^{-1} Y$.

- Review: $Cov(AY, BZ) = A\, Cov(Y, Z) B^T$

- If $\mathbf{Y} \sim MVN_n(\mathbf{X}\beta, \mathbf{V})$, the distribution of the MLE $\hat{\beta}$ is

$$\begin{aligned} V(\hat{\beta}) &= Cov(\hat{\beta}, \hat{\beta}) = B_V\, Cov(Y, Y) B_V^T \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1}\, V\, V^{-1} X (X V^{-1} X)^{-1} \\ &= (X^T V^{-1} X)^{-1} \end{aligned}$$

$$\hat{\beta}_{MLE} \sim MVN_n(\beta, (X^{-1} V^{-1} X)^{-1})$$

- If $\mathbf{Y} \sim MVN_n(\mathbf{X}\beta, \mathbf{V})$, the distribution of the OLE $\hat{\beta}_{OLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is

$$V(\hat{\beta}_{OLE}) = B\; V(Y) B^T = (X^T X)^{-1} X^T V X (X^T X)^{-1}$$

$$\hat{\beta}_{OLE} \sim MVN_n(\beta, V(\hat{\beta}_{OLE}))$$

Testing

- Consider $\mathbf{Y} \sim MVN(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Interest in testing H_0 $\mathbf{C}\beta = \mathbf{c}$ (most often $\mathbf{c} = \mathbf{0}$), where $\mathbf{C}$ is of rank r Then under H_0, $\mathbb{C} \in \mathbb{R}^{r,p}$, $rank(\mathbb{C}) = r$

$$\mathbf{C}\hat{\beta} - \mathbf{c} \sim MVN_r(0, \sigma^2 \mathbb{C}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{C}^T),\quad \mathbb{c} \in \mathbb{R}^r$$

H_0 is called general linear hypothesis (GLH).

so,

$$\frac{(\mathbf{C}\hat{\beta} - \mathbf{c})^T(\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T)^{-1}(\mathbf{C}\hat{\beta} - \mathbf{c})}{\sigma^2} \sim \chi^2_r.$$

  - Test statistic: under the normality and H_0

$$F = \frac{(\mathbb{C}\hat{\beta} - \mathbb{c})^T(\mathbb{C}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{C}^T)^{-1}(\mathbb{C}\hat{\beta} - \mathbb{c})/r}{SSE/(n-p)} \sim F_{r,\,n-p},$$

where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbb{Y}^T(I - H)\mathbb{Y}$,

Reject H_0 at level α if $F_{calc} > F_\alpha(r, n-p)$ (upper quantile),

where p = # parameters (including intercept β_0).

$\hat{\sigma}^2 = SSE/(n-p)$ is an unbiased estimator of σ^2.

Note $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, not unbiased

Testing (continued)

- For $\mathbf{c} = \mathbf{0}$:

  i. If $\mathbf{C} = \mathbf{a}$, a single vector with $a_j = 0, j \neq k$ and $a_j = 1, j = k$. Then it is a single covariate test, which is exactly same as the t-test for $\beta_k = 0$,

$$t = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \sim t_{n-p},$$

$SE(\hat{\beta}_k) = s\sqrt{(\mathbb{X}^T\mathbb{X})^{-1}_{kk}}$

$$F_k = \frac{(\hat{\beta}_k)^2}{s^2(\mathbb{X}^T\mathbb{X})^{-1}_{kk}}$$

  and $t^2 = F_k$ where $F_k = \frac{(\hat{\beta}_k)^2}{s^2(\mathbf{X}^T\mathbf{X})^{-1}_{kk}} \sim F_{1,n-p}$.

  ii. If $\mathbf{C} = \text{diag}(0, 1, 1, \quad, 1)$. Then it is an overall F test (i.e. $\beta_2 = \quad = \beta_p = 0$).

  iii. If $\mathbf{C} = \text{diag}(0, a_2, a_3, \quad, a_p)$ with $a_j = 0$ or $a_j = 1$. Then it is a test for subset of covariates (i.e. some of β_j are significant).

  iv. Otherwise, it is a general linear hypothesis testing (GLH).

Testing (continued)

- We can use "full model" and "reduced model (under H_0)"

$$F = \frac{\{SSR(\text{full model}) - SSR(\text{reduced model})\}/\Delta df R}{SSE(\text{full model})/df E(\text{full model})} \sim F_{\Delta df R, df E}.$$

or

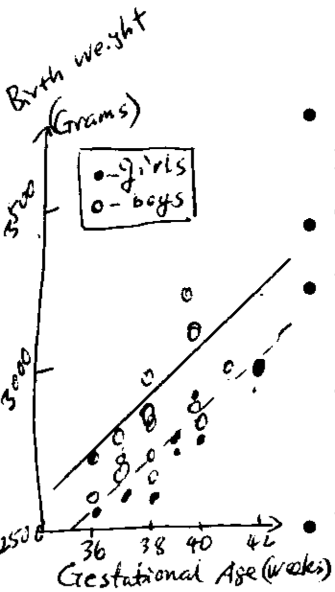
$$F = \frac{\{SSE(\text{reduced model}) - SSE(\text{full model})\}/\Delta df E}{SSE(\text{full model})/df E(\text{full model})} \sim F_{\Delta df E, df E}.$$

  - Exactly same result as previous!

Note: $SSE(\text{Reduced model under } H_0) - SSE(\text{Full model})$
$= (C\hat{\beta} - c)^T [C (X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - c)$

$\Delta df\ R = \Delta df\ E$

Example: Birthweight Data



- D&B, Table 2.3 on page 24. Birthweight and gestational age. See `Figure2_3and4 sas`. and Table 2-4 and 5 sas and Table 2-5. R
- Whether the rate of increase of birthweight is the same for boys and girls?
- The mean birthweight for boys is greater than that for girls. There is linear increasing trend, the girls tend to weigh less than the boys of the same gestational age. A general linear model:

$$E(Y_{ik}) = \mu_{jk} = \alpha_j + \beta_j x_{jk}. \; j = 1, \quad , J, \; k = 1, \quad , K.$$

$\beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix}$

- The question of interest can be formulated as

$$H_0: \; \beta_1 = \beta_2 = \beta, \; \text{(two lines are parallel).}$$

$C = (0, 0, 1, -1)$

$C\beta = 0$

$\Longleftrightarrow \beta_1 = \beta_2, \; \text{rank}(C) = 1$

versus

$$H_1 \quad \beta_1 \neq \beta_2.$$

- There are two possible models:
  - Model 0: (H_0 is true): $E(Y_{ik}) = \mu_{jk} = \alpha_j + \beta x_{jk}$, $Y_{jk} \sim N(\mu_{jk}, \sigma^2)$. → Common slope
  - Model 1. (H_1 is true): $E(Y_{ik}) = \mu_{jk} = \alpha_j + \beta_j x_{jk}$, $Y_{jk} \sim N(\mu_{jk}, \sigma^2)$. → different slope
- Model 1 is a full model, Model 0 is a reduced model. The F-test statistic for H_0 is

$\hat{S}_0 = SSE(\text{Reduced Model})$, $\hat{S}_1 = SSE(\text{Full Model})$

$$F = \frac{(\hat{S}_0 - \hat{S}_1)/\sigma^2}{J-1} \quad \frac{\hat{S}_1/\sigma^2}{JK-2J} \sim F_{(J-1, JK-2J)} = F_{1,20}.$$

Note: here $J = 2$ and $K = 12$. $J - 1 = \{JK - (\underset{\alpha_j}{J} + \underset{\beta}{1})\} - \{JK - \underset{\alpha_j+\beta_j}{2J}\}$. df:

$J-1 = \text{dfE} = \text{dfR}$, Since Full model has J slops and Reduced model has 1 slope

- From regression ANOVA table, see Table 2.5 on page 27 we have

$$F_{calc} = \frac{(658770.8 - 652424.5)/1}{652424.5/20} = 0.19.$$

Hence, p-value$=Pr(F_{1,20} > 0.19) = 0.6676$.

- The large p-value indicates that the data do not provide enough evidence against the hypothesis H_0: $\beta_1 = \beta_2$. Thus, Model 0 is preferable.

Response: Aptitude (ability to do something).

- The analysis is also called ANCOVA Analysis of Covariance (the covariates X's have both categorical and continuous variables).

ANCOVA Table: Table 6.14, Page 116. 3rd Edt. Textbook.

	Source	df	SS	MS	F	P
Model(6.14) ←	mean and covariates ($\mu + \gamma x_{jk}$)	2	853.766			
Residual for Red. model(6.14) ←	Factor levels (μ_j)	2	16.932	8.466	13.97	
Residual for Full model (6.13) ←	Residual	17	10.302	0.606		

Model (6.14) has 2 param. (μ, γ), Model (6.13) has 4 reg. parameters (μ_j, γ)

- See Table 6.13 on page 115 for another example. Data are shown in `Table6_12 sas`, 1st Edt.
- In this example, we want to compare three training methods, taking into account differences in initial aptitude (x_{jk}) between the three groups of subjects (μ_j).
- To test the hypothesis that there are no differences in mean achievement scores among the three training methods, after adjustment for initial aptitude. Let's compare the saturated (full) model

$$E(Y_{jk}) = \mu_j + \gamma x_{jk} \quad j = 1, 2, 3, \; k = 1, \quad , 7$$

with the reduced model

$$E(Y_{jk}) = \mu + \gamma x_{jk}.$$

The SAS program `Table6_12 sas` produces the results shown in Table 6.14 on page 116. (For plot and Full model) Table6-12-Reduced.sas for Reduced model (6.14)

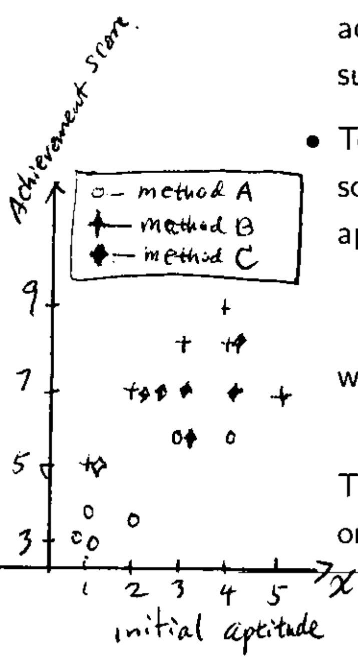


Fig 6.2

Diagnostics: Violation of Assumptions

- Assumptions:
  - Linearity: $E[\mathbf{Y}] = \mathbf{X}\beta$ with $\epsilon = \mathbf{Y} - \mathbf{X}\beta$.
  - Normality.
  - Equal variance (homoscedasticity).
  - independence:

$$\epsilon \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}).$$

STAT 635-GLM-Lecture Notes 3, Review Linear Regression, Fall 2017

Diagnostics: Violation of Assumptions

- Diagnostic (using residuals):
  - Linearity:
    - Check: Partial regression plot, residual plot, LOF test
    - Remedy: Transformation, GLM.
  - Normality:
    - Check: Normal probability plot, Shapiro-Wilks Test
    - Remedy: Transformation, GLM.
  - Equal variance:
    - Check: Residual plot
    - Remedy: Transformation, WLSE, GLM.
  - Independence:
    - Check: Done by intuition (e.g., repeatedly measured..)
    - Remedy: GLSE, Time series, longitudinal analysis.

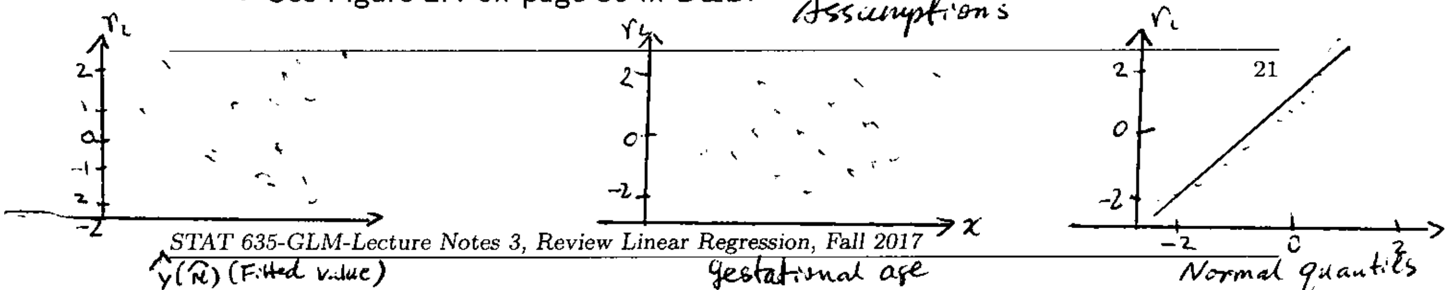
Model Checking: Birthweight Data

h_{ii} is the ith element on the diagonal of the projection or hat matrix $H = X(X^TX)^{-1}X^T$

- Residual plots:
  - (Studentized) standardized residuals (see the textbook, p.93):

$$r_i = (y_i - \hat{\mu}_i)/\{\hat{\sigma}(1-h_{ii})^{1/2}\} \sim N(0,1).$$

  - Residuals vs. fitted values $\hat{y}_i = \mu_i$ to detect changes in variance.
  - Residuals vs. existing explanatory variables or other potential explanatory variables to check apparent pattern in the plot, for example, linearity of relationships between variables, and associations with other potential explanatory variables.
  - Ordered residuals vs. their expected values (Normal quantiles) (Q-Q plot or normal probability plot) to assess the normality assumption.
- See Figure 2.4 on page 30 in D&B. Residual plots to check model Assumptions



Notation and Coding for Explanatory Variables

- In linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $\mathbf{X}$ is often called the **design matrix**, and $\mathbf{X}\beta$ is the **linear component** of the model.
- Various ways of defining the elements of $\mathbf{X}$ are illustrated in the following examples. If some of X are categorical variables

Example: Simple Linear Regression for Two Groups

- In the birthweight data, the model is

$$E(Y_{jk}) = \mu_{jk} = \alpha_j + \beta_j x_{jk}; \;\; Y_{jk} \sim N(\mu_{jk}, \sigma^2).$$

Then use two different intercepts

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \\ Y_{1K} \\ Y_{21} \\ \\ Y_{2K} \end{bmatrix}, \beta = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & 0 & x_{11} & 0 \\ 1 & 0 & x_{12} & 0 \\ & & & \\ 1 & 0 & x_{1K} & 0 \\ 0 & 1 & 0 & x_{21} \\ & & & \\ 0 & 1 & 0 & x_{2K} \end{bmatrix},$$

Example: Comparing the Means of Two Groups

- There are several alternative ways of formulating the linear components for comparing means of two groups: $Y_{11}, \quad , Y_{1K_1}$ and $Y_{21}, \quad , Y_{2K_2}$

(a). $E(Y_{1k}) = \beta_1$, and $E(Y_{2k}) = \beta_2$. In this case, $\beta = (\beta_1, \beta_2)^T$ and the rows of $\mathbf{X}$ are Called cell treatment model

Group 1. [1 0]

Group 2: [0 1]

(b). $E(Y_{1k}) = \mu + \alpha_1$, and $E(Y_{2k}) = \mu + \alpha_2$. In this version, μ represents the overall mean and α_1 and α_2 are differences from μ. Here $\beta = (\mu, \alpha_1, \alpha_2)^T$ and the rows of $\mathbf{X}$ are Called effects model

Group 1. [1 1 0] Rank(X) = 2

Group 2: [1 0 1]

However there are too many parameters as only two parameters can be estimated.

(c). $E(Y_{1k}) = \mu$, and $E(Y_{2k}) = \mu + \alpha$. This is equivalent to (b) subject to constraint $\alpha_1 = 0$ and $\alpha_2 = \alpha$. For this version $\beta = (\mu, \alpha)^T$ and the rows of $\mathbf{X}$ are

Group 1. [1 0]

Group 2: [1 1]

Group 1 is a reference category called the "corner point" and this is an example of **corner point parameterization**.

(d). $E(Y_{1k}) = \mu + \alpha$, and $E(Y_{2k}) = \mu - \alpha$. This is equivalent to (b) subject to constraint $\alpha_1 + \alpha_2 = 0$ and $\alpha_1 = \alpha$. For this version $\beta = (\mu, \alpha)^T$ and the rows of $\mathbf{X}$ are $\alpha_2 = -\alpha_1 = -\alpha$

μ α

Group 1. [1 1]

Group 2: [1 -1]

This is an example of **sum-to-zero** constraint.

- **HW**· Different software uses different constraints. Check out what constraints SAS and R use.