STAT 331 Notes

Thomas Liu

December 3, 2024

Contents

1 Introduction to Regression **4**
- 1.1 Regreession Analysis . . . 4
  - 1.1.1 Sample v.s. Population . . . 4

2 Simple Linear Regression **4**
  - 2.0.1 Population Model . . . 4
  - 2.0.2 Observe Sample . . . 5
- 2.1 Lease Square Estimation (LSE) . . . 5
  - 2.1.1 Person Correlation Coefficient . . . 6
  - 2.1.2 Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$ . . . 6
  - 2.1.3 Properties of c_i . . . 6
  - 2.1.4 Properties of r_i . . . 6
- 2.2 The Estimation of σ^2 . . . 6
- 2.3 Confidence Interval and Hypothesis Testing . . . 7
  - 2.3.1 Confidence Interval . . . 7
  - 2.3.2 Hypothesis Testing . . . 7
- 2.4 Inference of $\mu_0 = \beta_0 + \beta_1 x_0$ for some Predictor value x_0 . . . 8
- 2.5 Prediction of Future Value . . . 8
- 2.6 Analysis of Variance (ANOVA) for Testing $H_0 : \beta_1 = 0$ . . . 8
  - 2.6.1 Properties under $H_0 : \beta_1 = 0$ . . . 9
- 2.7 F-Statistic . . . 9
- 2.8 ANOVA Table . . . 9
- 2.9 Cochram's Theorem . . . 9
- 2.10 Coefficient of Determination . . . 9

3 Random Vector and Linear Regression **10**
- 3.1 Multivariate Normal Distribution . . . 11
  - 3.1.1 Properties . . . 12
- 3.2 Multiple Linear Regression (MLR) . . . 12
  - 3.2.1 Matrix Form Representation . . . 12
- 3.3 LSE of $\vec{\beta}$ . . . 13

3.3.1 Properties of $\hat{\vec{\beta}}$ . . . 13
3.4 Results of MLR . . . 13
3.4.1 Results of $\hat{\vec{\beta}}$ . . . 14
3.4.2 Confidence Interval . . . 14
3.5 Linear Combination of β_i's . . . 14
3.6 Prediction of y . . . 15
3.7 Analysis of Variance (ANOVA) . . . 15
3.7.1 ANOVA Table . . . 16
3.7.2 Total Coefficients of Determination . . . 16
3.8 Geometric Interpolation of LSE . . . 16
3.8.1 Column Space of X . . . 16
3.9 Test Linear Constraints . . . 16
3.9.1 Additional Sum of Squares Principle . . . 16
3.9.2 Summary . . . 17

4 Regression Model Specification **17**
4.1 Special Cases . . . 17
4.1.1 Piecewise Constants . . . 17
4.1.2 Piecewise Linear . . . 18
4.1.3 Piecewise Linear but Continuous . . . 18
4.1.4 One Sample Problem . . . 18
4.1.5 Two Sample Problem . . . 18
4.1.6 K-sample Problem . . . 19
4.1.7 ANOVA Table . . . 20

5 Model Checking **21**
5.1 Model Checking . . . 21
5.1.1 Studentized Residual . . . 21
5.1.2 Residual Plots for Checking $E(\epsilon_i) = 0$ . . . 22
5.1.3 Residuals vs x_j . . . 22
5.1.4 Residuals vs $\hat{y}$ . . . 23
5.1.5 Studentized Residuals vs $\hat{y}$ . . . 23
5.1.6 Residual Plots for Checking Variance $V(\epsilon_i) = \sigma^2$ . . . 23
5.1.7 Durbin-Waston Test . . . 24
5.1.8 Q-Q Plot . . . 24
5.2 Leverage . . . 24
5.3 Cook's Distance . . . 25
5.4 PRESS Residuals . . . 25

6 Model Selection **25**
6.1 Akaike's Information Criterion (AIC) . . . 25
6.2 Bayesian Information Criterion (BIC) . . . 26
6.3 Note . . . 26

6.4 Backward Elimination with p-value 26
6.5 Forward Selection with p-value 26
6.6 Variance Stablizing Transformation 26
6.7 Linear Dependency / Multicollinearity 27
6.7.1 Perfect Multicollinearity 27
6.7.2 Multicollinearity 27
6.7.3 Detection of Multicollinearity 28
6.8 Variance Inflation Factor 28

1 Introduction to Regression

1.1 Regreession Analysis

A statistical methodology that models the functional relationship between response variable y and one or more explanatory variables $x_1, x_2, \cdots, x_p$

$$y = f(x_1, x_2, \cdots, x_p) + \epsilon$$

- y: dependent / response variable
- $x_1, \cdots, x_p$: covariates, explanatory / independent variables, predictors
- ϵ: random error term

In this course, we focus on simplest form of regression: linear models

$$\begin{aligned} y &= f(x_1, \cdots, x_p) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \end{aligned}$$

β's are the regression parameters (coefficients)
We check if the model is linear by checking the derivative with respect to β

1.1.1 Sample v.s. Population

- sample: collection of units (people, animals, cities, fields) that is actually measured or surveyed in study
- population: large group of units we are interested in, which sample was selected

2 Simple Linear Regression

2.0.1 Population Model

$$y = \beta_0 + \beta_1 x + \epsilon$$

- y is response
- β_0, β_1 are regression coefficients
- x is predictor
- $\beta_0 + \beta_1 x$ is systematic component
- ϵ is random error

2.0.2 Observe Sample

Suppose we have n pairs of observations $(x_i, y_i), i = 1, 2, \cdots$, then

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- x_i's: fixed, known
- β's: fixed, unknown
- ϵ_i's: random, unknown
- y_i's: random, known

We usually make the following assumption

- $E(\epsilon_i) = 0$
- $\epsilon_1, \cdots, \epsilon_n$ are statistically independent
- $Var(\epsilon_i) = \sigma^2$, $Var(y_i) = \sigma^2$
- ϵ_i is normally distributed $\rightarrow \epsilon_i \sim N(0, \sigma^2)$, which means $y_i \sim N(\beta_0 + \beta_1 x_i, sigma^2)$

Assumption $i), ii), iii)$ are called Gauss-Markov assumptions

2.1 Lease Square Estimation (LSE)

Given $(x_i, y_i), i = 1, \cdots, n$, estimate (β_0, β_1) as $(\hat{\beta_0}, \hat{\beta_1})$ such that value of

$$r_i = y_i - \hat{\beta_0} - \hat{\beta_1} x_i = y_i - \hat{y_i}$$

- r_i is residual
- $\hat{y_i}$ is fitted value

want r_i to be small
Define

$$s(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1 x_i})^2 = \sum_{i=1}^{n} r_i^2$$

Want to minimize $s(\beta_0, \beta_1)$, so we have two normal equations

- $\hat{\beta_0} = \overline{y} - \hat{\beta_1} \overline{x}$
- $\hat{\beta_1} = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \dfrac{S_{xy}}{S_{xx}}$

2.1.1 Person Correlation Coefficient

$$P_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

2.1.2 Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

- LSE are unbiased, which means $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$
- $Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}})$, $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$
- $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \dfrac{\overline{x}}{S_{xx}}$

2.1.3 Properties of c_i

Let c_i be $\dfrac{x_i - \overline{x}}{S_{xx}}$

- $\sum_{i=1}^{n} c_i = 0$
- $\sum_{i=1}^{n} c_i x_i = 1$
- $\sum_{i=1}^{n} c_i^2 = \dfrac{1}{S_{xx}}$

2.1.4 Properties of r_i

Under least square fit

- $\sum_{i=1}^{n} r_i = 0$
- $\sum_{i=1}^{n} r_i x_i = 0$
- $\sum_{i=1}^{n} r_i \hat{y} = 0$
- The point $(\overline{x}, \overline{y})$ is always on the fitted regression line

2.2 The Estimation of σ^2

Notice that

$$\begin{cases} \epsilon_i = y_i - (\beta_0 + \beta_1 x_i) \\ r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \end{cases}$$

recall $\epsilon_i \sim N(0, \sigma^2)$

If ϵ_i's are known, we can estimate σ^2 as $\dfrac{1}{n} \sum_{i=1}^{n} \epsilon_i$

However, $E(\dfrac{1}{n} \sum_{i=1}^{n} r_i^2) = \sigma^2$

We define $s^2 = \dfrac{1}{n-2}\sum_{i=1}^n r_i^2$, then $E(s^2) = \sigma^2$. Note $(n-2)$ because we estimate both β_0, β_1
Recall if $y_i \sim N(\mu, \sigma^2)$, sample variance estimator is

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

and

$$E(\hat{\sigma}^2) = \sigma^2$$

2.3 Confidence Interval and Hypothesis Testing

2.3.1 Confidence Interval

Under the assumption that ϵ_i are independent and normally distributed, we have

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

If σ^2 is known,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

As we may not know σ^2, replace σ^2 with s^2

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{s^2}{S_{xx}}}} \sim t_{n-2}$$

And $100(1-\alpha)\%$ confidence interval for β_1 is

$$P_r(-t_{n-2,\alpha/2}) < \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} < t_{n-2,\alpha/2}$$

or

$$P_r(\hat{\beta}_1 - t_{n-2,\alpha/2} se(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + t_{n-2,\alpha/2} se(\hat{\beta}_1)) = 1 - \alpha$$

where $se(\hat{\beta}_1) = \sqrt{\dfrac{s^2}{S_{xx}}}$

2.3.2 Hypothesis Testing

We have $H_0 : \beta_1 = \beta_1^*$ vs $H_a : \beta_1 \neq \beta_1^*$ Under H_0,

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)} \sim t_{n-2}$$

If $|t| = |\frac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)}| >= t_{n-2,\alpha/2}$, we reject H_0 at the significance level α
Alternatively, we compute the p-value

$$p = P(|T| \geq |t|) \quad T \sim t_{n-2}$$

and rejct H_0 if $p \leq \alpha$

2.4 Inference of $\mu_0 = \beta_0 + \beta_1 x_0$ for some Predictor value x_0

To estimate μ_0, compute $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$
We get $\hat{\mu}_0 = \sum_{i=1}^n d_i y_i$ where $d_i = \frac{1}{n} + \frac{(x_0 + \overline{x})(x_i - \overline{x})}{S_{xx}}$

- $E(\hat{\mu_0}) = \mu_0$
- $Var(\hat{\mu_0}) = [\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}]\sigma^2$

2.5 Prediction of Future Value

Q: What's the prediction of y given that $x = x_p$?
We use the existing data point for the model, and use $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$ to predict

Some Result for $\hat{y}_p$

- $E(y_p - \hat{y}_p) = 0$
- $Var(y_p - \hat{y}_p) = [1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{S_{xx}}]\sigma^2 = Var(\hat{\mu}_p) + Var(\epsilon_p)$
- $\frac{y_p - \hat{y}_p}{se(y_p - \hat{y}_p)} \sim t_{n-2}$ where $se(y_p - \hat{y}_p) = \sqrt{[1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{S_{xx}}]s^2}$

The $100(1-\alpha)\%$ prediction interval for y_p is $\hat{y}_p \pm t_{n-2,\frac{\alpha}{2}} se(y_p - \hat{y}_p)$

2.6 Analysis of Variance (ANOVA) for Testing $H_0 : \beta_1 = 0$

The total variation of y_i's is measured by the total sum of squaress (SST)

$$\begin{aligned} SST &= \sum_{i=1}^n (y_i - \overline{y})^2 \\ &= \sum_{i=1}^n r_i^2 + \sum_{i=1}^n (\hat{y}_i - \overline{y})^2 \\ &= SSE + SSR \end{aligned}$$

If $H_0 : \beta_1 = 0$ is true ($y_i = \beta_0 + \epsilon_i$), SSR should be relatively small with respect to SSE

2.6.1 Properties under $H_0 : \beta_1 = 0$

- $SSR/\sigma^2 \sim \chi^2_1$
- $SST/\sigma^2 \sim \chi^2_{n-1}$
- $SSE/\sigma^2 \sim \chi^2_{n-2}$, and is independent of SSR

2.7 F-Statistic

$$F = \frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{SSR}{SSE/(n-2)} = \frac{MSR}{MSE} \sim F(1, n-2)$$

Where MSR stands for mean square of regression, MSE stands for mean square of error
To test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 1$, we reject H_0 at level α if

$$F > F_\alpha(1, n-2)$$

where α represents the upper α quantile
Recall that $\dfrac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)} \sim t_{n-2}$ can test $H_0 : \beta_1 = \beta_1^*$, where β_1^* is some value we are interested
In Simple Linear Regression, testing $H_0 : \beta_1 = 0$ using t-test and F-test are equivalent

2.8 ANOVA Table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F-Statistic
Regression	SSR	1	MSR=$\frac{SSR}{1}$	$\frac{MSR}{MSE}$
Residual	SSE	n-2	MSE=$\frac{SSE}{n-2}$	
Total	SST	n-1		

2.9 Cochram's Theorem

Suppose $U_i \overset{iid}{\sim} N(0,1)$ for $i = 1, 2, \cdots, n$, and $\sum_{i=1}^n U_i^2 = Q_1 + Q_2$
Let d_1 and d_2 be the degree of freedom of Q_1 and Q_2, which are the number of linearly independent linear combination of y_i's in Q_1 and Q_2
If $d_1 + d_2 = n$, Q_1 and Q_2 are independent, and $Q_1 \sim \chi^2_{d_1}$ and $Q_2 \sim \chi^2_{d_2}$

2.10 Coefficient of Determination

$$R^2 = \frac{SSR}{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

In SLR,

$$R^2 = \frac{\hat{\beta}_1 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

And

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Thus, $R^2 = r^2$

3 Random Vector and Linear Regression

Notation

- Capital letter for vector / matrix: A, X
- lower case for scalar: a, c
- lower direction vector / matrix: $\vec{a}, \vec{c}$
- vector is column vector: $\vec{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$

Definition

Suppose $Y = \begin{bmatrix} y_1, \cdots, y_n \end{bmatrix}^T$ is $n \times 1$ vector of random variable with $E(y_i) = \mu_i$, $Var(y_i) = \sigma_i^2$, $Cov(y_i, y_j) = \sigma_{ij}$
Then

$$E(Y) = \begin{bmatrix} E(y_1), \cdots, E(y_n) \end{bmatrix}^T = \begin{bmatrix} \mu_1, \cdots, \mu_n \end{bmatrix}^T$$

And

$$\begin{aligned} Var(Y) &= E([Y - E(Y)][Y - E(Y)]^T) \\ &= \begin{bmatrix} Var(y_1) & Cov(y_1, y_2) & \cdots & Cov(y_1, y_n) \\ Cov(y_2, y_1) & Var(y_2) & \cdots & Cov(y_2, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(y_n, y_1) & Cov(y_n, y_2) & \cdots & Var(y_n) \end{bmatrix} \\ &= (\sigma_{ij})_{n \times n} \end{aligned}$$

If $y_1, \cdots, y_n$ are independent ($Cov(y_i, y_j) = 0$, $Var(y_i) = \sigma^2$), then $Var(Y) = \sigma^2 I$

Basic Properties

Suppose $A = (a_{ij})_{m \times n}$, $\vec{b} = \begin{bmatrix} b_1, \cdots, b_m \end{bmatrix}^T$ and $\vec{c} = \begin{bmatrix} c_1, \cdots, c_n \end{bmatrix}^T$ are matrix and vector of constants

- $E(AY + \vec{b}) = AE(Y) + \vec{b}$
- $Var(Y + \vec{c}) = Var(Y)$
- $Var(AY) = AVar(Y)A^T$
- $Var(AY + \vec{b}) = AVar(Y)A^T$

Differentiation over Linear and Quadratic Forms, scalr w.r.t. vector

- $f = f(Y) = f(y_1, \cdots, y_n)$, then $\frac{df}{dY} = (\frac{df}{dy_1}, \cdots, \frac{df}{dy_n})$
- $f = \vec{c}^T Y = \sum_{i=1}^n c_i y_i$, then $\frac{df}{dY} = \vec{c}^T$

Matrix Result

- Trace
$$trace(A_{m\times m}) = \sum_{i=1}^m a_{ii}$$
$$trace(BC) = trace(CB)$$
- Rank
$$rank(A_{m\times n}) = \text{max num of linearly independent columns / rows}$$
- vectors are linearly independent iff
$$c_1 Y_1 + \cdots + c_n Y_n = 0 \rightarrow c_1 = \cdots = c_n = 0$$
is the only solution
- orthogonality
  - two vectors X and Y are orthogonal if $Y^T X = 0$
  - a square matrix is orthogonal iff $A^T A = AA^T = I_{n\times n}$
- Eigenvalue and Eigenvector of square matrix
  - non-zero vector $\vec{v}_i$ is eigenvector of $A_{m\times m}$ if
$$A\vec{v}_i = \lambda_i \vec{v}_i$$
- Idempotent Matrix $A_{m\times m}$ is idempotent if $A^2 = A$
  - if A is idempotent then all its eigenvalues are 0 or 1
  - if A is idempotent and symmetric, there exists orthogonal matrix P such that $A = P\Lambda P^T$ where Λ is a zero matrix but with the diagonal fill with $rank(A)$ 1's, $tr(A) = rank(A) = tr(\Lambda) =$ number of eigenvalues being 1

3.1 Multivariate Normal Distribution

The random vector $Y = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^T$ follow a multivariate normal distribution with pdf
$$f(Y) = [\frac{1}{\mu}]^{n/2} |\Sigma|^{-1/2} exp\{-\frac{1}{2}(Y - \vec{\mu})^T \Sigma^T (Y - \vec{\mu})\}$$
where $\vec{\mu} = E(Y)$, $\Sigma = Var(Y) = (\sigma_{ij})_{n\times n}$ and $|\Sigma|$ is the determinant of Σ
Denote it as $Y \sim MVN(\vec{\mu}, \Sigma)$

3.1.1 Properties

- $y_1, \cdots, y_n$ are independent iff Σ is diagonal
- marginal normality: $y_i \sim N(\mu_i, \sigma_{ii})$
- if $Y \sim MVN(\vec{\mu}, \Sigma)$, and $Z = AY$, then $Z \sim MVN(A\vec{\mu}, A\Sigma A^T)$
- if $Y \sim MVN(\vec{0}, \sigma^2 I)$, then $\frac{Y^TY}{\sigma^2} \sim \chi^2_n$
- Let $U = AY$, $W = BY$, $Y \sim MVN(\vec{\mu}, \Sigma)$, U and W are independent iff $A\Sigma B^T = \vec{0}$, and $Cov(AY, BY) = ACov(Y, Y)B^T = A\Sigma B^T$

3.2 Multiple Linear Regression (MLR)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where

- $x_{i1}, \cdots, x_{ip}$ are fixed and known predictor variable
- $\beta_0, \cdots, \beta_p$ are fixed but unkown regression parameters
- ϵ_i is random and unkown error
- y_i is random and observable response

We make the assumptions that

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = \sigma^2$
- ϵ_i are independent
- $\epsilon_1, \cdots, \epsilon_n \sim N(0, \sigma^2) \to y_i \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2)$

β_j: the average increase (or decrease) in response when the jth predictor x_j increase (or decrease) by 1 unit while holding all other predictors fixed/constant
$H_0 : \beta_j = 0$ means x_j is NOT linearly related to y, given all other predictors in the model fixed

3.2.1 Matrix Form Representation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$Y = X\vec{\beta} + \vec{\epsilon}$ where $\vec{\epsilon} \sim MVN(\vec{0}, \sigma^2 I)$ and $Y \sim MVN(X\vec{\beta}, \sigma^2 I)$

3.3 LSE of $\vec{\beta}$

We have

$$S(\vec{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = Y^T Y - 2Y^T X\vec{\beta} + \vec{\beta}^T X^T X\vec{\beta}$$

Derive and minimize to get

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

3.3.1 Properties of $\hat{\vec{\beta}}$

- $\hat{\vec{\beta}}$ is unbiased
- $Var(\hat{\vec{\beta}}) = \sigma^2 (X^T X)^{-1}$

3.4 Results of MLR

- fitted values: $\hat{Y} = X\hat{\vec{\beta}} = X(X^T X)^{-1} X^T Y = HY$, H is idempotent and symmetric
- residuals: $\vec{r} = Y - \hat{Y} = Y - HY = (I - H)Y$
  - matrix $I - H$ is idenpotent and symmetric
  - $\sum_{i=1}^{n} r_i = 0$
  - $X^T \vec{r} = \vec{0}$
  - $\hat{Y}^T \vec{r} = \vec{0}$
  - $E(\vec{r}) = \vec{0}$
  - $Var(\vec{r}) = \sigma^2 (I - H)$
  - estimation of σ^2:
  $$\hat{\sigma}^2 = \frac{\sum r_i}{n - p - 1}$$
- inference of a single β
  $$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 \nu_{ii}}}$$
  where ν_{ii} is the ith diagonal element of $(X^T X)^{-1}$
  If $Z \sim N(0, 1)$, $W \sim \chi^2_\nu$, Z and W are independent, then
  $$\frac{Z}{\sqrt{W/\nu}} \sim t_\nu$$

3.4.1 Results of $\hat{\vec{\beta}}$

Assume $Y \sim MVN(X\vec{\beta}, \sigma^2 I)$, then

- $\hat{\vec{\beta}} \sim MVN(\vec{\beta}, \sigma^2(X^T X)^{-1})$
- $\hat{\vec{\beta}}$ and $\hat{\sigma}^2$ are independent
- $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$

3.4.2 Confidence Interval

Recall $\hat{\vec{\beta}} \sim MVN(\vec{\beta}, \sigma^2(X^T X)^{-1})$, then $\hat{\beta}_1 \sim N(\beta_i, \sigma^2 \nu_{ii})$, where ν_{ii} are the ith diagonal element of $(X^X)^{-1}$

$$\frac{\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 \nu_{ii}}}}{\sqrt{\frac{(n-p-1)\hat{\sigma}^2/\sigma^2}{n-p-1}}} \sim t_{n-p-1}$$

Then

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}_i^2 \nu_i i}} \sim t_{n-p-1}$$

The denominator is the $s.e.(\hat{\beta}_i)$

3.5 Linear Combination of β_i's

Let $\vec{a} = (a_0, a_1, \cdots, a_p)^T$ be a $(p+1)$ dimension vector of constants. We are interested in

$$\theta = \vec{a}^T vec\beta = \sum_{i=0}^{p} a_i \beta_i$$

We estimate θ as $\hat{\theta} = \vec{a}^T \hat{\vec{\beta}}$
Recell that $\hat{\vec{\beta}} \sim MVN(\vec{\beta}, \sigma^2(X^T X)^{-1})$, then $\hat{\theta} = \vec{a}^T \hat{\vec{\beta}} \sim N(\theta, \sigma^2 \vec{a}^T (X^T X)^{-1} \vec{a})$

$$\frac{\hat{\theta} - \theta}{\sqrt{\sigma^2 \vec{a}^T (X^T X)^{-1} \vec{a}}} \sim N(0, 1)$$

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\sigma}^2 \vec{a}^T (X^T X)^{-1} \vec{a}}} \sim t_{n-p-1}$$

3.6 Prediction of y

Let $\vec{a}_p = (1, x_1, \cdots, x_p)^T$, $y_p = \vec{a}_p\vec{\beta} + \epsilon_p$

$$Var(y_p - \hat{y}_p) = \sigma^2[1 - \vec{a}_p^T(X^TX)^{-1}X^T\vec{a}_p]$$

Then

$$\frac{y_p - \hat{y}_p}{\sqrt{\hat{\sigma}^2[1 - \vec{a}_p^T(X^TX)^{-1}X^T\vec{a}_p]}} \sim t_{n-p-1}$$

A $100(1-\alpha)\%$ prediction interval for y_p is

$$\hat{y}_p \pm t_{n-p-1,\alpha/2}\sqrt{\hat{\sigma}^2[1 - \vec{a}_p^T(X^TX)^{-1}X^T\vec{a}_p]}$$

3.7 Analysis of Variance (ANOVA)

The sum of square of residuals (error) is

$$SSE(\hat{\beta}) = \sum_{i=1}^{n} r_i^2 = \vec{r}^T\vec{r} = (Y - X\hat{\beta})^T(Y - X\hat{\beta})$$

Test $H_0^* : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
Under H_0^*, the full model reduces to

$$y_i = \beta_0 + \epsilon_i$$

The LSE under reduced model is

$$\hat{\beta}_0 = \overline{y}$$

$$SSE(\hat{\beta}_0) = \sum_{i=1}^{n}(y_i - \overline{y})^2 = SST$$

The difference is

$$\begin{aligned} SSE(\hat{\beta}_0) - SSE(\hat{\beta}) &= SST - SSE \\ &= SSR \\ &= \vec{\hat{\beta}}^T X^T X \vec{\hat{\beta}} - \overline{y}^T\overline{y} \end{aligned}$$

Under $H_0^* : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$$\frac{SSR}{\hat{\sigma}^2} \sim \chi_p^2$$

The F-test Statistics

$$F = \frac{SSR/p}{SSE/n-p-1} = \frac{MSR}{MSE} \sim F_{p,n-p-1}$$

We reject H_0^* at level α if

$$F > F_{\alpha,(p,n-p-1)}$$

3.7.1 ANOVA Table

Source of Variation	Sum of Squares	df	Mean Square	F-Statistic
Regression	SSR=$\hat{\vec{\beta}}^T X^T X \hat{\vec{\beta}}$	p	MSR=$\frac{SSR}{p}$	$\frac{MSR}{MSE}$
Residual	SSE=$(Y - X\hat{\vec{\beta}})^T (Y - X\hat{\vec{\beta}})$	n-p-1	MSE=$\frac{SSE}{n-p-1}$	
Total	SST=$\sum (y_i - \overline{y})^2$	n-1		

3.7.2 Total Coefficients of Determination

$$R^2 = \frac{SSR}{SSE} \quad 0 \leq R^2 \leq 1$$

3.8 Geometric Interpolation of LSE

3.8.1 Column Space of X

$C(X)$ is all vectors that can be constructed as a linear combination of columns of X
$C(X)$ spans a $p+1$ dimensional subspace inside the n dimensional space

$$\text{LSE minimize } \sum r_i^2 \iff \text{minimize the length of residual vector}$$

$\hat{Y} = HY$ is the perpendicular projection of Y onto $C(X)$

$$\vec{r} \perp C(X) \quad \vec{r} \perp x_i, \forall i = 0, \cdots, p$$

3.9 Test Linear Constraints

Suppose we have l linear constraints, A is a $l \times (p+1)$ matrix

3.9.1 Additional Sum of Squares Principle

Recall $C(X) = \{\beta_0 \vec{\ell} + \beta_1 \vec{x}_1 + \cdots + \beta_0 \vec{x}_p\}$
Define $C_A(X) = \{\beta_0 \vec{\ell} + \beta_1 \vec{x}_1 + \cdots + \beta_0 \vec{x}_p | A\vec{\beta} = \vec{0}\}$ as subspace of $C(X)$ subject to the restriction $A\vec{\beta} = \vec{0}$
Let $\hat{Y}$ be the orthogonal projection of Y onto $C_A(X)$, and $\hat{Y}_A$ be the orthogonal projection of Y onto $C_A(X)$
If $H_0 : A\vec{\beta} = \vec{0}$ is true, we expect $\hat{Y}$ and $\hat{Y}_A$ to be close. The squared distance

$$||\hat{Y} - \hat{Y}_A||^2 = (\hat{Y} - \hat{Y}_A)^T (\hat{Y} - \hat{Y}_A) = SSE_A - SSE$$

is the additional sum of squares

Theory

Under $H_0 : A\vec{\beta} = \vec{0}$ where A is $l \times (p+1)$ matrix, we have

- $$\frac{||\hat{Y} - \hat{Y}_A||^2}{\sigma^2} \sim \chi^2_p$$
- $||\hat{Y} - \hat{Y}_A||^2$ is independent of $\hat{\sigma}^2 = \dfrac{(Y - \hat{Y})^T (Y - \hat{Y})}{n - p - 1}$
  F-statistic
  $$F = \frac{||\hat{Y} - \hat{Y}_A||^2 / l}{\hat{\sigma}^2} \sim F_{l,n-p-1}$$

We reject H_0 at level α if $F > F_{\alpha,(l,n-p-1)}$, or $p - value = P(F_{l,n-p-1} > F)$

3.9.2 Summary

To test $H_0 : A\vec{\beta} = 0$

- fit full model without restriction
- compute SSE
- fit restricted model with $A\vec{\beta} = 0$
- compute SSE_A
- compute $F = \dfrac{(SSE_A - SSE)/l}{SSE/(n - p - 1)}$

4 Regression Model Specification

In MLR, $Y = X\beta + \epsilon$, $E(Y) = X\beta$

4.1 Special Cases

4.1.1 Piecewise Constants

Naive Model

$$\begin{cases} y = \beta_0 & \text{if } x \leq a \\ y = \beta_1 & \text{if } x > a \end{cases}$$

We can rewrite it in linear way

$$y = \beta_0 I(x < a) + \beta_1 I(x \geq a)$$

where

$$X = \begin{bmatrix} I(x_1 < a) & I(x_1 \geq a) \\ \vdots & \vdots \\ I(x_n < a) & I(x_n \geq a) \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Or we can write it as

$$y = \beta_0 + \beta_2 I(x \geq a)$$

where $\beta_2 = \beta_1 - \beta_0$

4.1.2 Piecewise Linear

$$y = \beta_0 I(x < a) + \beta_1 x_1 I(x < a) + \beta_2 I(x \geq a) + \beta_3 x_1 I(x \geq a)$$

where

$$X = \begin{bmatrix} I(x_1 < a) & x_1 I(x_1 < a) & I(x_1 \geq a) & x_1 I(x_1 \geq a) \\ \vdots & \vdots & \vdots & \vdots \\ I(x_n < a) & x_n I(x_n < a) & I(x_n \geq a) & x_n I(x_n \geq a) \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

4.1.3 Piecewise Linear but Continuous

We have

$$\begin{cases} y = \beta_0 + \beta_1 x & \text{if } x < a \\ y = \beta_0 + \beta_1 x + \beta_3(x - a) & \text{if } x \geq a \end{cases}$$

4.1.4 One Sample Problem

We have $y_i = \beta_0 + \epsilon_i$ with $E(y_i) = \beta_0$

$E(Y) = X\beta$ with $X = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1}$ and $\beta = \beta_0$

4.1.5 Two Sample Problem

Cell Means Model

$y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, 2$, $j = 1, \cdots, n$

$$E\left[\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n} \\ -- \\ y_{21} \\ \vdots \\ y_{2n} \end{bmatrix}\right] = \begin{bmatrix} 1 \\ -- \\ 0 \end{bmatrix} \mu_1 + \begin{bmatrix} 0 \\ -- \\ 1 \end{bmatrix} \mu_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

Then

$$X^T X = \begin{bmatrix} (1_{n_1 \times 1})^T & \vec{0}^T \\ \vec{0}^T & (1_{n_2 \times 1})^T \end{bmatrix}_{2 \times n} \begin{bmatrix} 1_{n_1 \times 1} & 0 \\ 0 & 1_{n_2 \times 1} \end{bmatrix}_{n \times 2} = \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}$$

$$X^TY = \begin{bmatrix} (1_{n_1\times 1})^T & \vec{0}^T \\ \vec{0}^T & (1_{n_2\times 1})^T \end{bmatrix}_{2\times n} \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n_1} y_{ij} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix}$$

$$\hat{\beta} = (X^TX)^{-1}X^TY = \begin{bmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{j=1}^{n_1} y_{ij} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix} = \begin{bmatrix} \overline{y_{1+}} \\ \overline{y_{2+}} \end{bmatrix}$$

where $\hat{\mu}_1 = \overline{y_{1+}}$ and $\hat{\mu}_2 = \overline{y_{2+}}$

Effects Model

$E(y_i) = \beta_0 + \beta_1 x_i$ where $x_i = I$[observation i is in group 2]

$$E\left[\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{bmatrix}\right] = \begin{bmatrix} 1_{n_1\times 1} & 0_{n_1\times 1} \\ 1_{n_2\times 1} & 1_{n_2\times 1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} n & n_2 \\ n_2 & n_2 \end{bmatrix} \quad X^TY = \begin{bmatrix} \sum_{j=1}^{n_1} y_{ij} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix}$$

Then

$$\hat{\beta} = \frac{1}{n_1} \begin{bmatrix} 1 & -1 \\ -1 & \frac{n}{n_2} \end{bmatrix} \begin{bmatrix} \sum_{j=1}^{n_1} y_{ij} \\ \sum_{j=1}^{n_2} y_{2j} \end{bmatrix} = \begin{bmatrix} \overline{y_{1+}} \\ \overline{y_{2+}} - \overline{y_{1+}} \end{bmatrix}$$

4.1.6 K-sample Problem

Cell Means Model

$y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, \cdots, k$

$$E\left[\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{bmatrix}\right] = \begin{bmatrix} 1_{n_1\times} & 0_{n_1\times} & \cdots & 0_{n_1\times} \\ 0_{n_2\times} & 1_{n_2\times} & \cdots & 0_{n_2\times} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_k\times} & 0_{n_k\times} & \cdots & 1_{n_k\times} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}$$

$$X^TX = \begin{bmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_k \end{bmatrix}$$

$$\hat{\beta} = (X^TX)^{-1}XY = \begin{bmatrix} \overline{y_{1+}} \\ \vdots \\ \overline{y_{k+}} \end{bmatrix}$$

where $\overline{y_{i+}} = \frac{1}{n_1}\sum_{j=1}^{n_i} y_{ij}$

$$\hat{Y} = X\hat{\beta} = \begin{bmatrix} \overline{y_{1+}} \\ \vdots \\ \overline{y_{1+}} \\ \vdots \\ \overline{y_{k+}} \\ \vdots \\ \overline{y_{k+}} \end{bmatrix}$$

Effects Model

$E(y_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ij}$ where $x_{ij} = I[\text{obs i in group j}]$

$$E(Y) = \begin{bmatrix} \vec{1} & \vec{x}_2 & \cdots & \vec{x}_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

where $\vec{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \end{bmatrix}$ It can be shown that

$$\hat{\beta} = (X^TX)^{-1}X^TY = \begin{bmatrix} \overline{y_{1+}} \\ \overline{y_{2+}} - \overline{y_{1+}} \\ \vdots \\ \overline{y_{k+}} - \overline{y_{1+}} \end{bmatrix}$$

4.1.7 ANOVA Table

Source	**Sum of Squares**	**df**	**Mean Squares**	**F-statistic**
Regression	$SSR = \sum_{i=1}^{k}(\hat{y}_i - \bar{y})^2$	$k-1$	$MSR = \frac{SSR}{k-1}$	$\frac{MSR}{MSE}$
Residual	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-k$	$MSE = \frac{SSE}{n-k}$	
Total	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

5 Model Checking

Recall that $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$
We assume

- $E(\epsilon_i) = 0$
- $Var(\epsilon_i) = \sigma^2$
- ϵ_i are independent
- $\epsilon_i \sim N(0, \sigma^2)$

We hope to use $r_i = y_i - \hat{y}_i$ to approximate $\epsilon_i = y_i - E(y_i)$
If $n >> p$, and model is correctly specified, then

$$r_i \approx \epsilon_i$$

Recall that

$$\vec{r} = (I - H)\vec{\epsilon}$$

H is idempotent and symmetric, then

- $h_{ii} = (H)_{ii} = (HH)_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji}$
- $0 \leq h_{ii}(1 - h_{ii}) \leq \frac{1}{4}$
- off-diagonal elements cannot be large
- $\sum h_{ii} = p + 1$, the average of h_{ii} value is $\frac{p+1}{n}$
- if $n >> p$, all elements of H is small
$$\vec{r} \approx \vec{\epsilon}$$
- if $n = p + 1$, average of h_{ii} is 1, then $\vec{r} = \vec{0}$

5.1 Model Checking

5.1.1 Studentized Residual

Standardized residual

$$r_i^s = \frac{r_i}{\hat{\sigma}} \quad i = 1, \cdots, n$$

Studentized residual

$$d_i = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \quad i = 1, \cdots, n$$

where h_{ii} is the ith diagonal element of H. Under assumputions of random errors, $d_i \sim N(0, 1)$

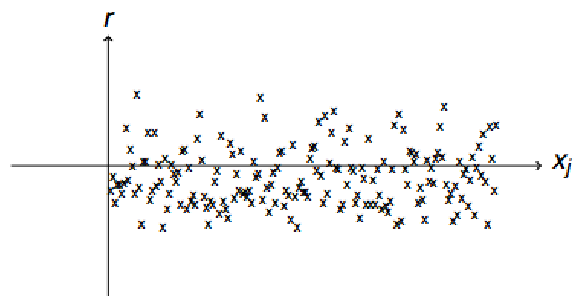
5.1.2 Residual Plots for Checking $E(\epsilon_i) = 0$

The most important assumption for linear regression models is $E(\epsilon_i) = 0$
The violation of this assumption can be

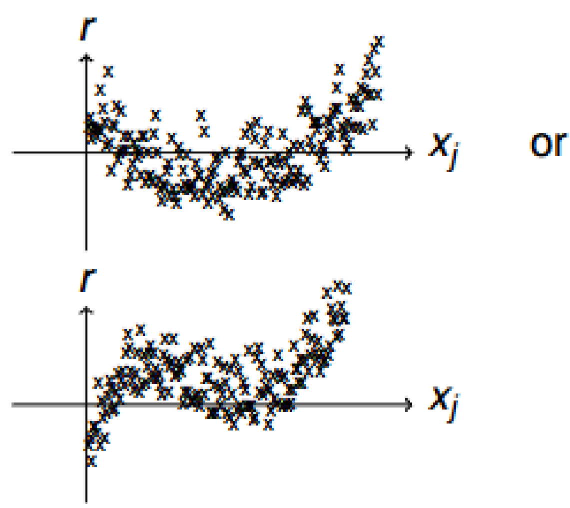
- effect of predictors on response variable is not in fact linear
- omission of some important predictors

5.1.3 Residuals vs x_j

If a linear effect on y, then we expect to see a random pattern, points fall into a horizontal band around 0

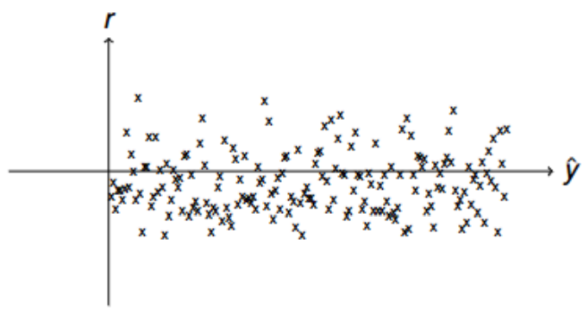


If we see any obvious non-random pattern, it suggests the non-linearity



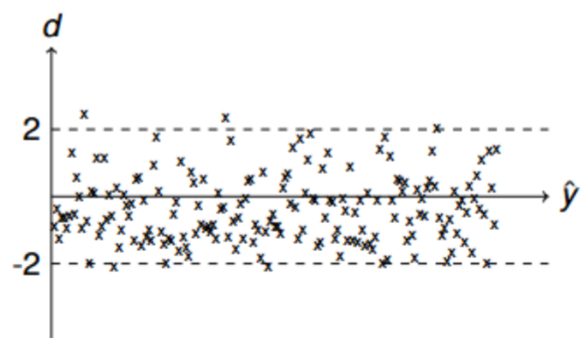
5.1.4 Residuals vs $\hat{y}$

If model is adequate $E(\epsilon_i) = 0$, we have $Cov(\epsilon_i, \hat{y}_i) = 0$
The residuals should lie within a horizontal band around zero, no special pattern



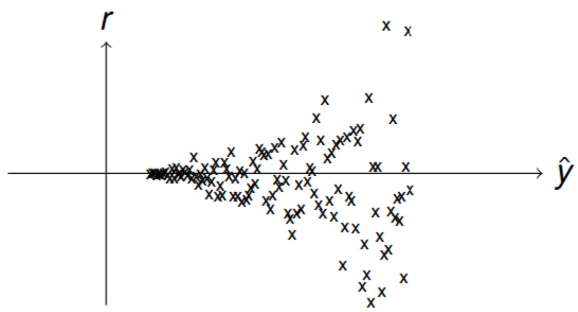
5.1.5 Studentized Residuals vs $\hat{y}$

The studendized residuals should lie within a horizontal band around zero, no special pattern
Approx 95% of studentized residuals should lie within $(-2, 2)$, and almost all of them should be within $(-3, 3)$



5.1.6 Residual Plots for Checking Variance $V(\epsilon_i) = \sigma^2$

The constant variance assumption is violated if there is a pattern

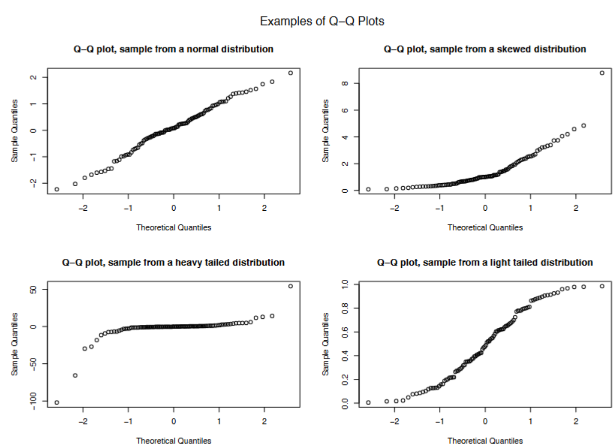


5.1.7 Durbin-Waston Test

The Durbin-Waston statistic tests $H_0 : \rho = 0$ vs $H_a : \rho \neq 0$

$$d = \sum_{i=2}^{n}(r_i - r_{i-1})^2 / \sum_{i=1}^{n} r_i^2 \approx 2(1-\rho)$$

5.1.8 Q-Q Plot



5.2 Leverage

Recall $H = X(X^TX)^{-1}X^T = (h_{ij})_{n\times n}$, and $\hat{Y} = HY$

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$$

Leverage of the ith observed predictor is defined as h_{ii}
It reflects the distance between the ith observation $(x_{i1}, \cdots, x_{ip})$ and the other observations

The leverage h_{ii} is small for cases $(x_{i1}, \cdots, x_{ip})$ near the centroid $(\overline{x_1}, \cdots, \overline{x_p})$ that is determined by all cases. Large if $(x_{i1}, \cdots, x_{ip})$ is far from the centroid
Case i is potentially influential if

$$h_{ii} > 2\frac{p+1}{n}$$

5.3 Cook's Distance

It can be shown

$$D_i = \frac{h_{ii} d_i^2}{(1-h_{ii})(p+1)}$$

where d_i is the studentized residual
Cook's distance is an overall measure

- if $|h_{ii}|$ is large, but d_i is small, then influence will be small
- if $|d_i|$ is large, but h_{ii} is small, then influence will be small

A large value indicates that the observation has a large influence on the results
Cook suggested that a Cook's Distance is significantly large when it is greater than $F_{0.5}(p+1, n-p+1)$

5.4 PRESS Residuals

prediction error

$$r_{(-i)} = y_i - x_i^T \hat{\beta}_{(-i)} = \frac{r_i}{1-h_{ii}}$$

PRESS residuals is

$$\sum_{i=1}^{n} r_{(-i)}^2 = \sum_{i=1}^{n} \frac{r_i^2}{(1-h_{ii})^2}$$

6 Model Selection

R^2 may only be appropriate for comparing two models with same number of predictors
Adjusted R^2 is

$$R_{adj}^2 = 1 - (1-R^2)\frac{n-1}{n-p-1}$$

n is sample size, p is number of covariates

6.1 Akaike's Information Criterion (AIC)

$$AIC = -2\log(L) + 2(p+1)$$

where L is the likelihood of the model
In general a smaller value of AIC is preferred

6.2 Bayesian Information Criterion (BIC)

$$BIC = -2\log(L) + \log(n)(p+1)$$

6.3 Note

For R^2 and R^2_{adj}, the larger the better. For AIC and BIC, the smaller the better

6.4 Backward Elimination with p-value

1. Start with all p potential explanatory variables in the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

2. For each explanatory variable, calculate the p-value for testing $H_0 : \beta_j = 0$

3. If largest p-value is greater than α, remove the variable with the largest p-value

4. Repeat step 2 and 3 until all p-values are less than α

6.5 Forward Selection with p-value

1. Fit p simple linear models, each with only a single explanatory variable v_j. There are p t-test statistics and p-values for testing $H_0 : \beta_j = 0$. The most significant predictors is the one with the smallest p-value, denote by v_k
   If the smallest p-value greater than α, stop. Otherwise, set $x_1 = v_k$ and fit the model

2. Start from model

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

   Enter the remains $p-1$ variables one at a time, and fit $p-1$ models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 v_j + \epsilon$$

   and let p_k denote the smallest p-value, v_k denote the most significant explanatory variable
   If $p_k > \alpha$, stop. Otherwise, set $x_2 = v_k$ and fit the model

3. Continue until no new explanatory variables can be added

6.6 Variance Stablizing Transformation

Consider general model

$$y_i = \mu_i + \epsilon_i$$

where $\mu_i = E(y_i) = f(x_i, \beta)$, the mean of response
Suppose that

$$V(y_i) = V(\epsilon_i) = h^2(\mu_i)\sigma^2$$

for some function h
Task: find a transformation $g(y_i)$ such that variance of $g(y_i)$ is constant We approximate $g(y_i)$ by a first order Taylor expansion around μ_i

$$g(y_i) \approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$$

Then

$$V(g(y_i)) \approx g'(\mu_i)^2 V(y_i) = g'(\mu_i)^2 h^2(\mu_i)\sigma^2$$

The common form of h is power function
Let $h^2(\mu_i) = \mu_i^\alpha$, want $g'(\mu_i) = \frac{1}{h(\mu_i)} = \mu_i^{-\alpha/2}$
Thus

$$g(y_i) = \begin{cases} y_i^{\alpha/2} & \alpha \neq 2 \\ \log(y_i) & \alpha = 2 \end{cases}$$

Special Case:

- $h^2(\mu_i) = \mu_i \Rightarrow Var(y_i) = \mu_i\sigma^2$, variance is proportional to mean, $\alpha = 1$ and $g(y_i) = \sqrt{y_i}$
- $h^2(\mu_i) = \mu_i^2 \Rightarrow Var(y_i) = \mu_i^2\sigma^2$, variance is proportional to square of mean, $\alpha = 2$ and $g(y_i) = \log(y_i)$

6.7 Linear Dependency / Multicollinearity

6.7.1 Perfect Multicollinearity

The columns of design matrix (predictors) $1, X_1, \cdots, X_p$ are linearly dependent, or have perfect multicollinearity if one column can be expressed as a linear combination of the other columns.

6.7.2 Multicollinearity

If there exists constants $c_0, c_1, \cdots, c_p$ not all zero such that $c_0 1 + c_1 X_1 + \cdots + c_p X_p \approx 0$, but maybe not exactly linearly dependent, then we say the predictors have multicollinearity

- If there is perfect multicollinearity, then $|X^T X| = 0$ and $(X^T X)^{-1}$ does not exist, thus $\hat{\beta}$ does not exist
- If there is multicollinearity, then $|X^T X| \approx 0$ and $(X^T X)^{-1}$ is large. Consequently, the variances of the estimated regression coefficients $\hat{\beta}_0, \cdots, \hat{\beta}_p$ are large

If multicollinearity exists

- The variance of $\hat{\beta}$ is large
- Important predictors become insignificant in the model
- Hard to distinguish the effect of each predictor

6.7.3 Detection of Multicollinearity

First check pairwise sample correlation coefficient

$$r_{lm} = \frac{\sum_{i=1}^{n}(x_{il} - \overline{x}_l)(x_{im} - \overline{x}_m)}{\sqrt{\sum_{i=1}^{n}(x_{il} - \overline{x}_l)^2 \sum_{i=1}^{n}(x_{im} - \overline{x}_m)^2}}$$

If $|r_{lm}| \approx 1$, then X_l and X_m are highly correlated, no need for both in the model

6.8 Variance Inflation Factor

A formal check: VIF

- x_k is regressed on the the remaining $p - 1$ x's:

$$x_{ij} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{k-1} x_{i(k-1)} + \beta_{k+1} x_{i(k+1)} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- The result

$$R_k^2 = \frac{SSR}{SST}$$

  is a measure of how strongly x_k is linearly related to the rest of x's

$$VIF_k = \frac{1}{1 - R_k^2}$$

- If $VIF_k > 10$, strong evidence of multicollinearity
- IF $VIF_k > 5$, some evidence of multicollinearity
- If $VIF_k < 5$, dont worry