第一章 线性模型的统计推断

Overview

- Linear models
  - The error structure
  - The systematic component
  - Parameter estimation of β
  - The rank of the design matrix
  - Properties of the OLS estimate (moments,equivalence,distributions)
  - Estimable functions
  - Estimation of σ^2

Generalized linear models

- Traditional regression and analysis of variance methods assume the data to be independent, normal with constant variance.
- This can be a poor approximation in practice:
  - The data may be non-normal, e.g, we may observe proportions,counts.
  - The data that may be independent, but not identically distributed. One way to handle this is to transform the data,and then fit the model. This can be hard to interpret on the original measurement scale.

The course aims

- This course aims to introduce the statistical theory to extend regression and analysis of variance to non-normal data.
- By the end of the course, you should be able to use generalized linear models to model data.
- We will focus on model identification, building, diagnostics and inference.
- Time permitting we shall extend to more complex generalized linear models.
- We start by considering linear models.

Linear models

- Suppose we observe a set of data $\{\mathrm{Y_i} : \mathrm{i} = 1, ..., \mathrm{n}\}$.
- Consider the model

$$\mathrm{Y_i} = \mathbf{x_i}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{\mathrm{i}}$$

$\mathbf{x_i}$ is the p-dimensional **covariate** vector for observation i.

$\boldsymbol{\beta}$ is a p-dimensional **parameter** vector.

$\{\epsilon_{\mathrm{i}} : \mathrm{i} = 1, ..., \mathrm{n}\}$ is a set of independent normal **errors** with mean 0 and variance σ^2.

The linear model expressed using matrix notation

- Let $\mathbf{Y} = (\mathrm{Y}_1, ..., \mathrm{Y}_n)^{\mathrm{T}}$ denote the vector associated with the observations **y**. Let

$$\mathrm{X} = \begin{pmatrix} \mathbf{x_1}^{\mathbf{T}} \\ \mathbf{x_2}^{\mathbf{T}} \\ \dots \\ \mathbf{x_n}^{\mathbf{T}} \end{pmatrix}$$

  be the **design** or **covariate** matrix and the random vector of errors be $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^{\mathrm{T}}$.

- Then the model becomes

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ &= \boldsymbol{\mu} + \boldsymbol{\epsilon} \end{aligned}$$

  where we let $\boldsymbol{\mu} = E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$.

The error structure

- We have that

$$\boldsymbol{\epsilon} \sim \boldsymbol{N}_{\mathrm{n}}(\mathbf{0}, \sigma^2 \boldsymbol{I}_{\mathrm{n}}),$$

  where $\boldsymbol{N}_{\mathrm{n}}(.,.)$ denotes a n-variate multivariate normal distribution and $\boldsymbol{I}_{\mathrm{n}}$ is the $\mathrm{n} \times \mathrm{n}$ identity matrix.

- It follows that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_{\mathrm{n}}$.

- We deduce that

$$\boldsymbol{Y}|\mathbf{X} \sim \boldsymbol{N}_{\mathrm{n}}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_{\mathrm{n}}).$$

The systematic component

- The systematic component or the linear predictor of the linear model is $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.
- The design matrix can contain either or both of the following terms
  - simple or multiple linear regression terms.
  - analysis of variance terms.
- Interpretation of the coeffcients $\boldsymbol{\beta}$ will depend on the form of $\mathbf{X}$.

Parameter estimation of β

- The least squares method involves calculating the residual sum of squares

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta).$$

- We minimize the RSS with respect to β.
- Graphically we find the line that minimizes the sum of squared vertical distances between $(\boldsymbol{x}_{\mathrm{i}}, \mathrm{y_i})$ and the line.
- Differentiating with respect to β and setting the expression equal to zero we obtain the normal equations for the least squares estimate,$\hat{\beta}$:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

The rank of the design matrix

- If $\mathbf{X}$ is of full column rank then $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ is also of full rank and thus

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

  This is the ordinary least squares estimate (OLS) of $\boldsymbol{\beta}$.

- If $\mathbf{X}$ is not of full column rank then we use generalized inverses or a numerical method to solve the normal equations.

  The **generalized inverse**,A^{-},of a matrix A are the solutions A^{-} of the equation $AA^{-}A = A$.

  The OLS for the non-full rank case is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

Properties of the OLS estimate

- Consider the full rank case.
- We have that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and

$$cov(\boldsymbol{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2.$$

  Thus the OLS estimator is an unbiased estimate of $\boldsymbol{\beta}$.
- From the Gauss-Markov theorem, the OLS estimator has the **minimum** variance of all unbiased estimators of $\boldsymbol{\beta}$
  (We say that $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate (BLUE)).
- None of these results need the assumption of normality. We only need assume that $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $cov(\mathbf{Y}) = \sigma^2\boldsymbol{I}_{\mathrm{n}}$.

Equivalence of the OLS and maximum likelihood estimates.

- Suppose that $\boldsymbol{Y} \sim \boldsymbol{N}_{\mathrm{n}}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_{\mathrm{n}})$.Then the **likelihood** of **Y** given the parameters $\boldsymbol{\beta}$ and σ^2 is

$$\begin{aligned} \mathrm{L}(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2) &= (2\pi)^{-\mathrm{n}/2} \det(\sigma^2\boldsymbol{I}_{\mathrm{n}})^{-1/2} \times \exp(-\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}) \\ &= (2\pi)^{-\mathrm{n}/2}(\sigma^2)^{-\mathrm{n}/2} \exp(-\frac{RSS(\boldsymbol{\beta})}{2\sigma^2}). \end{aligned}$$

Thus the **log-likelihood** is

$$\mathrm{l}(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2) = -\mathrm{n}/2\log(2\pi) - \mathrm{n}/2\log(\sigma^2) - \frac{RSS(\boldsymbol{\beta})}{2\sigma^2}.$$

Equivalence (cont.)

- Minimizing RSS(β) is equivalent to maximizing the likelihood with respect to β.
- Thus the maximum likelihood (ML) estimate of β is equal to the OLS estimate of β.
- We can also obtain the maximum likelihood estimate of σ^2.It is

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{\mathrm{n}},$$

where $\hat{\beta}$ is OLS(ML) estimate.

Sampling distribution of the OLS estimate

- If we want to make inferences we need some distributional assumptions.
- When **Y** has a multivariate normal distribution we have

$$\hat{\boldsymbol{\beta}} \sim \boldsymbol{N}_{\mathrm{p}}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}).$$

- Without normality, this result still holds asymptotically for large sample sizes, n.

Estimable functions

- When $\mathbf{X}$ is not of full rank, there may be many estimates of β.Regardless, it is possible to estimate certain unbiased linear combinations of β.
- A linear function of β,$\mathbf{a}^{\mathrm{T}}\beta$, is **estimable** if there exists some vector $\mathbf{c}$ such that

$$E(\mathbf{c}^{\mathrm{T}}\mathbf{Y}) = \mathbf{a}^{\mathrm{T}}\beta,$$

  for all β.
- It can be shown that $\mathbf{a}^{\mathrm{T}}\beta$ is estimable if and only if the vector $\mathbf{a}$ belongs to the space spanned by the columns of $\mathbf{X}$ (equivalently the space spanned by the columns of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$).
- Thus $\mathbf{a}^{\mathrm{T}}\beta$ is estimable for all vectors $\mathbf{a} \in \mathbb{R}^{\mathrm{p}}$,if the columns of $\mathbf{X}$ are linearly independent.

Estimable functions (cont.)

- More correct version of the Gauss-Markov theorem:

  Suppose that $\mathbf{a}^{\mathrm{T}}\beta$ is estimable. Then

  1. $\mathbf{a}^{\mathrm{T}}\hat{\beta}$ is the BLUE of $\mathbf{a}^{\mathrm{T}}\beta$.

- Corollary: If $\mathbf{a}^{\mathrm{T}}\beta$ is estimable then

$$\mathbf{a}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \mathbf{a}^{\mathrm{T}},$$

  for any generalized inverse $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}$ of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

- It also follows that

$$\mathrm{cov}(\mathbf{a}^{\mathrm{T}}\hat{\beta}) = \sigma^2\mathbf{a}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{a}.$$

Estimation of σ^2

- The ML estimate of σ^2 is a **biased** estimate.
- On the other hand, the estimator

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n - p}$$

  is an **unbiased** estimate of σ^2.
- When **Y** is multivariate normal

$$(n - p)\frac{\hat{\sigma}^2}{\sigma^2} \sim {\chi^2}_{n-p},$$

  where ${\chi^2}_{n-p}$ denotes a chi squared distribution with $n - p$ degrees of freedom.
- Also, $\hat{\sigma}^2$ is independent of the OLS estimate of β, $\hat{\beta}$.

Questions:

证明上述结论:

- When **Y** is multivariate normal

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p},$$

where χ^2_{n-p} denotes a chi squared distribution with $n-p$ degrees of freedom.

- $\hat{\sigma}^2$ is independent of the OLS estimate of β, $\hat{\beta}$.

第一题

$x_1, \ldots x_n$独立同分布于正太分布$N(u, \sigma^2)$,证明:

（1）$\bar{x}$与s^2相互独立

（2）$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

证明：构造正交矩阵A

$$A = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\cdot 1}} & \frac{-1}{\sqrt{2\cdot 1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\cdot 2}} & \frac{1}{\sqrt{3\cdot 2}} & \frac{-2}{\sqrt{3\cdot 2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{bmatrix} \tag{1}$$

第一题

令$Y = AX$，则Y服从正态分布，其中：

$$EY = A \cdot EX = (\sqrt{n}\mu, 0, ...0)$$

$$Var(Y) = AVar(X)A^T = A\sigma^2 I_n A^T = \sigma^2 I_n$$

则$\sum_{i=1}^{n} y_i^2 = Y^T Y = X^T A^T A X = X^T X = \sum_{i=1}^{n} x_i^2$

$$\begin{aligned}(n-1)s^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \sum_{i=1}^{n} x_i^2 - (\sqrt{n}\bar{x})^2 = \sum_{i=1}^{n} y_i^2 - y_1^2 = \sum_{i=2}^{n} y_i^2\end{aligned}$$

其中$\bar{x} = \frac{1}{\sqrt{n}} y_1$,则$\bar{x}$与$s^2$相互独立

$$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=2}^{n}(\frac{y_i}{\sigma})^2 \ \chi^2(n-1)$$

第二题

方差分析中，证明：在H_0成立下，$F = \frac{SSA/r-1}{SSE/n-r} \sim F(r-1, n-r)$

(1)$\frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{\sigma^2}$，由于$y_{ij} \sim N(\mu_i, \sigma^2)$,

则$\frac{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{\sigma^2} \sim \chi^2(n_i-1) i = 1...r$

各总体之间相互独立，由卡方分布可加性，

$\frac{SSE}{\sigma^2} \sim \chi^2(n-r)$

第二题

(2)对每个i,$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$与$\bar{y}_i$相互独立，则$\bar{y}_1,\bar{y}_2...\bar{y}_r$与SSE相互独立.

$SSA=\sum_{i=1}^{r}n_i(\bar{y}_i-\bar{y})^2$ 为$\bar{y}_1,\bar{y}_2...\bar{y}_r$的函数，则SSE与SSA相互独立

(3)由于$\frac{SST}{\sigma^2}=\frac{SSA}{\sigma^2}+\frac{SSE}{\sigma^2}$

即$\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}{\sigma^2}=\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{\sigma^2}=\frac{\sum_{i=1}^{r}n_i(\bar{y}_i-\bar{y})^2}{\sigma^2}$

在H_0成立下,$y_{ij}\sim N(\mu_i,\sigma^2)$

则$\frac{SST}{\sigma^2}\sim\chi^2(n-1)$，又因为SSA与SSE相互独立，$\frac{SSA}{\sigma^2}\sim\chi^2(r-1)$

Testing the significance of β

- Suppose we wish to test the following set of hypotheses:

$$H_0 : A\beta = c, \text{ versus } H_1 : A\beta \neq c.$$

Assume that :

—A is of full rank, r say.

—Each row of A, a_i, is such that $a_i^T \beta$ is estimable.

Special cases of the hypothesis

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

1. Testing all slope parameters are zero (the model F test)

   We suppose that the design matrix is of the form

$$X = \begin{bmatrix} 1 & \dots \\ \vdots & \vdots \\ 1 & \dots \end{bmatrix}$$

   and that $A = diag(0, 1, 1, ..., 1)$

2. Testing that the design ith parameter is significant.

   Let A be a single vector with $A_i = 1$ for some i, and 0 for all other entries.

3. Multiple partial F test (testing a subset of coefficients are zero).

   Let $A = diag(a_1, a_2, ..., a_n)$ where some a_i values are 0 and some are 1.

Distributional results

- We have that

$$A\hat{\beta} \sim N_r(A\beta, \sigma^2 A(X^T X)^- A^T).$$

and since $A\beta = c$,

$$A\hat{\beta} - c \sim N_r(0, \sigma^2 A(X^T X)^- A^T).$$

- Now let $V = A(X^T X)^- A^T$, and let L be some square root matrix of V (e.g., a Cholesky factor).
Then

$$(\sigma^2)^{-1/2} L^{-1}(A\hat{\beta} - c) \sim N_r(0, I_r).$$

and we conclude that

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

Obtaining the F statistic

- In practice we do not know σ^2. We estimate it by

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n-p},$$

and it is possible to show that

$$\frac{(A\hat{\beta}-c)^T V^{-1}(A\hat{\beta}-c)}{r\hat{\sigma}^2} \sim F_{r,n-p}.$$

This follows because the quadratic form $(A\hat{\beta}-c)^T(A\hat{\beta}-c)$ is independent of the quadratic form $RSS(\hat{\beta})$.

The ANOVA table

- The table is:

Source	SS	df	MS
Model	$SSM = SST - RSS(\hat{\beta})$	$p-1$	$MSM = \frac{SSM}{p-1}$
Error	$RSS(\hat{\beta})$	$n-p$	$MSE = \frac{RSS(\hat{\beta})}{n-p}$
Total	$SST = \boldsymbol{Y}^T\boldsymbol{Y} - n^{-1}\boldsymbol{Y}^T\boldsymbol{J}_n\boldsymbol{Y}$	$n-1$	$MST = \frac{SST}{n-1}$

- In the above table, $\boldsymbol{J}_n$ is an $n \times n$ matrix of ones.

问题1

- 方差分析为研究K个总体均值是否一致所构造的F统计量与线性回归模型对参数$\beta = 0$的检验所构造的F统计量是否一致？
- 即：方差分析是一种特殊的线性模型

方差分析

- 用于两个及以上样本均值差别的显著性检验
- 对于一个因子存在K个水平，即K个总体，分别抽取$n_1, n_2, \ldots, n_k$个样本，$n_1 + \cdots + n_k = n$，存在假定：

(1)每个总体为正态总体，即为$N(\mu_i, \sigma_i^2), i = 1, \ldots, k$；

(2)各总体方差相同，记为$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$；

(3)每个总体抽取的样本都是相互独立的，即所有实验结果y_{ij}都相互独立。

方差分析

原假设$H_0: \mu_1 = \cdots = \mu_k$

统计模型：
$$\begin{cases} y_{ij} = \mu_i + \varepsilon_{ij} & i = 1, \cdots, k \quad j = 1, \cdots, n \\ \varepsilon_{ij} \sim N(0, \sigma^2) \quad iid \end{cases}$$

设：$\mu = \frac{1}{k}(\mu_1 + \cdots + \mu_k) = \frac{1}{k}\sum_{i=1}^{k} \mu_i \quad a_i = \mu_i - \mu \quad \sum_{i=1}^{k} a_i = 0$

统计模型：
$$\begin{cases} y_{ij} = \mu_i + a_i + \varepsilon_{ij} & i = 1, \cdots, k \quad j = 1, \cdots, n \\ \sum_{i=1}^{k} a_i = 0 \\ \varepsilon_{ij} \sim N(0, \sigma^2) \quad iid \end{cases}$$

此时，原假设为$H_0 : a_1 = \cdots = a_k = 0$

构造F统计量：

$$F = \frac{S_A/f_A}{S_e/f_e} = \frac{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2/(k-1)}{\sum_{i=1}^{k}\sum_{i=1}^{n_i}(y_i - \bar{y}_i)^2/(n-k)}$$

方差分析

引入K-1个虚拟变量构建方差分析的线性回归模型

$Y = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$

原假设为$H_0 : \beta_1 = \cdots = \beta_k = 0$

构造F统计量：

$$F = \frac{\sum_{i=1}^{k}\sum_{i=1}^{n_i}(\hat{y}_{ij} - \bar{y})^2/(k-1)}{\sum_{i=1}^{k}\sum_{i=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2/(n-k)}$$

其中，$\hat{y}_{ij} = x_{ij}\hat{\beta}_i = \hat{\beta}_i$

$\hat{\beta}_i = (X_i^T X_i)^{-1} X_i^T Y_i = \frac{1}{n_i}(y_{i1} + \cdots + y_{in_i}) = \bar{y}_i$

$\Rightarrow \hat{y}_{ij} = \bar{y}_i$

问题2

多元线性回归模型：$Y = X\beta + \varepsilon$

原假设：$H_0 : \beta = 0$

$$W = \sqrt{n}(\hat{\beta} - 0)^T (Cov(\sqrt{n}\hat{\beta}))^{-1} \sqrt{n}(\hat{\beta} - 0)$$

$$F = \frac{\sum_{i=1}^{k} (\hat{y}_i - \bar{y})^2 / k}{\sum_{i=1}^{k} (y_i - \hat{y}_i)^2 / (n-k-1)}$$

判断F统计量与W的关系

当$Cov(\sqrt{n}\hat{\beta})$已知时，W渐近服从$\chi^2(k)$

当$Cov(\sqrt{n}\hat{\beta})$未知时，构造F统计量：

$$F = \sqrt{n}(\hat{\beta} - 0)^T(\hat{Cov}(\sqrt{n}\hat{\beta}))^{-1}\sqrt{n}(\hat{\beta} - 0)\frac{n-k-1}{k}$$

$$
\begin{aligned}
F &= \sqrt{n}(\hat{\beta}-0)^T(\hat{Cov}(\sqrt{n}\hat{\beta}))^{-1}\sqrt{n}(\hat{\beta}-0)\frac{n-k-1}{k} \\
&= \frac{\hat{\beta}^T(X^TX)\hat{\beta}/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)} \\
&= \frac{Y^TX(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^TY/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)} \\
&= \frac{Y^TX(X^TX)^{-1}X^TY/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)} \\
&= \frac{Y^THY/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)} \\
&= \frac{\sum_{i=1}^k \hat{y}_i^2/k}{\sum_{i=1}^k (y_i-\hat{y}_i)^2/(n-k-1)}
\end{aligned}
$$

对于$Y = \beta_0 + X_1\beta_1 + \cdots + X_k\beta_k$

令$A = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ $U = \begin{bmatrix} A & X \end{bmatrix}$ 其中X为$n \times p$维矩阵

$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$ $\gamma = \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix}$

则$Y = U\gamma$，令$M = [0, I_p]$，其中I_p为p维单位阵。

此时原假设为：$H_0 : \beta = 0$

构造F统计量：

$$F = \frac{(\hat{\gamma}-\gamma_0)^T M^T [M(U^T U)^{-1} M^T]^{-1} M(\hat{\gamma}-\gamma_0)/k}{\hat{\sigma}^T \hat{\sigma}/(n-k-1)}$$

其中

$$[M(U^T U)^{-1} M^T]^{-1} = X^T [I_n - \tfrac{1}{n} J] X$$
$$\hat{\beta} = M(\hat{\gamma} - \gamma_0) = (X^T [I_n - \tfrac{1}{n} J] X)^{-1} X^T [I_n - \tfrac{1}{n} J] Y$$

此时F统计量为：

$$F = \frac{Y^T[I_n - \frac{1}{n}J]X(X^T[I_n - \frac{1}{n}J]X)^{-1}X^T[I_n - \frac{1}{n}J]Y/k}{\hat{\sigma}^T\hat{\sigma}/(n-k-1)}$$

The coefficient of determination, R^2

- The coefficient of determination, R^2 is defined by

$$R^2 = \frac{SSM}{SST} = \frac{SST - RSS(\hat{\beta})}{SST} = 1 - \frac{RSS(\hat{\beta})}{SST}.$$

  It measures the proportional reduction in the total variation when we use the set of variables, described by X, in our linear model.

- $0 \leq R^2 \leq 1$
- A small R^2 does not always imply a bad fit to the model (there may just be many potential predictors which account for the variation in the response).
- For simple linear regression, $R^2 = r^2$, where r is Pearson's correlation coefficient.

The adjusted R^2

- Adding more variables to X will always increase the R^2 value.
- The adjusted R^2 is given by

$$adjR^2 = 1 - \frac{MSE}{MST} = 1 - \left(\frac{n-1}{n-p}\right)\frac{RSS(\hat{\beta})}{SST}.$$

- The adjusted R^2 can become smaller if the variable added does not add "greatly" to the model.

Fitted values and residuals

- The fitted values are

$$\hat{Y} = X\hat{\beta} = X(X^T X)^- XY = HY,$$

  where we let $H = X(X^T X)^- X$ denote the **hat matrix**.

- The residuals are

$$\hat{\epsilon} = Y - \hat{Y} = (I_n - H)Y.$$

- The covariance matrix for the residuals are

$$cov(\hat{\epsilon}) = \hat{\sigma}^2(I_n - H)(I_n - H)^T.$$

  This depends on σ^2. Shows that the residuals are correlated through the hat matrix, H.

Transformed residuals

1. The **standardized residuals** are $\hat{\epsilon}/\hat{\sigma}$.
2. The **studentized residuals** are

$$\frac{\hat{\epsilon}}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

   where h_{ii} is the (i,i) element of H.
3. The **jack-knife or deleted residuals** are

$$\frac{\hat{\epsilon}}{1-h_{ii}},$$

   which is equivalent to calculating a residual for case i from the model which does not include the ith case.
4. The **studentized jack-knife residuals** are

$$\hat{\epsilon}\left(\frac{n-p-1}{RSS(\hat{\beta})(1-h_{ii})+e_i^2}\right)^{1/2}.$$

Residual plots

- Plot residuals versus fitted:

  —Check for appropriateness of the fit.

  —Do we need to transform the response?

  —Check for constancy of the variance of errors.

  —Look for outliners.

- Residuals versus time or collection order.

  —Check for systematic problems in the residuals(e.g., serial correlation).

- Normal Q-Q plot of residuals.

  —Check distribution assumption of the errors.

Residual plots(cont.)

- Plot residuals versus each predictor/covariate in the model.

  —Check adequacy of fit for each predictor.

  —Curvature may indicate the need to transform predictors.

- Plot residuals versus potential predictors NOT in the model.

  —Have any variables been omitted from the model?

Using the leverage values to detect outliners

- We can detect whether a case is outlying with respect to its X values using the diagonal elements of the hat matrix, h_{ii}, $i = 1, ..., n$. They are called the **leverage values**.
- Properties of h_{ii}:
  —$0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^{n} h_{ii} = p$.
- Why is leverage important?
  —The large h_{ii} value, the more important it is in determining $\hat{Y}_i$ from Y_i (remember that $\hat{Y} = HY$).
  —The larger the value of h_{ii}, the smaller the studentized residuals will be. Indicates Y_i and $\hat{Y}_i$ are closer together.

Leverage (cont.)

- Different rules of thumb:

  —A leverage value h_{ii} is considered large relative to the mean leverage value $\bar{h} = \sum_{i=1}^{n} h_{ii}/n = p/n$, if $h_{ii} > 2\bar{h} = 2p/n$.

  —$0.2 < h_{ii} < 0.5$ indicates moderate leverage, and $h_{ii} > 0.5$ indicates high leverage.

  —Look for outlying h_{ii} relative to the other values, $h_{jj}, j \neq i$.

Influential points and outliers

- A case is considered **influential** if its exclusion from the model induces large changes in the fit of the model.
- Not all outlying values are influential.
- Some measures of influence are:
  —DFFITS (measures the influence on a single fitted value).
  —Cook's distance (measures the influence on all fitted values).
  —DFBETAS (measures the influence on the parameters).
- Notation: Let $\hat{Y}_i^{(-j)}$ denote the fitted value for Y_i in a model where case j is deleted.

Influence on a single fitted value

- The influence that case i has on $\hat{Y}_i$ is:

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_i^{(-i)}}{\sqrt{MSE^{(-i)} h_{ii}}}$$

  Here $MSE^{(-i)}$ is the mean squared error in the model with case i removed.

- Rule of thumb: consider case i influential if the absolute value of $(DFFITS)_i$ is larger than 1 for small datasets, and $2\sqrt{p/n}$ for larger datasets.
- Possible to calculate $(DFFITS)_i$ without removing cases (see Neter, et.al.(1996))

Influence on all fitted values

- Cook's distance demonstrates the influence of case i on all the fitted values is

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_j^{(-i)} \right)^2}{pMSE}$$

- Rule of thumb: We compare D_i to a $F(p, n-p)$ distribution. If percentile value near 50% or more consider influential.
- Possible to calculate D_i without removing cases (see Neter, et.al.(1996)).

Influence on parameters.

- To measure the influence of the ith case on the j regression parameter, calculate

$$(DFBETA)_{i,j} = \frac{\hat{\beta}_j - \hat{\beta}_j^{(-i)}}{\sqrt{MSE^{(-i)} c_{jj}}}$$

  Here $\hat{\beta}_j^{(-i)}$ is the estimate of the j parameter when we remove case i from the model, and c_{jj} is the (j,j) element of the $\left(X^T X\right)^{-}$ matrix.

- Rule of thumb: consider case i influential if the absolute value of $(DFBETA)_{i,j}$ is larger than 1 for small datasets, and $2/\sqrt{n}$ for larger datasets.