Applied Linear Regression

杜金鸿,15338039

January 14, 2018

Content

Part I Simple Regression Models

1 Regression Model

1.1 Relations

1. Functional Relation between Two Variables

$$Y = f(X)$$

2. Statistical Relation between Two Variables

$$Y = f(X) + \varepsilon$$

   Statistical relationship generally does not imply causality.

1.2 Definition

1.2.1 Basic Concepts

1. Simple Models

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, 2, \cdots, n$$

   where

   (1) Variables:

   $\varepsilon_i = Y_i - \mathbb{E}Y$: $\mathbb{E}\varepsilon_i = 0$, $Var\varepsilon_i = \mathbb{E}\varepsilon^2 = \sigma^2$.

   Y: response variable, output, dependent variable....

   $e_i = Y_i - \hat{Y}_i$: residual.

   X: known constants, predictor variable, input, independent variable....

   X_i: the ith level of X.

   Y_i: the ith level of Y.

   $Q = \sum_{i=1}^{n} \varepsilon_i^2$: sum of the squared errors.

   $SSE = \sum_{i=1}^{n} e_i^2$: sum of the squared residuals.

   (2) Parameters:

   σ^2: the variance of ε.

   β_0: the intercept of the regression line.

   β_1: the slope of the regression line.

(3) Estimators:

b_0: the estimator of β_0.

b_1: the estimator of β_1.

$\hat{Y}_i = b_0 + b_1 X$: fitted predicted value.

$$Cov(\varepsilon_i, \varepsilon_j) = 0$$

$$\mathbb{E}Y_i = \beta_0 + \beta_1 X_i$$

$$VarY_i = \sigma^2$$

$$Cov(Y_i, Y_j) = 0$$

$$SS_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

$$SS_{XY} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{XY}$$

$$SS_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$$

$$\mathbb{E}(SS_{XY}) = \beta_1 SS_{XX}$$

$$\mathbb{E}(SS_{YY}) = (n-1)\sigma^2 + \beta_1^2 SS_{XX}$$

There are 3 unknown parameters in the model we are interested in - β_0, β_1 and σ^2.

2. Alternative Models

$$Y_i = \beta_0^* + \beta_1(X_i - \overline{X}) + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

where $\beta_0^* = \beta_0 + \beta_1 \overline{X}$.

3. Datas $\begin{cases} \text{Observational Data} \\ \text{Experimental Data} \end{cases}$

1.2.2 Goals

To model a statistical relationship between X and Y.

1. Estimation

2. Prediction

1.2.3 Use of Regression Analysis

1. Description

2. Control

3. Prediction

* Always need to consider scope of the model.

1.3 Kinds

1. Simple Linear Regression Model with Distribution of Error Terms Unspecified - LSE

2. Normal Error Regression Model - MLE

2 Estimators

2.1 Least Square Estimators

2.1.1 Derivation

$$b_0,\ b_1 = \underset{\beta_0,\ \beta_1}{\arg\min}\, Q$$

Let

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = 0 \end{cases}$$

we have

$$\begin{cases} b_0 = \overline{Y} - b_1\overline{X} \\ b_1 = \dfrac{SS_{XY}}{SS_{XX}} \end{cases}$$

Estimation of Error Terms Variance σ^2 is

$$\begin{aligned} s^2 &= \frac{SSE}{n-2} \\ &= MSE \end{aligned}$$

2.1.2 Properties

1. Relationship

$$b_0,\ b_1 = \underset{\beta_0,\ \beta_1}{\arg\min}\, SSE$$

$$\overline{Y} = b_0 + b_1\overline{X}$$

2. $BLUE$ $\boldsymbol{(Gauss - Markov\ Theorem)}$

   Least square estimators b_0 and b_1 are $\boldsymbol{best\ linear\ unbiased\ estimators}$ of β_0 and β_1 respectively.

   (1) Linearity

$$\begin{aligned} b_1 &= \frac{SS_{XY}}{SS_{XX}} \\ &= \sum_{i=1}^{n} k_i Y_i \\ b_0 &= \sum_{i=1}^{n} \left(\frac{1}{n} - k_i\overline{X} \right) Y_i \end{aligned}$$

where $k_i = \dfrac{X_i - \overline{X}}{SS_{XX}}$

(2) Unbiased

$$\mathbb{E}b_0 = \beta_0$$
$$\mathbb{E}b_1 = \beta_1$$

(3) Variance and convariance

$$Varb_0 = \frac{\sum_{i=1}^{n} X_i^2}{nSS_{XX}}\sigma^2$$
$$Varb_1 = \frac{\sigma^2}{SS_{XX}}$$
$$Cov(b_0, b_1) = -\frac{\overline{X}}{SS_{XX}}\sigma^2$$

i.e.

$$Cov(b_0, b_1) = \frac{\sigma^2}{SS_{XX}} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix}$$

3. UE

MSE is an unbiased estimator of σ^2.

2.2 Maximum Likelihood Estimator

2.2.1 Derivation

Let $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \sim N(0, \sigma^2)$. Because of the uncorrelatedness, $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. Then $Y_1, Y_2, \cdots, Y_n \overset{iid}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$. The likelihood

$$L(\beta_0, \beta_1, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 X_i)]^2}$$

and let

$$\begin{cases} \dfrac{\partial \ln L}{\partial \beta_0} = 0 \\ \dfrac{\partial \ln L}{\partial \beta_1} = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = 0 \end{cases}$$

We have

$$\begin{cases} \hat{\beta}_0 = b_0 = \overline{Y} - b_1\overline{X} \\ \hat{\beta}_1 = b_1 = \dfrac{SS_{XY}}{SS_{XX}} \\ \hat{\sigma}^2 = \dfrac{n-2}{n} MSE \end{cases}$$

$\hat{\beta}_0, \hat{\beta}_1$ are the same as the solution of LSE.

2.2.2 Properties

(1) *BLUE*

Maximum likelihood estimators b_0 and b_1 are ***best linear unbiased estimators*** of β_0 and β_1 respectively and

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{SS_{XX}} \begin{pmatrix} \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}\right)$$

From 5.2, we know that $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, which implies the linearity of b_0 (or b_1) and Y_i.

(2) *Asymptotically UE*

Maximum likelihood estimators $\hat{\sigma}^2$ is ***asymptotically unbiased estimators*** of σ^2 and

$$\lim_{n\to+\infty} E\hat{\sigma}^2 = \sigma^2$$

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

(Fisher's Theorem)

(3) Independence

$(\hat{\beta}_0, \hat{\beta}_1, \overline{Y})$ and $\hat{\sigma}^2$(or SSE) are independent.

Proof

3 Inferences

Below we only dicuss with normal error regression model.

3.1 Parameter Inferences

3.1.1 β_1

$\because$

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{XX}}\right)$$

$\therefore$

$$\frac{b_1-\beta_1}{\sigma\{b_1\}} \sim N(0,1)$$

where

$$\sigma\{b_1\} = \sqrt{\frac{\sigma^2}{SS_{XX}}}$$

$\because$

$$\frac{SSE}{\sigma^2} = \frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$$

$$b_1 \perp MSE$$

$\therefore$

$$\frac{\dfrac{b_1-\beta_1}{\sigma\{b_1\}}}{\sqrt{\dfrac{(n-2)MSE}{\sigma^2(n-2)}}} = \frac{b_1-\beta_1}{s\{b_1\}} \sim t(n-2)$$

where

$$s\{b_1\} = \sqrt{\frac{MSE}{SS_{XX}}}$$

(1) Confident Intervals

Given a levle of significance of α,

$$\mathbb{P}\left\{\left|\frac{b_1-\beta_1}{s\{b_1\}}\right| < t\left(1-\frac{\alpha}{2}; n-2\right)\right\} = 1-\alpha$$

(2) Hypothesis Test

The null hypothesis and alternative hypothesis are

$$H_0: \beta_1 = \beta_{10} \qquad H_a: \beta_1 \neq \beta_{10}$$

The test statistic is

$$t^* = \frac{b_1-\beta_{10}}{s\{b_1\}} \overset{H_0}{\sim} t(n-2)$$

The decision rule is

$$\text{If } |t^*| \leqslant t\left(1-\frac{\alpha}{2};n-2\right), \text{ accept } H_0;$$
$$\text{If } |t^*| > t\left(1-\frac{\alpha}{2};n-2\right), \text{ reject } H_0.$$

The P-value is

$$1-\mathbb{P}\{t(n-2) \leqslant |t^*|\}$$

3.1.2 β_0

$\because$

$$b_0 \sim N\left(\beta_0, \sigma^2 \frac{\sum_{i=1}^{n} X_i^2}{nSS_{XX}}\right) = N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{SS_{XX}}\right)\right)$$

$\therefore$

$$\frac{b_0-\beta_0}{\sigma\{b_0\}} \sim N(0,1)$$

where

$$\sigma\{b_0\} = \sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{SS_{XX}}\right)}$$

$\because$

$$\frac{SSE}{\sigma^2} = \frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$$
$$b_0 \perp MSE$$

$\therefore$

$$\frac{\dfrac{b_0-\beta_0}{\sigma\{b_0\}}}{\sqrt{\dfrac{(n-2)MSE}{\sigma^2(n-2)}}} = \frac{b_0-\beta_0}{s\{b_0\}} \sim t(n-2)$$

where

$$s\{b_0\} = \sqrt{MSE\left(\frac{1}{n}+\frac{\overline{X}^2}{SS_{XX}}\right)}$$

(1) Confident Intervals

Given a level of significance of α,

$$\mathbb{P}\left\{\left|\frac{b_0-\beta_0}{s\{b_0\}}\right| < t\left(1-\frac{\alpha}{2};n-2\right)\right\} = 1-\alpha$$

(2) Hypothesis Test

The null hypothesis and alternative hypothesis are

$$H_0: \beta_0 = \beta_{00} \qquad H_a: \beta_0 \neq \beta_{00}$$

The test statistic is

$$t^* = \frac{b_0-\beta_{00}}{s\{b_0\}} \overset{H_0}{\sim} t(n-2)$$

The decision rule is

$$\text{If } |t^*| \leqslant t\left(1-\tfrac{\alpha}{2};n-2\right), \text{ accept } H_0;$$
$$\text{If } |t^*| > t\left(1-\tfrac{\alpha}{2};n-2\right), \text{ reject } H_0.$$

The P-value is

$$1-\mathbb{P}\{t(n-2)\leqslant |t^*|\}$$

3.1.3 Considerations

(1) Effects of Departures from Normality

With sufficiently large samples, the above inferences still apply even if Y depart far from normality and t value can be replaced by z value.

(2) Spacing of the X Levels

The variances of b_0 and b_1 (for a given n and σ^2) depend on the spacing of X. The larger is SS_{XX}, the smaller is the variance.

(3) Power of Tests

The power of β_i $(i=0,1)$ is given by

$$Power = \mathbb{P}\left\{|t^*|>t\left(1-\frac{\alpha}{2};n-2\right)\Big|\delta\right\}$$

where

$$\delta=\frac{|\beta_i-\beta_{i0}|}{\sigma\{b_i\}}$$

is the noncentrality measure, i.e.

$$t^* \overset{H_a}{\sim} t(n-2;\delta)$$

3.2 Other Inferences

3.2.1 $\mathbb{E}Y_h$

Given X_h, a fixed level of X within the scope of the model. (It can be either in the sample or not.)

$$Y_{h(new)}=\beta_0+\beta_1X_h+\varepsilon_{h(new)}$$

(The subscript h may relate to the word "held fixed", meaning that the linear regression model has been fitted.)

$\because$

$$\hat{Y}_h=b_0+b_1X_h$$

$\therefore$

$$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 X_h, \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)\right)$$

Proof

We are interested in the mean response of $\hat{Y}_h$. We have

$$\frac{\hat{Y}_h - \mathbb{E}Y_h}{s\{\hat{Y}_h\}} \sim t(n-2)$$

where

$$s\{\hat{Y}_h\} = \sqrt{MSE\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)}$$

Since

$$\mathbb{P}\left\{\left|\frac{\hat{Y}_h - \mathbb{E}Y_h}{s\{\hat{Y}_h\}}\right| < t\left(1 - \frac{\alpha}{2}; n-2\right)\right\} = 1 - \alpha$$

the confidence interval of $\mathbb{E}(Y_h)$ is given by

$$\left(\hat{Y}_h - t\left(1 - \frac{\alpha}{2}; n-2\right)s\{\hat{Y}_h\}, \hat{Y}_h + t\left(1 - \frac{\alpha}{2}; n-2\right)s\{\hat{Y}_h\}\right)$$

3.2.2 Prediction

(1) Prefiction Interval for $Y_{h(new)}$

When the paramter is known, the prediction interval of $Y_{h(new)}$ is given by

$$\left(\mathbb{E}Y_h - z\left(1 - \frac{\alpha}{2}\right)\sigma, \mathbb{E}Y_h + z\left(1 - \frac{\alpha}{2}\right)\sigma\right)$$

When the parameter is unknown and given X_h, the predicting new (future) observation is

$$Y_{h(new)} = \beta_0 + \beta_1 X_h + \varepsilon_{h(new)}$$

$\because$

$$Y_{h(new)} \sim N(\beta_0 + \beta_1 X_h, \sigma^2)$$
$$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 X_h, \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)\right)$$

$\therefore$

$$\frac{Y_{h(new)} - \hat{Y}_h}{s\{pred\}} \sim t(n-2)$$

where

$$s\{pred\} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)}$$

Since

$$\mathbb{P}\left\{\left|\frac{Y_{h(new)} - \hat{Y}_h}{s\{pred\}}\right| < t\left(1 - \frac{\alpha}{2}; n-2\right)\right\} = 1 - \alpha$$

the prediction interval of $Y_{h(new)}$ is given by

$$\left(\hat{Y}_h - t\left(1 - \frac{\alpha}{2}; n-2\right)s\{pred\}, \hat{Y}_h + t\left(1 - \frac{\alpha}{2}; n-2\right)s\{pred\}\right)$$

The prediction interval of $Y_{h(new)}$ is wider than the confidence interval of $\mathbb{E}Y_h$ in the same significance level.

(2) Prefiction Interval for $\overline{Y}_{h(new)}$

Given m observations in the same level of X, the mean of these m observations is $\overline{Y}_{h(new)}$.

the $1-\alpha$ prediction interval of $\overline{Y}_{h(new)}$ is given by

$$\left(\hat{Y}_h - t\left(1-\frac{\alpha}{2}\right)s\{predmean\}, \hat{Y}_h + t\left(1-\frac{\alpha}{2}\right)s\{predmean\}\right)$$

where

$$\begin{aligned} s\{predmean\} &= \frac{MSE}{m} + s^2\{\hat{Y}_h\} \\ &= MSE\left(\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right) \end{aligned}$$

3.3 Simultaneous Inferences

3.3.1 β_0 and β_1

(1) Bonferroni Correction for Simultaneous Inference

Want to have family confidence level $100(1-\alpha)\%$ or family type I error rate α. Consider g CIs or hypothesis tests, each using $\alpha^* = \frac{\alpha}{g}$.

Use confidence level $1-\frac{\alpha}{g}$ for each CI. Use significance level $\frac{\alpha}{g}$ for each hypothesis test.

Increasingly conservative as g increases.

(2) Joint Confidence Intervals for β_0 and β_1

Family confidence level $100(1-\alpha)\%$ for joint estimation of β_0 and β_1 is given by

$$\left(b_0 - t\left(1-\frac{\alpha}{4}; n-2\right)s\{b_0\}, b_0 + t\left(1-\frac{\alpha}{4}; n-2\right)s\{b_0\}\right)$$
$$\left(b_1 - t\left(1-\frac{\alpha}{4}; n-2\right)s\{b_1\}, b_1 + t\left(1-\frac{\alpha}{4}; n-2\right)s\{b_1\}\right)$$

3.3.2 $\mathbb{E}Y_h$

(1) **Working – Hotelling confidence band**. The confidence interval of $\mathbb{E}(Y_h)$ is given by

$$\left(\hat{Y}_h - t\left(1-\frac{\alpha}{2}; n-2\right)s\{\hat{Y}_h\}, \hat{Y}_h + t\left(1-\frac{\alpha}{2}; n-2\right)s\{\hat{Y}_h\}\right)$$

Replace $t\left(1-\frac{\alpha}{2}; n-2\right)$ with Working-Hotelling value

$$W = \sqrt{2F(1-\alpha; 2, n-2)}$$

and get the simultaneous confidence band at $(1-\alpha)$ level

$$(\hat{Y}_h - Ws\{\hat{Y}_h\}, \hat{Y}_h + Ws\{\hat{Y}_h\})$$

Working-Hotelling confidence band is the narrowest at $\overline{X}$.

Working-Hotelling confidence band is narrower than prdiction Intervals and wider than confidence intervals for $\mathbb{E}Y_h$ at the same confidence level.

(2) **Bonferroni Method**. The Bonferroni simultaneous prediction limits at $(1-\alpha)$ level for g $\mathbb{E}Y_h$ are given by

$$(\hat{Y}_h - Bs\{\hat{Y}_h\}, \hat{Y}_h + Bs\{\hat{Y}_h\})$$

where

$$B = t\left(1 - \frac{\alpha}{2g}; n-2\right)$$

3.3.3 Prediction

(1) **Scheffé Simultaneous Prediction Procedure**. The simultaneous confidence limits for g $Y_{h(new)}$ is given by

$$\left(\hat{Y}_h - Ss\{pred\}, \hat{Y}_h + Ss\{pred\}\right)$$

where

$$S = gF\left(1-\alpha; g, n-2\right)$$

(2) **Bonferroni Method**. The Bonferroni simultaneous preidction limits at $(1-\alpha)$ level for g $\mathbb{E}Y_h$ are given by

$$(\hat{Y}_h - Bs\{pred\}, \hat{Y}_h + Bs\{pred\})$$

where

$$B = t\left(1 - \frac{\alpha}{2g}; n-2\right)$$

3.4 Variance Analysis

3.4.1 F test

From

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$
$$SSTO = SSE + SSR$$

we have

(1) $SSTO$: the total sum of squares

$$\frac{SSTO}{\sigma^2} \sim \chi^2(n-1, \delta)$$

(2) SSE : Error (unexplained / residual)

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

(3) SSR : Model (explained by regression)

$$\begin{aligned} SSR &= \sum_{i=1}^{n}(b_0 + b_1 X_i - b_0 - b_1\overline{X})^2 \\ &= b_1^2 SS_{XX} \end{aligned}$$

$$SSR \perp SSE$$

and

$$\frac{SSR}{\sigma^2} \sim \chi^2(1,\delta)$$

where

$$\delta = \frac{\beta_1^2}{\frac{\sigma^2}{SS_{XX}}}$$

The mean squares is given by

$$MSE = \frac{SSE}{n-2}$$
$$MSR = \frac{SSR}{1}$$

Then

$$F^* = \frac{MSR}{MSE} \sim F(1,n-2,\delta)$$

F-test for

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0$$

The ANOVA test Statistic is

$$F^* \overset{H_0}{\sim} F(1,n-2)$$

When H_0 is false, MSR is much bigger than MSE.

Desicion rule is

$$\text{If } F^* \leqslant F(1-\alpha;1,n-2), \text{ accept } H_0;$$
$$\text{If } F^* > F(1-\alpha;1,n-2), \text{ reject } H_0.$$

The p-value is

$$\mathbb{P}\{F(1,n-2) > F^*\}$$

3.4.2 Equivalence of F test and two-sided t test

When $\beta_{10} = 0$, F test and t test are equivalent.

Equivalence of test statistics

$$F^* = t^{*2}$$

$$t^2(n-2) \sim F(1,n-2)$$
$$t^2\left(1-\frac{\alpha}{2};n-2\right) = F(1-\alpha;1,n-2)$$

Equivalence of rejection regions

$$F^* > F(1-\alpha;1,n-2) \Longleftrightarrow |t^*| > t\left(1-\frac{\alpha}{2};n-2\right)$$

Equivalence of p values

$$\mathbb{P}\{F(1,n-2) > F^*\} = \mathbb{P}\{t(n-2) > |t^*|\}$$

ANOVA table is given by

Source of Variation	SS	df	MS	$\mathbb{E}MS$
Regression	$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 SS_{XX}$
Error	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$	$n-1$	$MSTO = \frac{SSTO}{n-1}$	$\sigma^2 + \frac{\beta_1^2}{n-1} SS_{XX}$

3.5 General Linear Test

The full model or unrestricted model have more parameters than the reduced model or restricted model. The null hypothesis and alternative hypothesis are

$$H_0 : \text{Reduced model} \qquad H_a : \text{Full model}$$

The test statistic is

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} \overset{H_0}{\sim} F(df_R - df_F, df_F)$$

The decision rule is

$$\begin{aligned} &\text{If } F^* \leqslant F(1-\alpha; df_R - df_F, df_F), \text{ accept } H_0; \\ &\text{If } F^* > F(1-\alpha; df_R - df_F, df_F), \text{ reject } H_0. \end{aligned}$$

The P-value is

$$\mathbb{P}\{F(df_R - df_F, df_F) > F^*\}$$

3.6 Correlation Analysis

3.6.1 Coefficient of Determination

The proportion of total variation in Y explained by X.

$$\begin{aligned} R^2 &= \frac{SSR}{SSTO} \\ &= 1 - \frac{SSE}{SSTO} \end{aligned}$$

$$0 \leqslant R^2 \leqslant 1$$

Limitation

(1) High R^2 does not necessarily mean that useful predictions can be made or regression line is a good fit.

(2) Low R^2 does not necessarily mean that X and Y are not related.

3.6.2 Coefficient of Correlation

Pearson's product-moment correlation coefficient measures the strength of the linear relationship between two variables

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

The sample coefficient of correlation is defined by

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}}$$

$$-1 \leqslant r \leqslant 1$$

For simple regression model,

$$r = \pm\sqrt{R^2}$$

where the sign is the same as b_1. Proof

3.6.3 Normal Correlation Model

Normal correlation model uses bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho_{12})$

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} e^{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]}$$

The marginal distributions are

$$Y_1 \sim N(\mu_1, \sigma_1^2)$$
$$Y_2 \sim N(\mu_2, \sigma_2^2)$$

The conditional distributions are

$$Y_1|Y_2 = y_2 \sim N(\alpha_{1|2} + \beta_{12}y_2, \sigma_{1|2}^2)$$
$$Y_2|Y_1 = y_2 \sim N(\alpha_{2|1} + \beta_{21}y_1, \sigma_{2|1}^2)$$

where

$$\alpha_{1|2} = \mu_1 - \mu_2\rho_{12}\frac{\sigma_1}{\sigma_2}$$
$$\alpha_{2|1} = \mu_2 - \mu_1\rho_{12}\frac{\sigma_2}{\sigma_1}$$
$$\beta_{12} = \rho_{12}\frac{\sigma_1}{\sigma_2}$$
$$\beta_{21} = \rho_{12}\frac{\sigma_2}{\sigma_1}$$
$$\sigma_{1|2}^2 = \sigma_1^2(1-\rho_{12}^2)$$
$$\sigma_{2|1}^2 = \sigma_2^2(1-\rho_{12}^2)$$

The inferences about the parametres in conditional distribution are the same as simple linear regression model.

Inferences on ρ_{12}.

(1) Confident Intervals

Make the **Fisher z transformation**

$$z' = \frac{1}{2}\ln\left(\frac{1+r_{12}}{1-r_{12}}\right)$$

When n is large ($n > 25$), approximately,

$$z' \dot{\sim} N\left(\xi, \frac{1}{n-3}\right)$$

where

$$\xi = \frac{1}{2}\ln\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right)$$

Since $\rho_{12} = \dfrac{e^{2\xi}-1}{e^{2\xi}+1}$ increases as ξ increases, we can first seek for the confidence interval of ξ,

$$\left(z' - z\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{1}{n-3}}, z' + z\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{1}{n-3}}\right) = (c_1, c_2)$$

Then the $100(1-\alpha)\%$ confidence interval for ρ_{12} is

$$\left(\frac{e^{2c_1}-1}{e^{2c_1}+1}, \frac{e^{2c_2}-1}{e^{2c_2}+1}\right)$$

(2) Hypothesis Test

Under the bivariate normal model, the MLE of ρ_{12} is

$$\hat{\rho}_{12} = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} = r$$

The null hypothesis and alternative hypothesis are

$$H_0 : \rho_{12} = 0 \qquad H_a : \rho_{12} \neq 0$$

The test statistic is

$$t^* = \frac{r_{12}}{\sqrt{\frac{1-r_{12}^2}{n-2}}} = \frac{b_1}{s\{b_1\}} \overset{H_0}{\sim} t(n-2)$$

Proof

The decision rule is

$$\text{If } |t^*| \leqslant t\left(1-\tfrac{\alpha}{2}; n-2\right), \text{ accept } H_0;$$
$$\text{If } |t^*| > t\left(1-\tfrac{\alpha}{2}; n-2\right), \text{ reject } H_0.$$

The P-value is

$$\mathbb{P}\{t(n-2) > |t^*|\}$$

3.6.4 Spearman's Correlation Method

For non-normal model, we can use Spearman's correlation method. Given two groups of data,

(1) Rank $(Y_{11},\cdots,Y_{n1})$ from 1 to n (smallest to largest) and label: $(R_{11},\cdots,R_{n1})$.

(2) Rank $(Y_{12},\cdots,Y_{n2})$ from 1 to n (smallest to largest) and label: $(R_{12},\cdots,R_{n2})$.

(3) Compute Spearman's rank correlation coefficient:

$$r_S = \frac{\sum\limits_{i=1}^{n}(R_{i1}-\overline{R}_1)(R_{i2}-\overline{R}_2)}{\sqrt{\left[\sum\limits_{i=1}^{n}(R_{i1}-\overline{R}_1)^2\right]\left[\sum\limits_{i=1}^{n}(R_{i2}-\overline{R}_2)^2\right]}}$$

The null hypothesis and alternative hypothesis are

$$H_0: \text{No association between } Y_1, Y_2 \qquad H_a: \text{Association exists}$$

The test statistic is

$$t^* = \frac{r_S}{\sqrt{\frac{1-r_S^2}{n-2}}} \mathrel{\dot\sim} t(n-2)$$

when H_0 holds.

The decision rule is

$$\begin{aligned}&\text{If } |t^*| \leqslant t\left(1-\tfrac{\alpha}{2};n-2\right), \text{ accept } H_0;\\ &\text{If } |t^*| > t\left(1-\tfrac{\alpha}{2};n-2\right), \text{ reject } H_0.\end{aligned}$$

The P-value is

$$\mathbb{P}\{t(n-2) > |t^*|\}$$

3.7 Regression Through the Origin

Model

$$Y_i = \beta_1 X_i + \varepsilon_i$$

where $\varepsilon_1,\cdots,\varepsilon_n \overset{i.i.d.}{\sim} N(0,\sigma^2)$. The LSE or MLE of β_1 is

$$b_1 = \frac{\sum\limits_{i=1}^{n} X_i Y_i}{\sum\limits_{i=1}^{n} X_i^2} \sim N\left(\beta_1, \frac{\sigma^2}{\sum\limits_{i=1}^{n} X_i^2}\right)$$

we have

$$\begin{aligned}MSE &= \frac{SSE}{n-1}\\ s^2\{b_1\} &= \frac{MSE}{\sum\limits_{i=1}^{n} X_i^2}\end{aligned}$$

$$s^2\{\hat{Y}_h\} = \frac{X_h^2}{\sum_{i=1}^{n} X_i^2} MSE$$

$$s^2\{pred\} = \left(1 + \frac{X_h^2}{\sum_{i=1}^{n} X_i^2}\right) MSE$$

(1) Confident Intervals

Given a level of significance of α, the $(1-\alpha)100\%$ confident interval for β_1 is

$$\left(b_1 - t\left(1-\frac{\alpha}{2};n-1\right)s\{b_1\}, b_1 + t\left(1-\frac{\alpha}{2};n-1\right)s\{b_1\}\right)$$

The $(1-\alpha)100\%$ confident interval for $\mathbb{E}Y_h$ is

$$\left(\hat{Y}_h - t\left(1-\frac{\alpha}{2};n-1\right)s\{\hat{Y}_h\}, \hat{Y}_h + t\left(1-\frac{\alpha}{2};n-1\right)s\{\hat{Y}_h\}\right)$$

The $(1-\alpha)100\%$ confident interval for $Y_{h(new)}$ is

$$\left(\hat{Y}_h - t\left(1-\frac{\alpha}{2};n-1\right)s\{pred\}, \hat{Y}_h + t\left(1-\frac{\alpha}{2};n-1\right)s\{pred\}\right)$$

(2) Hypothesis Test

The null hypothesis and alternative hypothesis are

$$H_0 : \beta_1 = \beta_{10} \qquad H_a : \beta_1 \neq \beta_{10}$$

The test statistic is

$$t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}} \overset{H_0}{\sim} t(n-1)$$

The decision rule is

$$\text{If } |t^*| \leqslant t\left(1-\tfrac{\alpha}{2};n-1\right), \text{ accept } H_0;$$
$$\text{If } |t^*| > t\left(1-\tfrac{\alpha}{2};n-1\right), \text{ reject } H_0.$$

The P-value is

$$\mathbb{P}\{t(n-1) > |t^*|\}$$

4 Diagnostics and Remedial Measures

4.1 Graphics Diagnostics for Predictor Variables

Graphical Diagnostics for X:

(1) Dot plot

(2) Histogram or stem-and-leaf plot

(3) Box plot

(4) Sequence plot

4.2 Graphics Diagnostics for Residuals

4.2.1 Residuals

The residuals are defined by

$$e_i = Y_i - \hat{Y}_i$$

The studentized residuals and semi-studentized residuals are defined later.

The properties for residuals:

(1) $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} \hat{Y}_i e_i = 0$

(2) For simple normal regression model,

$$e_i \sim N\left(0, (1-h_{ii})\sigma^2\right)$$
$$Cov(e_i, e_j) = -h_{ij}\sigma^2$$

where

$$h_{ij} = \frac{1}{n} + \frac{(X_i - \overline{X})(X_j - \overline{X})}{SS_{XX}}$$

4.2.2 L.I.N.E.

To be studied by residuals,

(1) Linearity of the regression function

Plot Y versus X. Random cloud around regression line indicates linear relation while U-shape or inverted U-shape indicates nonlinear relation.

Plot residuals versus X. Random cloud around $e = 0$ indicates linear relation while U-shape or inverted U-shape indicates nonlinear relation.

(2) Independence of the error terms

Sequential observations can exhibit observable trends in error distribution.

(3) Normality of the error terms

Distribution plots of residuals.

Check proportion that lie within 1 standard deviation.

Normal probability plot of residuals like **Normal Quantile – Quantile(Q – Q) plot**.

(a) Sort the residuals by ascending order and get the sample quantile $e_{(1)}, \cdots, e_{(n)}$

(b) Let

$$p_i = \frac{i-a}{n+1-2a}$$

where $a \in [0, \frac{1}{2}]$. In R, $a = \begin{cases} \frac{3}{8} & , n \leqslant 10 \\ \frac{1}{2} & , n > 10 \end{cases}$.

(c) Calculate the theoretical normal quantile

$$q_i = \Phi^{-1}(p_i)$$

where $\Phi(x)$ is the distribution function of standard normal distribution.

(d) Plot $(q_i, e_{(i)})$. If the points are almost lying at the line $y = x$, then the departures from normality are small.

Equivalently, we can calculate the expected value of residuals by

$$e'_i = \sqrt{MSE} z(p_i)$$

and plot (e'_i, e_i).

(4) Equality of variance of the error terms

Plot residuals versus X. Funnel shape indicates error terms have non-constant variance.

Plot absolute/squared residuals versus X. Positive association indicates error terms have non-constant variance.

4.2.3 Outliers

The point (X_i, Y_i) is an outliers if the semi-studentized residuals

$$e_i^* > 4$$

.

4.3 Statistics Tests involving Residuals

4.3.1 Tests for Linearity

Testing lack of fit. Assuming that the observations Y for given X are independent, normally distributed and the distributions of Y have the same variance σ^2. It requires repeat obeservations at one or more X levels.

The null hypothesis and alternative hypothesis are

$$H_0 : \mathbb{E}Y_i = \beta_0 + \beta_1 X_i \qquad H_a : \mathbb{E}Y_i = \mu_i \neq \beta_0 + \beta_1 X_i$$

(1) Define X levels as $X_1, \cdots, X_c$ with n_j replicates respectively and $\sum\limits_{j=1}^{c} n_j = n$. Y_{ij} is the ith replicate at X_j.

(2) The SSE for the reduced model is

$$SSE = SSE(R) = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_{ij})^2 \qquad df_R = n-2$$

The SSE for the full model is

$$SSPE = SSE(F) = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2 \qquad df_F = n-c$$

Since

$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\text{Error Deviation}} = \underbrace{Y_{ij} - \overline{Y}_j}_{\text{Pure Error Deviation}} + \underbrace{\overline{Y}_j - \hat{Y}_{ij}}_{\text{Lack of Fit Deviation}}$$

we have

$$\begin{aligned} SSE &= SSPE + SSLE \\ SSLE &= \sum_{j=1}^{c}\sum_{i=1}^{n_c}(\overline{Y}_j - \hat{Y}_{ij})^2 \\ &= \sum_{j=1}^{c} n_j(\overline{Y}_j - \hat{Y}_j)^2 \end{aligned}$$

Since all Y_{ij} obeservations at the level X_j have the same fitted value $\hat{Y}_j$.

The test statistic is

$$\begin{aligned} F^* &= \frac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{\dfrac{SSE(F)}{df_F}} \\ &= \frac{MSLE}{MSPE} \overset{H_0}{\sim} F(c-2, n-c) \end{aligned}$$

The decision rule is

$$\text{If } F^* \leqslant F(1-\alpha; c-2, n-c), \text{ accept } H_0;$$
$$\text{If } F^* > F(1-\alpha; c-2, n-c), \text{ reject } H_0.$$

The P-value is

$$\mathbb{P}\{F(c-2, n-c) > F^*\}$$

ANOVA table is given by

Source of Variation	SS	df	MS	$\mathbb{E}MS$
Regression	$SSR = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(\hat{Y}_{ij}-\overline{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2$
Error	$SSE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij}-\hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	$\sigma^2 + \frac{c-2}{n-2}\sum_{j=1}^{c} n_j[\mu_j-(\beta_0+\beta_1X_j)]^2$
Lack of Fit	$SSLE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(\overline{Y}_j-\hat{Y}_{ij})^2$	$c-2$	$MSLE = \frac{SSLE}{c-2}$	$\sigma^2 + \frac{\sum_{j=1}^{c} n_j[\mu_j-(\beta_0+\beta_1X_j)]^2}{c-2}$
Pure Error	$SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$	σ^2
Total	$SSTO = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y})^2$	$n-1$		

Noting that MSE is no longer an unbiased estimator of σ^2 without linearity assumption of the model. That means when H_a holds, the model is not linear and therefore $\mathbb{E}MSE > \sigma^2$.

4.3.2 Tests for Normality of Error

Correlation Test

(1) Calculate the correlation between observed residuals and expected value of residuals.

(2) Compare correlation with critical value based on α-level. For $\alpha = 0.05$, critical value is $1.02 - \frac{1}{\sqrt{10n}}$.

(3) Reject the null hypothesis of normal errors if correlation is smaller than the critical value.

Shapiro – Wilk Test

4.3.3 Tests for Constancy of Error Variance

Brown – Forsythe Test. The null hypothesis and alternative hypothesis are

$$H_0: \text{the error variance is constant} \qquad H_a: \text{the error variance is not constant}$$

(1) Devide dataset into 2 groups based on levels with sample size n_1, n_2. Compute

$$d_{ij} = |e_{ij} - \widetilde{e}_j| \qquad j = 1, 2$$

where $\widetilde{e}_j$ is the median of the jth group.

(2) Compute the test statistic

$$t_{BF}^* = \frac{\overline{d}_1 - \overline{d}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \stackrel{.}{\sim} t(n_1+n_2-2)$$

when H_0 holds, where

$$\overline{d}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} d_{ij}$$

$$s_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(d_{ij}-\overline{d}_j)^2$$

$$s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

(3) The decision rule is

$$\text{If } |t_{BF}^*| < t\left(1-\frac{\alpha}{2}; n_1+n_2-2\right), \text{ accept } H_0;$$
$$\text{If } |t_{BF}^*| \geqslant t\left(1-\frac{\alpha}{2}; n_1+n_2-2\right), \text{ reject } H_0;$$

Breusch – Pagan Test or **Cook – Weisberg Test**. For the normal regression model, Assuming that the error terms are independent and normally distributed and

$$\ln \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \cdots + \gamma_p X_{ip}$$

Testing constance of error variance is equivalent to test

$$H_0 : \gamma_1 = \cdots = \gamma_p = 0 \qquad H_a : \gamma_i \neq 0$$

(1) Compute $SSE = \sum_{i=1}^{n} e_i^2$ from the original regression.

(2) Fit regression of e_i^2 on $X_{i1}, \cdots, X_{ip}$ and obtain SSR^*. Test statistic is

$$X_{BP}^2 = \frac{\frac{SSR^*}{2}}{\left(\frac{SSE}{n}\right)^2} \dot{\sim} \chi^2(p)$$

The decision rule is

$$\text{If } X_{BP}^2 \leqslant \chi^2(1-\alpha; p), \text{ accept } H_0;$$
$$\text{If } X_{BP}^2 > \chi^2(1-\alpha; p), \text{ reject } H_0;$$

4.4 Remedial Measures

4.4.1 Overview

If the simple regression model is not appropriate, then

(1) Use other appropriate model.

(2) Employ some transformation.

4.4.2 Box Cox Transforms

Consider the power transformation

$$Y' = \begin{cases} Y^\lambda & ,\lambda \neq 0 \\ \ln Y & ,\lambda = 0 \end{cases}$$

The model becomes

$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \varepsilon_i$$

(1) Standardize Y_i by

$$W_i = \begin{cases} K_1(Y_i^{\lambda} - 1) & ,\lambda \neq 0 \\ K_2 \ln Y_i & ,\lambda = 0 \end{cases}$$

where

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}$$
$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

(2) Compute the SSE for some choices of λ and choose the best model.

5 Matrix Approach to Simple Linear Regression

5.1 Special Matrixes

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & & & \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

for constant matrix **A** and random design matrix

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

$$\mathbb{E}\{\mathbf{AX}\} = \mathbf{A}\mathbb{E}\mathbf{X}$$

$$\sigma^2\{\mathbf{AX}\} = \mathbf{A}\mathbb{E}\mathbf{X}\mathbf{A}^T$$

5.2 Simple Linear Regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{SS_{XX}} \begin{bmatrix} \frac{SS_{XX}}{n} + \overline{X}^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{Xb} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{HY} \end{aligned}$$

where the hat matrix is given by

$$\begin{aligned}
\mathbf{H} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
\mathbf{e} &= (\mathbf{I}-\mathbf{H})\mathbf{Y} \\
\sigma^2\{\mathbf{e}\} &= \sigma^2(\mathbf{I}-\mathbf{H}) \\
s^2\{\mathbf{e}\} &= MSE(\mathbf{I}-\mathbf{H})
\end{aligned}$$

5.3 Hat Matrix

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

We will discuss hat matrix for multiple linear regression later in 6.1.2.

5.4 Analysis of Variance

$$\begin{aligned}
SSTO &= \sum_{i=1}^{n}(Y_i-\overline{Y})^2 \\
&= \mathbf{Y}^T\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y} \\
SSE &= \mathbf{e}^T\mathbf{e} \\
&= \mathbf{Y}^T\left(\mathbf{I}-\mathbf{H}\right)\mathbf{Y} \\
SSR &= \mathbf{Y}^T\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y}
\end{aligned}$$

we have

$$\begin{aligned}
rank\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right) &= n-1 \\
rank\left(\mathbf{I}-\mathbf{H}\right) &= n-2 \\
rank\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right) &= 1
\end{aligned}$$

therefore,

$$\begin{aligned}
\frac{SSTO}{\sigma^2} &\sim \chi^2(n-1,\delta) \\
\frac{SSE}{\sigma^2} &\sim \chi^2(n-2) \\
\frac{SSR}{\sigma^2} &\sim \chi^2(1,\delta)
\end{aligned}$$

where

$$\begin{aligned}
\delta &= \frac{1}{\sigma^2}(\mathbf{X}\boldsymbol{\beta})^T\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)(\mathbf{X}\boldsymbol{\beta}) \\
&= \frac{\beta_1^2}{\frac{\sigma^2}{SS_{XX}}}
\end{aligned}$$

5.5 $\boldsymbol{\beta}$

$\because$

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

$\therefore$

$$s^2\{\mathbf{b}\} = MSE(\mathbf{X}^T\mathbf{X})^{-1}$$

5.6 $\mathbb{E}Y_h$

$\because$

$$\hat{Y}_h \sim N(\mathbb{E}Y_h, \sigma^2\mathbf{X_h}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X_h})$$

$\therefore$

$$s^2\{Y_h\} = MSE\mathbf{X_h}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X_h}$$

5.7 $Y_{n(new)}$

$\because$

$$\begin{aligned}
\hat{Y}_h &\sim N(\mathbb{E}Y_h, \sigma^2\mathbf{X_h}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X_h}) \\
Y_{h(new)} &\sim N(\mathbb{E}Y_h, \sigma^2) \\
\hat{Y}_h &\perp Y_{h(new)}
\end{aligned}$$

$\therefore$

$$\begin{aligned}
\hat{Y}_h - Y_{h(new)} &\sim N(0, \sigma^2[1 + \mathbf{X_h}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X_h}]) \\
s^2\{pred\} &= MSE[1 + \mathbf{X_h}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X_h}]
\end{aligned}$$

Part II Multiple Regression Models

6 Multiple Regression

6.1 General Linear Regression Model

6.1.1 Definition

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots, \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

or

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1,p-1} \\ 1 & X_{21} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{n,p-1} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Both qualitative predictor variable and quantitive predictor variable can be in a general regression model.

Polynomial regression (with interaction effects) can be transformed to be the general regression model by substituting some terms with nex predictor variables.

Linearity means $\mathbb{E}Y$ can be express as the linear combination of the parameters in the model.

6.1.2 Hat Matrix

6.1.3 Definition

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

$$h_{ij} = X_i^T(\mathbf{X}^T\mathbf{X})^{-1}X_j$$

where

$$X_i = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}$$

6.1.4 Properties

(1) Projection

$$\mathbf{HY} = \hat{\mathbf{Y}}$$

$$\mathbf{HX} = \mathbf{X}$$
$$\mathbf{He} = \mathbf{0}$$

(2) Symmetric

$$\mathbf{H}^T = \mathbf{H}$$
$$(\mathbf{I} - \mathbf{H})^T = \mathbf{I} - \mathbf{H}$$

Proof

(3) Idempotent

$$\mathbf{HH} = \mathbf{H}$$
$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H}$$

Proof

We discuss hat matrix later in outlying X cases.

Suppose that X is centered. Let

$$\mathbf{X}_r = \begin{bmatrix} X_{11} & \cdots & X_{1,p-1} \\ X_{21} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots \\ X_{n1} & \cdots & X_{n,p-1} \end{bmatrix}$$
$$\mathbf{H}_r = \mathbf{X}_r(\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r^T$$
$$\mathbf{H}_0 = \frac{1}{n}\mathbf{1}\mathbf{1}^T$$

we have

$$\mathbf{H} = \mathbf{H}_r + \mathbf{H}_0$$

Proof

6.1.5 Residuals

We discuss residuals later in outlying Y observations.

6.2 Analysis of Variance

ANOVA table is given by

Source of Variation	SS	df	MS	$\mathbb{E}MS$
Regression	$SSR = \mathbf{Y}^T\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$	$p-1$	$MSR = \frac{SSE}{p-1}$	$\sigma^2 + \frac{1}{p-1}\sum_{i=1}^{p-1}\sum_{j=1}^{p-1} SS_{ij}\beta_i\beta_j$
Error	$SSE = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$	σ^2
Total	$SSTO = \mathbf{Y}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$	$n-1$	$MSTO = \frac{SSTO}{n-1}$	

where

$$SS_{ij} = \sum_{m=1}^{n} (X_{mi} - \overline{X}_i)(X_{mj} - \overline{X}_j)$$

we have

$$\frac{SSR}{\sigma^2} \sim \chi^2(p-1, \delta)$$
$$\frac{SSE}{\sigma^2} \sim \chi^2(n-p)$$
$$\frac{SSTO}{\sigma^2} \sim \chi^2(n-1, \delta)$$

and

$$\delta = \frac{1}{\sigma^2} \sum_{i=1}^{p-1} \sum_{j=1}^{p-1} SS_{ij} \beta_i \beta_j$$

The F statistic is given by

$$F^* = \frac{MSR}{MSE}$$

F-test for general linear regression model

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \qquad H_a : \text{not all } \beta_k (k = 1, \cdots, p-1) \text{ euqal } 0$$

The ANOVA test Statistic is

$$F^* \overset{H_0}{\sim} F(p-1, n-p)$$

When H_0 is false, MSR is much bigger than MSE.

Desicion rule is

$$\text{If } F^* \leqslant F(1-\alpha; p-1, n-p), \text{ accept } H_0;$$
$$\text{If } F^* > F(1-\alpha; p-1, n-p), \text{ reject } H_0.$$

The p-value is

$$\mathbb{P}\{F(p-1, n-p) > F^*\}$$

6.3 Correlation Analysis

6.3.1 Coefficient of Multiple Determination

The coefficient of multiple determination is given by

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

The adjusted coefficient of multiple determination is given by

$$R_a^2 = 1 - \frac{MSE}{MSTO}$$

6.3.2 Coefficient of Multiple Correlation

The coefficient of multiple correlation is given by

$$R = \sqrt{R^2}$$

For the simple regression model,

$$R = |r|$$

where r is the correlation coefficients.

6.4 Inference about Regression Parameters

6.4.1 β_k

$\because$

$$\mathbf{b} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

$\therefore$

$$\frac{b_k - \beta_k}{\sigma\{b_k\}} \sim N(0,1)$$

where

$$\sigma^2\{b_k\} = \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{k+1,k+1}$$

$\therefore$

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim t(n-p)$$

where

$$s^2\{b_0\} = MSE[(\mathbf{X}^T\mathbf{X})^{-1}]_{k+1,k+1}$$

(1) Confident Intervals

Given a level of significance of α,

$$\mathbb{P}\left\{\left|\frac{b_k - \beta_k}{s\{b_k\}}\right| < t\left(1 - \frac{\alpha}{2}; n-p\right)\right\} = 1 - \alpha$$

(2) Hypothesis Test

The null hypothesis and alternative hypothesis are

$$H_0 : \beta_k = \beta_{k0} \qquad H_a : \beta_k \neq \beta_{k0}$$

The test statistic is

$$t^* = \frac{b_k - \beta_{k0}}{s\{b_k\}} \overset{H_0}{\sim} t(n-p)$$

The decision rule is

$$\text{If } |t^*| \leqslant t\left(1 - \tfrac{\alpha}{2}; n-p\right), \text{ accept } H_0;$$
$$\text{If } |t^*| > t\left(1 - \tfrac{\alpha}{2}; n-p\right), \text{ reject } H_0.$$

The P-value is

$$1 - \mathbb{P}\{t(n-p) \leqslant |t^*|\}$$

6.4.2 Joint Inferences

The Bonferroni simultaneous confidence limits for g parameters with family confidence coefficient $1-\alpha$ are given by

$$(b_k - Bs\{b_k\}, b_k + Bs\{b_k\})$$

where

$$B = t\left(1 - \frac{\alpha}{2g}; n-p\right)$$

6.5 Other Inferences

6.5.1 $\mathbb{E}Y_h$

(1) Confident Interval for $\mathbb{E}Y_h$

Given X_h, a fixed level of X within the scope of the model. (It can be either in the sample or not.)

$$Y_{h(new)} = \mathbf{X}_h\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(The subscript h may relate to the word "held fixed", meaning that the linear regression model has been fitted.) where

$$\mathbf{X}_h = \begin{bmatrix} 1 & X_{h1} & X_{h2} & \cdots & X_{h,p-1} \end{bmatrix}$$

$\because$

$$\hat{Y}_h = \mathbf{X}_h\mathbf{b}$$

$\therefore$

$$\hat{Y}_h \sim N\left(\mathbf{X}_h\boldsymbol{\beta}, \sigma^2\mathbf{X}_h(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h^T\right)$$

We are interested in the mean response of $\hat{Y}_h$. We have

$$\frac{\hat{Y}_h - \mathbb{E}Y_h}{s\{\hat{Y}_h\}} \sim t(n-p)$$

where

$$\begin{aligned} s\{\hat{Y}_h\} &= \sqrt{MSE\mathbf{X}_h(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h^T} \\ &= \sqrt{\mathbf{X}_h s^2\{\mathbf{b}\}\mathbf{X}_h^T} \end{aligned}$$

Since

$$\mathbb{P}\left\{\left|\frac{\hat{Y}_h - \mathbb{E}Y_h}{s\{\hat{Y}_h\}}\right| < t\left(1-\frac{\alpha}{2}; n-p\right)\right\} = 1-\alpha$$

the confidence interval of $\mathbb{E}(Y_h)$ is given by

$$\left(\hat{Y}_h - t\left(1-\frac{\alpha}{2}; n-2\right)s\{\hat{Y}_h\}, \hat{Y}_h + t\left(1-\frac{\alpha}{2}; n-2\right)s\{\hat{Y}_h\}\right)$$

(2) Bonferroni Simultaneous Confidence Intervals for several $\mathbb{E}Y_h$ in g different levels

$$\left(\hat{Y}_h - Bs\{\hat{Y}_h\}, \hat{Y}_h + Bs\{\hat{Y}_h\}\right)$$

where

$$B = t\left(1 - \frac{\alpha}{2g}; n-2\right)$$

(3) Working-Hotelling Simultaneous Confidence Region Bounds for several $\mathbb{E}Y_h$ in g different levels

$$\left(\hat{Y}_h - Ws\{\hat{Y}_h\}, \hat{Y}_h + Ws\{\hat{Y}_h\}\right)$$

where

$$W = \sqrt{pF\left(1-\alpha; p, n-p\right)}$$

6.5.2 Prediction

(1) Prediction Interval for $Y_{h(new)}$

When the paramter is known, the prediction interval of $Y_{h(new)}$ is given by

$$\left(\mathbb{E}Y_h - z\left(1-\frac{\alpha}{2}\right)\sigma, \mathbb{E}Y_h + z\left(1-\frac{\alpha}{2}\right)\sigma\right)$$

When the parameter is unknown and given X_h, the predicting new (future) observation is

$$Y_{h(new)} = \mathbf{X}_h\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{h(new)}$$

$\because$

$$Y_{h(new)} \sim N(\beta_0 + \beta_1 X_h, \sigma^2)$$
$$\hat{Y}_h \sim N\left(\mathbf{X}_h\boldsymbol{\beta}, \sigma^2\mathbf{X}_h(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h^T\right)$$

$\therefore$

$$\frac{Y_{h(new)} - \hat{Y}_h}{s\{pred\}} \sim t(n-p)$$

where

$$s\{pred\} = \sqrt{MSE\left(1 + \mathbf{X}_h(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h^T\right)}$$

Since

$$\mathbb{P}\left\{\left|\frac{Y_{h(new)} - \hat{Y}_h}{s\{pred\}}\right| < t\left(1-\frac{\alpha}{2}; n-p\right)\right\} = 1-\alpha$$

the prediction interval of $Y_{h(new)}$ is given by

$$\left(\hat{Y}_h - t\left(1-\frac{\alpha}{2}; n-p\right)s\{pred\}, \hat{Y}_h + t\left(1-\frac{\alpha}{2}; n-p\right)s\{pred\}\right)$$

The prediction interval of $Y_{h(new)}$ is wider than the confidence interval of $\mathbb{E}Y_h$ in the same significance level.

(2) Bonferroni Simultaneous Prediction Limits for several $Y_{h(new)}$ in g different levels

$$(\hat{Y}_h - Bs\{pred\}, \hat{Y}_h + Bs\{pred\})$$

where

$$B = t\left(1-\frac{\alpha}{2g}; n-p\right)$$

(3) Scheffé Simultaneous Prediction Limits for several $Y_{h(new)}$ in g different levels

$$(\hat{Y}_h - Ss\{pred\}, \hat{Y}_h + Ss\{pred\})$$

where

$$S^2 = gF\left(1-\alpha; g, n-p\right)$$

(4) Prefiction Interval for $\overline{Y}_{h(new)}$

Given m observations in the same level of X, the mean of these m observations is $\overline{Y}_{h(new)}$.

the $1-\alpha$ prediction interval of $\overline{Y}_{h(new)}$ is given by

$$\left(\hat{Y}_h - t\left(1-\frac{\alpha}{2}\right)s\{predmean\}, \hat{Y}_h + t\left(1-\frac{\alpha}{2}\right)s\{predmean\}\right)$$

where

$$\begin{aligned} s\{predmean\} &= \frac{MSE}{m} + s^2\{\hat{Y}_h\} \\ &= MSE\left(\frac{1}{m} + \mathbf{X}_h(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h^T\right) \end{aligned}$$

7 Extra Sum of Squares

7.1 Definition

An extra sum of squares involves the difference between the error sum of squares for ther regression model containing the X variable(s) already in the model and the error sum of squares for the regression model containing both the original X vairables(s) and the new X variabel(s).

$$\begin{aligned} SSR(X_1|X_2) &= SSE(X_2) - SSE(X_1,X_2) \\ &= SSR(X_1,X_2) - SSR(X_2) \end{aligned}$$

We can define extra sums of squares similarly for three or more variables.

Decomposition of SSR into extra sums of squares:

$$\underbrace{SSR(X_1,X_2,\cdots,X_n)}_{df=n} = \underbrace{SSR(X_1)}_{1} + \underbrace{SSR(X_2|X_1)}_{1} + \underbrace{SSR(X_3|X_1,X_2)}_{1} + \cdots + \underbrace{SSR(X_n|X_1,X_2,\cdots,X_{n-1})}_{1}$$

ANOVA table containing decomposition of SSR ($p=4$):

Source of Variation	SS	df	MS
Regression	$SSR(X_1,X_2,X_3)$	3	$MSR(X_1,X_2,X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2\|X_1$	$SSR(X_2\|X_1)$	1	$MSR(X_2\|X_1)$
$X_3\|X_1,X_2$	$SSR(X_3\|X_1,X_2)$	1	$MSR(X_3\|X_1,X_2)$
Error	$SSE(X_1,X_2,X_3)$	$n-4$	$MSE(X_1,X_2,X_3)$
Total	$SSTO$	$n-1$	

7.2 Inferences about Regression Coefficients

Extra sum of squares provides a method to conduct tests about regression coefficients.

7.2.1 Overall F test

Test whether all (or several) $\beta_k = 0$.

The null hypothesis and alternative hypothesis are

$$H_0: \beta_0 = \cdots = \beta_{p-1} = 0 \qquad H_a: \text{not all } \beta_k \text{ equal } 0$$

The test statistic is

$$F^* = \frac{\dfrac{SSE(R)-SSE(F)}{df_R - df_F}}{\dfrac{SSE(F)}{df_F}}$$

$$= \frac{\frac{SSR(X_1, \cdots, X_{p-1})}{p-1}}{\frac{SSE(X_1, \cdots, X_{p-1})}{n-p}}$$

$$= \frac{MSR}{MSE} \overset{H_0}{\sim} F(p-1, n-p)$$

The decision rule is

$$\text{If } F^* \leqslant F(1-\alpha; p-1, n-p), \text{ accept } H_0;$$
$$\text{If } F^* > F(1-\alpha; p-1, n-p), \text{ reject } H_0.$$

The P-value is

$$1 - \mathbb{P}\{F(p-1, n-p) \leqslant F^*\}$$

7.2.2 Partial F test

Test whether a signle $\beta_k = 0$.

The null hypothesis and alternative hypothesis are

$$H_0 : \beta_k = 0 \qquad H_a : \beta_k \neq 0$$

The test statistic is

$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

$$= \frac{\frac{SSE(X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_{p-1}) - SSE(X_1, \cdots, X_{p-1})}{(n-p+1)-(n-p)}}{\frac{SSE(X_1, \cdots, X_{p-1})}{n-p}}$$

$$= \frac{MSR(X_k | X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_{p-1})}{MSE(X_1, \cdots, X_{p-1})} \overset{H_0}{\sim} F(1, n-p)$$

The decision rule is

$$\text{If } F^* \leqslant F(1-\alpha; 1, n-p), \text{ accept } H_0;$$
$$\text{If } F^* > F(1-\alpha; 1, n-p), \text{ reject } H_0.$$

The P-value is

$$1 - \mathbb{P}\{F(1, n-p) \leqslant F^*\}$$

The partial F test is equivalent to t test.

7.2.3 Test whether some $\beta_k = 0$

The null hypothesis and alternative hypothesis are

$$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \qquad H_a : \text{not all of the } \beta_k \text{ in the } H_0 \text{ equal } 0$$

The test statistic is

$$F^* = \frac{MSR(X_q, \cdots, X_{p-1}|X_1, \cdots, X_{q-1})}{MSE(X_1, \cdots, X_{p-1})} \overset{H_0}{\sim} F(p-q, n-p)$$

The decision rule is

$$\text{If } F^* \leqslant F(1-\alpha; p-q, n-p), \text{ accept } H_0;$$
$$\text{If } F^* > F(1-\alpha; p-q, n-p), \text{ reject } H_0.$$

The P-value is

$$1 - \mathbb{P}\{F(p-q, n-p) \leqslant F^*\}$$

7.3 Correlation Analysis

7.3.1 Coefficients of Partial Determination

For two variables,

$$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$$
$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

The generalization of coefficients of partial determination is similar. They can only take value in $[0,1]$.

A coefficients of partial determination $R^2_{Y1|2\cdots(p-1)}$ can be interpreted as a coefficient of simple determination R^2 for regressing residuals of predicting Y as function of $X_2, \cdots, X_{p-1}$

$$e_i(Y|X_2, \cdots, X_{p-1}) = Y_i - \hat{Y}_i(X_2, \cdots, X_{p-1})$$

on residuals of predicting X_1 as function of $X_2, \cdots, X_{p-1}$

$$e_i(X_1|X_2, \cdots, X_{p-1}) = X_{i1} - \hat{X}_{i1}(X_2, \cdots, X_{p-1})$$

Proof

7.3.2 Coefficients of Partial Correlation

For two variables,

$$r_{Y1|2} = sign(b_1)\sqrt{R^2_{Y1|2}}$$
$$r_{Y2|1} = sign(b_2)\sqrt{R^2_{Y2|1}}$$

8 Standardize Regression Model

8.1 Correlation Transformation

Centering and scaling of the data is given by

$$\frac{Y_i - \overline{Y}}{s_Y}$$
$$\frac{X_{ik} - \overline{X}_k}{s_k}$$

where

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}}$$
$$s_k = \sqrt{\frac{\sum_{i=1}^{n} (X_{ik} - \overline{X}_k)^2}{n-1}}$$

The correlation transformation is a simple function of the Standardized variables

$$Y_i^* = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \overline{Y}}{s_Y} \right)$$
$$X_{ik}^* = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \overline{X}_k}{s_k} \right)$$

8.2 Standardized Regression Model

The standardized regression model is given by

$$Y_i^* = \beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*$$

8.3 Properties

8.3.1 Regression Coefficients

The relationship between $\beta_1^*, \cdots, \beta_{p-1}^*$ and $\beta_0, \cdots, \beta_{p-1}$ is

$$\beta_k = \left(\frac{s_Y}{s_k} \right) \beta_k^* \qquad (k = 1, \cdots, p-1)$$
$$\beta_0 = \overline{Y} - \beta_1 \overline{X}_1 - \cdots - \beta_{p-1} \overline{X}_{p-1}$$

8.3.2 Estimated Regression Coefficients

Let

$$\mathbf{X}^* = \begin{bmatrix} X_{11}^* & \cdots & X_{1,p-1}^* \\ X_{21}^* & \cdots & X_{2,p-1}^* \\ \vdots & \ddots & \vdots \\ X_{n1}^* & \cdots & X_{n,p-1}^* \end{bmatrix} \qquad \mathbf{Y}^* = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_{p-1}^* \end{bmatrix} \qquad \mathbf{b}^* = \begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{p-1}^* \end{bmatrix}$$

The correlation matrix of $\mathbf{X}^*$ is given by

$$\begin{aligned} \mathbf{r}_{X^*X^*} &= \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{12} & 1 & \cdots & r_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix} \\ &= \mathbf{X}^{*T}\mathbf{X}^* \end{aligned}$$

The correlation matrix of $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$\begin{aligned} \mathbf{r}_{Y^*X^*} &= \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Y,p-1} \end{bmatrix} \\ &= \mathbf{X}^{*T}\mathbf{Y}^* \end{aligned}$$

We have

$$\mathbf{b}^* = \mathbf{r}_{X^*X^*}^{-1}\mathbf{r}_{Y^*X^*}$$

Then

$$\begin{aligned} b_k &= \left(\frac{s_Y}{s_k}\right) b_k^* \qquad (k = 1, \cdots, p-1) \\ b_0 &= \overline{Y} - b_1\overline{X}_1 - \cdots - b_{p-1}\overline{X}_{p-1} \end{aligned}$$

9 Regression Models for Quantitative and Qualitative Predictors

9.1 Quantitative Predictor

9.1.1 Polynomial Regression Models

2^{nd} order model for one predictor variable is given by

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

where

$$x = X - \overline{X}$$

2^{nd} order model for two predictors variables is given by

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} + \varepsilon$$

where

$$x_1 = X_1 - \overline{X}_1$$
$$x_2 = X_2 - \overline{X}_2$$

9.1.2 Interaction Regression Models

The cross-product terms like β_{12} are called the interaction effect coefficient.

The interaction effect between two quantitative variables is of an reinforcement or synergistic type if the slope of the response function against one of the predictor variables increases for higher levels of the other predictor variable.

The interaction effect between two quantitative variables is of an interference or antagonistic type if the slope of the response function against one of the predictor variables decreases for higher levels of the other predictor variable.

9.2 Qualitative Predictor

A qualitative variable with c classes will be represented by $c-1$ indicator variables (dummy variables or binary variables).

9.3 Modeling Interactions between Quantitative and Qualitative Predictors

Suppose that we have two Models

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$Y_i' = \beta_0' + \beta_1' X_i' + \varepsilon_i'$$

with respect to two similar but different data sets $\{X_1, \cdots, X_{n_1}\}$ and $\{X_1', \cdots, X_{n_2}'\}$. We can use qulitative predictors to model these two data sets into a model. First we re-arrange the dataset (It's umimportant of the order.)

$$\begin{cases} X_{11}^* = X_1 \\ \vdots \\ X_{n_1 1}^* = X_{n_1} \\ X_{n_1+1,1}^* = X_1' \\ \vdots \\ X_{n_1+n_2,1}^* = X_{n_2}' \end{cases}$$

Then we use a indicator

$$X_{i2}^* = \begin{cases} 1 & \text{,if } X_{i1}^* \text{ is in data set one} \\ 0 & \text{,if } X_{i1}^* \text{ is in data set two} \end{cases}$$

Finally, the first-order regression model with an added interaction term is given by

$$Y_i^* = \beta_0^* + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \beta_3^* X_{i1}^* X_{i2}^* + \varepsilon_i^*$$

When $X_{i2} = 0$,

$$\mathbb{E}Y = \beta_0^* + \beta_1^* X_1$$

When $X_{i2} = 1$

$$\mathbb{E}Y' = \beta_0^* + \beta_1^* X_1' + \beta_2^* + \beta_3 X_1'$$

Therefore,

$$\begin{cases} \beta_0^* = \beta_0 \\ \beta_1^* = \beta_1 \\ \beta_2^* = \beta_0' - \beta_0 \\ \beta_3^* = \beta_1' - \beta_1 \end{cases}$$

A test of whether $\beta_0 = \beta_0'$ can be carried out by testing

$$H_0 : \beta_2^* = 0 \qquad\qquad H_a : \beta_2^* \neq 0$$

A test of whether $\beta_1 = \beta_1'$ can be carried out by testing

$$H_0 : \beta_3^* = 0 \qquad\qquad H_a : \beta_3^* \neq 0$$

A test of whether $\beta_0 = \beta_0', \beta_1 = \beta_1'$ can be carried out by testing

$$H_0 : \beta_2^* = \beta_3^* = 0 \qquad\qquad H_a : \text{not both } \beta_2^* = 0 \text{ and } \beta_3^* = 0$$

Part III Building the Regression Model

10 Model Selection and Validation

10.1 Model Selection Criteria

The number of X variables will be denoted by $p-1$ where

$$1 \leqslant p \leqslant P$$

and

$$n > P$$

R_p^2 is used to decide the best model for a fixed size, while $R_{a,p}^2$, C_p, AIC_p, BIC_p and $PRESS_p$ are used to decide the appropriate subset size.

10.1.1 R_p^2 or SSE_p

The subscript p indicates that there are p parameters in the regression model.

$$\begin{aligned} R_p^2 &= \frac{SSR_p}{SSTO} \\ &= 1 - \frac{SSE_p}{SSTO} \end{aligned}$$

R_p^2 won't decrease when additional variables are included in the model. R_p^2 will be a maximum when all $p-1$ potential X variables are included in the regression model.

10.1.2 $R_{a,p}^2$ or MSE_p

$$\begin{aligned} R_{a,p}^2 &= 1 - \frac{\frac{SSR_p}{n-p}}{\frac{SSTO}{n-1}} \\ &= 1 - \frac{MSE_p}{MSTO} \end{aligned}$$

10.1.3 C_p

Mallows′ $\mathbf{C_p}$ Criterion considers the total mean squared error of the n fitted values for each subset regression model. The criterion measure is

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2}\mathbb{E}\sum_{i=1}^{n}(\hat{Y}_i - \mu_i)^2 \\ &= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(\mathbb{E}\hat{Y}_i - \mu_i)^2 + \sum_{i=1}^{n} Var\hat{Y}_i\right] \end{aligned}$$

$$= \frac{\mathbb{E}SSE_p}{\sigma^2} - (n-2p)$$

where μ_i is the true mean response of X_i. Proof

It can be estimated by

$$C_p = \frac{SSE_p}{MSE(X_1, \cdots, X_{P-1})} - (n-2p)$$

Replacing $\mathbb{E}SSE_p$ by the estimator SSE_p and using $MSE(X_1, \cdots, X_{P-1})$ as an unbiased estimator of σ^2, when the model is unbiased, $\mathbb{E}\hat{Y}_i \equiv \mu_i$,

$$\mathbb{E}C_p \approx p$$

Therefore, the model with C_p most near p will be best. Especially,

$$C_P = P$$

10.1.4 AIC_p

Akaike′s Information Criterion ($\mathbf{AIC_p}$) is given by

$$AIC_p = n \ln SSE_p - n \ln n + 2p$$

10.1.5 BIC_p

Bayesian Information Criterion ($\mathbf{BIC_p}$) or **Schwarz′ Bayesian Criterion** ($\mathbf{SBC_p}$) is given by

$$BIC_p = SBC_p = n \ln SSE_p - n \ln n + p \ln n$$

If $n \geqslant 8$, the penality for BIC_p - $p \ln n$ is larger than that for AIC_p.

10.1.6 $PRESS_p$

The **Prediction Sum of Squares Criterion** ($\mathbf{PRESS_p}$) is given by

$$\begin{aligned} PRESS_p &= \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2 \\ &= \sum_{i=1}^{n} d_i^2 \\ &= \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}} \right)^2 \end{aligned}$$

where $\hat{Y}_{i(i)}$ and d_i are defined in Deleted Residuals

10.2 Automatic search Procedures for Model Selection

10.2.1 "Best" Subsets Algorithm

Consider all the possible subset. For each of the model, evaluate the criteria.

Time-saving algorithms have been developed, which require the calculation of only a small fraction of all possible models.

Still, if $P > 30$, it requires excessive computer time.

Several regression models can be identified as "good" for final consideration, depending on which criteria we use.

10.2.2 Forward Selection

(1) Choose a significance level to enter the model (e.g. $SLE = 0.20$, generally .05 is too high, causing too few variables to be entered).

(2) Start with no variables.

(3) Add one variable with highest t or F-value (only if P-value $< SLE$).

(4) Add the next variable with highest partial F-value given the previous variables in the model (only if P-value $< SLE$).

(5) Continue until no new predictors have P-value$\geqslant SLE$.

(R uses model based criteria: AIC, SBC instead.)

10.2.3 Backward Elimination

(1) Select a significance level to stay in the model (e.g. SLS=0.20, generally .05 is too low, causing too many variables to be removed).

(2) Start with all the variables. Fit the full model with all possible predictors.

(3) Consider the predictor with lowest t-statistic (highest P-value).

(4) If P-value $> SLS$, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change).

(5) If P-value $\leqslant SLS$, stop and keep current model.

(6) Continue until all predictors have P-values below SLS.

(R uses model based criteria: AIC, SBC instead.)

10.2.4 (Forward) Stepwise Regression

(1) Select SLS and SLE ($SLE < SLS$).

(2) Starts like Forward Selection (Bottom up process).

(3) New variable must have smallest P-value and P-value$< SLE$ to enter.

(4) "Old variable"that has already been entered with biggest P-value and P-value$> SLS$ will be removed.

(5) Continues until no new variables can be entered and no old variables need to be removed.

(R uses model based criteria: AIC, SBC instead.)

10.3 Model Validation

Validation set is used to test model with a new set of data we have.

Mean Square Prediction Error when training model is applied to validation sample:

$$MSPR = \frac{1}{n_V}\sum_{i=1}^{n_V}(Y_{Vi} - \hat{Y}_{Vi})^2$$

K-fold cross-validation procedure is useful with small data sets.

11 Diagnostics

In this chapter, we discuss how to find out outliers and whether these outliers are influential or not. First we prppose a graphical method - added-variable plot. Then we indentify outlying Y observations and X cases by residuals and hat matrixes respectively. Finally, we propose some methods to indentify whether an outlier is influential.

11.1 Added-Variable Plots

11.1.1 Definition

Added-variable Plots, also called partial regression plots or adjusted variable plots, is given by

1. Fit regression of Y on X_2 , obtain residuals $e_i(Y|X_2)$
2. Fit regression of X_1 on X_2 , obtain residuals $e_i(X_1|X_2)$
3. Plot $e_i(Y|X_2)$ (vertical axis) versus $e_i(X_1|X_2)$ (horizontal axis)

It can be generalized to more than 2 variables.

11.1.2 Properties

Here we use the fact that for Coefficients of Partial Determination

$$R^2_{Y1|2\cdots(p-1)} = R^2_{e(Y|X_2,\cdots,X_{p-1})e(X_1|X_2,\cdots,X_{p-1})}$$

(1) $(0,0)$ is on the regression line $e_i(Y|X_2)$ with respect to $e_i(X_1|X_2)$.

(2) Good linearity of the regression function indicates that the variable should be included into the linear regression model.

(3) Non-linearity of the regression function indicates that the variable can be included into the linear regression model with non-linear form.

11.2 Outlying Y Observations

11.2.1 Residuals

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \\ \mathbf{e} &\sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))\end{aligned}$$

11.2.2 Studentized Residuals

The studentized residuals are defined by

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

11.2.3 Semi-Studentized Residuals

The semi-studentized residuals are defined by

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

11.2.4 Deleted Residuals

$$\begin{aligned} d_i &= Y_i - \hat{Y}_{i(i)} \\ &= \frac{e_i}{1-h_{ii}} \end{aligned}$$

where $\hat{Y}_{i(i)}$ is the fitted value of X_i when regression without (X_i, Y_i). We have

$$\begin{aligned} Vard_i &= VarY_i + Var\hat{Y}_{i(i)} \\ &= \sigma^2[1 + X_i^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}X_i] \\ s^2\{d_i\} &= MSE_{(i)}[1 + X_i^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}X_i] \\ &= \frac{MSE_{(i)}}{1-h_{ii}} \end{aligned}$$

(Proof) where $\mathbf{X}_{(i)}$ is the predictor without X_i. Here we use $MSE_{(i)}$ to estimate σ^2 in prediction for $Y_{h(new)}$ and $X_h = X_i$.

11.2.5 Studentized Deleted Residuals

$$\begin{aligned} t_i &= \frac{d_i}{s\{d_i\}} \\ &= \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \end{aligned}$$

Since

$$\begin{aligned} SSE &= (n-p)MSE \\ &= (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}} \end{aligned}$$

(Proof) we have

$$t_i = \frac{e_i\sqrt{n-p-1}}{\sqrt{SSE(1-h_{ii}) - e_i^2}} \sim t(n-p-1)$$

11.3 Outlying X Cases

11.3.1 Hat Matrix

$$\begin{aligned}
\mathbf{H} = (h_{ij})_{n\times n} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
h_{ij} &= h_{ji} \\
\sum_{i=1}^{n} h_{ii} &= p \\
\sum_{i=1}^{n} h_{ij} &= 1 \\
0 \leqslant h_{ii} &\leqslant 1 \\
\sum_{j=1}^{n} h_{ij}e_j &= 0
\end{aligned}$$

Proof

11.3.2 Leverage

h_{ii} is called the leverage of the ith case. It is a measure of distance between X_i and $\overline{X}$. The mean leverage value is given by

$$\overline{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$$

Leverage values greater than $2\overline{h} = \dfrac{2p}{n}$ will be considered as being an outlier. The larger is h_{ii}, the smaller is the variance of the residuals e_i.

Leverage values for new observations:

$$h_{new,new} = \mathbf{X}_{new}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{new}$$

New cases with leverage values larger than those in original dataset are extrapolations.

11.4 Indentifying Influential Cases

11.4.1 $DFFITS$

$DFFITES$ meansures the influence on single fitted value,

$$\begin{aligned}
(DFFITS)_i &= \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} \\
&= e_i\left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right]^{\frac{1}{2}}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}} \\
&= t_i\left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}}
\end{aligned}$$

The ith case is considered to be influential if the absolute value of $DFFITS$ exceeds 1 for small to medium data sets and $2\sqrt{\frac{p}{n}}$ for large data sets.

11.4.2 Cook's Distance

Cook's distance meansures the influence on all fitted values,

$$\begin{aligned} D_i &= \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE} \\ &= \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})^T(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{pMSE} \\ &= \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \\ &= e_i^{*2}\frac{h_{ii}}{p(1-h_{ii})^2} \end{aligned}$$

The ith case is considered to be influential if $D_i \geqslant F(\frac{1}{2}; p, n-p)$.

11.4.3 DFBETAS

$DFBETAS$ measures the influence on the regression coefficients, $\forall\ k \in \{1, 2, \cdots, p-1\}$,

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}$$

where c_{kk} is the kth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

The ith case is considered to be influential if the absolute value of $DFFITS$ exceeds 1 for small to medium data sets and $\dfrac{2}{\sqrt{n}}$ for large data sets.

11.5 Multicollinearity

11.5.1 VIF

The diagonal element of $\mathbf{r}_{XX}^{-1}$ is called variance inflation factor (VIF), which satisfies

$$\begin{aligned} \sigma\{b_k^*\} &= \sigma^{*2}(VIF)_k \\ \sigma\{\mathbf{b}^*\} &= \sigma^{*2}\mathbf{r}_{X^*X^*}^{-1} \end{aligned}$$

where $\mathbf{b}^*$ is related to standardized regression model. It can be shown that

$$(VIF)_k = \frac{1}{1-R_k^2} \geqslant 1$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the $p-1$ other X variables.

The largest VIF value among all X variables is often used as an indicator of the severity of multicollinearity. $\max_k\{(VIF)_k\} > 10$ indicates a multicollinearity problem.

The mean of the VIF values is given by

$$\begin{aligned}
\overline{VIF} &= \frac{\sum_{k=1}^{p-1} (VIF)_k}{p-1} \\
&= \frac{\mathbb{E}\left[\sum_{k=1}^{p-1} (b_k^* - \beta_k^*)^2\right]}{\sigma^{*2}(p-1)}
\end{aligned}$$

$\overline{VIF}$ considerably larger than 1 is indicative of serious multicollinearity problems.

12 Remedial Measures

12.1 Weighted Least Squares

Weighted Least Squares Method is used for remedying unequal error variances, i.e., $\varepsilon_i \sim N(0, \sigma_i^2)$.

Weighted least squares criterion is given by

$$\begin{aligned}
\mathbf{b} &= \min_{\beta} Q_w \\
&= \min_{\beta} \sum_{i=1}^{n} w_i (Y_i - \beta_0 - \cdots - \beta_{p-1} X_{i,p-1})^2 \\
&= \min_{\beta} (\mathbf{Y} - \beta \mathbf{X})^T \mathbf{W} (\mathbf{Y} - \beta \mathbf{X})
\end{aligned}$$

where

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

R^2 has no clear-cut meaning for weighted least squares.

12.1.1 Known σ_i^2

We can set $w_i = \dfrac{1}{\sigma_i^2}$.

The solution for the optimization problem is given by

$$\begin{aligned}
\mathbf{b}_w &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \\
\mathbf{b}_w &\sim N(\mathbf{0}, (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1})
\end{aligned}$$

For σ_i^2 known up to proportionality constant, we have

$$\begin{aligned}
\mathbf{b}_w &= (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \\
\mathbf{b}_w &\sim N(\mathbf{0}, k(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}) \\
s^2\{\mathbf{b}_w\} &= MSE_w (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \\
MSE_w &= \frac{\sum_{i=1}^{n} w_i (Y_i - \hat{Y}_i)^2}{n-p} \\
&= \frac{\sum_{i=1}^{n} w_i e_i^2}{n-p}
\end{aligned}$$

12.1.2 Unknown σ_i^2

1. Use estimated variance or standard deviation funtion to obtain the weights.

   Iteratively reweighted least squares:

(1) Fit the regression model by unweighted least squares and analyze the residuals.

(2) Estimate the variance function or the standard deviation function by regression either the squared residuals or the absolute residuals on the approriate predictor(s).

(3) Use the fitted values from the estimated variance or standard deviation function to obtain the weights w_i.

(4) Estimate the regression coefficients using these weights.

$$w_i = \frac{1}{\hat{s}_i^2}$$
$$w_i = \frac{1}{\hat{v}_i^2}$$

where $\hat{s}_i^2$ and $\hat{v}_i^2$ are the fitted values from standard deviation function and variance function respectively.

(5) Repeat until convergence.

2. Use of replicates or near replicates.

Use the sample variance of the replicates as the estimators for the variances.

12.1.3 Relationship between OLS and WLS

With unequal error variances, $\mathbf{b}$ is unbiased and consistent, but $b_i (i = 1, 2, \cdots, p-1)$ are no longer minimum variance estimators. The variance-convariance matrix of ordinary least squares becomes

$$\sigma^2\{\mathbf{b}\} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\sigma^2\{\varepsilon\}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}$$

The White estimator is

$$S^2\{\mathbf{b}\} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T S_0 \mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}$$

where

$$S_0 = \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$$

12.2 Ridge Regression

Ridge Regression is used for remedying multicollinearity.

It can given by the method of penalize least squares. Penalize least squares criterion is given by

$$\begin{aligned} \mathbf{b}^R &= \min_{\beta} Q \\ &= \min_{\beta} \sum_{i=1}^{n} (Y_i^* - \beta_0^* - \cdots - \beta_{p-1}^* X_{i,p-1}^*)^2 + \lambda \sum_{j=1}^{p-1} \beta_j^{*2} \\ &= \min_{\beta} (\mathbf{Y}^* - \beta^*\mathbf{X}^*)^T(\mathbf{Y}^* - \beta^*\mathbf{X}^*) + \lambda \beta^{*T}\beta^* \end{aligned}$$

where $\mathbf{X}^*, \mathbf{Y}^*, \beta^*$ and $\mathbf{b}^*$ are given by the standardized regression model.

Equivalently, the ridge estimaors are given by

$$\mathbf{b}^R = (\mathbf{r}_{X^*X^*} + \lambda \mathbf{I})^{-1} \mathbf{r}_{Y^*X^*}$$

To choose proper $\lambda \geqslant 0$, we can use the ridge trace and the $(VIF)_k$. As λ increases, we can choose the one such that VIF value near 1 and the estimated regression coefficients appear to have become reasonably stable. VIF values are the diagonal elements of $(\mathbf{r}_{X^*X^*} + \lambda \mathbf{I})^{-1} \mathbf{r}_{X^*X^*} (\mathbf{r}_{X^*X^*} + \lambda \mathbf{I})^{-1}$.

$$SSTO_R = 1$$
$$R_R^2 = 1 - SSE_R$$

12.3 Robust Regression

Robust Regression is used for remedying influential cases.

12.3.1 LAR or LAD Regression

Least absolute residuals (LAR) or least absolute deviations (LAD) regression, also called minimum L_1-norm regression minimizes

$$L_1 = \sum_{i=1}^{n} |Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1})|$$

12.3.2 LMS Regression

Least median of squares (LMS) regression minimizes

$$median\{[Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1})]^2\}$$

12.3.3 IRLS Robust Regression

Iteratively reweighted least squares (IRLS) robust regression is given by

(1) Choose a weight function for weighting the cases.

Huber weight function:

$$w = \begin{cases} 1 & , |u| \leqslant 1.345 \\ \dfrac{1.345}{|u|} & , |u| > 1.345 \end{cases}$$

Bisquare weight funtion:

$$w = \begin{cases} \left[1 - \left(\dfrac{u}{4.685}\right)^2\right]^2 & , |u| \leqslant 4.685 \\ 0 & , |u| > 4.685 \end{cases}$$

where u is the scaled residual given by

$$u_i = \frac{e_i}{MAD}$$

$$
\begin{aligned}
MAD &= \frac{1}{\Phi(0.75)} median\{|e_i - median\{e_i\}|\} \\
&= \frac{1}{0.6745} median\{|e_i - median\{e_i\}|\}
\end{aligned}
$$

(2) Obtain starting weights for all cases.

For Huber weight function, the initial weights can be those obtain from OLS. For bisquare function, they can be those obtained from Huber function.

(3) Use the starting weights in weighted least squares and obtain the residuals from the fitted regression function.

(4) Use the residuals in step (3) to obtain revised weights.

(5) Continue the iterations until convergence is obtained.

12.4 Nonparametric Regression

12.4.1 Lowess Method

12.4.2 Regression Trees

12.5 Bootstrap

Bootstrap is used for evaluateing precision in nonstandard situations.

12.5.1 Bootstrap Sampling

1. Fixed X Sampling

   (Model is good fit, constant variance, predictor variables are fixed, i.e. controlled experiment)

   (1) Fit the regression, obtain all fitted values and residuals

   (2) Keeping the corresponding X-level(s) and fitted values, re-sample the n residuals (with replacement)

   (3) Add the bootstrapped residuals to the fitted values, and re-fit the regression (repeat process many times)

   $$Y_i^* = \hat{Y}_i + e_i^*$$

   Then regress Y^* values on the original X variables to obtain the bootstrap estimate b_1^*.

2. Random X Sampling

   (Not sure of adequacy of model fit, variance, random predictor variables)

   After fitting regression, and estimating quantities of interest, sample n cases (with replacement) and re-estimate quantities of interest with “new”datasets (repeat many times)

12.5.2 Bootstrap Confidence Intervals

A relative simple procedure for setting up a $1-\alpha$ approximate confidence interval is the reflection method. The reflection method confidence interval for β_1 is based on the $\left(\frac{\alpha}{2}\right)100$ and $\left(1-\frac{\alpha}{2}\right)100$ percentiles of the bootstrap distribution of b_1^*. Let

$$
\begin{aligned}
d_1 &= b_1 - b_1^*\left(\frac{\alpha}{2}\right) \\
d_2 &= b_1^*\left(1-\frac{\alpha}{2}\right) - b_1
\end{aligned}
$$

where $b_1^*\left(\frac{\alpha}{2}\right)$ and $b_1^*\left(1-\frac{\alpha}{2}\right)$ are the $\left(\frac{\alpha}{2}\right)100$ and $\left(1-\frac{\alpha}{2}\right)100$ percentiles of the bootstrap distribution of b_1^* respectively. Then the approximate $1-\alpha$ confidence interval for β_1 is given by

$$(b_1 - d_2, b_1 + d_1)$$

Bootstrap confidence intervals by the reflection method require a large number of bootstrap sampl than do bootstrap estimates of precision.

Appendix

A Some Useful Results in Linear Algebra

A.1 Trace

Cyclic Property. The trace is invariant under cyclic permutations, i.e.,

$$\mathrm{tr}(\mathbf{ABCD}) = \mathrm{tr}(\mathbf{BCDA}) = \mathrm{tr}(\mathbf{CDAB}) = \mathrm{tr}(\mathbf{DABC})$$

if all these products of matrixes are well defined and squared.

A.2 Spectral Decomposition

If $\mathbf{A}$ is $n \times n$ symmetric matrix, then $\mathbf{A}$ can be decomposed as follow:

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^T = \sum_{i=1}^{n} \lambda_i v_i v_i^T$$

where $\mathbf{V} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_n$ and $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of $\mathbf{A}$.

From the cyclic property of trace,

$$\begin{aligned} \mathrm{tr}(\mathbf{A}) &= \mathrm{tr}(\mathbf{V}\Lambda\mathbf{V}^T) \\ &= \mathrm{tr}(\Lambda\mathbf{V}^T\mathbf{V}) \\ &= \mathrm{tr}(\Lambda) \\ &= \sum_{i=1}^{n} \lambda_i \\ |\mathbf{A}| &= \prod_{i=1}^{n} \lambda_i \end{aligned}$$

A.3 Idempotent Matrix

A squared matrix $\mathbf{A}$ is called idempotent if

$$\mathbf{A}^2 = \mathbf{A}$$

The eigenvalues of $\mathbf{A}$ are either 0 or 1.

$$rank(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$$

A.4 Cochran's Theorem

Suppose $U_1, \cdots, U_n$ are i.i.d. standard normally distributed random variables, and an identity of the form

$$\sum_{i=1}^{r} U_i^2 = Q_1 + \cdots + Q_k$$

can be written, where each Q_i is a quadratic form in $U_1, \dots, U_n$. Further suppose that

$$r_1 + \cdots + r_k = r$$

where r_i is the rank of Q_i.

Cochran's theorem states that the Q_i are independent, and each Q_i has a chi-squared distribution with r_i degrees of freedom. Here the rank of Q_i should be interpreted as meaning the rank of the matrix $B^{(i)}$, with elements $B^{(i)}_{j,l}$, in the representation of Q_i as a quadratic form:

$$Q_i = \sum_{j=1}^{N} \sum_{\ell=1}^{N} U_j B^{(i)}_{j,\ell} U_\ell.$$

Less formally, it is the number of linear combinations included in the sum of squares defining Q_i, provided that these linear combinations are linearly independent.

B Proof of Some Statements

2.2.2(3)

$(\hat{\beta}_0, \hat{\beta}_1, \overline{Y})$ is independent with SSE.

Proof.

Let

$$\mathbf{b} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \; SSE = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Without using the property of square matrix's exchangeability between transposed transformation and inversion transformation, we can show that for an invertible symmetric matrix, its inverse matrix is also symmetric.

Let $\mathbf{A}$ be an invertible symmetric matrix and $\mathbf{C} = \mathbf{A}^{-1}$, then

$$\mathbf{A}^T = \mathbf{A}$$

$$\mathbf{AC} = \mathbf{I} \tag{1}$$

$$\mathbf{CA} = \mathbf{I} \tag{2}$$

Take transposed transformation of (1), we have

$$\mathbf{C}^T\mathbf{A}^T = \mathbf{C}^T\mathbf{A} = \mathbf{I}$$

From the uniqueness of inverse matrix, we have

$$\mathbf{C}^T = \mathbf{C}$$

Since

$$(\mathbf{X}^T\mathbf{X})^T = \mathbf{X}^T\mathbf{X}$$

we have

$$[(\mathbf{X}^T\mathbf{X})^{-1}]^T = (\mathbf{X}^T\mathbf{X})^{-1}$$

$\because$

$$\begin{aligned}
(\mathbf{I} - \mathbf{H})\mathbf{Y} &\sim N((\mathbf{I} - \mathbf{H})\mathbf{X}\beta, (\mathbf{I} - \mathbf{H})\sigma^2) \\
\mathbf{b} &\sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \\
Cov[(\mathbf{I} - \mathbf{H})\mathbf{Y}, \mathbf{b}] &= Cov[(\mathbf{I} - \mathbf{H})\mathbf{Y}, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\
&= (\mathbf{I} - \mathbf{H})Cov(\mathbf{Y}, \mathbf{Y})[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\
&= (\mathbf{I} - \mathbf{H})Cov(\mathbf{X}\beta + \varepsilon, \mathbf{X}\beta + \varepsilon)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{I} - \mathbf{H})Cov(\varepsilon, \varepsilon)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{I} - \mathbf{H})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X} - \mathbf{HX})(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X} - \mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

$$= 0$$

$\therefore \quad (\mathbf{I}-\mathbf{H})\mathbf{Y}$ and $\mathbf{b}$ are independent

$\therefore \quad SSE = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y} = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})\mathbf{Y}$ and $\mathbf{b}$ are independent

$\because$

$$\begin{aligned}
\overline{Y} &\sim N(\frac{1}{n}\mathbf{1}^T\mathbf{X}\beta, \frac{1}{n}\sigma^2) \\
(\mathbf{I}-\mathbf{H})\mathbf{Y} &\sim N((\mathbf{I}-\mathbf{H})\mathbf{X}\beta, (\mathbf{I}-\mathbf{H})\sigma^2) \\
Cov((\mathbf{I}-\mathbf{H})\mathbf{Y}, \overline{Y}) &= Cov[(\mathbf{I}-\mathbf{H})\mathbf{Y}, \frac{1}{n}\mathbf{1}^T\mathbf{Y}] \\
&= (\mathbf{I}-\mathbf{H})Cov(\mathbf{Y},\mathbf{Y})\left(\frac{1}{n}\mathbf{1}^T\right)^T \\
&= (\mathbf{I}-\mathbf{H})Cov[\mathbf{X}\beta+\varepsilon, \mathbf{X}\beta+\varepsilon]\frac{1}{n}\mathbf{1} \\
&= (\mathbf{I}-\mathbf{H})Cov(\varepsilon,\varepsilon)\frac{1}{n}\mathbf{1} \\
&= \frac{1}{n}\sigma^2(\mathbf{I}-\mathbf{H})\mathbf{1} \\
&= \frac{1}{n}\sigma^2(\mathbf{1}-\mathbf{H}\mathbf{1}) \\
&= \frac{1}{n}\sigma^2(\mathbf{1}-\mathbf{1}) \\
&= 0
\end{aligned}$$

$\therefore \quad (\mathbf{I}-\mathbf{H})\mathbf{Y}$ and $\overline{Y}$ are independent

$\therefore \quad SSE = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y} = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})\mathbf{Y}$ and $\overline{Y}$ are independent □

3.2.1

$$\hat{Y}_h \sim N\left(\beta_0+\beta_1X_h, \sigma^2\left(\frac{1}{n}+\frac{(X_h-\overline{X})^2}{SS_{XX}}\right)\right)$$

Proof.

$\because \quad \overline{Y}, b_1(X_h-\overline{X})$ are normal random variables

$$\hat{Y} = \overline{Y} + b_1(X_h-\overline{X})$$

$\therefore \quad \hat{Y}$ is also a normal random variable

$$\begin{aligned}
\mathbb{E}\hat{Y} &= \mathbb{E}(b_0+b_1X_h) \\
&= \beta_0+\beta_1X_h
\end{aligned}$$

$\because$

$$\overline{Y} \perp b_1$$

$\therefore$

$$
\begin{aligned}
Var\hat{Y} &= Var[\overline{Y} + b_1(X_h - \overline{X})] \\
&= Var(\overline{Y}) + Var[b_1(X_h - \overline{X})] \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2(X_h - \overline{X})^2}{SS_{XX}} \\
&= \sigma^2\left(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{SS_{XX}}\right)
\end{aligned}
$$

□

3.6.2

For simple linear regression,

$$r = \pm\sqrt{R^2}$$

Proof.

$\because$

$$
\begin{aligned}
SSR &= \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \\
&= \sum_{i=1}^{n}(b_0 + b_1X_i - b_0 - b_1\overline{X})^2 \\
&= b_1^2 SS_{XX} \\
b_1 &= \frac{SS_{XY}}{SS_{XX}}
\end{aligned}
$$

$\therefore$

$$SSR = \frac{SS_{XY}^2}{SS_{XX}}$$

$\because$

$$
\begin{aligned}
r &= \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} \\
&= sign(SS_{XY})\sqrt{\frac{SS_{XY}^2}{SS_{XX}SS_{YY}}} \\
SSTO &= SS_{YY}
\end{aligned}
$$

$\therefore$

$$
\begin{aligned}
r &= sign(SS_{XY})\sqrt{\frac{SSR}{SSTO}} \\
&= \pm\sqrt{R^2}
\end{aligned}
$$

□

3.6.3(2)

$$\frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{b_1}{s\{b_1\}}$$

Proof.

$\because$

$$\begin{aligned} r &= \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} \\ 1-r^2 &= \frac{SSE}{SSTO} \\ &= \frac{(n-2)MSE}{SS_{YY}} \end{aligned}$$

$\therefore$

$$\begin{aligned} \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} &= \frac{\sqrt{\frac{SS_{XX}}{SS_{YY}}}b_1}{\sqrt{\frac{(n-2)MSE}{SS_{YY}}}} \\ &= \frac{b_1}{s\{b_1\}} \end{aligned}$$

□

6.1.2(2)

$$\begin{aligned} \mathbf{H}^T &= \mathbf{H} \\ (\mathbf{I}-\mathbf{H})^T &= \mathbf{I}-\mathbf{H} \end{aligned}$$

Proof.

Without using the property of square matrix's exchangeability between transposed transformation and inversion transformation, we can show that for an invertible symmetric matrix, its inverse matrix is also symmetric.

Let $\mathbf{A}$ be an invertible symmetric matrix and $\mathbf{C} = \mathbf{A}^{-1}$, then

$$\begin{aligned} \mathbf{A}^T &= \mathbf{A} \\ \mathbf{AC} &= \mathbf{I} \qquad (1) \\ \mathbf{CA} &= \mathbf{I} \qquad (2) \end{aligned}$$

Take transposed transformation of (1), we have

$$\mathbf{C}^T\mathbf{A}^T = \mathbf{C}^T\mathbf{A} = \mathbf{I}$$

From the uniqueness of inverse matrix, we have

$$\mathbf{C}^T = \mathbf{C}$$

Since

$$(\mathbf{X}^T\mathbf{X})^T = \mathbf{X}^T\mathbf{X}$$

we have

$$[(\mathbf{X}^T\mathbf{X})^{-1}]^T = (\mathbf{X}^T\mathbf{X})^{-1}$$

$$\begin{aligned}\mathbf{H}^T &= [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\ &= (\mathbf{X}^T)^T[(\mathbf{X}^T\mathbf{X})^{-1}]^T\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{H}\end{aligned}$$

$\because$

$$\begin{aligned}\mathbf{I}^T &= \mathbf{I} \\ \mathbf{H}^T &= \mathbf{H}\end{aligned}$$

$\therefore$

$$(\mathbf{I}-\mathbf{H})^T = \mathbf{I}-\mathbf{H}$$

□

6.1.2(3)

$$\begin{aligned}\mathbf{H}\mathbf{H} &= \mathbf{H} \\ (\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) &= \mathbf{I}-\mathbf{H}\end{aligned}$$

Proof.

$$\begin{aligned}\mathbf{H}\mathbf{H} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{H} \\ (\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) &= \mathbf{I}-2\mathbf{H}+\mathbf{H}\mathbf{H} \\ &= \mathbf{I}-2\mathbf{H}+\mathbf{H} \\ &= \mathbf{I}-\mathbf{H}\end{aligned}$$

□

6.1

$$\mathbf{H} = \mathbf{H}_r + \mathbf{H}_0$$

Proof.

$\because$ X is centered, i.e. $\forall\ j \in \{1,2,\cdots,p-1\}$,

$$\sum_{j=1}^{n} X_{ij} = 0$$

$\therefore$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathbf{X}_r^T\mathbf{X}_r & \\ 0 & & & \end{bmatrix}$$

$\therefore$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & (\mathbf{X}_r^T\mathbf{X}_r)^{-1} & \\ 0 & & & \end{bmatrix}$$

$\because$

$$\begin{aligned} \mathbf{H} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \begin{bmatrix} \frac{1}{n}\mathbf{1} & \mathbf{X}_r(\mathbf{X}_r^T\mathbf{X}_r)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{1}^T \\ \mathbf{X}_r^T \end{bmatrix} \\ &= \frac{1}{n}\mathbf{1}\mathbf{1}^T + \mathbf{X}_r(\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r^T \\ &= \mathbf{H}_0 + \mathbf{H}_r \end{aligned}$$

□

7.3.1

$$R^2_{Y1|2\cdots(p-1)} = R^2_{e(Y_1|X_2,\cdots,X_{p-1})e(X_1|X_2,\cdots,X_{p-1})}$$

Proof.

$R^2_{Y1|2\cdots(p-1)}$ is the coefficient of partial determination of

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1}X_{p-1} + \varepsilon \tag{1}$$

$R^2_{e(Y_1|X_2,\cdots,X_{p-1})e(X_1|X_2,\cdots,X_{p-1})}$ is the coefficient of determination of

$$e(Y_1|X_2,\cdots,X_{p-1}) = \beta_0' + \beta_1' e(X_1|X_2,\cdots,X_{p-1}) + \varepsilon' \tag{2}$$

We only need to show $SSE = SSE'$ since

$$\begin{aligned} R^2_{Y1|2\cdots(p-1)} &= 1 - \frac{SSE}{SSTO} \\ R^2_{e(Y_1|X_2,\cdots,X_{p-1})e(X_1|X_2,\cdots,X_{p-1})} &= 1 - \frac{SSE'}{SSTO'} \\ SSTO' &= \sum_{i=1}^{n} e(Y_1|X_2,\cdots,X_{p-1})^2 \\ &= SSTO \end{aligned}$$

Without loss of generality, we assume $X_1,\cdots,X_{p-1}$ are centered (so that we can use the conclusion of 6.1). Let

$$\begin{aligned} &\mathbf{X}_2 = \begin{pmatrix} X_2 & \cdots & X_{p-1} \end{pmatrix} \\ &\mathbf{H}_0 = \frac{1}{n}\mathbf{1}\mathbf{1}^T \\ &\mathbf{H}_1 = X_1(X_1^T X_1)^{-1} X_1^T \\ &\mathbf{H}_2 = \mathbf{X}_2(\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T \\ &\mathbf{H}_e = e(X_1|X_2,\cdots,X_{p-1}) \\ &\qquad [e(X_1|X_2,\cdots,X_{p-1})^T e(X_1|X_2,\cdots,X_{p-1})]^{-1} e(X_1|X_2,\cdots,X_{p-1})^T \end{aligned}$$

then

$$\begin{aligned} e(X_1|X_2,\cdots,X_{p-1}) &= (\mathbf{I} - \mathbf{H}_0 - \mathbf{H}_2)X_1 \\ &= (\mathbf{I} - \mathbf{H}_2)X_1 \\ e(Y|X_2,\cdots,X_{p-1}) &= (\mathbf{I} - \mathbf{H}_0 - \mathbf{H}_2)Y. \end{aligned}$$

$\therefore$

$$\mathbf{H}_e = (\mathbf{I} - \mathbf{H}_2)X_1[X_1^T(\mathbf{I} - \mathbf{H}_2)X_1]^{-1}X_1^T(\mathbf{I} - \mathbf{H}_2)$$

$\because$ $\mathbf{H}_1, \mathbf{H}_2$ are projection matrixes onto $Range(X_1) \subset Range(\mathbf{X}_r)$, $Range(\mathbf{X_2}) \subset Range(\mathbf{X}_r)$ respectively, where $\mathbf{X}_r = \begin{bmatrix} X_1 & \mathbf{X_2} \end{bmatrix}$, $Range(\mathbf{A})$ denotes the column space of matrix $\mathbf{A}$
$\therefore$

$$\begin{aligned} Range(\mathbf{X}_r) &= Range(X_1) \oplus Range((\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2) \\ &= Range(\mathbf{X}_2) \oplus Range((\mathbf{I} - \mathbf{H}_2)X_1) \end{aligned}$$

$\therefore$

$$\begin{aligned} \mathbf{H}_r &= \begin{bmatrix} X_1 & \mathbf{X} \end{bmatrix} \begin{bmatrix} X_1^T X_1 & X_1^T\mathbf{X} \\ \mathbf{X}^T X_1 & \mathbf{X}^T\mathbf{X} \end{bmatrix}^{-1} \begin{bmatrix} X_1^T \\ \mathbf{X}^T \end{bmatrix} \\ &= \mathbf{H}_1 + (\mathbf{I} - \mathbf{H}_1)\mathbf{X}[\mathbf{X}^T(\mathbf{I} - \mathbf{H}_1)\mathbf{X}]^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{H}_1) \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{H}_2 + (\mathbf{I}-\mathbf{H}_2)[X_1^T(\mathbf{I}-\mathbf{H}_2)X_1]^{-1}X_1^T(\mathbf{I}-\mathbf{H}_2) \\
&= \mathbf{H}_2 + \mathbf{H}_e
\end{aligned}
$$

$\because$ from the property of residuals,

$$
\begin{aligned}
\mathbf{1}^T e(Y|X_2,\cdots,X_{p-1}) &= \mathbf{1}^T e(X_1|X_2,\cdots,X_{p-1}) \\
&= \mathbf{0}
\end{aligned}
$$

from model (2), $\forall\ j \in \{2,\cdots,p-1\}$,

$$
X_j^T e(X_1|X_2\cdots,X_{p-1}) = \mathbf{0}
$$

i.e.

$$
\mathbf{X}_2 e(X_1|X_2\cdots,X_{p-1}) = \mathbf{0}
$$

$\therefore$

$$
\begin{aligned}
\mathbf{H}_e\mathbf{H}_0 &= \mathbf{0} \\
\mathbf{H}_e\mathbf{H}_2 &= \mathbf{0}
\end{aligned}
$$

$\because$

$$
\begin{aligned}
(\mathbf{I}-\mathbf{H}_0)\mathbf{H}_0 &= \mathbf{0} \\
(\mathbf{I}-\mathbf{H}_2)\mathbf{H}_2 &= \mathbf{0} \\
\mathbf{H}_0\mathbf{H}_2 = \mathbf{H}_2\mathbf{H}_0 &= \mathbf{0}
\end{aligned}
$$

$\therefore$

$$
\begin{aligned}
SSE' &= e(Y|X_2,\cdots,X_{p-1})^T(\mathbf{I}-\mathbf{H}_0-\mathbf{H}_e)e(Y|X_2,\cdots,X_{p-1}) \\
&= e(Y|X_2,\cdots,X_{p-1})^T(\mathbf{I}-\mathbf{H}_e)e(Y|X_2,\cdots,X_{p-1}) \\
&= \mathbf{Y}^T(\mathbf{I}-\mathbf{H}_0-\mathbf{H}_2)(\mathbf{I}-\mathbf{H}_e)(\mathbf{I}-\mathbf{H}_0-\mathbf{H}_2)\mathbf{Y} \\
&= \mathbf{Y}^T(\mathbf{I}-\mathbf{H}_0-\mathbf{H}_2-\mathbf{H}_e)(\mathbf{I}-\mathbf{H}_0-\mathbf{H}_2)\mathbf{Y} \\
&= \mathbf{Y}^T(\mathbf{I}-\mathbf{H}_0-\mathbf{H}_2-\mathbf{H}_e)\mathbf{Y} \\
&= \mathbf{Y}^T(\mathbf{I}-\mathbf{H}_0-\mathbf{H}_r)\mathbf{Y} \\
&= SSE
\end{aligned}
$$

$\square$

10.1.3

$$
\begin{aligned}
\Gamma_p &= \frac{1}{\sigma^2}\left[\sum_{i=1}^n(\mathbb{E}\hat{Y}_i-\mu_i)^2+\sum_{i=1}^n Var\hat{Y}_i\right] \\
&= \frac{\mathbb{E}SSE_p}{\sigma^2}-(n-2p)
\end{aligned}
$$

Proof.
$\because$

$$SSE_p = \mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

$\therefore$

$$\begin{aligned}
\mathbb{E}SSE_p &= \mathbb{E}[\mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbf{Y}] \\
&= \mathrm{tr}[(\mathbf{I}-\mathbf{H})]Var\mathbf{Y} + \mathbb{E}\mathbf{Y}^T(\mathbf{I}-\mathbf{H})\mathbb{E}\mathbf{Y} \\
&= (n-p)\sigma^2 + \mu^T(\mathbf{I}-\mathbf{H})\mu \\
&= (n-p)\sigma^2 + [(\mathbf{I}-\mathbf{H})\mu]^T[(\mathbf{I}-\mathbf{H})\mu] \\
&= (n-p)\sigma^2 + (\mu-\mathbb{E}\hat{\mathbf{Y}})^T(\mu-\mathbb{E}\hat{\mathbf{Y}}) \\
&= (n-p)\sigma^2 + \sum_{i=1}^{n}(\mathbb{E}\hat{Y}_i-\mu_i)^2
\end{aligned}$$

$\because$

$$\begin{aligned}
Var\hat{Y}_i &= Var(\mathbf{X}_i^T\mathbf{b}) \\
&= Var(\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}) \\
&= \mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TVar(\mathbf{Y})(\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TVar(\mathbf{X}\beta+\varepsilon)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i \\
&= \mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^TVar(\varepsilon)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i \\
&= \sigma^2\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i \\
&= \sigma^2\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i \\
&= h_{ii}\sigma^2
\end{aligned}$$

$\therefore$

$$\begin{aligned}
\Gamma_p &= \frac{1}{\sigma^2}\mathbb{E}\sum_{i=1}^{n}(\hat{Y}_i-\mu_i)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}(\hat{Y}_i-\mu_i)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}(\hat{Y}_i-\mathbb{E}\hat{Y}_i+\mathbb{E}\hat{Y}_i-\mu_i)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}[\mathbb{E}(\hat{Y}_i-\mathbb{E}\hat{Y}_i)^2+\mathbb{E}(\mathbb{E}\hat{Y}_i-\mu_i)^2+2\mathbb{E}(\hat{Y}_i-\mathbb{E}\hat{Y}_i)\mathbb{E}(\mathbb{E}\hat{Y}_i-\mu_i)] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}[Var\hat{Y}_i+(\mathbb{E}\hat{Y}_i-\mu_i)^2] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(\mathbb{E}\hat{Y}_i-\mu_i)^2+\sum_{i=1}^{n}Var\hat{Y}_i\right] \\
&= \frac{1}{\sigma^2}\left[\mathbb{E}SSE_p-(n-p)\sigma^2+\sum_{i=1}^{n}h_{ii}\sigma^2\right] \\
&= \frac{\mathbb{E}SSE_p}{\sigma^2}-(n-2p)
\end{aligned}$$

□

11.2.4

(1)

$$d_i = \frac{e_i}{1-h_{ii}}$$

Proof.

Lemma : Woodbury Matrix Identity If A, C are invertible, then

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}$$

$\because$ let $A = \mathbf{X}^T\mathbf{X}$, $U = X_i^T$, $C = -1$ and $V = X_i$,

$$\begin{aligned}
(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1} &= (\mathbf{X}^T\mathbf{X} - X_iX_i^T)^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1} - (\mathbf{X}^T\mathbf{X})^{-1}X_i^T(-1+X_i(\mathbf{X}^T\mathbf{X})^{-1}X_i^T)^{-1}X_i(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}X_i^TX_i(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_{ii}}
\end{aligned}$$

$\therefore$

$$\begin{aligned}
d_i &= Y_i - \hat{Y}_{i(i)} \\
&= Y_i - X_i^T\beta_{(i)} \\
&= Y_i - X_i^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^TY_{(i)} \\
&= Y_i - X_i^T\left[(\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}X_i^TX_i(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_{ii}}\right]\mathbf{X}_{(i)}^TY_{(i)} \\
&= \frac{1}{1-h_{ii}}\left[(1-h_{ii})Y_i - (1-h_{ii})X_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{(i)}^TY_{(i)} - h_{ii}X_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{(i)}^TY_{(i)}\right] \\
&= \frac{1}{1-h_{ii}}[(1-h_{ii})Y_i - X_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{(i)}^TY_{(i)}] \\
&= \frac{1}{1-h_{ii}}[(1-h_{ii})Y_i - X_i^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^TY_i - X_iY_i)] \\
&= \frac{1}{1-h_{ii}}[(1-h_{ii})Y_i - X_i^T\hat{\beta} + h_{ii}Y_i] \\
&= \frac{Y_i - X_i^T\hat{\beta}}{1-h_{ii}} \\
&= \frac{e_i}{1-h_{ii}}
\end{aligned}$$

□

(2)

$$s^2\{d_i\} = \frac{MSE_{(i)}}{1-h_{ii}}$$

Proof.

Let $A = 1$, $U = X_i^T$, $V = X_i$ and $C = -(\mathbf{X}^T\mathbf{X})^{-1}$,

$$s^2\{d_i\} = MSE_{(i)}[1+X_i^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}X_i]$$

$$
\begin{aligned}
&= MSE_{(i)}[1 + X_i^T(\mathbf{X}^T\mathbf{X} - X_iX_i^T)^{-1}X_i] \\
&= MSE_{(i)}\left\{1 + X_i^T\left[(\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}X_i^TX_i(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_{ii}}\right]X_i\right\} \\
&= MSE_{(i)}\left(1 - h_{ii} + \frac{h_{ii}^2}{1-h_{ii}}\right) \\
&= \frac{MSE_{(i)}}{1-h_{ii}}
\end{aligned}
$$

□

11.2.5

$$SSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

Proof.

Similar to 12.2.4 (1), we have $\forall\ j \neq i$,

$$
\begin{aligned}
Y_j - \hat{Y}_{j(i)} &= \frac{1}{1-h_{ii}}\left[(1-h_{ii})Y_j - (1-h_{ii})X_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{(i)}^TY_{(i)} - h_{ij}X_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{(i)}^TY_{(i)}\right] \\
&= \frac{1}{1-h_{ii}}\left[(1-h_{ii})Y_i - (1-h_{ii})(\hat{Y}_j - h_{ij}Y_i) - h_{ii}(\hat{Y}_i - h_{ii}Y_i)\right] \\
&= (Y_j - \hat{Y}_j) + \frac{h_{ij}}{1-h_{ii}}(Y_i - \hat{Y}_i) \\
&= e_j + \frac{h_{ij}}{1-h_{ii}}e_i
\end{aligned}
$$

$\because$ from 12.3.1 we have

$$
\begin{aligned}
\sum_{j=1}^n h_{ij}^2 &= \sum_{j=1}^n h_{ij}h_{ji} \\
&= h_{ii}
\end{aligned}
$$

$$\sum_{j=1}^n h_{ij}e_j = 0$$

$\therefore$

$$
\begin{aligned}
SSE_{(i)} &= \sum_{\substack{j=1\\ j=\neq i}}^n (Y_j - \hat{Y}_{j(i)})^2 \\
&= \sum_{\substack{j=1\\ j\neq i}}^n (e_j + \frac{h_{ij}}{1-h_{ii}}e_i)^2 \\
&= \sum_{\substack{j=1\\ j\neq i}}^n \left[e_j^2 + \frac{h_{ij}^2}{(1-h_{ii})^2}e_i^2 + 2\frac{h_{ij}e_j}{1-h_{ii}}e_i\right] \\
&= \sum_{\substack{j=1\\ j\neq i}}^n e_j^2 + \frac{h_{ii}-h_{ii}^2}{(1-h_{ii})^2}e_i^2 - 2\frac{h_{ii}}{1-h_{ii}}e_i^2
\end{aligned}
$$

$$= \sum_{\substack{j=1 \\ j \neq i}}^{n} e_j^2 - \frac{h_{ii}}{1-h_{ii}} e_i^2$$

$$= SSE - \frac{e_i^2}{1-h_{ii}}$$

$\square$

11.3.1

(1)

$$h_{ij} = h_{ji}$$

Proof.

$\because$

$$\mathbf{H}^T = \mathbf{H}$$

$\therefore$

$$h_{ij} = h_{ji}$$

$\square$

(2)

$$\sum_{i=1}^{n} h_{ii} = p$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} h_{ii} &= trace(\mathbf{H}) \\
&= trace(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\
&= trace((\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})) \\
&= trace(\mathbf{I}_p) \\
&= p
\end{aligned}$$

$\square$

(3)

$$\mathbf{JH} = \mathbf{J}$$

$$\mathbf{H1} = \mathbf{1}$$

$$\mathbf{1}^T\mathbf{H} = \mathbf{1}^T$$

$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$$

Proof.

$\because$

$$\mathbf{HX} = \mathbf{X}$$
$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{X}_r \end{bmatrix}$$

$\therefore$ consider the coefficient of X_{i1},

$$\sum_{j=1}^{n} h_{ij} = 1$$

i.e.

$$\mathbf{H1} = \mathbf{1}$$

$\because$

$$\mathbf{H}^T = \mathbf{H}$$

$\therefore$

$$\sum_{j=1}^{n} h_{ij} = 1$$

i.e.

$$\mathbf{1}^T\mathbf{H} = \mathbf{1}^T$$

$\therefore$

$$\mathbf{JH} = \mathbf{J}$$

$\square$

(4)

$$0 \leqslant h_{ii} \leqslant 1$$

Proof.

$\because$

$$\mathbf{HH} = \mathbf{H}$$

$\therefore$

$$h_{ii} = \sum_{i=1}^{n} h_{ij}h_{ji} = \sum_{i=1}^{n} h_{ij}^2 \geqslant 0$$

$\because$

$$(\mathbf{I} - \mathbf{H})^2 = \mathbf{I} - \mathbf{H}$$

$\therefore$

$$1 - h_{ii} = \sum_{i=1}^{n} (\delta_{ij} - h_{ij})^2 \geqslant 0$$

i.e.

$$h_{ii} \leqslant 1$$

$\square$

(5)

$$\sum_{j=1}^{n} h_{ij} e_j = 0$$

$Proof.$

$\because$

$$\begin{aligned}\mathbf{He} &= \mathbf{H}(\mathbf{Y} - \hat{\mathbf{Y}}) \\ &= \hat{\mathbf{Y}} - \hat{\mathbf{Y}} \\ &= \mathbf{0}\end{aligned}$$

$\therefore \quad \forall\ i,$

$$\sum_{j=1}^{n} h_{ij} e_j = 0$$

$\square$