STAT 425 Note 1

Wenxiao Yang*

*Department of Mathematics, University of Illinois at Urbana-Champaign

2021

目录

1 Review of statistics

1.1 Random Vectors

1.1.1 Mean

$$\mu = \mathbb{E}(\mathbf{Z}) = \begin{pmatrix} \mathbb{E}(Z_1) \\ \mathbb{E}(Z_2) \\ \cdots \\ \mathbb{E}(Z_m) \end{pmatrix}$$

1.1.2 Variance-Covariance matrix Σ

$$\Sigma_{m\times m} = Cov(\mathbf{Z}) = \mathbb{E}((\mathbf{Z}-\mu)(\mathbf{Z}-\mu)^T) = \begin{bmatrix} Var(Z_1) & \cdots & Cov(Z_1, Z_m) \\ \cdots & \cdots & \cdots \\ Cov(Z_m, Z_1) & \cdots & Var(Z_m) \end{bmatrix}$$

1.2 Affine Transformation

(1)

$$\mathbf{W} = \mathbf{a}_{n\times 1} + \mathbf{B}_{n\times m}\mathbf{Z}_{m\times 1}$$

$$\mathbb{E}(\mathbf{W}) = \mathbf{a} + \mathbf{B}\mu,\ Cov(\mathbf{W}) = \mathbf{B}\Sigma\mathbf{B}^T$$

(2)

$$\mathbf{W} = \mathbf{v}^T\mathbf{Z} = v_1Z_1 + ... + v_mZ_m$$

$$\mathbb{E}(\mathbf{W}) = \mathbf{v}^T\mu = \sum_{i=1}^{m} v_i\mu_i$$

$$Var(\mathbf{W}) = \mathbf{v}^T\Sigma\mathbf{v} = \sum_{i=1}^{m} v_i^2 Var(Z_i) + 2\sum_{i<j} v_iv_jCov(Z_i, Z_j)$$

$$\text{i.e. } \mathbb{E}(\mathbf{AZ}) = \mathbf{A}\mathbb{E}(Z);\ Var(\mathbf{AZ}) = \mathbf{A}Var(\mathbf{Z})\mathbf{A}^T$$

(3)

$$Cov(\mathbf{AX}, \mathbf{BY}) = \mathbb{E}[(\mathbf{AX}-\mathbf{A}\mathbb{E}(X))(\mathbf{BY}-\mathbf{B}\mathbb{E}(Y))^T] = \mathbf{A}\mathbb{E}[(\mathbf{X}-\mathbb{E}(X))(\mathbf{Y}-\mathbb{E}(Y))^T]\mathbf{B}^T = \mathbf{A}Cov(\mathbf{X}, \mathbf{Y})\mathbf{B}^T$$

2 Regression Analysis (SLR)

It is a "tool" used to examine the relationship between a **Dependent Variable** or **Response** Y, and one (or more) **Independent Variables** or **Regressors** or **Predictors** $X_1, X_2, ..., X_p$.

2.1 Simple Linear Regression

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

β_0 is the *intercept*; β_1 is the *slope*. One Response $\mathcal{Y}$; One Predictor $\mathcal{X}$ The data come in pairs:

$$\begin{matrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{matrix}$$

Y is a RANDOM VARIABLE that has a distribution for every level of the independent variable.

2.2 Simple Linear Regression Model

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

where the *intercept* β_0, the *slope* β_1, and the *error variance* σ^2 are the *model parameters*.

2.2.1 Assumptions of errors ε: 1. Mean zero, 2. umcorrelated, 3. homoscedastic

The *errors* $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are assumed to
–have *mean zero*: $E(\varepsilon_i) = 0$
–be *uncorrelated*: $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$
–be *homoscedastic*: $Var(\varepsilon_i) = \sigma^2$ does not depend on i.

The last two could be combined and written as:

$$Cov(\varepsilon_1, \varepsilon_j) = \sigma^2 \delta_{ij}$$

where $\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$

2.2.2 Assumptions on $Y|X$

Based on the SLR model moment assumptions on the error terms, we have the following assumptions for the moments of Yconditioning on X:
1. $E(y_i|x_i) = \beta_0 + \beta_1 x_i$
2. $Var(y_i|x_i) = \sigma^2$
3. $Cov(y_i, y_j|x_i, x_j) = 0, \ i \neq j$

2.2.3 Interpretation of β_1, β_0

β_1 is the **change in the mean** of the probability distribution function of y per unit change in x. When $x = 0$, β_0 is the **mean** of the probability distribution function of y(at $x = 0$), otherwise β_0 **has no particular meaning**.

2.3 Least Squares

We want to find estimates of β_0, β_1 to minimize:

$$\min[y_i - E(y_i)] \Leftrightarrow \min[y_i - (\beta_0 + \beta_1 x_i)]$$

minimize the **Residual Sum of Squares (RSS)**

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1)}{\operatorname{argmin}} RSS$$

$$\frac{\partial RSS}{\partial \beta_0} = 0 \Leftrightarrow -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Leftrightarrow \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\frac{\partial RSS}{\partial \beta_1} = 0 \Leftrightarrow -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\Leftrightarrow \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

2.3.1 LS Estimators

Then we can solve that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Alternative Representation of $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r_{xy} \sqrt{\frac{S_{yy}}{S_{xx}}}$$

Where

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}); \qquad S_{xx} = \sum_i (x_i - \bar{x})^2$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2; \qquad r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

2.3.2 Fitted Values & Residuals

The *Prediction of y_i* or the *fitted value at x_i*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}) = \bar{y} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(x_i - \bar{x})$$

The *i^{th} residual*

$$r_i = y_i - \hat{y}_i$$

2.3.3 Properties of residuals

1. $\sum_i r_i = 0$
2. $RSS = \sum_i r_i^2$ is minimized
3. $\sum_{i=1} y_i = \sum_{i=1} \hat{y}_i$
4. $\sum_i x_i r_i = 0$ 一阶导条件（proof: $\sum_i x_i r_i = \sum_i x_i(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})) = \sum_i x_i y_i - n\bar{x}\bar{y} - \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}(\sum_i x_i^2 - n\bar{x}^2) = 0$)
5. $\sum_i \hat{y}_i r_i = 0$ (inferred from 4)
6. The regression line always goes through the point $(\bar{x}, \bar{y})$.

2.3.4 Degree of freedom

The **degree of freedom(df)** of the residuals is

$$df = \textit{(Sample size)} - \textit{(\# of parameters)}$$

$df = 2$ in this case.

2.3.5 (Sample) Error variance

The error variance is estimated by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i r_i^2$$

2.4 Goodness of Fit: $R-$square

2.4.1 TSS, RSS, FSS

$TSS : \sum_i (y_i - \bar{y})^2$
$RSS : \sum_i r_i^2$
$FSS : \sum_i (\hat{y}_i - \bar{y})^2$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (r_i + \hat{y}_i - \bar{y})^2$$
$$= \sum_i r_i^2 + \sum_i (\hat{y}_i - \bar{y})^2$$
$$TSS = RSS + FSS$$

2.4.2 Cofficient of Determination(R^2)

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{FSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$0 \leq R^2 \leq 1$

It measures the effect of X in reducing the variation in Y.

The larger R^2 is, the more the total variation of y is reduced by reducing the independent variable x.

R^2 can also represent the degree of linear association between X and Y.

$r_{xy} = \pm\sqrt{R^2}$, where the sign is the sign of the slope.

$$\begin{aligned} r_{xy}^2 &= \frac{(\sum_i (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} = \frac{(\sum_i (x_i - \bar{x})(r_i + \hat{y}_i - \bar{y}))^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} \\ &= \frac{(\sum_i (x_i - \bar{x})(\hat{y}_i - \bar{y}))^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} = \frac{(\sum_i \left(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}\right)(\hat{y}_i - \bar{y}))^2}{\sum_i \left(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}\right)^2 \sum_i (y_i - \bar{y})^2} \\ &= \frac{(\sum_i (\hat{y}_i - \bar{y})^2)^2}{\sum_i (\hat{y}_i - \bar{y})^2 \sum_i (y_i - \bar{y})^2} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = R^2 \end{aligned}$$

2.5 Affine Transformations

Suppose we have a SLR model of Y on X, i.e. $y_i = \beta_0 + \beta_1 x_i$

2.5.1 $\tilde{y}_i = a y_i + b$

1. Rescale y_i by $\tilde{y}_i = a y_i + b$ and then regress $\tilde{y}_i$ on x_i. How would the LS estimates and R^2 be affected?

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(a y_i + b - a\bar{y} - b)}{\sum_{i=1}^n (x_i - \bar{x})^2} = a\hat{\beta}_1 \\ \tilde{\beta}_0 &= a\bar{y} + b - \tilde{\beta}_1 \bar{x} = a\hat{\beta}_0 + b \\ \tilde{R}^2 &= \frac{\sum_i (a\hat{y}_i + b - a\bar{y} - b)^2}{\sum_i (a y_i + b - a\bar{y} - b)^2} = R^2 \end{aligned}$$

2.5.2 $\tilde{x_i} = ax_i + b$

2. Rescale y_i by $\tilde{x_i} = ax_i + b$ and then regress y_i on $\tilde{x_i}$. How would the LS estimates and R^2 be affected?

$$\tilde{\beta_1} = \frac{\sum_{i=1}^{n} (ax_i + b - a\bar{x} - b)(y_i - \bar{y})}{\sum_{i=1}^{n} (ax_i + b - a\bar{x} - b)^2} = \frac{\hat{\beta}_1}{a}$$

$$\tilde{\beta_0} = \bar{y} - \tilde{\beta_1}(a\bar{x} + b) = \hat{\beta}_0 - \frac{b}{a}\hat{\beta}_1$$

$$\tilde{R}^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = R^2$$

2.5.3 Regress x on y instead

3. Regress x on y instead

$$x = \tilde{\beta}_0 + \tilde{\beta}_1 y$$

$$\tilde{\beta}_1 = \frac{S_{xy}}{S_{yy}};\ \tilde{\beta}_0 = \bar{x} - \tilde{\beta}_1 \bar{y};\ \tilde{R}^2 = r_{xy}^2 = R^2$$

2.6 Regression Through the Origin

$$y_i \approx \beta_1 x_i$$

(1) $\hat{\beta}_1$:
By LS: $\min_{\hat{\beta}_1} RSS = \sum_i (\hat{\beta}_1 x_i - y_i)^2$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = \sum_i 2x_i(\hat{\beta}_1 x_i - y_i) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(2) R^2:
negative R^2 is possible since $R^2 = 1 - \frac{RSS}{TSS}$ and RSS may be larger than TSS.
We use a modified R-square

$$\sum_i y_i^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i \hat{y}_i^2$$

$$\tilde{R}^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2} = 1 - \frac{RSS}{\sum_i y_i^2}$$

2.7 LS Estimators Properties

2.7.1 Unbiasedness of LS Estimators $E(\hat{\beta}_1) = \beta_1, E(\hat{\beta}_0) = \beta_0$

x_i's ($\mathcal{X}$) are already known.

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}}\left(\hat{\beta}_1\right) &= \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x})\, y_i}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x}) \cdot \mathbb{E}(y_i)}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{\sum_i (x_i - \bar{x}) \cdot \mathbb{E}(\beta_0 + \beta_1 x_i)}{\sum_i (x_i - \bar{x})^2} = \sum_i c_i (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \sum c_i + \beta_1 \sum c_i x_i = \beta_1 \ , where\ c_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_0\right) &= \mathbb{E}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \\
&= \mathbb{E}(\bar{y}) - \bar{x} \cdot \mathbb{E}\left(\hat{\beta}_1\right) = \frac{1}{n}\sum_i \mathbb{E}(y_i) - \bar{x} \cdot \beta_1 \\
&= \frac{1}{n}\sum_i \mathbb{E}(\beta_0 + \beta_1 x_i) - \bar{x} \cdot \beta_1 \\
&= \beta_0 + \bar{x} \cdot \beta_1 - \bar{x} \cdot \beta_1 = \beta_0
\end{aligned}$$

2.7.2 Mean squared error(MSE) of LS Estimators = $Var(\hat{\beta}_1) = \sigma^2 \frac{1}{S_{xx}}$, $Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

Mean squared error(MSE)= $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Note that since both estimators are unbiased $\Rightarrow$ MSE = Variance.

1. MSE for slope

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_1\right) &= \mathrm{Var}\left[\frac{\sum_i (x_i - \bar{x})\, y_i}{\sum_i (x_i - \bar{x})^2}\right] = \mathrm{Var}\left(\sum_i c_i y_i\right) (c_i \text{ as before }) \\
&= \sum_i c_i^2 \cdot \mathrm{Var}(y_i) = \sum_i c_i^2 \sigma^2 (\text{ from model assumption }) \\
&= \sigma^2 \cdot \left(\frac{\sum_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right)^2 = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \sigma^2 \frac{1}{S_{xx}}
\end{aligned}$$

2. MSE for intercept

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \mathrm{Var}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

2.8 Normality Assumption

Additionally, we assume that

$$\varepsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2)$$

Equivalently,

$$y_i \sim^{iid} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

(y_i are a linear shift of the ε_i, so it is also normally distributed)
(The y_i's are jointly normal, and so are linear combinations of the y_i's, since the errors are normally distributed and uncorrelated/independent.)

2.9 Distribution of LS Estimators

$\hat{\beta}_1$ and $\hat{\beta}_0$ are jointly normally distributed with

$$\mathbb{E}\left(\hat{\beta}_1\right) = \beta_1 \quad \text{Var}\left(\hat{\beta}_1\right) = \sigma^2 \frac{1}{\mathrm{S}_{xx}}$$
$$\mathbb{E}\left(\hat{\beta}_0\right) = \beta_0 \quad \text{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$
$$\text{Cov}\left(\hat{\beta}_1, \hat{\beta}_0\right) = -\sigma^2 \frac{\bar{x}}{S_{xx}}$$

$RSS = \sum_i (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi^2_{n-2}$ which implies that

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \mathbb{E}\left(\frac{RSS}{n-2}\right) = \frac{\sigma^2(n-2)}{n-2} = \sigma^2$$

$\left(\hat{\beta}_0, \hat{\beta}_1\right)$ and RSS are independent.

2.10 Hypothesis Testing (T-test)

2.10.1 Testing for the Slope

$$\begin{cases} H_0 : \beta_1 = c(null) \\ H_\alpha : \beta_1 \neq c \text{ (alternative)} \end{cases}$$

where c is an known constant. The test statistics is

$$t = \frac{\hat{\beta}_1 - c}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{S_{xx}}}$$

The distribution of t under the null is T_{n-2}. The p-value is twice the area under the T_{n-2} distribution more extreme than the observed statistic t.

2.10.2 Testing for the Intercept

$$\begin{cases} H_0 : \beta_0 = c(null) \\ H_\alpha : \beta_0 \neq c \text{ (alternative)} \end{cases}$$

The test statistics is

$$t = \frac{\hat{\beta}_0 - c}{\sqrt{\text{Var}\left(\hat{\beta}_0\right)}}$$

The distribution of t under the null is T_{n-2}. The p-value is twice the area under the T_{n-2} distribution more extreme than the observed statistic t.

2.11 ANOVA Table & F-Test

2.11.1 Degrees of Freedom

$df_{TSS} = n - 1$: one df is lost, because the sample mean is used to estimate the population mean.
$df_{RSS} = n - 2$: two df are lost, because the two parameters are estimated in obtaining the fitted values $\hat{y}$
$df_{FSS} = 1$: there are n deviations $\hat{y}_i - \bar{y}$, but all the fitted values are associated with the same regression line.

$$df_{TSS} = df_{RSS} + df_{FSS}$$

Sum of Squares	Expression	df
TSS	$\sum_i (y_i - \bar{y})^2$	$n-1$
FSS	$\sum_i (\hat{y}_i - \bar{y})^2$	1
RSS	$\sum_i (y_i - \hat{y}_i)^2$	$n-2$

2.11.2 ANOVA Table

Source	SS	df	MS	F
Regression (model)	FSS	1	MSReg $= \frac{FSS}{1}$	$F = \frac{MSReg}{MSE}$
Error	RSS	$n-2$	$MSE = \frac{RSS}{n-2}$	
Total	TSS	$n-1$		

图 1:

2.11.3 F-Test (equivalent to t-test)

An alternative way to test for the model parameters is using the F test:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_\alpha : \beta_1 \neq 0 \end{cases}$$

- Under H_0, the F-test statistic is

$$F = \frac{\text{MSReg}}{MSE} = \frac{FSS}{RSS/(n-2)} \sim F_{1,n-2}$$

- It can be shown that the F-test statistic is equal to the square of the t-test statistic and their p-values are the same. So, **this test is equivalent to the t-test before**.

2.12 Estimation and Prediction

2.12.1 Estimation (always reported for a parameter: $\beta_0 + \beta_1 x^* = \mathbb{E}(Y|x^*)$)

1. Estimation: We want to estimate the mean response at x^*. This is equivalent to estimate: $\beta_0 + \beta_1 x^*$

2. Accuracy of the estimation: is measured by the expected value of the squared difference between the point estimate and the target.

- For estimation the target is $\beta_0 + \beta_1 x^\star$:

$$\begin{aligned}
&\mathbb{E}\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star\right)^2 \\
=&\operatorname{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star\right) \\
=&\operatorname{Var}\left(\hat{\beta}_0\right) + (x^\star)^2 \operatorname{Var}\left(\hat{\beta}_1\right) + 2x^\star \operatorname{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\
=&\sigma^2\left(\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right) \\
=&\sigma^2\left(\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

3. Confidence interval: An $(1-\alpha)100\%$ Confidence Interval for the Mean Response when $x = x^\star$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star \pm T_{n-2}(\alpha/2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}}$$

2.12.2 Prediction (is reported for the value of a random variable Y^*)

1. Prediction: of an outcome of random variable Y^* at a given value x^* , where $Y^* \sim N(\beta_0 + \beta_1 x^*, \sigma^2)$

2. For prediction the target is $Y^\star = \beta_0 + \beta_1 x^\star + e^\star$, where $e^\star \sim N\left(0, \sigma^2\right)$ This new error $e^\star$ is independent of the previous n data points, i.e. is independent of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - Y^\star\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star - e^\star\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star\right)^2\right] + \mathbb{E}\left[(e^\star)^2\right] \\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

3. Prediction interval: An $(1-\alpha)100\%$ Prediction Interval for $\hat{Y}^\star$ when $x = x^\star$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star \pm T_{n-2}(\alpha/2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}}$$

2.12.3 Simultaneous Confidence

$\mu^* = \mathbb{E}[y|x^*] = \beta_0 + \beta_1 x^*$'s Confidence Interval:

$$I(x^*) = (\hat{\mu}^* \pm T_{n-2}(\frac{\alpha}{2}) se(\hat{\mu}^*))$$

Where

$$\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* \text{ and } se(\hat{\mu}^*) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

If we want confidence intervals at multiple points $(x_1^*, x_2^*, \cdots, x_m^*)$, we can use formula (1) to have confidence intervals at the m points: $I(x_1^*), I(x_2^*), \cdots, I(x_m^*)$.
We know that:

$$\mathbb{P}\left(\mu_i^* \in I\left(x_i^*\right)\right) = (1 - \alpha)$$

This is the point-wise coverage probability for μ_i^* and formula (1) gives the point-wise CI.
What about the simultaneous coverage probability? i.e.:

$$\mathbb{P}\left(\mu_i^* \in I\left(x_i^*\right), \text{ for } i = 1, \ldots, m\right) = ?$$

To make sure that (for example):

$$\mathbb{P}\left(\mu_i^* \in I\left(x_i^*\right), \text{ for } i = 1, \ldots, m\right) = .95$$

we need to set $\alpha = 5\%/m$, which is known as the **Bonferroni correction**.
Let A_k denotes the event that the k th confidence interval covers μ_k^* with:

$$\mathbb{P}\left(\mathrm{A}_k\right) = (1 - \alpha)$$

Then we can show:

$$\begin{aligned}
&\mathbb{P}\left(\text{ All Cls cover the corresponding } \mu_k^* \text{ values }\right) \\
&= \mathbb{P}\left(\mathrm{A}_1 \cap \mathrm{A}_2 \ldots \cap \mathrm{A}_m\right) \\
&= 1 - \mathbb{P}\left(\mathrm{A}_1^c \cup \mathrm{A}_2^c \ldots \cup \mathrm{A}_m^c\right) \\
&\geq 1 - \mathbb{P}\left(\mathrm{A}_1^c\right) - \ldots - \mathbb{P}\left(\mathrm{A}_m^c\right) \\
&= 1 - m\alpha
\end{aligned}$$

If we choose α/m instead of α, the simultaneous coverage probability will be $(1 - \alpha)$

2.12.4 Confidence Band (Larger than CI)

Ideally we would like to construct a simultaneous confidence band (i.e., $m = \infty$) across all x^* 's. (Scheffé's Theorem - 2959). Let

$$I(x) = (\hat{r}(x) - c\hat{\sigma}, \hat{r}(x) + c\hat{\sigma})$$

where

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, c\hat{\sigma} = \sqrt{2F(\alpha, 2, n-2)}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}}$$

Then,

$$\mathbb{P}(r(x) \in I(x) \text{ for all } x) \geq 1 - \alpha$$

Can we construct a simultaneous prediction band? No!

Are confidence bands always wider than point-wise confidence intervals? Yes! For SLR, at a location x^*, we have

$$\begin{aligned} \text{band} &: \hat{\mu}^* \pm \sqrt{2F(\alpha, 2, n-2)}\, \text{se}\,(\hat{\mu}^*) \\ \text{interval} &: \hat{\mu}^* \pm T_{n-2}(\alpha/2)\, \text{se}\,(\hat{\mu}^*) \end{aligned}$$

$$\sqrt{2F(\alpha, 2, n-2)} > T_{n-2}(\alpha/2) = \sqrt{2F(\alpha, 1, n-2)}$$

In fact, for any α, we have

$$T_m(\alpha/2) = \sqrt{2F(\alpha, 1, m)} < \sqrt{kF(\alpha, k, m)}$$

2.13 Maximum likelihood estimators with normal error terms

We start with the statistical model, which is the Gaussian-noise simple linear regression model, defined as follows:

1. The distribution of X is arbitrary (and perhaps X is even non-random).
2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients", "parameters") β_0 and β_1, and some random noise variable ϵ.
3. $\epsilon \sim N\left(0, \sigma^2\right)$, and is independent of X.
4. ϵ is independent across observations.

$$p(y_i|x_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma})^2}$$

Given any data set $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$, we can now write down the probability density, under the model, of seeing that data:

$$\prod_{i=1}^{n} p(y_i|x_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma})^2} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

Take the **log-likelihood**

$$L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Then we can compute the **Maximum likelihood estimators** $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$:

(1) $(\hat{\beta}_0, \hat{\beta}_1)$,

Obviously, maximizing $L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ is as same as minimizing $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$, then the **Maximum likelihood estimators** is exactly the **LS estimators**:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

(2) $\hat{\sigma}^2$,

And the $\hat{\sigma}^2$ is exactly the in-sample mean squared error:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

3 Multiple Linear Regression

3.1 Basic

$x_1, x_2, \ldots, x_p$ be p predictors of a response y.

The data will be of the form:

$$\begin{array}{ccccc} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{array}$$

Model Equation:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n$$

where we denote $\mathbf{x_i} = (x_{i1,\ldots,x_ip})^T$, with $x_{i1} = 1$

$\left(\beta_1, \beta_2, \ldots, \beta_p; \sigma^2\right)$ are unknown true parameters.

β_1 is the intercept.

$\beta_2, \beta_3, \ldots, \beta_p$ are partial slopes.

σ^2 is the error variance

3.1.1 Assumptions of errors

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are the random errors. They usually assumed to satisfy the same conditions as in simple linear regression:

- zero mean: $\mathbb{E}(\varepsilon_i) = 0$
- uncorrelated: $\mathrm{Cov}(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}) = 0, \mathrm{i} \neq \mathrm{j})$, and
- homoscedastic: $\mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2$ does not depend on i).

3.1.2 Matrix Representation

Matrix Representation of the MLR Model:

$$\begin{array}{ccccccc} \mathbf{y}_{n\times 1} & = & \mathbf{X}_{n\times p} & & \beta_{p\times 1} & + & \varepsilon_{n\times 1} \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ \text{Response} & & \text{Design} & & \text{Coefficients} & & \text{Error} \\ & & \text{Matrix} & & & & \text{Term} \end{array}$$

- n : sample size
- p : number of predictors or columns of X
- By default the intercept is included in the model in which case the first column of X is a vector of

1 's.

We set $\mathbb{E}(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 \mathbf{I}_n$, then we can infer that

$$\mathbb{E}(y) = \mathbf{X}\beta,\ Cov(y) = \sigma^2 \mathbf{I}_n$$

3.2 Parameter Estimation $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$

- We want to estimate β, i.e. obtain:

$$\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)^T$$

- The LS estimator of β minimizes the sum of squared residuals:

$$RSS = \|y - \mathbf{X}\beta\|^2 = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)$$

In order to minimize $RSS = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)$, we take derivatives with respect to β 's and set to zero (as before).

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T_{p\times n}(y - \mathbf{X}\beta)_{n\times 1} = \mathbf{0}_{p\times 1}$$

$\mathbf{X}^T(y - \mathbf{X}\beta) = \mathbf{0} \longrightarrow$ Normal Equations
$\left(\mathbf{X}^T\mathbf{X}\right)\beta = \mathbf{X}^T y$
$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y \rightarrow$ LS Estimators

Remarks

1. We assume that the rank of X is p, i.e. no columns of X is a linear combinations of the other columns of X.
2. Since $\mathbf{X}$ has rank p, the inverse of $\left(\mathbf{X}^T\mathbf{X}\right)$ exists.
3. if $\left(\mathbf{X}^T\mathbf{X}\right)$ is singular the LS solutions is not unique (identifiability problem)

3.2.1 Fitted Values $\hat{y}_{n\times 1} = \mathbf{H}_{n\times n} y_{n\times 1}$

$$\begin{aligned}
\hat{y}_{n\times 1} &= \mathbf{X}\hat{\beta} \\
&= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y \\
&= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top y \\
&= \mathbf{H}_{n\times n} y_{n\times 1}
\end{aligned}$$

$\mathbf{H}_{n\times n} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top$ is called the hat matrix, since it returns the "y-hat" values.

3.3 Residuals $\mathbf{r} = (\mathbf{I} - \mathbf{H})y$, (Sample) error variance $\hat{\sigma}^2 = \frac{\mathbf{r}^T\mathbf{r}}{n-p}$

$$\mathbf{r}_{n\times 1} = y - \hat{y} = y - \mathbf{H}y = (\mathbf{I} - \mathbf{H})y$$

The residuals $\mathbf{r}$ are used to estimate the error variance:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_i r_i^2 = \frac{\mathbf{r}^T\mathbf{r}}{n-p} = \frac{RSS}{n-p}$$

3.4 Properties of residuals

The LS estimator is the β that satisfies the normal equations, that is

$$\mathbf{X}^T(y - \hat{y}) = \mathbf{X}^T(y - \mathbf{X}\hat{\beta}) = \mathbf{0}$$

This implies the following properties for the residuals, $r_{n\times 1} = y - \mathbf{X}\hat{\beta}$:

3.4.1 $\mathbf{X}^T\boldsymbol{r} = 0$

1. The cross-products between the residual vector r and each column of $\mathbf{X}$ are zero, i.e.

$$\begin{aligned}\mathbf{X}^T\boldsymbol{r} &= \mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\mathbf{X}\hat{\beta} \\ &= \mathbf{X}^T\boldsymbol{y} - \left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y} = 0\end{aligned}$$

3.4.2 $\hat{y}^T\boldsymbol{r} = 0$

2. The cross-product between the fitted value $\hat{y}$ and the residual vector r is zero, i.e.

$$\hat{y}^T r = \hat{\beta}^T X^T r = 0$$

This implies that the residual vector r is **orthogonal** to each column of X and $\hat{y}$.

3.5 Properties of H

3.5.1 $\mathbf{HX} = \mathbf{X}$

Let c be any linear combination of the columns of $\mathbb{X}$, then

$$\mathbf{H}c = c$$

since $\mathbf{HX} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X}$

3.5.2 Symmetric: $\mathbf{H}^T = \mathbf{H}$

Symmetric, since $\mathbf{H}^T = \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{H}$

3.5.3 Idempotent: $\mathbf{HH} = \mathbf{H}\mathbf{H}^T = \mathbf{H}^\top\mathbf{H} = \mathbf{H}$

Idempotent, i.e. $\mathbf{HH} = \mathbf{H}\mathbf{H}^T = \mathbf{H}^\top\mathbf{H} = \mathbf{H}$. Indeed

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

3.5.4 $\mathbf{H}(\mathbf{I}-\mathbf{H})=\mathbf{0}$

This also implies that $\mathbf{H}(\mathbf{I}-\mathbf{H})=\mathbf{0}_{n\times n}$

3.5.5 $(\mathbf{I}-\mathbf{H})$ is also symmetric and idempotent

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

3.5.6 trace($\mathbf{H}$) = p

trace($\mathbf{H}$) = p, the number of LS coefficients we estimated.

3.6 Geometric Representation of LS

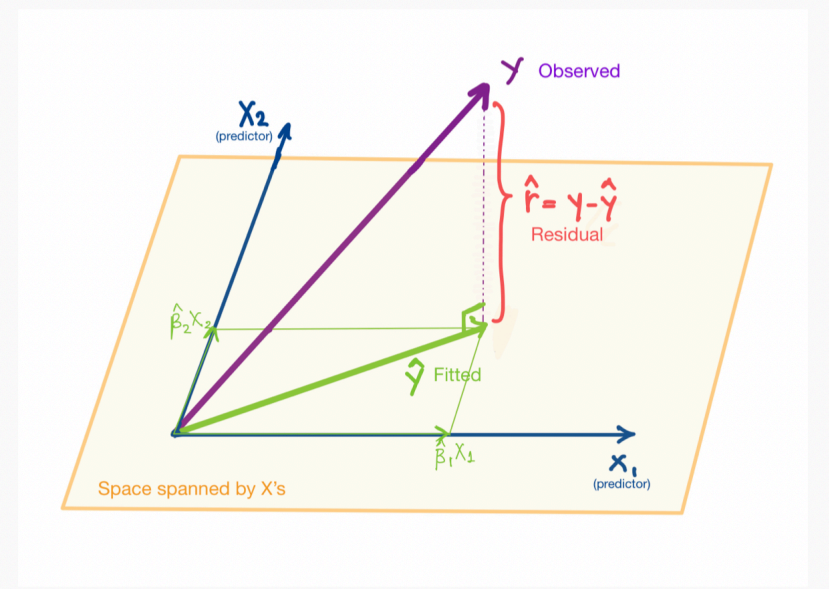

图 2:

3.6.1 Estimation Space

- The columns of $\mathbf{X}$ span a p-dimensional subspace in $\mathbb{R}^n$. This is a subspace that consists of vectors that can be written as linear combinations of the columns of X.
- The LS squares estimator $\hat{\beta}$ is obtained by minimizing the Euclidean distance between the vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$, i.e. $\|y-\hat{y}\|^2$. $\hat{y}$ is the projection of y onto the estimation space.
- $\mathbf{H}_{n\times n}$, projection/hat matrix is symmetric, unique, and idempotent.

3.6.2 Error Space

- The error space is an $(n-p)$-dimensional space that is orthogonal to the estimation space. The projection matrix of the error space is $(\mathbf{I}-\mathbf{H})$.
- The residual $\mathbf{r}$ is the projection of $\boldsymbol{y}$ onto the error space, orthogonal to the estimation space. So, $\mathbf{r}$ is orthogonal to any vector in the estimation space, including each column of X.
- When the intercept is included in the model, then

$$\sum_{i=1}^{n} r_i = 0$$

In general, $\sum_{i=1}^{n} r_i X_{ij} = 0, j = 1, \ldots, p$ due to the normal equations.

3.7 Coefficient of determination, $R-$Square

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

An equivalent definition is

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

$0 \leq R^2 \leq 1$

3.8 Properties of LS Estimators

3.8.1 Unbiased: $\mathbb{E}(\hat{\beta}) = \beta$

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}(y) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta\end{aligned}$$

3.8.2 Variance-Covariance Matrix of $\hat{\beta}$: $\mathrm{Cov}(\hat{\beta}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}) &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\mathrm{Cov}(\mathbf{y})\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right)^\top \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\sigma^2\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\end{aligned}$$

$$se(\hat{\beta}_i) = \hat{\sigma}\sqrt{((\mathbf{X}^T\mathbf{X})^{-1})_{ii}}$$

3.8.3 $\hat{y}$: $\mathbb{E}(\hat{y}) = \mathbf{X}\beta$, $Cov(\hat{y}) = \sigma^2\mathbf{H}$

$$\mathbb{E}(\hat{y}) = \mathbb{E}(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top y) = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta = \mathbf{X}\beta$$
$$Cov(\hat{y}) = Cov(\mathbf{H}y) = \mathbf{H}Cov(y)\mathbf{H}^T = \sigma^2\mathbf{H}\mathbf{H}^T = \sigma^2\mathbf{H}$$

3.8.4 r: $\mathbb{E}(\mathbf{r}) = 0$, $Cov(\mathbf{r}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$

$$\mathbb{E}(\mathbf{r}) = \mathbb{E}(y - \hat{y}) = 0$$
$$Cov(\mathbf{r}) = Cov(y - \hat{y}) = Cov((\mathbf{I} - \mathbf{H})y) = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{I} - \mathbf{H})$$

3.8.5 $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n-p}\mathbb{E}(\mathbf{r}^T\mathbf{r}) = \frac{1}{n-p}\sigma^2(n-p) = \sigma^2$$

$n-p$ 来自于: 有 p 个元素在 H 对角中为 1，其余为 0，所以 $\mathbf{I}_n - \mathbf{H}$ 对角中有 $n-p$ 个 1。所以只有 $n-p$ 个 $r_i^2 = \sigma^2$

$$\frac{\mathbf{r}^T\mathbf{r}}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$$

3.9 The Gauss-Markov Theorem: LS estimator is the BLUE(best linear unbiased estimator)

If the errors are 1. Mean zero, 2. umcorrelated, 3. homoscedastic, the LS estimators have **the lowest variance within the class of linear estimators.**

Suppose we are interested in estimating a linear combination of β of the form:

$$\theta = \mathbf{c}^T\beta = \sum_{j=1}^{p} c_j\beta_j$$

LS estimators:

$$\hat{\theta}_{LS} = \mathbf{c}^T\hat{\beta} = \mathbf{c}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$$

This is a **linear** (linear combination of $y_1, y_2, ..., y_n$) and **unbiased estimator** of θ. Its mean square error can be calculated as:

$$MSE(\hat{\theta}_{LS}) = \mathbb{E}(\hat{\theta}_{LS} - \theta)^2 = Var(\hat{\theta}_{LS})$$

Theorem 1 (Gauss-Markov Theorem)**.** *$\hat{\theta}_{LS} = \mathbf{c}^T\hat{\beta}$ is the **BLUE**(best linear unbiased estimator) of the parameter $\mathbf{c}^T\beta$ for any vector $\mathbf{c} \in \mathbb{R}^p$*

3.10 Maximum Likelihood Estimation, Distribution of LS estimates

3.10.1 $\mathbf{y} \sim \mathbf{N}_n\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$

Recall the normality assumption for the regression model:

$$y_i = \mathbf{x}_i^T\beta + \varepsilon_i, \quad i = 1, \ldots, n, \text{ with } \varepsilon_i \sim N\left(0, \sigma^2\right)$$

This implies that $\mathbf{y} \sim \mathbf{N}_n\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$

3.10.2 LS estimator is the Maximum Likelihood Estimator (MLE)

We can show that the likelihood function can be written as:

$$L\left(\beta, \sigma^2 \mid \mathbf{y}\right) \propto \frac{RSS^{-\frac{n}{2}}}{n}$$

The value of β that maximizes the Likelihood function is *the Maximum Likelihood Estimator (MLE)* of β. This estimator is equal to the LS estimate of β.

3.10.3 $\hat{\beta} \sim \mathbf{N_p}\left(\beta, \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$, $\hat{\mathbf{y}} \sim \mathbf{N_n}\left(\mathbf{X}\beta, \sigma^2\mathbf{H}\right)$, $\hat{\mathbf{r}} \sim \mathbf{N_n}\left(\mathbf{0}, \sigma^2\left(\mathbf{I_n} - \mathbf{H}\right)\right)$

Recall the assumption for the linear regression model:

$$\mathbf{y} \sim \mathbf{N_n}\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$$

Any affine transformation of $\mathbf{y}$ will also have a Normal distribution [2]. We can show that:

$$\hat{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \sim \mathbf{N_p}\left(\beta, \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$$

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \sim \mathbf{N_n}\left(\mathbf{X}\beta, \sigma^2\mathbf{H}\right)$$

$$\hat{\mathbf{r}} = \left(\mathbf{I_n} - \mathbf{H}\right)\mathbf{y} \sim \mathbf{N_n}\left(\mathbf{0}, \sigma^2\left(\mathbf{I_n} - \mathbf{H}\right)\right)$$

3.11 r Residuals'Properties

3.11.1 $\mathbf{r} \in \mathbb{R}^{n-p}$

Although $\mathbf{r}$ is a vector of dimension n, it always lies in a subspace of dimension $(n-p)$.

3.11.2 $\hat{\sigma}^2 \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}$

r behaves like a random vector with a distribution $\mathbf{N}_{n-p}\left(\mathbf{0}, \sigma^2\mathbf{I}_{n-p}\right)$, so we have:

$$\hat{\sigma}^2 = \frac{\|\mathbf{r}\|^2}{n-p} \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}$$

3.11.3 $\hat{\mathbf{y}}$ and $\boldsymbol{r}$ are independent

It can be show that $\hat{\mathbf{y}}$ and $\boldsymbol{r}$ are uncorrelated since they are in orthogonal spaces. Since they also have a joint normal distribution, they are independent.

3.12 Testing Predictors (Coefficients)

3.12.1 Testing a Single Predictor $H_0 : \beta_j = 0$: $t-$test

Suppose you have a p predictors in your regression model and you want to test the hypothesis:

$$H_0 : \beta_j = c \text{ vs. } H_\alpha : \beta_j \neq c$$

- The t-test statistic we use is:

$$t = \frac{\hat{\beta}_j - c}{se\left(\hat{\beta}_j\right)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{\left[(\mathbf{X}^\top \mathbf{X})^{-1}\right]_{jj}}} \sim T_{n-p}$$

under the null hypothesis H_0.

- p-value = $2\times$ the area under the curve of a T_{n-p} distribution more extreme than the observed statistic.

- The p-value returned by the Im function command is for $c = 0$.

3.12.2 Review the degree of freedom

The degrees of freedom of a t-test are determined by the denominator of the estimated variance $\hat{\sigma}^2$. Consider the following situations:

- In STAT 400: Test for $\theta = \alpha$, where $Z_1, \ldots, Z_n \sim \mathcal{N}\left(\theta, \sigma^2\right)$

$$\frac{\hat{\theta} - \alpha}{\text{se}(\hat{\theta})} = \frac{\overline{\mathbf{Z}} - \alpha}{\sqrt{\hat{\sigma}^2/n}} \sim T_{n-1}, \quad \hat{\sigma}^2 = \frac{\sum_i \left(Z_i - \bar{Z}\right)^2}{n-1}$$

- In SLR: Test for $\beta_1 = c$, we have

$$\frac{\hat{\beta}_1 - c}{\text{se}\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{S_{XX}}} \sim T_{n-2}, \quad \hat{\sigma}^2 = \frac{RSS}{n-2}$$

- In MLR with p predictors (including the intercept): Test for $\beta_j = c$,

$$\frac{\hat{\beta}_j - c}{\text{se}\left(\hat{\beta}_j\right)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{\left[(\mathbf{X}^T \mathbf{X})^{-1}\right]_{jj}}} \sim T_{n-p}, \quad \hat{\sigma}^2 = \frac{RSS}{n-p}$$

3.12.3 Testing all predictors: $F-$test

Testing all predictors

$$\begin{cases} H_0 : \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\ H_a : \beta_j \neq 0, \quad \text{for some } j, j = 2, \ldots, p \end{cases}$$

- Under the Null hypothesis, the test statistic:

$$\begin{aligned} F &= \frac{FSS\left(X_2, \ldots, X_p\right)}{p-1} \div \frac{RSS\left(X_2, \ldots, X_p\right)}{n-p} \\ &= \frac{MS(Reg)}{MS(\text{ Error })} \sim F_{p-1, n-p} \end{aligned}$$

Large values of F lead to conclusion H_α.

- This is the overall F test of whether or not there is a regression relation between the response variable Y and the set of X variables.

Source	df	SS	MS	F-test
Regression	$p-1$	FSS	FSS/$(p-1)$	MS(Reg)/MSE
Error	$n-p$	RSS	RSS/$(n-p)$	
Total	$n-1$	TSS		

图 3:

3.12.4 Partial $F-$test

In general, consider the following partition of the design matrix into two sub-matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, that is

$$\mathbf{X}_{n\times p} = \left(\mathbf{X}_{1n\times(p-q)}, \mathbf{X}_{2n\times q}\right)$$

The corresponding partition of the regression parameter is:

$$\beta^T = \left(\beta_1^T, \beta_2^T\right)$$

where β_1 is $(p-q)\times 1$ and β_2 is $q\times 1$

This partition is used to test the hypothesis:

$$\begin{cases} H_0 : \beta_2 = \mathbf{0}, \text{ i.e.,} & \mathbf{y} = \mathbf{X}_1\beta_1 + \text{ error} \\ H_\alpha : \beta_2 \neq \mathbf{0}, \text{ i.e.,} & \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \text{ error} \end{cases}$$

To test this hypothesis, the test statistic is:

$$F = \frac{\left(RSS_0 - RSS_\alpha\right)/q}{RSS_\alpha/(n-p)} \sim F_{q,n-p}$$

where RRS_0 = **Residual sum of squares for the model under** H_0 **;** RRS_a = **Residual sum of squares for the model under** H_α

Numerator: variation in the data not explained by the reduced model, but explained by the full model.

Denominator: variation in the data not explained by the full model (i.e., not explained by either model), which is used to estimate the error variance.
Reject H_0, if F test statistic is large, that is, the variation missed by the reduced model, when being compared with the error variance, is significantly large.
F-test statistic can be rewirtten:

$$F = \frac{(RSS_0 - RSS_\alpha)/q}{RSS_\alpha/(n-p)} = \frac{((1-R_R^2) - (1-R_F^2))/(df_F - df_R)}{(1-R_F^2)/df_F} = \frac{(R_F^2 - R_R^2)/(df_R - df_F)}{(1-R_F^2)/df_F}$$

Note that this test statistic is not appropriate when the full and reduced regression models do not contain the intercept term β_0 .

3.13 Permutation Tests (When the normal distribution hypothesis doesn't hold)

The distribution of the data is *unknown*. - A test statistic is a function of the data; denote it $g($ data $)$.
- The test statistic tends to take extreme values under the alternative hypothesis H_α.

3.13.1 Procedure

Procedure to conduct a permutation test
1. Form the test statistic $g($ data $)$ which tends to take extreme values under the alternative hypothesis.
2. Evaluate the test statistic on the observed data, denoted by g_0.
3. Find the distribution of $g($ data $)$, when data are generated from H_0.
4. Calculate the p-value, that is the following probability:

$$\mathbb{P}\left(g(\text{ data }) \text{ is more extreme than the observed } g_0 \mid \text{data } \sim H_0\right)$$

3.13.2 Calculation of the p-value: Monte Carlo method

We can obtain an approximation of $\mathbb{E}(Y)$ as follows:
1. Generate $N = 1000$ samples from this distribution, $Y_1, \ldots, Y_N$
2. Approximate the mean by

$$\mathbb{E}(Y) \approx \frac{1}{N}\sum_{i=1}^{N} Y_i$$

That is, population mean $\approx$ sample mean (when N is large).

This method also works if we want to approximate the expected value of a *function* of a random variable:

$$\mathbb{E}(f(Y)) \approx \frac{1}{N}\sum_{i=1}^{N} f\left(Y_i\right)$$

3.14 Confidence Intervals for β_j, Confidence Region for β

3.14.1 Confidence Intervals for β_j

Recall that

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} \sim \mathbf{N_p}\left(\beta, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$$

An $(1-\alpha)100\%$ CI(Confidence Interval) for β_j can be written as

$$(\hat{\beta}_j \pm T_{n-p}(\frac{\alpha}{2}) se(\hat{\beta}_j)) = (\hat{\beta}_j \pm T_{n-p}(\frac{\alpha}{2})\hat{\sigma}\sqrt{[(\mathbf{X}^\top \mathbf{X})^{-1}]_{jj}})$$

Justification: The vector β's confidence interval is a family/joint interval for all the betas, so it will be wider than the individual β_j's intervals.

3.14.2 Confidence Region for β

β is the entire vector,

$$\beta - \hat{\beta} \sim \mathbf{N_p}\left(0, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$$

Thus the quadratic form:

$$\frac{(\beta - \hat{\beta})^T \mathbf{X}^T \mathbf{X} (\beta - \hat{\beta})}{p\hat{\sigma}^2} \sim F_{p,n-p}$$

We can construct a $(1-\alpha)100\%$ confidence region for β to be all the points in the following ellipsoid,

$$\frac{(\beta - \hat{\beta})^T \mathbf{X}^T \mathbf{X} (\beta - \hat{\beta})}{p\hat{\sigma}^2} < F(\alpha; p, n-p)$$

Where $F(\alpha; p, n-p)$ is defined to be the point such that

$$\mathbb{P}(F_{p,n-p} > F(\alpha; p, n-p)) = \alpha$$

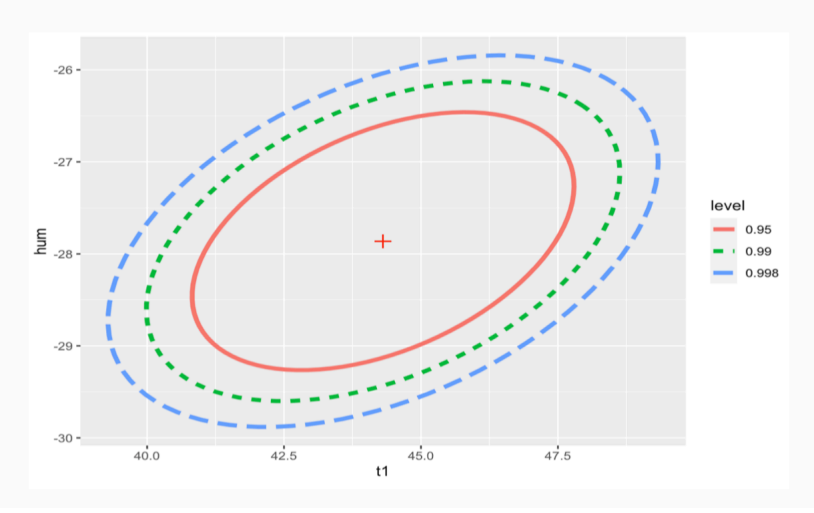

图 4:

3.15 Confidence/Prediction Intervals for New Observations

Set $\mathbb{E}(Y|x^*) = \mu^* = (x^*)^T\beta$

3.15.1 Confidence Interval for $\mu^* = (x^*)^T\beta$

The Gauss-Markov theorem tells us that the BLUE (Best Linear Unbiased Estimate) of μ^* is:

$$\hat{\mu}^* = (x^*)^T\hat{\beta} = (x^*)^T(X^TX)^{-1}X^Ty$$

Then,

$$\begin{aligned}\mathbb{E}[(\hat{\mu}^* - \mu^*)^2] &= Var(\hat{\mu}^*) \\ &= \sigma^2(x^*)^T(X^TX)^{-1}x^* \\ se(\hat{\mu}^*) &= \hat{\sigma}\sqrt{(x^*)^T(X^TX)^{-1}x^*}\end{aligned}$$

Where

$$\frac{\hat{\mu}^* - \mu}{se(\hat{\mu}^*)} \sim t(n-p)$$

A $(1-\alpha)100\%$ CI (confidence interval) for μ^* is:

$$\hat{\mu}^* \pm T_{n-p}(\frac{\alpha}{2})se(\hat{\mu}^*)$$

3.15.2 Prediction Interval for $y^* = (x^*)^T\beta + e^*$

The best estimate for y^* at a future observation x^* is also

$$\hat{y}^* = (x^*)^T\hat{\beta}$$

Then,

$$\begin{aligned}Var(\hat{y}^*) &= \mathbb{E}[((x^*)^T\hat{\beta} - y^*)^2] \\ &= \mathbb{E}[((x^*)^T\hat{\beta} - ((x^*)^T\beta + e^*))^2] \\ &= \mathbb{E}[((x^*)^T\hat{\beta} - ((x^*)^T\beta)^2] + \mathbb{E}[(e^*)^2] \\ &= Var(\hat{\mu}^*) + Var(e) \\ &= \sigma^2[1 + (x^*)^T(X^TX)^{-1}x^*] \\ se(\hat{y}^*) &= \hat{\sigma}\sqrt{1 + (x^*)^T(X^TX)^{-1}x^*}\end{aligned}$$

Where

$$\frac{\hat{y}^* - y^*}{se(\hat{y}^*)} \sim t(n-p)$$

A $(1-\alpha)100\%$ PI (prediction interval) for y^* is:

$$\hat{y}^* \pm T_{n-p}(\frac{\alpha}{2})se(\hat{y}^*)$$

3.15.3 Mahalanobis distance

$$\mathbf{X}_{n\times p} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}^T$$

For any observation vector $\mathbf{x}_{p\times 1} = \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}$, where $\mathbf{z}$ denotes the value of predictors without the intercept.
We would like to use **Mahalanobis distance** to quantify the distance between observation vector $\mathbf{x}_{p\times 1}$ and its sample meaning $\bar{\mathbf{x}}$.
The sample convariance matrix of $\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_n \end{bmatrix}$ is:

$$\hat{\Sigma}_{(p-1)\times(p-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$

Then the $(x^*)^T(X^TX)^{-1}x^*$ can be written as

$$(x^*)^T(X^TX)^{-1}x^* = \frac{1}{n} + \underline{\frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})}$$

The second term in the right hand side ($\frac{1}{n-1}(\mathbf{z}^*-\bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^*-\bar{\mathbf{z}})$) is the so-called Mahalanobis distance from $\mathbf{z}^*$ to the center of the data $\bar{\mathbf{z}}$ (the sample mean).
Then we can write,

$$\begin{aligned}
\text{se}\left(\hat{\mu}^*\right) &= \hat{\sigma}\sqrt{\mathbf{x}^{*\top}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}^*} \\
&= \hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{n-1}\left(\mathbf{z}^* - \bar{\mathbf{z}}\right)^T\hat{\Sigma}^{-1}\left(\mathbf{z}^* - \bar{\mathbf{z}}\right)} \\
se\left(\hat{y}^*\right) &= \hat{\sigma}\sqrt{1 + \left(\mathbf{x}^*\right)^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}^*} \\
&= \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{1}{n-1}\left(\mathbf{z}^* - \bar{\mathbf{z}}\right)^T\hat{\Sigma}^{-1}\left(\mathbf{z}^* - \bar{\mathbf{z}}\right)}
\end{aligned}$$

Since $\text{se}\left(\hat{y}^*\right)$ has an extra 1, when the sample size n goes to infinity,

$$\begin{aligned}
\text{se}\left(\hat{\mu}^*\right) &\to 0 \\
\text{se}\left(\hat{y}^*\right) &\to \sigma
\end{aligned}$$

4 MLR: unusual observations

Recall, that we can write the MLR model as:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\beta, \sigma^2\mathbf{I}_n\right)$$

– Error: assumed to be iid, $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$
– Model: assumed to be linear in the parameters, i.e., $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$
We might have unusual observations.

4.1 High leverage points: $h_i \geq \frac{2p}{n}$

4.1.1 Leverage Points

The diagonal elements of $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top$,

$$h_i = H_{ii} = \mathbf{x}_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i^\top = \frac{Var(x_i^T\hat{\beta})}{\sigma^2}$$

are called **leverages** and are very useful diagnostics. h_i gives a measure (invariant under any affine transformation of $\mathbf{X}$) of how far the i-th observation is from the center of the data (in the X-space). For simple linear regression:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

In general:

$$\begin{aligned} h_i &= \mathbf{x_i^T}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x_i} \\ &= \frac{1}{n} + \frac{1}{n-1}(\mathbf{z}_i - \overline{\mathbf{z}})^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{z}_i - \overline{\mathbf{z}}) \end{aligned}$$

where

$$\hat{\Sigma}^{-1}_{(p-1)\times(p-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z} - \overline{\mathbf{z}})(\mathbf{z} - \overline{\mathbf{z}})^T$$

is the sample covariance of the $(p-1)$ predictor variables. The second term in the right hand side $(\frac{1}{n-1}(\mathbf{z}_i - \overline{\mathbf{z}})^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{z}_i - \overline{\mathbf{z}}))$ is the so-called **Mahalanobis distance** from $\mathbf{z}_i$ to the data center $\overline{\mathbf{z}}$

4.1.2 Properties of the Leverage: $0 < h_i < 1, \quad \sum_i h_i = p$

Recall that the hat matrix is idempotent $\mathbf{H} = \mathbf{H}\mathbf{H}^\top$ and has $\text{tr}(\mathbf{H}) = p$
These imply that

$$\sum_i h_i = p \text{ and } \sum_j H_{ij}^2 = h_i$$

For a given i we can decompose the last sum as follows:

$$\begin{aligned} \sum_j H_{ij}^2 &= H_{ii}^2 + \sum_{i\neq j} H_{ij}^2 = h_i \\ \Rightarrow \sum_{i\neq j} H_{ij}^2 &= h_i(1-h_i) \Rightarrow h_i(1-h_i) > 0 \end{aligned}$$

From this we can conclude the following properties of h_i :

$$0 < h_i < 1, \quad \sum_i h_i = p$$

4.1.3 Fitted Values and Leverage: $Var(\hat{y}_i) = \sigma^2 h_i,\ Var(r_i) = \sigma^2(1-h_i)$

Recall the equation $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.

$$\begin{aligned}\hat{y}_i &= \sum_{t=1}^{n} H_{it} y_t \\ &= h_i y_i + \sum_{t \neq i}^{n} H_{it} y_t\end{aligned}$$

This means that $h_i = \frac{d\hat{y}_i}{dy_i}$
When h_i is **large (close to 1)**, $\hat{y}_i$ relies heavily on y_i (instead of using the information from other data points), therefore $\hat{y}_i$ will be "forced"to be **close** to the observed y_i.
Consequently, the variance for the residual r_i will be small, and the variance for the fit $\hat{y}_i$ will be large (since the fit from another data set would be quite different).

$$Var(\hat{y}_i) = \sigma^2 h_i,\ Var(r_i) = \sigma^2(1-h_i)$$

4.1.4 High-leverage Points: $h_i \geq \frac{2p}{n}$

Good high-leverage points: its y point follows the pattern of the rest of the data, but with an x_i value that is far away from the sample mean.
Bad high-leverage points: its y value does not follow the pattern suggested by the rest of the data; the LS fitting might change a lot if we remove this point.

4.2 Residuals: Standardized Residuals vs. Studentized residuals

The residuals $r_i = y_i - \hat{y}_i$ do not have a constant variance. So they need to be standardized.

4.2.1 Difference between ε and r

ε : true residuals (our theoretical quantities)
r : estimated residuals - Both residuals are normally distributed, but:

$$\varepsilon \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right), \quad \mathbf{r} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2\left(\mathbf{I}_n - \mathbf{H}\right)\right)$$

where **H** is the projection/hat matrix.
- The errors ε_i 's have equal variance and are independent, while the residuals r_i 's have unequal variance and are correlated.
$-\mathbb{E}(\varepsilon) = \mathbb{E}(\mathbf{r}) = \mathbf{0}$. But

$$\sum_i \varepsilon_i \neq 0, \quad \sum_i r_i = 0$$

(by default we assume an intercept is included in the model)

4.2.2 Standardized Residuals: $r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}$

Since $r_i \sim \mathcal{N}\left(0, \sigma^2\left(1-h_i\right)\right)$, it is reasonable to consider a standardization of the residuals in this form:

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}, \quad i = 1, \ldots, n$$

- $\sum_i r_i^*$ is no longer zero.
- Since the r_i is not independent of $\hat{\sigma}$, each r_i^* is **not distributed as a T distribution**.
- As an approximation, we can view the r_i^* 's as iid $\mathcal{N}(0,1)$ random variables, although they are not Normally distributed and they are slightly correlated.

4.2.3 Studentized Residuals: $t_i = r_i^* \left(\frac{n-p-1}{n-p-r_i^{*2}}\right)^{1/2}$

- The studentized residuals are based on the idea of leave-one-out (also know as jackknife residuals).
- Here is the leave-one-out idea:

1. Run a regression model on the $(n-1)$ samples with the i-th sample (x_i, y_i) removed.
2. Denote the leave-one-out estimates of the regression coefficient and error variance by $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$, where the notation (i) means "excluding the i-th observation."
3. Then, check the discrepancy between observations y_i and the fitted value $\hat{y}_{(i)} = \mathbf{x}^T \hat{\beta}_{(i)}$

- Define the Studentized Residuals as:

$$t_i = \frac{y_i - \hat{y}(i)}{\hat{\sigma}_{(i)}\left(1 + x_i^T\left(\mathbf{X}_{(\mathbf{i})}^\top \mathbf{X}_{(i)}\right)^{-1} x_i\right)^{1/2}} = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

which follows a T_{n-p-1} distribution if $y_i \sim \mathcal{N}\left(\mathbf{x}_i^T\beta, \sigma^2\right)$

- One can also show that r_i^* and t_i are a monotone transformation of each other.
- We do not need to run the model n times to get the estimates $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$ since it can be shown that:

$$t_i = r_i^*\left(\frac{n-p-1}{n-p-r_i^{*2}}\right)^{1/2}$$

4.3 Outlier (Large $|r_i^*|$)

4.3.1 Outlier test

Outliers are observations that do not fit the model, but Outliers are not necessarily observations with large residuals.

We need to used the studentized residuals for the outlier test.

Under the Null hypothesis H_0,

$$t_i \sim T_{n-p-1}$$

We can use t-test to test the i^{th} observation: form a PI at x_i.

(an example of data snooping)

Large $|r_i^*| \Rightarrow$ Large $|t_i| \Rightarrow$ Reject Null hypothesis $\Rightarrow$ Outlier

4.3.2 Bonferroni Correction

Suppose we are testing m hypothesis sinultaneously.
For each test, we use a significant level α. That is, the chance to make a **overall** Type I error is α.
Suppose we want to control the overall type I error rate (for all m tests) to be 95%.
We should set the individual significance levels to be $\alpha = 5\%/m$
当我们检验 outliers 时，由于 T 分布是双侧，我们需要置信度：$\alpha/(2 * n)$

4.3.3 What we should do with outliers?

Points should not be routinely deleted simply because they do not fit the model. No data snooping! Outliers, as well as other unusual observations discussed here, often flag potential problems of the current model. **Instead of dropping them, maybe, try a new alternative model.**

4.4 Highly Influential Points: Large $D_i = \frac{r_i^{*2}}{p}(\frac{h_i}{1-h_i})$ $(D_i \geq 1)$

4.4.1 Influential observations

Observations whose removal greatly affects the regression analysis are called **influential observations**.
An **influential observations** may be (or may not) an **outlier** or a **high-leverage observation**; or may be both: an outlier and a high-leverage observation.
We will use the **Cook's distance** to detect influential observations.

$$D_i = \frac{\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_{(i)}\right\|^2}{p\hat{\sigma}^2} = \frac{\left\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\right\|^2}{p\hat{\sigma}^2} = \frac{r_i^{*2}}{p}\left(\frac{h_i}{1-h_i}\right)$$

which indicates that highly influential points are either outliers (large $|r_i^*|$) or high-leverage points (large h_i) or both.
A rule-of-thumb: observations with $D_i \geq 1$ are highly influential.

5 Diagnostics: Checking Assumptions

5.1 Classical Linear Model (CLM) assumption: Gauss-Markov Assimption + $\varepsilon \sim^{IID} \mathcal{N}(0, \sigma^2)$

1. Constant Variance (Homoscedasticity)
2. Normality
3. Uncorrelated errors (No-Autocorrelation)
4. Linearity: $\mathbb{E}(y) = \mathbf{X}\beta$
5. Random Sampling

(1-5 calls Gauss-Markov Assimption)

6. $\mathbf{Y} = \beta\mathbf{X} + \varepsilon,$ where $\varepsilon \sim^{IID} \mathcal{N}(0, \sigma^2)$

(1-6 calls Classical Linear Model (CLM) assumption)

5.2 Check Constancy of Variance(Homoscedasticity)

5.2.1 Method 1: graph *residuals* against *Fitted Values* $\hat{y}$

SLR:

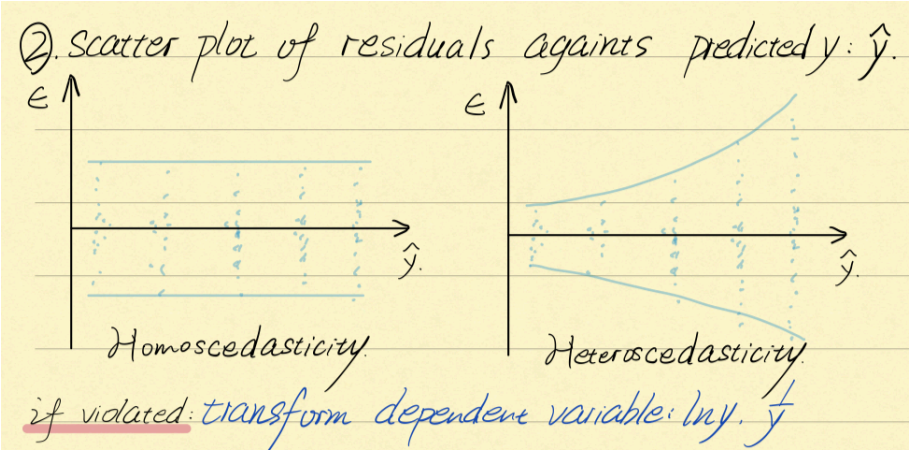

图 5:

MLR:

If the variance is constant, the residuals will look like a football-shaped cloud. Check residual plots and look for a "fan"type shape or trends.

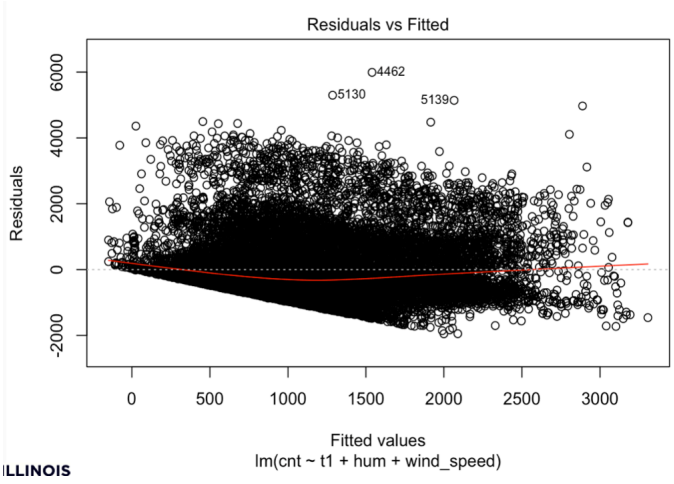

图 6:

5.2.2 Method 2: Breusch-Pagan Test

$$\text{BP} = nR^2$$

where R^2 is the *coefficient of Determination* between the **squared residuals** r_i^2 of LS regression and the **covariates** (or a sub-set) $X_1, X_2, ..., X_p$.

$$H_0 : \text{BP} \sim \chi^2_{p-1} \text{ (asymptotically)}$$

5.2.3 What happen if Heteroscedasticity

Unchanges:

1. The estimaor will still be *unbiased* ($\mathbb{E}(\hat{\beta}) = \beta$), *consistent* ($\lim_{n\to\infty} \hat{\beta} = \beta$), but *inefficient*.
2. Interpretation of R^2 is not changed.

Changes:

1. $\hat{\beta}$ will be *inefficient*.
2. invalidates variance formulas for OLS estimators ($Var(\hat{\beta}_i)$).
3. usual $F-$test, $t-$test are not valid.
4. *Gauss-Markov Theorem* doesn't hold: OLS is no longer the BLUE (best linear unbiased estimator).

5.2.4 What can we do if Heteroscedasticity

(i) Take logarithms of each of the variables
(ii) Use suitably modified standard errors
(iii) Use a generalised least squares (GLS) procedure

5.2.5 Remedial measure: Variance Stabilizing Transformations $\sqrt{Y}$, $\log Y$, $\frac{1}{Y}$ or $\frac{1}{Y+1}$

SLR:

If violates Homoscedasticity: Transform *dependent variable*: $\ln y, \frac{1}{y}$

MLR:

Find a transformation of the response, $h(Y)$, to achieve constant variance.

How does it work?

- Suppose h is a smooth function.
- Using Taylor's theorem, the expansion of $h(Y)$ around $\text{E}(Y)$ is:

$$h(Y) = h(\mathbf{E}(Y)) + h'(\mathbf{E}(Y))(Y - \mathbf{E}(Y)) + \text{ Remainder}$$

- The remainder is assumed small with high probability and we can ignore it:

$$\text{Var(h(Y))} \approx \left(\text{h}'(\text{E(Y)})\right)^2 \text{Var(Y)}$$

- We want to choose a transformation h such that Var(h(Y)) is approximately constant.

Example 1:

- Suppose that the variance of Y is proportional to the mean of Y, i.e., $\text{Var}(Y) \propto E(Y)$
- Select h such that:

$$h'(z) = \frac{1}{\sqrt{z}} \Rightarrow h(z) \propto \sqrt{z}$$

- When plugging-in the value of $h'(z)$ evaluated at $E(Y)$ in the variance of $h(Y)$ equation, the variance of $h(Y)$ will be approximately constant. Indeed,

$$\text{Var}(\sqrt{Y}) \approx \left(\frac{1}{\sqrt{E(Y)}}\right)^2 \text{Var}(Y) = \frac{\text{Var}(Y)}{E(Y)} \approx \text{ const}$$

Example 2:

- Suppose that the variance of Y is proportional to the squared mean of Y, i.e., $\text{Var}(Y) \propto (E(Y))^2$.
- Select h such that:

$$h'(z) = \frac{1}{z} \Rightarrow h(z) = \log(z)$$

- Then,

$$\text{Var}(\log Y) \approx \frac{1}{(\text{E(Y)})^2} \text{Var(Y)} \approx \text{ const}$$

Example 3:

$\text{Var}(Y) \propto (E(Y))^4$.

$$h(Y) = \frac{1}{Y} \text{ or } \frac{1}{Y+1}$$

5.3 Check Normality

5.3.1 Method 1: Histogram, graph *residuals* against its frequency

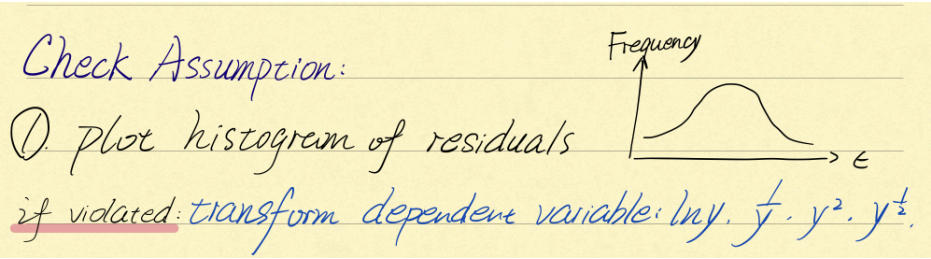

图 7:

If violates Normality: Transform *dependent variable*: $\ln y, \frac{1}{y}, y^2, \sqrt{y}$

5.3.2 Method 2: QQ-Plot, graph *residuals* against its frequency

- Suppose that we have a sample $z_1, z_2, \ldots, z_n$.
- We wish to examine the hypothesis that the z's are a sample from a normal distribution with mean μ and variance σ^2.

QQ-Plot:

1. Order the z's: $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$.
2. Compute $u_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$, where Φ is the cdf of the $N(0,1)$ and i is the order if the data $(i = 1, 2, \ldots, n)$.
3. Plot $z_{(i)}$ against u_i.

$\Rightarrow$ If the z's are normal, the plot should be approximately a straight line.

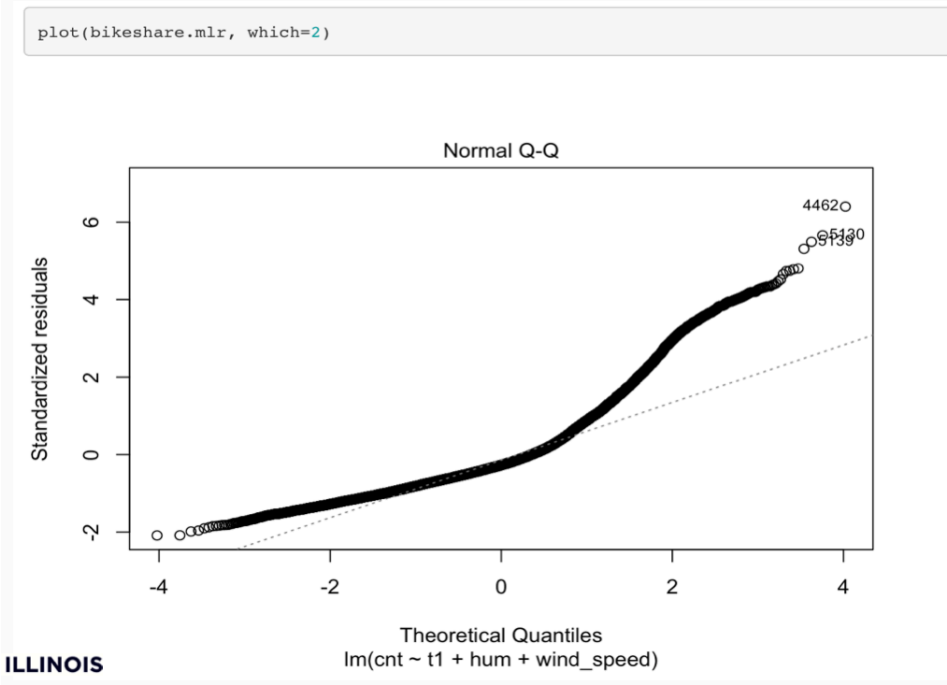

图 8:

5.3.3 Method 3: Shapiro-Wilk Test (good for $n \leq 50$)

$$W = \frac{(\sum_{i=1}^{n} a_i r_{(i)})^2}{\sum_{i=1}^{n}(r_i - \bar{r})^2}$$

where $r_{(i)}$ is the ith largest value of the r_i's and the a_i terms are calculated using the means, variances, and covariances of the r_is.

Small values of W will lead to rejection of the null hypothesis.

5.3.4 Method 4: Kolmogorov-Smirnov Test (good for $n > 50$)

$$D_n = \sup_x |F_n(x) - \Phi(x)|$$

where $\Phi(x)$ is the cdf of the Normal and F_n the empirical distribution function F_n for n i.i.d. ordered observations X_i is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[-\infty,x]}(X_i)$$

Small values of D will lead to rejection of the null hypothesis.

5.3.5 Remedial measure: Box-Cox Transformations of Y

Suppose each $y_i > 0$, and consider the following transformation:

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

Choose λ that maximizes the likelihood of the data, under the assumption that the transformed data $g_\lambda(\mathbf{y})$ has a normal distribution:

$$g_\lambda(\mathbf{y}) = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{l}\right)$$

- The log-likelihood function for $\lambda \neq 0$ is:

$$L(\lambda) = -\frac{n}{2}\log\left(RSS_\lambda/n\right) + (\lambda - 1)\sum_{i=1}^{n}\log\left(y_i\right)$$

where RSS_λ is the RSS when $g_\lambda(\mathbf{y})$ is the response, and for $\lambda = 0$ is:

$$L(0) = -\frac{n}{2}\log\left(RSS_0/n\right) - \sum_{i=1}^{n}\log\left(y_i\right)$$

The second term in these log-likelihood function corresponds to the Jacobian of the transformation.

In **R**, we can graph the log-likelihood as a function of $\lambda(L(\lambda))$ versus $\lambda \in (-2, 2)$ and then pick the maximizer $\hat{\lambda}$.

It is common to round $\hat{\lambda}$ to a nearby value like:

$$-1, -0.5, 0, 0.5, \text{ or } 1$$

then the transformation defined by $\hat{\lambda}$ is easier to interpret.
To answer the question whether we really need the transformation g_λ, we can do hypothesis testing $(H_0 : \lambda = 1)$, or equivalently construct a Confidence Interval for λ as follows:

$$\{\lambda : L(\lambda) > L(\hat{\lambda}) - \frac{1}{2}\chi_1^2(1-\alpha)\}$$

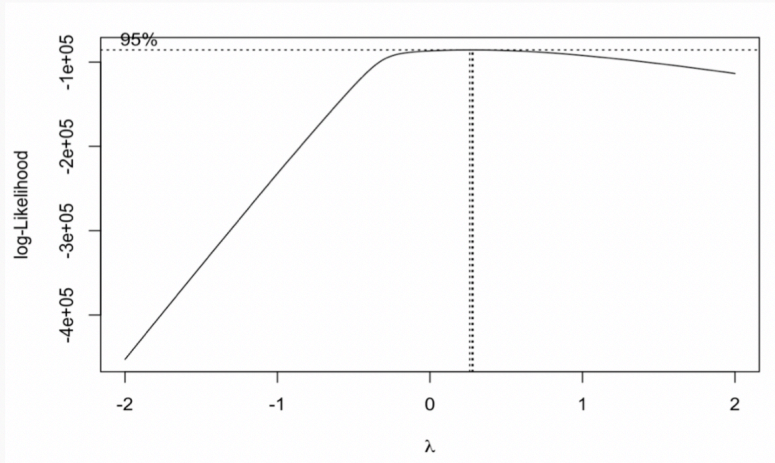

图 9:

5.4 Checking Serial Dependence

5.4.1 Method 1: graph *residuals* against index variable(time or case number)

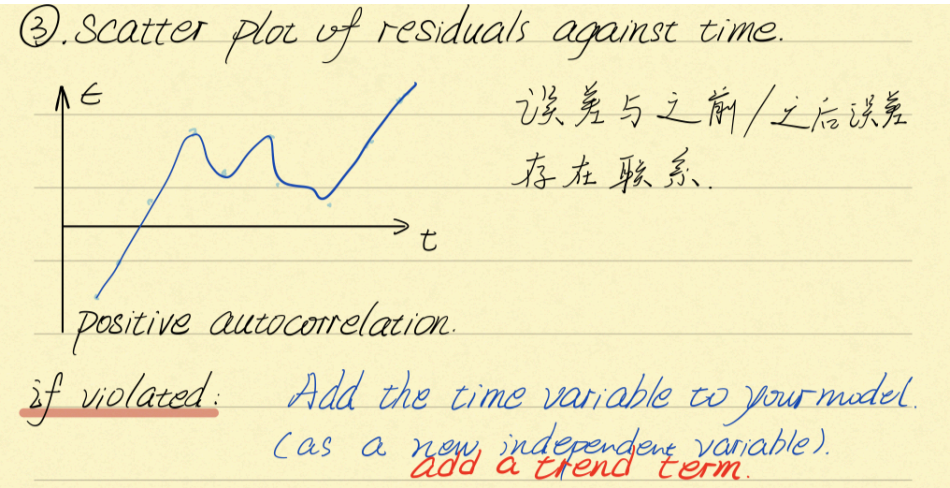

图 10:

If violates No-Autocorrelation: add a new independent vaiable (t).

5.4.2 Method 2: Durbin Watson test

SLR:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Compare d with d_L and d_U.

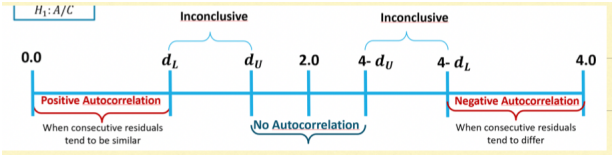

图 11:

MLR:

$$DW = \frac{\sum_{k=1}^{n-1}(r_k - r_{k+1})^2}{\sum_{k=1}^{n} r_k^2}$$

if $DW < 2$, then there is evidence for positive serial dependence.

5.5 Checking Non-Linearity

5.5.1 Method 1: Partial Regression Plots

We want to know **the relationship between the response** Y **and a predictor** X_k after the effect of the other predictors has been removed.
To remove the effect of the other predictors, run the following two regression models:

$$Y \sim X_1 + \ldots + X_{i-1} + X_{i+1} + \ldots \quad (1)$$
$$X_i \sim X_1 + \ldots + X_{i-1} + X_{i+1} + \ldots \quad (2)$$

Get the following residuals:

$$\mathbf{r}_y = \text{ residuals from (1)}$$
$$\mathbf{r}_k^X = \text{ residuals from (2)}$$

Plot $\mathbf{r}_y$ vs. $\mathbf{r}_k^X$: For a valid model, the added-variable plot should produce points randomly scattered around a line through the origin with slope $\hat{\beta}_k$. This is also a useful plot to detect *high influential* data points.

5.5.2 Remedial measure: Linearizing Transformations

1. $\log(Y)$ vs. $\log(X)$, suitable when $\mathbb{E}(Y) = \alpha X_1^{\beta_1} ... X_p^{\beta_p}$
2. $\log(Y)$ vs. X, suitable when $\mathbb{E}(Y) = \alpha \exp \sum_j X_j \beta_j$
3. $\frac{1}{Y}$ vs. X, suitable when $\mathbb{E}(Y) = \frac{1}{\alpha + \sum_j X_j \beta_j}$

5.5.3 Remedial measure: Box-Cox Transformations of Y also works

6 Diagnostics: Collinearity

6.1 Exact Collinearity/ linearly dependent

There exists a set of constants $c_1, c_2, ..., c_p$ (at least one of them is non-zero) s.t.

$$\sum_{j=1}^{p} c_j \mathbf{X}_{.j} = 0$$

then the columns of X are called *linearly dependent* and there is *exact collinearity.*

6.2 What happens if exact collinearity

1. $(\mathbf{X}^T\mathbf{X})^{-1}$ does not exists.
2. The LS estimate $\hat{\beta}$ is not unique.
3. The corresponding linear model is not identifiable.

6.3 Approximate Collinearity

We generally do not need to worry about exact collinearity (**R** can detect it and fix it automatically), but *approximate collinearity.*

$$\sum_{j=1}^{p} c_j \mathbf{X}_{.j} \approx 0$$

$$\mathbf{X}_{.k} \approx -\sum_{j \neq k} c_j \mathbf{X}_{.j}/c_k$$

A simple diagnostic for this is to obtain the regression of $\mathbf{X}_{.k}$ on the remaining predictors, and if the corresponding R_k^2 is close to 1, we would diagnose approximate collinearity.

$$\mathbf{X}_{.k} \sim \mathbf{X}_{.1} + \ldots + \mathbf{X}_{.k-1} + \mathbf{X}_{.k+1} + \ldots \Rightarrow R_k^2$$

6.4 What happens if approximate collinearity: based on $\left(\frac{1}{1-R_k^2}\right)$ ($k-$th variance inflation factor (VIF), $VIF > 10$)

In a multiple regression $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e$, the LS estimate $\hat{\beta}_k$ is unbiased with variance:

$$\mathrm{Var}\left(\hat{\beta}_{\mathrm{k}}\right) = \sigma^2 \left(\frac{1}{1-\mathrm{R}_{\mathrm{k}}^2}\right) \left(\frac{1}{\sum_{\mathrm{i}=1}^{n} \left(\mathrm{x}_{\mathrm{ik}} - \overline{\mathrm{x}}_{\cdot \mathrm{k}}\right)^2}\right)$$

where R_k^2 is the R-square from the regression of $\mathbf{X}_{.k}$ on the remaining predictors. When R_k^2 is close to 1 , the variance of $\hat{\beta}_k$ is large.

Consequently we will have:

1. large Mean Square Error

2. large (inflated) p-value to the corresponding t-test, i.e, we could miss a significant predictor.
The quantity $\left(\frac{1}{1-R_k^2}\right)$ is called the k-th variance inflation factor (VIF).
$VIF > 10$ infers conllinearity

6.5 Possible symptoms of collinearity

1. high pair-wise (sample) correlation between predictors
2. high VIF
3. high condition number
4. R^2 is relatively large but none of the predictor is significant.

6.6 Global Measure of Collinearity: *condition number* of $\mathbf{X}^T\mathbf{X}$

Condition number of $\mathbf{X}^T\mathbf{X}$:

$$\kappa = (\text{largest eigenvalue/smallest eigenvalue})^{1/2}$$

An empirical rule for declaring collinearity is $\kappa \geq 30$
Note that κ is *not scale-invariant*, so we should *standardized* each column of X (i.e. each column should have *zero mean* and *sample variance* equal to 1, before calculating the condition number).

6.7 What to do with collinearity

1. Remove some predictors from highly correlated groups of predictors.
2. Regularize the model using penalized Least Squares estimation.