Linear Regression Analysis Notes

J S

2024.03.13

Preface

This is my personal study notes of Linear Regression Analysis course, instructed by Zaiying Zhou, at Tsinghua University, during the spring semester of 2024. The notes is formulated throughout the semester, so it's also a personal study profile. The textbook is *Applied Linear Statistical Models*[a], although the slides are used as the main resource. The last update is on 2024.7.13.

Initially the notes are prepared as the "cheat sheet" for the partly open-book final exam, but I find it helpful to type the formulae by hand and devise the interpretations on my own understanding. On the whole, the notes focus on applications and necessary intuition rather than rigorous theoretical proof. A few mathematical proofs of interest are attached in the appendix, and the rest are either straightforward or accessible on the Internet. Apart from my limited contributions, a large proportion of explanatory texts are copied from the course slides. Moreover, this work is greatly inspired by v1ncent19's notes and vicayang's notes. Sincere thanks!

[a] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li, 5th Edition

Contents

Notations

Notation	Explanation
X_i	The value of the i-th predictor variable
Y_i	The value of the i-th response variable
β_0	The intercept of the regression line (Simple Linear Regression)
β_1	The slope of the regression line (Simple Linear Regression)
ε_i	Uncorrelated random error term.
$b_0/\hat{\beta}_0$	Point estimator of β_0
$b_1/\hat{\beta}_1$	Point estimator of β_1
$\hat{Y}_i$	Estimated value of the i-th response variable
e_i	Residual, $Y_i - \hat{Y}_i$
S_{XX}	$\sum_{i=1}^{n}(X_i - \bar{X})^2$
$s^2\{\cdot\}$	The sample variance of $\cdot$
s^2/MSE	Estimated variance of random error term
t_n	A t distribution with n degrees of freedom
F_{n_1,n_2}	An F distribution with n_1 and n_2 degrees of freedom
$t_{n;1-\alpha}$	The $1-\alpha$ percentile of a t distribution with n degrees of freedom
$F_{n_1,n_2;1-\alpha}$	The $1-\alpha$ percentile of an F distribution with n_1 and n_2 degrees of freedom
$\mathbf{H}$	The hat/projection matrix
$\underset{a\times b}{\mathbf{X}}$	A matrix $\mathbf{X}$ with dimension $a \times b$
OLS	Ordinary Least Squares method
WLS	Weighted Least Squares method

1 Simple Linear Regression

1.1 Model

- Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

  Parameters: β_0, β_1, σ^2, 3 parameters in total.

- Solution[1]:

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$
$$b_0 = \hat{\beta}_0 = \bar{Y} - b_1 \bar{X}$$

- Assumptions: LINE - Linear Function, Independent, Normally distributed, Equal Variance.
  - **L**inearity: $\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_i$
  - **I**ndependence: $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are independent.
  - **N**ormality: $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n \sim \mathcal{N}(0, \sigma^2)$
  - **E**qual Variance: $\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$

- Properties:
  - The regression line always goes through $(\bar{X}, \bar{Y})$.
  - $\sum_{i=1}^n e_i = \sum_{i=1}^n X_i e_i = \sum_{i=1}^n \hat{Y}_i e_i = 0$

1.2 Parameter Estimation

1.2.1 Estimation of σ^2

We denote the unbiased estimator of σ^2 as s^2, then[2]

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2}{n-2}$$

1.2.2 Estimation of β_1

$$b_1 = \beta_1 + \sum_{i=1}^n \frac{X_i - \bar{X}}{S_{XX}} \varepsilon_i$$
$$b_1 \sim \mathcal{N}(\beta_1, \sigma^2 / S_{XX})$$

To infer b_1,

$$s^2\{b_1\} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} / \sum_{i=1}^n (X_i - \bar{X}_i)^2 = \frac{\mathrm{MSE}}{S_{XX}}$$
$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}$$

[1] Proof

[2] Proof of unbiasedness

1.2.3 Estimation of β_0

$$b_0 = \beta_0 + \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{S_{XX}} \right) \varepsilon_i$$

$$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2 (\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}) \right)$$

To infer b_0,

$$s^2\{b_0\} = \text{MSE}(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}})$$

$$\frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{n-2}$$

1.2.4 Estimation of $\hat{Y}_h$

For a new X_h, we wish to predict the corresponding Y_h. We denote the mean of Y_h as $\hat{\mu}_h$.

$$\hat{\mu}_h = b_0 + b_1 X_h = \beta_0 + \beta_1 X_h + \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(X_i - \bar{X})(X_h - \bar{X})}{S_{XX}} \right) \varepsilon_i$$

$$\hat{\mu}_h \sim \mathcal{N}\left(\beta_0 + \beta_1 X_h, \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right) \right)$$

To infer $\hat{\mu}_h$,

$$s^2\{\hat{\mu}_h\} = s^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right)$$

$$\frac{\hat{\mu}_h - \beta_0 - \beta_1 X_h}{s\{\hat{\mu}_h\}} \sim t_{n-2}$$

1.2.5 Estimation of Y_h

$$Y_h = \hat{\mu}_h + d_h$$

$$Y_h \sim \mathcal{N}\left(\beta_0 + \beta_1 X_h, \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right) \right)$$

To infer Y_h,

$$s^2\{d_h\} = s^2 \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right)$$

$$\frac{Y_h - \hat{\mu}_h}{s\{d_h\}} \sim t_{n-2}$$

For m new observations,

$$s^2\{\text{predmean}\} = s^2 \left(\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}} \right)$$

1.2.6 Confidence Band for Entire Regression Line

We want to get a confidence band: $\{(x, y) : L(x) < y < U(x), x \in R\}$.

For a fixed x, the $100\%(1-\alpha)$ confidence interval is

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

Let $W = \max\left(\frac{\mu_h - Y_h}{s\{\hat{\mu}_h\}} \right)$, then the confidence band is

$$\left(\hat{\mu}_h - W \cdot s\{\hat{\mu}_h\}, \hat{\mu}_h + W \cdot s\{\hat{\mu}_h\} \right)$$

And it can be calculated that $W^2 = 2F_{2,n-2;1-\alpha}$.

So the confidence band for entire regression line is:

$$\Big(L(x), U(x)\Big) = \hat{Y}_h \pm W \cdot s(\hat{Y}_h)$$

where $W^2 = 2F_{2,n-2;1-\alpha}$.

It can be proved that $W > t_{n-2;1-\frac{\alpha}{2}}$, so the confidence band is wider than confidence interval.

1.3 ANOVA

1.3.1 Concepts

- Total sum of squares(SST):

  $$\text{SST} = \sum (Y_i - \bar{Y})^2$$

  Degree of freedom: $\text{df}_T = n - 1$. Note that $\text{SST} = \text{SSE} + \text{SSR}$.

- Variation due to error(SSE):

  $$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$$

  Degree of freedom: $\text{df}_E = n - 2$.

- Variation due to regression(SSR):

  $$\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum (X_i - \bar{X})^2$$

  Degree of freedom: $\text{df}_R = 1$.

- Mean square of error(MSE):

  $$\text{MSE} = \frac{\text{SSE}}{\text{df}_E} = \frac{\text{SSE}}{n-2}$$
  $$\mathbb{E}(\text{MSE}) = \sigma^2$$

- Mean square of regression(MSR):

  $$\text{MSR} = \frac{\text{SSR}}{\text{df}_R} = \text{SSR}$$
  $$\begin{aligned}\mathbb{E}(\text{MSR}) = \mathbb{E}(\text{SSR}) &= \mathbb{E}(b_1^2) \sum (X_i - \bar{X})^2 \\ &= \big(\mathbb{E}(b_1)^2 + \text{Var}(b_1)\big) \sum (X_i - \bar{X})^2 \\ &= \left(\beta_1^2 + \frac{\sigma^2}{S_{XX}}\right) \sum (X_i - \bar{X})^2 \\ &= \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2\end{aligned}$$

  MSR tends to be larger than MSE.

Table 1: ANOVA Table

Source	df	SS	MS	F-value
Regression	1	$\sum(\hat{Y}_i - \bar{Y}_i)^2$	SSR/df_R	MSR/MSE
Error	$n-2$	$\sum(Y_i - \bar{Y}_i)^2$	SSE/df_E	
Total	$n-1$	$\sum(Y_i - \bar{Y}_i)^2$	SST/df_T	

1.3.2 Descriptive Measures of Linear Association

- Coefficient of Determination:

$$R^2 = \frac{\text{SSR}}{\text{SST}} \in [0, 1]$$

R^2 measures the total variation that can be explained by the model.

- Pearson Correlation Coefficient:

$$\begin{aligned} r &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \\ &= \frac{\sqrt{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}\sqrt{\sum\left(Y_i - \bar{Y}\right)^2}}{\sum(\hat{Y}_i - \bar{Y})^2} \\ &\in [-1, +1] \end{aligned}$$

Pearson correlation coefficient measures the linear relationship between two variables. Under $H_0 : \beta_1 = 0$, $T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim t_{n-2}$.

In simple linear regression, $r^2 = R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$

1.3.3 F-test

Hypotheses:

$$H_0 : \beta_1 = 0 \Longleftrightarrow H_1 : \beta_1 \neq 0$$

Test statistic:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

Distribution:

$$F = \frac{\chi^2_{\text{df}_R}/\text{df}_R}{\chi^2_{\text{df}_E}/\text{df}_E} \overset{H_0}{\sim} \text{F}_{\text{df}_R, \text{df}_E} = \text{F}_{1, n-2}$$

Decision rule: If $F_0 > F_{1-\alpha,1,n-2}$, reject H_0; otherwise, accept H_0.

Note: Under $H_1 : \beta_1 \neq 0$, F follows a noncentral F-distribution.

1.3.4 General Linear Test

Hypotheses:

$$H_0 : \beta_1 = 0 \Longleftrightarrow H_1 : \beta_1 \neq 0$$

$$\textbf{R}\text{educed Model: } H_0 : \ Y_i = \beta_0 + \varepsilon_i$$

$$\textbf{F}\text{ull Model: } H_1 : \ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Test statistic:

$$F = \frac{\text{SSE}(R) - \text{SSE}(F)/(\text{df}_E(R) - \text{df}_E(F))}{\text{SSE}(F)/\text{df}_E(F)} \sim \text{F}_{\text{df}_E(R) - \text{df}_E(F), \text{df}_E(F)}$$

For the full model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$,

$$\text{SSE}(F) = \text{SSE}, \ \ \text{df}_E(F) = n - 2$$

For the reduced model $Y_i = \beta_0 + \varepsilon_i$,

$$\text{SSE}(R) = \text{SST}, \ \ \text{df}_E(R) = n - 1$$

GLT approach is a more general approach to test linear association.

1.4 Gauss-Markov Theorem

Suppose we have the linear relationship

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i \ \ (i = 1, 2, \cdots, n)$$

where β_0, β_1 are non-random but unobservable parameters, X_i are non-random observations, and ε_i are random noises.

Gauss-Markov assumptions:

- Zero mean: $\mathbb{E}(\varepsilon_i) = 0$
- Homoscedasticity: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- Uncorrelatedness: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \ (i \neq j)$

The theorem states that the OLS estimator b_0 and b_1 (or, in matrix notation, $\mathbf{b} = \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$) is a best linear unbiased estimator(BLUE).

- Best: $\mathbf{b}$ has the minumum variance among all linear estimators.
- Linear: $\mathbf{b}_j = c_{1j}y_1 + c_{2j}y_2 + \cdots + c_{kj}y_k, \forall j \in \{0, 1, 2, \cdots, p-1\}$
- Unbiased: $\mathbb{E}(\mathbf{b}) = \mathbf{b}$

1.5 Comments

- In this course, we treat model parameters as unknown constants.
- Y_i is a random variable, while X_i is not a random variable. For a random variable, there is a probability distribution, e.g. $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$. In contrast, X_i only has fixed value.
- The essence of regression is conditional expectation.
- Compared with time series model, regression models concern interpolation, while time series models concern extrapolation.
- Inference concerning β_0 is usually not of interest, because shift in X cause shift in β_0, and $X = 0$ may not be in the scope.
- If errors are not normal but are relatively symmetric, the tests and confidence intervals are reasonable approximations, and get improved when sample size increases (due to the central limit theorem).

2 Multiple Linear Regression

2.1 Model

- Scaler expression:

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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

- Matrix expression:

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

$$\underset{n\times 1}{\mathbf{Y}} = (Y_1, Y_2, \cdots, Y_n)^T$$

$$\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = (\beta_0, \beta_1, \cdots, \beta_{p-1})^T$$

$$\underset{n\times 1}{\boldsymbol{\varepsilon}} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$$

- Model assumptions:

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

2.2 Parameter Estimation

2.2.1 Estimation of $\boldsymbol{\beta}$

- OLS estimation:

$$Loss(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

  By minimizing $Loss(\boldsymbol{\beta})$, we have:

$$\underset{p\times 1}{\mathbf{b}} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$$

- Sampling distribution of $\boldsymbol{\beta}$:

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

$$\frac{b_i - \beta_i}{s\{b_i\}} \sim t_{n-p}$$

- Estimated covariance matrix of $\mathbf{b}$:

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

  where

$$s^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-p} = \frac{1}{n-p}\mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

2.2.2 Estimation of σ^2

-

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$$

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

- Random errors are i.i.d normally distributed:

$$\underset{n\times n}{\mathrm{Cov}\,(\boldsymbol{\varepsilon})} = \mathrm{Cov}\left([\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n]^T\right) = \underset{n\times n}{\sigma^2\mathbf{I}}$$

  Note that residuals are not i.i.d distributed $\left(\mathrm{Cov}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}) \neq \sigma^2\mathbf{I}\right)$. That's because we use regression method to fit the data, which influences the scale of residuals.

- Properties of residual:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$
$$\mathbf{e}^T\hat{\mathbf{Y}} = \mathbf{e}^T\mathbf{X} = \mathbf{0} \quad \mathbf{e}^T\mathbf{e} = \boldsymbol{\varepsilon}^T(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$$
$$\mathbb{E}(\mathbf{e}^T\mathbf{e}) = \mathbb{E}(\text{SSE}) = (n-p)\sigma^2$$
$$s^2 = \text{MSE} = \frac{\mathbf{e}^T\mathbf{e}}{n-p} = \frac{1}{n-p}\mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

- $\mathbf{b}$ and s^2 are independent.
- Sampling distribution of SSE:

$$\frac{\mathbf{e}^T\mathbf{e}}{\sigma^2} = \frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$$

2.2.3 Estimation of $\overrightarrow{Y}$

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{X}\mathbf{b}$$
$$\hat{\mathbf{Y}} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H})$$

2.2.4 Estimation of $\hat{Y}_h$

$$\begin{aligned}\hat{Y}_h &= \mathbf{X}_h^T\mathbf{b} \\ &= \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}_h^T\boldsymbol{\beta} + \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\end{aligned}$$
$$\hat{Y}_h \sim \mathcal{N}(\mathbf{X}_h^T\boldsymbol{\beta}, \sigma^2\mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h)$$

2.2.5 Estimation of Y_h

$$Y_h \sim \mathcal{N}\Big(\mathbf{X}_h^T\boldsymbol{\beta}, \sigma^2\left(1 + \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h\right)\Big)$$

2.2.6 Hat matrix H

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

Properties:

$\mathbf{H}^2 = \mathbf{H}$	$\mathbf{H}\mathbf{X} = \mathbf{X}$
$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} = \mathbf{H}\hat{\mathbf{Y}}$	$\mathbf{H}\mathbf{e} = \mathbf{0}$
$\text{Cov}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$	$\mathbf{e}^T\mathbf{e} = \boldsymbol{\varepsilon}^T(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$
$\text{Var}(e_i) = \sigma^2(1 - h_{ii})$	$\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$
$\text{rank}(\mathbf{I} - \mathbf{H}) = \text{tr}(\mathbf{I} - \mathbf{H}) = n - p$	$\text{rank}(\mathbf{H}) = p$
$h_{ii} \in [0, 1]$	

2.3 ANOVA

2.3.1 Concepts

- Total sum of squares(SST):

$$\text{SST} = \sum(Y_i - \bar{Y})^2 = (\mathbf{Y} - \bar{Y}\mathbf{1}_n)^T(\mathbf{Y} - \bar{Y}\mathbf{1}_n) = \mathbf{Y}^T(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{Y}$$
$$\text{SST} = \text{SSE} + \text{SSR}$$

Degree of freedom: $\text{df}_T = n - 1$.

- Sum of squared error(SSE):

$$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{e}^T \mathbf{e} = \mathbf{Y}^T (\mathbf{I} - \mathbf{H}) \mathbf{Y}$$

  Degree of freedom: $\text{df}_E = n - p$.

- Mean squared error(MSE):

$$\text{MSE} = s^2 = \frac{\mathbf{e}^T \mathbf{e}}{n - p} = \frac{1}{n - p} \mathbf{Y}^T (\mathbf{I} - \mathbf{H}) \mathbf{Y}$$

$$\mathbb{E}(\text{MSE}) = \sigma^2$$

- Sum of squares due to model(SSR):

$$\text{SSR} = \sum (Y_i - \hat{Y}_i)^2 = (\hat{\mathbf{Y}} - \bar{Y} \mathbf{1}_n)^T (\hat{\mathbf{Y}} - \bar{Y} \mathbf{1}_n) = \mathbf{Y}^T (\mathbf{H} - \frac{1}{n} \mathbf{1}\mathbf{1}^T) \mathbf{Y}$$

  Degree of freedom: $\text{df}_R = p - 1$.

- Mean of squares due to model(MSR):

$$\text{MSR} = \frac{\text{SSR}}{\text{df}_R}$$

$$\mathbb{E}(\text{MSR}) = \sigma^2 + \frac{1}{p - 1} (\mathbf{X}\boldsymbol{\beta})^T \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) \mathbf{X}\boldsymbol{\beta}$$

Table 2: ANOVA Table

Source	df	SS	MS	F-value
Rgression	$p - 1$	$\sum (\hat{Y}_i - \bar{Y}_i)^2$	SSR/df_R	MSR/MSE
Error	$n - p$	$\sum (Y_i - \bar{Y}_i)^2$	SSE/df_E	
Total	$n - 1$	$\sum (Y_i - \bar{Y}_i)^2$	SST/df_T	

2.3.2 F-test

Hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = 0 \iff H_1 : \beta_k \neq 0 \text{ for at least one k in } 1, 2, \cdots, p - 1$$

Test statistic:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

Distribution:

$$F = \frac{\chi^2_{\text{df}_R} / \text{df}_R}{\chi^2_{\text{df}_E} / \text{df}_E} \overset{H_0}{\sim} \text{F}_{\text{df}_R, \text{df}_E} = \text{F}_{p-1, n-p}$$

Decision rule: If $F_0 > \text{F}_{p-1, n-p; 1-\alpha}$, reject H_0; otherwise, accept H_0.

2.3.3 General F-test

Hypotheses:

$$H_0 : \underset{q \times p}{\mathbf{C}} \boldsymbol{\beta} - \underset{q \times 1}{t} = \mathbf{0}$$

Test statistic:

$$F = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - t)^T (\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - t)}{q s^2}$$

Distribution:

$$F \overset{H_0}{\sim} \text{F}_{q, n-p}$$

Decision rule: If $F_0 > F_{q, n-p; 1-\alpha}$, reject H_0; otherwise, accept H_0.

2.3.4 General Linear Test: (Partial) F-test

Suppose a model with $p-1$ explanatory variables: $X_1, X_2, \cdots, X_{p-1}$.

Full Model: All variables are predictive.
Reduced Model: Some (extra) variables can be removed.

Hypotheses:

$$H_0 : \beta_{j1} = \beta_{j2} = \cdots = \beta_{jk} = 0$$
$$H_1 : \text{At least one of } \beta_{j1}, \beta_{j2}, \cdots, \beta_{jk} \neq 0$$

Test statistic:

$$F^* = \frac{\Big(\text{SSE}(R) - \text{SSE}(F)\Big) / \Big(\text{df}_E(R) - \text{df}_E(F)\Big)}{\text{SSE}(F)/\text{df}_E(F)}$$

Distribution:

$$F^* \overset{H_0}{\sim} F_{\text{df}_E(R)-\text{df}_E(F),\text{df}_E(F)} = \text{F}_{k,n-p}$$

2.3.5 Extra Sum of Squares

An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model.

Examples of extra SS:

- $\text{SSR}(X_1, X_2, X_3)$: The total variation explained by X_1, X_2, X_3 in a model.
- $\text{SSR}(X_1, X_4 | X_2, X_3)$: The additional variation explained by X_1, X_4 when added to a model already containing X_2, X_3.

Table 3: Three types of extra sum of squares

Term	Type 1: Sequential	Type 2: Hierarchical	Type 3: Unique
A	$\text{SSR}(A)$ $= \text{SSE}(1) - \text{SSE}(A)$	$\text{SSR}(A\|B)$ $= \text{SSE}(B) - \text{SSE}(A, B)$	$\text{SSR}(A\|B, AB)$ $= \text{SSE}(B, AB) - \text{SSE}(A, B, AB)$
B	$\text{SSR}(B\|A)$ $= \text{SSE}(A) - \text{SSE}(A, B)$	$\text{SSR}(B\|A)$ $= \text{SSE}(A) - \text{SSE}(A, B)$	$\text{SSR}(B\|A, AB)$ $= \text{SSE}(A, AB) - \text{SSE}(A, B, AB)$
AB	$\text{SSR}(AB\|A, B)$ $= \text{SSE}(A, B) - \text{SSE}(A, B, AB)$		$\text{SSR}(AB\|A, B)$ $= \text{SSE}(A, B) - \text{SSE}(A, B, AB)$

Suppressor variable: If $\text{SSR}(X_2|X_1) > \text{SSR}(X_2)$, then X_1 is called a suppressor variable. The general idea is that a suppressor variable will suppress irrelevant variance of other independent variables.

2.3.6 Partial Determination Coefficients

- Definition:

$$\begin{aligned} R^2_{Yk|1,\cdots,k-1,k+1,\cdots,q} &= \frac{\text{SSR}(X_k|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_q)}{\text{SSE}(X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_q)} \\ &= \frac{\text{SSR}(X_k|X_{-k})}{\text{SSE}(X_{-k})} \\ &= \frac{\text{SSE}(R) - \text{SSE}(F)}{\text{SSE}(R)} \\ &= R^2(Y|X_{-k}, X_k|X_{-k}) \\ &\in [0, 1] \end{aligned}$$

For example, $R^2_{Y1|23}$ represents the percentage of the leftover variation in (after regressing on X_2 and X_3) that is explained by X_1.

$$R^2_{Y1|23} = \frac{\text{SSR}(X_1|X_2, X_3)}{\text{SSE}(X_2, X_3)} = 1 - \frac{\text{SSE}(X_1, X_2, X_3)}{\text{SSE}(X_2, X_3)}$$

- A coefficient of partial determination measures the marginal contribution of one variable when all others are already included in the model. Or, the amount of remaining variation explained by a variable given other variables already in the model.
- If $0 < r_{12\cdot 3} < r_{12}$, then variable 3 partly explains the correlation between 1 and 2.
- Suppose two explanatory variables X_1 and X_2 are uncorrelated but both are correlated with Y. If we take linear regression of $Y \sim X_1$ and $Y \sim X_1 + X_2$, then b_2 are the same in both cases. And $R^2_{Y2|1} > R^2_{Y2}$.
- Procedure to find partial correlation:
  - Predict Y using other X's.
  - Predict X_k using other X's.
  - Find correlation between the two sets of residuals.
- $\beta_1, \beta_2, \cdots, \beta_{p-1}$ are called partial regression coefficients for the explanatory variables.

2.4 Standardized Regression Model

Suppose a linear regression model with $p-1$ explanatory variables:

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

Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $s_Y = \sqrt{\frac{1}{n-1}\sum_i^n (Y_i - \bar{Y})^2}$;
Let $\bar{X_k} = \frac{1}{n}\sum_{i=1}^n X_{ik}$, $s_{X_k} = \sqrt{\frac{1}{n-1}\sum_i^n (X_{ik} - \bar{X_k})^2}$, for $k \in 1, 2, \cdots, p-1$.

Standardization:

$$\begin{aligned}
\frac{1}{\sqrt{n-1}}\frac{Y_i - \bar{Y}}{s_Y} &= \frac{1}{\sqrt{n-1}}\frac{(\beta_0 + \sum_{k=1}^{p-1}\beta_k X_{ik} + \varepsilon_i) - (\beta_0 + \sum_{k=1}^{p-1}\beta_k \bar{X_k} + \bar{\varepsilon})}{s_Y} \\
&= \frac{1}{\sqrt{n-1}}\frac{\sum_{k=1}^{p-1}(X_{ik} - \bar{X_k}) + (\varepsilon_i - \bar{\varepsilon})}{s_Y} \\
&= \sum_{k=1}^{p-1}\frac{\beta_k s_{X_k}}{s_Y}\frac{(X_{ik} - \bar{X_k})}{\sqrt{n-1}s_{X_k}} + \frac{\varepsilon_i - \bar{\varepsilon}}{\sqrt{n-1}s_Y}
\end{aligned}$$

Let

$$Y_i^* = \frac{1}{\sqrt{n-1}}\frac{Y_i - \bar{Y}}{s_Y} \quad X_{ik}^* = \frac{(X_{ik} - \bar{X_k})}{\sqrt{n-1}s_{X_k}} \quad \beta_k^* = \frac{s_{X_k}}{s_Y}\beta_k$$

Then

$$Y_i^* = 0 + \sum_{k=1}^{p-1}\beta_k^* X_{ik}^* + \varepsilon_i^*$$

Properties:

- Let

$$\underset{n\times(p-1)}{\mathbf{X}^*} = \begin{bmatrix} X_{11}^* & \cdots & X_{1,p-1}^* \\ \vdots & \ddots & \vdots \\ X_{n1}^* & \cdots & X_{n,p-1}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^* & |\cdots| & \mathbf{X}_{p-1}^* \end{bmatrix}$$

Then

$$\mathbf{X}^{*T}\mathbf{X}^* = \underbrace{\begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix}}_{\text{Sample correlation matrix}}, \quad \mathbf{X}^{*T}\mathbf{Y}^* = \mathbf{r}_{YX} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \cdots \\ r_{Y,p-1} \end{bmatrix}$$

- $\beta_1^*, \beta_2^*, \cdots, \beta_{p-1}^*$ are scale/unit free, admit same interpretation, and can be directly compared with each other in an intuitive sense.
- Compared with ordinary method, standardized regression generates the same R^2. But ANOVA is different (due to the scale shift of Y).

3 Diagnostics

Diagnostics: Use visualization tools (graphs) and/or formal procedures to check whether the assumptions of the model are violated.

3.1 Diagnostic of X

- Important Statistics: mean, standard deviation, skewness, kurtosis, range.
- Useful plots: Stem-and-leaf plot, box plot, histogram plot, quantile-quantile plot.

3.2 Diagnostic of Residual

- Homoscedasticity test: Bartlett, Levene, Brown-Forsythe, Breusch-Pagan.
- Normality test: Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling.
- Dependence test: Durbin-Watson, Ljung-Box.

Heteroscedasticity of residual: Unbiasedness is retained, but minimum variance can't be guaranteed.

3.3 Multicollinearity

- Zero collinearity: Explanatory variables are not correlated to each other. In other words, their corresponding columns in $\mathbf{X}$ are orthogonal to each other. In this case,

$$b_j = \frac{\mathbf{X}_j\mathbf{Y}}{\|\mathbf{X}_j\|^2}, \quad \mathrm{Var}(b_j) = \frac{\sigma^2}{\|\mathbf{X}_j\|^2}$$

  Properties of Zero Collinearity:

  - Point estimation of b_j does not depend on the other explanatory variables.
  - P-values for testing β_j is affected by other explanatory variables. Reason: p-values depend on MSE, which further depends on which explanatory variables are included.
  - The contribution of explanatory variable X_j to SST is clear-cut, and type I and III (or II) of will be the same.
  - Under the assumption that the linear regression model is true, orthogonal design is optimal.
    - No ambiguity between the explanatory variables.
    - Variances are the smallest possible (with high power).

- Linearly dependent: $\det(\mathbf{X}^T\mathbf{X}) = 0$, so that the linear regression fit $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ can not be calculated. It is possible to mathematically remove the redundancy by deleting one of the explanatory variables, so that the reduced model can be fitted. However, the confounding issue still remains.

- General Case: some correlation exists among explanatory variables.

  When the amount of correlation increases, problems arise:

  - Numerically: $\mathbf{X}^T\mathbf{X}$ is nearly singular and is therefore difficult to invert accurately via computer.
  - Statistically:
    * Ambiguity between explanatory variables increases, making it difficult to interpret regression coefficients.
    * Type I SS and Type II SS become different, leading to inconsistency between ANOVA F-test and t-tests.
    * $\mathrm{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, the variance of b_j increases.

  For prediction, increasing sample size is a viable remedy; for explanation, collinearity is problematic.

- Remedial measures:
  - Collect additional data under different experimental or observational conditions.
  - Principal component analysis.
  - Variable selection: remove some predictors.
  - Variable transformation.

- Case analysis: After omitting an important explanatory variable W, the estimation of σ^2 will increase, and $\mathrm{Var}(\mathbf{b})$ may decrease. Especially, if $W \perp \mathbf{X}$, $\mathbf{b}$ and $\mathrm{Var}(\mathbf{b})$ remain unchanged.

  Higher collinearity increases variance, but does not cause bias.

3.4 Limitations of R^2

- R^2 does not measure goodness of fit.
  - With a large σ^2, R^2 can be tiny even when the model is completely correct.
  - R^2 can still be high with a nonlinear term obviously missed.
  - R^2 can be anywhere between 0 and 1 just by changing the range of X.
- R^2 says nothing about predictive ability.
  - MSE is a better measure of prediction error.
- R^2 does not allow you to compare models using transformed responses.
  - R^2 can easily go down when the model assumptions are better fulfilled.
- R^2 does not measure how one variable explains another.
  - If we regress X on Y, we'd get exactly the same R^2.

3.5 Model Selection

3.5.1 K-fold Cross Validation

1. Randomly partition the original sample $\{\mathbf{X}, \mathbf{Y}\}$ into k equal sized subsamples: $\{\mathbf{X}_1, \mathbf{Y}_1\}, \{\mathbf{X}_2, \mathbf{Y}_2\}, \cdots, \{\mathbf{X}_k, \mathbf{Y}_k\}$
2. Iterate i in $\{1, 2, \cdots, k\}$:
   (a) Take $\{\mathbf{X}_i, \mathbf{Y}_i\}$ as test set and the other $k-1$ subsamples as training set.
   (b) Fit the model on the training set and evaluate it on the test set with certain criterion.
3. Average the k results to produce a single estimation.

3.5.2 Searching Methods

- For small p: exhaustive search.
- For large p: Automatic/greedy search in step/stage-wise fashion.
  1. Use t-test, F-test or p-value as criterion.
  2. Set prescribed values to add/delete a variable.
  3. Use forward/backward stepwise regression.
  4. End with the identification of a single regression model as "best".

  Drawbacks of stepwise regression procedure:
  - The final model may not be optimal.
  - The procedure yields a single final model, although there are often several equally good models
  - The procedure doesn't take domain knowledge into account.
  - Many-tests are conducted. The probability is high that we included some unimportant predictors or excluded some important predictors.

3.5.3 Evaluation Criteria

Suppose the model has $p-1$ variables, and the full model has $P-1$ variables.

- Coefficient of determination: $R^2 = 1 - \frac{\text{SSE}(p)}{\text{SST}}$.
- Adjusted R^2:
$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{\text{SSE}(p)}{\text{SST}} = 1 - \frac{\text{MSE}(p)}{\text{MST}}$$
  The full model maximizes R^2, but when unimportant explanatory variables are added, R_a^2 may decrease. To avoid redundancy, we can choose a model that maximizes R_a^2.
- Mallow's C_p:
$$C_p = \hat{\Gamma} = \frac{\text{SSE}(p)}{\hat{\sigma}^2} - (n-2p)$$
  - $\hat{\sigma}^2$ is the estimated residual variance by regressing on the complete set of variables.

  Interpretation:

  Consider a model of $p-1$ variables:
$$\hat{\mathbf{Y}}^p = \mathbf{X}_p(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{Y} = \mathbf{H}_p\mathbf{Y}$$
$$\mathbb{E}(\hat{\mathbf{Y}}^p) = \mathbf{H}_p\mathbb{E}(\mathbf{Y}) = \mathbf{H}_p\boldsymbol{\mu} \quad \text{Var}(\hat{\mathbf{Y}}^p) = \sigma^2\mathbf{H}_p$$
  where $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_n)^T$ is the true mean responses at the X_is.

  Bias-variance trade-off:
$$\mathbb{E}(\hat{\mathbf{Y}}_i^p - \mu_i)^2 + \text{Var}(\hat{\mathbf{Y}}_i^p)$$
  Variance part:
$$\sum_{i=1}^{n}\text{Var}(\hat{Y}_i^p) = \text{tr}(\text{Var}(\hat{\mathbf{Y}}^{\mathbf{p}})) = \text{tr}(\sigma^2\mathbf{H_p}) = p\sigma^2$$
  Bias part:
$$\left(\mathbb{E}(\hat{\mathbf{Y}}^{\mathbf{p}}) - \boldsymbol{\mu}\right)^T\left(\mathbb{E}(\hat{\mathbf{Y}}^{\mathbf{p}}) - \boldsymbol{\mu}\right) = (\mathbf{H_p}\mu - \mu)^{\mathbf{T}}(\mathbf{H_p}\mu - \mu) = \mathbb{E}\left(\text{SSE}(p)\right) - (n-p)\sigma^2$$

Total mean squared error:

$$\sum_{i=1}^{n} \mathbb{E}(\hat{Y}_i^p - \mu_i)^2 = \mathbb{E}\left(\mathrm{SSE}(p)\right) - (n-2p)\sigma^2$$

The model's performance measure, sum squared prediction error(SSPE):

$$\Gamma_p = \frac{\sum_{i=1}^{n} \mathbb{E}(\hat{Y}_i^p - \mu_i)}{\sigma^2} = \frac{\mathbb{E}\left(\mathrm{SSE}(p)\right)}{\sigma^2} - (n-2p)$$

Suppose the full model has $P-1$ variables, and the current model is exactly the full model, then $\Gamma_P = P$.

- When a model is unbiased, $C_p \approx p$.
- $C_p \gg p$ means underfitting, indicating missing relevant predictors.
- $C_p \ll p$ may indicate overfitting.
- Among all unbiased models, prefer the model with small C_p.

- Akaike Information Criterion(AIC):

$$\mathrm{AIC}(p) = -2\log(\hat{L}) + 2p$$

where $\hat{L}$ is the maximum likelihood under the model.

Under a linear regression model involving $p-1$ variables:

$$\mathrm{AIC}(p) = n \ln\left(\frac{\mathrm{SSE}(p)}{n}\right) + 2p$$

The AIC criterion select the model that minimizes $\mathrm{AIC}(p)$.

- Bayesian Information Criterion(BIC): Also known as Schwarz's Bayesian Criterion (SBC).

$$\mathrm{BIC}(p) = -2\log(\hat{L}) + \log(n)p$$

where $\hat{L}$ is the maximum likelihood under the model.

Under a linear regression model involving $p-1$ variables:

$$\mathrm{BIC}(p) = n \ln\left(\frac{\mathrm{SSE}(p)}{n}\right) + p\ln(n)$$

The BIC criterion select the model that minimizes $\mathrm{BIC}(p)$.

- PRESS: Prediction/Predicted/Predictive Residual Error Sum of Squares.

$$\mathrm{PRESS}(p) = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(-i)})^2$$

  - $\hat{Y}_{i(-i)} = [1, X_{i1}, X_{i2}, \cdots, X_{i,p-1}]\,\hat{\beta}_{(-i)}$
  - $\hat{\beta}_{(-i)} = (X_{(-i)}^T X_{(-i)})^{-1} X_{(-i)}^T Y_{(-i)}$
  - $X_{(-i)}$ is the design matrix without the i-th case.

  The PRESS criterion select the model that minimizes $\mathrm{PRESS}(p)$.

  - The PRESS criterion is a form of cross-validation.
  - It is noteworthy that

$$Y_i - \hat{Y}_{i(-i)} = \frac{e_i}{1-h_{ii}}$$

- R_p^2:

$$R_p^2 = 1 - \frac{\mathrm{PRESS}}{\mathrm{SST}}$$

  - If $R_p^2 \ll R^2$, the model is likely to be overfitting.
  - R_p^2 can be negative.

3.5.4 Partial Regression Plot

*Also called Added Variable Plot or Adjusted Variable Plot (AV plot).

Procedure of plotting AV plot for X_i:

1. Use the other X's to predict Y, get residuals $Y|X_{-i}$
2. Use the other X's to predict X_i, get residuals $X_i|X_{-i}$
3. Plot residuals $Y|X_{-i}$ against residuals $X_i|X_{-i}$.

These plots show the strength of the marginal relationship between Y and a single X_i in the full model.

3.6 Diagnostics of Influential

3.6.1 Leverage

Leverage: A measure of how far away the independent variable values of an observation are from those of the other observations. Also known as self-sensitivity or self-influence.

$$h_{ii} = \frac{\partial \hat{Y}_i}{\partial Y_i} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2} \in [0, 1]$$

$$\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2$$

- A point with zero leverage has no effect on the regression model.
- If a point has leverage equal to 1 the line must follow the point perfectly
- Leverage measures the "distance" of X from center.
- Leverage helps determine outlying/influential values.

A rule of thumb: When $h_{ii} > \frac{2p}{n}$, we declare case i has extreme X values.

Concept discrimination:

1. Outlier: A data point whose response does not follow the general trend of the rest of the data (high discrepancy, with large residual).
2. High leverage: Data point with "extreme" predictor values. A small change in this point has the potential to be influential.
3. Influential: A data point is influential if it unduly influences any part of a regression analysis. Removing the observation substantially changes the estimation.

3.6.2 Different Types of Residuals

- Standardized residual:

  $$\frac{e_i}{s(e_i)}$$

  Note that $s^2\{e_i\} = (1 - h_{ii})\mathrm{MSE}$.

- Deleted residual:

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

  $\hat{Y}_{i(-i)}$ is the predicted response for the i-th observation based on the estimated model with the i-th observation deleted.

- (Externally) Studentized residual: Standardized deleted residual without self-influence.

$$t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{\text{MSE}_{(-i)}(1-h_{ii})}} \sim t_{n-p-1}$$

- Internally Studentized residual: Standardized deleted residual. This is not commonly used.

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

3.6.3 Test for outliers

Hypothesis:

$$H_0: \text{ case } i \text{ is not an outlier.}$$

Test statistic:

$$t_i = \frac{e_i}{\sqrt{\text{MSE}_{(-i)}(1-h_{ii})}} \text{ (Externally studentized residual)}$$

Under H_0:

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

Decision Rule: If $|t_i| > t_{n-1-p}(1-\alpha/2n)$, reject H_0 (Note: Bonferroni Adjustment is adopted).

3.6.4 Measurements of Influential

- Cook's Distance: Influence of the i-th case on all predicted values.

$$D_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(-i)})^2}{ps^2} = \frac{e_i^2}{ps^2}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

- DFFIT: Difference caused to fitted values.

$$\text{DFFIT}_i = \hat{Y}_i - \hat{Y}_{i(-i)} = \frac{h_{ii}}{1-h_{ii}} e_i$$

- DFFITS: Studentized DFFIT.

$$\text{DFFITS}_i = \frac{\text{DFFIT}_i}{s(\hat{Y})} = t_i \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

  We consider it influential if $|\text{DFFITS}| > 1$(small to medium data) or $|\text{DFFITS}| > 2\sqrt{\frac{p}{n}}$(large data).

- DFBETAS: Difference in beta estimates.

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

  - c_{kk} is the k-th diagonal value of $(X^T X)^{-1}$.

  We consider it influential if $|\text{DFBETAS}| > 1$(small to medium data) or $|\text{DFBETAS}| > \frac{2}{\sqrt{n}}$(large data).

- VIF: Variance Inflation Factor.

$$\text{VIF}_k = \frac{1}{1-R_k^2}$$

  where R_k^2 is from regressing X_k against the other $p-2$ explanatory variables.

  Average VIF:

$$\overline{\text{VIF}} = \frac{1}{p-1}\sum_{i=1}^{p-1} \text{VIF}_k$$

  There is excessive multicollinearity when the largest VIF exceeds 10, or the average VIF is considerably larger than 1.

3.7 Lack of Fit Test

When we have repeat observations at various values of X, the error term will be partitioned into pure error (error within replicates) and a lack of fit.

Suppose there are n observations in total, partitioned into c groups with distinct Xs, and each group has n_i observations. Denote the j-th response in the i-th group as Y_{ij} and the mean response in the i-th group as $\bar{Y}_i$.

Decomposition of Error Deviation:

Source of Variation	SS	df	MS
Regression	$\text{SSR} = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$\text{MSR} = \frac{\text{SSR}}{1}$
Error	$\text{SSE} = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$
Lack of Fit	$\text{SSLF} = \sum\sum(\hat{Y}_{ij} - \bar{Y}_i)^2$	$c-2$	$\text{MSLF} = \frac{\text{SSLF}}{c-2}$
Pure Error	$\text{SSPE} = \sum\sum(Y_{ij} - \bar{Y}_i)^2$	$n-c$	$\text{MSPE} = \frac{\text{SSPE}}{n-c}$
Total	$\text{SSTO} = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

- Prerequisite: Replicates.
- Properties:

$$\text{SSE} = \text{SSPE} + \text{SSLF}$$

$$\mathbb{E}(\text{MSPE}) = \sigma^2 \quad \mathbb{E}(\text{MSLF}) = \sigma^2 + \frac{1}{c-2}\sum_{i=1}^{c} n_i(\mu_i - \beta_0 - \beta_1 X_i)^2$$

- Hypothesis test:

$$\text{Full model: } Y_{ij} = \mu_i + \varepsilon_{ij}$$

$$\text{Reduced model: } Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$$

Hypothesis:

$$H_0 : \mu_i = \beta_0 + \beta_1 X_i \iff H_1 : \mu_i \neq \beta_0 + \beta_1 X_i$$

Test statistic:

$$F = \frac{\text{MSLF}}{\text{MSPE}}$$

Distribution:

$$F \overset{H_0}{\sim} \text{F}_{c-2,n-c}$$

4 Remedies

Remedial Measures: Use transformation and other methods to fix/adjust the problems, or change to different models/analytic strategies to accommodate violations.

4.1 Variable Transformation

The goal of transformation:

- Stablize Variance
- Improve Normality
- Simplify the Model

4.1.1 Variance Stabilizing Transformations

If $\mathbb{E}(Y) = \mu_x$ and $\text{Var}(Y) = h(\mu_x)$, to stabilize variance, conduct transformation on Y.

$$f(\mu) = \int \frac{cd\mu}{\sqrt{h(\mu)}}$$

Examples:

- When $h(\mu) = \mu^2$, $f(\mu) = \log(\mu)$.
- When $h(\mu) = \mu^{2v}, v \neq 1$, $f(\mu) = c\mu^{1-v}$.

4.1.2 Box-Cox Transformation

$$Y^* = \frac{Y^\lambda - 1}{\lambda}$$

Standarized transformed Y is

$$h(Y_i, \lambda) = \begin{cases} K_1 \left(Y_i^\lambda - 1\right), \text{if } \lambda \neq 0 \\ K_2 \log(Y_i), \text{if } \lambda = 0 \end{cases}$$

- $K_2 = (\prod Y_i)^{\frac{1}{n}} \quad K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$

4.2 Weighted Regression

4.2.1 Model

- Model: Suppose Y_i's have different but known variances, then the linear regression model can be written as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

- Likelihood function:

$$L = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1}}{\sigma_i}\right)^2\right]$$

- Loss function: Maximizing the likelihood function is equivalent to minimizing the loss function:

$$Q = \sum w_i e_i^2 = \sum \frac{1}{\sigma_i^2} \left(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1}\right)^2$$

- Solution: Denote

$$\mathbf{W} = \begin{bmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{bmatrix} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n^2} \end{bmatrix} \quad \mathbf{Y}^* = \mathbf{W}^{\frac{1}{2}}\mathbf{Y} \quad \boldsymbol{\varepsilon}^* = \mathbf{W}^{\frac{1}{2}}\boldsymbol{\varepsilon}$$

Then

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}_w$$
$$\mathbf{b}_w = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$
$$\text{Var}(\mathbf{b}_w) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$$
$$\mathbf{W}^{\frac{1}{2}}\mathbf{Y} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}^{\frac{1}{2}}\boldsymbol{\varepsilon} \Longrightarrow \mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$
$$\text{Var}(\mathbf{Y}^*) = \text{Var}(\mathbf{W}^{\frac{1}{2}Y}) = \text{Var}(\boldsymbol{\varepsilon}^*) = \text{Var}(\mathbf{W}^{\frac{1}{2}}\boldsymbol{\varepsilon}) = \mathbf{I}$$

When only relative magnitude of variances are known (σ_i^2 unknown, σ_i^2/σ_j^2 known), denote the unknown $\sigma^2 = k$, then

$$\mathbf{W} = k \begin{bmatrix} \frac{1}{\sigma_1^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_n^2} \end{bmatrix}$$

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

$$\text{Var}(\mathbf{b}_w) = k(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$$

$$\text{Var}(\mathbf{Y}^*) = k\mathbf{I}$$

The estimator of the proportionality constant k is

$$\hat{k} = \text{MSE}_{\text{wls}} = \frac{\sum w_i e_i^2}{n-p}$$

- Hat matrix: $\underset{n\times n}{\mathbf{H}} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}$

4.2.2 OLS with Unequal Error Variances

If one uses $\mathbf{b}$ (rather than $\mathbf{b}_w$) with unequal error variances, the ordinary least squares estimators of the regression coefficients are still unbiased and consistent, but they are no longer minimum variance estimators. The correct variance-covariance matrix is

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

To estimate $\text{Var}(\mathbf{b})$, use White estimator:

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

where

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} e_1^2 & & \\ & \ddots & \\ & & e_n^2 \end{bmatrix}$$

White's estimator is sometimes referred to as a robust covariance matrix, because it can be used to make appropriate inferences about the regression parameters based on OLS without having to specify the form of the non-constant error variance.

4.2.3 Generalized Least Squares Regression

When there is a certain degree of correlation between the residuals, denote the covariance as

$$\text{Cov}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{\Omega}$$

To estimate $\boldsymbol{\beta}$, minimize the squared Mahalanobis length of the residual vector $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$.

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{\Omega}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}^{-1}\mathbf{Y}\end{aligned}$$

It is equivalent to applying OLS to a linearly transformed version of the data:

$$\mathbf{C}^{-1}\mathbf{Y} = \mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}^{-1}\boldsymbol{\varepsilon}$$

- $\mathbf{C}$ satisfies $\mathbf{\Omega} = \mathbf{C}\mathbf{C}^T$
- Note that $\text{Var}(\mathbf{C}^{-1}\boldsymbol{\varepsilon}) = \mathbf{I}$

4.3 Penalized Regression

To avoid overfitting or numerical instability in regression models, we add penalty terms to the loss function. This approach is called penalized regression.

4.3.1 Ridge Regression

- Model: Ridge regression is to minimize

$$Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_i \beta_i^2$$

- Solution: $\mathbf{b} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^T\mathbf{X}^T\mathbf{Y}$
- Hat matrix: $\underset{n\times n}{\mathbf{H}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$
- Bayesian interpretation: By maximum likelihood approach, to estimate $\boldsymbol{\beta}$ is to maximize the (posterior) probability of $\boldsymbol{\beta}$:

$$P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y})$$

  In other words,

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \underset{\text{Posterior}}{P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y})} = \arg\max_{\boldsymbol{\beta}} \underset{\text{Likelihood}}{P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})} \underset{\text{Prior}}{P(\boldsymbol{\beta}|\mathbf{X})}$$

  If we assume each β_i follows a normal distribution $\mathcal{N}(0, c^2)$, then

  1. The prior part is

$$P(\boldsymbol{\beta}|\mathbf{X}) = \prod_{i=1}^{n} p(\beta_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}c} \exp\left(-\frac{\beta_i^2}{2c^2}\right)$$

  2. The likelihood part is

$$P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(Y_i|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

  3. So the posterior probability of $\boldsymbol{\beta}$ is

$$P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = P(\boldsymbol{\beta}|\mathbf{X})P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}) = \left(\frac{1}{2\pi c\sigma}\right)^n \exp\left(-\frac{1}{2c^2}\sum_{i=1}^{n}\beta_i^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\right)$$

  Thus, to maximize the posterior is to minimize

$$\begin{aligned} Q &= \frac{1}{2c^2}\sum \beta_i^2 + \frac{1}{2\sigma^2}\sum_i \varepsilon_i^2 \\ &= \frac{1}{2\sigma^2}\left(\sum \varepsilon_i^2 + \frac{\sigma^2}{c^2}\sum \beta_i^2\right) \\ &= \frac{1}{2\sigma^2}\left(\sum \varepsilon_i^2 + \lambda \sum \beta_i^2\right) \end{aligned}$$

  The form is equivalent to ridge regression.[3]

[3] Note that $\sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$

4.3.2 LASSO Regression

- Model: LASSO (Least Absolute Shrinkage and Selection Operator) regression is to minimize

$$Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_i |\beta_i|$$

- Bayesian interpretation: Similar to ridge regression, LASSO regression sets the prior distribution of each β_i as Laplace distribution:

$$f(\beta_i) = \frac{1}{2b} \exp\left(-\frac{|\beta_i|}{b}\right)$$

To estimate $\boldsymbol{\beta}$ is to maximize the (posterior) probability of $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} \underbrace{P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y})}_{\text{Posterior}} = \arg\max_{\beta} \underbrace{P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta})}_{\text{Likelihood}} \underbrace{P(\boldsymbol{\beta}|\mathbf{X})}_{\text{Prior}}$$

1. The prior part is

$$P(\boldsymbol{\beta}|\mathbf{X}) = \prod_{i=1}^{n} p(\beta_i) = \prod_{i=1}^{n} \frac{1}{2b} \exp\left(-\frac{|\beta_i|}{b}\right)$$

2. The likelihood part is

$$P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(Y_i|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

3. So the posterior probability of β is

$$P(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y}) = P(\boldsymbol{\beta}|\mathbf{X})P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}) = \left(\frac{1}{2\sqrt{2\pi}b\sigma}\right)^n \exp\left(-\frac{1}{b}\sum_{i=1}^{n}|\beta_i| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\right)$$

Thus, to maximize the posterior is to minimize

$$\begin{aligned} Q &= \frac{1}{b}\sum|\beta_i| + \frac{1}{2\sigma^2}\sum_i \varepsilon_i^2 \\ &= \frac{1}{2\sigma^2}\left(\sum \varepsilon_i^2 + \frac{2\sigma^2}{b}\sum|\beta_i|\right) \\ &= \frac{1}{2\sigma^2}\left(\sum \varepsilon_i^2 + \lambda\sum|\beta_i|\right) \end{aligned}$$

The form is equivalent to LASSO regression.

4.3.3 Elastic Net

Elastic net combines L_1 and L_2 losses linearly.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1 \sum_i |\beta_i| + \lambda_2 \sum_i \beta_i^2\right]$$

4.3.4 Comments

- In penalized regression, the estimator of coefficients are biased. But they are more stable and yield interpretability.
- Standardization is necessary in convenience of the choice of magnitude of λ.
- LASSO regression tends to estimate the regression coefficient as 0, which serves as a sparse selection process.

4.4 Other Regression Methods

- Robust Regression:
  - Least Absolute Deviations (LAD/LAR): Minimize sum of absolute values of residuals.
  - Least Median of Squares (LMS): Minimize the median of the squares of residuals.
  - Iteratively Reweighted Least Squares (IRLS): Weighted regression with weights based on residuals.
- Nomparametric Regression:
  - Local Polynomial Regression (LOWESS, locally weighted scatterplot smoothing): Predict in an continuous interval based on every sample with certain weights.

5 Analysis of Variance

5.1 Overview

- Goal: ANOVA focuses on the mean responses for the different factor levels. The analysis usually proceeds in two steps:
  - Determine whether or not the factor level means are the same.
  - If the factor level means differ, examine how they differ and what the implications of the diffences are.
- Terms:
  - Factor: The explanatory variable X is categorical/qualitative/discrete. We call it a factor.
  - Level: The possible values of a factor are called levels. We often refer to these levels as groups or treatments.
- Assumptions:
  - The response variable has a mean that may depend on the level of the factor.
  - The responses for each factor level are random selections from the corresponding probability distribution and are independent of any other factor level.
  - Each probability distribution is normal.
  - Each probability distribution has the same variance.
- Relationship with regression:
  - ANOVA is regression.
  - ANOVA does not consider the quantitative differences in the X levels or their statistical relation to expected Y.
  - Regression models are concerned with the statistical relation between predictor variables and quantitative response variable.

5.2 One-Way ANOVA

Suppose we have N samples, divided into r groups, and each group has n_i observations. By Y_{ij} we use i to denote the level of the factor and j to denote the j-th observation at factor level i.

Assume μ_i be the theoretical mean of the i-th level/group, ε_{ij} be random disturbance, and

$$\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad \bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{N}\sum_{i=1}^{r} n_i \bar{Y}_{i\cdot}$$

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 \quad s^2 = \frac{\sum_{i=1}^{r}(n_i-1)s_i^2}{\sum_{i=1}^{r}(n_i-1)}$$

$$\varepsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$$

5.2.1 Cell Means Model

- Model:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

$$\mathbf{Y} = \begin{bmatrix} \underset{n_1\times 1}{\mathbf{1}} & 0 & \cdots & 0 & 0 \\ 0 & \underset{n_2\times 1}{\mathbf{1}} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \underset{n_{r-1}\times 1}{\mathbf{1}} & 0 \\ 0 & 0 & \cdots & 0 & \underset{n_r\times 1}{\mathbf{1}} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{r-1} \\ \mu_r \end{bmatrix} + \varepsilon$$

  The i-th explanatory variable is equal to 1 if the observation is from the i-th group.[4] Note that there is no intercept.

  The parameters of the model are

$$\sigma^2, \mu_1, \mu_2, \cdots, \mu_r$$

- Parameter estimation:

$$\hat{\mu}_i = \bar{Y}_{i.} = \sum_{j=1}^{n_i} Y_{ij}/n_i \text{ (Sample mean)}$$

$$s_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2/(n_i - 1) \text{ (Sample variance)}$$

$$\hat{\sigma}^2 = s^2 = \sum_{i=1}^{r}\left((n_i-1)s_i^2\right) / \left(\sum_{i=1}^{r}(n_i-1)\right) \text{ (Pooled estimate)}$$

- ANOVA table:

Table 4: ANOVA Table

Source	df	SS	MS	F-value
Regression	$r-1$	$\sum_{ij}(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$\mathrm{SSR}/\mathrm{df}_R$	MSR/MSE
Error	$N-r$	$\sum_{ij}(Y_{ij} - \bar{Y}_{i.})^2$	$\mathrm{SSE}/\mathrm{df}_E$	
Total	$N-1$	$\sum_{ij}(Y_{ij} - \bar{Y}_{..})^2$	$\mathrm{SST}/\mathrm{df}_T$	

  Properties:

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$$

$$\mathbb{E}(\mathrm{MSE}) = \sigma^2 \quad \mathbb{E}(\mathrm{MSTR}) = \sigma^2 + \frac{\sum n_i(\mu_i - \mu_.)^2}{r-1}$$

  - $\mu_. = \frac{1}{N}\sum n_i\mu_i$

- Hypothesis test:

  - Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r = \text{constant} \iff H_1 : \text{not all } \mu_i\text{'s are the same}$$

[4] A "cell" refers to a level of the factor.

- Test statistic:
$$F^* = \frac{\text{MSR}}{\text{MSE}}$$
- Distribution:
$$F^* \overset{H_0}{\sim} \text{F}_{r-1,N-r}$$

5.2.2 Factor Effects Model

- Model:
$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$
$$\mathbf{Y} = \begin{bmatrix} \underset{n_1\times 1}{\mathbf{1}} & \underset{n_1\times 1}{\mathbf{1}} & 0 & \cdots & 0 \\ \underset{n_2\times 1}{\mathbf{1}} & 0 & \underset{n_2\times 1}{\mathbf{1}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \underset{n_{r-1}\times 1}{\mathbf{1}} & 0 & 0 & \cdots & \underset{n_{r-1}\times 1}{\mathbf{1}} \\ \underset{n_r\times 1}{\mathbf{1}} & \underset{n_r\times 1}{-\mathbf{1}} & \underset{n_r\times 1}{-\mathbf{1}} & \cdots & \underset{n_r\times 1}{-\mathbf{1}} \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{r-2} \\ \tau_{r-1} \end{bmatrix} + \varepsilon$$
  - Constraint: $\sum_{i=1}^{r} \tau_i = 0$.
  - To avoid multicollinearity, substitute τ_r by $-\tau_1 - \tau_2 - \cdots - \tau_{r-1}$(the last row).
  - $\mu = \frac{\sum_{i=1}^{r} \mu_i}{r}$ is the (unweighted) overall mean, a constant defined to fit the purpose of the study.
  - $\tau_i = \mu_i - \mu$ is called the i-th factor level effect or the i-th treatment effect, a constant for each factor level.

  The parameters of the model are
$$\sigma^2, \mu, \tau_1, \tau_2, \cdots, \tau_r$$
- Parameter estimation:
$$\hat{\mu} = \frac{\sum_{i=1}^{r} \bar{Y}_{i.}}{r} \quad \hat{\tau}_i = \bar{Y}_{i.} - \hat{\mu}$$
- Variation - Weighted case: The analysis becomes complicate when n_i's are not all equal. If we desire weighted means in the first place, define
$$\mu = \frac{\sum_{i=1}^{r} n_i \mu_i}{N} = \sum_{i=1}^{r} w_i \mu_i$$
  Then the constraint becomes
$$\sum_{i=1}^{r} w_i \tau_i = 0$$

5.2.3 Inference

- Inference on one level mean:

  Theoretical distribution: $\bar{Y}_{i.} \sim \mathcal{N}(\mu_i, \sigma^2/n_i)$

  Confidence interval: $\bar{Y}_{i.} \pm t_{N-r}(1-\alpha/2)s/\sqrt{n_i}$

  To get simutaneous confidence interval, use Bonferroni correction to reduce familywise error rate.
  - To get confidence interval for means on all r levels, use $1 - \alpha/r$ as the confidence level.
  - To test the null hypothesis that all μ_i's are equal, due to the $r(r-1)/2$ pairwise comparisons, use $1 - \alpha/\frac{r(r-1)}{2}$ as the confidence level.

- Inference on difference of two level means:

$$\text{Theoretical distribution: } \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \sim \mathcal{N}(\mu_i - \mu_j, \sigma^2/n_i + \sigma^2/n_j)$$

$$\text{Confidence interval: } \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm t_c s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \text{ where } t_c = t_{N-r}(\alpha/2)$$

  - Tukey's HSD (Honestly Significant Difference test): Use studentized range distribution instead of t to get the confidence interval:

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm t_c s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \text{ where } q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{s/\sqrt{n}} \text{ and } t_c \to q_c/\sqrt{2}$$

  - Scheffé's Method: The confidence interval is

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm t_c s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \text{ where } t_c \to \sqrt{(r-1)F_{r-1,N-r}(1-\alpha)}$$

5.2.4 Contrast

- Linear Combinations of Means: Consider the linear combination

$$L = \sum_{i=1}^{r} c_i \mu_i$$

  Parameter estimation:

$$\hat{L} = \sum_{i=1}^{r} c_i \bar{Y}_{i\cdot} \sim \mathcal{N}\left(L, \text{Var}(\hat{L})\right) \quad \text{Var}(\hat{L}) = \sum_{i=1}^{r} c_i^2 \text{Var}(\bar{Y}_{i\cdot}) \quad s^2\{\hat{L}\} = s^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

  Hypothesis test:

  - Hypothesis:

$$H_0 : L = L_0$$

  - Test statistic:

$$T = \frac{\hat{L} - L_0}{\sqrt{s^2\{\hat{L}\}}}$$

  - Distribution:

$$T \overset{H_0}{\sim} t_{N-r}$$

- Contrast: A special case of linear combination, with the constraint

$$\sum_{i=1}^{r} c_i = 0$$

  - Hypothesis:

$$H_0 : L = 0$$

  - Test statistic:

$$T = \frac{\hat{L}}{\sqrt{s^2\{\hat{L}\}}} = \frac{\sum_{i=1}^{r} c_i \hat{Y}_{i\cdot}}{s^2 \sum_{i=1}^{r} c_i^2/n_i} \quad T^2 = \frac{(\sum_{i=1}^{r} c_i \hat{Y}_{i\cdot})^2}{s^2 \sum_{i=1}^{r} c_i^2/n_i}$$

  - Distribution:

$$T \overset{H_0}{\sim} t_{N-r} \quad T^2 \overset{H_0}{\sim} F_{1,N-r}$$

Denote contrast sum of squares (SSC) as

$$\text{SSC} = \left(\sum_{i=1}^{r} c_i \hat{Y}_{i.}\right)^2 / \sum_{i=1}^{r} \left(c_i^2/n_i\right)$$

Then

$$T^2 = \text{SSC}/\text{MSE} \sim \text{F}_{1,N-r}$$

SSC represents the amount of variation due to this contrast.

5.3 Two-Way ANOVA

Consider two categorical explanatory variables or factors. Since groups can be classified in two ways, the analysis is called two-way ANOVA.

Suppose we have N samples, divided into a and b groups by each factor, thus classified into ab subgroups, and each subgroup has n_{ij} observations. By Y_{ijk} we use i to denote the level of factor 1, j to denote the level of factor 2, and k to denote the k-th observation at factor level i & j.

Assume $\mu_{i.}$ be the theoretical mean of the i-th level/group of factor 1, $\mu_{.j}$ be the theoretical mean of the j-th level/group of factor 2, μ_{ij} be the theoretical mean of the ij-th subgroup, ε_{ijk} be random disturbance, and

$$\bar{Y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \quad \bar{Y}_{...} = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$s_{ij}^2 = \frac{1}{n_{ij}-1} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij.})^2 \quad s^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (n_{ij}-1) s_{ij}^2 / \sum_{i=1}^{a} \sum_{j=1}^{b} (n_{ij}-1)$$

$$\varepsilon_{ijk} \overset{i.i.d}{\sim} N(0, \sigma^2)$$

Double factor cases can also be analyzed by one-way ANOVA (with ab levels), but to identify the underlying mechanism of two factors, we develop two-way ANOVA theory.

In this chapter we discuss two-way ANOVA with equal sample sizes ($n_{ij} = n$).

5.3.1 Cell Means Model

- Model:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

  Model parameters:

$$\sigma^2, \mu_{ij} \ \ (i = 1, 2, \cdots, a; j = 1, 2, \cdots, b)$$

- Parameter estimation:

$$\hat{\mu}_{ij} = \bar{Y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \ \text{(Sample mean)}$$

$$s_i^2 = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij.})^2 \ \text{(Sample variance)}$$

$$\hat{\sigma^2} = s^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (n_{ij}-1) s_{ij}^2 / \sum_{i=1}^{a} \sum_{j=1}^{b} (n_{ij}-1) \ \text{(Pooled estimate)}$$

5.3.2 Factor Effects Model

- Model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

Model parameters:

$$\sigma^2, \alpha_i, \beta_j, (\alpha\beta)_{ij} \ \ (i = 1, 2, \cdots, a; j = 1, 2, \cdots, b)$$

$\alpha_i = \mu_{i.} - \mu$ is the main effect of level i of factor 1.
$\beta_j = \mu_{.j} - \mu$ is the main effect of level j of factor 2.
$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$ is the interaction term of factor 1 and 2.

- Constraints:

$$\sum_{i=1}^{a} \alpha_i = 0 \ \ (\text{df} = a - 1) \quad \sum_{j=1}^{b} \beta_j = 0 \ \ (\text{df} = b - 1)$$

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \ \text{ for all } j \quad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0 \ \text{ for all } i \ \ (\text{df} = (a-1)(b-1))$$

Note there is $a + b - 1$ constraints on interaction terms, and 2 constraints on main effect terms. The total degree of freedom for the model is $(a - 1) + (b - 1) + (a - 1)(b - 1) = ab - 1$.

- Parameter estimation:

$$\hat{\mu}_{ij} = \bar{Y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \ \ \text{(Sample mean)}$$

$$s_{ij}^2 = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij.})^2 \ \ \text{(Sample variance)}$$

$$\hat{\sigma^2} = s^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (n_{ij} - 1) s_{ij}^2 / \sum_{i=1}^{a} \sum_{j=1}^{b} (n_{ij} - 1) \ \ \text{(Pooled estimate)}$$

5.3.3 ANOVA Table

Table 5: ANOVA Table

Source	df	SS	MS	F-value
A	$a - 1$	$\sum_{ijk} (\bar{Y}_{i..} - \bar{Y}_{...})^2$	SSA/df_A	MSA/MSE
B	$b - 1$	$\sum_{ijk} (\bar{Y}_{.j.} - \bar{Y}_{...})^2$	SSB/df_B	MSB/MSE
AB	$(a - 1)(b - 1)$	$\sum_{ijk} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$	$\text{SSAB}/\text{df}_{AB}$	MSAB/MSE
Error	$ab(n - 1)$	$\sum_{ijk} (Y_{ijk} - \bar{Y}_{ij.})^2$	SSE/df_E	
Total	$abn - 1$	$\sum_{ijk} (Y_{ijk} - \bar{Y}_{...})^2$	SST/df_T	

Properties:

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}$$

$$\mathbb{E}(\text{MSA}) = \sigma^2 + \frac{bn}{a-1} \sum_{i=1}^{a} \alpha_i^2 \quad \mathbb{E}(\text{MSB}) = \sigma^2 + \frac{an}{b-1} \sum_{j=1}^{b} \beta_j^2$$

$$\mathbb{E}(\text{MSE}) = \sigma^2 \quad \mathbb{E}(\text{MSAB}) = \sigma^2 + \frac{n}{(a-1)(b-1)} \sum_{i=1}^{a} \sum_{j=1}^{b} (\alpha\beta)_{ij}^2$$

Hypothesis test:

- Hypotheses:

$$H_{0A} : \alpha_i = 0 \text{ for all } i$$
$$H_{0B} : \beta_j = 0 \text{ for all } j$$
$$H_{0AB} : (\alpha\beta)_{ij} = 0 \text{ for all } (i, j)$$

- Test statistic:

$$F_A = \text{MSA}/\text{MSE}, \quad F_B = \text{MSB}/\text{MSE}, \quad F_{AB} = \text{MSAB}/\text{MSE}$$

- Distribution:

$$F_A \overset{H_0}{\sim} \text{F}_{a-1,ab(n-1)}, \quad F_B \overset{H_0}{\sim} \text{F}_{b-1,ab(n-1)}, \quad F_{AB} \overset{H_0}{\sim} \text{F}_{(a-1)(b-1),ab(n-1)}$$

6 Appendix

Collection of some (relatively) unimportant but interesting topics.

6.1 OLS Solution to Simple Linear Regression

We use least square criterion:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

To minimize the loss, we calculate partial derivatives of β_0 and β_1:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

We then set these partial derivatives equal to zero, using b_0 and b_1 to denote the particular values of β_0 and β_1 that minimize Q:

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$
$$b_0 = \bar{Y} - b_1 \bar{X}$$

A test of the second partial derivatives will show that a minimum is obtained with the least squares estimators b_0 and b_1.

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2 \sum_{i=1}^{n} 1 \qquad \frac{\partial^2 Q}{\partial \beta_1^2} = 2 \sum_{i=1}^{n} X_i^2 \qquad \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} = 2 \sum_{i=1}^{n} X_i$$

$$\implies \mathbf{H} = \det \left(\begin{bmatrix} \frac{\partial^2 Q}{\partial \beta_0^2} & \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 Q}{\partial \beta_1^2} \end{bmatrix} \right) = 4 \left[\left(\sum_{i=1}^{n} 1 \right) \left(\sum_{i=1}^{n} X_i^2 \right) - \left(\sum_{i=1}^{n} X_i \right)^2 \right] \geq 0$$

$$\implies \mathbf{H} \text{ (Hessian matrix) is semi-positive definite}$$

$$\implies (b_0, b_1) \text{ is a minimum point}$$

6.2 Parameter Estimation of Simple Linear Regression

1. **Proof of unbiasedness of s^2**

$$e_i = Y_i - \hat{Y}_i$$
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$$
$$\hat{Y}_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}}))$$
$$\sum Y_i \hat{Y}_i = \sum (\hat{Y}_i + e_i)\hat{Y}_i = \sum \hat{Y}_i^2 + \sum \hat{Y}_i e_i = \sum \hat{Y}_i^2$$

So

$$\begin{aligned}
\mathbb{E}\left[(n-2)s^2\right]) &= \mathbb{E}\left(\sum_{i=1}^{n} e_i^2\right) \\
&= \mathbb{E}\left(\sum (Y_i - \hat{Y}_i)^2\right) \\
&= \mathbb{E}\left(\sum Y_i^2 - 2\sum Y_i \hat{Y}_i + \sum \hat{Y}_i^2\right) \\
&= \mathbb{E}\left(\sum Y_i^2 - \sum \hat{Y}_i^2\right) \\
&= \sum \mathbb{E}(Y_i^2) - \sum \mathbb{E}(\hat{Y}_i^2) \\
&= \sum \left(\mathbb{E}(Y_i)^2 + \mathrm{Var}(Y_i) - \mathbb{E}(\hat{Y}_i)^2 - \mathrm{Var}(\hat{Y}_i\right) \\
&= \sum \left(\sigma^2 - \sigma^2(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}})\right) \\
&= n\sigma^2 - \sigma^2 - \sigma^2 \sum \frac{(X_i - \bar{X})^2}{S_{XX}} \\
&= (n-2)\sigma^2
\end{aligned}$$

Thus we have

$$\mathbb{E}(s^2) = \mathbb{E}(\frac{\sum_{i=1}^{n} e_i^2}{n-2}) = \sigma^2$$

2. **Estimation of β_1**

$$\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \\
&= \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) \\
&= \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i-\bar{X})Y_i - \frac{1}{S_{XX}}\cdot\bar{Y}\sum_{i=1}^{n}(X_i-\bar{X}) \\
&= \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i-\bar{X})Y_i - 0 \\
&= \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i-\bar{X})(\beta_0+\beta_1X_i+\varepsilon_i) \\
&= \frac{1}{S_{XX}}\left(\beta_0\sum_{i=1}^{n}(X_i-\bar{X}) + \beta_1\sum_{i=1}^{n}(X_i-\bar{X})X_i + \sum_{i=1}^{n}(X_i-\bar{X})\varepsilon_i\right) \\
&= \frac{1}{S_{XX}}\left(0 + \beta_1\sum_{i=1}^{n}(X_i^2-\bar{X}^2) + \sum_{i=1}^{n}(X_i-\bar{X})\varepsilon_i\right) \\
&= \beta_1 + \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i-\bar{X})\varepsilon_i
\end{aligned}$$

Given that

$$\mathrm{Var}\left(\frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i-\bar{X})\varepsilon_i\right) = \frac{1}{S_{XX}^2}\sum_{i=1}^{n}(X_i-\bar{X})^2\sigma^2 = \frac{1}{S_{XX}}\sigma^2$$

We have

$$b_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{XX})$$

3. **Estimation of β_0**

$$Y_i \sim \mathcal{N}(\beta_0+\beta_1X_i, \sigma^2)$$

So

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i \sim \mathcal{N}(\beta_0+\beta_1\bar{X}, \frac{\sigma^2}{n})$$

Combined with

$$b_0 = \bar{Y} - \bar{X}b_1$$
$$b_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{XX})$$

We have

$$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2(\frac{1}{n}+\frac{\bar{X}^2}{S_{XX}})\right)$$

4. **Estimation of $\hat{Y}_h$**

$$\hat{Y}_h = b_0 + b_1X_h = \bar{Y} + b_1(X_h-\bar{X})$$

Combined with

$$b_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{XX})$$
$$\bar{Y} \sim \mathcal{N}(\beta_0+\beta_1\bar{X}, \frac{\sigma^2}{n})$$

We have

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

5. **Proof of** $\mathrm{Cov}(\bar{Y}, b_1) = 0$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 \bar{X} + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$$

$$b_1 = \beta_1 + \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i - \bar{X})\varepsilon_i$$

So

$$\begin{aligned}
\mathrm{Cov}(\bar{Y}, b_1) &= \mathrm{Cov}\Big(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i, \frac{1}{S_{XX}}\sum_{i=1}^{n}(X_i - \bar{X})\varepsilon_i\Big) \\
&= \frac{1}{nS_{XX}}\sum_{i=1}^{n}\mathrm{Cov}\Big(\varepsilon_i, (X_i - \bar{X})\varepsilon_i\Big) \\
&= \frac{1}{nS_{XX}}\sum_{i=1}^{n}(X_i - \bar{X})\sigma^2 \\
&= 0
\end{aligned}$$

6.3 Properties of Hat Matrix H

In linear regression, the definition of hat matrix is the projection matrix $\underset{n\times n}{\mathbf{H}}$ which satisfies:

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \mathbf{HY}$$

1. **Hat matrix in different situations**:
   - **OLS**: $\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \Longrightarrow \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
   - **WLS**: $\hat{\mathbf{Y}} = \mathbf{Xb}_w = \mathbf{X}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{WY} \Longrightarrow \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W}$
   - **Ridge**: $\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \Longrightarrow \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$
2. **Association of leverage and hat matrix**

   By definition, the leverage of the i-th observation is $l_i = \dfrac{\partial \hat{Y}_i}{\partial Y_i}$.

   By the property of matrix differentiation,

$$\frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_j} = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)_{ij} \Longrightarrow \frac{\partial(\mathbf{HY})_i}{\partial \mathbf{Y}_j} = \left[\frac{\partial(\mathbf{HY})}{\partial \mathbf{Y}}\right]_{ij}$$

$$\frac{\partial \mathbf{Ax}}{\partial \mathbf{x}} = \mathbf{A}^T \Longrightarrow \frac{\partial(\mathbf{HY})}{\partial \mathbf{Y}} = \mathbf{H}^T$$

   Thus we have

$$l_i = \frac{\partial \hat{Y}_i}{\partial Y_i} = \frac{\partial(\mathbf{HY})_i}{\partial \mathbf{Y}_i} = \left(\frac{\partial(\mathbf{HY})}{\partial \mathbf{Y}}\right)_{ii} = \mathbf{H}_{ii}$$

   By convention, we state that the leverage of the i-th observation is h_{ii}.
3. **Variance of residual**

   Unless the case is WLS where $\mathbf{H}$ is not symmetric, we have

$$\begin{aligned}
\mathrm{Cov}(\mathbf{e}) = \mathrm{Cov}\left(\mathbf{Y} - \hat{\mathbf{Y}}\right) &= \mathrm{Cov}\Big((\mathbf{I} - \mathbf{H})\mathbf{Y}\Big) \\
&= (\mathbf{I} - \mathbf{H})\mathrm{Cov}(\mathbf{Y})(\mathbf{I} - \mathbf{H})^T \\
&= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})^T \\
&= \mathbf{I} - \mathbf{H}
\end{aligned}$$

   So

$$\mathrm{Var}(e_i) = (\mathbf{I} - \mathbf{H})_{ii} = \sigma^2(1 - h_{ii})$$
$$\mathrm{Cov}(e_i, e_j) = (\mathbf{I} - \mathbf{H})_{ij} = -\sigma^2 h_{ij} \ \ (i \neq j)$$

6.4 Regression Through the Origin

It is not suggested to force the regression line through the origin. Consequences are:

1. $\mathrm{SST} \neq \mathrm{SSE} + \mathrm{SSR}$
2. $\underset{n-1}{\mathrm{df}(\mathrm{SST})} \neq \underset{1}{\mathrm{df}(\mathrm{SSR})} + \underset{n-1}{\mathrm{df}(\mathrm{SSE})}$
3. $\sum e_i \neq 0$

6.5 Inverse prediction

Usually, the effect of regressing X on Y and regressing Y on X are not simply reciprocal. The angle between two regression lines is

$$\tan\theta = \frac{1 - r^2}{r\left(\frac{S_X}{S_Y} + \frac{S_Y}{S_X}\right)}$$

To predict X_i based on known response Y_i, we should use

$$\bar{X}_h = \frac{Y_h - b_0}{b_1}$$

rather than regress X on Y.