2.6 Indicator variables

(Example)

Response variable: salary

Response variable: Education(HS,BS,ADV)
Management status (MGT,None)
Year of experience

Regression analysis:

1. Fit separate regression models for different levels of the qualitative predictors (in case there is only one qualitative predictor)/ or different combinations of the levels of the qualitative predictors (in case there are many)

   (e.g.) **6 models(x_i: year of experience)**

   $y_i = \beta_{01} + \beta_{11} x_i + \varepsilon_{i1}$ HS-NONE
   $y_i = \beta_{02} + \beta_{12} x_i + \varepsilon_{i2}$ HS-MGT
   $\vdots$
   $y_i = \beta_{06} + \beta_{16} x_i + \varepsilon_{i6}$ ADV-NONE

2. **Model with dummy/ indicator variables**

$$E_{1i} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ obs falls into HS (for education)} \\ 0 & \text{o.w} \end{cases}$$

$$E_{2i} = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ obs falls into BS (for education)} \\ 0 & \text{o.w} \end{cases}$$

$$MGT_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ obs falls into MGT (for management status)} \\ 0 & \text{o.w} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \gamma_1 E_{1i} + \gamma_2 E_{2i} + \delta \cdot MGT_i + \varepsilon_i$$

The model is equivalent to

$$\begin{cases} y_i = (\beta_0 + \gamma_1) + \beta_1 x_i + \varepsilon_i & : \text{ HS-None} \\ y_i = (\beta_0 + \gamma_1 + \delta) + \beta_1 x_i + \varepsilon_i & : \text{ HS-MGT} \\ y_i = (\beta_0 + \gamma_2) + \beta_1 x_i + \varepsilon_i & : \text{ BS-None} \\ y_i = (\beta_0 + \gamma_2 + \delta) + \beta_1 x_i + \varepsilon_i & : BS - MGT \\ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i & : \text{ ADV-None} \\ y_i = (\beta_0 + \delta) + \beta_1 x_i + \varepsilon_i & : \text{ ADV-MGT} \end{cases}$$

Interpretation:

β_1: the increment of salary when x_i increases in 1 unit the other explanatory variables are fixed

γ_1: the increment of salary for HS compared to for ADV when the other explanatory variables are fixed

γ_2 the increment of salary for BS compared to for ADV when the other explanatory variables are fixed

δ: the increment of salary for MGT compared to for None when the other explanatory variables are fixed

3. **General models with interaction**

(i)

$$
\begin{aligned}
y_i = \beta_0 + \beta_1 x_i + \gamma_1 E_{1i} + \gamma_2 E_{2i} + \delta \cdot MGT_i \\
+ \alpha_1 (E_{1i} MGT_i) + \alpha_2 (E_{2i} MGT_i) + \varepsilon_i
\end{aligned}
$$

$$
\Leftrightarrow \quad \begin{cases}
y_i = (\beta_0 + \gamma_1) + \beta_1 x_i + \varepsilon_i \quad : \text{ HS-None} \\
y_i = (\beta_0 + \gamma_1 + \delta + \alpha_1) + \beta_1 x_i + \varepsilon_i \quad : \text{ HS-MGT} \\
y_i = (\beta_0 + \gamma_2) + \beta_1 x_i + \varepsilon_i \quad : \text{ BS-None} \\
y_i = (\beta_0 + \gamma_2 + \delta + \alpha_2) + \beta_1 x_i + \varepsilon_i \quad : \text{ BS-MGT} \\
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad : \text{ ADV-None} \\
y_i = (\beta_0 + \delta) + \beta_1 x_i + \varepsilon_i \quad : \text{ ADV-MGT}
\end{cases}
$$

$\therefore$ The magnitude of the salary difference between MGT and None also depends on the education level

(ii)

$$
\begin{aligned}
y_i = \beta_0 + \beta_1 x_i + \gamma_1 E_{1i} + \gamma_2 E_{2i} + \delta \cdot MGT_i \\
+ \alpha_1 (E_{1i} MGT_i) + \alpha_2 (E_{2i} MGT_i) \\
+ \xi_1 (x_i E_{1i}) + \xi_2 (x_i E_{2i}) + \xi_3 (x_i MGT_i) + \xi_4 (x_i MGT_i E_{1i}) \\
+ \xi_5 (x_i MGT_i E_{2i}) + \varepsilon_i
\end{aligned}
$$

(**<u>Homework</u>**)
Express respective regression models for the combinations: HS-None, HS-MGT, BS-None,BS-MGT,ADV-None, ADV-MGT
$\Rightarrow$ Notice that the slope vary over combinations of the levels

Regression Approach to ANOVA

- One-way ANalysis Of VAriance model: To explain the variation of the observation of a characteristic Y by a single factor,

$$Y_{ij} = \mu_j + \varepsilon_{ij}, \quad 1 \leq i \leq n_j, \quad 1 \leq j \leq K$$

  Each level of the factor is called a "treatment". In this convention, Y_{ij} is the i^{th} observation from the j^{th} treatment.

  One of the main interests in one-way ANOVA is to test whether there is no treatment effect (on mean) i.e. all $\mu_j's$ are equal.

- Introduce (K-1) indicator variables as follow:

$$X_1 = \begin{cases} 1 & \text{if the observation is from treatment 1} \\ 0 & \text{o.w} \end{cases}$$

$$X_{K-1} = \begin{cases} 1 & \text{if the observation is from treatment } (K-1) \\ 0 & \text{o.w} \end{cases}$$

  $\Rightarrow$ The one-way ANOVA model can be represented by

$$Y_{ij} = \beta_0 + \beta_1 x_{ij,1} + \cdots + \beta_{K-1} x_{ij,K-1} + \varepsilon_{ij}, \quad i \leq i \leq n_j, \quad j = 1, \cdots, K$$

  where $\beta_0 = \mu_K,\ \beta_j = \mu_j - \mu_K, \quad j = 1, \cdots, K-1$

  Therefore, testing whether all $\mu_j's$ are equal is equivalent to testing whether

$$\beta_1 = \beta_2 = \ldots = \beta_{K-1} = 0$$

.

  One may write the one-way ANOVA model as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

Where

$$Y = (Y_{11}, \cdots, Y_{n1,1}, \cdots, Y_{1K}, \cdots, Y_{n_K,K})^T$$

$$\beta = (\beta_0, \cdots, \beta_{K-1})^T$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11,1} & \cdots & x_{11,K-1} \\ 1 & x_{21,1} & \cdots & x_{21,K-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1 1,1} & \cdots & x_{n_1 1,K-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1K,1} & \cdots & x_{1K,K-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_K K,1} & \cdots & x_{n_K K,K-1} \end{pmatrix}$$

Then, it can be shown that

$$\hat{\boldsymbol{\beta}}_0 = \bar{Y}_K,\ \hat{\boldsymbol{\beta}}_j = \bar{Y}_j - \bar{Y}_K, \quad 1 \le j \le K-1 \quad \cdots \text{ (HW)}$$

- Notice that the main interest is to test whether $\beta_1 = \cdots = \beta_{K-1} = 0$ in the regression model.

  Compute

$$SSR = Y^T(H_{\boldsymbol{X}} - H_{\boldsymbol{1}})Y = \sum_{j=1}^{K} n_j(\bar{Y}_j - \bar{Y})^2$$

$$\because Y^T H_{\boldsymbol{X}} Y = \sum_{j=1}^{K} n_j(\bar{Y}_j)^2, \quad Y^T H_{\boldsymbol{1}} Y = N(\bar{Y})^2, \quad with\ N = \sum_{j=1}^{K} n_j$$

$$\& \quad SSE = \sum_{j=1}^{K}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$$

so that $F_0 = \dfrac{SSR/(K-1)}{SSE/(N-(K-1)-1)} \;\sim\; F(K-1, n-K)$ under $H_0 : \beta_1 = \cdots = \beta_{K-1} = 0$

2.7 Maximum Likelihood Estimation

Assume the normality of $\varepsilon_i' s$, Then

$$Y_i \;\sim\; N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2) : \text{ indep}$$

$$\Rightarrow \; L(\beta_0, \beta_1, \cdots, \beta_p, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\bigg(-\frac{1}{2\sigma^2}\Big(Y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\Big)^2\bigg)$$

$$= \bigg(\frac{1}{\sqrt{2\pi\sigma^2}}\bigg)^n \exp\bigg(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(Y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\Big)^2\bigg)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\bigg(-\frac{1}{2\sigma^2}\|Y - \boldsymbol{X}\beta\|^2\bigg) \equiv L(\beta, \sigma^2)$$

$$\begin{aligned} l(\beta, \sigma^2) &= \log L(\beta, \sigma^2) \\ &= -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Y - \boldsymbol{X}\beta\|^2 \end{aligned}$$

$\Rightarrow$ likelihood equation :

$$\begin{cases} \frac{\partial^1 l}{\partial \beta} = \frac{1}{\sigma^2}\boldsymbol{X}^T(Y - \boldsymbol{X}\beta) = 0 \\ \\ \frac{\partial^1 l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y - \boldsymbol{X}\beta\|^2 = 0 \end{cases}$$

$$\therefore \hat{\beta}^{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y = \hat{\beta}^{LSE} \qquad \hat{\sigma^2}^{MLE} = \frac{\|Y - \boldsymbol{X}\hat{\beta}\|^2}{n} \neq \hat{\sigma^2}^{LSE}$$

<u>Rao-Crammer lower bound:</u>

Assume σ^2 is fixed. Then,

$$\begin{aligned} \frac{\partial^1 l}{\partial \beta} &= \frac{1}{\sigma^2}\boldsymbol{X}^T(Y - \boldsymbol{X}\beta) \\ \frac{\partial^2 l}{\partial \beta^2} &= -\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X} \end{aligned}$$

$I_n(\beta) = -E\left(\frac{\partial^2 l}{\partial \beta^2}\right) = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}$ so that $I_n^{-1}(\beta) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Because $Var(\hat{\beta}) = I^{-1}(\beta)$, $\hat{\beta}^{MLE} = \hat{\beta}^{LSE}$ is the minimum variance unbiased estimator.

Chapter 3

Model Adequacy & Regression Diagnostics

In this chapter, we will discuss the validity of the regression model. Especially, we focus on

(i) Linearity assumption

(ii) Independence assumption

(iii) Equal variance assumption

(iv) Normality assumption

(v) Leverage points

(vi) Influential points

3.1 Residuals

- Raw residual:

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

  Note that $e \sim N\left(0, (I - H_X)\sigma^2\right)$

- Standardized residual:

To make e_i have a unit variance, one may use $\dfrac{e_i}{\sigma\sqrt{1-h_{ii}}}$. The standardazed residual

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

can be obtained by replacing σ with $\hat{\sigma}$, where $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-p-1}$.

- Studentized residual:

  One may expect $\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ follows t_{n-p-1}. But this is not true because $\hat{\sigma}^2$ and e_i are not independent. The studentized residual is defined as

  $$\frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}},$$

  where $\hat{\sigma}^2_{(-i)}$ is the estimator i.e., MSE, from data without i^{th} obs. To compute $\hat{\sigma}_{(-i)}$, one might think of refitting the model to the data without the i^{th} observation. In fact, this is not necessary because

  $$\hat{\sigma}_{(-i)} = \frac{(n-p-1)\hat{\sigma^2} - e_i^2/(1-h_{ii})}{n-p-2}.$$

- PRESS residual:

  $$e_{i,-i} = Y_i - \hat{Y}_{i,-i},$$

  where $\hat{Y}_{i,-i} = X_i^T\hat{\beta}_{(-i)}$ is the estimated regression coefficient without i^{th} obs. It can be shown that $e_{i,-i} = \frac{e_i}{1-h_{ii}}$ so that we can easily compute $e_{i,-i}$ without refitting data.

- Standardized PRESS residual:

  $$\frac{e_{i,-i}}{\sqrt{Var(e_{i,-i})}} = \frac{e_i/(1-h_{ii})}{\sqrt{\sigma^2/(1-h_{ii})}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

  same as the studentized residual if replacing σ^2 with $\hat{\sigma}^2$

- Remark: some books define as follows:

  $\dfrac{e_i}{\sigma\sqrt{1-h_{ii}}}$: standardized residual
  $\dfrac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$: (Internally) studentized residual
  $\dfrac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$: (Externally) studentized residual

3.2 Residual Plots (for checking model assumptions)

- (Scaled) Residual r_i vs predicted response $\hat{Y}$ plot :

  Ideally,

  (i) no systematic pattern

  (ii) equal variance, i.e, variability of r_i seems to be constant, independent of $\hat{Y}_i$

  (iii) most $\hat{r}_i's$ fall between -2 and 2

- Normal-Quantile plot (normal probability plot):

  Plot of theoretical normal quantiles vs ordered studentized residuals

  If Q-Q plot is close to a straight line, this supports the normality of residuals, otherwise, we can say that the normality assumption is violated

- Quick remedies for violations

  (i) Residuals do not seem to have a constant variance. Especially, variance becomes larger as $\hat{Y}_i$ increases
  $\Rightarrow$ Transform Y_i into $log\ Y_i$ or $\sqrt{Y_i}$

  (ii) Residual plot show a certain pattern
  $\Rightarrow$ Transform x_i into some non-linear function of x_i, e.g.

  $$x_i' = log\ x_i, \quad e^{x_i}, \quad x_i^2, \cdots$$

  (iii) Q-Q plot shows a violation of normality
  $\Rightarrow$ It depends on a situation, but transforming Y_i into log Y_i is helpful in some cases

- Some advanced approaches:

Generalized Least Squares(GLS) regression:

when the errors do not have equal variance, or they are not independent, we may do better by slightly generalizing the least squares technique.

Let $Var(\varepsilon) = \sigma^2 V$, where V is not the identity matrix. Assume V is a positive definite matrix. Consider the following transformation of the model:

$$V^{-\frac{1}{2}}Y = V^{-\frac{1}{2}}\boldsymbol{X}\beta + V^{-\frac{1}{2}}\varepsilon$$

The least square estimator of β for the above transformed model is given by

$$\hat{\boldsymbol{\beta}}_G = (\boldsymbol{X}^T V^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T V^{-1} Y$$

Properties of GLS estimator $\hat{\boldsymbol{\beta}}_G$:

(i) $E(\hat{\boldsymbol{\beta}}_G) = \beta, \quad Var(\hat{\boldsymbol{\beta}}_G) = \sigma^2(\boldsymbol{X}^T V^{-1} \boldsymbol{X})^{-1}$

(ii) $\boldsymbol{X}\hat{\boldsymbol{\beta}}_G$ is the projection of Y on $C_{\boldsymbol{X}}$ when we endow R^n with a new norm $\| \cdot \|_{V^{-1}}$ defined by $\|u\|^2_{V^{-1}} = u^T V^{-1} u$.

Homework

1. Prove that $\| \cdot \|_{v-1}$ is a norm. (Hint: use the Cauchy- Schwarz inequality to verify this)
2. Prove (i) & (ii) in the above.

Weighted Least Squares (WLS) Regression:

Assume $V = \begin{pmatrix} \frac{1}{w_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{w_2} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & 0 \\ 0 & & \cdots & & \frac{1}{w_n} \end{pmatrix}$ is a diagonal matrix with $w_i > 0$, i.e. the error terms ε_i are uncorrelated but have unequal variances $Var(\varepsilon_i) = \frac{\sigma^2}{w_i}$.

Applying the GLS method in this special case is simply doing WLS that minimizes the weighted sum of squared errors

$$\sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

with the weight to each data point being inversely proportional to the variance of the corresponding response.

Variance stabilizing transformation:

It is considered to achieve common variance after transformation of the response. For example, if $Y|x_1, \cdots, x_p \sim Poisson\Big(\lambda(x_1, \cdots, x_p)\Big)$, it is suggested to take the square-root transformation $Y_i \to \sqrt{Y_i}$. A better way is to fit Poisson regression model, as a special case of generalized linear models. More examples will be given in Section 5.2

Box-Cox transformation:

It is considered to achieve the normality after transformation of the response:

$$Y' = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & if\ \lambda \neq 0 \\ \\ log\, Y, & if\ \lambda = 0, \end{cases}$$

where λ can be estimated from ML method

3.3 Leverage and Influence

Leverage is the i^{th} diagonal element of $H_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, that is,

$$h_{ii} = \boldsymbol{X}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i,$$

where $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1^T \\ \vdots \\ \boldsymbol{X}_n^T \end{pmatrix}$

What does h_{ii} measure?

Let us consider the simple linear regression model. Then,

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} & -\frac{\bar{x}}{S_{xx}} \\ -\frac{\bar{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{pmatrix}$$

so that

$$h_{ii} = (\mathbf{1} \quad x_i) \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} & -\frac{\bar{x}}{S_{xx}} \\ -\frac{\bar{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{pmatrix} \begin{pmatrix} \mathbf{1} \\ x_i \end{pmatrix} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

$\Rightarrow h_{ii}$ represents how far x_i is away from $\bar{x}$. In general, h_{ii} rpresents how far $\boldsymbol{X}_i$ is away from the center of $\boldsymbol{X}_i' s$

Properties of h_{ii}:

(i) $\frac{1}{n} \leq h_{ii} \leq 1$

(ii) $\sum_{i=1}^n h_{ii} = p + 1 \Rightarrow \bar{h} = \frac{1}{n}\sum_{i=1}^n h_{ii} = \frac{p+1}{n}$

(iii) $Var(\hat{Y}_i) = h_{ii}\sigma^2$

High leverage point:

If $h_{ii} > 2\bar{h} = \frac{2(p+1)}{n}$, then we call i^{th} observation "high leverage" point.
If i^{th} observation is a high leverage point,we can consider that this observation is unusual (in $\boldsymbol{X}$-space)
High leverage point is potentially dangerous for estimation of regression coefficients because a small change of the response variable corresponding to a high leverage can dramatically change the estimator. However, high leverage points are not always influential points.

Influence measure:

To see the influence of each data point, we should consider "How much would the regression results change if the i^{th} observation were deleted?"

$$\boldsymbol{X}_{(-i)} : (n-1) \times (p+1) \quad \text{design matrix without } i^{th} \text{ observation}$$

$$\begin{aligned}
\Rightarrow \hat{\beta}_{(-i)} &= \left(\boldsymbol{X}_{(-i)}^T \boldsymbol{X}_{(-i)}\right)^{-1} \boldsymbol{X}_{(-i)}^T \boldsymbol{Y} \\
&\underset{\uparrow}{=} \left(\boldsymbol{X}^T \boldsymbol{X} - \boldsymbol{X}_i \boldsymbol{X}_i^T\right)^{-1} \left(\boldsymbol{X}^T \boldsymbol{Y} - \boldsymbol{X}_i Y_i\right) \\
&\because \boldsymbol{X}^T \boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T \quad with \quad \boldsymbol{X}_i = (1, x_{i1}, \cdots, x_{ip})^T \quad \& \quad \boldsymbol{X}^T \boldsymbol{Y} = \sum_{i=1}^{n} \boldsymbol{X}_i Y_i
\end{aligned}$$

We use the formula $[A + BCB^T]^{-1} = A^{-1} - A^{-1}B(C^{-1} + B^T A^{-1} B)^{-1} B^T A^{-1}$ with taking $A = \boldsymbol{X}^T \boldsymbol{X},\ B = \boldsymbol{X}_i\ and\ C = -1$, which gives

$$\begin{aligned}
&[\boldsymbol{X}^T \boldsymbol{X} - \boldsymbol{X}_i \boldsymbol{X}_i^T]^{-1} \\
&= (\boldsymbol{X}^T \boldsymbol{X})^{-1} - (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_i (-1 + \boldsymbol{X}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i^T (\boldsymbol{X}^T \boldsymbol{X})^T \\
&= (\boldsymbol{X}^T \boldsymbol{X})^{-1} + \frac{1}{1 - h_{ii}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_i \boldsymbol{X}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\hat{\beta}_{(-i)} &= \left(\boldsymbol{X}_{(-i)}^T \boldsymbol{X}_{(-i)}\right)^{-1} \left(\boldsymbol{X}^T Y - \boldsymbol{X}_i Y_i\right) \\
&= \left[(\boldsymbol{X}^T \boldsymbol{X})^{-1} + \frac{1}{1 - h_{ii}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_i \boldsymbol{X}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1}\right] \times [\boldsymbol{X}^T Y - \boldsymbol{X}_i Y_i] \\
&= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Y - (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_i Y_i + \frac{1}{1 - h_{ii}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_{(-i)} \boldsymbol{X}_{(-i)}^T \hat{\beta} \\
&- \frac{1}{1 - h_{ii}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_i h_{ii} Y \\
&= \hat{\beta} - \left[1 + \frac{h_{ii}}{1 - h_{ii}}\right] (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_i (Y_i - \hat{Y}_i) \\
&= \hat{\beta} - \frac{1}{1 - h_{ii}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_i e_i
\end{aligned}$$

Now we can prove that the identity related to the PRESS residuals and the raw residuals, more precisely,

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}}.$$

To prove this, we have

$$
\begin{aligned}
e_{i,-i} &= Y_i - \hat{Y}_{i,-i} \\
&= Y_i - \boldsymbol{X}_i^T \hat{\beta}_{(-i)}, \quad \text{where} \quad \hat{\beta}_{(-i)} = \hat{\beta} - \frac{1}{1-h_{ii}} (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i e_i \\
&= Y_i - X_i^T \hat{\beta} + \frac{h_{ii}}{1-h_{ii}} e_i = \frac{e_i}{1-h_{ii}}
\end{aligned}
$$

- DFFITS

$$
\begin{aligned}
(\text{DFFITS})_i &= \frac{\hat{Y}_i - \hat{Y}_{i,-i}}{\hat{\sigma}_{(-i)}\sqrt{h_{ii}}}, \quad \text{where} \;\; \hat{Y}_i = \boldsymbol{X}_i^T\hat{\beta}, \; \hat{Y}_{i,-i} = \boldsymbol{X}^T \hat{\beta}_{(-i)} \\
&\underset{\uparrow}{=} \underbrace{\frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}}_{\text{studentized residual}} \times \underbrace{\sqrt{\frac{h_{ii}}{1-h_{ii}}}}_{\text{leverage measure}}
\end{aligned}
$$

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

Rule of thumb: If $|(DFFITS)_i| > 2\sqrt{\frac{p+1}{n-p-1}}$ then i^{th} observation considered to be influential.

- Cook's distance:

$$
\begin{aligned}
C_i &= \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^\top \boldsymbol{X}^\top \boldsymbol{X} (\hat{\beta} - \hat{\beta}_{(-i)})}{\hat{\sigma}^2(p+1)} : \quad \text{F-statistic-like measure} \\
&= \frac{\left[\frac{1}{1-h_{ii}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i e_i\right]^T \boldsymbol{X}^T\boldsymbol{X} \left[\frac{1}{1-h_{ii}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i e_i\right]}{\hat{\sigma}^2(p+1)} \\
&= \frac{\frac{1}{(1-h_{ii})^2}\; e_i^2 \;\boldsymbol{X}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i}{\hat{\sigma}^2(p+1)}, \quad \text{where} \quad h_{ii} : \boldsymbol{X}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i \\
&= \underbrace{\left(\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\right)^2}_{\text{(Internally) studentized residual}} \times \frac{1}{p+1} \times \underbrace{\frac{h_{ii}}{1-h_{ii}}}_{\text{leverage measure}}
\end{aligned}
$$

Rermark

(i)

$$C_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j,-i}^2)}{\hat{\sigma}^2(p+1)},$$

where $\hat{Y}_{j,-i} = \boldsymbol{X}_j^T\hat{\beta}_{(-i)}$.

(ii) In practice, if $C_i > 1$, i^{th} obs is considered to be influential

Chapter 4

Multicollinearity

4.1 Multicollinearity

-A set of predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ is said to have "multicollinearity" if there exist linear or near-linear dependencies among predictors.

-In case there exists a linear dependency among the predictors, the columns of $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ are linearly dependent, or equivalently, the centered columns $\boldsymbol{x}_1 - \bar{x}_1\mathbf{1}, \ldots, \boldsymbol{x}_p - \bar{x}_p\mathbf{1}$ are linearly dependent, so that the matrix $\boldsymbol{X}$ and $\boldsymbol{X}^\top\boldsymbol{X}$ are not of full rank.

Multicollinearity not only makes the computation of the parametric estimates erratic, but also increase the variance of the estimates

$$\sum_{j=0}^{p} Var(\hat{\beta}_j) = \mathrm{tr}(Var(\hat{\boldsymbol{\beta}})) = \sigma^2\mathrm{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}) = \sigma^2\sum_{j=0}^{p}\frac{1}{\kappa_j},$$

where κ_j's are eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$.

Let $S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ and R_j^2 denote the coefficient of determinant in regressing the jth predictor x_j on the remaining $(x_k : k \neq j)$.Then,

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2}\frac{\sigma^2}{S_{jj}}, \quad 1 \leq j \leq p$$

Proof. Assume j=1 without loss of generality. Recalling that

$$\hat{\beta}_A = (\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T(Y - \boldsymbol{X}_B\hat{\beta}_B), \quad \hat{\beta}_B = (\boldsymbol{X}_{B,\perp\!\!\!\perp}^\top\boldsymbol{X}_{B,\perp\!\!\!\perp})^{-1}\boldsymbol{X}_{B,\perp\!\!\!\perp}^\top\boldsymbol{Y}$$

in the regression $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, where

$$\beta = (\beta_A^T, \beta_B^T)^T, \quad \boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B) \quad \text{with } \boldsymbol{X}\beta = \boldsymbol{X}_A\beta_A + \boldsymbol{X}_B\beta_B, \quad \hat{\beta} = (\hat{\beta}_A^T, \hat{\beta}_B^T)^T.$$