Chapter 6. Transformation of the Linear Regression Model

6.1 The Use of Dummy Variables

6.1.1 One dummy variable case

(1) Motivating example

Problem : A company make advertisements either in the newspaper or the broadcast. They want to see the difference in the advertisement effect between the newspaper and the broadcast.

Solution : Let

$$Z = \begin{cases} 1 & \text{if newspaper} \\ 0 & \text{if broadcast} \end{cases}$$

and consider

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon,$$

where Z is called *dummy variable*(*indicator variable*). Hence,

$Z = 0: \; E[Y] = \beta_0 + \beta_1 X$

$Z = 1: \; E[Y] = (\beta_0 + \beta_2) + \beta_1 X$

Test : $H_0 : \beta_2 = 0$

Remark. If a categorical variable has c categories, then we must have $c-1$ dummy variables. For example, in this problem, we have two categories, so that we need $1(=2-1)$ dummy variable. Now, assume that we defined two dummy variables in this example, i.e.,

$$Z_1 = \begin{cases} 1 & \text{if newspaper} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if broadcast} \\ 0 & \text{otherwise} \end{cases}$$

Then, the corresponding model would be

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \varepsilon$$

then the design matrix might be, for example,

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1 & 1 & 0 \\ 1 & X_2 & 1 & 0 \\ 1 & X_3 & 1 & 0 \\ 1 & X_4 & 1 & 0 \\ 1 & X_5 & 1 & 0 \\ 1 & X_6 & 0 & 1 \\ 1 & X_7 & 0 & 1 \\ 1 & X_8 & 0 & 1 \\ 1 & X_9 & 0 & 1 \\ 1 & X_{10} & 0 & 1 \end{bmatrix}$$

which is not a full rank matrix because the sum of the 3rd and 4th columns become the 1st column. Therefore, $\boldsymbol{X}'\boldsymbol{X}$ is not invertible.

(2) Example

We have 4 types of seed; A, B, C, and D, and want to see the difference in products between seeds. Since this categorical variable has 4 categories, we must define $3(= 4 - 1)$ dummy variables. Hence, define

$$Z_1 = \begin{cases} 1 & \text{if A} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{if B} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{C} \\ 0 & \text{otherwise} \end{cases}$$

Note that seed D is given when $Z_1 = Z_2 = Z_3 = 0$. Hence, the model becomes

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \varepsilon$$

and the expectation of product of each seed is

$$\begin{aligned} A: &\quad E[Y] = \beta_0 + \beta_1 \\ B: &\quad E[Y] = \beta_0 + \beta_2 \\ C: &\quad E[Y] = \beta_0 + \beta_3 \\ D: &\quad E[Y] = \beta_0. \end{aligned}$$

For example, the difference of products between A and B is $\beta_2 - \beta_1$.

6.1.2 Extension of the use of dummy variables

(1) Interaction term

If we include the interaction term, then

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

so that

$$Z = 0: \ E[Y] = \beta_0 + \beta_1 X$$

$$Z = 1: \ E[Y] = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X$$

Here, both the slope and the intercept are different.

(2) Break (change) point

If we have a prior information on the change point, for example,

$$Z = \begin{cases} 1 & X > 20 \\ 0 & X \leq 20 \end{cases}$$

then, we can model like

$$Y = \beta_0 + \beta_1 X + \beta_2 (X - 20) Z + \varepsilon$$

so that

1. $X \leq 20 : E[Y] = \beta_0 + \beta_1 X$

2. $X > 20 : E[Y] = (\beta_0 - 20\beta_2) + (\beta_1 + \beta_2) X$

6.1.3 ANOVA using the regression model

Consider the following oneway ANOVA model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,, \quad i = 1,\, 2,\, \cdots,\, c\,;\; j = 1,\, 2,\, \cdots,\, n_i,$$

where Y_ij is the jth response of the ith group, μ is a overall effect, α_i is the effect of the ith group, and ε_{ij} is error term. For $c = 4$ and $n_i = 3$, we have

$$\boldsymbol{\beta} = (\mu,\, \alpha_1,\, \alpha_2,\, \alpha_3,\, \alpha_4)'$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

But, $\boldsymbol{X}$ is not a full rank matrix, so that it is not estimable. To overcome this difficulty, use dummy variables. Let

$$Z_1 = \begin{cases} 1 & \text{group 1} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_2 = \begin{cases} 1 & \text{group 2} \\ 0 & \text{otherwise} \end{cases}$$

$$Z_3 = \begin{cases} 1 & \text{group 3} \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \varepsilon_i\ , \quad i = 1,\ 2,\ \cdots,\ 12$$

Note that the test for $H_0 \ : \ \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ is equivalent to $H_0 \ : \ \beta_1 = \beta_2 = \beta_3 = 0$.

Ex. 6.3 (p.242)

6.2 Polynomial Regression

1st order polynomial regression

1 covariate : $Y = \beta_0 + \beta_1 X + \varepsilon$

2 covariates : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

k covariates : $Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon$

2nd order polynomial regression

1 covariate : $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

2 covariates : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon$

k covariates : $Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum \beta_{ii} X_i^2 + \sum_{i<j}^{k} \beta_{ij} X_i X_j + \varepsilon$

6.2.1 kth order polynomial regression

The kth order polynomial regression with 1 covariate is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ \cdots + \beta_k X^k + \varepsilon$$

(1) Estimation : Since the column vectors of $\boldsymbol{X}$ in the polynomial regression are highly correlated, we need to transform covariates. Among them orthogonal polynomial is often used, i.e.,

$$Y_i = \alpha_0 \phi_0(X_i) + \alpha_1 \phi_1(X_i) + \ \cdots + \alpha_k \phi_k(X_i) + \varepsilon_i,$$

where $\phi_r(X)$ is a rth degree orthogonal polynomial such that $\sum_{i=1}^n \phi_r(X_i)\phi_s(X_i) = 0,\ \forall r \neq s$. Then,

$$\hat{\alpha}_r = \frac{\sum_{i=1}^n \phi_r(X_i) Y_i}{\sum_{i=1}^n \phi_r^2(X_i)}\ , \quad r = 0,\ 1,\ \cdots,\ k$$

(2) Determination of k : Sequential test for the degree is often used, i.e,

STEP 1 : Fit $Y = \beta_0 + \beta_1 X + \varepsilon$, and test $H_0 \ : \ \beta_1 = 0$. If not rejected, then stop, i.e., $k = 1$. If rejected, go to STEP 2.

STEP 2 : Fit $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$, and test $H_0 \ : \ \beta_2 = 0$. If not rejected, then stop, i.e., $k = 2$. If rejected, go to STEP 3.

Ex. 6.4 (p.246)

6.2.2 Response surface analysis

Goal : analysis of response surface and finding the optimal condition for the response

(1) 2nd order model with 1 covariate case

Consider a fitted 2nd order model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

Then, the optimal condition can be obtained by the 1st derivative, i.e.,

$$\frac{d\hat{Y}}{dX} = \hat{\beta}_1 + 2\hat{\beta}_2 X = 0$$

and we obtain

$$X_m = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$$

which is called a stationary point and $\hat{Y}$ at this point becomes

$$\hat{Y}_m = \hat{\beta}_0 - \frac{\hat{\beta}_1^2}{4\hat{\beta}_2}$$

Also, the 2nd derivative is

$$\frac{d^2\hat{Y}}{dX^2} = 2\hat{\beta}_2.$$

Therefore, if $\hat{\beta}_2 > 0$, then $\hat{Y}_m$ is minimum, and if $\hat{\beta}_2 < 0$, then $\hat{Y}_m$ is maximum.

(2) 1st order model with 2 covariates case

Suppose that we have

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2.$$

Here, we cannot obtain stationary points X_1 and X_2 where Y is optimal, and we can only obtain the increasing (decreasing) direction. Hence, for a fixed Y, we have

$$X_2 = -\frac{\hat{\beta}_1}{\hat{\beta}_2} X_1 + c$$

Therefore, $\hat{\beta}_2$ / $\hat{\beta}_1$, the perpendicular direction of the slope, gives the maximum increase (decrease), and it is called $\textit{steepest ascent}(\textit{descent})$.

(3) 2nd order model with k covariates case

Suppose that we have

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum_{i=1}^{k} \beta_{ii} X_i^2 + \sum_{i<j}^{k} \beta_{ij} X_i X_j + \varepsilon$$

In matrix notation,

$$Y = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} + \varepsilon,$$

where $\boldsymbol{x} = (X_1,\ X_2,\ \cdots,\ X_k)'$, $\boldsymbol{\beta} = (\beta_1,\ \beta_2,\ \cdots,\ \beta_k)'$, and

$$
\boldsymbol{B} = \begin{bmatrix} \beta_{11} & \frac{\beta_{12}}{2} & \cdots & \frac{\beta_{1k}}{2} \\ \frac{\beta_{12}}{2} & \beta_{22} & \cdots & \frac{\beta_{2k}}{2} \\ \vdots & \vdots & & \vdots \\ \frac{\beta_{1k}}{2} & \frac{\beta_{2k}}{2} & \cdots & \beta_{kk} \end{bmatrix}
$$

Let the fitted model is

$$
\hat{Y} = \hat{\beta}_0 + \boldsymbol{x}'\hat{\boldsymbol{\beta}} + \boldsymbol{x}'\hat{\boldsymbol{B}}\boldsymbol{x}
$$

and the 1st derivative gives

$$
\frac{\partial \hat{Y}}{\partial \boldsymbol{x}} = \hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{B}}\boldsymbol{x} = \boldsymbol{0}
$$

so that the stationary point is

$$
\boldsymbol{x}_s = -\frac{1}{2}\hat{\boldsymbol{B}}^{-1}\hat{\boldsymbol{\beta}}
$$

The analysis of response at the stationary point is called *canonical analysis*, and two methods, eigenvalue analysis and contour plot, are often used for the canonical analysis. Now, the fitted value at the stationary point is

$$
\hat{Y}_s = \hat{\beta}_0 + \boldsymbol{x}_s'\hat{\boldsymbol{\beta}} + \boldsymbol{x}_s'\hat{\boldsymbol{B}}\boldsymbol{x}_s
$$

and let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{x}_s$. Then, we have

$$\begin{aligned}
\hat{Y} &= \hat{\beta}_0 + (\boldsymbol{x}_s + \boldsymbol{w})'\hat{\boldsymbol{\beta}} + (\boldsymbol{x}_s + \boldsymbol{w})'\hat{\boldsymbol{B}}(\boldsymbol{x}_s + \boldsymbol{w}) \\
&= \hat{\beta}_0 + \boldsymbol{x}_s'\hat{\boldsymbol{\beta}} + \boldsymbol{x}_s'\hat{\boldsymbol{B}}\boldsymbol{x}_s + \boldsymbol{w}'\hat{\boldsymbol{\beta}} + 2\boldsymbol{w}'\hat{\boldsymbol{B}}\boldsymbol{x}_s + \boldsymbol{w}'\hat{\boldsymbol{B}}\boldsymbol{w} \\
&= \hat{Y}_s + \boldsymbol{w}'\hat{\boldsymbol{B}}\boldsymbol{w} + \boldsymbol{w}'(\hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{B}}\boldsymbol{x}_s) \\
&= \hat{Y}_s + \boldsymbol{w}'\hat{\boldsymbol{B}}\boldsymbol{w}
\end{aligned}$$

Therefore, if $\hat{\boldsymbol{B}}$ is positive definite, then $\hat{Y}_s$ is minimum, and if $\hat{\boldsymbol{B}}$ is negative definite, then $\hat{Y}_s$ is maximum. Let $\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_k$ be eigenvalues of $\hat{\boldsymbol{B}}$. Then, we have

1. If all the $\lambda_j's$ are positive, then $\hat{Y}_s$ is minimum.
2. If all the $\lambda_j's$ are negative, then $\hat{Y}_s$ is maximum.
3. If some $\lambda_j's$ are positive and some $\lambda_j's$ are negative, then $\hat{Y}_s$ is neither minimum or maximum. The corresponding $\boldsymbol{x}_s$ is called a *saddle point*.

Fig. 6.5 (p. 250)

Ex. 6.5 (p.251)

6.3 Weighted Least Squares Method

So far we assumed $Cov(\boldsymbol{\varepsilon}) = \boldsymbol{I}\sigma^2$ which is called *homogeneity* assumption.

If the homogeneity assumption is false, then we often assume that

$$Cov(\varepsilon) = \sigma^2 \boldsymbol{W}^{-1},$$

where $\boldsymbol{W} = \text{diag}(w_1, \ \cdots, \ w_n)$, and it is called *heteroscedasticity*. Therefore, $Var(\varepsilon_i) = \sigma^2 / w_{ii}$. Now, there exists a non-singular matrix $\boldsymbol{P}$ s.t.

$$\boldsymbol{P}^2 = \boldsymbol{W}.$$

Let $\boldsymbol{\delta} = \boldsymbol{P\varepsilon}$, then $E(\boldsymbol{\delta}) = 0$ and

$$Cov(\boldsymbol{\delta}) = E(\boldsymbol{\delta}\boldsymbol{\delta}') = E(\boldsymbol{P\varepsilon\varepsilon}'\boldsymbol{P}) = \boldsymbol{PWP}\sigma^2 = \boldsymbol{I}\sigma^2$$

Now, in $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, multiply $\boldsymbol{P}$ on both sides, then

$$\boldsymbol{Py} = \boldsymbol{PX\beta} + \boldsymbol{P\varepsilon}$$

and let $\boldsymbol{Py} = \boldsymbol{z}, \ \boldsymbol{PX} = \boldsymbol{Q}$, then

$$\boldsymbol{z} = \boldsymbol{Q\beta} + \boldsymbol{\delta}$$

Hence, the normal equation becomes

$$\boldsymbol{Q}'\boldsymbol{Q}\hat{\boldsymbol{\beta}} \ = \boldsymbol{Q}'\boldsymbol{y}$$

and

$$X'WX\hat{\boldsymbol{\beta}} = X'W\boldsymbol{y}$$

gives

$$\hat{\boldsymbol{\beta}} = (X'WX)^{-1}X'W\boldsymbol{y}.$$

The above method is called *weighted* least squares method.

Ex. 6.6 (p.254)

6.4 Box-Cox Transformation Model

So far we assumed the normality of response, i.e. $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \sigma^2\boldsymbol{I})$. Hence, when the normality assumption is doubtful, Box and Cox (1964) suggested the followings; Box and Cox argued that there exists λ s.t.

$$w_i = \begin{cases} ({y_i}^{\lambda}-1)\ /\ \lambda, & \lambda \neq 0 \\ log\ y_i, & \lambda = 0 \end{cases}$$

is approximately normal. In matrix notation,

$$\boldsymbol{w}(\lambda) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

which is called $Box - Cox\ transformation\ model$. Here, we need to estimate λ in addition to $\boldsymbol{\beta}$ and σ^2, and we will obtain the MLEs for them. Now, the likelihood function is

$$L = (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{w} - \boldsymbol{X\beta})'(\boldsymbol{w} - \boldsymbol{X\beta})\right] J(\lambda)$$

where

$$J(\lambda) = \prod_{i=1}^{n} \frac{\partial w_i}{\partial y_i} = \prod_{i=1}^{n} y_i^{\lambda-1}$$

is the Jacobian from transforming y to $w(\lambda)$. To get the MLE of λ, $\boldsymbol{\beta}$, and σ^2, we first fix λ, and get the MLE for $\boldsymbol{\beta}$ and σ^2. Then,

$$\hat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{w}(\lambda),\; s^2(\lambda) = \boldsymbol{w}(\lambda)'(\boldsymbol{I} = \boldsymbol{H})\boldsymbol{w}(\lambda)/n$$

Now, we replace $\hat{\boldsymbol{\beta}}(\lambda)$ and $s^2(\lambda)$ in the likelihood, then the likelihood becomes a function of λ only, i.e.,

$$\begin{aligned} l(\lambda; \hat{\boldsymbol{\beta}}(\lambda),\, s^2(\lambda)) &= -\frac{n}{2} log\, s^2(\lambda) + log\; J(\lambda) \\ &= -\frac{n}{2} log\, s^2(\lambda) + (\lambda - 1)\sum log y_i \end{aligned}$$

Now, to get the MLE of λ, we need to solve $\partial l(\lambda; \hat{\boldsymbol{\beta}}(\lambda)/\partial\lambda = 0$, however, it does not give an explicit solution because it is nonlinear in λ. In this case, we are only to obtain the numerical result for the MLE through the grid search, say.

Ex. 6.7 (p.259)

6.5 Robust Regression

The distribution of error terms might have heavier tails than the normal distribution. For example, the double exponential (Laplace) distribution whose pdf is given by

$$f(\varepsilon) = \frac{1}{2\sigma} e^{-|\varepsilon|/\sigma}, \quad -\infty < \varepsilon < \infty$$

Also, the likelihood function becomes

$$\frac{1}{(2\sigma)^n} \exp[-\sum |\ y_i - \boldsymbol{x}_i' \boldsymbol{\beta}\ |\ /\ \sigma]$$

Now, to estimate $\boldsymbol{\beta}$, maximizing the likelihood is equivalent to minimizing $\sum |\ y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\ |$, which is called L_1-norm regression. In the normal distribution, we minimize $\sum(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$, which is $\boldsymbol{L}_2$-norm regression. In general, $\boldsymbol{L}_2$-norm regression minimizes $\sum(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^p$, $1 \le p \le 2$.

Fig. 6.10 (p. 261)

6.5.1 M-Estimation

For some function $\rho()$, if $\boldsymbol{\beta}$ is obtained by

$$min \sum_{i=1}^{n} \rho(\varepsilon_i) = min \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$$

then it is called M-estimator. If $\rho(u) = \frac{1}{2}u^2$, then is it is $\boldsymbol{L}_2$-norm regression. Also, if $\rho(u) = |u|$, then is it is $\boldsymbol{L}_1$-norm regression.

Now, to obatin the robust estimator of $\boldsymbol{\beta}$, we consider

$$min \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{s}\right) = min \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{s}\right)$$

where s is a robust estimator of σ, and it is given by

$$s = median|e_i - median(e_i)| \ / \ 0.6745$$

We can show that s is an unbiased estimator of σ if n is large and error terms have normal distribution. To obtain the minimum, we take 1st derivative of ρ w.r.t. $\beta_j, \ \ j = 0, \ 1, \ \cdots, \ p{-}1$. Then, we get

$$\sum_{i=1}^{n} x_{ij} \ \psi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{s}\right) = 0, \quad j = 0, \ 1, \ \cdots, \ p{-}1,$$

where x_{ij} is the jth component of $\boldsymbol{x}_i$ and $\psi = \rho'$. Since $|psi$ is nonlinear in $\boldsymbol{\beta}$, we cannot get explicit solution. Instead, we use an iterative method, called IRLS (Iterative Reweighted Least Squares). Let

$$w_{i\beta} = \frac{\psi\{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/s\}}{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/s}, \quad i = 1, \ 2, \ \cdots, \ n$$

and let $w_{i\beta} = 1$ if $y_i - \boldsymbol{x}_i'\boldsymbol{\beta} = 0$. Then, we may write

$$\sum_{i=1}^{n} x_{ij} w_{i\beta}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) = 0, \quad j = 0, \ 1, \ \cdots, \ p-1$$

and it is equivalent to

$$\sum_{i=1}^{n} x_{ij} w_{i\beta} \boldsymbol{x}_i'\boldsymbol{\beta} = \sum_{i=1}^{n} x_{ij} w_{i\beta} y_i, \quad j = 0, \ 1, \ \cdots, \ p-1.$$

In matrix notation, we have

$$\boldsymbol{X}'\boldsymbol{W}_\beta\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{W}_\beta\boldsymbol{y},$$

where $\boldsymbol{W}_\beta = \mathrm{diag}(w_{1\beta},\ w_{2\beta},\ \cdots,\ w_{n\beta})$. But, $\boldsymbol{W}_\beta$ contains unknown $\boldsymbol{\beta}$, $(\boldsymbol{X}'\boldsymbol{W}_\beta\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_\beta\boldsymbol{y}$ cannot be a solution of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_0$ be an initial value of $\boldsymbol{\beta}$, and replace $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_0$, then let $\hat{\beta}_1 = (\boldsymbol{X}'\boldsymbol{W}_{\hat{\beta}_0}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_{\hat{\beta}_0}\boldsymbol{y}$. We continue this process, i.e.,

$$\hat{\boldsymbol{\beta}}_{q+1} = (\boldsymbol{X}'\boldsymbol{W}_{\hat{\beta}_q}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_{\hat{\beta}_q}\boldsymbol{y}, \quad q = 0,\ 1,\ 2,\ \cdots$$

which is called IRLS.

6.5.2 Influence function

If let $u = (y - \boldsymbol{x}'\boldsymbol{\beta})/s$, then $w(u) = \{\partial\rho(u)/\partial u\}/u,\ u = (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})/s$, so that $\psi(u) = \partial\rho(u)/\partial u$, which is called $influence\ function$.

Table 6.8 (p. 264)

Fig. 6.11 (p. 265)

Ex. 6.8 (p. 265)

6.6 Inverse Regression

In the most cases of regression analysis, we predict response y at some $X = x$, however, we sometimes predict x_0 for a given response y_0. This problem is called *inverse regression* (*calibration, discrimination*).

Consider a simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

The estimate of x_0 for a given y_0 is

$$\hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1}.$$

Also, the $100(1-\alpha)\%$ C.I. for x_0 is obtained by solving

$$\frac{(y_0 - \hat{\beta}_0 - \hat{\beta}_1 x)^2}{s^2\ A^2} \leq t^2,$$

where

$$A^2 = 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}.$$

So, if we let $d = x - \bar{x}$, then the inequality becomes

$$d^2 \left\{ \hat{\beta}_1^2 - \frac{t^2 s^2}{\sum(x_i - \bar{x})^2} \right\} - 2d\hat{\beta}_1(y_0 - \bar{y}) + \left\{ (y_0 - \bar{y})^2 - t^2 s^2 (1 + \frac{1}{n}) \right\} = 0.$$

Let $d_1,\ d_2$ be solutions for d, and we have $X_L \leq X_0 \leq X_U$, where $X_L = \bar{X} + d_1,\ X_U = \bar{X} + d_2$.

Note that there are four types of solution for the 2nd degree polynomial inequality; $(a,\ b)$, $(-\infty,\ a)$, $(b,\ \infty)$, $(-\infty,\ \infty)$. Among them, only $(a,\ b)$ is meaningful. Recall that the solution $(a,\ b)$ can be obtained only when $a_2 > 0$ in $a_2x^2 + a_1x + a_0 < 0$. Note that $a_2 > 0$ is equivalent to $\hat{\beta}_1^2 - \frac{t^2s^2}{\sum(x_i - \bar{x})} > 0$ which is the rejection region for $H_0 : \beta_1 = 0$.

Ex. 6.9 (p.268)