8. Regression Problems

Linear regression model

- Let $x_1 < x_2 < \cdots < x_n$ be a set of n ordered numbers and Y_i a random variable related to x_i, $i = 1, \ldots, n$.
- Traditionally, a linear regression model assumes the mean of Y_i to be a linear function of x_i with a common slope β and intercept α for all i:

$$\mathrm{E}[Y_i] = \alpha + \beta x_i, \quad i = 1, \ldots, n. \tag{8.1}$$

- Equivalently, a linear regression model assumes

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n, \tag{8.2}$$

  where e_i is a random variable (called *random error*) with $\mathrm{E}[e_i] = 0$.
- In the standard (parametric) linear regression model, $e_1, \ldots, e_n$ are assumed to be i.i.d. with the $N(0, \sigma^2)$ distribution.
- We now consider model (8.2) with nonparametric random errors $e_1, \ldots, e_n$.

8.1 Nonparametric inference of linear regression

In model (8.2), $Y_1,\ldots,Y_n$ are called the *response variables*, and $x_1,\ldots,x_n$ are fixed points called the *independent variables*, *predictors*, *regressors* or *covariates*.

Assumption 8.1

(i) The response variables $Y_1,\ldots,Y_n$ are related to $x_1 < \cdots < x_n$ by model (8.2).

(ii) The random errors $e_1,\ldots,e_n$ in (8.2) are i.i.d. continuous random variables with median 0.

Theil test of the slope

We begin with a nonparametric test of the slope β in model (8.2)

Null hypothesis: $H_0 : \beta = \beta_0$, where β_0 is a specified known value. A special case of interest for β_0 is $\beta_0 = 0$, which indicates that the values of $x_1,\ldots,x_n$ have no effects on the response variables $Y_1,\ldots,Y_n$.

Alternative hypotheses: $H_1 : \beta > \beta_0$, $\beta < \beta_0$ or $\beta \neq \beta_0$.

Test statistic: Let

$$D_i = Y_i - \beta_0 x_i, \;\; i = 1, \ldots, n, \;\; \text{and} \;\; c(a) = \begin{cases} -1 & \text{if } a < 0 \\ 0 & \text{if } a = 0 \\ 1 & \text{if } a > 0 \end{cases} \tag{8.3}$$

Then the *Theil statistic* C is defined by

$$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c(D_j - D_i) = \sum_{1 \le i < j \le n} c(D_j - D_i) \tag{8.4}$$

and its standardized version is

$$\bar{C} = \frac{2C}{n(n-1)} \tag{8.5}$$

Distribution of C: Assume there are no ties among $D_1, \ldots, D_n$. As $x_1 < \cdots < x_n$, $(x_j - x_i)(D_j - D_i) > 0 \;\Leftrightarrow\; D_j - D_i > 0$ for $i < j$. Thus $c(D_j - D_i) = Q_{ij}$ in (7.5) with D_i in place of Y_i. As a result, the Theil statistic C has the same distribution as the Kendall statistic K in (7.6) under H_0. This leads to the following results.

Mean and variance of C: By (7.21) and (7.26), under H_0,

$$\mathrm{E}_0[C] = \mathrm{E}_0[K] = 0 \quad \text{and} \quad \mathrm{Var}_0(C) = \mathrm{Var}_0(K) = \frac{n(n-1)(2n+5)}{18} \tag{8.6}$$

Asymptotic distribution of C:

$$C^* = \frac{C - \mathrm{E}_0[C]}{\sqrt{\mathrm{Var}_0(C)}} = \frac{C}{\sqrt{n(n-1)(2n+5)/18}} \to_d N(0,1) \text{ as } n \to \infty \tag{8.7}$$

Rejection rule: The Theil test for the slope has the following rejection rules at level α based on the standardized Theil statistic $\overline{C}$ in (8.5):

- Reject $H_0 : \beta = \beta_0$ for $H_1 : \beta > \beta_0$ if $\overline{C} \geq k_\alpha$;
- Reject $H_0 : \beta = \beta_0$ for $H_1 : \beta < \beta_0$ if $\overline{C} \leq -k_\alpha$;
- Reject H_0 for $H_1 : \beta \neq \beta_0$ if $|\overline{C}| \geq k_{\alpha/2}$,

where k_α is the critical point of the Kendall test for independence.

Approximate rejection rule

The approximate rules to test H_0 at level α are as follows:

- Reject $H_0 : \beta = \beta_0$ for $H_1 : \beta > \beta_0$ if $C^* \geq z_\alpha$;
- Reject $H_0 : \beta = \beta_0$ for $H_1 : \beta < \beta_0$ if $C^* \leq -z_\alpha$;
- Reject H_0 for $H_1 : \beta \neq \beta_0$ if $|C^*| \geq z_{\alpha/2}$.

Ties: If there are ties among $D_1, \ldots, D_n$, formula (8.4) for the Theil statistic C remains valid with $c(0) = 0$ and the above rejection rules can still be applied, but the level α will be approximate.

Remark 8.1 The above test is *distribution-free*, or *nonparametric* with respect to $e_1, \ldots, e_n$, and hence $Y_1, \ldots, Y_n$, not to the relationship between x_i and Y_i in (8.2), which is *linear* – a special parametric form. The term *nonparametric regression* also includes a nonparametric (unspecified) relationship between x_i and Y_i. We will briefly discuss this issue later.

Example 8.1 Example 9.1 of the textbook (page 454) presents a case study on the effect of cloud seeding on rainfall. The data below include the measures (by the *double ratio*) of rainfall $Y_1,\ldots,Y_5$ over 5 years $(x_1,\ldots,x_5)=(1,2,3,4,5)$.

Years seeded x_i	1	2	3	4	5
Double ratio Y_i	1.26	1.27	1.12	1.16	1.03

Test $H_0:\beta=0$ (rainfall does not change with time) against $H_1:\beta<0$ (rainfall decreases over time). Under $H_0:\beta=0$, $D_i=Y_i$ and so $D_j-D_i=Y_j-Y_i$.

The values of $D_j-D_i=Y_j-Y_i$ and $c(D_j-D_i)$ for $i<j$ are shown below:

(i,j)	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
D_j-D_i	0.01	–0.14	–0.10	–0.23	–0.15	–0.11	–0.24	0.04	–0.09	–0.13
$c(D_j-D_i)$	1	–1	–1	–1	–1	–1	–1	1	–1	–1

By (8.4),

$$C=\sum_{1\le i<j\le n} c(D_j-D_i)=2-8=-6 \;\Rightarrow\; \bar{C}=\frac{2C}{n(n-1)}=\frac{2(-6)}{5(4)}=\frac{-12}{20}=-0.6$$

With $n=5$, $n(n-1)/2=5(4)/2=10$ and $n!=5!=120$. Hence $\Pr(K=k)$ for the Kendall statistic K at $k=10,8,6$ can be calculated as follows:

k	$k=10$	$k=9-1=8$	$k=8-2=6$
Permutation of 12345	12345	21345, 13245 12435, 12354	23145, 21435, 21354, 31245, 13425 13254, 14235, 12453, 12534
$\Pr(K=k)$	$1/120$	$4/120$	$9/120$

Since $C \sim K$, the p-value of the test is

$$\Pr(\bar{C} \le -0.6) = \Pr(C \le -6) = \Pr(K \ge 6) = \frac{1+4+9}{120} = \frac{14}{120} = 0.117$$

The approximate p-value by (8.7) is

$$\Pr\left(C^* \le \frac{-6}{\sqrt{5(5-1)(2\times 5+5)/18}} = -6\sqrt{\frac{3}{50}} = -1.47 \right) \approx 0.071$$

(not close enough to the exact p-value 0.117 due to small sample size).

The p-values point to some weak evidence for $\beta < 0$.

Estimation of the slope: Let

$$N = \text{Number of pairs } \{(i,j): 1 \le i < j \le n\} = \frac{n(n-1)}{2}$$

Define

$$S_{ij} = \frac{Y_j - Y_i}{x_j - x_i}, \quad 1 \le i < j \le n, \tag{8.8}$$

and order the values of $\{S_{ij}: 1 \le i < j \le n\}$ as

$$S_{(1)} \le S_{(2)} \le \cdots \le S_{(N)}$$

Then the slope β can be estimated by

$$\hat{\beta} = \text{median}\left\{S_{ij}, 1 \le i < j \le n\right\} = \begin{cases} S_{((N+1)/2)} & \text{if } N \text{ is odd;} \\ \dfrac{S_{(N/2)} + S_{(N/2+1)}}{2} & \text{if } N \text{ is even.} \end{cases} \tag{8.9}$$

Confidence interval of the slope: Let

$$C_\alpha = Nk_{\alpha/2} - 2, \quad M = \frac{N - C_\alpha}{2} \quad \text{and} \quad Q = \frac{N + C_\alpha}{2} = M + C_\alpha \tag{8.10}$$

Define

$$D_i = Y_i - \beta x_i, \quad i = 1, \ldots, n,$$

with the true slope β of the regression line,

$$C = \sum_{1 \le i < j \le n} c(D_j - D_i) \quad \text{and} \quad C^+ = \text{No.}\{i < j : c(D_j - D_i) = 1\}$$

Then $C = C^+ - (N - C^+)$ (if no ties) $\Rightarrow$ $2C^+ - N = C \sim K$ in (the Kendall statistic) under the null hypothesis of independence, and so $2C^+ \sim K + N$.

Moreover,

$$S_{ij} = \frac{Y_j - Y_i}{x_j - x_i} = \frac{D_j - D_i}{x_j - x_i} + \beta \Rightarrow \{S_{ij} > \beta\} = \{D_j > D_i\} = \{c(D_j - D_i) = 1\}$$

Hence $\text{No.}\{i<j : S_{ij} > \beta\} = C^+$ and so for $m \in \{1,\ldots,N\}$,

$$S_{(m)} > \beta \;\Leftrightarrow\; S_{(k)} > \beta \;\text{ for }\; k = m, m+1, \ldots, N \;\Leftrightarrow\; C^+ \geq N - m + 1$$

Note also that

$$2M = N - C_\alpha \quad \text{and} \quad 2Q = N + C_\alpha$$

Consequently,

$$\Pr\left(S_{(M)} > \beta\right) = \Pr\left(2C^+ \geq 2(N - M + 1)\right) = \Pr\left(C + N \geq 2N - 2M + 2\right)$$

$$= \Pr\left(K + N \geq 2N - (N - C_\alpha) + 2\right) = \Pr\left(K \geq C_\alpha + 2\right)$$

$$= \Pr\left(K \geq N k_{\alpha/2}\right) = \Pr\left(\bar{K} \geq k_{\alpha/2}\right) = \alpha/2$$

Furthermore,

$$S_{(Q+1)} > \beta \;\Leftrightarrow\; C^+ \geq N - (Q+1) + 1 = N - Q \;\Leftrightarrow\; 2C^+ \geq 2N - 2Q$$

$$\Leftrightarrow\; C = 2C^+ - N \geq 2N - 2Q - N = N - 2Q$$

This together with the symmetry of $C \sim K$ about 0 and $K \le k \Leftrightarrow K < k+2$ for integer k (if no ties) show that

$$\begin{aligned}\Pr\left(S_{(Q+1)} > \beta\right) &= \Pr\left(K \ge N - 2Q\right) = \Pr\left(K \le 2Q - N\right) = \Pr\left(K < 2Q - N + 2\right) \\ &= 1 - \Pr\left(K \ge 2Q - N + 2\right) = 1 - \Pr\left(K \ge N + C_\alpha - N + 2\right) \\ &= 1 - \Pr\left(K \ge C_\alpha + 2\right) = 1 - \Pr\left(K \ge N k_{\alpha/2}\right) = 1 - \alpha/2\end{aligned}$$

It follows that

$$\Pr\left(S_{(M)} < \beta < S_{(Q+1)}\right) = \Pr\left(S_{(Q+1)} > \beta\right) - \Pr\left(S_{(M)} > \beta\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

Thus a $100(1-\alpha)\%$ confidence interval for β is given by

$$(\beta_L, \beta_U) = \left(S_{(M)}, S_{(Q+1)}\right) \tag{8.11}$$

For large n, C_α can be approximated by

$$C_\alpha \approx z_{\alpha/2}\sqrt{\mathrm{Var}_0(C)} = z_{\alpha/2}\sqrt{\frac{n(n-1)(2n+5)}{18}} \tag{8.12}$$

Example 8.2 In Example 8.1, $n = 5 \Rightarrow N = 5(4)/2 = 10$ and the ordered values $S_{(1)} \le \cdots \le S_{(10)}$ of $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$ are given by

$$-0.15, -0.13, -0.08, -0.07, -0.0575, -0.055, -0.045, -0.033, 0.01, 0.04$$

Hence an estimate of β is given by

$$\hat{\beta} = \frac{S_{(5)} + S_{(6)}}{2} = \frac{-0.0575 - 0.055}{2} = -0.0563$$

From Example 8.1 we get

$$\Pr(K \ge 8) = \frac{1+4}{120} = \frac{5}{120} = \frac{1}{24} = 0.04167 \Rightarrow Nk_{1/24} = 8$$

Take $\alpha = 2/24 = 0.0833$. Then $1 - \alpha = 22/24 = 0.9167$ and by (8.10),

$$C_\alpha = Nk_{\alpha/2} - 2 = Nk_{1/24} - 2 = 8 - 2 = 6 \Rightarrow$$

$$M = \frac{N - C_\alpha}{2} = \frac{10 - 6}{2} = 2 \quad \text{and} \quad Q = M + C_\alpha = 2 + 6 = 8$$

Thus an exact 91.67% confidence interval of β is given by

$$(\beta_L, \beta_U) = (S_{(M)}, S_{(Q+1)}) = (S_{(2)}, S_{(9)}) = (-0.13, 0.01)$$

If we use the large-sample approximation, then by (8.12),

$$C_{0.10} \approx z_{0.05}\sqrt{\frac{5(n-1)(2n+5)}{18}} = 1.645\sqrt{\frac{5(4)(15)}{18}} = 6.72$$

Take

$$M = \frac{N - C_{0.05}}{2} \approx \frac{10 - 6.72}{2} = 1.64 \approx 2 \quad \text{and} \quad Q \approx \frac{10 + 6.72}{2} = 8.36 \approx 8$$

Then an approximate 90% confidence interval of β is also given by

$$(\beta_L, \beta_U) = (S_{(M)}, S_{(Q+1)}) = (S_{(2)}, S_{(9)}) = (-0.13, 0.01)$$

As shown above, the exact level of this confidence interval is 91.67%.

These results can also be obtained by R.

Estimation of the intercept: After $\hat{\beta}$ is calculated, let

$$A_i = Y_i - \hat{\beta} x_i, \quad i = 1, \ldots, n, \tag{8.13}$$

and $A_{(1)} \le \cdots \le A_{(n)}$ be ordered values of $A_1, \ldots, A_n$. Then the intercept α can be estimated by

$$\hat{\alpha} = \text{median}\{A_1, \ldots, A_n\} = \begin{cases} A_{((n+1)/2)} & \text{if } n \text{ is odd;} \\ \dfrac{A_{(n/2)} + A_{(n/2+1)}}{2} & \text{if } n \text{ is even.} \end{cases} \tag{8.14}$$

The median $m_Y(x^*)$ of Y when $x = x^*$ can be estimated (or *predicted*) by

$$\hat{m}_Y(x^*) = \hat{\alpha} + \hat{\beta} x^* \tag{8.15}$$

It should be noted that the prediction in (8.15) relies on the linear relationship in (8.2), and is subject to large error if x^* is far away from the domain of $x_1, \ldots, x_n$ (since the linear relationship may not hold for such x^*).

Example 8.3 For the data in Example 8.1, we have calculated $\hat{\beta} = -0.0563$ in Example 8.2. Hence by (8.13),

$$A_1 = Y_1 - \hat{\beta} x_1 = 1.26 - (-0.0563) = 1.3162, \quad A_2 = 1.27 - 2(-0.0563) = 1.3824,$$

and similarly, $A_3 = 1.2889$, $A_4 = 1.3852$, $A_5 = 1.3115$.

The ordered values of $A_1, \ldots, A_5$ are

$$\left(A_{(1)}, \ldots, A_{(5)}\right) = (1.2889,\ 1.3115,\ 1.3163,\ 1.3826,\ 1.3852)$$

By (8.14), $\hat{\alpha} = A_{(3)} = 1.3163$.

The median of the double ratio at the middle of year 4 ($x = 4.5$) is estimated by

$$\hat{m}_Y(4.5) = \hat{\alpha} + \hat{\beta}(4.5) = 1.3163 - 0.0563(4.5) = 1.06295$$

The median of the double ratio in year 6 ($x = 6$) is predicted by

$$\hat{m}_Y(6) = \hat{\alpha} + \hat{\beta}(6) = 1.3163 - 0.0563(6) = 0.9785$$

Test of equal slopes between regression lines

We now consider two or more regression lines:

$$\text{Line } i: \qquad Y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij}, \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,k. \tag{8.16}$$

The total sample size is $N = n_1 + \cdots + n_k$, $k \geq 2$.

The question of interest is whether the slopes of the k regression lines are equal, in other words, whether the k regression lines are parallel or not.

An equal slope of the k regression lines indicates the same effect of the regressors on the response variables in all k regression equations.

Assumption 8.2

(i) The response variables $Y_{i1},\ldots,Y_{in_i}$ are related to fixed points $x_{i1} < \cdots < x_{in_i}$ by model (8.16) for $i = 1,\ldots,k$.

(ii) The random errors $\{e_{ij} : j = 1,\ldots,n_i; i = 1,\ldots,k\}$ in (8.16) are i.i.d. continuous random variables with a common cdf F.

Hypotheses:

$$H_0 : \beta_1 = \cdots = \beta_k \text{ against } H_1 : \beta_1, \ldots, \beta_k \text{ are not all equal.} \qquad (8.17)$$

Test statistic: Let

$$\overline{\beta} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i) Y_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2} \qquad (8.18)$$

where

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad i = 1, \ldots, k.$$

The $\overline{\beta}$ in (8.18) is the least square estimate of the common slope β under the null hypothesis H_0. More details are provided in Appendix.

Define *aligned observations*:

$$Y_{ij}^* = Y_{ij} - \overline{\beta} x_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k. \qquad (8.19)$$

Order $Y_{i1}^*,\ldots,Y_{in_i}^*$ increasingly (assuming no ties) and let r_{ij}^* denote the rank of Y_{ij}^* among $Y_{i1}^*,\ldots,Y_{in_i}^*$. Then compute

$$C_i^2 = \sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 = \sum_{j=1}^{n_i}x_{ij}^2 - n_i\bar{x}_i^2 \quad \text{and} \tag{8.20}$$

$$T_i^* = \frac{1}{n_i+1}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)r_{ij}^*, \quad i=1,\ldots,k. \tag{8.21}$$

The *Sen-Adichie statistic* V for testing H_0 against H_1 in (8.17) is defined by

$$V = 12\sum_{i=1}^{k}\left(\frac{T_i^*}{C_i}\right)^2 = 12\sum_{i=1}^{k}\frac{(T_i^*)^2}{C_i^2} \tag{8.22}$$

Asymptotic rejection rule: Reject H_0 if $V \geq \chi^2_{k-1,\alpha}$.

Ties: If there are ties among $Y_{i1}^*,\ldots,Y_{in_i}^*$, assign average ranks to tied values for T_i^* in (8.21) and V in (8.22). Then the above rejection rule remains valid.

Example 8.4 Example 9.5 of the textbook (page 468) provides the following data for 4 regression lines, each with 5 pairs of (x, Y):

x_{1j}	Y_{1j}	x_{2j}	Y_{2j}	x_{3j}	Y_{3j}	x_{4j}	Y_{4j}
0	0	0	0	0	0	0	0
1.5	33.019	1.5	131.831	1.5	33.351	1.5	8.959
3	111.314	3	181.603	3	97.463	3	105.384
4.5	196.205	4.5	230.070	4.5	196.615	4.5	211.392
6	230.658	6	258.119	6	217.308	6	255.105

By (8.18) and (8.20), it is easy to calculate for:

$$\overline{x}_i = \frac{1}{5}(0+1.5+3+4.5+6) = 3 \quad \text{and} \tag{8.23}$$

$$C_i^2 = 1.5^2 + 3^2 + 4.5^2 + 6^2 - 5\times 3^2 = 22.5, \;\; i = 1,2,3,4. \tag{8.24}$$

It follows that

$$\sum_{i=1}^{4}\sum_{j=1}^{5}(x_{ij} - \overline{x}_i)^2 = \sum_{i=1}^{4} C_i^2 = 4(22.5) = 90 \tag{8.25}$$

Next, since $Y_{i1} = 0$, $x_{ij} = x_{1j}$ for $i = 1,\ldots,4$, $j = 1,\ldots,5$, and $x_{13} - \bar{x}_1 = 3 - 3 = 0$,

$$\sum_{i=1}^{4}\sum_{j=1}^{5}(x_{ij} - \bar{x}_i)Y_{ij} = \sum_{j=1}^{5}\sum_{i=1}^{4}(x_{ij} - \bar{x}_i)Y_{ij} = \sum_{j=1}^{5}(x_{1j} - \bar{x}_1)\sum_{i=1}^{4}Y_{ij}$$

$$\begin{aligned} &= 0 + (1.5-3)(33.019 + 131.831 + 33.351 + 8.959) + 0 \\ &\quad + (4.5-3)(196.205 + 230.070 + 196.615 + 211.292) \\ &\quad + (6-3)(230.658 + 258.119 + 217.308 + 255.105) \\ &= 3824.253 \end{aligned} \tag{8.26}$$

Thus by (8.18) and (8.25) – (8.26),

$$\bar{\beta} = \frac{3824.253}{90} = 42.49 \tag{8.27}$$

Then it follows from (8.19) that

$Y_{11}^* = 0$, $Y_{12}^* = 33.019 - 42.49(1.5) = -30.716$, $Y_{13}^* = 111.314 - 42.49(3) = -16.156$,

$Y_{14}^* = 196.205 - 42.49(4.5) = 5$ and $Y_{15}^* = 230.658 - 42.49(6) = -24.282$

Thus $Y_{12}^* < Y_{15}^* < Y_{13}^* < Y_{11}^* < Y_{14}^* \Rightarrow (r_{11}^*, r_{12}^*, r_{13}^*, r_{14}^*, r_{15}^*) = (4,1,3,5,2)$ and by (8.21),

$$T_1^* = \frac{(0-3)(4)+(1.5-3)(1)+(3-3)(3)+(4.5-3)(5)+(6-3)(2)}{5+1}$$

$$= \frac{-12-1.5+0+7.5+6}{6} = 0$$

Similarly to calculate $T_2^* = 0$, $T_3^* = -0.75$ and $T_4^* = 1.5$.

It then follows from (8.22) and (8.24) that

$$V = 12\sum_{i=1}^{k} \frac{(T_i^*)^2}{C_i^2} = 12\frac{0^2+0^2+(-0.75)^2+1.5^2}{22.5} = 1.5$$

Thus the *p*-value of the Sen-Adichie test is approximately $\Pr\left(\chi_{4-1}^2 \geq 1.5\right) = 0.682$.

This shows no evidence to reject H_0.

See Example 9.5 of the textbook page 468 for the meaning and interpretation of the data and the test result.

8.2 Multiple linear regression

In a multiple linear regression, the response variables depend on two or more covariates (regressors). Their relationship is modelled by

$$Y_i = \xi + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i, \quad i = 1, \ldots, n, \tag{8.28}$$

where $x_{1i}, \ldots, x_{pi}$ are p known covariates for Y_i, ξ is an unknown intercept, and $\beta_1, \ldots, \beta_p$ are unknown parameters called *regression coefficients* of covariates.

Write the data and parameters in vector or matrix form (with T for transpose):

$$\mathbf{Y} = [Y_1 \ \cdots \ Y_n]^{\mathsf{T}}, \quad \mathbf{x}_i = [x_{1i} \ \cdots \ x_{pi}]^{\mathsf{T}}, \quad i = 1, \ldots, n, \quad \mathbf{e} = [e_1 \ \cdots \ e_n]^{\mathsf{T}},$$

$$\boldsymbol{\xi} = \begin{bmatrix} \xi \\ \vdots \\ \xi \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}$$

Then (8.28) can be expressed as

$$Y_i = \xi + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + e_i, \quad i = 1,\ldots,n, \tag{8.29}$$

or in matrix form:

$$\mathbf{Y} = \xi + \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{8.30}$$

Assumption 8.3

(i) Response variables $Y_1,\ldots,Y_n$ are linearly related to covariates $\mathbf{x}_1,\ldots,\mathbf{x}_n$ by model (8.28) or (8.29), or by (8.30) in matrix form.

(ii) The errors $e_1,\ldots,e_n$ are continuous i.i.d. random variables with a symmetric distribution about 0 and density $f(t)$ satisfying

$$\int_{-\infty}^{\infty} f^2(t)dt < \infty$$

The null hypothesis H_0 of interest is that some of covariate coefficients are zeros, so that such covariates can be excluded if H_0 is accepted.

Hypotheses: For $1 \le q \le p$,

$$H_0 : \beta_1 = \cdots = \beta_q = 0 \text{ against } H_1 : \beta_j \neq 0 \text{ for some } j \in \{1, \ldots, q\}. \qquad (8.31)$$

$\beta_{q+1}, \ldots, \beta_p$ and ξ are unspecified under H_0.

This H_0 specifies that covariates $x_{1i}, \ldots, x_{qi}$ have no effects on Y_i, $i = 1, \ldots, n$. It is related to *variable selection* or *model selection*. q can be any integer between 1 and p, and $\beta_1, \ldots, \beta_q$ can be coefficients of any q covariates. For convenience we can place the covariates with zero coefficients under H_0 as the first q covariates without loss of generality as the order of the covariates does not matter. If H_0 is accepted, then the model can be simplified to fewer covariates $x_{q+1,i}, \ldots, x_{pi}$.

Hypotheses in (8.31) can be used to determine which covariates have significant effects on the response variable, and which ones are insignificant (hence can be ignored). This leads to selecting important covariates to obtain a parsimonious but sufficient model for data analysis.

Test statistic: The *Jaeckel-Hettmansperger-McKean test* statistic for testing the hypotheses in (8.31), denoted by *HM*, is calculated in 3 steps as follows.

Step 1. Let $R_i(\boldsymbol{\beta})$ denote the rank of $Y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ among $Y_1 - \mathbf{x}_1^{\mathsf{T}}\boldsymbol{\beta}, \ldots, Y_n - \mathbf{x}_n^{\mathsf{T}}\boldsymbol{\beta}$ and define the *Jaeckel dispersion* $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ by

$$D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \frac{\sqrt{12}}{n+1}\sum_{i=1}^{n}\left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2}\right](Y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) \tag{8.32}$$

Minimize $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ in (8.32) with respect to $\boldsymbol{\beta}$ without restriction to obtain an *unrestricted* estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

Step 2. Minimize $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ under the null hypothesis H_0. In other words, set $\beta_1 = \cdots = \beta_q = 0$ in $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ and minimize it with respect to $\beta_{q+1}, \ldots, \beta_p$. Denote the solution (minimizer) by $\hat{\boldsymbol{\beta}}_0$ and calculate

$$D_J^* = D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0) - D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{8.33}$$

Step 3. Obtain a consistent estimator $\hat{\tau}$ of the scale parameter

$$\tau = \frac{1}{\sqrt{12}\int_{-\infty}^{\infty} f^2(t)dt} \tag{8.34}$$

Then the test statistic *HM* is calculated by

$$HM = \frac{2D_J^*}{q\hat{\tau}} \tag{8.35}$$

Approximate distribution: For large samples, $HM \sim F_{q,n-p-1}$ approximately (the F distribution with q and $n-p-1$ degrees of freedom).

Rejection rule: The Jaeckel-Hettmansperger-McKean test rejects H_0 at the α level approximately if $HM \geq F_{q,n-p-1,\alpha}$, where $F_{q,n-p-1,\alpha}$ denotes the upper α percentile of the $F_{q,n-p-1}$ distribution.

A numerical example using R is provided in "R_for_STA3007.pdf".

Remark 8.2

1. A general form $D_J(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})=\sum_i a(R_i(\boldsymbol{\beta}))(Y_i-\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})$ was introduced as the Jaeckel dispersion, where $a(i)$ is nondecreasing in $i\in\{1,2,\ldots,n\}$ such that $\sum_i a(i)=0$. It is commonly to take $a(i)=\phi(i/(n+1))$, where $\phi(x)$ is a function on [0,1] such that $\phi(1-x)=-\phi(x)$, $\int_0^1\phi(x)dx=0$ and $\int_0^1\phi^2(x)dx=1$.

   The form in (8.32) is a special case with $\phi(x)=\sqrt{12}(x-1/2)$, which is known as the *Wilcoxon score function*.

2. The calculation of the test statistic *HM* generally requires numerical methods to compute $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_0$. The consistent estimation of the scale parameter τ in (8.34) also involves further theory and methods. These are only mentioned briefly in the textbook and will not be further discussed in this course.

   See Example 9.6 of the textbook (from page 479) for an application of the *HM* test and some numerical results by computer.

8.3 Nonparametric regression

- The regression problems we have considered so far are all under the *linear* regression models.
- In a nonparametric regression model, the relationship between the response variable Y_i and the regressor x_i is given by a nonparametric and unspecified function $\mu(x)$ in the form:

$$Y_i = \mu(x_i) + e_i, \;\; i = 1, \ldots, n, \tag{8.36}$$

  where $e_1, \ldots, e_n$ are i.i.d. continuous random variables with median 0.

- The problem now is on statistical inferences about the unknown regression function $\mu(x)$. This would involve a lot more theoretical and methodological development, including many research topics.
- We will only introduce some main ideas in nonparametric regression on how to estimate the regression function $\mu(x)$ by "smoothing" techniques referred to as "smoothers".

Running line smoother

Given $x_1,\ldots,x_n$, let $\delta = \delta_k(x) > 0$ be a positive number such that $|x_i - x| < \delta$ for k points $x_i \in \{x_1,\ldots,x_n\}$ and define

$$N_k(x) = \{i : |x_i - x| < \delta\} \tag{8.37}$$

Minimize the sum of squares (the method of least squares)

$$\sum_{i \in N_k(x)} (Y_i - \alpha - \beta x_i)^2$$

with respect to α and β to obtain the estimators $\hat{\alpha} = \hat{\alpha}_k(x)$ of α and $\hat{\beta} = \hat{\beta}_k(x)$ of β. Then estimate $\mu(x)$ by $\hat{\mu}(x) = \hat{\mu}_k(x) = \hat{\alpha} + \hat{\beta}x$. In particular,

$$\hat{\mu}(x_j) = \hat{\mu}_k(x_j) = \hat{\alpha} + \hat{\beta}x_j, \quad j = 1,\ldots,n.$$

This estimator of $\mu(x)$ is called the *running line smoother estimator*. It depends on the choice of k, which should not be too small or too large. It was suggested that $k \approx$ 10-15% of the sample size n is reasonable.

As an example, let $n = 20$ and $\{x_1, \ldots, x_{20}\} = \{1, 2, \ldots, 20\}$.

Take $k = 3$. Then for $x \in (8.5, 9.5)$, there are $k = 3$ points $\{x_8, x_9, x_{10}\} = \{8, 9, 10\}$ in $\{x_1, \ldots, x_{20}\}$ to satisfy $|x_i - x| < 1.5$. Hence $\delta = \delta_k(x) = \delta_3(x) = 1.5$ and

$$N_k(x) = N_3(x) = \{i : |x_i - x| < \delta = 1.5\} = \{8, 9, 10\} \quad \text{for } x \in (8.5, 9.5)$$

If $\{Y_8, Y_9, Y_{10}\} = \{54, 57, 66\}$, then

$$\sum_{i \in N_3(9)} (Y_i - \alpha - \beta x_i)^2 = (Y_8 - \alpha - \beta x_8)^2 + (Y_9 - \alpha - \beta x_9)^2 + (Y_{10} - \alpha - \beta x_{10})^2$$

$$= (54 - \alpha - 8\beta)^2 + (57 - \alpha - 9\beta)^2 + (66 - \alpha - 10\beta)^2$$

is minimized with respect to (α, β) by $\hat{\alpha} = 5$, $\hat{\beta} = 6$. Thus $\mu(x)$ is estimated by

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta} x = 5 + 6x \quad \text{for } x \in (8.5, 9.5)$$

For instance,

$$\hat{\mu}(9) = 5 + 6(9) = 59, \quad \hat{\mu}(9.2) = 5 + 6(9.2) = 60.2, \quad \text{and so on.}$$

Kernel regression smoother

A *kernel* (function) $K(t)$ is a density function. Commonly used kernels include:

- Tri-cube kernel: $K(t) = \frac{70}{81}(1-|t|^3)^3 I_{\{|t|\le 1\}}$;
- Standard normal density: $K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$; and
- Epanechnikov kernel: $K(t) = 0.75(1-t^2) I_{\{|t|\le 1\}}$

A *kernel regression smoother* estimates $\mu(x)$ by

$$\hat{\mu}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

where $K(t)$ is a kernel and $h > 0$ is referred to as the *bandwidth*, which controls the smoothness of the smoother and can be selected by certain criteria.

Locally regression smoother

Let $K(t)$ be a kernel and $N_k(x)$ be defined in (8.37). Take

$$w_i = w_i(x) = K\left(\frac{|x_i - x|}{\max\{|x_l - x| : l \in N_k(x)\}}\right), \quad i = 1, \ldots, n.$$

Minimize the weighed sum of squares (generalized least squares)

$$\sum_{i \in N_k(x)} w_i \left(Y_i - \alpha - \beta x_i\right)^2$$

with respect to α and of β to obtain the estimators $\hat{\alpha}$ of α and $\hat{\beta}$ of β. Then the *local regression smoother estimator* of $\mu(x)$ is given by.

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta} x$$

The tri-cube kernel $K(t) = \left(1 - t^3\right)^3 I_{\{0 \le t \le 1\}}$ is a popular choice of the kernel for the local regression smoother.

Spline regression smoother

The idea of spline regression smoother is to divide the domain of $x_1,\ldots,x_n$ into subintervals and approximate $\mu(x)$ by polynomials of a same degree d in each subinterval, then estimate the polynomials by multiple linear regression.

Specifically, let $\eta_1 < \cdots < \eta_K$ be K points (referred to as *knots*) in the domain of $x_1,\ldots,x_n$, $\eta_0 = -\infty$ and $\eta_{K+1} = \infty$. Approximate $\mu(x)$ by

$$\mu(x) \approx \sum_{k=1}^{K+1} \left(\beta_{k0} + \beta_{k1}x + \beta_{k2}x^2 + \cdots + \beta_{kd}x^d \right) I_{\{\eta_{k-1} \le x < \eta_k\}} \qquad (8.38)$$

($\eta_0 \le x$ is interpreted as $-\infty < x$). Common choices of the degree are $d = 1, 2, 3$.

Use multiple linear regression in (8.38) to obtain the estimators $\hat{\beta}_{k0}, \hat{\beta}_{k1},\ldots,\hat{\beta}_{kd}$ of $\beta_{k0}, \beta_{k1},\ldots,\beta_{kd}$. Then the *spline regression smoother estimator* of $\mu(x)$ is

$$\hat{\mu}(x) = \sum_{k=1}^{K+1} \left(\hat{\beta}_{k0} + \hat{\beta}_{k1}x + \hat{\beta}_{k2}x^2 + \cdots + \hat{\beta}_{kd}x^d \right) I_{\{\eta_{k-1} \le x < \eta_k\}}$$