LINEAR REGRESSION

1. Linear Regression Model

1.1. Model and Notations

Suppose there are N subjects in the system. For the ith subject, we could collect a continuous type response $y_i \in \mathbb{R}$ and an associated p-dimensional covariate vector $x_i = (x_{i1}, \cdots, x_{ip})^\top \in \mathbb{R}^p$. The linear regression model takes the following form,

$$y_i = x_i^\top \beta + \varepsilon_i. \tag{1.1}$$

Write $\mathbf{y} = (y_1, \cdots, y_N)^\top$, $\mathbf{X} = (x_1, \cdots, x_N)^\top \in \mathbb{R}^{N \times p}$, and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_N)^\top \in \mathbb{R}^N$. Then the linear regression can be written in a matrix form as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \tag{1.2}$$

Denote X_j as the jth column of $\mathbf{X}$. Then we can add an intercept term by letting $X_1 = \mathbf{1}$.

Comment:

1. **Quantitative inputs** & its transformations (log, squares) & basis expansions ($X_2 = X_1^2$, $X_3 = X_1^3$)
2. **Qualitative inputs***: dummy variable coding.
3. **Interaction between variables.**

1.2. Model Assumptions

(A1) The relationship between response ($\mathbf{y}$) and covariates ($\mathbf{X}$) is linear;

(A2) $\mathbf{X}$ is a non-stochastic matrix and $\text{rank}(\mathbf{X}) = p$;

(A3) $E(\varepsilon) = \mathbf{0}$. This implies $E(\mathbf{y}) = \mathbf{X}\beta$;

(A4) $\text{cov}(\varepsilon) = E(\varepsilon\varepsilon^\top) = \sigma^2 I_N$; (Homoscedasticity)

(A5) ε follows multivariate normal distribution $N(\mathbf{0}, \sigma^2 I_N)$ (Normality)

Remark: $\mathbf{X}$ could include intercept term. The assumption A2 can be further relaxed that X can be random matrix. The above assumptions can replaced by the following:

(A2*) $\mathbf{X}$ is a full rank matrix with probability 1;

(A3*) $E(\varepsilon|\mathbf{X}) = \mathbf{0}$;

(A4*) $E(\varepsilon\varepsilon^\top|\mathbf{X}) = \sigma^2 I_N$;

(A5*) $\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2 I_N)$.

A sufficient condition for (A2*) is that $\lambda_{\min}(\mathbf{X}^\top\mathbf{X}) \to \infty$ a.s.

2. Model Estimation

Ordinary least squares (OLS) estimation:

$$\text{RSS}(\beta) = \sum_{i=1}^N \big\{y_i - f(x_i)\big\}^2 = \sum_{i=1}^N \big\{y_i - \beta_0 - \sum_j x_{ij}\beta_j\big\}^2.$$

Comment:

1. This criterion is valid if y_i's are conditionally independently given the inputs x_i.

Rewrite $\text{RSS}(\beta)$ using a matrix form as

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta).$$

Differentiating with respect to β we require

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = -2(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{X} = 0.$$

Solution:

$$\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Comment:

1. Here we implicitly assume that $\mathbf{X}$ is full rank, hence $\mathbf{X}^\top\mathbf{X}$ is positive definite.

2. Fitted values:

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{H}\mathbf{y} \tag{2.1}$$

The residual vector $\mathbf{y}-\widehat{\mathbf{y}}$ is orthogonal to the column space of $\mathbf{X}$ (pls verify by yourself.) Hence $\widehat{\mathbf{y}}$ is the *orthogonal projection* of $\mathbf{y}$ onto the column space of $\mathbf{X}$. The matrix $\mathbf{H}$ is called "hat" matrix or projection matrix.

3. The residual sum of squares $\mathrm{RSS}(\beta)$ can be used as a goodness-of-fit measure;

4. If $\mathbf{X}$ is not of full rank (e.g., if two of the inputs are perfectly correlated)? Will $\widehat{\beta}$ or $\widehat{\mathbf{y}}$ change?

Homework: Prove that the OLS estimator $\widehat{\beta}$ is the same as the maximum likelihood estimator.

3. Statistical Inference

3.1. Mean and Variance of the OLS Estimator

Assume assumptions (A1)–(A4), we then have

$$E(\widehat{\beta}) = \beta, \quad \mathrm{cov}(\widehat{\beta}) = (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2.$$

Typically σ^2 is estimated by

$$\widehat{\sigma}^2 = \frac{1}{N-p}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2, \tag{3.1}$$

where $\widehat{y}_i = x_i^\top\beta$ is the fitted value of the ith subject.

Theorem 1. *(Gauss-Markov Theorem) Assume conditions A1–A4. Then $\widehat{\beta}$ is the best linear unbiased estimator (BLUE), provided it exists.*

It implies, $\widehat{\beta}$ has the smallest variance over all linear unbiased estimator $\widetilde{\beta}$, i.e., $\widetilde{\beta} = \sum_i w_i y_i$ and $E(\widetilde{\beta}) = \beta$. Here the smallest variance means that for any $\eta \in \mathbb{R}^p$ with $\|\eta\| = 1$, $\mathrm{var}(\eta^\top \widehat{\beta}) \le \mathrm{var}(\eta^\top \widetilde{\beta})$.

Homework:

(1) Prove the Gauss-Markov Theorem;

(2) Prove $E(\widehat{\sigma}^2) = \sigma^2$.

3.2. Sampling Properties

Assume conditions A1–A5. Then

$$\widehat{\beta} \sim N(\beta, (\mathbf{X}^\top \mathbf{X})^{-1}\sigma^2) \tag{3.2}$$

$$(N-p)\widehat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p} \tag{3.3}$$

In addition, $\widehat{\beta}$ is independent with $\widehat{\sigma}^2$. It is implied by (3.2) that $R(\widehat{\beta} - \beta) \sim N(\mathbf{0}, R(\mathbf{X}^\top \mathbf{X})^{-1} R^\top \sigma^2)$.

Homework: Prove (3.2) and (3.3).

Hypothesis test: $H_0 : \beta_j = 0$ v.s. $H_1 : \beta_j \neq 0$

Q: Suppose σ^2 is known. In this case, $R =?$.

Z-score:

$$z_j = \frac{\widehat{\beta}_j}{\widehat{\sigma}\sqrt{v_j}},$$

where v_j is the jth diagonal element of $(\mathbf{X}^\top \mathbf{X})^{-1}$.

Under the null: z_j follows t-distribution with $N-p$ degrees of freedom (if N is large we could also use normal quantiles because the differences between normal and t-distribution are negligible).

Test the significance of groups of coefficients simultaneously:

For instance, suppose we have p_1 covariates in total. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p_0} = 0$,

H_1 : there exists at least one j $(1 \leq j \leq p_0)$, such that $\beta_j \neq 0$. use F statistic:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/p_0}{\mathrm{RSS}_1/(N - p_1)}$$

The F statistic follows F distribution $F(p_0, N - p_1)$.

Q:

1. what is RSS_0 and what is RSS_1?

Answer: RSS_0 is the residual sum of squares when we drop $X_1, \cdots, X_{p_0}$; RSS_1 is defined for the full model.

2. prove the F statistic is equivalent to the t-test when dropping the single coefficient.

4. Goodness-of-fit

Define $\widehat{y}_i = x_i^\top \widehat{\beta}$. Let the intercept be included in the regression model. Define the total sum of squares (TSS) and explained sum of squares (ESS) as follows

$$\mathrm{TSS} = \sum_i (y_i - \overline{y})^2, \quad \mathrm{ESS} = \sum_i (\widehat{y}_i - \overline{y})^2.$$

It can be proved that

$$\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}.$$

HW: Prove the above equation.

Define the R-squares of the regression model as follows

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{ESS}}{\mathrm{TSS}}. \tag{4.1}$$

Adjusted R-squares:

$$\text{Adjusted } R^2 = 1 - \frac{\mathrm{RSS}/(n-p)}{\mathrm{TSS}/(n-1)}. \tag{4.2}$$

5. Model Selection

1. **Subset Selection**

1.1 Best-subset Selection: time consuming

1.2 Forward-stepwise selection (greedy algorithm): Starts with the intercept, and then sequentially adds into the model the predictor that most improves the fit.

1.3 Backward-stepwise selection: starts with the full model, and sequentially deletes the predictor that has the least impact on the fit. (can be only used when $N > p$).

1.4 Stepwise-selection: consider both forward and backward moves at each step, and select the "best" of the two (minimize AIC/BIC criterion).

$$\text{AIC} = -\frac{2}{N}\mathcal{L}(\beta) + 2\frac{d}{N}.$$
$$\text{BIC} = -2\mathcal{L}(\beta) + (\log N)d$$

where $\mathcal{L}(\beta)$ denotes the log-likelihood and d is the number of parameters to be estimated.

Q: Could you tell the difference here?

Comment:

1. There are other criterions including C_p and many others. See Chapter 7.4–7.7 for more details.
2. BIC can consistently select the true model.

2. **Shrinkage Methods**

2.1 Ridge Regression:

$$\widehat{\beta}^{ridge} = \arg\min_{\beta}\Big\{\sum_{i=1}^{N}\big(y_i - \sum_j x_{ij}\beta_j\big)^2 + \lambda\sum_j \beta_j^2\Big\}$$

Here λ is a tuning parameter which controls the amount of shrinkage.

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^\top\beta$$

Q: the solution takes the form?

$$\widehat{\beta}^{ridge} = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{y}.$$

Note: In the case of orthonormal inputs, the ridge estimates are scaled version of the least squares estimates, i.e., $\widehat{\beta}^{ridge} = \widehat{\beta}/(1+\lambda)$.

See Chapter 3.4.1 of Elements for interpretations of ridge regressions from the aspect of SVD decomposition.

2.2 Lasso Regression: $\sum_j \beta_j^2 \to \sum_j |\beta_j|$.