Multiple Linear Regression

Instructor: Junghye Lee

Department of Industrial Engineering
junghyelee@unist.ac.kr

Contents

Probabilistic Model

y_i: Observed value of the random variable Y_i depends on $x_{i1,} x_{i2}, \dots, x_{ip}$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1,2, \dots, n$$

$\beta_0, \beta_1, \dots, \beta_p$: unknown model parameters

n: number of observations

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

Fitting the Model

- LS provides estimates of the unknown model parameters, $\beta_0, \beta_1, \dots, \beta_p$ which minimizes Q

$$Q = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)\right]^2$$

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\right] = 0$$

$$\frac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij} \left[y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})\right] = 0$$

for $j = 1, 2, \dots, p$

Goodness of Fit of the Model

- Residuals $e_i = y_i - \hat{y}_i \quad (i = 1,2, \dots, n)$
- $\hat{y}_i$ are the fitted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}, \quad i = 1,2, \dots, n$$

- An overall measure of the goodness of fit R^2

✓ r^2 is R^2.
✓ It is satisfied only in simple linear regression.

Statistical Inference on Multiple Regression

- Determine which predictor variables have statistically significant effects.
- We test the hypotheses:

$$H_{0j}\colon \beta_j = 0, \qquad H_{aj}\colon \beta_j \neq 0$$

- If we can't reject H_{0j}, then x_j is not a significant predictor of y.

Statistical Inference on Multiple Regression

- The steps are similar to ones of simple linear regression.

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 V_{11}), \qquad SE\left(\hat{\beta}_1\right) = s\sqrt{V_{11}}$$

- What's V_{11}? Why $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 V_{11})$?

Statistical Inference on $\boldsymbol{\beta}$s

1. Mean: Recall from simple linear regression, the least squares estimators for the regression parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. Here, $\widehat{\boldsymbol{\beta}}$ of least squares estimators is also unbiased.

$$E(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} E(\hat{\beta}_0) \\ E(\hat{\beta}_1) \\ \vdots \\ E(\hat{\beta}_p) \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Statistical Inference on $\boldsymbol{\beta}$s

2. Variance: Constant Variance assumption: $V(\epsilon_i) = \sigma^2$

$$Var(Y) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I_n$$

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}Y = cY$$

$$Var\left(\widehat{\boldsymbol{\beta}}\right) = Var(cY) = cVar(Y)c^{T}$$

$$= \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}(\sigma^2 I_n)\left(\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\right)^{T} = \sigma^2\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}$$

Let V_{jj} be the jth diagonal of the matrix $\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}$

$$Var\left(\hat{\beta}_j\right) = \sigma^2 V_{jj}$$

Statistical Inference on $\boldsymbol{\beta}$s

- Derivation of confidence interval of β_j

$$P\left(-t_{n-(p+1),\frac{\alpha}{2}} \le \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \le t_{n-(p+1),\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(\hat{\beta}_j - t_{n-(p+1),\frac{\alpha}{2}} SE(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{n-(p+1),\frac{\alpha}{2}}\right) = 1 - \alpha$$

- The $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{n-(p+1),\frac{\alpha}{2}} SE(\hat{\beta}_j)$$

Statistical Inference on $\boldsymbol{\beta}$s

- An α-level test of hypotheses

$$H_{0j}: \beta_j = \beta_j^0, \qquad H_{1j}: \beta_j \neq \beta_j^0$$

$$P\left(Reject\ H_{0j} \middle| H_{0j}\ is\ true\right) = P\left(\left|t_j\right| \geq c\right) = \alpha$$

where $c = t_{n-(p+1),\alpha/2}$

- Reject H_{0j} if

$$\left|t_j\right| = \frac{\left|\hat{\beta}_j - \beta_j^0\right|}{SE(\hat{\beta}_j)} > t_{n-(p+1),\alpha/2}$$

Prediction of Future Observation

- Having fitted a multiple regression model, suppose we wish to predict the future value of Y^* for a **specified vector** of predictor variables

$$\boldsymbol{x}^* = \left(x_0^*, x_1^*, \ldots, x_p^*\right)^T$$

- First, you can see mean response $E(Y^*|\boldsymbol{x}^*)$ where $E(Y^*|\boldsymbol{x}^*) = (\boldsymbol{x}^*)^T\boldsymbol{\beta}$

$$\hat{Y}^* = \hat{E}(Y^*|\,\boldsymbol{x}^*) = (\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}}$$

Estimator for the mean response $E(Y^*|\boldsymbol{x}^*)$

$$E\left(\hat{Y}^*\right) = E\left((\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{x}^*)^T\boldsymbol{\beta}$$

Mean of the estimator (unbiased)

$$Var(\hat{Y}^*) = Var\left[(\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}}\right] = (\boldsymbol{x}^*)^T Var\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{x}^* = (\boldsymbol{x}^*)^T\sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}^*$$

Variance of the estimator

$$= \sigma^2(\boldsymbol{x}^*)^T\boldsymbol{V}\boldsymbol{x}^*, \text{where } \boldsymbol{V} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$$

Prediction of Future Observation

- Replacing σ^2 by its estimate $s^2 = MSE$, which has $n-(p+1)$ d.f., and using methods as in simple linear regression, a $(1-\alpha)$-level CI for $E(Y^*)$ is given by

$$\boxed{\hat{E}(Y^*|x^*)}^{(\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}}} - t_{n-(p+1),\frac{\alpha}{2}}\boxed{s}^{MSE}\sqrt{(\boldsymbol{x}^*)^T V(\boldsymbol{x}^*)} \le \boxed{E(Y^*|x^*)}^{(\boldsymbol{x}^*)^T\boldsymbol{\beta}}$$

$$\le \hat{E}(Y^*|x^*) + t_{n-(p+1),\frac{\alpha}{2}} s\sqrt{(\boldsymbol{x}^*)^T V(\boldsymbol{x}^*)}.$$

Mean response distribution

- The predicted response distribution is the predicted distribution of the residuals between future response (Y^*) and predicted future response ($(\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}}$) at the given point $\boldsymbol{x}^*$.
  - Definition of predicted response distribution
  - Predicted future response = Estimator of mean response

$$E\left(Y^* - (\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}}\right) = 0, Var\left(Y^* - (\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}}\right) = \sigma^2 + \sigma^2(\boldsymbol{x}^*)^T V(\boldsymbol{x}^*)$$

Predicted response distribution

$$(\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}} - t_{n-(p+1),\frac{\alpha}{2}} s\sqrt{\left(1+(\boldsymbol{x}^*)^T\boldsymbol{V}(\boldsymbol{x}^*)\right)} \le Y^*$$

$$\le (\boldsymbol{x}^*)^T\widehat{\boldsymbol{\beta}} + t_{n-(p+1),\frac{\alpha}{2}} s\sqrt{\left(1+(\boldsymbol{x}^*)^T\boldsymbol{V}(\boldsymbol{x}^*)\right)}$$

c.f.) Future response distribution

$Y^* \sim N((\boldsymbol{x}^*)^T\boldsymbol{\beta}, \sigma^2)$

F-Test for $\boldsymbol{\beta_j}$s

- Consider:

$$H_0: \beta_1 = \cdots = \beta_p = 0$$

$$H_1: At\ least\ one\ \beta_j \neq 0$$

  Here H_0 is the overall null hypothesis, which states that none of the X variables are related to y. The alternative one shows at least one is related.

- How to Build a F-Test
  - The test statistic $F = MSR/MSE$ follows F-distribution with p and $n - (p+1)$ d.f. The α-level test rejects H_0 if

$$F = \frac{MSR}{MSE} > f_{p,n-(p+1),\alpha}$$

Relation between F and R^2

- F can be written as a function of R^2.
- By using the formula:

$$SSR = R^2 SST; \quad SSE = (1 - R^2)SST.$$

- F can be as:

$$F = \frac{R^2[n - (p + 1)]}{p(1 - R^2)}$$

- We see that F is an increasing function of R^2 and test the significance of it.

Analysis of Variance (ANOVA)

- The relation between SST, SSR and SSE:

$$SST = SSR + SSE$$

- where they are respectively equals to:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- The corresponding degrees of freedom (d.f.) is:

$$d.f.(SST) = n - 1; d.f.(SSR) = p; d.f.(SSE) = n - (p + 1).$$

ANOVA Table for Multiple Regression

Source of Variation (Source)	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS)	F
Regression	SSR	p	MSR = SSR/p	F= MSR/MSE
Error	SSE	n – (p+1)	MSE = SSE/n-(p+1)	
Total	SST	n – 1		

- This table gives us a clear view of analysis of variance of Multiple Regression.

Extra Sum of Squares Method for Testing Subsets of Parameters

- Before, we consider the full model with *p* parameters. Now we consider the partial model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-m} x_{i,p-m} + \epsilon_i, \qquad i = 1,2, \dots, n$$

while the rest *m* coefficients are set to zero. And we could test these *m* coefficients to check out the significance:

$$H_0\colon \beta_{p-m+1} = \cdots = \beta_p = 0$$

$$H_1\colon At\ least\ one\ \beta_j \neq 0$$

where $j = p - m + 1, \dots, p$

Extra Sum of Squares Method for Testing Subsets of Parameters

- Full model with *p* parameters

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1,2,\ldots,n$$

- Now we consider the partial model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-m} x_{i,p-m} + \epsilon_i, \qquad i = 1,2,\ldots,n$$

while **the rest *m* coefficients are set to zero**. And we could test these *m* coefficients to check out the significance:

$$H_0: \beta_{p-m+1} = \cdots = \beta_p = 0$$

$$H_1: At\ least\ one\ \beta_j \neq 0$$

where $j = p - m + 1, \ldots, p$

Building F-test by Using Extra Sum of Squares Method

- Let $SSR_{(R)}$ and $SSE_{(R)}$ be the regression and error sums of squares for the partial model. Since SST is fixed regardless of the particular model, so:

$$SST = SSR_{(R)} + SSE_{(R)}$$

  then, we have:

$$SSR_{(R)} + SSE_{(R)} = SSR_{(F)} + SSE_{(F)}$$

- The α-level F-test rejects null hypothesis if

$$F = \frac{(SSR_{(F)} - SSR_{(R)})/m}{SSE_{(F)}/[n-(p+1)]} > f_{m,n-(p+1),\alpha}$$

Remarks on the F-test

- The numerator d.f. is m which is **the number of coefficients set to zero**. While the denominator d.f. is $n-(p+1)$ which is the **error d.f. for the full model**.
- The MSE in the denominator is the **normalizing factor**, which is an estimate of σ^2 for the full model. If the ratio is large, we reject H_0.

Control group
(You should do better than at least error)