Introduction to Statistics Note

2024 Spring Semester

21 CST H3Art

Chapter 11: Multiple Regression

11.1 Inference for Multiple Regression

The linear regression model in which the mean response, μ_y, is related to **one explanatory variable** x:

$$\mu_y = \beta_0 + \beta_1 x$$

The data for a **simple linear regression** problem consist of n observations (x_i, y_i) of **two variables**.

In multiple regression, the response variable y depends on p **explanatory variables** $x_1, x_2, \ldots, x_p$:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Data for **multiple linear regression** consist of the value of a response variable y and p **explanatory variables** $(x_1, x_2, \ldots, x_p)$ on each of n cases.

The **statistical model for multiple linear regression** is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

for $i = 1, 2, \ldots, n$

The **mean response** μ_y is a linear function of the explanatory variables:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

The **deviations** ε_i **（偏差）** are independent and Normally distributed $N(0, \sigma)$

The **parameters of the model（模型参数）** are $\beta_0, \beta_1, \ldots, \beta_p$ and σ.

The coefficient $\beta_i (i = 1, \ldots, p)$ represents **the average change in the response** when the variable x_i increases by one unit and all other x variables are held constant. **（每个x_i对应的系数β_i只代表了该变量变化单位值且其他变量x_j($j \neq i$)保持不变时，响应变量变化的值）**

Estimation of the Parameters（参数值估计）

Select a **random sample of n individuals** on which $p + 1$ variables $(x_1, \ldots, x_p, y)$ are measured. The **least-squares regression method** chooses $b_0, b_1, \ldots, b_p$ to minimize the sum of squared deviations $(y_i – \hat{y}_i)^2$, where:

$$\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$$

As with simple linear regression, the constant b_0 is the **y-intercept**.

The parameter σ^2 measures the variability of the responses about the population response mean. The estimator of σ^2 is:

$$s^2 = \frac{\sum e_i^2}{n - p - 1} = \frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}$$

Confidence Interval for β_i rely on the t-distribution, with $n-p-1$ **degrees of freedom**, a level C confidence interval for β_j is:

$$b_j \pm t^* \cdot \mathbf{SE}_{b_j}$$

where $\mathbf{SE}_{b_j}$ is the standard error of b_j and t^* is the t critical for the $t(n-p-1)$ distribution with area C between $-t^*$ and t^*

Significance Test for β_j **（**β_j**的显著性检验）**：

- Null hypothesis $H_0 : \beta_j = 0$, calculate the t statistic:

$$t = \frac{b_j}{\mathbf{SE}_{b_j}}$$

  which has the $t(n-p-1)$ distribution.
- Alternative hypothesis:
  - $H_a : \beta_j > 0$ is $P(T \geq t)$
  - $H_a : \beta_j < 0$ is $P(T \leq t)$
  - $H_a : \beta_j \neq 0$ is $2P(T \geq |t|)$
  - **Note**: Software typically provides **two-sided** P-values

Significance Test for β_j

- Suppose we test $H_0 : \beta_j = 0$ for each j and find that **none of the** p **tests is significant**, we **should not** conclude that **none of the explanatory variables is related to the response（不应得出结论认为所有解释变量均与响应无关）**.
- When we fail to reject $H_0 : \beta_j = 0$, this means that we **probably don't need** x_j **in the model with all the other variables（我们可能不需要在模型中包含**x_j**）**, so it merely means that it's **safe to throw away at least one of the variables（安全地丢弃至少一个变量）**.

ANOVA F**-test for Multiple Regression（对多变量回归的ANOVA** F**检验）**
In multiple regression, the ANOVA F statistic tests the hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ vs. } H_a : \text{ at least one } \beta_j \neq 0$$

by computing the F statistic $F = \mathbf{MSM}/\mathbf{MSE}$, When H_0 is true, F follows the $F(p, n-p-1)$ distribution. The P-value is $P(F \geq f)$.

p **is number of predictors（**p**是自变量/预测变量的数量）**

*A significant P-value doesn't mean that all p explanatory variables have a significant influence on y——**only that at least one does**.*

ANOVA Table for Multiple Regression

Source	Sum of Squares SS	df	Mean square MS	F	P-value
Model	$\sum(\hat{y}_i - \bar{y})^2$	p	$\mathbf{MSM} = \mathbf{SSM}/\mathbf{DFM}$	$\mathbf{MSM}/\mathbf{MSE}$	Tail area above F
Error	$\sum(y_i - \hat{y}_i)^2$	$n-p-1$	$\mathbf{MSE} = \mathbf{SSE}/\mathbf{DFE}$		
Total	$\sum(y_i - \bar{y})^2$	$n-1$			

$\mathbf{SSM}$ =model sum of squares, $\mathbf{SSE}$ =error sum of squares

$\mathbf{SST}$ =total sum of squares, $\mathbf{SST} = \mathbf{SSM} + \mathbf{SSE}$

$\mathbf{DFM} = p, \mathbf{DFE} = n - p - 1, \mathbf{DFT} = n - 1, \mathbf{DFT} = \mathbf{DFM} + \mathbf{DFE}$

Squared Multiple Correlation R^2（多变量相关系数R方）

R^2, the squared multiple correlation, is **the proportion of the variation in the response variable y** that is explained by the model.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\mathbf{SSM}}{\mathbf{SST}}$$

The square root of R^2, namely R, called the **multiple correlation coefficient（多变量相关系数）**, is the **correlation between the observations and the predicted values（观测值和预测值的相关性）**.