Introduction to Statistics Note

2024 Spring Semester

21 CST H3Art

Chapter 10: Inference for Regression

10.1 Simple Linear Regression

The **slope（斜率）** and **intercept（截距）** of the least-squares line are **statistics**.

These statistics would take somewhat different values if we repeated the data production process. To do inference, think of b_0 and b_1 as **estimates（估计量）** of unknown parameters β_0 and β_1 that describe the **population** of interest. **（记b_0和b_1为总体参数β_0和β_1的估计量）**

We have n **observations** on an **explanatory variable** x and a **response variable** y. Our goal is to study or predict the behavior of y for given values of x.

- For any fixed value of x, the **response y varies according to a Normal distribution**. Repeated responses y **are independent** of each other.
- The mean response μ_y has a straight line relationship with x given by a population regression line $\mu_y = \beta_0 + \beta_1 x$.
- The **slope** and **intercept** are **unknown parameters（斜率和截距是未知参数）**.
- The standard deviation of y (call it σ) is the same for all values of x. The value of σ is unknown.

In the **population**, the linear regression equation is:

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

Sample data fits **simple linear regression model**:

$$\begin{aligned}\textbf{data} &= \textbf{fit} + \textbf{error}\\ y_i &= (\beta_0 + \beta_1 x_i) + (\varepsilon_i)\end{aligned}$$

where the ε_i are **independent** and **Normally** distributed $N(0, \sigma)$.

The intercept b_0, the slope b_1, and the standard deviation σ of y are **the unknown parameters of the regression model**, and The value of $\hat{y}$ from the **least-squares regression line（最小二乘回归线）** is really a prediction of the mean value of $y(\mu_y)$ for a given value of x.

The least-squares regression line $(\hat{y} = b_0 + b_1 x)$ obtained from sample data is the **best estimate（最佳拟合）** of the true population regression line $(\mu_y = b_0 + b_1 x)$:

- $\hat{y}$ is an unbiased estimate for mean response μ_y
- b_0 is an unbiased estimate for intercept β_0
- b_1 is an unbiased estimate for slope β_1

The **population standard deviation σ** for y at any given value of x represents **the spread of the normal distribution** of the ε_i around the mean μ_y.

The **predicted values（预测值，预测值是均值，代表任何拥有自变量为x_i的值会得到预测值$\hat{y}_i$）** are $\hat{y}_i = b_0 + b_1 x_i, i = 1, \ldots, n$ and the **residuals（残差）** are $y_i - \hat{y}_i, i = 1, \ldots, n$. And the **regression standard error（回归标准差）**, s, for n sample data points is calculated from the residuals $(y_i - \hat{y}_i)$:

$$s=\sqrt{\frac{\sum(\mathbf{residual})^2}{n-2}}=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}$$

s is an essentially **unbiased estimate（无偏估计量）** of the regression standard deviation σ.

Before you can trust the results of **regression inference（回归推理）**, you must check the conditions for inference one by one:

- The **relationship is linear（关系是线性的）** in the population
- The **response varies Normally（因变量是关于回归线正态分布的）** about the population regression line
- Observations are **independent**
- The **standard deviation** of the responses is the **same** for all values of x

The **slope** β_1 of the population regression line $\mu_y=\beta_0+\beta_1 x$ is the **rate of change of the mean response** as the explanatory variable increases, the **confidence interval** for β_1 has the familiar form:

$$\mathbf{estimate}\pm t^*\cdot(\mathbf{standard\ deviation\ of\ estimate})$$

and because we use the statistic b as our estimate, the confidence interval is:

$$b_1\pm t^*\cdot\mathbf{SE}_{b_1}$$

Here t^* is the critical value for the t distribution with $df=n–2$ having area C between $–t^*$ and t^*.

To test the hypothesis $H_0:\beta_1=\mathbf{hypothesized\ value}$, compute the test statistic and use the t distribution with $df=n-2$:

$$t=\frac{b_1-\mathbf{hypothesized\ value}}{\mathbf{SE}_{b_1}}$$

We may look for evidence of a **significant relationship（关系显著性）**, we can test the hypothesis that the regression slope parameter β_1 is equal to zero:

$$H_0:\beta_1=0\ \text{vs.}\ H_a:\beta_1\neq 0$$

Testing $H_0:\beta_1=0$ is equivalent to testing the hypothesis of no correlation between x and y in the population:

$$\mathbf{slope}\ b_1=r\times\frac{s_y}{s_x}$$

We can also calculate a confidence interval for the **population mean** μ_y of all responses y:

$$\hat{\mu_y}\pm t^*\cdot\mathbf{SE}_{\hat{\mu}}$$

where t^* is the value such that the area under the $t(n–2)$ density curve between $–t^*$ and t^* is C.

To estimate an **individual response** y for a given value of x, we use a **prediction interval（预测区间）**, the level C prediction interval for a single observation on y is:

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

t^* is the critical value for the $t(n–2)$ distribution with area C between $–t^*$ and t^*.

10.2 More Detail about Simple Linear Regression*

From 10.1, The regression model is:

$$\mathbf{data}=\mathbf{fit}+\mathbf{error}$$
$$y_i=(\beta_0+\beta_1 x_i)+(\varepsilon_i)$$

where the ε_i are **independent** and **Normally** distributed $N(0,\sigma)$, and σ is the same for all values of x.

It resembles an **ANOVA（方差分析/ANalysis Of VAriance）**, which also assumes equal variance, where:

$$\mathbf{SST} = \mathbf{SS\ model} + \mathbf{SS\ error}$$
$$\mathbf{DFT} = \mathbf{DF\ model} + \mathbf{DF\ error}$$

SS means **Sum of Squares（平方和）**.

For a simple linear relationship, the ANOVA tests the hypotheses:

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

by comparing $\mathbf{MSM}$ (model) to $\mathbf{MSE}$ (error): $F = \dfrac{MSM}{MSE}$, when H_0 is true, F follows the $F(1, n-2)$ **distribution**. The P-value is $P(F \geq f)$.

The ANOVA test and the two-sided t-test for $H_0 : \beta_1 = 0$ yield the same P-value. （方差分析和双尾t检验对于$H_0 : \beta_1 = 0$的检验将会得到相同的p值）

The ANOVA Table:

Source	**Sum of Squares($\mathbf{SS}$)**	$\mathbf{DF}$	**Mean Square(MS)**	F	**P-value**
Model	$\mathbf{SSM} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$\mathbf{MSM} = \mathbf{SSM}/\mathbf{DFM}$	$\mathbf{MSM}/\mathbf{MSE}$	Tail area above F
Error	$\mathbf{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$\mathbf{MSE} = \mathbf{SSE}/\mathbf{DFE}$		
Total	$\mathbf{SST} = \sum_{i=1}^{n}(y_1 - \bar{y})^2$	$n-1$			

$$\mathbf{SST} = \mathbf{SSM} + \mathbf{SSE}$$
$$\mathbf{DFT} = \mathbf{DFM} + \mathbf{DFE}$$
$$F = \mathbf{MSM}/\mathbf{MSE}$$

The standard deviation, s, of the n residuals $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$, is calculated from the following quantity:

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\mathbf{SSE}}{\mathbf{DFE}} = \mathbf{MSE}$$

s is an approximately **unbiased estimate** of the regression standard deviation σ.

To assess variation in the estimates of β_0 and β_1, we calculate the standard errors for the estimated regression coefficients:

- The standard error of the slope estimate b_1 is:

$$\mathbf{SE}_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x}_i)^2}}$$

- The standard error of the intercept estimate b_0 is:

$$\mathbf{SE}_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x}_i)^2}}$$

To estimate mean responses or predict future responses, we calculate the following standard errors:

- The standard error of the estimate of the mean response μ_y is:

$$\mathbf{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

- The standard error for predicting an individual response y is:

$$\mathbf{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

To test the null hypothesis of no linear association, we have the choice of also using the correlation parameter ρ:

$$b_1 = r \times \frac{s_y}{s_x}$$

The test of significance for ρ uses the one-sample t-test for: $H_0 : \rho = 0$, compute the t statistic for sample size n and correlation coefficient r:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The P-value is the area under $t(n–2)$ for values of t as or more extreme than t in the direction of H_a.