Linear Regression & Analysis

12/14/22

Formulas from chapter 12.1:

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$E[aX + b] = aE[X] + b$$

$$Var[aX + b] = a^2 Var[X]$$

Chapter 12.2:

The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}},$$

where

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}, \qquad S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}.$$

The least squares estimate of the slope coefficient β_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \overline{y} - \hat{\beta}_1 \overline{x}.$$

SSE

The **error sum of squares** (equivalently, residual sum of squares), denoted by SSE, is

① $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$

② $$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2,$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}.$$

We can calculate SSE as

$$\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i = S_{yy} - \hat{\beta}_1 S_{xy}.$$

Chapter 12.2 Continued...

The **coefficient of determination**, denoted by r^2, is given by

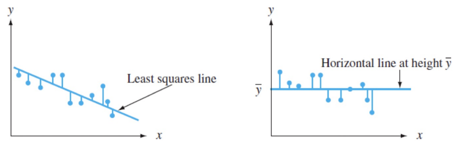
$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}.$$

It is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model.

SST

A quantitative measure of the total amount of variation in observed y values is given by the **total sum of squares**, denoted by SST is

$$\text{SST} = S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$



(a) (b)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 75.21243    2.98363  25.208 9.22e-12 ***
iodine      -0.20939    0.03109  -6.734 2.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.564 on 12 degrees of freedom
Multiple R-squared:  0.7908,	Adjusted R-squared:  0.7733
F-statistic: 45.35 on 1 and 12 DF,  p-value: 2.091e-05
```

$\hat{\beta}_0$ $\hat{\beta}_1$

SSR= $\sum(\hat{y}_i - \bar{y})^2$ =SST-SSE

Regression sum of squares is interpreted as the amount of total variation that is explained by the model.

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{(\text{SST}-\text{SSE})}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}$$

Chapter 12.3:

- The mean value of $\hat{\beta}_1$ is $E[\hat{\beta}_1] = \beta_1$, so $\hat{\beta}_1$ is an unbiased estimator for β_1 is (the distribution of $\hat{\beta}_1$ is always centered at the value of β_1).
- The variance and standard deviation of $\hat{\beta}_1$ are

$$\text{Var}(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}}, \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}.$$

Replacing σ by its estimate s gives an estimate for $\sigma_{\hat{\beta}_1}$ (the estimated standard deviation, i.e., estimated standard error, of $\hat{\beta}_1$):

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}.$$

- The estimator $\hat{\beta}_1$ has a normal distribution.
- $T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}}$ has a t distribution with $(n-2)$ df.

Hypothesis Testing for β_1

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

Null hypothesis: $H_0 : \beta_1 = \beta_{10}$

Test statistic: $T = \frac{\hat{\beta}_1 - \beta_{10}}{S_{\hat{\beta}_1}}$ $\quad S_{\hat{\beta}_1} = \frac{S}{\sqrt{S_{xx}}}$ $\quad s = \sqrt{\frac{SSE}{n-2}}$

Test statistic value: $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$ $\quad S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

Alternative Hypothesis	P-value Determination
$H_a : \beta_1 > \beta_{10}$	Area under the t_{n-2} curve to the right of t
$H_a : \beta_1 < \beta_{10}$	Area under the t_{n-2} curve to the left of t
$H_a : \beta_1 \neq \beta_{10}$	2(area under the t_{n-2} curve to the right of $\lvert t \rvert$)

A CI for the slope β_1 of the true regression line with a confidence level of $100(1-\alpha)\%$ is

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_1},$$

where $-$ gives the lower limit and $+$ the upper limit of the interval. An upper or lower confidence bound can also be obtained by retaining the appropriate sign ($+$ or $-$) and replacing $\frac{\alpha}{2}$ by α.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 75.21243    2.98363  25.208 9.22e-12 ***
iodine      -0.20939    0.03109  -6.734 2.09e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.564 on 12 degrees of freedom
Multiple R-squared:  0.7908,	Adjusted R-squared:  0.7733
F-statistic: 45.35 on 1 and 12 DF,  p-value: 2.091e-05
```

$\hat{\beta}_0$ $\hat{\beta}_1$ $s_{\hat{\beta}_1}$ P value

$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$

Chapter 12.4

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, where x^* is some fixed value of x.

1. The mean value of $\hat{Y}$ is $E[\hat{Y}] = E[\hat{\beta}_0 + \hat{\beta}_1 x^*] = \beta_0 + \beta_1 x^*$. Thus, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator for $\beta_0 + \beta_1 x^*$ (i.e., for $\mu_{Y|x^*}$).
2. The variance of $\hat{Y}$ is

$$\mathrm{Var}(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right],$$

   The estimated standard deviation of $\hat{Y}$, denoted by $s_{\hat{Y}}$, is:

$$s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}.$$

3. The estimator $\hat{Y}$ has a normal distribution.
4. $T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}}$ has a t distribution with $(n-2)$ df.

A CI for $\mu_{Y|x^*}$, the expected value of Y when $x = x^*$, with a confidence level of $100(1-\alpha)\%$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\frac{\alpha}{2}, n-2} s_{\hat{Y}},$$

where $-$ gives the lower limit and $+$ the upper limit of the interval. An upper or lower confidence bound can also be obtained by retaining the appropriate sign ($+$ or $-$) and replacing $\frac{\alpha}{2}$ by α.

A PI for a future Y observation to be made when $x = x^*$, with a confidence level of $100(1-\alpha)\%$ is

$$\begin{aligned} &\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2}, n-2} \sqrt{s^2 + s^2_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} \\ &= \hat{y} \pm t_{\frac{\alpha}{2}, n-2} \sqrt{s^2 + s^2_{\hat{Y}}}. \end{aligned}$$