Start-Tech Academy

Simple Linear Regression

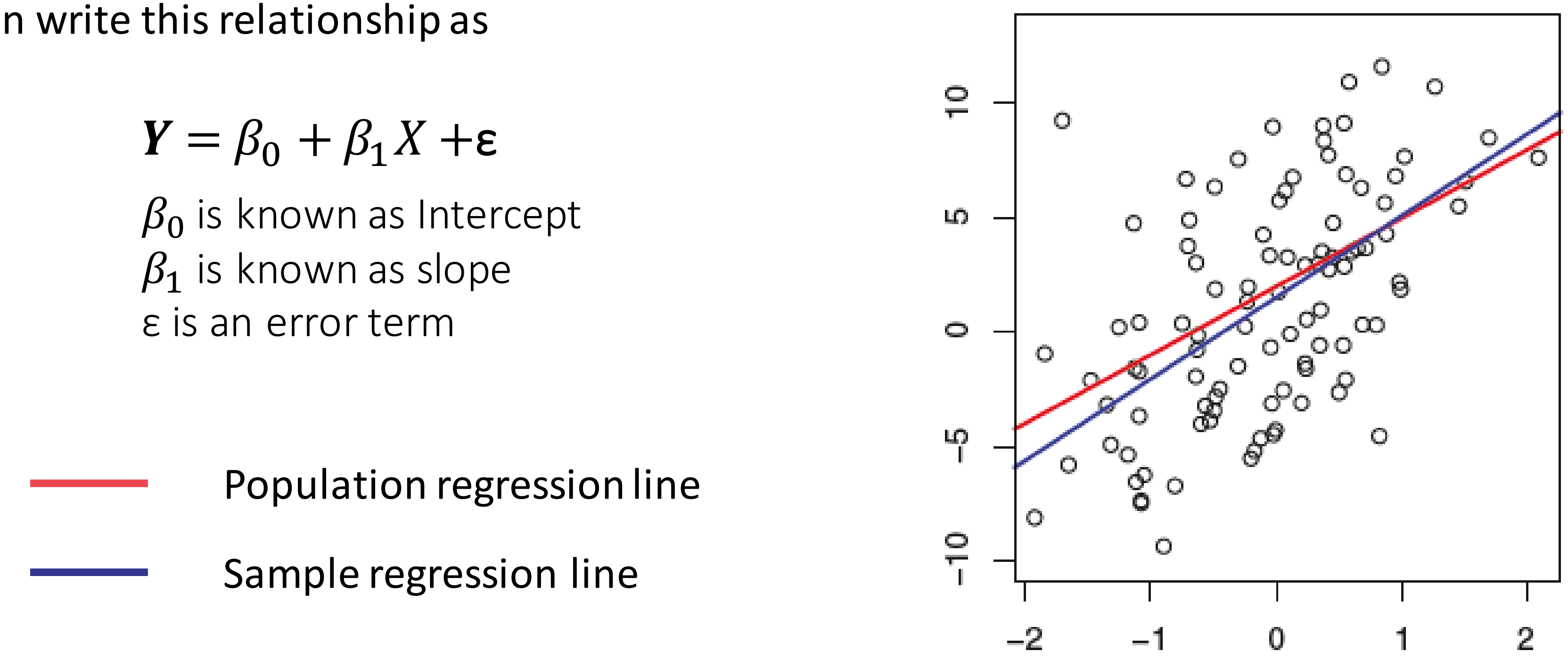
we assume that the true relationship between X and Y takes the form Y = f(X) + ε for some unknown function f, where ε is a mean-zero random error term.

Assessing the Accuracy

If f is to be approximated by a linear function, then we can write this relationship as

$$\boldsymbol{Y} = \beta_0 + \beta_1 X + \varepsilon$$

β_0 is known as Intercept
β_1 is known as slope
ε is an error term



Simple Linear Regression

Standard error In Coefficients

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] \qquad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

$$\sigma^2 = Var(\varepsilon)$$

$\sigma 2$ is not known, but can be estimated from the data. This estimate is known as the *residual standard error (RSE)*

$$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}.$$

There is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2\cdot\mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2\cdot\mathrm{SE}(\hat{\beta}_1)\right]$$

will contain the true value of β_1

Simple Linear Regression

Hypothesis tests

Is there any relationship between X and Y

$$\boldsymbol{Y} = \beta_0 + \beta_1 X$$

- If β_1 is zero, it means there is no relationship

Ho : There is no relationship between X and Y

Ha : There is some relationship between X and Y

$H : \beta_1 = 0$

$Ha : \beta_1 \neq 0,$

Simple Linear Regression

Hypothesis tests

- To disapprove Ho, we calculate T statistics

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- We also compute the probability of observing any value equal to $|t|$ or Larger
- We call this probability the *p-value*
- *A* small p-value means there is an association between the predictor and the response (typically less than 5% or 1 %)

```
Residuals:
    Min      1Q  Median      3Q     Max 
-23.336  -2.425   0.093   2.918  39.434 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -34.6592     2.6421  -13.12   <2e-16
room_num      9.0997     0.4178   21.78   <2e-16
```

Simple Linear Regression

Quality of Fit

RSE

The quality of a linear regression fit is typically assessed using two related quantities: the *residual standard error* (RSE) and the R^2 statistic.

Residual Standard Error

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

- RSE is the average amount that the response will deviate from the true regression line
- RSE is also considered as a measure of lack of fit of the model to the data

```
Residual standard error: 6.597 on 504 degrees of freedom
Multiple R-squared:  0.4848,    Adjusted R-squared:  0.4838
F-statistic: 474.3 on 1 and 504 DF,  p-value: < 2.2e-16
```

Simple Linear Regression

Quality of Fit R^2

The RSE provides an absolute measure of lack of fit of the model to the data.

R^2

- R^2 is the proportion of variance explained
- R^2 always takes on a value between 0 and 1,
- R^2 is independent of the scale of Y.

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

- *TSS - total sum of squares*
- *RSS - residual sum of squares*

```
Residual standard error: 6.597 on 504 degrees of freedom
Multiple R-squared:  0.4848,    Adjusted R-squared:  0.4838
F-statistic: 474.3 on 1 and 504 DF,  p-value: < 2.2e-16
```