Ch2. Simple Linear Regression

- Relationship between 2 variables
- The regression model
- Assumptions
- Estimation and method of least squares
- Inferences concerning β_1 and β_0
- Estimation of the mean of the response variable for a given level of x
- Prediction of new observation
- Analysis of variance approach to regression analysis
- Measures of linear association between x and y

Simple Linear Regression Model

dependent variable/response

slope

random error

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

intercept

independent variable (predictor, explanatory variable)

- Assumptions:
  - $E(\epsilon_i) = 0$,
  - $Var(\epsilon_i) = \sigma^2$
  - $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$
- In matrix notation.

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

Simple Linear Regression Equation

predicted value

estimation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- The simple linear regression equation provides an estimate of the population regression line
- $\hat{\beta}_0$ is the estimated average value of y when the value of x is zero
- $\hat{\beta}_1$ is the estimated change in the average values of y as a result of a one-unit change in x

Simple Linear Regression: an example

A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

- A random sample of 10 houses is selected
- $y=$ house price in \$1000s, $x=$ square feet

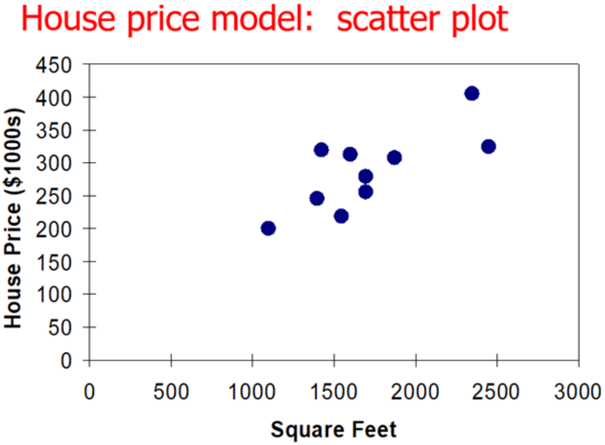
y	x
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

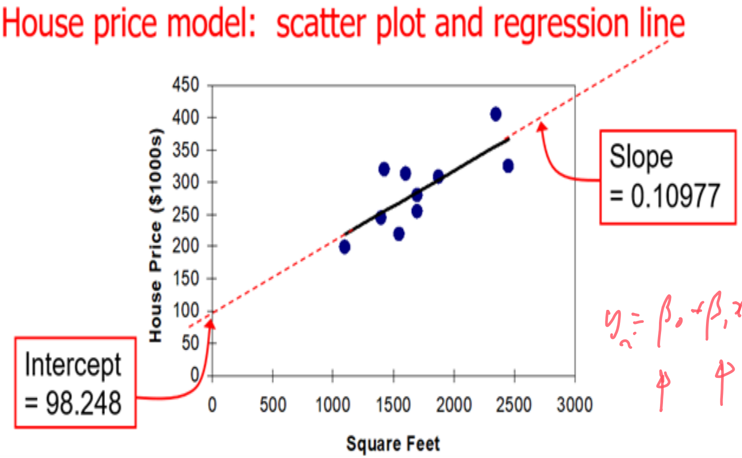
y = c(245, 312, ...)

x = c(1400, ...)

An example: Graphical Presentation



An example: Graphical Presentation



$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$\hat{y} = 98.248 + 0.10977x$$

An example: Interpretation of the intercept, $\hat{\beta}_0$

$$\hat{y} = 98.248 + 0.10977x$$

- $\hat{\beta}_0$ is the estimated average value of y when the value of x is zero (if $x = 0$ is in the range of observed x values)
- Here, no houses had 0 square feet, so $\hat{\beta}_0 = 98.248$ just indicates that, for houses within the range of sizes observed, \$98,248 is the portion of the house price not explained by square feet.

An example: Interpretation of the Slope Coefficient, $\hat{\beta}_1$

$$\hat{y} = 98.248 + 0.10977x$$

- $\hat{\beta}_1$ measures the estimated change in the average value of y as a result of a one-unit change in x
  - Here, $\hat{\beta}_1 = .10977$ tells us that the average value of a house increases by .10977(k)=$109.77, on average, for each additional one square foot of size.

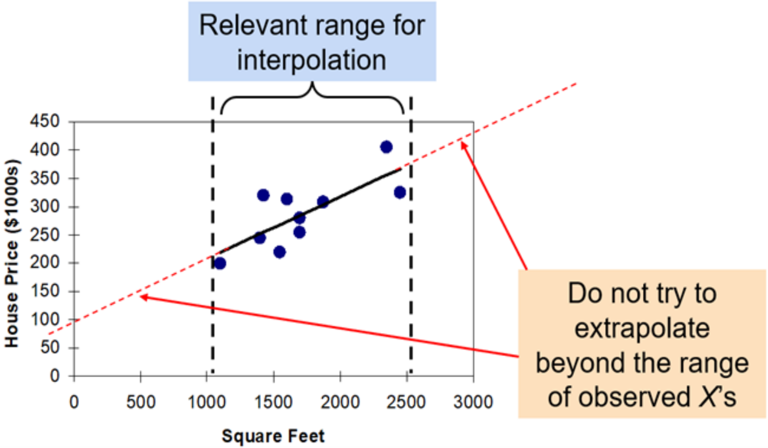
An example: Predictions using Regression Analysis

- Predict the price for a house with 2000 square feet:

$$\hat{y} = 98.25 + 0.10977 \times 2000 = 317.85$$

- The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

An example: Interpolation vs. Extrapolation



When using a regression model for prediction, only predict within the relevant range of data unless you have further information.

Estimation: Method of Least Squares

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by finding the values of β_0 and β_1 that minimize the sum of the squared differences between y and $\hat{y}$:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Remark 2.1

- Solutions:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Comparing $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ with $r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$

Estimation of error terms variance σ^2

$\hat{y}_i$

- The estimator of σ^2 is

$$S^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2}$$

- S^2 is an unbiased estimator of σ^2

$E\, S^2 = \sigma^2$

Estimation: Method of Maximum Likelihood

- The simple linear regression model with normal error

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2) \ i = 1, 2, \ldots, n,$$

Remark 2.2

- The likelihood of the above model
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by maximising the above likelihood
- MLEs:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- The estimator of σ^2 is $\frac{SSE}{n} = \frac{n-2}{n} S^2$.

approximate unbiased estimate.

Remark 2.2 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

$\Leftrightarrow y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \cdots, n$

MLE : step 1. $L = \prod_{i=1}^{n} p(y_i \mid \beta_0, \beta_1, \sigma^2)$

step 2. $\max_{\beta_0, \beta_1, \sigma^2} \log L \Leftrightarrow \max_{\beta_0, \beta_1, \sigma^2} \sum_{i=1}^{n} \log p(y_i \mid \beta_0, \beta_1, \sigma^2)$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{\bar{y}\overbrace{\sum_{i=1}^{n}(x_i - \bar{x})}^{0}}{S_{xx}}$$

$$= \sum_{i=1}^{n} c_i y_i,$$

$$c_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

$$\sum c_i = 0$$

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} c_i (E y_i) = \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \underbrace{\sum_{i=1}^{n} c_i}_{} + \beta_1 \sum_{i=1}^{n} c_i x_i$$

$$= \beta_1 \boxed{\frac{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = 1$$

$$= \beta_1$$

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i^2 \cdot Var(y_i)$$

$$= \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \sigma^2 / S_{xx}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right). \quad ☆$$

Similarly

$$☆ \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$$

prove by yourself ☺

$$\hat{\sigma}^2_{MLE} = S^2_{MLE} = \frac{\sum (y_i - \hat{y}_i)^2}{n} = \frac{n-2}{n} S^2$$

&

$$\frac{n \cdot S^2_{MLE}}{\sigma^2} = \frac{(n-2) S^2}{\sigma^2} \sim \chi^2_{n-2}$$

S^2 and $(\hat{\beta}_0, \hat{\beta}_1)$ are independent!

Proof will be discussed later in the course!

Estimation: Method of Maximum Likelihood

- MLE of β_0 = LSE of β_0 and is unbiased
- MLE of β_1 = LSE of β_1 and is unbiased
- MLE of σ^2 is less than the unbiased estimator of σ^2, but is asymptotically unbiased

Distribution of $\hat{\beta}_1$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

- Assumptions
  - x_i's are known constants,
  - $\epsilon_i \sim N(0, \sigma^2)$ independently for $i = 1, 2, \ldots, n$
- Therefore, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- $$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n c_i y_i$$

  where $c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$, and then $\hat{\beta}_1$ follows a normal distribution.
- $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / S_{xx})$.

Testing (Two-sided test of β_1)

C.I. of β_1 ?

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

- $$H_0 : \beta_1 = 0 \text{ (no linear relationship) } v.s.$$
  $$H_1 : \beta_1 \neq 0 \text{ (linear relationship does exist between } x \text{ and } y\text{)}$$

- Test statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{S/S_{xx}^{1/2}} \sim t_{n-2} \text{ if } H_0 \text{ is true}$$

Remark 2.3

- Decision rule: reject H_0 if $|t| > t_{\alpha/2, n-2}$.

Remark 2.3

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$$

or

$$\frac{\hat{\beta}_1 - \beta_1}{(\sigma^2/S_{xx})^{1/2}} \sim N(0,1)$$

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}, \quad \hat{\beta}_1 \text{ and } S^2 \text{ are independent}$$

$$\Rightarrow \frac{\hat{\beta}_1 - \beta_1}{(\sigma^2/S_{xx})^{1/2}} \Bigg/ \left(\frac{(n-2)S^2}{\sigma^2} \Big/ n-2\right)^{1/2} \sim t_{n-2}$$

$$\boxed{\frac{\hat{\beta}_1-\beta_1}{s/S_{xx}^{1/2}} \sim t_{n-2}}$$

Under H_0, $\beta_1=0$, $t=\dfrac{\hat{\beta}_1}{s/S_{xx}^{1/2}} \overset{H_0}{\sim} t_{n-2}$

Reject H_0 if $|t| > t_{\alpha/2,\,n-2}$

or

P-value $= Pr\left(|t_{n-2}| \geq t\right)$

C.I. of β_1: $P\left(\left|\dfrac{\hat{\beta}_1-\beta_1}{s/S_{xx}^{1/2}}\right| \leq t_{\alpha/2,\,n-2}\right) = 1-\alpha$

$\Rightarrow$ C.I. with $(1-\alpha)$: $\hat{\beta}_1 \pm t_{\alpha/2,\,n-2}\cdot\dfrac{s}{S_{xx}^{1/2}}$.

Two-sided test and confidence interval of β_1

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

- $$H_0 : \beta_1 = k \ v.s. H_1 : \beta_1 \neq k \text{ (} k \text{ is a constant)}$$
  - What are the test statistic and decision rule?
- What are the confidence interval of β_1?

Distribution of $\hat{\beta}_0$

R: lm(y ~ x)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \ \epsilon_i \sim N(0, \sigma^2) \ i = 1, 2, \ldots, n,$$

- $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$.
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ also follows a normal distribution

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right]\right)$$

C.I. of β_0 ?

Estimation of the mean of the response variable for a given level of x

- Example
  - y (in \$000) – house price, x (square feet) – house size
  - Estimate the average house price for houses with 2000 square feet.
- Let x_h be the level of x for which we wish to estimate the mean response, then

Remark 2.4

$$y_h = \beta_0 + \beta_1 x_h + \epsilon_h,$$

the mean response is $E(y_h) = \beta_0 + \beta_1 x_h$.
- The estimation of $E(y_h)$ is $\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$, with distribution

$$\hat{y}_h \sim N\left(\beta_0 + \beta_1 x_h, \underbrace{\sigma^2 \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]}\right)$$?

We want to estimate mean response of $\underbrace{\beta_0 + \beta_1 x_h}_{E(Y_h)}$

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h \sim N(\quad , \quad)$$

$$\frac{\hat{Y}_h - EY_h}{\left[\sigma^2\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)\right]^{1/2}} \bigg/ \left[\frac{(n-2)S^2}{\sigma^2} \Big/ n-2\right]^{1/2} \sim t_{n-2}$$

or:
$$\frac{\beta_0 + \beta_1 x_h - \hat{y}_h}{S\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]^{1/2}} \sim t_{n-2}$$

C. I. of $EY_h = \beta_0 + \beta_1 x_h$ is : $\hat{y}_h \pm t_{\alpha/2, n-2}\, S\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{xx}}\right]^{1/2}$
$(1-\alpha)$

Confidence interval for $\mathrm{E}(y_h)$

$$\mathrm{E}(y_h) - \hat{y}_h \sim N\left(0, \sigma^2 \left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]\right)$$

Two-sided $100(1-\alpha)\%$ C.I. for $\mathrm{E}(y_h)$ is

$$\left(\hat{y}_h - t_{\alpha/2,n-2} S \sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}, \hat{y}_h + t_{\alpha/2,n-2} S \sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}\right)$$

Prediction of a new observation y_h

- Example
  - y (in \$000) – house price, x (square feet) – house size
  - Estimate the house price for **an individual** house with 2000 square feet.
- It means we wish to estimate the response y_h given x_h

$$y_h = \beta_0 + \beta_1 x_h + \epsilon_h,$$

Remark 2.5

- The estimation of y_h is still $\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$, but

$$y_h - \hat{y}_h \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]\right) \quad (*)$$

Remark 2.5 Prediction of a new observation at $x = x_h$

$$y_h = \beta_0 + \beta_1 x_h + \varepsilon_h$$

$$= E y_h + \varepsilon_h, \qquad \varepsilon_h \sim N(0, \sigma^2)$$

$$Y_h - \hat{Y}_h = (EY_h - \hat{Y}_h) + \varepsilon_h$$

independent.

$$\sim N(0, \underbrace{Var(EY_h - \hat{Y}_h)}_{\text{Remark 2.4}} + \sigma^2)$$

Then, we proved (*)

* Prediction of $Y_{h(new)}$ and EY_h are the same.

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

* But the variances of

$$\left.\begin{array}{l} Y_{h(new)} - \hat{Y}_{h(new)} \\ \text{and} \quad E(Y_h) - \hat{Y}_h \end{array}\right\} \text{different!}$$

* Confidence interval for $Y_{h(new)}$ is wider than the c.I. for $E(Y_h)$

Confidence interval for a new observation y_h

Two-sided $100(1-\alpha)\%$ C.I. for y_h is

$$\left(\hat{y}_h - t_{\alpha/2,n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}},\right.$$

$$\left.\hat{y}_h + t_{\alpha/2,n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}\right)$$

We also call it as a **predictive interval**.

Analysis of variance approach to regression analysis

Remark 2.6

- Partitioning of Total Sum of Squares (SST)

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

where SSE=sum of squares of residual, SSR=sum of squares due to regression.

- OR

Total Variation = Unexplained Variation + Explained Variation

Remark 2.6

Total variation of y_i (without considering the model)

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

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

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

$$= \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\text{sum squares of residual}} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{\text{sum of squares due to regression}}$$

sum squares of residual — unexplained variation

sum of squares due to regression — explained variation by the model

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

To prove it by yourself!

$\stackrel{?}{=}$ SSE + SSR

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSE+SSR}$$

— % of the variation can be explained by the model

— coefficient of determination

Test $\quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (*)$

$H_0: \beta_1 = 0$ (the model with no covariate fits the data as well as the model (*))

$H_1: \beta_1 \neq 0$ (model (*) with the covariate fits the data better than the intercept-only model)

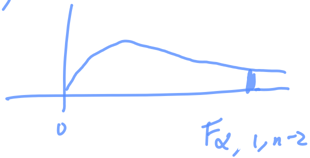
use a F-test or Analysis of Variance Table.

	(SS)	(d.f)	(M.S)	F
Regression	SSR	1	MSR = SSR/1	MSR/MSE
ERROR	SSE	n-2	MSE = SSE/n-2	
Total	SST	n-1		

ANOVA

under $H_0: \beta_1 = 0$ $\quad \dfrac{SSR/1}{SSE/(n-2)} = \dfrac{MSR}{MSE} \overset{H_0}{\sim} F_{1,n-2}$

Reject H_0 if $F > F_{(\alpha, 1, n-2)}$



OR

$p\text{-value} = P(F_{1,n-2} \geq F)$

Analysis of variance (ANOVA) table

	Sum of Squares (SS)	Degrees of freedom (df)	Mean squares (MS)	F
Regression	SSR	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	SSE	n-2	$MSE = \frac{SSE}{(n-2)}$	
Total	SST	n-1		

- Test $H_0 : \beta_1 = 0$ (**no linear relationship**) v.s. $H_1 : \beta_1 \neq 0$ (**linear relationship does exist** between y and x)
- Test statistics $F = \frac{MSR}{MSE} \sim_{H_0} F_{1,n-2}$
- Reject H_0 if $F > F_{\alpha,1,n-2}$.

The coefficient of determination

- The coefficient of determination OR R-squared is defined

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- The proportion of the variation can be explained by the model: $0 \leq R^2 \leq 1$.
- Coefficient of correlation (true for simple linear regression only)

$$r = \pm\sqrt{R^2}$$