Regression Analysis

Regression Analysis

(1) Relationship between degrees Fahrenheit and degrees Celsius:

$$F = \frac{9}{5}C + 32 \qquad \text{(deterministic)}$$

(2) Circumference $= \pi \times \text{diameter} \Rightarrow C = 2\pi r$ $\quad$ (deterministic)

(3) Height and weight of students (is there a perfect relationship?)

(4) Driving speed and gas mileage (is there a deterministic relation?)

(5) Fertilizer and crop yield (production)

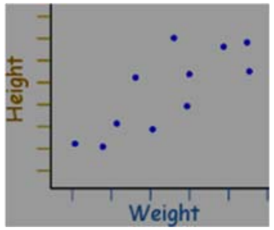
(6) Drug dosage and time to get cured

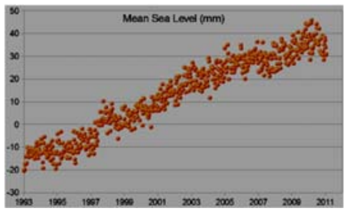
(7) Income and expenditure of a group of persons

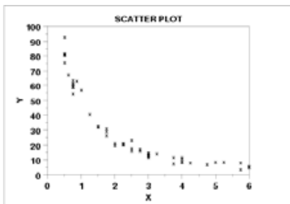
(8) Sunshine hours/temperature and icecream sale

(9) Age of car and its sale price

Let's Observe…linear or non-linear?

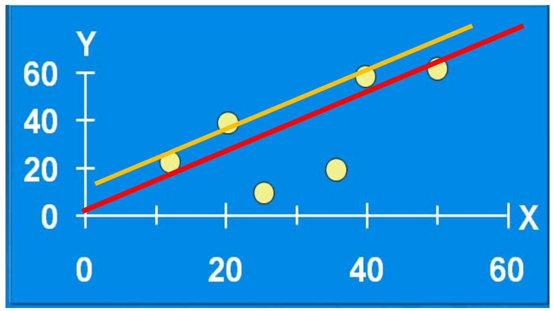






Purpose: Linear Regression

- For the following observed data, how can we get the "*best-fit*" line?
- What would be the equation?



How to Formulate?

(1) Suppose, X is NOT a RV, rather a mathematical variable.

e.g., let x: depth of water, Y: the water temperature.

Then can we model the water temperature Y, as a function of x?

(2) Armand's (pizza parlour) most successful locations are near college campuses. The manager thinks that their sale Y depends on the number of students (x).

(3) Thus, aren't we dealing with a conditional variable $Y|x$?

(4) This $Y|x$ will have a mean $\mu_{Y|x}$ (a function of x).

(5) Can I express (linearly) this as $\mu_{Y|x} = \beta_0 + \beta_1 x$?

Simple Linear Regression Model: $Y = \beta_0 + \beta_1 x + \varepsilon$

Simple Linear Regression Equation: $\mu_{Y|x} = \beta_0 + \beta_1 x$

Estimated Simple Linear Regression Equation: $\hat{\mu}_{Y|x}\,(or, \hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x$

Simple Linear Regression (SLR)

Simple linear regression (regression means 'act of going back', 'return', or 'reversion') is a statistical method that allows us to summarize and study relationships between **two continuous (quantitative) variables**:

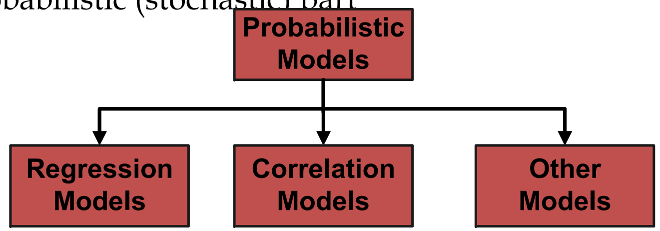
- One variable, denoted by x, is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted by y, is regarded as the response, outcome, or dependent variable.

Why is it called "Simple Linear Regression" model? What is a model?

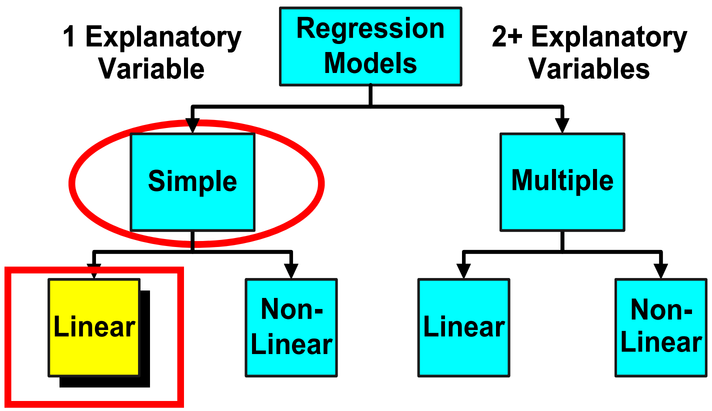
Why simple? Why linear? What is regression?

Mathematical Model

1. Often, they describe relationship between variables
2. Two parts:
   - Functional part
   - Probabilistic (stochastic) part

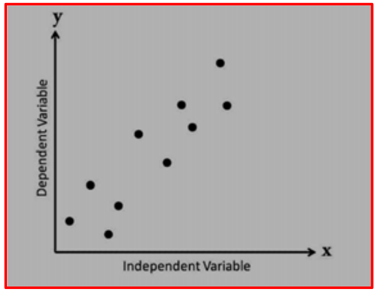


Regression Model



Formulation: Simple linear regression model (SLRM)

- In a regression study, it is useful to plot the data points in *xy*-plane. Such a plot is called the *scattergram (scatter diagram).*
- We do not expect the points to lie exactly on a straight line. However, if linear regression is applicable, then they should exhibit a linear trend.



Formulation: SLRM

- Since we do not know the true values of β_0 and β_1 (WHY??), we shall not know the true value of ε_i (the vertical distance from (x_i, y_i) to the true regression line).
- Letting $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimates of β_0 and β_1 respectively, the estimated line of regression takes the form,

$$\hat{\mu}_{Y|x}\left(or, \hat{y}\right) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Simple Linear Regression Model: $Y|x = \beta_0 + \beta_1 x + \varepsilon$

Simple Linear Regression Equation: $\mu_{Y|x} = \beta_0 + \beta_1 x$

$Var\left(Y|x\right) = Var\left(\beta_0 + \beta_1 x + \varepsilon\right) = \sigma^2$

Estimated Simple Linear Regression Equation: $\hat{\mu}_{Y|x}\left(or, \hat{y}\right) = \hat{\beta}_0 + \hat{\beta}_1 x$

SLE: Model Assumption

1. $E\left(\varepsilon_i\right) = 0$
2. $V\left(\varepsilon_i\right) = \sigma^2$, same for all values of x
3. ε_i and ε_j are uncorrelated. Thus for $i \neq j; \operatorname{cov}\left(\varepsilon_i, \varepsilon_j\right) = 0$. Thus

$E\left(y_i\right) = \beta_0 + \beta_1 x_i$; $\quad V\left(\varepsilon_i\right) = \sigma^2$; and y_i and y_j are uncorrelated

Under additional assumption

ε_i is normally distributed $\quad \varepsilon_i \sim N\left(0, \sigma^2\right)$

$\Rightarrow y$ is also normally distributed

Hence, ε_i and ε_j are not only uncorreleted but independent also

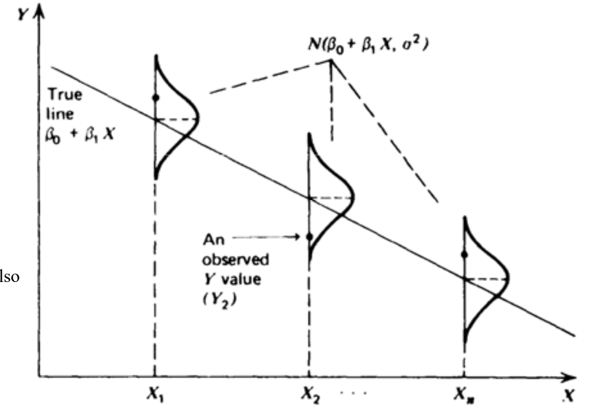
Least squares estimation

Under additional assumption that

ε_i is normally distributed $\quad \varepsilon_i \sim N\left(0, \sigma^2\right)$

$\Rightarrow y$ is also normally distributed

Hence, ε_i and ε_j are not only uncorreleted but independent also



Each response observation is assumed to come from a normal distribution cantered vertically at the level implied by the model, with identical variance σ^2

Least-Squares Estimation

- Parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ are determined by method of least squares.
- Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that we minimize the sum of the squares of the residuals.
- Sum of the squares of the residual errors about the estimated regression line is given by

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

Sum of squares of errors $(SSE) = S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

LS estimators of β_0, β_1, say $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0$$

- Simplifying, we get **normal equations**

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\underbrace{\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}}_{N} \underbrace{\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}}_{U}$$

Least-Squares Estimates for β_0 and β_1

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

$\because$ other term removed from numerator is $= 0; \sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\left.\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \end{aligned}\right\}$$

- Since the denominator of eq. for β_1 is the corrected sum of squares of the x_i and the numerator is the corrected sum of cross products of x_i and y_i, we may write these quantities in a more compact notation as

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

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

Useful properties of LS fit

- Sum of the residuals in any regression model that contains an intercept β_0 is always zero:

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

- Sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i$:

$$\boxed{\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{y}_i}$$

- Least-squares regression line always passes through the **centroid $(\bar{x}, \bar{y})$** of the data.
- Sum of the residuals weighted by the corresponding value of the regressor variable always equals zero

$$\boxed{\sum_{i=1}^{n} x_i\hat{e}_i = 0}$$

- Sum of the residuals weighted by the corresponding fitted value always equals zero

$$\boxed{\sum_{i=1}^{n} \hat{y}_i\hat{e}_i = 0}$$

Coefficient of Determination

Sum of Squares due to Error (SSE): $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Total Sum of Squares (SST): $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$

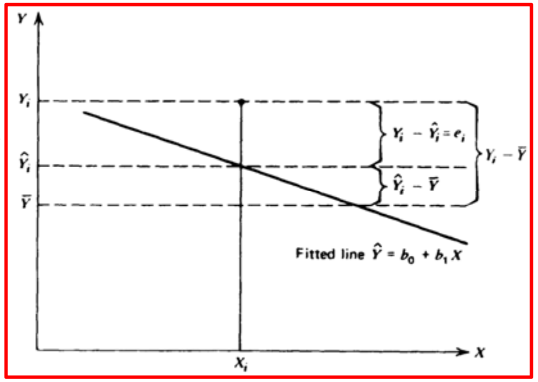
Sum of Squares due to Regression (SSR): $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

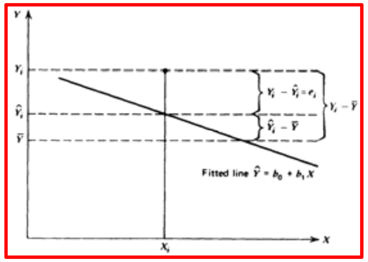
Relation: $SST = SSR + SSE$

$$\frac{SST}{SST} = 1 = \underbrace{\frac{SSR}{SST}}_{=r^2} + \frac{SSE}{SST}$$

Coefficient of determination: $r^2 = \frac{SSR}{SST}$

Sample correlation coefficient: $r = (\text{sign of } b_1)\sqrt{r^2}$





- We can write: $\boxed{(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)}$
- Squaring both sides and taking sum from i = 1 to n (and noting that the cross product term is equal to zero), we can write.

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

- Using relationship, cross-product term (CPT) = 0:

Using: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$; $\quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$;

$$\hat{y} = \bar{y} + \hat{\beta}_1(x - \bar{x})$$

$$\hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$$

$$y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})$$

$$\begin{aligned}
2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= 2\sum_{i=1}^{n}\hat{\beta}_1(x_i - \bar{x})\left[(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})\right] \\
&= 2\hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1\left[(x_i - \bar{x})((x_i - \bar{x}))\right] \\
&= 2\hat{\beta}_1\left[S_{xy} - \hat{\beta}_1 S_{xx}\right] \\
&= 0 \qquad \because \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}
\end{aligned}$$

Analysis and testing in regression

After obtaining the least-squares fit, a number of questions come to mind:

1. How well does this equation fit the data?
2. Is the model likely to be useful as a predictor?
3. Are any of the basic assumptions (such as constant variance and uncorrelated errors) violated, and if so, how serious is this?

Properties of LS estimator

- LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combination of the original observations.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1.

$$E\left[\hat{\beta}_0\right] = \beta_0 \qquad E\left[\hat{\beta}_1\right] = \beta_1$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i; \qquad c_i = \frac{(x_i - \bar{x})}{S_{xx}}$$

$$E\left[\hat{\beta}_1\right] = E\left[\sum_{i=1}^{n} c_i y_i\right] = \sum_{i=1}^{n} c_i E[y_i] = \sum_{i=1}^{n} c_i [\beta_0 + \beta_1 x_i]$$

$$\sum_{i=1}^{n} c_i [\beta_0 + \beta_1 x_i] = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \qquad \because E[\varepsilon_i] = 0$$

- It can be shown that

$$\sum_{i=1}^{n} c_i = 0; \sum_{i=1}^{n} c_i x_i = 1$$

$$\therefore E\left[\hat{\beta}_0\right] = \beta_0;\ E\left[\hat{\beta}_1\right] = \beta_1$$

- LS gives BLUE (**B**est **L**inear, **U**nbiased **E**stimators)

Variances of estimators

- Variances are give as

$$\sigma^2_{\hat{\beta}_1} = Var\left[\sum_{i=1}^{n} c_i y_i\right] = \sum_{i=1}^{n} c_i^2 \sigma^2_{y_i}; \qquad \text{Assuming } \sigma^2_{y_i} = \sigma^2$$

$$\sigma^2_{\hat{\beta}_1} = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{S_{xx}}\right]^2 = \frac{\sigma^2}{S_{xx}}$$

$$Var\left[\hat{\beta}_0\right] = Var\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] = Var[\bar{y}] + \bar{x}^2 Var\left[\hat{\beta}_1\right] - 2\bar{x}\,\mathrm{covar}\left[\bar{y}, \hat{\beta}_1\right]$$

$$\sigma^2_{\hat{\beta}_0} = \sigma^2_{\bar{y}} + \bar{x}^2 \sigma^2_{\hat{\beta}_1} - 2\bar{x}\,\underbrace{\sigma_{\bar{y}\hat{\beta}_1}}_{=0} \qquad \text{(can be shown)}$$

$$\sigma^2_{\hat{\beta}_0} = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad \because \sigma^2_{\bar{y}} = \frac{\sigma^2}{n}$$

Estimation of σ^2

- Estimate of σ^2 is required to test hypotheses and construct interval estimates pertinent to the regression model.
- Ideally, we would like this estimate not to depend on the *adequacy* of the fitted model. This is only possible:
  - when there are several observations on y for at least one value of x or
  - when prior information concerning σ^2 is available.
- When this approach cannot be used, estimate of σ^2 is obtained from the *residual* or *error sum of squares.*

$$SS_{\text{Res}} = \sum_{i=1}^{n} v_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

- It can be shown that an **unbiased estimator of σ^2**

$$\hat{\sigma}^2 = \frac{V^T P V}{n-2} = \frac{SS_{\text{Res}}}{n-2} = MS_{\text{Res}} \text{(Residual Mean Square)}$$

- Square root of $\hat{\sigma}^2$ is called **standard error of regression** and is model dependent.
- Because $\hat{\sigma}^2$ depends on the residual sum of squares, any violation of the assumptions on the model errors or any misspecification of the model form may seriously damage the usefulness of $\hat{\sigma}^2$as an estimate of σ^2. Because $\hat{\sigma}^2$ is computed from the regression model residuals, we say that it is a model-dependent estimate of σ^2 (*a posteriori reference variance*)

Testing for Significance

- We are often interested in testing hypotheses and constructing confidence intervals about the model parameters. It requires that we make the additional assumption that the model errors ε_i are normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with mean 0 and variance σ^2, abbreviated NID(0,σ^2). These assumptions are also checked through residual analysis.

- **Procedure to test the hypothesis that the slope equals a constant, say β_{10}.**
  - The appropriate hypotheses for a **two-tailed test** are:

$$H_0 : \beta_1 = \beta_{10}$$
$$H_1 : \beta_1 \neq \beta_{10}$$

  - Since the errors **ε_i ~ NID(0, σ^2)**, the observations **y_i ~ NID($\beta_0 + \beta_1 x_i$, σ^2).** Now $\hat{\beta}_1$ is a linear combination of the observations, so is normally distributed with mean β_1 and variance σ^2/S_{xx} using the mean and variance of $\hat{\beta}_1$found earlier.
  - The testing statistic

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1)$$

- If σ^2 were known, we could use Z_0 to test the above hypotheses. Typically, σ^2 is unknown. It can be shown that **MS_{Res}** is an unbiased estimator of σ^2. Further:
  - **$(n-2)\ MS_{Res}/\sigma^2$** follows a χ^2_{n-2} distribution and
  - **MS_{Res}** and $\hat{\beta}_1$ are independent.
- *The t* statistic given as t_0, with DoF same as associated with MS_{Res})

$$\boxed{t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{\sigma}_{\beta_1}} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \sim t_{n-2}}$$

  and implies that the null hypothesis H_0: $\beta_1 = \beta_{10}$ is true.
- Thus, the ratio t_0 is the test statistic used to test ***H_0: $\beta_1 = \beta_{10}$***.
- The test procedure computes t_0 and compares the observed value of t_0 from above equation with the upper $\alpha/2$ percentage point of the t_{n-2} distribution ($t_{\alpha/2,\, n-2}$). This procedure rejects the null hypothesis if

$$\boxed{|t_0| > t_{\alpha/2, n-2}}$$

- Alternatively, the p-value based approach can also be used.

- **Procedure for testing the intercept (same as for slope)**

$$\boxed{\begin{aligned} H_0 &: \beta_0 = \beta_{00} \\ H_1 &: \beta_0 \neq \beta_{00} \end{aligned}}$$

  - Use the statistic

$$\boxed{t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\hat{\sigma}_{\beta_1}} = \frac{\hat{\beta}_1 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}}$$

- **Testing for significance of regression**
  - A very important special case of the hypotheses

$$\boxed{\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}}$$

  - The test procedure for H_0: $\beta_1 = 0$ is developed by using *t-statistic* and simply using $\beta_{10} = 0$ or

$$\boxed{t_0 = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}}$$

  - The null hypothesis of significance of regression is rejected if

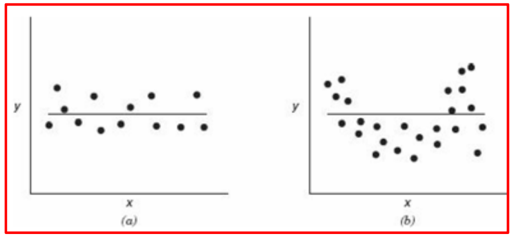
$$\boxed{|t_0| > t_{\alpha/2, n-2}}$$

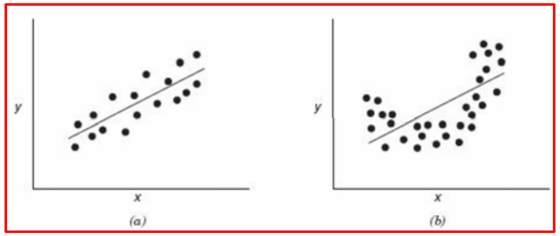
Confidence Interval for β_1

$$(1-\alpha)\times 100\% \text{ confidence interval for } \beta_1 \text{ is } \left(\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2}\, \sigma_{\hat{\beta}_1}\right)$$

v	Tail probability 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850

- Failing to reject H_0: $\beta_1 = 0$: Implies that there is no linear relationship between x and y. This may imply either that x is of little value in explaining the variation in y and that the best estimator of y for any x is $\hat{y} = \bar{y}$ (Left Figure a) or that the true relationship between x and y is not linear (Left Figure b). Therefore, failing to reject H_0: $\beta_1 = 0$ is equivalent to saying that there is no linear relationship between y and x.





Situations when H_0: $\beta_1 = 0$ is not rejected (Left fig) **Situations when H_0: $\beta_1 = 0$ is rejected (right fig)**

If H_0: $\beta_1 = 0$ is rejected: Implies that x is of value in explaining the variability in y. This is illustrated in Figure (Right). However, rejecting H_0: $\beta_1 = 0$ could mean either that the straight-line model is adequate (Right Figure a) or that even though there is a linear effect of x, better results could be obtained with the addition of higher order polynomial terms in x (Right Figure b).

Analysis of variance

- We may also use an **analysis-of-variance approach** to test significance of regression. It is based on a **partitioning of total variability in the response variable *y*** to draw inferences.
- To obtain this partitioning, begin with the identity $$\left(y_i-\bar{y}\right)=\left(\hat{y}_i-\bar{y}\right)+\left(y_i-\hat{y}_i\right)$$
- Squaring both sides and taking sum from $i = 1$ to n.

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2+\underbrace{2\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)\left(y_i-\hat{y}_i\right)}_{\text{cross-product term} = CPT}$$

- Now

$$\underbrace{2\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)\left(y_i-\hat{y}_i\right)}_{\text{cross-product term} = CPT}=2\sum_{i=1}^{n}\hat{y}_i\left(y_i-\hat{y}_i\right)-\bar{y}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)=2\sum_{i=1}^{n}\hat{y}_i e_i-\bar{y}\sum_{i=1}^{n}e_i=0 \qquad \because \sum_{i=1}^{n}e_i=0$$

- Hence

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

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}_{SS_T}=\underbrace{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}_{SS_R}+\underbrace{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}_{SS_{\text{Res}}}$$

$$SS_T=SS_R+SS_{\text{Res}}$$

LHS: **Corrected sum of squares of the observations**, **SS_T**, which measures the total variability in the observations.

Two components of **SS_T** measure, respectively:

(a) **Regression or model sum of squares (SS_R):** amount of variability in the observations y_i accounted for by the regression line

(b) Residual or error sum of squares (SS_{Res}): residual variation left unexplained by the regression line

Further:

$$SS_{\text{Res}}=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\hat{e}_i^2$$

Using: $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$

$$SS_{\text{Res}}=\sum_{i=1}^{n}\hat{e}_i^2=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0+\hat{\beta}_1x_i\right)^2=\sum_{i=1}^{n}y_i^2-n\bar{y}-\hat{\beta}_1S_{xy}$$

But: $SS_T=\sum_{i=1}^{n}y_i^2-n\bar{y}^2=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$

$\therefore$ $SS_{\text{Res}}=SS_T-\hat{\beta}_1S_{xy}\Rightarrow SS_R=\hat{\beta}_1S_{xy}$

Analysis of variance table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R=\hat{\beta}_1S_{xy}$	1	MS_R	MS_R/MS_{Res}
Residual	$SS_{Res}=SS_T-\hat{\beta}_1S_{xy}$	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

- Degree-of-freedom (DoF) breakdown.
  - Total sum of squares, SS_T: $df_T = (n - 1)$ because one degree of freedom is lost as a result of the constraint on the deviations $(y_i - \bar{y})$ $\sum_{i=1}^{n}\left(y_i-\bar{y}\right)$
  - Model or regression sum of squares, SS_R, has $df_R = 1$ degree of freedom because SS_R is completely determined by one parameter, namely, $\hat{\beta}_1$
  - SS_{Res} has $df_{Res} = (n - 2)$ because two constraints are imposed on deviations $\left(y_i-\hat{y}_i\right)$ as a result of estimating $\hat{\beta}_0$ and $\hat{\beta}_1$
  - Note that DoF have additive property: $$df_T=df_R+df_{\text{Res}}$$ $$\left(n-1\right)=1+\left(n-2\right)$$

- Now we can use usual **analysis-of-variance F test** to test the hypothesis H_0: $\beta_1 = 0$.
- It can be shown that

(i) $SS_{\text{Res}}=\left(n-2\right)\dfrac{MS_{\text{Res}}}{\sigma^2}\sim\chi^2_{n-2}$

(ii) If the null hypothesis $H_0:\beta_1=0$ is true, then $\dfrac{SS_{\text{R}}}{\sigma^2}\sim\chi^2_1$

(iii) SS_{Res} and SS_{R} are independent

- Therefore, by the definition of an ***F* statistics**, we have

$$F_0=\frac{SS_R/df_R}{SS_{\text{Res}}/df_{\text{Res}}}=\frac{SS_R/1}{SS_{\text{Res}}/(n-1)}=\frac{MS_R}{MS_{\text{Res}}}\sim F_{1,n-2}$$

$$E\left(MS_{\text{Res}}\right)=\sigma^2 \qquad E\left(MS_{\text{R}}\right)=\sigma^2+\beta_1^2S_{xx}$$

- It can also be shown that
- These expected mean squares indicate that if the observed value of F_0 is large, then it is likely that the slope $\beta_1 \neq 0$.

- It can also be shown that if $\beta_1 \neq 0$, then F_0 follows a non-central F distribution with 1 and $(n-2)$ degrees of freedom and a non-centrality parameter of

$$\boxed{\lambda = \frac{\beta_1^2 S_{xx}}{\sigma^2}}$$

- Non-centrality parameter also indicates that the observed value of F_0 should be large if $\beta_1 \neq 0$.
- Therefore, to test the hypothesis H_0: $\beta_1 = 0$, compute the test statistic F_0 and reject H_0 if

$$\boxed{F_0 \sim F_{\alpha,(1,n-2)}}$$

Source of variation	Sum of squares	Degrees of freedom	Mean square	F0
Regression	$SS_R = \beta_1^2 S_{xy}$	1	MS_R	MS_R / MS_{Res}
Residual	$SS_{\text{Res}} = SS_T - \beta_1^2 S_{xy}$	$(n-1)$	MS_{Res}	
Total	SS_T	$(n-2)$		

Interval estimation for β_0, β_1, and σ^2

- For β_0, β_1, and σ^2, we may also obtain confidence interval estimates of these parameters.
- The width of these confidence intervals is a measure of the overall quality of the regression line.
- If the errors are normally and independently distributed, then the sampling distribution of both $\boxed{\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \text{ and } \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\beta_0}}}$ is t_{n-2} with $(n - 2)$ DoF. Therefore, $(1 - \alpha)$ 100 percent CI are given as:

$$\boxed{\left(\hat{\beta}_1 \pm t_{\frac{\alpha}{2},n-2}\sigma_{\hat{\beta}_1}\right) \text{and} \left(\hat{\beta}_0 \pm t_{\frac{\alpha}{2},n-2}\sigma_{\hat{\beta}_0}\right)}$$

- These CIs have the usual frequentist interpretation. That is, if we were to take repeated samples of the same size at the same x levels and construct, for example, 95% CIs on the slope/intercept for each sample, then 95% of those intervals will contain the true value of β_1/β_0.
- If the errors are normally and independently distributed, sampling distribution $\boxed{(n-2)\frac{MS_{\text{Res}}}{\sigma^2} \sim \chi^2_{n-2}}$
- Hence, $\boxed{P\left(\chi^2_{1-\frac{\alpha}{2},n-2} \leq (n-2)\frac{MS_{\text{Res}}}{\sigma^2} \leq \chi^2_{n-2}\chi^2_{\frac{\alpha}{2},n-2}\right) = 1-\alpha}$
- Thus $(1 - \alpha)$ 100 percent CI on σ^2 is

$$\boxed{\left(\frac{(n-2)MS_{\text{Res}}}{\chi^2_{\frac{\alpha}{2},n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{\text{Res}}}{\chi^2_{1-\frac{\alpha}{2},n-2}}\right)}$$