Chapter 14: Simple Linear Regression Analysis
STAT 2601 – Business Statistics

Esam Mahdi

School of Mathematics and Statistics
Carleton University

January 24, 2026

Learning Objectives

By the end of this chapter, you should be able to:

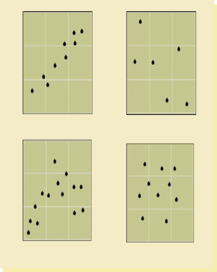
1. **Understand correlation:** Interpret scatter plots and calculate/interpret Pearson's r
2. **Formulate the simple linear regression model:** $y = \beta_0 + \beta_1 x + \varepsilon$ with LINE assumptions
3. **Estimate parameters:** Use least squares to find b_0, b_1; interpret slope and intercept
4. **Assess model fit:** Calculate $R^2 = SSR/SSTO$ and interpret explained variation
5. **Test significance:** Perform F-test ($H_0 : \rho^2 = 0$), t-test ($H_0 : \beta_1 = 0$), t-test ($H_0 : \rho = 0$)
6. **Make predictions:** Construct confidence intervals (mean response) and prediction intervals (individual response)
7. **Apply to business:** Use regression output for data-driven decisions; communicate results effectively

What is Correlation (Linear Relationship)?

Correlation measures the **strength (strong/weak)** and **direction (positive/negative)** of a linear relationship between two quantitative variables.

Scatter Plot: Types of Relationships

- **Positive Linear:** Change in X and Y tends to happen in the same direction ($X \uparrow \Rightarrow Y \uparrow$ and $X \downarrow \Rightarrow Y \downarrow$).
  - **Example:** *years of experience* and *salary*.
- **Negative Linear:** Change in X and Y tends to happen in the opposite direction ($X \uparrow \Rightarrow Y \downarrow$ and $X \downarrow \Rightarrow Y \uparrow$).
  - **Example:** *vehicle weight* and *fuel efficiency*.
- **Curvilinear:** The relationship between X and Y is not a straight line; instead, it follows a curve, such as a U-shape or an inverted U-shape (parabola).
  - **Example:** *age* and *physical strength*; strength increases through youth, peaks in adulthood, and gradually declines in old age.
- **No Relationship:** There is no discernible pattern between X and Y. Changes in X do not predict any specific change in Y, resulting in a random "cloud" of data points.
  - **Example:** *coffee consumption* and *shoe size*.



Population Correlation Coefficient (ρ, rho)

The population correlation coefficient (ρ, rho) measures (numerically) the linear relationship between two variables, X, and Y, in an entire population, calculated as the population covariance divided by the product of their population standard deviations:

$$\rho = \frac{\text{Covariability of X and Y}}{(\text{Standard Deviation of X}) \times (\text{Standard Deviation of Y})} = \frac{\text{Cov}(X, Y)}{\sigma_X \times \sigma_y}$$

Sample Correlation Coefficient: Pearson (r)

For a dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the **Pearson Correlation Coefficient**, denoted by r is given by:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{SS_{xy}}{\sqrt{SS_{xx}\, SS_{yy}}}$$

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

where

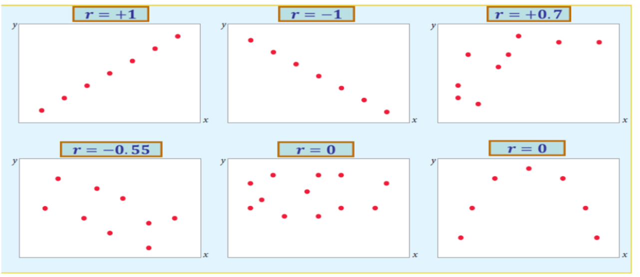
$$s_x^2 = \frac{SS_{xx}}{n-1}, \quad s_y^2 = \frac{SS_{yy}}{n-1}, \quad s_{xy} = \frac{SS_{xy}}{n-1}$$

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

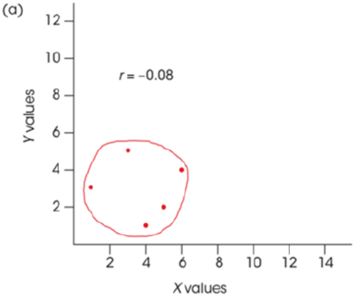
$$SS_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}, \quad SS_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Characteristics of Correlation Coefficient

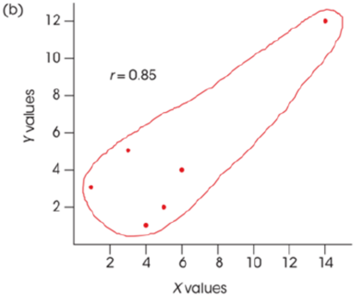
- Unit free and range between -1 and $+1$: $-1 \leq r \leq 1$
- Positive linear relationship: $r > 0$
  - Weak positive linear relationship: $0 < r < 0.5$
  - Strong positive linear relationship: $0.5 \leq r < 1$
- Negative linear relationship: $r < 0$
  - Weak negative linear relationship: $-0.5 < r < 0$
  - Strong negative linear relationship: $-1 < r \leq -0.5$
- Perfect linear relationship: $r = \pm 1$
  - Perfect positive linear relationship: $r = 1$ (all data points fall on a straight line with positive slope)
  - Perfect negative linear relationship: $r = -1$ (all data points fall on a straight line with negative slope)
- No linear relationship: $r = 0$



Outlier: Outliers produce a disproportionately large impact on the correlation coefficient. An outlier is an extremely deviant individual in the sample.



Original Data		
Subject	X	Y
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2

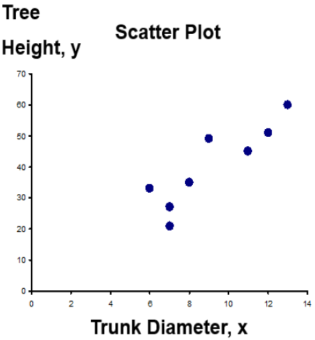


Data with Outlier Included		
Subject	X	Y
A	1	3
B	3	5
C	6	4
D	4	1
E	5	2
F	14	12

Example: Correlation between Trunk Diameter and Tree Height

Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=321	Σ=73	Σ=3142	Σ=14111	Σ=713

Example: Correlation between Trunk Diameter and Tree Height



$$
\begin{aligned}
r &= \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \\
&= \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{\sum x^2 - (\sum x)^2/n}\sqrt{\sum y^2 - (\sum y)^2/n}} \\
&= \frac{(3142) - (73)(321)/8}{\sqrt{713 - \frac{(73)^2}{8}}\sqrt{14111) - \frac{(321)^2}{8}}} \\
&= 0.886
\end{aligned}
$$

$\boldsymbol{r = 0.886}$ $\rightarrow$ relatively strong positive linear relationship between x and y

The Simple Linear Regression Model

Example: Predict the Yearly Revenue Based on Population Size

The Tasty Sub Shop is a restaurant chain that sells franchises to business entrepreneurs. Management is interested in the population regression model that relates yearly revenue (y) to the population size (x) in the franchise's market area. This model represents the average revenue that would be expected for all possible locations with a given population size and can be used to predict expected yearly revenue for future franchise sites.

Probabilistic Model:

$$\text{Yearly revenue} = \beta_0 + \beta_1(\text{Population size}) + \varepsilon$$

The error term ε represents all other influences on yearly revenue besides population size; such as competition, location quality, management skill, local income, advertising, and pure randomness.

Simple Linear Regression (SLR)

Let y and x be two variables observed on experimental units $i = 1, 2, \cdots, n$. **A Simple Linear Regression** model assumes that:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for all } i,$$

where

- β_0 & β_1 are two unknown parameters (regression coefficients) called **intercept** & **slope**, respectively.
- ε is the error term (random error).
- x: called **independent** variable, or **predictor** or **explanatory**.
- y: called **dependent** variable, or **response** or **covariate**.

The LINE Assumptions

1. **L**inearity: The expected value (average) of Y given $X = x$ is linear in the parameters β_0 and β_1

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i$$

2. **I**ndependence: Errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent
3. **N**ormality:

$$\varepsilon_i \sim N(\text{mean } = 0, \text{variance } = \sigma^2)$$

4. **E**qual variance (Homoscedasticity):

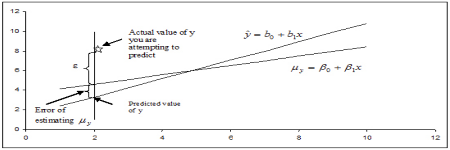
$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \text{for all } i$$

5. x_i's are observed without errors.

Least Squares Estimation (LSE)

Goal: Minimize Sum of Squared Errors (SSE)

$$SSE = \sum_{i=1}^{n}[y_i - \hat{y}_i]^2 = \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2$$



Sample (or Estimated or Predicted) Regression Line

$$\hat{y} = b_0 + b_1 x$$

where:

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

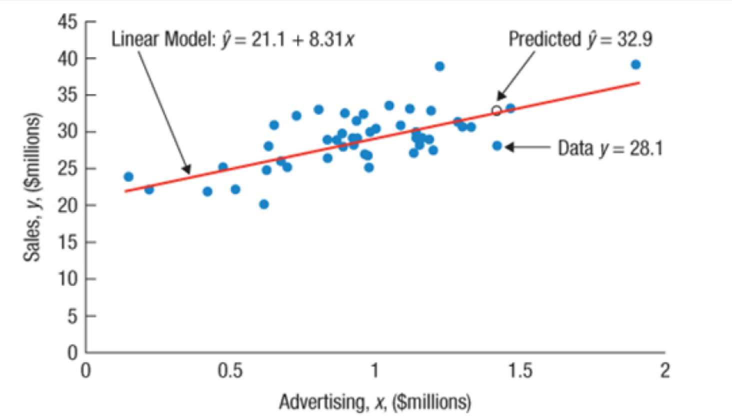
$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

Note: $\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$ is called **residual** and $\sum_{i=1}^{n} e_i = 0$.

Example: Sales Versus Advertising Data

For advertising expenses of $x = \$1.42$ million, the actual sales are $y = \$28.1$ million and the predicted sales are $\hat{y} = 21.1 + 8.31(1.42) = \$32.9$ million. The residual is

$$e = y - \hat{y} = 28.1 - 32.9 = -4.8$$



Example: Tasty Sub Shop

To estimate the population regression model, Tasty Sub Shop collects data from a sample of $n = 10$ existing franchise locations. For each site, they record:

- x: Population size of the market area (in thousands)
- y: Yearly revenue (in thousands of dollars)

Yearly Revenue (y_i) *(Thousands of Dollars)*	**Population Size** (x_i) *(Thousands of Residents)*
527.1	20.8
548.7	27.5
767.2	32.3
722.9	37.2
826.3	39.6
810.5	45.1
1040.7	49.9
1033.6	55.4
1090.3	61.7
1235.8	64.6

Table: Tasty Sub Shop Revenue and Population Data

Calculation Summary for Tasty Sub Shop ($n = 10$)

y_i	x_i	x_i^2	$x_i y_i$
527.1	20.8	432.64	10963.68
548.7	27.5	756.25	15089.25
767.2	32.3	1043.29	24780.56
722.9	37.2	1383.84	26891.88
826.3	39.6	1568.16	32721.48
810.5	45.1	2034.01	36553.55
1040.7	49.9	2490.01	51930.93
1033.6	55.4	3069.16	57261.44
1090.3	61.7	3806.89	67271.51
1235.8	64.6	4173.16	79832.68
$\sum y_i = 8603.1$	$\sum x_i = 434.1$	$\sum x_i^2 = 20,757.41$	$\sum x_i y_i = 403,296.96$

Example: Tasty Sub Shop (Cont.)

- Estimated Slope:

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{403,296.96 - \frac{(434.1)(8,603.1)}{10}}{120,757.41 - \frac{(434.1)^2}{10}} = 15.596$$

- Estimated Intercept:

$$\bar{y} = \frac{\sum y}{n} = \frac{8,603.1}{10} = 860.31, \quad \bar{x} = \frac{\sum x}{n} = \frac{434.1}{10} = 43.41$$

$$b_0 = \bar{y} - b_1\bar{x} = 860.31 - (15.596)(43.41) = 183.31$$

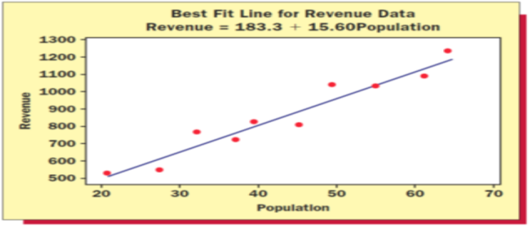
- Estimated (Sample) Regression Line:

$$\hat{y} = b_0 + b_1 x = 183.31 + 15.596x$$

- Prediction and Residual ($x = 20.8$):

$$\hat{y} = b_0 + b_1 x = 183.31 + 15.596(20.8) = 507.69 (\text{that is, } \$507,690)$$

$$\text{Residual: } y - \hat{y} = 527.1 - 507.69 = 19.41 (\text{that is, } \$19,410)$$



Interpreting b_0 and b_1

- **Intercept (b_0):** The predicted value of y when $x = 0$.
- **Slope (b_1):** The **direction** and **magnitude** of the relationship between x and y.

Positive Slope ($b_1 > 0$)

Direct Relationship: As x increases, y **increases**.

- *Example:* More hours studied (x) leads to higher exam scores (y).

Negative Slope ($b_1 < 0$)

Inverse Relationship: As x increases, y **decreases**.

- *Example:* More absences (x) lead to lower exam scores (y).

Interpretation of b_1: For every 1-unit increase in x, y is predicted to change by b_1 units.

Example1: Predicting Weekly Sales from Advertising Budget

Suppose we fit the regression model: $\hat{y} = 12 + 3x$, where y and x are the weekly sales and advertising budget (in thousands of dollars), respectively

Interpretation:

- **Intercept ($b_0 = 12$):** When the advertising budget is \$0, the predicted weekly sales are \$12,000.
- **Slope ($b_1 = 3$):** For each additional \$1,000 spent on advertising, weekly sales are predicted to increase by \$3,000 on average.

Interpreting b_0 and b_1

Example: Predicting Product Demand from Price

Suppose we fit the following regression model:

$$\hat{y} = 100 - 5x$$

where:

- y = Weekly demand (in hundreds of units)
- x = Price per unit (in dollars)

Interpretation:

- **Intercept ($b_0 = 100$):** When the price is \$0, the predicted weekly demand is 10,000 units.
- **Slope ($b_1 = -5$):** For each \$1 increase in price, weekly demand is predicted to **decrease** by 500 units on average.

Example: House Price Model

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.

- Dependent Variable (y): House Price in $1000s
- Independent Variable (x): Square Feet

House Price in $1000s ($y$)	**Square Feet** (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Sum of Squares

Partitioning of Variation

- Total sum of squares *(Total variation)* is given by

$$SSTO = SS_{yy} = \sum (y_i - \bar{y})^2.$$

- Sum of squares for regression *(Explained variation)* is given by

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

- Sum of squared errors *(Unexplained variation)* is given by

$$SSE = \sum (y_i - \hat{y}_i)^2.$$

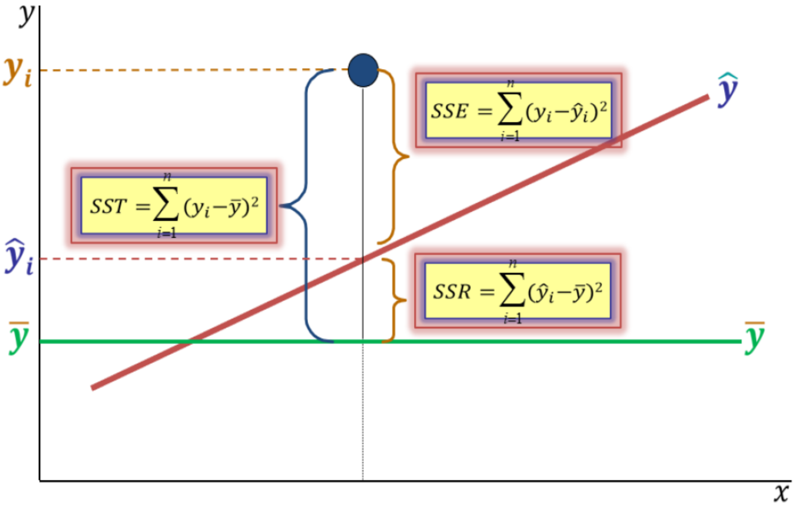
- Total variation is the sum of explained and unexplained variation. That is

$$SSTO = SSR + SSE.$$

Degrees of Freedom:

- df(SSR) $= 1$
- df(SSE) $= n - 2$
- df(SSTO) $= n - 1$

Sum of Squares (Graphically)



The Coefficient of Determination (R^2)

R^2 represents how well the regression line fits the data by measuring the percentage of the total variation in the response variable y that is explained by the predictor variable x.

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- **The Scale:** Ranges from **0 to 1** (or 0% to 100%).
- **Interpretation:**
  - $R^2 = 0.85$**:** 85% of the change in y is explained by x. The remaining 15% is due to other factors or random noise.
  - $R^2 = 0$**:** The model explains nothing; the mean is just as good a predictor.

Note:

- The closer the data points are to the regression line, the higher the R^2.
- In simple linear regression, R^2 is exactly the square of the correlation coefficient (r).

Example: House Price Model — Excel Regression Output

Regression Equation:

$$\hat{y} = 98.25 + 0.1098x$$

where y = House Price ($1000s), x = Square Feet

Model Summary

Multiple R	0.762
R Square	0.581
Adjusted R Square	0.528
Standard Error	41.33
Observations	10

ANOVA

Source	df	SS	MS	F	Significance F
Regression	1	18,934.93	18,934.93	11.08	0.0104
Residual	8	13,665.57	1,708.20		
Total	9	32,600.50			

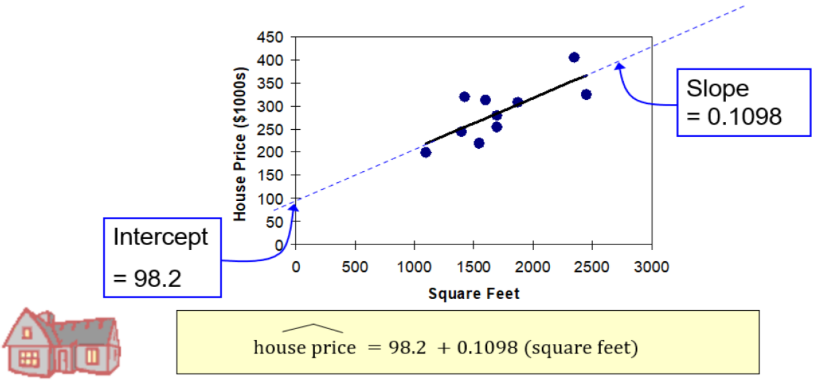
Coefficients

Term	Coefficient	Std Error	t Stat	P-value	Lower 95% / Upper 95%
Intercept	98.25	58.03	1.69	0.129	$[-35.58,\ 232.07]$
Square Feet (x)	0.1098	0.0330	3.33	0.010	$[0.034,\ 0.186]$

Example: House Price Model — Excel Regression Output

Key Interpretations:

- $b_0 = 98.2$ means that when the house size is 0 square feet, the predicted house price is $98,200.
- $b_1 = 0.1098$ indicates that for each additional one square foot of house size, the average house price increases by $109.8.
- $R^2 = 0.581$: about 58.1% of price variation is explained by house size.



$$\widehat{\text{house price}} = 98.2 + 0.1098\ (\text{square feet})$$

Testing the Significance

For test of significance of simple linear regression, the following tests are equivalent:

- **Test 1:** Test for significance of the coefficient of determination (R^2)
- **Test 2:** Test for significance of the regression slope coefficient (β_1)
- **Test 3:** Test for significance of the correlation coefficient (ρ)

Test 1: Test for significance of the coefficient of determination (R^2)

Hypotheses:

$$H_0 : \rho^2 = 0 \quad vs \quad H_a : \rho^2 \neq 0$$

In other words,

H_0 : The independent variable does not explain a significant portion of the variation in t

H_a : The independent variable explains a significant portion of the variation in the depe

Test Statistic:

$$F_0 = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} \sim F(df_1 = 1, df_2 = n-2)$$

Example: Housing Price Model

Recall Excel output of fitting the regression model to the housing price data gives us the ANOVA table as follows:

ANOVA

Source	**df**	**SS**	**MS**	**F**	**Significance F**
Regression	1	18,934.93	18,934.93	11.08	0.0104
Residual	8	13,665.57	1,708.20		
Total	9	32,600.50			

$$F_0 = \frac{SSR/1}{SSE/(n-2)} = \frac{18{,}934.93/1}{13{,}665.57/(10-2)} = 11.085$$

Also, at $\alpha = 0.05$ with $df_1 = 1, df_2 = 8$, the critical $F = 5.318$. Since $11.085 > 5.318$, we reject $H_0 : \rho^2 = 0$ and conclude that the model is statistically significant.

Test 2: Test for significance of the slope (β_1)

Hypotheses:

$$H_0 : \beta_1 = 0 \text{ (no linear relationship)}$$
$$H_a : \beta_1 \neq 0 \text{ (linear relationship does exist)}$$

Test Statistic:

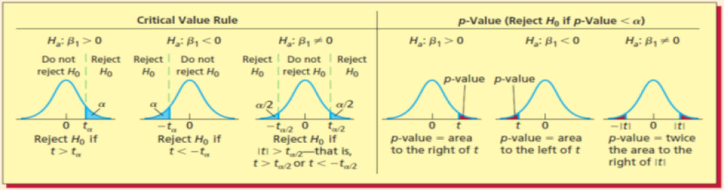
$$t_0 = \frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}$$

Confidence Interval around β_1:

$$b_1 \pm t_{\alpha/2,n-2} \frac{s}{\sqrt{SS_{xx}}}$$

where,

- $(\beta_1)_0 = 0$ is the hypothesized β_1 value
- $s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$, standard error of the slope
- $s = \sqrt{SSE/(n-2)}$, standard error of estimate



Here $t_{\alpha/2}$, t_α, and all p-values are based on $n-2$ degrees of freedom. **If we can reject H_0: $\beta_1 = 0$ at a given value of α, then we conclude that the slope (or, equivalently, the regression relationship) is significant at the α level.**

Example: Housing Price Model

Test Statistic: t = 3.329

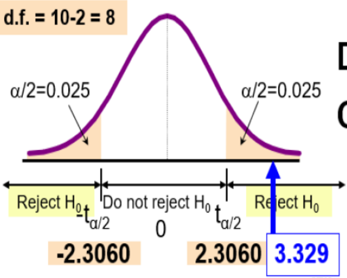
$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

From Excel output:

b_1 $\quad s_{b_1}$ $\quad t$

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039



Decision:
Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Example: Housing Price Model

95% Confidence Interval for the Slope, β_1

$$b_1 \pm t_{n-2} \times SE(b_1) = 0.10977 \pm 2.306 \times 0.03297 = (0.0337, 0.1858)$$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is \$1000s, we are 95% confident that the average impact on sales price is between \$33.70 and \$185.80 per square foot of house size

Confidence Interval Approach to test the slope: This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the 0.05 level of significance

Test 3: Test for significance of the correlation coefficient (ρ)

Pearson correlation is usually computed for sample data, but is also used to test hypotheses about the relationship in the population.

Hypotheses:

1. **Two-sided:** $H_0 : \rho = 0 \quad vs \quad H_a : \rho \neq 0$
2. **Right-tailed:** $H_0 : \rho \leq 0 \quad vs \quad H_a : \rho > 0$
3. **Left-tailed:** $H_0 : \rho \geq 0 \quad vs \quad H_a : \rho < 0$

where ρ denotes the population correlation coefficient.

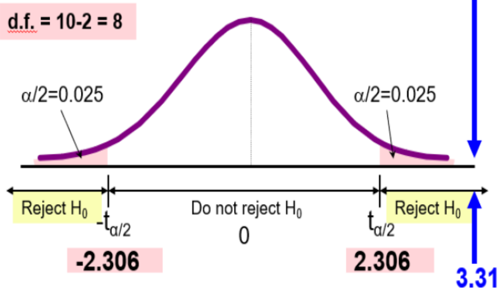
Test Statistic:

$$t_0 = \frac{r - \rho}{\sqrt{\dfrac{1 - r^2}{n - 2}}}$$

Critical Value: Use t table with $df = n - 2$.

Example: Housing Price Model

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.76}{\sqrt{\frac{1-0.76^2}{10-2}}} = 3.31$$



Decision:
Reject H_0

Conclusion:
There **is sufficient evidence** of a linear relationship at the 0.05 significance level

Confidence and Prediction Intervals

- The point on the regression line corresponding to a particular value of x_0 of the independent variable x, $\hat{y} = b_0 + b_1 x_0$, is deemed as the **point estimate of the mean value of** y and the **point prediction of an individual value of** y.
- We will assess the accuracy of $\hat{y}$ as both a point estimate and a point prediction.
- We can do this by calculating a confidence interval for the mean value of y and a prediction interval for an individual value of y.
- Both the confidence interval for the mean value of y and the prediction interval for an individual value of y employ a quantity called the distance value:

$$d = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}.$$

- The distance value is a measure of the distance between the value x_0 of x and $\bar{x}$. Notice that the further x_0 is from $\bar{x}$, the larger the distance value and hence wider the confidence interval.

Confidence and Prediction Intervals

- Confidence Interval for the Predicted Mean Value of Y given $x = x_0$ $(\mu_{y|x_0})$

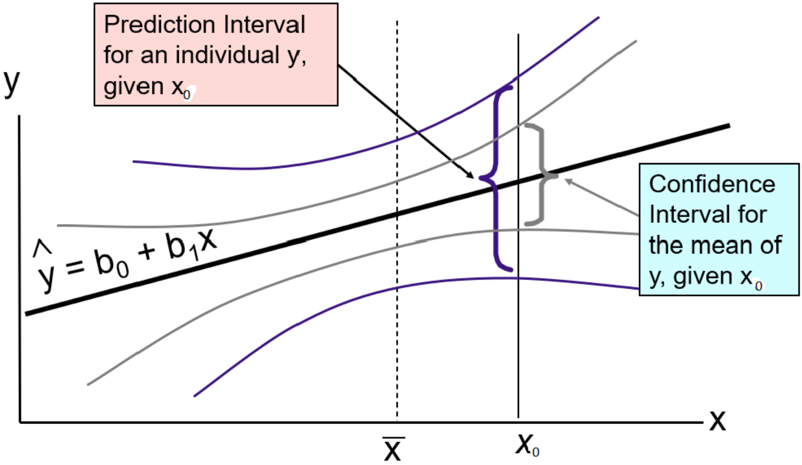
$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} \times SE(\hat{\mu}_0)$$
$$= \hat{y} \pm t_{\frac{\alpha}{2},n-2} \times s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = \hat{y} \pm t_{\frac{\alpha}{2},n-2} \times s\sqrt{\text{distance}}$$

- Prediction Interval for the Individual Value of Y given $x = x_0$ $(y|_{x_0})$

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2} \times SE(\hat{y})$$
$$= \hat{y} \pm t_{\frac{\alpha}{2},n-2} \times s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = \hat{y} \pm t_{\frac{\alpha}{2},n-2} \times s\sqrt{1 + \text{distance}}$$

where s is standard error of estimate and $SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$.

Confidence and Prediction Intervals



Example: Housing Price Model

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

Question: Predict the price for a house with 2000 square feet.

Solution:

$$\widehat{\text{house price}} = 98.25 + 0.1098(2000) = 317.85$$

The predicted price for a house with 2000 square feet is $317.85 \times \$1000s = \$317,850$

Housing Price Model: Confidence Interval Estimate for $E(y)|x_0$

Question: Find the 95% confidence interval for the average price of 2,000 square-foot houses.

Solution: Predicted Price $\hat{y}_i = 317.85$ ($1,000s)

$$\hat{y} \pm t_{\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = 317.85 \pm (2.306)(41.33)\sqrt{\frac{1}{10} + \frac{(2000-1715)^2}{1571500}}$$
$$= 317.85 \pm 37.12$$

The confidence interval endpoints are 280.73 – 354.97, or from $280,730 to $354,970

Housing Price Model: Prediction Interval Estimate for $y|x_0$

Question: Find the 95% prediction interval for the individual house price of 2,000 square-foot houses.

Solution: Predicted Price $\hat{y}_i = 317.85$ ($1,000s)

$$\hat{y} \pm t_{\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = 317.85 \pm (2.306)(41.33)\sqrt{1 + \frac{1}{10} + \frac{(2000-1715)^2}{1571500}}$$
$$= 317.85 \pm 102.28$$

The prediction interval endpoints are 215.57 – 420.13, or from $215,570 to $420,130

Formula Summary

Correlation & Sum of Squares

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

Regression Model

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

$$R^2 = \frac{SSR}{SSTO} = r^2$$

Variation Partitioning

$$SSTO = SSR + SSE$$

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

$$SSE = \sum(y_i - \hat{y}_i)^2$$

$$s = \sqrt{\frac{SSE}{n-2}}$$

Hypothesis Tests (All Equivalent)

F-test:

$$F_0 = \frac{SSR/1}{SSE/(n-2)}$$

$$H_0 : \rho^2 = 0$$

Degrees of Freedom:

$$df_R = 1,\ df_E = n-2$$

t-test (Slope):

$$t_0 = \frac{b_1 - (\beta_1)_0}{s_{b_1}}$$

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$$

$$H_0 : \beta_1 = (\beta_1)_0,\ \text{(usually } (\beta_1)_0 = 0)$$

t-test (Correlation):

$$t_0 = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$H_0 : \rho = 0$$

$$df = n-2$$

Formula Summary

Confidence & Prediction Intervals

Distance Value:

$$d = \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}$$

Confidence Interval (Mean):

$$\hat{y} \pm t_{\alpha/2,n-2} \cdot s\sqrt{d}$$

For $\mu_{y|x_0}$ (average of all y at x_0)

Prediction Interval (Individual):

$$\hat{y} \pm t_{\alpha/2,n-2} \cdot s\sqrt{1+d}$$

For y_{x_0} (single observation at x_0)
Note: Prediction interval is always wider than confidence interval

Key Relationships:

- $R^2 = r^2$ in simple linear regression
- Three significance tests are equivalent
- Prediction intervals $>$ Confidence intervals in width