PMDS504L: Regression Analysis and Predictive Models

Properties of Least Squares Estimators And Fitted Regression Model

Dr. Jisha Francis

Department of Mathematics
School of Advanced Sciences
Vellore Institute of Technology
Vellore Campus, Vellore - 632 014
India

Simple Linear Regression Model

Model Definition:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y: Dependent (study) variable
- X: Independent (explanatory) variable
- β_0: Intercept term (regression coefficient)
- β_1: Slope parameter (regression coefficient)
- ϵ: Error term

Assumptions of the Error Term and Model Properties

Assumptions of Error Term (ϵ):

- ϵ is an **independent and identically distributed (i.i.d.)** random variable.
- Mean of ϵ: $E(\epsilon) = 0$.
- Variance of ϵ (error variance): $\text{Var}(\epsilon) = \sigma^2$ (constant).

LSE of Parameters

$$\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Where:

$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i (x_i - \bar{x}), \quad s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Fitted Linear Regression Model and Residuals

- The fitted line or fitted linear regression model is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

- The predicted values of y are:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad \text{for } i = 1, 2, \ldots, n.$$

- The difference between the observed value y_i and the fitted (or predicted) value $\hat{y}_i$ is called a residual . The i-th residual is defined as:

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right).$$

Properties of the Least-Squares Estimators

- The least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the observations y_i.
- For $\hat{\beta}_1$, we have:

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

- The least-squares estimators are unbiased, i.e.:

$$E(\hat{\beta}_1) = \beta_1 \quad \text{and} \quad E(\hat{\beta}_0) = \beta_0.$$

Variance of the Least-Squares Estimators

- The variance of $\hat{\beta}_1$ is:

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{S_{xx}}.$$

- The variance of $\hat{\beta}_0$ is:

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(y) + x^2 \mathrm{Var}(\hat{\beta}_1) - 2x\mathrm{Cov}(y, \hat{\beta}_1).$$

- From the assumption of uncorrelated errors, $\mathrm{Cov}(y, \hat{\beta}_1) = 0$, so:

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

Gauss-Markov Theorem

- The Gauss-Markov theorem states that, for the linear regression model with the assumptions $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$, and uncorrelated errors, the least-squares estimators are unbiased and have the minimum variance among all unbiased estimators that are linear combinations of the y_i's.
- The least-squares estimators are therefore termed Best Linear Unbiased Estimators (BLUE).

Properties of the Residuals

- The sum of the residuals is always zero: $\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$.
- The sum of the observed values equals the sum of the fitted values: $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$.
- The regression line passes through the centroid $(\bar{y}, \bar{x})$ of the data.
- The sum of the residuals weighted by the corresponding regressor values equals zero: $\sum_{i=1}^{n} x_i e_i = 0$.
- The sum of the residuals weighted by the corresponding fitted values equals zero: $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

Estimation of σ^2

- Estimating σ^2 is crucial for:
  - Testing hypotheses.
  - Constructing interval estimates.

Estimation of σ^2

Ideal Conditions for Estimating σ^2

1. **Multiple y values for the same x:**
   - If we have several data points (observations) of y for at least one value of x, we can calculate the variation in y at that point directly, without depending on the model.
2. **Prior knowledge of σ^2:**
   - If we already know something about σ^2 (from theory or past studies), we can use that information instead of relying on the model.

Estimation of σ^2

When Ideal Conditions Are Absent

- If these conditions aren't met, we estimate σ^2 using the residuals:

$$\text{Residual} = \text{Actual value} - \text{Predicted value}.$$

- This method depends on how well the model fits the data.

Residual Sum of Squares (SS_{Res})

The residual sum of squares is given by:

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

Substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into the equation simplifies it further.

Computing Formula for SS_{Res}

The formula for SS_{Res} becomes:

$$SS_{\text{Res}} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

where:

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

Hence:

$$SS_{\text{Res}} = SS_T - \hat{\beta}_1 S_{xy}$$

Estimation of σ^2

- $\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n-2}$ is an unbiased estimate of the variance of residuals (σ^2).
- Dividing by $n - 2$ accounts for the loss of 2 DoF due to parameter estimation, ensuring accurate variance estimation.

What are Degrees of Freedom?

- Degrees of freedom (DoF) represent the number of independent observations available for estimating a parameter.
- In regression, DoF is the total number of observations (n) minus the number of parameters estimated.

Degrees of Freedom (DoF) Simplified

What Are Degrees of Freedom?

- Degrees of freedom (DoF) are the number of values in a dataset that are free to vary while calculating a statistic.
- When you estimate a parameter, you "use up" one degree of freedom.

Simple Example:

- Imagine you have 5 numbers, and their average is known.
- If 4 numbers are chosen, the 5th number is fixed to maintain the average.
- So, the degrees of freedom are $5 - 1 = 4$.

Degrees of Freedom in Regression

Why $n - 2$ in Simple Linear Regression?

- Simple linear regression estimates two parameters: β_0 (intercept) and β_1 (slope).
- The remaining $n - 2$ observations are used to compute the residual sum of squares (SS_{Res}).

Why DoF Matters?

- $\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n-2}$ is an unbiased estimate of the variance of residuals (σ^2).
- Dividing by $n - 2$ accounts for the loss of 2 DoF due to parameter estimation, ensuring accurate variance estimation.

Degrees of Freedom and Estimation of Variance

- The residual sum of squares (SS_{Res}) has $n-2$ degrees of freedom, as two degrees are associated with $\hat{\beta}_0$ and $\hat{\beta}_1$.
- The expected value of SS_{Res} is:

$$E(SS_{\text{Res}}) = (n-2)\sigma^2$$

- An unbiased estimator of σ^2 is:

$$\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n-2} = MS_{\text{Res}}$$

Residual Mean Square and Standard Error

- $MS_{\text{Res}} = \frac{SS_{\text{Res}}}{n-2}$ is the residual mean square.
- The square root of $\hat{\sigma}^2$ is called the **standard error of regression**, which has the same units as the response variable y.

Model Dependency of $\hat{\sigma}^2$

- $\hat{\sigma}^2$ is model-dependent because it is computed from regression model residuals.
- Any violation of model assumptions or misspecification can seriously affect the usefulness of $\hat{\sigma}^2$ as an estimate of σ^2.

Numerical Problem

SCUBA divers have specific maximum dive times that they must not exceed at various depths to ensure safety. The table below shows the relationship between the depth (in feet) and the corresponding maximum dive time (in minutes):

Depth (feet), X	50	60	70	80	90	100
Maximum Dive Time (minutes), Y	80	55	45	35	25	22

1. **Fit the Least Squares Regression Line:** Use the data to calculate the least squares regression equation of Y (maximum dive time) on X (depth).
2. **Predict:** Based on the regression equation, predict the maximum dive time for a depth of 110 feet.
3. **Plot the Data:** Create a scatter plot of the data points, draw the fitted regression line, and sketch the relationship between depth and maximum dive time.

Solution

To fit a simple linear regression line using the method of least squares, we calculate the coefficients β_0 and β_1 using the following formulas:

$$\hat{\beta}_1 = \frac{n \sum_i^n x_i y_i - \sum_i^n x_i \sum_i^n y_i}{n \sum_i^n x_i^2 - (\sum_i^n x_i)^2}$$

$$\hat{\beta}_0 = \frac{\sum_i^n y_i - \hat{\beta}_1 \sum_i^n x_i}{n}$$

Data and Calculations for Regression Line

Given data:

x	y	xy	x^2
50	80	4000	2500
60	55	3300	3600
70	45	3150	4900
80	35	2800	6400
90	25	2250	8100
100	22	2200	35500
450	262	17700	408

Summary of Values and Regression Coefficients

$$S_{xx} = 1750$$
$$S_{xy} = -1950$$
$$\beta_1 = -1.114285714$$
$$\beta_0 = 127.2380952$$

Summary of Values and Regression Coefficients

The estimated regression line is: $Y \approx 127.2380952 - 1.114285714X$

- **Intercept ($\beta_0 = 127.24$):** This represents the estimated maximum dive time (in minutes) when the depth (X) is 0 feet. While it has no physical meaning in the context of diving, it is a necessary component of the regression equation.
- **Slope ($\beta_1 = -1.11$):** For every 1-foot increase in depth, the maximum dive time decreases by approximately 1.11 minutes. This reflects the inverse relationship between depth and dive time due to safety constraints.

The relationship between depth (X) and maximum dive time (Y) is linear and negative, meaning that as the depth increases, the maximum allowable dive time decreases to mitigate risks associated with increased pressure and nitrogen absorption.

Alternate Form of the Model

Introduction to the Alternate Form:

- The simple linear regression model can be expressed in an alternate form by redefining the regressor variable as the deviation from its mean.
- Let the regressor variable x_i be rewritten as $x_i - \bar{x}$, where $\bar{x}$ is the mean of x.

Transformation of the Model:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + \epsilon_i \\
&= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + \epsilon_i \\
&= \beta_0' + \beta_1(x_i - \bar{x}) + \epsilon_i \qquad (2.20)
\end{aligned}
$$

Alternate Form of the Model

- In this form, $\beta_0' = \beta_0 + \beta_1 \bar{x}$.
- The regression equation is now centered around the mean of x, simplifying some interpretations.

Discussion Question:

- *How does redefining x_i as $x_i - \bar{x}$ affect the interpretation of the regression coefficients β_0 and β_1?*

Hypothesis Testing

- Hypothesis testing is discussed in this section.

Model Assumptions

- Additional assumption: Model errors ϵ_i are normally distributed.
- Complete assumptions: Errors are:
  - Normally and independently distributed.
  - Mean 0, variance σ^2.
- Abbreviation: $\text{NID}(0, \sigma^2)$.

Use of t-Tests

- Testing the hypothesis that the slope equals a constant, say β_{10}.
- Hypotheses:

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

- Two-sided alternative hypothesis specified.

Distribution of the Test Statistic

- Errors ϵ_i are NID$(0, \sigma^2)$.
- Observations y_i are NID$(\beta_0 + \beta_1 x_i, \sigma^2)$.
- The slope estimator $\hat{\beta}_1$ is:
  - Normally distributed with mean β_1 and variance σ^2 / S_{xx}.

Test Statistic for the Slope

- If $H_0 : \beta_1 = \beta_{10}$ is true:

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

- Since σ^2 is typically unknown, use MSRes as an unbiased estimator of σ^2.
- The test statistic:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\text{MSRes}}{S_{xx}}}} \sim t_{n-2}$$

- Degrees of freedom: $n - 2$.

Decision Rule for Hypothesis Testing

- Reject H_0 if:

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

- Alternatively, use the P-value approach for decision-making.

Standard Error of the Slope

- The denominator of t_0 is the estimated standard error of the slope:

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\text{MSRes}}{S_{xx}}}$$

Testing the Intercept

- To test $H_0 : \beta_0 = \beta_{00}, H_1 : \beta_0 \neq \beta_{00}$:

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\text{se}(\hat{\beta}_0)}$$

- Standard error of the intercept:

$$\text{se}(\hat{\beta}_0) = \sqrt{\text{MSRes}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

- Reject H_0 if:

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

Significance of Regression

- Special case of the hypotheses:

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$$

- These hypotheses assess the linear relationship between x and y.
- Failing to reject $H_0 : \beta_1 = 0$ implies:
  - No linear relationship between x and y.
  - Best estimator of y is $\hat{y} = \bar{y}$ (Figure 2.2a).
  - The true relationship might not be linear (Figure 2.2b).

Interpretation of Rejecting H_0

- If $H_0 : \beta_1 = 0$ is rejected:
  - x is valuable in explaining the variability in y.
  - This can imply:
    - A straight-line model is adequate (Figure 2.3a).
    - Better results might be achieved with higher-order polynomial terms (Figure 2.3b).

Test Procedure for $H_0 : \beta_1 = 0$

- Use the t-statistic:

$$t_0 = \frac{\hat{\beta}_1 - 0}{\mathrm{se}(\hat{\beta}_1)}$$

- The null hypothesis is rejected if:

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

- This procedure determines whether x contributes to the model.

Figures: $H_0 : \beta_1 = 0$

- $H_0 : \beta_1 = 0$ is **not rejected**.
- No linear relationship.
    - Figure 2.2a: Best estimator $\hat{y} = \bar{y}$.
    - Figure 2.2b: True relationship is not linear.

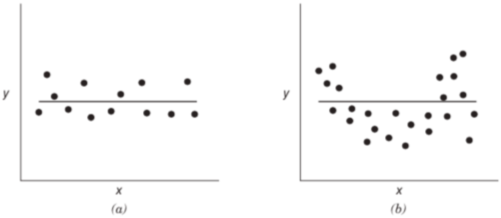

Figure 2.2 Situations where the hypothesis H_0: $\beta_1 = 0$ is not rejected.

Figures: $H_0 : \beta_1 = 0$

- $H_0 : \beta_1 = 0$ is **rejected**.
- Linear effect of x.
  - Figure 2.3a: Straight-line model is adequate.
  - Figure 2.3b: Polynomial terms might improve results.

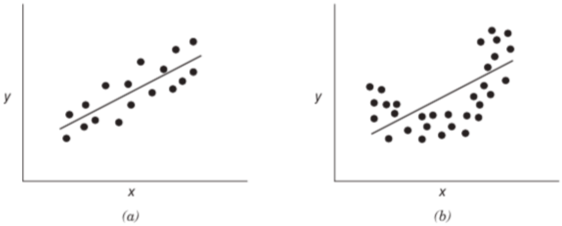

Figure 2.3 Situations where the hypothesis $H_0: \beta_1 = 0$ is rejected.

Summary of $H_0 : \beta_0 = \beta_{00}$

HYPOTHESIS TEST FOR β_0

One-sided test

$H_0 : \beta_0 = \beta_{00}$
(β_{00} is a specific value of β_0)

$H_a : \beta_0 > \beta_{00}$ or $\beta_0 < \beta_{00}$

Test statistic:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{00}}{\left[MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right]^{1/2}}$$

Rejection region:

$t > t_{\alpha,(n-2)}$ (upper tail region)
$t < -t_{\alpha,(n-2)}$ (lower tail region)

Two-sided test

$H_0 : \beta_0 = \beta_{00}$

$H_a : \beta_0 \neq \beta_{00}$

Test statistic:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{00}}{\left[MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right]^{1/2}}$$

Rejection region:

$|t| > t_{\alpha/2,(n-2)}$

Decision: If t_{β_0} falls in the rejection region, reject the null hypothesis at level of significance α.

Assumptions: Assume that the errors ε_i, $i = 1, \ldots, n$ are independent and normally distributed with $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$, and $Var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$.

Summary of $H_0 : \beta_1 = \beta_{10}$

HYPOTHESIS TEST FOR β_1

One-sided test

$H_0 : \beta_1 = \beta_{10}$ (β_{10} is a specific value of β_1)

$H_a : \beta_1 > \beta_{10}$ or $\beta_1 < \beta_{10}$

Test statistic:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

Rejection region:

$t > t_{\alpha,(n-2)}$ (upper tail region)
$t < -t_{\alpha,(n-2)}$ (lower tail region)

Two-sided test

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

$H_a : \beta_1 \neq \beta_{10}$

Test statistic:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

Rejection region:

$|t| > t_{\alpha/2,(n-2)}$

Decision: If t_{β_1} falls in the rejection region, reject the null hypothesis at confidence level α.

Assumptions: Assume that the errors $\varepsilon_i, i = 1, \ldots, n$ are independent and normally distributed with $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$, and $Var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$.

Confidence intervals for the slopes and for the intercept

PROCEDURE FOR OBTAINING CONFIDENCE INTERVALS FOR β_0 AND β_1

1. Compute $S_{xx}, S_{xy}, S_{xy}, \overline{y}$, and $\overline{x}$ as in the procedure for fitting a least-squares line.
2. Compute $\hat{\beta}_1, \hat{\beta}_0$ using equations $\hat{\beta}_1 = (S_{xy})/(S_{xx})$ and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$, respectively.
3. Compute SSE by $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$.
4. Define MSE (mean square error) to be

$$MSE = \frac{SSE}{n-2},$$

where $n =$ Number of pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$.

5. A $(1-\alpha)100\%$ confidence interval for β_1 is given by

$$\left(\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}\right)$$

where $t_{a/2}$ is the upper tail $\alpha/2$-point based on a t-distribution with $(n-2)$ degrees of freedom.

6. A $(1-\alpha)100\%$ confidence interval for β_0 is given by

$$\left(\hat{\beta}_0 - t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^2, \hat{\beta}_0 + t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right]^{1/2}\right).$$

Using the t-Table

The t-table is used to find critical values for the t-distribution in hypothesis testing and confidence interval construction.

Steps to Use the t-Table:

(1) **Determine the Degrees of Freedom (df):**
- For a single sample: $df = n - 1$
- For regression: $df = n - 2$

(2) **Select the Significance Level (α):**
- Common levels: $\alpha = 0.05$ (95% confidence), $\alpha = 0.01$ (99% confidence).
- For two-tailed tests: Divide α by 2.

(3) **Locate the Critical Value:**
- Find the row corresponding to df.
- Move across to the column for your $\alpha/2$ or α.
- The intersection provides $t_{\alpha/2,df}$.

Using the t-Table

(4) **Compare the Test Statistic:**

- For a two-tailed test: Reject H_0 if $|t| > t_{\alpha/2,df}$.
- Otherwise, fail to reject H_0.

Example 1

Use the method of least squares to fit a straight line to the accompanying data points. Give the estimates of β_0 and β_1. Plot the points and sketch the fitted least-squares line.
The observed data values are as follows:

$$x : -1,\ 0,\ 2,\ -2,\ 5,\ 6,\ 8,\ 11,\ 12,\ -3$$

$$y : -5,\ -4,\ 2,\ -7,\ 6,\ 9,\ 13,\ 21,\ 20,\ -9$$

Solution

x_i	y_i	$x_i y_i$	x_i^2
-1	-5	5	1
0	-4	0	0
2	2	4	4
-2	-7	14	4
5	6	30	25
6	9	54	36
8	13	104	64
11	21	231	121
12	20	240	144
-3	-9	27	9
$\sum x_i = 38$	$\sum y_i = 46$	$\sum x_i y_i = 709$	$\sum x_i^2 = 408$

Calculating the Coefficients

The formulas for S_{xx} and S_{xy} are:

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

$$S_{xx} = 408 - \frac{38^2}{10} = 408 - 144.4 = 263.6$$

$$S_{xy} = 709 - \frac{38 \cdot 46}{10} = 709 - 174.8 = 534.2$$

Calculating the Coefficients

The slope ($\hat{\beta}_1$) and intercept ($\hat{\beta}_0$) are:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{534.2}{263.6} = 2.0266$$

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

Using $\bar{x} = \frac{\sum x_i}{n} = 3.8$ and $\bar{y} = \frac{\sum y_i}{n} = 4.6$:

$$\hat{\beta}_0 = 4.6 - (2.0266)(3.8) = 4.6 - 7.6981 = -3.1011$$

The Least-Squares Line

The equation of the least-squares line is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -3.1011 + 2.0266x$$

Mean Square Error

Given Data:

$$S_{xx} = 263.6, \quad S_{xy} = 534.2, \quad \bar{y} = 4.6, \quad \bar{x} = 3.8,$$

$$\hat{\beta}_1 = 2.0266, \quad \hat{\beta}_0 = -3.1011.$$

Additional Calculations:

$$\sum_{i=1}^{n} y_i^2 = 1302, \quad SS_T = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

Substituting the values:

$$SS_T = 1302 - \frac{(46)^2}{10} = 1302 - 211.6 = 1090.4.$$

Mean Square Error

Error Sum of Squares (SS_{Res}):

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}.$$

Substituting the values:

$$SS_{Res} = 1090.4 - (2.0266)(534.2) = 1090.4 - 1082.60972 = 7.79028.$$

Mean Square Error (MSE):

$$MS_{Res} = \frac{SS_{Res}}{n-2} = \frac{7.79028}{8} = 0.973785.$$

Hypothesis Testing Example

Problem: Using the data given in above Example, test the hypothesis $H_0 : \beta_0 = -3$ versus $H_a : \beta_0 \neq -3$ at the 0.05 level of significance.

Solution: We test $H_0 : \beta_0 = -3$ versus $H_a : \beta_0 \neq -3$. Here $\beta_{00} = -3$.

From the calculations in the previous example, we have: $t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\text{MSE}\cdot\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$.

Substituting the values:

$$t_{\beta_0} = \frac{-3.1011 - (-3)}{\sqrt{(0.973785) \cdot \left(\frac{1}{10} + \frac{(3.8)^2}{263.6}\right)}} = -0.26041.$$

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Hypothesis Testing Example

$$t < -2.306 \quad \text{or} \quad t > 2.306.$$

Decision: Because the test statistic $t_{\beta_0} = -0.26041$ does not fall in the rejection region ($t < -2.306$ or $t > 2.306$), we do not reject H_0 at $\alpha = 0.05$.

Example: Hypothesis Testing for β_1

Problem: Using the data from Example 1, test the hypothesis $H_0 : \beta_1 = 2$ versus $H_a : \beta_1 \neq 2$ at the $\alpha = 0.05$ level of significance.

Solution: We test $H_0 : \beta_1 = 2$ versus $H_a : \beta_1 \neq 2$.

Rejection Region: For $\alpha = 0.05$ and $n = 10$, the rejection region is:

$$t < -2.306 \quad \text{or} \quad t > 2.306.$$

Example: Hypothesis Testing for β_1

Test Statistic: The test statistic is given by:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1^0}{\sqrt{\frac{\text{MSE}}{S_{xx}}}}.$$

Substituting the values:

$$\hat{\beta}_1 = 2.0266, \quad \beta_1^0 = 2, \quad \text{MSE} = 0.973785, \quad S_{xx} = 263.6,$$

$$t_{\beta_1} = \frac{2.0266 - 2}{\sqrt{\frac{0.973785}{263.6}}}.$$

Example: Hypothesis Testing for β_1

Simplifying:

$$t_{\beta_1} = \frac{2.0266 - 2}{0.0603} = 0.4376.$$

Decision: Since $t_{\beta_1} = 0.4376$ does not fall in the rejection region ($t < -2.306$ or $t > 2.306$), we do not reject H_0.

Conclusion: At $\alpha = 0.05$, the data support the null hypothesis that the true value of the slope β_1 of the regression line is equal to 2.

Confidence Intervals for β_0 and β_1

Problem: For the data from Example 1: (a) Construct a 95% confidence interval for β_0 and interpret. (b) Construct a 95% confidence interval for β_1 and interpret.

Solution: From Example 1, we have:

$$S_{xx} = 263.6, \quad S_{xy} = 534.2, \quad \bar{y} = 4.6, \quad \bar{x} = 3.8,$$

$$\hat{\beta}_1 = 2.0266, \quad \hat{\beta}_0 = -3.1011, \quad MSE = 0.973785, \quad t_{0.025,8} = 2.306.$$

Confidence Intervals for β_0 and β_1

(a) 95% Confidence Interval for β_0: The formula for the confidence interval is:

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{\text{MSE}\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}.$$

Substituting the values:

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{0.973785\left(\frac{1}{10}+\frac{(3.8)^2}{263.6}\right)}.$$

Confidence Intervals for β_0 and β_1

Simplifying:

$$\hat{\beta}_0 \pm 2.306\sqrt{0.973785\,(0.1 + 0.0548)} = -3.1011 \pm 2.306 \cdot 0.3817.$$

Calculating the interval:

$$(-3.1011 - 0.8811, -3.1011 + 0.8811) = (-3.9846, -2.2176).$$

Interpretation: With 95% confidence, the true value of the intercept β_0 lies between -3.9846 and -2.2176.

Confidence Intervals for β_0 and β_1

(b) 95% Confidence Interval for β_1: The formula for the confidence interval is:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{\text{MSE}}{S_{xx}}}.$$

Substituting the values:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{0.973785}{263.6}} = 2.0266 \pm 2.306 \cdot 0.0603.$$

Calculating the interval:

$$(2.0266 - 0.1392, 2.0266 + 0.1392) = (1.8864, 2.1668).$$

Confidence Intervals for β_0 and β_1

Interpretation: With 95% confidence, the true value of the slope β_1 lies between 1.8864 and 2.1668.

Reference

- Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.
- https://home.iitk.ac.in/ shalab/course5.htm

Thank You!

Thank you for your attention!