Probability and Statistical Techniques BCNS 1206C
Unit 3: Linear Regression

1 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable y and a single independent variable x. The assumption is that this relationship is linear, and that the observed values of y vary randomly around a true regression line.

1.1 The Simple Linear Regression Framework

We assume that the response variable y can be expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where:

- β_0 is the intercept,
- β_1 is the slope of the line,
- ϵ is a random error term with mean zero and constant variance σ^2.

The errors ϵ are assumed to be independent and normally distributed.

1.2 Estimating Parameters via Least Squares

Given n observed data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the model becomes:

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

To estimate β_0 and β_1, we minimize the sum of squared residuals:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Solving the normal equations yields:

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

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y respectively.
The predicted value for each x_i is:

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

and the residual is $e_i = y_i - \hat{y}_i$.

Notation

Let:

$$s_{xy} = \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i, \quad s_{xx} = \sum x_i^2 - \frac{1}{n} \left(\sum x_i\right)^2.$$

Then:

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

Exercise 1. *List and explain the standard assumptions made about the error term ϵ_i in a simple linear regression model.*

Exercise 2. *Show that minimizing the sum of squared residuals leads to the estimators:*

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

Exercise 3. *Given the following summary statistics from a dataset of 20 observations:*

$$\sum x_i = 23.92, \quad \sum y_i = 1843.21, \quad \sum x_i^2 = 29.2892,$$

$$\sum y_i^2 = 170044.5321, \quad \sum x_i y_i = 2214.6566,$$

compute the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

Exercise 4. *The table below shows the number of hours studied (x) and the corresponding exam scores (y) for 7 students. Fit a simple linear regression model to this data.*

x	2.0	3.5	1.0	0.5	0.4	3.0	5.0
y	65.5	70.0	50.5	45.5	75.0	68.0	85.5

1.3 Analysis of Variance (ANOVA) in Regression

To assess the significance of the regression model, we decompose the total variation in y as:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2,$$

or:

$$SST = SSR + SSE,$$

where:

- SST is the total sum of squares,
- SSR is the regression sum of squares,
- SSE is the error sum of squares.

To test $H_0 : \beta_1 = 0$, we compute the F-statistic:

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

We reject H_0 if $F_0 > F_{1,n-2}$ at the chosen significance level.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Regression	SSR	1	$MSR = SSR$	$F_0 = MSR/MSE$
Error	SSE	$n-2$	$MSE = SSE/(n-2)$	
Total	SST	$n-1$		

1.4 Tutorial

Problem 1. *A study was conducted to examine the relationship between excise duties (in pence per litre) on unleaded petrol (x) and diesel (y) across 10 European nations. The following summary statistics were recorded:*

$$\sum x_i = 375, \quad \sum y_i = 298, \quad \sum x_i^2 = 15193, \quad \sum y_i^2 = 9974, \quad \sum x_i y_i = 12191$$

(i) Construct the ANOVA table for this dataset.

(ii) At the 1% significance level, test whether there is a statistically significant linear relationship between x and y. Use the critical value $F_{1,8}(0.01) = 11.26$.

Problem 2. **(a)** *In the linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $i = 1, 2, \ldots, n$, list the standard assumptions made about the error term ϵ_i.*

(b) *Show that minimizing the sum of squared residuals leads to the least squares estimators:*

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

where $S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$ and $S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$.

(c) *A company is evaluating the link between consulting project duration (x in hours) and invoiced cost (y in dollars). The data collected are:*

x	1	1.5	3	3.5	3.5	4.5	5	6	8
y	40	74	80	140	180	220	175	209	331

(i) Compute the least squares estimates and write the fitted regression equation.

(ii) Interpret the slope coefficient in the context of this problem.

(iii) Predict the invoiced cost when $x = 4.5$ hours. Calculate the residual for this prediction.

Problem 3. *An oil refinery manager recorded the specific gravity of crude oil (x) and the corresponding yield of petroleum spirit (y) on seven occasions:*

x	30.2	32.8	32.9	35.1	42.3	45.5	46.0
y	6.8	10.1	14.3	19.3	10.2	20.0	23.7

(i) Fit a linear regression model and provide the equation.

(ii) Comment on the nature of the relationship between x and y.

(iii) Estimate the yield when the specific gravity is 40.

Problem 4. *A chemical experiment measured the specific heat H (in calories per gram) of a compound at various temperatures T (in $^\circ C$). The summary statistics are:*

$$\sum T = 900, \quad \sum H = 20.16, \quad \sum T^2 = 71000, \quad \sum H^2 = 33.8894, \quad \sum TH = 1519.9$$

(i) Construct the ANOVA table for this data.

(ii) Test whether the slope of the regression line is significantly different from zero. Interpret your findings.

Problem 5. *A medical researcher collected data on the age (x) and cholesterol level (y) of 20 patients. The summary statistics are:*

$$\sum x_i = 809, \quad \sum y_i = 68.3, \quad S_{xx} = 3630.95, \quad S_{xy} = 201.665, \quad S_{yy} = 12.9455$$

(i) Complete the ANOVA table using the given statistics.

(ii) Test the hypothesis $H_0 : \beta_1 = 0$ at the 0.1% significance level. Use $F_{1,18}(0.001) = 15.38$.

Problem 6. *During a jet engine test, temperature (x) and thrust (y) were recorded under consistent conditions. For 15 observations, the following statistics were obtained:*

$$\sum x_i = 540, \quad \sum x_i^2 = 21412, \quad \sum y_i = 33.0, \quad \sum y_i^2 = 78.54, \quad \sum x_i y_i = 1276.6$$

(i) Prepare the ANOVA table.

(ii) At the 2.5% significance level, test whether the slope β_1 is significantly different from zero. Use $F_{1,13}(0.025) = 6.4142$.

Problem 7. *The specific heat (y) of a compound was measured at various temperatures (x in $^\circ C$). The data are:*

x	50	60	70	80	90	100
y	1.60	1.65	1.67	1.70	1.72	1.74

(i) Fit a linear regression model and write the equation.

(ii) Describe the relationship between temperature and specific heat.

(iii) Predict the specific heat when $x = 85$.

(iv) Estimate the temperature at which the specific heat is 1.64 calories per gram.

Problem 8. *A researcher is investigating the relationship between advertising expenditure (x in thousands of rupees) and monthly sales revenue (y in lakhs of rupees) for a retail chain. The following summary statistics are available from 12 months of data:*

$$\sum x_i = 96, \quad \sum y_i = 144, \quad \sum x_i^2 = 832, \quad \sum y_i^2 = 1800, \quad \sum x_i y_i = 1224$$

(i) Compute the least squares estimates of the regression coefficients.

(ii) Construct the ANOVA table.

(iii) Test whether advertising expenditure significantly predicts sales at the 5% level.

Problem 9. *A nutritionist is studying the effect of daily protein intake (x in grams) on muscle mass gain (y in kilograms) over a 6-week training program. The data for 8 participants is summarized as follows:*

$$\sum x_i = 640, \quad \sum y_i = 48.2, \quad S_{xx} = 3200, \quad S_{xy} = 384, \quad S_{yy} = 24.5$$

(i) Fit a simple linear regression model and write the equation.

(ii) Test the hypothesis that protein intake has no effect on muscle gain at the 1% significance level.

(iii) Interpret the slope in the context of this study.

Problem 10. *A dataset records the number of hours employees spend in training (x) and their corresponding performance scores (y) on a standardized test. The regression equation obtained is:*

$$\hat{y} = 52.3 + 1.8x$$

(i) Predict the performance score for an employee who received 10 hours of training.

(ii) If the actual score was 72, calculate the residual.

(iii) Explain what the intercept and slope represent in this context.

Problem 11. *A biologist is analyzing the relationship between the concentration of a nutrient (x in mg/L) and the growth rate of a plant species (y in cm/week). The following data was collected:*

$$\sum x_i = 150, \quad \sum y_i = 210, \quad \sum x_i^2 = 3100, \quad \sum y_i^2 = 4700, \quad \sum x_i y_i = 4350, \quad n = 10$$

(i) Calculate the regression coefficients and write the fitted model.

(ii) Construct the ANOVA table and compute the F-statistic.

(iii) At the 5% level, test whether nutrient concentration significantly affects plant growth.