# example

October 2, 2025

## 1 Example Questions

`{note} The questions below are designed to be solved analytically by hand (except when otherwise stated), encouraging a deeper understanding of the underlying statistical concepts rather than relying solely on computational tools in R.`

---

1. **Linear Regression**

| SNo. | X1 | X2 | X3 | Y |
|---|---|---|---|---|
| 1 | 2.258 | 0.917 | 39 | 150.74 |
| 2 | 1.315 | 0.595 | 22 | 126.33 |
| 3 | 4.996 | 0.843 | 37 | 151.13 |
| 4 | 4.647 | 0.474 | 17 | 124.04 |
| 5 | 1.653 | 0.123 | 7 | 99.04 |
| 6 | 4.939 | 0.995 | 39 | 158.37 |
| 7 | 2.762 | 0.026 | 10 | 98.63 |
| 8 | 3.214 | 0.180 | 37 | 121.06 |
| 9 | 3.229 | 0.222 | 8 | 106.24 |
| 10 | 0.464 | 0.345 | 27 | 117.56 |
| 11 | 2.567 | 0.463 | 6 | 113.86 |
| 12 | 4.755 | 0.700 | 99 | 180.27 |
| 13 | 4.770 | 0.811 | 3 | 130.01 |
| 14 | 4.753 | 0.663 | 12 | 129.07 |
| 15 | 2.150 | 0.001 | 67 | 129.14 |
| 16 | 1.162 | 0.171 | 9 | 101.33 |
| 17 | 4.313 | 0.351 | 49 | 136.73 |
| 18 | 1.676 | 0.157 | 42 | 120.46 |
| 19 | 1.056 | 0.449 | 62 | 142.75 |
| 20 | 3.412 | 0.855 | 57 | 160.40 |
| 21 | 4.385 | 0.644 | 26 | 135.69 |
| 22 | 3.261 | 0.527 | 58 | 147.32 |
| 23 | 0.046 | 0.447 | 25 | 119.88 |
| 24 | 0.817 | 0.782 | 73 | 162.26 |
| 25 | 1.650 | 0.512 | 28 | 126.90 |
| 26 | 3.616 | 0.201 | 91 | 153.45 |
| 27 | 3.740 | 0.284 | 83 | 152.45 |

| SNo. | X1 | X2 | X3 | Y |
|---|---|---|---|---|
| 28 | 3.595 | 0.748 | 81 | 170.04 |
| 29 | 1.616 | 0.440 | 13 | 115.34 |

For the above-detailed dataset with exogenous variables $X_1, X_2, X_3$ and endogenous variable $Y$, develop the following linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ in R, and answer the questions below

`{tip} Save the dataset in an Excel/CSV file and import it into R using functions like `read.csv()` or `readxl::read_excel()`.`

a. Comment on the estimated value and significance of each coefficient - $\beta_0, \beta_1, \beta_2, \beta_3$
b. Fill in the above table, computing fitted values and corresponding errors
c. Compute the following model statistics

- Sum Squares Total
- Sum Squares Regression
- Sum Squares Error
- R-squared
- Adjusted R-squared

d. Perform ex-post analysis (compute correlation between $X_1$ and $X_2$; draw residuals plot) to comment upon the validity of the model.

`{tip} You can perform similar analysis using some standard datasets available in R, such as: - `mtcars`: Explore the relationship between car features (e.g., horsepower, weight) and miles per gallon. - `iris`: Predict petal length or width using sepal measurements. - `Boston` (from the `MASS` package): Model median house value based on socioeconomic and housing variables. - `airquality`: Analyze how temperature, wind, and solar radiation affect ozone levels. - `swiss`: Study fertility rates as a function of socio-economic indicators. These datasets are well-documented and suitable for hands-on linear regression practice.`

---

2. **Logistic Regression**

For the following dataset with exogenous variables $X_1, X_2, X_3$ and binary endogenous variable $Y$, develop the following logistic regression model in R

`{tip} Save the dataset in an Excel/CSV file and import it into R using functions like `read.csv()` or `readxl::read_excel()`.`

$$\log\left(\frac{\hat{P}_{Y=S}}{1-\hat{P}_{Y=S}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

a. Comment on the estimated value and significance of each coefficient $\beta_0, \beta_1, \beta_2, \beta_3$.

b. Compute estimated probabilities ($\hat{P}_{Y=S}$ and $\hat{P}_{Y=F}$)

c. Compute the following model statistics

- Log-Likelihood for
  - Equally-Likely Model
  - Market Share Model
- McFadden R-squared for the Estimated Model vs.
  - Equally-Likely Model
  - Market Share Model
- Adjusted McFadden R-squared for the Estimated Model vs.
  - Equally-Likely Model
  - Market Share Model

| SNo | $X_1$ | $X_2$ | $X_3$ | $Y$ |
|---|---|---|---|---|
| 1 | 1.8 | 10 | 0 | F |
| 2 | 2.1 | 14 | 0 | F |
| 3 | 2.3 | 13 | 1 | F |
| 4 | 2.5 | 15 | 0 | F |
| 5 | 2.7 | 18 | 1 | S |
| 6 | 2.9 | 16 | 0 | F |
| 7 | 3.0 | 20 | 1 | S |
| 8 | 3.1 | 17 | 0 | F |
| 9 | 3.2 | 22 | 1 | S |
| 10 | 3.3 | 21 | 0 | F |
| 11 | 3.4 | 19 | 1 | S |
| 12 | 3.5 | 25 | 0 | S |
| 13 | 3.6 | 23 | 1 | S |
| 14 | 3.7 | 26 | 0 | S |
| 15 | 3.8 | 24 | 1 | S |
| 16 | 4.0 | 28 | 0 | S |
| 17 | 2.2 | 12 | 1 | F |
| 18 | 2.6 | 15 | 1 | S |
| 19 | 2.8 | 19 | 0 | F |
| 20 | 3.0 | 18 | 1 | S |
| 21 | 3.2 | 22 | 0 | S |
| 22 | 3.4 | 20 | 1 | S |
| 23 | 3.6 | 27 | 0 | S |
| 24 | 3.8 | 29 | 1 | S |
| 25 | 2.4 | 14 | 0 | F |
| 26 | 2.7 | 16 | 1 | F |
| 27 | 2.9 | 18 | 1 | S |
| 28 | 3.1 | 21 | 0 | S |
| 29 | 3.5 | 23 | 1 | S |

| SNo | $X_1$ | $X_2$ | $X_3$ | $Y$ |
|---|---|---|---|---|
| 30 | 3.9 | 30 | 0 | S |

{tip} You can perform similar analysis using some standard datasets available in R, such as: - `Titanic` (from the `datasets` package): Predict survival based on passenger features. - `iris`: Convert species to a binary variable and predict using petal/sepal measurements. - `mtcars`: Predict whether a car has automatic or manual transmission (`am` variable) using other features. - `PimaIndiansDiabetes` (from the `mlbench` package): Predict diabetes status based on medical measurements. - `Default` (from the `ISLR` package): Predict default status using income, balance, and student status. These datasets are well-documented and suitable for hands-on logistic regression practice.