Predictive Analysis Assignment 2

SHRESTHA BAJAJ

2026-02-05

Problem Set 2: Linear Regression

Question 1. Problem to demonstrate that the population regression line is fixed, but least square regression line varies.

Suppose the population regression line is given by Y = 2 + 3x, while the data comes from the model y = 2 + 3x + ε.

Step 1: For x in the range [5,10] graph the population regression line.

Step 2: Generate x_i (i = 1, 2, .., n) from Uniform(5, 10) and ε_i (i = 1, 2, ..,n) from N(0,4^2). Hence, compute the y1, y2, .., yn.

Step 3: On the basis of the data (x_i,y_i) (i = 1, 2, .., n) generated in Step 2, report the least squares regression line.

Step 4: Repeat steps 2-3 five times. Graph the 5 least squares regression lines over the population regression line obtained in Step 1.

Interpret the findings.

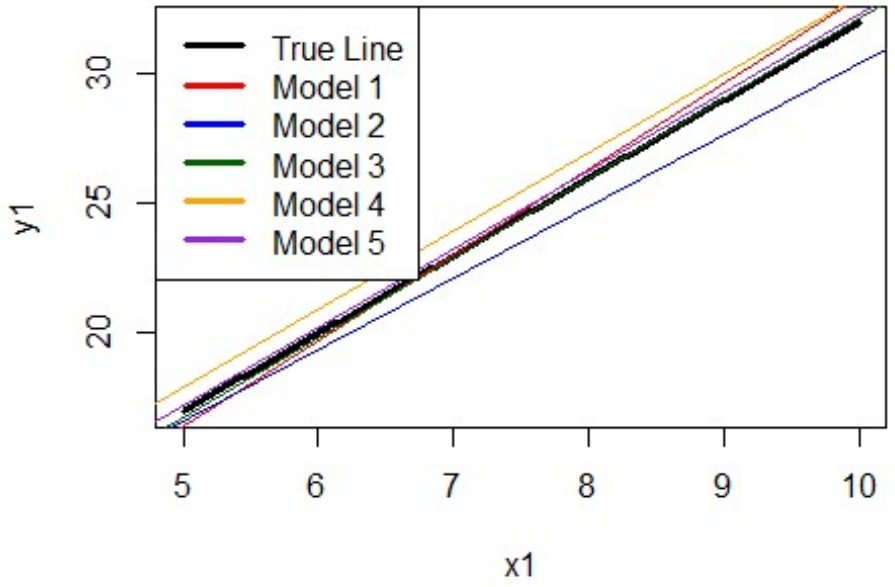
Take n = 50. Set the seed as seed=123.

```
set.seed(123)
x1=seq(5,10,length.out=200)
y1=2+3*x1
plot(x1,y1,type='l',col="black",lwd=3)
results = data.frame(iteration = 1:5, intercept = numeric(5), slope =
numeric(5))

cols = c("red", "blue", "darkgreen", "orange", "purple")

for (i in 1:5){
x = runif(50, 5, 10)
eps = rnorm(50, 0, 4)
y = 2 + 3*x + eps
mod = lm(y ~ x)
abline(mod, col = cols[i])
results$intercept[i] =coef(mod)[1]
results$slope[i]= coef(mod)[2]
}
```

```
legend("topleft",legend = c("True Line", paste("Model", 1:5)), col =
c("black", cols), lty = 1, lwd = 3)
```




```
results

##   iteration   intercept    slope
## 1         1 -0.09638929 3.305396
## 2         2  2.79218839 2.761042
## 3         3  1.39299737 3.073267
## 4         4  2.82308856 3.023608
## 5         5  2.03250638 3.028097
```

Interpretation:

- The black line represents the true underlying relationship between x_1 and y_1.
- Models 1–5 are different fitted models attempting to approximate this true line.
- All models show a positive linear relationship, so they've captured the direction of the effect correctly.
- However, they differ in how closely they match the true line — this reflects model bias and estimation error.

Question 2. Problem to demonstrate that $\hat{\beta}_0$ and $\hat{\beta}$ minimises RSS.

Step 1: Generate x_i from Uniform(5, 10) and mean centre the values. Generate ε_i from N(0, 1). Calculate y_i = 2 + 3x_i + ε_i, i = 1,2,.., n. Take n=50 and seed=123.

Step 2: Now imagine that you only have the data on (x_i,y_i), i = 1, 2, .., n, without knowing the mechanism that was used to generate the data in step 1. Assuming a linear regression of the type y_i = β_0 + βx_i + ε_i, and based on these data (x_i,y_i), i = 1, 2, .., n, obtain the least squares estimates of β0 and β.

Step 3: Take a large number of grid values of (β_0, β) that also include the least squares estimates obtained from step 2. Compute the RSS for each parametric choice of (β_0, β), where RSS = $(y_1 - \beta_0 - \beta x_1)^2$ + $(y_2 - \beta_0 - \beta x_2)^2$ + ….+$(y_n - \beta_0 - \beta x_n)^2$. Find out for which combination of (β0, β), RSS is minimum.

```
set.seed(123)

x = runif(50, 5, 10)
eps = rnorm(50, 0, 1)

x_centered = x - mean(x)
y = 2 + 3*x_centered + eps

mod1 = lm(y ~ x_centered)
b0_hat = coef(mod1)[1]
b1_hat = coef(mod1)[2]

b0_grid = seq(b0_hat - 0.5, b0_hat + 0.5, length.out = 51)
b1_grid = seq(b1_hat - 0.5, b1_hat + 0.5, length.out = 51)

RSS = matrix(NA, 51, 51)

for (i in 1:51) {
  for (j in 1:51) {
    RSS[i, j] <- sum((y - b0_grid[i] - b1_grid[j] * x_centered)^2)
  }
}

which(RSS == min(RSS), arr.ind = TRUE)

##      row col
## [1,]  26  26
```

Interpretation:

The RSS attains its minimum at grid index (26, 26), which corresponds to the least squares estimates $(\hat{\beta}_0, \hat{\beta})$.

Question 3. Problem to demonstrate that least square estimators are unbiased.

Step 1: Generate x_i (i = 1, 2, .., n) from Uniform(0, 1), ε_i (i = 1, 2, .., n) from N(0, 1) and hence generate y using $y_i = \beta_0 + \beta x_i + \varepsilon_i$.(Take β0 = 2, β = 3).

Step 2: On the basis of the data (x_i, y_i) (i = 1, 2, .., n) generated in Step 1, obtain the least square estimates of β_0 and β.

Repeat Steps 1-2, R = 1000 times. In each simulation obtain $\hat{\beta}_0$ and $\hat{\beta}$. Finally, the least-square estimates will be given by the average of these estimated values.

Compare these with the true β_0 and β and comment. Take n = 50 and seed=123.

```
set.seed(123)

result = data.frame(
  beta0_hat = numeric(1000),
  beta_hat = numeric(1000)
)

b0 = 2
b1 = 3

for (i in 1:1000){
  x = runif(50, 0, 1)
  eps = rnorm(50, 0, 1)
  y = b0 + b1*x + eps
  mod = lm(y ~ x)

  result$beta0_hat[i] = coef(mod)[1]
  result$beta_hat[i] = coef(mod)[2]
}

avg_b0 = mean(result$beta0_hat)
avg_b0
```

```
## [1] 2.013053
```

```
avg_b1 = mean(result$beta_hat)
avg_b1
```

```
## [1] 2.982112
```

```
var_beta0 = var(result$beta0_hat)
var_beta0
```

```
## [1] 0.07969639
```

```
var_beta1 = var(result$beta_hat)
var_beta1
```

```
## [1] 0.2360544
```

Interpretation:

The estimated values of $\hat{\beta}_0$ and $\hat{\beta}$ obtained, from running the simulations for 1000 times, are very close to value of true parameters β_0 = 2 and β = 3. This implies that the least square estimators are unbiased and consistent.

Question 4. Comparing several simple linear regressions.

Attach "Boston" data from MASS library in R. Select median value of owner- occupied homes, as the response and per capita crime rate, nitrogen oxides concentration, proportion of blacks and percentage of lower status of the population as predictors.

(a) Selecting the predictors one by one, run four separate linear regressions to the data. Present the output in a single table.

(b) Which model gives the best fit?

(c) Compare the coefficients of the predictors from each model and comment on the usefulness of the predictors.

```
#(a)
library(MASS)

## Warning: package 'MASS' was built under R version 4.5.2

data(Boston)
attach(Boston)
y = medv

# Fit separate linear models
model_crim  = lm(y ~ crim, data = Boston)
model_nox   = lm(y ~ nox, data = Boston)
model_black = lm(y ~ black, data = Boston)
model_lstat = lm(y ~ lstat, data = Boston)

results = data.frame(
  Predictor = c("crim", "nox", "black", "lstat"),
  Intercept = c(coef(model_crim)[1], coef(model_nox)[1],
                coef(model_black)[1], coef(model_lstat)[1]),
  Slope = c(coef(model_crim)[2], coef(model_nox)[2],
            coef(model_black)[2], coef(model_lstat)[2]),
  R_Squared = c(summary(model_crim)$r.squared,
                summary(model_nox)$r.squared,
                summary(model_black)$r.squared,
                summary(model_lstat)$r.squared),
  P_Value = c(summary(model_crim)$coefficients[2,4],
              summary(model_nox)$coefficients[2,4],
              summary(model_black)$coefficients[2,4],
              summary(model_lstat)$coefficients[2,4]),row.names = NULL
)
```

```
results
```

```
##   Predictor Intercept        Slope R_Squared      P_Value
## 1      crim  24.03311  -0.41519028 0.1507805 1.173987e-19
## 2       nox  41.34587 -33.91605501 0.1826030 7.065042e-24
## 3     black  10.55103   0.03359306 0.1111961 1.318113e-14
## 4     lstat  34.55384  -0.95004935 0.5441463 5.081103e-88
```

```
#(b)
best_model = results[which.max(results$R_Squared),]
best_model
```

```
##   Predictor Intercept      Slope R_Squared      P_Value
## 4     lstat  34.55384 -0.9500494 0.5441463 5.081103e-88
```

R-squared (R^2) measures how well the model fits the data. It tells us what proportion of the variability in the response variable is explained by the model.An R^2 close to 1 means the model explains most of the variation in the response.An R^2 close to 0 means the model explains very little. Here, the model with lstat as the predictor has highest R^2 (R^2 = 0.5441463). It means that this model explains around 54% of the variability in the response variable.

The coefficient of per capita crime rate (crim) is negative, implying that an increase in crime rate is associated with a decrease in median house value. However, the magnitude of the coefficient is relatively small and the corresponding R^2 value is low, indicating limited explanatory power when used alone.

The coefficient of nitrogen oxides concentration (nox) is also negative and larger in magnitude, suggesting that higher levels of air pollution significantly reduce house prices. Although statistically significant, its explanatory power remains moderate.

The coefficient of proportion of blacks (black) is positive, indicating a weak positive association with median house value. However, the low R^2 suggests that this variable alone does not explain much of the variation in house prices and hence is a weaker predictor.

The coefficient of percentage of lower status population (lstat) is strongly negative and has the largest magnitude among all predictors. This model also has the highest R^2 indicating that lstat explains the maximum variation in median house values and is the most useful predictor among the four.