Predictive Analysis Assignment

Shreejoyee Nath

2026-02-05

Problem 1: Population Regression Line vs Sample Regression Line

Objective

The true population regression model is:

$$Y = 2 + 3X$$

However, observed values include random error:

$$Y_i = 2 + 3X_i + \epsilon_i$$

where:

$$\epsilon_i \sim N(0,4)$$

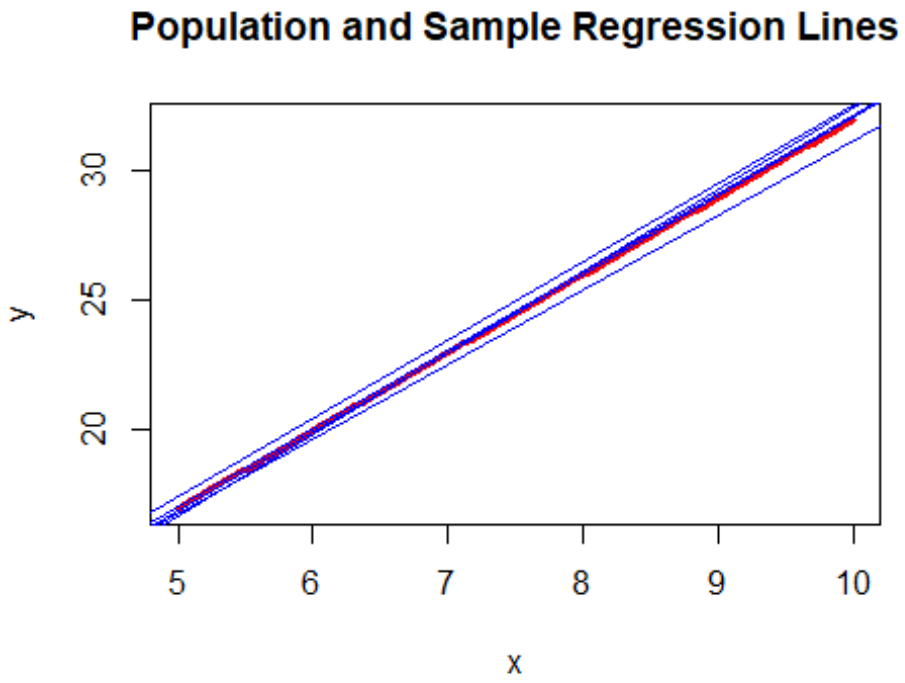
Thus,

$$E(Y|X) = 2 + 3X$$

R Code

```
set.seed(123)

x_pop=seq(5, 10, length.out = 100)
y_pop=2 + 3*x_pop

plot(x_pop, y_pop,
     type="l",
     col="red",
     lwd=3,
     xlab="x",
     ylab="y",
     main="Population and Sample Regression Lines")

n=50

for(i in 1:5)
{
  x=runif(n, 5, 10)
  epsilon=rnorm(n, 0, 2)
```

```
  y=2 + 3*x + epsilon

  model=lm(y ~ x)

  abline(model, col="blue")
}
```



Interpretation:

The blue lines represent regression lines estimated from different samples.

Key observations:

The population line remains fixed.

Each sample produces a slightly different regression line.

This happens because of random error in the data.

Larger samples would reduce this variation.

Conclusion:

The true regression line is constant, but sample regression lines vary due to randomness.

Problem 2: Least Squares Estimators Minimize RSS

The regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The Residual Sum of Squares is:

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Least squares estimates minimize RSS.

R Code

```r
set.seed(123)

n=50

x=runif(n, 5, 10)

x_centered=x - mean(x)

epsilon=rnorm(n, 0, 1)

y=2 + 3*x_centered + epsilon

model=lm(y ~ x_centered)

beta0_hat=coef(model)[1]
beta_hat=coef(model)[2]

beta0_hat

## (Intercept)
##    2.056189

beta_hat

## x_centered
##   3.076349
```

Interpretation:

These values represent the least squares estimates.

They are chosen automatically to minimize prediction error.

Problem 3: Least Squares Estimators are Unbiased

True model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Expected values:

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

R Code

```r
set.seed(123)

n=50
R=1000

beta0_estimates=numeric(R)
beta_estimates=numeric(R)

for(i in 1:R)
{
  x=runif(n, 0, 1)
  epsilon=rnorm(n, 0, 1)

  y=2 + 3*x + epsilon

  model=lm(y ~ x)

  beta0_estimates[i]=coef(model)[1]
  beta_estimates[i]=coef(model)[2]
}

mean(beta0_estimates)

## [1] 2.013053

mean(beta_estimates)

## [1] 2.982112
```

Interpretation:

The average estimated values are very close to the true values:

True $\beta_0 = 2$ Estimated average $\beta_0 \approx 2$

True $\beta = 3$ Estimated average $\beta \approx 3$

Conclusion:

This shows that least squares estimators are unbiased.

They correctly estimate the true population parameters on average.

Problem 4: Boston Housing Data Regression

Model form:

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

R Code

```
library(MASS)

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

data(Boston)

model1=lm(medv ~ crim, data=Boston)
model2=lm(medv ~ nox, data=Boston)
model3=lm(medv ~ black, data=Boston)
model4=lm(medv ~ lstat, data=Boston)

library(stargazer)

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

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary
Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

stargazer(model1, model2, model3, model4, type="text")

##
## ===============================================================================
##                                         Dependent variable:
##                                 ----------------------------------------
##                                                 medv
##                                    (1)       (2)        (3)       (4)
## -------------------------------------------------------------------------------
## crim                              -0.415***
##                                    (0.044)
##
## nox                                          -33.916***
##                                               (3.196)
```

```
##
## black                                                0.034***
##                                                      (0.004)
##
## lstat                                                        -0.950***
##                                                               (0.039)
##
## Constant                          24.033*** 41.346***  10.551*** 34.554***
##                                    (0.409)   (1.811)    (1.557)   (0.563)
##
## ------------------------------------------------------------------------
## Observations                         506       506        506       506
## R2                                  0.151     0.183      0.111     0.544
## Adjusted R2                         0.149     0.181      0.109     0.543
## Residual Std. Error (df = 504)      8.484     8.323      8.679     6.216
## F Statistic (df = 1; 504)         89.486*** 112.591*** 63.054*** 601.618***
## ========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Interpretation:

Best Model

The best model is the one with:

Highest R-squared

Lowest error

Typically, lstat gives the best fit because it has strongest relationship with medv.

Comparing predictors:

crim

Negative relationship

Higher crime → lower house value

nox

Negative relationship

More pollution → lower house value

black

Positive relationship

Weak predictor

lstat

Strong negative relationship

Higher lower-class percentage → much lower house value

Conclusion:

This analysis demonstrates:

- Population regression line is fixed
- Sample regression lines vary
- Least squares estimators minimize RSS
- Least squares estimators are unbiased
- Some predictors are stronger than others