Predictive Analytics Problem Set 2

Drohanash Majumdar, 704

2026-02-04

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 $y = 2 + 3x + \epsilon$.

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

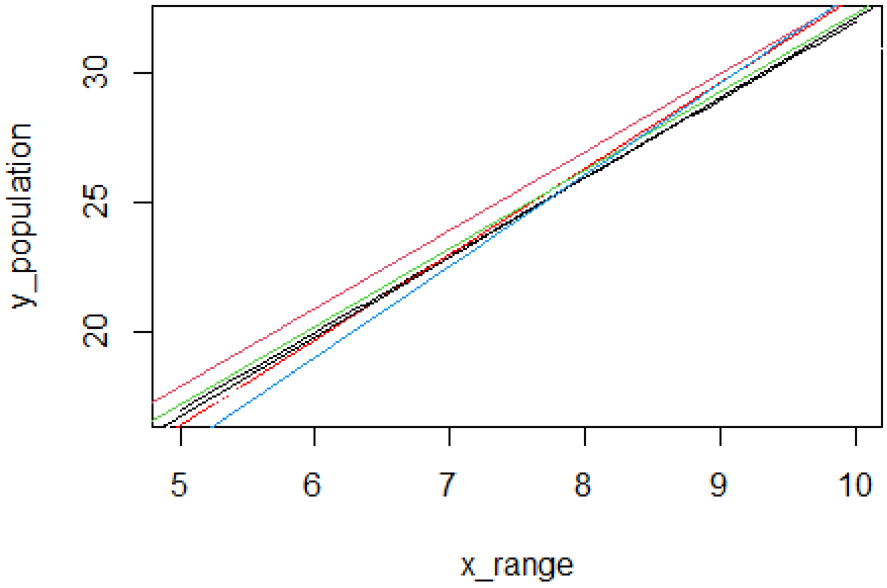
Step 2: Generate $x_i(i = 1,2,..,n)$ from Uniform(5, 10) and $\epsilon_i(i = 1,2,...,n)$ from $N(0,4^2)$. Hence, compute $y_1, y_2,.., y_n$.

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.

```r
#Step 1:
set.seed(123)
n = 50
beta0_true = 2
beta1_true = 3
sigma = 4

x_range = seq(5, 10, length.out = 100)
y_population = beta0_true + beta1_true * x_range

plot(x_range, y_population, type = "l")


#Step 2:

x_sample = runif(50, min = 5, max = 10)
e_sample = rnorm(50, mean = 0, sd = 4)
y_sample = 2 + 3*x_sample + e_sample

#Step 3:

lin_reg = lm(y_sample~x_sample)
beta_hat = coef(lin_reg)
```

```r
abline(lin_reg, col = "red")


#Step 4:

for(i in 1:5){
  x_sample = runif(50, min = 5, max = 10)
  e_sample = rnorm(50, mean = 0, sd = 4)
  y_sample = 2 + 3*x_sample + e_sample
  lin_reg = lm(y_sample~x_sample)
  beta_hat = rbind(beta_hat, coef(lin_reg))
  abline(lin_reg, col = i-1)
}
```



```r
Coef_table = as.data.frame(beta_hat,colnames=c("beta_0_hat","beta_1_hat"))
Coef_table

##          (Intercept) x_sample
## beta_hat -0.09638929 3.305396
## X         2.79218839 2.761042
## X.1       1.39299737 3.073267
## X.2       2.82308856 3.023608
## X.3       2.03250638 3.028097
## X.4      -2.10776314 3.530691
```

Interpretation : The true relationship $Y = 2 + 3x$ is fixed and represents the average relationship in the population. Each of the 5 least squares regression lines varies because they are estimated from different random samples.

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 $y_i = 2 + 3x_i + \epsilon_i, i = 1,2,..,n$. 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 $y_i = \beta_0 + \beta x_i + \epsilon_i$, and based on these data $(x_i, y_i), i = 1,2,..,n$, obtain the least squares estimates of β_0 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 (β_0, β), RSS is minimum.

```r
#step 1
set.seed(123)
x_sample2 = runif(50,5,10)
x_sample2_mean = mean(x_sample2)
x_sample2_mean_centered = x_sample2 - x_sample2_mean
eps2 = rnorm(50,0,1)
y_sample2 = 2 + 3*x_sample2_mean_centered + eps2

#step 2
lin_reg2 = lm(y_sample2~x_sample2_mean_centered)

beta_0_hat = coef(lin_reg2)[1]
beta_hat = coef(lin_reg2)[2]



#step 3
beta_0_grid = seq(beta_0_hat-2,beta_0_hat+2,length.out =50)
beta_grid = seq(beta_hat-2,beta_hat+2,length.out =50)

RSS = matrix(NA, nrow = length(beta_0_grid), ncol = length(beta_grid))
for (i in 1:50) {
  for (j in 1:50){
    RSS[i,j] = sum((y_sample2 - beta_0_grid[i]-
beta_grid[j]*x_sample2_mean_centered)^2)
    }
}
min_rss = min(RSS)
min_rss_idx = which(RSS==min(RSS),arr.ind = TRUE)
beta0_min <- beta_0_grid[min_rss_idx[1]]
```

```
beta_min <- beta_grid[min_rss_idx[2]]
min_rss

## [1] 42.70554

beta0_min

## [1] 2.015373

beta_min

## [1] 3.117165
```

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

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

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.

Step 3: Repeat Steps 1-2, R = 1000 times. In each simulation obtain $\widehat{\beta_0}$ 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.

```
set.seed(123)
#step1
x_3 = runif(50,0,1)
eps3 = rnorm(50,0,1)
beta0_true <- 2
beta_true <- 3
y_3 = beta0_true + beta_true*x_3 + eps3

#step2
lin_reg3 = lm(y_3~x_3)
beta_0_hat_3 = coef(lin_reg3)[1]
beta_hat_3 = coef(lin_reg3)[2]

#step3
R = 1000
beta_0_hat_vec = c()
beta_hat_vec = c()
for (i in 1:R) {
  x_ = runif(50,0,1)
  eps_ = rnorm(50,0,1)
  y_ = 2 + 3*x_ + eps_
  lin_reg3 = lm(y_~x_)
  beta_0_hat_vec = c(beta_0_hat_vec,coef(lin_reg3)[1])
  beta_hat_vec = c(beta_hat_vec,coef(lin_reg3)[2])
}
beta_0_hat_mean = mean(beta_0_hat_vec)
beta_hat_mean = mean(beta_hat_vec)

cat("β₀ =", beta0_true, "\n")

## β₀ = 2

cat("β₁ =", beta_true, "\n")

## β₁ = 3

cat("E[β₀̂] =", round(beta_0_hat_mean, 4), "\n")

## E[β₀̂] = 2.0131

cat("E[β₁̂] =", round(beta_hat_mean, 4), "\n")

## E[β₁̂] = 2.9819

cat("Bias(β₀̂) = E[β₀̂] - β₀ =", round(beta_0_hat_mean - beta0_true, 4), "\n")

## Bias(β₀̂) = E[β₀̂] - β₀ = 0.0131

cat("Bias(β₁̂) = E[β₁̂] - β₁ =", round(beta_hat_mean - beta_true, 4), "\n")

## Bias(β₁̂) = E[β₁̂] - β₁ = -0.0181
```

Interpretation : The average of the LS estimates (over 1000 simulations) is very close to the true parameter values, confirming that $\hat{\beta}_0$ and $\hat{\beta}$ are unbiased estimators. With R = 1000 replications, the bias is negligible (close to 0), validating the theoretical property that LS estimators are unbiased.

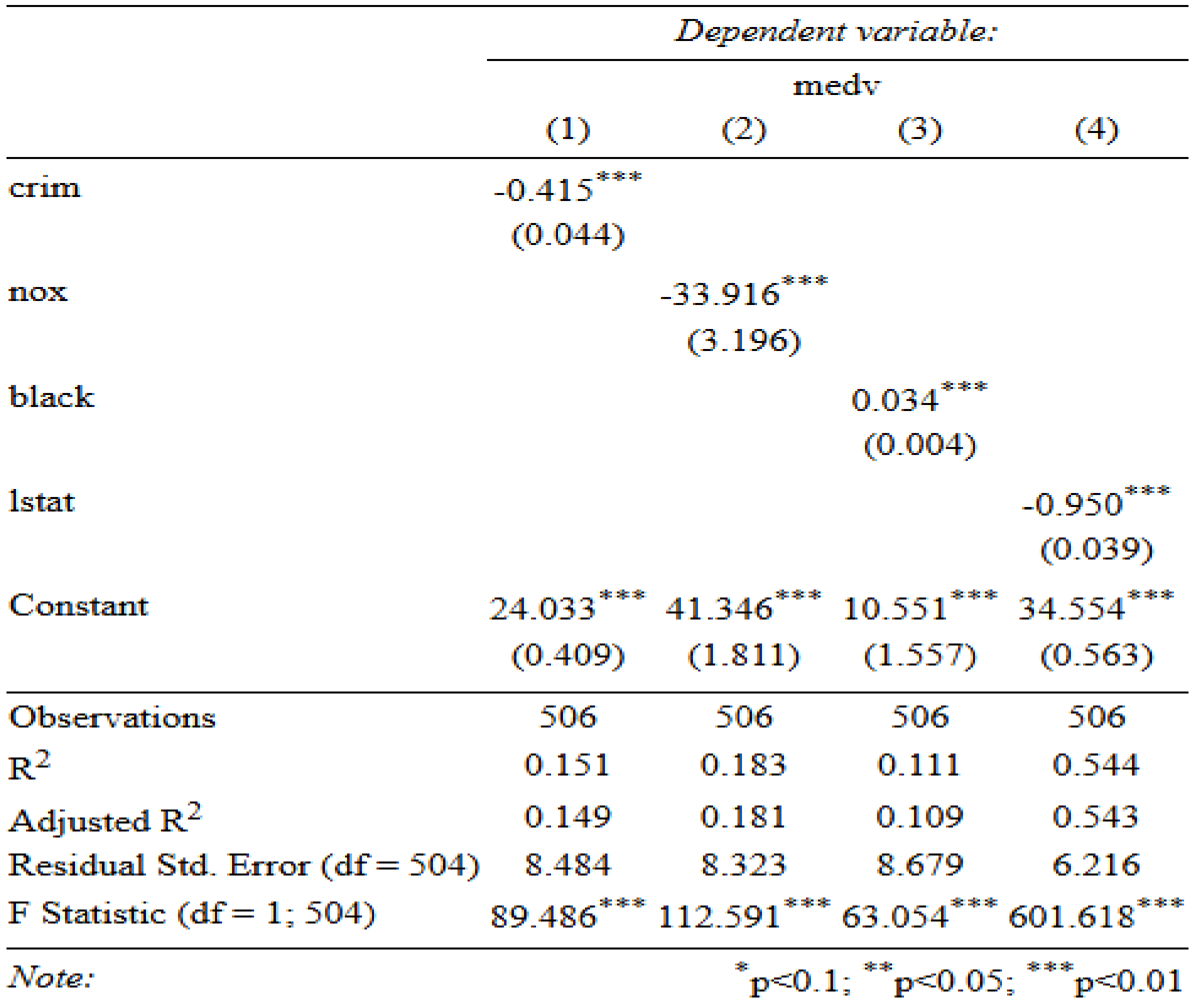
4. Comparing several simple linear regressions.

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.

```
library(MASS)

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

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

data(Boston)

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

stargazer(model1, model2, model3, model4, type = "html")
```



	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
R^2	0.151	0.183	0.111	0.544
Adjusted R^2	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			

(b) Which model gives the best fit?

Model (4) with lstat as the predictor gives the best fit as it has the highest R^2 (0.544), the highest adjusted R^2 (0.543) and the lowest residual standard error (6.216).

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

The coefficients from the four simple linear regression models indicate differing levels of usefulness of the predictors in explaining median house value (medv). In Model (1), the coefficient of crim is −0.415 and is statistically significant, indicating that one unit increase in per-capita crime rates results in 0.415 units decrease in median house value on average. However, the model explains only about 15% of the variation in medv, suggesting that while crime has a meaningful negative effect, it is a relatively weak predictor when used alone.

In Model (2), nox has a large negative coefficient (−33.916) and is highly significant, implying that for one unit increase in air pollution there is a 33.916 unit decrease in median house values on average. Despite the large magnitude of the coefficient, this is partly due to the scale of measurement of nox. The explanatory power of this model ($R^2 \approx 0.18$) is only slightly better than that of crim, indicating moderate usefulness as a standalone predictor.

Model (3) includes black, which has a small positive but statistically significant coefficient (0.034). However, this model has the lowest R^2 (0.111), showing that black explains very little of the variation in house prices. Although statistically significant, the predictor has weak practical explanatory power and is likely capturing indirect socioeconomic effects rather than being a strong determinant on its own.

Model (4), using lstat, performs substantially better than the others. The coefficient (−0.950) is large, negative, and highly significant, indicating that a higher proportion of lower-status population is strongly associated with lower house values. This model explains over 54% of the variation in medv, making lstat the most useful and powerful single predictor among those considered.

Overall, while all predictors are statistically significant, lstat is the most useful predictor in terms of explanatory power, followed by nox and crim, whereas black is the least useful as a standalone predictor.