Regression And Stats Models 52571 - Ex5

Contents

Q1.

a.

```r
lm(formula = medv ~ lstat + nox + dis + rm, data = Boston)
```

```
##
## Call:
## lm(formula = medv ~ lstat + nox + dis + rm, data = Boston)
##
## Coefficients:
## (Intercept)        lstat          nox          dis           rm
##     12.0360      -0.6587     -14.5379      -0.9670       4.8634
```

```r
values <- c(
 0.546920650, -0.00300380644, -0.267061900, -0.01476321472, -0.0481765600,
 -0.003003806, 0.00008930473, -0.001871825, 0.00004125515, 0.0004384591,
 -0.267061900, -0.00187182489, 0.420746305, 0.01463589900, 0.0002890320,
 -0.014763215, 0.00004125515, 0.014635899, 0.00111507017, 0.0003008943,
-0.048176560, 0.00043845907, 0.000289032, 0.00030089435, 0.0065757967)
XTX_inv_matrix <- matrix(values, nrow = 5, ncol = 5, byrow = TRUE)
rownames(XTX_inv_matrix) <- colnames(XTX_inv_matrix) <- c("Intercept", "lstat", "nox",
"dis", "rm")
```

Define the new observation vector x_0

$$x_0 = \begin{bmatrix} 1 \\ 0.13 \\ 0.5 \\ 3 \\ 4.5 \end{bmatrix}$$

```r
x0 <- c(1, 0.13, 0.5, 3, 4.5)
```

Define the estimated coefficients $\hat{\beta}$

From the regression output:

$$\hat{\beta} = \begin{bmatrix} 12.036 \\ -0.65865 \\ -14.53791 \\ -0.96699 \\ 4.86340 \end{bmatrix}$$

```
beta_hat <- c(12.03600, -0.65865, -14.53791, -0.96699, 4.86340)
```

Compute the predicted value $\hat{y}_0 = x_0^\top \hat{\beta}$

$$\hat{y}_0 = \sum_{i=1}^{5} x_{0i} \cdot \hat{\beta}_i$$

```
y_hat <- sum(x0 * beta_hat)
```

Define residual variance $\hat{\sigma}^2$

$$\hat{\sigma}^2 = (5.396)^2 = 29.123$$

```
sigma2 <- 5.396^2
```

Compute the variance of $\hat{y}_0$:

$$\mathrm{Var}(\hat{y}_0) = \hat{\sigma}^2 \cdot x_0^\top (X^\top X)^{-1} x_0$$

```
x0_matrix <- matrix(x0, nrow = 1)
var_y_hat <- x0_matrix %*% XTX_inv_matrix %*% t(x0_matrix) * sigma2
```

Standard error of $\hat{y}_0$

$$\mathrm{SE}(\hat{y}_0) = \sqrt{\mathrm{Var}(\hat{y}_0)}$$

```
se_y_hat <- sqrt(var_y_hat)
```

Get critical t-value for 95% CI (df = 501)

$$t_{0.975,501} \approx 1.964$$

```
t_crit <- qt(0.975, df = 501)
```

Compute confidence interval

$$\mathrm{CI} = \hat{y}_0 \pm t \cdot \mathrm{SE}$$

```
lower_bound <- y_hat - t_crit * se_y_hat
upper_bound <- y_hat + t_crit * se_y_hat
```

Return the result

```r
list(
  prediction = y_hat,
  se = se_y_hat,
  CI_95 = c(lower_bound, upper_bound)
)
```

```
## $prediction
## [1] 23.66575
##
## $se
##          [,1]
## [1,] 1.309877
##
## $CI_95
## [1] 21.09222 26.23928
```

b.

```r
# Define X'X inverse matrix from previous step
values <- c(
 0.546920650, -0.00300380644, -0.267061900, -0.01476321472, -0.0481765600,
 -0.003003806, 0.00008930473, -0.001871825, 0.00004125515, 0.0004384591,
 -0.267061900, -0.00187182489, 0.420746305, 0.01463589900, 0.0002890320,
 -0.014763215, 0.00004125515, 0.014635899, 0.00111507017, 0.0003008943,
 -0.048176560, 0.00043845907, 0.000289032, 0.00030089435, 0.0065757967)
XTX_inv_matrix <- matrix(values, nrow = 5, ncol = 5, byrow = TRUE)
rownames(XTX_inv_matrix) <- colnames(XTX_inv_matrix) <- c("Intercept", "lstat", "nox", "dis", "rm")
```

Define base values for the new observation

$$x_0 = \begin{bmatrix} 1 \\ 0.13 \\ 0.5 \\ 3 \\ 4.5 \end{bmatrix}$$

```r
x_base <- c(1, 0.13, 0.5, 3, 4.5)
```

Estimated coefficients from regression

$$\hat{\beta} = \begin{bmatrix} 12.036 \\ -0.65865 \\ -14.53791 \\ -0.96699 \\ 4.86340 \end{bmatrix}$$

```r
beta_hat <- c(12.03600, -0.65865, -14.53791, -0.96699, 4.86340)
```

Compute predicted value at baseline

$$\hat{y}_0 = x_0^\top \hat{\beta}$$

```r
y_base <- sum(x_base * beta_hat)
```

Simulate ±5% change in lstat

$$lstat_{+5\%} = 0.13 \cdot 1.05 = 0.1365 lstat_{-5\%} = 0.13 \cdot 0.95 = 0.1235$$

```r
x_lstat_up <- x_base;   x_lstat_up[2] <- 0.13 * 1.05
x_lstat_down <- x_base; x_lstat_down[2] <- 0.13 * 0.95

y_lstat_up <- sum(x_lstat_up * beta_hat)
y_lstat_down <- sum(x_lstat_down * beta_hat)
```

Simulate ±5% change in dis

$$dis_{+5\%} = 3 \cdot 1.05 = 3.15 dis_{-5\%} = 3 \cdot 0.95 = 2.85$$

```r
x_dis_up <- x_base;   x_dis_up[4] <- 3 * 1.05
x_dis_down <- x_base; x_dis_down[4] <- 3 * 0.95

y_dis_up <- sum(x_dis_up * beta_hat)
y_dis_down <- sum(x_dis_down * beta_hat)
```

Summarize Results

```r
tibble::tibble(
  Scenario = c(
    "Baseline",
    "+5% lstat", "-5% lstat",
    "+5% dis", "-5% dis"
  ),
  Predicted_Value = c(
    y_base,
    y_lstat_up, y_lstat_down,
    y_dis_up, y_dis_down
  ),
  Change_vs_Base = c(
    0,
    y_lstat_up - y_base,
    y_lstat_down - y_base,
    y_dis_up - y_base,
    y_dis_down - y_base
  )
)
```

```
## # A tibble: 5 x 3
##   Scenario  Predicted_Value Change_vs_Base
##   <chr>               <dbl>          <dbl>
## 1 Baseline             23.7        0
## 2 +5% lstat            23.7       -0.00428
## 3 -5% lstat            23.7        0.00428
## 4 +5% dis              23.5       -0.145
## 5 -5% dis              23.8        0.145
```

c.

Test statistic F-test for the joint null hypothesis

We test the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

Against the alternative:

$$H_1 : \text{At least one } \beta_j \neq 0$$

The test statistic is the **F-statistic**:

$$F = \frac{(R^2/k)}{((1 - R^2)/(n - k - 1))}$$

Where: - $R^2 = 0.6585$ - $k = 4$ (number of predictors: `lstat`, `nox`, `dis`, `rm`) - $n = 506$ (observations in Boston dataset) - $n - k - 1 = 501$ (degrees of freedom of residual)

```r
R2 <- 0.6585
k <- 4
n <- 506
df1 <- k
df2 <- n - k - 1

F_stat <- (R2 / k) / ((1 - R2) / df2)
F_stat
```

```
## [1] 241.5143
```

Determine the critical value at 5% significance level

We compare the test statistic to the critical value:

$$F_{0.95,4,501}$$

```r
F_crit <- qf(0.95, df1 = df1, df2 = df2)
F_crit
```

```
## [1] 2.389731
```

Decision rule

If:

$$F_{\text{stat}} > F_{\text{crit}} \Rightarrow \text{Reject } H_0$$

```r
reject_null <- F_stat > F_crit
reject_null
```

```
## [1] TRUE
```

Conclusion

If `reject_null` is TRUE, we reject the null hypothesis and conclude that **at least one coefficient** among `lstat`, `nox`, `dis`, or `rm` significantly contributes to the model.

d.

Testing a single coefficient β_j using a t-test

We test:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

The test statistic is:

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot [(X^\top X)^{-1}]_{(j+1)(j+1)}}}$$

Where: - $\hat{\beta}_j$ is the estimated coefficient - $\hat{\sigma}^2 = 5.396^2$ - $[(X^\top X)^{-1}]_{(j+1)(j+1)}$ is the $(j+1)$-th diagonal entry of the inverse matrix - Index shift of +1 is needed because indexing includes the intercept (e.g., `lstat` is index 2)

Calculate the t-statistic manually for `nox` (j = 2)

```
# Define estimated values
sigma2 <- 5.396^2
beta_hat <- c(12.03600, -0.65865, -14.53791, -0.96699, 4.86340)

# Index for "nox" is 3rd coefficient, so (j+1) = 3 + 1 = 4
j <- 3
numerator <- beta_hat[j + 1]  # -14.53791
denominator <- sqrt(sigma2 * XTX_inv_matrix[j + 1, j + 1])
t_stat_nox <- numerator / denominator
t_stat_nox
```

```
## [1] -5.366596
```

Compare to critical t-value

We compare to:

$$t_{0.975,501} \approx 1.964$$

```
t_crit <- qt(0.975, df = 501)
abs(t_stat_nox) > t_crit
```

```
## [1] TRUE
```

Conclusion

Since the **absolute t-statistic is larger than 1.964**, we **reject** $H_0 : \beta_{\text{nox}} = 0$ at the 5% significance level. This means the variable `nox` has a **statistically significant effect** on the median house price.

e.

We are comparing two cities that are **identical** except for the value of `dis`: - City A: `dis = 3` - City B: `dis = 2` All other values are:

$$x_{\text{base}} = \begin{bmatrix} 1 \\ 0.13 \\ 0.5 \\ \text{dis} \\ 4.5 \end{bmatrix}$$

We want to test whether the **difference in predicted medv** between these two cities is statistically significant at the 5% level.

Define predictor vectors

```r
# Base vector
x_cityA <- c(1, 0.13, 0.5, 3, 4.5)
x_cityB <- c(1, 0.13, 0.5, 2, 4.5)
```

Predicted difference:

$$\Delta\hat{y} = x_{\mathrm{A}}^{\top}\hat{\beta} - x_{\mathrm{B}}^{\top}\hat{\beta}$$

```r
beta_hat <- c(12.03600, -0.65865, -14.53791, -0.96699, 4.86340)

diff_pred <- sum(x_cityA * beta_hat) - sum(x_cityB * beta_hat)
diff_pred
```

```
## [1] -0.96699
```

Standard error of the difference:

We compute:

$$\mathrm{Var}(\Delta\hat{y}) = \hat{\sigma}^2 \cdot (x_A - x_B)^{\top}(X^{\top}X)^{-1}(x_A - x_B)$$

```r
sigma2 <- 5.396^2
diff_vec <- matrix(x_cityA - x_cityB, nrow = 1)

var_diff <- diff_vec %*% XTX_inv_matrix %*% t(diff_vec) * sigma2
se_diff <- sqrt(var_diff)
```

Test statistic:

$$t = \frac{\Delta\hat{y}}{\mathrm{SE}(\Delta\hat{y})}$$

```r
t_value <- diff_pred / se_diff
t_value
```

```
##           [,1]
## [1,] -5.366596
```

Critical value at 5% significance level:

```r
t_crit <- qt(0.975, df = 501)
t_crit
```

```
## [1] 1.96471
```

Conclusion:

```r
significant <- abs(t_value) > t_crit
significant
```

```
##      [,1]
## [1,] TRUE
```

If `significant = TRUE`, we reject the null hypothesis that the effect is zero, and we conclude that **the difference in distance has a statistically significant effect** on `medv`.

If the sign of `diff_pred` is **positive**, then City A (dis = 3) leads to higher `medv`. If **negative**, City B (dis = 2) is better.