Stat 570 HW1

Dongyang Wang

2022-10-06

Computation Part

The betas reproduced using matrix calculations are as follows.

```
##            [,1]
##    111.72848064
## x1  -1.26794109
## x2   0.06491817
## x3  -0.03927674
## x4  -3.18136579
## x5   0.51235896
## x6  -0.05205019
```

The standard errors reproduced using matrix calculations are as follows.

```
##           [,1]
##    47.31810073
## x1  0.62117952
## x2  0.01574825
## x3  0.01513274
## x4  1.81501910
## x5  0.36275507
## x6  0.16201386
```

The t-statistics and p-values reproduced using matrix calculations are as follows.

```
##         [,1]
##     2.361221
## x1 -2.041183
## x2  4.122245
## x3 -2.595482
## x4 -1.752800
## x5  1.412410
## x6 -0.321270
```

```
##            [,1]
##    0.0240867374
## x1 0.0490557189
## x2 0.0002277862
## x3 0.0138461970
## x4 0.0886503978
## x5 0.1669175999
## x6 0.7499724652
```

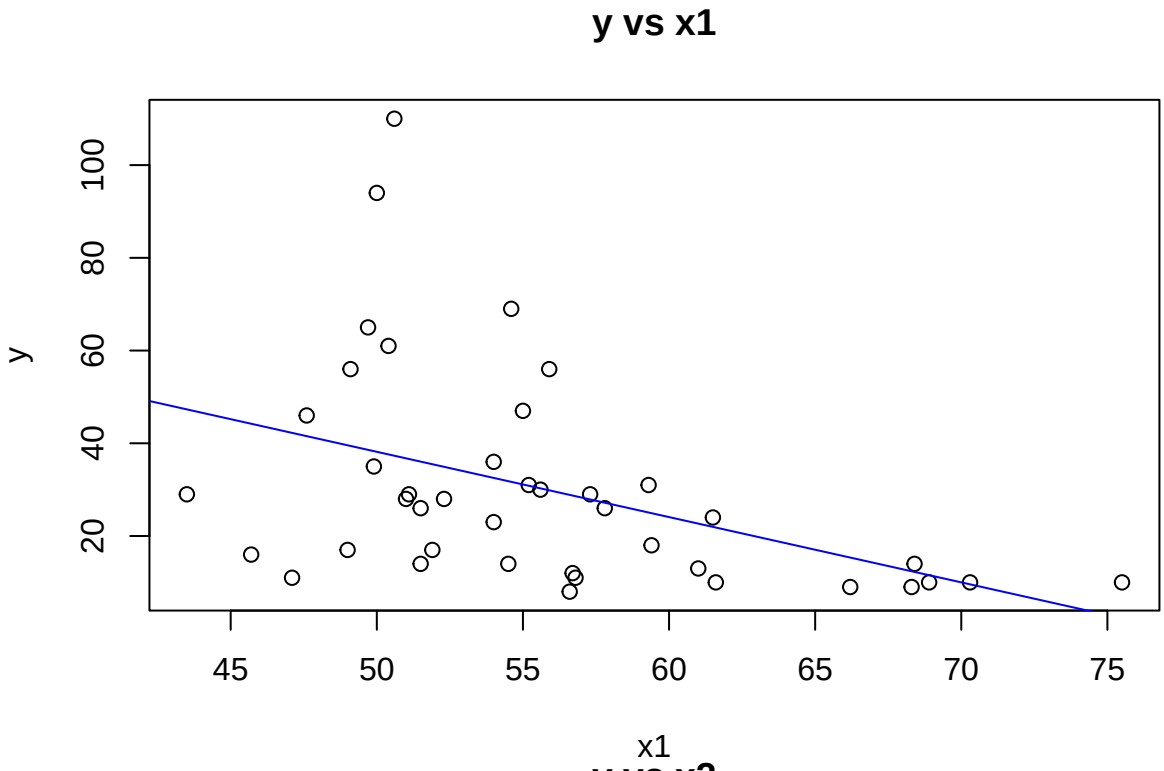
Therefore, the entire replicated result is as follows.

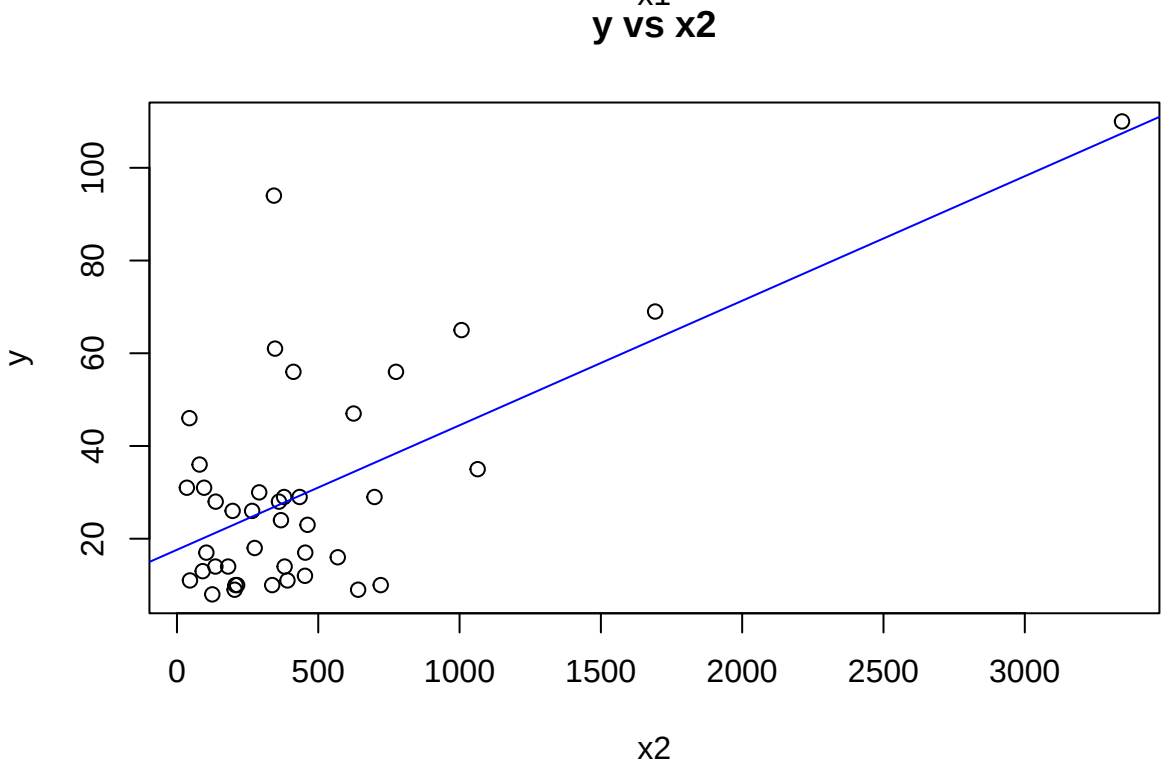
```
##                 Estimate  Std. Error   t value      p-value
```

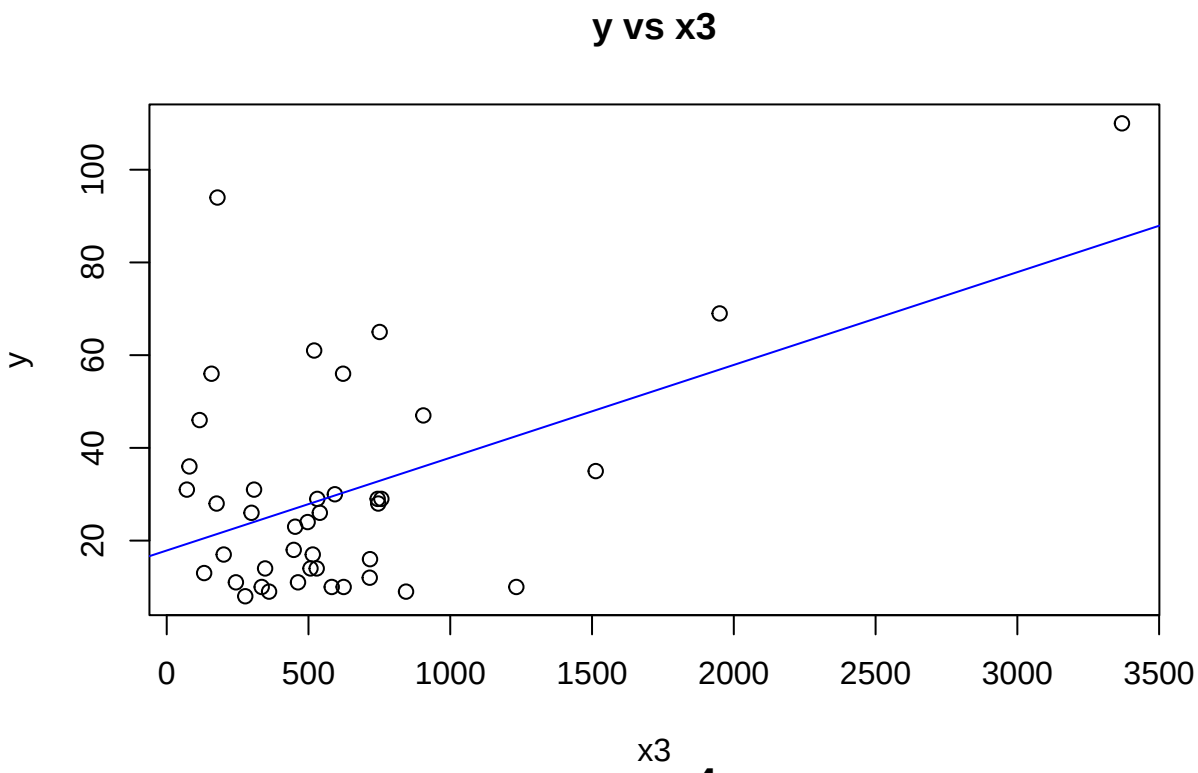
```
## (Intercept) 111.72848064 47.31810073  2.361221 0.0240867374
## x1           -1.26794109  0.62117952 -2.041183 0.0490557189
## x2            0.06491817  0.01574825  4.122245 0.0002277862
## x3           -0.03927674  0.01513274 -2.595482 0.0138461970
## x4           -3.18136579  1.81501910 -1.752800 0.0886503978
## x5            0.51235896  0.36275507  1.412410 0.1669175999
## x6           -0.05205019  0.16201386 -0.321270 0.7499724652
```

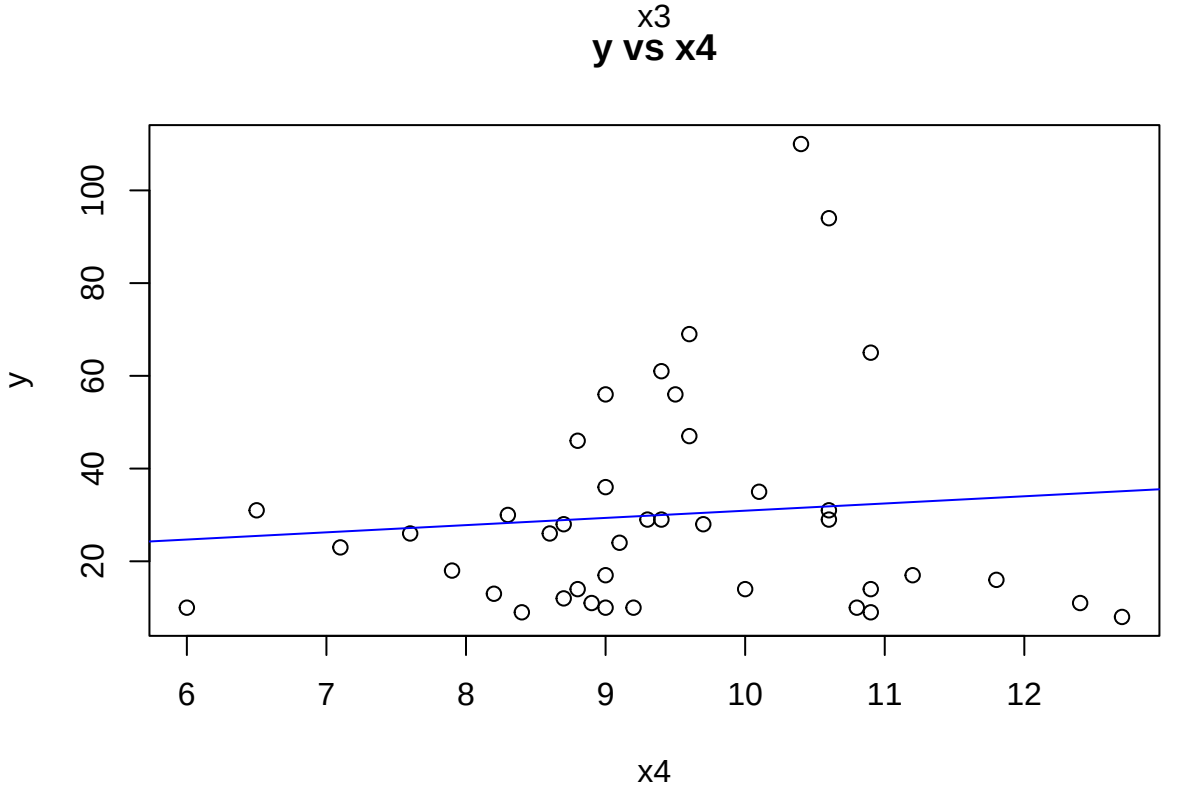
Interpretation Part

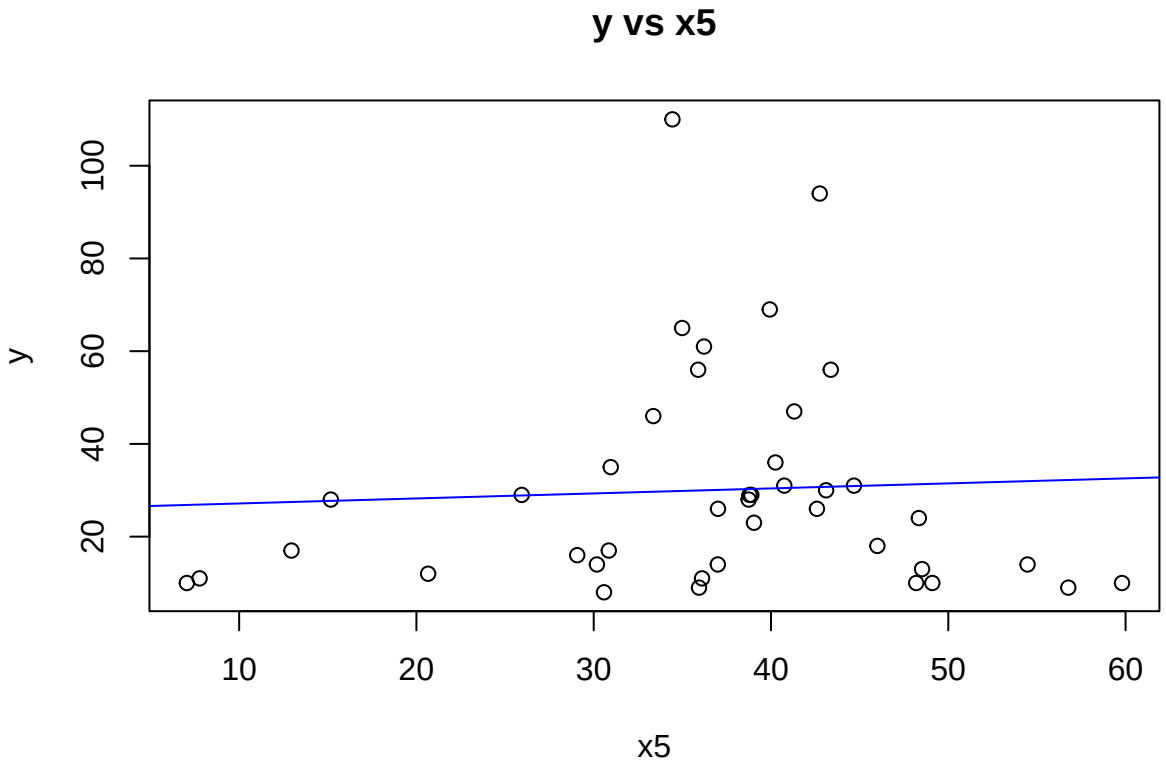
1

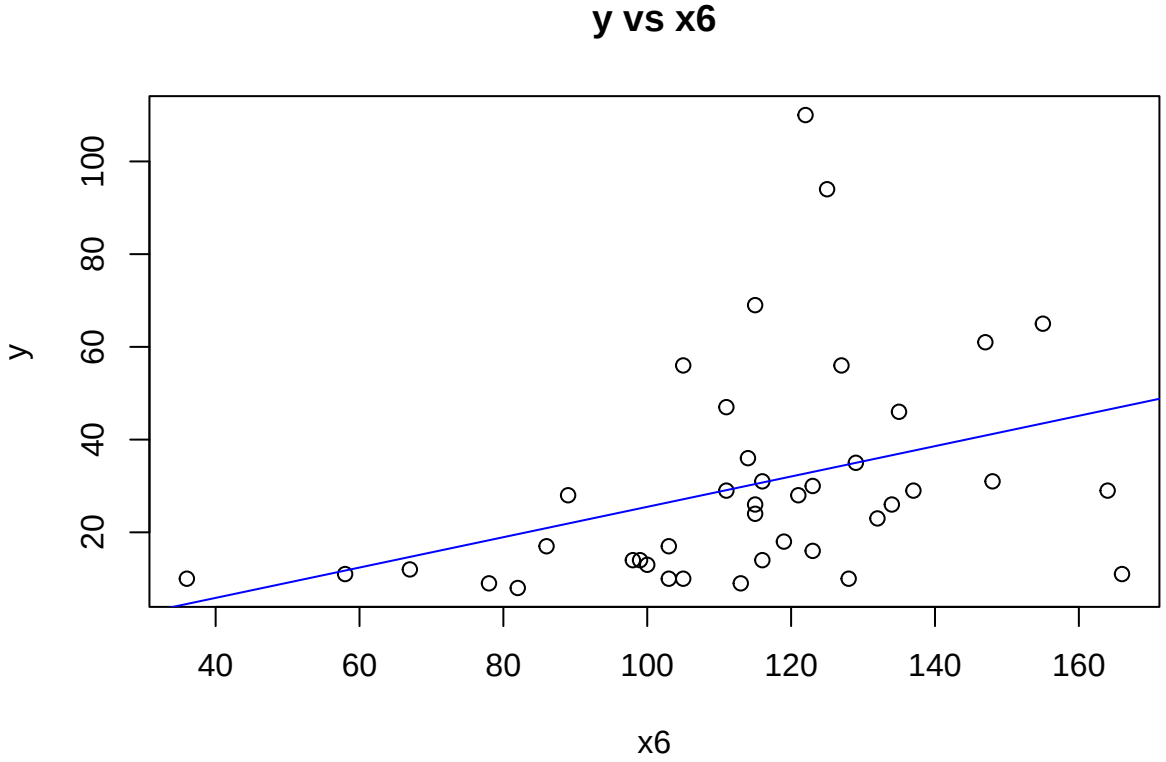

y vs x1
y
20
40
60
80
100
45
50
55
60
65
70
75
x1


y vs x2
y
20
40
60
80
100
0
500
1000
1500
2000
2500
3000
x2


y vs x3
y
20
40
60
80
100
0
500
1000
1500
2000
2500
3000
3500
x3


y vs x4
y
20
40
60
80
100
6
7
8
9
10
11
12
x4





Based on the plots above, we can observe the relationship between the response variable and x1 is negatively linear. The relationship between the response variable and x2, x3 can be observed as linear (positive correlation), although the regression lines seem tilted by some outliers. The relationship between the response variable and x4, x5 does not appear significant at all. The relationship between response and x6 seems positively linearly correlated but is also affected by extreme values.

2

The interpretation of the parameters is as follows.

intercept For the intercept, it means that the SO2 level is predicted to be 111.72848064 if all other variables are 0. This would practically very unlikely though.

x1 For $\hat{\beta}_1$, -1.26794109 is the change in SO2 level if there is 1 degree increase in average annual temperature in F.

x2 For $\hat{\beta}_2$, 0.06491817 is the change in SO2 level if there is 1 additional manufacturer employing > 20 workers.

x3 For $\hat{\beta}_3$, -0.03927674 is the change in SO2 level if there is 1 additional thousand in population.

x4 For $\hat{\beta}_4$, -3.18137 is the change in SO2 level if average annual wind speed goes up by 1 miles per hour.

x5 For $\hat{\beta}_5$, 0.51236 is the change in SO2 level if average annual rainfall increases by 1 inch.

x6 For $\hat{\beta}_6$, -0.05205 is the change in SO2 level if average number of days rainfall per year increases by 1.

Assumptions Part

1

$E(\hat{\beta}) = E((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^TE(X\beta) = \beta$. Therefore, $\hat{\beta}_j$ will always be unbiased for β_j. No assumption needed.

2

$Cov(\hat{\beta}|X) = Cov((X^TX)^{-1}X^TY|X) = (X^TX)^{-1}X^TCov(Y|X)X^T(X^TX)^{-1}$. If $Cov(Y|X) = \sigma^2 I_n$, then we would have $Cov(\hat{\beta}|X) = \sigma^2 I_n(X^TX)^{-1}$, which can be estimated by $RSS/(n-p-1)$. For $Cov(\hat{\beta}|X) = \sigma^2 I_n(X^TX)^{-1}$ to hold, we would need the assumption that $\epsilon \sim^{iid} N(0, \sigma^2)$.

3

We need the Central Limit Theorem to hold. Assumptions include $E(\hat{\beta}) = \beta$, $\hat{Var}(\hat{\beta})$ is consistent with respect to $Var(\beta)$, as well as $\hat{\beta}$ follows normal distribution. $E(\hat{\beta}) = \beta$ is true under all cases so only the latter two assumptions are required.

4

Similar to the situation above. We need the Central Limit Theorem to hold. Assumptions include $E(\hat{\beta}) = \beta$, $\hat{Var}(\hat{\beta})$ is consistent with respect to $Var(\beta)$, as well as $\hat{\beta}$ follows normal distribution. $E(\hat{\beta}) = \beta$ is true under all cases so only the latter two assumptions are required. The difference is that we are using different distributions to construct the interval. There will be a slight difference if the sample size is small.

5

We need the assumption that $E(\hat{Y}|X = x_0) = E(Y|X = x_0)$ meaning that the linear model should be correct in order to get an accurate prediction.

Appendix

```
library(gamlss.data)
data(usair)
lmod <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = usair)
summary(lmod)

Y <- as.matrix(usair[,1])
X <- as.matrix(usair[,-1])
ones <- as.matrix(rep(1, length(Y)), nrow = length(Y))
X <- cbind(ones, X)
beta <- solve(t(X) %*% X) %*% t(X) %*% Y

beta

sigma_hat <- sum((Y - X %*% beta)^2)/(nrow(X) - ncol(X))
sd_hat <- sqrt(diag(sigma_hat * solve(t(X) %*% X)))
sd_hat <- as.matrix(sd_hat, ncol = 1)

sd_hat

t_stat <- beta/sd_hat
p_val <- 2*(1-pt(abs(t_stat), df = nrow(X) - ncol(X)))

t_stat
p_val

res <- cbind(beta, sd_hat, t_stat, p_val)
rownames(res) <- c("(Intercept)", "x1", "x2", "x3", "x4", "x5", "x6")
colnames(res) <- c("Estimate", "Std. Error", "t value", "p-value")

res

attach(usair)
p1 <- plot(x1, y, main="y vs x1")
abline(lm(y ~ x1, data = usair), col = "blue")

p2 <- plot(x2, y, main="y vs x2")
abline(lm(y ~ x2, data = usair), col = "blue")

plot(x3, y, main="y vs x3")
abline(lm(y ~ x3, data = usair), col = "blue")

plot(x4, y, main="y vs x4")
abline(lm(y ~ x4, data = usair), col = "blue")

plot(x5, y, main="y vs x5")
abline(lm(y ~ x5, data = usair), col = "blue")

plot(x6, y, main="y vs x6")
abline(lm(y ~ x6, data = usair), col = "blue")

detach(usair)
```