Homework2

Nam Jun Lee

10/14/2021

Q1. Do Question 14. of Chapter 3 of the ISLR book. (Page 125).

a. Perform the following commands in R:

```
set.seed(1)
x1 = runif(100)
x2 = 0.5 * x1 + rnorm(100)/10
y = 2 + 2 * x1 + 0.3 * x2 + rnorm(100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

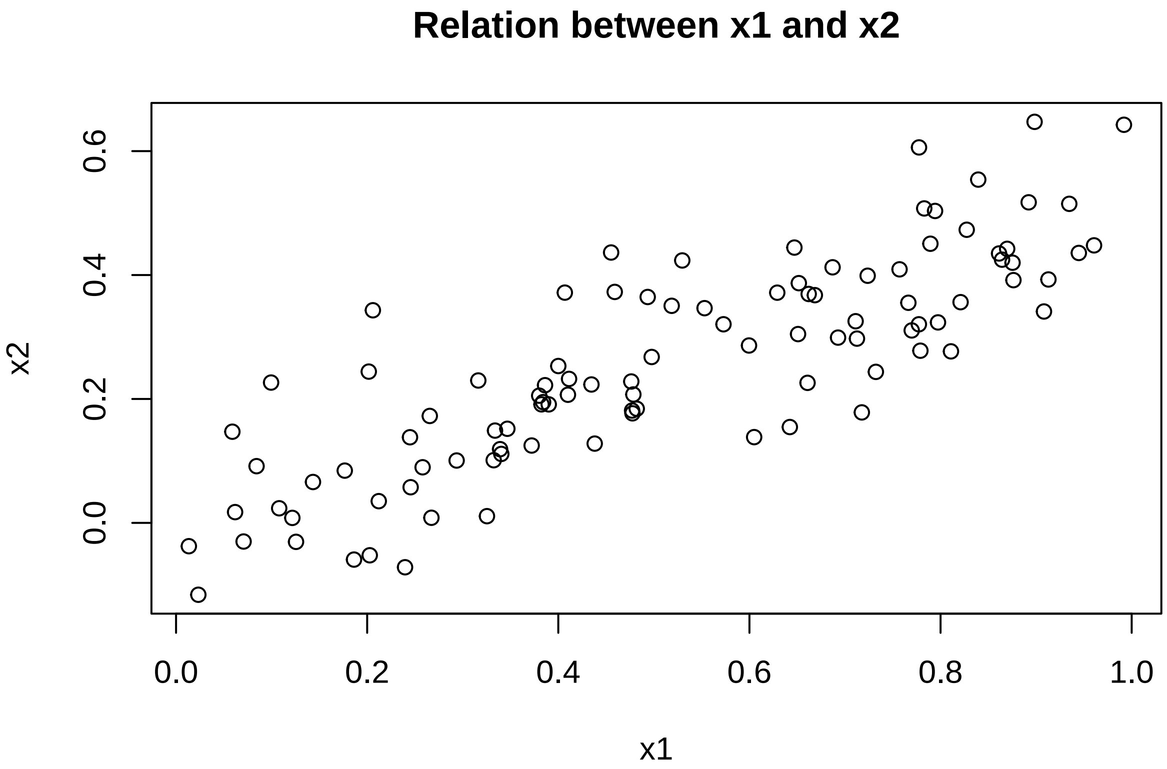
Regression Coeffcients:
$\beta0 = \mathbf{2 + rnorm(100)}$
$\beta1 = \mathbf{2}$
$\beta3 = \mathbf{0.3}$

b. What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

```
# correlation between x1 & x2
cor(x1, x2)
```

```
## [1] 0.8351212
```

```
# scatterplot x1 and x2
plot(x1, x2, main = "Relation between x1 and x2", xlab = "x1", ylab = "x2")
```


Relation between x1 and x2

Correlation between x1 and x2 is **0.8351212**. It means between x1 and x2 have a **high positive correlation**.

c. Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\beta 0$, $\beta 1$, and $\beta 2$? How do these relate to the true $\beta 0$, $\beta 1$, and $\beta 2$? Can you reject the null hypothesis H0 : $\beta 1 = 0$? How about the null hypothesis H0 : $\beta 2 = 0$?

```
lm.fit <- lm(y ~ x1 + x2)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

Coefficient estimates: $\beta 0 =$ **2.1305**, $\beta 1 =$ **1.4396**, $\beta 2 =$ **1.0097** It means there are poor estimates.
$\beta 1$, **can only reject the null hypothesis** at a 99 % lv of confindence. (less than 0.05)
$\beta 2$, **may not reject the null hypothesis**. (higher than 0.05)

d. Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis H0 : $\beta 1 = 0$?

```
lm.fit1 <- lm(y ~ x1)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

In this case x1 is highly significant as p-value is very lower, therefore, **may reject** $\beta 0$.

e. Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis H0 : $\beta 1 = 0$?

```
lm.fit2 <- lm(y ~ x2)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

In this case x2 is highly significant as p-value is very lower, therefore, **may reject** $\beta 0$.

f. Do the results obtained in (c)–(e) contradict each other? Explain your answer.

In part (c), it was found that **x1 and x2 were not significant** based on the multiple linear regression model.
However, part (d) and part (e) show that **each x1 and x2 is actually highly significant**. Therefore, this was *contradictory*.

g. Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```
x1 = c(x1, 0.1)
x2 = c(x2, 0.8)
y = c(y, 6)
```

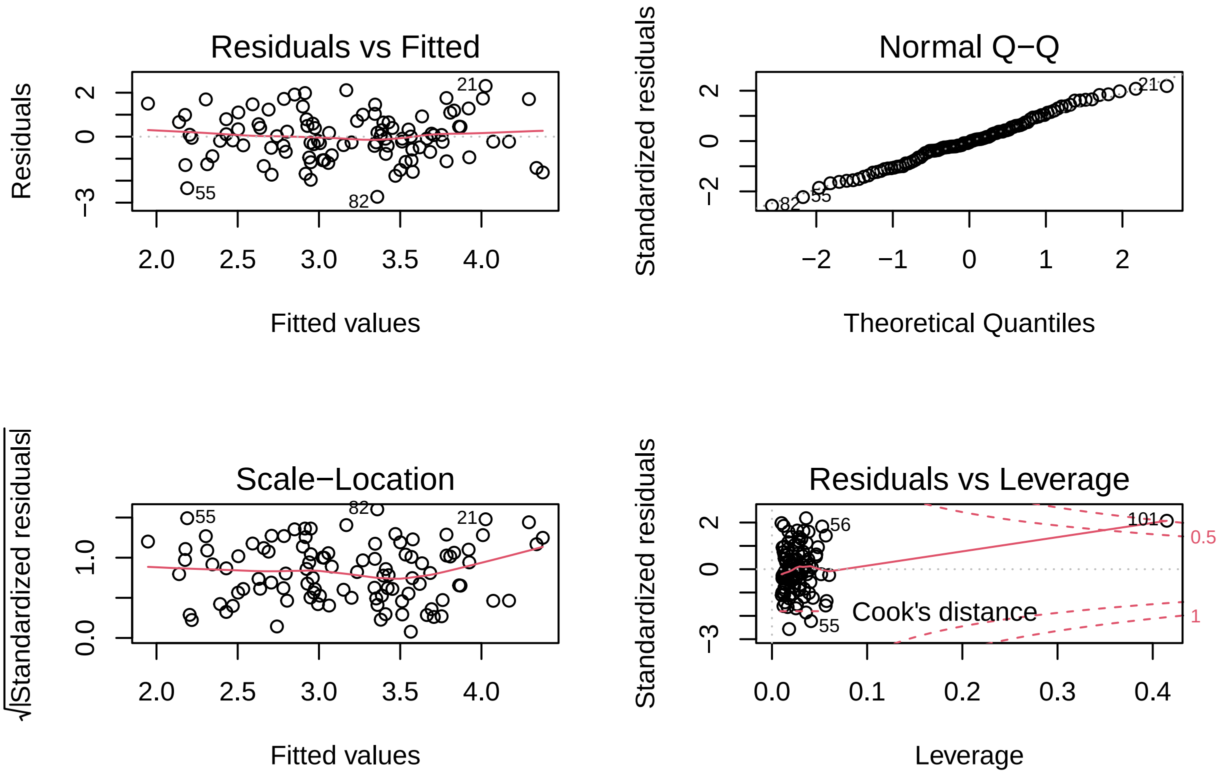
Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

```
# refit (c)
lm.fit3 = lm(y ~ x1 + x2)
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

```
par(mfrow = c(2, 2))
plot(lm.fit3)
```



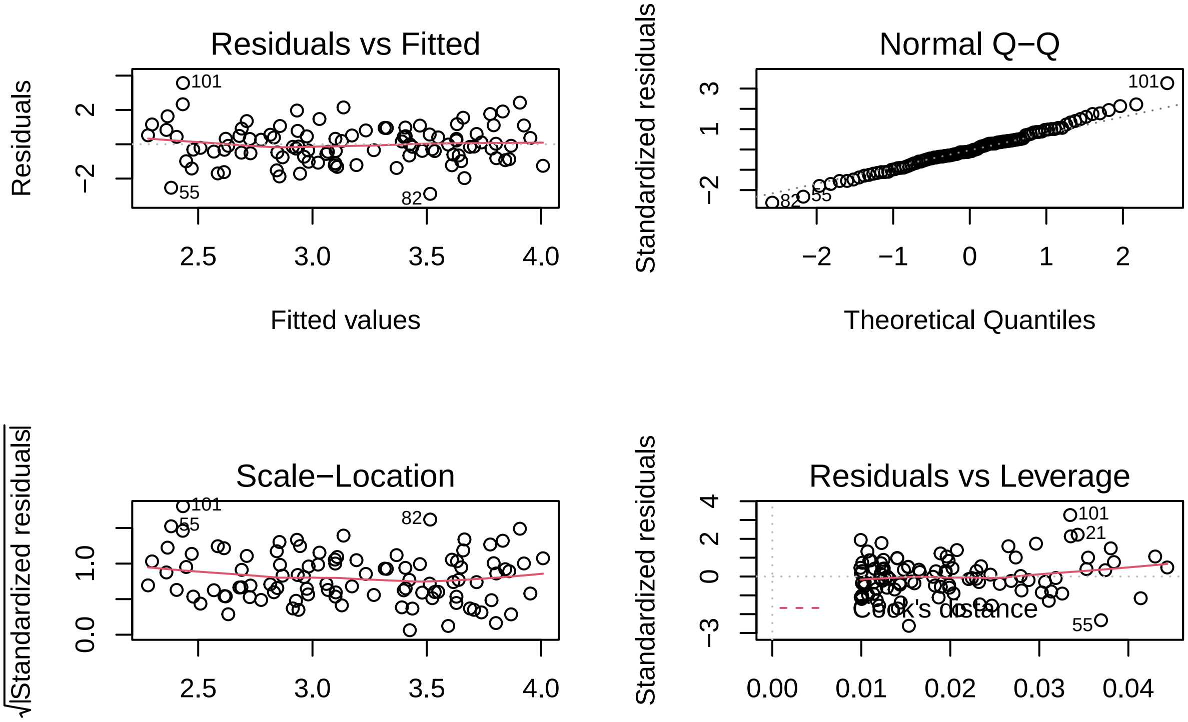
Although the adjusted R-square value has risen slightly, it is still significantly lower. (20.29%)
Compared to the previous model, it can be seen that the x2 variable is statistically introduced, and conversely, the x1 variable is not significant.
The Residuals versus Fitted plot does follow a normal distribution with constant variance. The Scale-Location plot has several outliers, and the figure of Residuals versus Leverage plot indicates that there are several leverage points.

```
# refit (d)
lm.fit4 = lm(y ~ x1)
summary(lm.fit4)
```

```
##
```

```
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

```
par(mfrow = c(2, 2))
plot(lm.fit4)
```



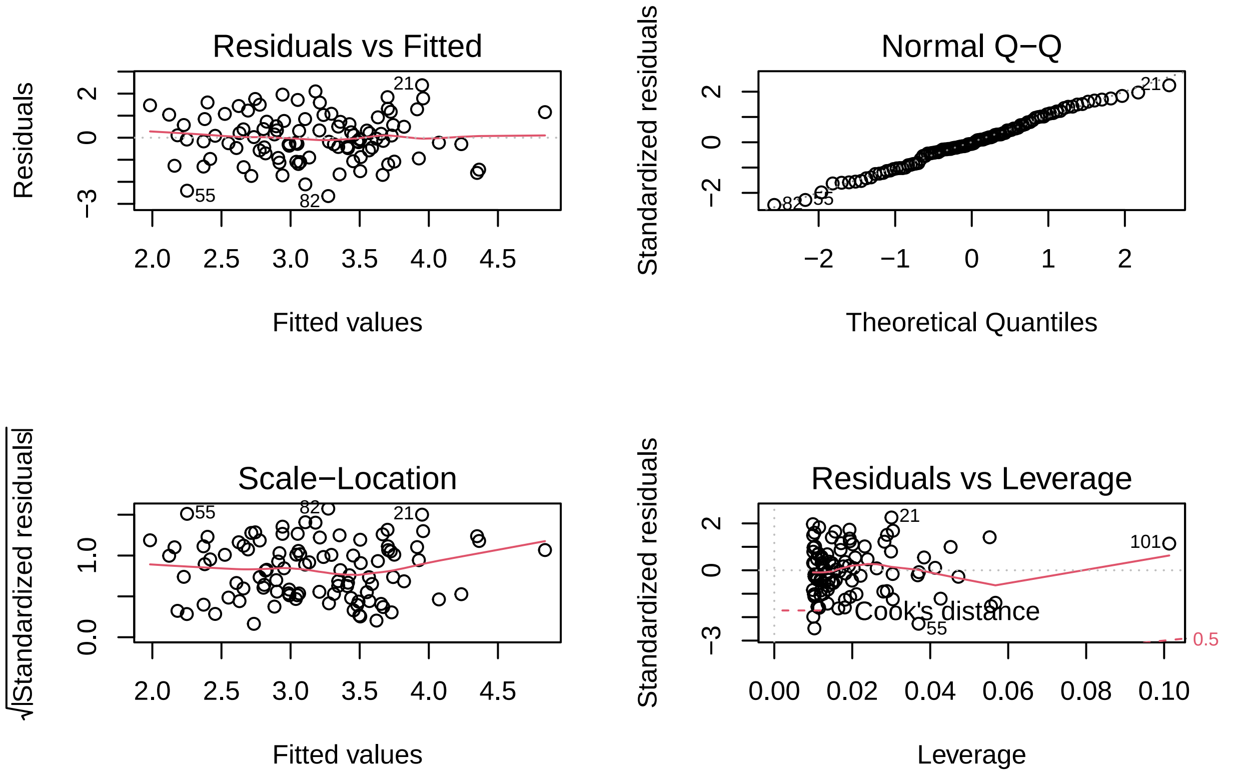
In both the previous model and the changed model, x1 is statistically significant for the fluctuation of y. However, the adjusted R-square value was lowered and this shows that the changed model is a worse model than the previous one. (19.42% -> 14.77%)

The Residuals versus Fitted plot does follow a normal distribution with constant variance but has several outliers. The Scale-Location plot has also several outliers, and the figure of Residuals versus Leverage plot indicates that there are several leverage points.

```
# refit (e)
lm.fit5 = lm(y ~ x2)
summary(lm.fit5)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

```
# diagnostic plots of the least squares regression (lm.fit5)
par(mfrow = c(2, 2))
plot(lm.fit5)
```

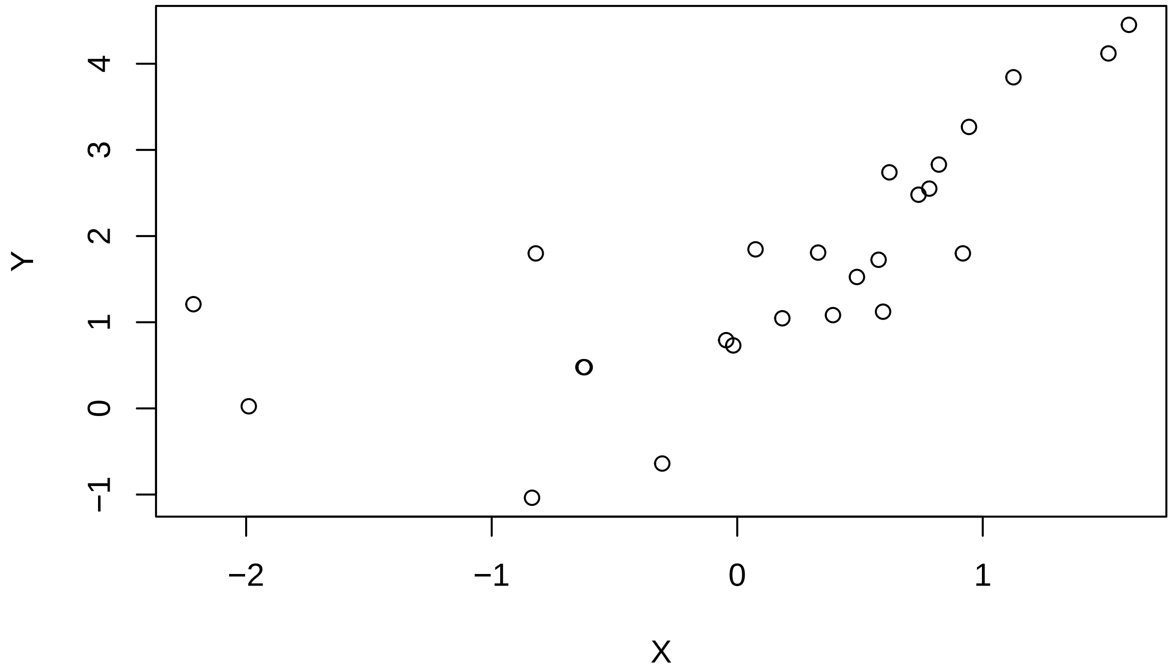


In both the previous and changed models, x2 is statistically significant for fluctuations in y. However, the adjusted R-square value increased, indicating that the changed model is a better model than the previous model. (16.79% -> 20.42%)
The Residuals versus Fitted plot does follow a normal distribution with constant variance but has several outliers. The Scale-Location plot has also several outliers, Normal Q-Q plot has also some outliers, and the figure of Residuals versus Leverage plot indicates that there are several leverage points.

Q2. Before attempting this question, set the seed number in R by using set.seed(1) to ensure consistent results.

a. Simulate a training data set of $n = 25$ observations as $y = exp(x) + \epsilon$ where x and ϵ are generated via a normal distribution with mean zero and standard deviation one. (use rnorm() to simulate these variables).

```
set.seed(1)
n = 25
X <- rnorm(n, mean = 0, sd = 1)
E <- rnorm(n, mean = 0, sd = 1)
Y = exp(X) + E
plot(X, Y)
```



b. Fit the following four linear regression models to the above training data set (using the lm() function in R).

(i) y = $\beta 0$ **+** $\beta 1$**x +** ϵ **.**

```
fit.first <- lm(Y ~ X)
summary(fit.first)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81736 -0.66341 -0.04611  0.64348  2.06862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5029     0.1965   7.647 9.22e-08 ***
## X             1.0666     0.2077   5.135 3.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9669 on 23 degrees of freedom
## Multiple R-squared:  0.5341, Adjusted R-squared:  0.5138
## F-statistic: 26.36 on 1 and 23 DF,  p-value: 3.343e-05
```

(ii) $\mathbf{y} = \beta 0 + \beta 1\mathbf{x} + \beta 2x^2 + \epsilon$**.**

```
fit.second <- lm(Y ~ X + I(X^2))
summary(fit.second)
```

```
##
## Call:
## lm(formula = Y ~ X + I(X^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21546 -0.25219 -0.05994  0.37896  1.61475
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.9123     0.1824   5.002 5.24e-05 ***
## X             1.3803     0.1591   8.676 1.50e-08 ***
## I(X^2)        0.6007     0.1212   4.958 5.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6794 on 22 degrees of freedom
## Multiple R-squared:  0.7799, Adjusted R-squared:  0.7599
## F-statistic: 38.99 on 2 and 22 DF,  p-value: 5.859e-08
```

(iii) $\mathbf{y} = \beta 0 + \beta 1\mathbf{x} + \beta 2x^2 + \beta 3x^3 + \epsilon$**.**

```
fit.third <- lm(Y ~ X + I(X^2) + I(X^3))
summary(fit.third)
```

```
##
## Call:
## lm(formula = Y ~ X + I(X^2) + I(X^3))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15835 -0.23662 -0.05723  0.34600  1.68884
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.91380    0.18625   4.906 7.48e-05 ***
## X            1.47285    0.32494   4.533 0.000182 ***
## I(X^2)       0.56796    0.15885   3.575 0.001785 **
## I(X^3)      -0.04049    0.12318  -0.329 0.745636
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6936 on 21 degrees of freedom
## Multiple R-squared:  0.7811, Adjusted R-squared:  0.7498
## F-statistic: 24.97 on 3 and 21 DF,  p-value: 4.007e-07
```

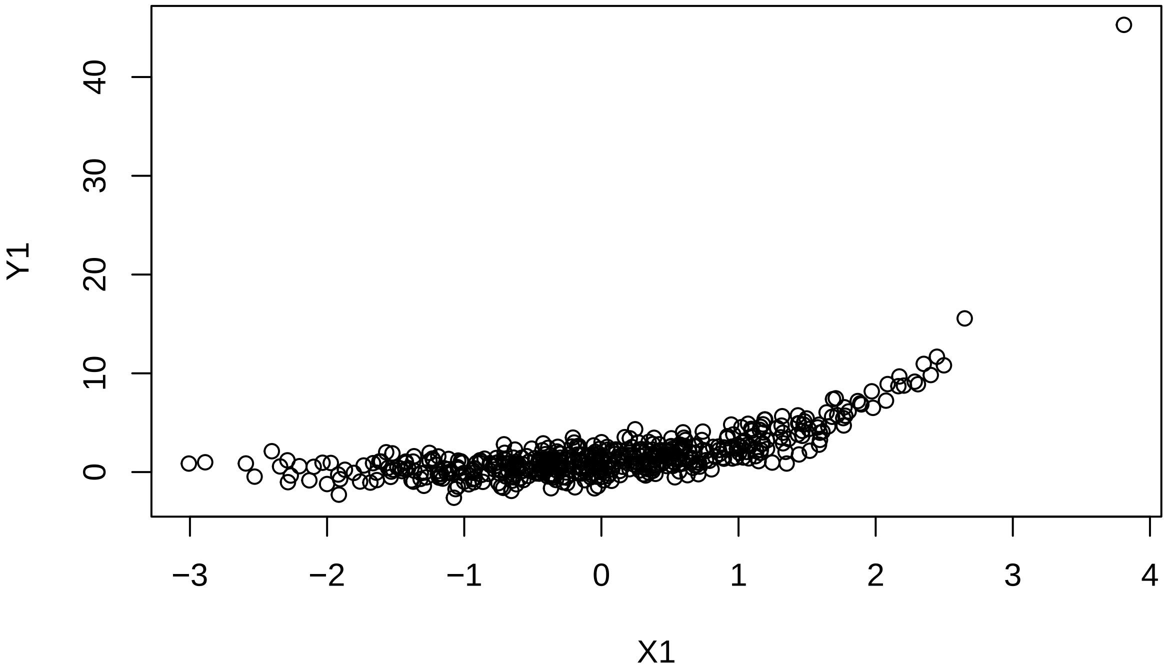
(iv) y = $\beta 0$ + $\beta 1$x + $\beta 2x^2$ + $\beta 3x^3$ + $\beta 4x^4$ + ϵ.

```r
fit.fourth <- lm(Y ~ X + I(X^2) + I(X^3) + I(X^4))
summary(fit.fourth)
```

```
##
## Call:
## lm(formula = Y ~ X + I(X^2) + I(X^3) + I(X^4))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17341 -0.23125 -0.02382  0.36720  1.65807
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.88069    0.25537   3.449 0.002540 **
## X            1.48465    0.33812   4.391 0.000282 ***
## I(X^2)       0.68004    0.59763   1.138 0.268607
## I(X^3)      -0.06437    0.17583  -0.366 0.718130
## I(X^4)      -0.03329    0.17078  -0.195 0.847431
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7101 on 20 degrees of freedom
## Multiple R-squared:  0.7815, Adjusted R-squared:  0.7378
## F-statistic: 17.88 on 4 and 20 DF,  p-value: 2.188e-06
```

c. Now simulate a testing data set with n = 500 observations from the model in part (a), by generating new values of x and ϵ.

```r
# new values of x and E
n1 = 500
X1 <- rnorm(n1, mean = 0, sd = 1)
E1 <- rnorm(n1, mean = 0, sd = 1)
Y1 = exp(X1) + E1
plot(X1, Y1)
```



d. Use the estimated coefficients in Part (b) to compute the test error, i.e. the MSE $= \frac{1}{n}\sum(y_i - \widehat{y_i})^2$ of the testing data set for each of the four models computed in part (b).

First Model:

```
b = rep(1, n1)
x1 = matrix(c(b, X1), byrow = F, nrow = n1)
b1 = as.matrix(fit.first$coefficients)
widehat1 = x1 %*% b1
MSE1 = sum((Y1 - widehat1)^2)/n1
MSE1
```

```
## [1] 5.70084
```

Second Model:

```
x2 = matrix(c(b, X1, I(X1^2)), byrow = F, nrow = n1)
b2 = as.matrix(fit.second$coefficients)
widehat2 = x2 %*% b2
MSE2 = sum((Y1 - widehat2)^2)/n1
MSE2
```

```
## [1] 3.232397
```

Third Model:

```
x3 = matrix(c(b, X1, I(X1^2), I(X1^3)), byrow = F, nrow = n1)
b3 = as.matrix(fit.third$coefficients)
widehat3 = x3 %*% b3
MSE3 = sum((Y1 - widehat3)^2)/n1
MSE3
```

```
## [1] 3.631079
```

Fourth Model:

```
x4 = matrix(c(b, X1, I(X1^2), I(X1^3), I(X1^4)), byrow = F, nrow = n1)
b4 = as.matrix(fit.fourth$coefficients)
widehat4 = x4 %*% b4
MSE4 = sum((Y1 - widehat4)^2)/n1
MSE4
```

```
## [1] 4.733067
```

e. Based on your results of Part (b), which model would you reccommend as the 'best fit model'? is the conclusion suprising?

See above all four models MSE values.
First model MSE: **5.70084**
Second model MSE: **3.232397**
Third model MSE: **3.631079**
Fourth model MSE: **4.733067**
Second model MSE value is 3.232397; this is lowest MSE among that 4 models. Therefore, best fit model is [**Second model**].
Equation: $Y = 0.9123 + 1.3803 * X + 0.6007 * X^2$.

Q3. Consider the Hitters data in the ISLR package, our objective here is to predict the salary variable as the response using the remaining variables.

```
# Hitters dataset into hitters
hitters <- ISLR::Hitters
# columns in hitters
names(hitters)
```

```
##  [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"       "Walks"
##  [7] "Years"     "CAtBat"    "CHits"     "CHmRun"    "CRuns"     "CRBI"
## [13] "CWalks"    "League"    "Division"  "PutOuts"   "Assists"   "Errors"
## [19] "Salary"    "NewLeague"
```

```
# find NA in hitters Salary variable
sum(is.na(hitters$Salary))
```

```
## [1] 59
```

```
# Remove all NA in hitters dataset
hitters <- na.omit(hitters)
```

a. Split the data into a training and testing data set.

```
# Split the dataset into train and test
set.seed(1)
train = sample(c(TRUE, FALSE), nrow(hitters), rep = TRUE)
test = (!train)
hitters_train = hitters[train, ]
hitters_test = hitters[test, ]
```

b. Fit a linear model using least squares on the training set and report the test error obtained.

```
# linear model
lm.fit_hit = lm(Salary ~ ., data = hitters_train)
lm.pred = predict(lm.fit_hit, hitters_test)
test_error <- mean((lm.pred - hitters_test$Salary)^2)
test_error
```

```
## [1] 142238.2
```

c. Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

```
# set up matrices neede for glmnet fucntions
train_mo = model.matrix(Salary ~ ., data = hitters_train)
test_mo = model.matrix(Salary ~ ., data = hitters_test)
# cross validation by ridge regression alpha = 0
set.seed(1)
cv1 <- cv.glmnet(train_mo, hitters_train$Salary, alpha = 0)
lam1 = cv1$lambda.min
ridge_model = glmnet(train_mo, hitters_train$Salary, alpha = 0)
ridge_pred = predict(ridge_model, s = lam1, newx = test_mo)
ridge_test_error <- mean((ridge_pred - hitters_test$Salary)^2)
ridge_test_error
```

```
## [1] 145635.3
```

d. Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficients estimates.

```
# cross validation by lasso model alpha = 1
set.seed(1)
cv2 <- cv.glmnet(train_mo, hitters_train$Salary, alpha = 1)
lam2 = cv2$lambda.min
lasso_model = glmnet(train_mo, hitters_train$Salary, alpha = 1)
lasso_pred = predict(lasso_model, s = lam2, newx = test_mo)
lasso_test_error = mean((lasso_pred - hitters_test$Salary)^2)
lasso_test_error
```

```
## [1] 141599.6
```

e. Commment on the results obtained. How accurately can we predict the number of college applications recieved? Is there much difference among the test errors resulting from these three approaches?

```
# compute total average
total_avg = sum((mean(hitters_test$Salary) - hitters_test$Salary)^2)
# ridge
total_ridge = sum((ridge_pred - hitters_test$Salary)^2)
# lasso
total_lasso = sum((lasso_pred - hitters_test$Salary)^2)
# linear
total_linear = sum((lm.pred - hitters_test$Salary)^2)
# compute ridge R square value
ridge_mod <- 1 - (total_ridge)/(total_avg)
# compute lasso R square value
lasso_mod <- 1 - (total_lasso)/(total_avg)
# compute linear R square value
linear_mod <- 1 - (total_linear)/(total_avg)
# show each MSE values
ridge_test_error
```

```
## [1] 145635.3
```

```
lasso_test_error
```

```
## [1] 141599.6
```

```
test_error
```

```
## [1] 142238.2
```

```
# show each ridge, lasso, and linear R square values
ridge_mod
```

```
## [1] 0.2869518
```

```
lasso_mod
```

```
## [1] 0.306711
```

```
linear_mod
```

```
## [1] 0.3035847
```

Compare MSE for each three models:
Ridge model MSE: **145635.3**.
Lasso model MSE: **141599.6**.
Linear model MSE: **142238.2**.
Best MSE value:

$$Lasso > Linear > Ridge$$

Compare R square for each three models:
Test for ridge model R square value: **0.2869518**. (28.70 %)
Test for lasso model R square value: **0.306711**. (30.67 %)
Test for linear model R square value: **0.3035847**. (30.36 %)
Best R square value:

$$Lasso > Linear > Ridge$$

Therefore, the **Lasso model seems to be the most accurately predictable**. However, there **are not many differences** in the three approaches and they **are similar**.