Homework 3 - STAT 4510/7510

Yang, Anton – #14405729

Due Wednesday, Feb. 14, 11:30 pm

Instructions: Please list your name and student number clearly. In order to receive credit for a problem, your solution must show sufficient detail so that the grader can determine how you obtained your answer.

Submit a single pdf generated using R Markdown. All R code should be included, as well as all output produced. Upload your work to the Canvas course site.

Problem 1

In this problem, we will create a simulated data set to study regression. Generate data with the following code.

```
set.seed(1)
x = runif(100,0,1)
y = 51 + 10*x + rnorm(100, mean=0, sd=.5)
SimulatedData = data.frame(cbind(x,y))
```

(a) What are the true values of β_0 and β_1? What is the true value of σ^2, the true variance of the residuals? *Hint: Look up the* `rnorm()` *function in the help.*

```
set.seed(1)
x = runif(100,0,1)
y = 51 + 10*x + rnorm(100, mean=0, sd=.5)
SimulatedData = data.frame(cbind(x,y))
```

We see that the x and y are highly correlated since the value of y depends on x significantly. Therefore, the true value of $\beta_0 = 51, \beta_1 = 10$ and true variance $\sigma^2 = 0.5^2 = 0.25$.

(b) Fit a linear model using y as the dependent variable (response) and x as the independent variable (predictor).

- Produce a summary of this analysis.
- Are the estimates of β_0, β_1, and σ^2 close to the true values you listed above?
- What proportion of the variance in y is explained by the model?

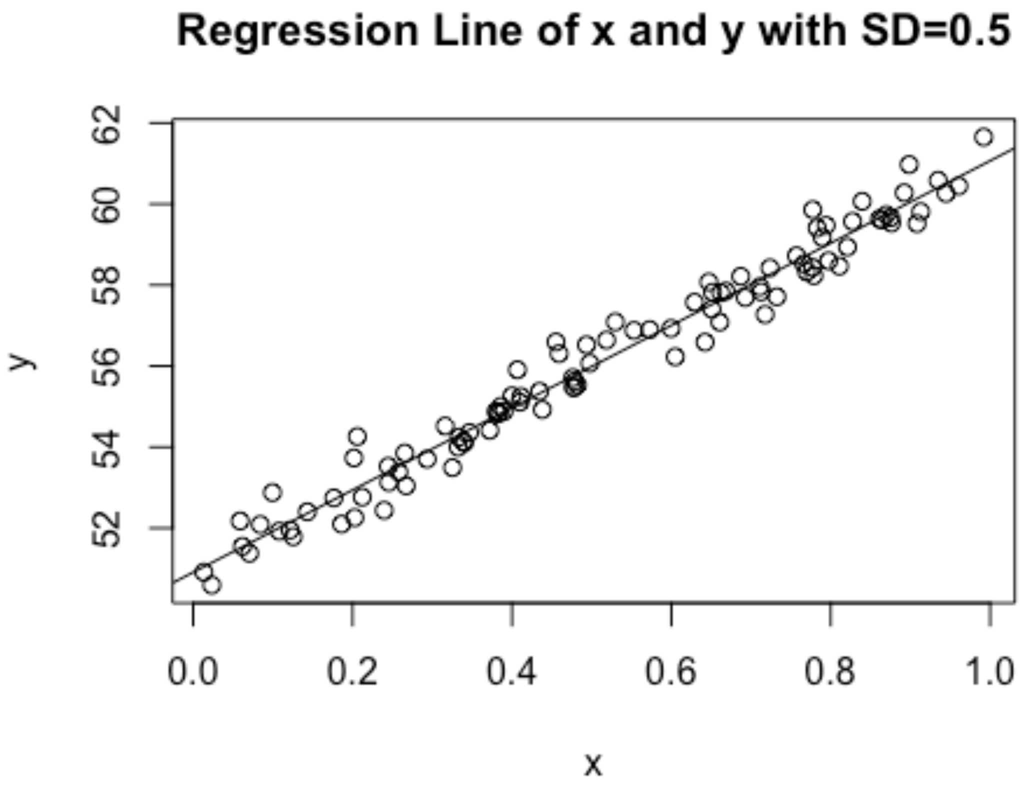
```
model<-lm(y~x, data=SimulatedData)
summary(model)

##
## Call:
## lm(formula = y ~ x, data = SimulatedData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92489 -0.28111 -0.04353  0.26214  1.25830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.9103     0.1029  494.73   <2e-16 ***
## x            10.1562     0.1767   57.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4705 on 98 degrees of freedom
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.9709
## F-statistic:  3303 on 1 and 98 DF,  p-value: < 2.2e-16
```

Yes, according to the summary, we see that the intercept is 50.9103 and x is 10.1562, which is really close to the true value. We also see that the sum of Standard error is about 0.2796 which is also really close to the true variance. We see that $R^2 = 0.9712$, which means that this model explains 97.12% of the variance in y.

(c) Produce a scatterplot of x and y, and add the regression line. Label your axes.

```
plot(SimulatedData,
     main="Regression Line of x and y with SD=0.5",
     xlab = "x",
     ylab = "y")
abline(model)
```



(d) Repeat parts (b) and (c) changing to sd = 8 in your simulated data.

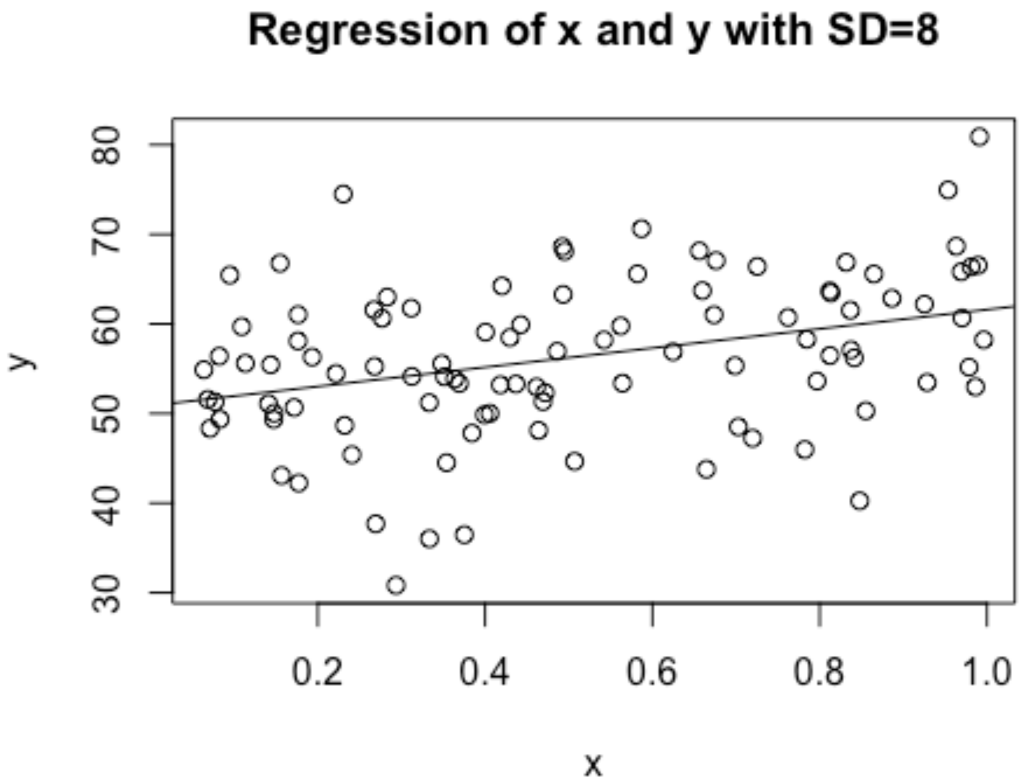
```
x1 = runif(100,0,1)
y1 = 51 + 10*x1 + rnorm(100, mean=0, sd=8)
SimulatedData2 = data.frame(cbind(x1,y1))

model2<-lm(y1~x1, data=SimulatedData2)
summary(model2)

##
## Call:
## lm(formula = y1 ~ x1, data = SimulatedData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1978  -3.9052  -0.0572   5.1365  21.1524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.875      1.671  30.437  < 2e-16 ***
## x1            10.721      2.883   3.719 0.000333 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.314 on 98 degrees of freedom
## Multiple R-squared:  0.1237, Adjusted R-squared:  0.1147
## F-statistic: 13.83 on 1 and 98 DF,  p-value: 0.0003335
```

```
plot(SimulatedData2,
     main="Regression of x and y with SD=8",
     xlab="x",
     ylab="y")
abline(model2)
```



As we can see from the model that the coefficient is little father away from the true β_0, β_1 but are still approximately the same. However, the σ^2 is not close to the true σ^2 by having an estimated variance approximately 4.554 compared to the true $\sigma^2 = 8^2 = 64$. The summary shows that the $R^2 = 0.1237$, which means that this model explains 12.37% of the variance.

(e) Fit a quadratic, cubic, and 7th degree polynomial to the noisier data you used in part (d). (In each model, include terms of lower degrees upto the indicated degree.) How does the R^2 value change as the degree of your polynomial increases? Are these models better than the linear model? Discuss.

```
model_quad<-lm(y1~poly(x1, degree =2, raw=TRUE),data=SimulatedData2)
model_cubic<-lm(y1~poly(x1,degree=3,raw=TRUE), data=SimulatedData2)
model_7th<-lm(y1~poly(x1,degree=7,raw=TRUE),data=SimulatedData2)
summary(model_quad)
```

```
##
## Call:
## lm(formula = y1 ~ poly(x1, degree = 2, raw = TRUE), data = SimulatedData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.8031  -4.2312   0.4097   4.9375  21.0402
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         53.839      2.863  18.806   <2e-16 ***
## poly(x1, degree = 2, raw = TRUE)1   -5.064     12.728  -0.398    0.692
## poly(x1, degree = 2, raw = TRUE)2   14.802     11.627   1.273    0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.287 on 97 degrees of freedom
## Multiple R-squared:  0.1381, Adjusted R-squared:  0.1203
## F-statistic: 7.769 on 2 and 97 DF,  p-value: 0.0007419
```

```
summary(model_cubic)
```

```
##
## Call:
## lm(formula = y1 ~ poly(x1, degree = 3, raw = TRUE), data = SimulatedData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6822  -4.2440   0.3691   4.9694  21.1349
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         54.304      4.479  12.123   <2e-16 ***
## poly(x1, degree = 3, raw = TRUE)1   -9.513     35.252  -0.270    0.788
## poly(x1, degree = 3, raw = TRUE)2   25.070     76.720   0.327    0.745
## poly(x1, degree = 3, raw = TRUE)3   -6.514     48.103  -0.135    0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.33 on 96 degrees of freedom
## Multiple R-squared:  0.1382, Adjusted R-squared:  0.1113
## F-statistic: 5.133 on 3 and 96 DF,  p-value: 0.002459
```

```
summary(model_7th)
```

```
##
## Call:
## lm(formula = y1 ~ poly(x1, degree = 7, raw = TRUE), data = SimulatedData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3481  -4.4263   0.3212   4.9460  21.5898 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          46.48      31.07   1.496    0.138
## poly(x1, degree = 7, raw = TRUE)1     96.06     769.90   0.125    0.901
## poly(x1, degree = 7, raw = TRUE)2    322.91    6919.13   0.047    0.963
## poly(x1, degree = 7, raw = TRUE)3  -7401.39   30231.00  -0.245    0.807
## poly(x1, degree = 7, raw = TRUE)4  30933.15   71287.11   0.434    0.665
## poly(x1, degree = 7, raw = TRUE)5 -55034.19   92583.41  -0.594    0.554
## poly(x1, degree = 7, raw = TRUE)6  44953.81   62251.11   0.722    0.472
## poly(x1, degree = 7, raw = TRUE)7 -13854.20   16909.04  -0.819    0.415
## 
## Residual standard error: 8.191 on 92 degrees of freedom
## Multiple R-squared:  0.2014, Adjusted R-squared:  0.1406 
## F-statistic: 3.314 on 7 and 92 DF,  p-value: 0.003439
```

The summary of these models clearly show that as the degree increase, the R^2 increase. It's hard to state whether the model because as we can see that the standard error of 7th degree polynomial reached to 16909.04, which is really far off from the true variance. In addition, it's also really hard to state since there's not a huge difference in R^2 between linear and polybnomial models. To test whether these models are better, we need first to know what's the goal in using the models. If the goal is to mainly predict, then we need to test the model with MSE of the test data set to know whether what model is best at predicting. If the goal is not mainly predicting but also interpretation, then the linear model might be the best since it's not R^2 is not substantially different from the polynomial models, and it's also the easiest to interpret.

Problem 2

The `UScereal` data frame is part of the `MASS` library. The data come from the 1993 ASA Statistical Graphics Exposition, and are taken from the mandatory FDA food label. The data have been normalized here to a portion of one American cup. The data contains the following columns:

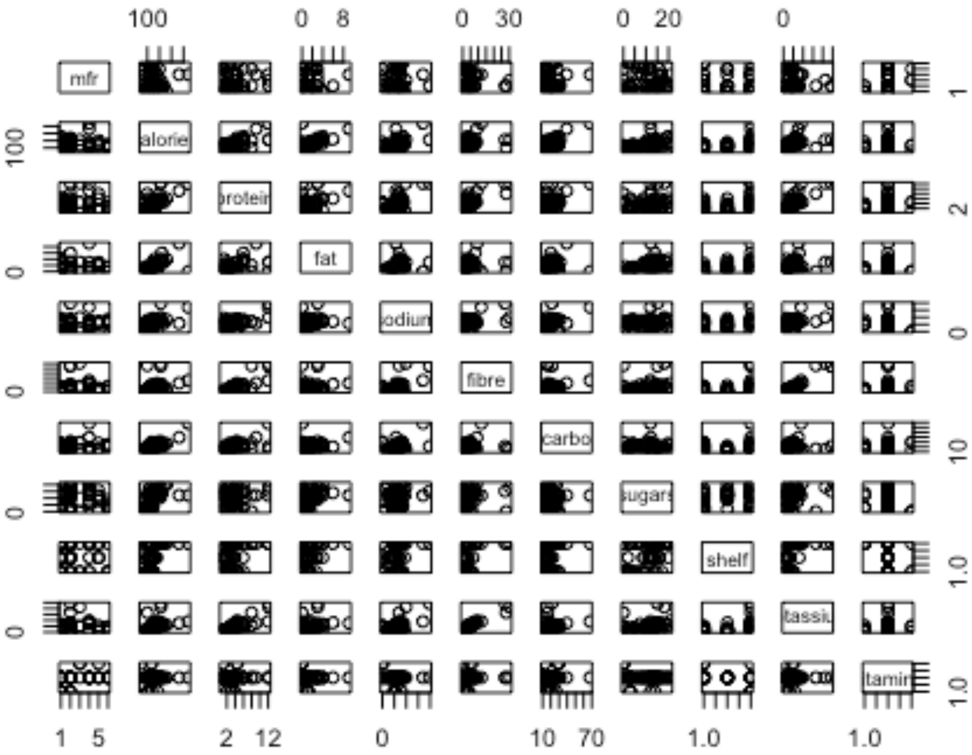
- `mfr` - Manufacturer, represented by its first initial (G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina)
- `calories` - number of calories in one portion
- `protein` - grams of protein in one portion
- `fat` - grams of fat in one portion
- `sodium` - milligrams of sodium in one portion
- `fibre` - grams of dietary fibre in one portion
- `carbo` - grams of complex carbohydrates in one portion
- `sugars` - grams of sugars in one portion
- `shelf` - display shelf (1, 2, or 3, counting from the floor)
- `potassium` - grams of potassium
- `vitamins` - vitamins and minerals (none, enriched, or 100%)

(a) Load the MASS library using library(MASS). Use data("UScereal") to import the dataset.

```
library(MASS)
data<-UScereal
```

(b) Change shelf to a factor variable. Produce a scatterplot matrix of all the variables. Do you see any relationships which you would expect? Any you wouldn't expect?

```
data$shelf<-as.factor(data$shelf)
pairs(data)
```



I expected that the calories and fat will have a moderate of high correlation, and sugar and calories have a moderate or high correlation, which is displayed in the scatterplots. I didn't expect that there's not any variables shows significance on what shelf level.

(c) For Y = calories, X1 = carbo, and X2 = sugars, fit a multiple linear regression.

```
model_data<-lm(calories~carbo+sugars, data=data)
```

(d) Produce a summary of your linear model fit from part (c).

- What is the function $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ found by this model?
- What proportion of the variance is explained by the model?
- Find 95% confidence intervals for the coefficients in the model.

- Comment on the coefficient estimates and the confidence intervals. Are you surprised by any of the results?

```
summary(model_data)
```

```
##
## Call:
## lm(formula = calories ~ carbo + sugars, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.335 -10.629  -3.636   8.160  86.518
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.6350     7.8381  -3.398  0.00119 **
## carbo         5.9718     0.2950  20.240  < 2e-16 ***
## sugars        5.6513     0.4282  13.198  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.97 on 62 degrees of freedom
## Multiple R-squared:  0.9008, Adjusted R-squared:  0.8976
## F-statistic: 281.5 on 2 and 62 DF,  p-value: < 2.2e-16
```

```
confint(model_data, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -42.303160 -10.966823
## carbo         5.382011   6.561583
## sugars        4.795400   6.507269
```

The function is $\hat{Y} = -26.6350 + 5.9718X_1 + 5.6513X_2$. The summary shows $R^2 = 0.9008$, which means this model explains 90.08% of the variance. The 95% confidence interval is between -42.30316 and -10.966823, carbo is between 5.382011 and 6.561583, and sugars is between 4.7954 and 6.507269. I am surprised on the interval of the intercept's confidence interval since the model shows a high value of R^2. I assume that the model with high R^2 will have a low standard error of each coefficient, but it showed a large intervals of Intercept, which means that the standard error of intercept is also really high.