Problem Set 4 Report

Minh Duong

04 – 30 – 2023

1. **Medical Expenses Dataset**
   a. **Using the data for all countries estimate a regression model that explains medical expenses as a function of income, education, country, and location (rural vs. urban).**
   Using R with the lm() function, we get the following equation:

$$\begin{aligned} medicalExpense \\ &= 8.84034 + 0.33494 * income + 0.50136 * education \\ &- 12.64338 * isRural - 1.57098 * isUSA - 0.11764 * isCanada \end{aligned}$$

   In the equation, isRural, isUSA and isCanada are dummy variables.

   b. **Test the null hypothesis that country and location (urban vs rural) jointly have no effect on medical expenses. Interpret your results using complete sentences.**
   We have the following hypotheses:

$$H_0: \beta_{isRural} = 0 \; and \; \beta_{isUSA} = 0 \; and \; \beta_{isCanada} = 0$$
$$H_1: \beta_{isRural} \; or \; \beta_{isUSA} \; or \; \beta_{isCanada} \neq 0$$

   Using the linearHypothesis() function in R, we get a p-value of 2.301e-09 < 0.05. This result means that we reject the null hypothesis and accept the alternative hypothesis that country and urban jointly affect medical expenses.

   c. **Using the estimated model, construct a 95% confidence interval for expected medical expenses for a rural Mexican household with an income of $50,000 and 12 years of education. Compare this to that of an urban US household with the same level of income and education. Do both intervals overlap?**
   The result from R indicates a 95% confidence interval of $[11381.07, 22117.24]$ in medical expenses of a rural Mexican household with an income of $50,000 and 12 years of education. An urban US household with the same income and education level has a medical expense within a 95% confidence interval of $[11393.69, 22126.77]$. These two confidence intervals bear similarities and they do overlap.

2. **Scores Dataset**

a. The School District, which encompasses both schools, claims that average test scores increased for both schools between 2014 and 2016. Does the data support the School District's claim? Verify the claim one school at a time.

Using regression analysis, we get the result for both schools as follow:

i. Kennedy High School: The resulting equation is $score = 62.580 + 9.380 * is2016$, with is2016 being a dummy variable for the year 2016. The hypotheses here are $H_0: \beta_{isYear} = 0, H_1: \beta_{isYear} > 0$. As a one-tailed test using pt() gives us a p-value of 0.00598 < 0.05, so at a 5% significance level we reject the null hypothesis and conclude that the scores in Kennedy High School do increase from the year 2014 to 2016

ii. Lincoln High School: The resulting equation is $score = 51.32 + 12.30 * is2016$. The hypotheses here are $H_0: \beta_{isYear} = 0, H_1: \beta_{isYear} > 0$. As we run pt() to execute a one-tailed test, we get the p-value of 0.0094 < 0.05. That means at a 5% significance level, we reject the null hypothesis and conclude that the scores in Lincoln High School do increase from 2014 to 2016.

b. A county official claims that the performance of fourth graders does not differ between both schools. Does the data support the claim? Verify the claim for 2014 and 2016 separately.

Using regression analysis, we get the result for the 2 years as follow:

i. 2014: The resulting equation is $score = 51.32 + 11.26 * isKennedy$, with isKennedy being a dummy variable for the Kennedy High School. Running the 2-tailed test using pt() with the hypotheses of $H_0: \beta_{isKennedy} = 0, H_1: \beta_{isKennedy} \neq 0$ gives us a p-value of 0.029 < 0.05. That means at a 5% significance level, we reject the null hypothesis and conclude that there is a difference between the two schools.

ii. 2016: The resulting equation is $score = 63.62 + 8.34 * isKennedy$, with isKennedy being a dummy variable for the Kennedy High School. Running the 2-tailed test using pt() with the hypotheses of $H_0: \beta_{isKennedy} = 0, H_1: \beta_{isKennedy} \neq 0$ gives us a p-value of 0.029 < 0.05. That means at a 5% significance level, we reject the null hypothesis and conclude that there is a difference between the two schools.

It should be noted there that the two-tailed test only tests if there are differences between the scores of the two schools. To test if a school is statistically significantly higher, we need to execute a one-tailed test for each of the years.

c. Using data for both schools for both years, estimate the given model. Using estimates from this model test the hypothesis that average test scores are the same for both schools and for both years against the alternative that they are not.

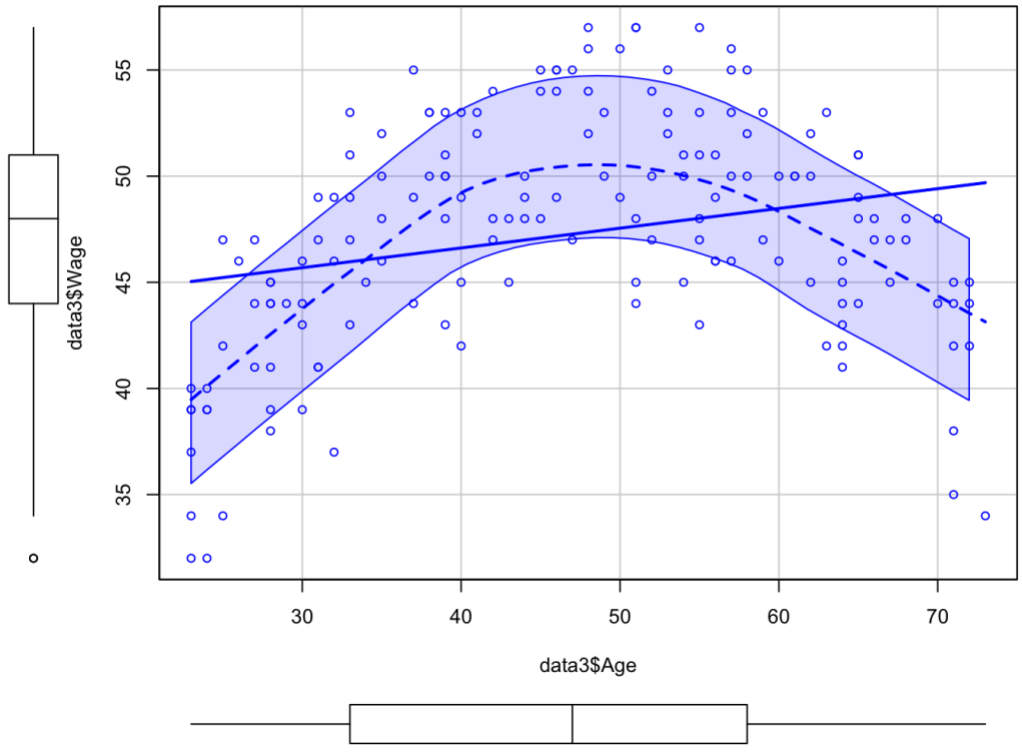
Executing the regression analysis on R gives an equation of

$$score = 52.050 + 9.800 * isKennedy + 10.840 * is2016 (+error)$$

After that, to test if the scores are similar in both scores in both years, we create hypotheses H_0: $\beta_{isKennedy} = 0\ and \beta_{isYear} = 0$, H_1: $\beta_{isKennedy}\ or\ \beta_{isYear} \neq 0$.

It is followed that the linearHypothesis() function gives a p-value of 3.762e-05 < 0.05. Thus we reject the null hypothesis and accept that there is a difference between the scores between the two schools and/or years.

3. **Wages Dataset (Chapter 8)**
   a. **Draw a scatter plot for the data with age on the x-axis and wages on the y-axis. What can you infer from this chart?**
   By running the scatterplot() function from R, we get the following plot between age (x-axis) and wage (y-axis):



It is apparent from the plot that a linear line of best fit may not produce the best result (one with the lowest residual sum of squares). In this situation, a non-linear regression may result in a better model. The plot also suggests a polynomial, most probably quadratic, relationship between age and wage, with a high point (local maxima) at around 47- 48 years old.

   b. **There are 160 observations in the dataset. Partition the sample into two sub-samples: the first 140 observations (training set) and the last 20 observations**

(validation set). Using the first 140 observations, fit two regression models. In model 1, regress wages on graduate and age. In model 2, regress wages on graduate, age, and age^2. Summarize estimated coefficients, standard errors, p-values, R-square, and Adjusted R-square from both models in a well-formatted table. Which model would you choose? Justify your choice.
The first model gives the equation:

$$wage = 40.14946 + 6.10610 * graduate + 0.07249 * age$$

The graduate variable has a p-value of 1.35e-12, and age has a p-value of 0.00628. The R-squared is 0.3348, while the Adjusted R-squared is 0.325.

The second model gives the equation:

$$wage = -1.3118416 + 6.3599795 * graduate + 2.0088529 * age - 0.0205010 * age^2$$

The graduate, age, and age^2 variables all have p-values of < 2e-16, with an R-squared of 0.8595 and an Adjusted R-squared of 0.8564.

Here is a table of the statistics:

	1st Model	2nd Model
Coefficient of graduate	6.10610	6.3599795
Coefficient of age	0.07249	2.0088529
Coefficient of age^2	-	0.0205010
p-value of graduate	1.35e-12	< 2e-16
p-value of age	0.00628	< 2e-16
p-value of age^2	-	< 2e-16
R-squared	0.3348	0.8595
Adjusted R-squared	0.325	0.8564

We can easily observe that the explanatory variables on the second model have much better p-value, and the second model has a much higher R-squared and adjusted R-squared. That means that the second model has a much better explanatory power, and should be preferred over the first model. This finding also complies with our observations from the plot that there is a quadratic relationship between wage and age.

In conclusion, the second model should be preferred over the first model.

c. Using estimated models, predict expected wages for a 30-year-old individual with a graduate degree.

Using the predict() function in R, we can apply the model to predict the wage of such an individual:

i. The first model gives a wage prediction of 48.43022 (wage unit).
ii. The second model gives a wage prediction of 46.86278 (wage unit).

d. According to estimated model 2, at what age are wages at the maximum? Do these calculations in R.

The wage will be at the maximum if at the local maxima on the equation of wage over age. We can calculate it by taking the derivative of wage over age. That gives us $age = -\text{Coefficient of age}/(2 * \text{Coefficient of age}^2)$. Using R, that gives us a result in age of $48.99391 \approx 49$ years old.

e. Cross-validation. Calculate, MSE, RMSE, MAE, and MAPE for both models over the training set and over the validation set. Comment on these cross-validation results. Based on these results, which model would you choose?

Using R, we can obtain the following table regarding the two models:

	1st Model	2nd Model
Training Set (Observations 1 – 140)		
R-square	0.3348	0.8595
Adjusted R-square	0.325	0.8564
$[corr(wages, \widehat{wages})]^2$	0.3347565	0.8595431
MSE	20.10457	4.244801
RMSE	4.483812	2.060291

MAE	3.664206	1.667441
MAPE	8.092382	3.609098
Validation Set (Observations 141 – 160):		
$[corr(wages, \widehat{wages})]^2$	0.3804186	0.8686315
MSE	2.588916	0.8076116
RMSE	1.609011	0.8986721
MAE	3.422674	1.866065
MAPE	7.665038	4.095992

From the observed information, we can see that the second model has better accuracy and predictive power.

Firstly, the second model has a better R-square, Adjusted R-square, and the square of the correlation between the predicted wage and observed wage. That means that the second model has better explanatory power.

MSE, RMSE, and MAE are measures of the error between the predicted values and the actual values. The second model has lower values of MSE, RMSE, and MAE, which indicates that the second model's predictions are closer to the actual values.

MAPE is a measure of the percentage difference between the predicted and actual values. The second model has a lower value of MAPE, meaning that the model's predictions are more accurate.

Overall, the second model has better R-square, MSE, RMSE, MAE, and MAPE during cross-validation. This suggests that the second model is more accurate and has a better fit to the data, and thus has better predictive power.

ECON 453, Spring 2023, Pset 4
Minh Duong

Addendum

Project Repository: https://github.com/MykeDuong/econ453

R Script code:

```
## ----setup, include=FALSE-------------------------------------------------
--
knitr::opts_chunk$set(echo = TRUE)

rm(list = ls()) # clear the workspace

getwd()
setwd("./")

library(readxl)
library(car) # loads car library that does tests of linear restrictions

###----------------------------------------------------------------------------
---
### Question 1
data1 = read_excel("./data/pset1_data.xlsx", sheet="medical_expenses")

## 1A
data1$D_RURAL=ifelse((data1$location =="RURAL"), 1, 0)
data1$D_USA=ifelse((data1$country =="USA"), 1, 0)
data1$D_CANADA=ifelse((data1$country =="CANADA"), 1, 0)

model1 <-lm(
  data = data1,
  medicalexpn ~ income + education + D_RURAL + D_USA + D_CANADA
)

summary(model1)

## 1B
linearHypothesis(model1, c("D_RURAL = 0","D_USA=0","D_CANADA=0"))
# 2.301e-09< 0.05 => Reject the Null hypothesis

## 1C

predict(
  model1,
  data.frame(income=50000, education=12, D_RURAL=1, D_USA=0, D_CANADA=0),
  interval="confidence", level = 0.95
)

predict(
  model1,
  data.frame(income=50000, education=12, D_RURAL=0, D_USA=1, D_CANADA=0),
  interval="confidence", level = 0.95
```

```
)

###-----------------------------------------------------------------------
---
### Question 2
data2 = read_excel("./data/pset1_data.xlsx", sheet="scores")

data2$dken=ifelse((data2$school =="Kennedy"), 1, 0)
data2$d2016=ifelse((data2$year ==2016), 1, 0)

## 2A
# Kennedy
model2_ak = lm(
  data = data2[data2$dken==1,],
  score ~ d2016
)

summary(model2_ak)
Output <- summary(model2_ak)
coef(Output)

pt(coef(Output)[2,3], 98, lower = FALSE)

# Lincoln

model2_al = lm(
  data = data2[data2$dken==0,],
  score ~ d2016
)

summary(model2_al)
Output <- summary(model2_al)
coef(Output)

pt(coef(Output)[2,3], 98, lower = FALSE)
# 0.00943 < 0.05, reject the Null even at 5% significance level

## 2B
# 2014
model2_b2014 = lm(
  data = data2[data2$d2016 == 0,],
  score ~ dken
)

linearHypothesis(model2_b2014, c("dken = 0"))

summary(model2_b2014)
Output <- summary(model2_b2014)
coef(Output)
2 * (1 - pt(coef(Output)[2, 3], 98))
# 0.02915<0.05, reject the Null even at 5% significance level
```

```
# 2016
model2_b2016 = lm(
  data = data2[data2$d2016 == 1,],
  score ~ dken
)

linearHypothesis(model2_b2016, c("dken = 0"))

summary(model2_b2016)
Output <- summary(model2_b2016)
coef(Output)
2 * (1 - pt(coef(Output)[2, 3], 98))
# 0.02855 < 0.05, reject the Null even at 5% significance level

## 2C

model2_c = lm(
  data = data2,
  score ~ dken + d2016
)
summary(model2_c)
linearHypothesis(model2_c, c("dken = 0", "d2016 = 0"))
# 3.762e-05< 0.05, reject the Null.

###-------------------------------------------------------------------------
---
### Question 3
data3<- read_excel("./data/jaggia_ba_2e_ch08_data.xlsx", sheet = "Wages")

## 3A
scatterplot(data3$Wage~data3$Age)
# The plot is not linear

## 3B
data3_T = data3[1:140,]
data3_V = data3[141:160,]

model3_1 = lm(
  Wage~ Graduate + Age ,
  data = data3_T
)
summary(model3_1);

model3_2 = lm(
  Wage  ~ Graduate + Age + I(Age^2),
  data = data3_T
)
summary(model3_2);

## 3C
```

```
predict.lm(model3_1, data.frame(Graduate = 1, Age = 30))

predict.lm(model3_2, data.frame(Graduate = 1, Age = 30))

## 3D
Output = summary(model3_2)
-coef(Output)[3,1]/(2*coef(Output)[4,1])

## 3E
evaluate_model <- function(model, data) {
  summary(model)

  data$yhat = predict.lm(model, data)
  data$res = data$Wage - data$yhat

  print("Correlation-Square:")
  print((cor(data$Wage, data$yhat)) ^ 2)

  MSE = (sum((data$Wage - data$yhat) ^ 2) / (140))
  print("MSE:")
  print(MSE)

  RMSE = MSE ^ 0.5
  print("RMSE: ")
  print(RMSE)

  MAE = mean(abs(data$Wage - data$yhat))
  print("MAE: ")
  print(MAE)
  mean(abs(data$res));

  MAPE = mean(abs(data$Wage - data$yhat) / data$Wage * 100)
  print("MAPE: ")
  print(MAPE)

  RSE = (sum((data$Wage - data$yhat) ^ 2) / (140 - 4)) ^ 0.5
  print("RSE: ")
  print(RSE)
}

evaluate_model(model3_1, data3_T)
evaluate_model(model3_2, data3_T)
evaluate_model(model3_1, data3_V)
evaluate_model(model3_2, data3_V)
```