STAT 4130: Homework (Template)

Your name

2023-05-27

```r
knitr::opts_chunk$set(echo = TRUE)
# if you are using libraries, it's good practice to load them here
```

Question 1

```r
# please do your coding inside a code chunk
# unless otherwise stated, feel free to do all computations in R
# commented code is always appreciated

#print("Hello")
```

1a. [Write out your response as text, outside of code chunks/comments] The expectation of X E[X] = 1/2. The expectation of Y E[Y] = k The expectation of XY E[XY] = 5k/4. 1) X and Y are correlated as E[XY] is not equal to E[X]*E[Y]. 2) X and Y are not independent because Y = k* X^2. 3) X and Y are not orthogonal because E[XY] is not equal to 0.

1b. [...] The expectation of X E[X] = 1/2. The expectation of Y E[Y] = k + c. The expectation of XY E[XY] = 5k/4 + c/2. 1) X and Y are correlated as E[XY] is not equal to E[X]*EY 2) X and Y are not independent because Y = k* X^2 + c. 3) X and Y are not orthogonal because E[XY] is not equal to 0.

1c. [...] The expectation of X E[X] = 0. The expectation of Y E[Y] = k/3. The expectation of XY E[XY] = 0. 1) X and Y are uncorrelated as E[XY] = E[X] * E[Y] = 0. 2) X and Y are not independent because Y = k * X^2. 3) X and Y are orthogonal because E[XY] = 0.

1d. Conclusion: If X and Y are uncorrelated, X and Y are not necessarily orthogonal. If the expectation of X is 0 or the expectation of Y is 0, then X and Y are orthogonal. If X and Y are orthogonal, X and Y are uncorrelated.

Question 2

2a. In matrix notation, we can calculate SSR as follows:

$$
\begin{aligned}
e'e &= (y - Xb)'(y - Xb) \\
&= (y' - (Xb)')\,(y - Xb) \\
&= y'y - y'Xb - (Xb)'y + (Xb)'Xb \\
&= y'y - 2b'X'y + b'X'Xb
\end{aligned}
$$

2b. We now take the derivative of the last expression:

$$
\begin{aligned}
&-2X'y + 2X'Xb = 0 \\
\Rightarrow \quad &-X'y + X'Xb = 0 \\
&\Rightarrow \quad X'Xb = X'y
\end{aligned}
$$

As a result, b is given by:

$$b = (X'X)^{-1} X'y \tag{1}$$

2c.

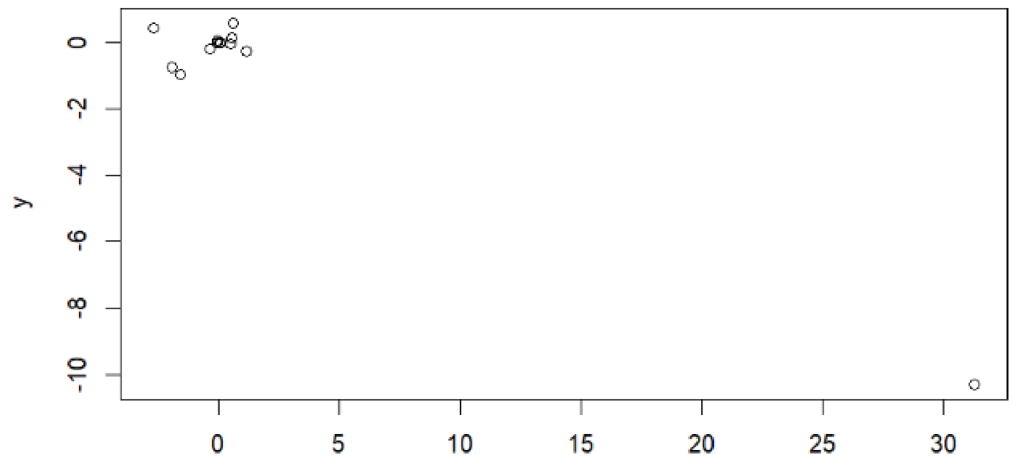
$$
\begin{aligned}
\hat{\beta} &= (X'X)^{-1} X'(X\beta + u) \\
&= (X'X)^{-1} (X'X) \beta + (X'X)^{-1} X'u \\
&= \beta + (X'X)^{-1} X'u
\end{aligned}
$$

It is an unbiased estimator, as:

$$E(\hat{\beta}) = E\left[\beta + (X'X)^{-1} X'u\right] = \beta + E\left[(X'X)^{-1} X'u\right] \tag{2}$$

Question 3

```
# Your Code

#3a.
library(MASS)
data1 = Boston
lm1 = lm(crim ~ zn, data = data1)
lm2 = lm(crim ~ indus, data = data1)
lm3 = lm(crim ~ chas, data = data1)
lm4 = lm(crim ~ nox, data = data1)
lm5 = lm(crim ~ rm, data = data1)
lm6 = lm(crim ~ age, data = data1)
lm7 = lm(crim ~ dis, data = data1)
lm8 = lm(crim ~ rad, data = data1)
lm9 = lm(crim ~ tax, data = data1)
lm10 = lm(crim ~ ptratio, data = data1)
lm11 = lm(crim ~ black, data = data1)
lm12 = lm(crim ~ lstat, data = data1)
lm13 = lm(crim ~ medv, data = data1)

#3b.
lm14 = lm(crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat + medv,


#3c.
x = c(-0.07393, 0.5098, -1.893, 31.25, -2.684, 0.1078, -1.551, 0.6179, 0.02974, 1.152, -0.03628, 0.5488

y = c(0.044855, -0.063855, -0.749134, -10.313535, 0.430131, 0.001452, -0.987176, 0.588209, -0.003780, -

plot(x,y)
```




```r
#3d.
lm1full = lm(crim ~ zn + I(zn^2) + I(zn^3), data = data1)
lm2full = lm(crim ~ indus + I(indus^2) + I(indus^3), data = data1)
lm3full = lm(crim ~ chas + I(chas^2) + I(chas^3), data = data1)
lm4full = lm(crim ~ nox + I(nox^2) + I(nox^3), data = data1)
lm5full = lm(crim ~ rm + I(rm^2) + I(rm^3), data = data1)
lm6full = lm(crim ~ age + I(age^2) + I(age^3), data = data1)
lm7full = lm(crim ~ dis + I(dis^2) + I(dis^3), data = data1)
lm8full = lm(crim ~ rad + I(rad^2) + I(rad^3), data = data1)
lm9full = lm(crim ~ tax + I(tax^2) + I(tax^3), data = data1)
lm10full = lm(crim ~ ptratio + I(ptratio^2) + I(ptratio^3), data = data1)
lm11full = lm(crim ~ black + I(black^2) + I(black^3), data = data1)
lm12full = lm(crim ~ lstat + I(lstat^2) + I(lstat^3), data = data1)
lm13full = lm(crim ~ medv + I(medv^2) + I(medv^3), data = data1)

anova(lm1,lm1full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ zn
## Model 2: crim ~ zn + I(zn^2) + I(zn^3)
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    504 35862
## 2    502 35187  2    674.56 4.8118 0.008512 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm2,lm2full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ indus
## Model 2: crim ~ indus + I(indus^2) + I(indus^3)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    504 31187
## 2    502 27662  2    3525.1 31.987 8.409e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm3,lm3full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ chas
## Model 2: crim ~ chas + I(chas^2) + I(chas^3)
##   Res.Df   RSS Df Sum of Sq F Pr(>F)
## 1    504 37247
## 2    504 37247  0         0
```

```r
anova(lm4,lm4full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ nox
## Model 2: crim ~ nox + I(nox^2) + I(nox^3)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    504 30742
## 2    502 26267  2    4474.6 42.758 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm5,lm5full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ rm
## Model 2: crim ~ rm + I(rm^2) + I(rm^3)
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    504 35567
## 2    502 34831  2    736.69 5.3088 0.005229 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm6,lm6full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ age
## Model 2: crim ~ age + I(age^2) + I(age^3)
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1    504 32714
## 2    502 30853  2      1861 15.14 4.125e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm7,lm7full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ dis
## Model 2: crim ~ dis + I(dis^2) + I(dis^3)
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1    504 31977
## 2    502 26983  2    4994.5 46.46 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm8,lm8full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ rad
## Model 2: crim ~ rad + I(rad^2) + I(rad^3)
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    504 22745
## 2    502 22417  2    328.06 3.6733 0.02608 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm9,lm9full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ tax
## Model 2: crim ~ tax + I(tax^2) + I(tax^3)
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1    504 24674
## 2    502 23581  2    1093.5 11.64 1.144e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm10,lm10full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ ptratio
## Model 2: crim ~ ptratio + I(ptratio^2) + I(ptratio^3)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    504 34222
## 2    502 33112  2    1110.2 8.4155 0.0002542 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm11,lm11full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ black
## Model 2: crim ~ black + I(black^2) + I(black^3)
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    504 31823
## 2    502 31765  2    58.495 0.4622 0.6302
```

```r
anova(lm12,lm12full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ lstat
## Model 2: crim ~ lstat + I(lstat^2) + I(lstat^3)
##   Res.Df   RSS Df Sum of Sq     F  Pr(>F)
## 1    504 29607
## 2    502 29221  2    386.39 3.319 0.03698 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm13,lm13full)
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ medv
## Model 2: crim ~ medv + I(medv^2) + I(medv^3)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    504 31730
## 2    502 21663  2     10066 116.63 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3b. When testing the null hypothesis, we should apply the two-tailed t-test. Firstly we should calculate the t-statistic by dividing β_j by the statistical error of β_j. Then we can reject H_0 if t is smaller than $-t_\alpha/2$ or greater than $t_\alpha/2$.

3c. Results from (1) are relatively inaccurate and one-sided, while results from (2) are more comprehensive and persuasive. For example, the column of "chas" is constant 0 in this example, which make lm3 meaningless.

3d. anova(lm1, lm1full): P-value = 0.008512 < 0.05 It fits more in the cubic model for "zn". anova(lm2, lm2full): P-value = 8.409e-1 < 0.05 It fits more in the cubic model for "indus" anova(lm3, lm3full): P-value = 0 Because all of the data for "chas" is 0 anova(lm4, lm4full): P-value < 2.2e-16 It fits more in the cubic model for "nox" anova(lm5, lm5full): P-value = 0.005229 < 0.05 It fits more in the cubic model for "rm" anova(lm6, lm6full): P-value = 4.125e-07 < 0.05 It fits more in the cubic model for "age" anova(lm7, lm7full): P-value < 2.2e-16 It fits more in the cubic model for "dis" anova(lm8, lm8full): P-value = 0.02608 < 0.05 It fits more in the cubic model for "rad" anova(lm9, lm9full): P-value = 1.144e-05 < 0.05 It fits more in the cubic model for "tax" anova(lm10, lm10full): P-value = 0.0002542 < 0.05 It fits more in the cubic model for "ptratio" anova(lm11, lm11full): P-value = 0.6302 > 0.05 It fits more in the linear model for "black" anova(lm12, lm12full): P-value = 0.03698 < 0.05 It fits more in the cubic model for "lstat" anova(lm13, lm13full): P-value < 2.2e-16 It fits more in the cubic model for "medv"