Lab 1: Simple Regression

Saiwen Yu 2020210610**
State Key Laboratory of Cognitive Neuroscience and Learning

11/11/2020

Context

Locate and download the data set. SAVE YOUR DATA FILE. Run a simple linear regression with "FUND" as the independent variable and "NUM_ARTICLE" as the dependent variable. ,

The output navigator will appear displaying the results. Use these results to answer the following questions.

We fit two models:

Model C:

$$Y_i = \beta_0 + \varepsilon_i \,.$$

Model A:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \,.$$

Prior work should be done to ensure the latter analysis finished as expected.

```
fa <- read.csv('/Users/syen/Documents/data/reg.csv')

fund <- fa$FUND
nar <- fa$NUM_ARTICLE
```

`read.csv` is a function to upload the data from a csv file, which were assigned to the variable `fa`. Then variables `fund` and `nar` were further assigned to the value of fa's FUND and NUMBER_ARTICLE respectively.

From the printout, find:

$$\begin{aligned} SSE(C) &= 331419083.935 \\ PC &= 1 \\ SSE(A) &= 129271881.948 \\ PA &= 2 \end{aligned}$$

SSE is the sum of square error, a estimation of error in the model. P is the number of the parameters, with the compact and augmented model gaining one and two respectively, namely β_0 and β_0, β_1.

For Model C:

The R script is below:

```
conmod <- lm(formula=nar~fund-fund, data=fa)
summary(conmod)
```

$$b_0 = 4530.258$$
$$\hat{Y}_c = 4530.258$$

b_0 is identical to $\hat{Y}_c$, which is the average of the number of the articles.

For Model A:

The R script is below:

```
lnrmd <- lm(formula=nar~fund, data=fa)
summary(lnrmd)
```

$$b_0 = 2386.430149$$
$$b_1 = 0.038536$$
$$\hat{Y}_A = 0.039X + 2386.430$$
$$SSR = 202147202$$
$$\text{Multiple Corr} = 0.781$$
$$PRE = 0.6099$$
$$adjusted\ PRE \text{ or } \eta^2 = 0.5965$$

Problem 1

Question: Test $H_o : \eta^2 = 0$ (Model C fits as well as Model A)

The test is transferred to whether the difference of the error, PRE, is significant. Then the F distribution is constructed:

$$\begin{aligned} F &= \frac{PRE/(PA - PC)}{(1 - PRE)/(n - PA)} \\ &= \frac{0.6099/(2-1)}{(1-0.6099)/(31-2)} \\ &= 45.34 \end{aligned}$$

Then the calculate the p value:

```
cumul_p <- 0.6099/(2-1)/(1-0.6099)*(31-2)
p_value <- 1 - pf(cumul_p,1,29)
print(p_value)
```

Error(A)<ERROR(C)($p < 0.01$)

Reject H_0

Problem 2

Question: For Model A, test: $H_o : \beta_1 = 0$

Problem 2 is identical to problem 1, because Model A is identical to Model C with β_1 being 0. To test null hypothesis whether β_1 is 0 is in line with the test to determine if η^2 is 0, as η^2 is to determine the difference of two models. But an alternative t test is accessible, by testing the correlation of two variables. If Model A succeeds to account for larger enough information compared with Model C, fund will exert significant influence on the number of articles, resulting in a significant r. The method for testing the significance of r is t test by constructing:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Correlation between the fund the number of articles is first calculated by the following R scrip:

```
cor(data.frame(fa$FUND,fa$NUM_ARTICLE))
```

Then $r = 0.7809893$, further $t = 6.347$. The significance of r, therefore is tested:

```
r_f_na <- cor(fa$FUND, fa$NUM_ARTICLE)
t <- r_f_na/sqrt((1-r_f_na^2)/(31-2))
pt_value <- (1 - pt(t,29))*2
print(pt_value)
```

Then $t = 6.374$ ($p < .001$). Thus H_0 is rejected, indicating $\beta_1 \neq 0$.

Problem 3

Question: For Model A, find a 95% confidence interval for β_1.

```
confint(lnrmd)
```

Then the interval of 95% confidence is $[0.02683187, 0.05023926]$

Problem 4

Question: For Model A, find $\hat{\sigma}_e = 2111.315$

$$\begin{aligned}\hat{\sigma}_e &= \sqrt{MSE} \\ &= \sqrt{\frac{SSE(A)}{n-2}} \\ &= \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}\end{aligned}$$

Problem 5

Question: Does FUND help in explaining the variability in NUMBER OF ARTICLES across the 31 province?

Yes. FUND accounts for more 61% of the variability of NUMBER OF ARTICLES compared with the absence of it. η^2 is calculated as 0.596, thought of as a big effect size (Cohen, 1977).

Problem 6

Question: Display the scatterplot of FUND and NUM_ARTICLE

I used ggplot2, a R packaging for generating gorgeous statistic images.The R code is:

```
ggplot(fa, aes(x=FUND, y=NUM_ARTICLE))+
  geom_point()+
  geom_smooth(method="rlm", formula = y~x)+
  theme_light()+
  labs(x="funds", y="numbers of articles")
```

`fa` is the vector of data. `rlm` is a method of fitting the linear regression with a package, MASS, loaded necessarily.

The output image is shown in Figure 1.

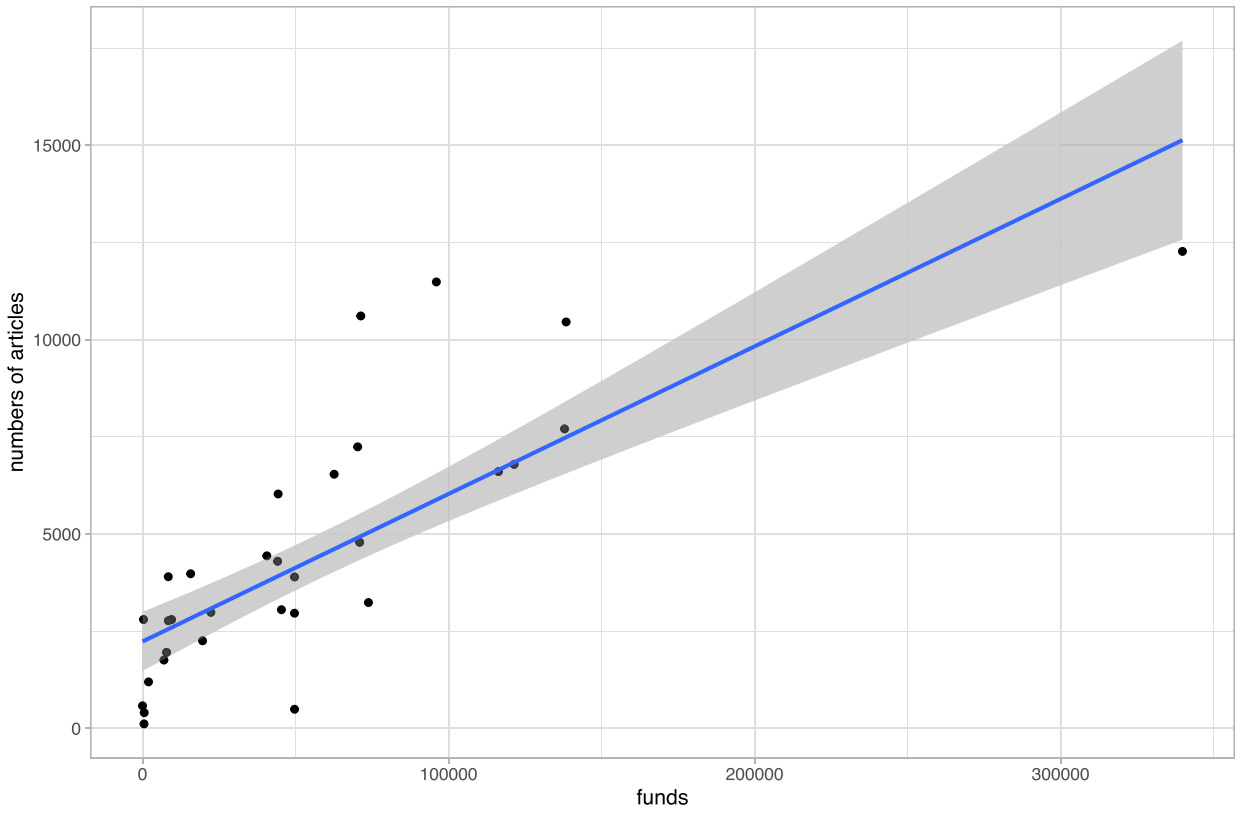

Figure 1: The scatter plot shows the correlation of funds and article numbers.

Source Code

Source code is supplied in Github. Download it by accessing: https://pan.bnu.edu.cn/l/noXqdn

Code Results

```
library(ggplot2)
library(MASS) #for rml method in geom_smooth()
```

```r
fa <- read.csv('/Users/syen/Documents/data/reg.csv')

fund <- fa$FUND
nar <- fa$NUM_ARTICLE

cumul_p <- 0.6099/(2-1)/(1-0.6099)*(31-2)
pf_value <- 1 - pf(cumul_p,1,29)
print(pf_value)
```

```
## [1] 2.176976e-07
```

```r
r_f_na <- cor(fa$FUND, fa$NUM_ARTICLE)
t <- r_f_na/sqrt((1-r_f_na^2)/(31-2))
pt_value <- (1 - pt(t,29))*2
print(pt_value)
```

```
## [1] 2.173318e-07
```

```r
conmod <- lm(formula=nar~fund-fund, data=fa)
lnrmd <- lm(formula=nar~fund, data=fa)

summary(lnrmd)
```

```
##
## Call:
## lm(formula = nar ~ fund, data = fa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3807.0 -1172.3  -260.8   737.1  5474.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.386e+03  4.951e+02   4.820 4.18e-05 ***
## fund        3.854e-02  5.722e-03   6.734 2.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2111 on 29 degrees of freedom
## Multiple R-squared:  0.6099, Adjusted R-squared:  0.5965
## F-statistic: 45.35 on 1 and 29 DF,  p-value: 2.173e-07
```

```r
summary(conmod)
```

```
##
## Call:
## lm(formula = nar ~ fund - fund, data = fa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4413.3 -2017.3  -633.3  2042.7  7739.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4530        597   7.589 1.84e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3324 on 30 degrees of freedom
```

```
#Scatter plot
options(scipen=200)
ggplot(fa, aes(x=FUND, y=NUM_ARTICLE))+
  geom_point()+
  geom_smooth(method="rlm", formula = y~x)+
  theme_light()+
  labs(x="funds", y="numbers of articles")

ggsave("scatter.pdf")
```

Reference

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev ed). *Englewood Cliffs, Mahwah. NJ.*

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, anova, and beyond, third edition.* Taylor & Francis.

Kabacoff, R. (2015). *R in action: Data analysis and graphics with r.* Manning.

Lab 3: Multiple Regression with Continuous Variables

Saiwen Yu 2020210610**
State Key Laboratory of Cognitive Neuroscience and Learning

11/25/2020

Instructions about the Variables

IVs:

X_1	Obigat	having obligation for friend's needs
X_2	Effic	having power to assist friends
X_3	Empathy	empathy for friend's obstacles
X_4	Sympath	sympathy personality

DV:

Y	Help	Effect of Helping Behavior

Part 1 Calculate the Redundancy

Description

We will begin by taking a look at some of the redundancy between the predictor variables. The approach for measuring redundancy among the predictor variables involves determining to what degree you can predict the scores on one predictor variables by using the scores on the other predictor variables.

First load the data:

```
lab3 <- read.csv("/Users/syen/Documents/data/lab3/lab3.csv")

obigat <- lab3$x1
effict <- lab3$x2
empathy <- lab3$x3
sympathy <- lab3$x4

help <- lab3$y
```

Problem 1

Regress the variable "empathy" on the other predictor variables. Use X3 as your dependent variable, and then use X_1, X_2, and X_4 as the predictor variables. Write down the value of the PRE for the overall model of regressing X_3 on X_1, X_2, and X_4. What information can we get from this PRE?

The model is:

$$X_{i4} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

Multiple regression was done by `lm` function:

```
lnrmd <- lm(formula=empathy~obigat+effict+sympathy, data=lab3)
summary(lnrmd)
```

```
##
## Call:
## lm(formula = empathy ~ obigat + effict + sympathy, data = lab3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00951 -0.59532  0.06847  0.68798  1.95228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.59353    0.22787  15.770  < 2e-16 ***
## obigat       0.05225    0.02177   2.400   0.0167 *
## effict       0.07966    0.04188   1.902   0.0577 .
## sympathy     0.14901    0.03156   4.721 3.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9017 on 533 degrees of freedom
## Multiple R-squared:  0.08262,    Adjusted R-squared:  0.07746
## F-statistic:    16 on 3 and 533 DF,  p-value: 5.655e-10
```

From the result can find the PRE, equivalent to R^2 in this case, is 0.08262. Then the inference work:

$$
\begin{aligned}
F &= \frac{PRE/(PA-PC)}{(1-PRE)/(n-PA)} \\
&= \frac{R^2/(p-1)}{(1-R^2)/(n-p)} \\
&= \frac{0.08262/(4-1)}{(1-0.08262)/(537-4)} \\
&= 16.00081
\end{aligned}
$$

Then calculate the p value:

```
p_value <- 1 - pf(16.00081, 3, 533)
print(p_value)
```

```
## [1] 5.656964e-10
```

Then $F = 16.00081$ ($p < .001$). PRE is significant enough in the general population, indicating that X_1, X_2, X_3, X_4 share in their ability to reduce error in the predictions of other index, say Y(Effect of Helping Behavior), with non-negligible redundancy among the predictors.

Problem 2

Compute "tolerance" according to the PRE and write it down.

$$
\begin{aligned}
tolerance &= 1 - PRE \\
&= 0.917
\end{aligned}
$$

Problem 3

Have R compute the tolerance of all of the predictor variables, and write them in a list. (Note that except for rounding differences the tolerance you computed for X3 should be the same as that provided by R).

CRAN offers a package, `olsrr`, whose function `ols_vif_tol` is for calculating VIF and tolerance. Then:

```
lnrmd_1 <- lm(formula=help~obigat+effict+empathy+sympathy, data=lab3)
library("olsrr")
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
vif_tol <- ols_vif_tol(lnrmd_1)
knitr::kable(vif_tol) #kable from knitr package is to display the R output as a figure
```

Variables	Tolerance	VIF
obigat	0.9034160	1.106910
effict	0.9253023	1.080728
empathy	0.9173789	1.090062
sympathy	0.8605729	1.162017

Part 2 Multiple Linear Regression

Problem 1

Write down Model C and Model A for adding variable X_1 to a model that already has the other three predictor variables in it.

For Model C:

$$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

For Model A:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

Problem 2

Write down H_0 and H_1 in terms of β.

$$H_0 : \beta_{1.234} = 0, \quad H_1 : \beta_{1.234} \neq 0$$

Problem 3

Write the values of some important statistics for adding X_1 to the model.

Define the two models:

```
lnrmd_1 <- lm(formula=help~obigat+effict+empathy+sympathy, data=lab3)
lnrmd_2 <- lm(formula=help~effict+empathy+sympathy, data=lab3)
```

Use functon `anova` to compare two models:

```
anova(lnrmd_1, lnrmd_2)
```

```
## Analysis of Variance Table
##
## Model 1: help ~ obigat + effict + empathy + sympathy
## Model 2: help ~ effict + empathy + sympathy
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    532 180.44
## 2    533 183.02 -1   -2.5861 7.6249 0.005956 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then:

	SS	df
SSR	2.5862	1
SSA	180.4359	532
SSC	183.0221	533

$$\begin{aligned} PRE &= 2.586/183.022 = 0.01412945 \\ \text{Partial correlation} &= \sqrt{r^2_{Y1.234}} \\ &= \sqrt{PRE} \\ &= 0.1188674 \end{aligned}$$

Problem 4

what's the partial correlation between X_1 and Y? Use partial correlation to compute the PRE and write it down.

Partial correlation coefficient is the correlation between X_1 and Y while holding X_2, X_3, X_4 constant.

$$\begin{aligned} PRE &= r^2_{Yp\cdot 123\cdots p-1} \\ &= 0.1188674^2 \\ &= 0.01412945 \end{aligned}$$

Problem 5

Write the value of t and p that go with that PRE.

$$t = \sqrt{F} = 2.761322$$

$p = 0.005956$, in line with the F test previously.

Problem 6

Would it be worthwhile to add X_1 to the model that already had the other predictors in it?

It is worthwhile to add X_1 to the model, for X_1 helps to reduce significant errors with an extra parameter($F = 7.6249,\ p < .01$).

Problem 7

Draw the P-P picture of Model A and check whether the normality of residual error is appropriate or not.

To draw pp plot shall we first calculate the residuals(e_i) in the Model A with the function `residuals()`. Then we shall construct frame named after `norm` and assign the values of `residuals()` to it and then use ggplot2 to draw the pp plot. Note that I use a package `qqplotr` to extend some `ggplot2` functionalities by permitting the drawing of both quantile-quantile (Q-Q) and probability-probability (P-P) points, lines, and confidence bands. Other features of the package are omitted.

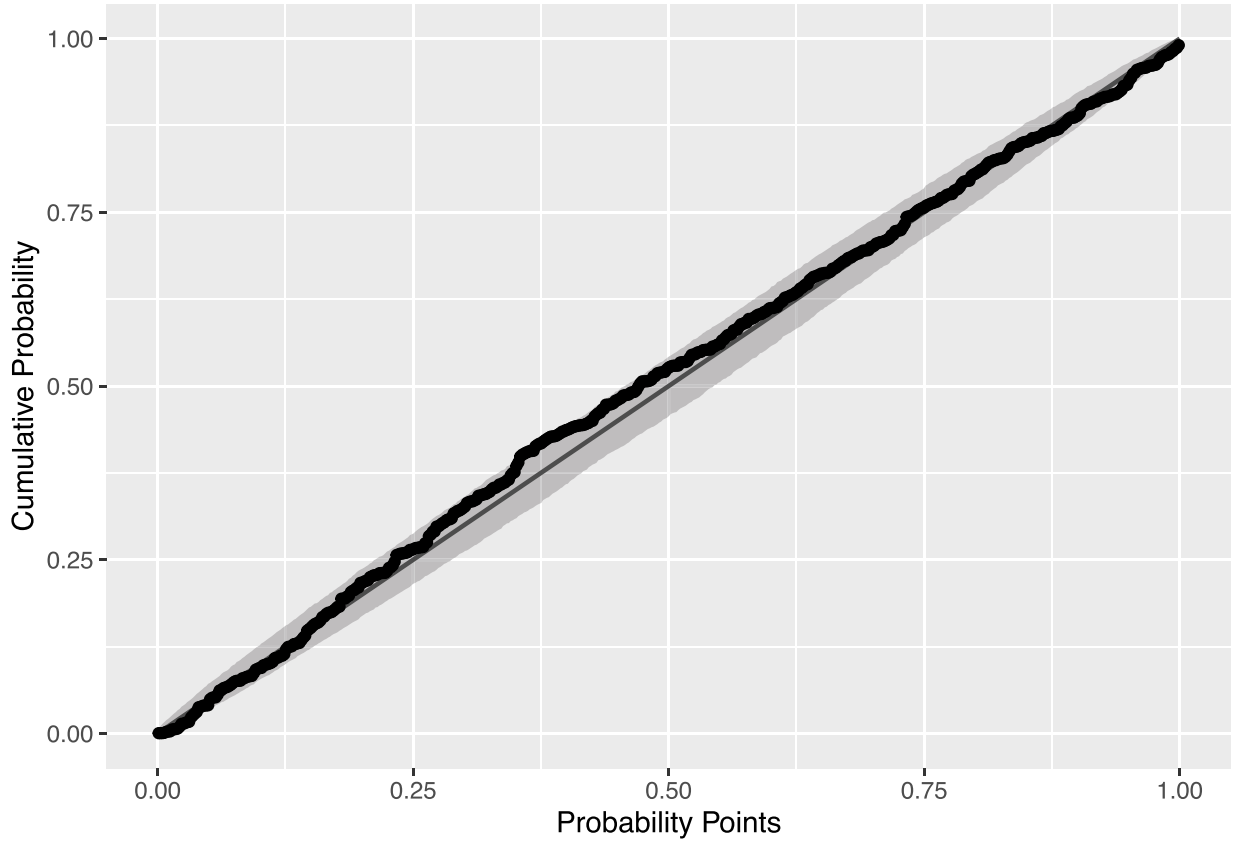

Figure 1: pp plot to test whether the residuals are normally distrubuted.

As we can see from figure 1, the points don't distribute well as a line, with modest points falling out of the confidence interval.

For the accurate conclusion shall we conduct the test of goodness of fit.

There are several methods for normality test such as Kolmogorov-Smirnov (K-S) normality test and Shapiro-Wilk's test. Shapiro-Wilk's method is widely recommended for normality test and it provides better power than K-S. It is based on the correlation between the data and the corresponding normal scores.

```
shapiro.test(res)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.98821, p-value = 0.0002526
```

Note that the null hypothesis of normality test is that sample distribution is normal. p value is found to be lower than 0.001, indicating the residuals is not normal.

Problem 8

Explain the relationship between PRE and partial correlation

In the cases where we compare two models in which the augmented model added an extra variable together with its parameter:

$$PRE = r^2_{Yp \cdot 123 \cdots p-1}$$

where PRE is proportional Reduction in Error and $r^2_{Yp \cdot 123 \cdots p-1}$, namely coefficient of partial determination, is square of partial correlation coefficient.

Problem 9

What does the part correlation imply?

It is the simple correlation between `help` and `obigat` when controlling or holding constant the other 3 predictors(`effict`, `empathy`, `sympathy`).

Source Code

Source code is supplied in cloud drive. Download it by accessing: https://pan.bnu.edu.cn/l/lu8ARR

Reference

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, anova, and beyond, third edition.* Taylor & Francis.

Kabacoff, R. (2015). *R in action: Data analysis and graphics with r.* Manning.

Lab 4 Oulier

Saiwen Yu 2020210610**
State Key Laboratory of Cognitive Neuroscience and Learning

12/1/2020

Problem 1

Open dataset LAB4_RESPONSETIME. Usging bivariate scatter plot to identify the most possible outlier. Show your figure below.

R script is blew with `ggplot2` to draw the scatter plot:

```
library(ggplot2)
library(MASS) #for rml method in geom_smooth()

lab4_1 <- read.csv("/Users/syen/Documents/data/lab4/LAB4_RESPONSE_TIME.csv")
lab4_2 <- read.csv("/Users/syen/Documents/data/lab4/LAB4_SAT.csv")

mi_len <- lab4_1$mission
res_t <- lab4_1$response
```

```
ggplot(lab4_1, aes(x=mi_len, y=res_t))+
  geom_point()+
  geom_smooth(method="rlm", formula = y~x)+
  theme_light()+
  labs(x="mission length", y="response time")
```

Clearly can we find, from figure1, a dot has fallen way from other dots as marked green, whose `id` is 5, with `mission length` and `response time` being 3.4 and 12.6 respectively.

Problem 2

Open dataset "LAB5_SAT". Regress "average total score" on other variables and use h_{ii} to find out the extreme value in three independent variable.

SPSS and R would generate different leverage, I would spell it on in the section `Test SPSS and R`.

Case Name = Missouri

R's $H_{ii} = 0.876$
SPSS's $H_{ii} = 0.855$

Regress `average total score` on `average student/teacher ratio`, in this case (the case name you write

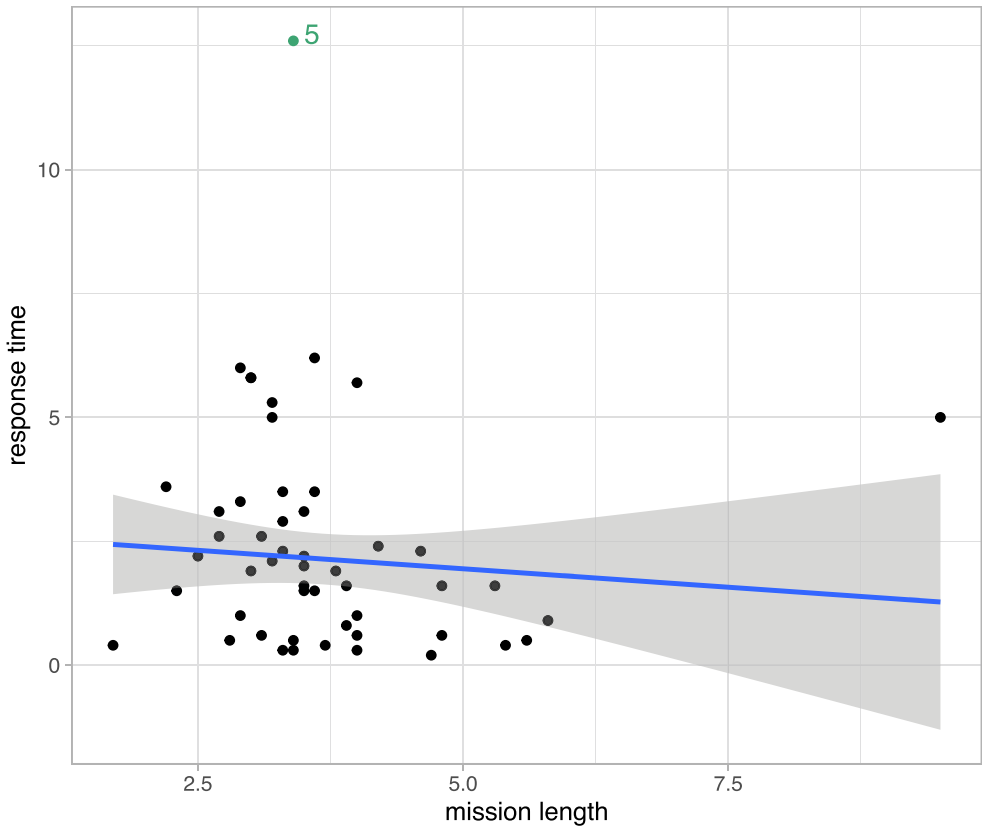

Figure 1: Scatter plot of mission length and response time, showcasing a dot falling far from the vast majority of the observations which was marked green

above)

$$\text{R's } H_{ii} = 0.874$$
$$\text{SPSS's } H_{ii} = 0.854$$

`hatvalues()` is a function to calculate the leverage. So R script is:

```
state <- lab4_2$name
per_elig_stu <- lab4_2$x1
avg_stu_tea_rat <- lab4_2$x2
an_sal_tea <- lab4_2$x3
avg_to_sco <- lab4_2$y

lm_1 <- lm(formula = avg_to_sco~per_elig_stu+avg_stu_tea_rat+an_sal_tea, data=lab4_2)
lev <- hatvalues(lm_1)
```

An outlier defined here is the data surpassing 2 or 3 times of the average hat value (p/n). Then

```
avg_lev <- 4/49 #4 parameters and 49 observations
ii <- 0
for (i in lev) {
  ii <- ii+1
  if(i >= 2*avg_lev) # high leverage is defined as two or three times of the average leverage
  cat(ii,"th observation ", lab4_2[ii,1], " is high leverage value (", i,")", sep="")
```

```
}
```

```
## 22th observation Missouri is high leverage value (0.8755981)
```

Problem 3

Open dataset "LAB4_SAT". Regress "average total score" on "annual salary of teachers" and use studentized deleted residual method to check over the Y_{29}(South Carolina).

Let Y_i stand for the ith observation of `average total score`. X_{i1} stand for the ith observation of `annual salary of teachers`. X_{i2} is an additional variable for the probable outlier, 29th observation, with the 29th value of it being 1 and the reset of it being 0.($1 \leq i \leq 49$)

Model C:

$$Y_i = 979.653083 - 0.822446X_{i1} + \varepsilon_i$$

Model A:

$$Y_i = 1078.468 - 3.090X_{i1} + 940.910X_{i2} + \varepsilon_i$$

where

$$X_{i2} = \begin{cases} 1 & i = 29 \\ 0 & i \neq 29 \end{cases}$$

$$PRE = 0.784$$
$$F = 166.680$$

Calculating process and R script are blew:

```
#generate a vector with the 29th being 1 and the rest being 0
x4 <- c(0)
for (i in 1:48) {
  if(i==28){
    x4 <- c(x4,1)
  }else{
  x4 <- c(x4,0)
  }#end of if
}

#generate a new data frame
lab4_3 <- data.frame(an_sal_tea, x4, avg_to_sco)
avg_to_sco_d <- lab4_3$avg_to_sco
an_sal_tea_d <- lab4_3$an_sal_tea
x4 <- lab4_3$x4

#regress & calculate the coefficients & compare the models
lm_3 <- lm(formula = avg_to_sco_d~an_sal_tea_d, data = lab4_3)
lm_4 <- lm(formula = avg_to_sco_d~an_sal_tea_d+x4, data = lab4_3)

lm_3$coefficients
lm_4$coefficients

SS <- anova(lm_3, lm_4)
print(SS)
```

```r
SS <- SS$RSS
PRE <- 1-min(SS)/max(SS)
print(PRE)
```

Problem 4

Cook's D for case 29(South Carolina)(considering the three independent variables)

$$D_{29} = 1.123$$

R script is blew:

```r
#calculate the 29th's Cook's d
d <- cooks.distance(lm_1)
round(d[29],3)
```

Problem 5

Delete the two outliers (change them into missing variables). Redo the multiple regression and write down your regression equation and the PRE value.

`Missouri`(22th observation) and `South Carolina`(29th observation), from the analysis above, are two outliers. Then the regression formula is:

$$Y_i = 1105.364 - 2.892X_{i1} - 6.943X_{i2} + 2.165X_{i3} - 845.303X_{i4} + 205.285X_{i5} + \varepsilon_i$$

where

$$X_{i4} = \begin{cases} 1 & i = 29 \\ 0 & i \neq 29 \end{cases} \qquad X_{i5} = \begin{cases} 1 & i = 22 \\ 0 & i \neq 22 \end{cases}$$

PRE is not made clear in this problem, which could be the PRE comparing the following models:

Model A:

$$Y_i = 1105.364 - 2.892X_{i1} - 6.943X_{i2} + 2.165X_{i3} - 845.303X_{i4} + 205.285X_{i5} + \varepsilon_i$$

where

$$X_{i4} = \begin{cases} 1 & i = 29 \\ 0 & i \neq 29 \end{cases} \qquad X_{i5} = \begin{cases} 1 & i = 22 \\ 0 & i \neq 22 \end{cases}$$

Model C:

$$Y_i = 902.563 - 4.425X_{i1} - 6.846X_{i2} + 8.989X_{i3} + \varepsilon_i$$

or:

Model A:

$$Y_i = 1105.364 - 2.892X_{i1} - 6.943X_{i2} + 2.165X_{i3} + \varepsilon_i$$

Model C:

$$Y_i = 967.404 + \varepsilon_i$$

For the first circumstance:

$$PRE = 0.939$$

For the second circumstance:

$$PRE = 0.827$$

Deleting two observations reduced significant error per parameter in the model ($p < .001$).

R script is below:

```
##############another outlier
x5 <- c(0)
for (i in 1:48) {
  if(i==21){
    x5 <- c(x5,1)
  }else{
    x5 <- c(x5,0)
  }#end of if
}

lab4_3 <- data.frame(per_elig_stu, avg_stu_tea_rat, an_sal_tea, x4, x5, avg_to_sco)
lm_5 <- lm(formula = avg_to_sco~per_elig_stu + avg_stu_tea_rat + an_sal_tea + x4 + x5)
lm_5$coefficients
```

Test SPSS and R

I found that SPSS and R would generate different leverage of a bulk of data. To demonstrate this, I tested the data from (Judd et al., 2017), whose chapter 13 contained their calculation of leverage. The original data and the results of the textbook, R and SPSS calculating leverage are:

sat	hsrank	lev from book	lev from R	lev from SPSS
42	90	0.08	0.0838626	0.0069395
48	87	0.08	0.0772589	0.0003359
58	85	0.08	0.0775476	0.0006246
45	79	0.10	0.1009310	0.0240079
45	90	0.08	0.0838626	0.0069395
86	48	0.76	0.7598152	0.6828922
51	83	0.08	0.0815892	0.0046661
56	99	0.15	0.1543375	0.0774144
51	81	0.09	0.0893837	0.0124606
58	94	0.11	0.1058025	0.0288795
42	86	0.08	0.0769342	0.0000111
55	99	0.15	0.1543375	0.0774144
61	99	0.15	0.1543375	0.0774144

The results indicated that results of textbook are in line with that of R and different from that of SPSS. Therefore for the Problem 2 I gave the results of both SPSS and R.

R(R version 4.0.2 (2020-06-22), run in x86_64-apple-darwin17.0) script is:

```
sat <- c(42,48,58,45,45,86,51,56,51,58,42,55,61)
hsrank <- c(90,87,85,79,90,48,83,99,81,94,86,99,99)

text_book <- lm(formula = sat~hsrank)
lev <- hatvalues(text_book)
lev
```

SPSS(version 25 64bit, run in MacOS 11.0.1) processes are:

```
Analyze-Regression-Save-Leverage values-continue-OK
```

References

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, anova, and beyond, third edition.* Taylor & Francis.

Kabacoff, R. (2015). *R in action: Data analysis and graphics with r.* Manning.

Lab 5 One-way ANOVA

Saiwen Yu 2020210610**
State Key Laboratory of Cognitive Neuroscience and Learning

12/21/2020

Problem 1

A psychological research on advertisement could be seen "lab5_1.csv", which mainly inquired the influence of the format of the advertisement's title on salse volume. The levels of the independent variable came in the following three values: news, inquiry and showoff. The dependent variable is sales volume of some product. Please code following the instructions below:

(1) Show the Helmert contrast codes table.

(2) Show any contrast codes table other than the previous one.

(3) Regress "lab5_1.csv" by Helmert contrast codes and showcase three indicators with concrete formula.

(1) The Helmert contrast codes of the `lab5_1.csv` are shown in Table 1.

	Category level		
code	news	inquiry	showoff
λ_{1k}	2	-1	-1
λ_{2k}	0	1	-1

Table 1: Helmert contrast codes of "lab5_1.csv"

(2) Following are the Dummy codes:

	Category level		
code	news	inquiry	showoff
λ_{1k}	0	1	0
λ_{2k}	0	0	1

Table 2: Dummy contrast codes of "lab5_1.csv"

(3) First we should organized the data into the specific format with Helmert contrast codes added to it.

```
lab5_1 <- read.csv("/Users/syen/Documents/data/lab5/lab5_1.csv")
news <- lab5_1$news
inquiry <- lab5_1$inquiry
showoff <- lab5_1$showoff

len <- nrow(lab5_1)

i1 <- c()
```

```r
for(i in 1:len){
  i1 <- c(i1, 2)
}#end of for
for(i in 1:len){
  i1 <- c(i1, -1)
}#end of for
for(i in 1:len){
  i1 <- c(i1, -1)
}#end of for

i2 <- c()
for(i in 1:len){
  i2 <- c(i2, 0)
}#end of for
for(i in 1:len){
  i2 <- c(i2, 1)
}#end of for
for(i in 1:len){
  i2 <- c(i2, -1)
}#end of for

salse <- c(news, inquiry, showoff)
```

Model A:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Then regress:

```r
lm_1 <- lm(formula = salse~i1+i2)
summary(lm_1)
```

```
##
## Call:
## lm(formula = salse ~ i1 + i2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.556  -7.889   1.278   6.694  29.444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67.0000     1.1092  60.405  < 2e-16 ***
## i1            3.4444     0.7843   4.392 2.69e-05 ***
## i2            7.0000     1.3585   5.153 1.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.53 on 105 degrees of freedom
## Multiple R-squared:  0.3039, Adjusted R-squared:  0.2906
## F-statistic: 22.92 on 2 and 105 DF,  p-value: 5.501e-09
```

We therefore arrived at the regression model:

$$Y_i = 67.000 + 3.444 X_{1i} + 7.000 X_{2i} + \varepsilon_i$$

The predictors via three levels are:

$$
\begin{aligned}
&\text{news:} && \hat{Y}_{\text{news}} = 67.000 + 3.444(2) + 7.000(0) && = 73.888 \\
&\text{inquiry:} && \hat{Y}_{\text{inquiry}} = 67.000 + 3.444(-1) + 7.000(1) && = 70.556 \\
&\text{showoff:} && \hat{Y}_{\text{showoff}} = 67.000 + 3.444(-1) + 7.000(-1) && = 56.556
\end{aligned}
$$

Problem 2

A researcher, in a hope for the knowledge of the effect of interference on anxiety, counted this emotion by anxious scale before, amid and after the therapy respectively for 30 participants. The results showcased that the average of pre-interference, interference and post-interference scored 15, 9 and 5 respectively. What will the partial regression coefficients result by orthogonal polynomial contrast codes?

Orthogonal polynomial contrast codes are:

	Category level		
code	before	amid	after
λ_{1k}	2	-1	-1
λ_{2k}	0	-1	1

Calculate the partial regression coefficients :

$$b_{X_1} = \frac{\sum\limits_k \lambda_k \bar{Y}_k}{\sum\limits_k \lambda_k^2} = \frac{\hat{Y}_{\text{before}} - (\frac{\bar{Y}_{\text{amid}} + \hat{Y}_{\text{after}}}{2})}{3} = \frac{(\frac{9+5}{2}) - 15}{3} = 2.667$$

$$b_{X_2} = \frac{\sum\limits_k \lambda_k \bar{Y}_k}{\sum\limits_k \lambda_k^2} = \frac{\hat{Y}_{\text{after}} - \hat{Y}_{\text{amid}}}{2} = \frac{5-9}{2} = -2$$

Problem 3

As can be seen in "lab5_3.csv", investigation was conducted for the influence of management methods on performances of final exam among students. Managements came in three ways, with DV being the improvement of the final exam. Coding methods are:

	Category level		
code	casual	military	enterprising
λ_{1k}	-1	0	1
λ_{2k}	-1	2	-1

(1) Interpret the meaning of two regression coefficients.

$$b = \frac{\sum\limits_k \lambda_k \bar{Y}_k}{\sum\limits_k \lambda_k^2}$$

$$\hat{\beta}_1 = b_1 = \frac{\hat{Y}_3 - \hat{Y}_1}{2}$$

In b_1, the numerator represents a comparison between casual management and enterprising management, and the denominator is a scaling dependent on the specific values for values used for contrast-code predictor.

$$\hat{\beta}_2 = b_2 = \frac{\hat{Y}_2 - (\frac{\hat{Y}_1 + \hat{Y}_3}{2})}{3}$$

In b_2, the numerator represents a comparison between military management and the rest two management methods, and the denominator is the same as below. Note this comparison regards casual and enterprising management as a whole, only to see if difference is exitent between $\hat{Y}_2$ and $\hat{Y}_1$, $\hat{Y}_3$ as a whole.

(2) Test if b_2 equals 0 by PRE and F test.

Model A:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Model C:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

Compare two models:

```
lab5_3 <- read.csv("/Users/syen/Documents/data/lab5/lab5_3.csv")
DV <- lab5_3$DV
lambda1 <- lab5_3$X1
lambda2 <- lab5_3$X2

lm_3_1 <- lm(formula = DV~lambda1+lambda2)
lm_3_2 <- lm(formula = DV~lambda1)
an <- anova(lm_3_1, lm_3_2)
knitr::kable(an)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
21	1171.500	NA	NA	NA	NA
22	1523.583	-1	-352.0833	6.311353	0.0202355

From the results can we see:

$$PRE = \frac{1523.6 - 1171.5}{1523.6} = 0.231$$
$$F = 6.311$$

$F = 6.311 (p < .05)$, Model C is rejected and Model A is received, indicating that b_2 does not equal 0.

Source Code

Source code is supplied in cloud drive. Download it by accessing: https://pan.bnu.edu.cn/l/b1kN0D

References

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, anova, and beyond, third edition.* Taylor & Francis.

Kabacoff, R. (2015). *R in action: Data analysis and graphics with r.* Manning.