Assignment 2

03/04/2022

Tiange Chang

```
# Assignment 2
# 03/04/2022
# Tiange Chang

?MCAS

# Simple Linear Regression
# 1
library(tidyverse)
# install.packages('Ecdat')
library(Ecdat)
View(MCAS)
```

```
# 2.a
df <- MCAS
anyNA(df)
summary(df)
# There are missing values in this dataset.
# Column 'spc', 'totsc8', 'avgsalary' has NAs.
```

```
# 2.b
# Missing data could reduce the statistical power, which refers to the
# probability that the test will reject the null hypothesis when it is false.
# It can cause bias in the estimation of parameters. It may also reduce the
# representativeness of the samples. It may complicate the analysis of the
# study. Each of these distortions may threaten the validity of the trials and
# can lead to invalid conclusions. (NCBI)
```

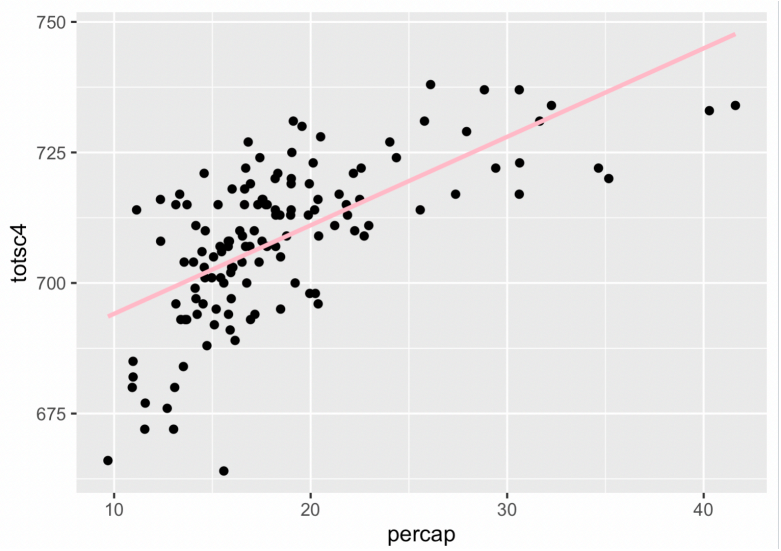
```
# 2.c
anyNA(df$spc) # TRUE
df$spc[is.na(df$spc)] = mean(df$spc, na.rm = TRUE)
anyNA(df$spc) # FALSE
```

```r
# 2.d
install.packages('simputation')
library('simputation')

anyNA(df$totsc8) # TRUE

cor(df$totsc8, df[4:17], use = 'complete.obs') # only column 4 to 17 are numeric
# percap has 0.78, totsc has 0.86

df2 <- impute_lm(df, totsc8 ~ totsc4 + percap)

anyNA(df2$totsc8) # FALSE
View(df2)
```

```r
# 2.e
anyNA(df$avgsalary) # TRUE

cor(df2$avgsalary, df2[4:17], use = 'complete.obs')
# percap has 0.62, regday has 0.52

df3 <- impute_lm(df2, avgsalary ~ regday + percap)

anyNA(df3$avgsalary) # FALSE
View(df3)

anyNA(df3) # FALSE
```

```r
# 3
set.seed(20)
train <- sample_frac(df3, 0.6) # Randomly selects 60% of rows from the dataset
valid <- setdiff(df3, train) # Sends the other rows to validation set, with no
                             # overlap
View(train)
View(valid)
```

```r
# 4
ggplot(train, aes(x = percap, y = totsc4)) + geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = 'pink')

# totsc4: 4th grade score (math + eng + science)
# percap: per capita income

# Per capita income can be used to determine the average per-person income for
# an area and to evaluate the standard of living and quality of life of the
# population. Per capita income for a nation is calculated by dividing the
# country's national income by its population. (Investopedia)

# From the plot we can tell the higher the per capita income, the higher the
# 4th grade score. It provides me an intuitive sense that people with higher
# score of math, English, and science tend to have a better quality of life.
```




```r
# 5
cor.test(train$percap, train$totsc4)
# By running the test, the correlation is Pearson's product-moment correlation
# The Pearson product-moment correlation coefficient is a measure of the
# strength and direction of association that exists between two variables
# measured on at least an interval scale. (Laerd)

# The correlation is quiet strong, with a sample estimate of 0.6535.

# The correlation is not significant, the p value is less than significance
# level (alpha = 0.05), we reject the null hypothesis.
```

```r
# 6
df4 <- lm(totsc4 ~ percap, data = train)
summary(df4)
```

```
> summary(df4)

Call:
lm(formula = totsc4 ~ percap, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-39.564  -6.944   1.131   8.040  21.443

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 677.1627     3.3375 202.897   <2e-16 ***
percap        1.6944     0.1721   9.844   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.37 on 130 degrees of freedom
Multiple R-squared:  0.4271,    Adjusted R-squared:  0.4226
F-statistic:  96.9 on 1 and 130 DF,  p-value: < 2.2e-16
```

```r
# 7
# The minimum residual value is -39.564
# The maximum residual value is 21.443
```

```r
# 7.a
train$residual <- df4$residuals
View(train)
# Berlin has the highest residual value 21.44
```

```r
# 7.b
# Boston has the lowest residual value -39.56

# 7.c
# It is unfair to say this because the observations are limited, there were
# only 220 observations in the SLR model. It cannot represent the whole
# district. The data were collected in 1997-1998, it is outdated right now.
```

```r
# 8
lm(totsc4 ~ percap, data = train)
# y = 677.163 + 1.694x
# Now, we can predict a y value for any x value (but we should stay within the
# range of x values that we used to build the model)
# set x = 15
677.163 + 1.694 * 15
# the predict result is 702.57



# 9
install.packages('forecast')
library('forecast')

regt <- df4
ta <- accuracy(predict(regt, train), train$totsc4)

regv <- lm(totsc4 ~ percap, data = valid)
va <- accuracy(predict(regv, valid), valid$totsc4)

ta
va

# RMSE is the root mean square deviation
# MAE is the mean absolute error
# By comparing the train & valid data frame, if the difference are small,
# our prediction is accurate.
11.28472-12.37612
8.986091-9.162471
# The difference between RMSE is 1.09, and MAE is 0.17. Our prediction is
# accurate.



# 10
sd(train$totsc4)
ta

# The standard deviation of 4th grade test score averages in the training set is
# 14.96351, the RMSE is 11.28472.

# RMSE is calculated on the estimated/predicted data by comparing it with the
# true values. Standard deviation captures the spread of your data around the
# mean. This is calculated on the already available data. (Quora)
# It teaches me that there will be an error compare to exist data when you use
# estimated/predicted data, compare to exist data.
```

```r
# Multiple Linear Regression
# 1.a
str(df3)
# 'code' is categorical but R currently sees as numeric.
df3$code <- as.factor(df3$code)

# 1.b
str(df3)
# The dataset now has three factor variables.

# 1.c
# From the str() function, we know the total number of records is 220
nlevels(unique(df3$code))/220
nlevels(unique(df3$municipa))/220
nlevels(unique(df3$district))/220
# The uniqueness quotient of all three factors are one.

# 1.d
# If a categorical variable has entirely unique value, it  means every factor
# in the dataset is unique. Since there is no repeated variable, it will not
# be useful for predictive purposes.
```

```r
# 2
ct <- cor(train[4:17], use = 'complete.obs')
ct
# From the correlation table, we can see that totday and regday are very highly
# correlated with each other (0.95300875).
sum(df3$regday)
sum(df3$totday)
# I decide to keep totday which is the total of spending per pupil, because it
# has the larger sum.
df5 = subset(df3, select = -c(regday))
df5
```

```r
# 3
train2 <- sample_frac(df5, 0.6)
valid2 <- setdiff(df5, train2)

df6 <- train2[4:16]
ml <- lm(df6$totsc4~., data = df6)

# 3.a
backward <- step(ml, direction = 'backward')
summary(backward)
```

```
> summary(backward)

Call:
lm(formula = df6$totsc4 ~ bilingua + lnchpct + totsc8 + avgsalary +
    pctel, data = df6)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8290  -4.3583  -0.1152   3.8449  19.5553

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.089e+02  4.077e+01  10.030  < 2e-16 ***
bilingua    -2.626e-04  1.530e-04  -1.717   0.0885 .
lnchpct     -1.781e-01  1.026e-01  -1.735   0.0852 .
totsc8       4.032e-01  6.283e-02   6.418 2.55e-09 ***
avgsalary    6.348e-01  2.835e-01   2.239   0.0269 *
pctel       -7.973e-01  3.547e-01  -2.248   0.0263 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.36 on 126 degrees of freedom
Multiple R-squared:  0.7313,    Adjusted R-squared:  0.7206
F-statistic: 68.57 on 5 and 126 DF,  p-value: < 2.2e-16
```

```r
# 4
library('car')
ml2 <- lm(df6)
vif(ml2)
```

```
> vif(ml2)
 bilingua  occupday    totday       spc    speced   lnchpct  tchratio    percap    totsc4    totsc8 avgsalary     pctel
 1.269421  1.412125  3.046727  1.195146  1.245047  3.991743  1.856394  4.021781  3.871906  6.688864  3.241631  2.046280
```

```
212  # 4.a
213  # totsc8 has value of 6.6888
214  df7 = subset(df6, select = -c(totsc8))
215  df7
216
217  ml3 <- lm(df7)
218  vif(ml3)
219  summary(ml3)
220:1  (Top Level)
```

Console | Terminal | Jobs

R 4.1.1 · ~/Desktop/

```
> summary(ml3)

Call:
lm(formula = df7)

Residuals:
    Min      1Q  Median      3Q     Max
-3769.4  -834.6  -128.2   707.5  6959.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.993e+03  1.014e+04   0.196 0.844555
bilingua     6.360e-03  2.860e-02   0.222 0.824380
occupday    -1.036e-02  6.275e-02  -0.165 0.869135
totday       1.372e+00  1.842e-01   7.451 1.57e-11 ***
spc          3.207e+00  4.737e+01   0.068 0.946125
speced      -1.468e+02  3.783e+01  -3.879 0.000172 ***
lnchpct      3.116e+01  1.807e+01   1.724 0.087204 .
tchratio     2.472e+01  7.155e+01   0.346 0.730297
percap       2.327e+01  3.345e+01   0.696 0.487893
totsc4       3.008e+00  1.423e+01   0.211 0.833021
avgsalary   -4.884e+01  6.711e+01  -0.728 0.468127
pctel       -9.945e+01  6.584e+01  -1.511 0.133529
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1311 on 120 degrees of freedom
Multiple R-squared:  0.5649,    Adjusted R-squared:  0.525
F-statistic: 14.16 on 11 and 120 DF,  p-value: < 2.2e-16
```

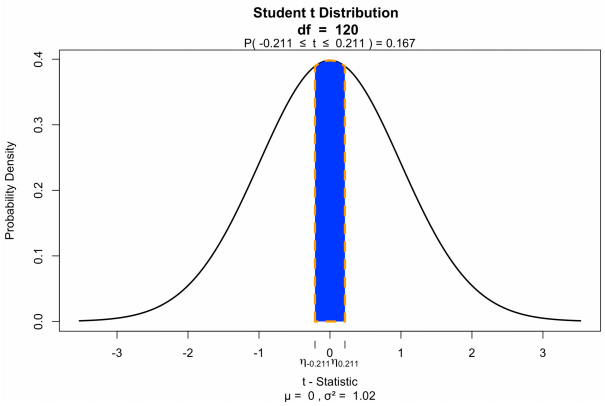
```
# 4.b
# The input I picked is totsc4
totsc4_vif <- lm(totsc4~avgsalary + pctel + bilingua + occupday + totday + spc +
                   speced + lnchpct + tchratio + percap, data = df7)
summary(totsc4_vif)
# Multiple R-squared: 0.666
# VIF = 1/(1-R^2) = 2.994254 # from vif(ml3)
# test: VIF = 1/(1-0.666)
1/(1-0.666) # 2.994012
```

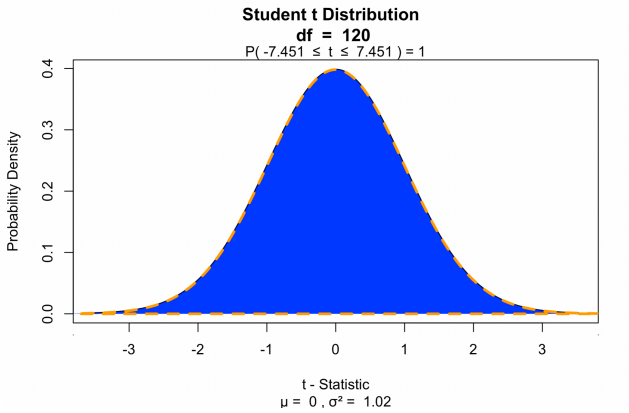
```
> vif(ml3)
 bilingua  occupday    totday       spc    speced   lnchpct  tchratio    percap    totsc4 avgsalary     pctel
 1.242510  1.354495  3.043272  1.185889  1.231265  3.533852  1.850789  2.875868  2.994254  3.241232  2.036938
```

```r
# 5
head(df7)

SSE <- sum(ml3$residuals^2)
SSE

SSR <- sum((fitted(ml3) - mean(df7$totsc4))^2)
SSR

SST <- SSR + SSE
SST

# My SST is 8906492877; My SSE is 206220375
```

```r
# 6
SSR
# My SSR is 8700272501.

# 7
R_sqt <- SSR/SST
R_sqt
# My SSR/SST is 0.9768461. We can also see this value in the summary of my model
# at the bottom: Multiple R-squared.
```

```r
# 8
library(visualize)
# F-statistic: 14.16 on 11 and 120 DF

# I chose totsc4 as my predictor. t value = 0.211
visualize.t(stat=c(-0.211, 0.211 ), df = 120, section="bounded")

# the range was too small, I chose a predictor, totday, t value = 7.451
visualize.t(stat=c(-7.451, 7.451 ), df = 120, section="bounded")

# 100% of the curve is shaded.
# As standard errors become smaller, relative to the coefficient value, that's
# going to make the t-value bigger. A bigger t-value will occupy more space on
# the distribution shown to the right.
```






```r
# 9
# My F-statistic: 14.16 on 11 and 120 DF, p-value: < 2.2e-16

# 10
```

```r
# 11
regt2 <- lm(totsc4 ~ percap, data = train2)
ta2 <- accuracy(predict(regt2, train2), train2$totsc4)

regv2 <- lm(totsc4 ~ percap, data = valid2)
va2 <- accuracy(predict(regv2, valid2), valid2$totsc4)

ta2
va2

# The RMSE in my train2 model is 10.88487, the MAE is 8.259777.
# The RMSE in my valid2 model is 12.88238, the MAE is 10.3279.

# The difference is around 2. The SLR model has 17 variables, my model only
# has 16, becasue I dropped 'regday' column. I also changed the data type
# of 'code' from int to factor. I also fill the NAs with some numbers that
# created myself based on the dataset. By dropping one column and filling NAs
# with created numbers, it will affect the accuracy. Another thing that may
# affect the accuracy is the sample size. The valid2 dataset only has 88
# objects. Small sample size could cause results of low accuracy.

```

273:5 (Top Level)

Console | Terminal | Jobs

R 4.1.1 · ~/Desktop/

```
> ta2
                    ME     RMSE      MAE          MPE     MAPE
Test set 1.754026e-13 10.88487 8.259777 -0.02379239 1.164697
> va2
                     ME     RMSE      MAE          MPE     MAPE
Test set -1.230329e-14 12.88238 10.3279 -0.03378302 1.465216
```

Reference

Kang, H. (2013, March 24). *NCBI - WWW Error Blocked Diagnostic.* NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/#:%7E:text=Missing%20data%20present%20various%20problems,the%20representativeness%20of%20the%20samples

Kenton, W. (2020, November 10). *How Per Capita Income is Calculated and Used by Companies.* Investopedia. https://www.investopedia.com/terms/i/income-per-capita.asp#:%7E:text=What%20Is%20Per%20Capita%20Income,of%20life%20of%20the%20population

Pearson's Product-Moment Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example. (n.d.). Laerd. https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php#:%7E:text=The%20Pearson%20product%2Dmoment%20correlation,at%20least%20an%20interval%20scale

Haroon, D. (2019). *What is the difference between RMSE and Standard Deviation?* Quora. https://www.quora.com/What-is-the-difference-between-RMSE-and-Standard-Deviation#:%7E:text=RMSE%20is%20calculated%20on%20the,on%20the%20already%20available%20data.&text=The%20difference%20is%20in%20where%20they%20are%20applied