MATH 4780 (MSSC 5780) Homework 6

Due Date: November 4, 2022 11:59 PM Henrique Medeiros Dos Reis

1 Homework Instruction and Requirement

- Homework 6 covers course materials of Week 1 to 12.
- Please submit your work in **one PDF** file including all parts to **D2L > Assessments > Dropbox**. *Multiple files or a file that is not in pdf format are not allowed.*
- In your homework, please number and answer questions **in order**.
- Your entire work should be completed by any word processing software (Microsoft Word, Google Docs, (R)Markdown, LaTex, etc) and your preferred programming language. Your document should be a PDF file.
- It is your responsibility to let me understand what you try to show. If you type your answers, make sure there are no typos. I grade your work based on *what you show, not what you want to show*. If you choose to handwrite your answers, write them neatly. If I can't read your sloppy handwriting, your answer is judged as wrong.

2 Reading and Writing

In this course, we have been using the classical or frequentist approach to do a variety of statistical inferences, marginal t test for β_j and F tests for model comparison for example. But Dr. Yu once said he never uses p-value in his own research, and what is taught in Intro Stats MATH 4720 is mostly problematic. In fact, the null hypothesis significance testing (NHST) paradigm and the p-value usage have been much criticized and shown to be problematic, misused, and resulting in reproducibility and replication crisis in scientific research. Please write a summary paper *at least two pages* including

- Interpretation of p-value
- List and discussion about the problems of the NHST and p-value method
- Possible solutions to those problems

Some references are

- Wikipedia: Misuse of p-values
- A. Reinhart (2015), "Statistics Done Wrong", *No Starch Press*, San Francisco.
- R. L. Wasserstein and Nicole A. Lazar (2016), "The ASA Statement on p-Values: Context, Process, and Purpose", *The American Statistician*, 70:2, 129-133.
- V. Amrhein, S. Greenland and B. McShane (2019), "Retire statistical significance", *Nature*, 567, 305-307.

- B. McShane, D. Gal, A. Gelman, C. Robert and Jennifer L. Tackett (2019), "Abandon Statistical Significance", *The American Statistician*, 73:sup1, 235-245.
- A. Gelamn and E. Loken (2014), "The Statistical Crisis in Science", *American Scientist*, 102, 460-465.
- R. L. Wasserstein, A. L. Schirm and N. A. Lazar (2019), "Moving to a World Beyond $p < 0.05$", *The American Statistician*, 73:sup1, 1-19.

There are lots of discussions and papers out there. You are welcome to google more resources to support your argument. The work should be entirely your effort. **You are not allowed to copy anyone's words, and you have to cite any resources you use, papers, blogs, videos, lecture notes, etc, or you violate Marquette academic misconduct policy.**

Check last page

3 Statistical Computing and Data Analysis

Please perform a data analysis using `R` or your preferred language. **Any results should be generated by computer outputs, and your work should be done entirely by a computer. Handwriting is not allowed. Relevant code should be attached.**

3.1 Gasoline Mileage Data

We use the same data set `mpg.csv` for data analysis. For the following analysis, if any regressor contains a missing value `NA`, remove the corresponding row of the data matrix.

```
mpgData <- read.csv("mpg.csv")
head(mpgData,3)
```

```
##      y  x1  x2  x3   x4   x5 x6 x7    x8   x9  x10 x11
## 1 18.9 350 165 260 8.00 2.56  4  3 200.3 69.9 3910   1
## 2 17.0 350 170 275 8.50 2.56  4  3 199.6 72.9 3860   1
## 3 20.0 250 105 185 8.25 2.73  1  3 196.7 72.2 3510   1
```

```
mpgData<-mpgData[c(-23,-25),]
```

1. Build a linear regression model relating gasoline mileage y to vehicle weight x_{10} and the type of transmission x_{11} (1 automatic; 0 manual). Does the type of transmission significantly affect the mileage performance?

   Solution:

```
mpg_wt_t <- lm(y~x10+x11, data=mpgData)
summary(mpg_wt_t)
```

```
##
## Call:
## lm(formula = y ~ x10 + x11, data = mpgData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0152 -2.4290 -0.1662  2.4488  6.8922
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.1807662  2.7095142  14.460 3.11e-14 ***
## x10         -0.0047650  0.0009898  -4.814 5.02e-05 ***
## x11         -2.5436905  2.0708234  -1.228     0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.3 on 27 degrees of freedom
## Multiple R-squared:  0.7418, Adjusted R-squared:  0.7227
## F-statistic: 38.79 on 2 and 27 DF,  p-value: 1.15e-08
```

It does not do a great job, we would need to include some interaction term and test it again. The p-value is a little high, and the standard error is also too big compared with the estimate of $x_1 1$.

2. Modify the model developed in (1) to include an interaction between vehicle weight and the type of transmission. What conclusions can you draw about the effect of the type of transmission on gasoline mileage? Interpret the parameters in this model.

   Solution:

```
mpg_wt_t <- lm(y~x10*x11, data=mpgData)
summary(mpg_wt_t)
```

```
##
## Call:
## lm(formula = y ~ x10 * x11, data = mpgData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.565 -1.736  0.169  1.339  4.828
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.290809   5.423544  10.932 3.21e-11 ***
## x10          -0.012905   0.002163  -5.967 2.68e-06 ***
## x11         -27.730076   6.444509  -4.303 0.000211 ***
## x10:x11       0.009395   0.002323   4.044 0.000417 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.635 on 26 degrees of freedom
## Multiple R-squared:  0.8415, Adjusted R-squared:  0.8232
## F-statistic: 46.01 on 3 and 26 DF,  p-value: 1.539e-10
```

Now it is relevant, since there is a correlation between weight and type of transmission. The interaction term resolves the problem that we had before. The type of transmission seems to impact gas mileage, but the weight of the car highly correlated with the type of transmission and gas mileage.

For the following questions (3) to (9), consider all the regressors except x_4, x_5 and x_{11}.

- y: MPG
- x_1: Displacement (cubic in.)
- x_2: Horsepower (ft-lb)
- x_3: Torque (ft-lb)
- x_6: Carburetor (barrels)
- x_7: No. of transmission speeds
- x_8: Overall length (in.)
- x_9: Width (in.)
- x_{10}: Weight (lb)

Normalize the response and predictors by unit length scaling before later analysis.

3. Obtain the correlation matrix of regressors. Does it give any indication of collinearity?

   Solution:

```r
mpg_modif <- mpgData[,c(-5,-6,-12)]

mpg_scale <- mpg_modif

for(i in 1:length(mpg_modif))
{
  mpg_scale[,i]<-mpg_scale[,i]/norm(as.matrix(mpg_scale[,i]),"2")
}
X<- mpg_scale[,-1]
(XtX <- cor(X))
```

```
##            x1         x2         x3         x6         x7         x8
## x1  1.0000000  0.9408473  0.9891628  0.6427984 -0.7719151  0.8623681
## x2  0.9408473  1.0000000  0.9643592  0.7614190 -0.6259445  0.8027387
## x3  0.9891628  0.9643592  1.0000000  0.6531263 -0.7461800  0.8641224
## x6  0.6427984  0.7614190  0.6531263  1.0000000 -0.2756386  0.4220680
```

```
## x7  -0.7719151 -0.6259445 -0.7461800 -0.2756386  1.0000000 -0.6552065
## x8   0.8623681  0.8027387  0.8641224  0.4220680 -0.6552065  1.0000000
## x9   0.7974811  0.7105117  0.7881284  0.3003862 -0.6551300  0.8831512
## x10  0.9515520  0.8878810  0.9434871  0.5203669 -0.7058126  0.9554541
##             x9        x10
## x1   0.7974811  0.9515520
## x2   0.7105117  0.8878810
## x3   0.7881284  0.9434871
## x6   0.3003862  0.5203669
## x7  -0.6551300 -0.7058126
## x8   0.8831512  0.9554541
## x9   1.0000000  0.8994711
## x10  0.8994711  1.0000000
```

There are indications of collinearity between different parameters.

4. Calculate the variance inflation factors (VIFs). Is there any evidence of collinearity?

   Solution:

```
full_mod <- lm(y~., data = data.frame(mpg_scale))
(vif_all <- car::vif(full_mod))
```

```
##         x1         x2         x3         x6         x7         x8         x9
## 113.523010  33.895263 115.933545   4.595686   4.309102  18.178509   7.468513
##        x10
##  78.555890
```

There is evidence of colinearity. x_1, x_2, x_3, x_8, x_10 all have variance inflation factors higher than 10, which is significant evidence of colinearity.

5. Find the eigenvectors associated with the smallest eigenvalues of $\mathbf{X}'\mathbf{X}$. Interpret the elements of these vectors. What can you say about the source of collinearity in these data?

   Solution:

```
eigen_XtX <- eigen(XtX)
(lambda <- eigen_XtX$values)
```

```
## [1] 6.351402513 0.928138000 0.441075405 0.130021286 0.095475961 0.037933061
## [7] 0.011856952 0.004096822
```

Collinearity seems to come from $\lambda_6, \lambda_7, \lambda_8, \lambda_9$. However, that does not tell us which x, because λs are ordered.

6. Compute the condition indices and variance-decomposition proportions. What statements can you make about collinearity in these data?

Solution:

```
max(lambda) / lambda
```

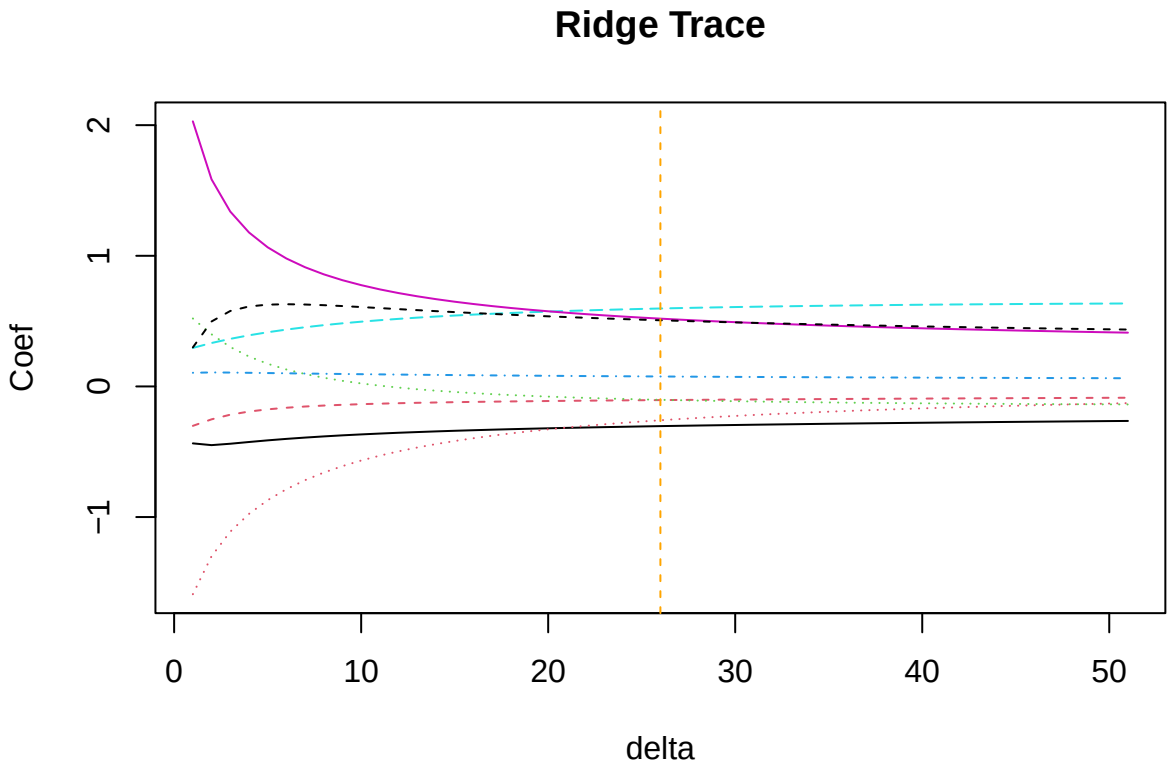
```
## [1]    1.000000    6.843166   14.399811   48.848944   66.523578  167.437120
## [7]  535.669069 1550.324241
```

Only k_8 has a value taht is higher than 1000, which implicates that there is colinearity in that specificy lambda value.

7. Fit the ordinary multiple regression and the ridge regression. Use the ridge trace to select an appropriate value of δ. Explain how the ridge regression coefficients change as the parameter δ increases. This gives you the idea of why the ridge regression is a *shrinkage* method.

Solution:

```
delta <- seq(0, 0.5, by = 0.01)
ridge_fit <- MASS::lm.ridge(y ~ .-1,
                            data = data.frame(mpg_scale),
                            lambda = delta)
matplot(coef(ridge_fit), type = "l",
        xlab = "delta",
        ylab = "Coef",
        main = "Ridge Trace")
abline(v = which(delta == 0.25),
       col = "orange", lty = 2)
```



It is possible to see that all the lines are converging to almost the same result. The ridge regression coefficients change as the parameter δ increases, since we are introducing some bias at the cost of R^2, but after some time it will converge to a stabilized tradeof between the bias and R^2.

8. Use the all-possible-regressions approach to find an appropriate regression model.

Solution:

```
olsrr_all <- olsrr::ols_step_all_possible(full_mod)
print(olsrr_all[which.max(olsrr_all$adjr),])
```

```
##    Index N Predictors  R-Square Adj. R-Square Mallow's Cp
## 90    37 3  x7 x8 x10 0.7886898     0.7643079   0.3427024
```

```
print(olsrr_all[which.max(olsrr_all$rsquare),])
```

```
##     Index N                Predictors  R-Square Adj. R-Square Mallow's Cp
## 255   255 8 x1 x2 x3 x6 x7 x8 x9 x10 0.8013887     0.7257273           9
```

```
print(olsrr_all[which.min(olsrr_all$cp),])
```

```
##   Index N Predictors  R-Square Adj. R-Square Mallow's Cp
## 1     1 1         x1 0.7606808     0.7521337  -0.6957811
```

```
olsrr::ols_step_best_subset(full_mod, metric = "adjr")
```

```
##         Best Subsets Regression
## ---------------------------------------
## Model Index    Predictors
## ---------------------------------------
##      1         x1
##      2         x1 x6
##      3         x7 x8 x10
##      4         x1 x6 x8 x10
##      5         x1 x6 x8 x9 x10
##      6         x1 x3 x6 x8 x9 x10
##      7         x1 x2 x3 x6 x8 x9 x10
##      8         x1 x2 x3 x6 x7 x8 x9 x10
## ---------------------------------------
##
##                                                    Subsets Regression Summary
## ------------------------------------------------------------------------------------------------
##                       Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)         AIC         SBIC          SBC
## ------------------------------------------------------------------------------------------------
##   1        0.7607      0.7521      0.7075    -0.6958    -127.2638    -211.7267    -123.0602
##   2        0.7741      0.7574      0.7116    -0.1174    -126.9984    -210.7290    -121.3936
##   3        0.7887      0.7643       0.703     0.3427    -126.9979    -209.6274    -119.9919
##   4        0.7915      0.7582      0.6991     2.0428    -125.4033    -207.1904    -116.9962
##   5        0.7944      0.7516      0.6406     3.7371    -123.8223    -204.6861    -114.0139
##   6        0.7969      0.7439      0.6188     5.4760    -122.1848    -202.0795    -110.9752
##   7        0.8001      0.7366      0.5879     7.1315    -120.6699    -199.4436    -108.0592
##   8        0.8014      0.7257      0.5341     9.0000    -118.8572    -196.6466    -104.8453
## ------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

The appropriate model is given by using the predictors x_7, x_8, x_10

9. Use stepwise regression to specify a subset regression model. Does this lead to the same model found in (8)?

Solution:

```
(olsrr_both <- olsrr::ols_step_both_aic(full_mod))
```

```
##
##
##                               Stepwise Summary
## ------------------------------------------------------------------------
## Variable     Method       AIC        RSS     Sum Sq     R-Sq      Adj. R-Sq
## ------------------------------------------------------------------------
## x1          addition    -127.264    0.021     0.066    0.76068      0.75213
## ------------------------------------------------------------------------
```

This does not lead to the same model, it suggests to only use x_1 as a predictor.

Henri Medeiros Dos Reis

Dr. Yu

MSSC 5780

9 December 2022

Problems and Misconceptions about p-values, how to fix them?

In order to discuss the problems of using p-values improperly, it is needed to first understand what a p-value is and how it should be used, not only use it because it is popular. Well, p-value is the likelihood, under the premise that some null hypothesis is true, such that the outcomes of a hypothesis test will be at least close to the observed findings. The alternative hypothesis is more likely to be supported by stronger evidence when the p-value is lower. P-value is also frequently utilized by organizations to support the legitimacy of their studies or findings.

Some of the problems and misconceptions that usually happen when using p-value are:

P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. Most people frequently tend to interpret a p-value as a statement about the validity of the null hypothesis, or the likelihood that the observed data were generated by chance. Which is false for both assumptions.

Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold, also the fact that the user is setting the threshold is problematic by itself. Using this threshold implies that the conclusion instantaneously changes from true to false in very small differences, such as 0.049999 is true, while 0.0500001 is false.

Proper inference requires full reporting and transparency, p-values and associated analyses shouldn't be reported in an arbitrary manner. Instead, they should be used in addition to

one another. P-value could support a massive analysis process, while the process and the results have equal importance.

A p-value, or statistical significance, does not measure the size of an effect or the importance of a result. Any effect, no matter how little, can result in a low p-value if sample size or measurement accuracy are high enough, while huge effects can result in unimpressive p-values if sample size or measurement accuracy are low. Statistical significance is not the same as significance in the subject we are studying. Smaller p-values do not always indicate more significant results, while larger p-values do not always indicate less significant or even no effects.

By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis .A p-value without context or any supporting data only provides limited information, and people need to be aware of this. For these reasons, when other methods are suitable and practical, data analysis should not stop with the determination of a p-value, instead, further analysis and the use of other approaches are also needed.

One of the most common solutions that people use are replacing p-values completely with methods that emphasize the estimation of the model rather than the results. Some of the approaches used to achieve this are: Bayesian methods, measuring evidence, confidence, credibility, and other similar approaches. All the approaches listed above do not rely on only one value to tell whether the result is good or bad. They do need a lot of further assumptions, and deal more adequately with the size of the effects, or even the validity of the hypothesis, making it more reliable in general.

In conclusion, the results of some research should not simply be qualified over a simple constant, it should be rather qualified under the principles of good design and many different tests, either graphical or numerical that can support what that specific study is trying to answer.

References

Wikipedia: Misuse of p-values

A. Reinhart (2015), “Statistics Done Wrong”, No Starch Press, San Francisco.

R. L. Wasserstein and Nicole A. Lazar (2016), “The ASA Statement on p-Values: Context, Process, and Purpose”, The American Statistician, 70:2, 129-133.

.