SMML I Exam 3

John Kubale

2024-12-3

Notes

- Label each answer with the appropriate question number in your R Markdown document (e.g., 1.1, 1.2, etc.).
- Include all relevant code and output with your answer next to the appropriate question number. **Do not** provide all code/output at the end of the document.
- Clearly demonstrate your work. Where applicable, include any R code pertinent to your answer.
- Submit a single pdf file via Canvas by the deadline (12pm December 10).
- You may consult any reference materials **except tools that utilize AI (e.g., Chat GPT, Github Copilot, etc.)**
- This is not a group assignment so **do not** consult with your classmates. Your answers should be based on your own, individual work.
- The point value for each question is given in square brackets.

Please answer whether the following statements (Questions 1-3) are True or False.

1. The Box-Cox transformation can help you identify what transformation of a predictor might benefit the model. [2pt]

2. A limitation of leave one out (LOO) cross-validation is that it can be computationally intensive. [2pt]

3. We can use cross-validation methods to estimate the test error of a model. [2pt]

4. You fit a model and are assessing for the presence of collinearity by estimating VIF for the predictors. The VIF for one of the predictors is 212.678. Roughly

how much higher would you expect the standard error of this predictor to be than if there were no collinearity (you must show your work)? [2pt]

5. List 3 ways (discussed in class) that variable selection methods (forward/backward/stepwise) can be problematic. [3pt]

6. "PATH_FINAL.csv" is provided on Canvas.

- Read in the data ("PATH_FINAL.csv").
- The data comes from a project called "Positive Attitudes Towards Health" in 2017 that targeted persons who inject drugs in Southeast Michigan.
- This dataset includes 409 cases and the following variables:
  - SAMPLEID: ID of respondents
  - AGE: Age in years
  - MALE: 1. Male; 0. Female
  - BLACK: 1. Race black; 0. Other race than black
  - EDUC: 1. <High school (HS); 2. HS Education; 3. >HS Education
  - LIFESAT: Life satisfaction score summarized from 5 questions; Higher scores mean higher satisfaction
  - INJECTMULTIPLE: 1. Inject multiple drugs; 0. Injects only one drug
  - AGE_DIFF: Age one feels (from a question, "How old do you feel?") minus actual age. Positive values mean feeling older than actual age. Negative values mean feeling younger than actual age.

6.1. Regress AGE_DIFF on the remaining variables (except for SAMPLEID). Examine collinearity of the predictors and report your conclusions [9pt]

6.2. From the regression model in Q6.1, select predictors using AIC-based stepwise regression. Report the changes in AIC and the selected predictors through all steps. [10pt]

6.3. Based on the model selected in Q6.2, check whether the OLS assumptions are met. What do you conclude and what are the implications? [10pt]

6.4. Would a transformation of the outcome variable help the model selected in Q6.3? If so, which transformation? [5pt]

6.5. Focus on the slope coefficient for AGE in the model from Q6.3. Report its standard error and 95% confidence interval using the following three approaches.

What do you observe? [20pt] a. OLS b. Heteroscedasticity consistent variance estimator c. Bootstrap method (use set.seed(97) and bootstrap 1000 times; do not use boot() or boot.ci(); do not use quantile-based confidence intervals for this question)

6.6. Among the three CI's in Q6.5, which one would you recommend to a non-statistician and how would you explain your rationale? [5pt]

7. Questions 7.1-7.3 do not require you to do any coding, but instead ask you to interpret different types of output. PATH researchers conducted an experiment about the location of self-rated health question as follows: * The literature has debated on the location of self-rated health (SRH) question, which asks respondents, "Would you say your health in general is excellent, very good, good, fair, or poor?" * PATH researchers asked SRH at 3 different locations in the questionnaire (described below under SRH_LOC) and randomly assigned 410 respondents to one of these 3 locations. In other words, respondents were asked SRH only once, but at different places in the questionnaire. The researchers analyzed the data for this experiment which included the following variables: + SAMPLEID: ID of respondents + SRH: SRH score; 1. Poor … 5. Excellent; the higher, the better health + SRH_LOC: Location of SRH question 1. SRH asked before any health and well-being questions 2. SRH asked after chronic condition questions 3. SRH asked after a global life satisfaction question + GLOBALSAT: Global life satisfaction asked in 1 question 1. Satisfied 0. Not satisfied

7.1. PATH researchers analyzed SRH with the experimental factor, SRH_LOC}. The outputs are as follows. Report all hypotheses tested in these outputs mathematically or in text and describe the substantive conclusions in layman's terms. Use outputs from both summary() and anova(). [9pt]

```r
PATHQ7 <- read.csv("data/PATH_FINALQ7.csv")
summary(lm(SRH ~ SRH_LOC, PATHQ7))
```

```
##
## Call:
## lm(formula = SRH ~ SRH_LOC, data = PATHQ7)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8088 -0.6544  0.1912  0.5580  2.5580
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.65441    0.07625  34.812   <2e-16 ***
## SRH_LOC2.After chronic condition  0.15441    0.10783   1.432   0.1529
## SRH_LOC3.After life satisfaction -0.21238    0.10744  -1.977   0.0487 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8892 on 407 degrees of freedom
## Multiple R-squared:  0.02808,    Adjusted R-squared:  0.02331
## F-statistic:  5.88 on 2 and 407 DF,  p-value: 0.003038
```

```r
anova(lm(SRH ~ SRH_LOC, PATHQ7))
```

```
## Analysis of Variance Table
##
## Response: SRH
##            Df Sum Sq Mean Sq F value   Pr(>F)
## SRH_LOC     2   9.30  4.6495  5.8801 0.003038 **
## Residuals 407 321.82  0.7907
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

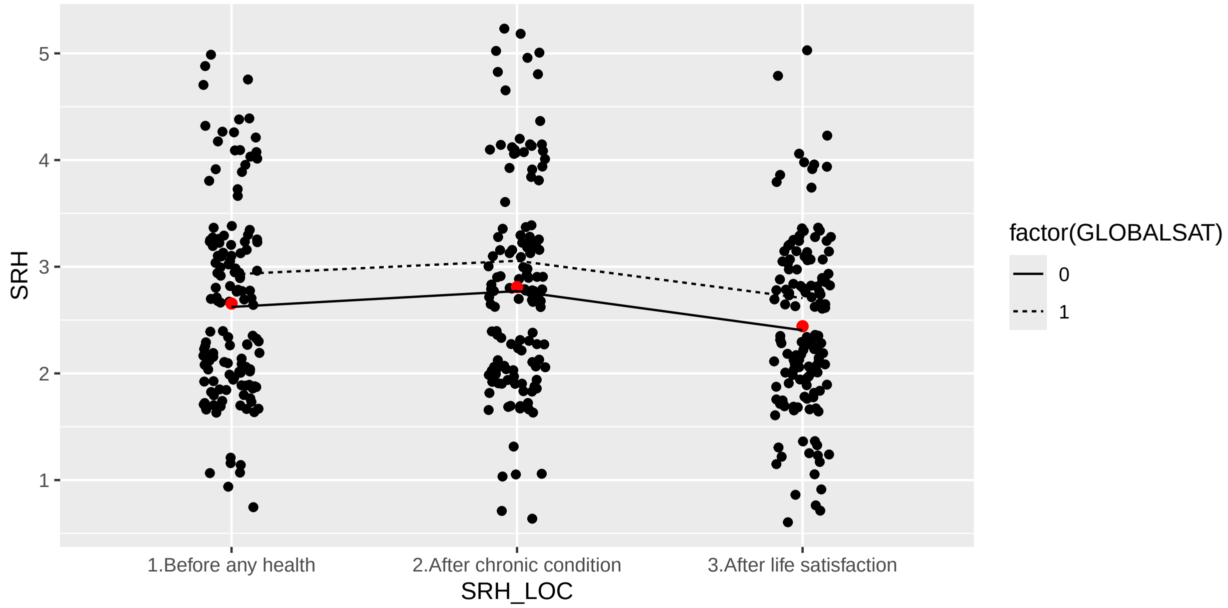
7.2. PATH researchers also analyzed SRH as a function of SRH_LOC as follows. Interpret SRH_LOC’s impact on SRH and the hypothesis tested mathematically or in text. Describe the substantive conclusions in layman’s terms. How does the hypothesis tested (for SRH_LOC) differ in this model when compared to a fixed effects model? [9pt]

```r
library(lme4)
summary(lmer(SRH ~ (1|SRH_LOC), PATHQ7))
```

```
## Linear mixed model fit by REML [’lmerMod’]
## Formula: SRH ~ (1 | SRH_LOC)
##    Data: PATHQ7
##
## REML criterion at convergence: 1074.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0007 -0.7322  0.2485  0.5908  2.8400
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  SRH_LOC  (Intercept) 0.02817  0.1678
##  Residual             0.79072  0.8892
## Number of obs: 410, groups:  SRH_LOC, 3
```

```
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.6349     0.1064   24.77
```

7.3. PATH researchers were curious about whether responses to the global life satisfaction (GLOBALSAT) interact with SRH_LOC and conducted the following analysis. Report all hypotheses tested in the output of anova() below (mathematically or in text). Describe the substantive conclusions of each hypothesis in layman's terms. Be sure to incorporate your observations about the plot. [12pt]



```
anova(lm(SRH ~ SRH_LOC*GLOBALSAT, PATHQ7))
```

```
## Analysis of Variance Table
##
## Response: SRH
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## SRH_LOC             2   9.30  4.6495  5.9062 0.002963 **
## GLOBALSAT           1   3.78  3.7821  4.8044 0.028957 *
## SRH_LOC:GLOBALSAT   2   0.00  0.0018  0.0023 0.997695
## Residuals         404 318.04  0.7872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```