Lab 4 Covid model selection

Kent Codding

2023-09-27

# Set up

setwd("C:/Users/Kent Codding/Desktop/Adv Biostat")  
all\_data <- read.csv(file = "Lab 2/all\_data.csv")  
vote <- read.csv(file = "Lab 3/vote.csv")

## make new df

all\_data <- all\_data[, c("total\_death\_rate","perc\_one\_dose", "perc\_fully\_vax","med\_income","nuclear", "Row.names")]  
  
vote <- vote[, c("called", "dem\_percent")]  
df <- merge(all\_data, vote, by = 0)[-1] #-1 gets rid of duplicated row.names in result

## Warning in merge.data.frame(all\_data, vote, by = 0): column name 'Row.names' is  
## duplicated in the result

## change rownames to state name

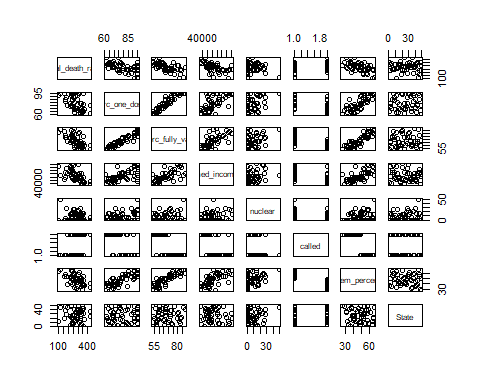
df$State <- df$Row.names  
df <- df[-6]

# Data cleaning and selection

## Outliers

### Graph it

plot(df)



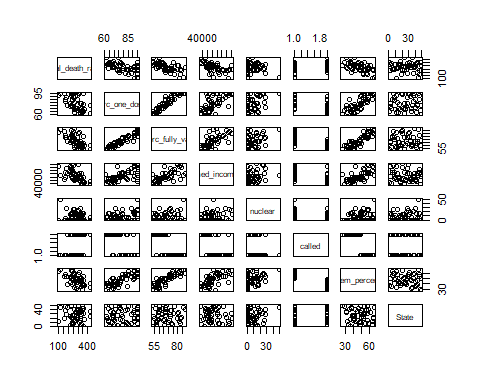
### Question 1

option 1. Remove outliers from dataset. option 2. Apply data transformation (logarithmic transformation reduced outliers in logworms case as shown in class).

## Collinearity

### Graph it

plot(df)



cor(df$perc\_one\_dose, df$perc\_fully\_vax) #collinear - might be redundant

## [1] 0.9440005

### Question 2

Yes, some variables show collinearity like percent one dose and percent fully vax. This is because they are measuring a very similar metric. I can verify collinearity by looking at visualizations of the data like the plot above or running tests like the Pearson R above… which should be taken with a grain of salt. After running a linear model, if collinear variables are still present within, redundant categorical variables could give an NA result as shown with the SPECIES and NATIVE categorical variables in the crayfish example. If numerical variables like percent one dose and percent fully vax are kept in this case, a linear model may incorrectly assume which variable is the predictor. As follows, the summary of the linear model would give different results each time pertaining to which variable is an accurate predictor and which is not.

### remove variables

df <- df[-c(2,7)]

# Interactions

## Question 3

because we do not have a specific hypothesis pertaining to the remaining variables within df, it may be unwise to include an interaction term willy-nilly because it can make the model unnecessarily complicated. It may be useful to add an interaction term after playing with the barebones model first. ## Model it

mod1 <- lm(formula = total\_death\_rate ~ perc\_fully\_vax + med\_income + nuclear + called, data = df)  
summary(mod1)

##   
## Call:  
## lm(formula = total\_death\_rate ~ perc\_fully\_vax + med\_income +   
## nuclear + called, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -103.896 -29.340 4.671 31.406 105.557   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 656.421433 96.963278 6.770 2.24e-08 \*\*\*  
## perc\_fully\_vax -3.679973 1.338629 -2.749 0.00857 \*\*   
## med\_income -0.002321 0.001263 -1.838 0.07268 .   
## nuclear -0.179946 0.750795 -0.240 0.81167   
## calledR -13.810502 23.887199 -0.578 0.56604   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 53.64 on 45 degrees of freedom  
## Multiple R-squared: 0.4457, Adjusted R-squared: 0.3964   
## F-statistic: 9.047 on 4 and 45 DF, p-value: 1.89e-05

## Question 4

THe overall model is significant as the p-value is 1.89e-05. However, only perc\_fully\_vax is significant. The other variables will have to be removed by top-down model selection in the next step. The adjusted R-squared is about 40%, showing that this barebones model explains about 40% of variation in the response variable.

## perform “top-down” model selection using drop1 until all predictors are significant

drop1(mod1, test = "F")

## Single term deletions  
##   
## Model:  
## total\_death\_rate ~ perc\_fully\_vax + med\_income + nuclear + called  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 129472 402.96   
## perc\_fully\_vax 1 21743.6 151215 408.72 7.5573 0.008573 \*\*  
## med\_income 1 9718.9 139191 404.58 3.3780 0.072680 .   
## nuclear 1 165.3 129637 401.02 0.0574 0.811672   
## called 1 961.7 130433 401.33 0.3343 0.566042   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod2 <- lm(total\_death\_rate ~ perc\_fully\_vax + med\_income + called, data = df)  
summary(mod2)

##   
## Call:  
## lm(formula = total\_death\_rate ~ perc\_fully\_vax + med\_income +   
## called, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -101.643 -29.582 2.992 30.350 106.763   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 656.005501 95.949359 6.837 1.6e-08 \*\*\*  
## perc\_fully\_vax -3.706301 1.320376 -2.807 0.00731 \*\*   
## med\_income -0.002316 0.001250 -1.854 0.07023 .   
## calledR -13.208149 23.509990 -0.562 0.57697   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 53.09 on 46 degrees of freedom  
## Multiple R-squared: 0.445, Adjusted R-squared: 0.4088   
## F-statistic: 12.29 on 3 and 46 DF, p-value: 4.942e-06

drop1(mod2, test = "F")

## Single term deletions  
##   
## Model:  
## total\_death\_rate ~ perc\_fully\_vax + med\_income + called  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 129637 401.02   
## perc\_fully\_vax 1 22205.3 151842 406.93 7.8793 0.00731 \*\*  
## med\_income 1 9681.8 139319 402.62 3.4355 0.07023 .   
## called 1 889.5 130526 399.37 0.3156 0.57697   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod3 <- lm(total\_death\_rate ~ perc\_fully\_vax + med\_income, data = df)  
summary(mod3)

##   
## Call:  
## lm(formula = total\_death\_rate ~ perc\_fully\_vax + med\_income,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -103.202 -31.441 4.955 37.050 104.617   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 612.149172 55.382267 11.053 1.15e-14 \*\*\*  
## perc\_fully\_vax -3.265862 1.054683 -3.097 0.0033 \*\*   
## med\_income -0.002174 0.001215 -1.790 0.0800 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 52.7 on 47 degrees of freedom  
## Multiple R-squared: 0.4412, Adjusted R-squared: 0.4174   
## F-statistic: 18.55 on 2 and 47 DF, p-value: 1.149e-06

drop1(mod3, test = "F")

## Single term deletions  
##   
## Model:  
## total\_death\_rate ~ perc\_fully\_vax + med\_income  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 130526 399.37   
## perc\_fully\_vax 1 26629 157155 406.65 9.5885 0.003298 \*\*  
## med\_income 1 8895 139421 400.66 3.2029 0.079951 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod4 <- lm(total\_death\_rate ~ perc\_fully\_vax, data = df)  
summary(mod4)

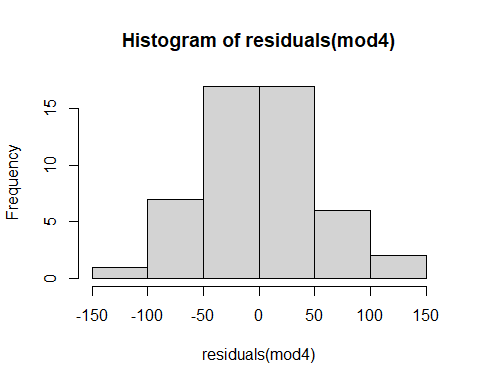
##   
## Call:  
## lm(formula = total\_death\_rate ~ perc\_fully\_vax, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -117.823 -36.869 -0.244 32.146 116.443   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 586.257 54.672 10.723 2.45e-14 \*\*\*  
## perc\_fully\_vax -4.538 0.797 -5.694 7.34e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 53.89 on 48 degrees of freedom  
## Multiple R-squared: 0.4031, Adjusted R-squared: 0.3907   
## F-statistic: 32.42 on 1 and 48 DF, p-value: 7.337e-07

# Model validation

## Assumption of normality

### Graph it

hist(residuals(mod4))

 ### Question 5

shapiro.test(mod4$residuals) #take with grain of salt because it is better to visualize

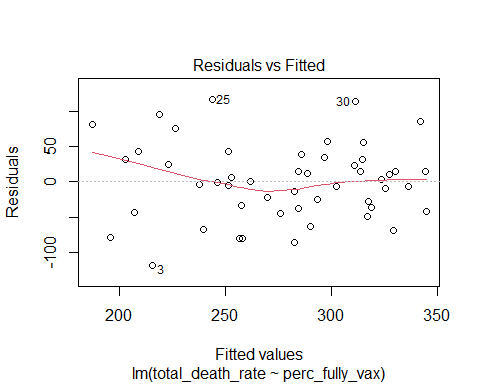
##   
## Shapiro-Wilk normality test  
##   
## data: mod4$residuals  
## W = 0.987, p-value = 0.8531

There is no evidence of violating the assumption according to both the histogram of residuals and the W, which is close to 1.

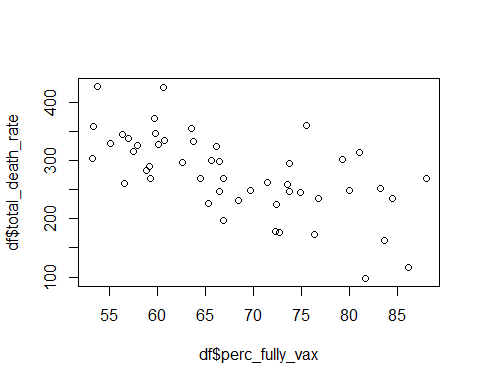
## Assumption of homogeneity

### Graph it

plot(mod4, which = 1)

 ### Question 6 my model passes the test for this assumption. If my model showed heterogeneity, it may need to undergo a transformation like a log transformation. ## Assumption of fixed X ### Question 7

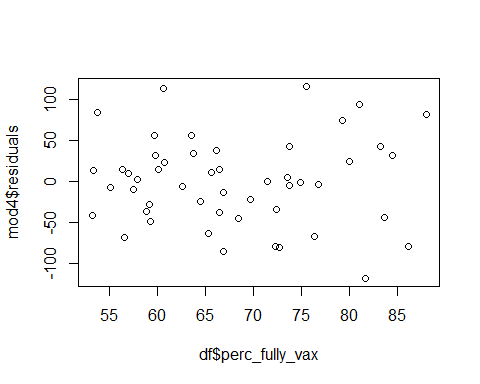
plot(df$total\_death\_rate ~ df$perc\_fully\_vax)

 No, I do not think that my predictor variable meets this assumption because noise exists within the X variable: perc\_fully\_vax. But then again, what model can meet this assumption?

## Assumption of Independence

### Graph it

plot(mod4$residuals ~ df$perc\_fully\_vax)

 ### Question 8 This plot shows no obvious relationship between values of y being affected by any other values of X, so it provides evidence for independence. ### Graph it

y\_pred <- predict(mod4, newdata = df)  
  
plot(df$total\_death\_rate ~ df$perc\_fully\_vax,  
 bg = "coral",  
 pch = 21)  
lines(y\_pred ~ df$perc\_fully\_vax, col = "green")

