Group 4 Stat-512 Project Markdown

Alex Scarcelli
Ryan Ellis
Ali Erkan

```
[1] "Loading libraries"
[1] "/usr/local/lib/R/site-library:/usr/lib/R/site-library:/usr/lib/R/library"
[1] "/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games:/snap/bin:/home/rstudio/.local/bin:/home/rstudio/bin"
[1] "Loading MLR3"

[1] "Loading required package: lpsolve"

[1] "Loading required package: Rcpp"

[1] "Loading required package: ggplot2"

[1] "Loading required package: data.table"

[1] "Loading MASS"
[1] "Loading glmnet"

[1] "Loading required package: Matrix"

[1] "Loading required package: survival"

[1] "Loading required package: lattice"

[1] "Loading required package: parallel"

[1] "Loading required package: mlr"

[1] "Attaching package: 'lattice'"

The following object is masked from 'package:ggplot2':

    qq

[1] "Loading rpart"
[1] "Loading rpart.plot"
[1] "Loading rsample"
[1] "Loading ranger"
```

```
[1] "/home/ubuntu/final_deaths/datasets/cdc_100k.csv"
```

```
## Dim:
## [1] 100000 [1] Total_deaths = (positive_100k, data = CDC100)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.00000 -0.00000  0.00000  0.00000  0.00000
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.000000   0.000000   0.000   1.000
## positive_100k 1.000000   0.000000   1.000   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0 on 99998 degrees of freedom
## Multiple R-squared:  1,\tAdjusted R-squared:  1
## F-statistic: 1.0E+36 on 1 and 99998 DF,  p-value: <2.2e-16
```

```
/home/ubuntu/final_deaths/datasets/cdc_100k.csv
```

```
## Analysis of Variance Table
##
## Response: Total_deaths
##           Df Sum Sq Mean Sq F value Pr(>F)
## positive_100k  1 100000  100000    Inf <2.2e-16 ***
## Residuals 99998      0       0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
/home/ubuntu/final_deaths/datasets/cdc_100k.csv
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.2000 -0.5000 -0.1000  0.3000  1.8000
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7000     0.7000  -1.000    0.321
x_pf_100k     0.2000     0.1000   2.000    0.046 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9 on 18 degrees of freedom
Multiple R-squared,  0.2222,	Adjusted R-squared:  0.1667
F-statistic: 4.25 on 1 and 18 DF,  p-value: 0.0448
```

```
>>>
```

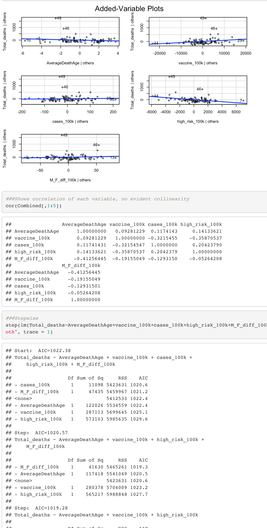
```
  2  3  4  5  6  7  8  9
    5
```

```
[1]"DATA/STATA/STATA_DATA/BATHURSTIL2.
Lambda
```

```
## [1] -0.4999999
```

```
[1]"DATA/STATA/STATA_DATA/BATHURSTIL2.
#GAM(formula=log(LC)~ZCOVERSPIN2_dbarbs*F_ZCOVERSPIN2+ZCOVERSPIN2_SYN+ZCOVERSPIN2_SYN4+
+ZCOVERSPIN2_SYN5+
+ZCOVERSPIN2_SYN6+
+ZCOVERSPIN2_SYN7+
+ZCOVERSPIN2_SYN8+
+ZCOVERSPIN2_SYN9+
+ZCOVERSPIN2_SYN10+
+ZCOVERSPIN2_SYN11+
+ZCOVERSPIN2_SYN12+
+ZCOVERSPIN2_SYN13+
+ZCOVERSPIN2_SYN14+
+ZCOVERSPIN2_SYN15+
+ZCOVERSPIN2_SYN16+
+ZCOVERSPIN2_SYN17+
+ZCOVERSPIN2_SYN18+
+ZCOVERSPIN2_SYN19+
+ZCOVERSPIN2_SYN20+
+ZCOVERSPIN2_SYN21+
+ZCOVERSPIN2_SYN22+
+ZCOVERSPIN2_SYN23+
+ZCOVERSPIN2_SYN24+
+ZCOVERSPIN2_SYN25+
+ZCOVERSPIN2_SYN26+
+ZCOVERSPIN2_SYN27+
+ZCOVERSPIN2_SYN28+
+ZCOVERSPIN2_SYN29+
+ZCOVERSPIN2_SYN30+
+ZCOVERSPIN2_SYN31+
+ZCOVERSPIN2_SYN32+
+ZCOVERSPIN2_SYN33+
+ZCOVERSPIN2_SYN34+
+ZCOVERSPIN2_SYN35+
+ZCOVERSPIN2_SYN36+
+ZCOVERSPIN2_SYN37+
+ZCOVERSPIN2_SYN38+
+ZCOVERSPIN2_SYN39+
+ZCOVERSPIN2_SYN40+
+ZCOVERSPIN2_SYN41+
+ZCOVERSPIN2_SYN42+
+ZCOVERSPIN2_SYN43+
+ZCOVERSPIN2_SYN44+
```

```
avPlots(modelall)
```




```
###Show correlation of each variable, no evident collinearity
cor(Combined[,i:5])
```

```
##                   AverageDeathAge vaccine_100k cases_100k high_risk_100k
## AverageDeathAge            1.00000      0.09281229  0.11374143     0.14133621
## vaccine_100k               0.09281229   1.00000000 -0.12154655    -0.25870537
## cases_100k                 0.11374143  -0.12154655  1.00000000     0.29422790
## high_risk_100k             0.14133621  -0.25870537  0.29422379     1.00000000
## M_F_diff_100k             -0.41256445  -0.19535049 -0.12931550    -0.05264208
##                   M_F_diff_100k
## AverageDeathAge       -0.41256445
## vaccine_100k          -0.19535049
## cases_100k            -0.12931550
## high_risk_100k        -0.05264208
## M_F_diff_100k          1.00000000
```

```
###Stepwise
step(lm(Total_deaths~AverageDeathAge+vaccine_100k+cases_100k+high_risk_100k+M_F_diff_100k, Combined), method = "b
oth", trace = 1)
```

```
## Start:  AIC=1022.38
## Total_deaths ~ AverageDeathAge + vaccine_100k + cases_100k +
##     high_risk_100k + M_F_diff_100k
##
##                Df Sum of Sq    RSS    AIC
## - cases_100k    1      3199 5423631 1020.4
## - M_F_diff_100k 1     47435 5439967 1021.2
## <none>                      5423531 1022.4
## - AverageDeathAge  1  122020 5545551 1022.4
## - vaccine_100k  1     287113 5709645 1025.1
## - high_risk_100k  1   573103 5996635 1029.6
##
## Step:  AIC=1020.37
## Total_deaths ~ AverageDeathAge + vaccine_100k + high_risk_100k +
##     M_F_diff_100k
##
##                Df Sum of Sq    RSS    AIC
## - M_F_diff_100k 1     41630 5465261 1019.3
## - AverageDeathAge  1  118438 5541969 1020.5
## <none>                      5423631 1020.4
## - vaccine_100k  1     280378 5704009 1023.2
## - high_risk_100k  1   565217 5988848 1027.7
##
## Step:  AIC=1019.28
## Total_deaths ~ AverageDeathAge + vaccine_100k + high_risk_100k
##
##                Df Sum of Sq    RSS    AIC
## <none>                      5465261 1019.3
## - AverageDeathAge  1  210025 5675286 1020.8
## - vaccine_100k  1     259976 5725237 1021.4
## - high_risk_100k  1   587519 6052779 1026.7
```

```
##
## Call:
## lm(formula = Total_deaths ~ AverageDeathAge + vaccine_100k +
##     high_risk_100k, data = Combined)
##
## Coefficients:
##     (Intercept)  AverageDeathAge     vaccine_100k  high_risk_100k
##       2.145e+03       -2.134e+01        7.319e-03       -4.131e-02
```

```
##Best Subset (Both validated same model)
bw<-bestsubh(Combined[,i:5], Combined$Total_deaths, num = 1)
bw
```

```
##   p L 2 S 4 5    SSEp         r2      r2.adj       Cp     AICp       SBCp  PRESEp
## 1 2 2 0 0 0 1 5644542 0.3392075 0.1609876 5.182638 1021.744 1026.808 4475488
## 2 3 0 1 0 1 0 5675286 0.3369914 0.1789463 4.174899 1020.746 1028.311 4333743
## 3 4 1 1 0 1 0 5465261 0.2267083 0.2003461 2.837799 1019.276 1029.363 4324010
## 4 5 1 1 0 1 1 5423631 0.2325986 0.1973158 4.176342 1020.573 1033.182 4392298
## 5 6 1 1 1 1 1 5412533 0.2341969 0.1996429 6.000000 1022.386 1037.515 4447246
```

```
##Best Model
modelbest<-lm(Total_deaths~AverageDeathAge+vaccine_100k+high_risk_100k, Combined)
summary(modelbest)
```

```
##
## Call:
## lm(formula = Total_deaths ~ AverageDeathAge + vaccine_100k +
##     high_risk_100k, data = Combined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151.96 -110.40  -41.45   39.92 1664.39
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.145e+03  9.524e+02   2.317  0.01366 *
## AverageDeathAge -2.153e+01  1.171e+01  -1.839  0.06929 .
## vaccine_100k  7.319e-03  2.441e-03   2.010  0.04747 *
## high_risk_100k -4.131e-02  1.346e-02  -3.074  0.00281 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 249.2 on 88 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2003
## F-statistic:  8.6 on 3 and 88 DF,  p-value: 4.553e-05
```

```
anova(modelbest)
```

```
## Analysis of Variance Table
##
## Response: Total_deaths
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## AverageDeathAge  1  287624  287624  6.4312 0.0141332 *
## vaccine_100k     1  727625  727625 16.7160 0.0009424 ***
## high_risk_100k   1  587519  587519  9.4520 0.0028089 **
## Residuals       88 5465261   62105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
###BP test on best model (No violations)
CombinedGroup=cut_number(CombinedVaccine_100k,5)
CombinedBresidual<-modelbestbresidual
bf.test(residual~group, Combined)
```

```
##
##  Brown-Forsythe Test (alpha = 0.05)
## -------------------------------------------------------
##  data: residual and group
##
##  statistic  : 0.4962466
##  num df     : 4
##  denom df   : 38.20202
##  p.value    : 0.7385306
##
##  Result     : Difference is not statistically significant.
## -------------------------------------------------------
```

```
###Shapiro test best model (Violates normality)
data.model<-residuals(modelbest)
shapiro.test(data.model)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data.model
## W = 0.62979, p-value = 6.878e-14
```

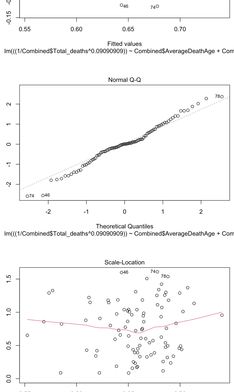
```
###Transform Y due to violation on normality
modelbesttrans=lm((1/Combined$Total_deaths^0.09090909)~Combined$AverageDeathAge+CombinedVaccine_100k+Combined
$high_risk_100k)
###Check assumptions on violations
CombinedBresidual<-modelbesttransbresidual
bf.test(residual~group, Combined)
```

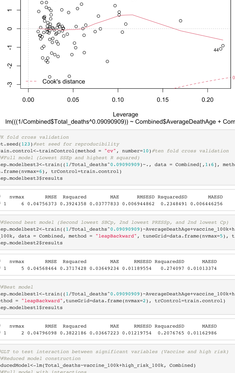
```
##
##  Brown-Forsythe Test (alpha = 0.05)
## -------------------------------------------------------
##  data: residual and group
##
##  statistic  : 0.7658535
##  num df     : 4
##  denom df   : 71.05246
##  p.value    : 0.5559905
##
##  Result     : Difference is not statistically significant.
## -------------------------------------------------------
```

```
data.model<-residuals(modelbesttrans)
shapiro.test(data.model)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data.model
## W = 0.98823, p-value = 0.4484
```

```
###Explain transformed model
plot(modelbesttrans)
```






```
### 5fold cross validation
set.seed(123)#set seed for reproducibility
train.control=trainControl(method = "cv", number=10)#ten fold cross validation
###Full model (lowest SSEp and highest R squared)
step.modelbest3<-train(((1/Total_deaths^0.09090909))~., data = Combined[,i:6], method = "leapBackward", tuneGrid=da
ta.frame(nvmax=5), trControl=train.control)
step.modelbest3$results
```

```
##  nvmax     RMSE  Rsquared ~ MAE     RMSESD RsquaredSD     MAESD
## 1     5 0.04756373 0.3924358 0.03377833 0.006944862  0.2348491 0.006446256
```

```
###Second best model (second lowest SSEp, 2nd lowest PRESSp, and 2nd lowest Cp)
step.modelbest2<-train(((1/Total_deaths^0.09090909))~AverageDeathAge+vaccine_100k+high_risk_100k+cases_100k+M_F_dif
f_100k, data = Combined, method = "leapBackward", tuneGrid=data.frame(nvmax=5), trControl=train.control)
step.modelbest2$results
```

```
##Best model
step.modelbest1<-train(((1/Total_deaths^0.09090909))~AverageDeathAge+vaccine_100k+high_risk_100k, data = Combined,
method = "leapBackward", tuneGrid=data.frame(nvmax=2), trControl=train.control)
step.modelbest1$results
```

```
##  nvmax     RMSE  Rsquared      MAE   RMSESD RsquaredSD     MAESD
## 1     2 0.04796698 0.3822186 0.03667223 0.01219754  0.2076765 0.01162986
```

```
###SLS to test interaction between significant variables (Vaccine and high risk)
###Reduced model constructing
reducedModel<-lm(Total_deaths~vaccine_100k+high_risk_100k, Combined)
###Full model with interactions
fullModel=lm(Total_deaths~vaccine_100k+high_risk_100k+(vaccine_100k*high_risk_100k), Combined)
anova(reducedModel, fullModel)
```

```
## Analysis of Variance Table
##
## Model 1: Total_deaths ~ vaccine_100k + high_risk_100k
## Model 2: Total_deaths ~ vaccine_100k + high_risk_100k + (vaccine_100k *
##     high_risk_100k)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     89 5675286
## 2     88 5667625  1    7650.9 0.1188 0.7312
```

```
###No significant interaction
```