Homework 10: ISYE 6501 - Introduction to Analytics Modeling

Question 14.1

Prompt

The breast cancer data set `breast-cancer-wisconsin.data.txt` from http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
(description at http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29) has missing values.

1. Use the mean/mode imputation method to impute values for the missing data.
2. Use regression to impute values for the missing data.
3. Use regression with perturbation to impute values for the missing data.
4. (Optional) Compare the results and quality of classification models (e.g., SVM, KNN) build using
   (1) the data sets from questions 1,2,3;
   (2) the data that remains after data points with missing values are removed; and
   (3) the data set when a binary variable is introduced to indicate missing values.

Solution

The data provided is the Diagnostic Wisconsin Breast Cancer Database provided by the UCI Machine Learning Repository [1].

- `Sample_code_number`: (ID, categorical)
- `Clump_thickness`: (feature, integer)
- `Uniformity of cell`: size (feature, integer)
- `Uniformity of cell`: shape (feature, integer)
- `Marginal adhesion`: (feature, integer)
- `Single epithelial cell size`: (feature, integer)
- `Bare nucleai`: (feature, integer)
- `Bland chromatin`: (feature, integer)
- `Normal nucleoli`: (feature, integer)
- `Mitoses`: (feature, integer)
- `Class` (Target) Binary 2 = benign 4 = malignant

Load Data Set

The first step is to load the data. A few packages are loaded for basic data wrangling. A seed is set so the results are reproducible. The working directory is defined as the `HW10` folder so that the data can be loaded for this assignment. The provided data from `breast-cancer-wisconsin.data.txt` is loaded into a table and the first few rows are printed to confirm the data has loaded properly.

```r
# Load packages
library(dplyr)
library(tidyr)

# Set seed so results are reproducible
set.seed(123)

# Set the working directory
setwd("~/projects/ISYE6501/HW10")

# Load the Wisconsin breast cancer database (original)
data <- read.table("data/breast-cancer-wisconsin.data.txt",
    stringsAsFactors = FALSE, header = FALSE, sep = ","
)

cat(paste("\nWisconsin Breaast Cancer Data:\n"))
```

```
##
## Wisconsin Breaast Cancer Data:
```

```r
head(data, 5)
```

```
##        V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11
## 1 1000025  5  1  1  1  2  1  3  1   1   2
## 2 1002945  5  4  4  5  7 10  3  2   1   2
## 3 1015425  3  1  1  1  2  2  3  1   1   2
## 4 1016277  6  8  8  1  3  4  3  7   1   2
## 5 1017023  4  1  1  3  2  1  3  1   1   2
```

Inspect the Data

After inspecting the data, there are a few rows initially spotted that do not contain integers and have values of `?`. To determine the number of rows with missing values in data the first approach is to look at the sum of all rows that contain `?` in each column. To confirm these findings, the second approach is to print the rows that have values with `?`. Since we can see that the only `?` values are in the `V7` column, we can look for all rows that contain the value and print each row. If you count the rows, there are 16, which matches the first approach that shows `V7` has 16 rows with `?`.

```r
# Inspect the data (contains "?" values)
# Counting missing values that are "?" in each column
missing_values_count <- sapply(data, function(x) sum(x == "?", na.rm = TRUE))
cat(paste("\nNumber of Missing Values by Column:\n"))
```

```
##
## Number of Missing Values by Column:
```

```r
print(missing_values_count)
```

```
##  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11
##   0   0   0   0   0   0  16   0   0   0   0
```

```r
# To confirm the missin values in V7 print the missing value rows (16)
cat(paste("\nRows with Missing Values:\n"))
```

```
##
## Rows with Missing Values:
```

```r
print(data[which(data$V7 == "?"), ])
```

```
##          V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11
## 24   1057013  8  4  5  1  2  ?  7  3   1   4
## 41   1096800  6  6  6  9  6  ?  7  8   1   2
## 140 1183246  1  1  1  1  1  ?  2  1   1   2
## 146 1184840  1  1  3  1  2  ?  2  1   1   2
## 159 1193683  1  1  2  1  3  ?  1  1   1   2
## 165 1197510  5  1  1  1  2  ?  3  1   1   2
## 236 1241232  3  1  4  1  2  ?  3  1   1   2
## 250  169356  3  1  1  1  2  ?  3  1   1   2
## 276  432809  3  1  3  1  2  ?  2  1   1   2
## 293  563649  8  8  8  1  2  ?  6 10   1   4
## 295  606140  1  1  1  1  2  ?  2  1   1   2
## 298   61634  5  4  3  1  2  ?  2  3   1   2
## 316  704168  4  6  5  6  7  ?  4  9   1   2
## 322  733639  3  1  1  1  2  ?  3  1   1   2
## 412 1238464  1  1  1  1  1  ?  2  1   1   2
## 618 1057067  1  1  1  1  1  ?  1  1   1   2
```

Inspecting the data with missing values, we must consider two things.

The first is bias. It does not seem there is bias occurring when `V7` has a missing value. None of the features have all the same values and leads to my assumption that the data does not indicate a bias.

The second is amount of missing data. The number of observations is 699 and there are 16 missing values. This would be about 2% of all observations missing data, which is assumed to be acceptable to perform data imputation.

```r
missing_percentage <- (nrow(data[which(data$V7 == "?"),]) / nrow(data)) * 100
print(missing_percentage)
```

```
## [1] 2.288984
```

Finally, to wrap it up, we determine the location of the missing values prior to resolving the issue.

```r
# Store location of missing values
missing <- which(data$V7 == "?", arr.ind = TRUE)
missing
```

```
##  [1]  24  41 140 146 159 165 236 250 276 293 295 298 316 322 412 618
```

Part A: Mean/Mode Imputation

The 9 features (`V2` through `V10`) in the data set have values between 1 and 10. This means that `V7` is a categorical variable with encoding to use with numerical values. Missing values in categorical variables can be imputed using the **mode**.

```
# Function to find the mode for a vector, v
get_mode <- function(v) {
    # Store only the unique values from the vector
    uniqv <- unique(v)
    # Find mode with which.max() using the tabulated counts of unique values
    uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Impute missing values with duplicate data set
data_imputed <- data
# Replace "?" with NA in the entire dataset
data_imputed[data_imputed == "?"] <- NA

# Calculate and print the mode for the V7column
mode_V7 <- get_mode(data_imputed$V7)
cat("V7 Mode after imputation:", mode_V7, "\n")
```

```
## V7 Mode after imputation: 1
```

A mode value of 1 means that 1 is the most common value in `V7`, appearing more frequently than any value in the range of 1 to 10.

The final step is to use data imputation to replace rows with missing data for `V7` with values from `mode_V7`.

```
data_imputed[data_imputed == "NA"] <- mode_V7
data_imputed$V7 <- as.integer(data_imputed$V7)
```

Part B: Regression Imputation

To prepare the data for regression imputation, the data must be modified. The data set should only include features, or V2 to V10 (`2:10`). We must also consider removing the rows where `V7` has missing values.

Next, an initial linear model is developed to predict `V7` using all potential predictors V2 to V10.e

```
# Prepare modified data without response variable or ID
# removing rows where there are missing values
data_modified <- data[-missing, 2:10]
# Discrete response variable
data_modified$V7 <- as.integer(data_modified$V7)

# Build initial linear model
initial_model <- lm(V7 ~ V2 + V3 + V4 + V5 + V6 + V8 + V9 + V10, data = data_modified)
summary(initial_model)
```

```
##
## Call:
```

```
## lm(formula = V7 ~ V2 + V3 + V4 + V5 + V6 + V8 + V9 + V10, data = data_modified)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7316 -0.9426 -0.3002  0.6725  8.6998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.616652   0.194975  -3.163  0.00163 **
## V2           0.230156   0.041691   5.521 4.83e-08 ***
## V3          -0.067980   0.076170  -0.892  0.37246
## V4           0.340442   0.073420   4.637 4.25e-06 ***
## V5           0.339705   0.045919   7.398 4.13e-13 ***
## V6           0.090392   0.062541   1.445  0.14883
## V8           0.320577   0.059047   5.429 7.91e-08 ***
## V9           0.007293   0.044486   0.164  0.86983
## V10         -0.075230   0.059331  -1.268  0.20524
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.274 on 674 degrees of freedom
## Multiple R-squared:  0.615,  Adjusted R-squared:  0.6104
## F-statistic: 134.6 on 8 and 674 DF,  p-value: < 2.2e-16
```

After reviewing the initial model, a backward stepwise regression approach is implemented. Using `step()` refines the model and allows for selection of significant predictors.

```
# Use backward stepwise regression for variable selection
model_selected <- step(initial_model)
```

```
## Start:  AIC=1131.43
## V7 ~ V2 + V3 + V4 + V5 + V6 + V8 + V9 + V10
##
##        Df Sum of Sq    RSS    AIC
## - V9    1     0.139 3486.8 1129.5
## - V3    1     4.120 3490.8 1130.2
## - V10   1     8.317 3495.0 1131.0
## <none>              3486.6 1131.4
## - V6    1    10.806 3497.5 1131.5
## - V4    1   111.227 3597.9 1150.9
## - V8    1   152.482 3639.1 1158.7
## - V2    1   157.657 3644.3 1159.6
## - V5    1   283.119 3769.8 1182.8
##
## Step:  AIC=1129.45
## V7 ~ V2 + V3 + V4 + V5 + V6 + V8 + V10
##
##        Df Sum of Sq    RSS    AIC
## - V3    1     4.028 3490.8 1128.2
## - V10   1     8.179 3495.0 1129.0
## <none>              3486.8 1129.5
## - V6    1    11.211 3498.0 1129.7
## - V4    1   114.768 3601.6 1149.6
```

```
## - V2    1   158.696 3645.5 1157.8
## - V8    1   160.776 3647.6 1158.2
## - V5    1   285.902 3772.7 1181.3
##
## Step:  AIC=1128.24
## V7 ~ V2 + V4 + V5 + V6 + V8 + V10
##
##        Df Sum of Sq    RSS    AIC
## - V6    1     8.606 3499.4 1127.9
## - V10   1     8.889 3499.7 1128.0
## <none>              3490.8 1128.2
## - V4    1   153.078 3643.9 1155.6
## - V2    1   155.308 3646.1 1156.0
## - V8    1   157.123 3647.9 1156.3
## - V5    1   282.133 3772.9 1179.3
##
## Step:  AIC=1127.92
## V7 ~ V2 + V4 + V5 + V8 + V10
##
##        Df Sum of Sq    RSS    AIC
## - V10   1     5.562 3505.0 1127.0
## <none>              3499.4 1127.9
## - V2    1   159.594 3659.0 1156.4
## - V8    1   169.954 3669.4 1158.3
## - V4    1   206.785 3706.2 1165.1
## - V5    1   295.807 3795.2 1181.3
##
## Step:  AIC=1127.01
## V7 ~ V2 + V4 + V5 + V8
##
##        Df Sum of Sq    RSS    AIC
## <none>              3505.0 1127.0
## - V2    1    155.70 3660.7 1154.7
## - V8    1    172.42 3677.4 1157.8
## - V4    1    201.22 3706.2 1163.1
## - V5    1    290.68 3795.7 1179.4
```

Last, the final model is built by using only a few select predictors. The selected predictors are `V2`, `V4`, `V5`, and `V8`.

```
# Build refined model based on selected variables
final_model <- lm(V7 ~ V2 + V4 + V5 + V8, data = data_modified)
summary(final_model)
```

```
##
## Call:
## lm(formula = V7 ~ V2 + V4 + V5 + V8, data = data_modified)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8115 -0.9531 -0.3111  0.6678  8.6889
##
## Coefficients:
```

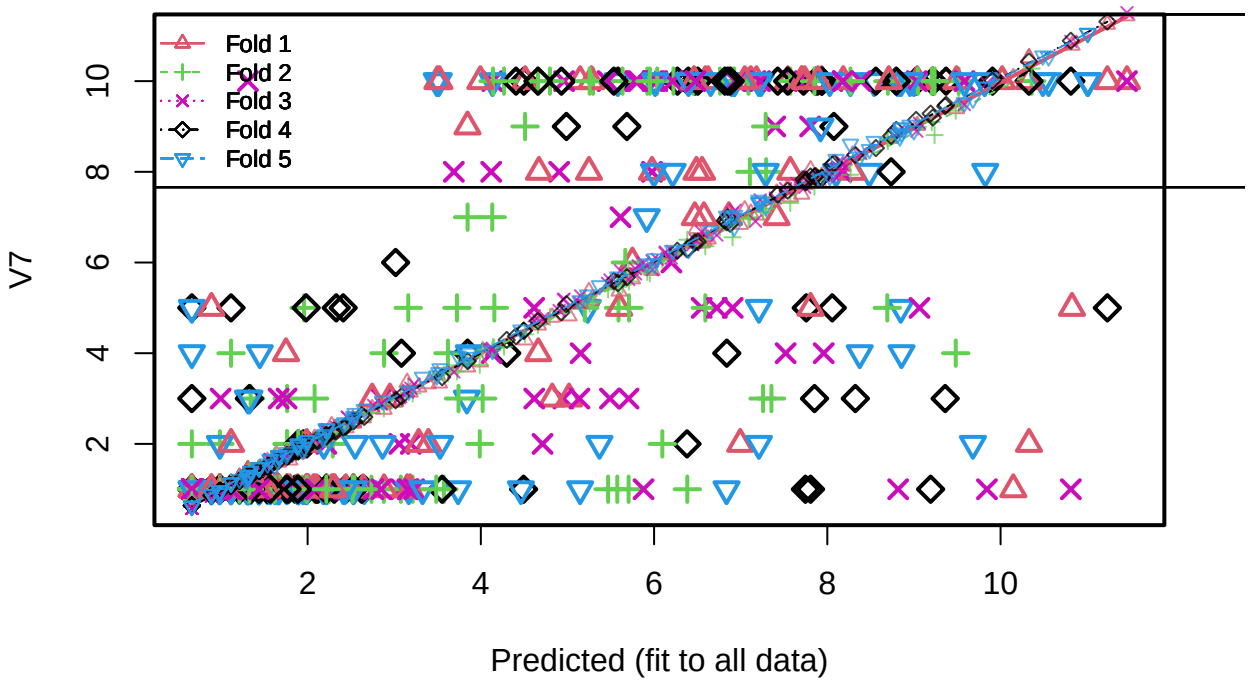
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.53601    0.17514  -3.060   0.0023 **
## V2           0.22617    0.04121   5.488 5.75e-08 ***
## V4           0.31729    0.05086   6.239 7.76e-10 ***
## V5           0.33227    0.04431   7.499 2.03e-13 ***
## V8           0.32378    0.05606   5.775 1.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.274 on 678 degrees of freedom
## Multiple R-squared:  0.6129, Adjusted R-squared:  0.6107
## F-statistic: 268.4 on 4 and 678 DF,  p-value: < 2.2e-16
```

The DAAG package is loaded consider cross-validation. Using 5-fold cross-validation, the R-Squared value is obtained.

```
# Load packages
library(DAAG)

# cv model 5-fold
model_cv <- cv.lm(data_modified, final_model, m = 5)
```

```
## Warning in cv.lm(data_modified, final_model, m = 5):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
##
## fold 1
## Observations in test set: 136
##                      4         6        12        21        22        29
## Predicted    4.6629181 10.018398  1.213452  6.623331 6.5754051  1.213452
## cvpred       4.6191932 10.058192  1.255391  6.517183 6.4654585  1.255391
## V7           4.0000000 10.000000  1.000000 10.000000 7.0000000  1.000000
## CV residual -0.6191932 -0.058192 -0.255391  3.482817 0.5345415 -0.255391
##                     36        37        51        55        56        57
## Predicted    1.213452 10.146951  5.017641 7.5722028  5.5975534 5.7498291
## cvpred       1.255391 10.074662  4.821406 7.4469069  5.3690859 5.6380545
## V7           1.000000  1.000000  3.000000 8.0000000  5.0000000 6.0000000
## CV residual -0.255391 -9.074662 -1.821406 0.5530931 -0.3690859 0.3619455
##                     59        64        67        74        80        81
## Predicted    3.497886  3.393772  1.9895750  7.703744  1.213452  3.149661
## cvpred       3.525040  3.324441  1.9818912  7.504717  1.255391  3.354155
## V7          10.000000  2.000000  1.0000000 10.000000  1.000000  1.000000
## CV residual  6.474960 -1.324441 -0.9818912  2.495283 -0.255391 -2.354155
##                     82        87         89         91        104          109
## Predicted    1.9980563 4.670420  1.9895750  1.3110677  4.823962  0.98728319
## cvpred       1.9540254 4.751053  1.9818912  1.4246636  4.792298  1.06964846
## V7           1.0000000 8.000000  1.0000000  1.0000000  3.000000  1.00000000
## CV residual -0.9540254 3.248947 -0.9818912 -0.4246636 -1.792298 -0.06964846
##                    118       123        125       127       129       130
## Predicted    7.544252  7.182002  7.4146942  6.591389  4.512626 0.6634986
## cvpred       7.641050  7.046719  7.5240472  6.569453  4.291032 0.7146333
```

```
## V7          10.000000 10.000000  7.0000000 10.000000 10.000000 1.0000000
## CV residual  2.358950  2.953281 -0.5240472  3.430547  5.708968 0.2853667
##                   131         138       145       156       160        166
## Predicted    2.526532  1.11583682  1.213452  5.146219  7.935668 1.98957497
## cvpred       2.459995  1.08611828  1.255391  5.098134  7.898157 1.98189115
## V7           1.000000  1.00000000  1.000000 10.000000 10.000000 2.00000000
## CV residual -1.459995 -0.08611828 -0.255391  4.901866  2.101843 0.01810885
##                   169         173        179      187        194        195
## Predicted    1.7634059  0.98728319  1.9895750 6.488540  1.3110677  1.7634059
## cvpred       1.7961486  1.06964846  1.9818912 6.746289  1.4246636  1.7961486
## V7           1.0000000  1.00000000  1.0000000 8.000000  1.0000000  1.0000000
## CV residual -0.7961486 -0.06964846 -0.9818912 1.253711 -0.4246636 -0.7961486
##                   197        200       202       210       215        221
## Predicted   6.4703123  1.4396214  8.183315  2.215744 11.459052  1.6433336
## cvpred      6.4245843  1.4411335  8.118688  2.167634 11.438991  1.7518131
## V7          7.0000000  1.0000000 10.000000  1.000000 10.000000  1.0000000
## CV residual 0.5754157 -0.4411335  1.881312 -1.167634 -1.438991 -0.7518131
##                   226       231       238      248      255        259
## Predicted    0.98728319 6.8637366  6.994273 3.846111 5.975998  1.7634059
## cvpred       1.06964846 6.8827567  6.840034 3.695926 5.823797  1.7961486
## V7           1.00000000 7.0000000  2.000000 9.000000 8.000000  1.0000000
## CV residual -0.06964846 0.1172433 -4.840034 5.304074 2.176203 -0.7961486
##                   272        275        281        302        303        304
## Predicted    2.215744  1.7634059  1.7634059  1.3110677  9.4909008  1.3110677
## cvpred       2.167634  1.7961486  1.7961486  1.4246636  9.3924971  1.4246636
## V7           1.000000  1.0000000  1.0000000  1.0000000 10.0000000  1.0000000
## CV residual -1.167634 -0.7961486 -0.7961486 -0.4246636  0.6075029 -0.4246636
##                   318       324       327      331       341      350
## Predicted    8.2904156  7.116354  3.994159 6.553928  5.276300 5.248350
## cvpred       8.3606274  7.085435  3.806377 6.430670  5.163783 5.357926
## V7           8.0000000 10.000000 10.000000 8.000000 10.000000 8.000000
## CV residual -0.3606274  2.914565  6.193623 1.569330  4.836217 2.642074
##                   354        358       364       376         377         378
## Predicted    7.706992  9.0385627 2.9499156 0.6634986  0.98728319  0.98728319
## cvpred       7.822960  9.0210121 2.9250899 0.7146333  1.06964846  1.06964846
## V7          10.000000 10.0000000 3.0000000 1.0000000  1.00000000  1.00000000
## CV residual  2.177040  0.9789879 0.0749101 0.2853667 -0.06964846 -0.06964846
##                   381       388        397       403         408       414
## Predicted   0.6634986  3.182583  1.7634059  2.526532  0.98728319  2.533030
## cvpred      0.7146333  3.142159  1.7961486  2.459995  1.06964846  2.491322
## V7          1.0000000  1.000000  1.0000000  1.000000  1.00000000  1.000000
## CV residual 0.2853667 -2.142159 -0.7961486 -1.459995 -0.06964846 -1.491322
##                   417         418       421       426         429       439
## Predicted    7.739458  0.98728319 2.7452241 11.232883  0.98728319  2.641111
## cvpred       7.719337  1.06964846 2.7741358 11.253249  1.06964846  2.573536
## V7          10.000000  1.00000000 3.0000000 10.000000  1.00000000  1.000000
## CV residual  2.280663 -0.06964846 0.2258642 -1.253249 -0.06964846 -1.573536
##                   447        455        459        473        477        482
## Predicted   0.6634986 0.88966773  1.8854614  1.7943441  1.6592923  2.882259
## cvpred      0.7146333 0.90037577  1.7812914  1.6433458  1.5955489  2.762740
## V7          1.0000000 1.00000000  1.0000000  1.0000000  1.0000000  1.000000
## CV residual 0.2853667 0.09962423 -0.7812914 -0.6433458 -0.5955489 -1.762740
##                   492        498       500       501       502       511
## Predicted    7.045971  1.3420059  1.665790  2.441913  1.665790 0.6634986
```

```
## cvpred       7.158648  1.2718608  1.626876  2.353376  1.626876 0.7146333
## V7          10.000000  1.0000000  1.000000  1.000000  1.000000 1.0000000
## CV residual  2.841352 -0.2718608 -0.626876 -1.353376 -0.626876 0.2853667
##                   524        528        533       536       537       539
## Predicted    7.892476  1.9895750  1.3110677  2.277906  2.215744  1.665790
## cvpred       7.811935  1.9818912  1.4246636  2.399189  2.167634  1.626876
## V7          10.000000  1.0000000  1.0000000  1.000000  1.000000  1.000000
## CV residual  2.188065 -0.9818912 -0.4246636 -1.399189 -1.167634 -0.626876
##                   543       544        551        552       556        563
## Predicted    1.5681750  1.665790  1.4396214  1.3110677  2.313360  1.3110677
## cvpred       1.4576033  1.626876  1.4411335  1.4246636  2.336906  1.4246636
## V7           1.0000000  1.000000  1.0000000  1.0000000  1.000000  1.0000000
## CV residual -0.4576033 -0.626876 -0.4411335 -0.4246636 -1.336906 -0.4246636
##                   566       567       569       570       571        580
## Predicted   10.3282067  2.080692  3.522326 10.824479  6.461831  1.3110677
## cvpred      10.5102786  2.119837  3.340910 10.791615  6.452450  1.4246636
## V7          10.0000000  1.000000 10.000000  5.000000 10.000000  1.0000000
## CV residual -0.5102786 -1.119837  6.659090 -5.791615  3.547550 -0.4246636
##                   590        597       616        617      626       628
## Predicted    1.5681750  1.9830769  2.300363  1.4396214 1.750410 0.8896677
## cvpred       1.4576033  1.9505641  2.274252  1.4411335 1.733495 0.9003758
## V7           1.0000000  1.0000000  1.000000  1.0000000 4.000000 5.0000000
## CV residual -0.4576033 -0.9505641 -1.274252 -0.4411335 2.266505 4.0996242
##                   636       638        641        642       649        663
## Predicted    2.067696  3.282182  2.0065377  1.4396214 10.328207  1.6218560
## cvpred       2.057183  3.252239  1.9261597  1.4411335 10.510279  1.7170248
## V7           1.000000  2.000000  1.0000000  1.0000000  2.000000  1.0000000
## CV residual -1.057183 -1.252239 -0.9261597 -0.4411335 -8.510279 -0.7170248
##                   665        673       676        677       679       682
## Predicted    2.104153  1.5372368  2.293865  1.3045696 0.6634986  8.709284
## cvpred       2.095432  1.6104061  2.242925  1.3933366 0.7146333  8.735203
## V7           1.000000  1.0000000  1.000000  1.0000000 1.0000000 10.000000
## CV residual -1.095432 -0.6104061 -1.242925 -0.3933366 0.2853667  1.264797
##                   688       695        696       699
## Predicted    1.4396214 1.1158368 0.88966773  7.806135
## cvpred       1.4411335 1.0861183 0.90037577  8.041412
## V7           1.0000000 2.0000000 1.00000000  5.000000
## CV residual -0.4411335 0.9138817 0.09962423 -3.041412
##
## Sum of squares = 675.12    Mean square = 4.96    n = 136
##
## fold 2
## Observations in test set: 137
##                     3        16         17       26         27       40
## Predicted   1.7634059  5.575097  1.6657904 3.847115  1.4396214 4.131480
## cvpred      1.7311017  5.539097  1.6353391 3.810260  1.4395826 4.058919
## V7          2.0000000  1.000000  1.0000000 7.000000  1.0000000 7.000000
## CV residual 0.2688983 -4.539097 -0.6353391 3.189740 -0.4395826 2.941081
##                   53       54        62        66        73        77       79
## Predicted    5.589072 7.105889 0.9872832  3.987660  3.573045  1.939142 1.763406
## cvpred       5.466072 6.994128 1.0480697  3.741518  3.533016  2.093875 1.731102
## V7           5.000000 8.000000 2.0000000  2.000000  1.000000  1.000000 3.000000
## CV residual -0.466072 1.005872 0.9519303 -1.741518 -2.533016 -1.093875 1.268898
##                   83       85         92         93        99      102
```

```
## Predicted    2.215744 7.288124  1.4481027  1.9895750 5.6671932 3.162109
## cvpred       2.122615 7.299818  1.4782107  1.9268582 5.7330826 3.280073
## V7           1.000000 9.000000  1.0000000  1.0000000 6.0000000 5.000000
## CV residual -1.122615 1.700182 -0.4782107 -0.9268582 0.2669174 1.719927
##                  111      112       115       120        139        141
## Predicted    2.290903 4.510642 2.0806923 1.7634059  1.9830769  1.1158368
## cvpred       2.252774 4.415984 2.0797034 1.7311017  1.9839408  1.1480635
## V7           2.000000 9.000000 3.0000000 2.0000000  1.0000000  1.0000000
## CV residual -0.252774 4.584016 0.9202966 0.2688983 -0.9839408 -0.1480635
##                  143       150       152       154        172        176
## Predicted    5.710148  7.089906  4.267511 1.342006  1.3110677  6.760627
## cvpred       5.582008  6.958012  4.142971 1.343820  1.3395887  6.778146
## V7           5.000000 10.000000 10.000000 3.000000  1.0000000 10.000000
## CV residual -0.582008  3.041988  5.857029 1.656180 -0.3395887  3.221854
##                  178       182       183       198       199       204
## Predicted    6.383423 0.6634986  2.441913  3.212542 0.6634986  2.215744
## cvpred       6.503363 0.7565506  2.318371  3.113056 0.7565506  2.122615
## V7           1.000000 1.0000000  1.000000  1.000000 1.0000000  1.000000
## CV residual -5.503363 0.2434494 -1.318371 -2.113056 0.2434494 -1.122615
##                  208        211        218       220       222        225
## Predicted    1.3110677 10.3591449  1.3110677  3.076486  6.905688  7.572203
## cvpred       1.3395887 10.2762193  1.3395887  3.015575  6.556611  7.312616
## V7           1.0000000 10.0000000  1.0000000  1.000000 10.000000 10.000000
## CV residual -0.3395887 -0.2762193 -0.3395887 -2.015575  3.443389  2.687384
##                  232       233       235       237       242        243
## Predicted   7.2936180  5.469000  2.080692  6.223645  2.427938  1.5372368
## cvpred      7.1881655  5.404706  2.079703  6.163275  2.391396  1.5353452
## V7          8.0000000  1.000000  1.000000 10.000000  1.000000  1.0000000
## CV residual 0.8118345 -4.404706 -1.079703  3.836725 -1.391396 -0.5353452
##                  244       254        261       265       267       268
## Predicted   1.958637  6.924895  9.2055566  7.257904  5.921338  4.659956
## cvpred      1.922627  6.952010  9.0959715  7.114295  5.796219  4.621732
## V7          5.000000 10.000000 10.0000000  3.000000 10.000000 10.000000
## CV residual 3.077373  3.047990  0.9040285 -4.114295  4.203781  5.378268
##                  271       274        279        282        287        288
## Predicted    4.797970 3.6199415  1.3110677  1.8695027  9.5163449  1.4396214
## cvpred       4.801495 3.5599328  1.3395887  1.8654923  9.5016559  1.4395826
## V7          10.000000 4.0000000  1.0000000  1.0000000 10.0000000  1.0000000
## CV residual  5.198505 0.4400672 -0.3395887 -0.8654923  0.4983441 -0.4395826
##                  289       290       291        292       294       297
## Predicted   3.724055  6.451822 0.6634986  1.3110677  5.635994 4.1559197
## cvpred      3.598613  6.468172 0.7565506  1.3395887  5.506419 4.1202332
## V7          5.000000 10.000000 1.0000000  1.0000000 10.000000 5.0000000
## CV residual 1.401387  3.531828 0.2434494 -0.3395887  4.493581 0.8797668
##                  299       306       312       317        319        323
## Predicted    2.246682  5.295246 0.6634986  4.140940  1.3110677  1.7634059
## cvpred       2.126846  5.137644 0.7565506  4.138688  1.3395887  1.7311017
## V7           1.000000 10.000000 1.0000000 10.000000  1.0000000  1.0000000
## CV residual -1.126846  4.862356 0.2434494  5.861312 -0.3395887 -0.7311017
##                  328       333       344       349       351       353
## Predicted    0.98728319  2.541512 0.6634986  5.705658  2.232707  4.020868
## cvpred       1.04806967  2.509844 0.7565506  5.748231  2.199871  4.077322
## V7           1.00000000  1.000000 1.0000000  1.000000  1.000000  3.000000
## CV residual -0.04806967 -1.509844 0.2434494 -4.748231 -1.199871 -1.077322
```

```
##                   362        366       368        371       374        395
## Predicted    6.035941  1.2134523  9.046065  1.9830769  2.224225  1.6218560
## cvpred       5.982667  1.2438261  8.991746  1.9839408  2.161243  1.7452732
## V7          10.000000  1.0000000 10.000000  1.0000000  1.000000  1.0000000
## CV residual  4.017333 -0.2438261  1.008254 -0.9839408 -1.161243 -0.7452732
##                   398       404       413       416       428       432
## Predicted    1.342006 1.115837  9.482419  3.742022  6.095091  2.880276
## cvpred       1.343820 1.148064  9.347144  3.730439  6.063282  2.782909
## V7           1.000000 4.000000  4.000000  3.000000  2.000000  1.000000
## CV residual -0.343820 2.851936 -5.347144 -0.730439 -4.063282 -1.782909
##                   433       441       442       444       450        452
## Predicted    1.8919595  9.238047 2.882259 0.6634986  8.698819  1.5681750
## cvpred       1.8310956  8.810558 2.878620 0.7565506  8.680054  1.5395765
## V7           1.0000000 10.000000 4.000000 2.0000000 10.000000  1.0000000
## CV residual -0.8310956  1.189442 1.121380 1.2434494  1.319946 -0.5395765
##                   454        471        475        481        495       505
## Predicted    7.851074  1.4396214  1.5681750  1.5681750  5.1996390 0.6634986
## cvpred       7.672929  1.4395826  1.5395765  1.5395765  5.1510214 0.7565506
## V7          10.000000  1.0000000  1.0000000  1.0000000  5.0000000 1.0000000
## CV residual  2.327071 -0.4395826 -0.5395765 -0.5395765 -0.1510214 0.2434494
##                    518        519        525       531       541        549
## Predicted    0.98728319  1.7653891  1.4396214  5.255827 1.8919595  1.1158368
## cvpred       1.04806967  1.8268124  1.4395826  5.094784 1.8310956  1.1480635
## V7           1.00000000  1.0000000  1.0000000 10.000000 2.0000000  1.0000000
## CV residual -0.04806967 -0.8268124 -0.4395826  4.905216 0.1689044 -0.1480635
##                   550       553       557       558        560        564
## Predicted    6.591389  2.736743  2.541512  2.232707  1.8919595  1.4396214
## cvpred       6.358290  2.701370  2.509844  2.199871  1.8310956  1.4395826
## V7           5.000000  1.000000  1.000000  1.000000  1.0000000  1.0000000
## CV residual -1.358290 -1.701370 -1.509844 -1.199871 -0.8310956 -0.4395826
##                    574        577         579       593       600        601
## Predicted    0.98728319  1.8919595  0.98728319  5.951297  2.520034  1.4396214
## cvpred       1.04806967  1.8310956  1.04806967  5.759310  2.585382  1.4395826
## V7           1.00000000  1.0000000  1.00000000 10.000000  1.000000  1.0000000
## CV residual -0.04806967 -0.8310956 -0.04806967  4.240690 -1.585382 -0.4395826
##                   607       611        615       624       633       640
## Predicted    1.6742718  8.266694  1.2134523 0.6634986 0.6634986  2.232707
## cvpred       1.6739671  7.936001  1.2438261 0.7565506 0.7565506  2.199871
## V7           1.0000000 10.000000  1.0000000 1.0000000 1.0000000  1.000000
## CV residual -0.6739671  2.063999 -0.2438261 0.2434494 0.2434494 -1.199871
##                   644       658        662        664       666       670
## Predicted   0.6634986  3.484890  1.9895750  1.6218560 0.6634986  8.692321
## cvpred      0.7565506  3.517022  1.9268582  1.7452732 0.7565506  8.737137
## V7          1.0000000  1.000000  1.0000000  1.0000000 1.0000000  5.000000
## CV residual 0.2434494 -2.517022 -0.9268582 -0.7452732 0.2434494 -3.737137
##                   683       685        691       697
## Predicted    2.215744 0.6634986  1.3280304  7.354776
## cvpred       2.122615 0.7565506  1.4168448  7.377920
## V7           1.000000 1.0000000  1.0000000  3.000000
## CV residual -1.122615 0.2434494 -0.4168448 -4.377920
##
## Sum of squares = 671.95    Mean square = 4.9    n = 137
##
## fold 3
```

```
## Observations in test set: 137
##                      7         10         11       15         23        30
## Predicted    1.311068  1.6657904  1.3110677 7.801359  1.4396214  1.298072
## cvpred       1.210609  1.7024006  1.2106087 7.903720  1.4329981  1.171744
## V7          10.000000  1.0000000  1.0000000 9.000000  1.0000000  1.000000
## CV residual  8.789391 -0.7024006 -0.2106087 1.096280 -0.4329981 -0.171744
##                     32        42        45        46       50        61
## Predicted    1.5372368  4.952516  8.817888 0.9872832 4.904067  5.137738
## cvpred       1.4800112  5.178445  8.885669 0.8941930 4.940760  5.119914
## V7           1.0000000  3.000000  1.000000 1.0000000 8.000000  3.000000
## CV residual -0.4800112 -2.178445 -7.885669 0.1058070 3.059240 -2.119914
##                     63        65       69        71        78        84
## Predicted   5.975998 0.9872832 7.398711  2.526532  2.224225  3.058544
## cvpred      6.038679 0.8941930 7.421061  2.565770  2.303640  3.015076
## V7          8.000000 1.0000000 9.000000  1.000000  1.000000  2.000000
## CV residual 1.961321 0.1058070 1.578939 -1.565770 -1.303640 -1.015076
##                      86        88         94         97       101       105
## Predicted    4.12596072  5.982521 0.9872832  1.2219336 4.6157352 10.811483
## cvpred       4.08955102  5.842760 0.8941930  1.1790165 4.8231646 10.876689
## V7           4.00000000 10.000000 1.0000000  1.0000000 5.0000000  1.000000
## CV residual -0.08955102  4.157240 0.1058070 -0.1790165 0.1768354 -9.876689
##                     106       108       124        136      147       149
## Predicted    4.616739  7.170035  4.132459  2.2157441 3.688602  3.075507
## cvpred       4.713484  6.908853  4.108983  2.2882188 3.585599  3.045918
## V7           3.000000 10.000000 10.000000  2.0000000 8.000000  1.000000
## CV residual -1.713484  3.091147  5.891017 -0.2882188 4.414401 -2.045918
##                     151       162       164       167       175       184
## Predicted    1.3110677  1.989575 0.9957645  6.445324  6.131548  8.057748
## cvpred       1.2106087  2.018816 0.9096140  6.392975  6.229116  7.955621
## V7           1.0000000  1.000000 3.0000000 10.000000 10.000000 10.000000
## CV residual -0.2106087 -1.018816 2.0903860  3.607025  3.770884  2.044379
##                     185       186       201       203       207       219
## Predicted    6.125050  1.5372368  7.648341  1.3110677  6.549413  7.958150
## cvpred       6.209684  1.4800112  7.740738  1.2106087  6.674771  7.912619
## V7          10.000000  1.0000000 10.000000  1.0000000  5.000000  4.000000
## CV residual  3.790316 -0.4800112  2.259262 -0.2106087 -1.674771 -3.912619
##                     240       241       245       249       260       269
## Predicted    4.960997  3.191064  1.3110677  1.989575 4.119463  7.518783
## cvpred       5.193866  3.229443  1.2106087  2.018816 4.070119  7.628030
## V7          10.000000  2.000000  1.0000000  1.000000 8.000000  4.000000
## CV residual  4.806134 -1.229443 -0.2106087 -1.018816 3.929881 -3.628030
##                     270       296       307       315       325       330
## Predicted    1.3110677  7.740462  1.3110677 0.9872832  1.3110677  7.224957
## cvpred       1.2106087  7.658638  1.2106087 0.8941930  1.2106087  7.381321
## V7           1.0000000 10.000000  1.0000000 1.0000000  1.0000000 10.000000
## CV residual -0.2106087  2.341362 -0.2106087 0.1058070 -0.2106087  2.618679
##                     334       335       337       340       346       357
## Predicted    5.787290  5.462502  6.036920  5.816245 0.6634986 2.8503169
## cvpred       5.748734  5.541999  6.068411  5.928122 0.5777774 2.8821854
## V7          10.000000 10.000000 10.000000 10.000000 1.0000000 3.0000000
## CV residual  4.251266  4.458001  3.931589  4.071878 0.4222226 0.1178146
##                     359       360       363       367       369       373
## Predicted    5.151713 5.612533 1.756908  9.5830222  1.298072  1.9830769
## cvpred       5.264448 5.818674 1.729981  9.4826633  1.171744  1.9993839
```

```
## V7           4.000000 7.000000 3.000000 10.0000000  1.000000  1.0000000
## CV residual -1.264448 1.181326 1.270019  0.5173367 -0.171744 -0.9993839
##                    375       380       382       384       385       387
## Predicted    1.7569078  3.167603  6.937630 0.8896677 0.8896677  6.214184
## cvpred       1.7299814  3.179169  7.154044 0.8471799 0.8471799  6.241276
## V7           1.0000000  1.000000 10.000000 1.0000000 1.0000000 10.000000
## CV residual -0.7299814 -2.179169  2.845956 0.1528201 0.1528201  3.758724
##                    389        391       415       427       437       446
## Predicted    1.2134523  1.3195491  5.784303  3.154607  5.878383 0.8896677
## cvpred       1.1635956  1.2260297  5.862427  3.140304  5.991665 0.8471799
## V7           1.0000000  1.0000000 10.000000  1.000000  1.000000 1.0000000
## CV residual -0.1635956 -0.2260297  4.137573 -2.140304 -4.991665 0.1528201
##                    448       449       451       461       463       464
## Predicted    1.5681750 0.6634986  2.330322  2.232707  2.458876  1.3420059
## cvpred       1.6553874 0.5777774  2.366074  2.319061  2.588463  1.3859849
## V7           1.0000000 1.0000000  1.000000  1.000000  1.000000  1.0000000
## CV residual -0.6553874 0.4222226 -1.366074 -1.319061 -1.588463 -0.3859849
##                    472        479       480       484        487       489
## Predicted    2.458876  1.5681750  8.178825  8.387027  1.4396214  5.705658
## cvpred       2.588463  1.6553874  8.052908  8.401150  1.4329981  5.626893
## V7           1.000000  1.0000000 10.000000 10.000000  1.0000000  3.000000
## CV residual -1.588463 -0.6553874  1.947092  1.598850 -0.4329981 -2.626893
##                    491       494       497       507        512        513
## Predicted   0.6634986  9.033069 0.6634986  9.064007  1.8919595  1.5681750
## cvpred      0.5777774  8.896845 0.5777774  9.072221  1.9718031  1.6553874
## V7          1.0000000 10.000000 1.0000000  5.000000  1.0000000  1.0000000
## CV residual 0.4222226  1.103155 0.4222226 -4.072221 -0.9718031 -0.6553874
##                    514       521        522       523       529        530
## Predicted    1.4396214 0.6634986  1.3420059  6.907671  2.761183  1.6657904
## cvpred       1.4329981 0.5777774  1.3859849  7.084337  2.850593  1.7024006
## V7           1.0000000 1.0000000  1.0000000  5.000000  1.000000  1.0000000
## CV residual -0.4329981 0.4222226 -0.3859849 -2.084337 -1.850593 -0.7024006
##                    534        542        546       548        555       561
## Predicted    1.4396214  1.1158368  1.8919595 0.8896677  1.1158368  2.215744
## cvpred       1.4329981  1.1165824  1.9718031 0.8471799  1.1165824  2.288219
## V7           1.0000000  1.0000000  1.0000000 1.0000000  1.0000000  1.000000
## CV residual -0.4329981 -0.1165824 -0.9718031 0.1528201 -0.1165824 -1.288219
##                    568       585        588       598        603       605
## Predicted   1.665790  3.229504  1.8919595  2.850317  1.6657904  6.477814
## cvpred      1.702401  3.314571  1.9718031  2.882185  1.7024006  6.490137
## V7          3.000000  1.000000  1.0000000  1.000000  1.0000000 10.000000
## CV residual 1.297599 -2.314571 -0.9718031 -1.882185 -0.7024006  3.509863
##                    606        614       622       627        630       634
## Predicted    8.1618372  1.2134523  4.712371  6.200209  1.3420059  5.490477
## cvpred       8.2374171  1.1635956  4.764508  6.096742  1.3859849  5.615717
## V7           8.0000000  1.0000000  2.000000  6.000000  1.0000000  3.000000
## CV residual -0.2374171 -0.1635956 -2.764508 -0.096742 -0.3859849 -2.615717
##                    637        639        650        655        656       659
## Predicted    9.842661  1.3420059  1.4396214  1.7634059  1.4396214  8.177821
## cvpred       9.954897  1.3859849  1.4329981  1.7494138  1.4329981  8.162589
## V7           1.000000  1.0000000  1.0000000  1.0000000  1.0000000 10.000000
## CV residual -8.954897 -0.3859849 -0.4329981 -0.7494138 -0.4329981  1.837411
##                    660       661       672       675       681       684
## Predicted   0.6634986 0.9872832  2.095672 0.9872832 11.45905 0.6634986
```

```
## cvpred      0.5777774 0.8941930  2.081250 0.8941930 11.50952 0.5777774
## V7          1.0000000 1.0000000  1.000000 1.0000000 10.00000 1.0000000
## CV residual 0.4222226 0.1058070 -1.081250 0.1058070 -1.50952 0.4222226
##                   686       690       692        693        694
## Predicted   0.6634986 0.6634986  6.724170  1.1158368  1.4396214
## cvpred      0.5777774 0.5777774  6.604831  1.1165824  1.4329981
## V7          1.0000000 1.0000000  5.000000  1.0000000  1.0000000
## CV residual 0.4222226 0.4222226 -1.604831 -0.1165824 -0.4329981
##
## Sum of squares = 834.59    Mean square = 6.09    n = 137
##
## fold 4
## Observations in test set: 137
##                     2         5          8        20         34         35
## Predicted    4.496667  2.654107  1.8545232  2.441913  1.8695027  1.7569078
## cvpred       4.496317  2.595512  1.8472985  2.383609  1.8299937  1.7458263
## V7          10.000000  1.000000  1.0000000  1.000000  1.0000000  1.0000000
## CV residual  5.503683 -1.595512 -0.8472985 -1.383609 -0.8299937 -0.7458263
##                    39       47         48        52        58        70
## Predicted    6.473300 4.987707 0.98728319 3.8471146  4.493680  1.311068
## cvpred       6.442201 5.081462 0.96300316 3.8273985  4.495312  1.284331
## V7          10.000000 9.000000 1.00000000 4.0000000  1.000000  1.000000
## CV residual  3.557799 3.918538 0.03699684 0.1726015 -3.495312 -0.284331
##                    72         75        76         90         95       103
## Predicted    6.380199  4.2984487 1.9521387  1.5457182  1.5372368  2.306861
## cvpred       6.292682  4.2923672 1.9487707  1.5086659  1.5041866  2.287010
## V7           2.000000  4.0000000 2.0000000  1.0000000  1.0000000  1.000000
## CV residual -4.292682 -0.2923672 0.0512293 -0.5086659 -0.5041866 -1.287010
##                   110       116       119        122        132       133
## Predicted   5.685708 0.6634986 0.6634986 1.98957497  1.5372368  7.427142
## cvpred      5.675677 0.6416753 0.6416753 1.94389789  1.5041866  7.497800
## V7          9.000000 5.0000000 3.0000000 2.00000000  1.0000000 10.000000
## CV residual 3.324323 4.3583247 2.3583247 0.05610211 -0.5041866  2.502200
##                   155        163       168       170         171        177
## Predicted   0.6634986  1.7634059  9.192560 0.9957645  1.11583682  1.5372368
## cvpred      0.6416753  1.7240423  9.283302 0.9674824  1.08138657  1.5041866
## V7          1.0000000  1.0000000  1.000000 1.0000000  1.00000000  1.0000000
## CV residual 0.3583247 -0.7240423 -8.283302 0.0325176 -0.08138657 -0.5041866
##                   181        188        193        206        212       213
## Predicted    1.311068  9.2185528  1.8919595  9.0245872  8.7362557  1.311068
## cvpred       1.284331  9.1961661  1.8424257  9.1435282  8.7910644  1.284331
## V7           1.000000 10.0000000  1.0000000 10.0000000  8.0000000  1.000000
## CV residual -0.284331  0.8038339 -0.8424257  0.8564718 -0.7910644 -0.284331
##                   216        217       223       227       228       234
## Predicted    7.756421 0.98728319 2.330322  7.866028  8.056744  6.268583
## cvpred       7.822602 0.96300316 2.274184  7.905763  8.156755  6.260035
## V7           5.000000 1.00000000 5.000000 10.000000  5.000000 10.000000
## CV residual -2.822602 0.03699684 2.725816  2.094237 -3.156755  3.739965
##                   239       247       252       253       263       264
## Predicted   8.0756904  9.370829  7.936411  4.405550  7.724479  7.936411
## cvpred      8.1919766  9.477288  7.867287  4.373061  7.830948  7.867287
## V7          9.0000000 10.000000 10.000000 10.000000 10.000000 10.000000
## CV residual 0.8080234  0.522712  2.132713  5.626939  2.169052  2.132713
##                   273       283       300       305       309       310
```

```
## Predicted    4.659956  5.583603  6.394174  6.504238  7.852053 2.410975
## cvpred       4.707608  5.545474  6.300635  6.459113  7.897811 2.366698
## V7          10.000000 10.000000 10.000000 10.000000  3.000000 5.000000
## CV residual  5.292392  4.454526  3.699365  3.540887 -4.897811 2.633302
##                   311       321       332       336       338        339
## Predicted    1.2134523  4.930059  2.215744 0.6634986  1.311068 0.98728319
## cvpred       1.1828588  4.900807  2.163754 0.6416753  1.284331 0.96300316
## V7           1.0000000 10.000000  1.000000 1.0000000  1.000000 1.00000000
## CV residual -0.1828588  5.099193 -1.163754 0.3583247 -0.284331 0.03699684
##                   342       343       345       347       348        352
## Predicted    1.311068 0.8896677  7.888510  2.217727 0.6634986  1.5372368
## cvpred       1.284331 0.8615309  7.841417  2.190017 0.6416753  1.5041866
## V7           1.000000 1.0000000 10.000000  1.000000 1.0000000  1.0000000
## CV residual -0.284331 0.1384691  2.158583 -1.190017 0.3583247 -0.5041866
##                   355       356         361        365       372        379
## Predicted   0.98728319  1.66579  9.90680671  1.5372368  1.298072  2.436419
## cvpred      0.96300316  1.62257 10.01051881  1.5041866  1.327899  2.380136
## V7          1.00000000  1.00000 10.00000000  1.0000000  1.000000  1.000000
## CV residual 0.03699684 -0.62257 -0.01051881 -0.5041866 -0.327899 -1.380136
##                   386       390       392        393       394         402
## Predicted    1.7653891 1.8919595  7.542244  1.4396214 0.6634986  1.11583682
## cvpred       1.7503055 1.8424257  7.584435  1.4027144 0.6416753  1.08138657
## V7           1.0000000 2.0000000 10.000000  1.0000000 1.0000000  1.00000000
## CV residual -0.7503055 0.1575743  2.415565 -0.4027144 0.3583247 -0.08138657
##                   405        406        407        410        411        420
## Predicted    1.3280304 0.98728319  1.9830769  1.7569078 0.98728319 0.8896677
## cvpred       1.2932895 0.96300316  1.9656819  1.7458263 0.96300316 0.8615309
## V7           1.0000000 1.00000000  1.0000000  1.0000000 1.00000000 1.0000000
## CV residual -0.2932895 0.03699684 -0.9656819 -0.7458263 0.03699684 0.1384691
##                   423        431       434       436        438       443
## Predicted    2.624148 0.98728319  2.097655  6.849736  1.3420059 1.328030
## cvpred       2.630122 0.96300316  2.076113  6.993694  1.3012422 1.293289
## V7           1.000000 1.00000000  1.000000 10.000000  1.0000000 3.000000
## CV residual -1.630122 0.03699684 -1.076113  3.006306 -0.3012422 1.706711
##                   445        453       456       457       458       462
## Predicted    3.553289  1.7803686 3.016307  8.560519  9.360364 1.115837
## cvpred       3.471461  1.7330008 2.963488  8.522768  9.446546 1.081387
## V7           1.000000  1.0000000 6.000000 10.000000  3.000000 5.000000
## CV residual -2.471461 -0.7330008 3.036512  1.477232 -6.446546 3.918613
##                   465        474       483        485        488        490
## Predicted    1.3420059  1.3420059 11.232883  1.8854614 10.8114830 3.0829841
## cvpred       1.3012422  1.3012422 11.312741  1.8642097 10.8899413 3.0480488
## V7           1.0000000  1.0000000  5.000000  1.0000000 10.0000000 4.0000000
## CV residual -0.3012422 -0.3012422 -6.312741 -0.8642097 -0.8899413 0.9519512
##                   496        503        504         506        509        516
## Predicted    1.4396214  1.9980563  1.9895750  1.11583682  1.5681750  6.877737
## cvpred       1.4027144  1.9483771  1.9438979  1.08138657  1.5210978  6.856043
## V7           1.0000000  1.0000000  1.0000000  1.00000000  1.0000000 10.000000
## CV residual -0.4027144 -0.9483771 -0.9438979 -0.08138657 -0.5210978  3.143957
##                   520        526        527        535       538        540
## Predicted    6.817819  1.4481027  1.3420059  1.2134523  2.533030  2.118129
## cvpred       6.925262  1.4071937  1.3012422  1.1828588  2.506865  2.062281
## V7          10.000000  1.0000000  1.0000000  1.0000000  1.000000  1.000000
## CV residual  3.074738 -0.4071937 -0.3012422 -0.1828588 -1.506865 -1.062281
```

```
##                  554        559       562       572        573        578
## Predicted   1.983077  1.2134523  2.215744  8.795431  1.4396214 0.98728319
## cvpred      1.965682  1.1828588  2.163754  8.922667  1.4027144 0.96300316
## V7          5.000000  1.0000000  1.000000 10.000000  1.0000000 1.00000000
## CV residual 3.034318 -0.1828588 -1.163754  1.077333 -0.4027144 0.03699684
##                  581         584       586       589        591        595
## Predicted    2.209246  1.11583682 0.6634986  8.323337  7.806135  5.535677
## cvpred       2.185538  1.08138657 0.6416753  8.372744  7.898204  5.596383
## V7           1.000000  1.00000000 1.0000000  3.000000  1.000000 10.000000
## CV residual -1.185538 -0.08138657 0.3583247 -5.372744 -6.898204  4.403617
##                  599       604        609        632        646       667
## Predicted    1.4396214  7.746960 10.3282067  1.8919595  1.4396214  2.217727
## cvpred       1.4027144  7.766602 10.4333189  1.8424257  1.4027144  2.190017
## V7           1.0000000  1.000000 10.0000000  1.0000000  1.0000000  1.000000
## CV residual -0.4027144 -6.766602 -0.4333189 -0.8424257 -0.4027144 -1.190017
##                  668        674        678        689       698
## Predicted    1.7634059  1.8854614  1.5681750  1.3420059  6.839297
## cvpred       1.7240423  1.8642097  1.5210978  1.3012422  6.886173
## V7           1.0000000  1.0000000  1.0000000  1.0000000  4.000000
## CV residual -0.7240423 -0.8642097 -0.5210978 -0.3012422 -2.886173
##
## Sum of squares = 778.77    Mean square = 5.68    n = 137
##
## fold 5
## Observations in test set: 136
##                    1         9         13       14        18        19
## Predicted    2.215744 0.8896677  3.8386332 1.311068  1.989575  7.235422
## cvpred       2.330188 0.8826526  3.9095676 1.299865  2.072607  7.338340
## V7           1.000000 1.0000000  3.0000000 3.000000  1.000000 10.000000
## CV residual -1.330188 0.1173474 -0.9095676 1.700135 -1.072607  2.661660
##                   25         28         31        33        38         43
## Predicted    1.3110677  1.8919595  1.4396214  7.209978  3.737051  6.924895
## cvpred       1.2998648  1.9927914  1.4776299  7.321236  3.937354  6.781496
## V7           1.0000000  1.0000000  1.0000000  5.000000  1.000000 10.000000
## CV residual -0.2998648 -0.9927914 -0.4776299 -2.321236 -2.937354  3.218504
##                   44        49        60        68         96        98
## Predicted    5.146219  2.654107  5.369401  3.491388  1.3110677  2.215744
## cvpred       5.157252  2.758803  5.489977  3.501266  1.2998648  2.330188
## V7           1.000000  1.000000  2.000000 10.000000  1.0000000  1.000000
## CV residual -4.157252 -1.758803 -3.489977  6.498734 -0.2998648 -1.330188
##                  100       107       113        114       117        121
## Predicted    8.540046  8.851813  8.266694  8.4856218  3.528824  1.9606200
## cvpred       8.672285  8.858217  8.611345  8.4923139  3.658718  1.9208564
## V7          10.000000 10.000000 10.000000  8.0000000  2.000000  1.0000000
## CV residual  1.327715  1.141783  1.388655 -0.4923139 -1.658718 -0.9208564
##                  126        128        134        135        137       142
## Predicted   0.98728319  1.7634059  1.4396214  1.4396214  1.6657904 0.8896677
## cvpred      0.96246831  1.8150263  1.4776299  1.4776299  1.7352106 0.8826526
## V7          1.00000000  1.0000000  1.0000000  1.0000000  1.0000000 1.0000000
## CV residual 0.03753169 -0.8150263 -0.4776299 -0.4776299 -0.7352106 0.1173474
##                  144       148       153        157        158       161
## Predicted   0.6634986 0.9872832  8.847847  1.3045696  1.5372368  6.894675
## cvpred      0.6250719 0.9624683  8.965820  1.2403621  1.5574455  6.989541
## V7          5.0000000 2.0000000  5.000000  1.0000000  1.0000000 10.000000
```

```
## CV residual 4.3749281 1.0375317 -3.965820 -0.2403621 -0.5574455  3.010459
##                    174       180        189        190       191       192
## Predicted   10.5543758  3.514849  8.1007036  1.9456406  9.823167  8.837838
## cvpred      10.5384684  3.572171  8.2761283  1.8556524  9.870407  8.771671
## V7          10.0000000 10.000000  8.0000000  1.0000000  8.000000 10.000000
## CV residual -0.5384684  6.427829 -0.2761283 -0.8556524 -1.870407  1.228329
##                    196        205        209        214      224        229
## Predicted    1.989575  1.3110677  1.3110677 10.4876985 6.214184  1.3110677
## cvpred       2.072607  1.2998648  1.2998648 10.5566022 6.270600  1.2998648
## V7           1.000000  1.0000000  1.0000000 10.0000000 8.000000  1.0000000
## CV residual -1.072607 -0.2998648 -0.2998648 -0.5566022 1.729400 -0.2998648
##                    230        246        251       256        257        258
## Predicted    8.809406  2.5480100 0.98078507  4.134442  1.1158368  1.4396214
## cvpred       8.829710  2.6732857 0.90296568  4.062755  1.1402334  1.4776299
## V7          10.000000  2.0000000 1.00000000 10.000000  1.0000000  1.0000000
## CV residual  1.170290 -0.6732857 0.09703432  5.937245 -0.1402334 -0.4776299
##                    262       266        277        278      280       284
## Predicted    8.999143  3.167603  1.4396214 0.98728319 5.91484  5.580591
## cvpred       8.908496  3.163869  1.4776299 0.96246831 5.97165  5.682354
## V7          10.000000  1.000000  1.0000000 1.00000000 7.00000 10.000000
## CV residual  1.091504 -2.163869 -0.4776299 0.03753169 1.02835  4.317646
##                    285       286       301        308       313       314
## Predicted    6.927621 11.00671  8.374031  1.3110677  6.836504 0.6634986
## cvpred       7.044805 11.05363  8.325952  1.2998648  7.024492 0.6250719
## V7          10.000000 10.00000  4.000000  1.0000000  1.000000 1.0000000
## CV residual  2.955195 -1.05363 -4.325952 -0.2998648 -6.024492 0.3749281
##                    320        326          329        370       383        396
## Predicted    5.2333701  1.7569078  3.861090053  1.6218560  2.412958  1.4396214
## cvpred       5.2851677  1.7555237  4.001815629  1.5182559  2.436018  1.4776299
## V7           5.0000000  1.0000000  4.000000000  1.0000000  1.000000  1.0000000
## CV residual -0.2851677 -0.7555237 -0.001815629 -0.5182559 -1.436018 -0.4776299
##                    399        400       401        409        419        422
## Predicted    1.4396214  1.2980715 7.920713  2.1867891  2.8652964  9.8146854
## cvpred       1.4776299  1.1808595 7.843249  2.1784372  2.9511795  9.8647051
## V7           1.0000000  1.0000000 9.000000  2.0000000  2.0000000 10.0000000
## CV residual -0.4776299 -0.1808595 1.156751 -0.1784372 -0.9511795  0.1352949
##                    424        425        430       435       440       460
## Predicted    2.526532  1.1158368  1.2134523 5.998480  1.568175  2.202748
## cvpred       2.548579  1.1402334  1.2200491 5.964919  1.655395  2.211183
## V7           1.000000  1.0000000  1.0000000 8.000000  1.000000  1.000000
## CV residual -1.548579 -0.1402334 -0.2200491 2.035081 -0.655395 -1.211183
##                    466       467       468        469        470        476
## Predicted    8.856328  7.207995  6.652547  1.3420059  1.3045696  1.1158368
## cvpred       8.971521  7.375037  6.699215  1.3978142  1.2403621  1.1402334
## V7           4.000000 10.000000 10.000000  1.0000000  1.0000000  1.0000000
## CV residual -4.971521  2.624963  3.300785 -0.3978142 -0.2403621 -0.1402334
##                    478       486        493        499       508       510
## Predicted    1.3420059 1.328030  1.6657904  1.6657904 0.6634986 0.8896677
## cvpred       1.3978142 1.311268  1.7352106  1.7352106 0.6250719 0.8826526
## V7           1.0000000 3.000000  1.0000000  1.0000000 4.0000000 1.0000000
## CV residual -0.3978142 1.688732 -0.7352106 -0.7352106 3.3749281 0.1173474
##                    515       517       532       545       547       565
## Predicted    8.9549474 0.6634986  1.983077  2.180291  9.583022  1.989575
## cvpred       9.0299941 0.6250719  2.013104  2.118935  9.526279  2.072607
```

```
## V7          10.0000000 1.0000000  1.000000  1.000000 10.000000  1.000000
## CV residual  0.9700059 0.3749281 -1.013104 -1.118935  0.473721 -1.072607
##                   575       576       582       583      587       592
## Predicted    7.209978  2.533030  8.027790  6.011476 11.00671  6.397423
## cvpred       7.321236  2.608082  7.896307  6.083924 11.05363  6.289883
## V7           2.000000  1.000000 10.000000 10.000000 10.00000 10.000000
## CV residual -5.321236 -1.608082  2.103693  3.916076 -1.05363  3.710117
##                   594        596        602       608       610       612
## Predicted    1.8854614  1.8919595 0.98728319 0.6634986  1.568175  9.680638
## cvpred       1.9332888  1.9927914 0.96246831 0.6250719  1.655395  9.606095
## V7           1.0000000  1.0000000 1.00000000 1.0000000  1.000000  2.000000
## CV residual -0.9332888 -0.9927914 0.03753169 0.3749281 -0.655395 -7.606095
##                   613        619        620        621       623       625
## Predicted   11.00671  1.6657904  1.8919595  1.4396214  3.326116  2.224225
## cvpred      11.05363  1.7352106  1.9927914  1.4776299  3.472042  2.335889
## V7          10.00000  1.0000000  1.0000000  1.0000000  1.000000  1.000000
## CV residual -1.05363 -0.7352106 -0.9927914 -0.4776299 -2.472042 -1.335889
##                   629       631        635        643       645        647
## Predicted   0.8896677  2.428917  1.1158368  1.4396214 0.8896677 0.98078507
## cvpred      0.8826526  2.468763  1.1402334  1.4776299 0.8826526 0.90296568
## V7          1.0000000  1.000000  1.0000000  1.0000000 1.0000000 1.00000000
## CV residual 0.1173474 -1.468763 -0.1402334 -0.4776299 0.1173474 0.09703432
##                   648      651        652        653        654        657
## Predicted    1.3280304 1.448103  1.6518150  1.8919595  1.6657904  1.8919595
## cvpred       1.3112675 1.483331  1.6486639  1.9927914  1.7352106  1.9927914
## V7           1.0000000 4.000000  1.0000000  1.0000000  1.0000000  1.0000000
## CV residual -0.3112675 2.516669 -0.6486639 -0.9927914 -0.7352106 -0.9927914
##                   669       671       680       687
## Predicted    4.462741 7.2881239 0.8896677 0.6634986
## cvpred       4.513455 7.2336597 0.8826526 0.6250719
## V7           1.000000 8.0000000 1.0000000 1.0000000
## CV residual -3.513455 0.7663403 0.1173474 0.3749281
##
## Sum of squares = 591.02    Mean square = 4.35    n = 136
##
## Overall (Sum over all 136 folds)
##       ms
## 5.199782
```

```
# Calculate SST
SST <- sum((as.numeric(data[-missing,]$V7) - mean(as.numeric(data[-missing,]$V7)))^2)
# R-squared
R2_cv <- 1 - attr(model_cv, "ms") * nrow(data[-missing,]) / SST
R2_cv
```

```
## [1] 0.607808
```

Next we obtain the predictions for missing V7 values.

```
# Get predictions for missing V7 values.
V7_hat <- predict(final_model, newdata = data[missing,])
V7_hat
```

```
##        24        41       140       146       159       165       236       250
## 5.4585352 7.9816106 0.9872832 1.6218560 0.9807851 2.2157441 2.7152652 1.7634059
##       276       293       295       298       316       322       412       618
## 2.0741942 6.0866099 0.9872832 2.5265324 5.2438347 1.7634059 0.9872832 0.6634986
```

Finally, data imputation is performed.

```
# Copy of original data set
reg_imputation <- data
# Replace the missing values with the predicted value and round for int
reg_imputation[missing,]$V7 <- round(V7_hat)
# Determine values are numeric
reg_imputation$V7 <- as.numeric(reg_imputation$V7)

# Maintain V7 values stay within the original range [1, 10]
reg_imputation$V7 <- pmin(pmax(reg_imputation$V7, 1), 10)
```

Part C: Regression with Perturbation

Regression with perturbation is for perturbing the predicted values for `V7`. A random normal distribution is used and the standard deviation is of the predicted value.

```
# Perturb missing V7 value predictions
# Use random normal distribution std. dev. of predicted value
V7_hat_pert <- rnorm(nrow(data[missing,]), V7_hat, sd(V7_hat))
V7_hat_pert
```

```
##  [1]  6.93560160  3.14378473  0.08579384 -0.78969862  0.95708275  3.31797605
##  [7]  0.79160608 -0.57705244  6.79889284  7.76637879  2.78578227  2.98556238
## [13]  6.48492561  3.80582469 -0.73027027 -1.57838969
```

Finally, data imputation is performed.

```
# Copy of original data set
reg_imputation_pert <- data
# Replace the missing values with the predicted value and round for int
reg_imputation_pert[missing,]$V7 <- round(V7_hat_pert)
reg_imputation_pert$V7 <- as.numeric(reg_imputation_pert$V7)

# Maintain V7 values stay within the original range [1, 10]
reg_imputation_pert$V7 <- pmin(pmax(reg_imputation_pert$V7, 1), 10)
```

Question 15.1

Prompt

Describe a situation or problem from your job, everyday life, current events, etc., for which optimization would be appropriate. What data would you need?

Solution

Optimization is very important in the National Football League (NFL). Given a game scenario, it may be important to consider the optimal play-calling strategy. NFL coaches must determine which play to call considering game clock, down and distance, score, and field position. The play call could be a number of variations, whether run or pass. The run could be inside, outside, sweeps, etc. The pass could be a screen pass, short pass, long pass, etc. Optimization would be useful to maximize the likelihood of scoring or game control.

Data Needed:

- **Play-by-Play Data:** Historical play-by-play data can be obtained through the R package, `nflfastR`. The data includes details such as down, distance, play type, yards gained and success.
- **Player Performance Data:** This would include player performance metrics for an individual. For example, running backs would consider yards per carry, yards after catch and catch percentage. This would help determine likely outcomes of certain plays based on player personnel currently on the field.
- **Game Scenario Data:** Previous game scenario data will help inform models similar to the current situation and identify what the most effective play types were under those specific conditions.
- **Defensive Metrics for Opposing Team:** Defensive stats about the upcoming opponent to assist in identifying weaknesses. This could include efficiency metrics for the defense in different situations like run, pass, blitz, etc.
- **Player Injury or Health Status:** Exposing weaker opponent matchups could allow for big success plays especially further in the game for fatigued players.

References

[1] U. M. L. Repository, "Breast cancer wisconsin (original)." http://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original, 1995.