

Data Preparation

Zachary Levonian

11/02/2018

```
library(alr4)
```

```
## Loading required package: car
## Loading required package: carData
## Loading required package: effects
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(mice) # for multiple imputation
```

```
## Loading required package: lattice
##
## Attaching package: 'mice'
##
## The following objects are masked from 'package:base':
##
##      cbind, rbind
```

```
library(BaylorEdPsych) # For Little's MCAR test
library(polycor) # To compute correlation between heterogenous variables
library(plotrix) # For side-along histograms
library(caret) # For Cross Validation
```

```
## Loading required package: ggplot2
```

```
library(glmnet) # For generalized linear models
```

```
## Loading required package: Matrix
## Loading required package: foreach
## Loaded glmnet 2.0-16
```

```
library(xgboost) # For gradient-boosted decision trees
```

Load data

```
train <- read.csv("../data/raw/train.csv", stringsAsFactors=FALSE, na.strings = c("NA", ""))
test <- read.csv("../data/raw/test.csv", stringsAsFactors=FALSE, na.strings = c("NA", ""))
```

Combine the data into a single dataframe to make it easier to work with. I denote data in the test set with Survived = 2.

```
test$Survived = 2
df <- rbind(train, test)
```

Load other derived data

```
new_cols <- read.csv("../data/derived/levon003_new_cols.csv", stringsAsFactors=TRUE)
df <- merge(df, new_cols, by="PassengerId", all.y=TRUE, all.x=FALSE)
```

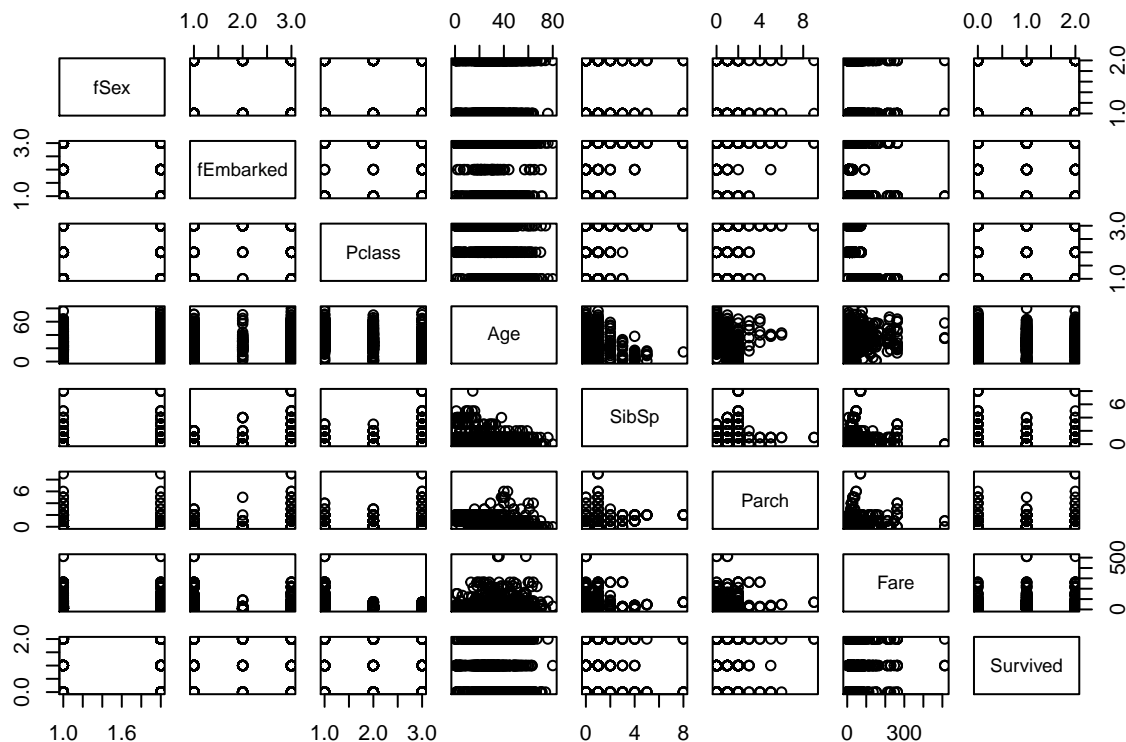
Data exploration

Build factors from data

```
df$fSex = factor(df$Sex)
df$fEmbarked = factor(df$Embarked)
```

High-level summaries and visualization

```
pairs(df[c("fSex", "fEmbarked", "Pclass", "Age", "SibSp", "Parch", "Fare", "Survived")])
```



Missing data

```
sapply(df, function(x) sum(is.na(x)))
```

```
##      PassengerId      Survived      Pclass
##           0           0           0
##      Name      Sex      Age
##           0           0      263
```

```
##          SibSp          Parch          Ticket
##          0            0            0
##          Fare          Cabin          Embarked
##          1          1014            2
##  ticket_category cabin_first_letter      name_title
##          0            0            0
##  name_title_raw  name_word_length  name_char_length
##          0            0            0
##          fSex          fEmbarked
##          0            2
```

It looks like the only data that's missing is Age and Cabin data. In addition, a single instance of the missing Fare data (in the test set) and two instances of the Embarked data are missing.

Fare missing data

```
# print the row where Fare info is missing
df[is.na(df["Fare"])]
```

```
## [1] "1044"          "2"            "3"
## [4] "Storey, Mr. Thomas" "male"         "60.50"
## [7] "0"            "0"            "3701"
## [10] NA              NA              "S"
## [13] "digit"        "n"            "Mr."
## [16] "Mr."          "2"            "12"
## [19] "male"         "S"
```

We need to impute this value, but as there's only a single missing value it's impossible to determine if the data is missing at random or not.

We will assume the data is missing at random and impute a value for Thomas Storey's fare using *mice*.

It is imputed alongside the Age data below.

fEmbarked missing data

Is imputed alongside the Age data below.

Age missing data

263 passengers (20%) are missing age data.

First, we want to determine if the data are missing at random (MAR) or completely at random (MCAR).

```
age_little_df <- df[,c("fSex", "fEmbarked", "Pclass", "Age", "SibSp", "Parch", "Survived")]
mcar <- LittleMCAR(age_little_df)
```

```
## Loading required package: mvnmle
## Warning in nlm(lf, startvals, ...): NA/Inf replaced by maximum positive
## value
## Warning in nlm(lf, startvals, ...): NA/Inf replaced by maximum positive
## value
## this could take a while
```

```
mcars$missing.patterns
```

```
## [1] 3
```

```
mcars$amount.missing
```

```
##           fSex  fEmbarked Pclass           Age SibSp Parch Survived
## Number Missing    0 2.000000000    0 263.0000000    0    0      0
## Percent Missing    0 0.001527884    0  0.2009167    0    0      0
```

```
mcars$p.value
```

```
## [1] 0
```

Little's MCAR test generates a test statistic against the null hypothesis that the missing data are MCAR. Thus, we have evidence that we ought to reject the null hypothesis and the missing age data are MAR [2].

```
age_little_df <- df[,c("fSex", "fEmbarked", "Pclass", "SibSp", "Parch", "Fare", "Survived", "name_title", "name_word_length", "name_char_length", "cabin_first_letter", "ticket_category", "AgeMissing")]
age_little_df$AgeMissing = as.numeric(is.na(df["Age"]))
hetcor(age_little_df)
```

```
## Warning in hetcor.data.frame(age_little_df): could not compute polyserial correlation between variable 'AgeMissing' and 'Survived'
## Message: Error in optim(rho, f, control = control, hessian = TRUE, method = "BFGS") :
## initial value in 'vmmin' is not finite
```

```
##
```

```
## Two-Step Estimates
```

```
##
```

```
## Correlations/Type of Correlation:
```

```
##           fSex  fEmbarked  Pclass  SibSp  Parch
## fSex           1 Polychoric Polyserial Polyserial Polyserial
## fEmbarked      0.1682      1 Polyserial Polyserial Polyserial
## Pclass         0.1528    0.1998      1  Pearson  Pearson
## SibSp         -0.1364    0.1073    0.06015      1  Pearson
## Parch         -0.2627    0.07867    0.0176    0.3733      1
## Fare          -0.2267   -0.2732   -0.5579    0.161    0.2223
## Survived      -0.2838   -0.1685   -0.1531   -0.04387    0.03514
## name_title    -0.03472    0.008406   -0.1445   -0.1749   -0.0454
## name_word_length -0.2822    0.0944   -0.245    0.1557    0.175
## name_char_length  0.03937    0.1071   -0.1551    0.1008    0.05296
## cabin_first_letter  0.1931    0.2492    0.5813   -0.0315   -0.05094
## ticket_category -0.08557   -0.2866   -0.2275   -0.1568   -0.08901
## AgeMissing     0.08103   -0.1729    0.2078   -0.008244   -0.08266
##           Fare  Survived name_title name_word_length
## fSex      Polyserial Polyserial Polychoric  Polyserial
## fEmbarked Polyserial Polyserial Polychoric  Polyserial
## Pclass     Pearson  Pearson Polyserial  Pearson
## SibSp      Pearson  Pearson Polyserial  Pearson
## Parch      Pearson  Pearson Polyserial  Pearson
## Fare        1  Pearson Polyserial  Pearson
## Survived    0.123      1 Polyserial  Pearson
## name_title  0.006022  0.003144      1  Polyserial
## name_word_length  0.1589    0.0741    0.2333      1
## name_char_length  0.08975  0.0007376   -0.1018    0.6734
## cabin_first_letter <NA>   -0.1033   -0.09417   -0.1793
## ticket_category  0.1855    0.03557    0.05789    0.002628
## AgeMissing   -0.13   -0.02776   -0.01838   -0.1834
```

```

##          name_char_length cabin_first_letter ticket_category
## fSex          Polyserial          Polychoric          Polychoric
## fEmbarked      Polyserial          Polychoric          Polychoric
## Pclass          Pearson          Polyserial          Polyserial
## SibSp           Pearson          Polyserial          Polyserial
## Parch           Pearson          Polyserial          Polyserial
## Fare            Pearson          Polyserial          Polyserial
## Survived        Pearson          Polyserial          Polyserial
## name_title      Polyserial          Polychoric          Polychoric
## name_word_length Pearson          Polyserial          Polyserial
## name_char_length      1          Polyserial          Polyserial
## cabin_first_letter    -0.124          1          Polychoric
## ticket_category      0.01601        -0.1717          1
## AgeMissing         -0.1536          0.1566        -0.01558
##          AgeMissing
## fSex          Polyserial
## fEmbarked      Polyserial
## Pclass          Pearson
## SibSp           Pearson
## Parch           Pearson
## Fare            Pearson
## Survived        Pearson
## name_title      Polyserial
## name_word_length Pearson
## name_char_length Pearson
## cabin_first_letter Polyserial
## ticket_category Polyserial
## AgeMissing      1
##
## Standard Errors:
##          fSex fEmbarked Pclass SibSp Parch Fare
## fSex
## fEmbarked      0.04427
## Pclass          0.03418      0.0327
## SibSp           0.03395      0.04096 0.02758
## Parch           0.03282      0.03934 0.02767 0.02383
## Fare            0.0337      0.03366 0.01907 0.02696 0.02631
## Survived        0.03239      0.03413 0.02703 0.02763 0.02765 0.02726
## name_title      0.03679      0.03814 0.02919 0.02814 0.0296 0.02967
## name_word_length 0.03205      0.03507 0.02602 0.02701 0.02683 0.02698
## name_char_length 0.03548      0.03561 0.02702 0.0274 0.0276 0.02746
## cabin_first_letter 0.04239      0.04053 0.02044 0.03578 0.03451 0
## ticket_category 0.0416      0.03992 0.03036 0.02978 0.03104 0.02851
## AgeMissing      0.03597      0.03178 0.02649 0.02768 0.02749 0.02721
##          Survived name_title name_word_length name_char_length
## fSex
## fEmbarked
## Pclass
## SibSp
## Parch
## Fare
## Survived
## name_title      0.03025
## name_word_length 0.02753      0.02781

```

```

## name_char_length      0.02768      0.02957          0.01514
## cabin_first_letter    0.03471      0.03619          0.03293          0.03382
## ticket_category       0.0322      0.03506          0.03225          0.03216
## AgeMissing            0.02766      0.03017          0.02675          0.02703
##
##          cabin_first_letter ticket_category
## fSex
## fEmbarked
## Pclass
## SibSp
## Parch
## Fare
## Survived
## name_title
## name_word_length
## name_char_length
## cabin_first_letter
## ticket_category          0.04029
## AgeMissing              0.03535          0.03257
##
## n = 1306
##
## P-values for Tests of Bivariate Normality:
##          fSex  fEmbarked          Pclass
## fSex
## fEmbarked          0.02482
## Pclass          4.375e-213 2.008e-251
## SibSp              0          0          0
## Parch              0          0          0
## Fare              0          0          0
## Survived          7.6e-208 1.453e-169          0
## name_title          0  1.407e-11          3.584e-270
## name_word_length    7.798e-211 7.348e-214          0
## name_char_length    0.02409  4.976e-05          5.873e-209
## cabin_first_letter    0.1081  1.308e-13 1.75000000000102e-312
## ticket_category      7.638e-08  8.61e-43          0
## AgeMissing           0          0          0
##
##          SibSp Parch Fare          Survived
## fSex
## fEmbarked
## Pclass
## SibSp
## Parch              0
## Fare              0          0
## Survived          0          0          0
## name_title          0          0          0          5.74e-279
## name_word_length    0          0          0          0
## name_char_length    4.94065645841247e-324          0          0          7.475e-167
## cabin_first_letter    0          0          0 8.825999999999999e-256
## ticket_category      0          0          0          5.4e-240
## AgeMissing           0          0          0          0
##
##          name_title name_word_length name_char_length
## fSex
## fEmbarked
## Pclass

```

```
## SibSp
## Parch
## Fare
## Survived
## name_title
## name_word_length 2.336e-286
## name_char_length 3.349e-50 2.079e-210
## cabin_first_letter 1.16e-05 4.414e-283 1.52e-82
## ticket_category 0.06065 2.269e-270 2.473e-71
## AgeMissing 0 0 0
## cabin_first_letter ticket_category
## fSex
## fEmbarked
## Pclass
## SibSp
## Parch
## Fare
## Survived
## name_title
## name_word_length
## name_char_length
## cabin_first_letter
## ticket_category 1.254e-14
## AgeMissing 0 0
```

A missing age value is correlated positively with passenger class ($r = 0.2082$) and negatively with point of embarkment ($r = -0.1672$) and passenger fare ($r = -0.1306$). All other correlations are < 0.1 .

I'm inclined to think that the true mediator of missing age (among the covariates in the dataset) is passenger class, which embarkment and fare both correlate with.

```
t.test(df[is.na(df["Age"]), "Fare"], df[!is.na(df["Age"]), "Fare"])
```

```
##
## Welch Two Sample t-test
##
## data: df[is.na(df["Age"]), "Fare"] and df[!is.na(df["Age"]), "Fare"]
## t = -6.9669, df = 852.61, p-value = 6.481e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -21.61344 -12.11208
## sample estimates:
## mean of x mean of y
## 19.82332 36.68608
```

We use a univariate t -test to evaluate the fare, since it's numeric, and at the 99% confidence level we reject the null hypothesis which suggests the missing data are MAR (rather than MCAR).

Now, we turn to tangible estimation of the missing data estimates.

```
age_df <- df[,c("Age", "fSex", "fEmbarked", "Pclass", "SibSp", "Parch", "Fare", "Survived", "name_title")]
imp <- mice(age_df, print=FALSE, m=20, seed=1, maxit=20)
imputed_df <- complete(imp)

reg_imp <- mice(age_df, print=FALSE, m=20, seed=1, maxit=20, method="norm.nob")
```

```
## Warning: Type mismatch for variable(s): fEmbarked
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

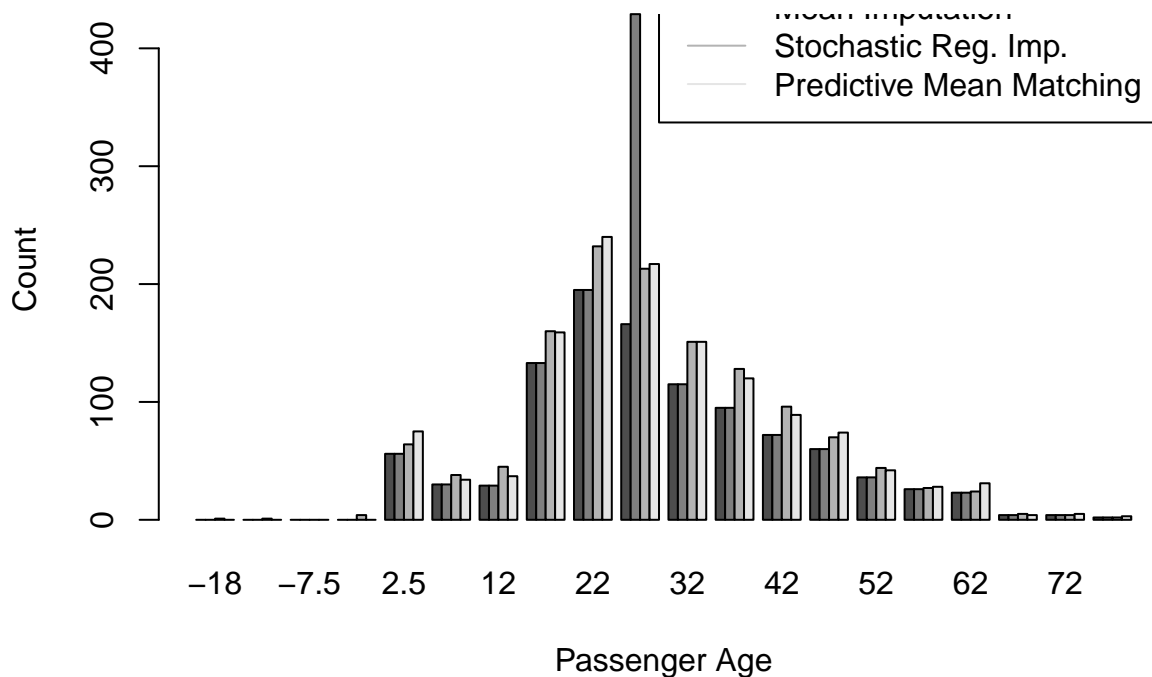

```
## Warning in Ops.factor(p$r, 2): '^' not meaningful for factors
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
## Warning in Ops.factor(p$r, 2): '^' not meaningful for factors
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
## Warning in Ops.factor(p$r, 2): '^' not meaningful for factors
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
## Warning in Ops.factor(p$r, 2): '^' not meaningful for factors
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
## Warning in Ops.factor(p$r, 2): '^' not meaningful for factors
reg_imputed_age <- complete(reg_imp)$Age

#plot(imp)
#fit <- with(imp, lm(Survived ~ Age))
#pool(fit)

age_pre <- imp$data$Age

mean_imputed <- age_pre
mean_imputed[is.na(mean_imputed)] = mean(mean_imputed, na.rm=TRUE)

# plot of the change in the distribution of Age
# after imputation of NA values
colors=c("grey30", "grey50", "grey70", "grey90")
multihist(list(imp$data$Age, mean_imputed, reg_imputed_age, complete(imp)$Age), xlab="Passenger Age", ylab="Count",
           legend=50, 500, legend=c("Missing Excluded", "Mean Imputation", "Stochastic Reg. Imp.", "Predictive Mean Matching"))
```



Following the guidance of [1], we utilize multiple imputation with $m = 20$ imputations.

TODO I should compare the performance of models where Age is imputed vs when it is removed via complete case analysis.

Cabin missing data

I choose not to handle the Cabin data right now, since I think it needs a more elaborate extraction into multiple additional columns.

We could add a binary indicator variable for the presence of Cabin, but such indicator variables can result in biased regression estimates [1].

Overwrite the original dataframe with the imputed values

```
df$fEmbarked <- imputed_df$fEmbarked
df$Age <- imputed_df$Age
df$Fare <- imputed_df$Fare
sapply(df, function(x) sum(is.na(x)))
```

```
##      PassengerId      Survived      Pclass
##           0           0           0
##      Name         Sex         Age
##           0           0           0
##      SibSp      Parch      Ticket
##           0           0           0
##      Fare      Cabin      Embarked
##           0      1014           2
## ticket_category cabin_first_letter name_title
##           0           0           0
## name_title_raw  name_word_length name_char_length
##           0           0           0
##      fSex      fEmbarked
##           0           0
```

Save the cleaned-up data

Now, we save all the columns to be used as potential features to a file.

```
toWrite <- df[,c("PassengerId", "Survived", "fSex", "fEmbarked", "Pclass", "Age", "SibSp", "Parch", "Fare")]
write.table(toWrite, file="../../data/derived/factorized_data.csv",
            row.names=FALSE, col.names=TRUE,
            sep="," , quote=FALSE)
```

References

1. Stef van Buuren. 2018. *Flexible Imputation of Missing Data, Second Edition*. Chapman; Hall/CRC. <https://doi.org/10.1201/9780429492259>
2. Craig K. Enders. 2010. *Applied Missing Data Analysis*. Guilford Press. Retrieved from <http://www.appliedmissingdata.com/>