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
```

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)
```

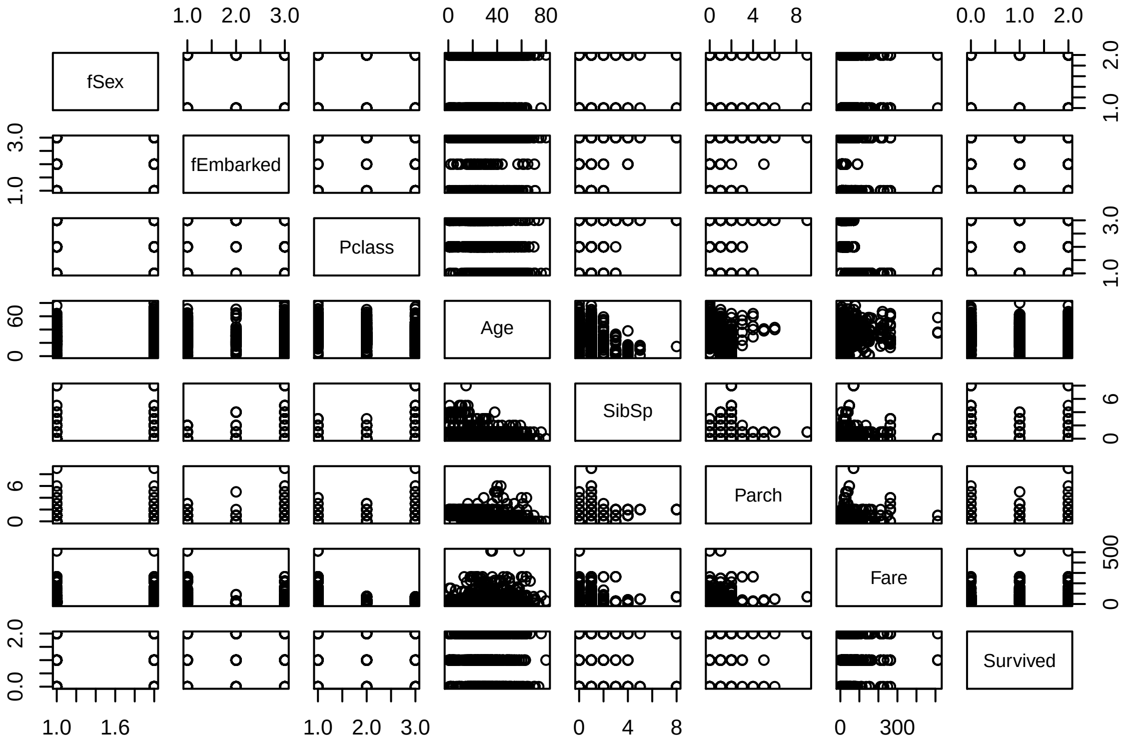
Data exploration

Build factors from data

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

High-level summaries and visualization

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



Missing data

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

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1        1014           2
##        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

```r
# 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] "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.

```
#TODO use mice to impute the data
```

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
```

```
mcar$missing.patterns
```

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

```
mcar$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
```

```
mcar$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")]
age_little_df$AgeMissing = as.numeric(is.na(df["Age"]))
hetcor(age_little_df)
```

```
##
## Two-Step Estimates
##
## Correlations/Type of Correlation:
##               fSex  fEmbarked     Pclass      SibSp      Parch       Fare
## fSex             1 Polychoric Polyserial Polyserial Polyserial Polyserial
## fEmbarked   0.1682          1 Polyserial Polyserial Polyserial Polyserial
```

```
## Pclass      0.1528     0.1998          1     Pearson     Pearson     Pearson
## SibSp      -0.1364     0.1073    0.06015           1     Pearson     Pearson
## Parch      -0.2627    0.07867     0.0176      0.3733           1     Pearson
## Fare       -0.2267    -0.2732    -0.5579       0.161      0.2223           1
## Survived   -0.2838    -0.1685    -0.1531    -0.04387     0.03514       0.123
## AgeMissing 0.08103    -0.1729     0.2078   -0.008244    -0.08266       -0.13
##              Survived AgeMissing
## fSex       Polyserial Polyserial
## fEmbarked  Polyserial Polyserial
## Pclass        Pearson    Pearson
## SibSp         Pearson    Pearson
## Parch         Pearson    Pearson
## Fare          Pearson    Pearson
## Survived            1    Pearson
## AgeMissing   -0.02776          1
##
## Standard Errors:
##               fSex fEmbarked  Pclass   SibSp   Parch    Fare Survived
## 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
## AgeMissing 0.03597   0.03178 0.02649 0.02768 0.02749 0.02721  0.02766
##
## n = 1306
##
## P-values for Tests of Bivariate Normality:
##                  fSex  fEmbarked Pclass SibSp Parch Fare Survived
## fSex
## fEmbarked     0.02482
## Pclass     4.375e-213 2.008e-251
## SibSp               0          0      0
## Parch               0          0      0     0
## Fare                0          0      0     0     0
## Survived     7.6e-208 1.453e-169      0     0     0    0
## AgeMissing          0          0      0     0     0    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.

```r
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
```

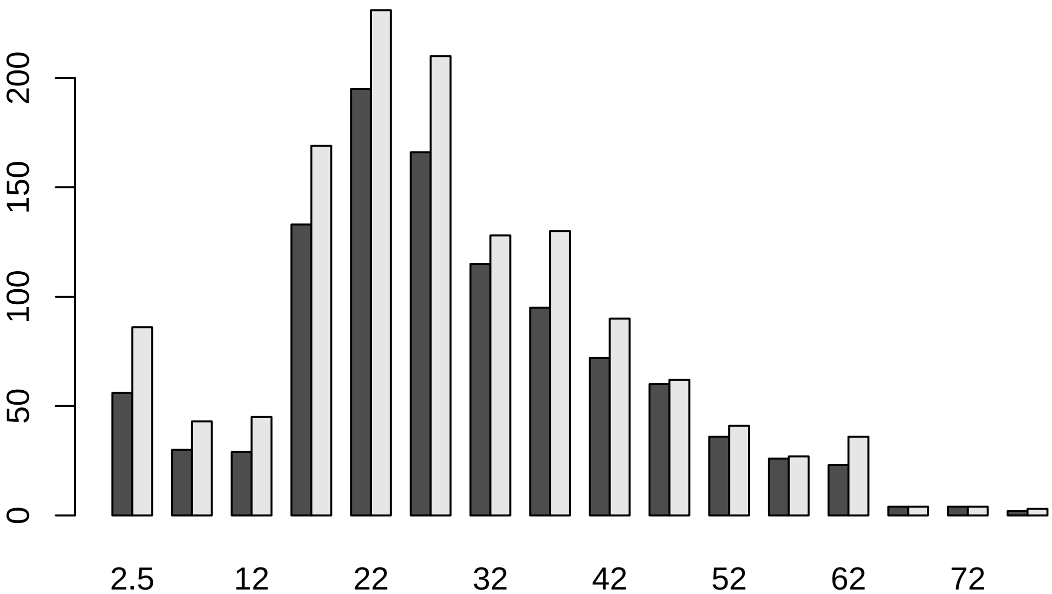
We use a univariate t-test to evaluate the fare, since it's numeric, and at the 99% confidence level we reject the null hypothesis which suggests the missing data are MAR (rather than MCAR).

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

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

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

```r
# plot of the change in the distribution of Age
# after imputation of NA values
multhist(list(imp$data$Age, complete(imp)$Age))
```



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        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0        1014           2
##        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", "fSex", "fEmbarked", "Pclass", "Age", "SibSp", "Parch", "Fare", "Surviv
write.table(toWrite, file="../../data/derived/factorized_data.csv",
            row.names=FALSE, col.names=TRUE,
            sep=",", quote=FALSE)
```

Train a Model

The code below demonstrates: - Reading in the features - Splitting the data into the training and test data - Training a regression model - Predicting the test set based on the trained model - Saving the predictions in the Kaggle format for submission

```
df <- read.csv("../../data/derived/factorized_data.csv", stringsAsFactors=TRUE)
```

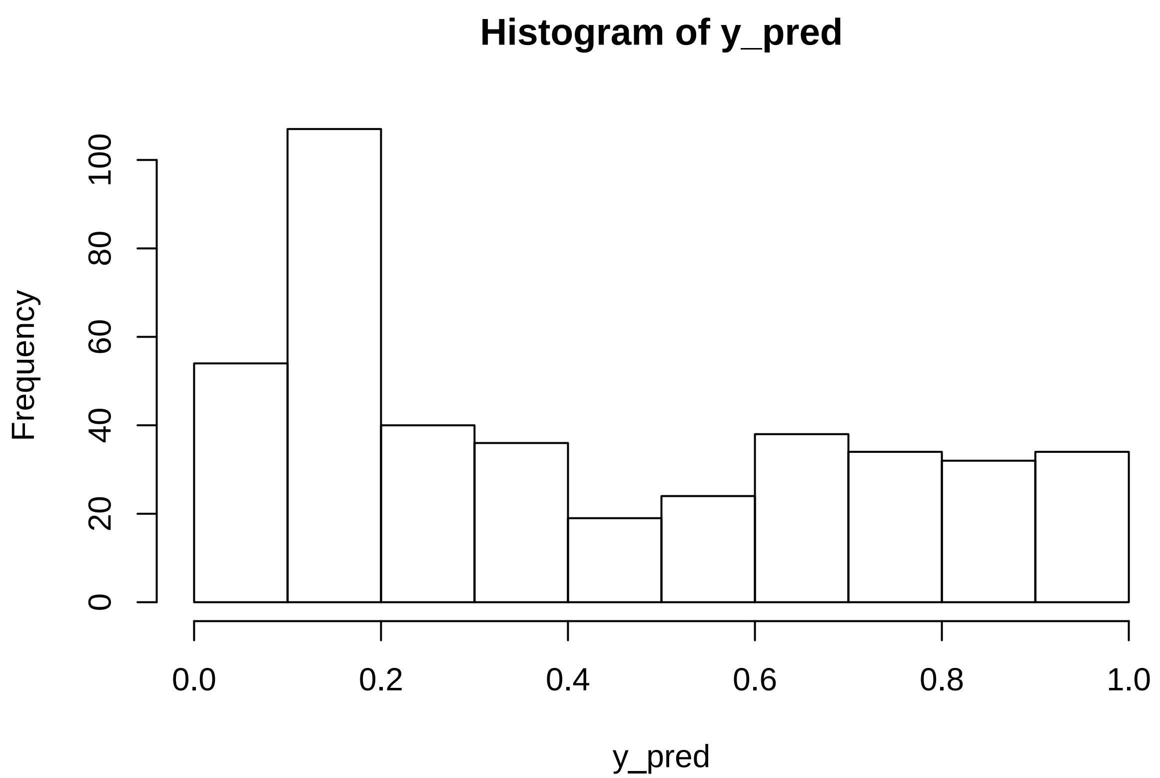
```
train <- df[df$Survived != 2, ]
test <- df[df$Survived == 2, ]
```

```
md <- glm(Survived ~ fSex + fEmbarked + Pclass + Age + SibSp + Parch + Fare, family="binomial", data=tra
summary(md)
```

```
##
## Call:
## glm(formula = Survived ~ fSex + fEmbarked + Pclass + Age + SibSp +
##     Parch + Fare, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5932  -0.6032  -0.4052   0.6229   2.6962
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.012094   0.539993   9.282  < 2e-16 ***
## fSexmale    -2.690097   0.199378 -13.492  < 2e-16 ***
## fEmbarkedQ   0.101319   0.388682   0.261  0.79434
## fEmbarkedS  -0.369079   0.235813  -1.565  0.11755
## Pclass      -1.088240   0.143465  -7.585 3.31e-14 ***
```

```
## Age         -0.034139   0.007121  -4.794 1.63e-06 ***
## SibSp       -0.335815   0.110199  -3.047  0.00231 **
## Parch       -0.093231   0.117841  -0.791  0.42885
## Fare         0.002394   0.002427   0.986  0.32403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  787.91  on 882  degrees of freedom
## AIC: 805.91
##
## Number of Fisher Scoring iterations: 5
```

```r
# predict on the test set
y_pred <- predict(md, test, type="response")
hist(y_pred)
```




```r
# save the predictions to a file that can be submitted to Kaggle
test$Survived = as.numeric(y_pred > 0.5)
# while there shouldn't be NAs in the test output,
# here we apply a nasty hack to ensure any NAs are numeric in the output
test[is.na(test)] <- 0
toWrite <- test[,c("PassengerId", "Survived")]
write.table(toWrite, file="../../data/derived/testModelPredictions.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/