Coding_Sample_1

Martin Munoz

2023-10-14

We will load a dataset that includes 50 observations for 16 variables. Each observation is indexed to a year. Observations track statistics for the Tour de France cycling race aggregated across all cyclists. There are 10 control variables that measure race statistics (e.g., winner time), 5 variables of interest that are all dummy variables that measure whether an anti-doping policy was in place, and 1 response variable that measures the percentage of total cyclists that tested positive for performance enhancing drugs ('ped_tot').

```r
# loading dataset and packages
library(glmnet)
library(ggplot2)
library(sandwich)
library(lmtest)
library(dplyr)
library(reshape2)
library(regclass)
library(stargazer)
library(ggfortify)
library(hdnom)
library(knitr)
tdf <- read.csv("/Users/martrinmunoz/Desktop/EconPredoc/Writing Samples/TdF/tdf_cleaned.csv")
```

```r
# Let's load a table showing summary statistics (e.g., median, mean, min, max)
# for each variable.
summary(tdf)
```

```
##       year         ped_tot        gen_ad_test   amph_test       epo_test
##  Min.   :1968   Min.   :0.1060   Min.   :1    Min.   :0.00   Min.   :0.00
##  1st Qu.:1980   1st Qu.:0.3145   1st Qu.:1    1st Qu.:1.00   1st Qu.:0.00
##  Median :1992   Median :0.4125   Median :1    Median :1.00   Median :0.00
##  Mean   :1992   Mean   :0.3729   Mean   :1    Mean   :0.88   Mean   :0.34
##  3rd Qu.:2005   3rd Qu.:0.4490   3rd Qu.:1    3rd Qu.:1.00   3rd Qu.:1.00
##  Max.   :2017   Max.   :0.5400   Max.   :1    Max.   :1.00   Max.   :1.00
##   bio_passport   night_test        ooct         num_stages      tot_length
##  Min.   :0.0   Min.   :0.00   Min.   :0.00   Min.   :20.50   Min.   :3278
##  1st Qu.:0.0   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:20.50   1st Qu.:3529
##  Median :0.0   Median :0.00   Median :0.00   Median :21.50   Median :3734
##  Mean   :0.2   Mean   :0.06   Mean   :0.24   Mean   :21.65   Mean   :3754
```

```
##  3rd Qu.:0.0   3rd Qu.:0.00   3rd Qu.:0.00   3rd Qu.:22.50   3rd Qu.:3982
##  Max.   :1.0   Max.   :1.00   Max.   :1.00   Max.   :25.50   Max.   :4492
##    avg_speed     num_entrants   num_finishers   first_prize_money
##  Min.   :33.41   Min.   :100.0   Min.   : 53.0   Min.   : 18006
##  1st Qu.:36.23   1st Qu.:150.0   1st Qu.: 97.0   1st Qu.: 49364
##  Median :38.93   Median :209.0   Median :135.5   Median :433754
##  Mean   :38.21   Mean   :186.1   Mean   :127.1   Mean   :290396
##  3rd Qu.:39.91   3rd Qu.:219.0   3rd Qu.:152.5   3rd Qu.:468621
##  Max.   :41.65   Max.   :229.0   Max.   :174.0   Max.   :520725
##  tot_prize_money   tot_time_winner  second_lag_time
##  Min.   : 482781   Min.   : 82.09   Min.   :0.00200
##  1st Qu.: 701118   1st Qu.: 87.70   1st Qu.:0.02800
##  Median :2104928   Median : 92.79   Median :0.07150
##  Mean   :1991232   Mean   : 98.55   Mean   :0.08872
##  3rd Qu.:3043916   3rd Qu.:109.07   3rd Qu.:0.12075
##  Max.   :3702938   Max.   :133.83   Max.   :0.29800
```

```r
# notice that the 'gen_ad' variable is constant throughout. This will create
# problems for our analysis, so let's drop it. Note that 'gen_ad' was not
# included in my description of the dataset above.
vars_to_remove <- c('gen_ad_test')
tdf <- tdf[, !(colnames(tdf) %in% vars_to_remove)]
```

```r
# let's run a linear regression of ped_tot on all the predictor variables
model1 <- lm(ped_tot ~., data = tdf)
# let's display a summary of the unrestricted linear regression using the
# stargazer function. The summary will display coefficient estimates and the
# standard error in brackets next to the coefficient estimate. P-values are
# indicated with asteriks. Everything is rounded to 2 digits.
stargazer(model1, type = 'latex', title = "Unrestricted OLS Regression",
header=FALSE, digits = 2, digits.extra = 10, intercept.bottom = FALSE,
single.row = TRUE, dep.var.labels   = "Percentage of cyclists
tested positive for PEDs", table.placement="H",
covariate.labels = c("Intercept", "Year", "Amphetamine Test",
                     "EPO Test", "Biological Passport",
                     "Night Test", "Out of Competition Testing",
                     "Number of Stages", "Total Length",
                     "Average Speed", "Number of Entrants",
                     "Number of Finishers", "First Prize Money",
                     "Total Prize Money", "Total Time Winner",
                     "Lag Time Between Winner and Runner Up"))
```

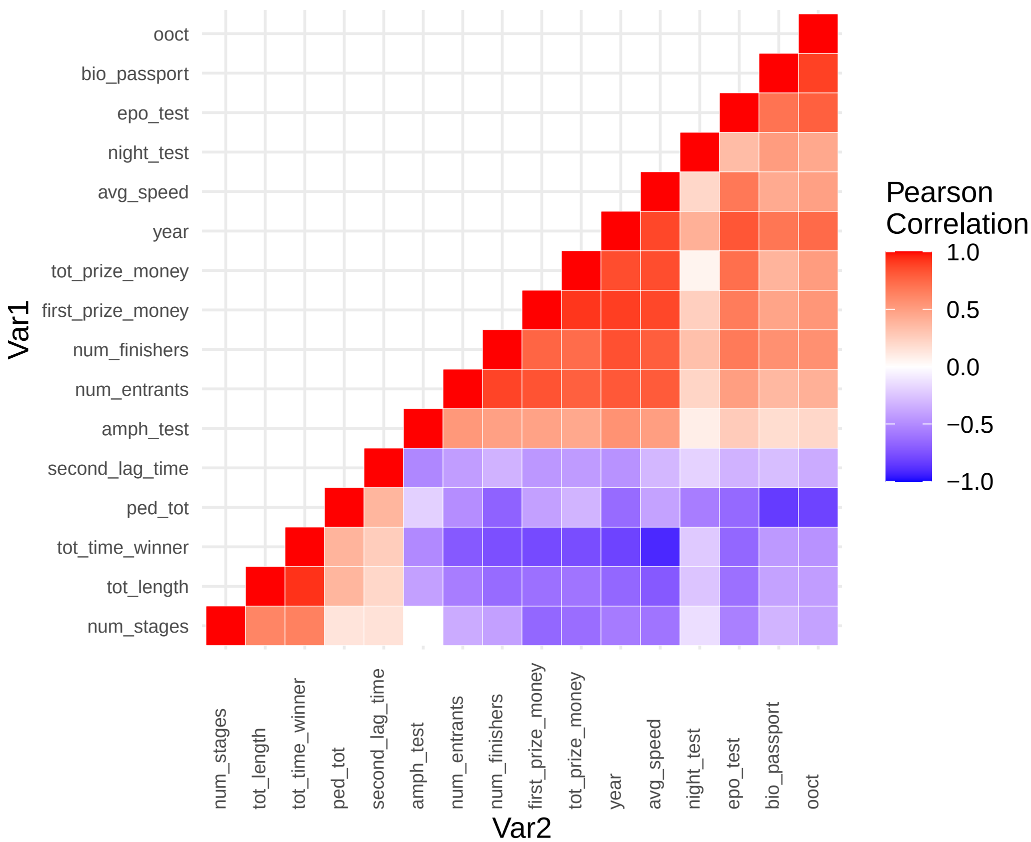
Table 1: Unrestricted OLS Regression

	Dependent variable:
	Percentage of cyclists tested positive for PEDs
Intercept	−2.11 (5.24)
Year	0.002 (0.003)
Amphetamine Test	0.06 (0.04)
EPO Test	−0.07* (0.04)
Biological Passport	−0.09** (0.03)
Night Test	−0.01 (0.03)
Out of Competition Testing	−0.09** (0.03)
Number of Stages	−0.02** (0.01)
Total Length	0.0001 (0.0001)
Average Speed	−0.01 (0.02)
Number of Entrants	−0.001 (0.001)
Number of Finishers	−0.002*** (0.001)
First Prize Money	−0.0000001 (0.0000001)
Total Prize Money	0.0000001*** (0.00000002)
Total Time Winner	−0.003 (0.004)
Lag Time Between Winner and Runner Up	0.33*** (0.10)
Observations	50
R^2	0.94
Adjusted R^2	0.91
Residual Std. Error	0.04 (df = 34)
F Statistic	36.04*** (df = 15; 34)
Note:	*p<0.1; **p<0.05; ***p<0.01

There are two potential issues with the linear regression. Firstly, there may be multicollinearity between predictor variables and secondly there may be too many variables for the number of observations which could lead to overfitting. So let's examine whether there is multicollinearity. If there is, we may be able to drop some variables to prevent overfitting.

```
# let's create a correlation matrix
heatmap <- function(df, vars) {
  cormat <- round(cor(na.omit(df)),2)
  # set up hierarchical clustering
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <- cormat[hc$order, hc$order]
  # remove redundant information
  cormat[lower.tri(cormat)] <- NA
  # melt cormat
  melted_cormat <- melt(cormat, na.rm = TRUE)
  # create matrix
  heat_plot <- ggplot(melted_cormat, aes(Var2, Var1, fill = value)) +
    geom_tile(color = "white") +
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
```

```r
                        name = "Pearson\nCorrelation") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0, size = 7,
                                     hjust = 0),
    axis.text.y = element_text(size = 7)) +
    coord_fixed()
    colnames(melted_cormat) <- c('Var1', 'Var2', 'correlation')
    melted_cormat <- melted_cormat[order(melted_cormat$Var1),]
    return(list(heat_plot = heat_plot, cormat = melted_cormat))
}
# create cormat for tdf dataset
heatmap(df = tdf, vars = colnames(tdf))$heat_plot
```




```r
# Just a quick look at the correlation matrix indicates that there may be
# quite severe multicolinearity between some variables.

# Now let's look at the Variable Inflation Factor for the unrestricted model
VIF(model1)
```

```
##              year         amph_test          epo_test      bio_passport
##         61.087819          6.763028         12.088262          7.680833
##        night_test              ooct        num_stages        tot_length
##          2.675428          7.591460          5.543018         44.738136
##         avg_speed      num_entrants     num_finishers first_prize_money
##         45.418148         18.704689          9.910155         18.211458
##   tot_prize_money   tot_time_winner   second_lag_time
##         16.225575        126.913025          2.220194
```

```r
# looks like some of these independent variables have severe VIF (i.e., > 10),
# so let's see if we can drop any. First, notice that (i) total prize amount
# (tot_prize_amount) and first prize amount (first_prize_amount) and
# (ii) number of entrants (num_entrants) and number of finishers (num_finishers)
# are pairs of variables that are likely very highly correlated such that one of
# the pair can be dropped. Let's confirm this by printing their correlations
# below
cor(tdf$tot_prize_money, tdf$first_prize_money)
```

```
## [1] 0.9178642
```

```r
cor(tdf$num_entrants, tdf$num_finishers)
```

```
## [1] 0.8825201
```

```r
# So we will drop one of the variables from each pair. I will drop the one with
# the higher VIF. Let's take a look at some other variables with very high VIF.
# Notice that year has a very high VIF. It will also be a problem if year is
# correlated with the variables of interest because then we will imprecisely
# estimate the coefficients on the variable of interest. Let's check if year is
# correlated with variables of interest.
cor(tdf[-1], tdf$year)
```

```
##                         [,1]
## ped_tot           -0.6362180
## amph_test          0.5629625
## epo_test           0.8206518
## bio_passport       0.6929589
## night_test         0.4114216
## ooct               0.7398777
## num_stages        -0.5672752
## tot_length        -0.6625301
## avg_speed          0.8743058
## num_entrants       0.8145740
## num_finishers      0.8447573
## first_prize_money  0.8985264
## tot_prize_money    0.8531097
## tot_time_winner   -0.8095467
## second_lag_time   -0.4696109
```

```r
# so year is severely correlated with epo_test and ooct which are variables of
# interest. Given that year also has one of the highest VIFs we will drop that
# too. Finally, I will also drop avg_speed  since it was unclear how this
# was measured and it also has a high VIF. Let's remove the variables now
model1_remove <- c('avg_speed', 'num_entrants', 'first_prize_money',
                   'tot_time_winner', 'year')
tdf <- tdf[, !(colnames(tdf) %in% model1_remove)]
```

```r
# let's run a linear regression on the restricted model
model2 <- lm(ped_tot ~., data = tdf)
# let's display a regular summary of this restricted OLS regression
# showing coefficient estimates, standard error, t-value, and P-values for
# all variables.
summary(model2)
```

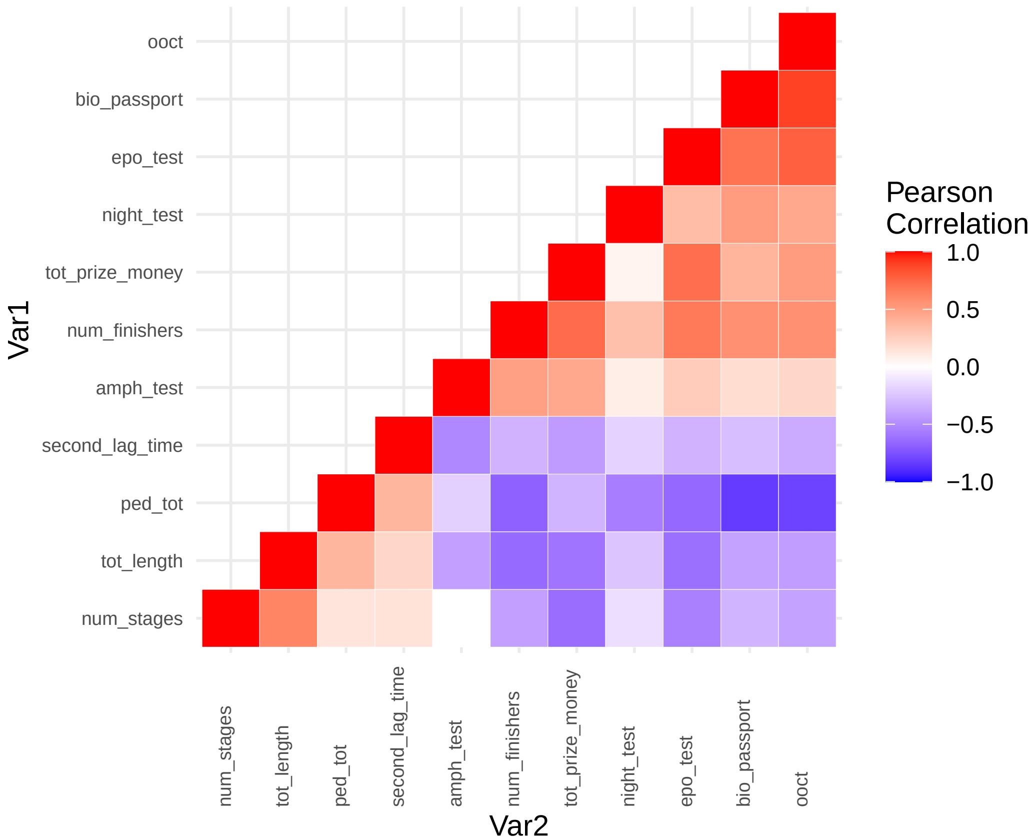
```
##
```

```
## Call:
## lm(formula = ped_tot ~ ., data = tdf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07017 -0.02248 -0.00172  0.02308  0.07161
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.470e-01  1.476e-01   6.417 1.37e-07 ***
## amph_test        8.394e-02  2.335e-02   3.595 0.000899 ***
## epo_test        -4.166e-02  2.334e-02  -1.785 0.082084 .
## bio_passport    -8.214e-02  2.981e-02  -2.756 0.008860 **
## night_test      -2.456e-02  2.656e-02  -0.925 0.360856
## ooct            -8.229e-02  3.068e-02  -2.682 0.010675 *
## num_stages      -2.436e-02  7.314e-03  -3.331 0.001902 **
## tot_length       2.890e-05  3.042e-05   0.950 0.347940
## num_finishers   -2.339e-03  3.054e-04  -7.659 2.72e-09 ***
## tot_prize_money  4.199e-08  1.030e-08   4.078 0.000217 ***
## second_lag_time  4.110e-01  9.215e-02   4.460 6.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0354 on 39 degrees of freedom
## Multiple R-squared:  0.9314, Adjusted R-squared:  0.9139
## F-statistic: 52.99 on 10 and 39 DF,  p-value: < 2.2e-16
```

```r
# let's look at VIF of the restricted model
VIF(model2)
```

```
##       amph_test        epo_test    bio_passport      night_test            ooct
##        2.297097        4.879392        5.673003        1.587879        6.852144
##      num_stages      tot_length   num_finishers tot_prize_money second_lag_time
##        2.894875        2.970654        3.562655        5.503620        1.693550
```

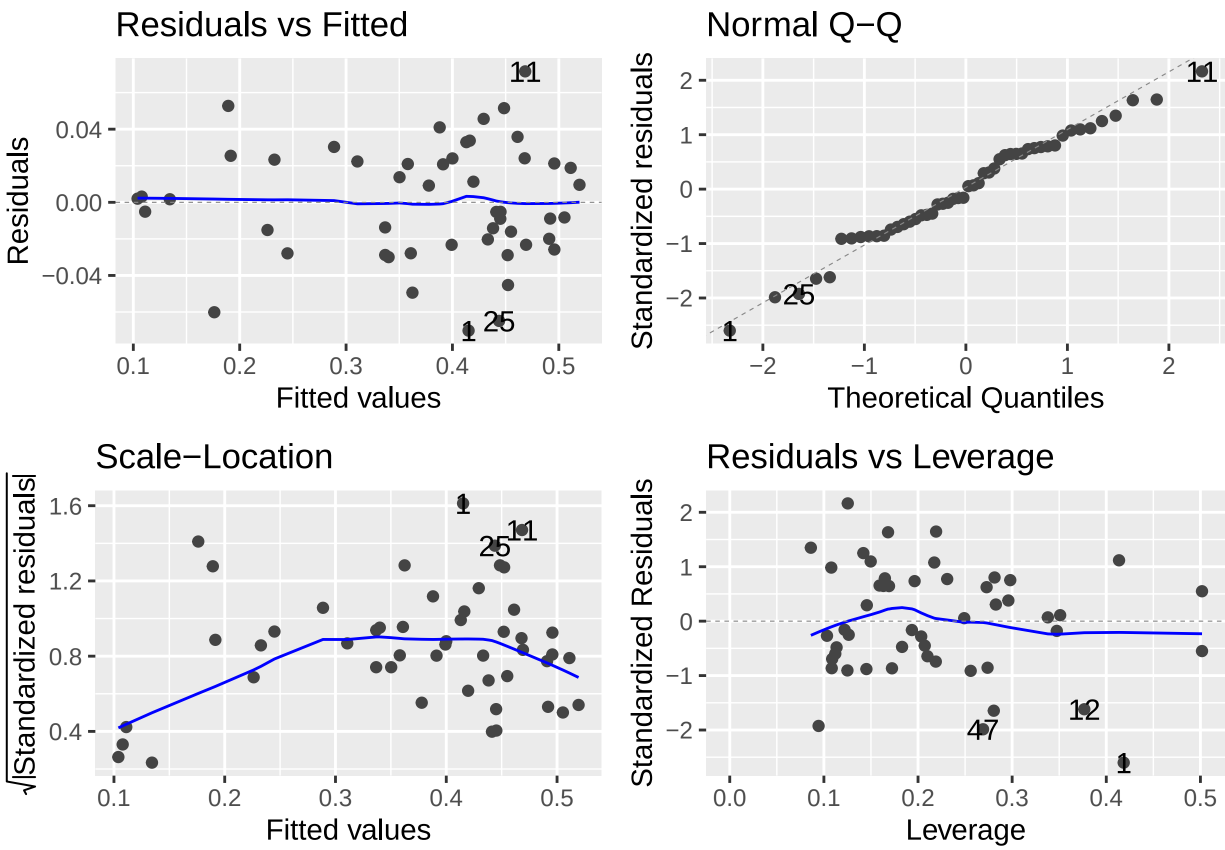
```r
# the VIF values look much better. They are all under 10 now. Let's create a
# correlation matrix again and then list all the correlations between variables
# that are higher than |0.7|
heatmap(df = tdf, vars = colnames(tdf))$heat_plot
```




```r
tdf_cor <- heatmap(df = tdf, vars = colnames(tdf))$cormat
for(i in 1:nrow(tdf_cor)) {
  if (abs(tdf_cor[i, 'correlation']) > 0.7 &
      tdf_cor[i, 'Var1'] != tdf_cor[i, 'Var2']) {
    print(as.name(paste(' ', tdf_cor[i, 'Var1'], 'and', tdf_cor[i, 'Var2'],
                        'have correlation', tdf_cor[i, 'correlation'])))
  }
}
```

```
## `  ped_tot and bio_passport have correlation -0.84`
## `  ped_tot and ooct have correlation -0.81`
## `  num_finishers and tot_prize_money have correlation 0.73`
## `  tot_prize_money and epo_test have correlation 0.72`
## `  epo_test and ooct have correlation 0.78`
## `  bio_passport and ooct have correlation 0.89`
```

```r
# looks like we have 0.78 cor between epo_test and ooct and 0.89 between
# bio_passport and ooct, which means that these coefficients could be
# imprecisely estimated. I will ignore this potential issue for now.
# Let's do model diagnostics for the restricted model
autoplot(model2)
```




```
# it appears that there is heteroskedasticity (non-horizontal line on bottom
# left plot), so let's get heteroskedastic robust standard errors using a
# function from the 'sandwich' library.
se1 <- vcovHC(model2, type = "HC1")
robust_se1 <- sqrt(diag(se1))
# let's display a summary of the unrestricted linear regression using the
# stargazer function again.
stargazer(model2, type = 'latex', title = "Restricted OLS Regression",
header=FALSE, se = list(NULL, robust_se1), digits = 2, digits.extra = 10,
intercept.bottom = FALSE, single.row = TRUE,
dep.var.labels = "Percentage of cyclists tested positive for PEDs",
table.placement="H", covariate.labels = c("Intercept", "Amphetamine Test",
                                         "EPO Test", "Biological Passport",
                                         "Night Test", "Out of Competition Testing",
                                         "Number of Stages", "Total Length",
                                          "Number of Finishers", "Total Prize Money",
                                          "Lag Time Between Winner and Runner Up"))
```

Table 2: Restricted OLS Regression

	Dependent variable:
	Percentage of cyclists tested positive for PEDs
Intercept	0.95*** (0.15)
Amphetamine Test	0.08*** (0.02)
EPO Test	−0.04* (0.02)
Biological Passport	−0.08*** (0.03)
Night Test	−0.02 (0.03)
Out of Competition Testing	−0.08** (0.03)
Number of Stages	−0.02*** (0.01)
Total Length	0.00003 (0.00003)
Number of Finishers	−0.002*** (0.0003)
Total Prize Money	0.00000004*** (0.00)
Lag Time Between Winner and Runner Up	0.41*** (0.09)
Observations	50
R^2	0.93
Adjusted R^2	0.91
Residual Std. Error	0.04 (df = 39)
F Statistic	52.99*** (df = 10; 39)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
# One issue with the OLS regression is that the dependent variable is bounded
# between 0 and 1, which means the OLS regression might imprecisely estimate the
# standard errors of coefficients and give predictions of the dependent variable
# that are above 1 or below 0. This is less of an issue if most of the data from
# the dependent variable is not close to the boundary, which is the case with
# ped_tot as we can see from the summary statistic table below.
summary(tdf$ped_tot)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1060  0.3145  0.4125  0.3729  0.4490  0.5400
```

```r
# Nevertheless, let's run a fractional logistic regression to cover our bases.
# We will run the regression on the same set of variables that we used in the
# previous OLS regression.
logistic1 <- glm(ped_tot~., data = tdf, family = quasibinomial('logit'))
```

```r
# Let's also get standard error estimates that are heteroskedastic robust
se2 <- vcovHC(logistic1, type="HC1")
robust_se2 <- sqrt(diag(se2))
# Let's print the results of the fractional logistic model in a table
# using stargazer.
stargazer(logistic1, type = 'latex', title = "Fractional Logistic Regression",
header=FALSE, se = list(NULL, robust_se2), digits = 2, digits.extra = 10,
intercept.bottom = FALSE, single.row = TRUE,
dep.var.labels = "Percentage of cyclists tested positive for PEDs",
table.placement="H", covariate.labels = c("Intercept", "Amphetamine Test",
                                          "EPO Test", "Biological Passport",
                                          "Night Test", "Out of Competition Testing",
                                          "Number of Stages", "Total Length",
                                           "Number of Finishers", "Total Prize Money",
```

```r
                                        "Lag Time Between Winner and Runner Up"))
```

Table 3: Fractional Logistic Regression

	Dependent variable:
	Percentage of cyclists tested positive for PEDs
Intercept	1.85*** (0.67)
Amphetamine Test	0.34*** (0.10)
EPO Test	−0.19* (0.10)
Biological Passport	−0.48*** (0.14)
Night Test	−0.42** (0.17)
Out of Competition Testing	−0.35** (0.14)
Number of Stages	−0.10*** (0.03)
Total Length	0.0001 (0.0001)
Number of Finishers	−0.01*** (0.001)
Total Prize Money	0.0000002*** (0.00000005)
Lag Time Between Winner and Runner Up	1.74*** (0.41)
Observations	50
Note:	*p<0.1; **p<0.05; ***p<0.01

Now let's take a different approach. Instead of choosing what predictor variables to include in our model, let's try an automatic method that selects predictor variables for us based on an algorithm.

We will use LASSO to do this.

```r
# First we load the dataset again
tdf <- read.csv("/Users/martrinmunoz/Desktop/EconPredoc/Writing Samples/TdF/tdf_cleaned.csv")
#Let's remove the 'gen_ad' variable again
vars_to_remove <- c('gen_ad_test')
tdf <- tdf[, !(colnames(tdf) %in% vars_to_remove)]
# Now we create a training set
X_train <- model.matrix(ped_tot~., data = tdf)[,-1]
Y_train <- tdf$ped_tot
# We create a list that will store the mean-squared errors from cross-fold
# validation
MSEs <- NULL
# Now we run LASSO using 5-fold cross-validation and we use a for loop to repeat
# this algorithm 100 times to try and guard against the stochastic nature of the
# algorithm, which is especially a problem when the dataset is small as it is
# here. Since the dataset is small we also don't separate out a training and
# testing set. We just use the whole dataset to train the model. Alpha=1
# indicates that this is a LASSO regression.
# Each time the loop is run, we append MSEs to the MSE list. The LASSO algorithm
# estimates the lambda parameter that results in the lowest MSE. The lambda
# parameter is a penalty term that is used in generating the model.
for (i in 1:100){
  cv <- cv.glmnet(x = X_train, y = Y_train, alpha=1, nfolds=5,
                  standardize = TRUE)
  MSEs <- cbind(MSEs, cv$cvm)
```

```r
}
# we name the rows of the MSE list based on the lambda estimated through the
# LASSO model
rownames(MSEs) <- cv$lambda
# finally, we choose the model based on the lambda
# that is the minimum lambda plus 1 standard error
model3 <- coef(cv, s = cv$lambda.1se)
# Note that we cannot immediately represent the object 'model3' through
# stargazer because it is a 'dgCMatrix' as the next line of code shows.
class(model3)
```

```
## [1] "dgCMatrix"
## attr(,"package")
## [1] "Matrix"
```

```r
# It is possible to convert it into an object appropriate for Stargazer, but
# this process is involved. Let's just present it in a nice table. First, we
# convert the dgCMatrix object into a data frame.
model3 <-data.frame(Variable_Name = model3@Dimnames[[1]][model3@i + 1],
                    Variable_Estimate = model3@x)
# Let's rename the rows in the data frame
model3[,1] <- c("Intercept", "Amphetamine Test", "EPO Test", "Biological Passport",
                "Night Test", "Out of Competition Testing", "Number of Stages",
                "Number of Finishers", "Total Prize Money",
                "Lag Time Between Winner and Runner Up")
# Let's also round the coefficient estimates to two significant figures
model3[,2] <- signif(model3[,2],2)
# Now we represent the data frame using the kable function from the knitr package.
# We have two columns: one shows the name of the variable and the other shows the
# coefficient estimate. Note that some variables are not included. This means
# they were dropped by the LASSO regression.
knitr::kable(model3, caption = "LASSO Regression", format = "markdown",
             digits = 10)
```

Table 4: LASSO Regression

Variable_Name	Variable_Estimate
Intercept	8.9e-01
Amphetamine Test	3.3e-02
EPO Test	-9.5e-03
Biological Passport	-1.0e-01
Night Test	-4.4e-02
Out of Competition Testing	-8.0e-02
Number of Stages	-1.6e-02
Number of Finishers	-1.7e-03
Total Prize Money	2.0e-08
Lag Time Between Winner and Runner Up	2.0e-01