Lab02

Minh Tam Hoang

9/27/2019

```
#' Return the predictions from the linear models
#' @param model_fit
#' @param testData a compatible new data
#' @return

prediction_reg <- function(model_fit, testData){
  return(predict(model_fit,testData))
}

#'  Fit linear models and return the predictions from them.
#' @param y target variable
#' @param x covariate
#' @param x_2 covariate
#' @param data data set
#' @param type type of regression
#' @param horizon a compatible new data
model_fit <- function(y, x,x_2, data, type, horizon){


  if(!missing(x_2) && type == "MLR"){

    formula <- paste0(y, "~", x, "+", x_2)
    reg <- lm(as.formula(formula), data = data)
  }else{
    if(type == "SLR"){
      formula <- paste0(y, "~", x)
      reg <- lm(as.formula(formula), data = data)
    }else{
      formula <- paste0(y, "~", x, "+", "I(", x,"^2) + ", "I(", x, "^3)")
      reg <- lm(as.formula(formula), data = data)

    }
  }
  prediction_reg(reg,horizon)
}
```

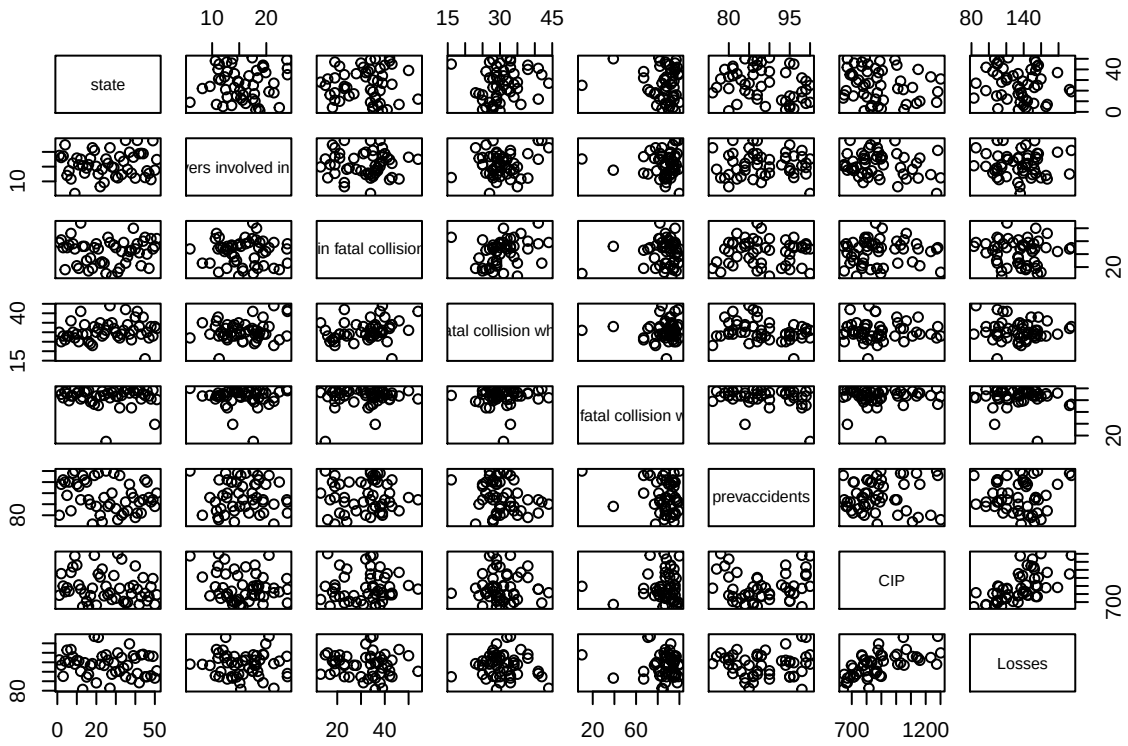
```
library(ggplot2)
bad_drivers_data <- read.csv('./data/bad-drivers.csv')
head(bad_drivers_data)
```

```
##        State
## 1    Alabama
## 2     Alaska
## 3    Arizona
## 4   Arkansas
## 5 California
## 6   Colorado
##   Number.of.drivers.involved.in.fatal.collisions.per.billion.miles
```

```
## 1                                                                 18.8
## 2                                                                 18.1
## 3                                                                 18.6
## 4                                                                 22.4
## 5                                                                 12.0
## 6                                                                 13.6
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Speeding
## 1                                                                  39
## 2                                                                  41
## 3                                                                  35
## 4                                                                  18
## 5                                                                  35
## 6                                                                  37
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Alcohol.Impaired
## 1                                                                          30
## 2                                                                          25
## 3                                                                          28
## 4                                                                          26
## 5                                                                          28
## 6                                                                          28
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Were.Not.Distracted
## 1                                                                       96
## 2                                                                       90
## 3                                                                       84
## 4                                                                       94
## 5                                                                       91
## 6                                                                       79
##   Percentage.Of.Drivers.Involved.In.Fatal.Collisions.Who.Had.Not.Been.Involved.In.Any.Previous.Accid
## 1
## 2
## 3
## 4
## 5
## 6
##   Car.Insurance.Premiums....
## 1                     784.55
## 2                    1053.48
## 3                     899.47
## 4                     827.34
## 5                     878.41
## 6                     835.50
##   Losses.incurred.by.insurance.companies.for.collisions.per.insured.driver....
## 1                                                                       145.08
## 2                                                                       133.93
## 3                                                                       110.35
## 4                                                                       142.39
## 5                                                                       165.63
## 6                                                                       139.91
```

```r
names(bad_drivers_data) <- c("state", "Number of drivers involved in fatal collisions",
                             "%of drivers involved in fatal collision who were speeding",
                             "%of drivers involved in fatal collision who were alcohol impaired",
                             "%of drivers involved in fatal collision who were not distracted",
                             "prevaccidents",
```

```
                                "CIP","Losses")
plot(bad_drivers_data)
```



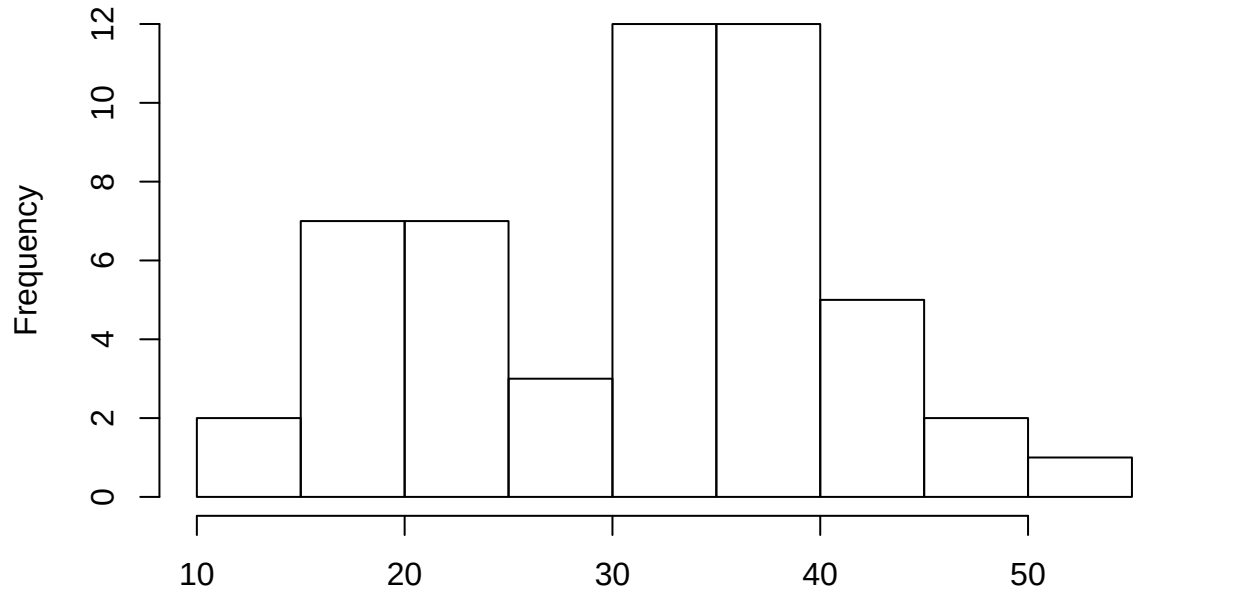

```
cor(x = bad_drivers_data$Losses, y = bad_drivers_data$CIP)
```

```
## [1] 0.6231164
```

```
cor(x = bad_drivers_data$prevaccidents, y = bad_drivers_data$CIP)
```

```
## [1] 0.07553314
```

```
# Plot the estimated distribution of _Percentage of Drivers Involved in Fatal Collisions who were speed
hist(bad_drivers_data$`%of drivers involved in fatal collision who were speeding`)
```




```
reg01 <- lm(bad_drivers_data$CIP~bad_drivers_data$Losses)

reg02 <- lm(bad_drivers_data$CIP~bad_drivers_data$Losses+bad_drivers_data$prevaccidents)

reg03 <- lm(bad_drivers_data$CIP~bad_drivers_data$Losses + I(bad_drivers_data$Losses^2) + I(bad_drivers_

summary(reg01)
```

```
##
## Call:
## lm(formula = bad_drivers_data$CIP ~ bad_drivers_data$Losses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -213.33  -96.75  -40.11  112.24  379.97
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              285.3251   109.6689   2.602   0.0122 *
## bad_drivers_data$Losses    4.4733     0.8021   5.577 1.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 140.9 on 49 degrees of freedom
## Multiple R-squared:  0.3883, Adjusted R-squared:  0.3758
## F-statistic:  31.1 on 1 and 49 DF,  p-value: 1.043e-06
```

```
summary(reg02)
```

```
##
```

```
## Call:
## lm(formula = bad_drivers_data$CIP ~ bad_drivers_data$Losses +
##     bad_drivers_data$prevaccidents)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -209.9 -101.3  -41.9  109.8  365.9
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   176.0405   274.8870   0.640    0.525
## bad_drivers_data$Losses         4.4583     0.8096   5.507 1.41e-06 ***
## bad_drivers_data$prevaccidents  1.2545     2.8889   0.434    0.666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 142 on 48 degrees of freedom
## Multiple R-squared:  0.3907, Adjusted R-squared:  0.3653
## F-statistic: 15.39 on 2 and 48 DF,  p-value: 6.863e-06
```

```
summary(reg03)
```

```
##
## Call:
## lm(formula = bad_drivers_data$CIP ~ bad_drivers_data$Losses +
##     I(bad_drivers_data$Losses^2) + I(bad_drivers_data$Losses^3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -225.35  -99.12  -41.12  117.53  379.40
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.270e+03  1.890e+03   0.672    0.505
## bad_drivers_data$Losses     -1.861e+01  4.252e+01  -0.438    0.664
## I(bad_drivers_data$Losses^2)  1.745e-01  3.110e-01   0.561    0.577
## I(bad_drivers_data$Losses^3) -4.260e-04  7.401e-04  -0.576    0.568
##
## Residual standard error: 143.3 on 47 degrees of freedom
## Multiple R-squared:  0.3929, Adjusted R-squared:  0.3541
## F-statistic: 10.14 on 3 and 47 DF,  p-value: 2.896e-05
```

```
coef(reg01)
```

```
##             (Intercept) bad_drivers_data$Losses
##              285.325089                4.473333
```

```
coef(reg02)
```

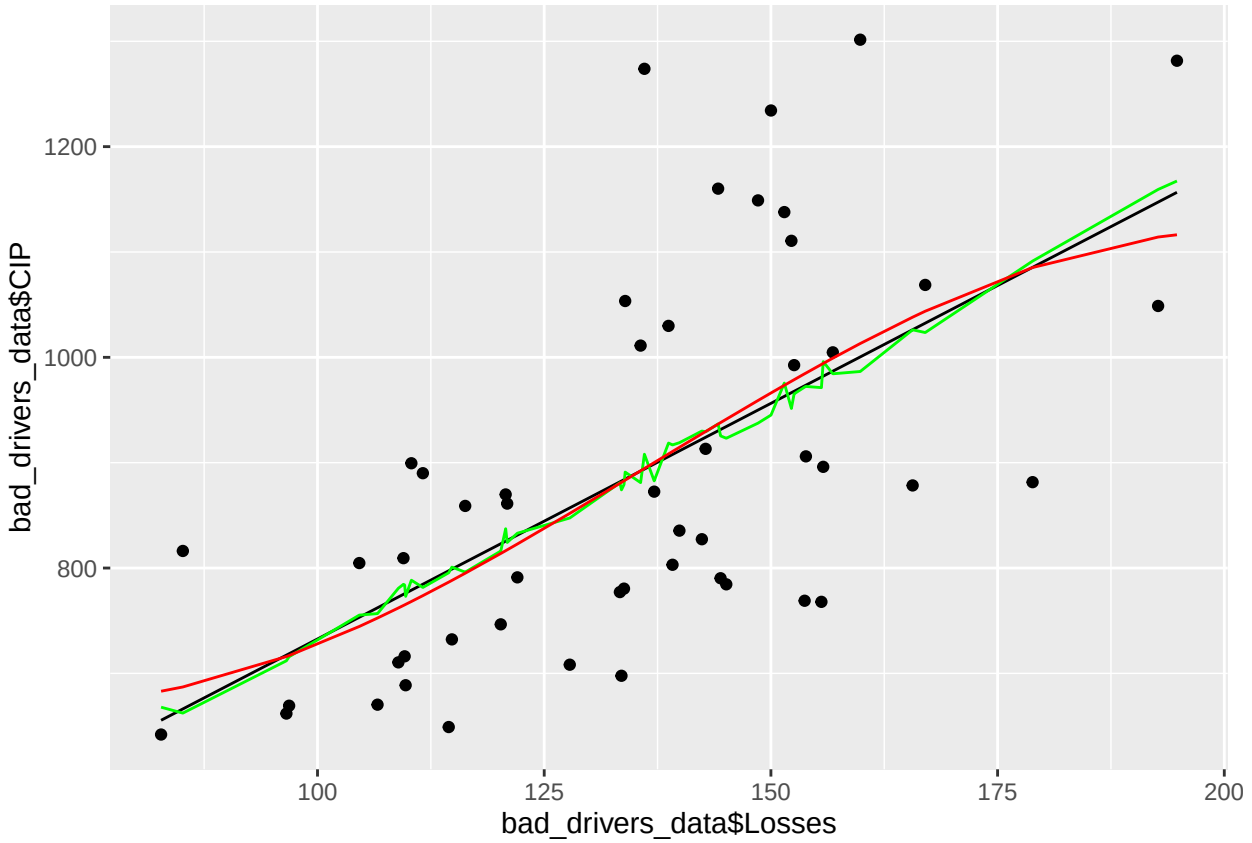
```
##                     (Intercept)        bad_drivers_data$Losses
##                      176.040549                      4.458296
## bad_drivers_data$prevaccidents
##                        1.254508
```

```
coef(reg03)
```

```
##                   (Intercept)      bad_drivers_data$Losses
```

```
##                 1.269960e+03                -1.861436e+01
## I(bad_drivers_data$Losses^2) I(bad_drivers_data$Losses^3)
##                 1.744903e-01                -4.259848e-04
```

```
ggplot2::ggplot()+
  geom_point(mapping = aes(x = bad_drivers_data$Losses, y = bad_drivers_data$CIP))+
  geom_line(mapping = aes(x = bad_drivers_data$Losses, y = fitted(reg01)))+
  geom_line(mapping = aes(x = bad_drivers_data$Losses, y = fitted(reg02)), col = "green")+
  geom_line(mapping = aes(x = bad_drivers_data$Losses, y = fitted(reg03)), col = "red")
```



Hold-out

```
random <- sample(1:nrow(bad_drivers_data), size = 10)
hold_out <- bad_drivers_data[random,]
head(hold_out)
```

```
##             state Number of drivers involved in fatal collisions
## 34 North Carolina                                           16.8
## 2          Alaska                                           18.1
## 40   Rhode Island                                           11.1
## 35   North Dakota                                           23.9
## 7     Connecticut                                           10.8
## 51        Wyoming                                           17.4
##    %of drivers involved in fatal collision who were speeding
## 34                                                        39
## 2                                                         41
## 40                                                        34
## 35                                                        23
## 7                                                         46
```

```
## 51                                                          42
##    %of drivers involved in fatal collision who were alcohol impaired
## 34                                                                31
## 2                                                                 25
## 40                                                                38
## 35                                                                42
## 7                                                                 36
## 51                                                                32
##    %of drivers involved in fatal collision who were not distracted
## 34                                                              94
## 2                                                               90
## 40                                                              92
## 35                                                              99
## 7                                                               87
## 51                                                              81
##    prevaccidents     CIP Losses
## 34            81  708.24 127.82
## 2             94 1053.48 133.93
## 40            79 1148.99 148.58
## 35            86  688.75 109.72
## 7             82 1068.73 167.02
## 51            90  791.14 122.04
```

```
train <- bad_drivers_data[-random,]
head(train)
```

```
##        state Number of drivers involved in fatal collisions
## 1    Alabama                                           18.8
## 3    Arizona                                           18.6
## 4   Arkansas                                           22.4
## 5 California                                           12.0
## 6   Colorado                                           13.6
## 8   Delaware                                           16.2
##   %of drivers involved in fatal collision who were speeding
## 1                                                        39
## 3                                                        35
## 4                                                        18
## 5                                                        35
## 6                                                        37
## 8                                                        38
##   %of drivers involved in fatal collision who were alcohol impaired
## 1                                                                30
## 3                                                                28
## 4                                                                26
## 5                                                                28
## 6                                                                28
## 8                                                                30
##   %of drivers involved in fatal collision who were not distracted
## 1                                                              96
## 3                                                              84
## 4                                                              94
## 5                                                              91
## 6                                                              79
## 8                                                              87
##   prevaccidents     CIP Losses
```

```
## 1            80  784.55 145.08
## 3            96  899.47 110.35
## 4            95  827.34 142.39
## 5            89  878.41 165.63
## 6            95  835.50 139.91
## 8            99 1137.87 151.48
```

```
REG_01 <- lm(train$CIP~ train$Losses)

REG_02  <- lm(train$CIP~train$Losses+train$prevaccidents)

REG_03 <- lm(train$CIP~train$Losses + I(train$Losses^2)+ I(train$Losses^3))
m <- 0
for (type in c("SLR", "MLR", "CR")){



    pred_val <- model_fit("CIP"
                                ,"Losses"
                                ,"prevaccidents"
                                ,data = train
                                ,type = type
                                ,horizon = hold_out)

    MSE <- mean((hold_out$CIP - pred_val)^2)

   print(type)
   print(MSE)

   m = m+1
}
```

```
## [1] "SLR"
## [1] 28505.6
## [1] "MLR"
## [1] 35704.15
## [1] "CR"
## [1] 28876.56
```

Cross_validation

```
# Split your data into 5 training/testing sets
train_test <- list()
j = 1
for( i in 1:4){

  train_test[[i]] <- bad_drivers_data[j:(j+9),]
  j <- j+10
}
train_test[[5]] <- tail(bad_drivers_data, 11)

#Obtain a list containing train data and a list containing hold-out data
train_valid <- list()
hold_out_set <- list()
train_set <- list()
```

```
for(i in 1 : length(train_test)){
  train_test_f <- train_test
  hold_out_set[[i]] <- train_test[[i]]

  train_test_f[[i]] <- NULL
  train_valid[[i]] <- train_test_f

  for( j in 1: length(train_valid[[i]])){
    a <- train_valid[[i]]
    a <-rbind.data.frame(a[[1]], a[[2]], a[[3]], a[[4]])

  }
  train_set[[i]] <- a
}




# Create an empty data.frame called `crossValResults` that has 3 columns (one for each model) and 5 row
crossValResults <- as.data.frame(matrix(NA, nrow = 5, ncol = 3))
colnames(crossValResults) <- c("SLR", "MLR", "CR")
k <- 1
for (model_type in c("SLR", "MLR", "CR")){

  for( i in 1:length(train_set)){

      prediction_val <- model_fit("CIP"
                                  ,"Losses"
                                  ,"prevaccidents"
                                  ,data = train_set[[i]]
                                  ,type = model_type
                                  ,horizon = hold_out_set[[i]])

      MSE <- mean((hold_out_set[[i]]$CIP - prediction_val)^2)


      crossValResults[i,k] <- MSE

  }
  k = k+1
}

print(crossValResults)
```

```
##        SLR       MLR       CR
## 1 35386.66 43554.736 36780.99
## 2 10042.99  9951.381 19457.30
## 3 14242.15 17481.728 15270.90
## 4 32705.55 39891.888 33274.10
## 5 11891.89 11684.206 14428.63
```

```
CV_errors <- c(mean(crossValResults[,1]), mean(crossValResults[,2]), mean(crossValResults[,3]))
print(" CV scores for SLR, MLR, CB:")
```

```
## [1] " CV scores for SLR, MLR, CB:"
```

```
print(CV_errors)
```

```
## [1] 20853.85 24512.79 23842.38
```

Brief Report

Paragraph 1: Describe the dataset

In this lab, we worked on the dataset on driving incidents that was collected for all 52 states in the States. This data set contains information on insurance pay-outs, insurance charges, and number of car crashes due to speeding, alcohol, etc... across the country. There are 51 samples and 7 feature-variables in this data set. The purpose of this lab is to explore whether or not state-level factors influence car insurance premiums in the States.

Paragraph 2: Describe the models you'll use to make predictions.

a) Exploratory data analysis:

According to the draftsman plot, I notice that there exists a faily strong linear relationship between CIPS and losses incurred by insurance companies for collisions per insured driver. CIPS and Percentage Of Drivers Involved In Fatal Collisions Who Had Not Been Involved In Any Previous Accidents are positively related. No obvious, clear trends are observed in other pairs of variables. The distribution of the covariate as an explanatory variable has no effect on the model fit since covariates are not random variables and are fixed.

b) Regression analysis

The target variable is CIPS, and we decided to choose 'losses incurred by insurance companies for collisions per insured driver' as a covariate in our regression (since the value of correlation of CIPS and losses is largest, it seems that 'losses' is most related to the target variable)

+) Single linear regression: As the variable 'Losses' does relate to y linearly, fitting a straight line to this regression can capture the relationship between these two variables. Based on this single linear regression, it is estimated that each increase in mean 'Losses' of one dollar is associated with an increase of about 4.473 in the mean CIP. Also, a hypothesis test indicates that there exists an association/relationship between mean 'losses' and mean 'CIP'(since p-value for the test is much smaller than the significance level, providing evidence against the null hypothesis). The squared correlation R^2 is 0.3758, which indicates that the inclusion of 'losses' in the regression accounts for 37.58% of the variability in the data.

Q:How does your multiple regression model compare the your simple linear regression, and how would communicate these results to an audience?

The adjusted R-square is 36.53 and only one variable, 'losses' is significant by t-test. The inclusion of both 'losses' and percentage of drivers involved in fatal collision who were had not been involved in previous accidents accounts for 36.53% of the variability in the data, which does not show any improvement over the smaller model.

As it is difficult and unconvincing to visually assess the model fit, we use the test metric (MSE) to evaluate the model fits and determine which model fit has a better performance in predicting the target variable.

Paragraph 3: Summarize the methods you'll use to decide which model is best. Make sure to define and describe the test metric (MSE)

Hold-out

In this method, we randomly selected and removed 10 states from the data set and stored them in "Hold-out set". Also, we stored the remaining states in "training". We re-trained single linear regression, multiple linear regression and cubic regression on the training set and computed and evaluated the MSE on the hold-out set.

The model with the smallest MSE is the single linear regression, which means that the difference between model predictions made by the single linear model and the empirical data is the smallest. Hence, single linear model shows a better performance in predicting the CIPs in comparison with the multiple linear regression and cubic regression models.

Cross-validation

In this method, we splitted the data set into 5 pieces of data. Then, we trained the models on four pieces of the data, made predictions on the test-set, and computed the MSE on each test set. For each model, we averaged the MSEs over the test sets and obtained the CV-scores.

How does the CV error compare to the hold-out error?

Since we randomly split the data into train and hold-out sets in the hold-out method, it is hard to tell whether or not CV error is smaller than the hold-out error. This is because the estimate of test error in the hold-out method is heavily dependent on how we divide the data.

Cross-validation is more reliable than hold-out method in evaluating and assessing the model fit. This is because in the hold out method, we randomly split our data set into training set and test set. The evaluation greatly depend on the end points of training data and test data, which means that the estimates of test error are significantly different due to how the divisions of the data is made. Meanwhile, in cross-validation method, we split the data set into x pieces, and repeat the hold-out method x times. Thus, the evaluation no long depends on how the data is divided, and therefore, the variance of the estimate will decrease.

How does the Cross-validation MSE compare between your simple and multiple regression?

The Cross-validation MSE of simple regression is slightly smaller that of multiple regression, which suggests that the simple linear model gives a better fit than the bigger model. We might want to try multiple regression with different covariates than the precentage of drivers involved in previous accidents to see if there exist some covariates that contribute significantly to the prediction of CIPs in combination with Losses incurred by insurance companies for collisions per insured driver.