Contents

[Linear Regression 1](#_Toc517179687)

[Workflow 1](#_Toc517179688)

[Import Data 1](#_Toc517179689)

[Standardize 2](#_Toc517179690)

[Dummy Variables 2](#_Toc517179691)

[Model 2](#_Toc517179692)

[Find Outliers 2](#_Toc517179693)

[Cook’s Distance 2](#_Toc517179694)

[Histogram 2](#_Toc517179695)

[P-Value 2](#_Toc517179696)

[RMSE (Root Mean Sqaure Error) 4](#_Toc517179697)

# Linear Regression

## Workflow

1. Understand Data
2. Created a fitted model and evaluate prediction quality
3. Simplify the model (by removing some predictors)
4. Locate and remove outliers
5. Predict Reponses to new data

Summary

1. Measuring overall accuracy: RMSE ˅ R2 ˄
2. Identifying outliers: Leverage, Cook’s distance, Probability plot, residual histogram
3. Identifying useful variables: p-value, Added-Variable plot, plotSlice

## Import Data

#### Matlab

load patients

X = [X\_norm,dummy\_gender(:,end), Smoker,dummy\_location, dummy\_health];

#### R

patients <- read.csv("~/Desktop/SEIS 763/patients.csv")

X = patients[,-2]

X = X[,-4]

## Standardize

#### Matlab

X\_pre = [Age, Height, Weight];

X\_norm = zscore(X\_pre);

#### R

X$Age =scale(X$Age)

#### Python

preprocessing.sale(X\_train)

## Dummy Variables

#### Matlab

location\_group = nominal(Location);

dummy\_location = dummyvar(location\_group);

#### R

#### Python

Pandas.get\_dummies(x)

## Model

#### Matlab

lm = fitlm(X, y)

plot(lm)

#### R

fit = lm(Systolic~., X)

summary(fit)

## Find Outliers

Use various plot residuals methods

### Cook’s Distance

A point likely be an outlier IF it’s cook’s distance > (3 X average Cook’s distance).

The dashed line is 3xAvg of Cooks Distance

#### Matlab

plotDiagnostics(lm, 'cookd')

#### R

cooksd = cooks.distance(fit)

plot(cooksd, col="lightblue", pch=19, cex=2)

text(cooksd, labels=rownames(X), cex=0.9, font=2)

### Histogram

#### Matlab

plotResiduals(lm)

R

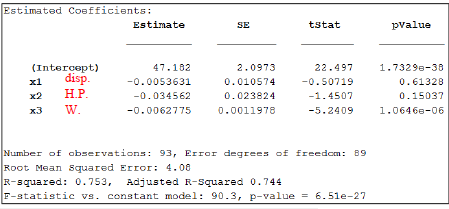
### P-Value

If p Value is very small 🡪 good predictor to y

If p value > 0.01 🡪 not a good predictor to y.

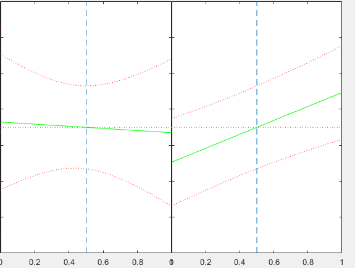
#### Matlab

plot(lm)



Disp has a very high p-value so it might an outlier

plotSlice(lm)

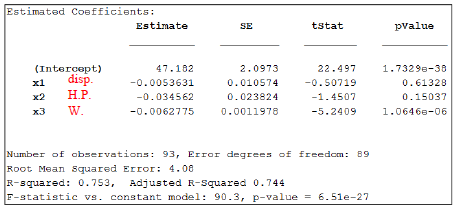


If the line is straight then it has no impact on Y so we don’t have to use that predictor whereas if the line is at an angle it has a significant impact on Y.

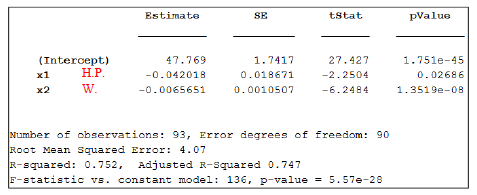
### RMSE (Root Mean Sqaure Error)

After removing outliers you should see if it effects the RMSE which will tell you if it was effective. It should low the RMSE.

Before removing disp and outlier

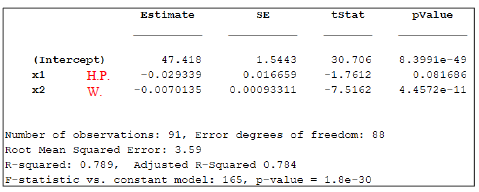


Remove only disp



There seems to be no change at all.

After removing disp and outlier



# Regularization

## Cross validation

1. Prepare original data
2. Divide data into training and test sets
3. Build a model using training data
4. Validate the model using test set

