Homework 01

Zixin Ouyang

Exercise 1

```
hw01_data<-read.csv('hw01-data.csv')
set.seed(42)
train_index = sample(1:nrow(hw01_data), size = round(0.5 * nrow(hw01_data)))
train_data = hw01_data[train_index, ]
test_data = hw01_data[-train_index, ]
```

```
rmse = function(actual, predicted) {
  sqrt(mean((actual - predicted) ^ 2))
}
```

```
get_rmse = function(model, data, response) {
  rmse(actual = data[, response],
       predicted = predict(model, data))
}
```

```
get_complexity = function(model) {
  length(coef(model))
}
```

```
mod_1<-lm(y~., data=train_data)
mod_2<-lm(y~. + I(a ^ 2) + I(b ^ 2) + I(c ^ 2), data=train_data)
mod_3<-lm(y~. ^ 2 + I(a ^ 2) + I(b ^ 2) + I(c ^ 2), data=train_data)
mod_4<-lm(y~a * b * c * d + I(a ^ 2) + I(b ^ 2) + I(c ^ 2), data=train_data)
```

```
predicted = predict(mod_1, data=train_dat)
```

```
model_list = list(mod_1, mod_2, mod_3, mod_4)

trn_rmse = sapply(model_list, get_rmse, data = train_data, response = "y")
tst_rmse = sapply(model_list, get_rmse, data = test_data, response = "y")
model_complexity = sapply(model_list, get_complexity)
```

Model	Train RMSE	Test RMSE	Predictors
mod_1	1.4381782	1.4286911	5
mod_2	1.1242482	1.1526319	8
mod_3	0.5105619	0.5206716	14
mod_4	0.5082713	0.5211251	19

Based on these results, Model 3 is the best model for prediction.

Exercise 2

```
library(tibble)
library(readr)
library(MASS)
```

```r
data(Boston)
Boston = as_tibble(Boston)
```

```r
set.seed(42)
boston_index = sample(1:nrow(Boston), size = 400)
train_boston = Boston[boston_index, ]
test_boston  = Boston[-boston_index, ]
```

```r
fit = lm(medv ~ . ^ 2, data = train_boston)
fit_smaller<-lm(medv ~ ., data = train_boston)
fit_larger<-lm(medv ~ . ^ 2 + crim:zn:rm, data = train_boston)
```

```r
models = list(fit_smaller, fit, fit_larger)

train_rmse = sapply(models, get_rmse, data = train_boston, response = "medv")
test_rmse = sapply(models, get_rmse, data = test_boston, response = "medv")
nparameters = sapply(models, get_complexity)
```

Model	Train RMSE	Test RMSE	Predictors
fit_smaller	4.6084282	5.0538877	14
fit	2.6239518	3.2202003	92
fit_larger	2.5984973	3.3013487	93

Exercise 3

```r
newtrain1=train_boston[abs(rstandard(fit)) <= 2,]
newtrain2=train_boston[abs(rstandard(fit)) <= 3,]
```

```r
fit_2=lm(medv ~ . ^ 2, data = newtrain1)
fit_3=lm(medv ~ . ^ 2, data = newtrain2)
```

```r
model_ = list(fit,fit_2, fit_3)

rmse_tests = sapply(model_, get_rmse, data = test_boston, response = "medv")
```

Model	Test RMSE	# Obs removed
fit	3.2202003	0
fit_2	3.0088003	19
fit_3	3.0560078	5

Model that removes observations from the training data with absolute standardized residuals greater than 2 performs the best, it removes largest amount of observations. So modifying training data is justified.

```r
newdata=tibble(crim=0.02763, zn=75.0, indus=3.95, chas=0, nox=0.4280,   rm=6.595,   age=22.8,   dis=5.4
predict(fit_2, newdata, interval = "prediction", level = 0.99)
```

```
##        fit      lwr      upr
## 1 27.52639 21.03786 34.01491
```