36-402 Homework 2

Jacky Liu

2/19/21

Problem 1

Problem 1 (a)

```r
housetrain = read.csv("housetrain.csv", sep=",")
housetest = read.csv("housetest.csv", sep=",")
res = cor(housetrain)
round(res, 4)
```

```
##                          Population Latitude Longitude Median_house_value
## Population                   1.0000  -0.1615   -0.1614             0.0537
## Latitude                    -0.1615   1.0000    0.7282            -0.4658
## Longitude                   -0.1614   0.7282    1.0000            -0.5391
## Median_house_value           0.0537  -0.4658   -0.5391             1.0000
## Median_household_income      0.1218  -0.1354   -0.1835             0.6465
## Mean_household_income        0.0861  -0.1249   -0.1715             0.6941
##                          Median_household_income Mean_household_income
## Population                                0.1218                0.0861
## Latitude                                 -0.1354               -0.1249
## Longitude                                -0.1835               -0.1715
## Median_house_value                        0.6465                0.6941
## Median_household_income                   1.0000                0.9493
## Mean_household_income                     0.9493                1.0000
```

Problem 1 (b)

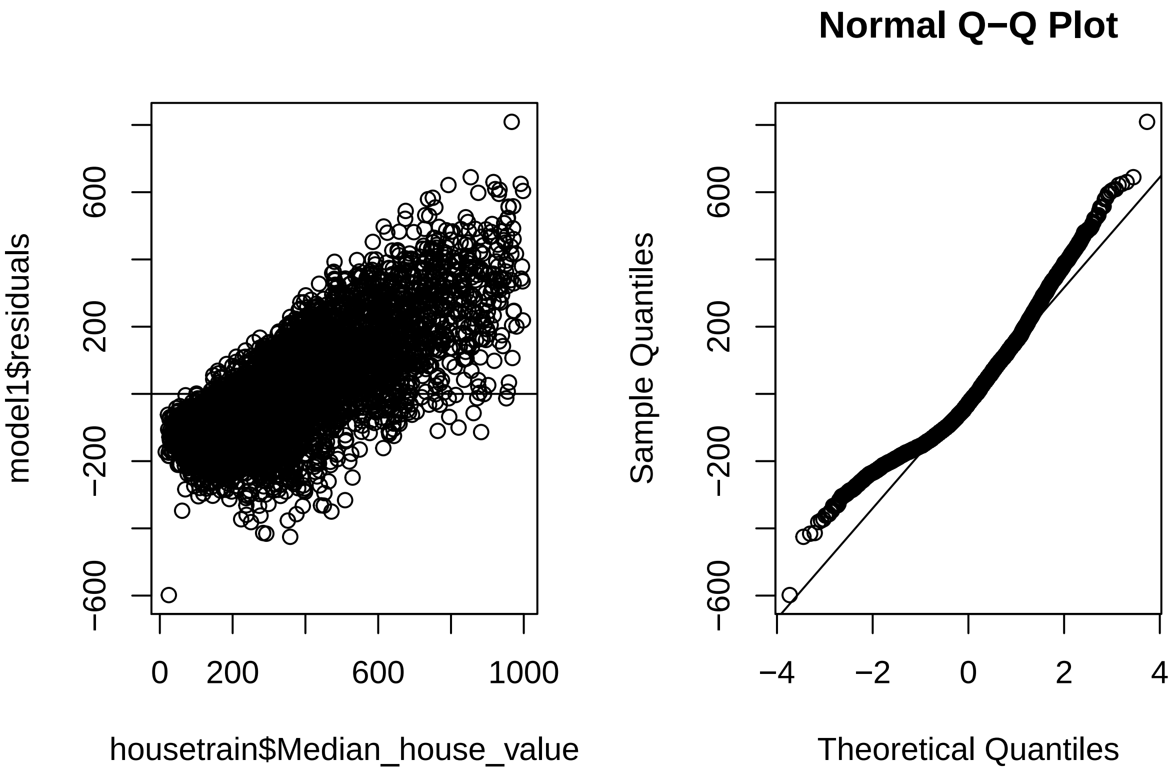
```r
model0 = lm(Median_house_value ~ 1, data=housetrain)
model0
```

```
##
## Call:
## lm(formula = Median_house_value ~ 1, data = housetrain)
##
## Coefficients:
## (Intercept)
##       344.5
```

```r
model1 = lm(Median_house_value ~ Median_household_income, data=housetrain)
model1
```

```
##
## Call:
## lm(formula = Median_house_value ~ Median_household_income, data = housetrain)
##
## Coefficients:
##             (Intercept)  Median_household_income
##                30.28233                  0.00517
```

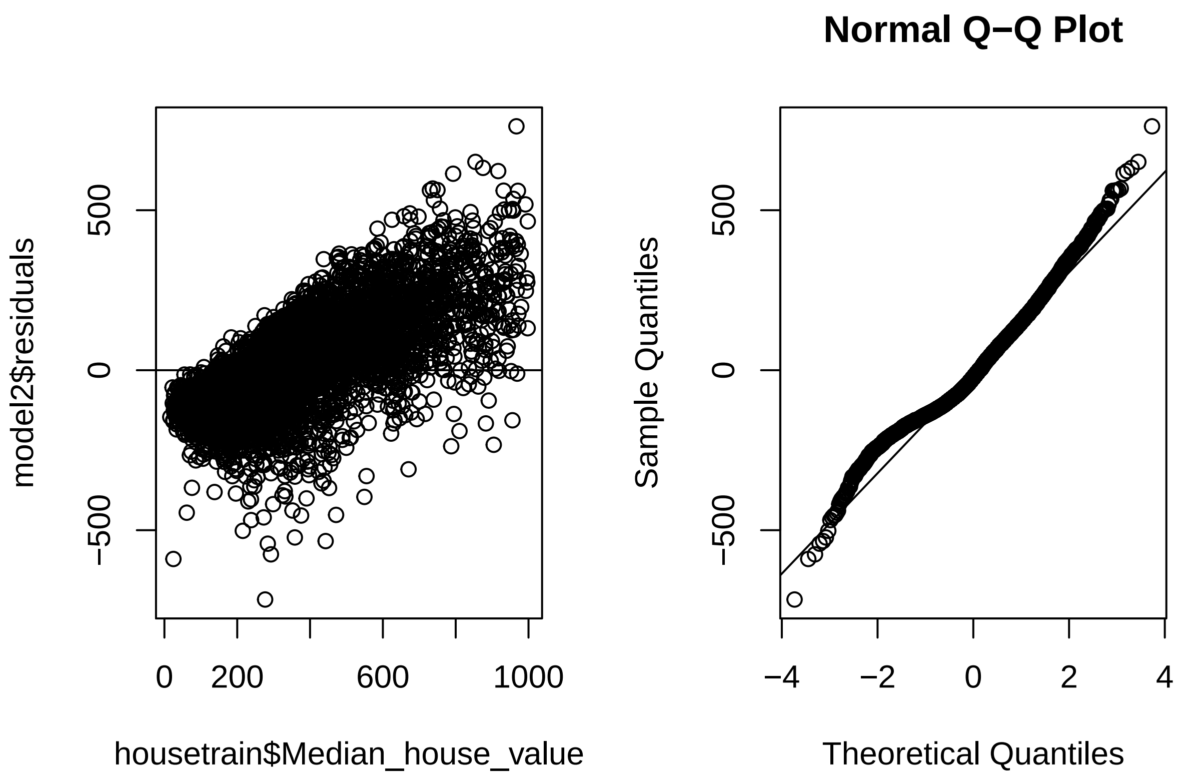
```r
par(mfrow=c(1,2))
plot(housetrain$Median_house_value, model1$residuals); abline(h=0)
qqnorm(model1$residuals); qqline(model1$residuals)
```



```r
model2 = lm(Median_house_value ~ Mean_household_income, data=housetrain)
model2
```

```
##
## Call:
## lm(formula = Median_house_value ~ Mean_household_income, data = housetrain)
##
## Coefficients:
##           (Intercept)  Mean_household_income
##             -4.523268               0.004658
```

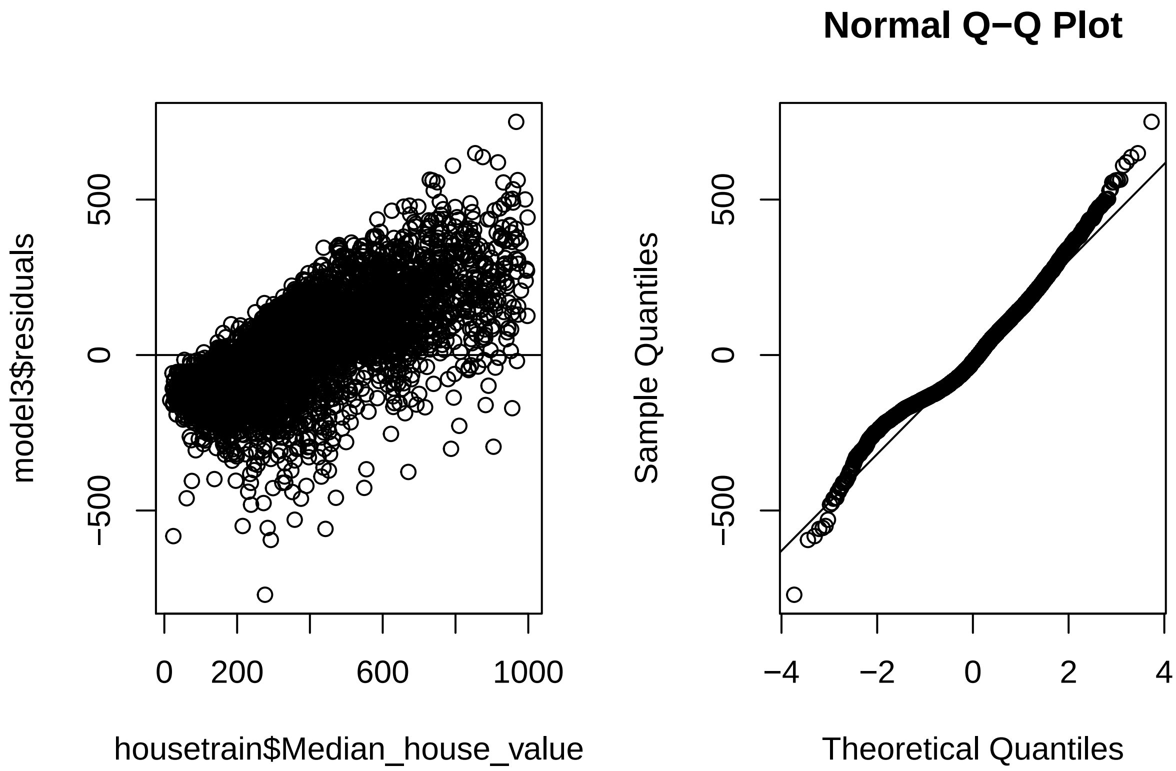
```
par(mfrow=c(1,2))
plot(housetrain$Median_house_value, model2$residuals); abline(h=0)
qqnorm(model2$residuals); qqline(model2$residuals)
```



```
model3 = lm(Median_house_value ~ Median_household_income + Mean_household_income, data=housetrain)
model3
```

```
##
## Call:
## lm(formula = Median_house_value ~ Median_household_income + Mean_household_income,
##     data = housetrain)
##
## Coefficients:
##             (Intercept)  Median_household_income    Mean_household_income
##               -3.432593                -0.001003                 0.005458
```

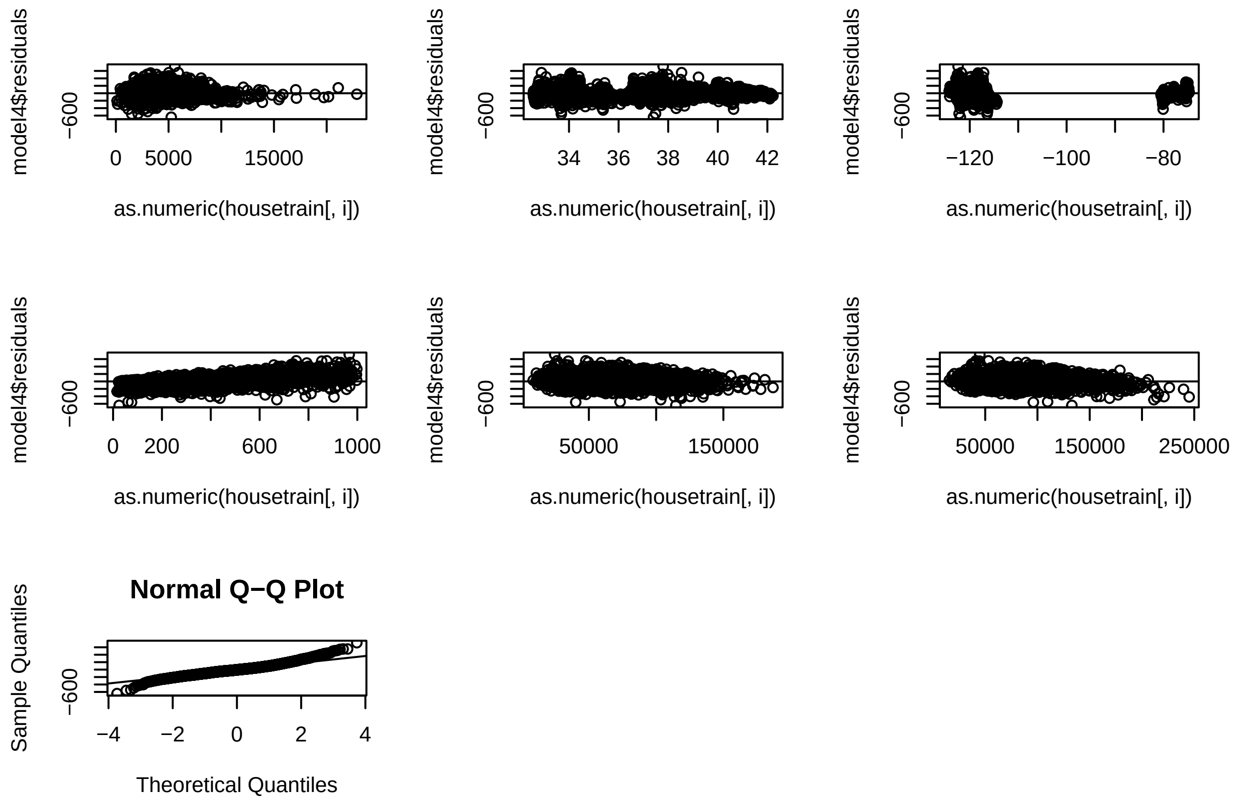
```
par(mfrow=c(1,2))
plot(housetrain$Median_house_value, model3$residuals); abline(h=0)
qqnorm(model3$residuals); qqline(model3$residuals)
```



```
model4 = lm(Median_house_value ~ Median_household_income + Mean_household_income +
              Population*Median_household_income + Latitude*Median_household_income +
              Population*Latitude + Mean_household_income*Longitude + Latitude*Longitude, data=housetrai
summary(model4)
```

```
##
## Call:
## lm(formula = Median_house_value ~ Median_household_income + Mean_household_income +
##     Population * Median_household_income + Latitude * Median_household_income +
##     Population * Latitude + Mean_household_income * Longitude +
##     Latitude * Longitude, data = housetrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -642.67  -65.73   -6.41   58.58  721.84
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -2.075e+03  6.513e+02  -3.186  0.00145 **
## Median_household_income               -9.616e-03  1.140e-03  -8.435  < 2e-16 ***
## Mean_household_income                 -4.304e-04  4.061e-04  -1.060  0.28916
## Population                            -7.020e-02  1.238e-02  -5.670 1.50e-08 ***
## Latitude                               5.282e+01  1.607e+01   3.287  0.00102 **
## Longitude                             -2.903e+01  5.594e+00  -5.190 2.18e-07 ***
## Median_household_income:Population     4.573e-08  2.975e-08   1.537  0.12437
## Median_household_income:Latitude    2.026e-04  2.980e-05   6.797 1.18e-11 ***
## Population:Latitude                 1.645e-03  3.203e-04   5.137 2.90e-07 ***
## Mean_household_income:Longitude    -5.702e-05  3.623e-06 -15.736  < 2e-16 ***
## Latitude:Longitude                  7.429e-01  1.389e-01   5.349 9.20e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114.7 on 5292 degrees of freedom
## Multiple R-squared:  0.7039, Adjusted R-squared:  0.7033
## F-statistic:  1258 on 10 and 5292 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(3,3))
for(i in c(1:6)){
  plot(as.numeric(housetrain[,i]), model4$residuals)
  abline(h=0)
}
qqnorm(model4$residuals); qqline(model4$residuals)
```



For model 4, we simulated these extra covariates by computing a regression model for all covariates and looking at significance.

The assumptions of the linear model (iid, linearity, constant variance, Gaussian noise) seem relatively but not perfectly plausible. Although the residuals appear to mostly have constant variance, there may be some heteroskedasticity, especially in the first three plots. The residuals appear roughly Gaussian but some have heavier tails than we'd expect under normality. Overall the data show some modest signs of betraying the linear model assumptions, but not too extreme.

Problem 1 (c)

The reason the coefficients of Median_household_income and Mean_household_income in Model 3 are both different from the coefficients of the same predictors in Models 1 and 2 is because doing multiple linear regression is not the same thing as doing two simple linear regression on both variables. Having separate univariate models causes correlations to be ignored.

Problem 1 (d)

```r
train_error_3 = mean(model3$residuals^2)
train_error_4 = mean(model4$residuals^2)
train_error_3
```

```
## [1] 22891.66
```

```r
train_error_4
```

```
## [1] 13122.21
```

The training error on model 4 is lower than the training error on model 3. I would expect the training error to decrease more if we added more covariates as it overfits.

Problem 1 (e)

```r
test_error_0 = mean((housetest$Median_house_value - predict.lm(model0, housetest)) ^ 2)
test_error_1 = mean((housetest$Median_house_value - predict.lm(model1, housetest)) ^ 2)
test_error_2 = mean((housetest$Median_house_value - predict.lm(model2, housetest)) ^ 2)
test_error_3 = mean((housetest$Median_house_value - predict.lm(model3, housetest)) ^ 2)
test_error_4 = mean((housetest$Median_house_value - predict.lm(model4, housetest)) ^ 2)

test_error_0
```

```
## [1] 44062.46
```

```r
test_error_1
```

```
## [1] 26359.29
```

```r
test_error_2
```

```
## [1] 22985.77
```

```r
test_error_3
```

```
## [1] 22813.26
```

```
test_error_4
```

```
## [1] 13387.69
```

The highest test error was model 0 followed by model 1. Models 2 and 3 had similar errors. The lowest test error was model 4 which makes it the best model here.