Data Analysis/Regression

Anthony Hu

2024-01-21

```
library(stats)
library(glmnet)
```

```
## Loading required package: Matrix
```

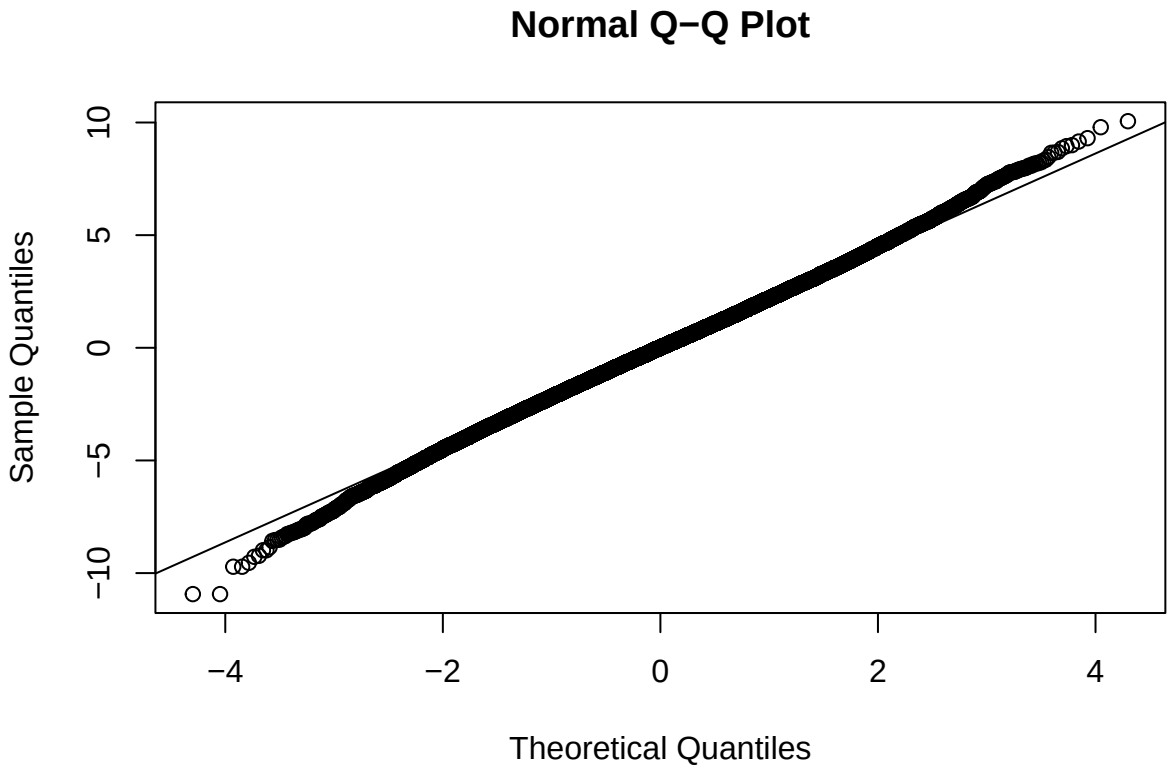
```
## Loaded glmnet 4.1-8
```

```
data <- read.csv("datathon_2024_dataset_model.csv")[,-c(2, 3, 5, 6)]
# Linear Regression
mod_linear <- lm(run_diff~as.factor(is_day_game_x)+single_diff+double_diff+triple_diff+hr_diff+pa_diff+
                   free_base_diff+home_k+away_k+field_diff+days_since_last_home_game+days_since_last_aw
                   home_distance+away_distance, data = data)
summary(mod_linear)
```

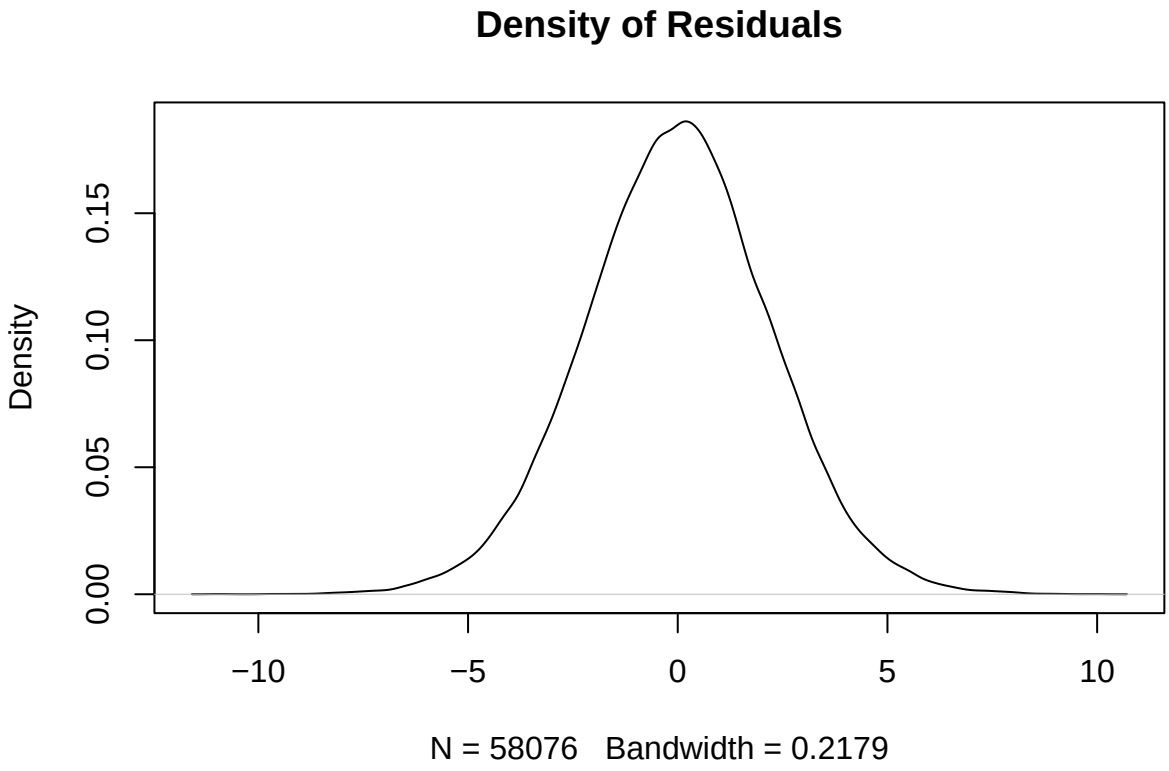
```
##
## Call:
## lm(formula = run_diff ~ as.factor(is_day_game_x) + single_diff +
##     double_diff + triple_diff + hr_diff + pa_diff + free_base_diff +
##     home_k + away_k + field_diff + days_since_last_home_game +
##     days_since_last_away_game + home_distance + away_distance,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9324  -1.4612   0.0132   1.4495  10.0571
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.420e-01  3.787e-02   3.749 0.000177 ***
## as.factor(is_day_game_x)1   6.015e-04  2.026e-02   0.030 0.976317
## single_diff                 5.991e-01  1.491e-02  40.187  < 2e-16 ***
## double_diff                 8.799e-01  1.552e-02  56.680  < 2e-16 ***
## triple_diff                 1.205e+00  2.121e-02  56.805  < 2e-16 ***
## hr_diff                     1.521e+00  1.587e-02  95.875  < 2e-16 ***
## pa_diff                    -8.565e-02  1.132e-02  -7.569 3.81e-14 ***
## free_base_diff             -4.742e-01  1.496e-02 -31.704  < 2e-16 ***
```

```
## home_k                    -1.808e+00  5.189e-01  -3.485 0.000492 ***
## away_k                     1.564e+00  5.581e-01   2.802 0.005086 **
## field_diff                 3.032e+00  5.268e-01   5.756 8.63e-09 ***
## days_since_last_home_game -7.510e-03  3.470e-03  -2.164 0.030437 *
## days_since_last_away_game  5.331e-03  3.457e-03   1.542 0.123100
## home_distance             -1.097e-05  1.186e-05  -0.925 0.354983
## away_distance             -4.189e-06  1.186e-05  -0.353 0.723964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.227 on 58061 degrees of freedom
## Multiple R-squared:  0.7439, Adjusted R-squared:  0.7438
## F-statistic: 1.205e+04 on 14 and 58061 DF,  p-value: < 2.2e-16
```

```r
# Testing for normality of residuals
qqnorm(mod_linear$residuals)
qqline(mod_linear$residuals)
```




```r
plot(density(mod_linear$residuals), main = "Density of Residuals")
```

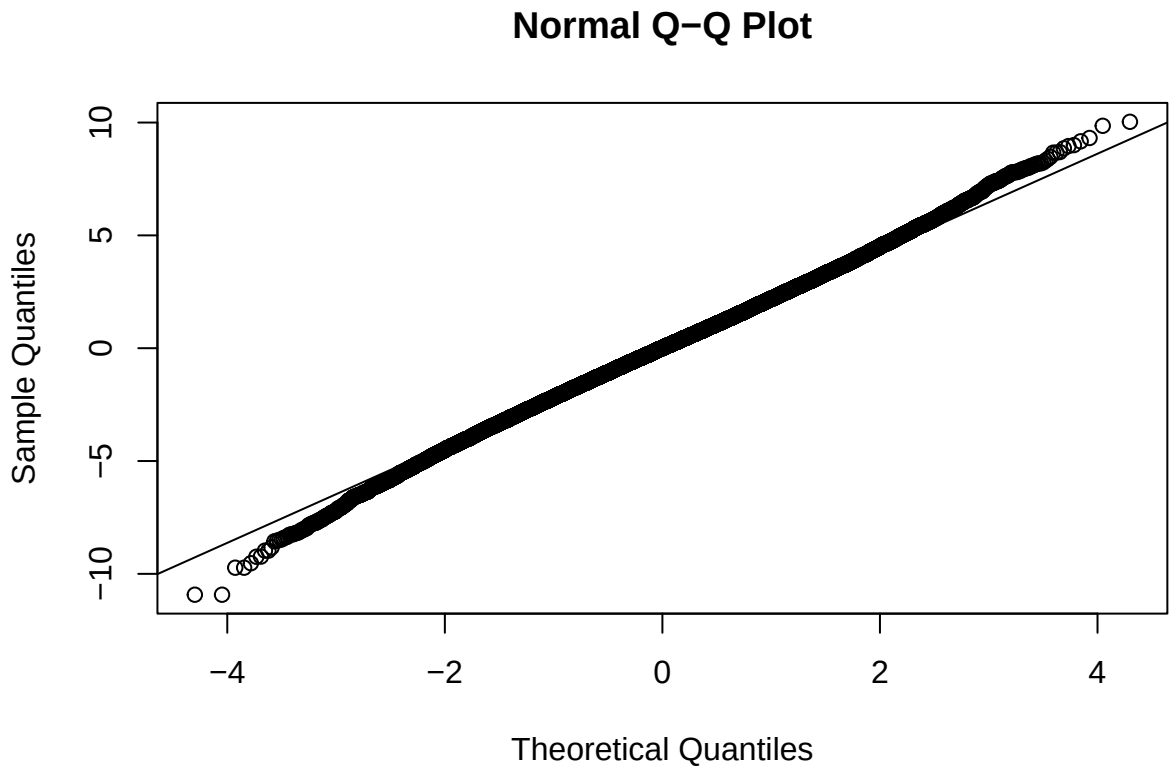



```
# Linear Regression, but drop all variables with insignificant p-values
mod_linear_2 <- lm(run_diff~single_diff+double_diff+triple_diff+hr_diff+pa_diff+free_base_diff + home_k
                   field_diff+days_since_last_home_game, data = data)
summary(mod_linear_2)
```

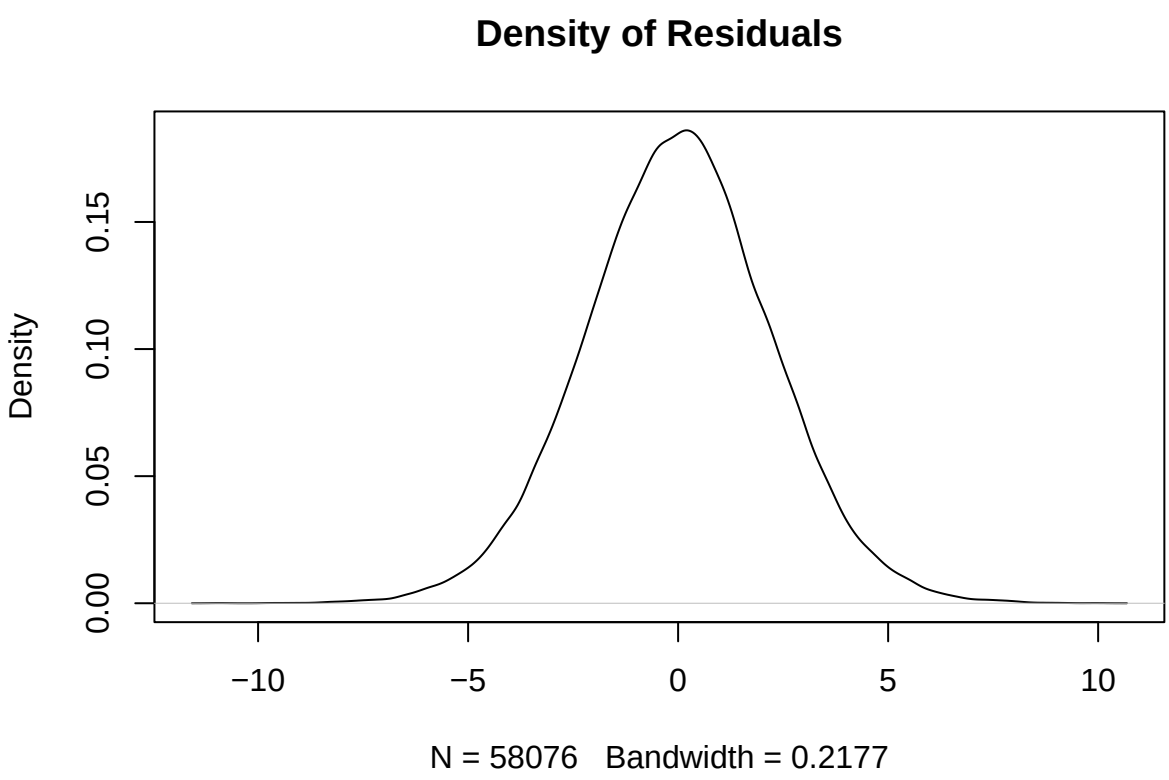
```
##
## Call:
## lm(formula = run_diff ~ single_diff + double_diff + triple_diff +
##     hr_diff + pa_diff + free_base_diff + home_k + away_k + field_diff +
##     days_since_last_home_game, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9240  -1.4581   0.0124   1.4497  10.0324
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.124147   0.033801   3.673 0.000240 ***
## single_diff               0.599032   0.014907  40.183  < 2e-16 ***
## double_diff               0.879780   0.015524  56.673  < 2e-16 ***
## triple_diff               1.204326   0.021206  56.793  < 2e-16 ***
## hr_diff                   1.521084   0.015866  95.869  < 2e-16 ***
## pa_diff                  -0.085590   0.011316  -7.564 3.97e-14 ***
## free_base_diff           -0.474149   0.014956 -31.703  < 2e-16 ***
## home_k                   -1.806685   0.518836  -3.482 0.000498 ***
## away_k                    1.565188   0.558077   2.805 0.005039 **
```

```
## field_diff                 3.033269   0.526758   5.758 8.54e-09 ***
## days_since_last_home_game -0.005155   0.003066  -1.681 0.092693 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.227 on 58065 degrees of freedom
## Multiple R-squared:  0.7439, Adjusted R-squared:  0.7438
## F-statistic: 1.686e+04 on 10 and 58065 DF,  p-value: < 2.2e-16
```

```r
# Testing for normality of residuals
qqnorm(mod_linear_2$residuals)
qqline(mod_linear_2$residuals)
```




```r
plot(density(mod_linear_2$residuals), main = "Density of Residuals")
```



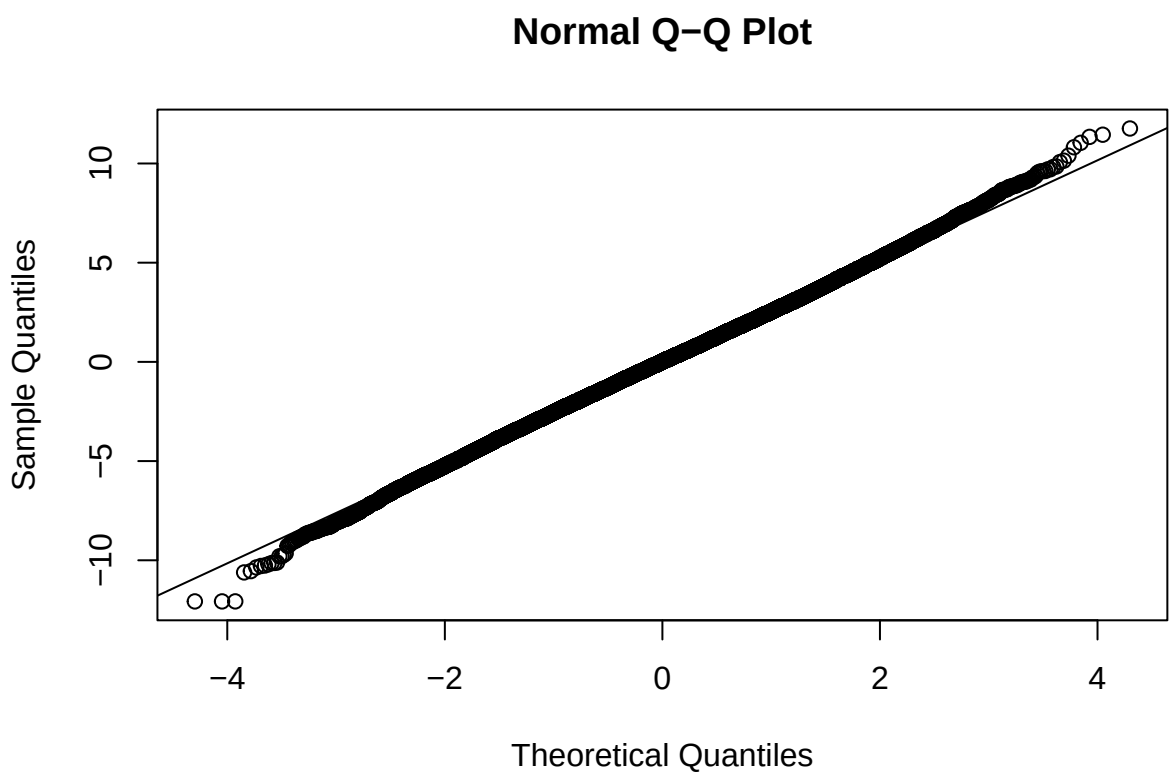

```
mod_linear_3 <- lm(run_diff~ I(single_diff+double_diff+triple_diff+hr_diff)+ free_base_diff+pa_diff+I(ho
                     field_diff+days_since_last_home_game +
                     I((single_diff + double_diff + triple_diff + hr_diff)* (home_distance + away_distan
                     I((home_k - away_k) * (home_distance + away_distance)) +
                     I(field_diff + (home_distance + away_distance)) + home_distance + away_distance
                   ,data = data)

summary(mod_linear_3)
```

```
##
## Call:
## lm(formula = run_diff ~ I(single_diff + double_diff + triple_diff +
##     hr_diff) + free_base_diff + pa_diff + I(home_k - away_k) +
##     field_diff + days_since_last_home_game + I((single_diff +
##     double_diff + triple_diff + hr_diff) * (home_distance + away_distance)) +
##     I((home_k - away_k) * (home_distance + away_distance)) +
##     I(field_diff + (home_distance + away_distance)) + home_distance +
##     away_distance, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0727  -1.7103   0.0101   1.7151  11.7624
##
## Coefficients: (1 not defined because of singularities)
##                                                                                 Estimate
```

```
## (Intercept)                                                                    3.774e-02
## I(single_diff + double_diff + triple_diff + hr_diff)                           8.025e-01
## free_base_diff                                                                -5.271e-01
## pa_diff                                                                       -1.319e-01
## I(home_k - away_k)                                                            -8.994e-01
## field_diff                                                                     1.802e+00
## days_since_last_home_game                                                     -7.078e-03
## I((single_diff + double_diff + triple_diff + hr_diff) * (home_distance + away_distance)) -1.292e-06
## I((home_k - away_k) * (home_distance + away_distance))                         1.123e-04
## I(field_diff + (home_distance + away_distance))                               -7.210e-06
## home_distance                                                                 -2.899e-06
## away_distance                                                                         NA
##                                                                              Std. Error
## (Intercept)                                                                    2.299e-02
## I(single_diff + double_diff + triple_diff + hr_diff)                           1.710e-02
## free_base_diff                                                                 1.694e-02
## pa_diff                                                                        1.283e-02
## I(home_k - away_k)                                                             6.198e-01
## field_diff                                                                     5.950e-01
## days_since_last_home_game                                                      3.566e-03
## I((single_diff + double_diff + triple_diff + hr_diff) * (home_distance + away_distance))  1.107e-06
## I((home_k - away_k) * (home_distance + away_distance))                         4.909e-05
## I(field_diff + (home_distance + away_distance))                                1.384e-05
## home_distance                                                                  2.562e-05
## away_distance                                                                         NA
##                                                                                t value
## (Intercept)                                                                      1.642
## I(single_diff + double_diff + triple_diff + hr_diff)                            46.938
## free_base_diff                                                                -31.113
## pa_diff                                                                       -10.285
## I(home_k - away_k)                                                             -1.451
## field_diff                                                                      3.029
## days_since_last_home_game                                                      -1.985
## I((single_diff + double_diff + triple_diff + hr_diff) * (home_distance + away_distance))  -1.166
## I((home_k - away_k) * (home_distance + away_distance))                          2.287
## I(field_diff + (home_distance + away_distance))                                -0.521
## home_distance                                                                  -0.113
## away_distance                                                                      NA
##                                                                              Pr(>|t|)
## (Intercept)                                                                   0.10068
## I(single_diff + double_diff + triple_diff + hr_diff)                          < 2e-16
## free_base_diff                                                                < 2e-16
## pa_diff                                                                       < 2e-16
## I(home_k - away_k)                                                            0.14675
## field_diff                                                                    0.00246
## days_since_last_home_game                                                     0.04716
## I((single_diff + double_diff + triple_diff + hr_diff) * (home_distance + away_distance))  0.24342
## I((home_k - away_k) * (home_distance + away_distance))                        0.02218
## I(field_diff + (home_distance + away_distance))                               0.60237
## home_distance                                                                 0.90992
## away_distance                                                                      NA
##
## (Intercept)
## I(single_diff + double_diff + triple_diff + hr_diff)                          ***
```

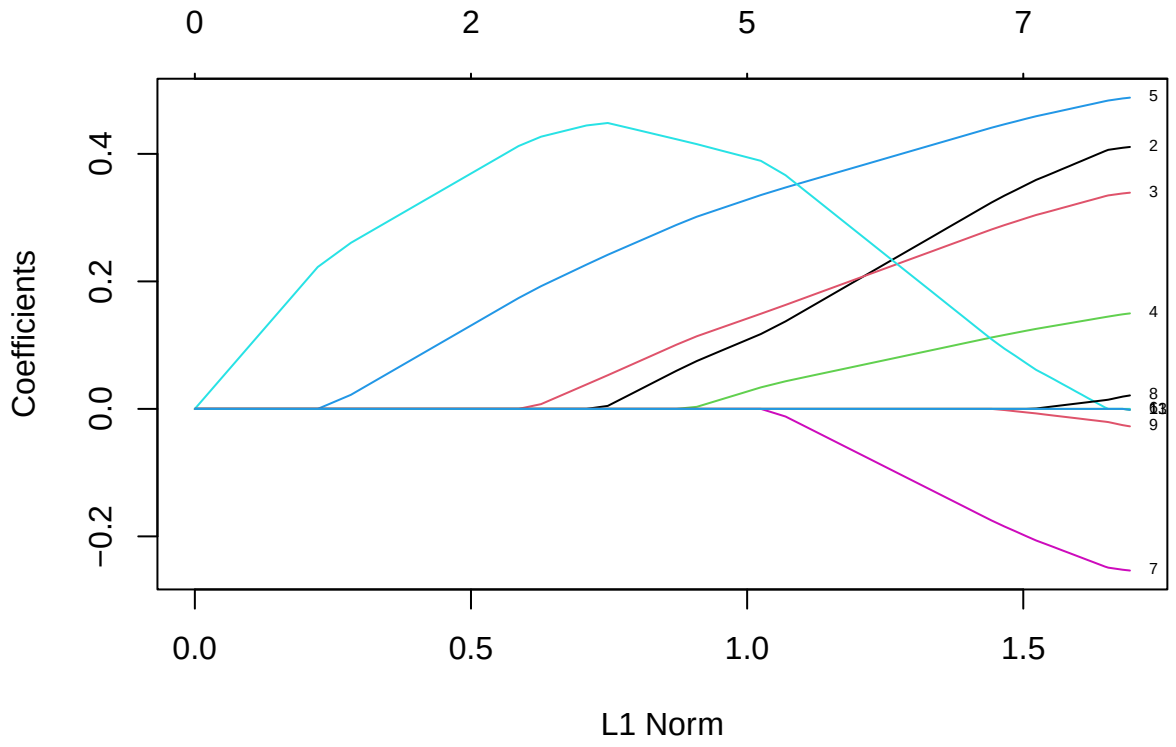
```
## free_base_diff                                                                  ***
## pa_diff                                                                         ***
## I(home_k - away_k)
## field_diff                                                                      **
## days_since_last_home_game                                                       *
## I((single_diff + double_diff + triple_diff + hr_diff) * (home_distance + away_distance))
## I((home_k - away_k) * (home_distance + away_distance))                          *
## I(field_diff + (home_distance + away_distance))
## home_distance
## away_distance
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.59 on 58065 degrees of freedom
## Multiple R-squared:  0.6535, Adjusted R-squared:  0.6535
## F-statistic: 1.095e+04 on 10 and 58065 DF,  p-value: < 2.2e-16
```

```r
qqnorm(mod_linear_3$residuals)
qqline(mod_linear_3$residuals)
```

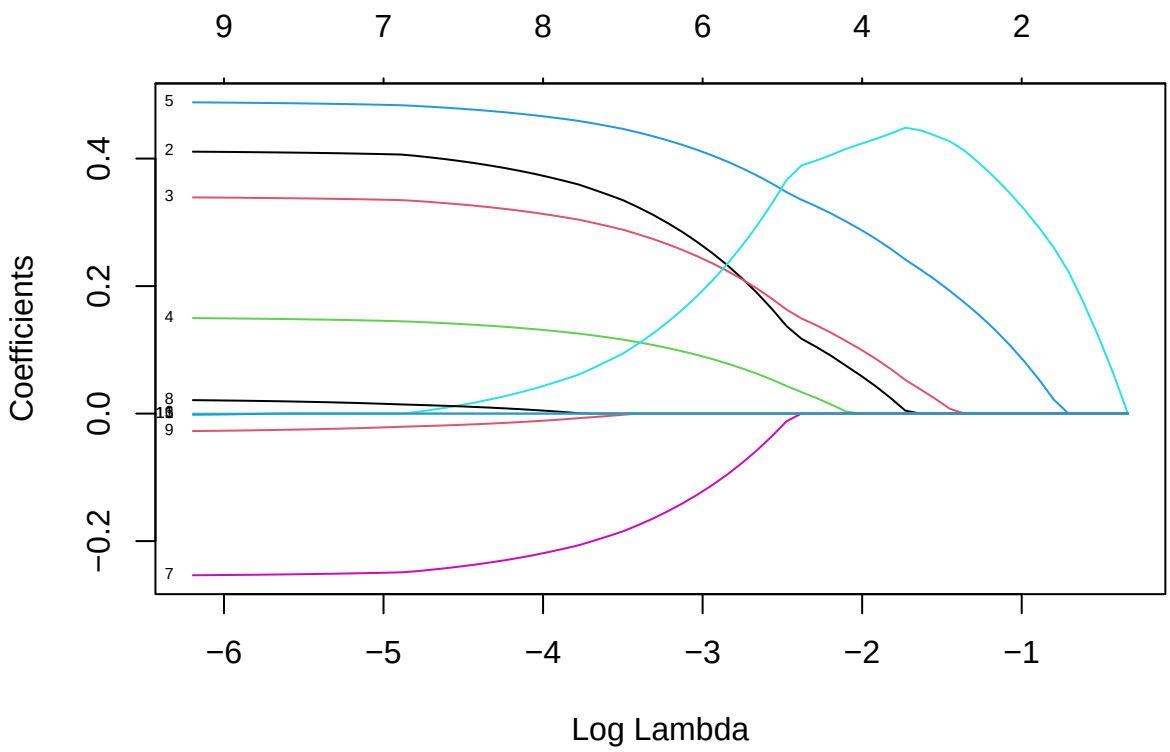



```r
# LASSO Regression
data_reg <- data[,-1]
X <- scale(as.matrix(data_reg[,-2]), center = TRUE, scale = TRUE)
Y <- scale(as.matrix(data_reg[,2]), center = TRUE, scale = TRUE)
```

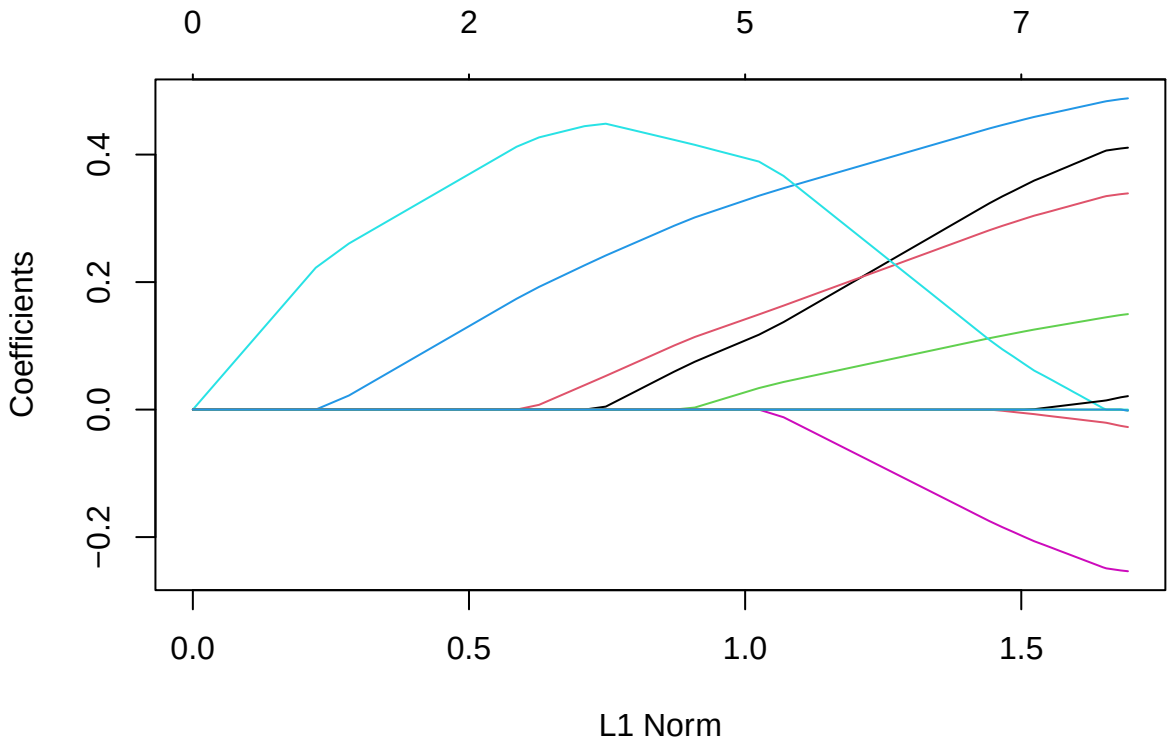
```
LASSO_cv <- cv.glmnet(X, Y, family = "gaussian",
                      type.measure = "mse")

opt_Lambda <- LASSO_cv$lambda.min
plot(LASSO_cv$glmnet.fit, "norm", label = TRUE)
```




```
plot(LASSO_cv$glmnet.fit, "lambda", label = TRUE)
```



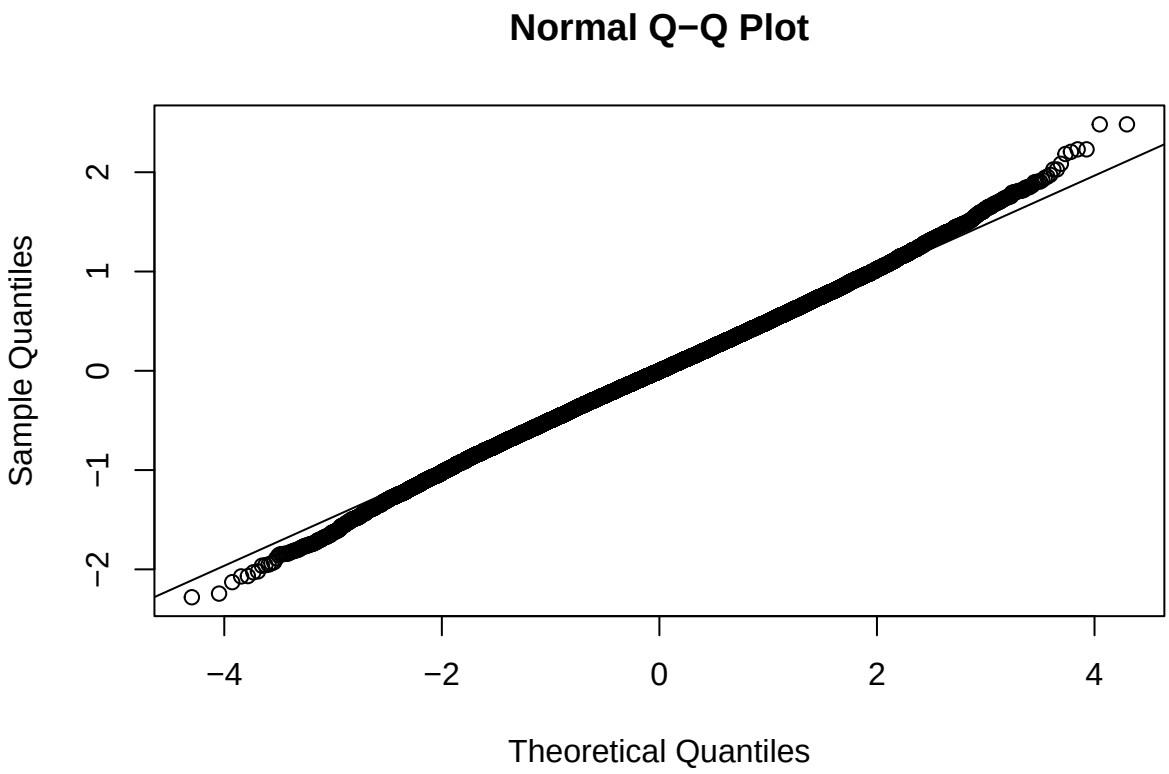

```
LASSO_mod <- glmnet(X, Y, family = "gaussian",
                    type.measure = "mse")
plot(LASSO_mod)
```




```
LASSO_coef <- predict(LASSO_mod, type = "coefficients", s = opt_Lambda)
LASSO_coef
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                      s1
## (Intercept)                -2.192194e-18
## is_day_game_x               .
## single_diff                 4.108599e-01
## double_diff                 3.391468e-01
## triple_diff                 1.497662e-01
## hr_diff                     4.882256e-01
## pa_diff                     .
## free_base_diff             -2.535833e-01
## home_k                      2.109266e-02
## away_k                     -2.741313e-02
## field_diff                  .
## days_since_last_home_game -1.473889e-03
## days_since_last_away_game   .
## home_distance              -1.577190e-03
## away_distance               .
```

```
LASSO_pred <- predict(LASSO_mod,X, type = "response", s = opt_Lambda)
LASSO_MSE <- mean((LASSO_pred - Y) ^ 2)
qqnorm(LASSO_pred - Y)
qqline(LASSO_pred - Y)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−4
−2
0
2
4