MATH3424 HW2

Chow, Hau Cheung Jasper (hcjchow / 20589533)

September 27, 2021

```
# Q1 TEST
setwd("/Users/jchow/Downloads/MATH3424 R") # need to set the starting directory as this
data <- read.table(file="Supervisor.txt",header=TRUE)
head(data)
```

```
##    Y X1 X2 X3 X4 X5 X6
## 1 43 51 30 39 61 92 45
## 2 63 64 51 54 63 73 47
## 3 71 70 68 69 76 86 48
## 4 61 63 45 47 54 84 35
## 5 81 78 56 66 71 83 47
## 6 43 55 49 44 54 49 34
```

```
Sx2_y <- sum( (data$X2-mean(data$X2)) * (data$Y-mean(data$Y)) )
Sx2_x2 <- sum((data$X2-mean(data$X2))^2)
beta_1_hat <- Sx2_y/Sx2_x2
beta_0_hat <- mean(data$Y) - mean(data$X2)*beta_1_hat

mean(data$Y)
```

```
## [1] 64.63333
```

```
mean(data$X2)
```

```
## [1] 53.13333
```

```
Sx2_y
```

```
## [1] 1840.467
```

```
Sx2_x2
```

```
## [1] 4341.467
```

```
beta_1_hat
```

```
## [1] 0.4239274
```

```
beta_0_hat
```

```
## [1] 42.10866
```

```
sup_model_1 = lm(data$Y ~ data$X2, data = data)
summary(sup_model_1)
```

```
##
## Call:
## lm(formula = data$Y ~ data$X2, data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.9357  -5.7397  -0.1691   5.6026  23.3582
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.1087     9.2661   4.544 9.63e-05 ***
## data$X2       0.4239     0.1701   2.492   0.0189 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.21 on 28 degrees of freedom
## Multiple R-squared:  0.1816, Adjusted R-squared:  0.1523
## F-statistic: 6.212 on 1 and 28 DF,  p-value: 0.01888
```

```
e_y_x2 <- resid(sup_model_1)
```

```
# Q1-2 TEST
Sx2_x1 <- sum( (data$X2-mean(data$X2)) * (data$X1-mean(data$X1)) )
Sx2_x2 <- sum((data$X2-mean(data$X2))^2)
c_1_hat <- Sx2_x1/Sx2_x2
c_0_hat <- mean(data$X1) - mean(data$X2)*c_1_hat

mean(data$X1)
```

```
## [1] 66.6
```

```
mean(data$X2)
```

```
## [1] 53.13333
```

```
Sx2_x1
```

```
## [1] 2637.6
```

```
Sx2_x2
```

```
## [1] 4341.467
```

```
c_1_hat
```

```
## [1] 0.6075366
```

```
c_0_hat
```

```
## [1] 34.31955
```

```
sup_model_2 = lm(data$X1 ~ data$X2, data = data)
summary(sup_model_2)
```

```
##
## Call:
## lm(formula = data$X1 ~ data$X2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.8361  -5.7851  -0.9813   7.4500  25.3036
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  34.3196     9.2950   3.692 0.000953 ***
## data$X2       0.6075     0.1706   3.561 0.001346 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.24 on 28 degrees of freedom
## Multiple R-squared:  0.3117, Adjusted R-squared:  0.2871
## F-statistic: 12.68 on 1 and 28 DF,  p-value: 0.001346
```

```r
e_x1_x2 <- resid(sup_model_2)
```

```r
# Q1-3 TEST
S_e_y_x2 <- sum( (e_y_x2-mean(e_y_x2)) * (e_x1_x2-mean(e_x1_x2)) )
S_e_x1_x2 <- sum((e_x1_x2-mean(e_x1_x2))^2)
d_1_hat <- S_e_y_x2/S_e_x1_x2
d_0_hat <- mean(e_y_x2) - mean(e_x1_x2)*d_1_hat

mean(e_y_x2)
```

```
## [1] -5.932754e-17
```

```r
mean(e_x1_x2)
```

```
## [1] -1.444446e-16
```

```r
S_e_y_x2
```

```
## [1] 2761.449
```

```r
S_e_x1_x2
```

```
## [1] 3538.761
```

```r
d_1_hat
```

```
## [1] 0.7803434
```

```r
d_0_hat
```

```
## [1] 5.338888e-17
```

```r
sup_model_3 = lm(e_y_x2 ~ e_x1_x2)
summary(sup_model_3)
```

```
##
## Call:
## lm(formula = e_y_x2 ~ e_x1_x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7887  -5.6893  -0.0284   6.2745   9.9726
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.927e-16  1.273e+00   0.000        1
## e_x1_x2      7.803e-01  1.172e-01   6.656 3.19e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.974 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.6127, Adjusted R-squared:  0.5989
## F-statistic:  44.3 on 1 and 28 DF,  p-value: 3.192e-07
```

```r
# Q2a
exm_data <- read.table(file="Examination_Data.txt",header=TRUE)
head(exm_data)
```

```
##    F P1 P2
## 1 68 78 73
## 2 75 74 76
## 3 85 82 79
## 4 94 90 96
## 5 86 87 90
## 6 90 90 92
```

```r
q2_model1 <- lm(F ~ P1, data=exm_data)
q2_model2 <- lm(F ~ P2, data=exm_data)
q2_model3 <- lm(F ~ P1+P2, data=exm_data)
```

```r
# Q2a cont.
summary(q2_model1)
```

```
##
## Call:
## lm(formula = F ~ P1, data = exm_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.844 -2.020 -0.587  4.043  7.938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22.3424    11.5640  -1.932   0.0676 .
## P1            1.2605     0.1399   9.008 1.78e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.081 on 20 degrees of freedom
## Multiple R-squared:  0.8023, Adjusted R-squared:  0.7924
## F-statistic: 81.14 on 1 and 20 DF,  p-value: 1.779e-08
```

```r
summary(q2_model2)
```

```
##
## Call:
## lm(formula = F ~ P2, data = exm_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4323  -1.5027   0.5421   2.2580   7.5165
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.85355    7.56181  -0.245    0.809
## P2           1.00427    0.09059  11.086 5.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.275 on 20 degrees of freedom
## Multiple R-squared:   0.86,  Adjusted R-squared:  0.853
## F-statistic: 122.9 on 1 and 20 DF,  p-value: 5.442e-10
```

```r
summary(q2_model3)
```

```
##
## Call:
## lm(formula = F ~ P1 + P2, data = exm_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7328 -2.1703  0.3938  2.6443  6.3660
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.5005     9.2356  -1.570  0.13290
## P1            0.4883     0.2330   2.096  0.04971 *
## P2            0.6720     0.1793   3.748  0.00136 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.953 on 19 degrees of freedom
## Multiple R-squared:  0.8863, Adjusted R-squared:  0.8744
## F-statistic: 74.07 on 2 and 19 DF,  p-value: 1.069e-09
```

2a) The fitted models are:

Model 1: $\hat{F} = -22.3424 + 1.2605 P_1$

Model 2: $\hat{F} = -1.85355 + 1.00427 P_2$

Model 3: $\hat{F} = -14.5005 + 0.4883 P_1 + 0.672 P_2$

2b) From the above summaries, the t-statistic for $H_0 : \beta_0 = 0, H_1 : \beta_0 \neq 0$ has the values -1.932 for model 1, -0.245 for model 2, -1.570 for model 3.

```r
q2_alpha = 0.05
qt(1-q2_alpha/2, df=22-2) # for models 1 and 2
```

```
## [1] 2.085963
```

```r
qt(1-q2_alpha/2, df=22-3) # for model 3
```

```
## [1] 2.093024
```

For model 1, since $|-1.932| \leq t_{1-\alpha/2,20}$, we fail to reject the null hypothesis that $H_0 : \beta_0 = 0$.

For model 2, since $|-0.245| \leq t_{1-\alpha/2,20}$, we fail to reject the null hypothesis that $H_0 : \beta_0 = 0$.

For model 3, since $|-1.57| \leq t_{1-\alpha/2,19}$, we fail to reject the null hypothesis that $H_0 : \beta_0 = 0$

2c) Comparing the output of the first two models, we see that the multiple R-squared value (aka correlation to F) for the 2nd model with predictor P_2 is higher than the multiple R-squared value for the 1st model with predictor P_1. As such, P_2 is a better predictor of F than P_1.

2d) Let us consider the hypothesis test where H_0 : reduced model (q2_model2) is adequate, against H_1 : full model (q2_model3) is adequate. Construct the F statistic as follows:

```r
# Q2d
sse_rm <- sum((exm_data$F - q2_model2$fitted.values)^2)
sse_fm <- sum((exm_data$F - q2_model3$fitted.values)^2)
p = 2 # since 2 predictors in full model
k = 2 # = no. of parameters in reduced model = beta_0, beta_1
n = length(exm_data$F) # number of datapoints
q2_F = ((sse_rm-sse_fm)/(p+1-k)) / (sse_fm/(n-p-1))
q2_F
```

```
## [1] 4.392948
```

```r
alpha <- 0.05 # choose a confidence level
F_alpha <- qf(1-alpha, df1=p+1-k, df2=n-p-1, lower.tail=TRUE)
F_alpha
```

```
## [1] 4.38075
```

Hence at confidence level $\alpha = 0.05$, since the F-statistic is greater than the critical value, we reject the null hypothesis that the reduced model is adequate, as such we will use the full model with both predictors.

```r
# Q2d continued
q2_pred_int_conf = 0.05 # chosen value of alpha

x_0_df = data.frame(P1=78, P2=85) # add new datapoint
y_0_hat <- predict(q2_model3, newdata=x_0_df)
y_0_hat # our prediction
```

```
##        1
## 80.71282
```

```r
X <- data.matrix(exm_data[,c('P1','P2')]) # convert predictor columns into matrix
X <- cbind(rep(1,n), X) # add a column of 1's before the actual data

x_0 <- data.matrix(x_0_df[,c('P1','P2')])
x_0 <- cbind(rep(1,1), x_0)

t_stat <- qt(q2_pred_int_conf/2, df=n-3, lower.tail=FALSE)
sigma_hat <- (sse_fm/(n-3))^0.5 # since 3 coefficients in full model
se_y_0_hat <- sigma_hat*(1 + (x_0 %*% solve(t(X) %*% X) %*% t(x_0)))^0.5

# the lower and upper limits of the prediction interval are as follows:
y_0_hat - t_stat*se_y_0_hat
```

```
##          [,1]
## [1,] 71.79724
```

```r
y_0_hat + t_stat*se_y_0_hat
```

```
##         [,1]
## [1,] 89.6284
```

```r
# check answers
predict(q2_model3, newdata=x_0_df, interval="prediction", level=1-q2_pred_int_conf)
```

```
##        fit      lwr     upr
## 1 80.71282 71.79724 89.6284
```

We can observe that the prediction interval at the confidence level $\alpha = 0.05$ is [71.79724, 89.6284]. (You may modify the R code by changing the value of q2__pred__int__conf to see the prediction interval for other values of alpha.)

```r
# Q4a
q4a_rm <- lm(Y-0.5*(X1+X3) ~ 1, data=data)
q4a_fm <- lm(Y-0.5*(X1+X3) ~ X1+X3, data=data)
anova(q4a_rm, q4a_fm)
```

```
## Analysis of Variance Table
##
## Model 1: Y - 0.5 * (X1 + X3) ~ 1
## Model 2: Y - 0.5 * (X1 + X3) ~ X1 + X3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     29 1469.6
## 2     27 1254.7  2    214.93 2.3126 0.1183
```

```r
qf(1-0.05, df1=2, df2=27, lower.tail=TRUE) # crit value for 4a
```

```
## [1] 3.354131
```

```r
qf(1-0.05, df1=2, df2=26, lower.tail=TRUE) # crit value for 4b
```

```
## [1] 3.369016
```

Here, the trick is to notice that under the null hypothesis $H_0 : \beta_1 = \beta_3 = 0.5$, the model is $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$. Then, let $\alpha_0 = \beta_0, \alpha_1 = \beta_1 - 0.5, \alpha_3 = \beta_3 - 0.5$ and write $Y - 0.5(X_1 + X_3) = \beta_0 + (\beta_1 - 0.5)X_1 + (\beta_3 - 0.5)X_3 + \epsilon$ and write $Y - 0.5(X_1 + X_3) = \alpha_0 + \alpha_1 X_1 + \alpha_3 X_3 + \epsilon$ (this is the new full model)

We may also re-express the null hypothesis as follows: $H_0 : \alpha_1 = \alpha_3 = 0$. Under H_0, the model is $Y = \alpha_0 + \epsilon$ (reduced model) and the alternative hypothesis is that at least one of $\{\alpha_1, \alpha_3\}$ is nonzero. In general, the `lm(Y~1)` command fits an intercept-only model to the response variable Y.

From the ANOVA table, we see that the p-value = Pr(>F)=0.1183, so we fail to reject the null hypothesis at any significance level <0.1183 (eg 0.05.) From the ANOVA table, we see that the F-statistic is 2.3126 < critical value of 3.35, so we fail to reject the null hypothesis at any significance level 0.05.

```r
# Q4b
q4b_rm <- lm(Y-0.5*(X1+X3) ~ X2, data=data)
q4b_fm <- lm(Y-0.5*(X1+X3) ~ X1+X2+X3, data=data)
anova(q4b_rm, q4b_fm)
```

```
## Analysis of Variance Table
##
## Model 1: Y - 0.5 * (X1 + X3) ~ X2
## Model 2: Y - 0.5 * (X1 + X3) ~ X1 + X2 + X3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 1410.7
## 2     26 1224.6  2    186.08 1.9753  0.159
```

We use the same idea as in Q4a. Notice that under the null hypothesis $H_0 : \beta_1 = \beta_3 = 0.5$, the model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$. Then, let $\alpha_0 = \beta_0, \alpha_1 = \beta_1 - 0.5, \alpha_2 = \beta_2, \alpha_3 = \beta_3 - 0.5$ and write $Y - 0.5(X_1 + X_3) = \beta_0 + (\beta_1 - 0.5)X_1 + \beta_2 X_2 + (\beta_3 - 0.5)X_3 + \epsilon$ and write $Y - 0.5(X_1 + X_3) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \epsilon$ (this is the new full model)

We may also re-express the null hypothesis as follows: $H_0 : \alpha_1 = \alpha_3 = 0$. Under H_0, the model is $Y = \alpha_0 + \alpha_2 X_2 + \epsilon$ (reduced model) and the alternative hypothesis is that at least one of $\{\alpha_1, \alpha_3\}$ is nonzero. As such, we use the anova command on the reduced and full models.

From the ANOVA table, we see that the p-value = Pr(>F)=0.159, so we fail to reject the null hypothesis at any significance level <0.159 (eg 0.05.) Alternatively, we see that the F-statistic is 1.9753 < critical value of

3.37, so we fail to reject the null hypothesis at any significance level 0.05.

```
# Q5
qf(0.05, df1=4, df2=88, lower.tail=FALSE)
```

```
## [1] 2.475277
```

```
t_stat = 2.16
pt(abs(t_stat), df=88, lower.tail=FALSE)
```

```
## [1] 0.01674573
```

```
qt(1-0.05, df=88, lower.tail=TRUE)
```

```
## [1] 1.662354
```

```
# Q6
qf(0.05, df1=3, df2=88, lower.tail=FALSE)
```

```
## [1] 2.708186
```

```
# Q7a
cr_data <- read.table(file="Computer_Repair.txt",header=TRUE)
cr_model = lm(Minutes ~ Units, data=cr_data)
summary(cr_model)
```

```
##
## Call:
## lm(formula = Minutes ~ Units, data = cr_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.603 -14.801  -0.045  17.335  29.092
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2127     7.9853    4.66  0.00012 ***
## Units         9.9695     0.7218   13.81 2.56e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.75 on 22 degrees of freedom
## Multiple R-squared:  0.8966, Adjusted R-squared:  0.8919
## F-statistic: 190.7 on 1 and 22 DF,  p-value: 2.556e-12
```

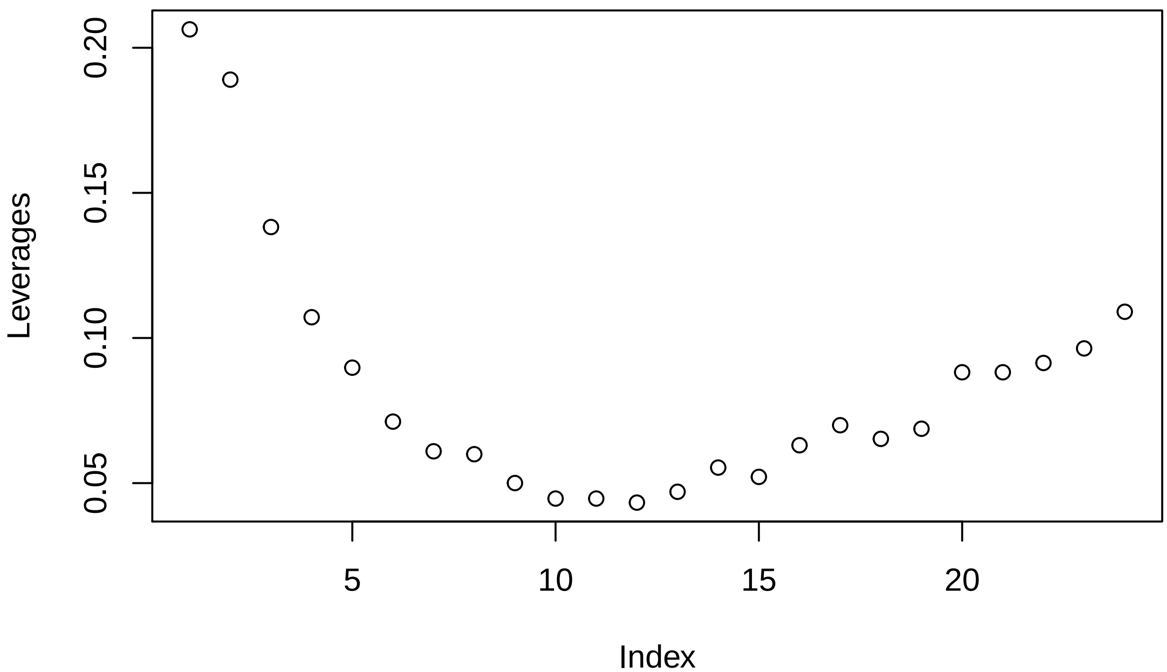
The regression equation is $\hat{Y} = 37.21 + 9.9695X_1$ where Y denotes Minutes, X_1 denotes number of Units.

The standard regression assumptions are: (1) Assumption of linearity (2.1) Assumption that errors are independent of each other (2.2) Assumption that errors are normally distributed (2.3) Assumption that errors each have mean 0 (2.4) Assumption that errors each have common variance σ^2 (3.1) Assumption that predictor variables $X_1, X_2, ...X_n$ are nonrandom. (3.2) Assumption that predictor values $x_{1j}, x_{2j}, ..., x_{nj}$ are measured without error. (3.3) Assumption that predictors $X_1, X_2, ...X_n$ are independent of each other. (4) Assumption that all observations are equally reliable and have an approximately equal role in determining regression results.

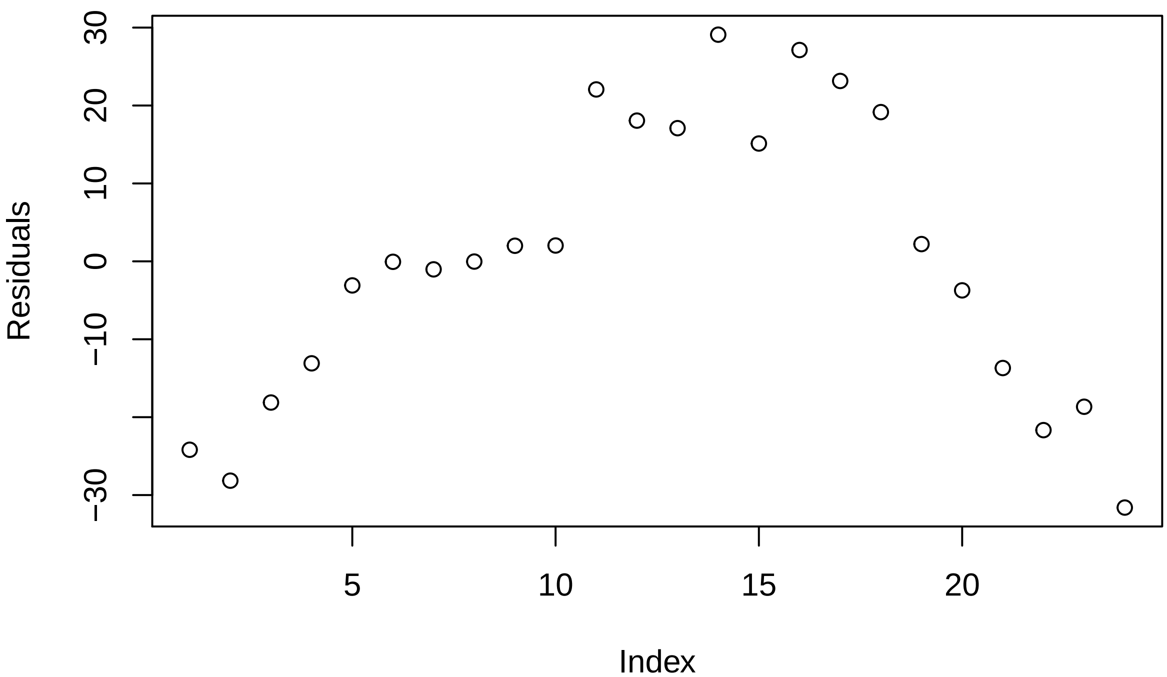
Of these assumptions, (3.2) and (3.1) are difficult to assume. Also, there is only 1 predictor so (3.3) is automatically satisfied. Therefore, we will only examine the rest.

```
# Q7b
X <- as.matrix(cr_data[,-1])
```

```
X <- cbind(rep(1,length(cr_data$Units)),X)
hat_mat <- X %*% solve(t(X) %*% X) %*% t(X)
leverages <- diag(hat_mat)
plot(seq(1,length(cr_data$Units)), leverages, xlab="Index", ylab="Leverages")
```



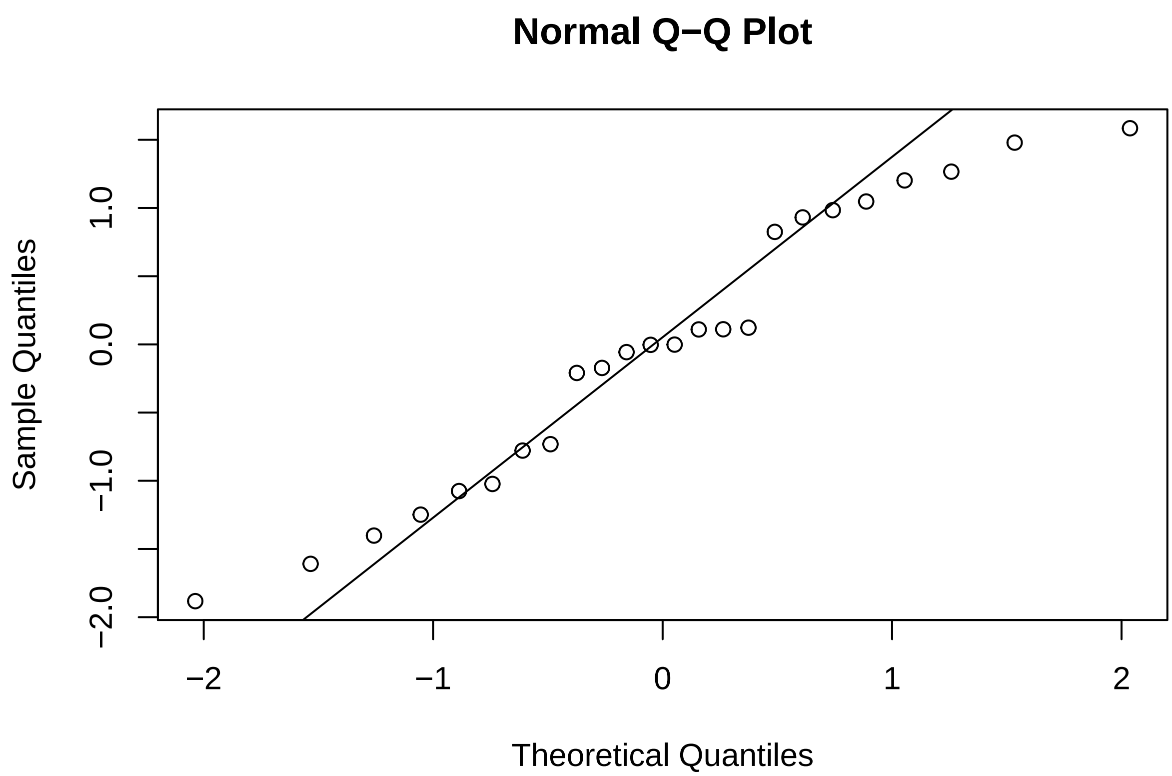

```
plot(seq(1,length(cr_data$Units)), residuals(cr_model), xlab="Index", ylab="Residuals")
```



As seen here, the leverages do not follow a random pattern (they seem to be quadratic) - observations no. 6-15 have low leverages while observations 1-5 and 16-25 have substantially higher leverages.

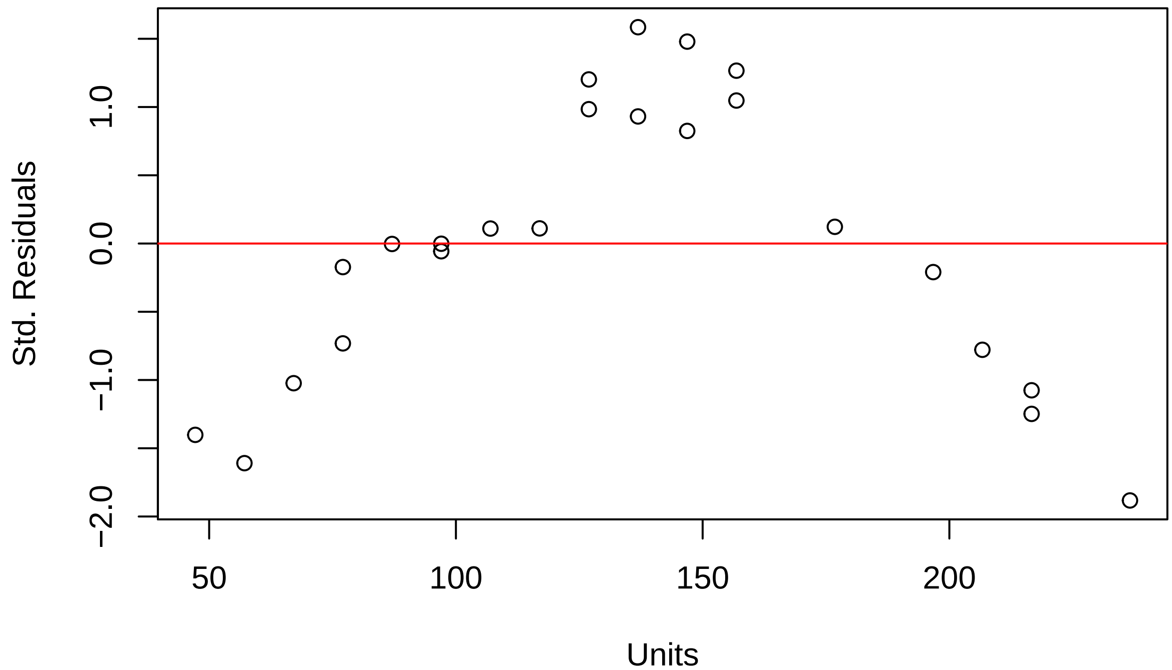
If the residuals were independent of each other, we would expect that the scatter plot of residuals by observation number would have a random pattern (2.1). However, this is not the case, as in the residuals-index plot, the residuals seem to follow a quadratic pattern. A quadratic model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$ may be more suitable. Hence (2.1) is violated.

```
# Q7b cont.
cr_model.stdres = rstandard(cr_model)
qqnorm(cr_model.stdres, ylab="Sample Quantiles", xlab="Theoretical Quantiles")
qqline(cr_model.stdres)
```



Under the normality of errors assumption (2.2), the ordered residuals should be approximately form a straight line with slope 1 and intercept 0 with the quantiles of the standard normal distribution. From examining the Q-Q plot of the standardised residuals against the standard normal distribution, we notice that there appears to be a fair amount of deviation (especially in the first two and last four points), hence it is likely (2.2) is violated.

```
# Q7b cont.
# remember this is plot(x, y) so stdres needs to be 2nd argument
plot(fitted.values(cr_model), cr_model.stdres, ylab="Std. Residuals", xlab="Units")
abline(a=0, b=0, col="red")
```




```
mean(cr_model.stdres)
```

```
## [1] -0.02214834
```

In the case of simple linear regression, the plot of standard residuals against the predictor and the plot of standard residuals against the response are identical.

The standard residuals should be uncorrelated with the predictor variables/response. As such when plotting the std. residuals against each of the predictors, we expect to see a random scatter of points, which is NOT the case, since fitted values in the 10-15 region seem to indicate a negative standard residual, while fitted values < 5 seem to indicate a slightly positive standard residual. Hence the linearity assumption (1) is not satisfied despite the mean of the standardised residuals (-0.022) being approximately close to 0. At each value of Units, the mean of the standardised residuals differs and is not necessarily close to 0, we can say (2.3) (errors having mean 0) is violated.

At every fitted value, the spread of the residuals is roughly the same (between 1 and -1). Hence the constant variance assumption (2.4) is satisfied.

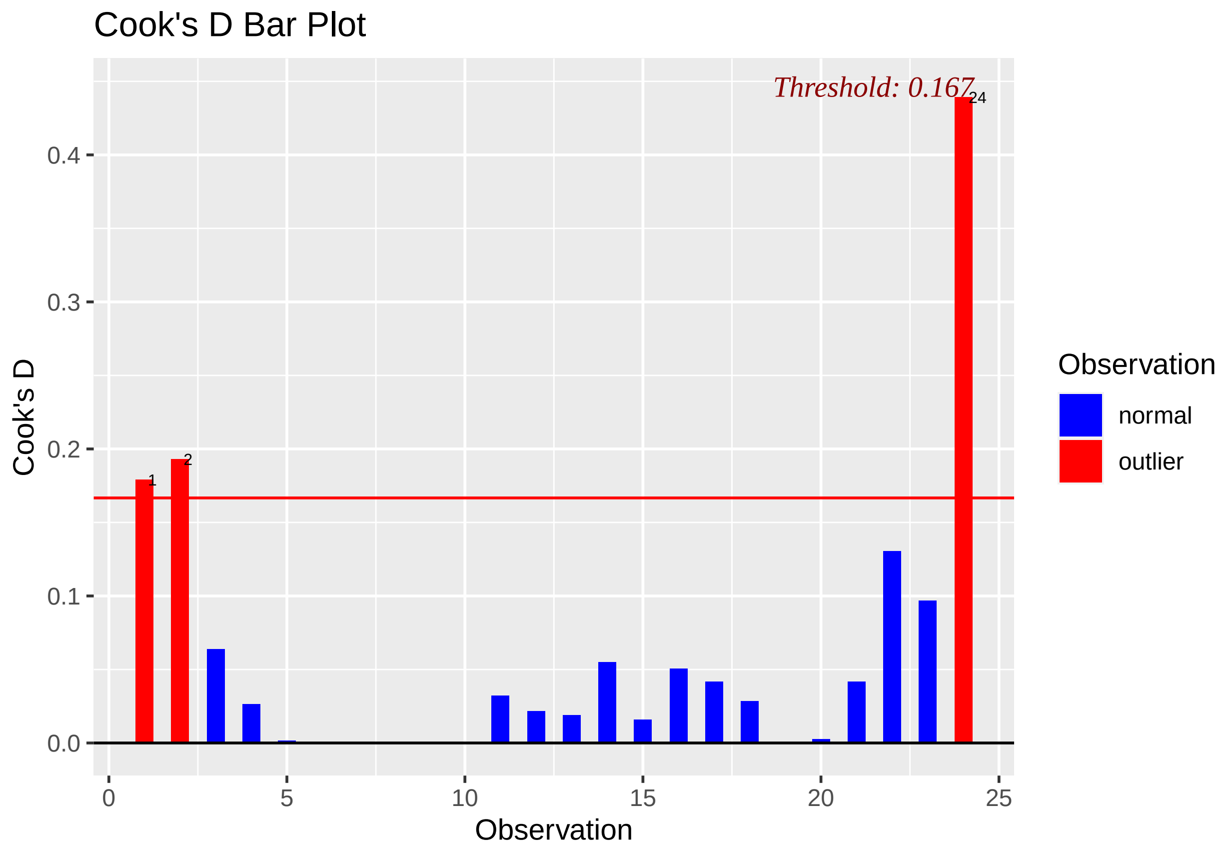
```
# Q7b cont.
# this didn't work
#update.packages(checkBuilt=TRUE)
#install.packages("car", dependencies=TRUE)
#install.packages("olsrr")

# to actually install a package go to Tools > Install Packages > type in package's name
library(olsrr)
```

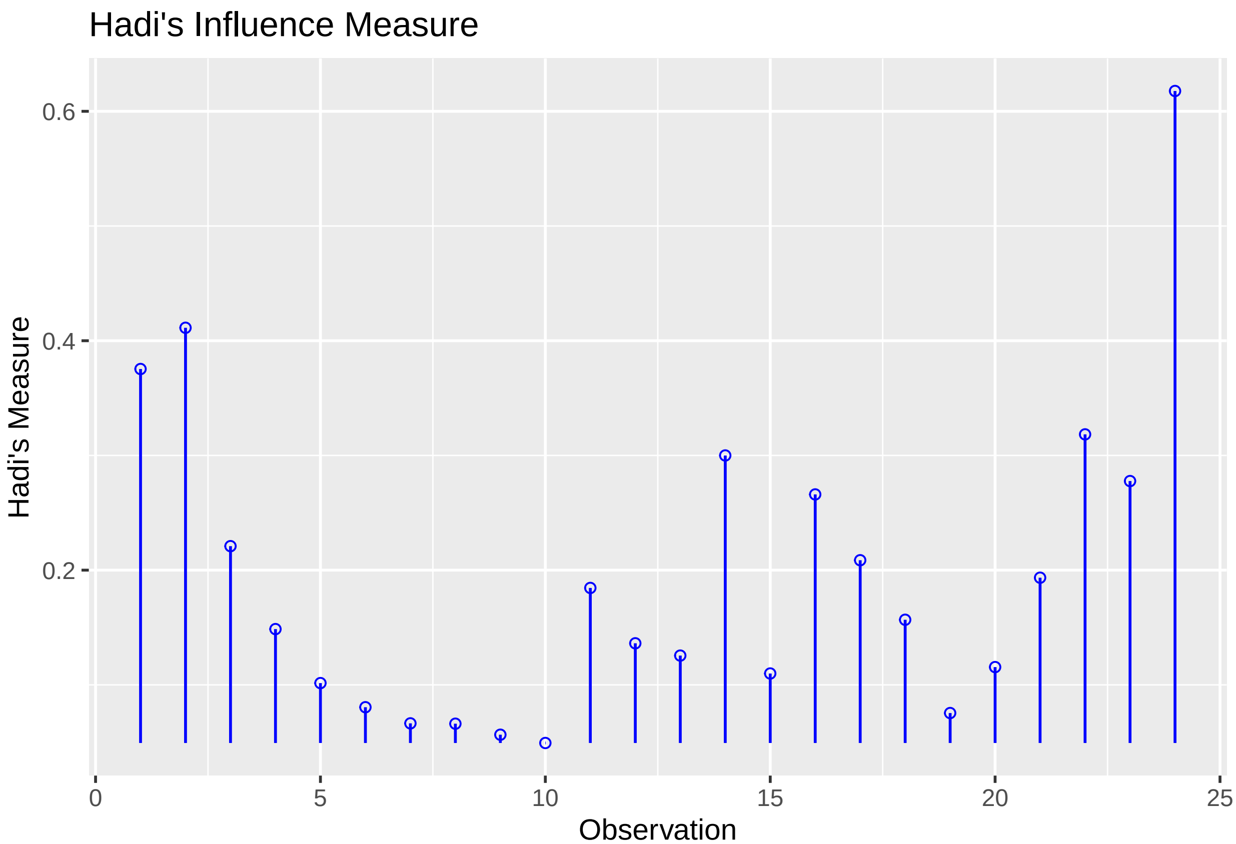
```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

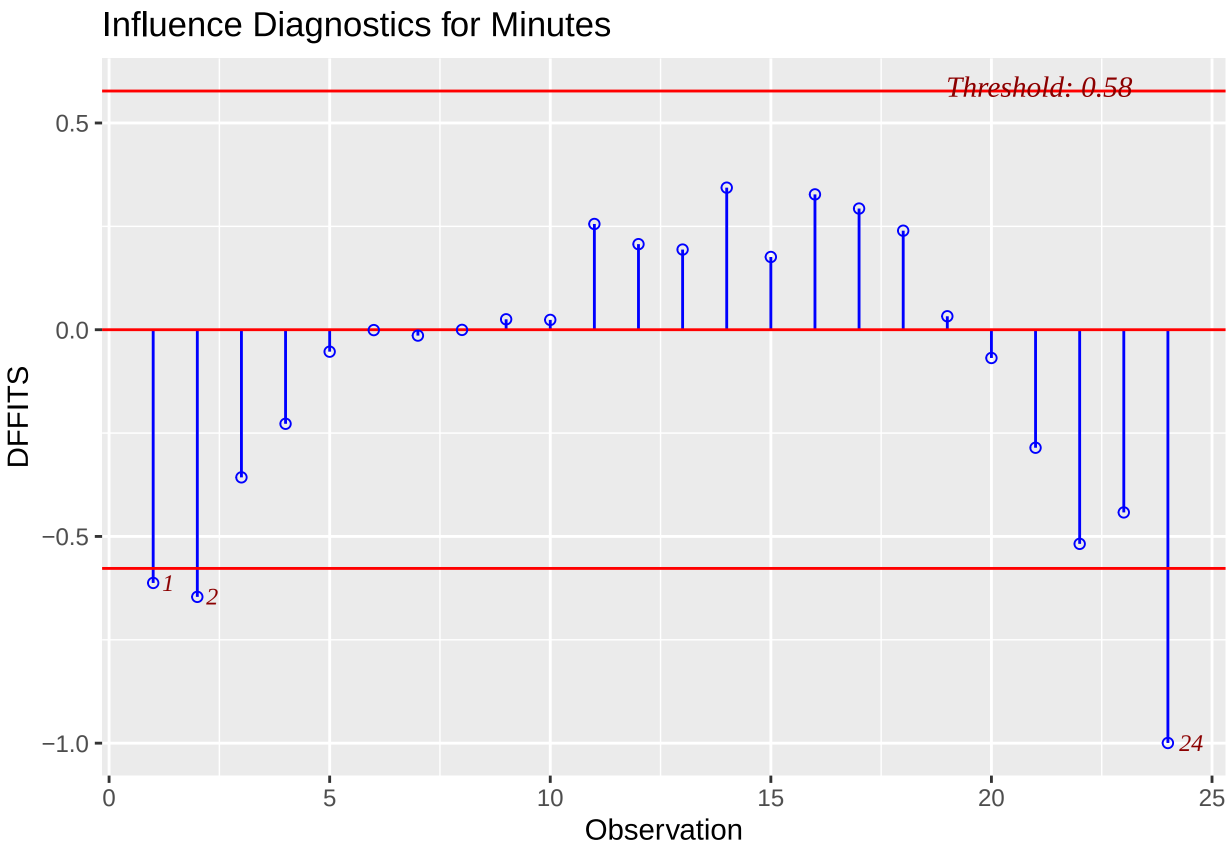
```
ols_plot_cooksd_bar(cr_model)
```



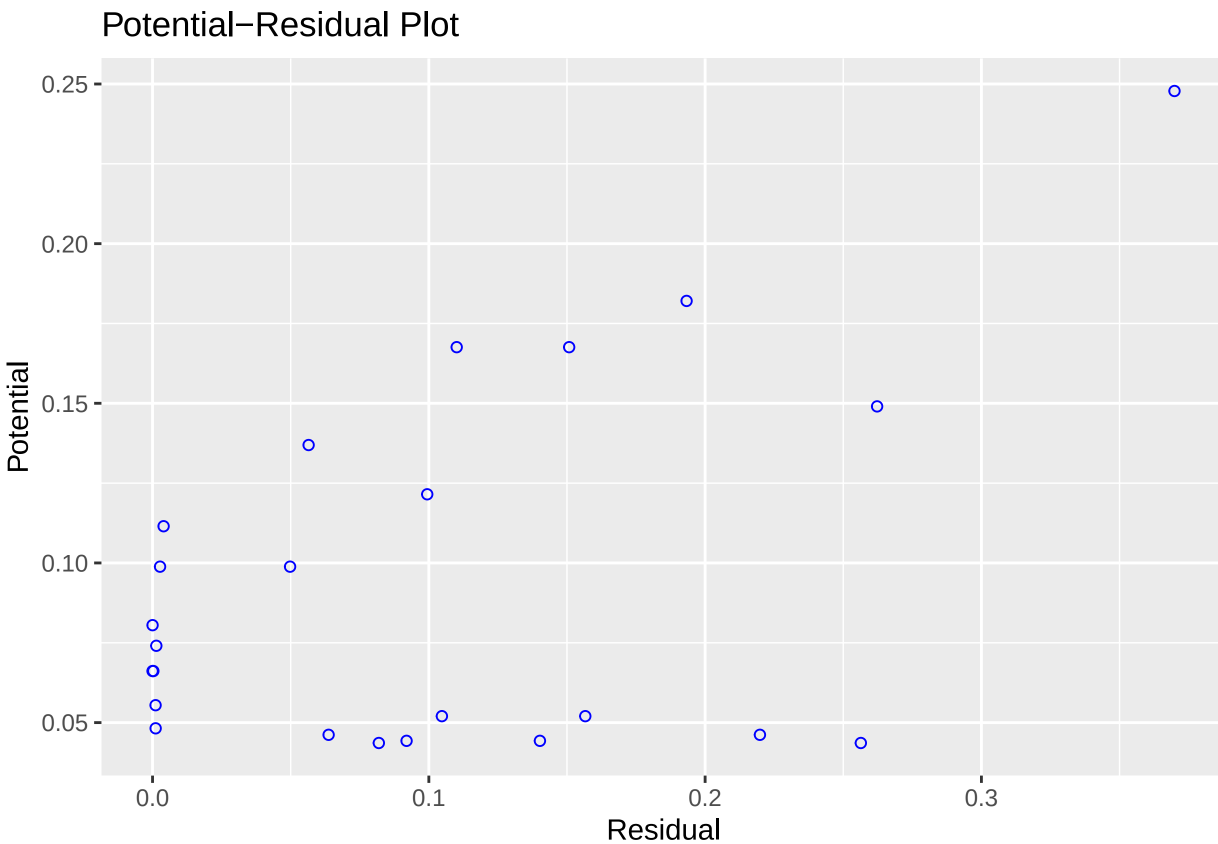

```
ols_plot_hadi(cr_model)
```




```
ols_plot_dffits(cr_model)
```




```
ols_plot_resid_pot(cr_model)
```



From observing the various plots (Cook's distance, DFITS, Hadi, potential-residual), we notice that in all cases, observation #24 far exceeds the threshold or is significantly different than the other observations. In the potential-residual plot, #24 has high residual and decently high potential. Hence obs. #24 is an influential point (so are #1 and #2 in Cook's distance plot.) Therefore assumption (4) is violated since not all assumptions have an equal role in determining the regression result.

Q10: According to the notes, outliers in the response are those whose standardised residual $r_i \geq 3$ and outliers in the predictor space are those whose leverage value $p_{ii} \geq \dfrac{2(p+1)}{n}$.

```
#install.packages(scatterplot3d)
#library(scatterplot3d)
#scatterplot3d(exm_data)
#pairs(exm_data[,1:3], lower.panel=NULL)
```

In models 1 and 2, high leverage points have $p_{ii} \geq 2(1+1)/22 = 4/22$ meaning the potential function is threshold $4/22/(1-4/22) = 2/9$. In model 3, high leverage points have $p_{ii} \geq 2(2+1)/22 = 6/22$ meaning the potential function is threshold $6/22/(1-6/22) = 3/8$.

```
# Q10a
q2_model1.stdres = rstandard(q2_model1)
abs(q2_model1.stdres)>=3
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    14    15    16    17    18    19    20    21    22
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
q2_model2.stdres = rstandard(q2_model2)
abs(q2_model2.stdres)>=3
```

```
##      1      2      3      4      5      6      7      8      9     10     11     12     13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     14     15     16     17     18     19     20     21     22
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

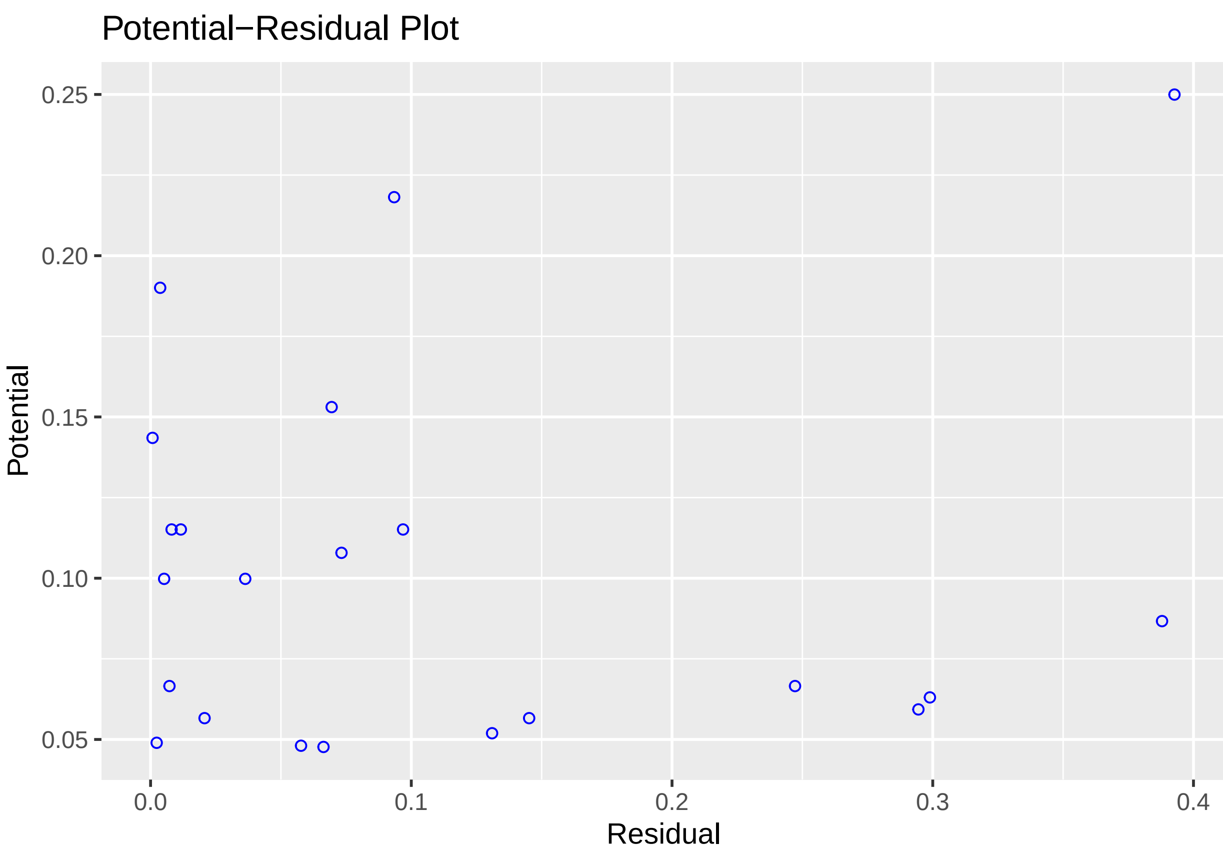
```
q2_model3.stdres = rstandard(q2_model3)
abs(q2_model3.stdres)>=3
```

```
##      1      2      3      4      5      6      7      8      9     10     11     12     13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     14     15     16     17     18     19     20     21     22
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

However, listing the standardised residuals of each model, we see no observation's std. residual exceeds 3 (although in models 2 and 3, observation no. 9 has $|r_i| \geq 2$ so it is likely an outlier in response space.)

First look at points whose potential function exceeds the threshold, those are high-leverage points aka outliers in the predictor space.

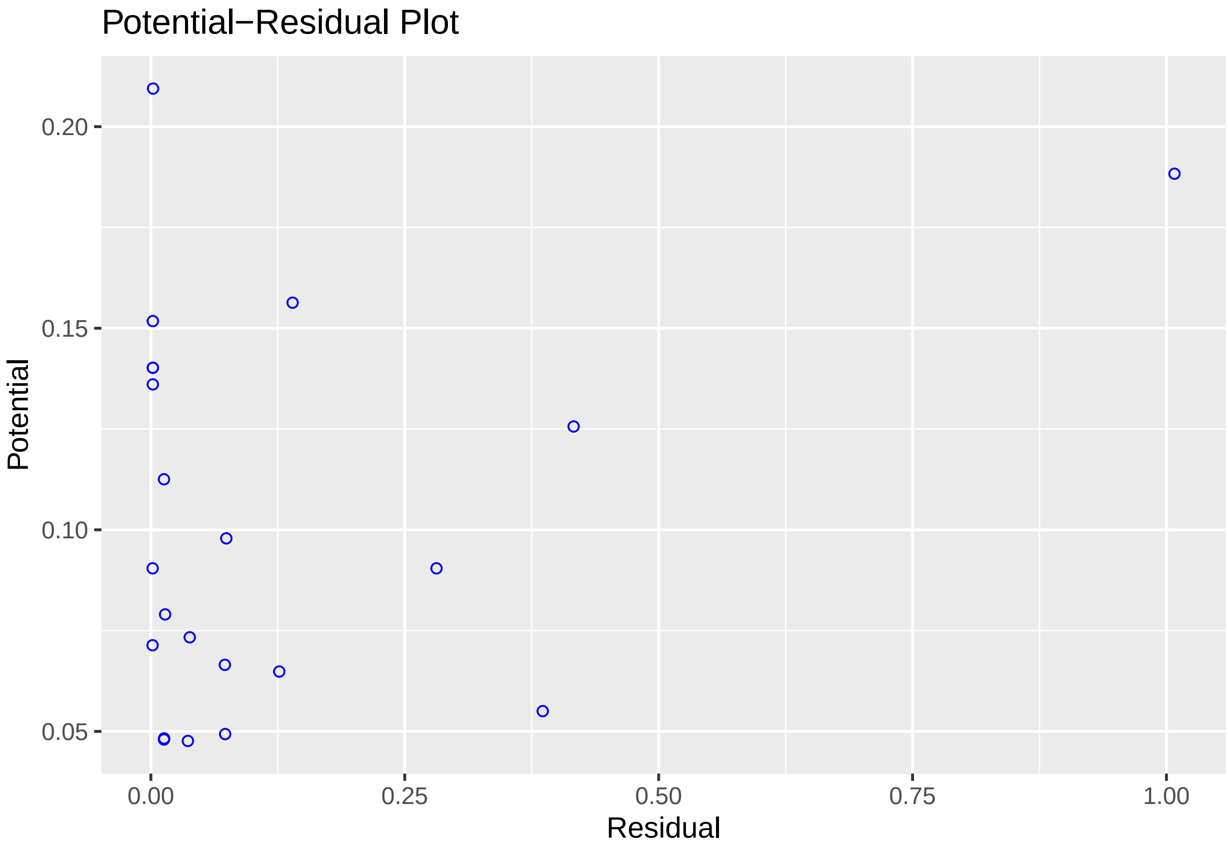
```
q2_model1_pr <- ols_plot_resid_pot(q2_model1)
```




```
q2_model1_pr[["data"]][["pot"]]>=2/9
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

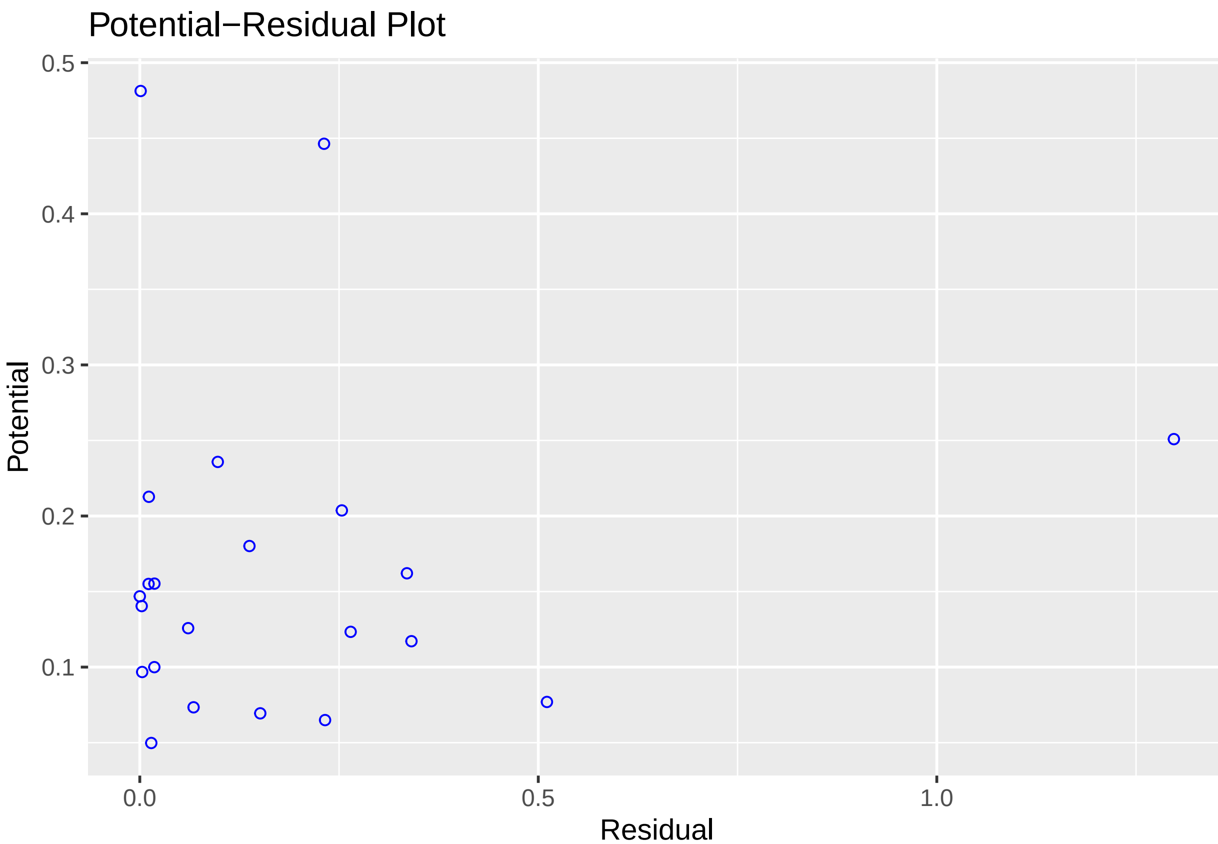
```
q2_model2_pr <- ols_plot_resid_pot(q2_model2)
```




```
q2_model2_pr[["data"]][["pot"]]>=2/9
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
q2_model3_pr <- ols_plot_resid_pot(q2_model3)
```




```
q2_model3_pr[["data"]][["pot"]]>=3/8
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

In conclusion:

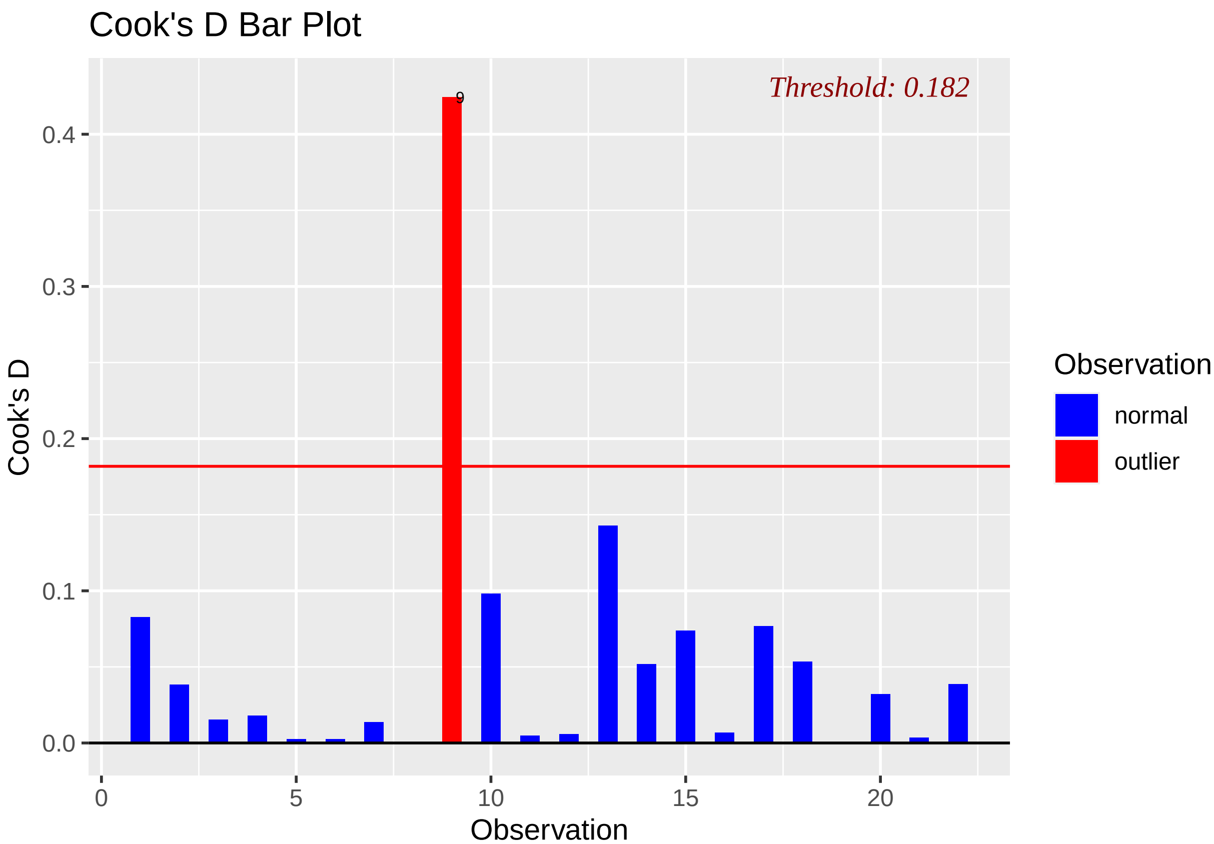
In model 1, points 9 are X-outliers.

In model 2, no X-outliers.

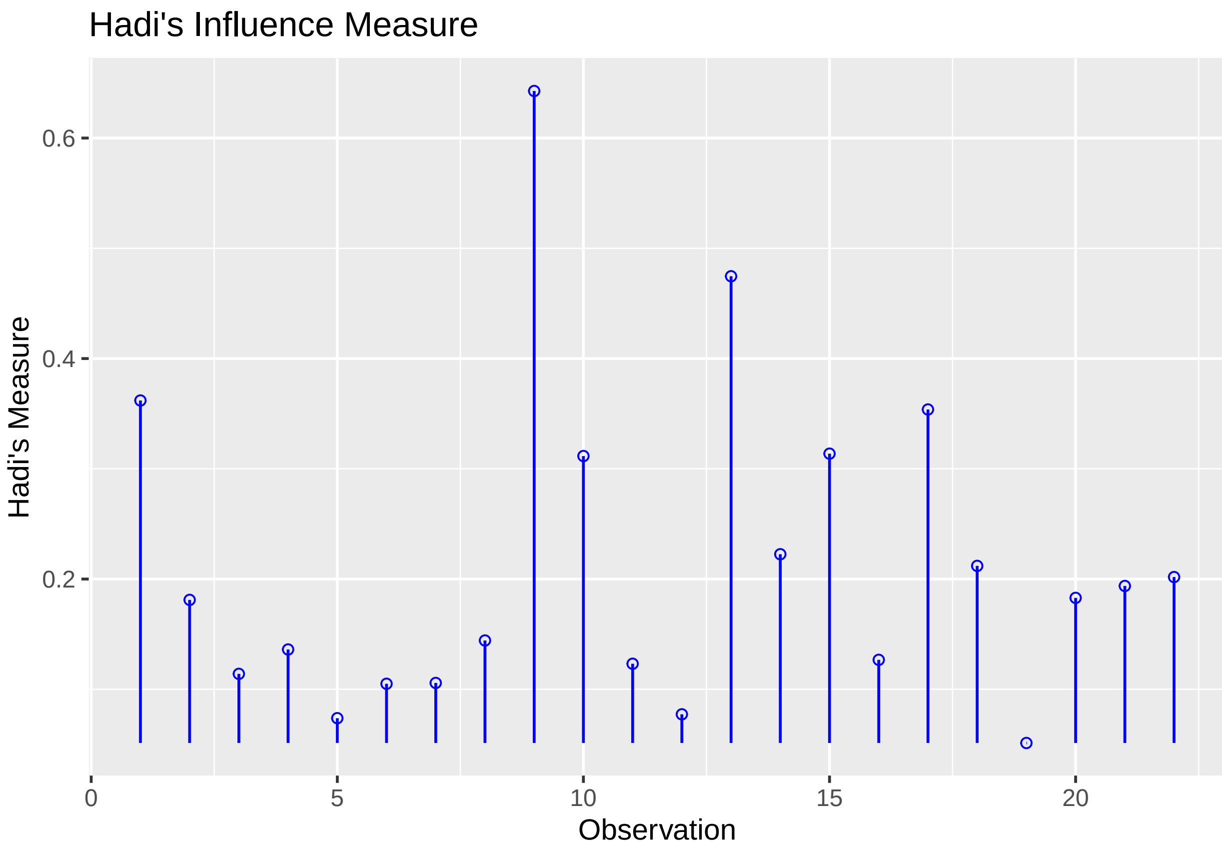
In model 3, points 7 and 15 are X-outliers.

Let's combine this knowledge with graphical methods (Cook's, DFITS, Hadi).

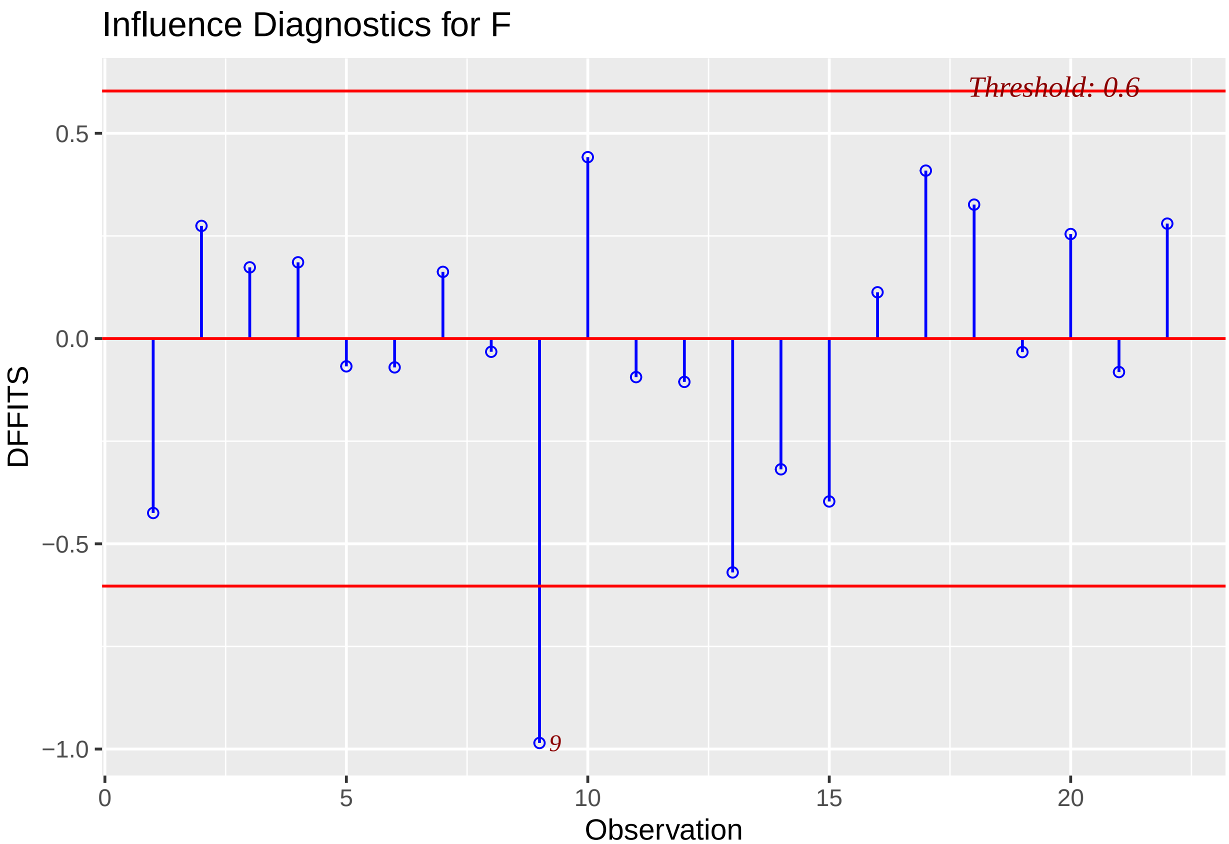
```
# Q10a cont.
ols_plot_cooksd_bar(q2_model1)
```



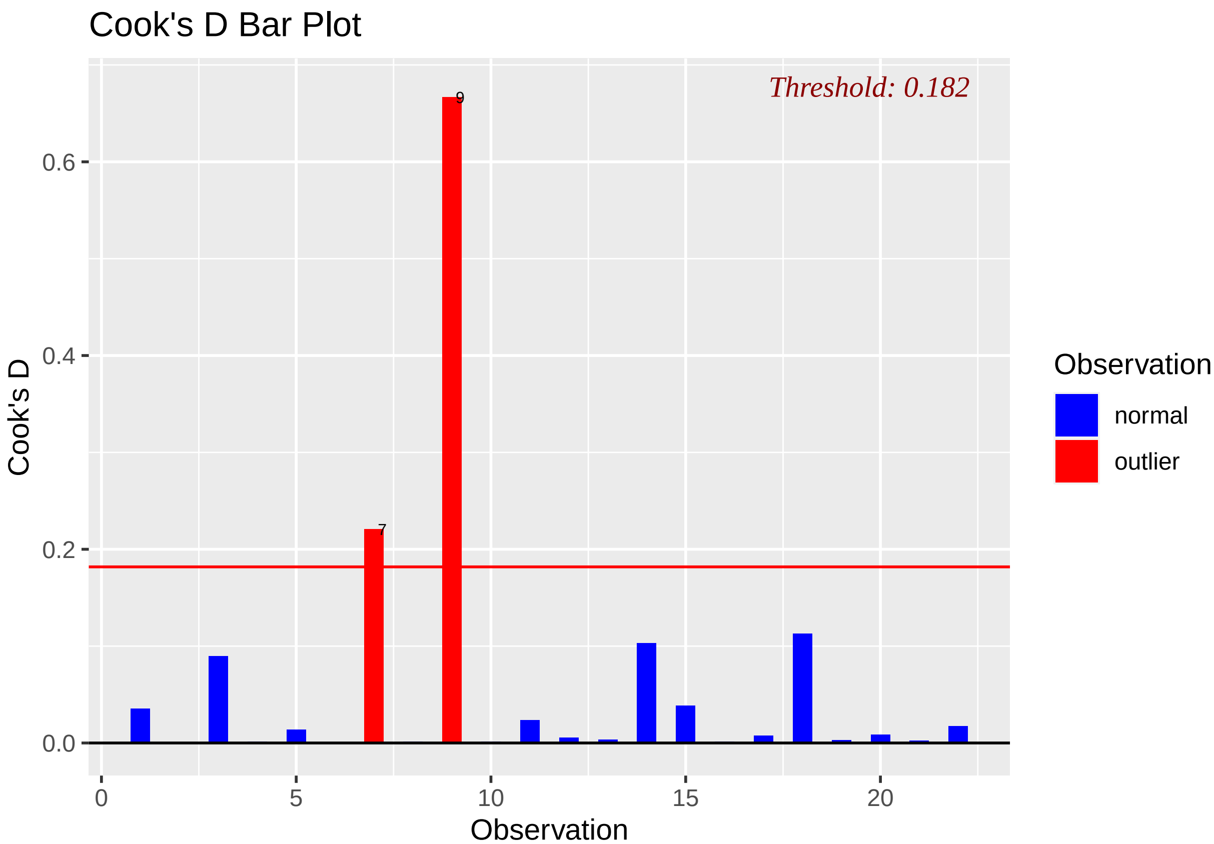

```
ols_plot_hadi(q2_model1)
```



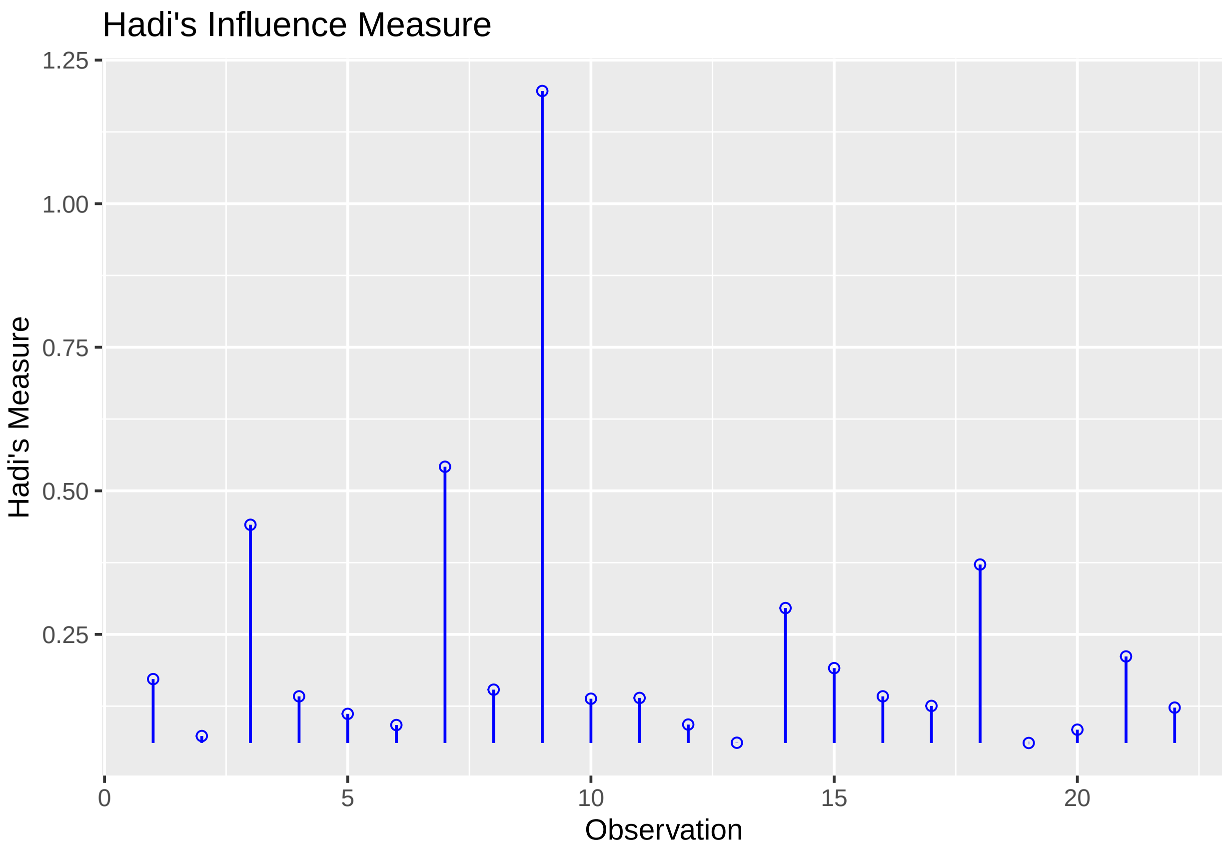

```
ols_plot_dffits(q2_model1)
```



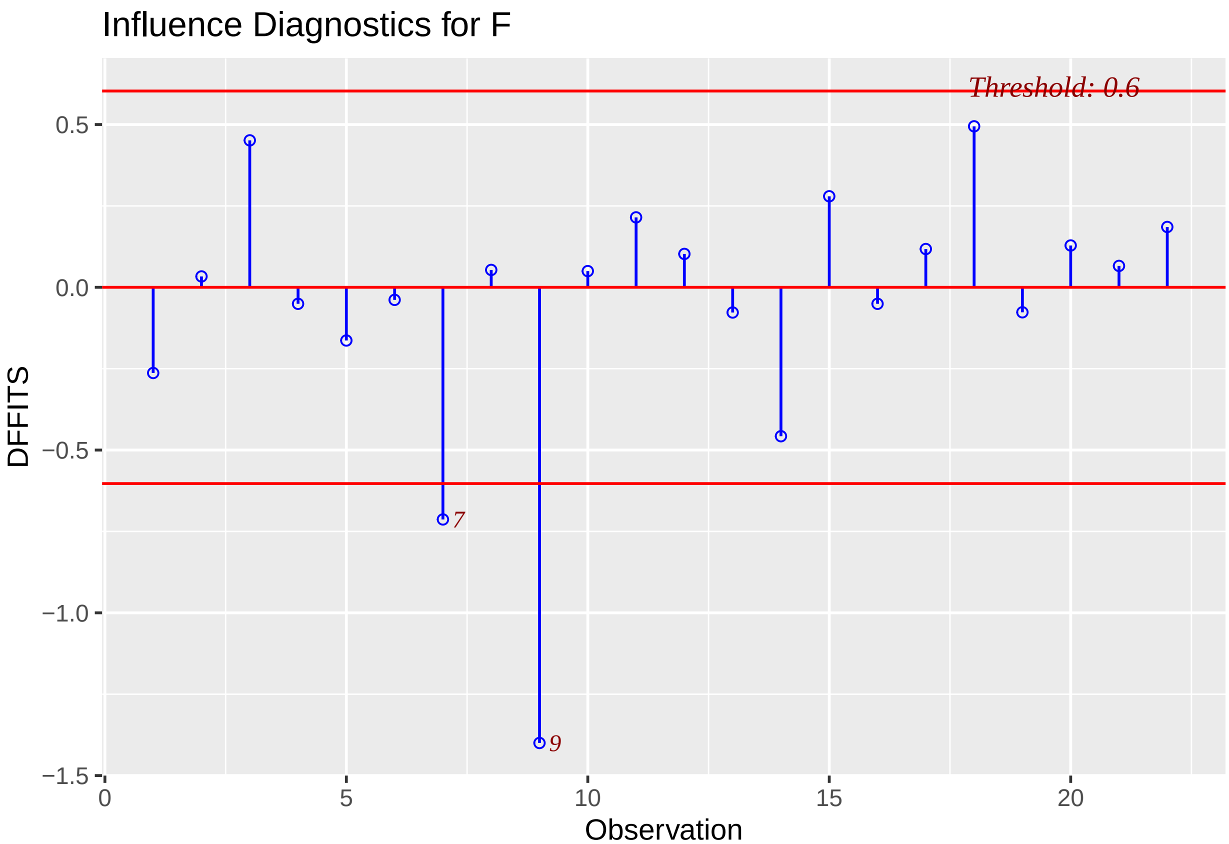

```
ols_plot_cooksd_bar(q2_model2)
```



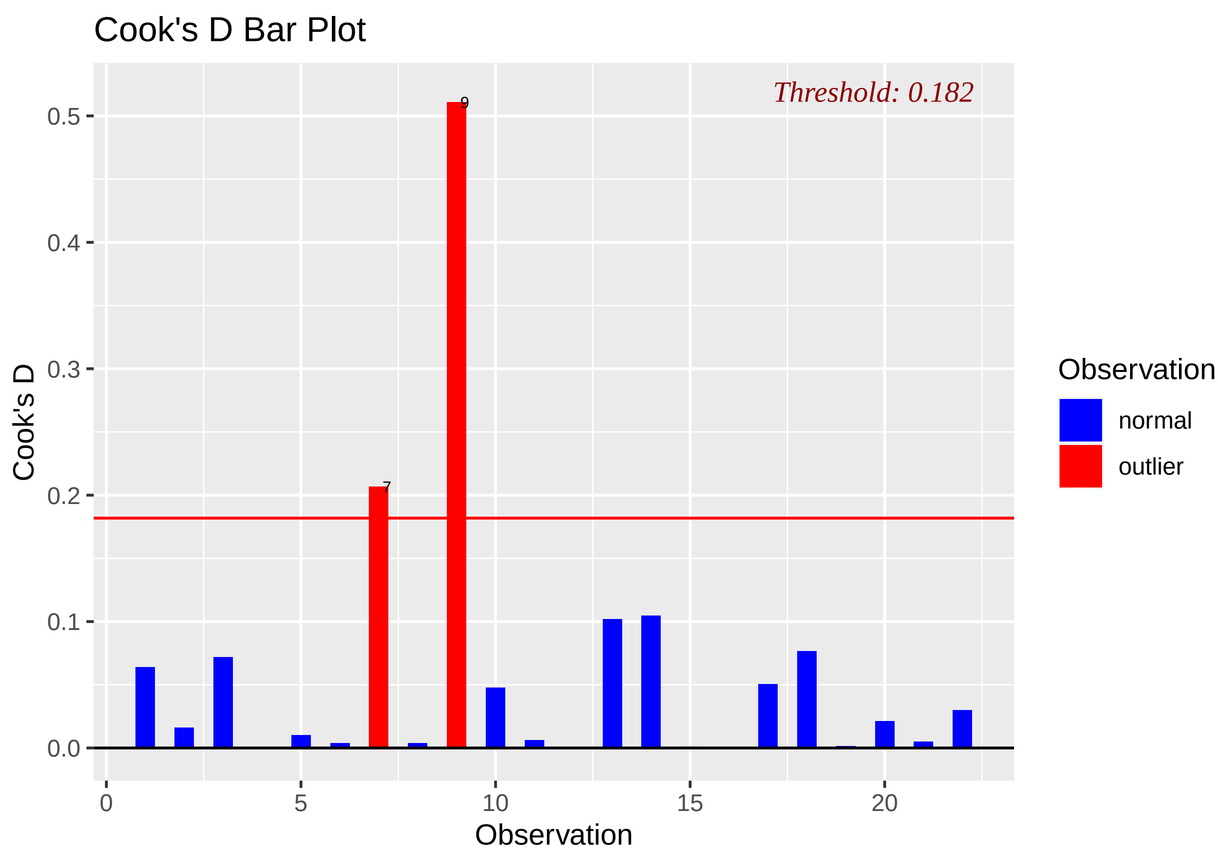

```
ols_plot_hadi(q2_model2)
```



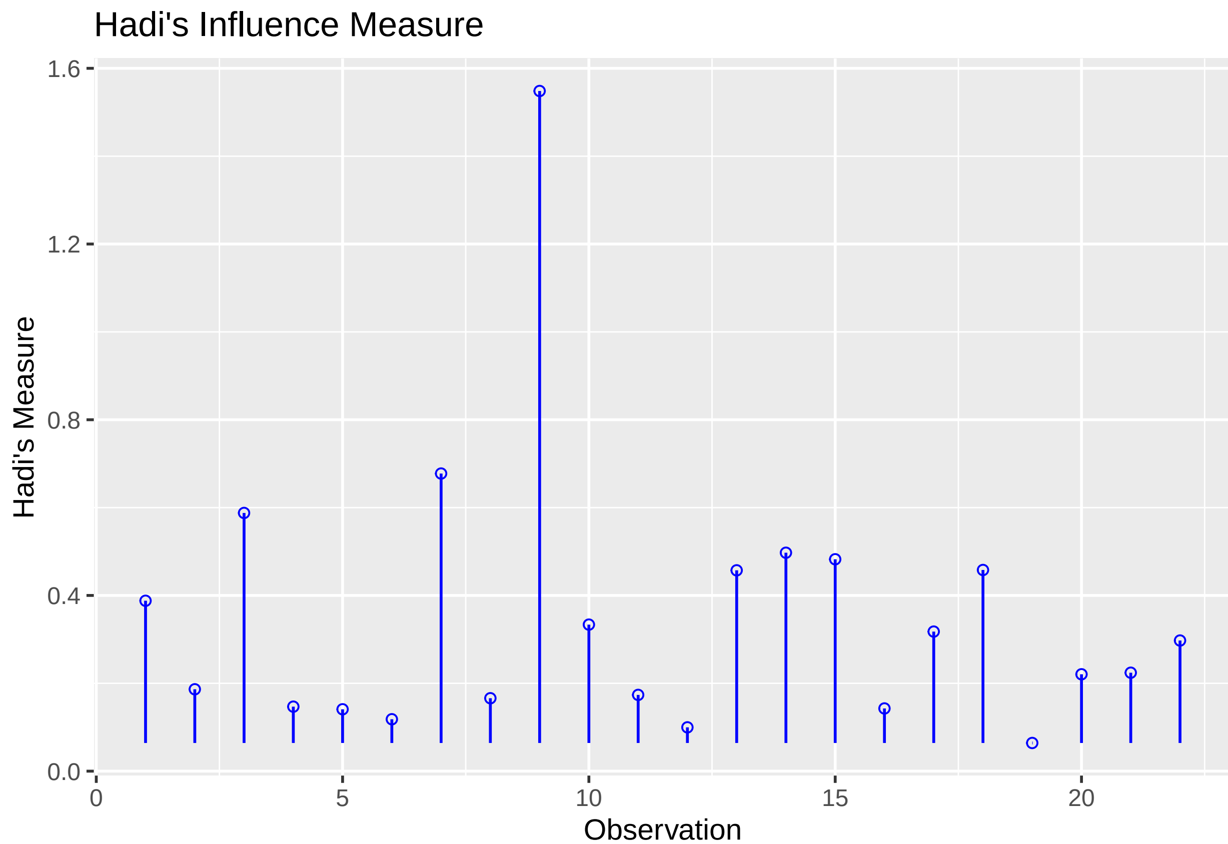

```
ols_plot_dffits(q2_model2)
```



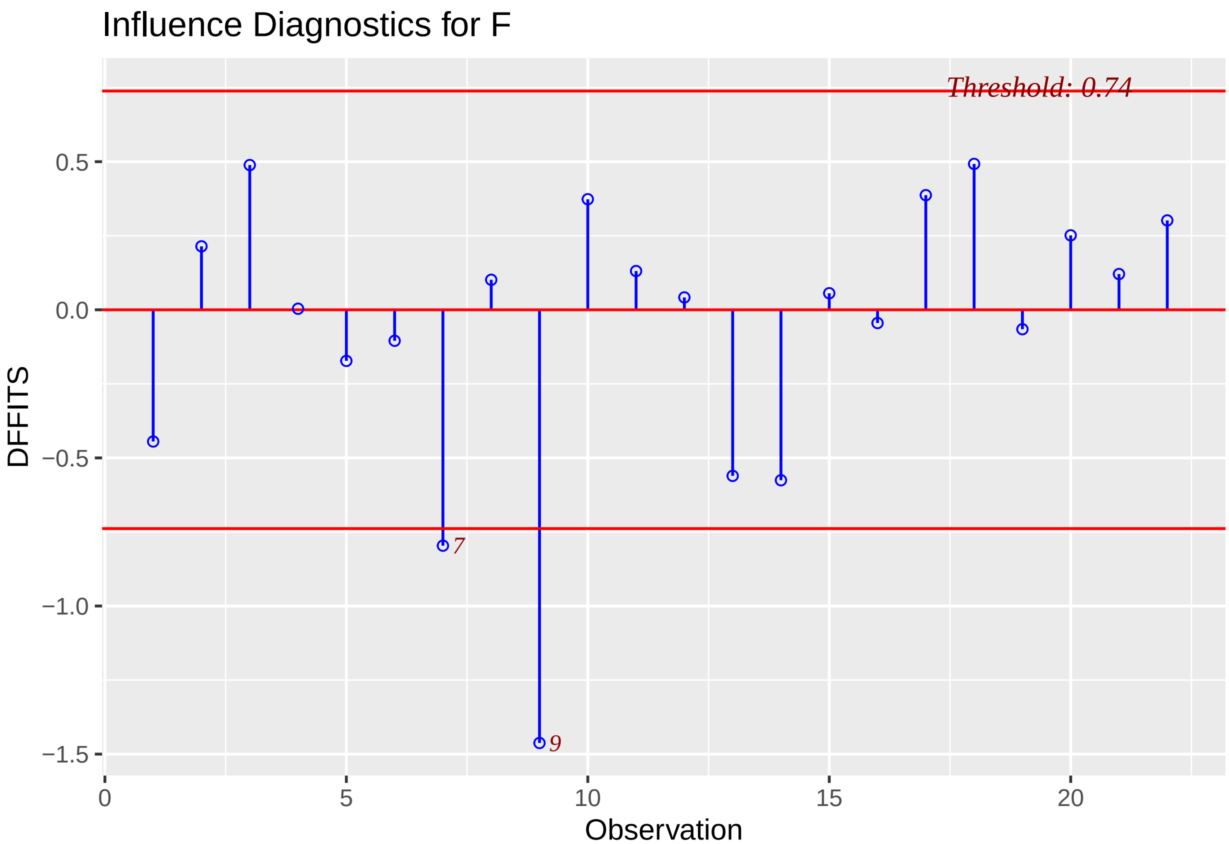

```
ols_plot_cooksd_bar(q2_model3)
```




```
ols_plot_hadi(q2_model3)
```




```
ols_plot_dffits(q2_model3)
```



According to the plots, in model 1, Cook's and DFITS show points 9 to be highly influential. In both model 2 and model 3, Cook's and DFITS show points 7 and 9 to be highly influential.

In conclusion:

In model 1, points 9 are X-outliers, point 9 is influential.

In model 2, no X-outliers. Points 7 and 9 are influential.

In model 3, points 7 and 15 are X-outliers. Points 7 and 9 are influential.

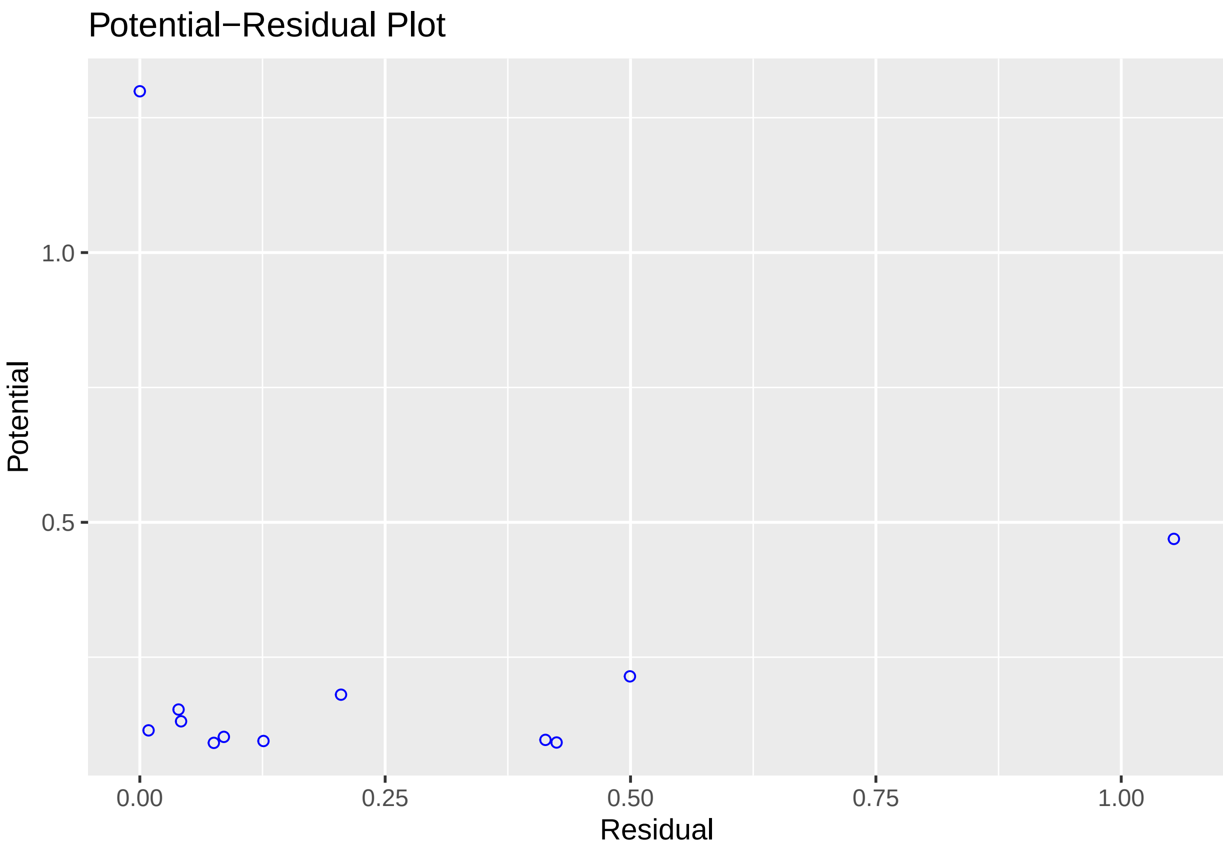
Q10b) According to the Hadi influence plots in Q10a, the model with the most even spread of influence is Model 1 (the other two have point 9's influence measure significantly higher than that of the others, at values near 1.1 and 1.6 respectively.) At such, models 2 and 3 may be less reliable when it comes to measuring the response variable if the predictor/response is an outlier, hence we should choose model 1 to predict F.

```
# Q11
q11_data <- data.frame(Y=c(8.11, 11, 8.2, 8.3, 9.4, 9.3, 9.6, 10.3, 11.3, 11.4, 12.2, 12.9),
                       X=c(0, 5, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24))
q11_model = lm(Y ~ X, data=q11_data)
summary(q11_model)
```

```
##
## Call:
## lm(formula = Y ~ X, data = q11_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7852 -0.8997 -0.1394  0.7607  2.2730
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.10970    1.05167   7.711 1.62e-05 ***
## X            0.12347    0.05826   2.119   0.0601 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.399 on 10 degrees of freedom
## Multiple R-squared:  0.3099, Adjusted R-squared:  0.2409
## F-statistic: 4.491 on 1 and 10 DF,  p-value: 0.0601
```

```r
q11_model_pr <- ols_plot_resid_pot(q11_model)
```




```r
# Q11
q11_model_pr[["data"]][["res"]]
```

```
##  [1] 2.163244e-08 1.053640e+00 4.132994e-01 4.246553e-01 7.542294e-02
##  [6] 1.259963e-01 8.564684e-02 8.901899e-03 4.202069e-02 3.947754e-02
## [11] 2.049993e-01 4.993842e-01
```

```r
q11_model_pr[["data"]][["pot"]]
```

```
##  [1] 1.29900332 0.46921444 0.09667195 0.09182707 0.09113844 0.09459032
##  [7] 0.10226187 0.11433172 0.13108859 0.15294902 0.18048448 0.21446121
```