final_analysis

Ilia Nikiforov

2022-11-29

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'stringr' was built under R version 4.1.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.1.3

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(olsrr)
```

```
## Warning: package 'olsrr' was built under R version 4.1.3

##
## Attaching package: 'olsrr'

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

```
library(boot)
library(MASS)
```

```
##
## Attaching package: ’MASS’

## The following object is masked from ’package:olsrr’:
##
##     cement

## The following object is masked from ’package:dplyr’:
##
##     select
```

```
library(car)
```

```
## Warning: package ’car’ was built under R version 4.1.3

## Loading required package: carData

## Warning: package ’carData’ was built under R version 4.1.3

##
## Attaching package: ’car’

## The following object is masked from ’package:boot’:
##
##     logit

## The following object is masked from ’package:dplyr’:
##
##     recode

## The following object is masked from ’package:purrr’:
##
##     some
```

```
setwd("D:\\UCSB\\MAT265\\Final")
data = read.csv("checkouts_sample_custom_variables.csv")
```

EDA - variables solo

```
response = ggplot(data) + geom_boxplot(aes(y=duration)) +
  ggtitle("Boxplot of duration")

ggsave("response.png",plot=response)
```

```
## Saving 6.5 x 4.5 in image
```

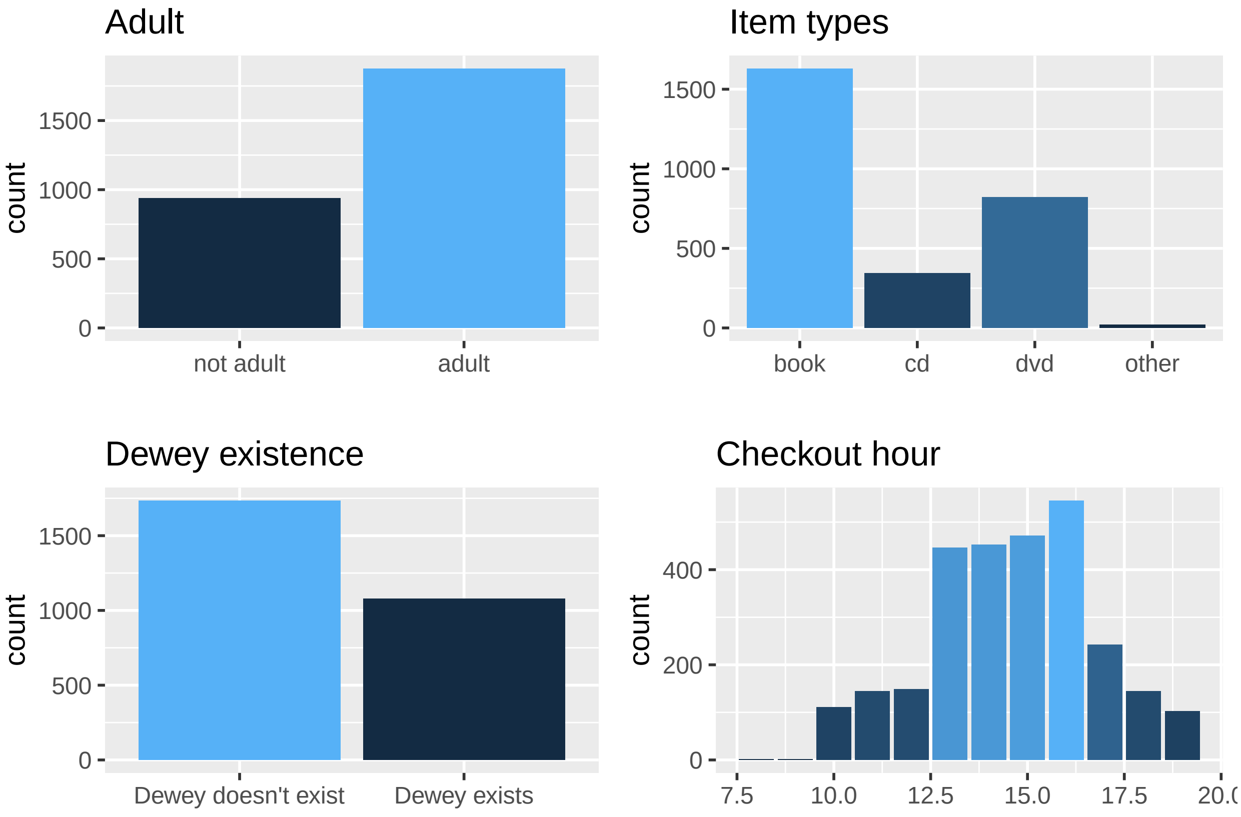
```
pred_adult = data %>% ggplot(aes(x=factor(adult))) + geom_bar(aes(fill=..count..),show.legend = FALSE) +
  scale_x_discrete(labels=c("0"="not adult","1"="adult"))+
  ggtitle("Adult")+
  xlab("")

pred_types = data %>% ggplot(aes(x=item_type)) + geom_bar(aes(fill=..count..),show.legend = FALSE)+
  ggtitle("Item types")+
  xlab("")


pred_dewey = data %>% ggplot(aes(x=factor(dewey_exists))) + geom_bar(aes(fill=..count..),show.legend =
  scale_x_discrete(labels=c("0"="Dewey doesn't exist","1"="Dewey exists"))+
  ggtitle("Dewey existence")+
  xlab("")

pred_hour = data %>% ggplot(aes(x=checkout_hour)) + geom_bar(aes(fill=..count..),show.legend = FALSE)+
  ggtitle("Checkout hour")+
  xlab("")

preds = grid.arrange(pred_adult,pred_types,pred_dewey,pred_hour)
```



```
ggsave("preds.png",plot=preds)
```

```
## Saving 6.5 x 4.5 in image
```

EDA - reponse vs predictors

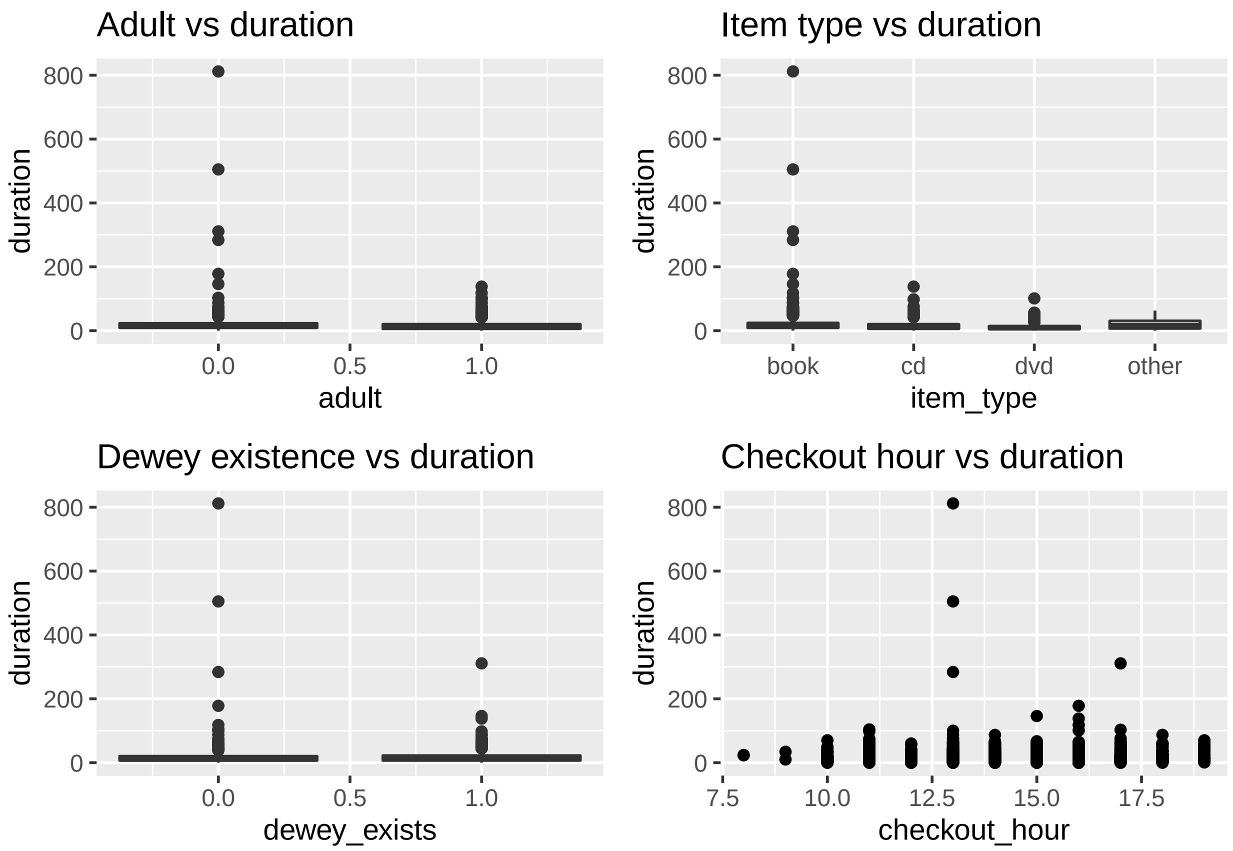
```
a =data %>% ggplot(aes(x=adult,y=duration,group=adult)) + geom_boxplot() +
  ggtitle("Adult vs duration")

b = data %>% ggplot(aes(x=item_type,y=duration,group=item_type)) + geom_boxplot() +
  ggtitle("Item type vs duration")

c = data %>% ggplot(aes(x=dewey_exists,y=duration,group=dewey_exists)) + geom_boxplot() +
  ggtitle("Dewey existence vs duration")

d= data %>% ggplot(aes(x=checkout_hour,y=duration)) + geom_point() +
  ggtitle("Checkout hour vs duration")

preds_response = grid.arrange(a,b,c,d)
```



```
ggsave("preds_response.png",plot=preds_response)
```

```
## Saving 6.5 x 4.5 in image
```

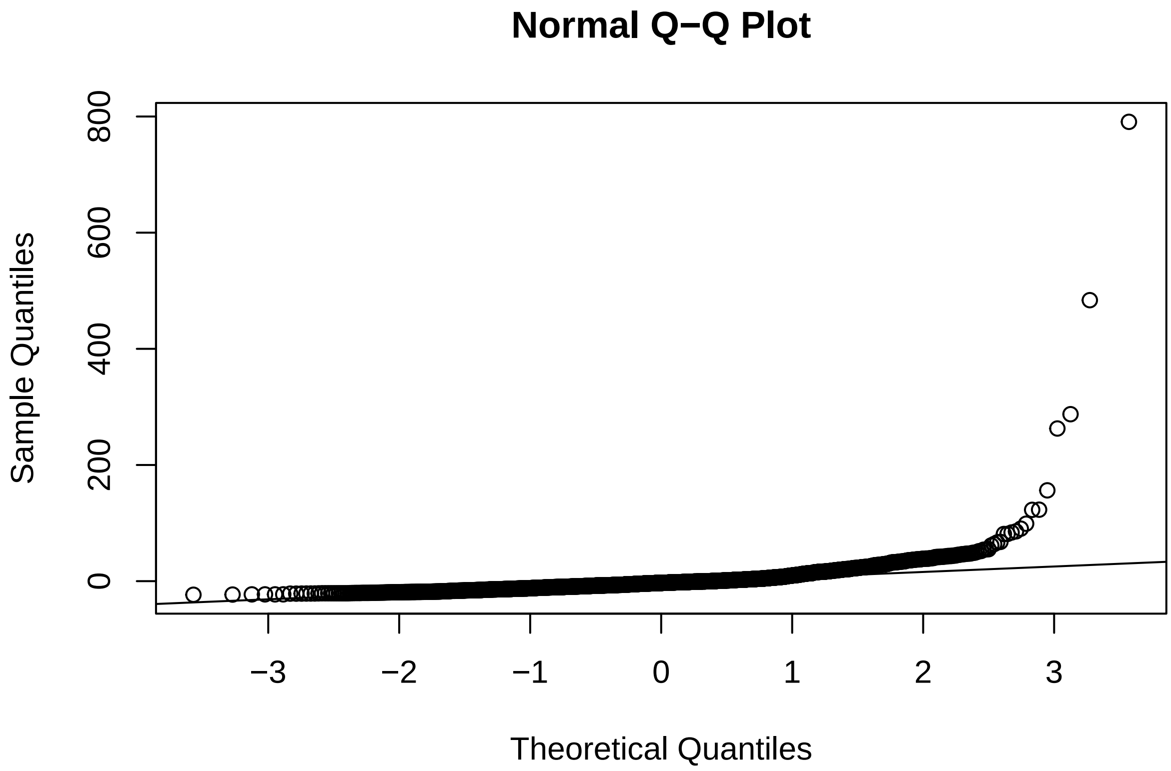
Initial regression

```
fit = lm(duration ~ adult + factor(item_type) + checkout_hour + dewey_exists,data=data)
summary(fit)
```

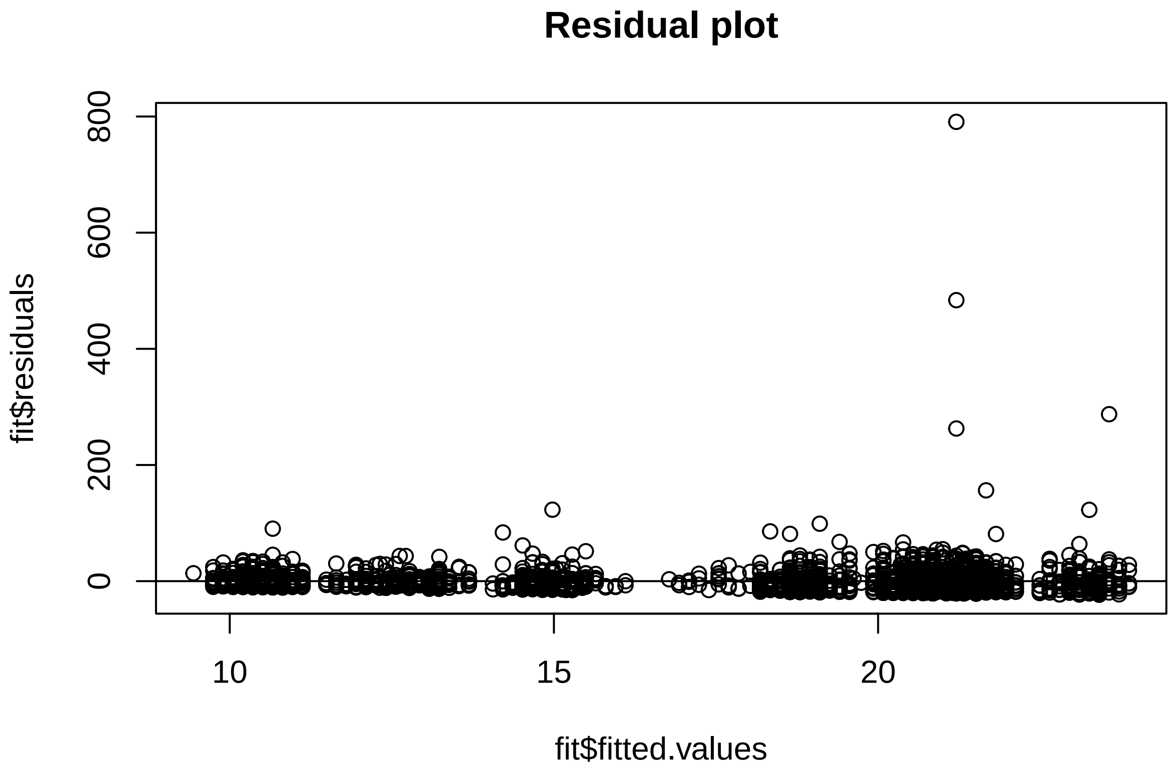
```
##
## Call:
## lm(formula = duration ~ adult + factor(item_type) + checkout_hour +
##     dewey_exists, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23.41  -9.36  -3.08   3.33 790.79
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              19.2186     3.1658   6.071 1.44e-09 ***
## adult                    -2.5652     1.0761  -2.384   0.0172 *
## factor(item_type)cd      -5.8676     1.4730  -3.983 6.96e-05 ***
## factor(item_type)dvd     -8.4369     1.1823  -7.136 1.22e-12 ***
## factor(item_type)other   -0.3418     5.4896  -0.062   0.9504
## checkout_hour             0.1529     0.2105   0.727   0.4675
## dewey_exists              1.7445     1.0720   1.627   0.1038
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 23.77 on 2806 degrees of freedom
## Multiple R-squared:  0.03666,    Adjusted R-squared:  0.0346
## F-statistic:  17.8 on 6 and 2806 DF,  p-value: < 2.2e-16
```

Diagnostics

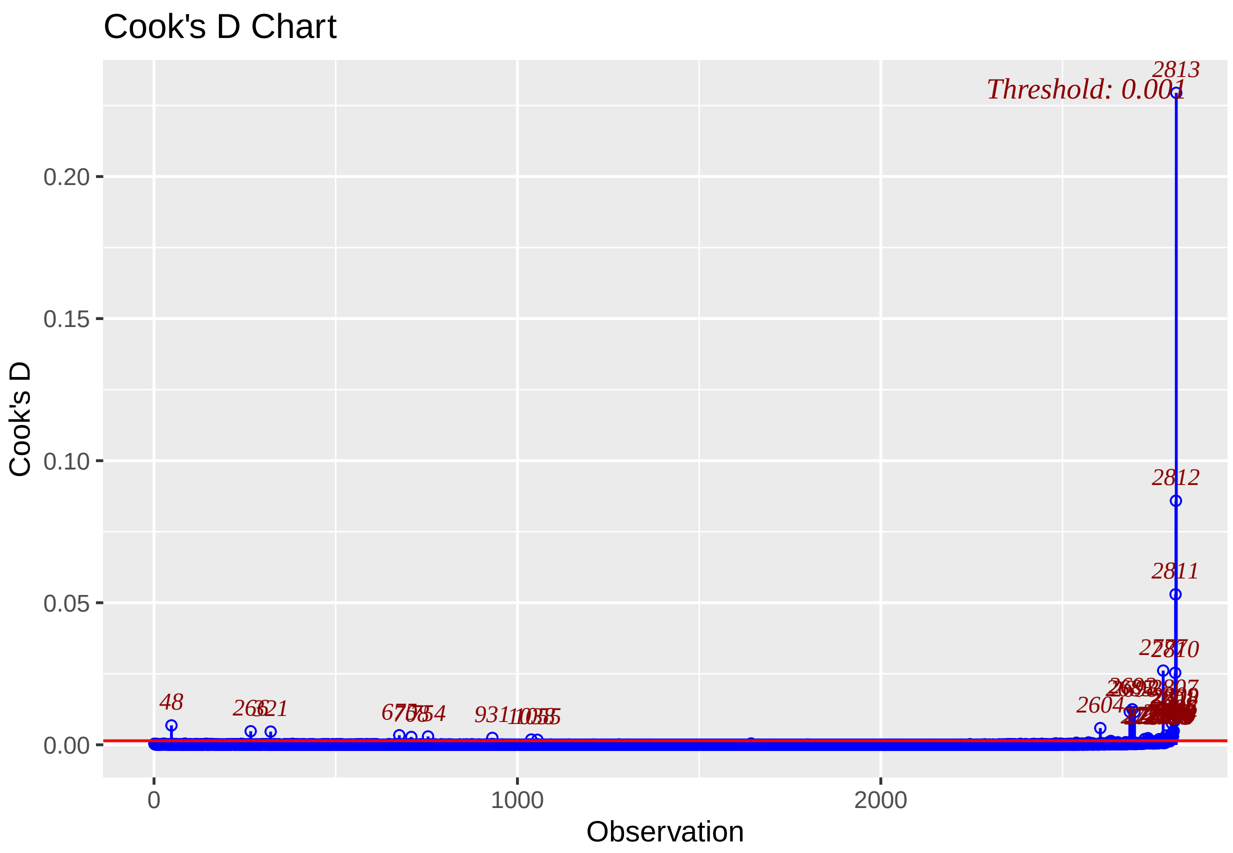
```
qqnorm(fit$residuals)
qqline(fit$residuals)
```



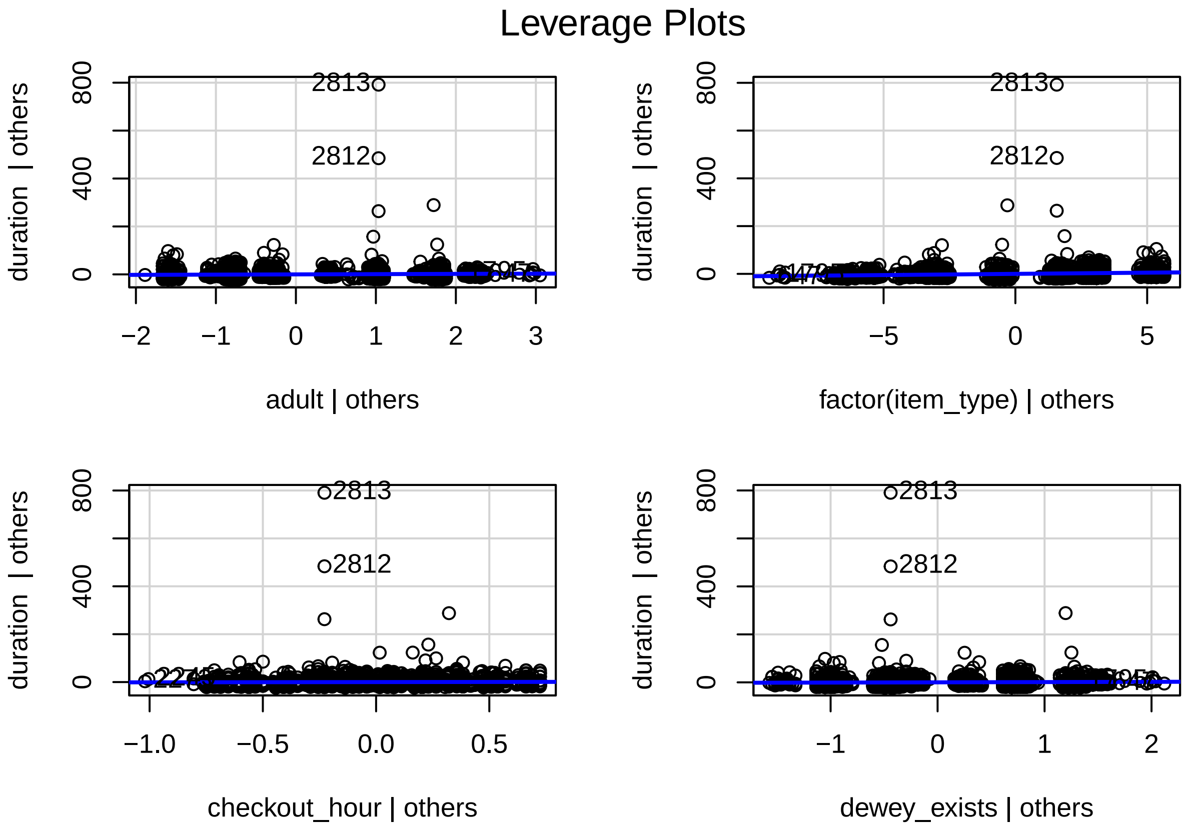
```
plot(fit$fitted.values,fit$residuals, main="Residual plot")
abline(0,0)
```



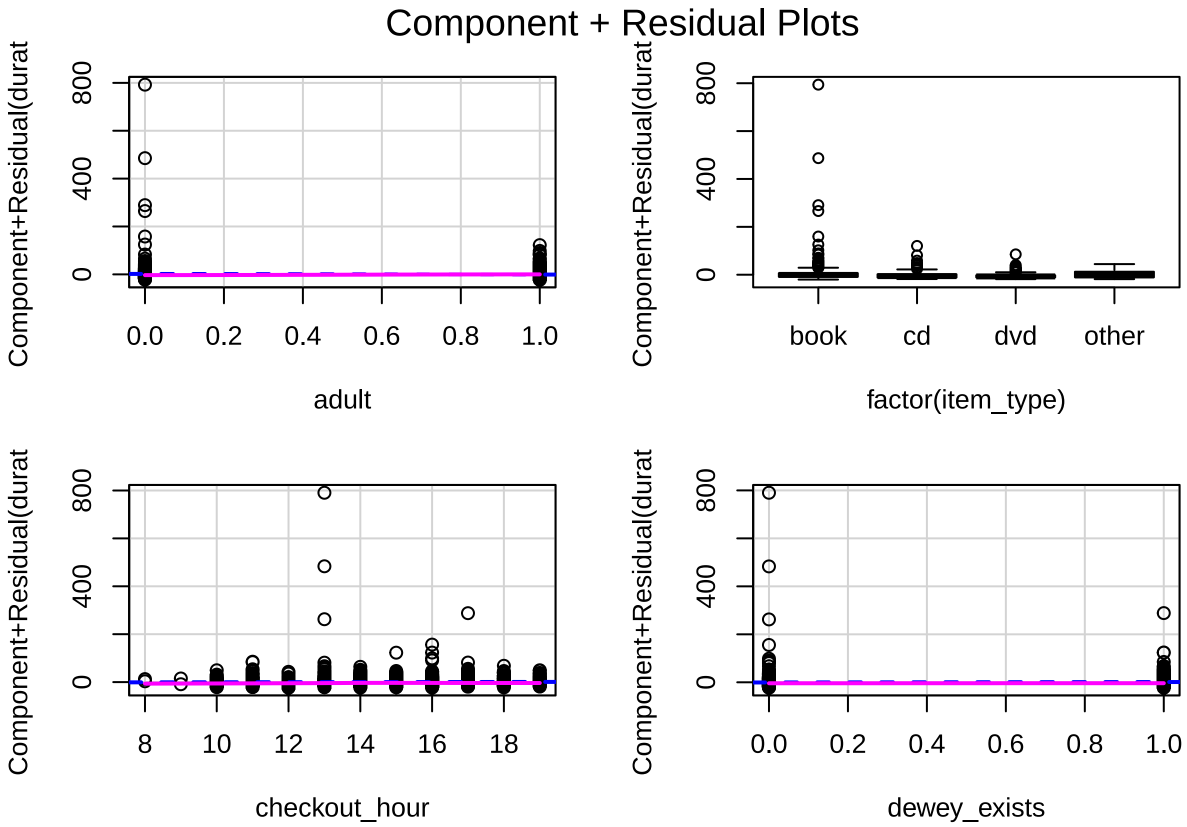
```
ols_plot_cooksd_chart(fit)
```



```
leveragePlots(fit)
```


Leverage Plots
duration | others
800
400
0
2813
2812
−2
−1
0
1
2
3
adult | others
duration | others
800
400
0
2813
2812
−5
0
5
factor(item_type) | others
duration | others
800
400
0
2813
2812
−1.0
−0.5
0.0
0.5
checkout_hour | others
duration | others
800
400
0
2813
2812
−1
0
1
2
dewey_exists | others

```
crPlots(fit)
```



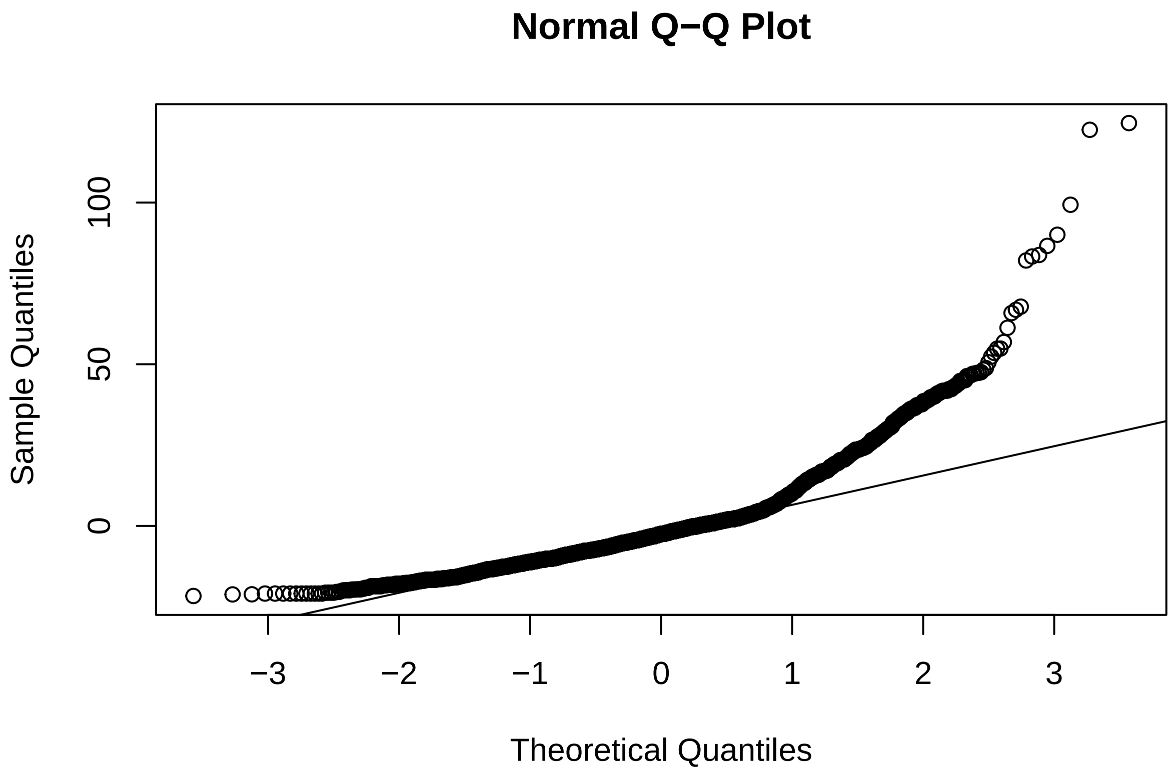
Dropping influential observations (outliers)

```
fitminus = lm(duration ~ adult + factor(item_type) + checkout_hour + dewey_exists,data=data,subset=-c(2
summary(fitminus)
```

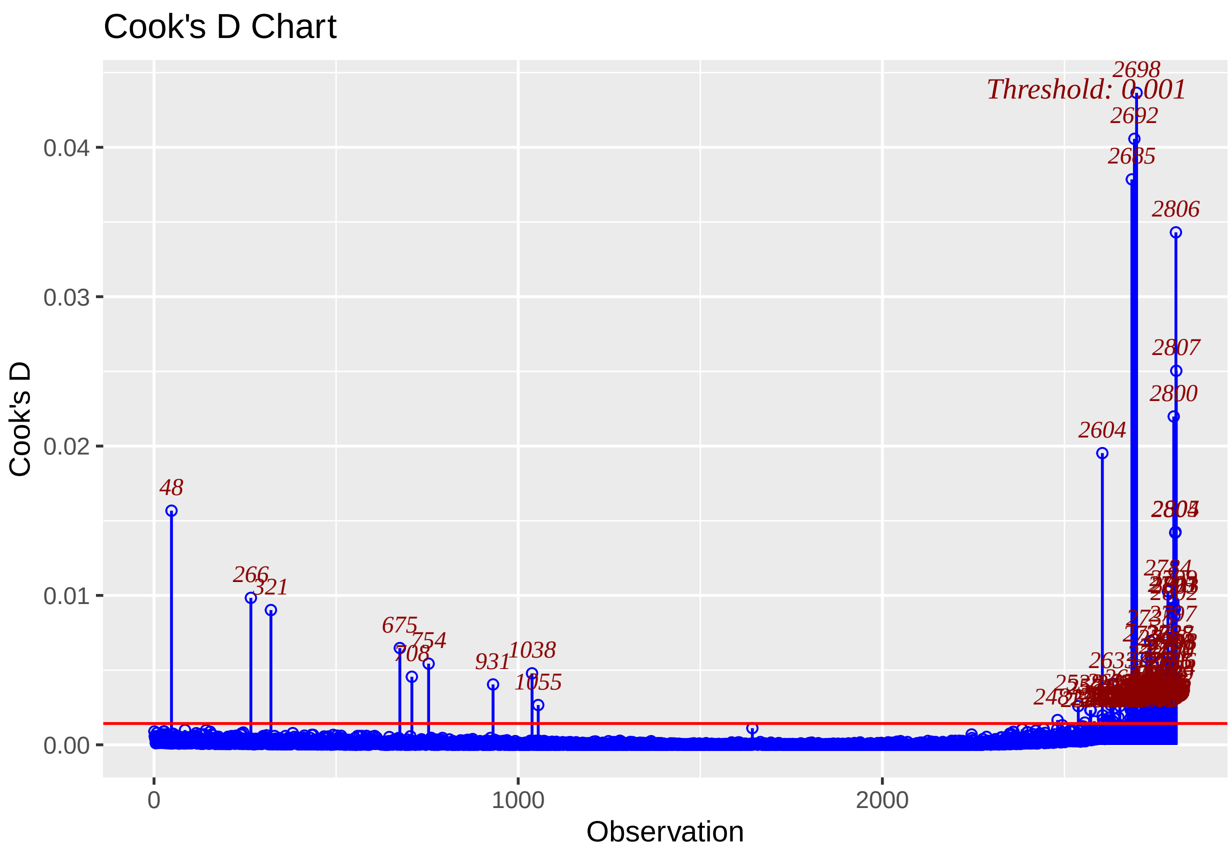
```
##
## Call:
## lm(formula = duration ~ adult + factor(item_type) + checkout_hour +
##     dewey_exists, data = data, subset = -c(2777, 2809, 2810,
##     2811, 2812, 2813))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.676  -8.663  -2.494   3.567 124.581
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              15.3192     1.8579   8.246 2.50e-16 ***
## adult                    -0.7472     0.6316  -1.183 0.236843
## factor(item_type)cd      -5.4342     0.8636  -6.293 3.61e-10 ***
## factor(item_type)dvd     -7.7318     0.6932 -11.153  < 2e-16 ***
## factor(item_type)other   -1.0652     3.3053  -0.322 0.747281
## checkout_hour             0.2566     0.1235   2.078 0.037804 *
## dewey_exists              2.2503     0.6290   3.578 0.000352 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 13.94 on 2800 degrees of freedom
## Multiple R-squared:  0.07798,    Adjusted R-squared:  0.076
## F-statistic: 39.47 on 6 and 2800 DF,  p-value: < 2.2e-16
```

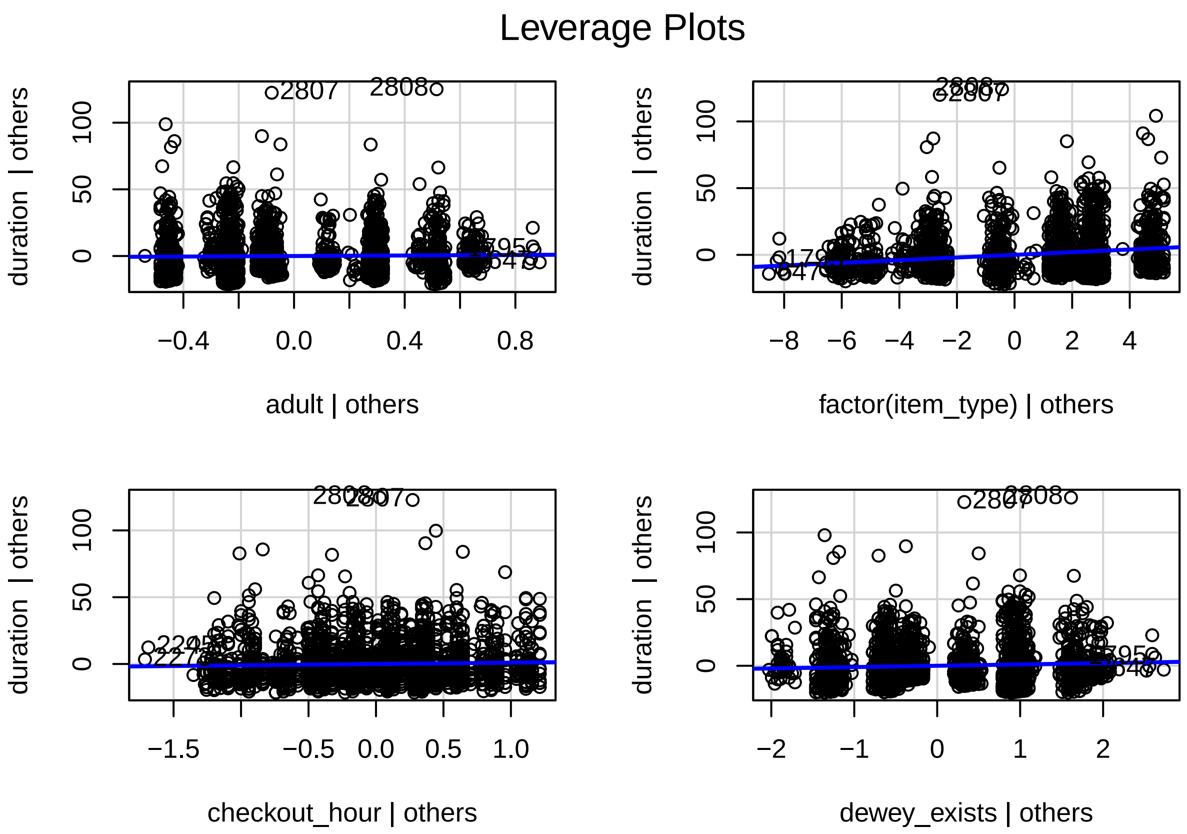
```
qqnorm(fitminus$residuals)
qqline(fitminus$residuals)
```



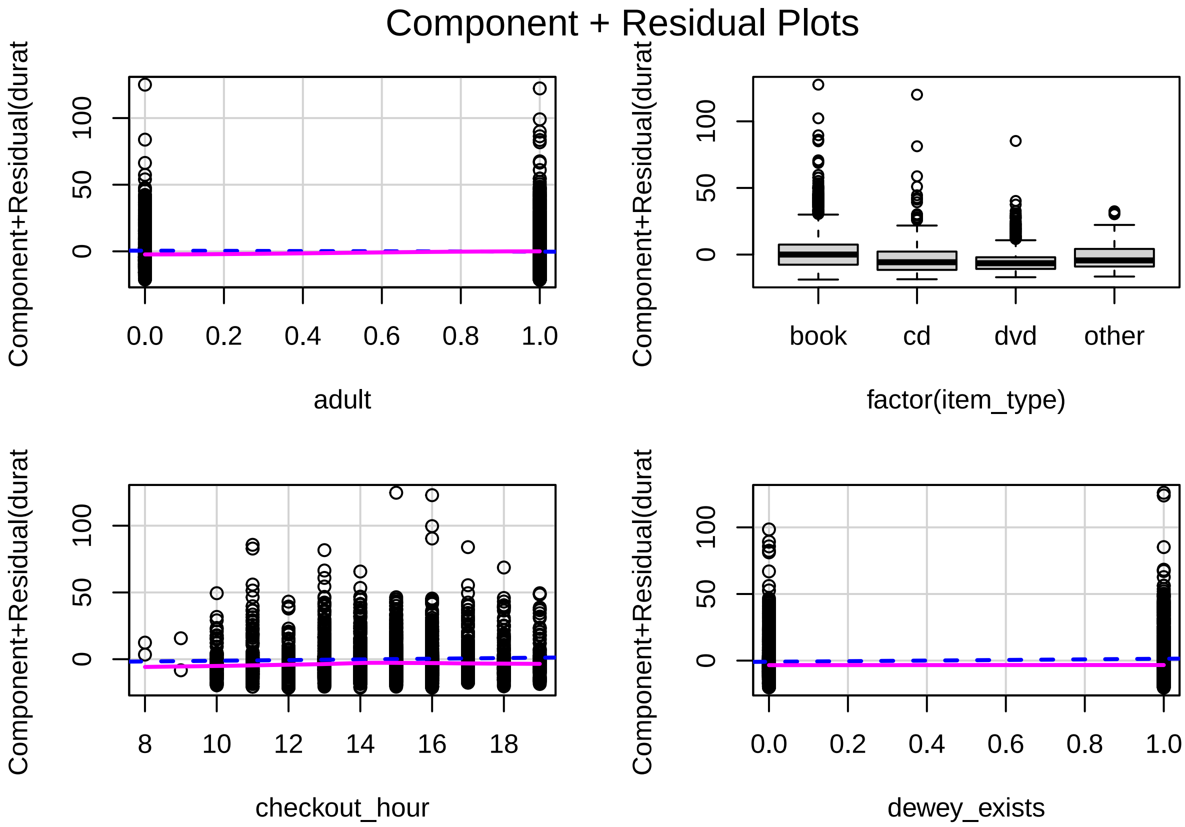
```
ols_plot_cooksd_chart(fitminus)
```



```
leveragePlots(fitminus)
```



```
crPlots(fitminus)
```

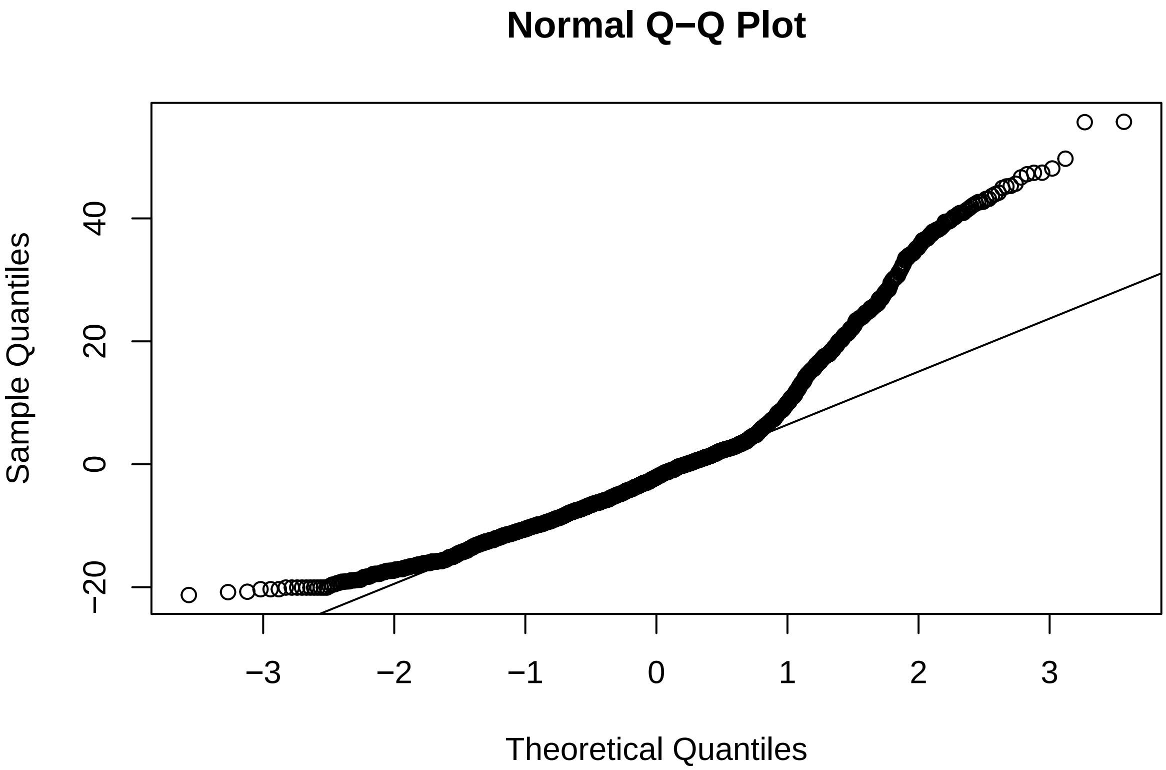


```
cooksd <- cooks.distance(fit)
sample_size <- nrow(data)
influential <- as.numeric(names(cooksd)[(cooksd > (4/sample_size))])
data_2 = data[-influential, ]
```

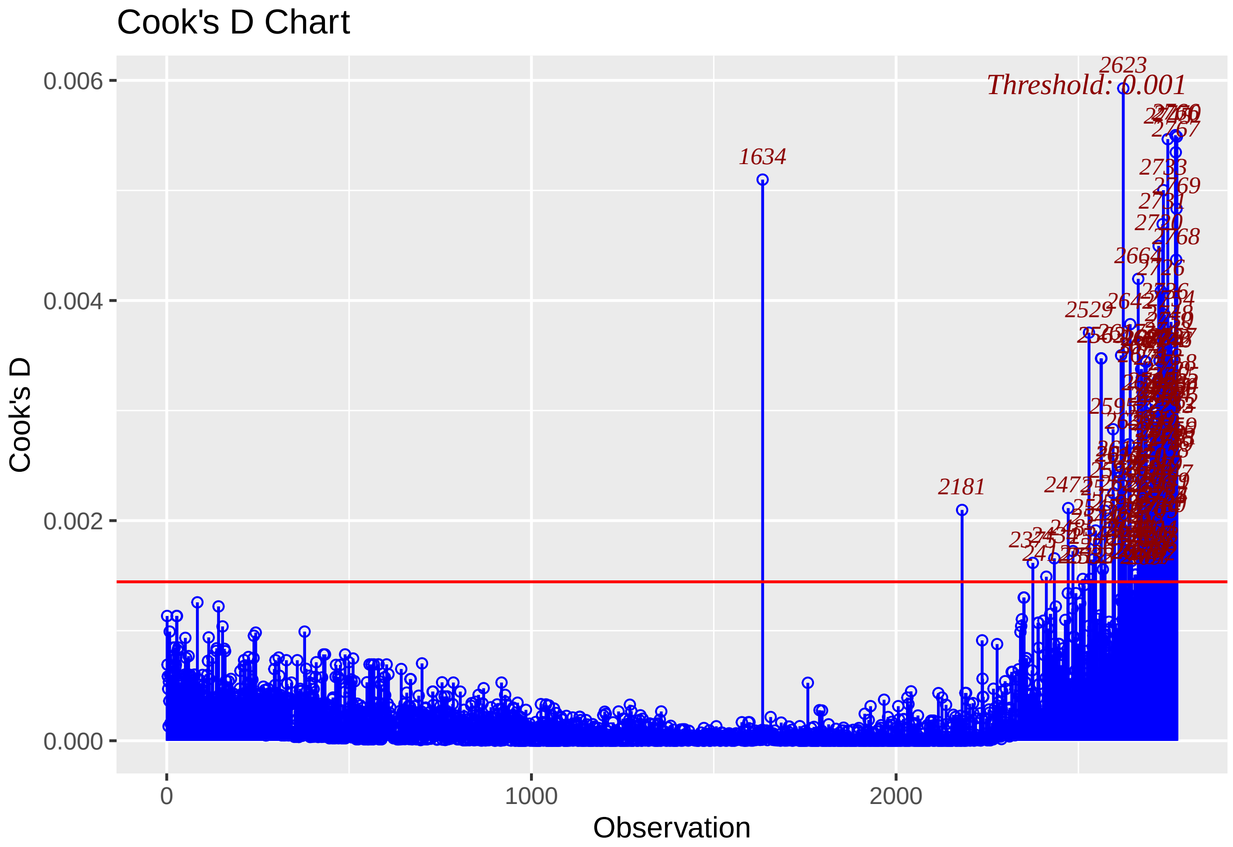
```
fit_2 = lm(duration ~ adult + factor(item_type) + checkout_hour + dewey_exists,data=data_2)
summary(fit)
```

```
##
## Call:
## lm(formula = duration ~ adult + factor(item_type) + checkout_hour +
##     dewey_exists, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23.41  -9.36  -3.08   3.33 790.79
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               19.2186     3.1658   6.071 1.44e-09 ***
## adult                     -2.5652     1.0761  -2.384   0.0172 *
## factor(item_type)cd       -5.8676     1.4730  -3.983 6.96e-05 ***
## factor(item_type)dvd      -8.4369     1.1823  -7.136 1.22e-12 ***
## factor(item_type)other    -0.3418     5.4896  -0.062   0.9504
## checkout_hour              0.1529     0.2105   0.727   0.4675
## dewey_exists             1.7445     1.0720   1.627   0.1038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.77 on 2806 degrees of freedom
## Multiple R-squared:  0.03666,    Adjusted R-squared:  0.0346
## F-statistic:  17.8 on 6 and 2806 DF,  p-value: < 2.2e-16
```

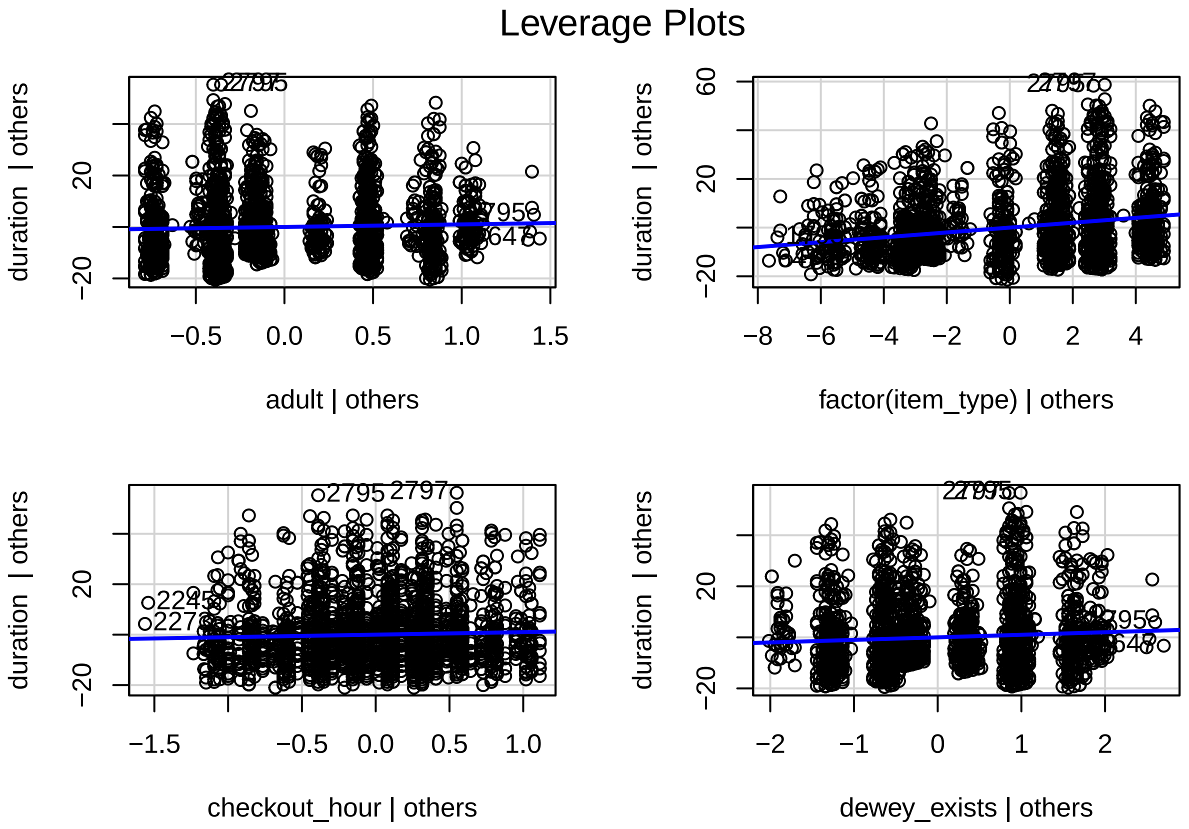
```
qqnorm(fit_2$residuals)
qqline(fit_2$residuals)
```



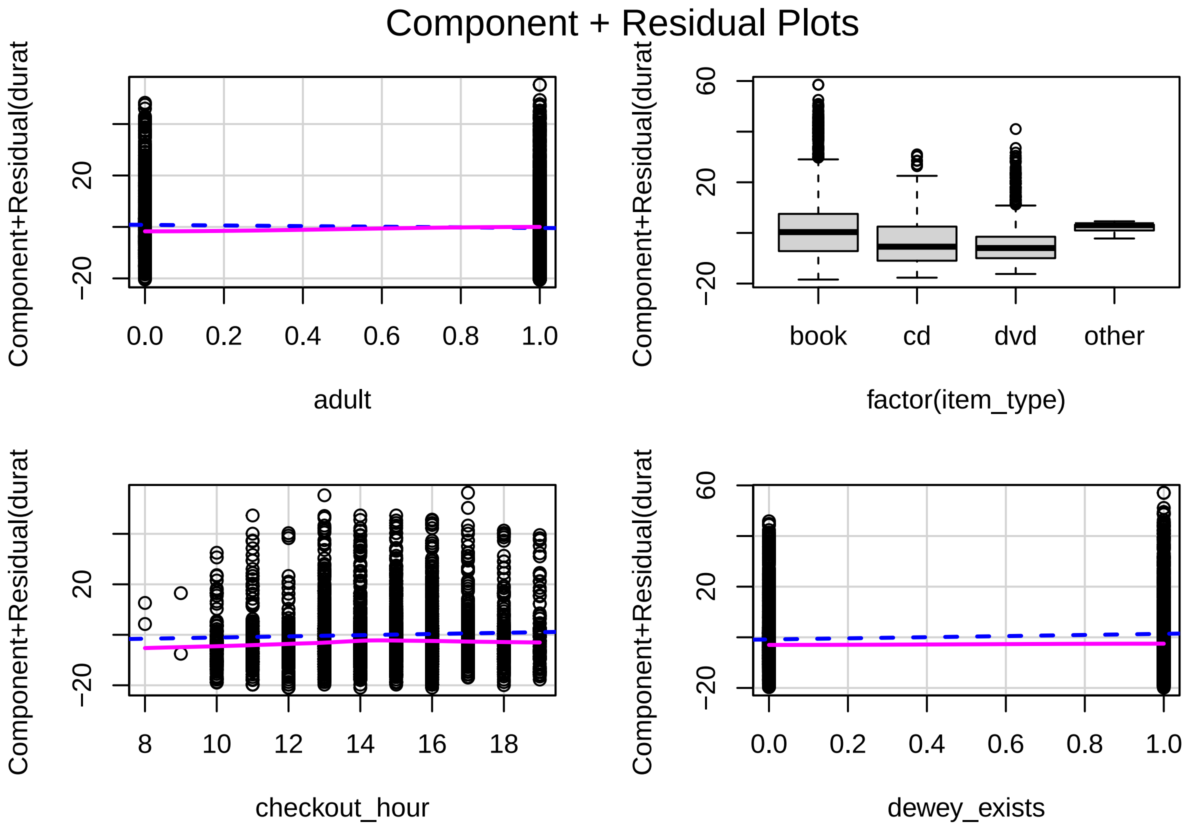
```
ols_plot_cooksd_chart(fit_2)
```



```
leveragePlots(fit_2)
```



```
crPlots(fit_2)
```


Component + Residual Plots
Component+Residual(durat
20
−20
0.0
0.2
0.4
0.6
0.8
1.0
adult
Component+Residual(durat
60
20
−20
book
cd
dvd
other
factor(item_type)
Component+Residual(durat
20
−20
8
10
12
14
16
18
checkout_hour
Component+Residual(durat
60
20
−20
0.0
0.2
0.4
0.6
0.8
1.0
dewey_exists