practice14

Qiang Liu

2022-11-29

```
> library(dplyr)
```

```
##
##    'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

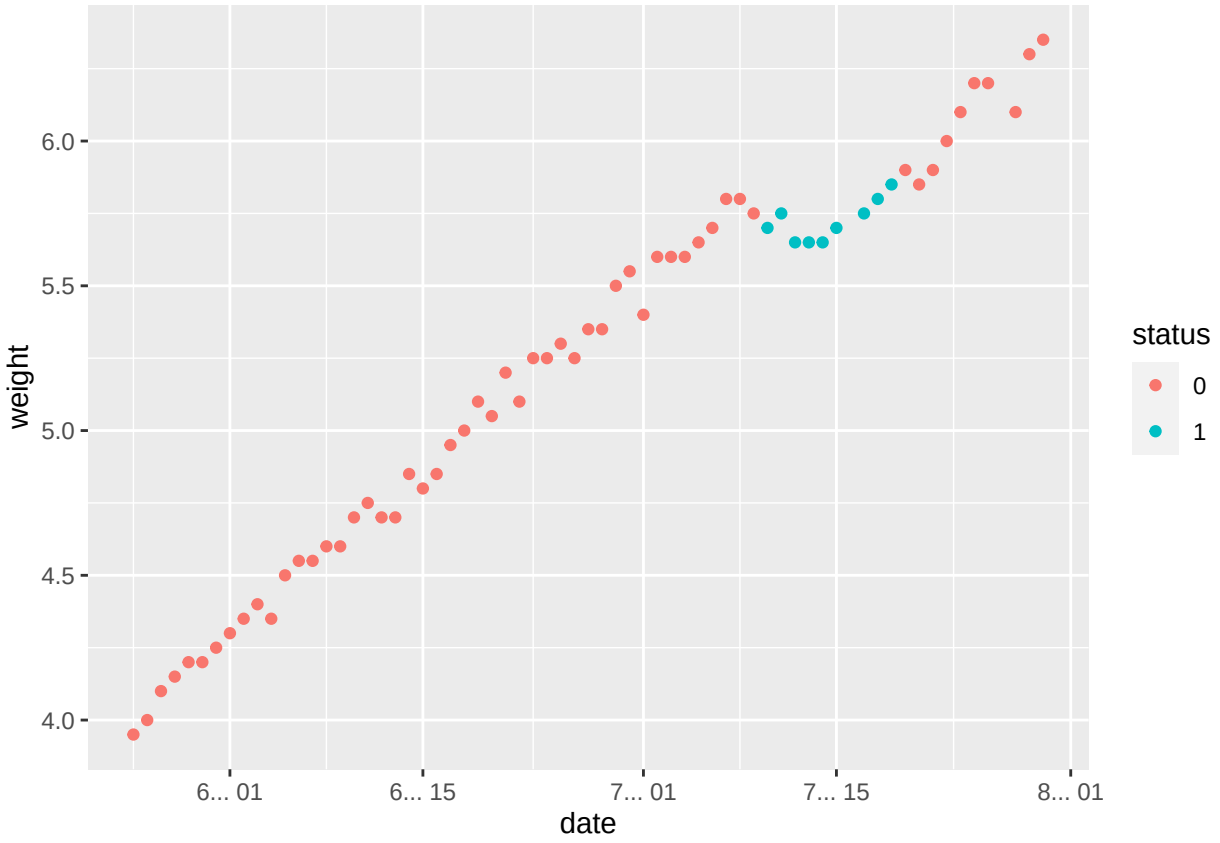
```
> my_data <- read.csv("weight.csv")
> my_data <- my_data %>% mutate(date=as.Date(date), status=as.factor(status))
> ## take a look at the data
> summary(my_data)
```

```
##       date                weight      status
##  Min.   :2020-05-25   Min.   :3.950   0:57
##  1st Qu.:2020-06-10   1st Qu.:4.700   1:10
##  Median :2020-06-27   Median :5.300
##  Mean   :2020-06-27   Mean   :5.214
##  3rd Qu.:2020-07-13   3rd Qu.:5.750
##  Max.   :2020-07-30   Max.   :6.350
##                       NA's   :2
```

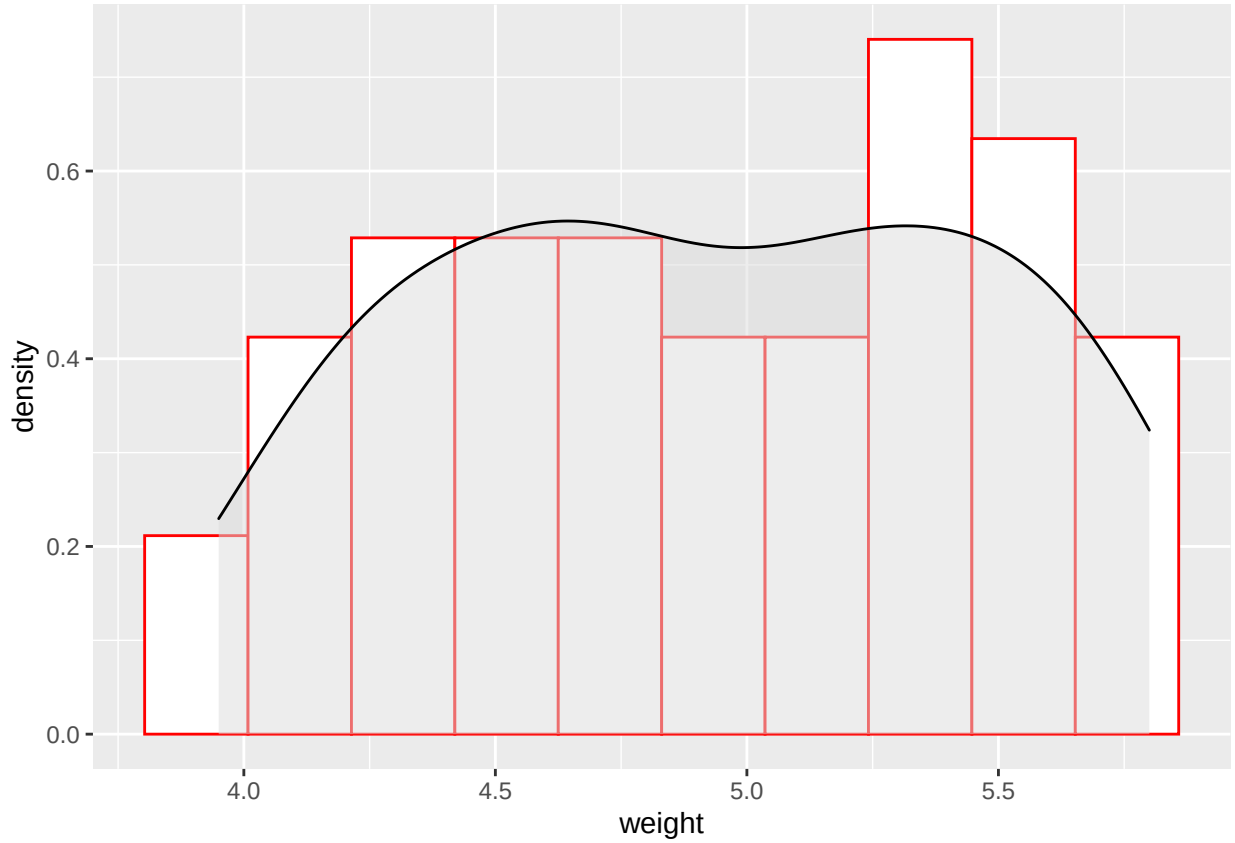
```
> head(my_data)
```

```
##         date weight status
## 1 2020-05-25   3.95      0
## 2 2020-05-26   4.00      0
## 3 2020-05-27   4.10      0
## 4 2020-05-28   4.15      0
## 5 2020-05-29   4.20      0
## 6 2020-05-30   4.20      0
```

```
> ## visual plot
> library(ggplot2)
> library(dplyr)
> my_data %>% ggplot(aes(date, weight, color=status)) + geom_point()
```




```
> ## subset data
> sub_data <- my_data %>% filter(date < date[min(which(status == 1))])
> ## hist and density plot
> ggplot(sub_data, aes(x=weight)) +
+   geom_histogram(aes(y=..density..),
+                  bins = 10, colour="red", fill="white")+
+   geom_density(alpha=.5, fill="#dddddd")
```



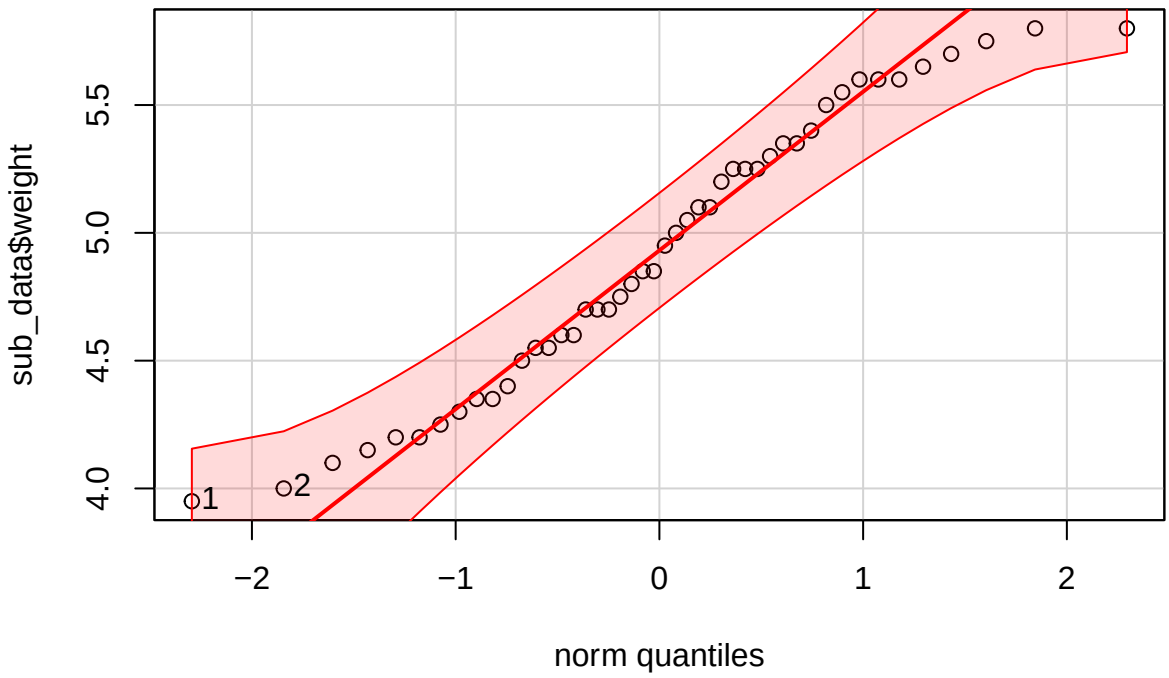

```
> ## qq-plot
> library("car")
```

```
##      carData

##
##     'car'

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

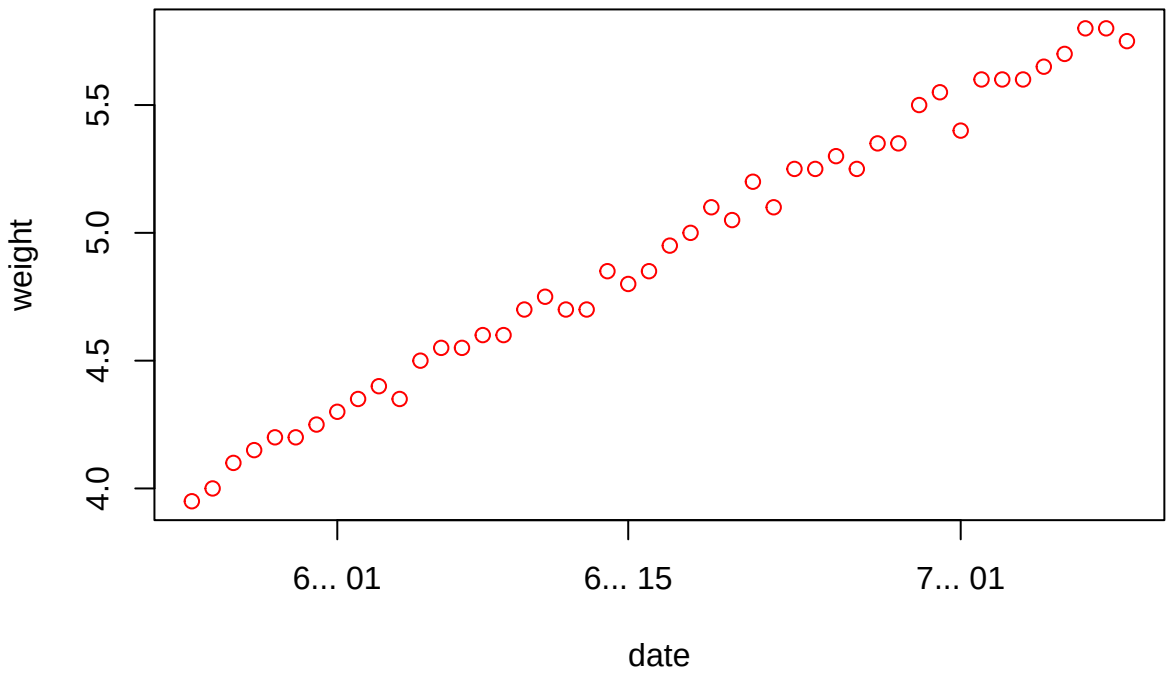
```
> out <- qqPlot(sub_data$weight, col.lines = 'red')
```



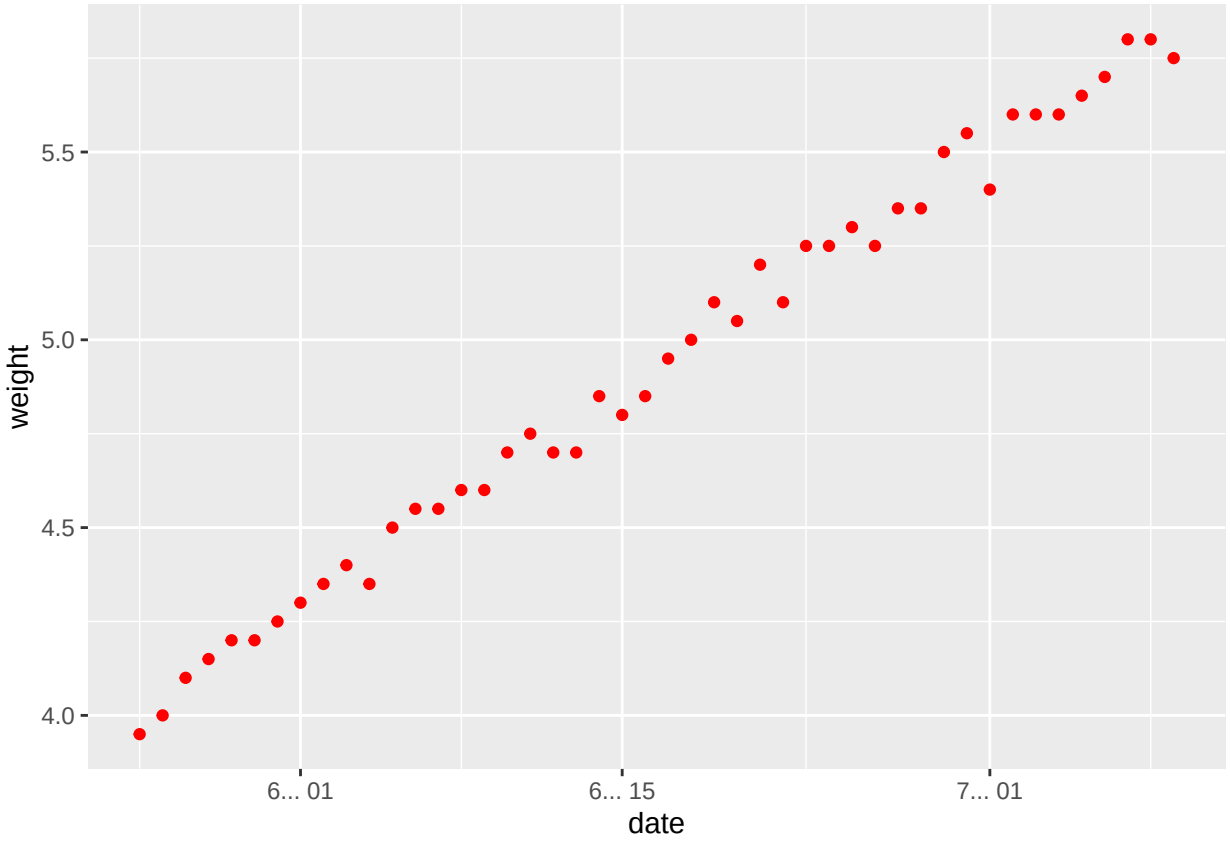

```
> ## shapiro.test()
> shapiro.test(sub_data$weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sub_data$weight
## W = 0.95491, p-value = 0.07265
```

```
> plot(weight ~ date, col="red", data = sub_data)
```



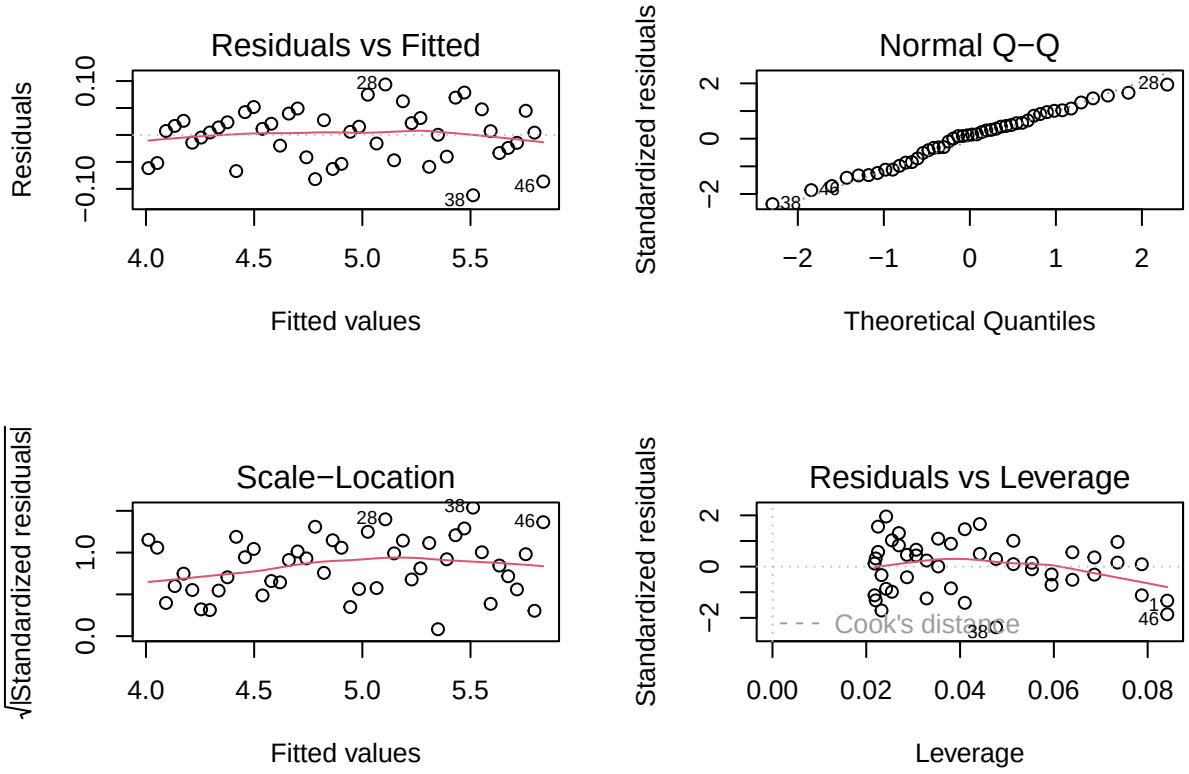

```
> sub_data %>% ggplot(aes(date, weight)) + geom_point(color="red")
```



```
> ## linear regression model
> growth.lm <- lm(weight ~ date, data = sub_data)
> summary(growth.lm)
```

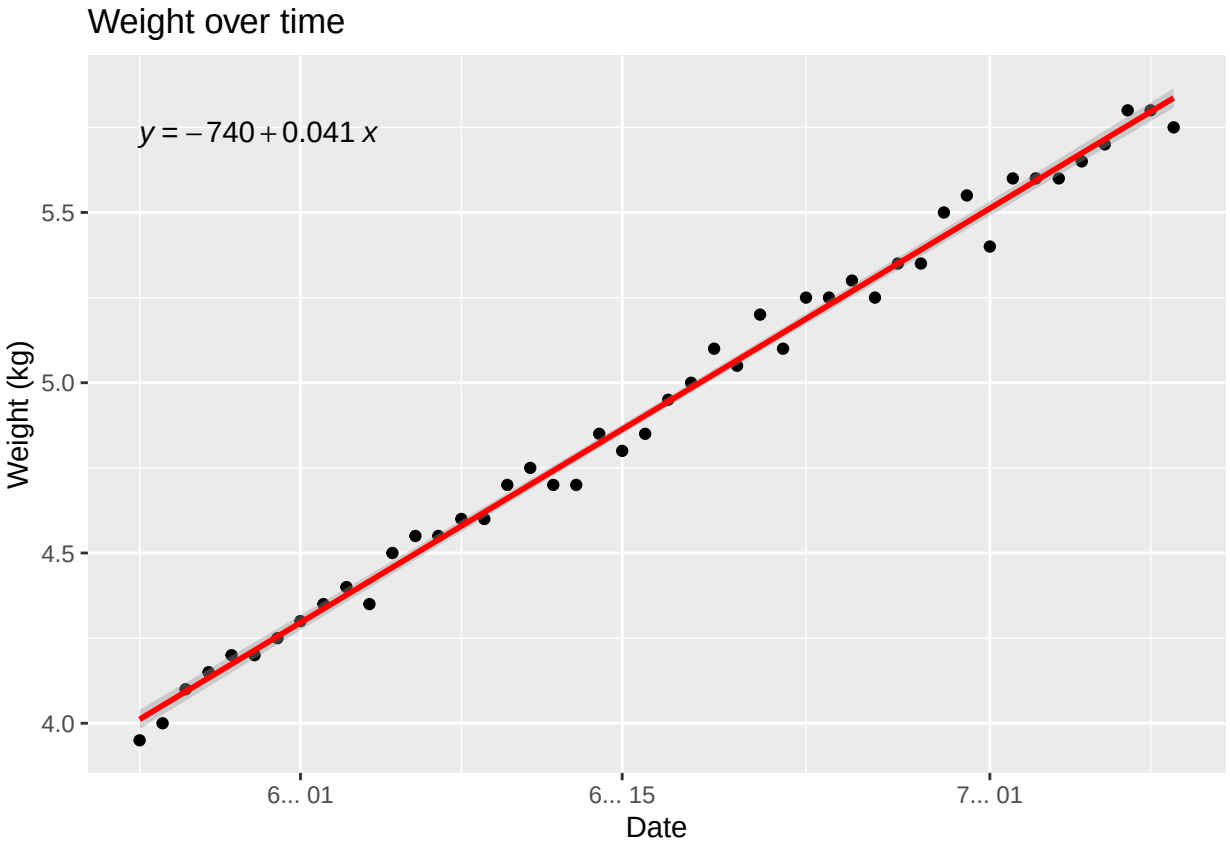
```
##
## Call:
## lm(formula = weight ~ date, data = sub_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.111872 -0.038548  0.006422  0.030431  0.093617
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.424e+02  9.915e+00  -74.87   <2e-16 ***
## date         4.055e-02  5.380e-04   75.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04844 on 44 degrees of freedom
## Multiple R-squared:  0.9923, Adjusted R-squared:  0.9921
## F-statistic:  5681 on 1 and 44 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(growth.lm)
```

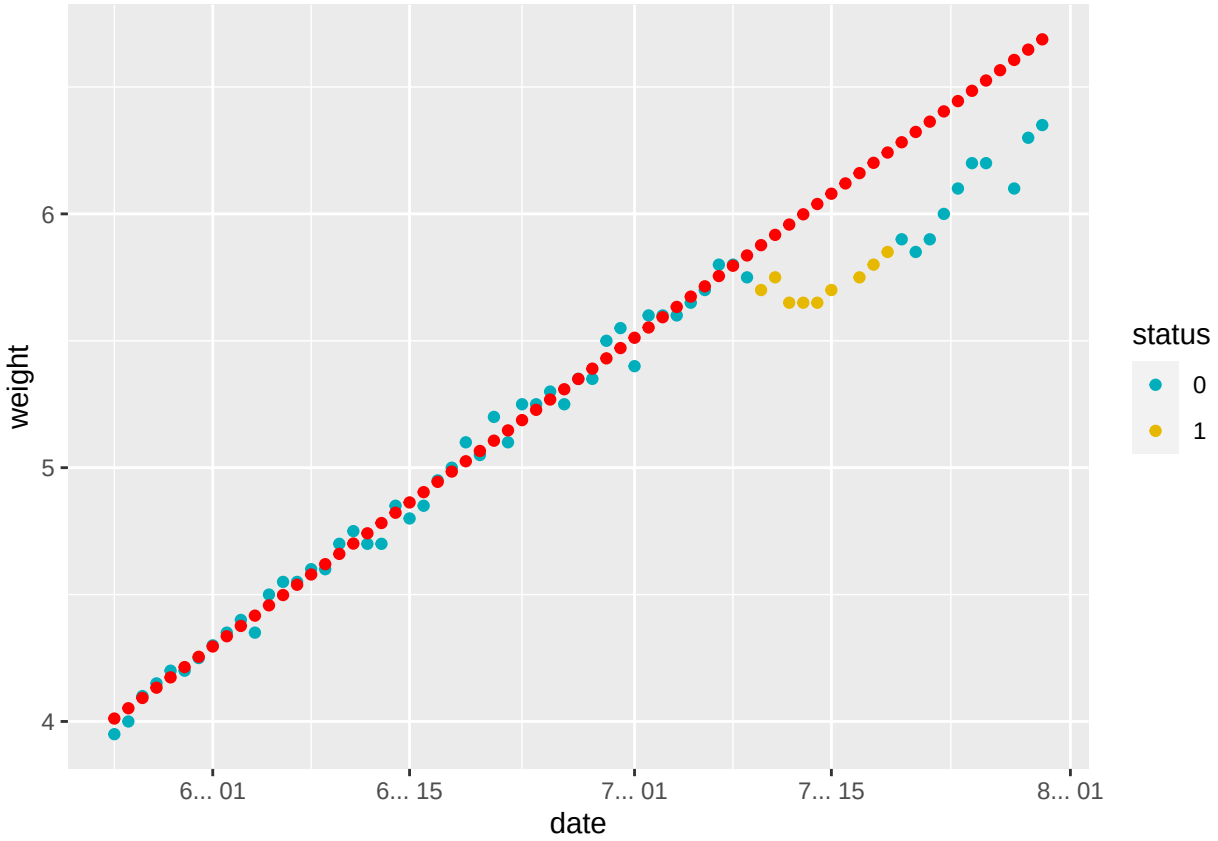



```
> par(mfrow=c(1,1))
> library(ggpubr)
> sub_data %>% ggplot(aes(date, weight)) + geom_point() +
+   geom_smooth(method="lm", col="red") +
+   stat_regline_equation() +
+   labs(title = "Weight over time",
+        x = "Date",
+        y = "Weight (kg)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



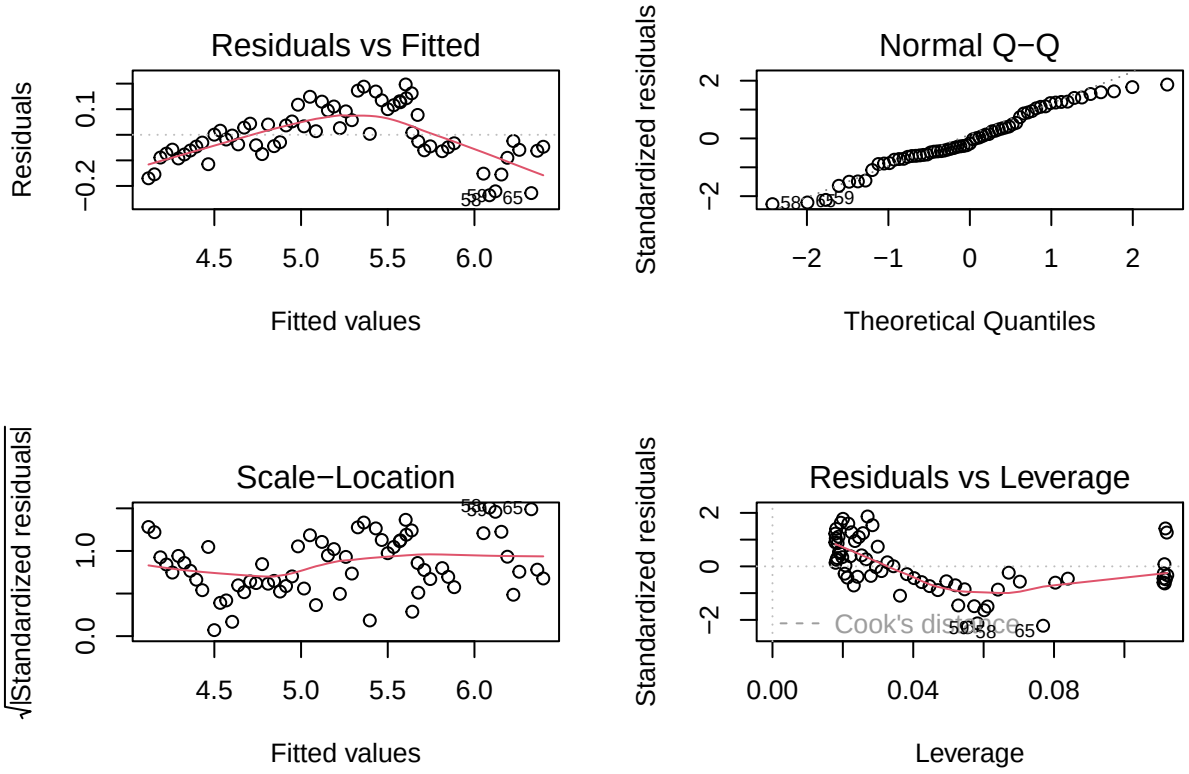

```
> my_data$predict1 <- predict(growth.lm, my_data)
> my_data %>% ggplot(aes(date, weight, color=status)) + geom_point() +
+   geom_point(aes(date, predict1), color="red") +
+   scale_color_manual(values = c("#00AFBB", "#E7B800"))
```



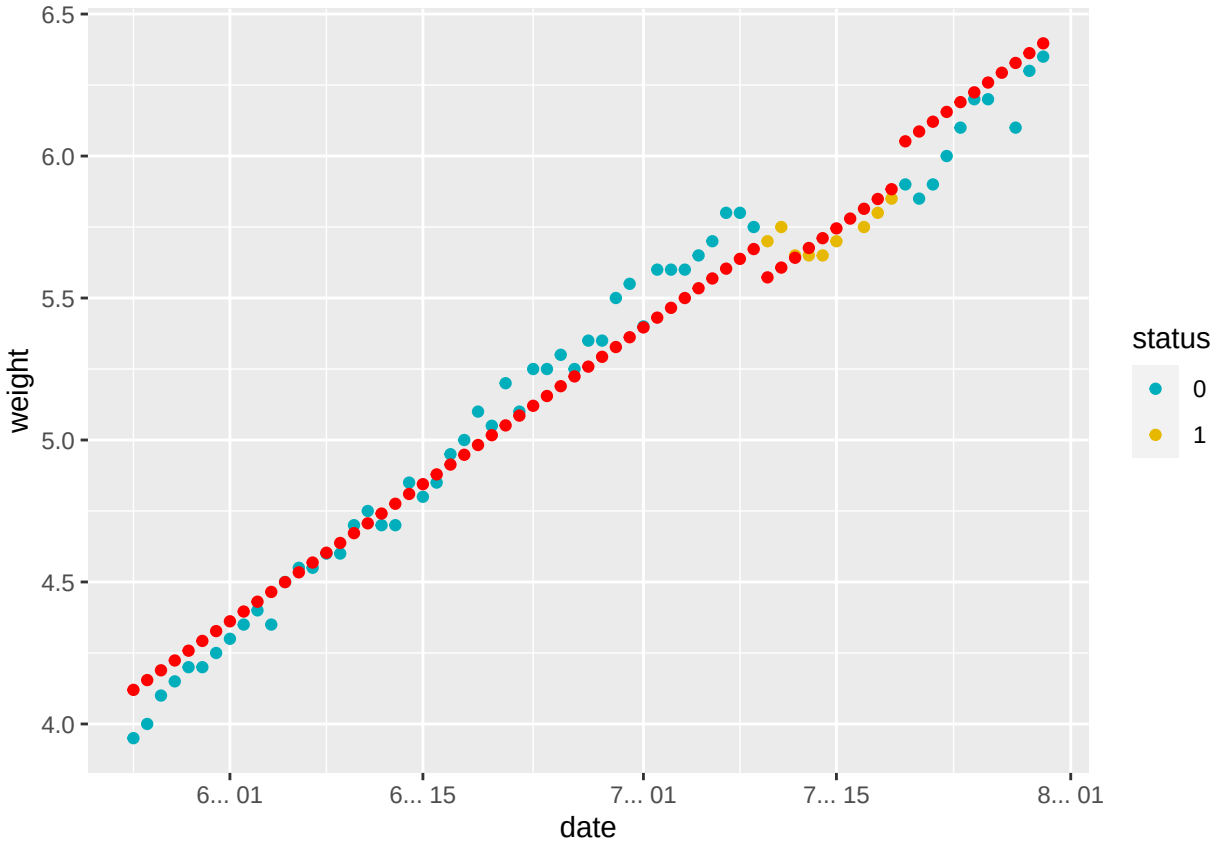

```
> growth.lm2 <- lm(weight ~ date + status, data = my_data)
> summary(growth.lm2)
```

```
##
## Call:
## lm(formula = weight ~ date + status, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23638 -0.06149 -0.01847  0.09157  0.19659
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.309e+02  1.381e+01 -45.669  < 2e-16 ***
## date         3.450e-02  7.493e-04  46.040  < 2e-16 ***
## status1     -1.342e-01  4.148e-02  -3.235  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1069 on 62 degrees of freedom
##   (   2    )
## Multiple R-squared:  0.9744, Adjusted R-squared:  0.9735
## F-statistic:  1178 on 2 and 62 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(growth.lm2)
```




```
> par(mfrow=c(1,1))
> my_data$predict2 <- predict(growth.lm2, my_data)
> my_data %>% ggplot(aes(date, weight, color=status)) + geom_point() +
+   geom_point(aes(date, predict2), color="red") +
+   scale_color_manual(values = c("#00AFBB", "#E7B800"))
```



```
> model1 <- lm(weight ~ date, data = my_data)
> model1.sum <- summary(model1)
> model1.sum$r.squared
```

```
## [1] 0.970033
```

```
> model2 <- lm(weight ~ date + status, data = my_data)
> model2.sum <- summary(model2)
> model2.sum$r.squared
```

```
## [1] 0.9743601
```