SF_JUL

2022-06-15

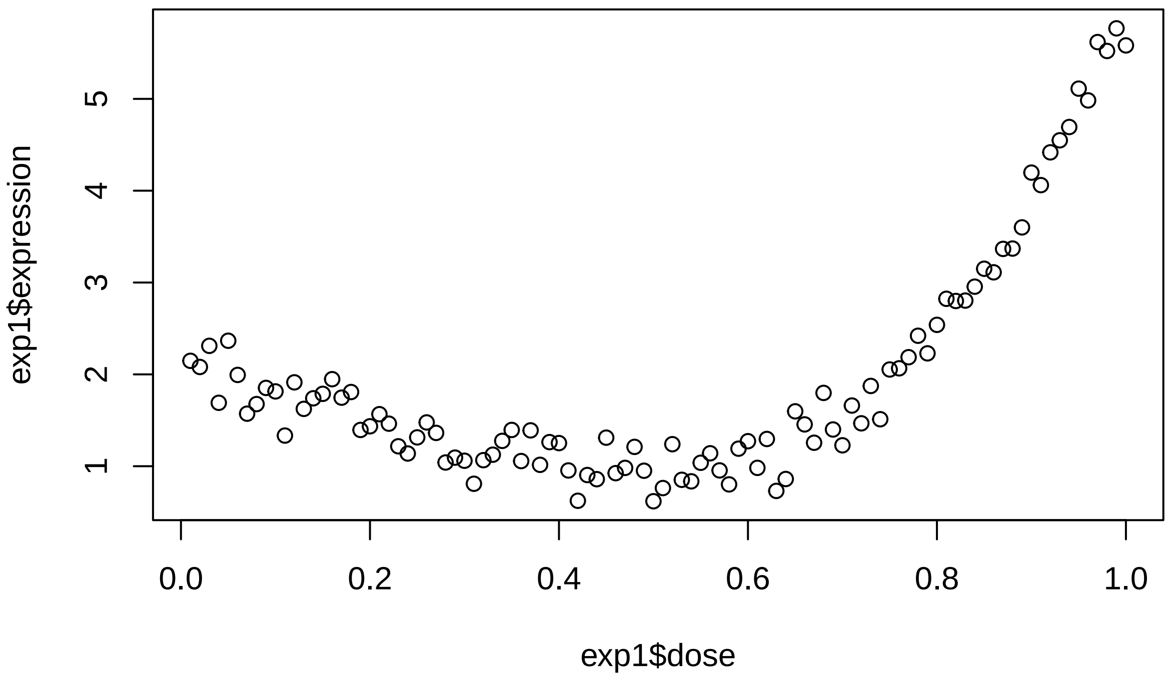
The data frame "exp1" contains 100 measurements of the expression of a gene at 100 doses of a drug.

```
load("C:/Users/seren/Desktop/test_2107.RData")
head(exp1)
```

```
##   dose expression
## 1 0.01   2.147618
## 2 0.02   2.080570
## 3 0.03   2.310649
## 4 0.04   1.690693
## 5 0.05   2.366353
## 6 0.06   1.993866
```

1. Plot the dependence of the expression on the dose.

```
plot(exp1$expression ~ exp1$dose)
```



2. Using linear regression, determine the polynomial regression function that best describes the dependence of the expression on the dose. Plot the data together with your best regression function.

```
lm_1 <- lm(exp1$expression ~ exp1$dose) #superflual, it was said polynomial

lm_2 <- lm(exp1$expression ~ poly(exp1$dose, degree = 2, raw = TRUE))
lm_3 <- lm(exp1$expression ~ poly(exp1$dose, degree = 3, raw = TRUE))
lm_4 <- lm(exp1$expression ~ poly(exp1$dose, degree = 4, raw = TRUE))
```

ANOVA test between models will tell which one is the best. Don't look at R^2 because obviously will always increase by adding new terms.

```
an_1 <- anova(lm_3, lm_2)
an_1
```

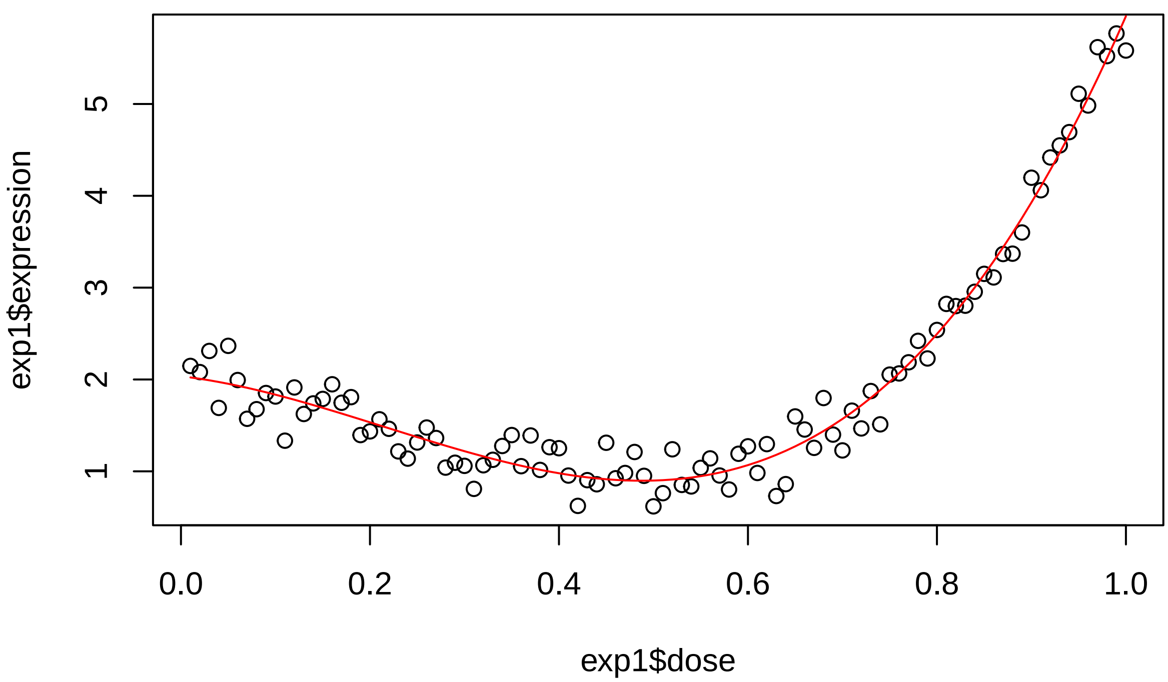
```
## Analysis of Variance Table
##
## Model 1: exp1$expression ~ poly(exp1$dose, degree = 3, raw = TRUE)
## Model 2: exp1$expression ~ poly(exp1$dose, degree = 2, raw = TRUE)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     96  4.1528
## 2     97 11.6386 -1   -7.4858 173.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
an_2 <- anova(lm_4, lm_3)
an_2
```

```
## Analysis of Variance Table
##
## Model 1: exp1$expression ~ poly(exp1$dose, degree = 4, raw = TRUE)
## Model 2: exp1$expression ~ poly(exp1$dose, degree = 3, raw = TRUE)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     95 4.0853
## 2     96 4.1528 -1 -0.067579 1.5715 0.2131
```

lm_3 is the best model because of its smaller P-value. Let's plot it together with data:

```r
plot(exp1$expression ~ exp1$dose)
lines(exp1$dose, lm_3$fitted.values, col = "red")
```



3. The experiment was then repeated with identical conditions, and the results are in the data frame "exp2". Can you use these data to show that - **using a polynomial of lower degree than your best model determined above leads to decreased accuracy as measured in the new experimental data? - using a polynomial of higher degree than your best model determined above does not improve the accuracy in the new experimental data?**

Let's treat "exp2" as a testing dataset:

```r
express_exp2_lm2 <- predict(lm_2, newdata = exp2)
express_exp2_lm4 <- predict(lm_4, newdata = exp2)

r2 <- function(y, y_pred) {
  1 - sum((y - y_pred)^2)/sum((y - mean(y))^2)
}

r2_lm3 <- r2(exp1$expression, lm_3$fitted.values)
#equal to calling R^2 as "summary(lm_3)$r.squared" instead
r2_lm3
```

```
## [1] 0.9733747
```

```r
r2_lm2 <- r2(exp1$expression, lm_2$fitted.values)
r2_lm4 <- r2(exp1$expression, lm_4$fitted.values)

R2 <- c(r2_lm2,  r2_lm3, r2_lm4)
names(R2) <- c("lm_2", "lm_3", "lm_4")

R2[order(R2)]
```

```
##      lm_2      lm_3      lm_4
## 0.9253806 0.9733747 0.9738079
```