Homework1

Cheng-Han Yu

July 17, 2017

Load the data

```
data <- read.csv("../Data/priemelDataReconstruction.csv", header = TRUE)
```

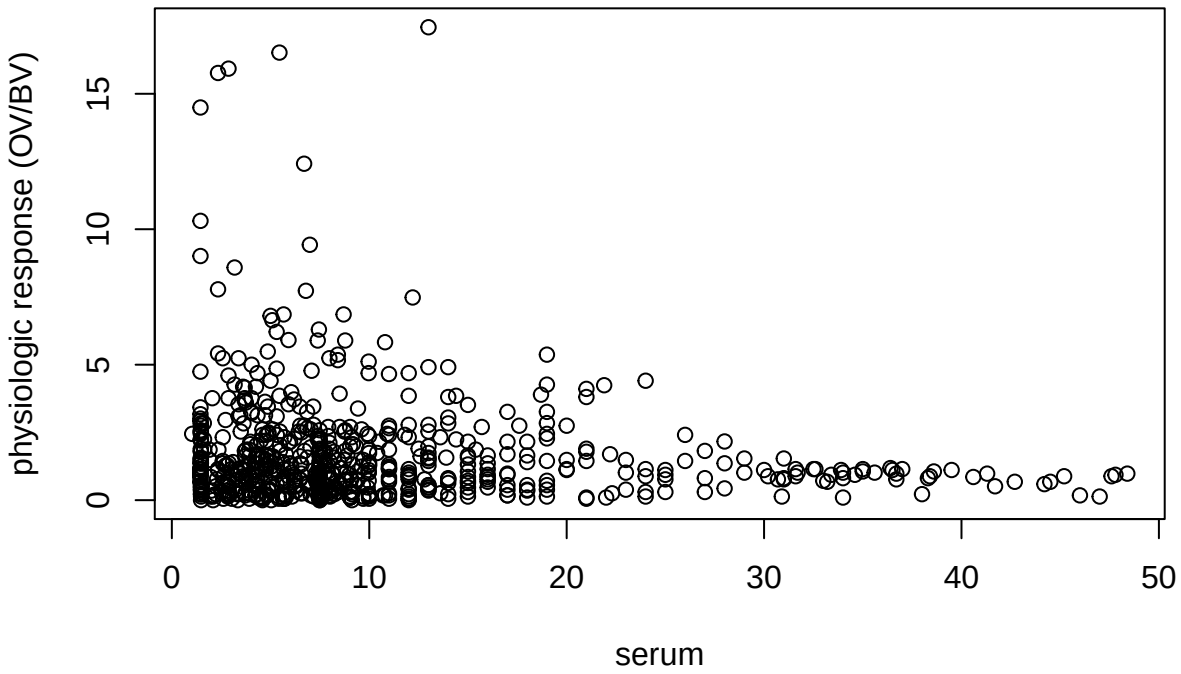
Show the data

```
head(data)
```

```
##   SerumLevelInNgPerMl OV.BV
## 1                7.49     0
## 2               12.00     0
## 3                7.49     0
## 4                2.09     0
## 5                9.12     0
## 6                4.59     0
```

Plot the physiologic response (OV/BV) as a function of serum level.

```
plot(data$SerumLevelInNgPerMl, data$OV.BV,
     xlab = "serum", ylab = "physiologic response (OV/BV)")
```



Describe the trend

The higher the serum level is, the lower the physiologic response is. It seems that the physiologic response and serum level are negatively correlated.

Logistic Regression

```
# create a binary variable
data$OV.BV.bin <- 0
data$OV.BV.bin[data$OV.BV > 2] <- 1
# View(data)
out <- glm(OV.BV.bin ~ SerumLevelInNgPerMl,
           family = binomial(link = "logit"), data = data)
summary(out)
```

```
##
## Call:
## glm(formula = OV.BV.bin ~ SerumLevelInNgPerMl, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8963  -0.8260  -0.7576   1.4874   2.0360
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.63719    0.13751  -4.634 3.59e-06 ***
## SerumLevelInNgPerMl -0.04646    0.01257  -3.696 0.000219 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 768.35  on 674  degrees of freedom
## Residual deviance: 751.53  on 673  degrees of freedom
## AIC: 755.53
##
## Number of Fisher Scoring iterations: 4
```

```
data2 <- data[data$SerumLevelInNgPerMl > 10, ]
out2 <- glm(OV.BV.bin ~ SerumLevelInNgPerMl,
           family = binomial(link = "logit"), data = data2)
summary(out2)
```

```
##
## Call:
## glm(formula = OV.BV.bin ~ SerumLevelInNgPerMl, family = binomial(link = "logit"),
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9827  -0.8103  -0.5375  -0.1977   2.1759
##
## Coefficients:
```

```
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.58182    0.49743   1.170 0.242138
## SerumLevelInNgPerMl -0.10181    0.02837  -3.588 0.000333 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 217.25  on 207  degrees of freedom
## Residual deviance: 197.88  on 206  degrees of freedom
## AIC: 201.88
##
## Number of Fisher Scoring iterations: 5
```

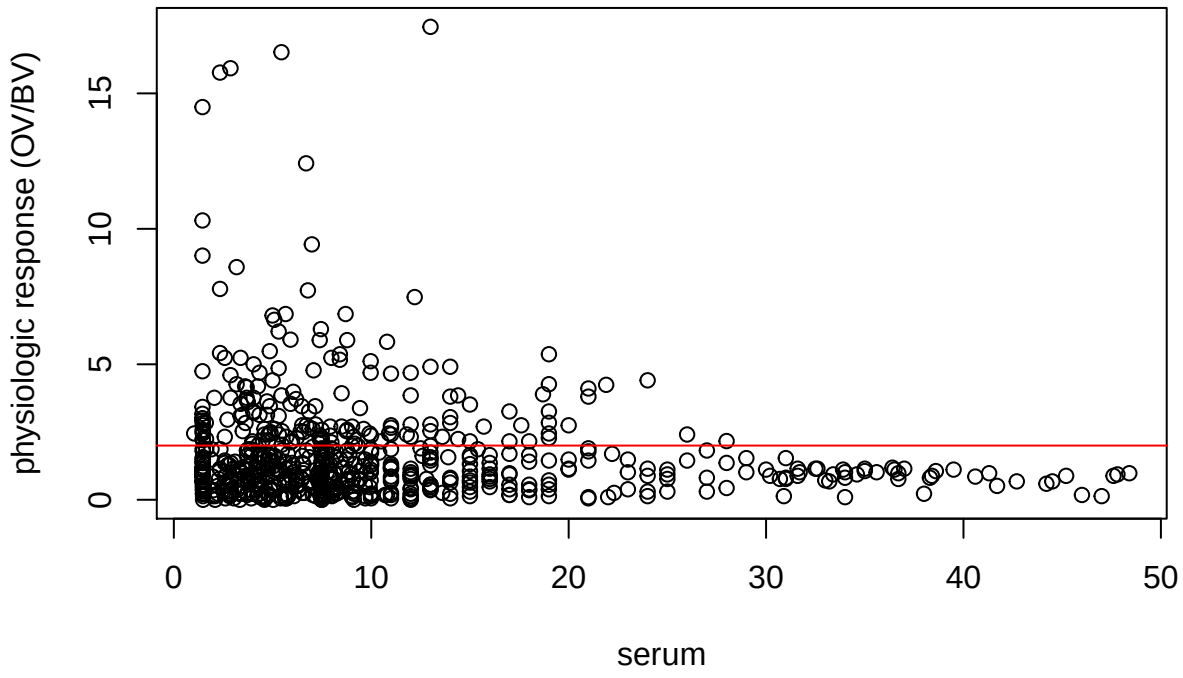
The answer changes much if we restrict the set of patients to exclude those with serum values far from the target level (e.g., 10 nmol/L or less). I would trust the result that excludes higher serum level.

Describe the fit in text. State what the coefficient values are inline, rounded to 3 decimal places.

The estimated coefficients of the first fit are -0.637 and -0.046. The estimated coefficients of the second fit are 0.582 and -0.102.

Invoke the earlier plot chunk to set things up, and then superimpose the regression fit.

```
plot(data$SerumLevelInNgPerMl, data$OV.BV,
     xlab = "serum", ylab = "physiologic response (OV/BV)")
abline(h = 2, col = "red")
```



Use abline to add a horizontal line at your estimate of the serum level at which 97.5% of the people would have their requirements met.