Undergrad Biostatistics - R Training - Week XI

Ege Ulgen

Linear Regression

Rationale behind linear regression

We'll use the pre/post dataset for this exercise. This dataset contains simulated data for pre-intervention measurements (`pre`) for 20 individuals together with their post-intervention measurements (`post`).

```
pre_post_df <- read.csv("../data/pre_post_data.csv")
dim(pre_post_df)
```

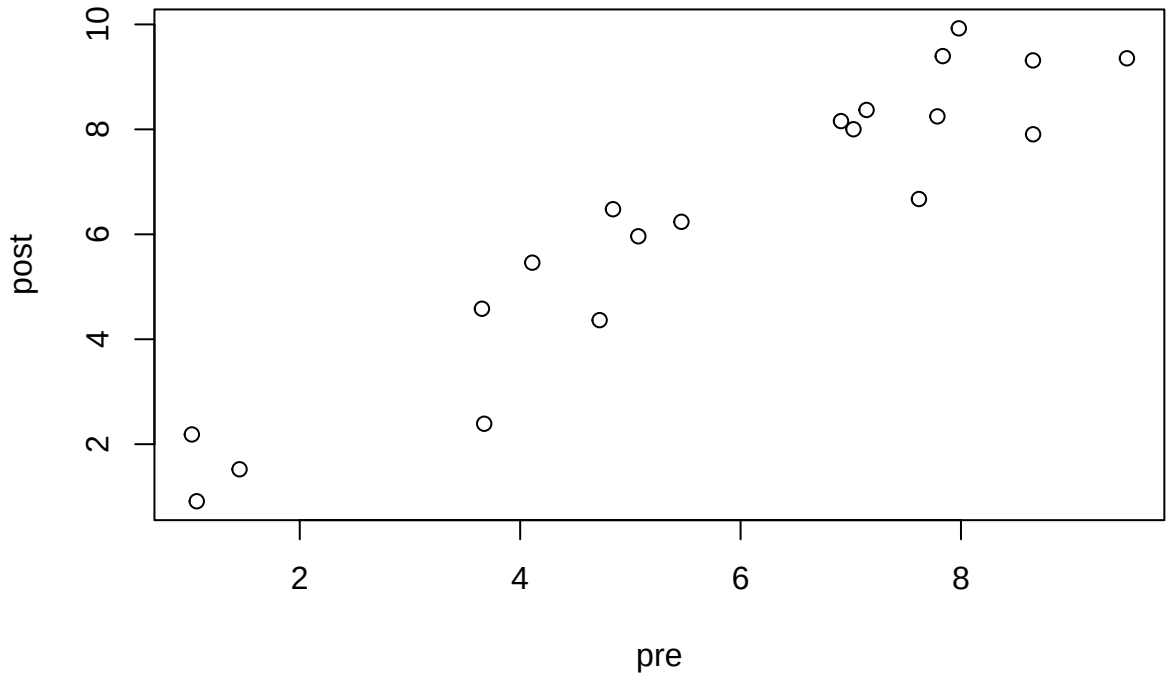
```
## [1] 20  2
```

```
head(pre_post_df, 3)
```

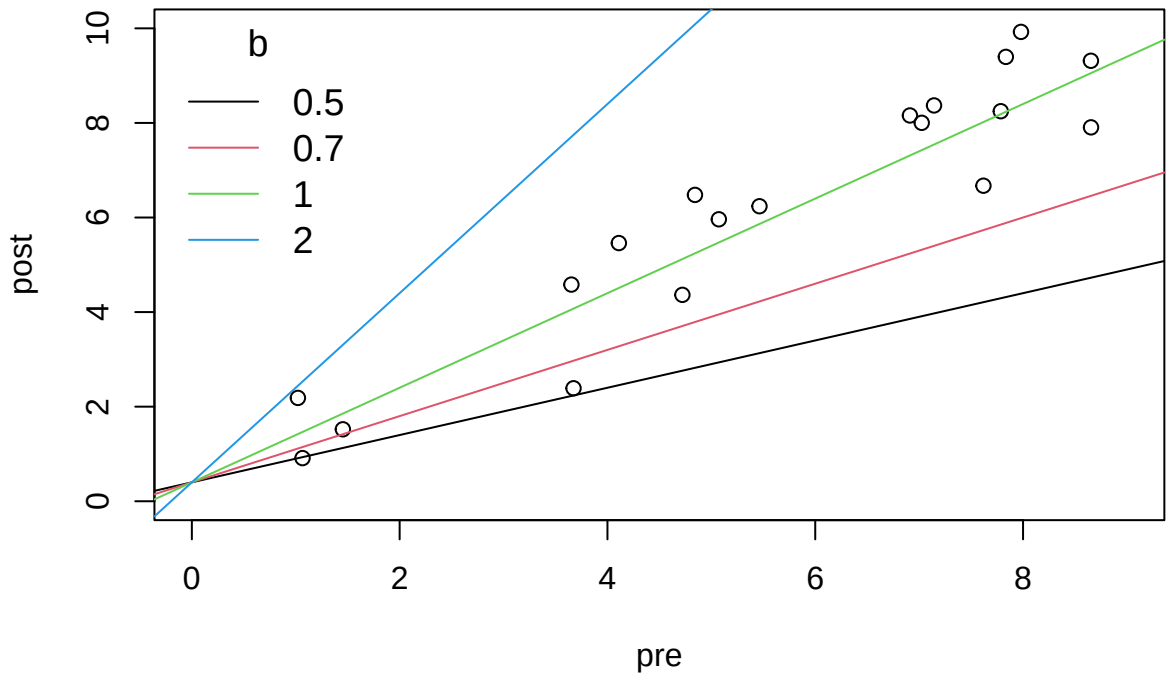
```
##      pre   post
## 1 7.7858 8.2474
## 2 7.6186 6.6725
## 3 8.6532 9.3145
```

Let's visualize the scatter plot to investigate any relationship between `pre` and `post` measurements:

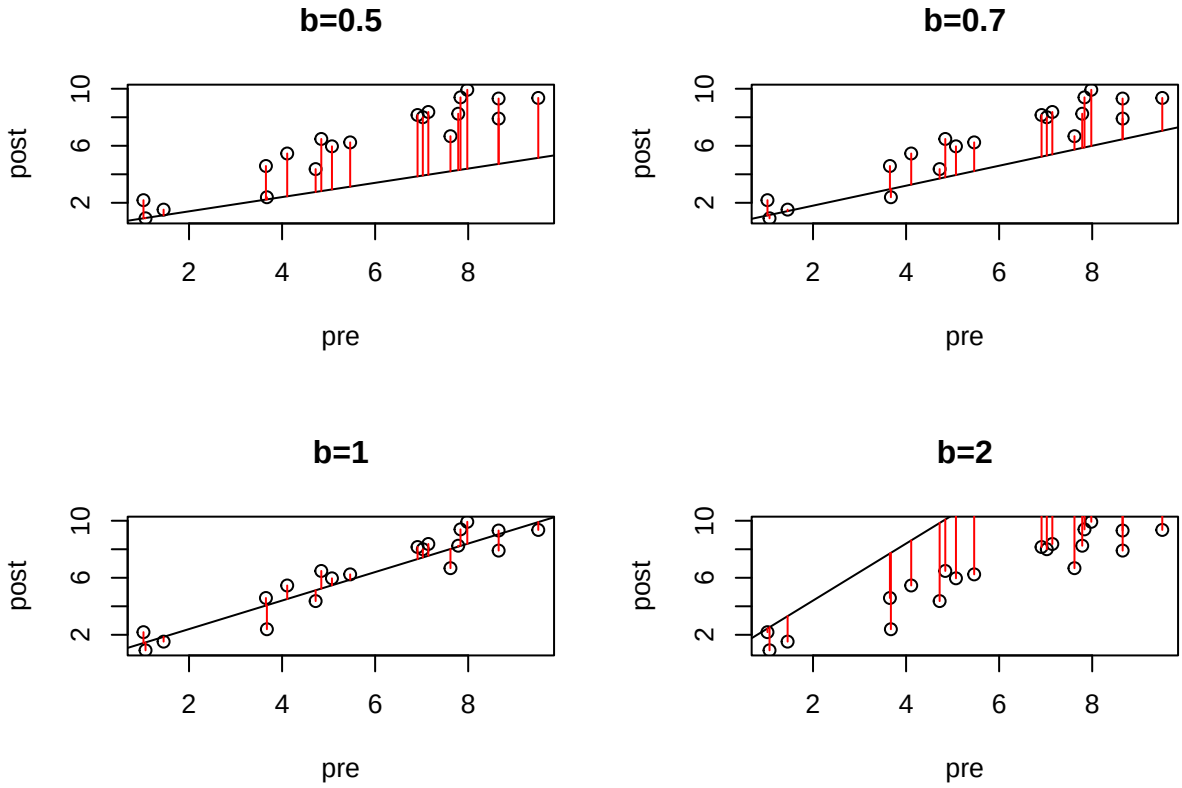
```
plot(post~pre, data = pre_post_df)
```



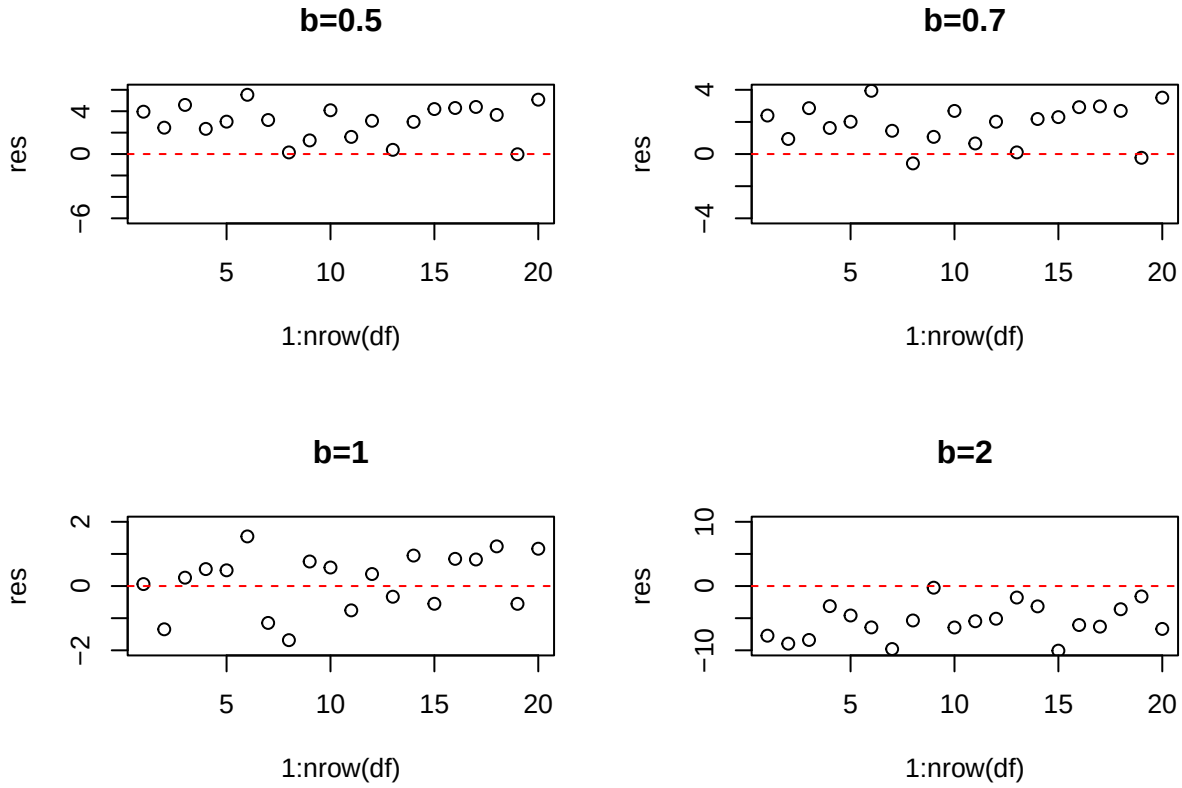
We can fit many lines with varying slope (and intercept, which is kept constant here):



Our aim is to minimize the distance of the residuals to the line (residuals = errors):



The residuals should fluctuate around 0:



Examples

Simple Linear Regression

```
fit_simple <- lm(post~pre, pre_post_df)
summary(fit_simple)
```

```
##
## Call:
## lm(formula = post ~ pre, data = pre_post_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.811 -0.670  0.278  0.668  1.342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4616     0.5165    0.89     0.38
## pre           1.0177     0.0826   12.32  3.3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.94 on 18 degrees of freedom
## Multiple R-squared:  0.894,  Adjusted R-squared:  0.888
## F-statistic:  152 on 1 and 18 DF,  p-value: 3.32e-10
```

```
### prediction
# what is the estimated (predicted) 'post' value given a 'pre' value of 3.2?
0.4616 + 1.0177 * 3.2
```

```
## [1] 3.7182
```

```
# more compactly for multiple new data points
new_data <- data.frame(pre = c(3.2, 1.8, 8.2))
predict(fit_simple, new_data)
```

```
##      1      2      3
## 3.7182 2.2935 8.8067
```

Multiple Linear Regression

We'll use the prostate cancer dataset for this exercise. The main aim of collecting this data set was to inspect the associations between prostate-specific antigen (PSA) and prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomy.

```
prca_df <- read.csv("../data/prostate_cancer.csv")

dim(prca_df)
```

```
## [1] 97  8
```

```
head(prca_df)
```

```
##     PSA    vol     wt age BPH invasion penetration Gleason
## 1 0.651 0.5599 15.959  50   0        0           0       6
## 2 0.852 0.3716 27.660  58   0        0           0       7
## 3 0.852 0.6005 14.732  74   0        0           0       7
## 4 0.852 0.3012 26.576  58   0        0           0       6
## 5 1.448 2.1170 30.877  62   0        0           0       6
## 6 2.160 0.3499 25.280  50   0        0           0       6
```

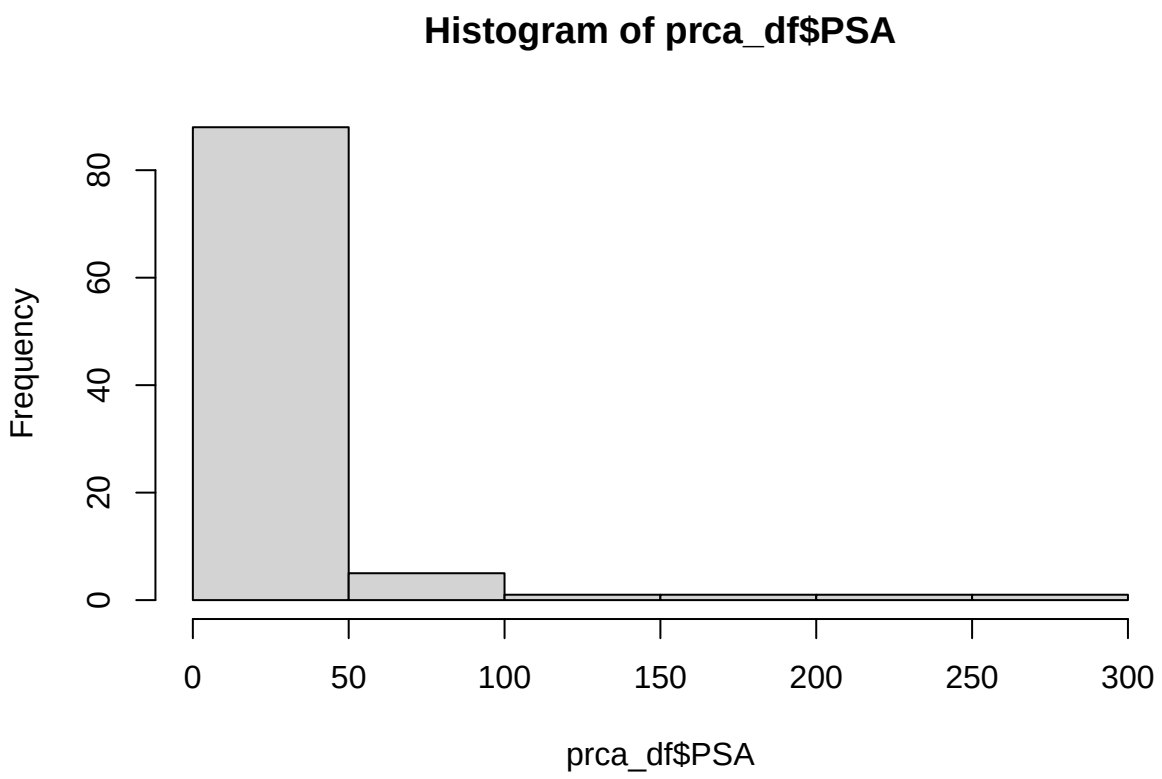
```
# fix factor (categorical) variables
prca_df$invasion <- as.factor(prca_df$invasion)
prca_df$Gleason <- as.factor(prca_df$Gleason)

# summary for all variables
summary(prca_df)
```

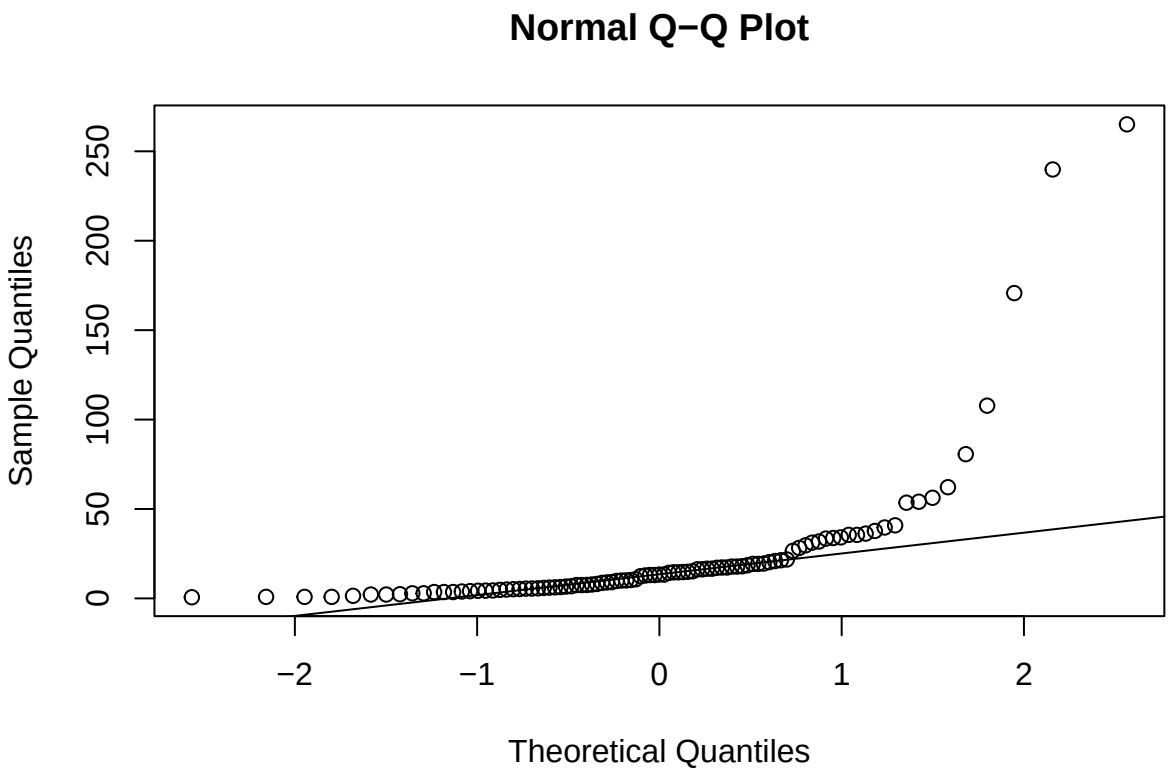
```
##       PSA               vol               wt              age
##  Min.   :  0.651   Min.   : 0.259   Min.   : 10.7   Min.   :41.0
##  1st Qu.:  5.641   1st Qu.: 1.665   1st Qu.: 29.4   1st Qu.:60.0
##  Median : 13.330   Median : 4.263   Median : 37.3   Median :65.0
##  Mean   : 23.730   Mean   : 6.999   Mean   : 45.5   Mean   :63.9
##  3rd Qu.: 21.328   3rd Qu.: 8.415   3rd Qu.: 48.4   3rd Qu.:68.0
##  Max.   :265.072   Max.   :45.604   Max.   :450.3   Max.   :79.0
##       BPH        invasion  penetration     Gleason
##  Min.   : 0.00   0:76     Min.   : 0.000   6:33
##  1st Qu.: 0.00   1:21     1st Qu.: 0.000   7:43
##  Median : 1.35            Median : 0.449   8:21
##  Mean   : 2.53            Mean   : 2.245
##  3rd Qu.: 4.76            3rd Qu.: 3.254
##  Max.   :10.28            Max.   :18.174
```

We will log-transform PSA (outcome of interest, dependent variable) so that it is normally distributed.

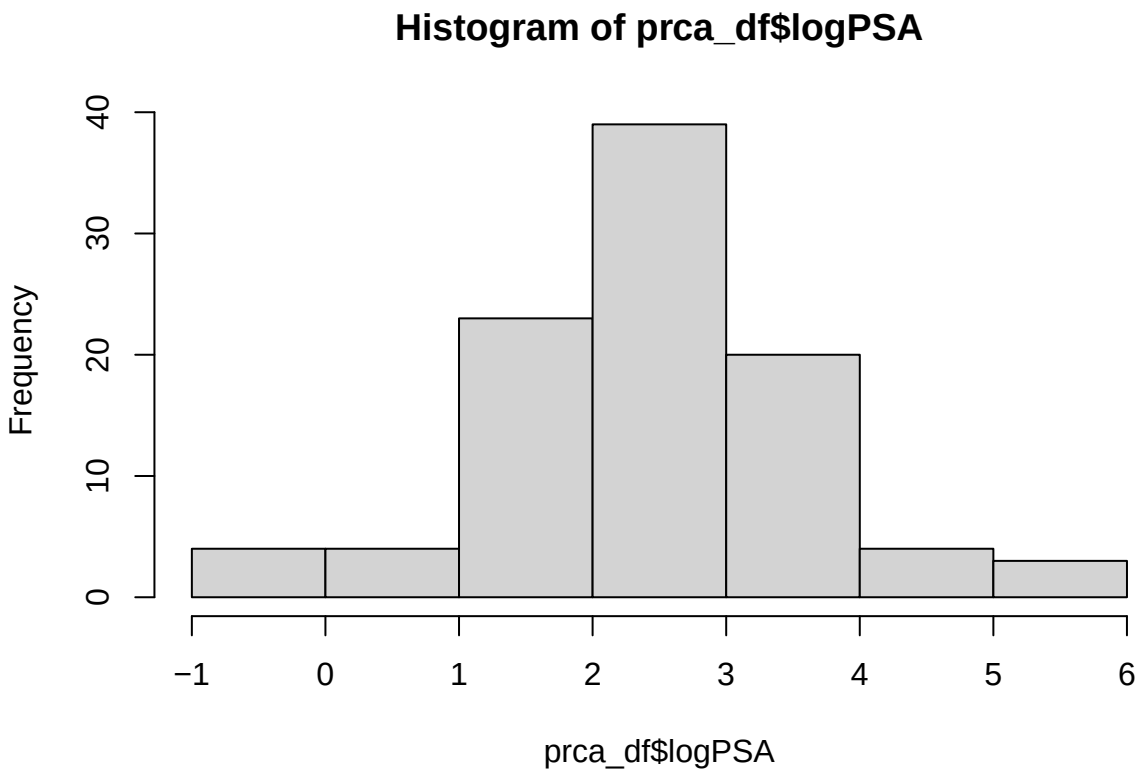
```r
# check normality
hist(prca_df$PSA)
```



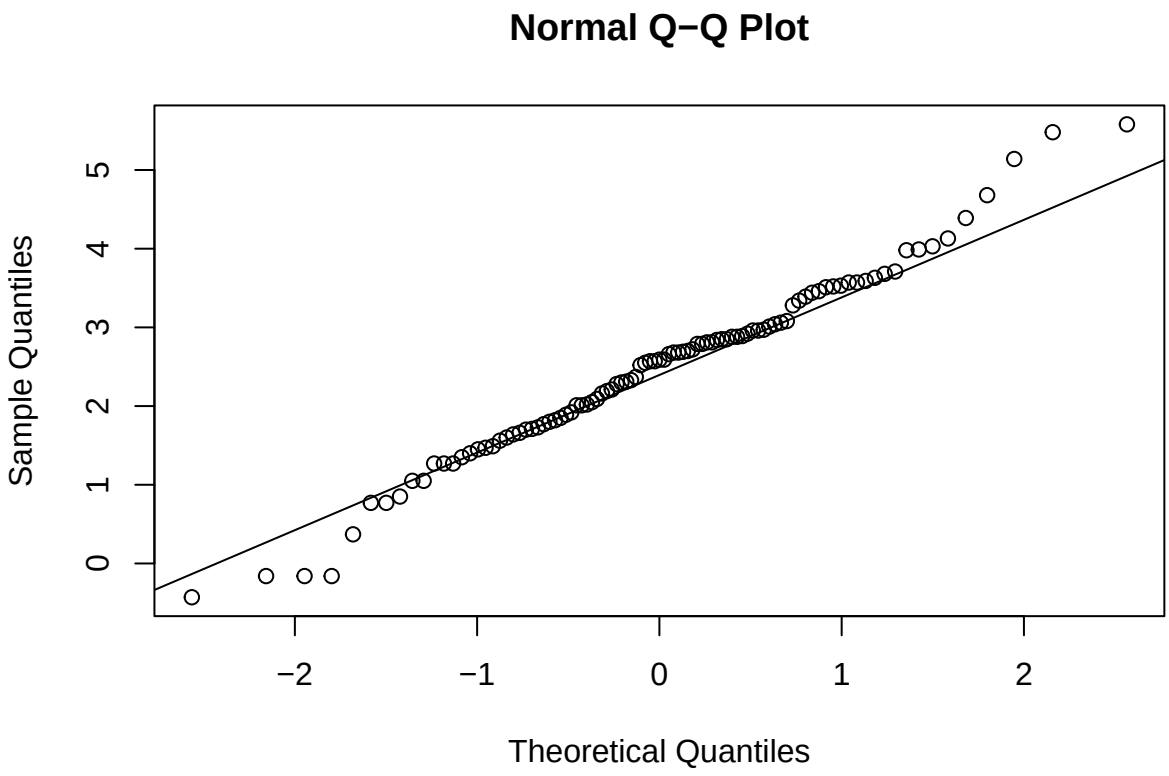

```r
qqnorm(prca_df$PSA)
qqline(prca_df$PSA)
```




```
prca_df$logPSA <- log(prca_df$PSA)
hist(prca_df$logPSA)
```




```
qqnorm(prca_df$logPSA)
qqline(prca_df$logPSA)
```




```
fit1 <- lm(logPSA~vol, data = prca_df)
summary(fit1)
```

```
##
## Call:
## lm(formula = logPSA ~ vol, data = prca_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.289 -0.659  0.149  0.577  1.961
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8055     0.1190    15.2  < 2e-16 ***
## vol           0.0962     0.0113     8.5  2.7e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.874 on 95 degrees of freedom
## Multiple R-squared:  0.432,  Adjusted R-squared:  0.426
## F-statistic: 72.2 on 1 and 95 DF,  p-value: 2.69e-13
```

```
# to make sense of the intercept
fit1_1 <- lm(logPSA~I(vol - min(vol)), data = prca_df)
summary(fit1_1)
```

```
##
```

```
## Call:
## lm(formula = logPSA ~ I(vol - min(vol)), data = prca_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.289 -0.659  0.149  0.577  1.961
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.8304     0.1171    15.6  < 2e-16 ***
## I(vol - min(vol))   0.0962     0.0113     8.5  2.7e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.874 on 95 degrees of freedom
## Multiple R-squared:  0.432,  Adjusted R-squared:  0.426
## F-statistic: 72.2 on 1 and 95 DF,  p-value: 2.69e-13
```

```
fit2 <- lm(logPSA~vol + invasion, data = prca_df)
summary(fit2)
```

```
##
## Call:
## lm(formula = logPSA ~ vol + invasion, data = prca_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.273 -0.626  0.120  0.641  1.610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8035     0.1141   15.81  < 2e-16 ***
## vol           0.0725     0.0133    5.43  4.4e-07 ***
## invasion1     0.7755     0.2541    3.05    0.003 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.838 on 94 degrees of freedom
## Multiple R-squared:  0.483,  Adjusted R-squared:  0.472
## F-statistic: 43.9 on 2 and 94 DF,  p-value: 3.42e-14
```

```
fit3 <- lm(logPSA~vol * invasion, data = prca_df)
summary(fit3)
```

```
##
## Call:
## lm(formula = logPSA ~ vol * invasion, data = prca_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1537 -0.5283  0.0843  0.5585  1.6663
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.6673     0.1289   12.94  < 2e-16 ***
```

```
## vol             0.1021     0.0191    5.35  6.2e-07 ***
## invasion1       1.3260     0.3588    3.70  0.00037 ***
## vol:invasion1  -0.0560     0.0262   -2.13  0.03540 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.823 on 93 degrees of freedom
## Multiple R-squared:  0.507,  Adjusted R-squared:  0.491
## F-statistic: 31.9 on 3 and 93 DF,  p-value: 2.87e-14
```

```
fit4 <- lm(logPSA~vol + Gleason, data = prca_df)
summary(fit4)
```

```
##
## Call:
## lm(formula = logPSA ~ vol + Gleason, data = prca_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.210 -0.584  0.107  0.559  1.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5523     0.1548   10.02  < 2e-16 ***
## vol           0.0758     0.0131    5.79  9.3e-08 ***
## Gleason7      0.4521     0.1928    2.34   0.0212 *
## Gleason8      0.9043     0.2747    3.29   0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.832 on 93 degrees of freedom
## Multiple R-squared:  0.496,  Adjusted R-squared:  0.48
## F-statistic: 30.5 on 3 and 93 DF,  p-value: 7.82e-14
```

```
prca_df$Gleason <- relevel(prca_df$Gleason, "7")
fit4_2 <- lm(logPSA~vol + Gleason, data = prca_df)
summary(fit4_2)
```

```
##
## Call:
## lm(formula = logPSA ~ vol + Gleason, data = prca_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.210 -0.584  0.107  0.559  1.674
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0044     0.1429   14.02  < 2e-16 ***
## vol           0.0758     0.0131    5.79  9.3e-08 ***
## Gleason6     -0.4521     0.1928   -2.34    0.021 *
## Gleason8      0.4522     0.2598    1.74    0.085 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.832 on 93 degrees of freedom
## Multiple R-squared:  0.496,  Adjusted R-squared:  0.48
## F-statistic: 30.5 on 3 and 93 DF,  p-value: 7.82e-14
```

```
prca_df$Gleason <- relevel(prca_df$Gleason, "6")

fit5 <- lm(logPSA~I(vol - min(vol)) + invasion + Gleason, data = prca_df)
summary(fit5)
```

```
##
## Call:
## lm(formula = logPSA ~ I(vol - min(vol)) + invasion + Gleason,
##     data = prca_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.155 -0.474  0.103  0.620  1.633
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.6199     0.1508   10.74   <2e-16 ***
## I(vol - min(vol))   0.0587     0.0144    4.07   0.0001 ***
## invasion1           0.6259     0.2519    2.49   0.0148 *
## Gleason7            0.3544     0.1918    1.85   0.0678 .
## Gleason8            0.7863     0.2716    2.90   0.0047 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.81 on 92 degrees of freedom
## Multiple R-squared:  0.528,  Adjusted R-squared:  0.507
## F-statistic: 25.7 on 4 and 92 DF,  p-value: 2.54e-14
```

Logistic Regression

The data we'll use is `birthwt` from the `MASS` package. The `birthwt` data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

```
# install.packages("MASS")
library(MASS)

data(birthwt)
?birthwt

dim(birthwt)
```

```
## [1] 189  10
```

```
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

```r
# turn categorical variables into factor
birthwt$low <- as.factor(birthwt$low)
birthwt$race <- as.factor(birthwt$race)
birthwt$smoke <- as.factor(birthwt$smoke)
birthwt$ht <- as.factor(birthwt$ht)
birthwt$ui <- as.factor(birthwt$ui)

summary(birthwt)
```

```
##  low          age            lwt      race   smoke        ptl        ht
##  0:130   Min.   :14.0   Min.   : 80   1:96   0:115   Min.   :0.000   0:177
##  1: 59   1st Qu.:19.0   1st Qu.:110   2:26   1: 74   1st Qu.:0.000   1: 12
##          Median :23.0   Median :121   3:67           Median :0.000
##          Mean   :23.2   Mean   :130                  Mean   :0.196
##          3rd Qu.:26.0   3rd Qu.:140                  3rd Qu.:0.000
##          Max.   :45.0   Max.   :250                  Max.   :3.000
##  ui           ftv             bwt
##  0:161   Min.   :0.000   Min.   : 709
##  1: 28   1st Qu.:0.000   1st Qu.:2414
##          Median :0.000   Median :2977
##          Mean   :0.794   Mean   :2945
##          3rd Qu.:1.000   3rd Qu.:3487
##          Max.   :6.000   Max.   :4990
```

We'll be using logistic regression to identify risk factors associated with low infant birth weight (birth weight less than 2.5 kg).

```r
fit6 <- glm(low~. - bwt, data = birthwt, family = binomial)
summary(fit6)
```

```
##
## Call:
## glm(formula = low ~ . - bwt, family = binomial, data = birthwt)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.895  -0.821  -0.532   0.982   2.212
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.48062    1.19689    0.40   0.6880
## age         -0.02955    0.03703   -0.80   0.4249
## lwt         -0.01542    0.00692   -2.23   0.0258 *
## race2        1.27226    0.52736    2.41   0.0158 *
## race3        0.88050    0.44078    2.00   0.0458 *
## smoke1       0.93885    0.40215    2.33   0.0196 *
## ptl          0.54334    0.34540    1.57   0.1157
## ht1          1.86330    0.69753    2.67   0.0076 **
## ui1          0.76765    0.45932    1.67   0.0947 .
## ftv          0.06530    0.17239    0.38   0.7048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 201.28  on 179  degrees of freedom
## AIC: 221.3
##
## Number of Fisher Scoring iterations: 4
```

```
fit7 <- glm(low~lwt + race + smoke + ht, data = birthwt, family = binomial)
summary(fit7)
```

```
##
## Call:
## glm(formula = low ~ lwt + race + smoke + ht, family = binomial,
##     data = birthwt)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.775  -0.875  -0.571   0.963   2.113
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3520     0.9244    0.38   0.7033
## lwt          -0.0179     0.0068   -2.63   0.0084 **
## race2         1.2877     0.5216    2.47   0.0136 *
## race3         0.9436     0.4234    2.23   0.0258 *
## smoke1        1.0716     0.3875    2.77   0.0057 **
## ht1           1.7492     0.6908    2.53   0.0113 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 208.25  on 183  degrees of freedom
## AIC: 220.2
##
## Number of Fisher Scoring iterations: 4
```

```
(exp(-0.0179) - 1) * 100
```

```
## [1] -1.7741
```

Poisson Regression

We will use the `epilepsy` data from the package `HSAUR`. We'll only analyze the first period (4 treatment periods exist) and investigate how various factors affect the number of seizures (recorded in `seizure.rate`).

```
library(HSAUR)
```

```
## Loading required package: tools
```

```
data("epilepsy")

head(epilepsy, 10)
```

```
##     treatment base age seizure.rate period subject
## 1     placebo   11  31            5      1       1
```

```
## 110   placebo   11  31            3      2       1
## 112   placebo   11  31            3      3       1
## 114   placebo   11  31            3      4       1
## 2     placebo   11  30            3      1       2
## 210   placebo   11  30            5      2       2
## 212   placebo   11  30            3      3       2
## 214   placebo   11  30            3      4       2
## 3     placebo    6  25            2      1       3
## 310   placebo    6  25            4      2       3
```

```
summary(epilepsy)
```

```
##      treatment        base            age        seizure.rate    period
##  placebo  :112   Min.   :  6.0   Min.   :18.0   Min.   :  0.00   1:59
##  Progabide:124   1st Qu.: 12.0   1st Qu.:23.0   1st Qu.:  2.75   2:59
##                  Median : 22.0   Median :28.0   Median :  4.00   3:59
##                  Mean   : 31.2   Mean   :28.3   Mean   :  8.26   4:59
##                  3rd Qu.: 41.0   3rd Qu.:32.0   3rd Qu.:  9.00
##                  Max.   :151.0   Max.   :42.0   Max.   :102.00
##
##     subject
##  1      :  4
##  2      :  4
##  3      :  4
##  4      :  4
##  5      :  4
##  6      :  4
##  (Other):212
```

```
## subset for the first period
epilepsy_1_df <- epilepsy[epilepsy$period == 1, ]

head(epilepsy_1_df, 10)
```

```
##    treatment base age seizure.rate period subject
## 1    placebo   11  31            5      1       1
## 2    placebo   11  30            3      1       2
## 3    placebo    6  25            2      1       3
## 4    placebo    8  36            4      1       4
## 5    placebo   66  22            7      1       5
## 6    placebo   27  29            5      1       6
## 7    placebo   12  31            6      1       7
## 8    placebo   52  42           40      1       8
## 9    placebo   23  37            5      1       9
## 10   placebo   10  28           14      1      10
```

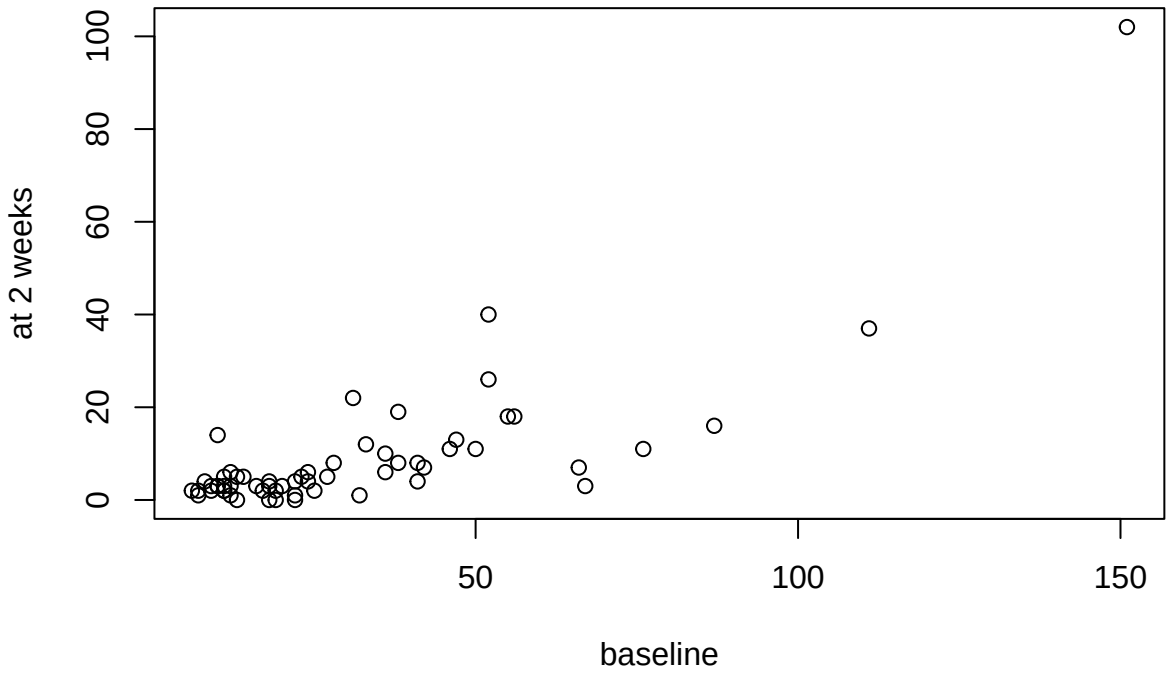
```
summary(epilepsy_1_df)
```

```
##      treatment       base            age        seizure.rate    period
##  placebo  :28   Min.   :  6.0   Min.   :18.0   Min.   :  0.00   1:59
##  Progabide:31   1st Qu.: 12.0   1st Qu.:23.0   1st Qu.:  2.00   2: 0
##                 Median : 22.0   Median :28.0   Median :  4.00   3: 0
##                 Mean   : 31.2   Mean   :28.3   Mean   :  8.95   4: 0
##                 3rd Qu.: 41.0   3rd Qu.:32.0   3rd Qu.: 10.50
##                 Max.   :151.0   Max.   :42.0   Max.   :102.00
##
```

```
##     subject
##  1      : 1
##  2      : 1
##  3      : 1
##  4      : 1
##  5      : 1
##  6      : 1
##  (Other):53
```

As usual, first perform exploratory data analysis:

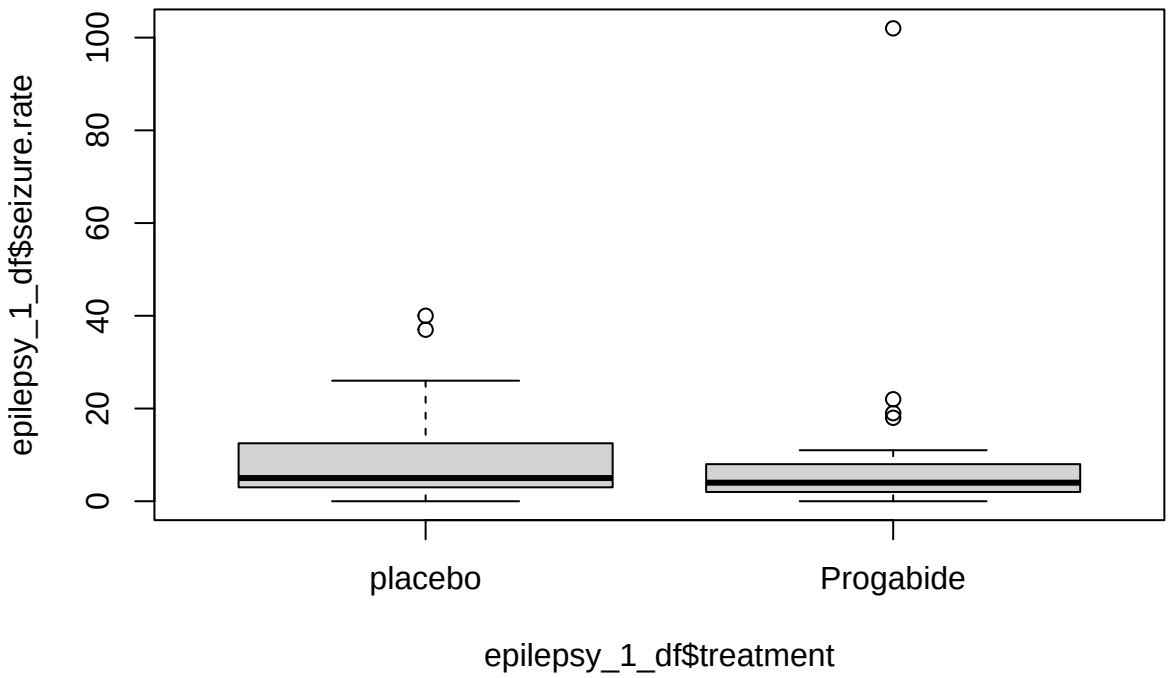
```
# visually compare number of seizures at baseline and at 2 weeks
plot(epilepsy_1_df$base, epilepsy_1_df$seizure.rate, xlab = "baseline", ylab = "at 2 weeks")
```



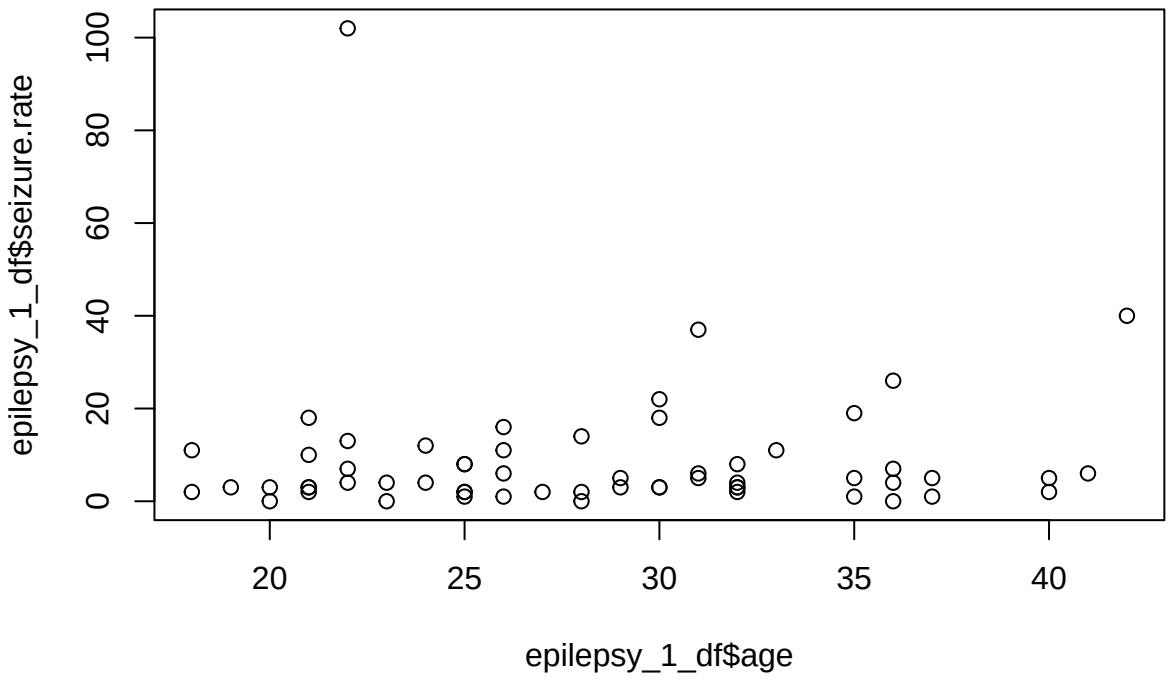

```
cor(epilepsy_1_df$seizure.rate,epilepsy_1_df$base, method = "spearman")
```

```
## [1] 0.66203
```

```
# compare bw/ treatment groups
boxplot(epilepsy_1_df$seizure.rate~epilepsy_1_df$treatment)
```




```
# correlation with age?
plot(epilepsy_1_df$seizure.rate~epilepsy_1_df$age)
```




```
cor(epilepsy_1_df$seizure.rate,epilepsy_1_df$age, method = "spearman")
```

```
## [1] 0.083624
```

We can now build our Poisson regression model:

```
min(epilepsy_1_df$base)
```

```
## [1] 6
```

```
min(epilepsy_1_df$age)
```

```
## [1] 18
```

```
pos_reg <- glm(seizure.rate ~ as.factor(treatment) + I(base - 6) + I(age - 18),
               data = epilepsy_1_df, family = poisson)
summary(pos_reg)
```

```
##
## Call:
## glm(formula = seizure.rate ~ as.factor(treatment) + I(base -
##     6) + I(age - 18), family = poisson, data = epilepsy_1_df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.226  -1.178  -0.526   0.397   4.882
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                    0.750838   0.140979    5.33  1.0e-07 ***
## as.factor(treatment)Progabide -0.118885   0.092641   -1.28      0.2
## I(base - 6)                    0.025736   0.000976   26.37  < 2e-16 ***
## I(age - 18)                    0.046528   0.007818    5.95  2.7e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 746.44  on 58  degrees of freedom
## Residual deviance: 187.38  on 55  degrees of freedom
## AIC: 391.9
##
## Number of Fisher Scoring iterations: 5
```

```
## A patient in placebo group, with 6 previous seizure, and aged 18 had
# approximately 2 seizures on average in the first two weeks after the trial was started
exp(0.750838)
```

```
## [1] 2.1188
```

```
## With 95% confidence, it could be said that there was no difference between
# placebo and progabide (p-value = 0.2). Negative estimate for beta1 indicates
# lowered mean number of seizures for progabide, but the difference from placebo
# was not significant.

## With 95% confidence, it could be said that previous number of seizures occurred
# in the 8-week interval prior to the study start and mean seizure rate was
# significantly associated (p-value < 2 × 10-16). One unit increase in previous
# seizure is associated with approximately 2.6% increase in the mean number of
# seizures in the first two weeks of the trial.
(exp(0.025736) - 1) * 100
```

```
## [1] 2.607
```

```
## With 95% confidence, it could be said that age and mean number of seizures
# was significantly associated (p-value = 2.66 × 10-9). One year increase in age
# was associated with approximately 4.8% increase in the seizure rate in the
# first two weeks of the trial.
(exp(0.046528) - 1) * 100
```

```
## [1] 4.7627
```