Introduction to R

Answers to Session 8 exercises

Statistical Consulting Centre

2 March, 2017

1 Linear regression

(i) Perform a linear regression between age (explanatory variable) and nerdy score (dependent variable).

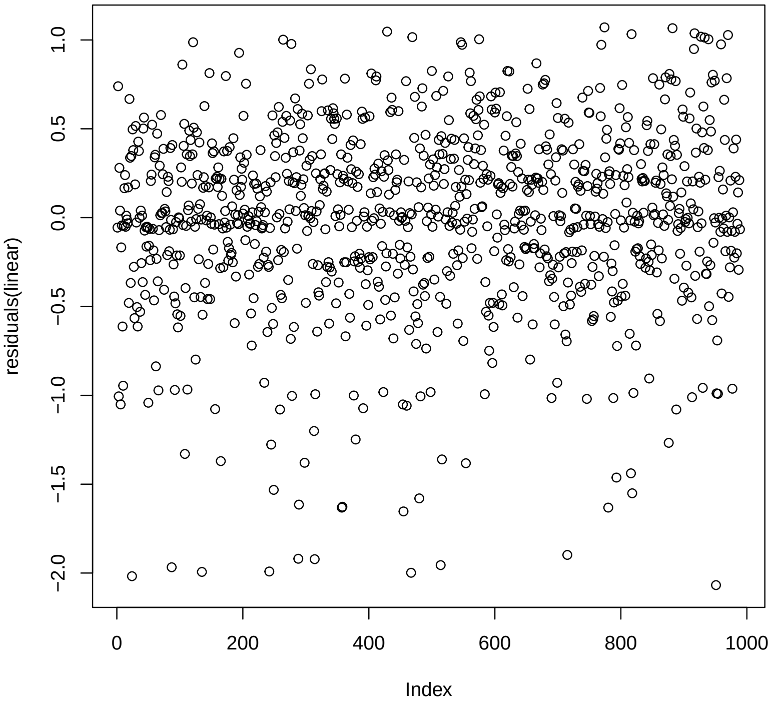
```
linear <- lm(nerdy.sc~age, data=sports.df)
```

(ii) Are the estimated intercept and slope significantly different from zero?

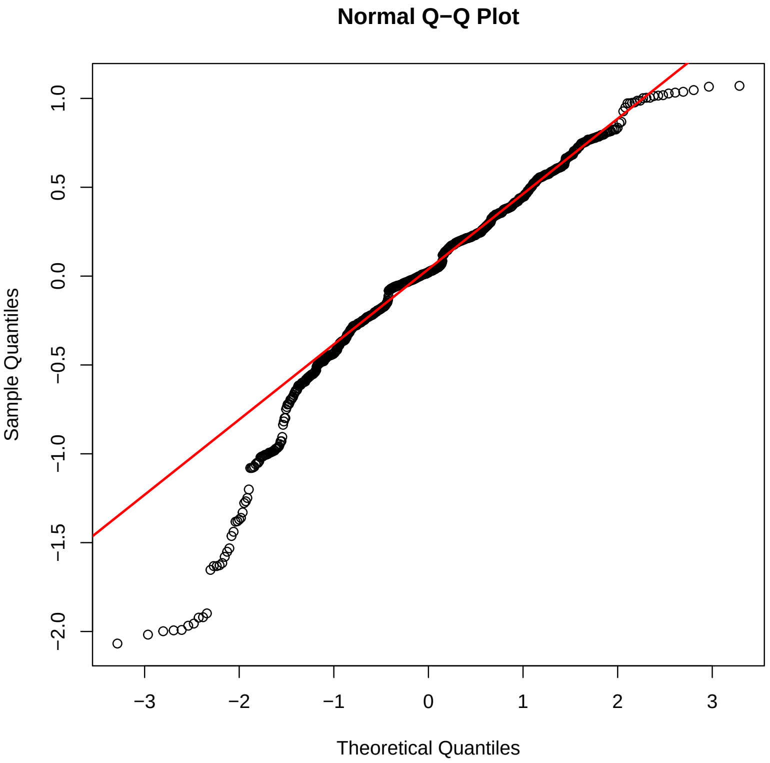
```
summary(linear)



Call:
lm(formula = nerdy.sc ~ age, data = sports.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.06790 -0.24615  0.02059  0.32493  1.07081 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.130076   0.050895  61.501   <2e-16 ***
age         -0.002391   0.000932  -2.566   0.0104 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5073 on 987 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.006626,Adjusted R-squared:  0.00562 
F-statistic: 6.584 on 1 and 987 DF,  p-value: 0.01044
```

(iii) Examine the residuals of the fitted linear model.

```
plot(residuals(linear))
```

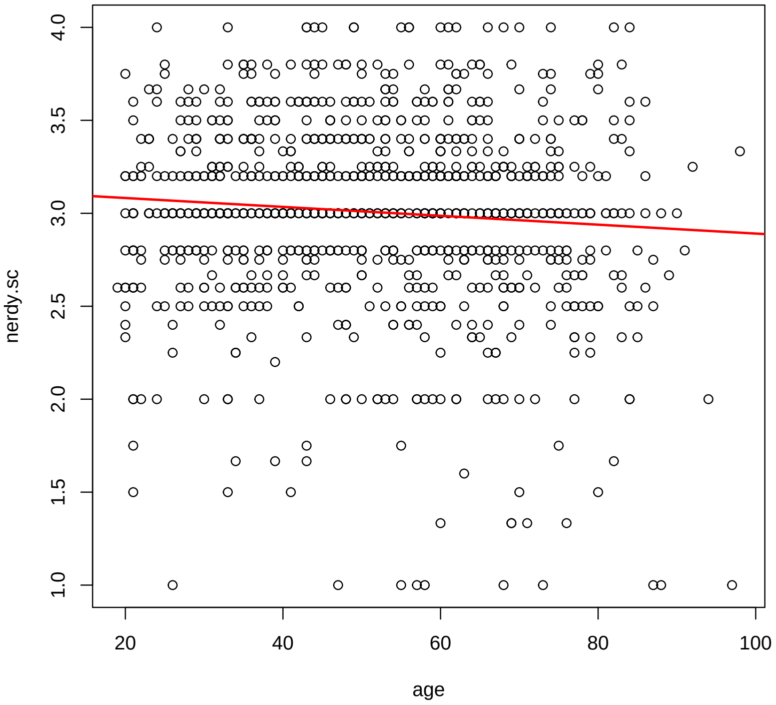


```r
qqnorm(residuals(linear))
qqline(residuals(linear), lwd = 2, col = 2)
```



(iv) Add the fitted line to the scatterplot of nerdy score against age.

```
estimated.intercept <- coef(linear)[1]
estimated.slope <- coef(linear)[2]
with(sports.df, plot(age, nerdy.sc), xlab = "Age", ylab = "Nerdy score")
abline(a = estimated.intercept, b = estimated.slope, lwd = 2, col = 2)
```



(v) What conclusions can you draw? Do you think age and nerdy score are linearly correlated?

2 Logistic Regression

2.1 Continuous explanatory variable

(i) Suppose we want to model the probability of being male, i.e., `gender = Male`. First, ensure that `gender` is a variable with a correct type.

```
# Check gender
class(sports.df$gender)

[1] "character"

table(sports.df$gender)


Female   Male
   535    461

sports.df$gender <- ifelse(sports.df$gender == "Male", 1, 0)
```

(ii) Fit a logistic model with `gender` as the response variable and `nerdy.sc` as the explanatory variable.

```
myglm <- with(sports.df, glm(gender˜nerdy.sc, family = binomial))
```

(iii) Perform an analysis of deviance to determine the overall significance of `nerdy.sc`.

```
anova(myglm, test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: gender

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       988     1365.0
nerdy.sc  1   14.677       987     1350.4 0.0001276 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(iv) Calculate the estimated slope of the logistic regression. What can you conclude about the slope?

```
summary(myglm)


Call:
glm(formula = gender ˜ nerdy.sc, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3205  -1.1092  -0.9144   1.2047   1.6877

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6417     0.4026  -4.078 4.55e-05 ***
nerdy.sc      0.4930     0.1315   3.749 0.000177 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1365.0  on 988  degrees of freedom
Residual deviance: 1350.4  on 987  degrees of freedom
```

```
  (7 observations deleted due to missingness)
AIC: 1354.4

Number of Fisher Scoring iterations: 4
```

2.2 Categorical explanatory variable

(i) We now want to model the probability of living with a partner given age group. `partner` is already of type `factor`. Now, generate a one-way table of `partner` to examine its contents.

```
table(sports.df$partner)
```

```

 No Yes 
122 229 
```

(ii) Set a response variable with `partner` = `Yes`.

```
sports.df$partner <- ifelse(sports.df$partner == "Yes", 1, 0)
```

(iii) Once again geenerate the one-way frequency table of `partner`.

```
table(sports.df$partner)
```

```

  0   1 
122 229 
```

(iv) Fit a logistic model with `partner` as the response variable and `age.group` as the explanatory variable.

```
myglm2 <- with(sports.df, glm(partner~age.group, family = binomial))
```

(v) Is `age.group` a significant predictor of whether or not an individual in particular age group has a partner?

```
anova(myglm2, test = "Chisq")
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: partner

Terms added sequentially (first to last)
```

```
          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        350     453.45
age.group  2  0.19833       348     453.25   0.9056
```

Not at the 5% level of significance since $p = 0.91$.

(vi) Generate a two-way frequency table of `partner` against `age.group`.

```
twoway.tab <- with(sports.df, table(partner, age.group))
twoway.tab
```

```
       age.group
partner Under 40 41 to 60 Over 61
      0       54       36      32
      1       96       69      64
```

(vii) Convert these frequencies to percentages of age group total. Does this table agree with your earlier conclusion?

```
round(100*prop.table(twoway.tab, 2), 1)
```

```
       age.group
partner Under 40 41 to 60 Over 61
      0     36.0     34.3    33.3
      1     64.0     65.7    66.7
```

Yes, since the percentages of `Yes` and `No` are approximately the same across age groups.