MLW / KUHeS Statistics and R short course

Session 5 - Practical (solutions)

Marc Henrion / Jessie Khaki

2026-02-04

Session 5 - Practical (Solutions)

Exercise 1

Using the `adolescent_small.csv` data from Session 4, fit a linear model regressing weight (variable `a104wt`) on age (variable `a12age`).

Test if the regression coefficient of age $\beta_{age} = 0$.

Note:

- deviance = sum of squares
- residual = error

Exercise 1 - Solution

```
ado<-read.csv("dataAndSupportDocs/adolescent_small.csv")

mod1<-glm(a104wt~a12age,data=ado)
# print(mod1)

summary(mod1)
##
## Call:
## glm(formula = a104wt ~ a12age, data = ado)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -12.2481     3.0327  -4.039 7.05e-05 ***
## a12age        3.5562     0.2213  16.067  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 95.29828)
##
##     Null deviance: 49760  on 265  degrees of freedom
## Residual deviance: 25159  on 264  degrees of freedom
##   (35 observations deleted due to missingness)
## AIC: 1971
##
## Number of Fisher Scoring iterations: 2
```

The output above contains all the information we need for this exercise.

The model equation for the model above is simply:

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon$$

where Y is the response variable weight (`a104wt`) and X is the predictor variable age (`a12age`).

We are meant to test the null hypothesis H_0 against the alternative H_1:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

In the table above, we get this from the section `Coefficients`, by finding the row for `a12age` and the p-value for the test of H_0 is the one given in the column `Pr(>|t|)`. Here this is $5.5524318 \times 10^{-41}$ (shown as `<2e-16` in the output; i.e. essentially 0).

Since this p-value is < 0.05, we reject H_0 at the 5% significance level – there is sufficient evidence that the coefficient of `a12age` is not equal to 0. In other words, we reject the null hypothesis of no association between `a104wt` and `a12age`.

Alternatively, you could also calculate the p-value also manually, using an F-test:

```
F<-((mod1$null.deviance-mod1$deviance)/1)/((mod1$deviance)/mod1$df.residual)
P<-1-pf(F,df1=1,df2=mod1$df.residual)
print(P)
```

```
[1] 0
```

Exercise 2

Repeat the ANOVA test from Session 4 as a linear regression model. Do you get the same results?

Note: `glm` will have the word *deviance* in its output. This is just another word for sum of squares.

Recall from Session 4:

`aov(cd41~as.factor(hosp),data=Tbreg)`

Exercise 2 - Solution

Yesterday

```
Tbreg <- read.csv("dataAndSupportDocs/btTBreg.csv")
summary(aov(cd41 ~ as.factor(hosp), data = Tbreg))
##                   Df  Sum Sq Mean Sq F value Pr(>F)
## as.factor(hosp)    4    2204   551.1   0.229  0.922
## Residuals       2995 7194212  2402.1
```

Today

```
mod2 <- glm(cd41 ~ as.factor(hosp), data = Tbreg)
summary(mod2)
##
## Call:
## glm(formula = cd41 ~ as.factor(hosp), data = Tbreg)
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      247.6520     2.0143 122.945   <2e-16 ***
## as.factor(hosp)2   1.8890     2.8427   0.665    0.506
## as.factor(hosp)3   1.8273     2.8634   0.638    0.523
## as.factor(hosp)4   1.7912     2.8011   0.639    0.523
## as.factor(hosp)5   0.1738     2.8427   0.061    0.951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2402.074)
##
##     Null deviance: 7196416  on 2999  degrees of freedom
## Residual deviance: 7194212  on 2995  degrees of freedom
## AIC: 31873
##
## Number of Fisher Scoring iterations: 2
```

You need to manually compute the F statistic:

RSS = (7196416-7194212) with d.f. k1=5-1=4

ESS = 7196212 with d.f. k2=3000-4-1=2995

$$F = \frac{RSS/k1}{ESS/k2} = 0.2294 \sim F_{4,2995}$$

```r
F<-((7196416-7194212)/4)/(7194212/2995)
P<-1-pf(F,df1=4,df2=2995)
print(P)
```

```
[1] 0.922007
```

Exercise 3

Using the `adolescent_small.csv` data, fit the following GLM model:

Weight `a104wt` as a function of

- age `a12age`
- height `a103ht`
- hiv `hiv`
- sex `a13sex`

Produce:

- a residuals vs. fitted values graph
- histogram of the residuals
- a QQ plot.

Exercise 3 - Solution

We fit a multiple linear regression model with an identity link function and normal distribution.

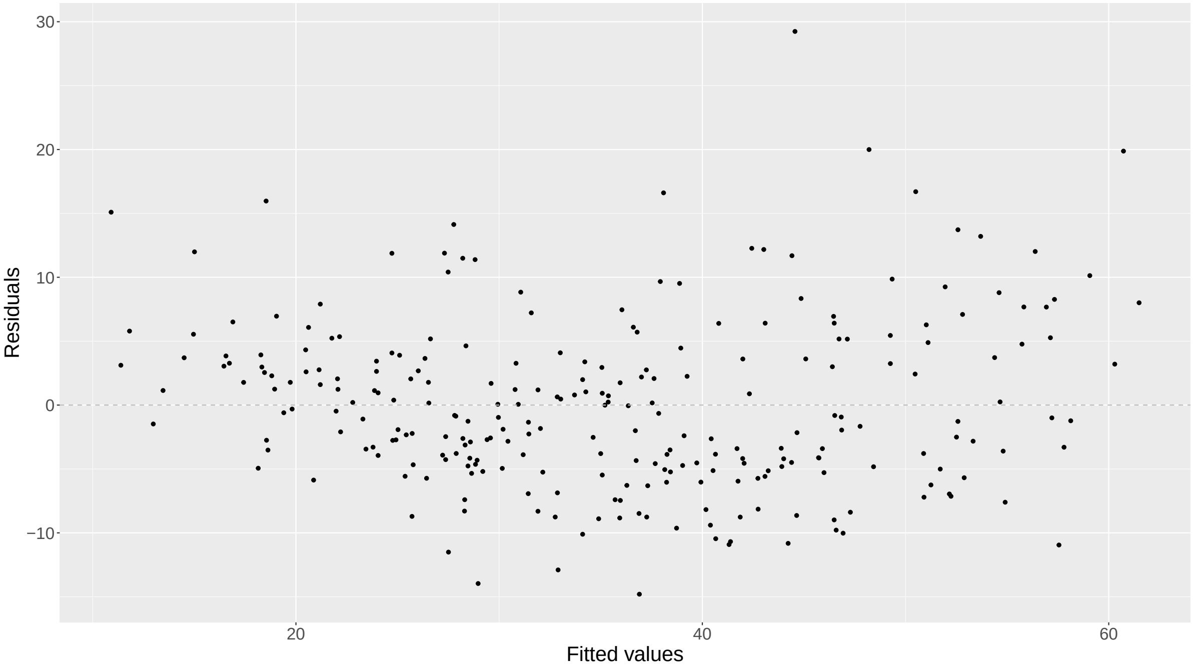
```
mod3 <- glm(a104wt ~ a12age + a103ht + hiv + a13sex, data = ado)
summary(mod3)
##
## Call:
## glm(formula = a104wt ~ a12age + a103ht + hiv + a13sex, data = ado)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50.87129    4.84188 -10.507  < 2e-16 ***
## a12age        1.32731    0.23284   5.700 3.26e-08 ***
## a103ht        0.49929    0.04457  11.203  < 2e-16 ***
## hivpositive  -6.41914    0.90752  -7.073 1.42e-11 ***
## a13sexMale   -1.26779    0.84335  -1.503    0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 45.36643)
##
##     Null deviance: 47549  on 262  degrees of freedom
## Residual deviance: 11705  on 258  degrees of freedom
##   (38 observations deleted due to missingness)
## AIC: 1756.6
##
## Number of Fisher Scoring iterations: 2
```

Having fitted the model we observe very strong association of weight with age (`a12age`), height (`a103ht`) and hiv (`hiv`) status, but no statistically significant association with sex (`a13sex`).

But the p-values in the table above are only correct if the model assumptions are correct, so we need to check these!

We start off by plotting residuals against fitted values.

```
dfResPred<-data.frame(
  pred=predict(mod3,data=ado),
  res=residuals(mod3)
)

dfResPred %>%
  ggplot(mapping=aes(x=pred,y=res)) +
  geom_point() +
  geom_hline(yintercept=0,lty=2,col="grey") +
  xlab("Fitted values") +
  ylab("Residuals") +
  theme(text=element_text(size=20))
```

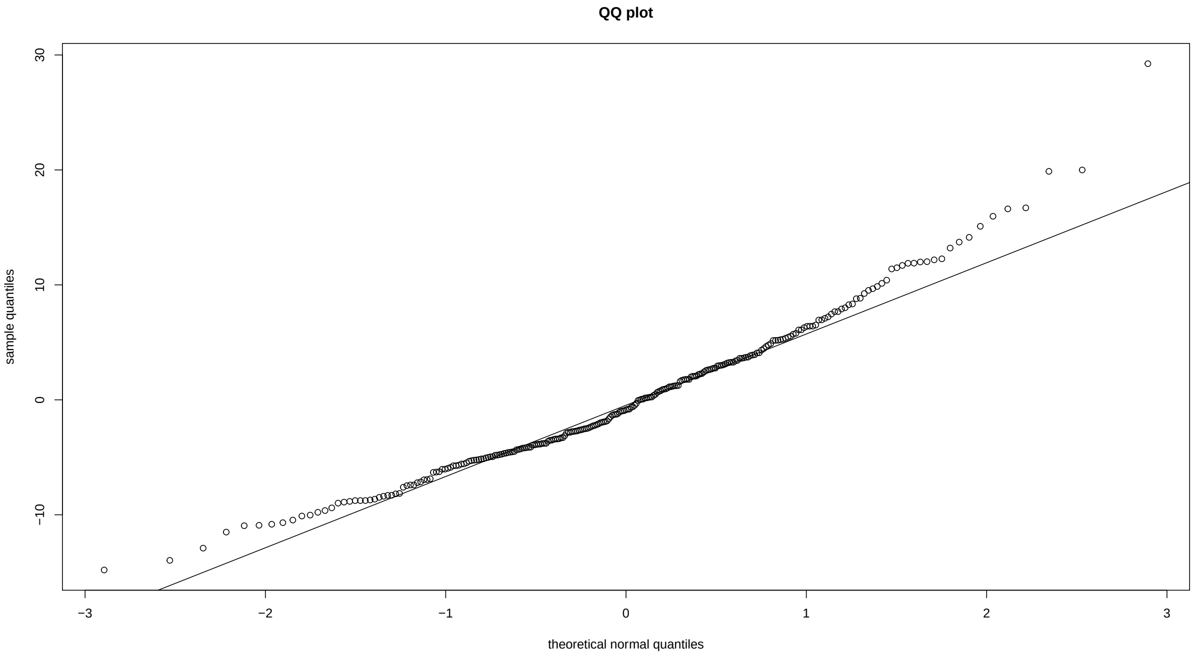


While this does not look too bad, you should note that at the lower and upper end there are fewer residuals below zero than above zero - indicating that the model is not fully fitting well and in fact we have a mild deviation from the assumed normality with mean 0 as clearly the residuals are not centered on zero over the range of fitted values. (If you simply typed `plot(mod3`, the graph will contain a helpful red trend line that makes this observation a bit clearer.)

There are also a few quite large residuals that indicate the model is not fitting well for these.

The slight inappropriateness of the normal distribution for the residuals is confirmed by a QQ-plot:

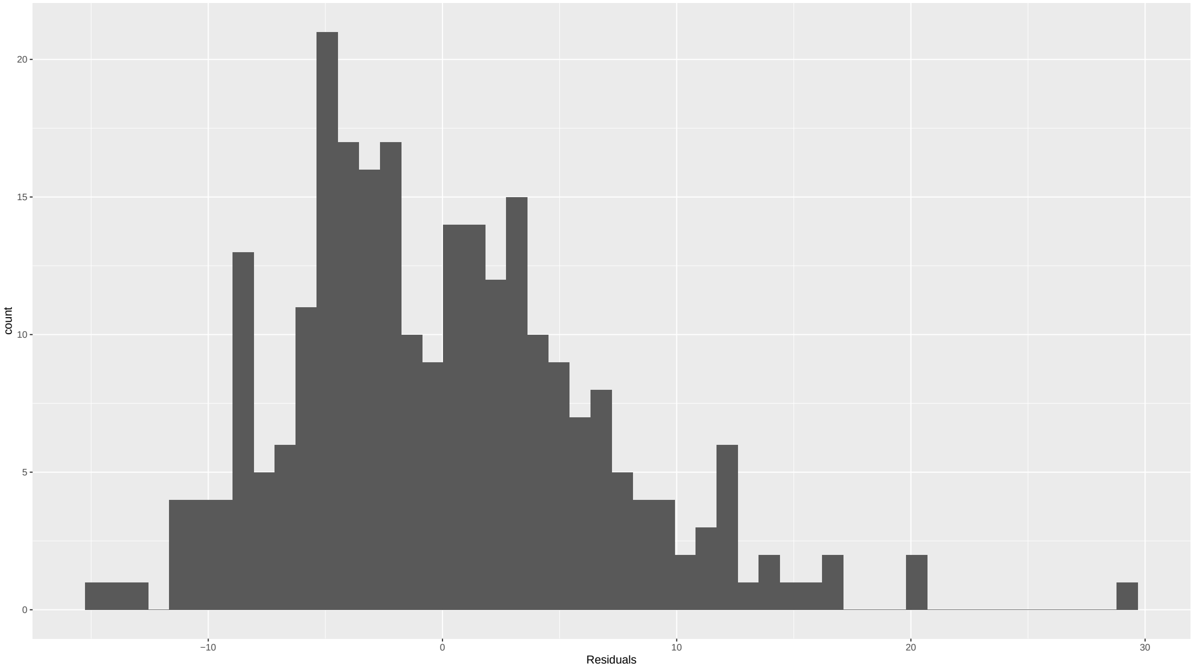
```
qqnorm(residuals(mod3),
     xlab="theoretical normal quantiles",
     ylab="sample quantiles",
     main="QQ plot")
qqline(residuals(mod3))
```



We got deviations from the diagonal at both the lower and upper end - an indication that the implied normal distribution of residuals is not fully valid.

We can also plot a histogram of the residuals:

```
dfResPred %>%
  ggplot(mapping=aes(x=res)) +
  geom_histogram(bins=50) +
  xlab("Residuals")
```



The histogram shows that the distribution of the residuals is slightly skew. This indicates that for some individuals, the model predicts much larger weights than were actually recorded.

In conclusion: the model does not fit that well. It is not a catastrophe, and given how string the association p-values were, we can probably still be confident that thes associations are robust. However, if we wanted to use the model for prediction, we should be careful as it is clear that the model will overestimate weights for some individuals. To improve the model, we could look whether there are additional variables that we should have included, whether we should model e.g. age non-linearly, or whether age and height are collinear with this causing issues.

Exercise 4

You are given the following data:

$$\mathbf{x} = (-6, -6, -4, -1, 0.5, 2, 8, 8, 11, 11.5)^T$$

$$\mathbf{y} = (-3.7, -4.3, -3.9, -4.6, 0.5, -6.9, 10.2, 16.1, 6, 19.5)^T$$

a. Fit a linear regression model to these data and show the model output.
b. Describe the resulting regression line:
    - What is the relationship between variables X and Y?
    - How much (on average) does Y change when X changes by 1?
    - What value does Y take (on average) when $X = 0$?
c. Compute the coefficient of determination R^2, the adjusted R^2, the likelihood and the AIC. Which of these tell you how good your model fits the data?
d. Compute the residuals $r_i = y_i - \hat{y}_i$ and do a normal distribution QQ plot.
e. What other diagnostic check(s) could you do? Do this and explain whether you think this is a good model.
f. Re-fit the model, but now including a term for X^2: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$. Check and discuss the resulting model and compare it to the previous one. Which model would you recommend for this dataset?

Exercise 4 - Solution

a.

Using R, we need to code up the data, then fit the model using either the `lm()` or the `glm()` function (`glm()` used below):

```
x <- c(-6, -6, -4, -1, 0.5, 2, 8, 8, 11, 11.5)
y <- c(-3.7, -4.3, -3.9, -4.6, 0.5, -6.9, 10.2, 16.1, 6, 19.5)

mod <- glm(y ~ x)

summary(mod)
##
## Call:
## glm(formula = y ~ x)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09255    1.87070   0.049  0.96175
## x            1.16560    0.27157   4.292  0.00264 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 30.74697)
##
##     Null deviance: 812.39  on 9  degrees of freedom
## Residual deviance: 245.98  on 8  degrees of freedom
## AIC: 66.405
##
## Number of Fisher Scoring iterations: 2
```

b.

- From the regression line, we conclude there is a *positive* relationship between X and Y since $\hat{\beta}_1 > 0$: as X increases, so Y increases. The p-value for testing the null hypothesis $h_0 : \beta_1 = 0$ against a two-sided alternative $H_1 : beta_1 \neq 0$ is low (0.0026), well below the usual statistical significance threshold of 0.05. We have therefore evidence that this positive relationship is real and not just due to random noise in the (finite) dataset.
- When X increases by 1, then Y increases - on average - by 1.17. Let $x_2 = x_1 + 1$, then:

$$\hat{y}_2 - \hat{y}_1 = 0.09 + 1.17x_2 - (0.09 + 1.17x_1) = 1.17(x_2 - x_1) = 1.17$$

- When $X = 0$, then, on average, we expect Y to be 0.09. Let $x = 0$, then:

$$\hat{y} = 0.09 + 1.17x = 0.09 + 1.17 * 0 = 0.09$$

c.

You can extract most of these quite conveniently from the R model object. `glm()` does not compute R^2, so you could just refit the model using `lm()` to get this. Likewise `lm()` will not compute a model likelihood (hence also no AIC), so you will need to get this from `glm()`. Recall that a simple linear regression does not make any distributional assumption and hence cannot yield a likelihood, only when you assume the errors/residuals to be normally distributed, i.e. by using `glm()`, will you be able to compute a likelihood. The log-likelihood can be extracted using the function `logLik()` on the GLM model object. To get the actual likelihood, just exponentiate.

```
modLm<-lm(y~x)
R2<-summary(modLm)$r.squared
R2adj<-summary(modLm)$adj.r.squared
likelihood<-exp(logLik(mod))
AIC<-mod$aic

print(paste(sep="","The coefficient of determination R2 is ",R2,"."))
```

```
[1] "The coefficient of determination R2 is 0.697219249916526."
```

```
print(paste(sep="","The adjusted R2 is ",R2adj,"."))
```

```
[1] "The adjusted R2 is 0.659371656156092."
```

```
print(paste(sep="","The model likelihood is ",likelihood,"."))
```

```
[1] "The model likelihood is 7.64129162540492e-14."
```

```
print(paste(sep="","The AIC is ",AIC,"."))
```

```
[1] "The AIC is 66.405249304221."
```

You can however also calculate all of these manually (same results up to rounding errors):

```
beta<-coef(mod)
r<-y-(beta[1]+beta[2]*x) # same as y<-resid(mod)

R2<-beta[2]^2*sum((x-mean(x))^2)/sum((y-mean(y))^2)
R2adj<-1-(1-R2)*(length(x)-1)/(length(x)-1-1)
    # p in the lecture notes is the number of predictors, not number of parameters
likelihood<-prod(dnorm(r,sd=sd(r)))
AIC<-(-2*log(likelihood))+2*(2+1)

print(paste(sep="","The coefficient of determination R2 is ",R2,"."))
```

```
[1] "The coefficient of determination R2 is 0.697219249916526."
```

```
print(paste(sep="","The adjusted R2 is ",R2adj,"."))
```

```
[1] "The adjusted R2 is 0.659371656156092."
```

```
print(paste(sep="","The model likelihood is ",likelihood,"."))
```

```
[1] "The model likelihood is 7.43920561909125e-14."
```

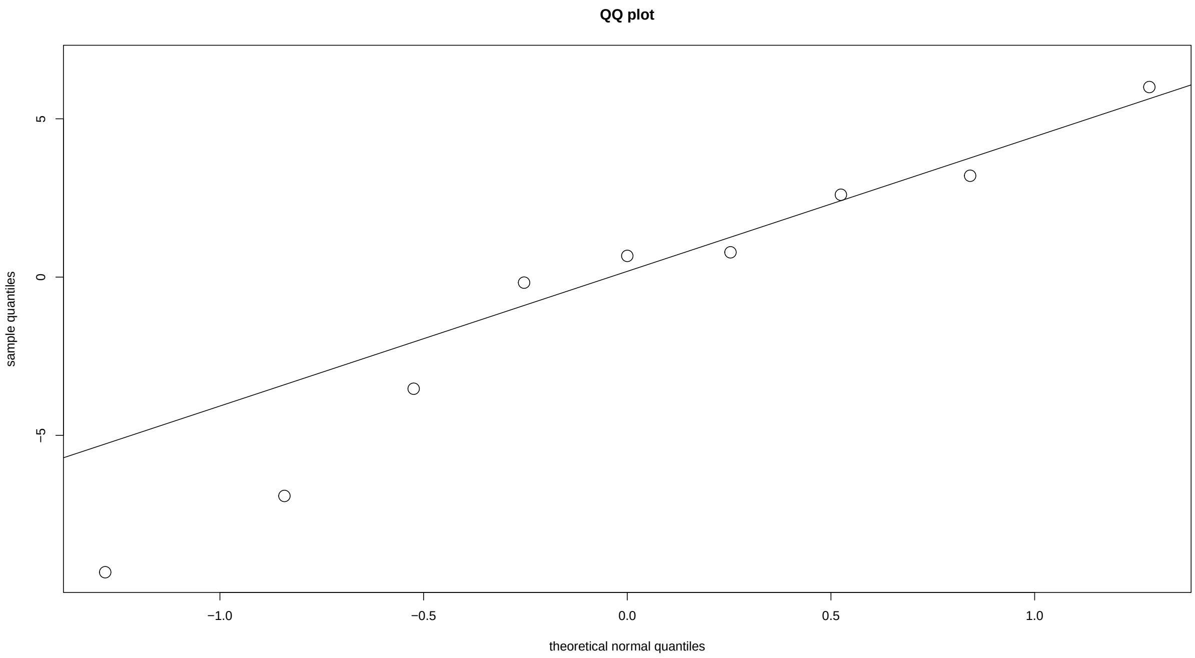
```
print(paste(sep="","The AIC is ",AIC,"."))
```

```
[1] "The AIC is 66.4588544607993."
```

d.

The individual residuals are listed in the table / calculation worksheet given under a. above. We can easily produce a QQ plot:

```
r<-resid(mod) # had already computed this manually above; just redoing it here
theoQ<-qnorm(order(order(r))/length(r))

plot(theoQ,r,
     xlab="theoretical normal quantiles",
     ylab="sample quantiles",
     main="QQ plot", cex=2)
qqline(r) # just adds the line
```



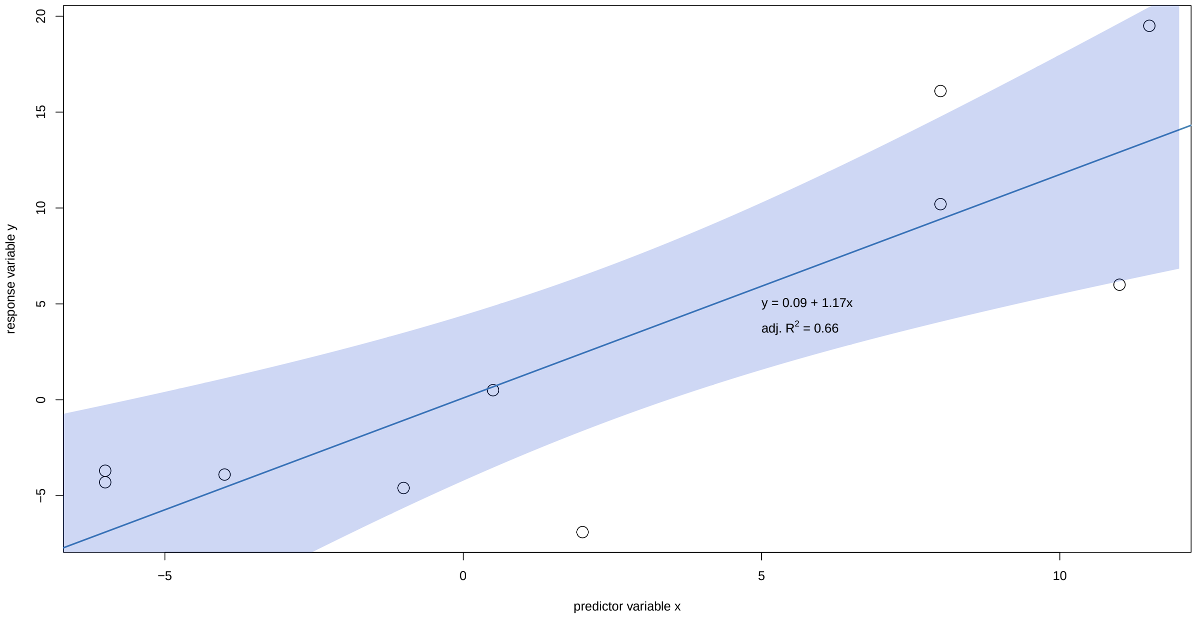
With only 10 data points, it is a bit difficult to interpret this. Overall it looks OK, but perhaps some deviation at the lower end - this could indicate that the normality of residuals assumption is not fully met.

e.

Goodness of fit

We have not yet actually looked at the actual model fit. This should be the first check you do: does the model seem to fit?

```
xx<-seq(-12,12,length=500)
fit<-as.data.frame(predict(modLm,newdata=data.frame(x=xx),interval="confidence"))

plot(x,y,cex=2,xlab="predictor variable x",ylab="response variable y")
abline(a=beta[1],b=beta[2],col="steelblue",lwd=2)
polygon(c(xx,xx[length(xx):1]),c(fit$lwr,fit$upr[length(xx):1]),
      border=NA,col=rgb(0,50,200,alpha=50,maxColorValue=255))
text(x=5,y=5,adj=0,
     labels=paste(sep="","y = ",round(beta[1],digits=2)," + ",round(beta[2],digits=2),"x"))
text(x=5,y=3.75,adj=0,
     labels=substitute(paste("adj. ",R^2," = ",R2adj),list(R2adj=round(R2adj,digits=2))))
```

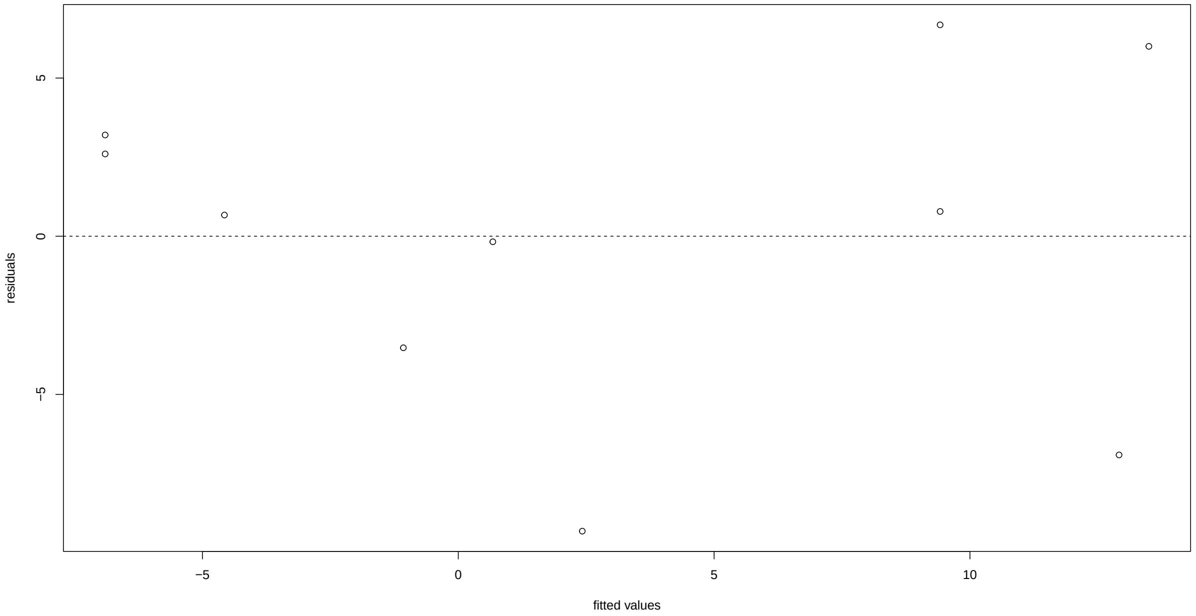


With so few data points, it's difficult to say much. Overall the line seems a reasonable fit, but you could argue that from $X = -5$, to $X = 5$, a better fit would be just a flat line, to be followed by a steeper increase in Y with X for $X > 5$.

Residuals vs. fitted values

Next, we can plot residuals against fitted values and check if the residuals look to be randomly distributed, are homoscedastic and that there are no obvious outliers.

```
yhat<-beta[1]+beta[2]*x # same as predict(mod)

plot(yhat,r,xlab="fitted values",ylab="residuals")
abline(h=0,lty=2)
```

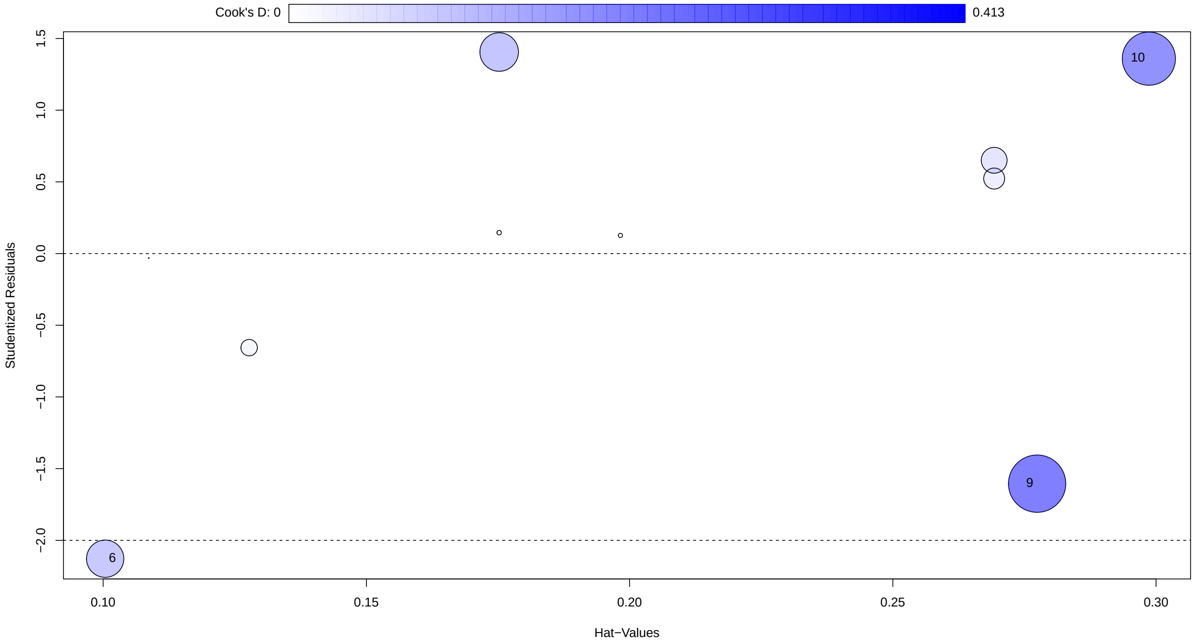


Too little data points to say anything definite, but it seems there is a wider spread around 0 for larger predicted values.

Influential observations

Finally we can check for outliers & influential observations.

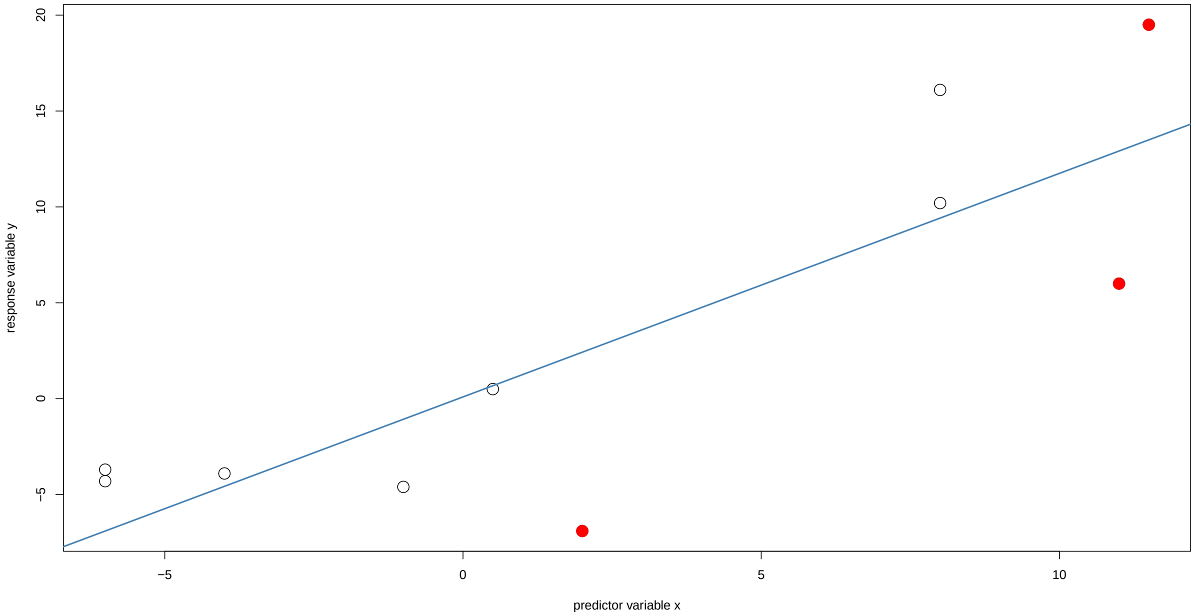
```r
require(car)
influencePlot(mod)
```



```
     StudRes       Hat     CookD
6  -2.128233 0.1003838 0.1753472
9  -1.604823 0.2774047 0.4130223
10  1.359498 0.2986328 0.3557561
```

This suggests 3 potentially influential observations: observations number 6, 9 and 10. We highlight them in red below.

```
plot(x,y,cex=2,xlab="predictor variable x",ylab="response variable y")
points(x[c(6,9,10)],y[c(6,9,10)],pch=19,cex=2,col="red")
abline(a=beta[1],b=beta[2],col="steelblue",lwd=2)
```



f.

Let's fit the model with a term for X^2:

```
df <- data.frame(y = y, x = x, x2 = x^2)
mod2 <- glm(y ~ x + x2, data = df)

summary(mod2)
##
## Call:
## glm(formula = y ~ x + x2, data = df)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.38487    2.70194  -0.883   0.4067
## x            0.77845    0.40903   1.903   0.0987 .
## x2           0.07179    0.05808   1.236   0.2563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 28.84427)
##
##     Null deviance: 812.39  on 9  degrees of freedom
## Residual deviance: 201.91  on 7  degrees of freedom
## AIC: 66.431
##
## Number of Fisher Scoring iterations: 2
```

We note that none of the regression coefficients for X or X^2 are statistically significant anymore (using $p < 0.05$ as threshold).

Goodness of fit

Let's compute goodness of fit metrics:

```
mod2Lm<-lm(y~x+x2,data=df)
R2<-summary(mod2Lm)$r.squared
R2adj<-summary(mod2Lm)$adj.r.squared
likelihood<-exp(logLik(mod2))
AIC<-mod2$aic

print(paste(sep="","The coefficient of determination R2 is ",R2,"."))
```

```
[1] "The coefficient of determination R2 is 0.751461526840908."
```

```
print(paste(sep="","The adjusted R2 is ",R2adj,"."))
```

```
[1] "The adjusted R2 is 0.680450534509739."
```

```
print(paste(sep="","The model likelihood is ",likelihood,"."))
```

```
[1] "The model likelihood is 2.05040639246456e-13."
```

```
print(paste(sep="","The AIC is ",AIC,"."))
```

```
[1] "The AIC is 66.4311363903853."
```

We note that according to R^2, the new model explains slightly more of the variation in the dataset (75% vs 70% on standard R^2, 68% vs. 66% on adjusted R^2).

The likelihood is also slightly better (larger) but the AICs are virtually the same, with the model without the X^2 term having a marginalyl better (lower) AIC.

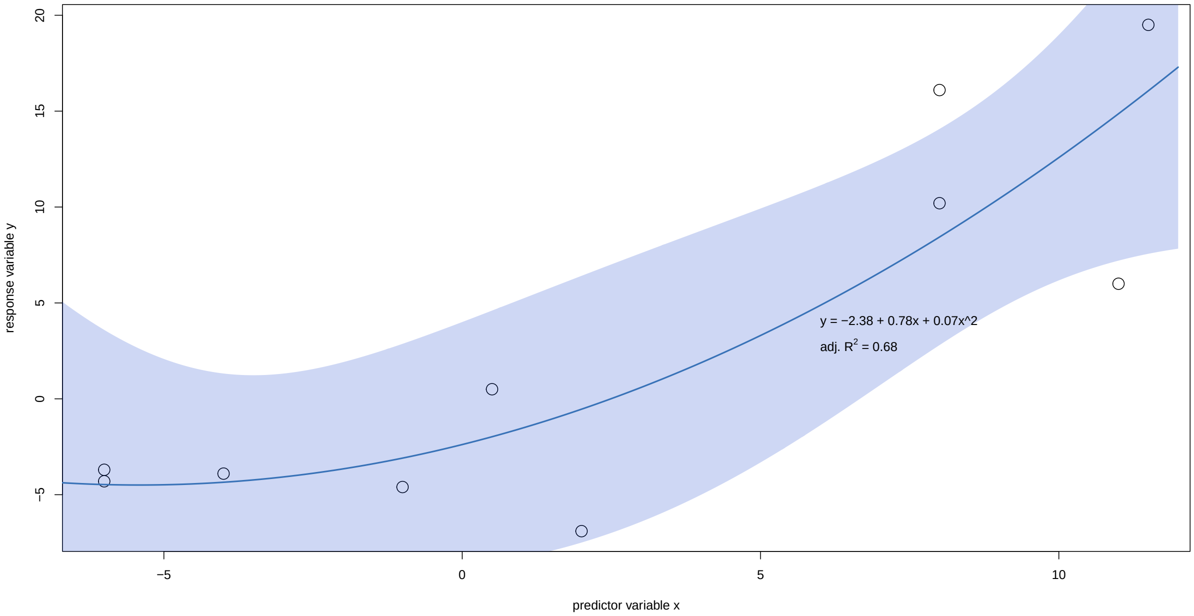
From this we conclude, that both models have similarly good fit, and in such a case we would usually prefer the more parsimonious (simpler) model. We would therefore prefer the model without the X^2 term.

We can probe this further, byt comparing both models (which are an example of nested models) by using a likelihood ratio test. This confirms (p=0.16) that the inclusion of the X^2 term does not significantly improve model fit.

```
library(lmtest)
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
lrtest(mod2, mod)
## Likelihood ratio test
##
## Model 1: y ~ x + x2
## Model 2: y ~ x
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -29.216
## 2   3 -30.203 -1 1.9741       0.16
```

Looking at model diagnostics for the second model, we can start by inspecting the model fit visually (we already computed various goodness-of-fit metrics above).

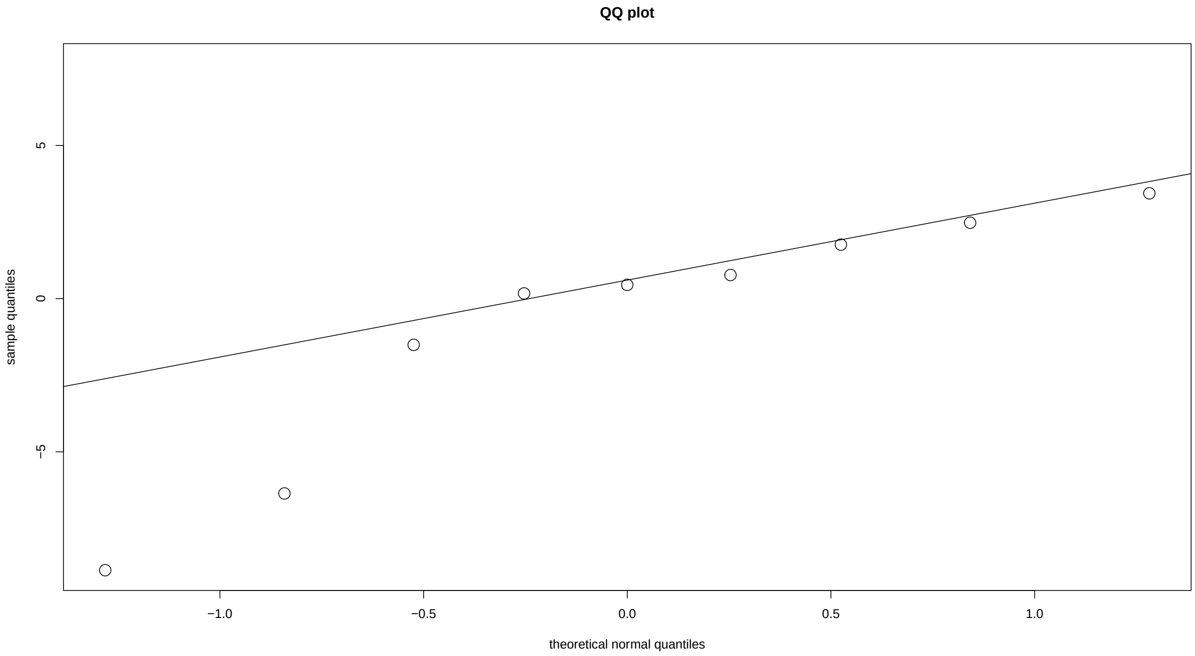
```
beta2<-round(digits=2,coef(mod2))
fit2<-as.data.frame(predict(mod2Lm,newdata=data.frame(x=xx,x2=xx^2),interval="confidence"))

plot(x,y,cex=2,xlab="predictor variable x",ylab="response variable y")
lines(xx,fit2$fit,col="steelblue",lwd=2)
polygon(c(xx,xx[length(xx):1]),c(fit2$lwr,fit2$upr[length(xx):1]),
        border=NA,col=rgb(0,50,200,alpha=50,maxColorValue=255))
text(x=6,y=4,adj=0,
     labels=substitute(paste("y = ",b1," + ",b2,"x + ",b3,"x^2"),list(b1=beta2[1],b2=beta2[2],b3=beta
text(x=6,y=2.75,adj=0,
     labels=substitute(paste("adj. ",R^2," = ",R2adj),list(R2adj=round(R2adj,digits=2))))
```



As previously, difficult to make definite statements with so few datapoints, but compared to earlier this model seems to slightly better capture the initial flat relationship between X and Y, followed by a positive increase at larger X values.

QQ plot

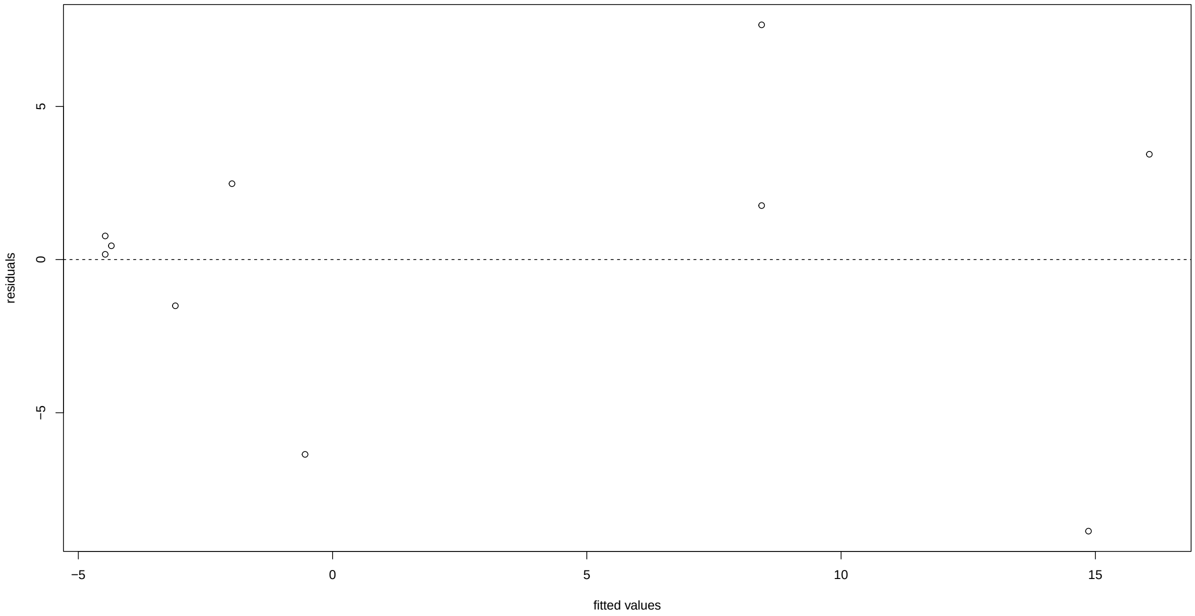
```
r2<-resid(mod2)
theoQ2<-qnorm(order(order(r2))/length(r2))

plot(theoQ2,r2,
     xlab="theoretical normal quantiles",
     ylab="sample quantiles",
     main="QQ plot", cex=2)
qqline(r2) # just adds the line
```



The residuals hug the diagonal even more closely at the upper end, but this just makes the departure from the line at the lower end more marked.

Residuals against predicted values

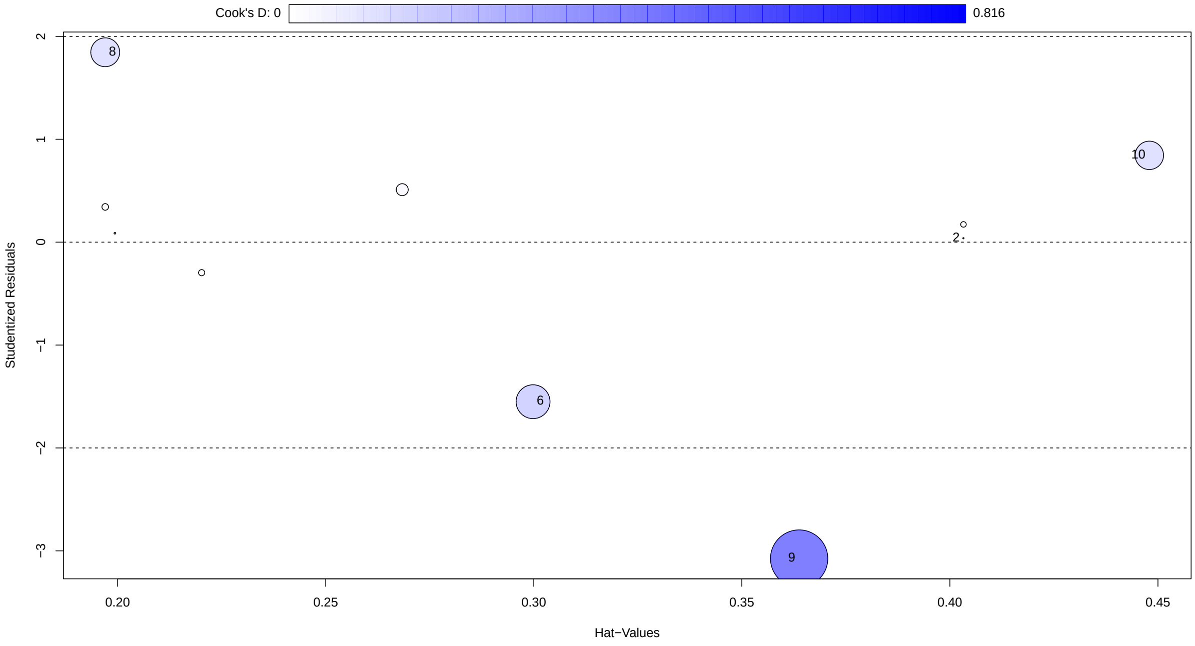
```
yhat2<-predict(mod2)

plot(yhat2,r2,xlab="fitted values",ylab="residuals")
abline(h=0,lty=2)
```



This looks fairly similar to the former model: overall OK, but perhaps more spread around 0 at larger fitted values.

Influential observations

```
influencePlot(mod2)
```



```
   StudRes       Hat       CookD
```

```
2   0.03815863 0.4032611 0.0003825671
6  -1.55046605 0.2998273 0.2858143733
8   1.84565972 0.1970122 0.2073184661
9  -3.07488307 0.3637716 0.8161750794
10  0.84363573 0.4479316 0.2007574248
```

In addition to the same 3 potentially observations, there is now a fourth one: observation 8.

As an overall conclusion, we would recommend the simpler model $Y = \beta_0 + \beta_1 X + \epsilon$ to the model with a term for X^2. Both models fit the data similarly well. One could argue that the more complex model better captures the real relationship between X and Y (steeper increase in Y as X gets larger), but there are too few data points to confirm this. Other model diagnostics are quite similar for both models. Sticking to the principle of parsimony, we recommend the simpler model.

Exercise 5

Download (from GitHub) and load the dataset `cuse.csv`.

This is a dataset on contraceptive use. `using`, `notUsing` lists how many people in each group implied by combinations of `age`, `education`, `wantsMore` are currently using contraceptives. `age`, `education` are self-explanatory. `wantsMore` lists whether individuals want more children or not.

Model the binary variable specified by the 2 columns `using`, `notUsing` in terms of `age`, `education`, `wantsMore`.

- Discuss your results.
- What can you say about the deviance? Does it look like this is a good model?
- What happens if you include an interaction term between the age variable and the desire for more children variable?

Exercise 5 - Solution

```
dat <- read.csv("dataAndSupportDocs/cuse.csv")
mod <- glm(cbind(using, notUsing) ~ age + education + wantsMore, data = dat, family = binomial)

summary(mod)
##
## Call:
## glm(formula = cbind(using, notUsing) ~ age + education + wantsMore,
##     family = binomial, data = dat)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.8082     0.1590  -5.083 3.71e-07 ***
## age25-29        0.3894     0.1759   2.214  0.02681 *
## age30-39        0.9086     0.1646   5.519 3.40e-08 ***
## age40-49        1.1892     0.2144   5.546 2.92e-08 ***
## educationlow   -0.3250     0.1240  -2.620  0.00879 **
## wantsMoreyes   -0.8330     0.1175  -7.091 1.33e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 165.772  on 15  degrees of freedom
## Residual deviance:  29.917  on 10  degrees of freedom
## AIC: 113.43
##
## Number of Fisher Scoring iterations: 4
```

The reference group are people <25 years of age, with high education level and who do not want more children.

For this group, the (average) probability of using contraceptives is given by

$\frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \frac{e^{-0.8082}}{1+e^{-0.8082}} = 0.31.$

Note: $log\left(\frac{p}{1-p}\right) = \beta_0 \Rightarrow p = \frac{e^{\beta_0}}{1+e^{\beta_0}}$

Relative to this reference group:

- Contraceptive use increases for older ager groups: compared to the <25 group, the odds ratio of using contraceptives are $e^{0.3894} = 1.48$, $e^{0.9086} = 2.48$, $e^{1.1892} = 3.28$ for the age groups $25 - 29$, $30 - 39$, $40 - 49$ respectively.
- Having a low level of education decreases contraceptive use ($OR = e^{-0.3250}$=0.72<1)
- Wanting more children decreases contraceptive use ($OR = e^{-0.8330} = 0.43 < 1$)

Note: you can directly compute ORs by typing `exp(coef(mod)[-1])`:

```
print(exp(coef(mod)[-1]))
##     age25-29     age30-39     age40-49 educationlow wantsMoreyes
##    1.4760678    2.4808804    3.2845805    0.7225312    0.4347628
```

You can also compute ORs for groups that involve several of the predictors: e.g. the OR for individuals aged 40-49, with high education and wanting more children is $e^{1.1892-0.833} = 1.43$.

We can even compute probabilities: $P(\text{contraceptive use}|\mathbf{x}) = \frac{e^{^T\mathbf{x}}}{1+e^{^T\mathbf{x}}}$

So for the above group, this is $p = \frac{e^{-0.8082+1.1892-0.833}}{1+e^{-0.8082+1.1892-0.833}} = 0.65$

Coefficients for all variables are statistically significant, telling us that there is evidence that the predictor variables are associated with the response variable.

The deviance of this model is 29.917 on 10 degrees of freedom. This is highly statistically signficant: $p = 8.84 \cdot 10^{-4}$.

```
1-pchisq(29.917,df=10)
```

```
[1] 0.0008838311
```

This means we reject the null hypothesis that our model is no different than the saturated model. There is evidence that we have not explained all the variation in the dataset. Our model does not fit well.

Let's add an interaction term between age and wanting more children.

```
mod2 <- glm(cbind(using, notUsing) ~ age + education + wantsMore + education + age:wantsMore,
    data = dat, family = binomial)
summary(mod2)
##
## Call:
## glm(formula = cbind(using, notUsing) ~ age + education + wantsMore +
##     education + age:wantsMore, family = binomial, data = dat)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.39630    0.29782  -4.688 2.75e-06 ***
## age25-29                 0.65378    0.35700   1.831  0.06705 .
## age30-39                 1.65933    0.32207   5.152 2.58e-07 ***
## age40-49                 1.94120    0.35076   5.534 3.13e-08 ***
## educationlow            -0.34065    0.12577  -2.709  0.00676 **
## wantsMoreyes            -0.06622    0.33071  -0.200  0.84130
## age25-29:wantsMoreyes -0.25918    0.40975  -0.633  0.52704
## age30-39:wantsMoreyes -1.11266    0.37404  -2.975  0.00293 **
## age40-49:wantsMoreyes -1.36167    0.48433  -2.811  0.00493 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 165.77  on 15  degrees of freedom
## Residual deviance:  12.63  on  7  degrees of freedom
## AIC: 102.14
##
## Number of Fisher Scoring iterations: 4
```

Some of the interaction terms we introduced are statistically significant - so it was good to include them in the model.

To interpret the interaction terms: the combined odds ratio (compared to individuals aged <25, having high education and not wanting more children), for being aged 30-39, having high education and wanting more children is $e^{1.6593-0.0662-1.1127} = 1.62$. Without the interaction term this would have been (coefficients from the previous model!) $e^{0.9086-0.8330} = 1.07$.

Our revised model now has a deviance of 12.63 on 7 degrees of freedom, $p = 0.08$. This is no longer statistically significant (though only just) - our model fits much better now.

```
1-pchisq(12.63,df=7)
```

```
[1] 0.08165321
```

Note that the coefficient for the individual term `wantsMore` is no longer statistically significant (as well as one for the age categories).

That is no reason to remove it from the model: the interaction terms would be difficult to interpret without it and also it can still contribute to predicting new data.