NCSU ST 503 HW 6

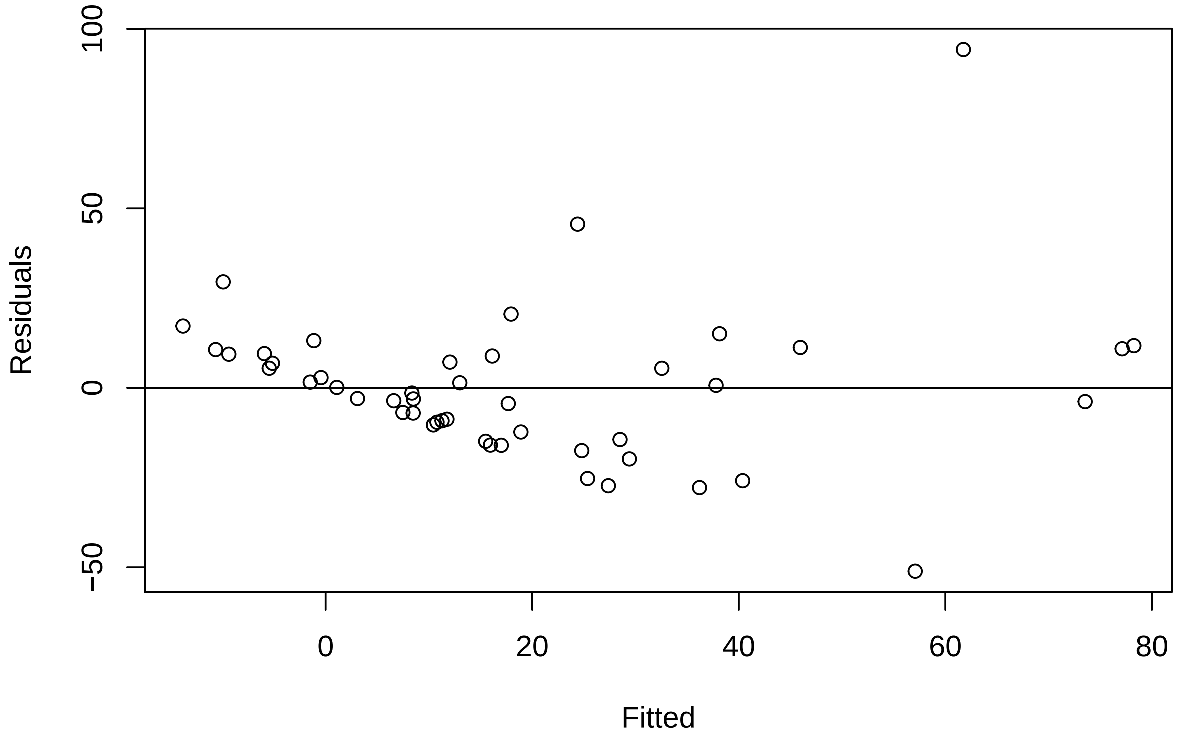
Probems 6.2,6.3,6.4,6.5,6.8 Faraway, Julian J. Linear Models with R, Second Edition Chapman & Hall / CRC Press.

Bruce Campbell

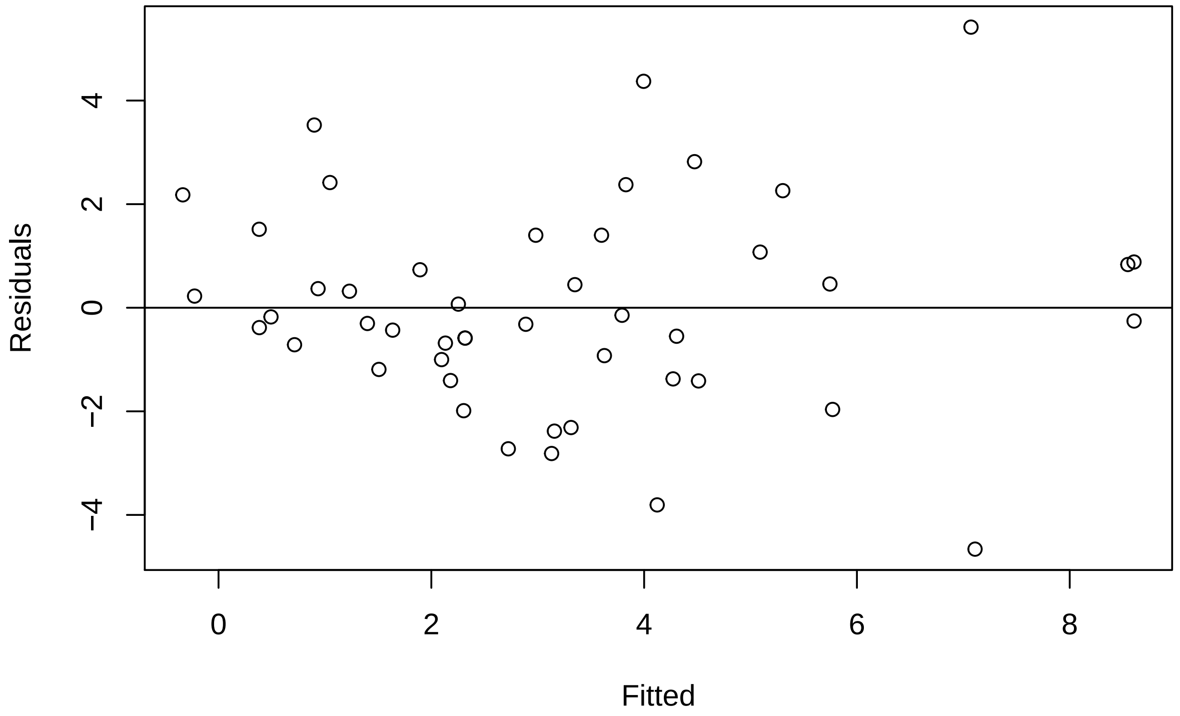
25 September, 2017

6.2 Using the teengamb dataset, fit a model with gamble as the response and the other variables as predictors.

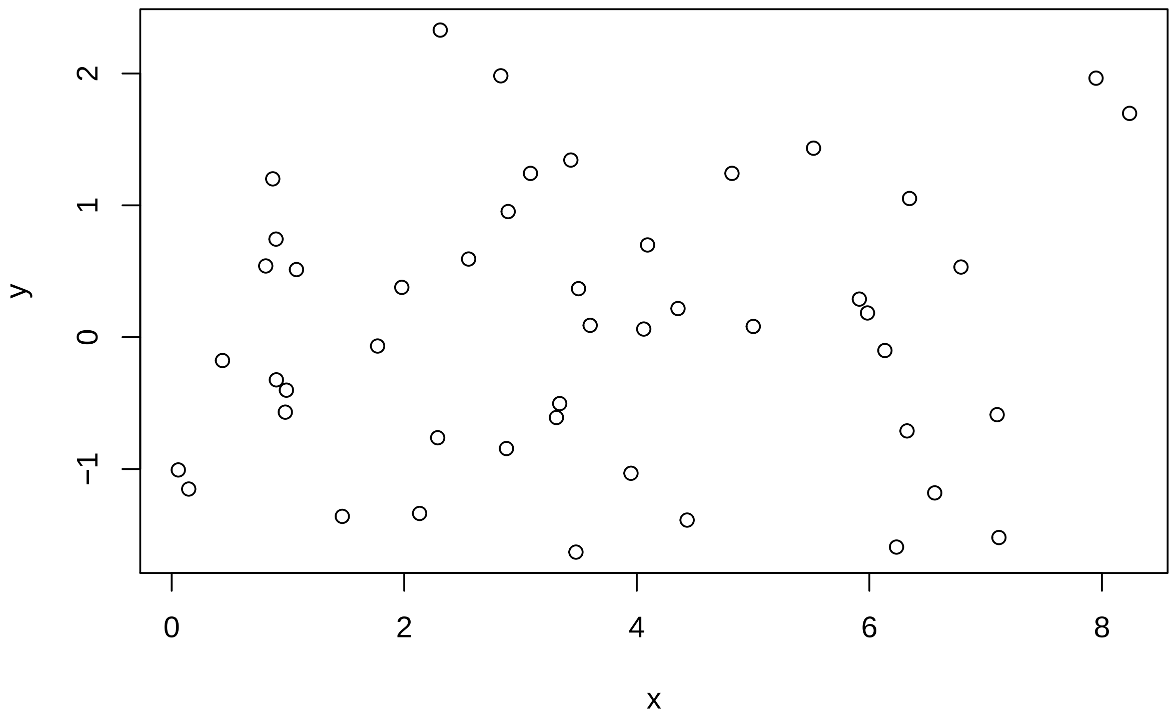
(a) Check the constant variance assumption for the errors.



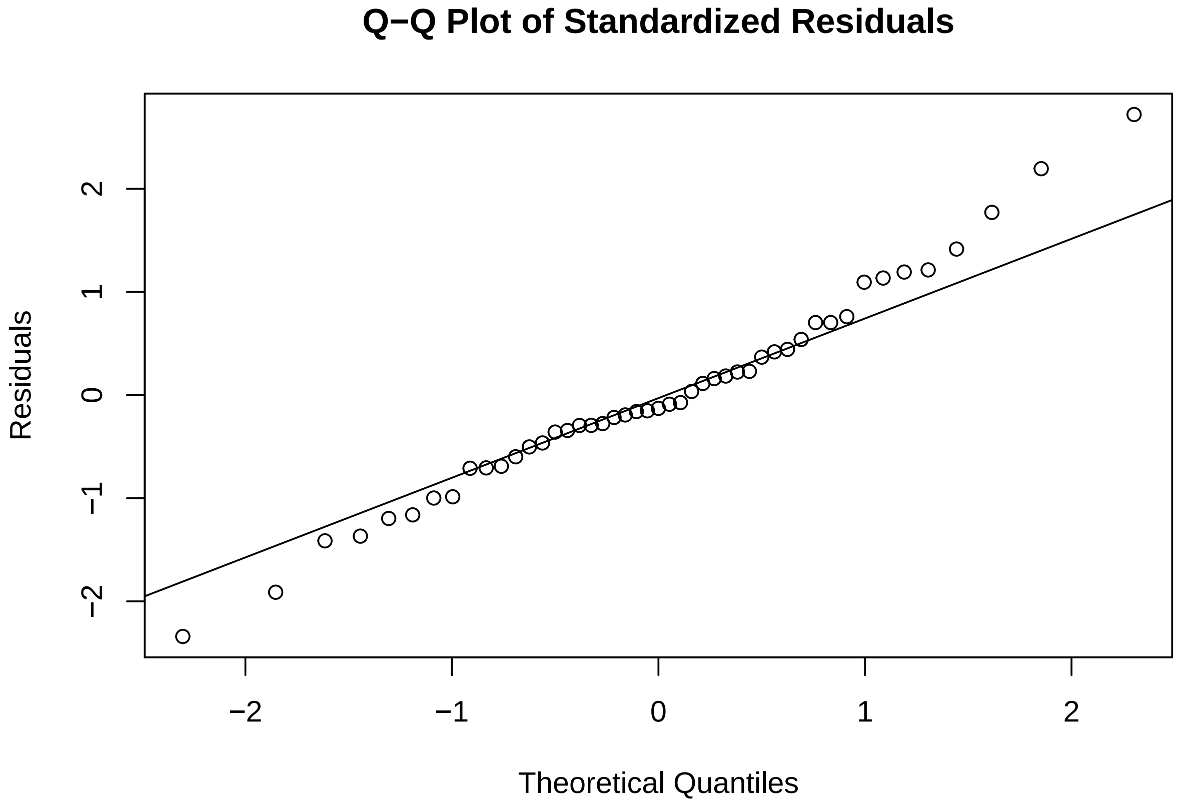
To check the assumption of constant variance we plot fitted values against the residuals - looking for any structure in the distribution of values about the theoretical mean value line $E[\epsilon] = 0$. There appears to be structure and heteroskedasticity in the plot. Below we plot the fitted values against the residuals for a model where the response has been transformed with the square root function. We see less structure and a more evenly distributed variance. We do see even with the transformed response evidence that the variance is not constant.



For reference we plot below what constant $N(0,1)$ error over the same range of the response would look like for the same number of data points. We ran this a number of times to get a good idea of what constant variance looks like with this number of points. It's helpful to calibrate this way when evaluating whether variance is constant for a small and medium data sets.


y
2
1
0
−1
0
2
4
6
8
x

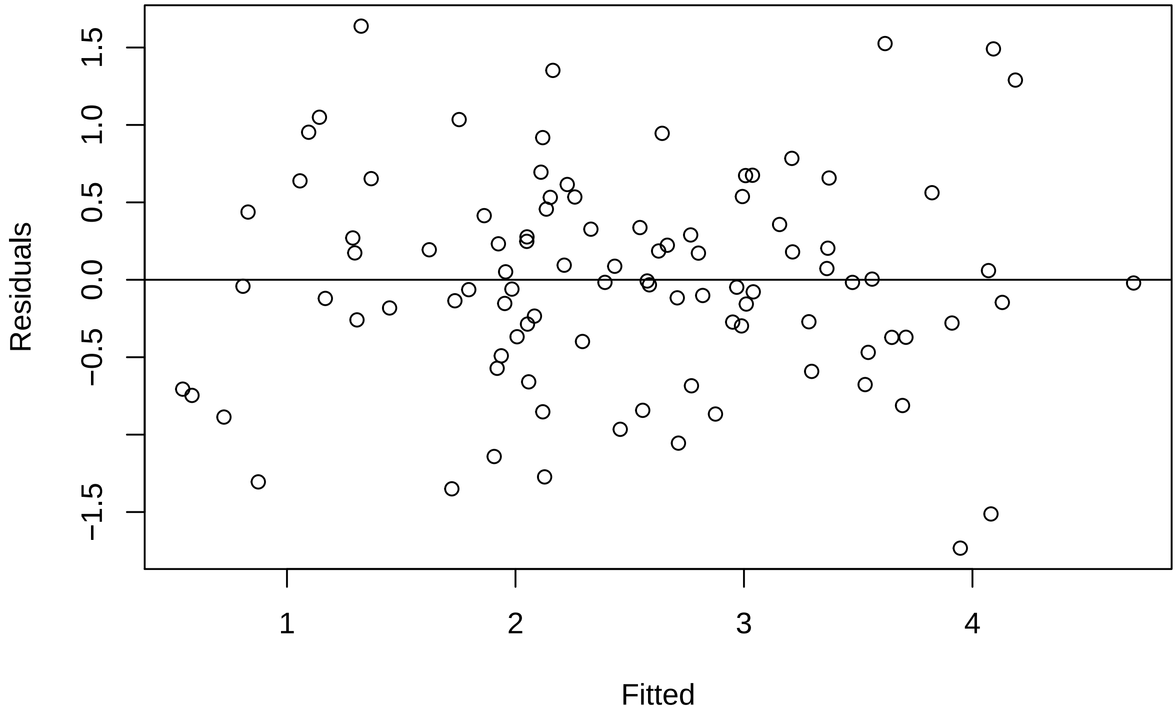
(b) Check the normality assumption.



We see clear evidence of long tails in the distribution of the residuals.

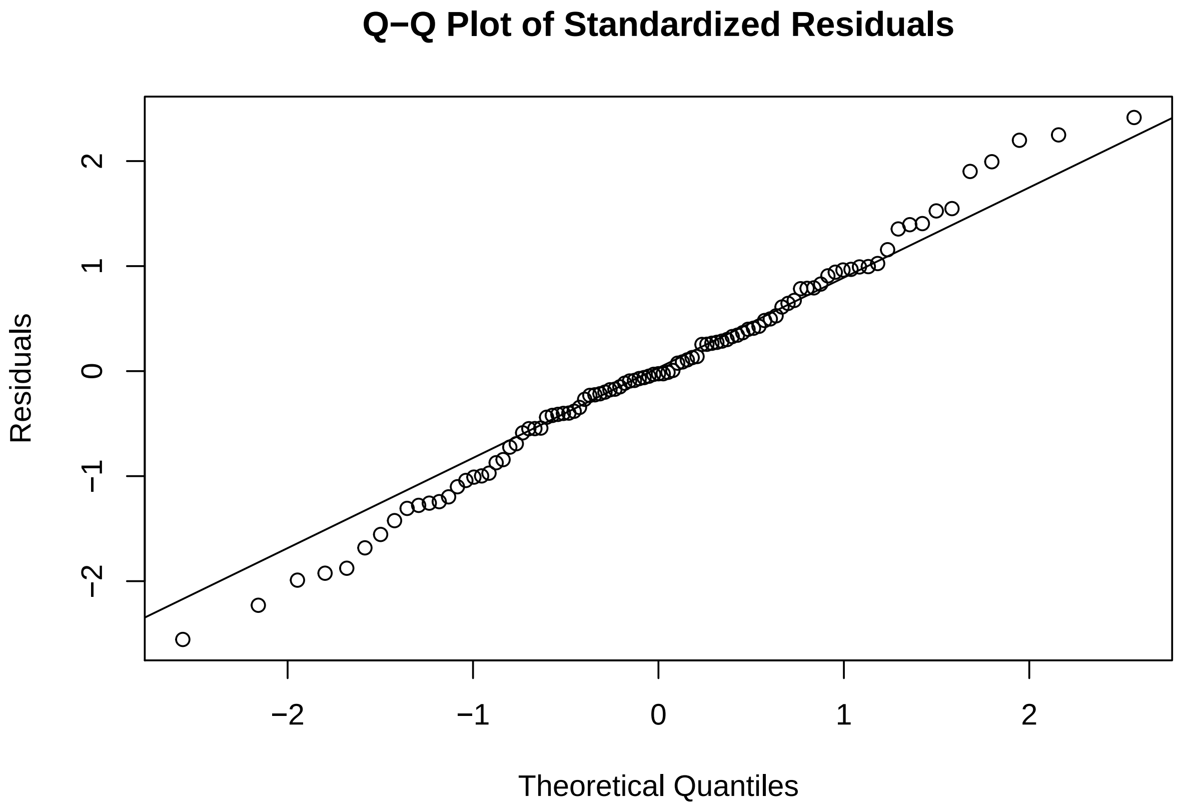
6.3 For the prostate data, fit a model with lpsa as the response and the other variables as predictors.

(a) Check the constant variance assumption for the errors.



The variance of the residuals appears constant over the range of the fitted values. We're comfortable claiming homoskedasticity of residuals for this data set.

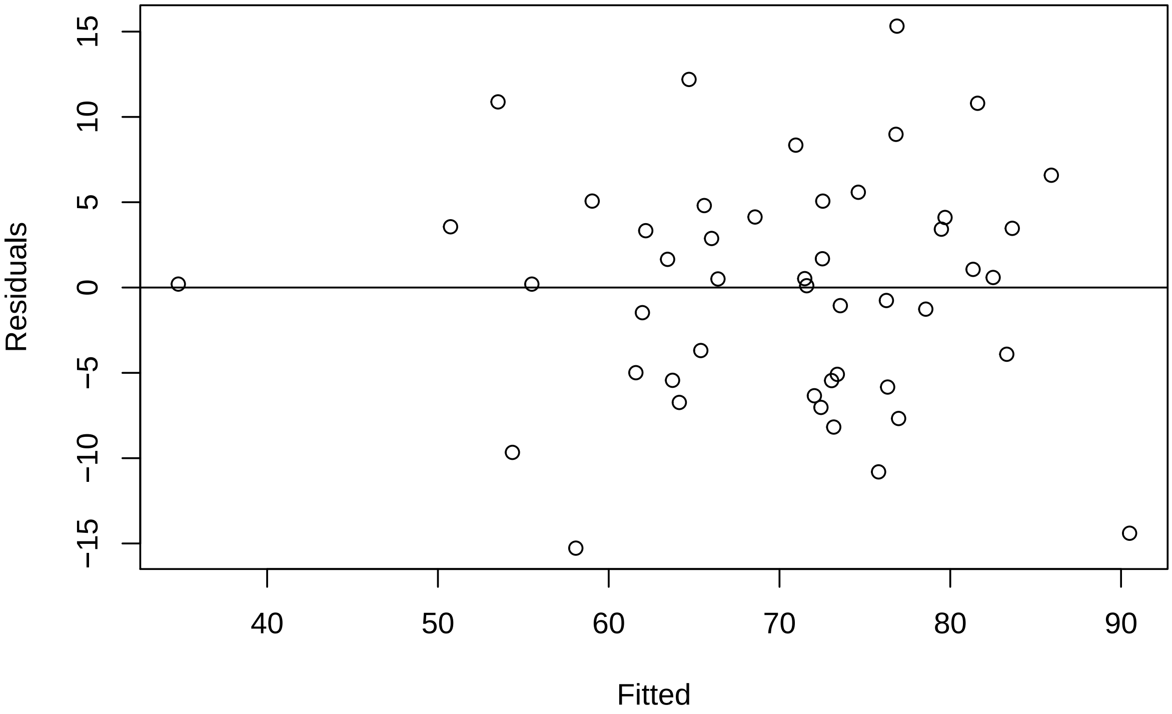
(b) Check the normality assumption.



The studentized residuals appear to be slightly long tailed.

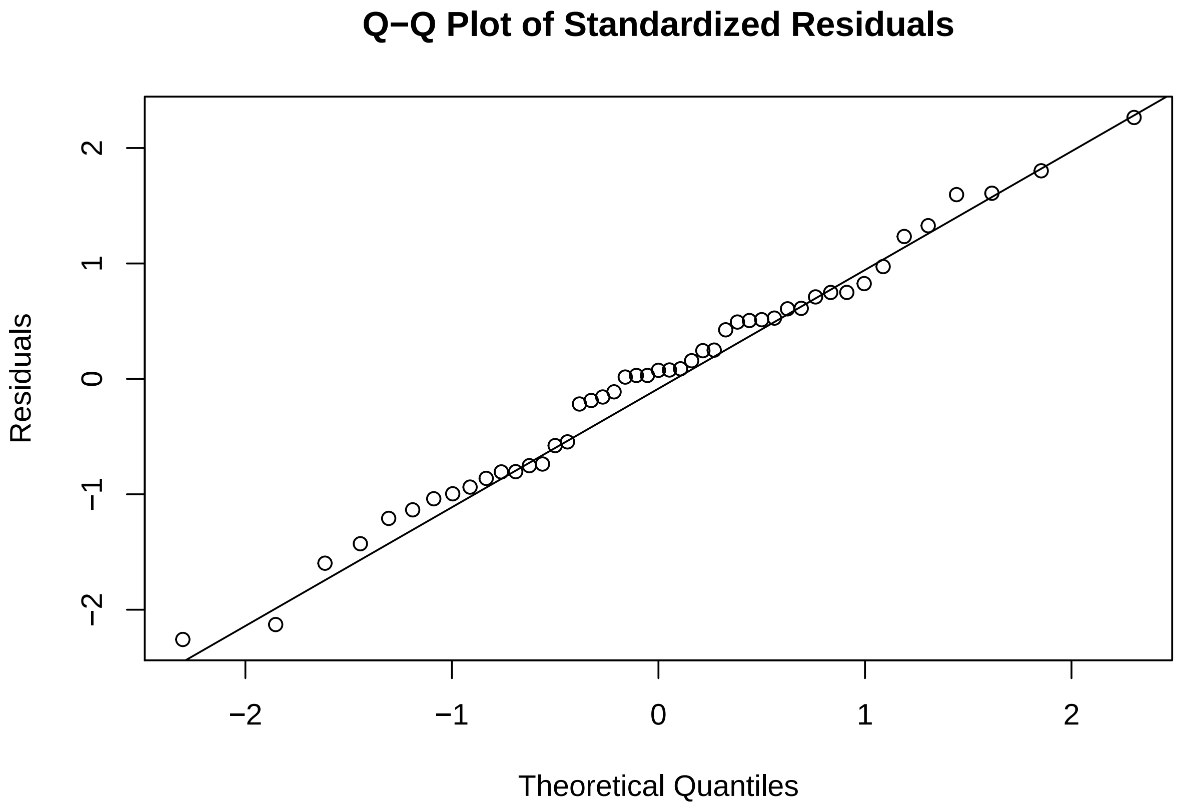
6.4 For the swiss data, fit a model with Fertility as the response and the other variables as predictors.

(a) Check the constant variance assumption for the errors.



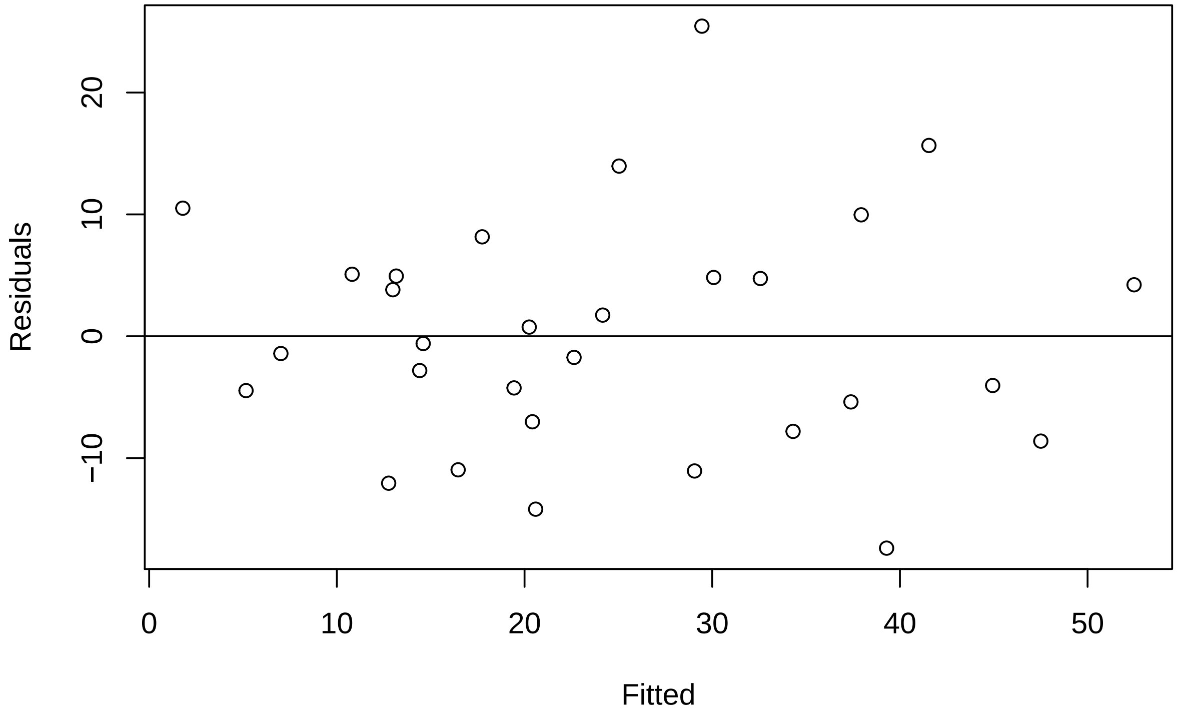
The variance of the residuals appears constant over the range of the fitted values. We're comfortable claiming homoskedasticity of residuals for this data set.

(b) Check the normality assumption.



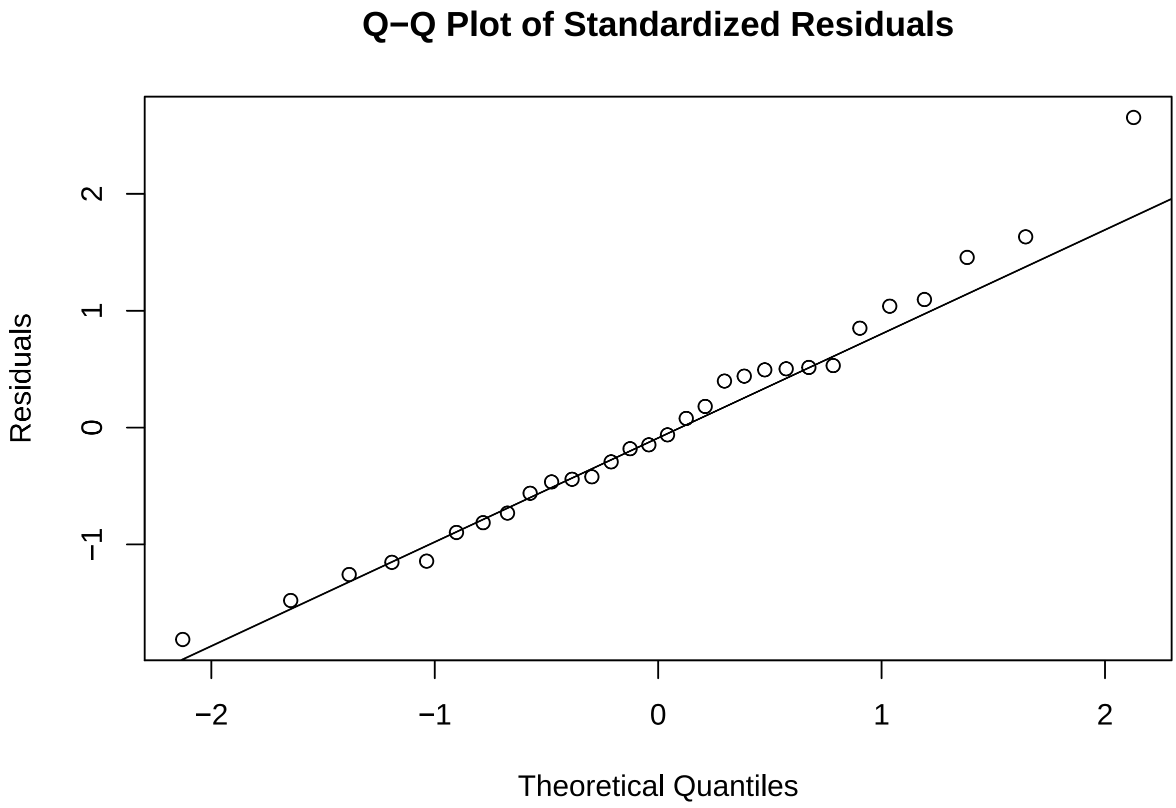
6.5 Using the cheddar data, fit a model with taste as the response and the other three variables as predictors.

(a) Check the constant variance assumption for the errors.



The variance of the residuals appears constant over the range of the fitted values. We're comfortable claiming homoskedasticity of residuals for this data set.

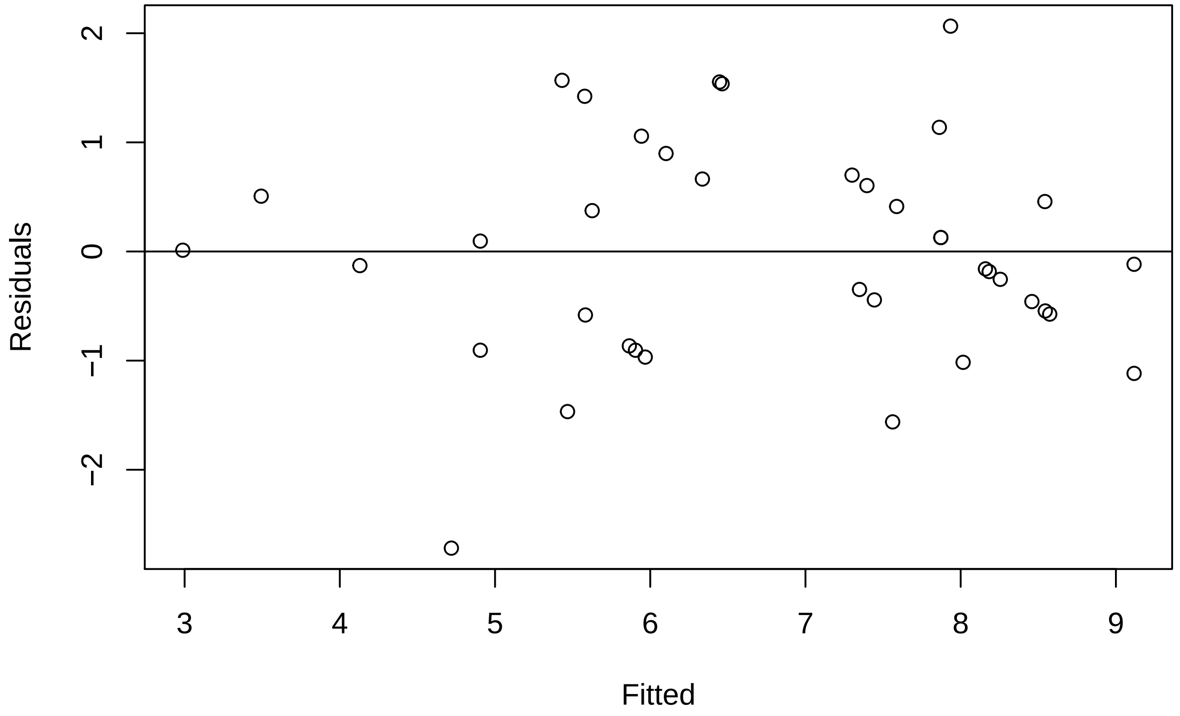
(b) Check the normality assumption.



The standardized residuals appear to be normally distributed.

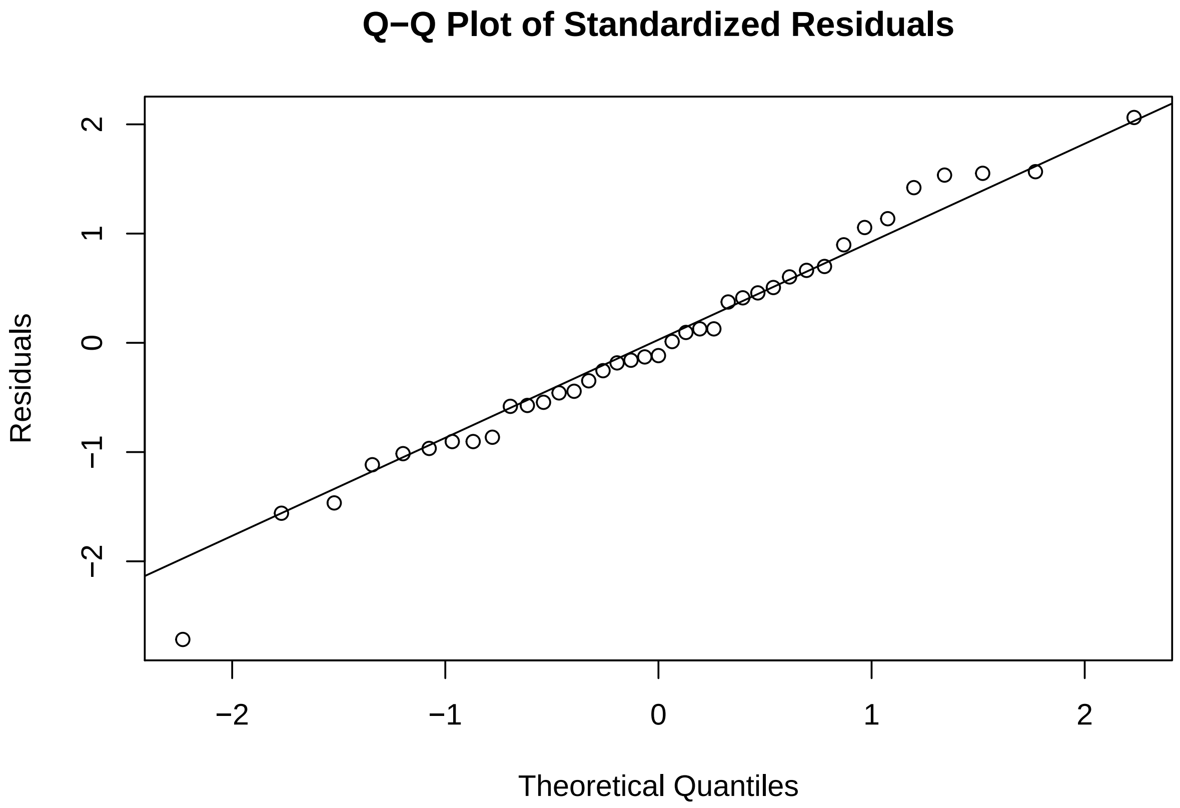
6.6 Using the happy data, fit a model with happy as the response and the other four variables as predictors.

(a) Check the constant variance assumption for the errors.



We see structure in the plot of the residuals. There is serial correlation in the residuals - which is a red flag for our model. The variance of the residuals appears slightly lower over the low end of the range of the response, and higher at the high end of the response. This is difficult to judge though since there are only a few points at the low end of the range of the response.

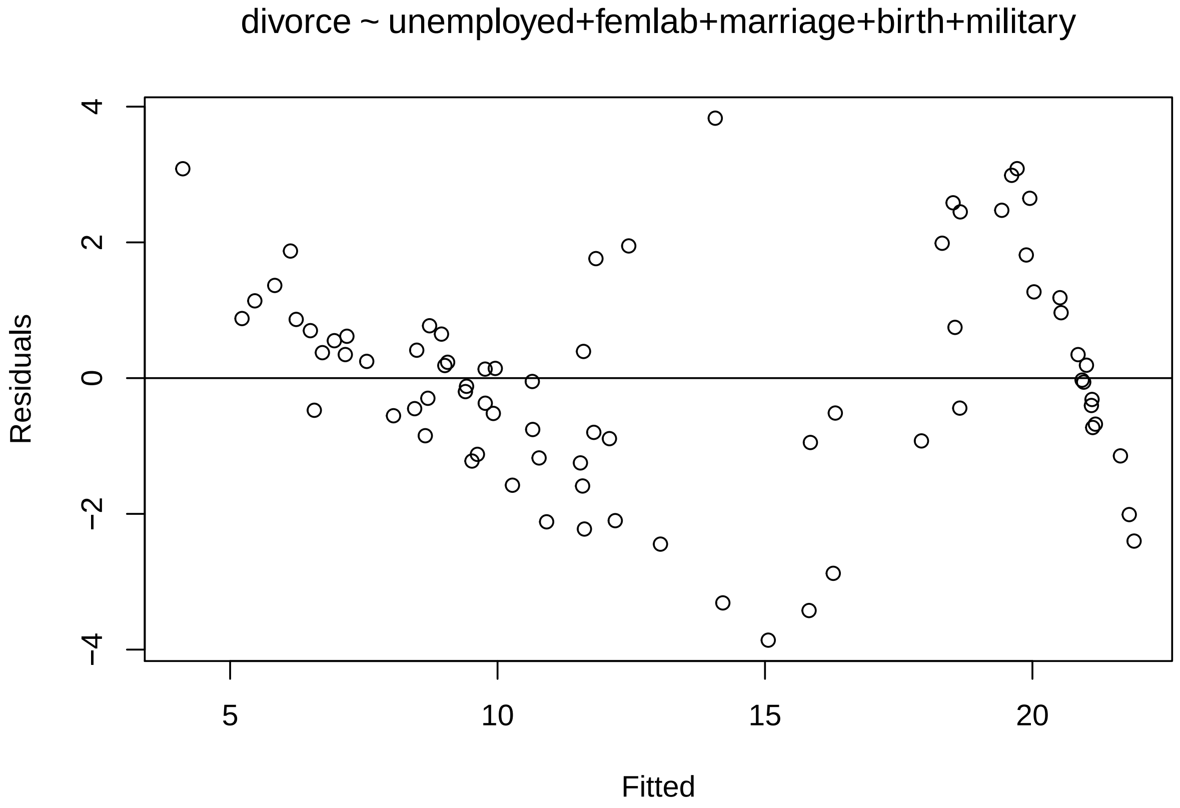
(b) Check the normality assumption.



The standardized residuals appear to be normally distributed. This is interesting in light of the results from part a).

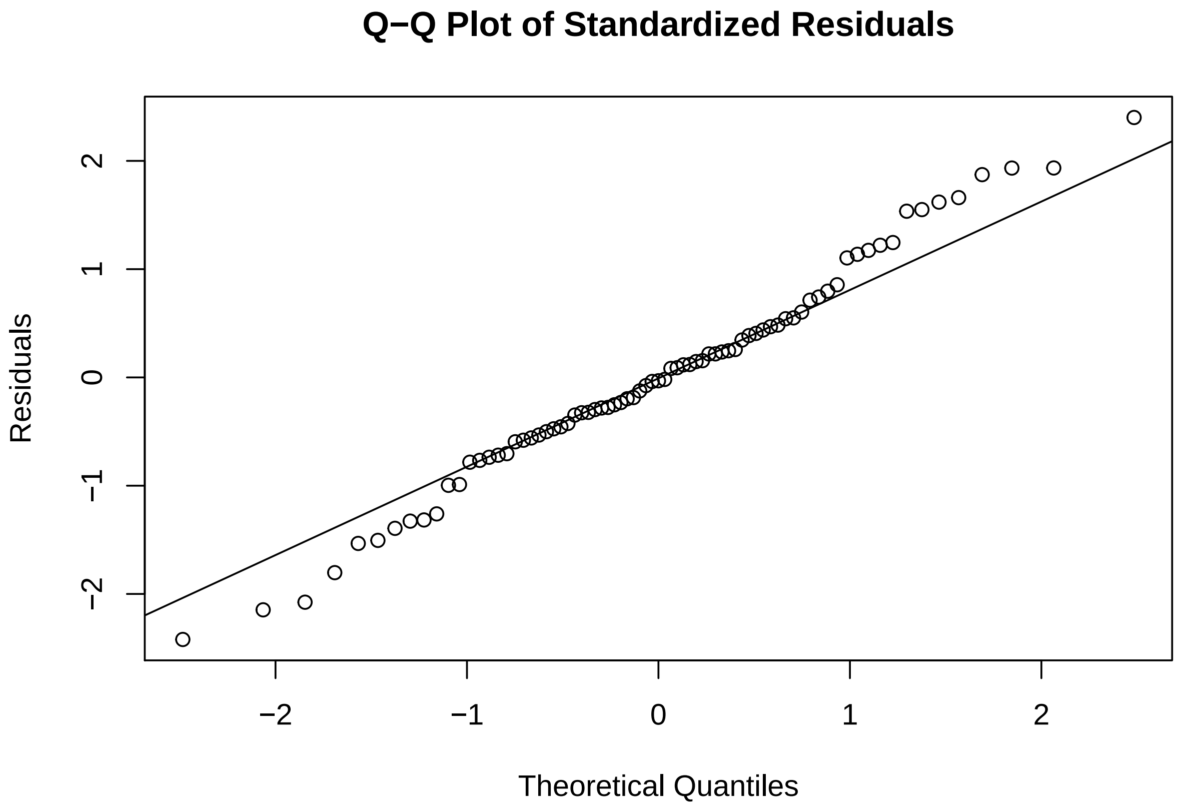
6.8 For the divusa data, fit a model with divorce as the response and the other variables, except year as predictors.

(a) Check the constant variance assumption for the errors.



We see clear structure and serial correlation in the residuals. We may want to plot the response against some of the predictors to look for which ones may be candidates for polynomial terms in the model.

(b) Check the normality assumption.



There is some evidence for mild long tail behavior in the residuals.

(c) Check for large leverage points.

Table 1: High Leverage Data Elements

	year	divorce	unemployed	femlab	marriage	birth	military
13	1932	6.1	23.6	24.46	56	81.7	1.96
14	1933	6.1	24.9	24.89	61.3	76.3	1.94
24	1943	11	1.9	35.7	83	94.3	66.15
25	1944	12	1.2	36.3	76.5	88.8	82.75
26	1945	14.4	1.9	35.8	83.6	85.9	86.64
27	1946	17.9	3.9	30.8	118.1	101.9	21.43

We've used the rule of thumb that points with a leverage greater than $\frac{2p}{n}$ should be looked at.

(d) Check for outliers.

Table 2: Range of Studentized residuals

range.residuals.left	range.residuals.right
-2.447	2.886

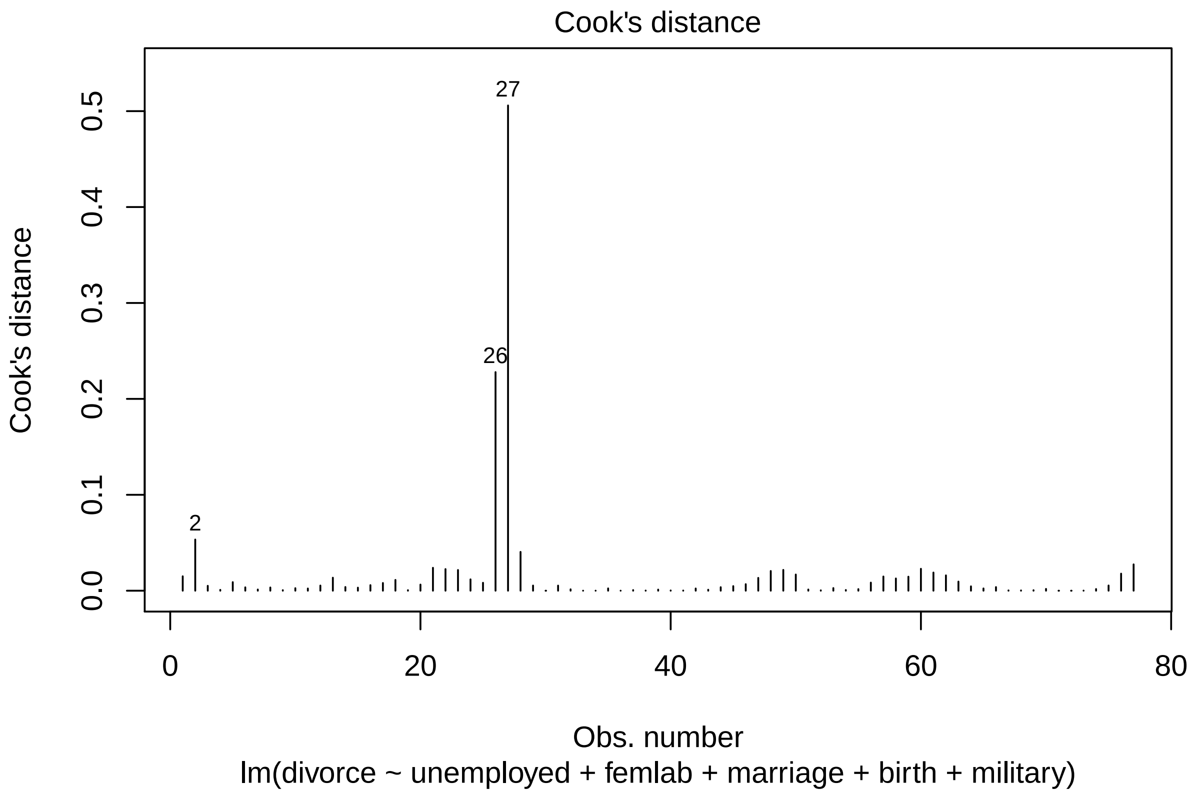
Table 3: Bonferroni corrected t-value

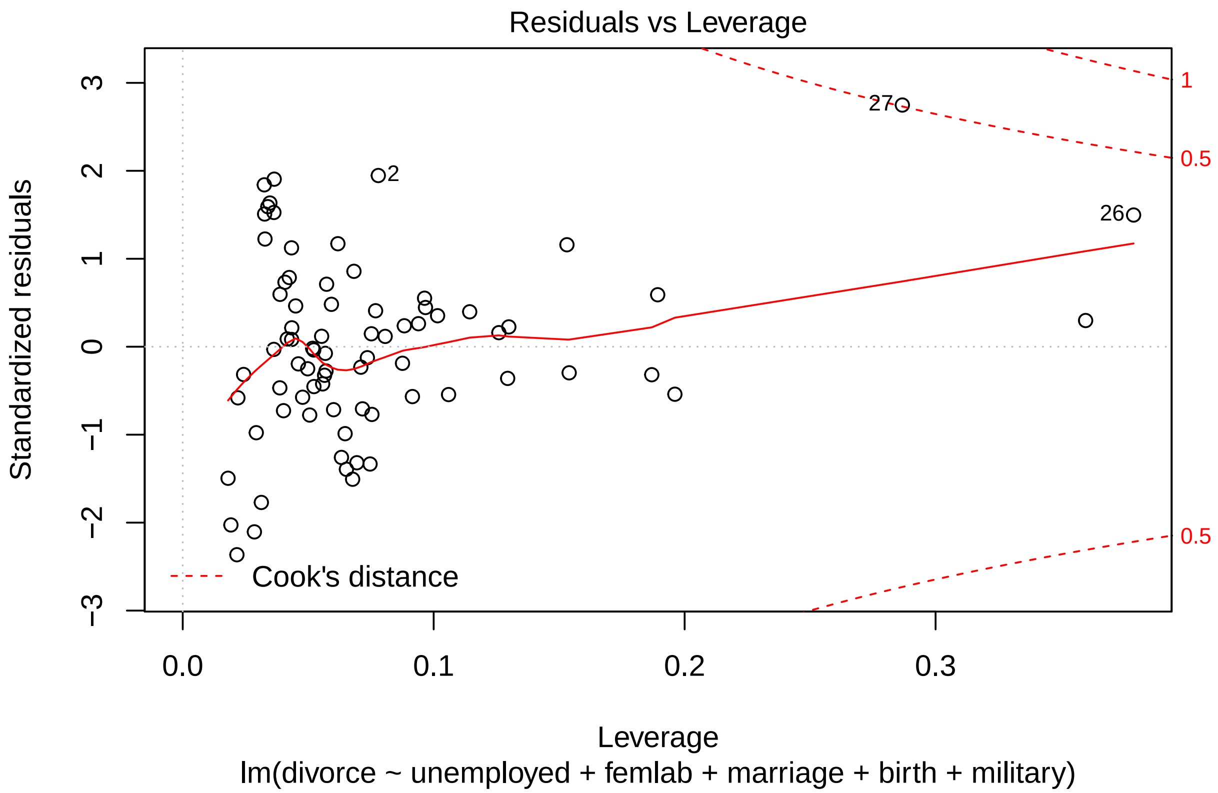
t.val.alpha
-3.57

Since none of the studentized residuals fall outside the interval given by the Bonferroni corrected t-values we claim there are no outliers in the dataset.

(e) Check for influential points.

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.


Cook's distance
Cook's distance
0.0
0.1
0.2
0.3
0.4
0.5
2
26
27
0
20
40
60
80
Obs. number
lm(divorce ~ unemployed + femlab + marriage + birth + military)



We see two clear high leverage points - elements 26 and 27. A third is labelled by R, but the leverage doesn't seem very large. The book does not discuss a criteria for selecting influential points from the Cook distances.

Some guidelines for selecting influential points; * points with a Cook distance more than three times the mean Cook distance
* points with a Cook distance greater than 4/n * points with a cook distance greater than 1

Here we select points with a Cook distance more than three times the mean Cook distance.

Table 4: Mean Cook distance

mean.cooks.distance
0.01712

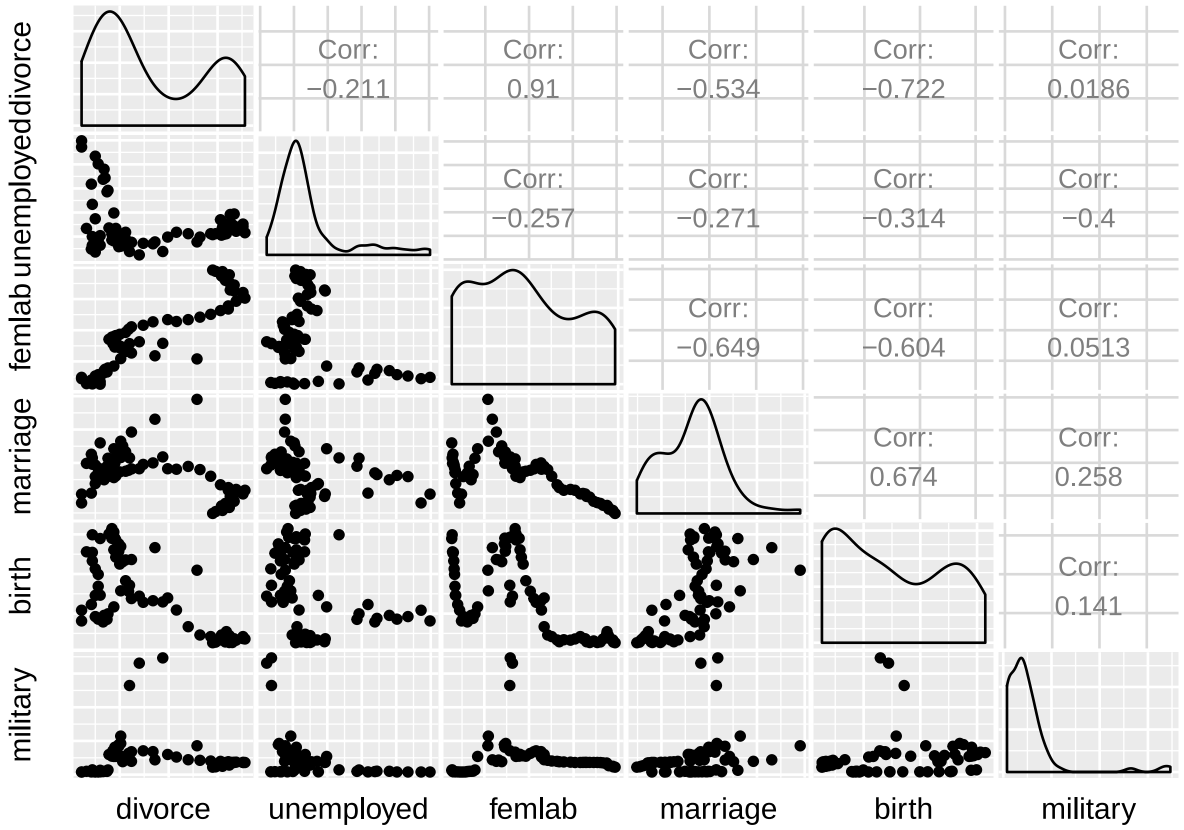
Table 5: Points with Cook distance greater than three times the mean Cook distance.

	cook.distance
2	0.05337
26	0.2279

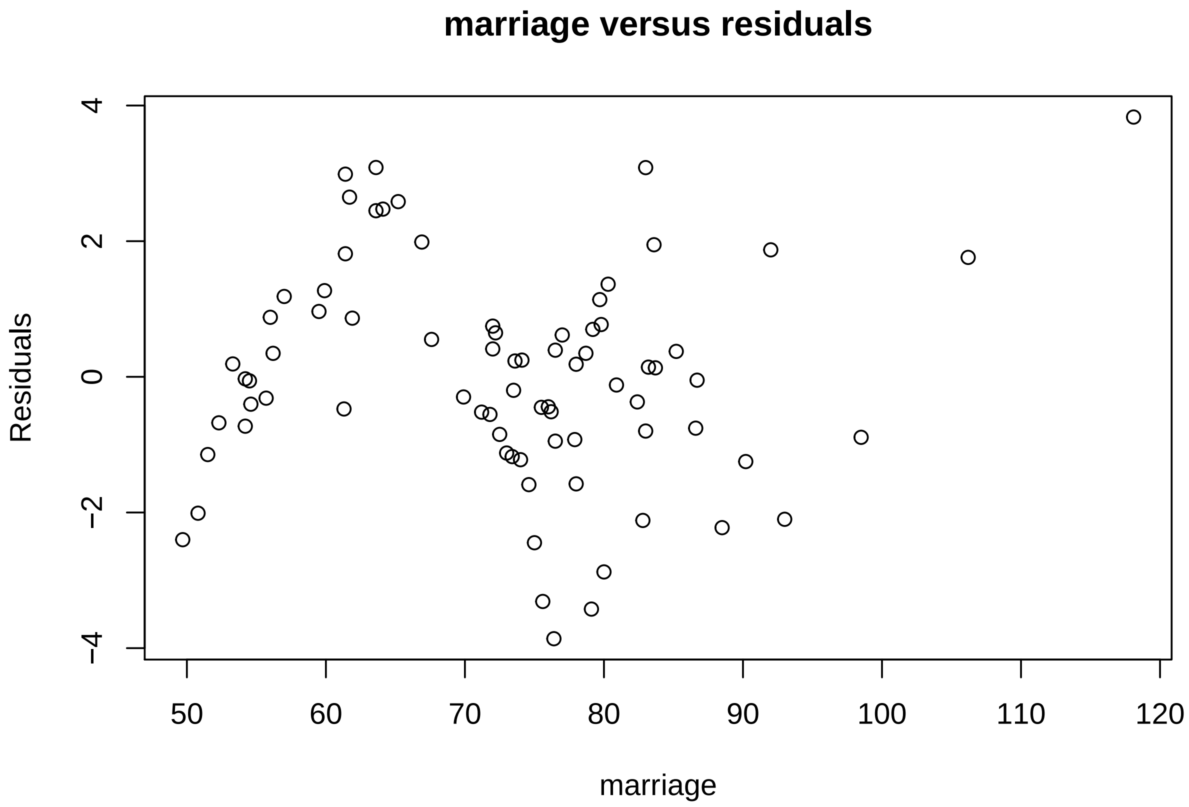
	cook.distance
27	0.5059

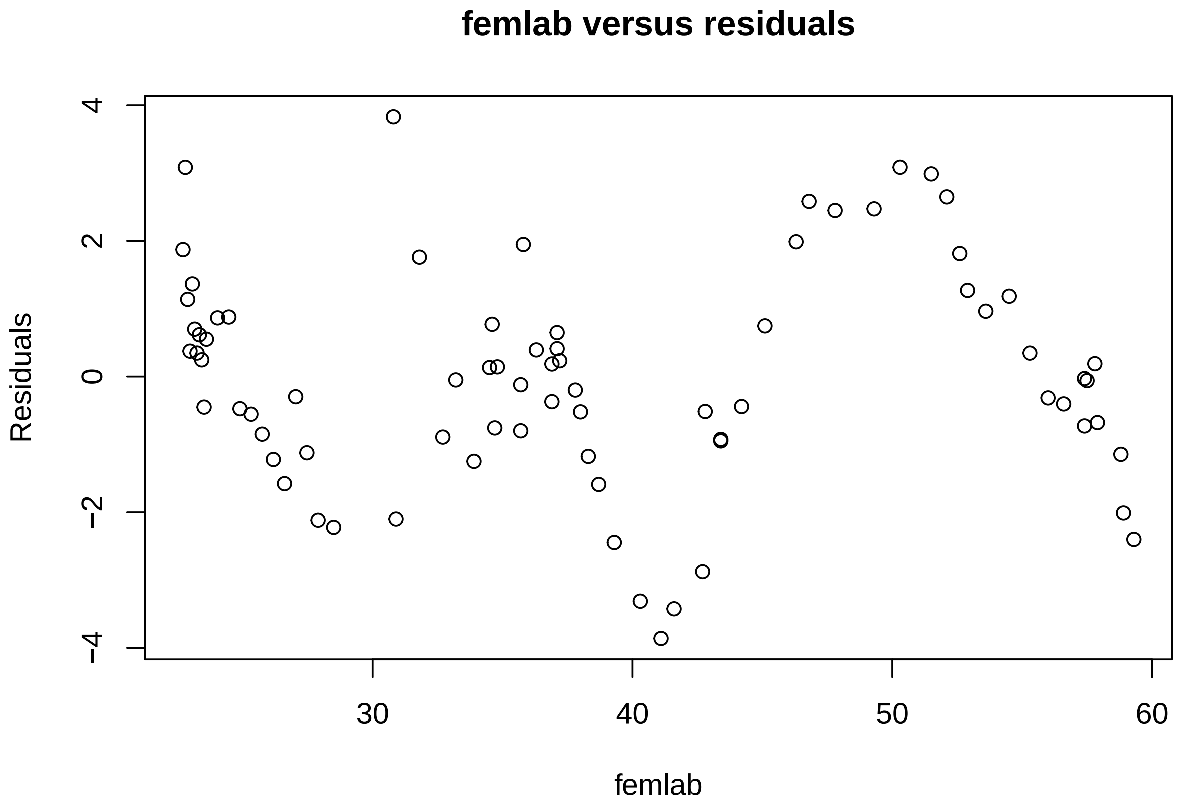
(f) Check for structure in the model.

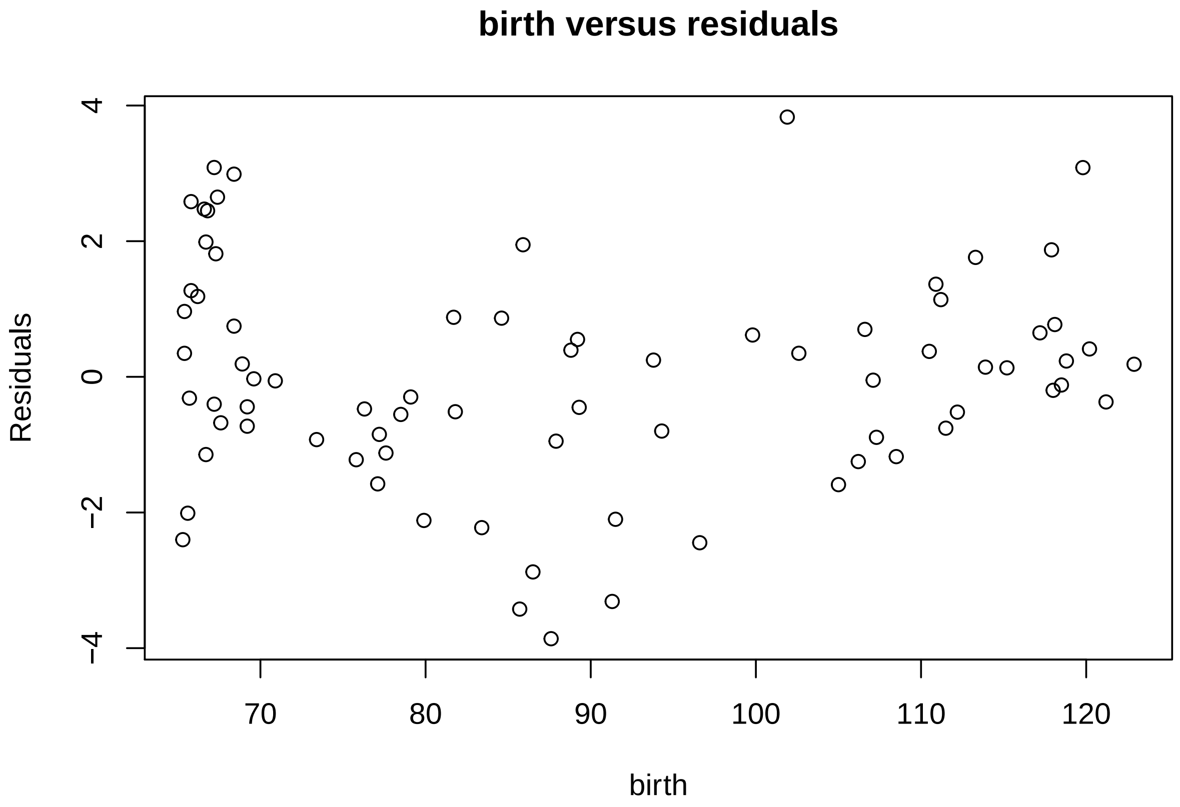
We saw evidence for additional structure not accounted for by the model. First a plot of the variables may help guide the next steps.



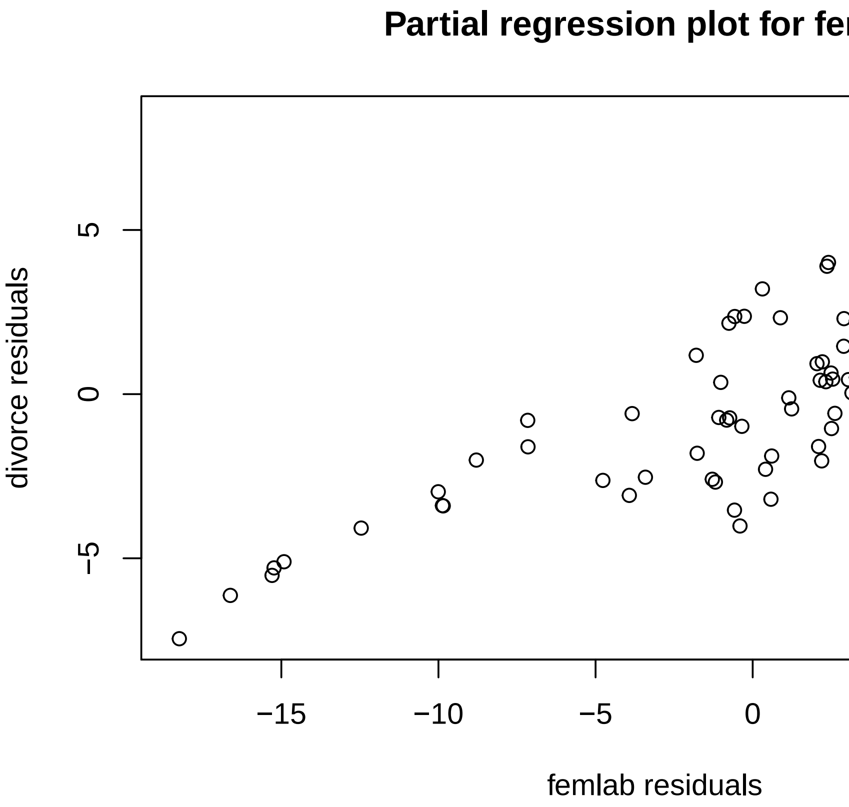
We plotted residuals against all the predictors and found that $femlab$ and $marriage$ had the most structure. These are the likely candidates for including additional terms in the regression. It's apparent that a third order polynomial would be appropriate. We plot *birth* versus residuals because we found out later that adding in polynomial terms for that reduced the structure we saw in the residuals versus fitted plot.


marriage versus residuals
Residuals
4
2
0
−2
−4
50
60
70
80
90
100
110
120
marriage


femlab versus residuals
Residuals
4
2
0
−2
−4
30
40
50
60
femlab

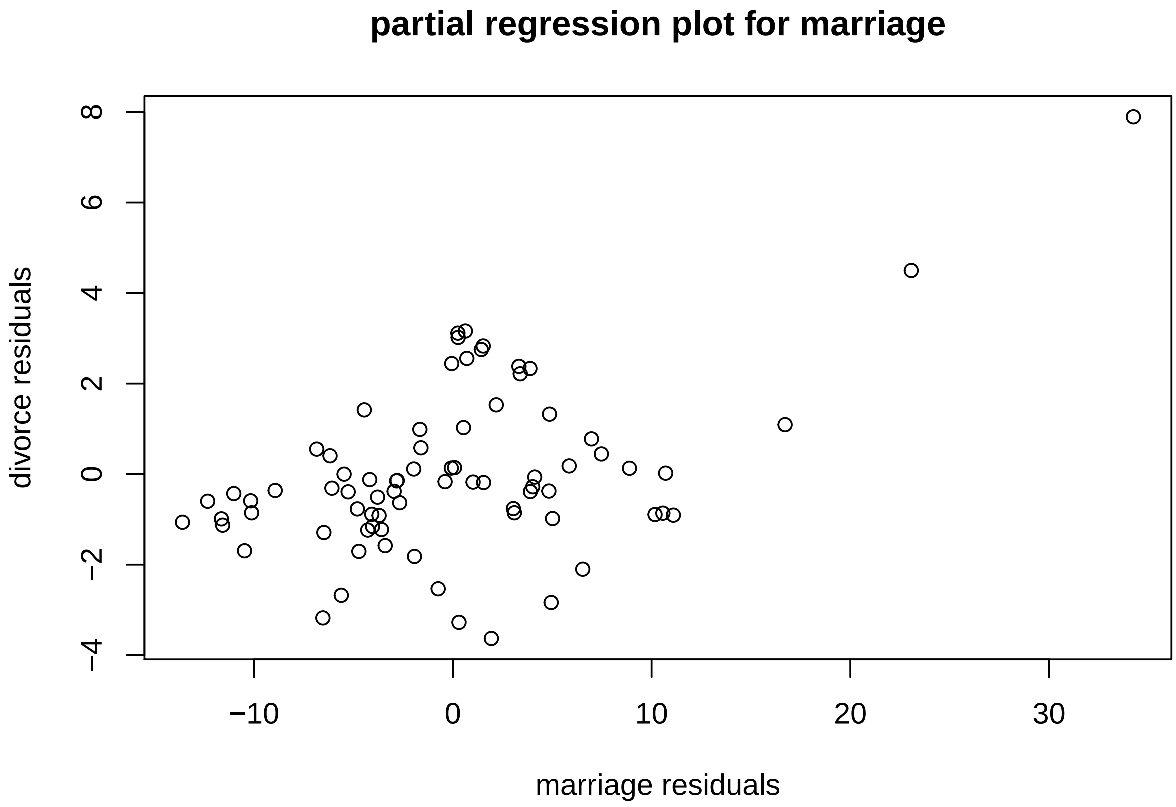


Before we try to remove the unexplained structure let's investigate the partial regression / added variable plot for these variables.

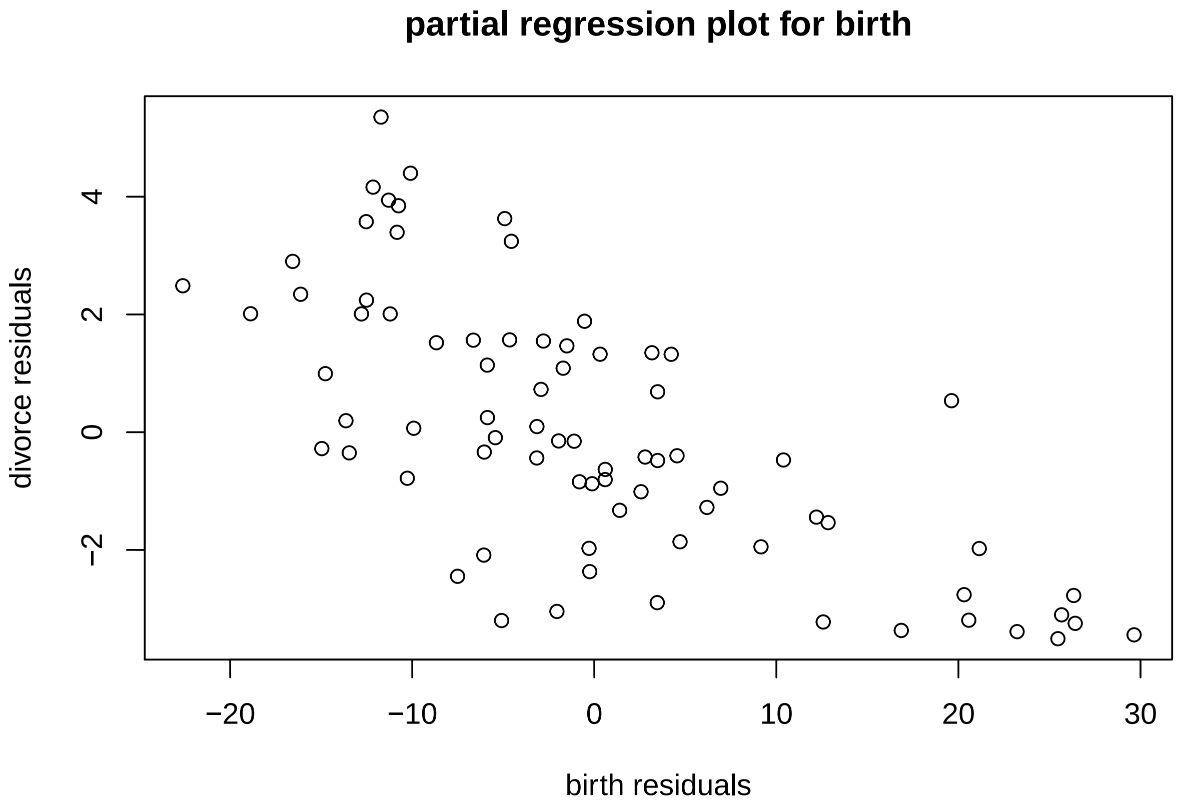


This is the partial regression plot for $femlab$

This is the partial regression plot for $marriage$

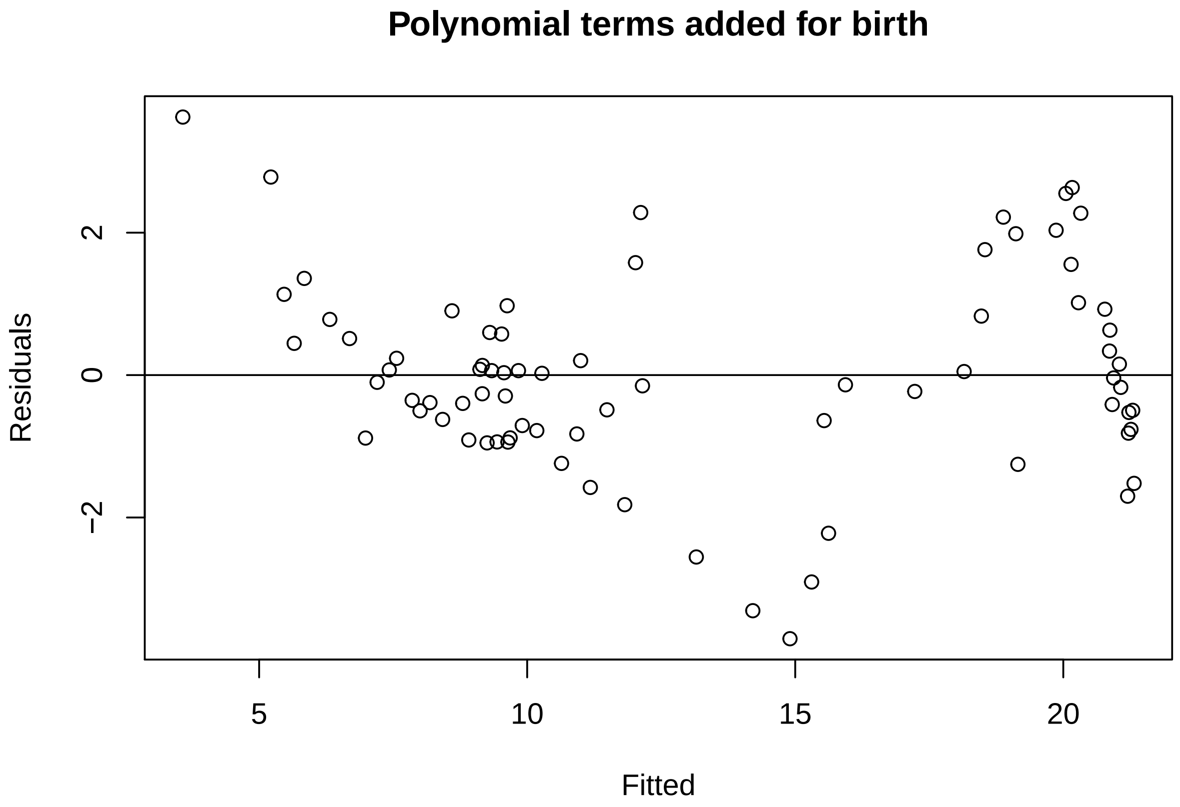


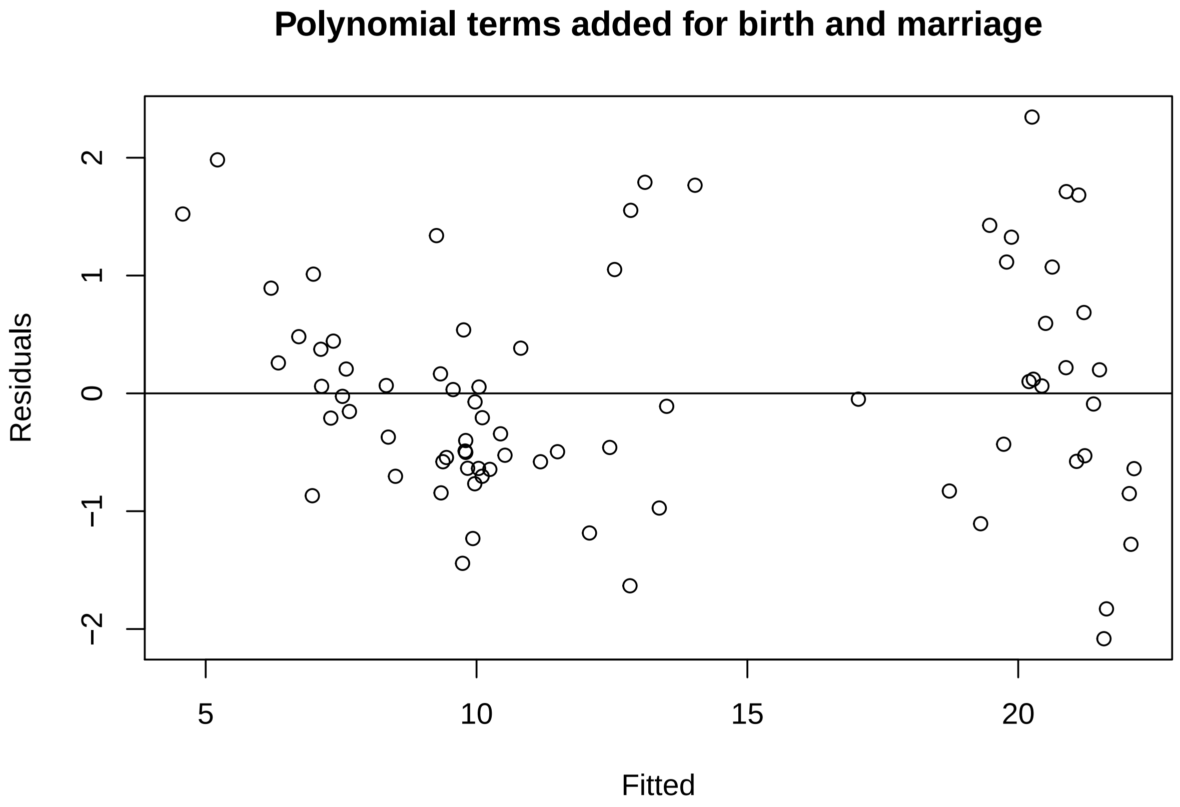
This is the partial regression plot for *birth*



I'm not sure why we don't see non-linearity in these plots. I'll return to the theory behind this and investigate - hopefully before the homework is due! for now let's see if introduction of polynomial terms reduces the structure in the residuals versus fitted plot.

We tried adding in polynomial terms for marriage and femlab. It was not until we added polynomial terms for birth and marriage that the structure in the residuals was reduced. The residuals versus fitted for the models with polynomial terms


Polynomial terms added for birth
Residuals
2
0
−2
5
10
15
20
Fitted


Polynomial terms added for birth and marriage
Residuals
2
1
0
−1
−2
5
10
15
20
Fitted