NCSU ST 503 HW 7

Probems 8.1, 8.6, 8.8 Faraway, Julian J. Linear Models with R, Second Edition Chapman & Hall / CRC Press.

Bruce Campbell

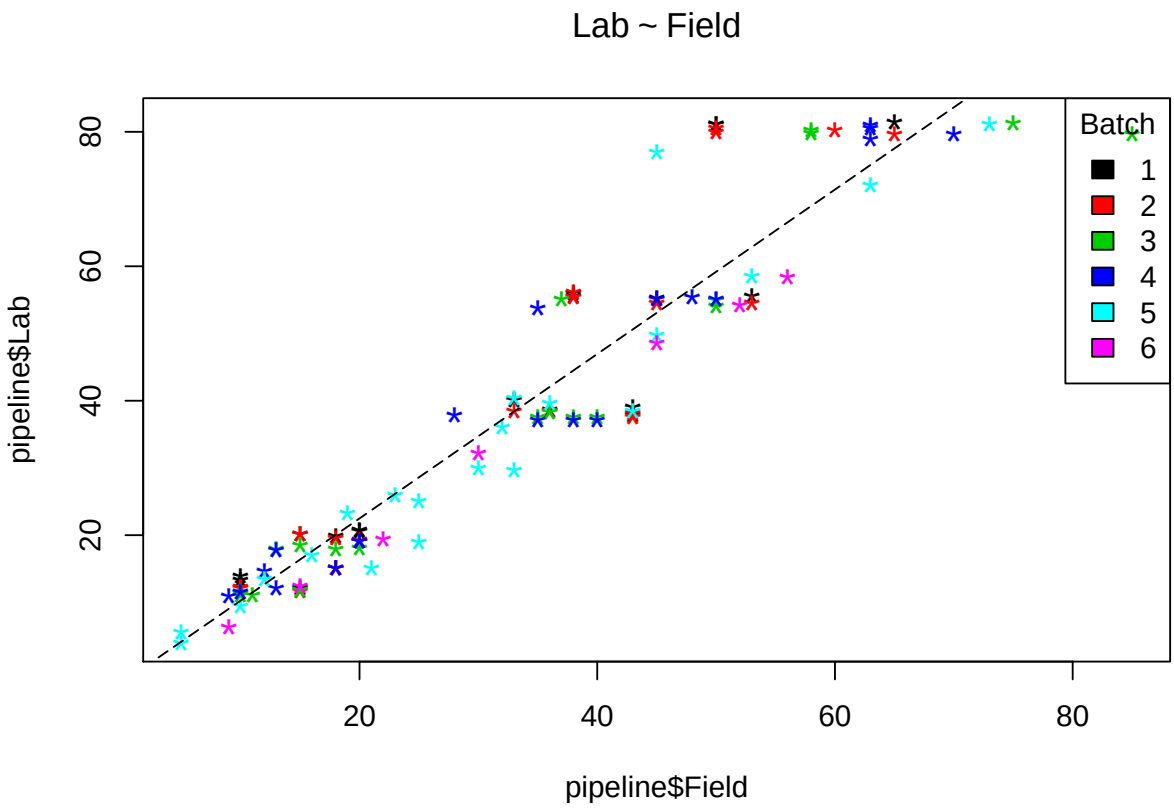
17 October, 2017

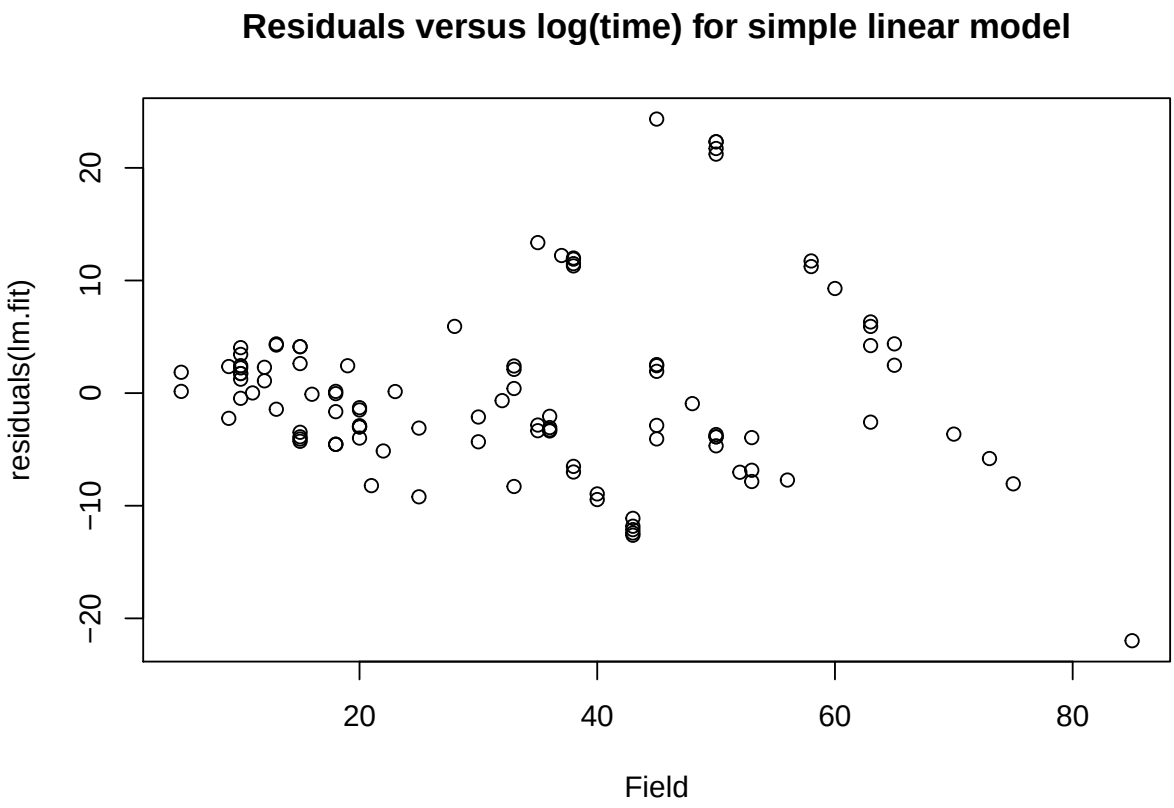
8.1 NIST pipeline Data

Researchers at National Institutes of Standards and Technology (NIST) collected pipeline data on ultrasonic measurements of the depth of defects in the Alaska pipeline in the field. The depth of the defects were then remeasured in the laboratory. These measurements were performed in six different batches. It turns out that this batch effect is not significant and so can be ignored in the analysis that follows. The laboratory measurements are more accurate than the in-field measurements, but more time consuming and expensive. We want to develop a regression equation for correcting the in-field measurements.

(a) Fit a regression model $Lab \sim Field$. Check for non-constant variance.

```
##
## Call:
## lm(formula = Lab ~ Field, data = pipeline)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.985  -4.072  -1.431   2.504  24.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96750    1.57479  -1.249    0.214
## Field        1.22297    0.04107  29.778   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.865 on 105 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8931
## F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```


Lab ~ Field
pipeline$Lab
pipeline$Field
20
40
60
80
Batch
1
2
3
4
5
6

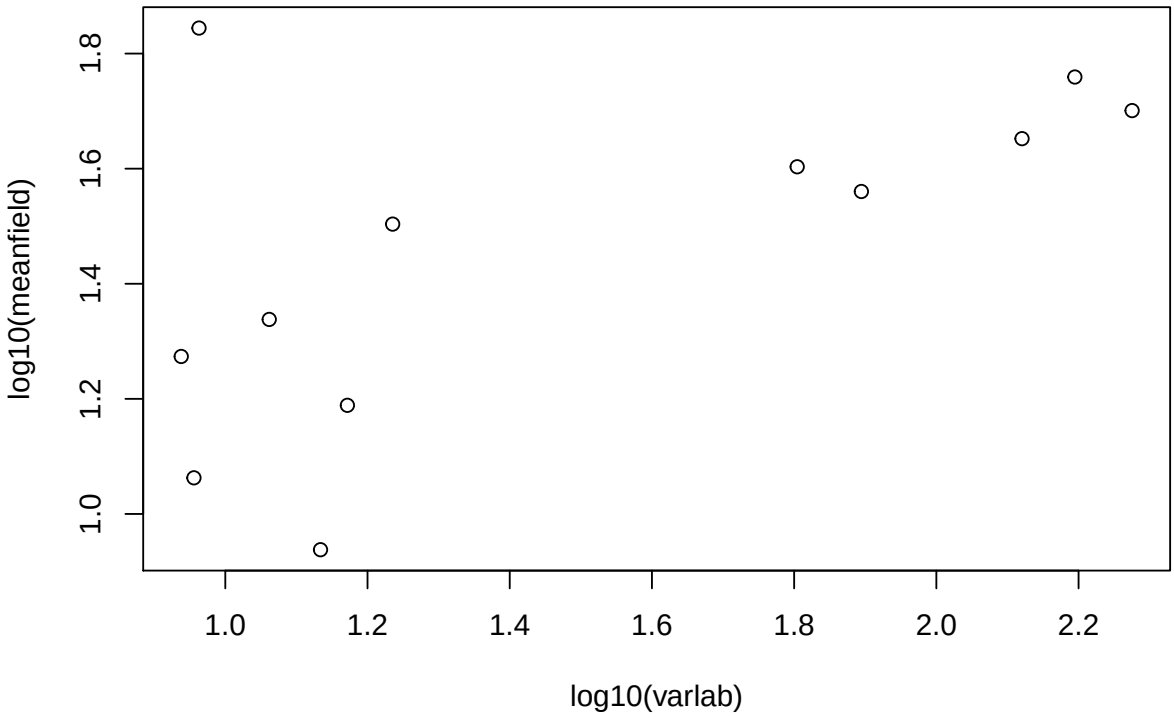


Based on the residual plot we see that we have evidence of non-constant variance. Variance increases with increasing predictor value.

(b) We wish to use weights to account for the non-constant variance.

Here we split the range of Field into 12 groups of size nine (except for the last group which has only eight values). Within each group, we compute the variance of Lab as varlab and the mean of Field as meanfield. Supposing pipeline is the name of your data frame, the following R code will make the needed computations:

Suppose we guess that the the variance in the response is linked to the predictor in the following way: $var(Lab) = a0\ Field^{a1}$ Regress $log(varlab)$ on $log(meanfield)$ to estimate $a0$ and $a1$. (You might choose to remove the last point.) Use this to determine appropriate weights in a WLS fit of Lab on Field. Show the regression summary.

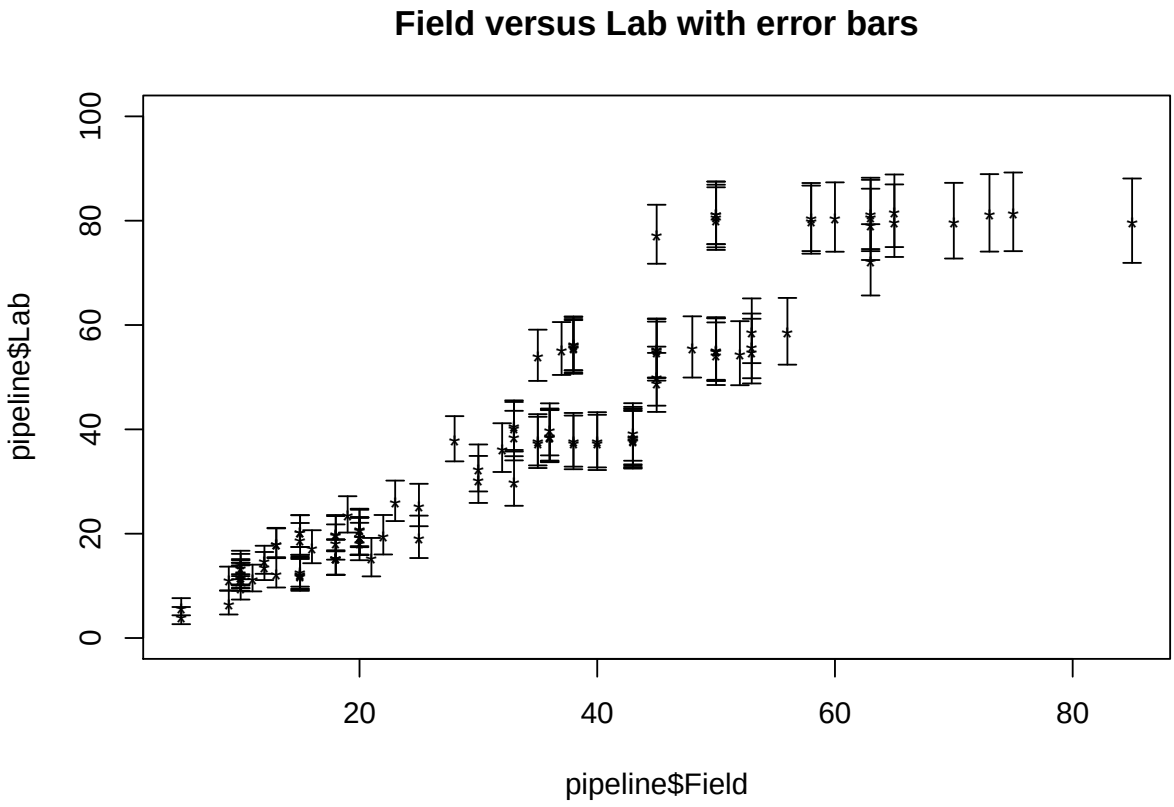



```
##
## Call:
## lm(formula = log(varlab) ~ log(meanfield), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2038 -0.6729  0.1656  0.7205  1.1891
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.3538     1.5715  -0.225   0.8264
## log(meanfield)   1.1244     0.4617   2.435   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 10 degrees of freedom
## Multiple R-squared:  0.3723, Adjusted R-squared:  0.3095
## F-statistic: 5.931 on 1 and 10 DF,  p-value: 0.03513
```

Since $var(Lab) = a0\ Field^{a1}$ we have that $log(var(Lab)) = log(a0) + a1\ log(Field)$ and from the model we fit our estimates of $a0$ and $a1$ are $10^{-0.3538} = 0.4427922$ and 1.1244.

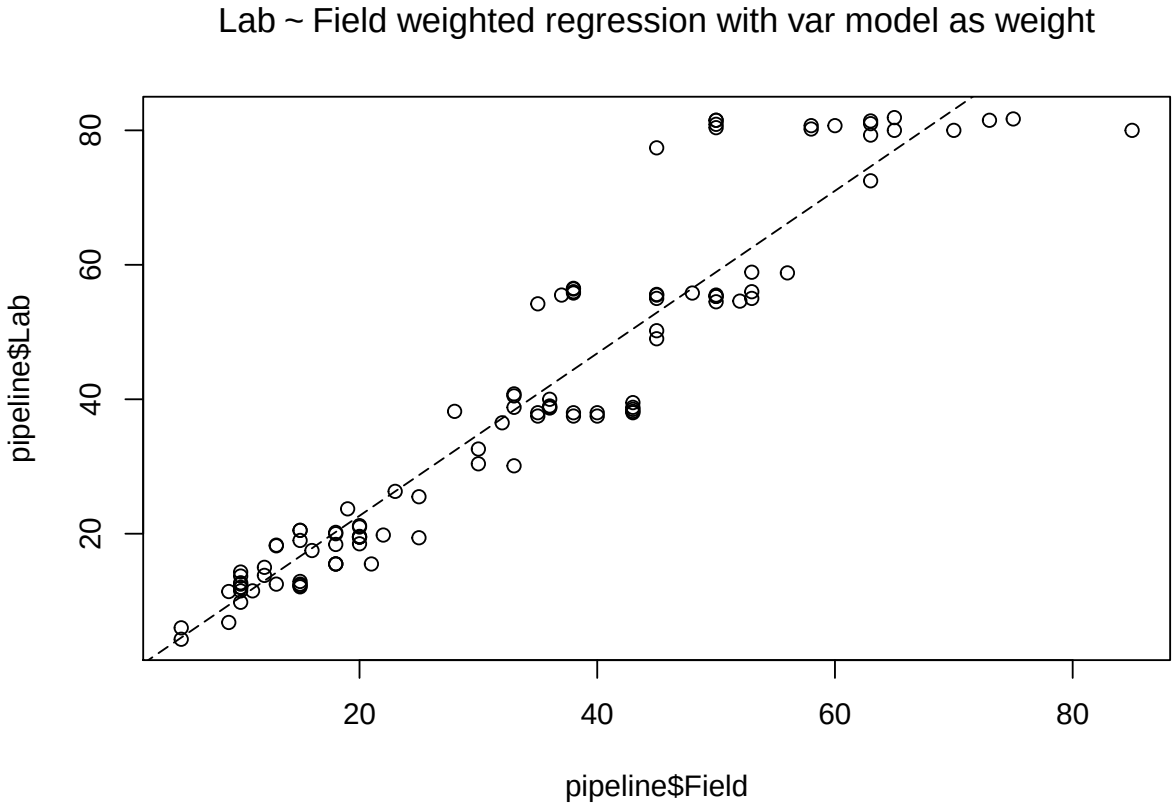
Now we calculate our weight vector with the variances obtained from our model.

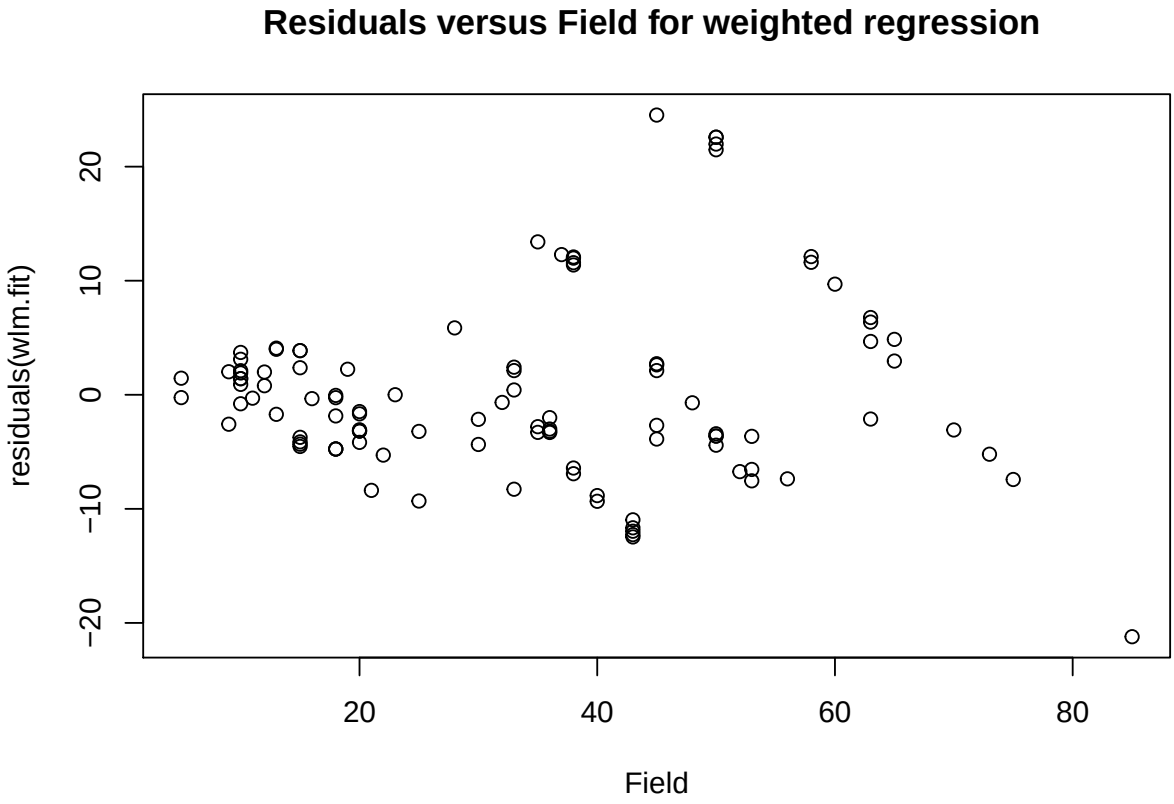
Let plot the data with some error bars from the estimated variances just to make sure everything looks reasonable.




```
## Generalized least squares fit by REML
##   Model: Lab ~ Field
##   Data: pipeline
##        AIC      BIC    logLik
##   708.1764 716.1383 -351.0882
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~var.lab
##
## Coefficients:
##                 Value Std.Error  t-value p-value
## (Intercept) -1.494365 0.9070661 -1.64747  0.1025
## Field        1.208276 0.0348839 34.63704  0.0000
##
##  Correlation:
##       (Intr)
```

```
## Field -0.809
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -1.7814680 -0.6930709 -0.2727923  0.5313567  2.9450736
##
## Residual standard error: 1.47198
## Degrees of freedom: 107 total; 105 residual
```

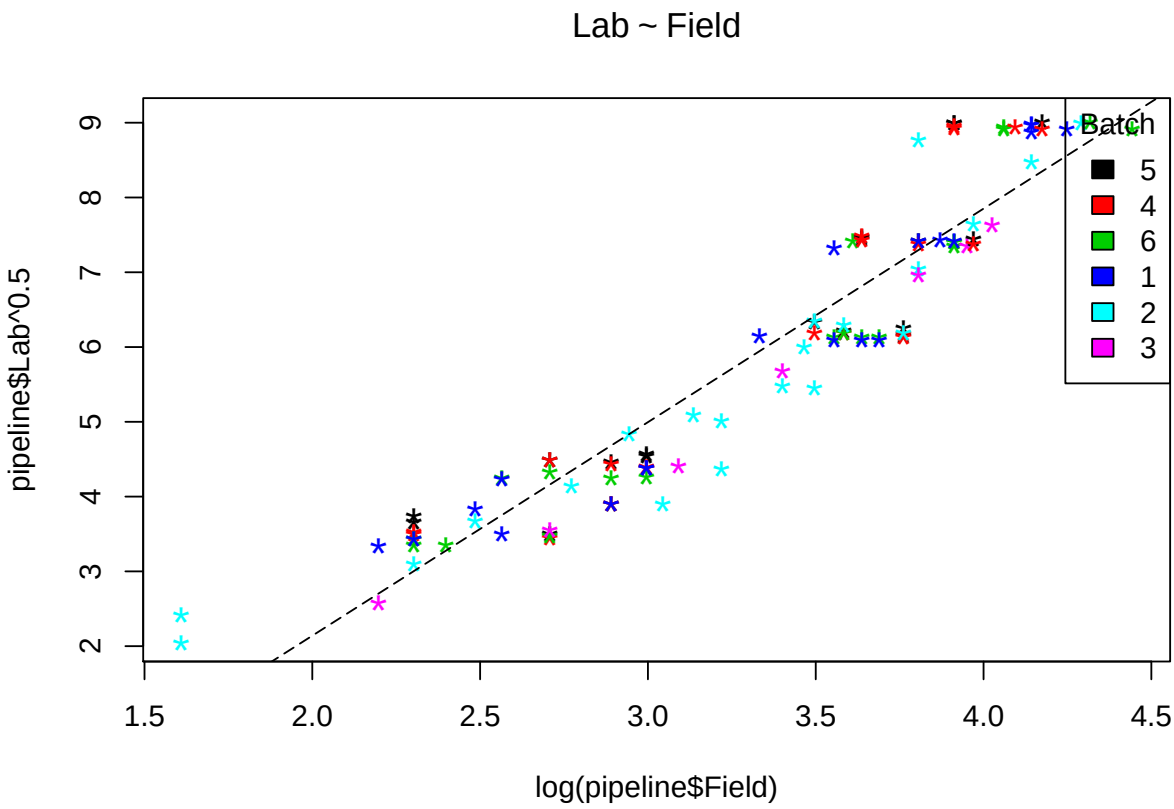


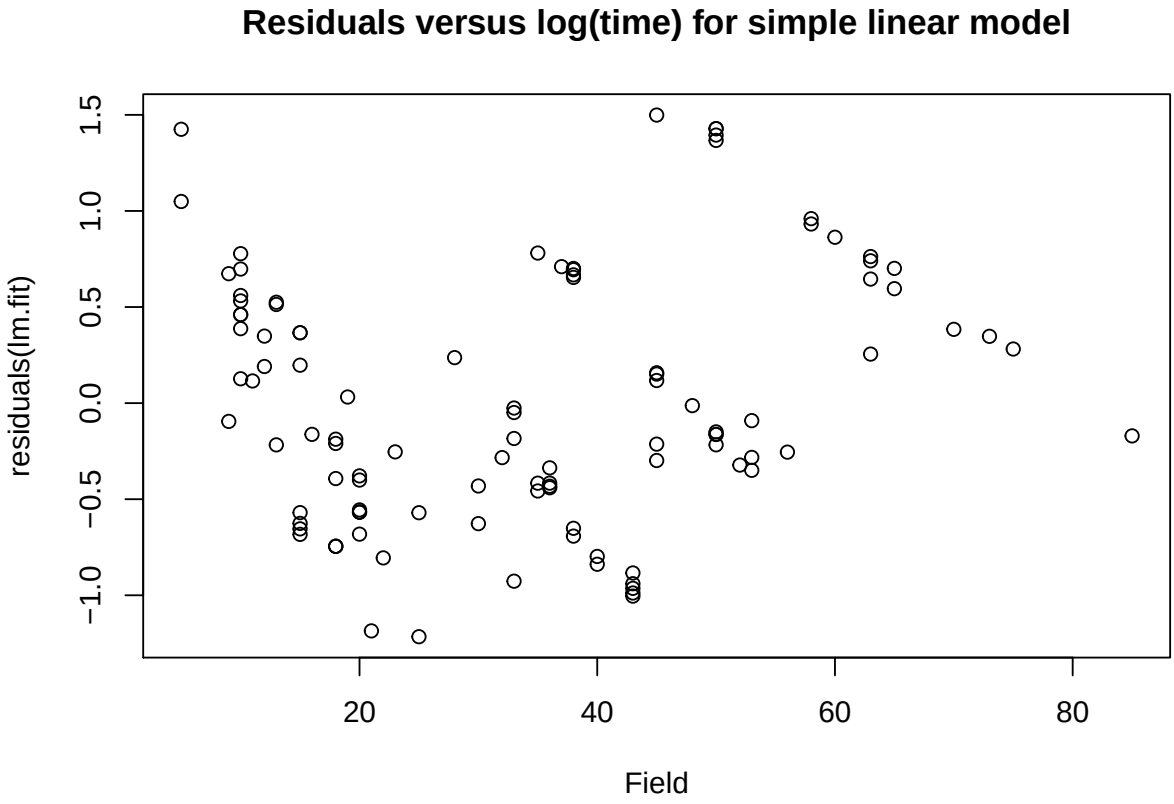


(c) An alternative to weighting is transformation. Find transformations on Lab and/or Field so that in the transformed scale the relationship is approximately linear with constant variance. You may restrict your choice of transformation to square root, log and inverse.

```
##
## Call:
## lm(formula = (Lab)^0.5 ~ log(Field), data = pipeline)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2157 -0.5066 -0.1625  0.5191  1.4991
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.57141    0.33200  -10.76   <2e-16 ***
## log(Field)   2.85554    0.09786   29.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.6553 on 105 degrees of freedom
## Multiple R-squared:  0.8902, Adjusted R-squared:  0.8892
## F-statistic: 851.4 on 1 and 105 DF,  p-value: < 2.2e-16
```





The RSE of this model is lower than the weighted model. I'm not sure I'd use that alone as a criteria for selecting a model. If we had a physical reason for the variance model we used - then we might opt to stick with the weighted regression.

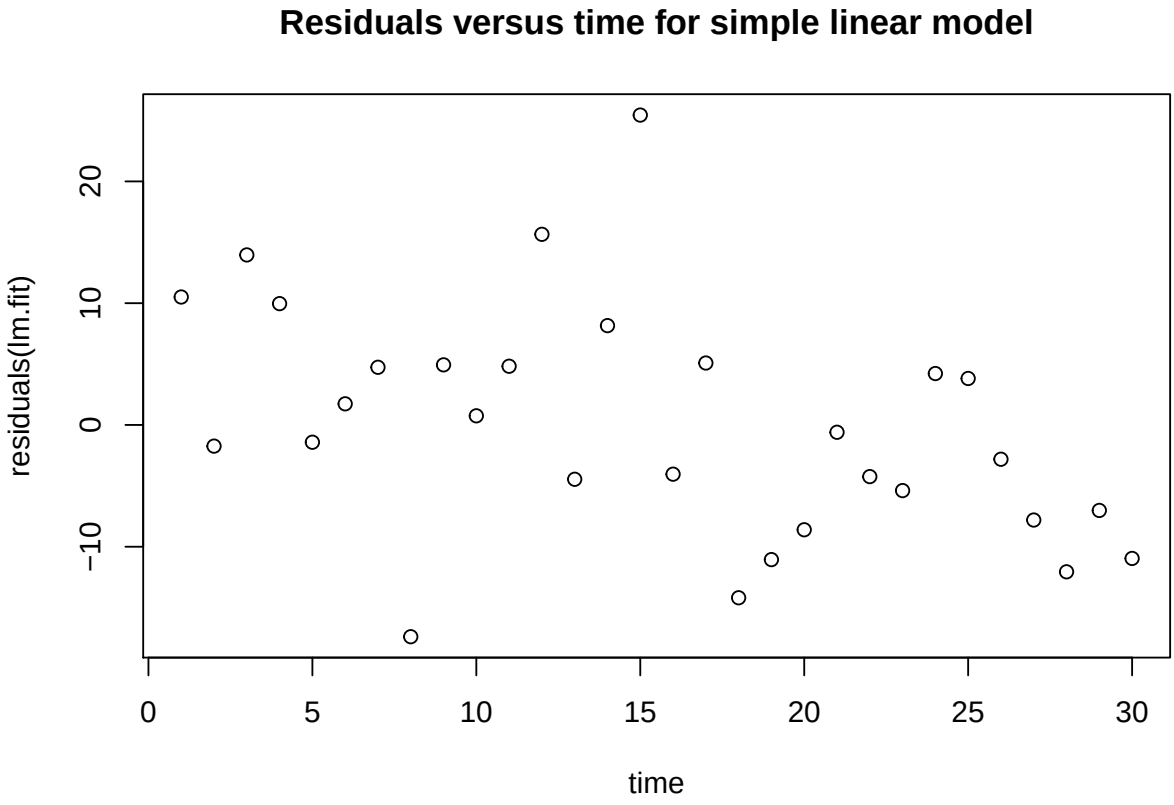
8.6 Analysis of cheddar data

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

```
##
## Call:
## lm(formula = taste ~ ., data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
```

```
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

(a) Suppose that the observations were taken in time order. Create a time variable. Plot the residuals of the model against time and comment on what can be seen.



Fitting a linear model we can get an estimate of the correlation.

```
##
## Call:
## lm(formula = X1 ~ X2, data = df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -21.0152  -5.2405  -0.7975   4.1784  25.2072
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.4910     3.2764   2.286   0.0300 *
## X2           -0.4833     0.1846  -2.619   0.0141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.749 on 28 degrees of freedom
## Multiple R-squared:  0.1967, Adjusted R-squared:  0.168
## F-statistic: 6.857 on 1 and 28 DF,  p-value: 0.01409
```

Here we calculate the lag(1) correlation on the residuals.

Table 1: lag(1) correlation on residuals

cor.residuals
0.1787

(b) Fit a GLS model with same form as above but now allow for an AR(1) correlation among the errors. Is there evidence of such a correlation?

```
## Generalized least squares fit by REML
##   Model: taste ~ Acetic + H2S + Lactic
##   Data: na.omit(cheddar)
##      AIC      BIC  logLik
##   214.94 222.4886 -101.47
##
## Correlation Structure: AR(1)
##  Formula: ~time
##  Parameter estimate(s):
##       Phi
## 0.2641944
##
## Coefficients:
##                   Value Std.Error   t-value p-value
## (Intercept) -30.332472 20.273077 -1.496195  0.1466
## Acetic        1.436411  4.876581  0.294553  0.7707
## H2S           4.058880  1.314283  3.088284  0.0047
## Lactic       15.826468  9.235404  1.713674  0.0985
##
##  Correlation:
```

```
##        (Intr) Acetic H2S
## Acetic -0.899
## H2S     0.424 -0.395
## Lactic  0.063 -0.416 -0.435
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -1.64546468 -0.63861716 -0.06641714  0.52255676  2.41323021
##
## Residual standard error: 10.33276
## Degrees of freedom: 30 total; 26 residual
```

(c) Fit a LS model but with time now as an additional predictor. Investigate the significance of time in the model.

```
##
## Call:
## lm(formula = taste ~ ., data = cheddar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3523  -4.9735  -0.5089   4.8531  23.1311
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.6127    17.9845  -2.036  0.05250 .
## Acetic        4.1275     4.2556   0.970  0.34139
## H2S           3.5387     1.1315   3.127  0.00444 **
## Lactic       17.9527     7.7875   2.305  0.02973 *
## time         -0.5459     0.2043  -2.672  0.01306 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.112 on 25 degrees of freedom
## Multiple R-squared:  0.7291, Adjusted R-squared:  0.6858
## F-statistic: 16.83 on 4 and 25 DF,  p-value: 8.205e-07
```

Time is significant in this model at a level of $\alpha = 0.05$. The coefficient tells us that when all other variables are held constant an increasing time value results in a decreasing value of taste. We see the value of the coefficient is close to the model we fit of residuals of the original model versus time ~ -0.5.

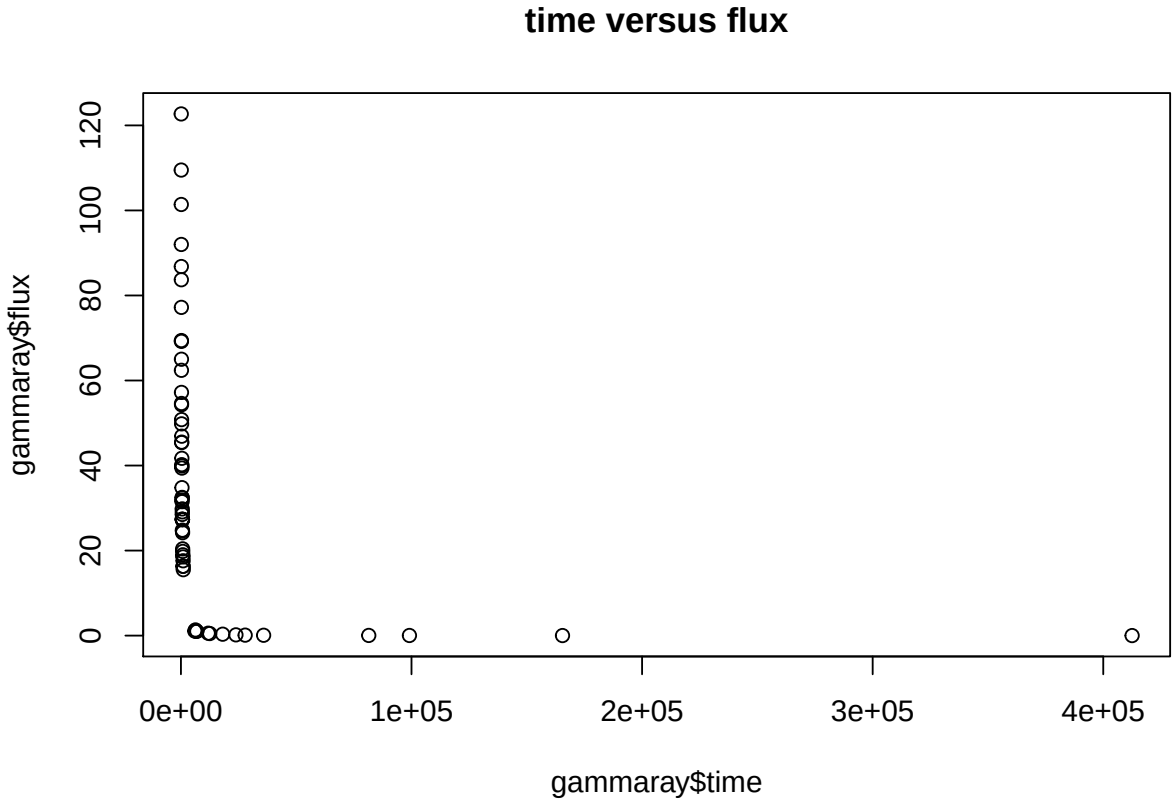
(d) The last two models have both allowed for an effect of time. Explain how they do this differently.

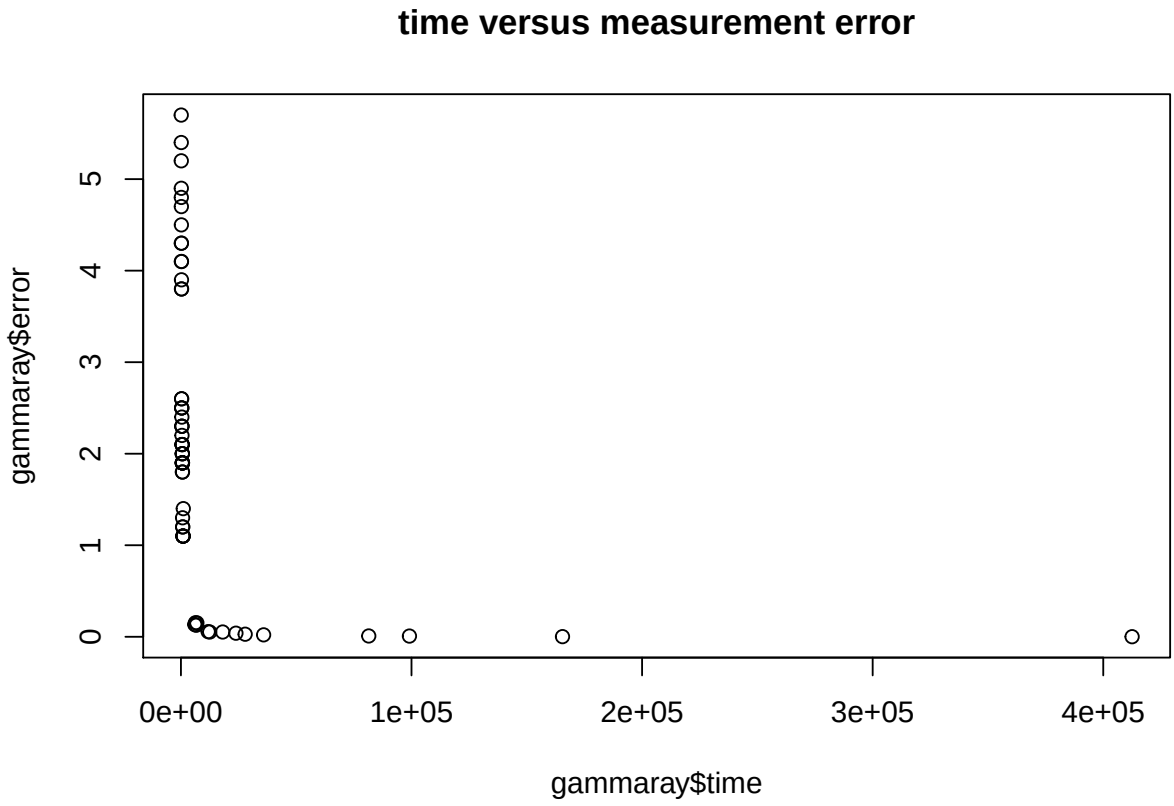
Obviously the LS model accounts for time by explicitly including it as a predictor. The GLS model we account for time through the error structure. We construct an estimate of the variance covariance matrix of the regression equation $\Sigma = S^t S$ from an AR(1) model of the residuals.

8.8 Gammaray analysis

The gammaray dataset shows the x-ray decay light curve of a gamma ray burst. Build a model to predict the flux as a function time that uses appropriate weights.

First we plot the data

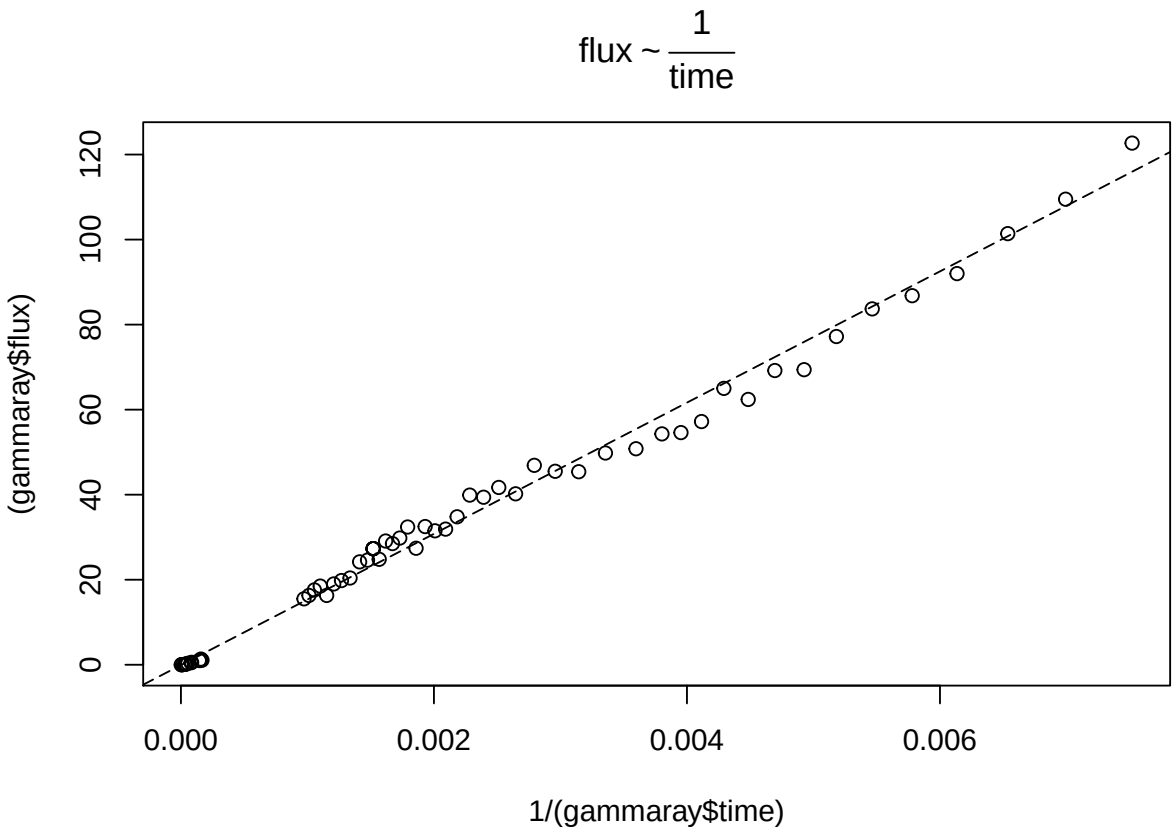


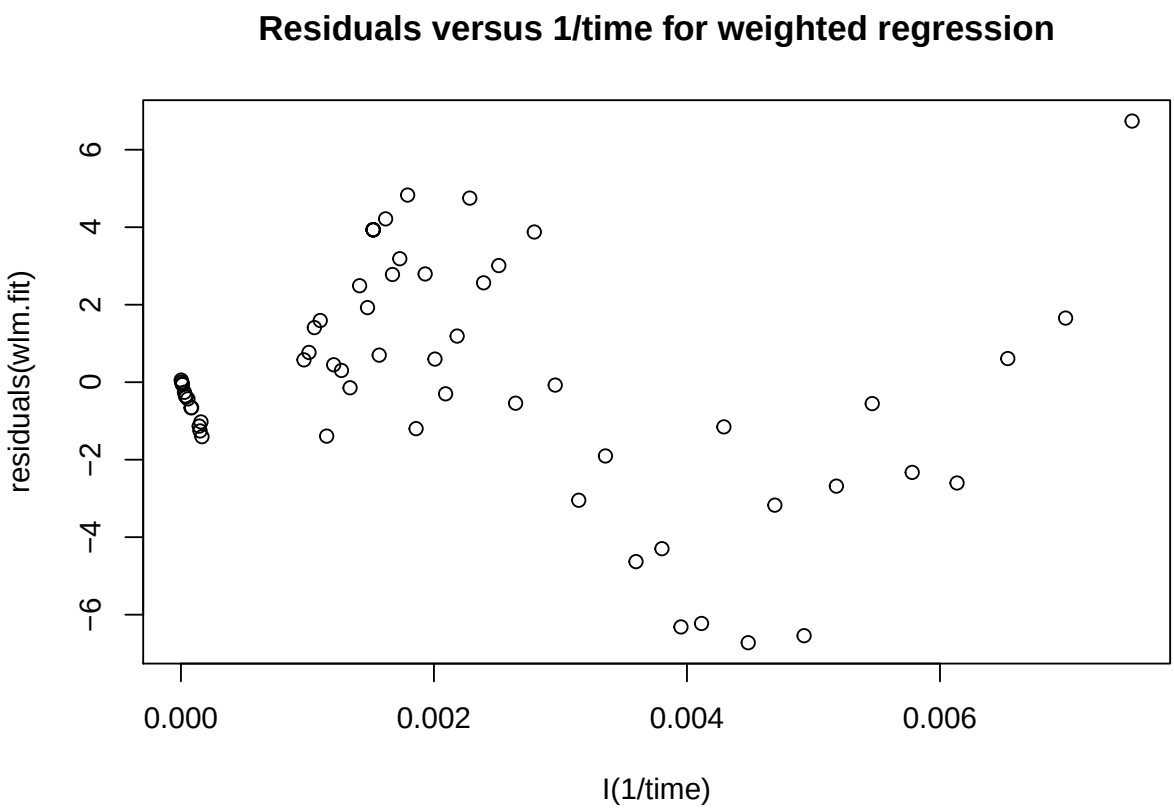


We see a hyperbolic relationship btween the predictor and the response so We fit a weighted regression $flux \sim \frac{1}{time}$ with weights equal to the error

```
## Generalized least squares fit by REML
##   Model: flux ~ I(1/time)
##   Data: gammaray
##        AIC     BIC    logLik
##   261.1843 267.517 -127.5922
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~error
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept)    -0.09   0.04068 -2.21963  0.0302
## I(1/time)   15434.73 171.88649 89.79603  0.0000
##
##  Correlation:
##           (Intr)
## I(1/time) -0.112
```

```
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -1.81209736 -0.87077410 -0.02224618  0.91280791  1.57197856
##
## Residual standard error: 2.120048
## Degrees of freedom: 63 total; 61 residual
```



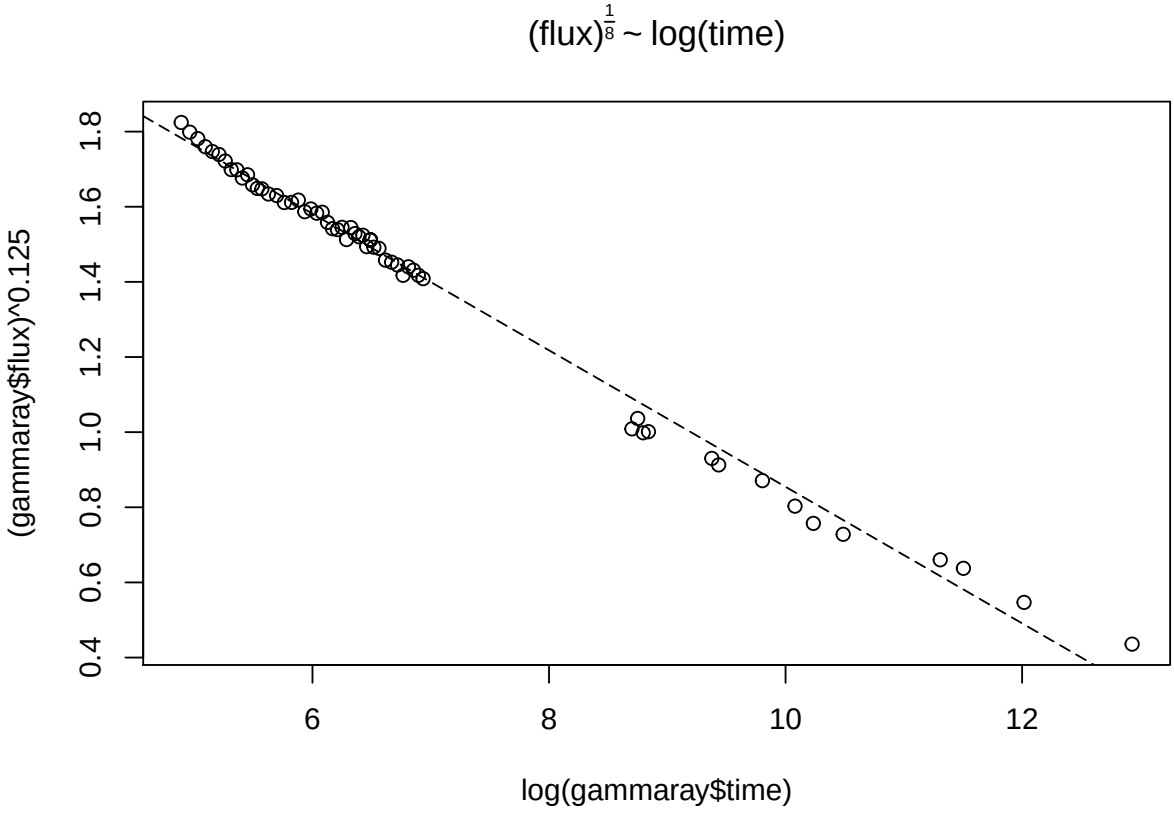


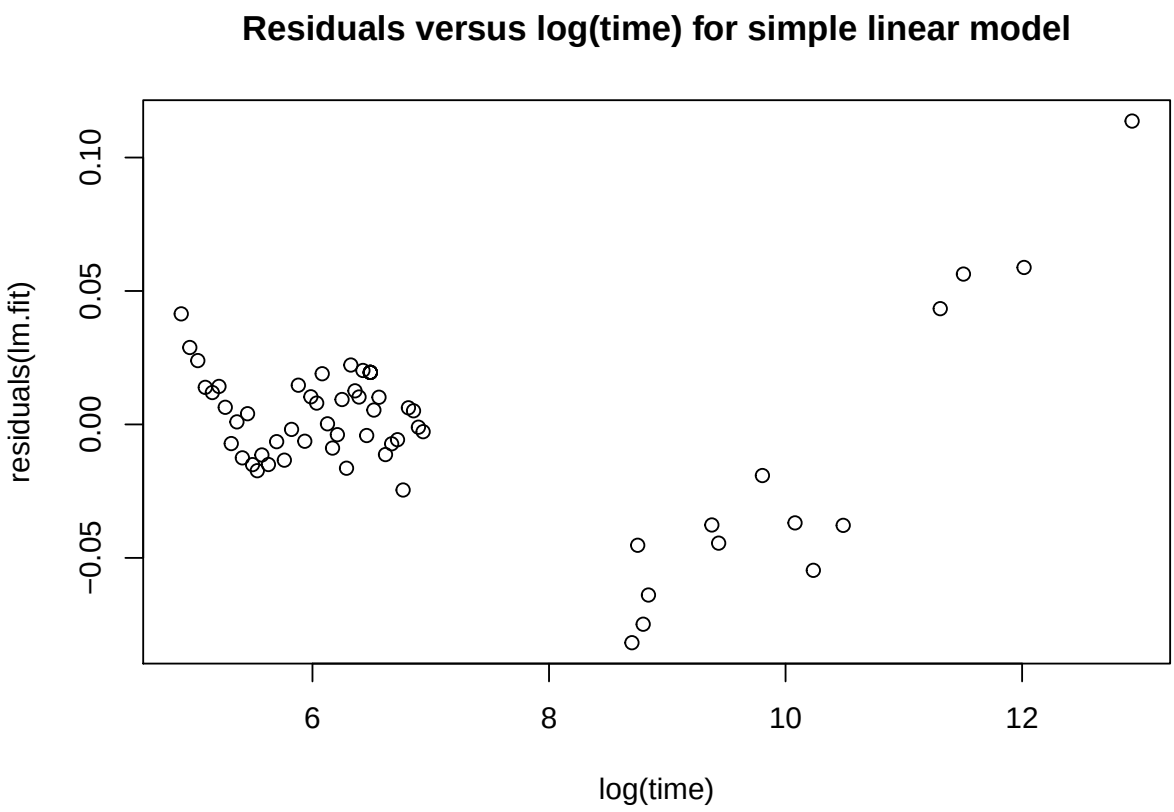
We try a trasformation - for reference, and for fun. We Chose this model after some experimenting

$$(flux)^{\frac{1}{8}} \sim log(time)$$

```
##
## Call:
## lm(formula = (flux)^0.125 ~ log(time), data = gammaray)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.081779 -0.012954  0.000958  0.016848  0.113664
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.671490   0.015088  177.06   <2e-16 ***
## log(time)   -0.181701   0.002094  -86.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0314 on 61 degrees of freedom
```

```
## Multiple R-squared:  0.992,  Adjusted R-squared:  0.9918
## F-statistic:  7532 on 1 and 61 DF,  p-value: < 2.2e-16
```




Residuals versus log(time) for simple linear model
residuals(lm.fit)
0.10
0.05
0.00
−0.05
6
8
10
12
log(time)