Homework 5

ST623

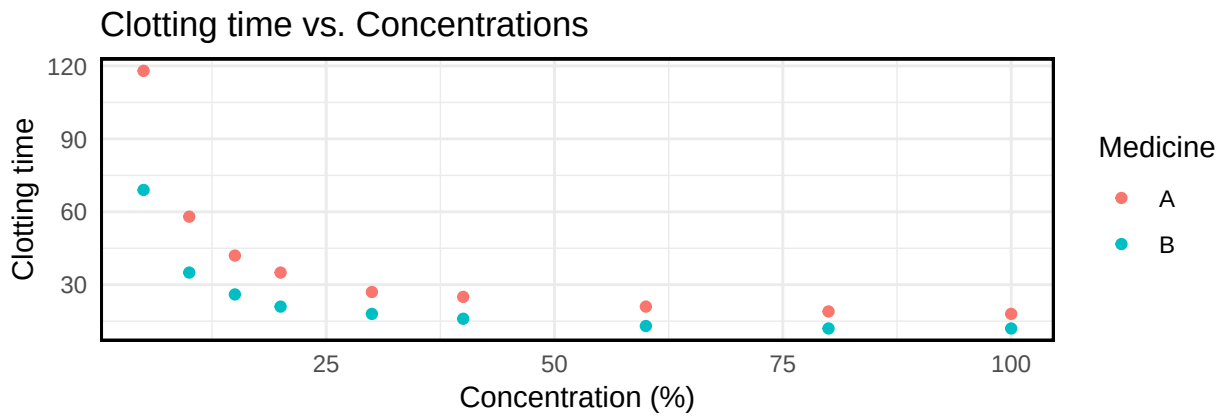
Nick Sun

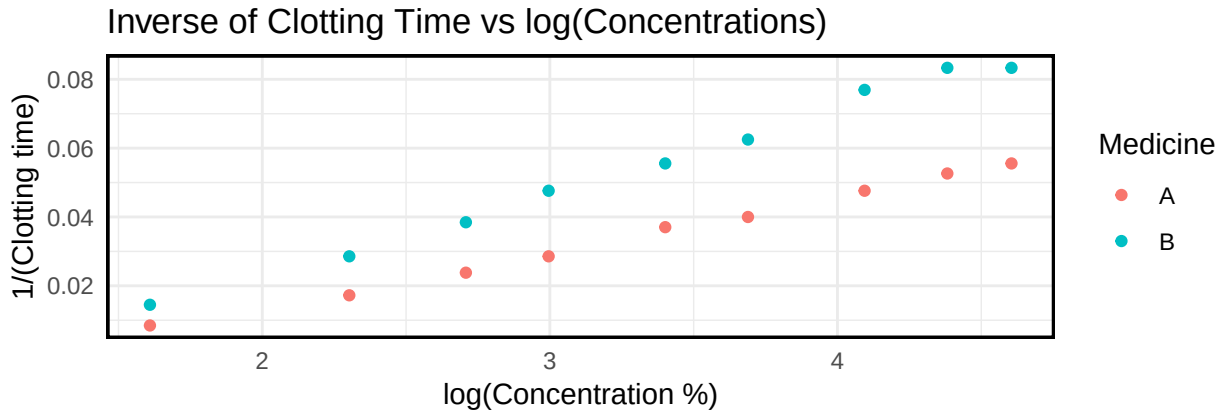
November 14, 2019

Question 1

Part a.

concentration	clottingTime	medicine
5	118	A
10	58	A
15	42	A
20	35	A
30	27	A
40	25	A





In the first plot, there appears to be an exponential curve. However, taking the mentioned transformations of the explanatory and response variable is shown in the second figure. We can see that it makes the relationship more linear! Nice.

Part b.

This suggests to us to use a generalized regression model with an inverse link. A common one is the gamma regression model.

We will model clotting as a function of `log(concentration) + as.factor(medicine) + log(concentration)*as.factor(` with no intercept term.

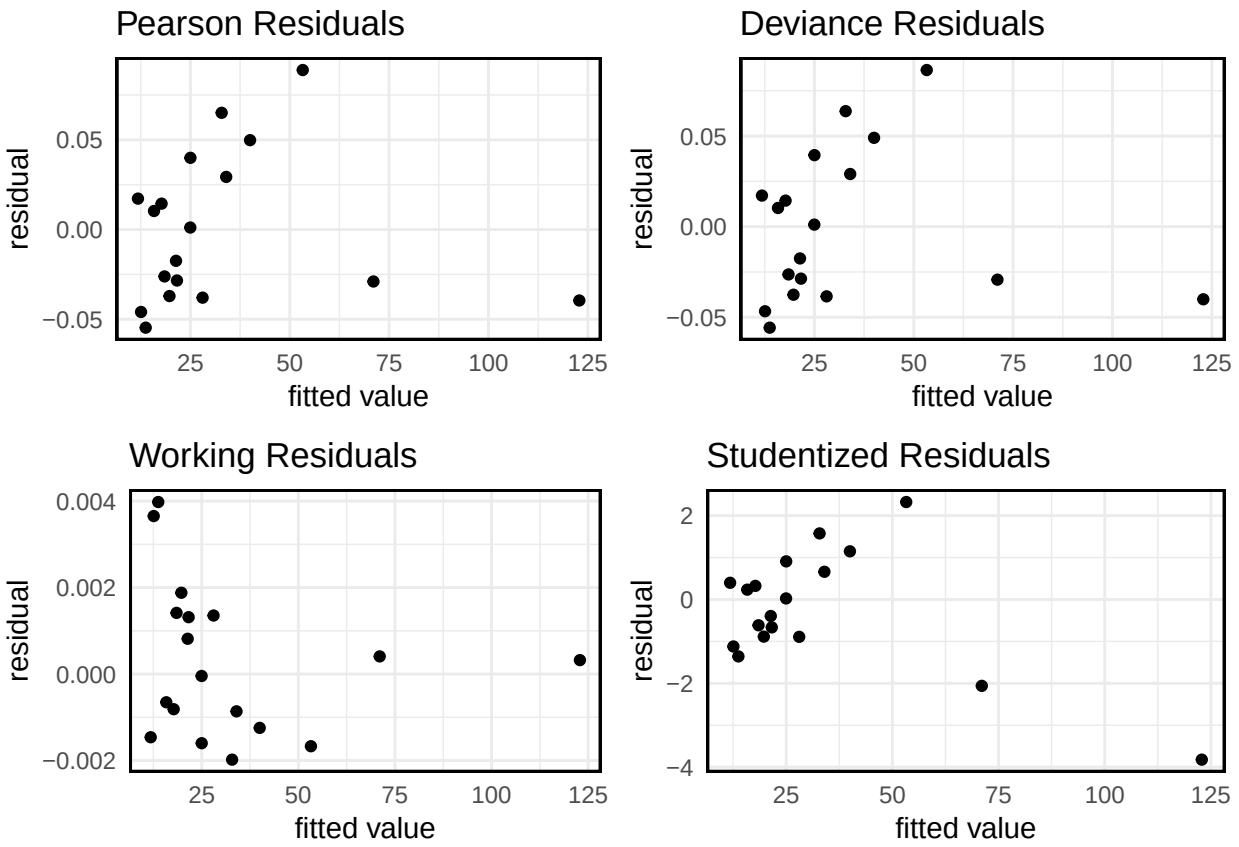
Table 2: Table continues below

	Estimate	Std. Error	t value
log(concentration)	0.01534	0.0003872	39.63
as.factor(medicine)A	-0.01655	0.0008655	-19.13
as.factor(medicine)B	-0.02391	0.001438	-16.63
log(concentration):as.factor(medicine)B	0.008256	0.0007353	11.23

	Pr(>|t|)
log(concentration)	8.851e-16
as.factor(medicine)A	1.967e-11
as.factor(medicine)B	1.289e-10
log(concentration):as.factor(medicine)B	2.184e-08

(Dispersion parameter for Gamma family taken to be 0.002129707)

Null deviance:	NaN on 18 degrees of freedom
Residual deviance:	0.0294 on 14 degrees of freedom



Overall the residual plots look pretty good. There is a pretty constant spread in the Pearson and Deviance residuals. My only concern are the two fitted points that are very far wara from the other points and result in relatively high studentized residuals.

Part c.

Now we are being asked to find the relative potency of one medicine vs. the other.

Our quantity of interest is $\frac{x_i}{x_j}$ where x_i is the dosage for medicine A and x_j is the dosage for medicine B. The equation we want to solve then is:

$$\alpha_i + \beta_i \log(x_i) = \alpha_j + \beta_j \log(x_j) + \beta_j^* \log(x_j) (\text{the interaction term})$$

$$\frac{x_i}{x_j} = \exp\left(\frac{\alpha_j - \alpha_i - (\beta_i - \beta_j - \beta_j^*}{\beta_i}\right)$$

Now all we have to do is plug in the appropriate coefficients.

```
## as.factor(medicine)A
##             9.701534
```

Question 2

Let's analyze the `bees.txt` dataset.

Case	Number	Time
1	34	9
2	13	10
3	11	12
4	32	13
5	39	14
6	36	15

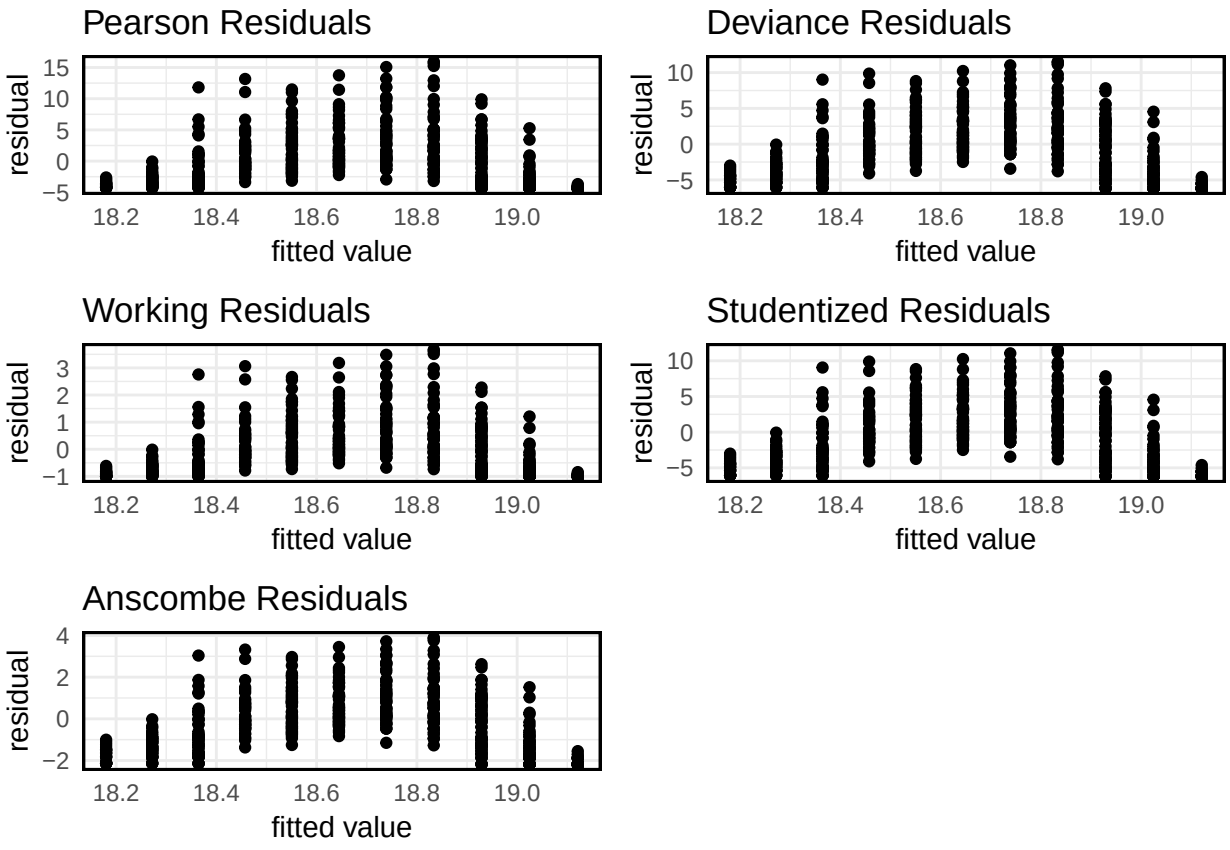
We can model the number of bees as a function of time several different ways while using the Poisson regression framework. We will go over a few of these ways here.

Part a: Linear in Time

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.986	0.0419	71.27	0
Time	-0.005046	0.00349	-1.446	0.1482

(Dispersion parameter for poisson family taken to be 1)

Null deviance:	9306 on 503 degrees of freedom
Residual deviance:	9303 on 502 degrees of freedom

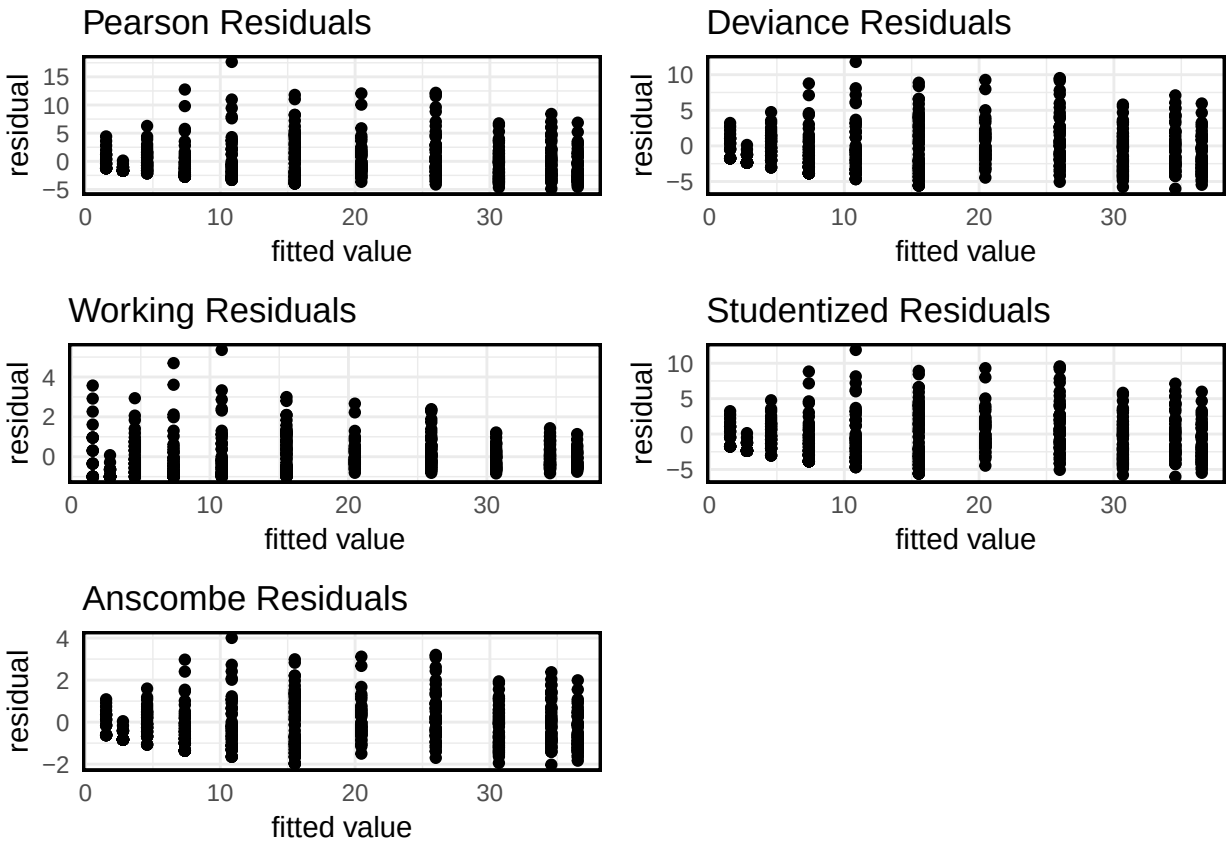


Part b: Quadratic in Time

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.24	0.2852	-42.9	0
Time	2.699	0.04929	54.76	0
I(Time^2)	-0.1149	0.002096	-54.84	0

(Dispersion parameter for poisson family taken to be 1)

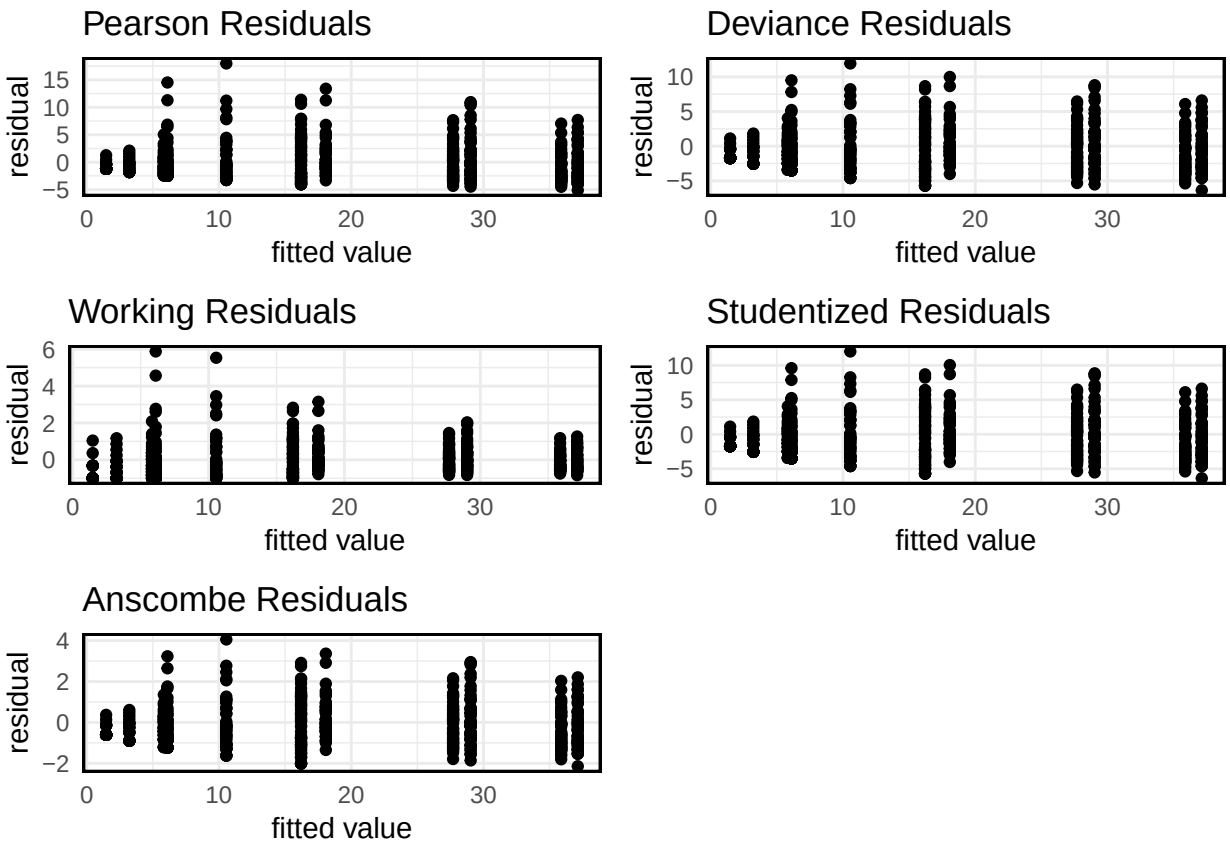
Null deviance:	9306 on 503 degrees of freedom
Residual deviance:	4879 on 501 degrees of freedom



Part c: Cubic in Time

Table 10: Fitting generalized (poisson/log) linear model: Number ~ I(Time^3) + I(Time^2) + Time

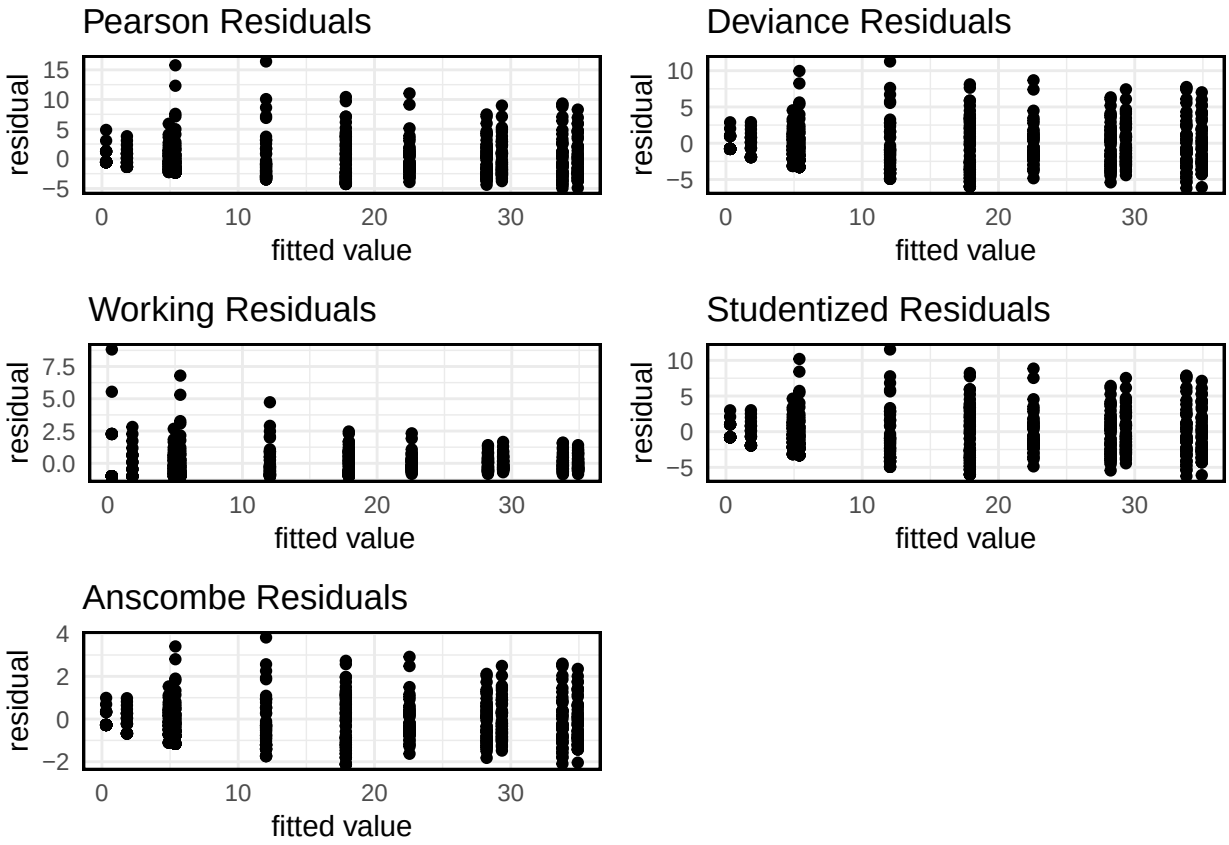
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-26.91	1.3	-20.69	4.071e-95
I(Time^3)	0.009383	0.0007866	11.93	8.309e-33
I(Time^2)	-0.4499	0.02843	-15.82	2.11e-56
Time	6.588	0.3362	19.6	1.63e-85



Part d: model using factor(Time)

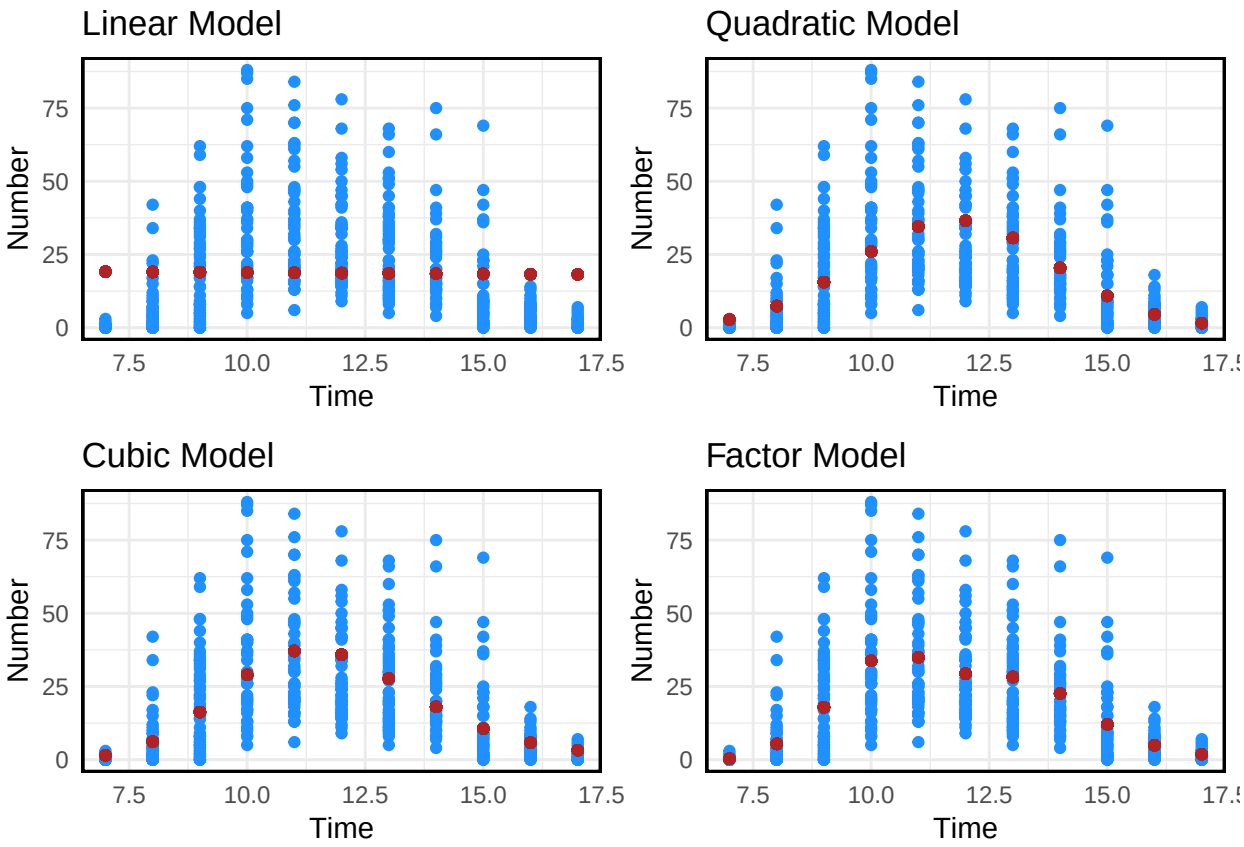
Table 11: Fitting generalized (poisson/log) linear model: Number ~ as.factor(Time)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.186	0.3015	-3.932	8.415e-05
as.factor(Time)8	2.87	0.3072	9.344	9.309e-21
as.factor(Time)9	4.07	0.3031	13.43	3.955e-41
as.factor(Time)10	4.706	0.3025	15.55	1.475e-54
as.factor(Time)11	4.739	0.3025	15.67	2.583e-55
as.factor(Time)12	4.565	0.3026	15.09	1.941e-51
as.factor(Time)13	4.526	0.3027	14.95	1.548e-50
as.factor(Time)14	4.302	0.3033	14.19	1.102e-45
as.factor(Time)15	3.674	0.3044	12.07	1.544e-33
as.factor(Time)16	2.776	0.3095	8.97	2.965e-19
as.factor(Time)17	1.792	0.3371	5.315	1.065e-07



Conclusion

Let's take a quick look at the plots of the fitted values:



From these fits, we see that we have two strong candidates for a decent fitting model: quadratic and factor.

Let's do an AIC calculation for good measure to compare the different models.

Linear	Quadratic	Cubic	Factor
11253	6831	6691	6453

We can immediately notice the model with Time as a factor variable has the lowest AIC value, which might suggest its the model that captures the relationship of the data the best. However, we can also see that there is some heteroskedasticity in the residuals, particularly in the Pearson and Deviance residuals.

The cubic model performs nearly as well on AIC, but we see that the residuals are also somewhat heteroskedastic.

We might also be interested in looking at the quadratic model with Time, since the AIC value of this model is similar. While there is also heteroskedasticity with this model, most noticeably in the Pearson residuals, however the other residual plots look better. The difference is most pronounched in the Working residuals I think. Furthermore, the quadratic model has less parameter estimates than the factor model.

For these reasons, I would go with using the quadratic model.