Solution to Exercise 17.2 (Version 1, 4/9/15)

from Statistical Methods in Biology: Design & Analysis of Experiments and Regression (2014) S.J. Welham, S.A. Gezan, S.J. Clark & A. Mead. Chapman & Hall/CRC Press, Boca Raton, Florida. ISBN: 978-1-4398-0878-8

Exercise 17.2 (Data: courtesy V. Buchanan-Wollaston (PRESTA), University of Warwick)

The microarray study described in full in Exercise 12.6 investigated gene expression associated with senescence of leaves. File SENESCENCE.DAT holds design information (*ID*, variate *Day*, factor BiolRep) and the expression value for three genes (variates *CATMA3A13560*, *CATMA2A31585* and *CATMA1A09000*) from each plant following normalization.

a) Can you reasonably use polynomial regression to predict the expression of genes *CATMA2A31585* or *CATMA1A09000* over time? Over what range are your predictions reliable?
b) Can you improve on these predictions by using non-linear models?

Solution 17.2

a) Polynomial models

In this experiment, we have biological replicates at each sampling date, so we can test for lack of fit (LoF, see Exercise 13.1). We follow the procedure described in Section 17.1.2: we start with linear regression and progressively add higher-order terms, at each step examining diagnostic plots and formally checking for LoF.

To fit polynomial models, we calculate powers of the *Day* variate, after correcting for the mean, with $Dayk = (Day - \text{mean}(Day))^k$. For testing LoF, we use factor fDay, defined to have a separate level for each sampling date.

Gene CATMA2A31585

We identified an outlying observation (unit 32) for this gene in Exercise 13.1 and we will continue to omit this observation.

The SLR model was fitted in Exercise 13.1 and showed significant LoF ($F_{9,32}$=4.08, $P < 0.001$) as well as evidence of curvature in the fitted value plot. The quadratic model takes symbolic form

Response:	*CATMA2A31585*
Explanatory component:	*[1] + Day + Day2*

and gives the composite set of diagnostic plots in Figure S17.2.1. There is still strong curvature left in the fitted value plot, and a plot of the fitted model with the data (not shown) shows systematic deviations of the fitted model from the overall pattern.

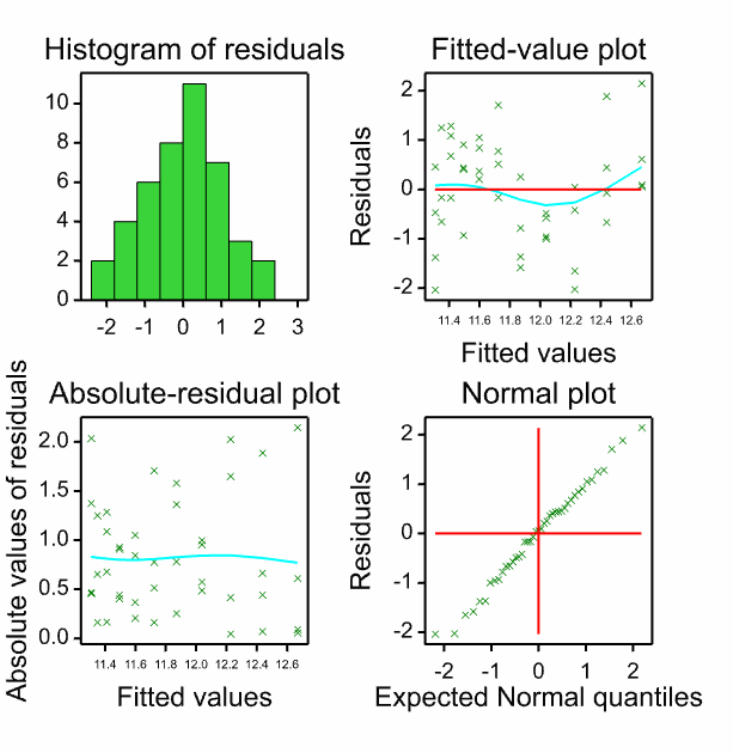

Figure S17.2.1 Composite set of residual plots based on standardized residuals for a quadratic polynomial model for gene CATMA2A31585 (omitting unit 32).

Table S17.2.1 A sequential ANOVA table for quadratic polynomial model for gene CATMA2A31585 (omitting unit 32), testing for lack of fit.

Term added	Incremental df	Incremental SS	Mean square	Variance ratio	*P*
+ *Day*	1	7.9222	7.9222	137.412	< 0.001
+ *Day2*	1	0.3832	0.3832	6.647	0.015
+ fDay	8	1.7349	0.2169	3.762	0.003
Residual	32	1.8449	0.0577		
Total	42	11.8853			

We can formally test LoF by adding factor fDay to the model and examining the resulting incremental ANOVA table (Table S17.2.1). There is evidence that adding the quadratic term has improved the model ($F_{1,32}$ = 6.65, *P* = 0.015) but there is still significant LoF ($F_{8,32}$ = 3.76, *P* = 0.003), as suggested by the diagnostic plots. We therefore proceed to the cubic model, which has symbolic form

Response: *CATMA2A31585*
Explanatory component: *[1] + Day + Day2 + Day3*

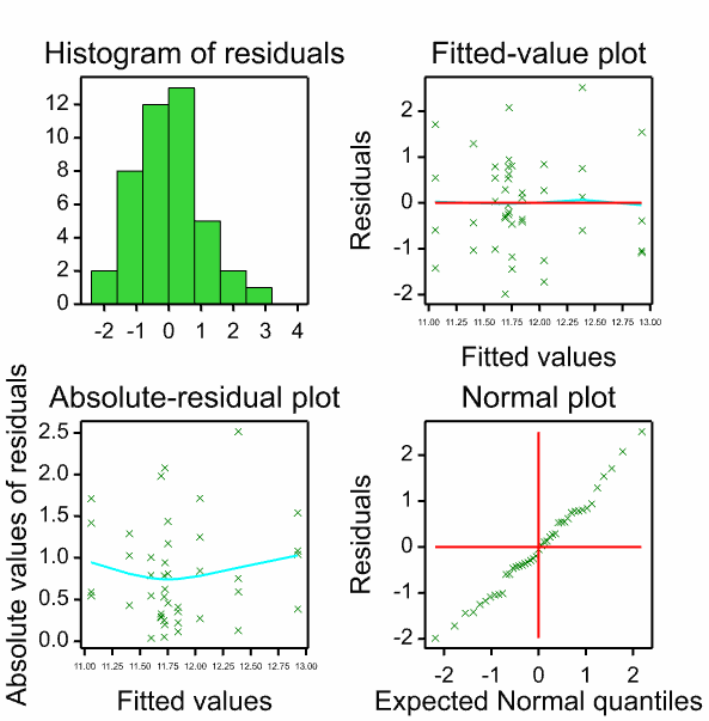

Figure S17.2.2 Composite set of residual plots based on standardized residuals for a cubic polynomial model for gene CATMA2A31585 (omitting unit 32).

Table S17.2.2 A sequential ANOVA table for cubic polynomial model for gene CATMA2A31585 (omitting unit 32), testing for lack of fit.

Term added	Incremental df	Incremental SS	Mean square	Variance ratio	*P*
+ *Day*	1	7.9222	7.9222	137.412	< 0.001
+ *Day2*	1	0.3832	0.3832	6.647	0.015
+ *Day3*	1	1.2285	1.2285	21.309	< 0.001
+ fDay	7	1.7349	0.0723	1.255	0.303
Residual	32	1.8449	0.0577		
Total	42	11.8853			

Table S17.2.3 Parameter estimates with standard errors (SE), t-statistics (t) and observed significance levels (*P*) for cubic polynomial model for gene CATMA2A31585.

Term	Parameter	Estimate	SE	t	*P*
[1]	α_1	11.869	0.444	26.734	< 0.001
Day	β_1	-0.0050	0.0152	-0.329	0.744
Day2	β_2	0.00268	0.00105	2.544	0.015
Day3	β_3	-0.000882	0.000195	-4.514	< 0.001

There is very little systematic pattern left in Figure S17.2.2, and this is supported by a formal test for LoF for this model, derived from the sequential ANOVA table (Table S17.2.2, LoF test with $F_{7,32}$ = 1.255, P = 0.303). A plot of the fitted cubic polynomial (Figure S17.2.3) shows that it passes through the trend in the data, so we accept this model, which accounts for 78.7% of the variation in expression (adjusted R^2 = 0.787). The parameter estimates are in Table S17.2.3, so the predictive model can be written as:

$$\begin{aligned}\hat{\mu}(Day) &= 11.87 - 0.005Day + 0.0027(Day-29)^2 - 0.00088(Day-29)^3 \\ &= 11.87 - 0.005Day + 0.0027(Day^2 - 58Day + 841) - 0.00088(Day^3 - 87Day^2 + 2523Day - 24389) \\ &= 35.64 - 2.387Day + 0.0794Day^2 - 0.00088Day^3\end{aligned}$$

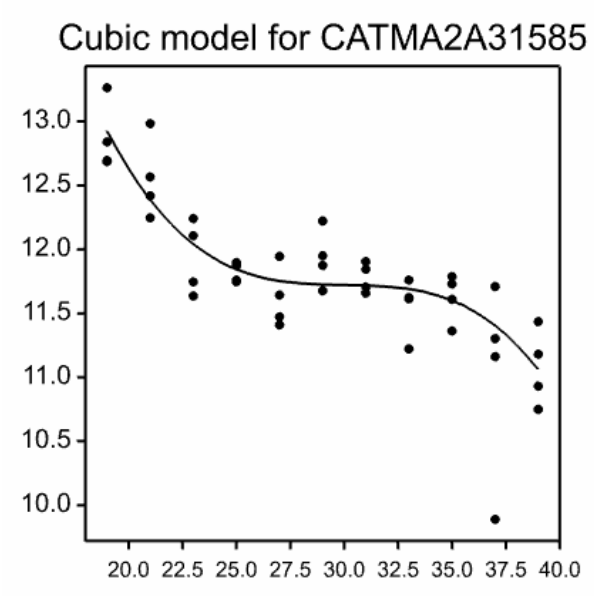

Figure S17.2.3 Fitted cubic polynomial model with observed data for gene CATMA2A31585.

Gene CATMA1A09000

We have not examined this data before, so we start by plotting it in Figure S17.2.4. The shape appears to be a straight line with curvature at the ends of the range, possibly a sigmoid shape. We start by fitting the SLR, written in symbolic form as:

Response:	*CATMA1A09000*
Explanatory component:	*[1] + Day*

Residual plots from this model are shown in Figure S17.2.5. As we might expect from Figure S17.2.4, there is evidence of curvature in the fitted values plot, and this is confirmed by a formal test for lack of fit. We repeat the previous procedure, adding a quadratic term, which still shows curvature in the fitted values plot and lack of fit, and so proceed to a cubic polynomial model. The sequential ANOVA table for this cubic polynomial model is in Table S17.2.4. This shows no formal lack of fit ($F_{7,33}$ = 1.325,

$P = 0.270$). However, a plot of the fitted model with the data suggests that the fit is not entirely satisfactory (Figure S17.2.6), as the expression values plateau at the end of the experiment, but the cubic polynomial fits a downwards trend after day 35. We would prefer to find a non-linear model that gives a better representation of the observed trend.

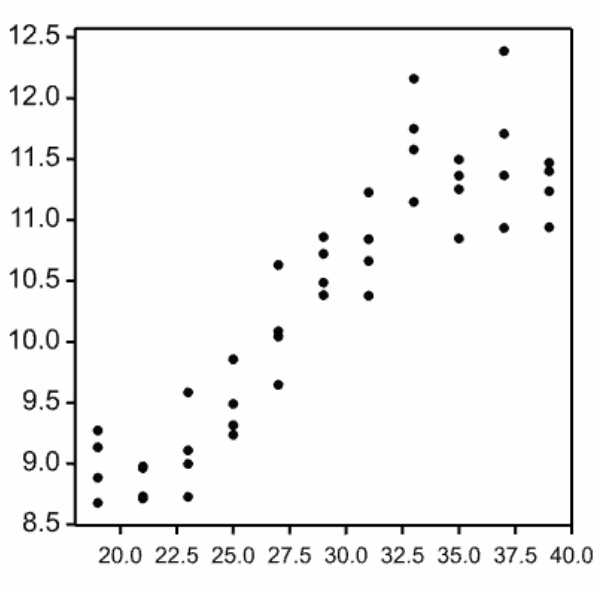

Figure S17.2.4 Observed progression of expression over time for gene CATMA1A09000.

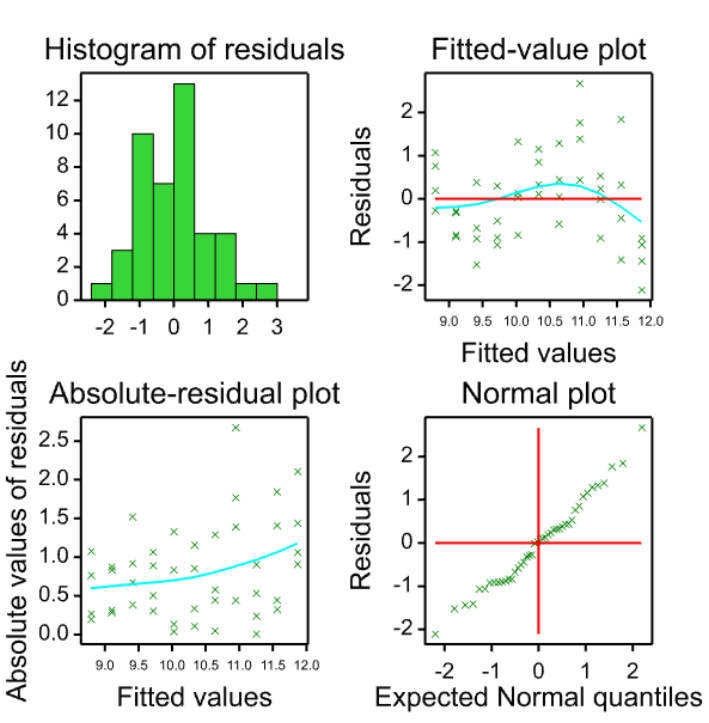

Figure S17.2.5 Composite set of residual plots based on standardized residuals for an SLR model for gene CATMA1A09000.

Table S17.2.4 A sequential ANOVA table for cubic polynomial model for gene CATMA1A09000, testing for lack of fit.

Term added	Incremental df	Incremental SS	Mean square	Variance ratio	*P*
+ *Day*	1	41.5441	41.5441	347.250	< 0.001
+ *Day2*	1	1.0084	1.0084	8.429	0.007
+ *Day3*	1	2.8287	2.8287	23.644	< 0.001
+ fDay	7	1.1094	0.1585	1.325	0.270
Residual	33	3.9480	0.1196		
Total	43	50.4386			

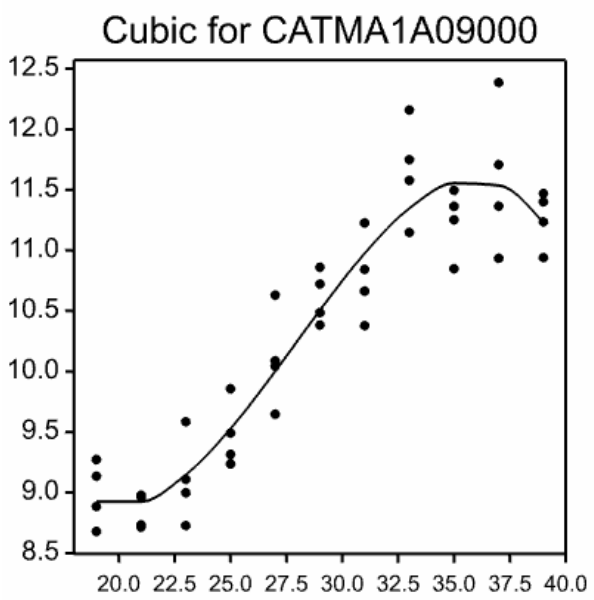

Figure S17.2.7 Fitted cubic polynomial model with observed data for gene CATMA1A09000.

b) Non-linear models

Gene CATMA2A31585

This gene does not really fit the shape of any of the standard non-linear curves, so we might not be hopeful of success. Out of the models covered in Chapter 17, we will try the exponential and ldl curves. In each case, we will test for LoF after fitting the non-linear model. Here, we will illustrate how to do this calculation by hand, as statistical program facilities for non-linear models rarely provide this facility. We start by getting quantities associated with pure error, namely the PESS, PEDF and PEMS. We can obtain these from the polynomial models above which include the LoF term, and deduce that PESS = 1.8449, PEDF = 32, PEMS = 0.0577.

We next fit the exponential model, which takes the form

$$Expression_i = \alpha + \beta \exp(-\gamma Day_i) + e_i$$

where $Expression_i$ is the i^{th} observation of expression values on gene CATMA2A31585, taken on Day_i with deviation e_i. The summary ANOVA table for this model is Table S17.2.5 and the model accounts for 71.3% of the variation (adjusted R^2 = 0.713). The fitted values and diagnostic plots from this model are in Figure S17.2.8. There is curvature in the fitted values plot as the fitted curve does not follow the observed trend well. To calculate a formal test of LoF, we need to obtain the LoF SS and df. We can calculate the LoFSS as the difference between the ResSS for the exponential model (= 3.253) and the PESS (= 1.845). The LoFDF is similarly the difference between the ResDF for the exponential model (= 40) and the PEDF (= 32). The F-statistic for testing LoF takes the form

$$F = \frac{(\text{ResSS} - \text{PESS}) / (\text{ResDF} - \text{PEDF})}{\text{PEMS}} = \frac{(3.253 - 1.845) / (40 - 33)}{0.05765} = 3.054$$

with 8 and 32 df. This indicates strong evidence (P = 0.011) of lack of fit to the exponential model.

Table S17.2.5 ANOVA table for exponential model for gene CATMA2A31585, excluding unit 32.

Term	df	SS	Mean square	Variance ratio	*P*
Model	2	8.632	4.3160	53.06	< 0.001
Residual	40	3.253	0.0813		
Total	42	11.885			

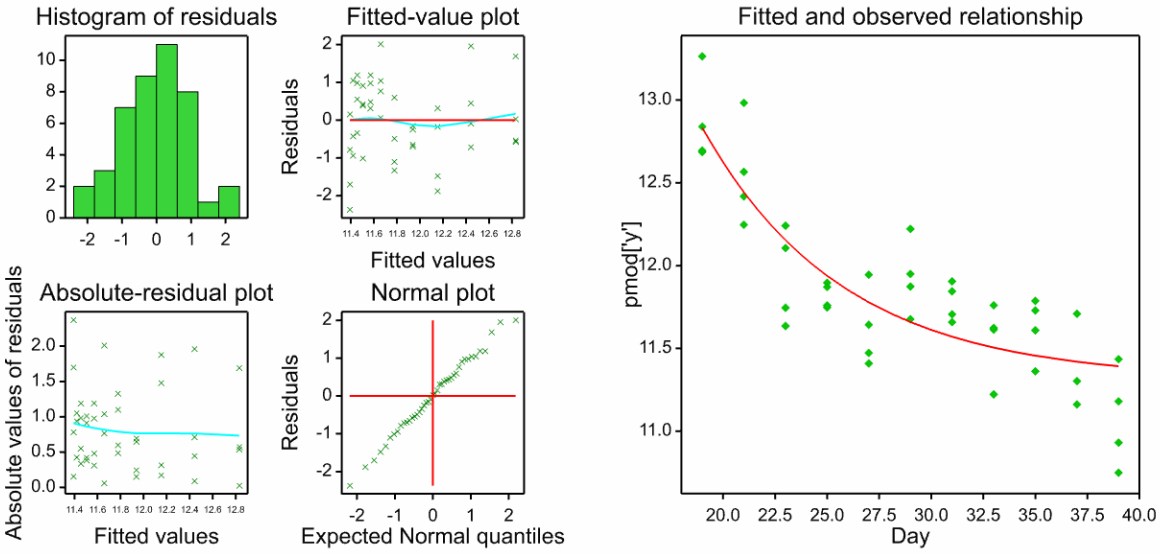

Figure S17.2.8 Exponential model for gene CATMA2A31585: (a) Composite set of residual plots based on standardized residuals; (b) fitted curve with data.

We can repeat this process for the ldl model, which takes the form:

$$Expression_i = \alpha + \frac{\beta}{1+\gamma\, Day_i} + e_i$$

with the summary ANOVA table in Table S17.2.6. This model accounts for 72.6% of the variation (adjusted R^2 = 0.726). Residual plots and the fitted model are shown in Figure S17.2.9. The fit is very similar to the exponential curve, as we might expect: there is still curvature in the fitted values plot as the fitted curve does not follow the observed trend well. The F-statistic for testing LoF takes the form

$$F = \frac{(\text{ResSS} - \text{PESS}) / (\text{ResDF} - \text{PEDF})}{\text{PEMS}} = \frac{(3.106 - 1.845) / (40 - 33)}{0.05765} = 2.734$$

with 8 and 32 df. This also gives evidence (P = 0.020) of lack of fit to the ldl model.

Both of the non-linear models we've tried show evidence of lack of fit (both formal and visual), whereas the cubic polynomial showed no evidence of lack of fit and appeared to follow the trend well. We would therefore prefer the cubic polynomial model for gene CATMA2A31585.

Table S17.2.6 ANOVA table for ldl model for gene CATMA2A31585, excluding unit 32.

Term	df	SS	Mean square	Variance ratio	*P*
Model	2	8.779	4.390	56.53	< 0.001
Residual	40	3.106			
Total	42	11.885			

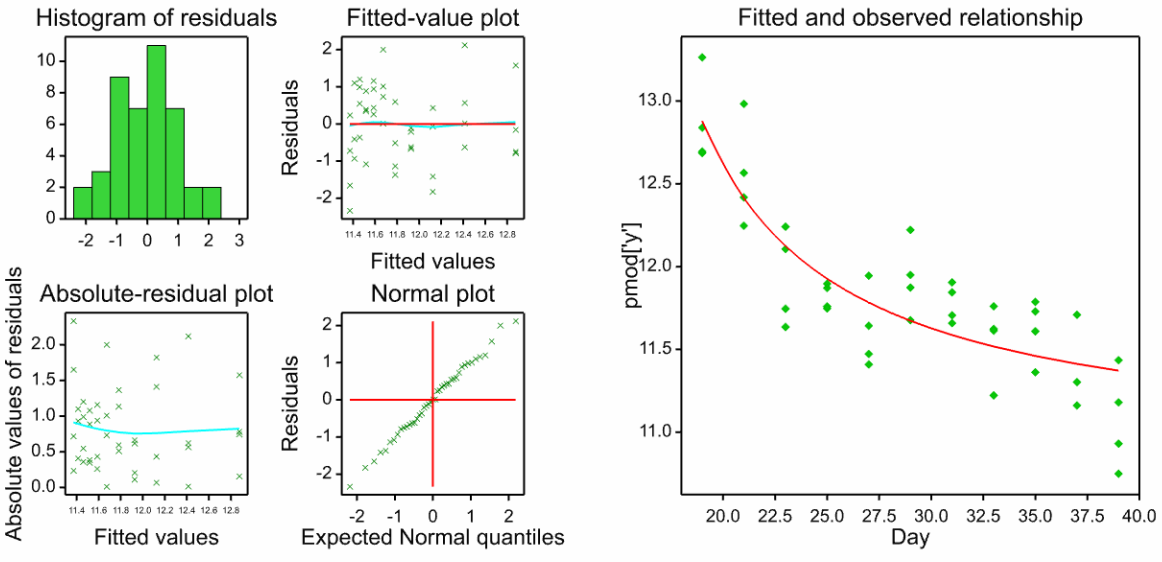

Figure S17.2.9 Ldl model for gene CATMA2A31585: (a) Composite set of residual plots based on standardized residuals; (b) fitted curve with data.

Gene CATMA1A09000

We follow the same process as above for this gene, but recognising the sigmoidal pattern, we fit the logistic and Gompertz models, again testing for LoF. From the polynomial regressions, we have the pure error (PE) estimates as PESS = 3.948, PEDF = 33, PEMS = 0.1196. We first fit the logistic model, which takes the form

$$Expression_i = \alpha + \frac{\beta}{1+\exp[-\gamma(\text{Day}_i - \delta)]} + e_i$$

with the summary ANOVA table in Table S17.2.7. This model accounts for 88.8% of the variation (adjusted R^2 = 0.888). Residual plots and the fitted model are shown in Figure S17.2.10. There is a slight suggestion of curvature in the fitted values plot for the largest fitted values, although the fitted curve follow the observed trend reasonably well. The F-statistic for testing LoF takes the form

$$F = \frac{(\text{ResSS} - \text{PESS}) / (\text{ResDF} - \text{PEDF})}{\text{PEMS}} = \frac{(5.272 - 3.948) / (41 - 33)}{0.1196} = 1.581$$

with 8 and 33 df. There is no formal evidence (P = 0.176) of lack of fit to this logistic model.

Table S17.2.7 ANOVA table for logistic model for gene CATMA1A09000.

Term	df	SS	Mean square	Variance ratio	*P*
Model	3	45.167	15.056	114.23	< 0.001
Residual	40	5.272	0.132		
Total	43	50.439			

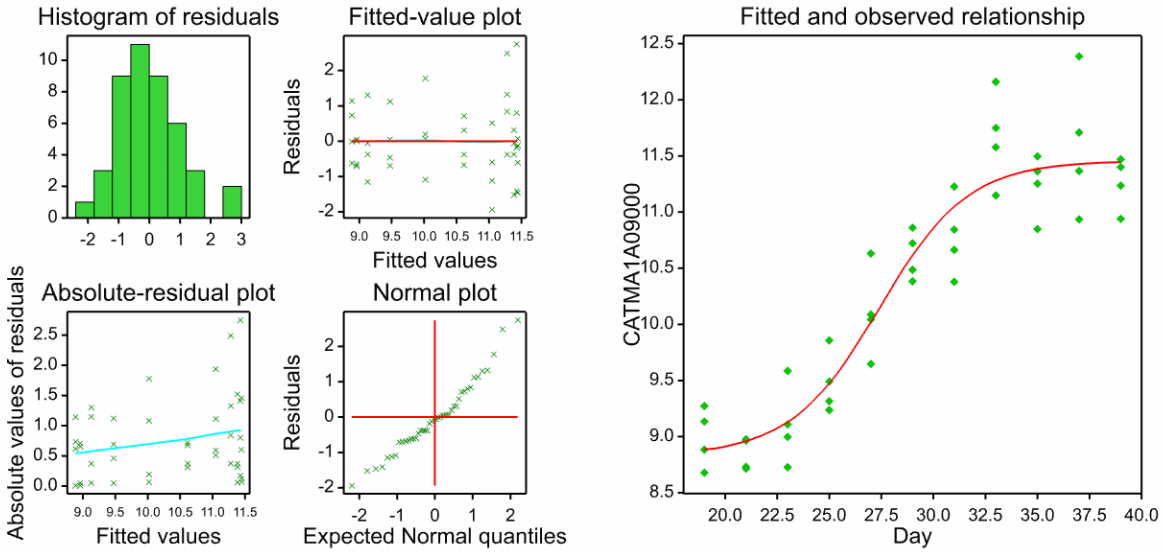

Figure S17.2.10 Logistic model for gene CATMA1A09000: (a) Composite set of residual plots based on standardized residuals; (b) fitted curve with data.

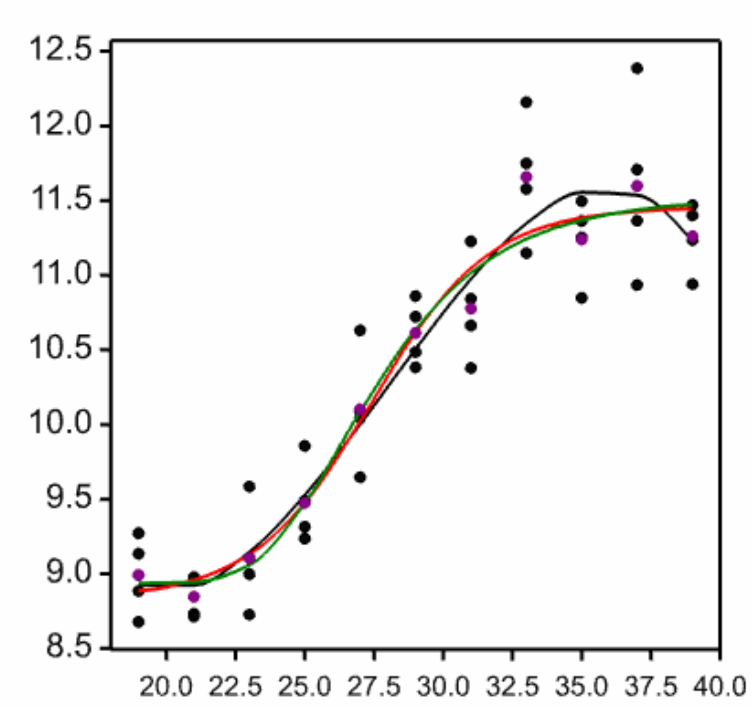

Figure S17.2.11 Observed data (black dots) for gene CATMA1A09000 with day means (purple dots) and fitted cubic (black line), logistic (red line) and Gompertz (green line) models.

The Gompertz model gives very similar results to the logistic model, and these models are plotted with the fitted cubic polynomial in Figure S17.2.11. The cubic and logistic curves are not nested, so we cannot use an F-test to quantify differences in their fit. We can calculate the AIC for each of these models as:

Cubic polynomial: AIC = 79.32
Logistic curve: AIC = 81.15
Gompertz curve: AIC = 81.18

In fact, as each of these models has 4 parameters, we could equally well compare their residual SS. Using the AIC criterion, the fit of the cubic polynomial appears slightly better than the logistic and Gompertz models. However, there is also a qualitative difference in the fits: the cubic polynomial suggests that the curve starts to decrease again at the end of the time period, whereas the logistic and Gompertz models fit an upper asymptote. Without prior knowledge (or other experimental evidence) on the behaviour of this gene, it is not possible to decide which is more plausible. All we can say is that the cubic model gives a slightly better fit.