

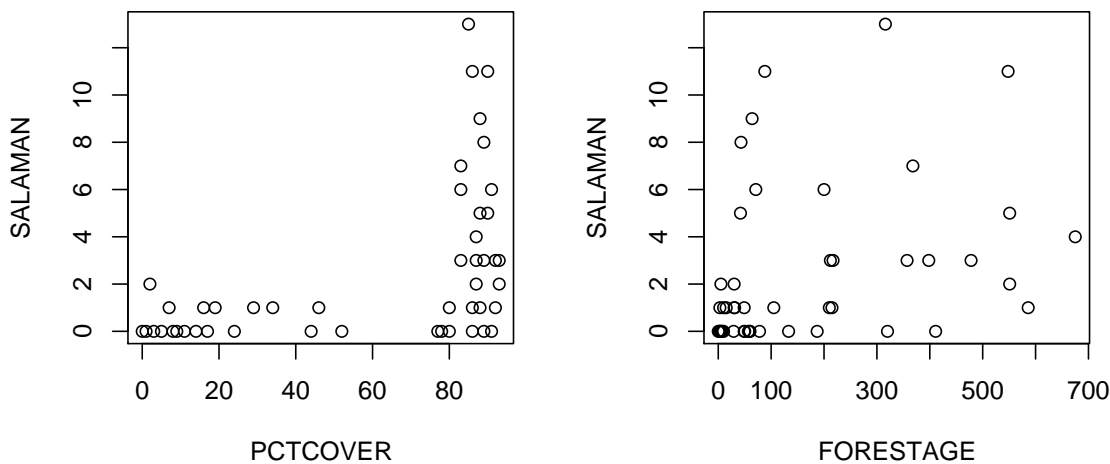
The second case study in Chapter 22 describes an observational study in which the number of salamanders in different locations are counted. The response is modeled as a Poisson random variable whose mean is a function of several explanatory variables. For this data set, *salamander count*, the response variable, takes on small interger values. The explanatory variables *percent canopy cover* and *forest age* are continuous variables. Here is a brief summary of the data set.

```
> str(case2202)

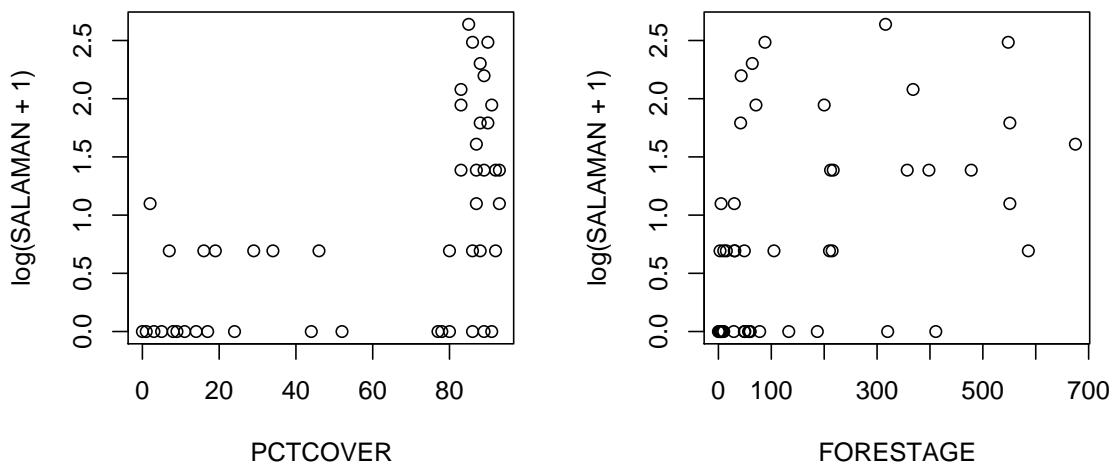
'data.frame':      47 obs. of  4 variables:
 $ SITE      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ SALAMAN   : int  13 11 11 9 8 7 6 6 5 5 ...
 $ PCTCOVER  : int  85 86 90 88 89 83 83 91 88 90 ...
 $ FORESTAGE: int  316 88 548 64 43 368 200 71 42 551 ...
```

We will begin by looking at several descriptive plots of the data. First, here are scatterplots of *salamander count* versus each *percent canopy cover* and *forest age*.

```
> par(mfrow = c(1, 2))
> plot(PCTCOVER, SALAMAN)
> plot(FORESTAGE, SALAMAN)
```



The log function is the conventional link function for Poisson regression. It would be good to plot the log of the response variables versus the explanatory variables to see the relationship. However, some counts are 0. I will plot the log of one plus the salamander counts instead.



These plots show an unusual relationship between *percent canopy cover* and *salamander count*. When the canopy coverage is less than 60 percent, the counts are always small, ranging from zero to two. When the canopy coverage is greater than 70 percent, the counts are much larger, ranging from zero to twelve. There are no sites with intermediate canopy coverage. There looks to be a great shift in variability of response when canopy coverage jumps from below 60 percent to above 70 percent.

On the other hand, there looks to be only a very weak relationship with *forest age* at either scale.

### Fitting a parameter-rich model

We will start our analysis of this data set by fitting a model with many parameters. including a dummy variable for whether or not the canopy cover is greater than 70 percent, quadratic effects and interactions. We will examine the residuals from this plot. If a parameter-rich model fits well, even if it has unnecessary variables, then the residuals ought to indicate a good fit. The model here is the one from the textbook. There are slight differences in the estimated coefficients and larger differences in the standard errors, although the deviance is computed identically.

```
> closed <- (PCTCOVER > 70)
> fit <- glm(SALAMAN ~ (PCTCOVER * FORESTAGE + I(PCTCOVER^2) +
+ I(FORESTAGE^2)) * closed, family = poisson)
> summary(fit)
```

```
Call:
glm(formula = SALAMAN ~ (PCTCOVER * FORESTAGE + I(PCTCOVER^2) +
I(FORESTAGE^2)) * closed, family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.5499	-1.0008	-0.3325	0.3949	2.4898

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.450e+00	8.750e-01	-1.657	0.097462 .

PCTCOVER	5.529e-02	1.328e-01	0.416	0.677200
FORESTAGE	4.675e-02	9.956e-02	0.470	0.638643
I(PCTCOVER^2)	-1.843e-03	4.064e-03	-0.453	0.650296
I(FORESTAGE^2)	-2.578e-03	2.880e-03	-0.895	0.370728
closedTRUE	-2.743e+02	6.447e+01	-4.254	2.10e-05 ***
PCTCOVER:FORESTAGE	2.808e-03	5.688e-03	0.494	0.621472
PCTCOVER:closedTRUE	6.447e+00	1.510e+00	4.270	1.95e-05 ***
FORESTAGE:closedTRUE	-7.931e-02	1.019e-01	-0.778	0.436312
I(PCTCOVER^2):closedTRUE	-3.626e-02	9.687e-03	-3.743	0.000182 ***
I(FORESTAGE^2):closedTRUE	2.575e-03	2.880e-03	0.894	0.371363
PCTCOVER:FORESTAGE:closedTRUE	-2.408e-03	5.693e-03	-0.423	0.672269

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 190.219 on 46 degrees of freedom  
 Residual deviance: 89.178 on 35 degrees of freedom  
 AIC: 198.24

Number of Fisher Scoring iterations: 4

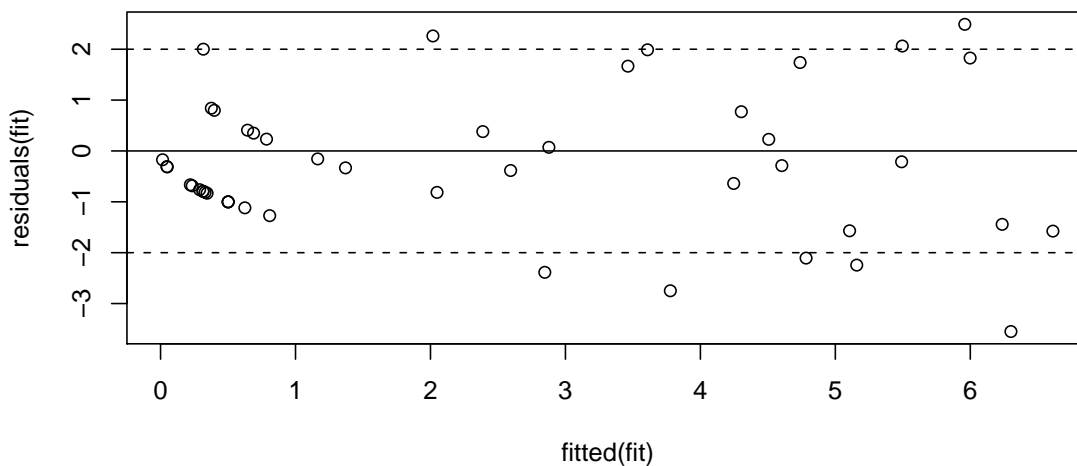
```
> 1 - pchisq(deviance(fit), df.residual(fit))
```

```
[1] 1.280605e-06
```

The last command above computes the p-value from a deviance goodness-of-fit test. This p-value is quite small which is an indicator that the model, even with all of its parameters, does not fit very well. One explanation is that there is extra-Poisson variation. The Poisson distribution assumes that the mean and variance are identical. If the actual distribution of counts had a variance larger than the mean (which could come from unmeasured variables or another level of variability in the model coefficients), the model would not fit well. This can also be apparent in a residual plot.

```
> plot(fitted(fit), residuals(fit))
> abline(h = 0)
> abline(h = 2, lty = 2)
> abline(h = -2, lty = 2)
> title("Residual plot from Poisson Regression")
```

### Residual plot from Poisson Regression



This residual plot has several deviance residuals larger than 2 in absolute value, but no real outliers. This is evidence that fitting a quasi-likelihood model is a good idea. If the deviance were large due to a small number of extreme outliers, extra-Poisson variation may not have been the best explanation.

### Eliminating Forest Age

In the previous fit, every term with *forest age* was insignificant. We can check to see if a model that drops all terms related to *forest age* fits nearly as well using a drop-in-deviance test.

```
> fit2 <- glm(SALAMAN ~ (PCTCOVER + I(PCTCOVER^2)) * closed, family = poisson)
> drop <- deviance(fit2) - deviance(fit)
> df1 <- df.residual(fit2) - df.residual(fit)
> df2 <- df.residual(fit)
> fstat <- (drop/df1)/(deviance(fit)/df2)
> fstat

[1] 0.5263966

> 1 - pf(fstat, df1, df2)

[1] 0.7843477
```

Notice that the textbook has an inconsequential typo for the p-value on page 659. It is justified to drop *forest age* from the model, and then to continue with the analysis.

### Quasi-likelihood approach

A quasi-likelihood model permits extra variation by multiplying the variance term by a factor. The fitted values will be identical to those from the Poisson log-linear regression, but the standard errors will be larger, and hence the p-values will be larger as well. In R, a quasi-likelihood model can be fit by using `family=quasipoisson` instead of `family=poisson` in the call to `glm`.

```
> fit3 <- glm(SALAMAN ~ (PCTCOVER + I(PCTCOVER^2)) * closed, family = quasipoisson)
> summary(fit2)
```

Call:

```
glm(formula = SALAMAN ~ (PCTCOVER + I(PCTCOVER^2)) * closed,
     family = poisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.3098	-0.9820	-0.7546	0.6956	2.9490

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.256e+00	8.072e-01	-1.556	0.119645
PCTCOVER	4.699e-02	8.363e-02	0.562	0.574191
I(PCTCOVER^2)	-8.328e-04	1.633e-03	-0.510	0.610081
closedTRUE	-2.186e+02	5.536e+01	-3.949	7.84e-05 ***
PCTCOVER:closedTRUE	5.054e+00	1.280e+00	3.948	7.89e-05 ***
I(PCTCOVER^2):closedTRUE	-2.852e-02	7.545e-03	-3.780	0.000157 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 190.219 on 46 degrees of freedom  
 Residual deviance: 97.225 on 41 degrees of freedom  
 AIC: 194.28

Number of Fisher Scoring iterations: 4

```
> summary(fit3)
```

Call:

```
glm(formula = SALAMAN ~ (PCTCOVER + I(PCTCOVER^2)) * closed,
     family = quasipoisson)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.3098	-0.9820	-0.7546	0.6956	2.9490

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.256e+00	1.176e+00	-1.068	0.29155
PCTCOVER	4.699e-02	1.218e-01	0.386	0.70166
I(PCTCOVER^2)	-8.328e-04	2.378e-03	-0.350	0.72804
closedTRUE	-2.186e+02	8.064e+01	-2.711	0.00974 **
PCTCOVER:closedTRUE	5.054e+00	1.865e+00	2.710	0.00977 **
I(PCTCOVER^2):closedTRUE	-2.852e-02	1.099e-02	-2.595	0.01307 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2.121420)

Null deviance: 190.219 on 46 degrees of freedom  
Residual deviance: 97.225 on 41 degrees of freedom  
AIC: NA

Number of Fisher Scoring iterations: 4

Notice that the fitted coefficients are identical, but that the standard errors and p-values are different. The quasi-likelihood fit estimates the extra variation to be more than twice as large as predicted by the Poisson distribution.