Statistics 149 - Generalized Linear Models

Mark E. Irwin
Department of Statistics
Harvard University

Spring Term

Thursday, February 2, 2006 –
Wednesday, May 24, 2006

Personnel

Instructor:	Mark Irwin
Office:	611 Science Center
Phone:	617-495-5617
E-mail:	irwin@stat.harvard.edu
Web-site:	<http://www.courses.fas.harvard.edu/~stat149/>
Lectures:	Tuesday, Thursday 10:00 - 11:30, Science Center 111
Office Hours:	Monday 1:00 - 2:00, Thursday 2:00 - 3:00, or by appointment
Teaching Fellow:	Alan Lenarcic
E-mail:	lenarcic@fas.harvard.edu
Section:	To be determined

Syllabus

- General Linear Model (Regression/ANOVA) review
- Generalized Linear Models
- Binary/Binomial responses - Logit & Probit regression
- Contingency tables - 2-, 3-, and higher-way tables
- Count data - Poisson regression
- Multicategory Logit models (Polychotomous logit and proportional odds models)
- Inference and diagnostics

References

Required text:

- Ramsey FL and Schafer DW (2002). The Statistical Sleuth: A Course in Methods of Data Analysis, 2nd edition. Duxbury. (Stat 139 text)

Other references:

- McCullagh P and Nelder JA (1989) Generalized Linear Models, 2nd edition. Chapman and Hall. (Stat 249 text)
- Agresti A (1996). An Introduction to Categorical Data Analysis. Wiley.
- Hosmer D and Lemeshow S (2000). Applied Logistic Regression 2nd edition. Wiley.
- Dobson AJ (1990). An Introduction to Generalized Linear Models. Chapman and Hall.

Grading

- Homework (30%): 5 or 6 during the term.
- Final Project (10%): Analysis of data and writing a report.
- Midterm (25%): Tuesday, March 21st, in class (Tentative).
- Final (35%): Exam Group: 12, 13. This implies the exam date should either be May 18th or May 24th.

Computing

The suggested package for the course will be **R**, a free implementation of the **S** language which has Windows, Linux, and Macintosh implementations. **S-Plus**, a commercial implementation from Insightful is also available in Windows and Linux formats.

Online material about **R**/**S-Plus** is available via the Stat Computing link on the course web site. Included on the **S-Plus**/**R** page are links to where you can download **R** for your system.

You may use other statistic packages to do your assignments. However you may not be able to get assistance from the TF or myself if you have problems.

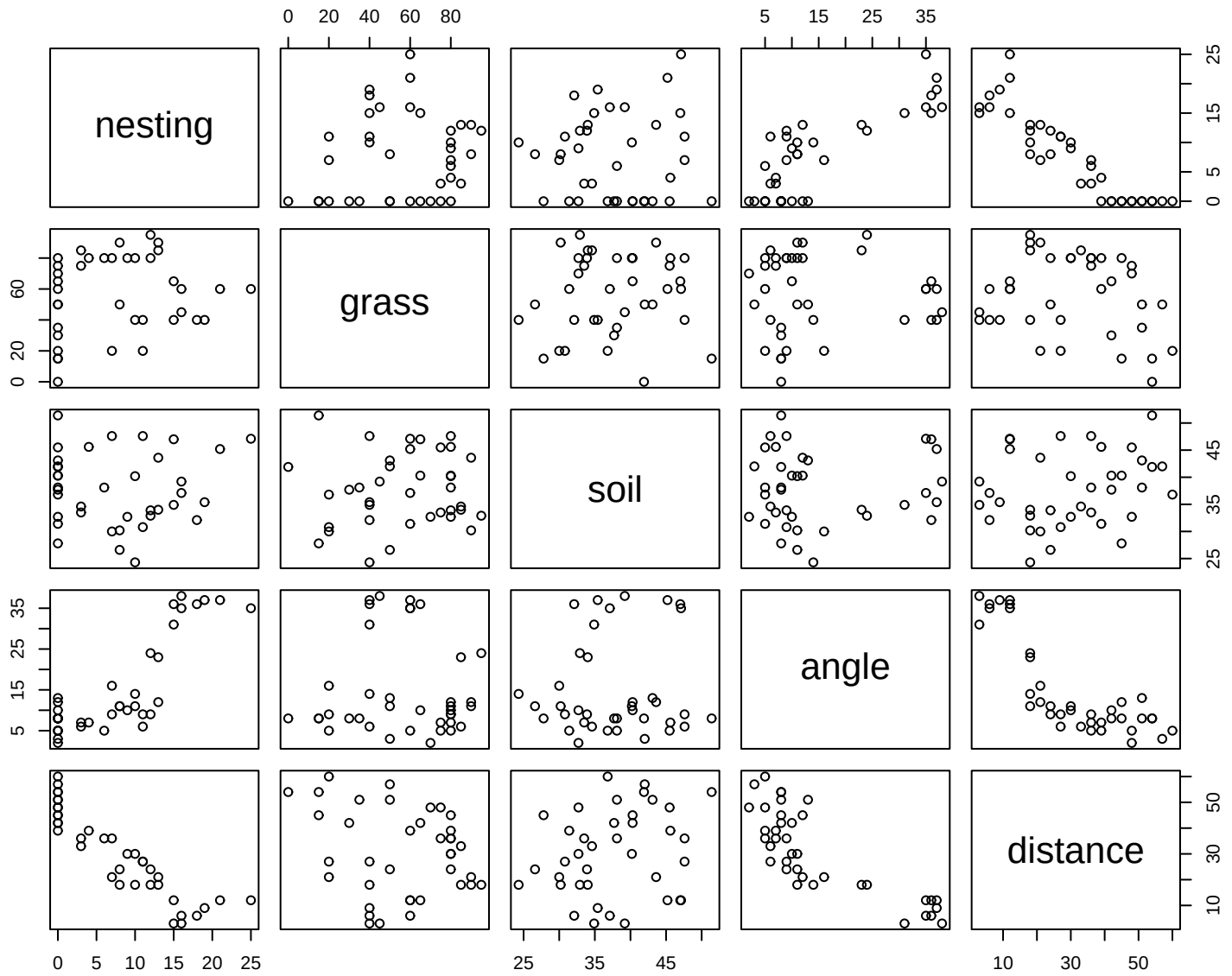
Motivating Example

Puffin Nesting: Based on the dataset from the article "Breeding Success of the Common Puffin on Difference Habitats at Great Island, Newfoundland"

Four variables where considered in trying to describe the nesting frequency of the common puffin in a $3m \times 3m$ grid of plots.

- nesting: number of nests per $9m^2$
- grass: grass cover percentage
- soil: mean soil depth in cm
- angle: angle of slope in degrees
- distance: distance from cliff edge in m

```
> summary(puffin)
    nesting           grass             soil            angle      
 Min.   : 0.000   Min.   : 0.00   Min.   :24.30   Min.   : 2.00  
 1st Qu.: 0.000   1st Qu.:40.00   1st Qu.:32.75   1st Qu.: 7.25  
 Median : 7.500   Median :60.00   Median :37.40   Median :10.00  
 Mean   : 7.684   Mean   :56.45   Mean   :37.72   Mean   :15.00  
 3rd Qu.:12.750   3rd Qu.:80.00   3rd Qu.:42.83   3rd Qu.:21.25  
 Max.   :25.000   Max.   :95.00   Max.   :51.40   Max.   :38.00  
    distance    
 Min.   : 3.00  
 1st Qu.:18.00  
 Median :30.00  
 Mean   :30.39  
 3rd Qu.:44.25  
 Max.   :60.00  
```


nesting
grass
soil
angle
distance
0 20 40 60 80
5 15 25 35
0 5 15 25
0 20 60
25 35 45
5 15 25 35
10 30 50
0 5 10 15 20 25

It appears from the scatterplot matrix, that puffins prefer to nest closer to the cliff edge, which tends to be more sloped. (Maybe its harder for predators to get to nests in these locations). In addition is appears that nesting tends to increase with increasing soil depth. Note that these comments are tentative due to the correlation between the predictors.

```
> print(cor(puffin), digits=2)
         nesting  grass  soil  angle distance
nesting    1.000  0.158 0.022  0.836    -0.91
grass      0.158  1.000 0.069 -0.017    -0.21
soil       0.022  0.069 1.000  0.066     0.21
angle      0.836 -0.017 0.066  1.000    -0.81
distance  -0.908 -0.205 0.212 -0.815     1.00
```

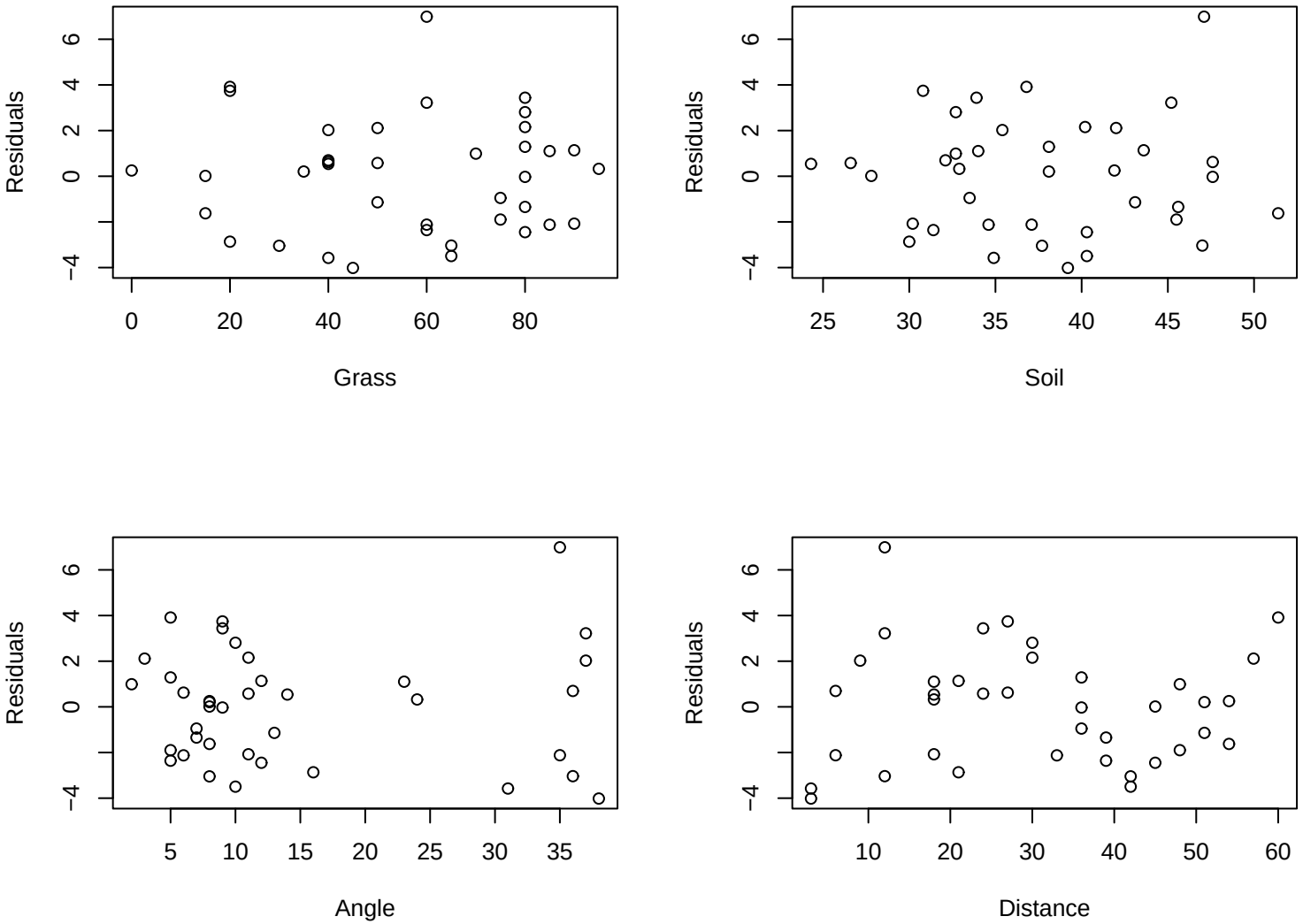
Lets fit the linear regression model where `nesting` is predicted by `grass`, `soil`, `angle`, and `distance`.

```
> summary(puffin.lm)

Call:
lm(formula = nesting ~ grass + soil + angle + distance,
  data = puffin)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0166 -2.1088  0.2293  1.2505  6.9881

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.117840   3.185028   3.177  0.00323 **
grass       -0.007408   0.019459  -0.381  0.70586
soil         0.209211   0.077238   2.709  0.01062 *
angle        0.082389   0.077796   1.059  0.29727
distance    -0.366571   0.057473  -6.378 3.18e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

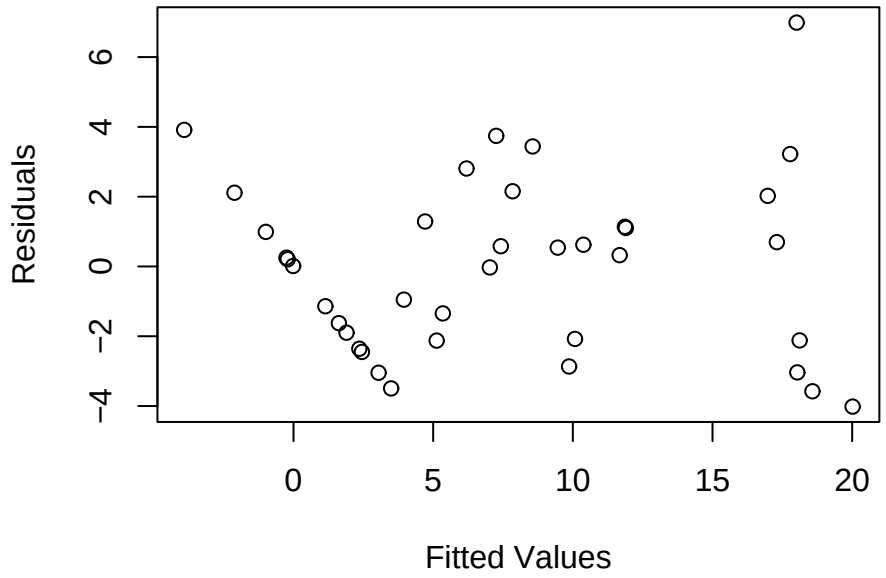
```
Residual standard error: 2.647 on 33 degrees of freedom
Multiple R-Squared: 0.8792,     Adjusted R-squared: 0.8645
F-statistic: 60.03 on 4 and 33 DF,  p-value: 1.113e-14

> anova(puffin.lm)
Analysis of Variance Table

Response: nesting
          Df  Sum Sq Mean Sq  F value    Pr(>F)
grass      1   48.08   48.08   6.8599   0.01321 *
soil       1    0.22    0.22   0.0313   0.86057
angle      1 1349.50 1349.50 192.5410 2.506e-15 ***
distance   1  285.12  285.12  40.6802 3.184e-07 ***
Residuals 33  231.29    7.01
```

Now lets look at some of the standard diagnostic plots.

Residuals vs Predictors



Generally these look ok, though maybe there is some slight evidence of nonconstant variance. We probably don't want to put much weight in this as it could be driven by a couple of observations.



This looks more problematic, particularly the straight line of points on the left side of the plot.

Lets ignore this and make forecasts at two possible nesting location configurations.

grass	soil	angle	distance	$\hat{y}$
50	35	20	15	13.22
95	25	5	60	-6.94

```
> newdata <- data.frame(grass=c(50,95), soil=c(35,25),
    angle=c(20,5), distance=c(15,60))
> predict(puffin.lm,newdata)
        1         2
13.219018 -6.938013
```

So we can get impossible forecasts using linear regression.

What model was actually being fit?

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_4 x_{4i} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$\begin{aligned} y_i | x_{1i}, \ldots, x_{4i} &\overset{ind}{\sim} N(\mu_i, \sigma^2) \\ \mu(y_i | x_{1i}, \ldots, x_{4i}) &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_4 x_{4i} = \mu_i \\ \mathrm{Var}(y_i | x_{1i}, \ldots, x_{4i}) &= \sigma^2 \end{aligned}$$

So under this model, y_i and μ_i can be any value, integer or non-integer, non-negative or negative.

What does the data look like?

Well the response variable here is a count variable, so

- $y_i \geq 0$ and must be an integer
- $\mu_i \geq 0$

So the normal linear regression model fit, can't be right. However it's possible it could be a reasonable approximation (though there are problems here).

Question: Can we find another modeling approach that satisfies the above two conditions needed for this data set.

Answer: Yes. A Generalized Linear Model could be used. One possibility is a Poisson regression model, which is an example of a generalized linear model.

Poisson Regression Model

$$
\begin{aligned}
y_i|x_{1i},\ldots,x_{4i} &\overset{ind}{\sim} Poisson(\mu_i) \\
\mu(y_i|x_{1i},\ldots,x_{4i}) &= \exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_4 x_{4i}) = \mu_i \\
\log\mu_i &= \beta_0+\beta_1 x_{1i}+\ldots+\beta_4 x_{4i} \quad \text{(equivalently)} \\
\mathrm{Var}(y_i|x_{1i},\ldots,x_{4i}) &= \mu_i
\end{aligned}
$$

This model satisfies the two conditions, the responses must be non-negative integers and the means are positive.

Lets look at some output from this model generated by **R**.

```
> puffin.glm <- glm(nesting ~ grass + soil + angle + distance,
    data=puffin, family=poisson())

> summary(puffin.glm)

Call:
glm(formula = nesting ~ grass + soil + angle + distance,
    family = poisson(), data = puffin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3263  -1.2984  -0.6617   0.8119   2.5304
```

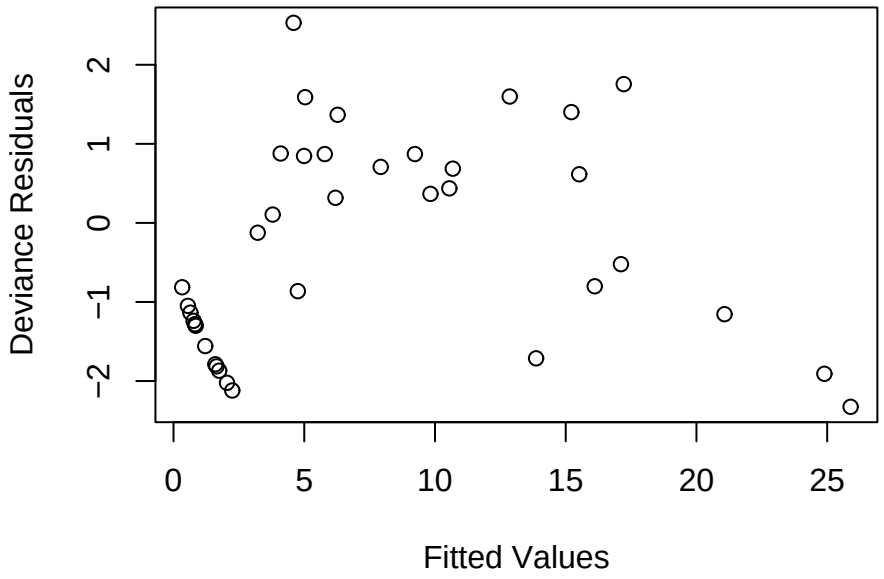
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.069973   0.452568   6.783 1.17e-11 ***
grass        0.005441   0.003104   1.753  0.07960 .
soil         0.033441   0.010822   3.090  0.00200 **
angle       -0.030077   0.010724  -2.805  0.00504 **
distance    -0.089399   0.010680  -8.371  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 310.427  on 37  degrees of freedom
Residual deviance:  68.765  on 33  degrees of freedom
AIC: 183.38

Number of Fisher Scoring iterations: 6
```

```
> anova(puffin.glm)
Analysis of Deviance Table

Model: poisson, link: log

Response: nesting

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev
NULL                        37    310.427
grass     1    6.393        36    304.033
soil      1    0.033        35    304.000
angle     1  159.343        34    144.657
distance  1   75.892        33     68.765
```

```
> predict(puffin.glm, newdata, type="response")
         1          2
13.0648650  0.3357275
```

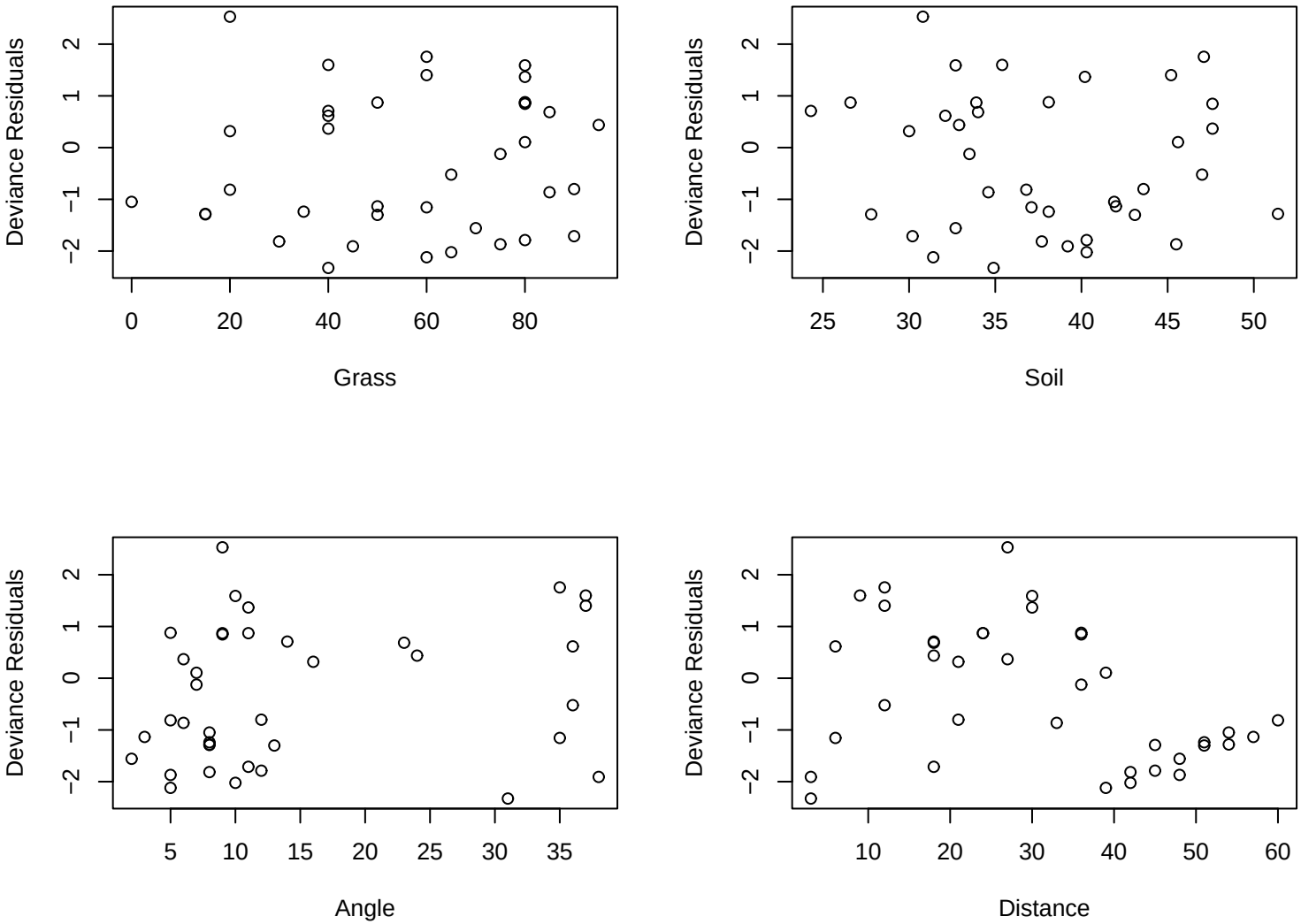
grass	soil	angle	distance	Normal $\hat{y}$	Poisson $\hat{\mu}$
50	35	20	15	13.22	13.06
95	25	5	60	-6.94	0.34

Lets look at a couple of residual plots. Note that the residuals being used are not standard residuals, but what are known as deviance residuals. However the goal with these plots is the same. Are there any patterns in the plots (hopefully not)?



Not too bad. Note that the line effect is here again, as it must be given so many observations with 0 nests. The following plots again don't look too bad

Deviance Residuals vs Predictors



Lets take a second look at the two models fit

- Normal model:

$$\begin{aligned} y_i|x_{1i},\ldots,x_{4i} &\overset{ind}{\sim} N(\mu_i,\sigma^2) \\ \mu(y_i|x_{1i},\ldots,x_{4i}) &= \beta_0+\beta_1 x_{1i}+\ldots+\beta_4 x_{4i}=\mu_i \\ \mathrm{Var}(y_i|x_{1i},\ldots,x_{4i}) &= \sigma^2 \end{aligned}$$

- Poisson model:

$$\begin{aligned} y_i|x_{1i},\ldots,x_{4i} &\overset{ind}{\sim} Poisson(\mu_i) \\ \mu(y_i|x_{1i},\ldots,x_{4i}) &= \exp(\beta_0+\beta_1 x_{1i}+\ldots+\beta_4 x_{4i})=\mu_i \\ \log\mu_i &= \beta_0+\beta_1 x_{1i}+\ldots+\beta_4 x_{4i} \\ \mathrm{Var}(y_i|x_{1i},\ldots,x_{4i}) &= \mu_i \end{aligned}$$

Differences:

- Additive vs Multiplicative models:

  The normal regression model is an example of an additive model. Changing x_1 say by 1 leads to μ shifting by β_1.

  The Poisson regression model is an example of a multiplicative model. Changing x_1 say by 1 leads to μ being multiplied by e^{β_1}.

  As we will see during the term, the generalized linear model will allow for both type of mean models to be used in a wide range of situations (plus others)

- Variability

  In the normal model, the residual variance can be anything (σ^2 is arbitrary). However in the Poisson case, the variance is fixed by the mean.

For the example, there is some evidence that this assumption isn't reasonable. Instead there is evidence of overdispersion where $\text{Var}(y_i|x_{1i}, \ldots, x_{4i}) > \mu_i$.

One approach is a quasi-likelihood approach where the following modeling assumptions are made

$$\begin{aligned} \mu(y_i|x_{1i}, \ldots, x_{4i}) &= \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_4 x_{4i}) = \mu_i \\ \log \mu_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \beta_4 x_{4i} \\ \text{Var}(y_i|x_{1i}, \ldots, x_{4i}) &= \phi\mu_i; \quad \phi \geq 1 \end{aligned}$$

If we fit this model to the data, an estimate of ϕ is 1.71.

```
> puffin.qglm <- glm(nesting ~ grass + soil + angle + distance,
    data=puffin, family=quasipoisson())
```

```
> summary(puffin.qglm)

Call:
glm(formula = nesting ~ grass + soil + angle + distance,
    family = quasipoisson(), data = puffin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3263  -1.2984  -0.6617   0.8119   2.5304

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.069973   0.592049   5.185 1.07e-05 ***
grass        0.005441   0.004060   1.340   0.1894
soil         0.033441   0.014157   2.362   0.0242 *
angle       -0.030077   0.014029  -2.144   0.0395 *
distance    -0.089399   0.013971  -6.399 3.00e-07 ***

(Dispersion parameter for quasipoisson family taken to be 1.711)
```

```
    Null deviance: 310.427  on 37  degrees of freedom
Residual deviance:  68.765  on 33  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

By making this change to the model, our inference changes a bit. The t-tests on the regression parameters are not as significant, though in this case most people will make the same conclusion.