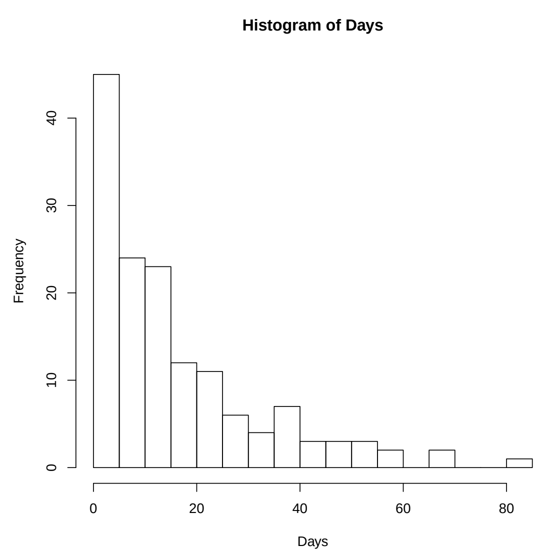
Slide 1

Example: Let's consider the Quine dataset from the MASS package, which reports absenteeism from school in rural New South Wales. Data on 146 children from Walgett, New South Wales, Australia, were obtained. The outcome of interest is the number of days absent from school in a particular school year. The variables in the data are:

- `Eth` – ethnic background: Aboriginal or Not ("A" or "N").
- `Sex` – Female or Male ("F" or "M").
- `Age` – age group: Primary ("F0"), or forms "F1," "F2" or "F3".
- `Lrn` – learner status: factor with levels Average or Slow learner ("AL" or "SL").
- `Days` – days absent from school in the year.

Slide 2

```
attach(quine)
hist(Days, breaks=20)
```



Slide 3

```
# main effects model
fit.main = glm(Days ~ Age+Sex+Eth+Lrn, data=quine, family="poisson")

> summary(fit.main)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.71538    0.06468  41.980  < 2e-16 ***
AgeF1       -0.33390    0.07009  -4.764 1.90e-06 ***
AgeF2        0.25783    0.06242   4.131 3.62e-05 ***
AgeF3        0.42769    0.06769   6.319 2.64e-10 ***
SexM         0.16160    0.04253   3.799 0.000145 ***
EthN        -0.53360    0.04188 -12.740  < 2e-16 ***
LrnSL        0.34894    0.05204   6.705 2.02e-11 ***
---
(Dispersion parameter for poisson family taken to be 1)
```

Slide 4

```
    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1696.7  on 139  degrees of freedom
AIC: 2299.2

> 1-pchisq(deviance(fit.main), df.residual(fit.main))
[1] 0
```

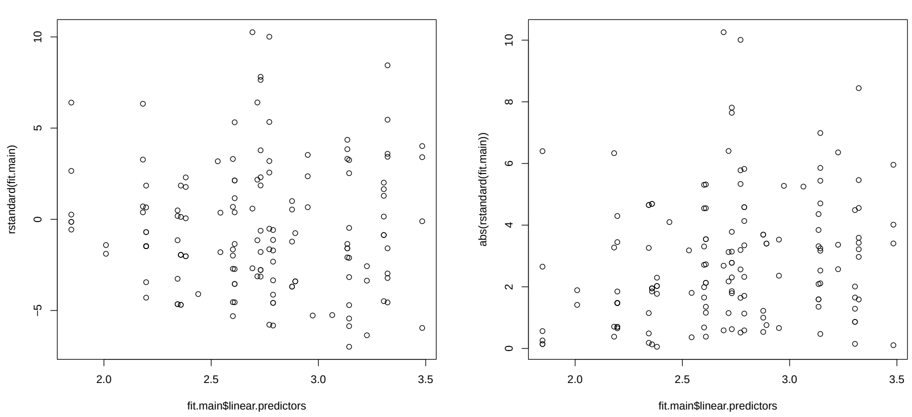
Slide 5

Check for outliers and influential points:

- The plot of standardized residuals versus linear predictors is randomly scattered around 0.
- Many of the standardized residuals have magnitude greater than 3.
- There are 3 potential influential observations based on Cook's distance.

```
plot(fit.main$linear.predictors, rstandard(fit.main))
plot(fit.main$linear.predictors, abs(rstandard(fit.main)))
```

Slide 6



Slide 7

```
> summary(influence.measures(fit.main))
Potentially influential observations of
 glm(formula = Days ~ Age + Sex + Eth + Lrn, family = "poisson",      data = quine) :

    dfb.1_ dfb.AgF1 dfb.AgF2 dfb.AgF3 dfb.SexM dfb.EthN dfb.LrSL dffit   cov.r   cook.d  hat
46   0.06   0.21    -0.04     0.02    -0.10    -0.17     0.10     0.46    0.85_*  0.60    0.04
59   0.11  -0.10     0.18     0.02    -0.25    -0.21     0.18     0.61    0.83_*  0.96_*  0.06
72   0.15  -0.50    -0.55    -0.29     0.23     0.31     0.41     0.81_*  0.72_*  2.02_*  0.07
104 -0.21   0.03     0.02     0.33     0.27     0.31     0.05     0.63    0.72_*  1.19_*  0.04
```

Slide 8

Check for over/underdispersion:

```
> deviance(fit.main, type="pearson")/df.residual(fit.main)
[1] 12.20652
fit.quasi = glm(Days ~ Age+Sex+Eth+Lrn, data=quine, family=quasi(link="log", variance="mu"))
> summary(fit.quasi)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7154     0.2347  11.569  < 2e-16 ***
AgeF1        -0.3339     0.2543  -1.313 0.191413
AgeF2         0.2578     0.2265   1.138 0.256938
AgeF3         0.4277     0.2456   1.741 0.083831 .
SexM          0.1616     0.1543   1.047 0.296914
EthN         -0.5336     0.1520  -3.511 0.000602 ***
LrnSL         0.3489     0.1888   1.848 0.066760 .
---
(Dispersion parameter for quasi family taken to be 13.16692)
```

Slide 9

Check for zero-inflation:

```
library(pscl)
fit.zip = zeroinfl(Days ~ Age+Sex+Eth+Lrn , data=quine, dist="poisson")

> summary(fit.zip)

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.71883    0.06480  41.956  < 2e-16 ***
AgeF1       -0.32048    0.06968  -4.599 4.24e-06 ***
AgeF2        0.24602    0.06212   3.960 7.49e-05 ***
AgeF3        0.43721    0.06781   6.447 1.14e-10 ***
SexM         0.18904    0.04253   4.445 8.78e-06 ***
EthN        -0.44061    0.04190 -10.517  < 2e-16 ***
LrnSL        0.34400    0.05155   6.674 2.50e-11 ***
```

Slide 10

```
Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7229     0.3442  -7.912 2.53e-15 ***
---
Log-likelihood: -1051 on 8 Df
```

Slide 11

Try a negative-binomial model:

```
fit.nb = glm.nb(Days ~ Age+Sex+Eth+Lrn, data=quine)

> summary(fit.nb)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.89458    0.22842  12.672  < 2e-16 ***
AgeF1       -0.44843    0.23975  -1.870 0.061425 .
AgeF2        0.08808    0.23619   0.373 0.709211
AgeF3        0.35690    0.24832   1.437 0.150651
SexM         0.08232    0.15992   0.515 0.606710
EthN        -0.56937    0.15333  -3.713 0.000205 ***
LrnSL        0.29211    0.18647   1.566 0.117236
---
(Dispersion parameter for Negative Binomial(1.2749) family taken to be 1)
```

Slide 12

```
    Null deviance: 195.29  on 145  degrees of freedom
Residual deviance: 167.95  on 139  degrees of freedom
AIC: 1109.2

              Theta:  1.275
          Std. Err.:  0.161

 2 x log-likelihood:  -1093.151

> 1-pchisq(deviance(fit.nb), df.residual(fit.nb))
[1] 0.04765619
```

Slide 13

Try a zero-inflated negative binomial model:

```
fit.zinb = zeroinfl(Days ~ Age+Sex+Eth+Lrn | 1 , data=quine, dist="negbin")

> summary(fit.zinb)
Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.89279    0.21880  13.221  < 2e-16 ***
AgeF1       -0.44524    0.22868  -1.947 0.051536 .
AgeF2        0.08903    0.23267   0.383 0.701988
AgeF3        0.36497    0.23840   1.531 0.125792
SexM         0.09974    0.15987   0.624 0.532691
EthN        -0.53585    0.15363  -3.488 0.000487 ***
LrnSL        0.29523    0.17570   1.680 0.092904 .
Log(theta)   0.37516    0.15751   2.382 0.017232 *

Zero-inflation model coefficients (binomial with logit link):
```

Slide 14

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.5621     0.8726  -4.082 4.46e-05 ***
---
Theta = 1.4552
Log-likelihood: -545.8 on 9 Df
```

Slide 15

Based on AIC we would opt for the negative binomial model.

```
> AIC(fit.main, fit.zip, fit.nb, fit.zinb)
         df      AIC
fit.main  7 2299.184
fit.zip   8 2117.268
fit.nb    8 1109.151
fit.zinb  9 1109.622
```

Slide 16

Several of the covariates do not appear significant in the negative binomial model. Let's refit the model with only `Age` and `Eth` as predictors and perform a likelihood ratio test:

```
fit.nb2 = glm.nb(Days ~ Age+Eth, data=quine)

> summary(fit.nb2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.0382     0.1957  15.529  < 2e-16 ***
AgeF1        -0.3855     0.2274  -1.695 0.090019 .
AgeF2         0.1846     0.2313   0.798 0.424744
AgeF3         0.2550     0.2407   1.060 0.289332
EthN         -0.5611     0.1547  -3.628 0.000286 ***
---
(Dispersion parameter for Negative Binomial(1.2491) family taken to be 1)
```

Slide 17

```
   Null deviance: 192.04  on 145  degrees of freedom
Residual deviance: 167.84  on 141  degrees of freedom
AIC: 1107.8

              Theta:  1.249
          Std. Err.:  0.157

 2 x log-likelihood:  -1095.801

> 1-pchisq(-2*(logLik(fit.nb2)-logLik(fit.nb)), 2)
'log Lik.' 0.2657784 (df=6)

> AIC(fit.nb, fit.nb2)
        df      AIC
fit.nb   8 1109.151
fit.nb2  6 1107.801
```

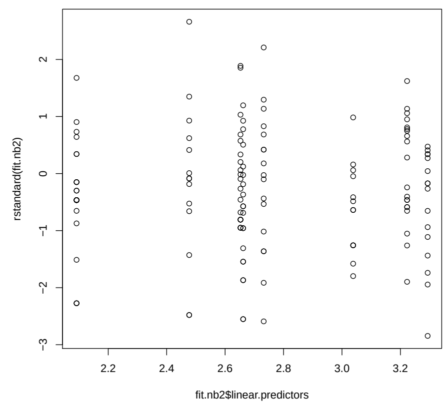
Slide 18

Let's check the residuals plot and identify outliers and influential points:

```
plot(fit.nb2$linear.predictors, rstandard(fit.nb2))
> which(abs(rstandard(fit.nb2)) > 3)
named integer(0)
> summary(influence.measures(fit.nb2))
Potentially influential observations of
 glm.nb(formula = Days ~ Age + Eth, data = quine, init.theta = 1.249142793,      link = log) :

    dfb.1_ dfb.AgF1 dfb.AgF2 dfb.AgF3 dfb.EthN dffit cov.r   cook.d hat
61  -0.09  -0.01     0.00    -0.32     0.23    -0.53  0.84_*  0.01   0.04
72   0.36  -0.39    -0.37    -0.36     0.21     0.53  0.87_*  0.23   0.04
92   0.08   0.01    -0.24     0.00    -0.20    -0.43  0.88_*  0.01   0.03
98   0.08   0.01     0.00    -0.28    -0.20    -0.47  0.88_*  0.01   0.04
127  0.08   0.01    -0.24     0.00    -0.20    -0.43  0.88_*  0.01   0.03
```

Slide 19


rstandard(fit.nb2)
fit.nb2$linear.predictors
2.2
2.4
2.6
2.8
3.0
3.2
−3
−2
−1
0
1
2