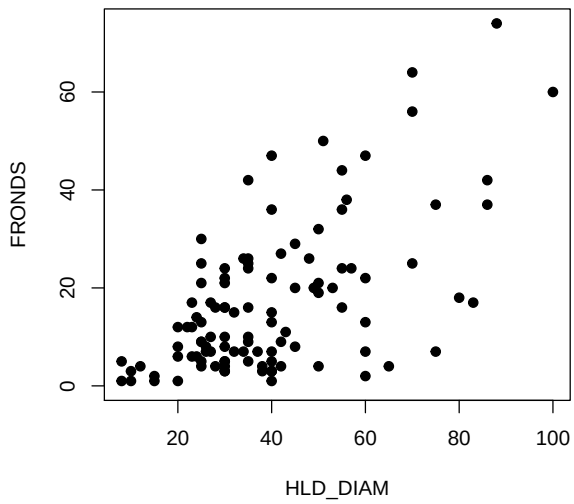
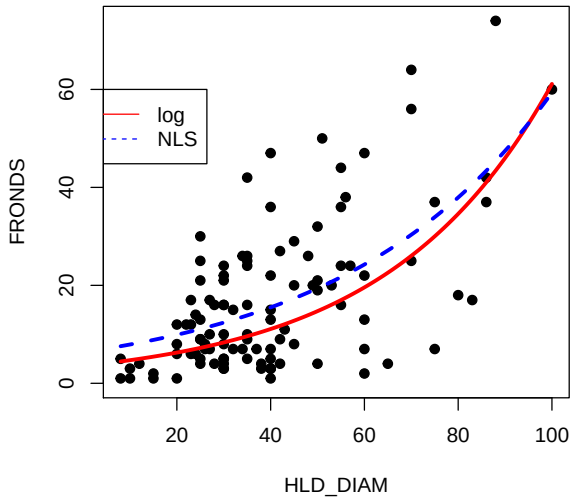
Generalized Linear Models

The Kelp Data

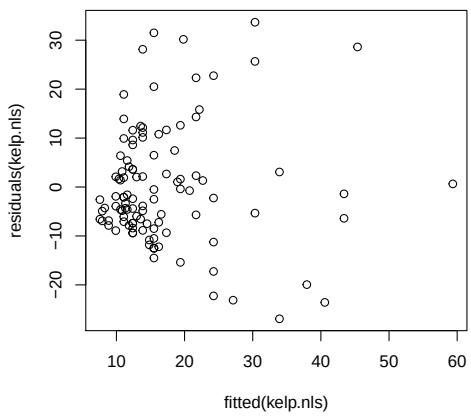


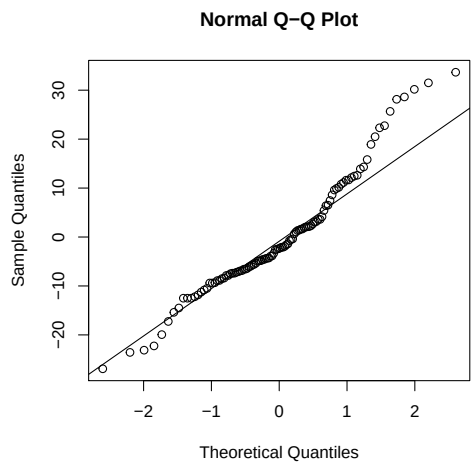
FRONDS are a count variable, cannot be < 0

Nonlinear Fits!



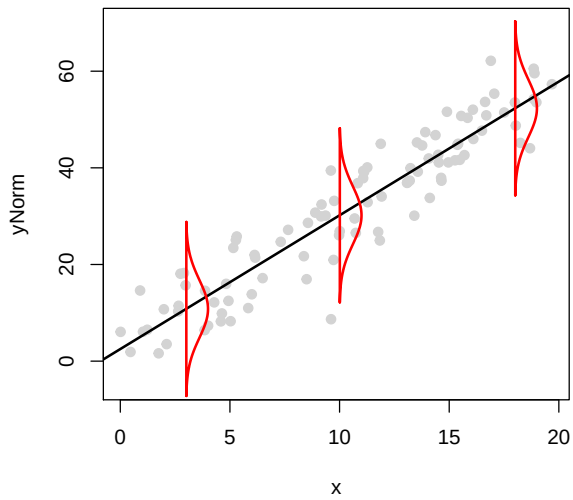
QQ Violations even in NLS



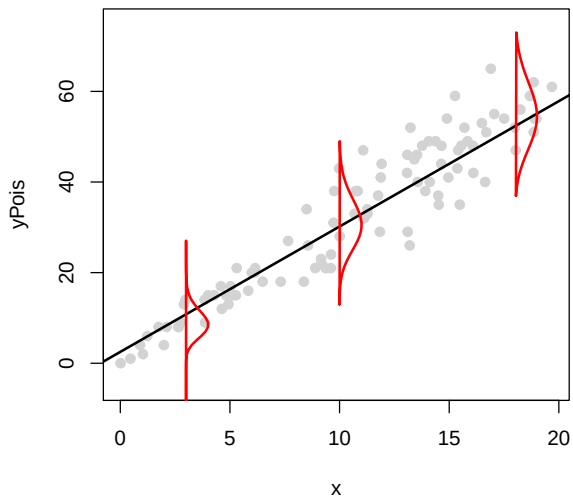


Maybe the error is wrong...

We've had a Normal Journey

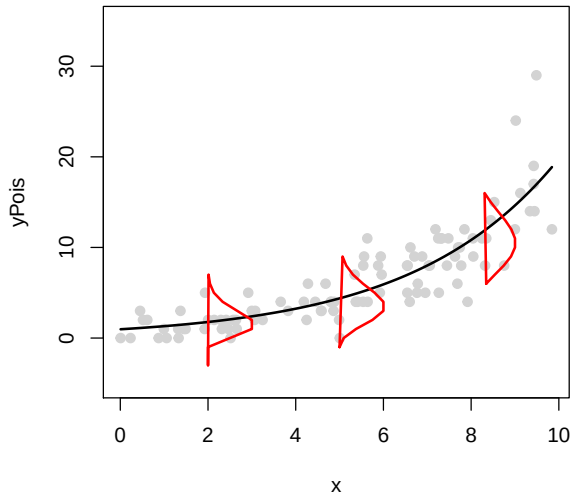


But Now it's Time To Swim Forward



This model has a Poisson error!

To the world of Generalized Linear Models



Poisson Error, Log Link (exponential function)

Generalized Linear Models: Link Functions

Basic Premise:

1. We have a linear predictor, $\eta_i = a + Bx$
2. That predictor is linked to the fitted value of Y_i, μ_i
3. We call this a link function, such that $g(\mu_i) = \eta_i$
   - For example, for a linear function, $\mu_i = \eta_i$
   - For an exponential function, $log(\mu_i) = \eta_i$

Some Common Links

- Identity: $\mu = \eta$ - e.g. $\mu = a + bx$
- Log: $log(\mu) = \eta$ - e.g. $\mu = e^{a+bx}$
- Logit: $logit(\mu) = \eta$ - e.g. $\mu = \frac{e^{a+bx}}{1+e^{a+bx}}$
- Inverse: $\frac{1}{\mu} = \eta$ - e.g. $\mu = (a + bx)^{-1}$

Generalized Linear Models: Error

Basic Premise:

1. The error distribution is from the *e*xponential family
   - e.g., Normal, Poisson, Binomial, and more.

2. For these distributions, the variance is a funciton of the fitted value on the curve: $var(Y_i) = \theta V(\mu_i)$
   - For a normal distribution, $var(\mu_i) = \theta * 1$ as $V(\mu_i) = 1$
   - For a poisson distribution, $var(\mu_i) = 1 * \mu_i$ as $V(\mu_i) = \mu_i$

Distributions, Canonical Links, and Dispersion

Distribution	Canonical Link	Variance Function
Normal	identity	θ
Poisson	log	μ
Quasipoisson	log	$\mu\theta$
Binomial	logit	$\mu(1-\mu)$
Quasibinomial	logit	$\mu(1-\mu)\theta$
Negative Binomial	log	$\mu+\kappa\mu^2$
Gamma	inverse	μ^2
Inverse Normal	$1/\mu^2$	μ^3

Distributions and Other Links

Distribution	Links
Normal	identity, log, inverse
Poisson	log, identity, sqrt
Quasipoisson	log, identity, sqrt
Binomial	logit, probit, cauchit, log, log-log
Quasibinomial	logit, probit, cauchit, log, log-log
Negative Binomial	log, identity, sqrt
Gamma	inverse, identity, log
Inverse Normal	$1/\mu^2$, inverse, identity, log

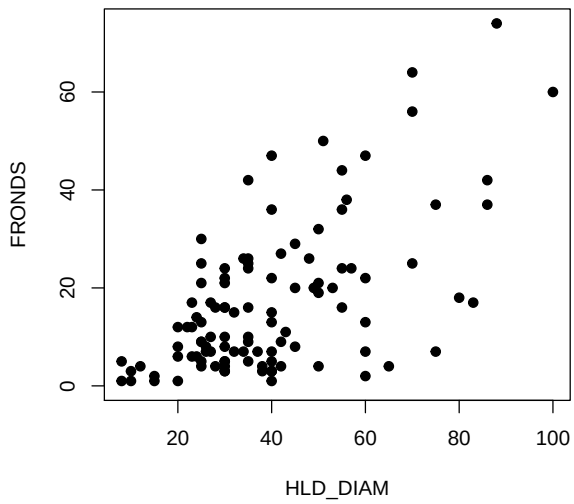
Fitting: Deviance and IWLS

Every GLM has a Deviance Function to be Minimized,
$2\theta(LL_{sat} - LLmod)$

i.e., for a normal distribution $D_M = \sum(y_i - \hat{\mu}_i)^2$

Models are Fit using Iteratively Weighted Least Squares algorithm
Confidence intervals are determined assuming a quadratic
Log-Likelihood Surface

The Kelp Data



FRONDS are a count variable, cannot be < 0

Fitting a GLM with a Poisson Error and Log Link

Fronds $\sim$ Poisson($\hat{Fronds}$)

$\hat{Fronds}$ = exp(a + b * holdfast diameter)

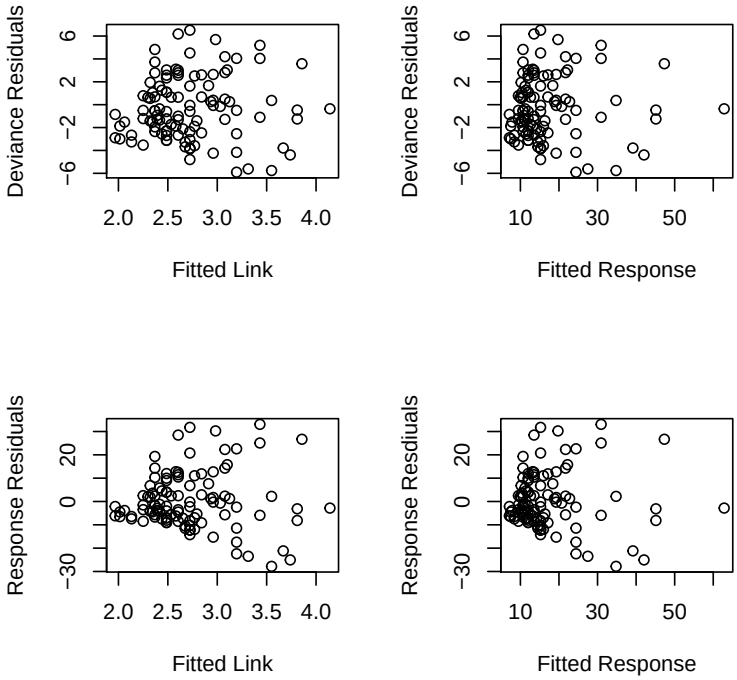
```
kelp.glm <- glm(FRONDS ~ HLD_DIAM, data=kelp,
                family=poisson(link="log"))
```


Different Types of Residuals

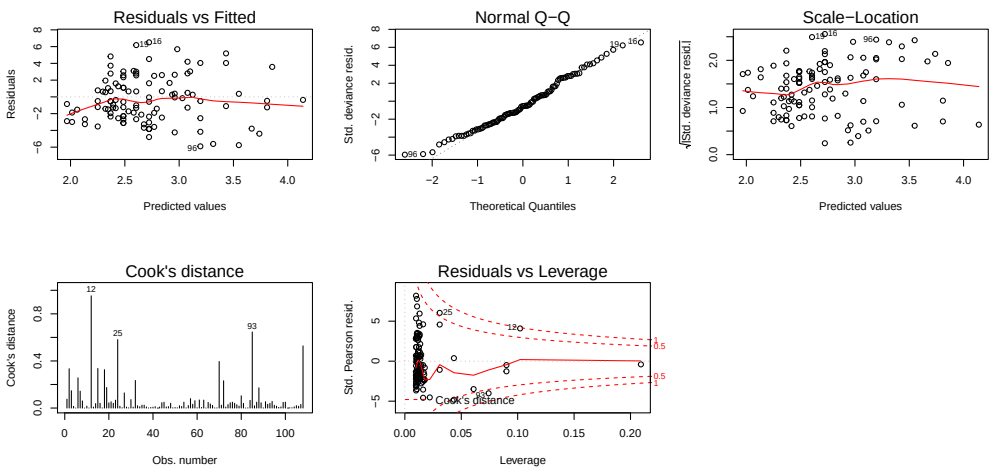
```
residuals(kelp.glm, type="deviance")
residuals(kelp.glm, type="pearson")
residuals(kelp.glm, type="response")
```

Deviance residuals are based on the density of an observation given its fit estimate
Response residuals are $Y_i - \mu_i$

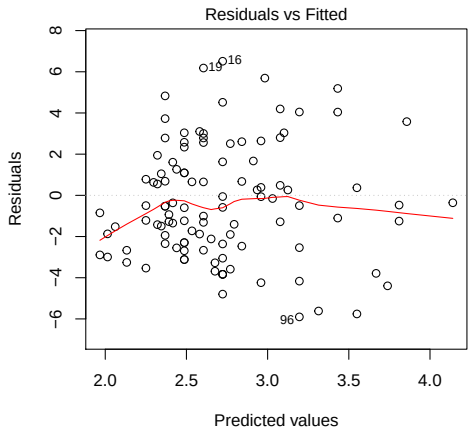
Evaluating Residuals and Fits

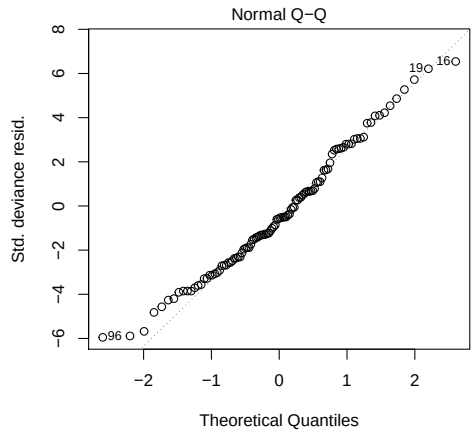


How do we Assess Meeting Assumptions?



How do we Assess Meeting Assumptions?





GLM Model Coefficients

```
#
# Call:
# glm(formula = FRONDS ~ HLD_DIAM, family = poisson(link = "log"),
#     data = kelp)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -5.9021  -2.3871  -0.5574   1.6132   6.5117
#
# Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept)  1.77806    0.05726   31.05   <2e-16
# HLD_DIAM     0.02362    0.00105   22.50   <2e-16
#
# (Dispersion parameter for poisson family taken to be 1)
#
#     Null deviance: 1289.17  on 107  degrees of freedom
# Residual deviance:  832.56  on 106  degrees of freedom
#   (32 observations deleted due to missingness)
```

Checking Fit - R^2_{corr}

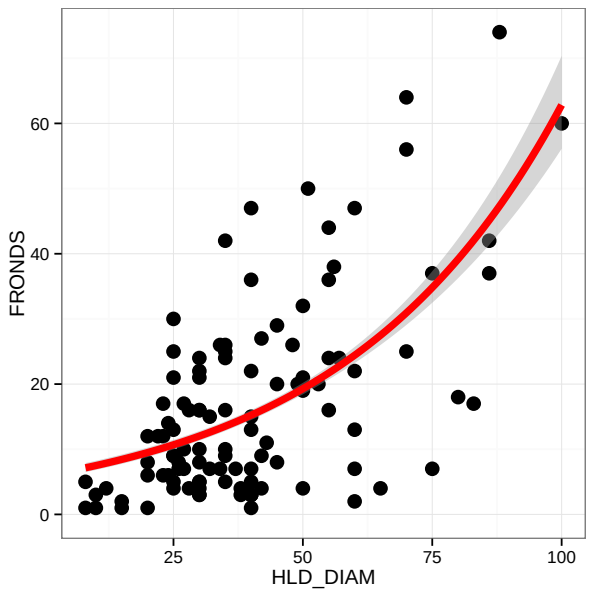
```
cor(fitted(kelp.glm),
    fitted(kelp.glm) + residuals(kelp.glm, type="response"))^2

# [1] 0.3649143

summary(kelp.lm)$r.squared

# [1] 0.2769708
```

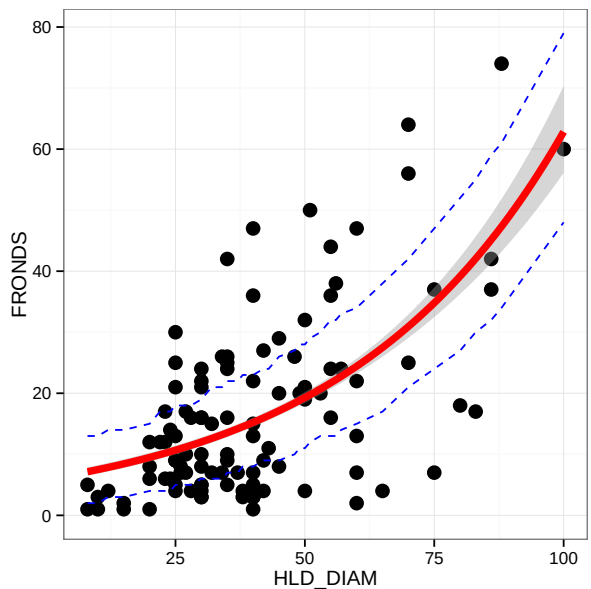

The Fitted Model



Prediction Confidence Intervals by Hand

```
upperCI <- qpois(0.975, lambda = round(fitted(kelp.glm)))
lowerCI <-  qpois(0.025, lambda = round(fitted(kelp.glm)))
HLD <- na.omit(kelp)$HLD_DIAM
#
kelp.ggplot +
  geom_line(mapping=aes(x=HLD,  y=upperCI), lty=2, col="blue") +
  geom_line(mapping=aes(x=HLD,  y=lowerCI), lty=2, col="blue")
```


Prediction Confidence Intervals by Hand



Overdispersion?

What is Overdispersion?

Greater variance than expected based on a modeled distribution.

- Since the variance increases faster than the mean, our variability is overdispersed
- This can be solved with different distributions whose variance have different properties
- OR, we can fit a model, then scale it's variance posthoc with a coefficient
- The likelihood of these latter models is called a Quasi-likelihood, as it does not reflect the true spread of the data

Which Overdispersed Distribution to Use?

Var(Negative Binomial) $= \mu + \kappa\mu^2$

Var(quasipoisson) $= \mu\theta$

Both would work for the link function and data type we have in this example (count data)
Which to use?
see Ver Hoef and Boveng 2007 Ecology

GLM with Negative Binomial

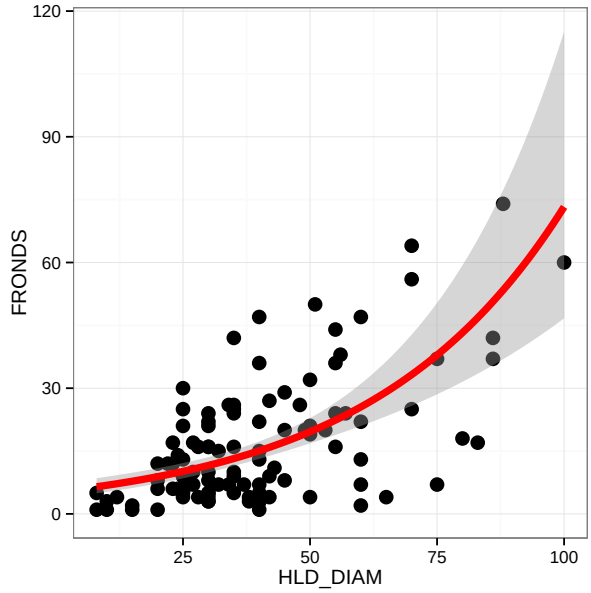
```
library(MASS)
#
kelp.glm.nb <- glm.nb(FRONDS ~ HLD_DIAM, data=kelp)
```


Negative Binomial Performs Better

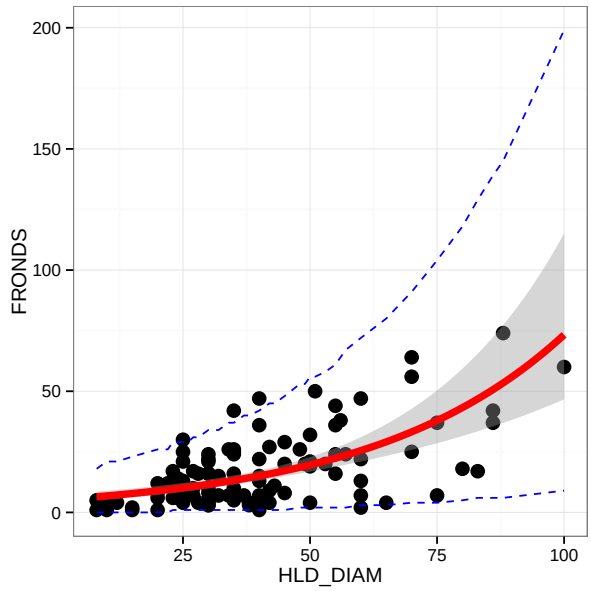
```
anova(kelp.glm, kelp.glm.nb)

# Analysis of Deviance Table
#
# Model 1: FRONDS ~ HLD_DIAM
# Model 2: FRONDS ~ HLD_DIAM
#   Resid. Df Resid. Dev Df Deviance
# 1       106     832.56
# 2       106     114.49  0   718.07
```

The Fitted Model



Fit with Prediction Error



QuasiPoisson Fit

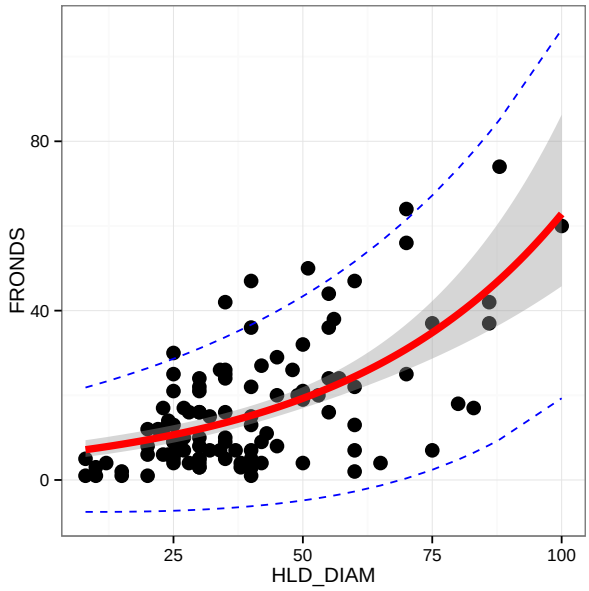
```
kelp.glm2 <- glm(FRONDS ~ HLD_DIAM, data=kelp,
                 family=quasipoisson(link="log"))
```


QuasiPoisson Summary with Dispersion Parameter

```
summary(kelp.glm2)

#
# Call:
# glm(formula = FRONDS ~ HLD_DIAM, family = quasipoisson(link = "log"),
#     data = kelp)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -5.9021  -2.3871  -0.5574   1.6132   6.5117
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 1.778059   0.160455  11.081  < 2e-16
# HLD_DIAM    0.023624   0.002943   8.027 1.45e-12
#
# (Dispersion parameter for quasipoisson family taken to be 7.852847)
....
```

Compare to Quasipoisson with Prediction Error



Example: Wolf Inbreeding and Litter Size: The Final Analysis

- The Number of Pups is a Count!
- Fit GLMs with different errors and links
- Which is the best model?
- Plot with fit and prediction error

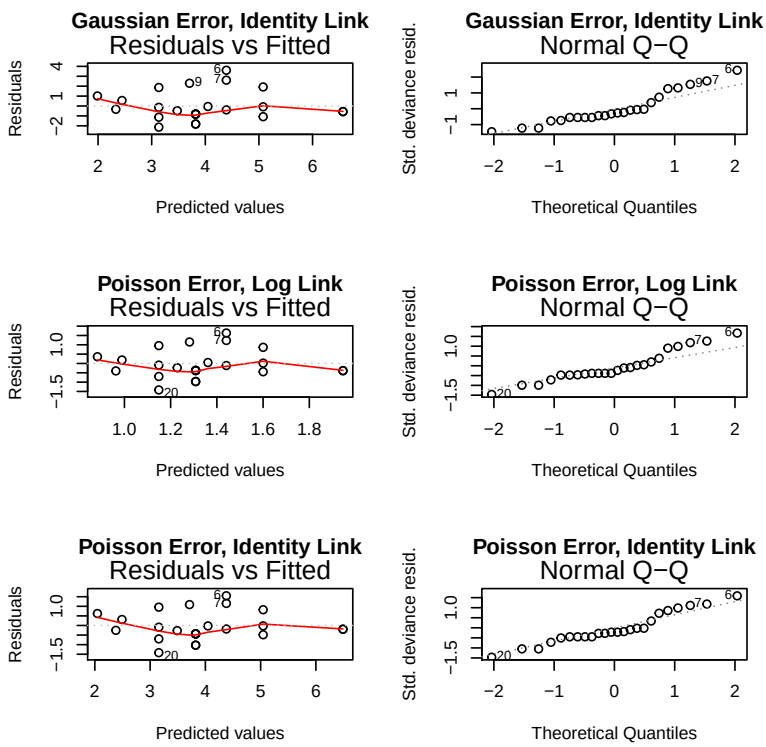
Three Models

```
a<-glm(pups ~ inbreeding.coefficient, data=wolves,
       family=gaussian(link="identity"))


b<-glm(pups ~ inbreeding.coefficient, data=wolves,
       family=poisson(link="log"))


d<-glm(pups ~ inbreeding.coefficient, data=wolves,
       family=poisson(link="identity"))
```


Which Fit Works?



Poisson Error, Identity Link!