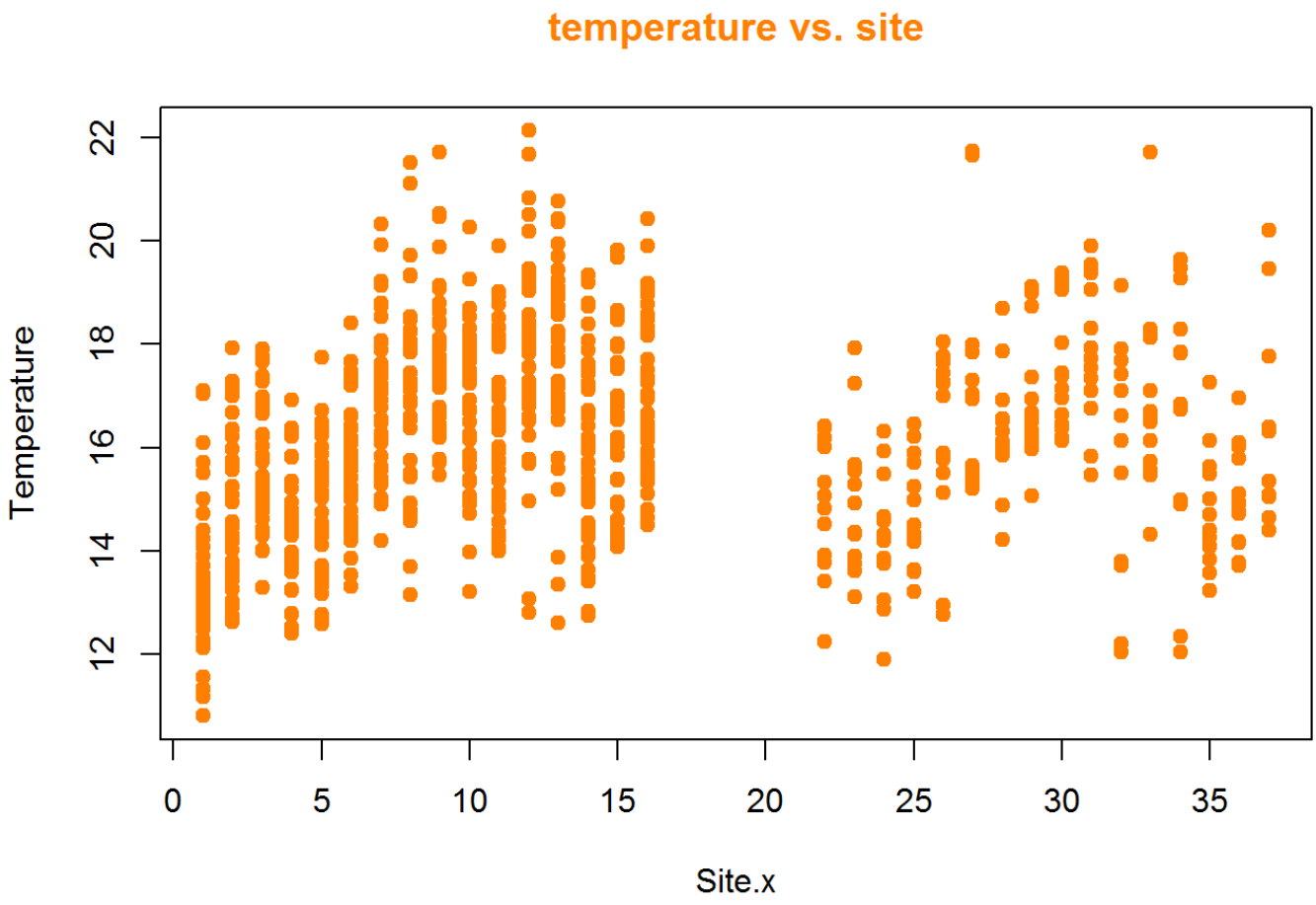
OCN 750 - HW6

Maia Kapur

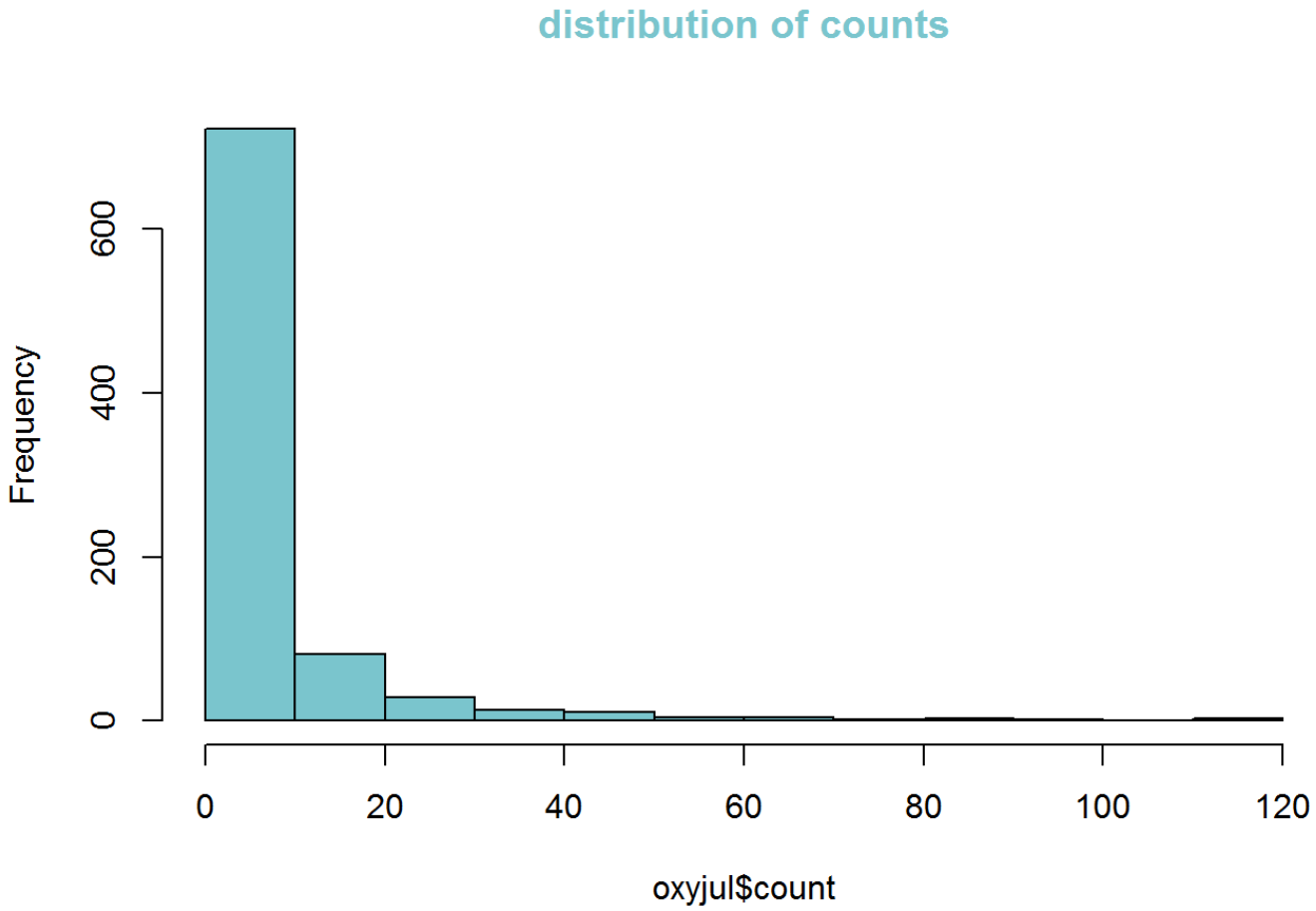
September 30, 2015

Make some exploratory plots of temperature vs Site, the distribution of all counts, and count vs temperature. What are your thoughts so far?

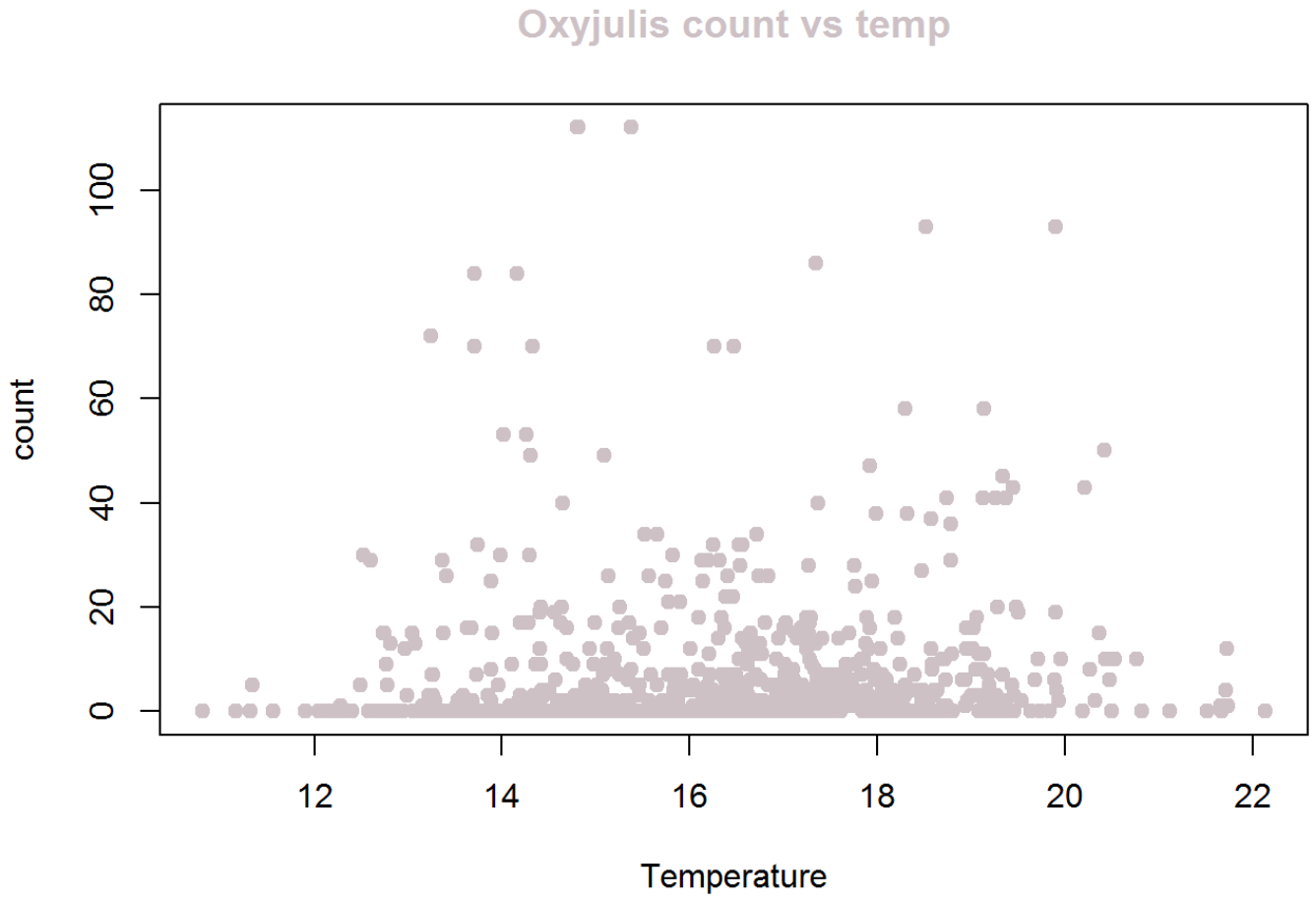
```
#setwd("~/Dropbox/2015 Fall/OCN 750/hw6")
oxyjul = read.csv("C:/Users/mkapur/Dropbox/2015 Fall/OCN 750/hw6/oxyjulis_subset.csv")
plot(Temperature ~ Site.x, data = oxyjul, main = "temperature vs. site", pch = 19, col =
"darkorange1", col.main = "darkorange1")
```




```
hist(oxyjul$count, col = "cadetblue3", main = "distribution of counts", col.main = "cadet
blue3")
```




```
plot(count ~ Temperature, data = oxyjul, pch = 19, col = "lavenderblush3", main = "Oxyjul
is count vs temp", col.main = "lavenderblush3")
```




```
##The counts are highly skewed with many zeros, possibly inflated. Even without the zeros there doesn't seem to be a big trend between count and temperature. There seem to be some slight trends amongst the two groups of sites. So...this is tricky.
```

This is count data, so starting with a Poisson GLM seems like a good idea. There are many samples per site, so we should include Site as a predictor (make sure to make it a factor!). Make a Poisson GLM with temperature and Site as predictors, and quantify how overdispersed the data is.

```
as.factor(oxyjul$Site.x) ##coerce the site vals to factor
```

```
site.modp = glm(count ~ Site.x + Temperature, data = oxyjul, family = "poisson") ##genera
te your poisson glm
summary(site.modp)
```

```
##
## Call:
## glm(formula = count ~ Site.x + Temperature, family = "poisson",
##     data = oxyjul)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0642  -3.3243  -2.4741  -0.0862  21.1793
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.255220   0.112622   2.266   0.0234 *
## Site.x      0.022628   0.001255  18.028   <2e-16 ***
## Temperature 0.077774   0.006784  11.464   <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: 13937  on 872  degrees of freedom
## Residual deviance: 13438  on 870  degrees of freedom
## AIC: 15211
##
## Number of Fisher Scoring iterations: 6
```

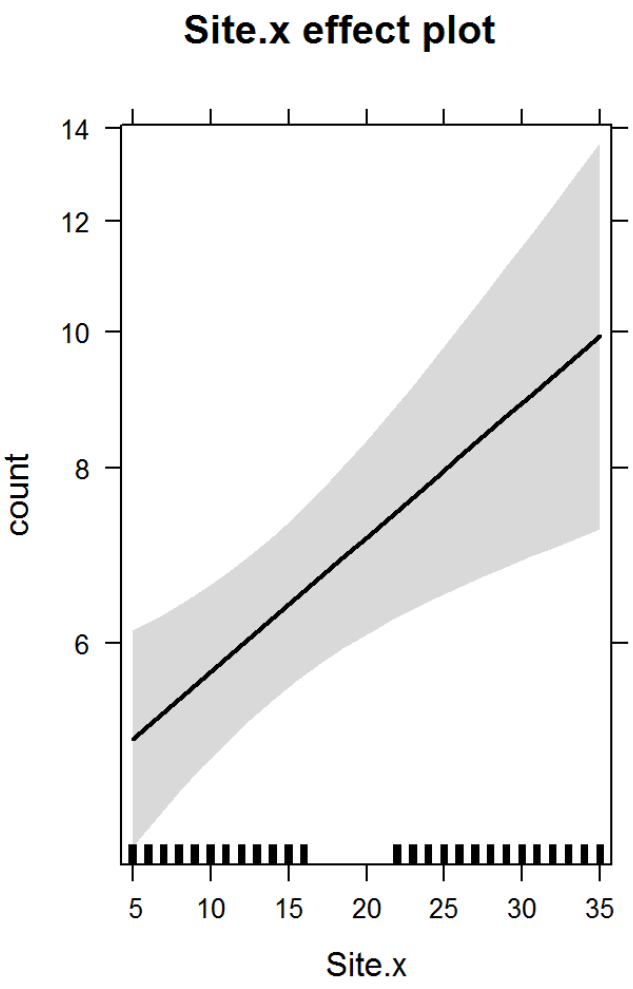
```
overdis  = function(model)  {
        sum(residuals(model,    type    = "pearson")^2)/(length(model$y) - length(model$c
oefficients))
} ##create the overdispersion function
overdis(site.modp) ##this parameter is very high, meaning overdispersion is low
```

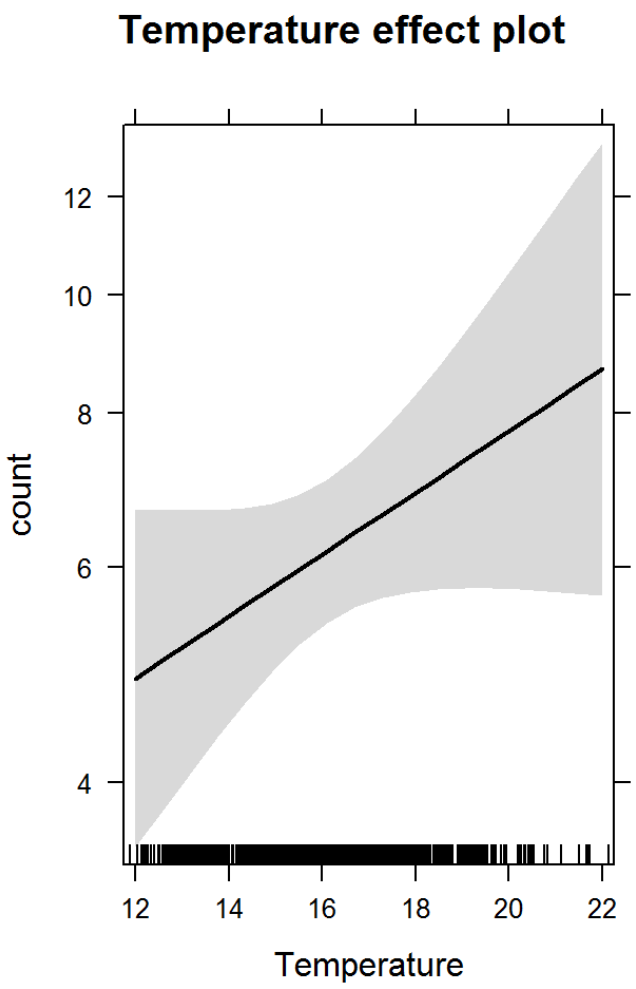
```
## [1] 29.30449
```

Based on the shape of the data, it seems like a good idea to consider zero inflation. But we'll want to compare a zero-inflated model to a non-inflated model, so first fit a negative binomial GLM with temperature and site as predictors. Plot the fitted effects, and report likelihood ratio tests for the predictors.

```
##fit the neg-binomial GLM
require(MASS)
nb.mod = glm.nb(count ~ Site.x + Temperature, data = oxyjul)

##Plot of fitted effects
require(effects)
plot(allEffects(nb.mod))
```






```
##Anova LRT for predictors
require(car)
Anova(nb.mod)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: count
##             LR Chisq Df Pr(>Chisq)
## Site.x        9.8753  1   0.001675 **
## Temperature   2.6996  1   0.100371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What's the estimate for theta for the negative binomial?

```
theta = overdis(nb.mod)
theta ##less than 1.5
```

```
## [1] 1.193846
```

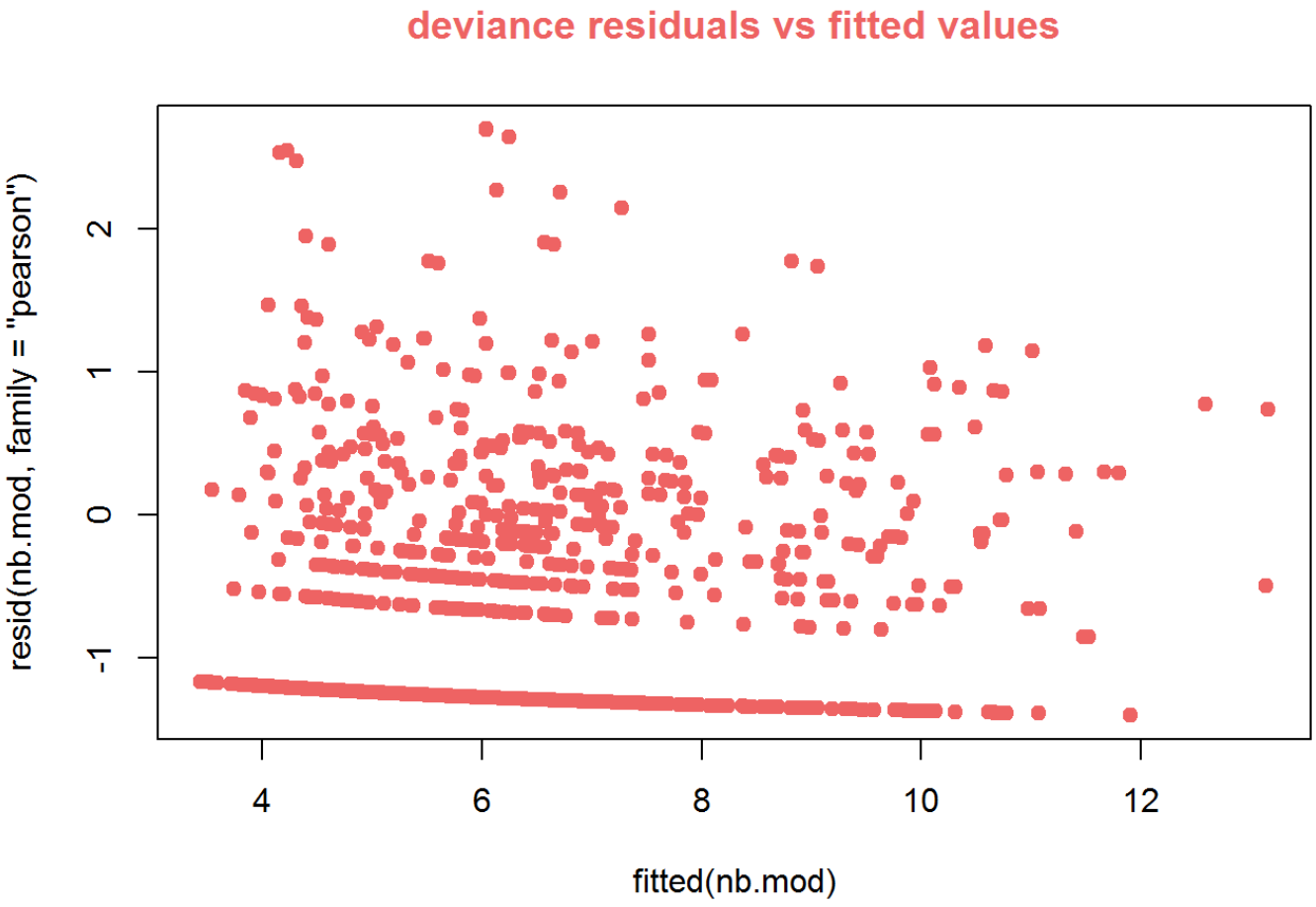
What does this model say about the relationship between abundance and

temperature?

```
##The relationship looks nearly linear though the CIs are wide. So Temperature is somewha
t positively correlated with the model, which isn't greatly overdispersed.
```

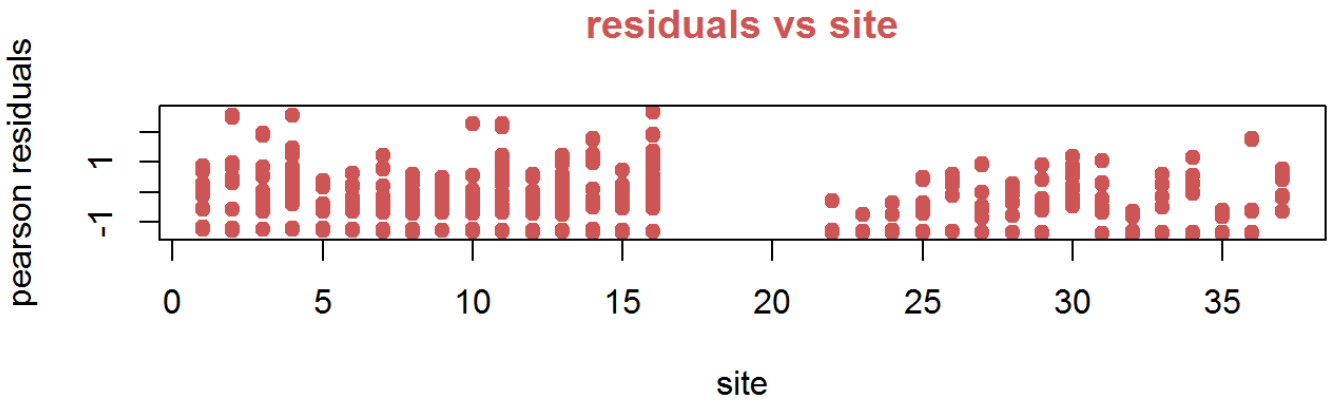
Plot the deviance residuals vs 1) the fitted values, 2) vs the predictors. Does a linear model for temperature (on the link scale) seem reasonable?

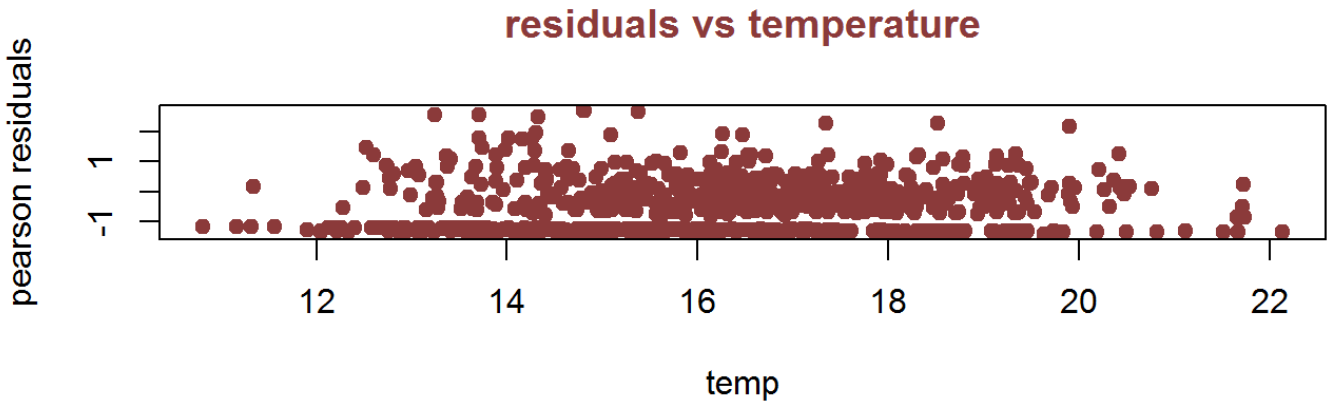
```
##plot of deviance residuals vs fitted values
plot(resid(nb.mod, family = "pearson") ~ fitted(nb.mod), main = "deviance residuals vs fi
tted values", pch = 19, col = "indianred2", col.main = "indianred2")
```




```
##plot of deviance residuals vs. predictors for two params
par(mfrow = c(2,1))
plot(resid(nb.mod, family = "pearson") ~ oxyjul$Site.x, main = "residuals vs site", pch =
19, col = "indianred3", col.main = "indianred3", ylab = "pearson residuals", xlab = "sit
e")

plot(resid(nb.mod, family = "pearson") ~ oxyjul$Temperature, main = "residuals vs tempera
ture", pch = 19, col = "indianred4", col.main = "indianred4",  ylab = "pearson residual
s", xlab = "temp")
```






```
dev.off()
```

```
## null device
##           1
```

```
##A linear model doesn't make sense to me. There seems to be a big asymptote/gap in the d
ata and the residuals show a slight negative trend as predictors increase.
```

For now let's keep it somewhat simple and make 5 models to compare: 1) the negative binomial model with temperature + site;

```
##See above.
##AIC for nb.mod. Store this for later comparison
nb.mod.aic = AIC(nb.mod)
```

2) a zero-inflated poisson with temperature + site as predictors only for the count model;

```
require(pscl)
as.data.frame(oxyjul)
modz = zeroinfl(count   ~ Temperature + Site.x  | 1, data = oxyjul, dist    = "poisson")
modz.aic = AIC(modz)
```

3) a zero-inflated negative binomial with temperature + site as predictors only for the count model;

```
modz.nb = zeroinfl(count  ~ Temperature + Site.x    | 1, data = oxyjul, dist    = "negbi
n")
modz.nb.aic = AIC(modz.nb)
```

4) a zero-inflated poisson with temperature + site as predictors both for the count model and also for the zero-inflation (binomial) model;

```
##similar to number 2, but now you have the predictors on both sides of the line, indicat
ing them both for the poisson and binomial parts
modz.both = zeroinfl(count  ~ Temperature + Site.x  | Temperature + Site.x, data = oxyju
l,  dist    = "poisson")
modz.both.aic = AIC(modz.both)
```

5) a zero-inflated negative binomial with temperature + site as predictors both for the count model and also for the zero-inflation (binomial) model.

```
##add in the "as factor" so R knows to separate the sites
modz.nb.both = zeroinfl(count  ~ Temperature + as.factor(Site.x)  | Temperature + as.fact
or(Site.x), data = oxyjul, dist    = "negbin")
modz.nb.both.aic = AIC(modz.nb.both)
```

Calculate AIC for all five models. Which is the ‘best’ model (lowest AIC)? What does it mean that this is the best model, compared to the other models?

```
aics = cbind (modz.aic, modz.nb.aic, modz.both.aic, modz.nb.both.aic)
names(aics[1:4]) = c("MODZ.Poisson", "MODZ.NegBin", "MODZ.Poisson.Both", "MODZ.NegBin.Bot
h")
min(aics) ##occurs at negative binomial model with predictors in both the count model and
binomial model.
```

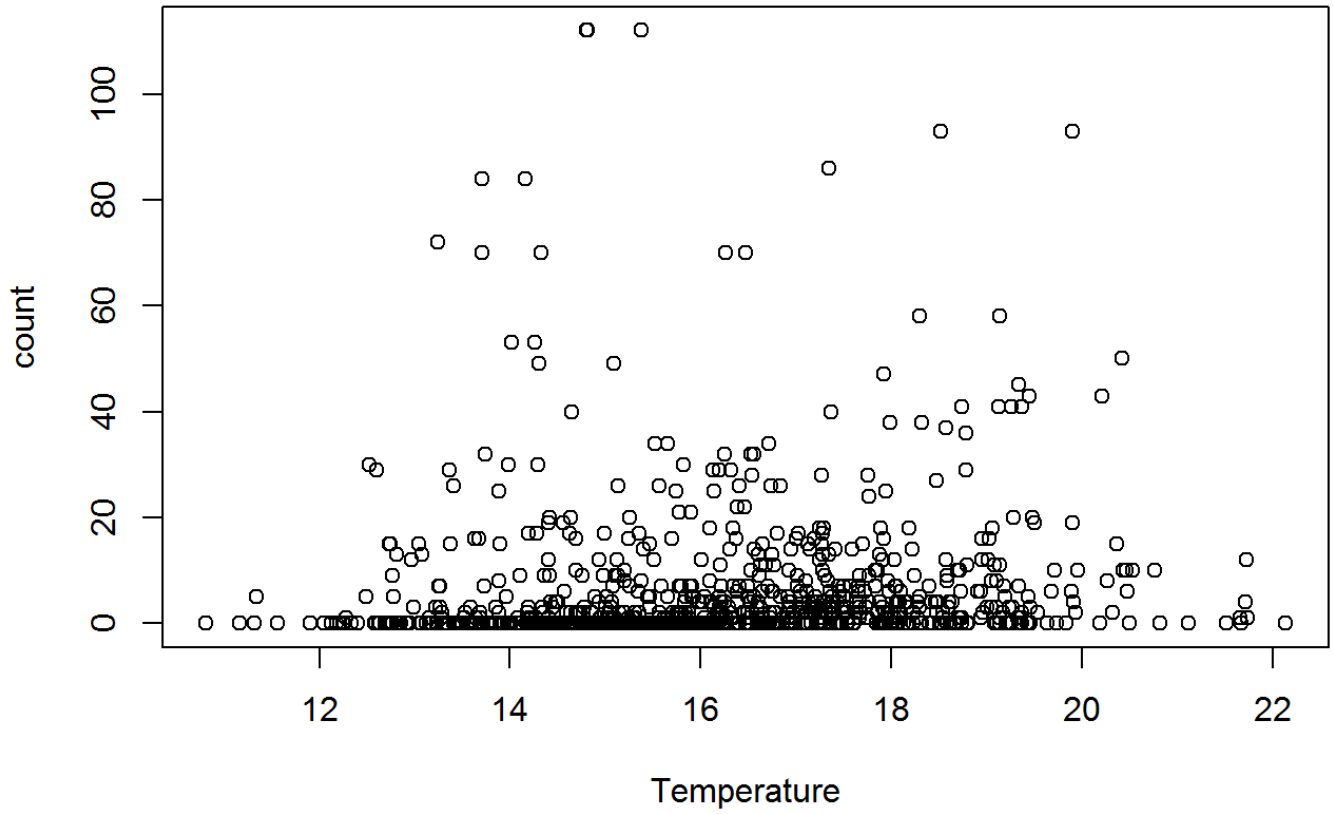
```
## [1] 4139.507
```

```
##this estimates the model's quality, as an estimate of the information lost by using a g
iven model. It also accounts for the complexity of the model in a sort of "tradeoff" betw
een good-fit and overfitting. This model does the best job in minimizing unexplained data
while being fairly simple. Normally we look for AICs under 10, I think.
```

What is the effect(s) of temperature in the best model? How do you interpret this result, statistically and biologically, compared to the negative binomial model

(i.e. non-inflated)?

```
summary(nb.mod) ##the negative binomial model only suggests significance for "site" upon the abundance
summary(modz.nb.both) ##whereas the best model shows significance in both the count model and zero-inflated binomial model. The effect is less significant and quite small in the former, however.
plot(count ~ Temperature, data = oxyjul)
```




```
##This suggests that the extra zeros may have masked the effect of temperature, which has as strong and significant effect without them.
```

Do you notice anything a little funny about the standard errors for the model coefficients for the best model?

```
##They seem really small
```

For model #5 in the above list, let's visualize the fitted effect of temperature on the probability of getting an 'extra' zero. Because the model has a factor for Site, that means the different sites will have different intercepts (but the same slope for temperature). So pick four sites and use curve() to plot the fitted logistic curve for those sites, all on the same plot, with the x-axis having the same range as the

range of temperature in the dataset.

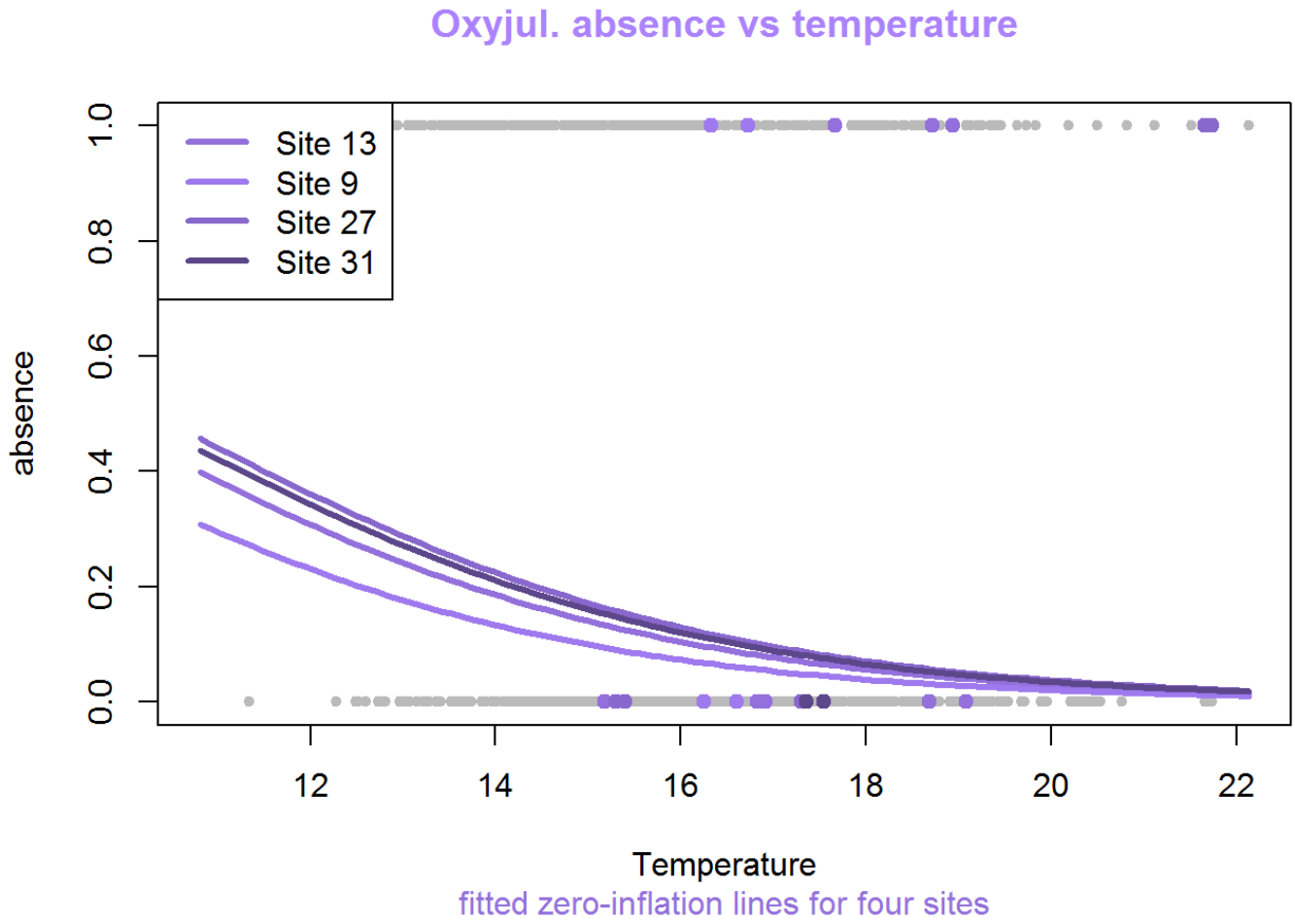
```
##plot presence/absence of counts vs temp. This is what was donoe for possum absence vs. stags in lec 10.
plot(absence ~ Temperature, data = oxyjul, pch = 20, col = "gray 73", main = "Oxyjul. absence vs temperature", sub = "fitted zero-inflation lines for four sites", col.main = "mediumpurple1", col.sub = "mediumpurple")

##highlight the sites' points, for reference
with(subset(oxyjul,  Site.x == 13), points(count  ~ Temperature,    pch = 19,   col = 'mediumpurple'))
with(subset(oxyjul,  Site.x == 9),  points(count  ~ Temperature,  pch   = 19,   col = "mediumpurple2"))
with(subset(oxyjul,  Site.x == 27), points(count  ~ Temperature,    pch  = 19,  col = "mediumpurple3"))
with(subset(oxyjul,  Site.x == 31), points(count  ~ Temperature,    pch  = 19,  col = "mediumpurple4"))

##function  for logistic    curve
logistic    = function(x) {exp(x)/(1+exp(x))}

##the offsets are pretty small for each, and the slope for temperature is minimal
curve(logistic(coef(modz.nb.both)["count_(Intercept)"] + coef(modz.nb.both)["count_as.factor(Site.x)13"] + coef(modz.nb.both)["zero_Temperature"] * x), add = TRUE, col = "mediumpurple",lwd = 3)
curve(logistic(coef(modz.nb.both)["count_(Intercept)"] + coef(modz.nb.both)["count_as.factor(Site.x)9"] + coef(modz.nb.both)["zero_Temperature"] * x), add = TRUE, col = "mediumpurple2",lwd = 3)
curve(logistic(coef(modz.nb.both)["count_(Intercept)"] + coef(modz.nb.both)["count_as.factor(Site.x)27"] + coef(modz.nb.both)["zero_Temperature"] * x), add = TRUE, col = "mediumpurple3",lwd = 3)
curve(logistic(coef(modz.nb.both)["count_(Intercept)"] + coef(modz.nb.both)["count_as.factor(Site.x)31"] + coef(modz.nb.both)["zero_Temperature"] * x), add = TRUE, col = "mediumpurple4", lwd = 3)

##add a legend
legend('topleft',  lty  = 1,    lwd = 3,    col = c('mediumpurple', 'mediumpurple2', 'mediumpurple3', 'mediumpurple4'), legend  = c('Site 13', 'Site 9', 'Site 27', 'Site 31'))
```



Roughly how much does temperature change the proportion of extra zeros?

##As temp increases, the proportion of obtaining extra zeros goes down along with oxyjul abundance. In other words, at low temperatures you are more likely to have an extra zero and/or observe fewer oxyjul, by a factor of about 2% for each degree increase of temp. (The slope varies for each site, but I got this by taking ~0.5/22.)