**Section 5.3 – Overdispersion**

We will focus on count responses arising in the Poisson setting in this section and briefly mention other types of responses at the end.

A limiting assumption for a Poisson distribution is that E(Y) = Var(Y) = μ. Sometimes, the sample variance of Y appears to be greater than the sample mean of Y for a data set. Evidence of this occurs in the horseshoe example. Below is a table that can be used as an ad-hoc assessment (Horseshoe\_Ch5.R):

> crab <- read.csv(file = "C:\\data\\horseshoe.csv")

> crab$groups <- cut(x = crab$width, c(20, seq(from = 23.25, to = 29.25, by = 1), 34))

> head(crab)

satellite width groups

1 8 28.3 (28.2,29.2]

2 0 22.5 (20,23.2]

3 9 26.0 (25.2,26.2]

4 0 24.8 (24.2,25.2]

5 4 26.0 (25.2,26.2]

6 0 23.8 (23.2,24.2]

> ybar <- aggregate(x = satellite ~ groups, data = crab, FUN = mean)

> yvar <- aggregate(x = satellite ~ groups, data = crab, FUN = var)

> data.frame(width = ybar$groups, mean = round(ybar$satellite, 2), var = round(yvar$satellite, 2))

width mean var

1 (20,23.2] 1.00 2.77

2 (23.2,24.2] 1.43 8.88

3 (24.2,25.2] 2.39 6.54

4 (25.2,26.2] 2.69 11.38

5 (26.2,27.2] 2.86 6.89

6 (27.2,28.2] 3.88 8.81

7 (28.2,29.2] 3.94 16.88

8 (29.2,34] 5.14 8.29

If the Poisson assumptions were satisfied, we would expect the mean.sat column to be approximately the same as the var.sat column. Obviously, this does not occur.

When the variance is larger than the mean, this is called overdispersion, and it is a violation of our model. Thus, inferences made using the model may be incorrect. What can you do when this occurs?

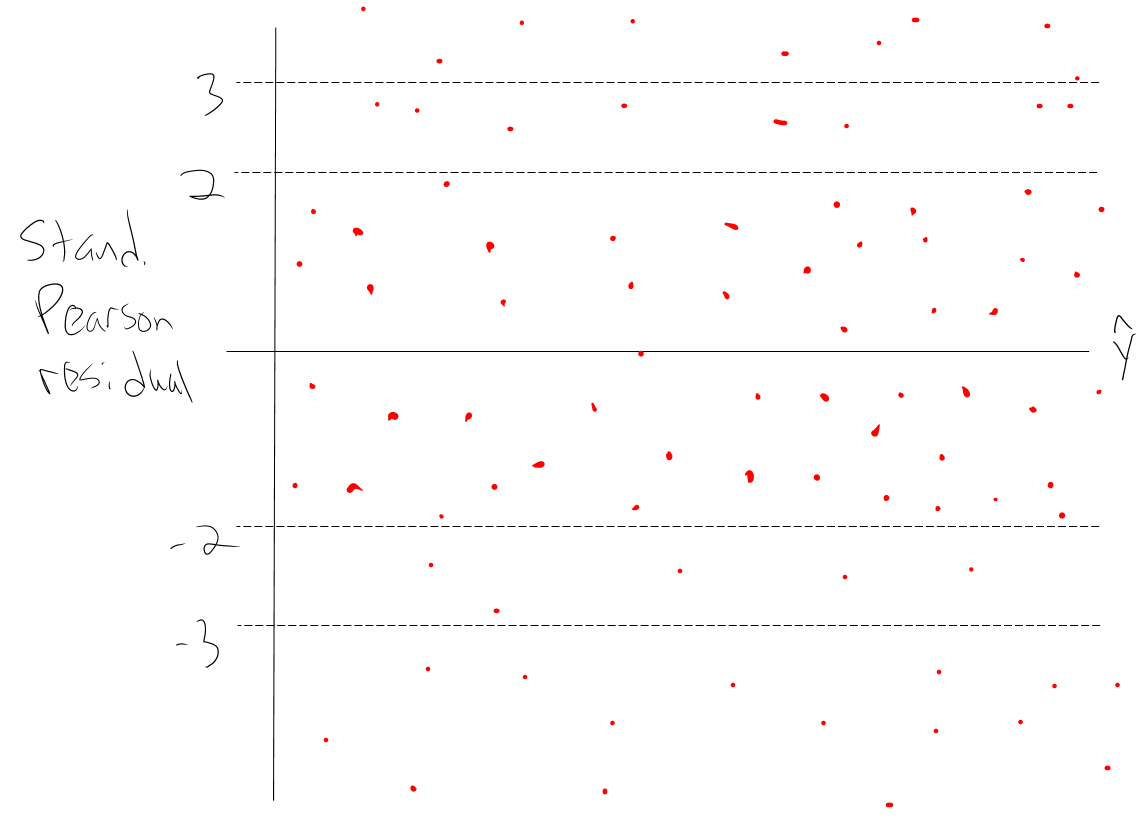
1. Find more explanatory variables that help explain the variability in the response variable! The additional variability could be due to not accounting for other explanatory variables. For example, perhaps crab weight plays an important role in estimating the mean number of satellites. By using width only in a model, there may be additional satellite variability at individual widths due to the omission of weight.
2. Check if it is a zero inflation problem (Chapter 4 topic).
3. Use Poisson generalized linear mixed models (Chapter 6 topic). These models are helpful if there is an identifiable cause for overdispersion, such as observations are collected over time on the same individuals.
4. Use quasi-Poisson regression models. These models do not assume a full parametric form for the model.
5. Use negative binomial regression models.

Options 4 and 5 should only be used when there is not an identifiable or measurable cause for the overdispersion.

Detection of overdispersion

We previously compared estimates of the mean counts to the variance of the counts to determine if overdispersion exists. This cannot always be done due there being more than one explanatory variable. Other approaches are:

1. Poor overall model fit as determined by deviance/df
2. More extreme standardized Pearson residuals than expected; this is shown in a residual plot by a fairly uniform trend of extreme standardized Pearson residual values across it.





Example: Horseshoe crabs and satellites (Horseshoe\_ch5.R, horseshoe.csv)

> mod.fit <- glm(formula = satellite ~ width, data = crab, family = poisson(link = log))

> summary(mod.fit)

Call:

glm(formula = satellite ~ width, family = poisson(link = log), data = crab)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.8526 -1.9884 -0.4933 1.0970 4.9221

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.30476 0.54224 -6.095 1.1e-09 \*\*\*

width 0.16405 0.01997 8.216 < 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: 632.79 on 172 degrees of freedom

Residual deviance: 567.88 on 171 degrees of freedom

AIC: 927.18

Number of Fisher Scoring iterations: 6

> mod.fit$deviance / mod.fit$df.residual

[1] 3.320927

> round(c(1 + 2\*sqrt(2/mod.fit$df.residual), 1 +

3\*sqrt(2/mod.fit$df.residual)), 2)

[1] 1.22 1.32

Poor model fit is indicated by a large deviance/df.

> mu.hat <- mod.fit$fitted.values

> stand.resid <- rstandard(model = mod.fit, type =

"pearson")

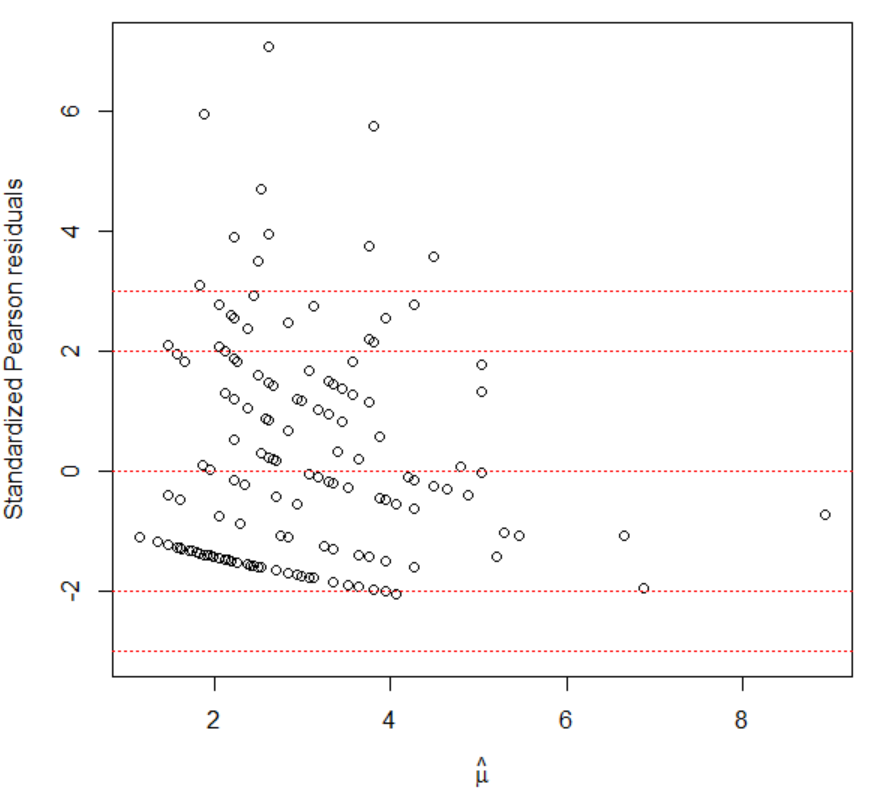
> plot(x = mu.hat, y = stand.resid, xlab =

expression(hat(mu)), ylab = "Standardized Pearson

residuals", ylim = c(min(c(-3, stand.resid)), max(c(3,

stand.resid))))

> abline(h = c(-3,-2,0,2,3), lty = "dotted", col = "red")



> sum(abs(stand.resid) > 3)

[1] 10

> length(stand.resid)

[1] 173

> sum(abs(stand.resid) > 3) / length(stand.resid)

[1] 0.05780347

> #Expected number of extreme residuals if N(0,1) works

> length(stand.resid)\*2\*(1-pnorm(q = 3))

[1] 0.4670647

The expected number of extreme residuals is larger than expected when using a standard normal approximation. Also, some residuals are MUCH larger than one would expect. Therefore, overdispersion appears to exist when only having the crab width available as an explanatory variable.

Quasi-Poisson regression models

To find maximum likelihood estimates for β0, …, βp, we need



for r = 0, …, p. These equations are set equal to 0 and solved for β0, …, βp. In general, equations that are set equal to 0 and then solved for parameter estimates are called “estimating equations.” This makes sense here because our “estimates” result from these “equations.”

Wedderburn (1974) suggested using quasi-likelihood methods to find regression parameter estimates for a generalized linear model. In this setting, one assumes a particular relationship between the mean and variance, BUT no particular distribution for the response variable. In the count data situation, we can proceed in a similar manner as in the past, but with some adjustments. Let the relationship between the mean and variance be

Var(Y) = γE(Y)

for some constant γ, but do not assume Y has a Poisson distribution. Notice that when γ > 1, we would have overdispersion for the regular Poisson distribution. Models of this type are called quasi-Poisson.

The estimating equations used for Poisson regression simplify to



for r = 0, …, p, where there are m = 1, …, M observations, xmr is the mth observed value for the rth explanatory variable, and μm = exp(β0 + β1xm1 + + βpxmp). The quasi-likelihood approach uses Var(Ym) = γμm leading to the estimating equations of



for r = 0, …, p.

What happens when you estimate the regression parameters with this set of estimating equations?

1. When the above equation is set equal to 0, γ does not play a role. The exact same estimates for  as with Poisson regression will result.
2. The estimated covariance matrix for  is the same as with the Poisson regression model except it is multiplied by γ.
3. Diagnostic measures, like standardized residuals, are appropriately adjusted. For example,



for the ith observation.

1. Because γ is a parameter, it needs to be estimated as well. The standard approach is to use  = X2/(M-) where X2 is the Pearson statistic from the regular likelihood approach model. The reason is because E(X2) is approximately equal to M-, where  is the number of regression parameters, when there is no overdispersion.
2. The -2log(Λ) statistic for a LRT is divided by  to account for the overdispersion. The distribution of this new statistic is approximated by an F-distribution. The numerator degrees of freedom is the number of parameters testing (q) and the denominator degrees of freedom is the sample size minus the total number of parameters in full model (M – ). Similarly, profile LR intervals are found using the F-distribution rather than the chi-square distribution.
3. The AIC can no longer be used because likelihood methods are not being used. Instead, the QAIC is used to compare quasi-Poisson regression models. This statistic is the same as the AIC, but now with the likelihood function divided by . Do not compare a QAIC to an AIC from a Poisson regression model. Also, you should use the same value of  when comparing different quasi-Poisson regression models. Please see the book for a discussion.

Example: Horseshoe crabs and satellites (Horseshoe\_ch5.R, horseshoe.csv)

The model is estimated by glm() again, but now the family option has changed to quasipoisson(link = log).

> mod.fit.quasi <- glm(formula = satellite ~ width, data =

crab, family = quasipoisson(link = log))

> summary(mod.fit.quasi)

Call:

glm(formula = satellite ~ width, family = quasipoisson(link = log), data = crab)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.8526 -1.9884 -0.4933 1.0970 4.9221

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.30476 0.96729 -3.417 0.000793 \*\*\*

width 0.16405 0.03562 4.606 7.99e-06 \*\*\*

---

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

(Dispersion parameter for quasipoisson family taken to be 3.182205)

Null deviance: 632.79 on 172 degrees of freedom

Residual deviance: 567.88 on 171 degrees of freedom

AIC: NA

Number of Fisher Scoring iterations: 6

> vcov(mod.fit.quasi)

(Intercept) width

(Intercept) 0.93565064 -0.034334475

width -0.03433447 0.001268475

> sum.fit.quasi <- summary(mod.fit.quasi)

> vcov(mod.fit)

(Intercept) width

(Intercept) 0.29402590 -0.0107895239

width -0.01078952 0.0003986151

> vcov(mod.fit) \* sum.fit.quasi$dispersion

(Intercept) width

(Intercept) 0.93565064 -0.034334475

width -0.03433447 0.001268475

> anova(mod.fit.quasi, test = "F")

Analysis of Deviance Table

Model: quasipoisson, link: log

Response: satellite

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev F Pr(>F)

NULL 172 632.79

width 1 64.913 171 567.88 20.399 1.168e-05 \*\*\*

---

> Anova(mod.fit.quasi, test.statistic = "F")

Analysis of Deviance Table (Type II tests)

Response: satellite

SS Df F Pr(>F)

width 64.91 1 20.399 1.168e-05 \*\*\*

Residuals 544.16 171

---

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

> # Wald interval with c = 1

> beta.ci <- confint.default(object = mod.fit.quasi, parm = "width", level = 0.95)

> beta.ci

2.5 % 97.5 %

width 0.09423969 0.2338505

> 100\*(exp(beta.ci) - 1)

2.5 % 97.5 %

width 9.882309 26.34556

> # Profile likelihood ratio interval

> beta.ci <- confint(object = mod.fit.quasi, parm =

"width", level = 0.95)

Waiting for profiling to be done...

> beta.ci

2.5 % 97.5 %

0.09363759 0.23324826

> 100\*(exp(beta.ci) - 1)

2.5 % 97.5 %

9.816169 26.269491

The estimated model is .

Notice the output from summary() is exactly the same as what we had for the regular Poisson regression model except for a few differences:

1. The “dispersion parameter” is the estimate of γ, say , and it is given to be 3.18.
2. The estimated standard deviations for the model parameter estimates are  = 0.96731 and  = 0.03562. The Poisson regression model had values of  = 0.5422 and  = 0.01996. Notice that

> 0.54222 \* sqrt(3.18)

[1] 0.9669168

> 0.01996 \* sqrt(3.18)

[1] 0.03559378

1. With the larger standard errors, Wald statistics are smaller and corresponding p-values are larger.
2. A LRT for H0: β1 = 0 vs. Ha: β1 ≠ 0 with an F-distribution approximation results in a small p-value indicating that width is an important explanatory variable.
3. The 95% confidence intervals for PC are wider here than when using the regular Poisson regression model. For example, the 95% profile LR interval was (13.28, 22.51) with a Poisson regression model.
4. There is no AIC listed because this is not a full likelihood method.
5. There are much fewer extreme residuals with this model than with the Poisson regression model:

