FES 524: Natural Resources Data Analysis

Reading 7.2

**Contents**

[Other readings to do before class 1](#_Toc64285934)

[Overdispersion 1](#_Toc64285935)

[Causes of overdispersion 2](#_Toc64285936)

[Alternatives 2](#_Toc64285937)

[Quasibinomial model 3](#_Toc64285938)

[Binary data and overdispersion 3](#_Toc64285939)

[Underdispersion 3](#_Toc64285940)

[Complete separation 4](#_Toc64285941)

[Causes of and solutions to complete separation 5](#_Toc64285942)

[Alternatives when you have complete separation 5](#_Toc64285943)

[Quasi-complete separation 6](#_Toc64285944)

[Fitting a binomial GLM 6](#_Toc64285945)

[Response variable 6](#_Toc64285946)

[Check for overdispersion 6](#_Toc64285947)

[Complete separation 7](#_Toc64285948)

[Role of residuals 7](#_Toc64285949)

[Hypothesis tests and confidence intervals 7](#_Toc64285950)

[Results 8](#_Toc64285951)

[Output from R 8](#_Toc64285952)

Today we will continue to learn about binomial generalized linear models and how to fit them. We take time to talk about additional complications that can arise with these models.

# Other readings to do before class

Read the description of the study that assignment 7 is based on in Handout 7.2 and come prepared to discuss aspects of the research question and study design.

# Overdispersion

Overdispersion is term that we use when the variance in the observed data is bigger than the variance we expect from the distribution we are using to model the data. Single parameter distributions are especially prone to overdispersion, since once we know the mean we also know the variance. However, any distribution where the variance is based on the mean can have issues where the distribution-defined variance doesn’t capture the variance in the data well.

Overdispersion is a common issue in the binomial distribution since the binomial distribution is a single parameter distribution. The variance is based directly on the mean without any other scale parameters. When discussing overdispersion and the binomial distribution you may see the term *extra-binomial variation* used.

**Distribution Mean Variance**

Binomial p p(1 - p)/m

Why do we care about overdispersion? Well, our model results are based on the distribution-defined variance. If the variance from the distribution is smaller than the actual variance observed in the data, our results will be based on an underestimate of the actual variance. Underestimating the variance means all standard errors are too small, which means any p-values are too small and the size of the confidence interval is underestimated. Our results will be *anticonservative*.

## Causes of overdispersion

A positive correlation among the *trials* within a study unit is one primary cause of overdispersion in binomial data. Positive correlation is likely a missing covariate problem. I’ll use this week’s motivating example to discuss this.

In the motivating example, overdispersion would indicate that individual fingerlings getting tumors within a single tank are not independent from other fish getting tumors in the same tank. If one fingerling has a tumor in a tank it is more likely that other fingerlings in the tank will have tumors, as well. Since dose is in the model, the presence of overdispersion would indicate that something other than dose, such as temperature, is causing the individual fingerlings getting tumors within a tank to be non-independent. We didn’t measure that variable, so our model is underspecified and our trials are not conditionally independent (i.e., independent after accounting for everything in the model).

You’ll see “a missing explanatory variable” also listed as a common cause of overdispersion, although to me this is tied up with positive correlation among trials as discussed above. For example, if we used blocking in our design but leave blocks out of the model we could see overdispersion. Again, this would be due to trials no longer being conditionally independent because, as you know, if blocking is successful then blocks will cause positive correlation among observations with blocks.

A complication with GLM’s is that we are no longer have the observation-level variation accounted for by the model. Tanks could simply be different from each other for unexplained reasons. We’ve always accounted for this *observation-level variation* when using linear models, but the variance-mean relationship of the one-parameter binomial distribution means we no longer have this by default in our model. There is no reason that we would suddenly no longer expect our study units to have unexplained variation among them when switching from LM’s to GLM’s, and this unexplained variation can manifest itself as overdispersion.

Finally, having excessive 0 values that aren’t modeled well by the binomial distribution can lead to overdispersion.

## Alternatives

There are several alternative approaches we can use if we have overdispersion in a binomial GLM, since having overdispersion means our model does not fit the data well. I am not listing these in order of importance or how useful they are.

*Quasibinomial*

In the quasibinomial model, we estimate the overdispersion and multiply all standard errors by the square root of this value. All standard errors are then larger than in the original model.

*Beta-binomial*

In the beta-binomial model, we model the mean of the binomial distribution with the beta distribution and then the observed data with a binomial distribution. This sort of model allows for more variation than the binomial GLM does. It also allows us to use covariates in both the parts of the model, as needed. The beta-binomial model is a type of *compound* or *mixture* model, which can be fit in R in, e.g., packages VGAM or glmmTMB.

*Observation-level random effect*

Since we are missing the observation-level random effect in our binomial GLM, we could add this to account for observation-level noise. This would mean fitting a generalized linear *mixed* model (GLMM).

None of the three listed options are useful if excessive 0 values are causing the overdispersion. In that case, other tools like *zero-inflated models* will likely be needed.

## Quasibinomial model

The quasibinomial model is a very simple approach to dealing with overdispersion, which is why this is what we will be using in the analysis example this week. However, since we use a single, overall correction based on a single estimate of overdispersion, this approach can be too simplistic for a lot of real analyses. For example, maybe we have different amounts of overdispersion in different groups, which we can’t model with the quasibinomial distribution.

The quasibinomial distribution is not actually a distribution (it is a “quasi” distribution) and it has no likelihood associated with it. If you want to use a statistic based on the likelihood, like an information criterion, it must be based on a quasilikelihood. Note the quasibinomial option is not available in some R packages when fitting GLMM’s.

We use F/t tests when reporting results from a model fit using a quasilikelihood approach such as a quasibinomial GLM. There is actually no theoretical reason for this, but it is considered to be a logical thing to do and is standard.

## Binary data and overdispersion

Truly binary data cannot be overdispersed. However, any variables known to cause variation in the response still need to be in model (fixed effects, blocks, etc.). Switching to working with binary data, such as working on the fish level instead of tank level in the motivating example, is not a way around the issue overdispersion. The model still won’t fit the data correctly even if you can no longer calculate the overdispersion, and an alternative model would still be needed.

## Underdispersion

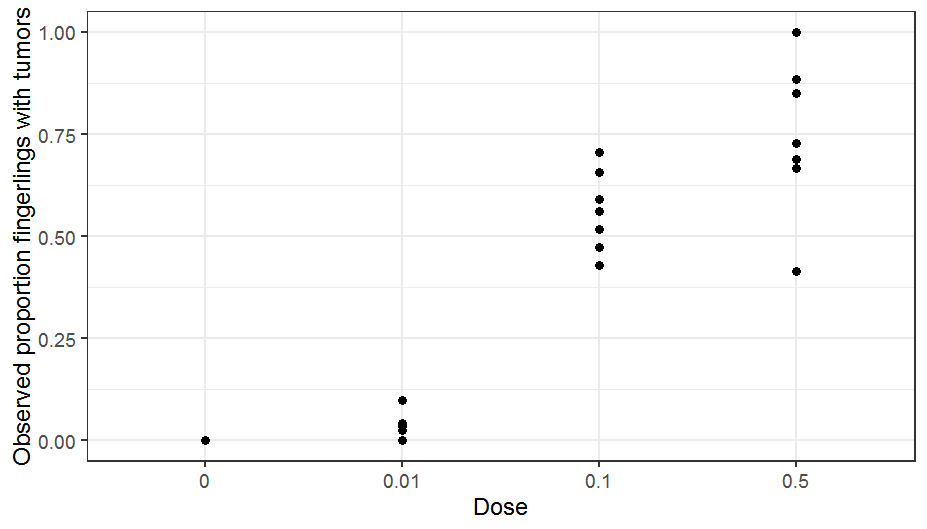
Underdispersion, where the variation in the data is *smaller* than what we expect given the distribution, is also a possibility. We can get underdispersion if we had negative correlation of trials within any study units, for example, although negative correlation is less common.

Historically researchers haven’t addressed underdispersion, and there aren’t modeling approaches that have been designed to deal with it. Certainly underdispersion indicates a lack of fit of the model to the data. If you ignore it, you are being overly conservative and understating the results. P-values are too large and confidence intervals will be too wide.

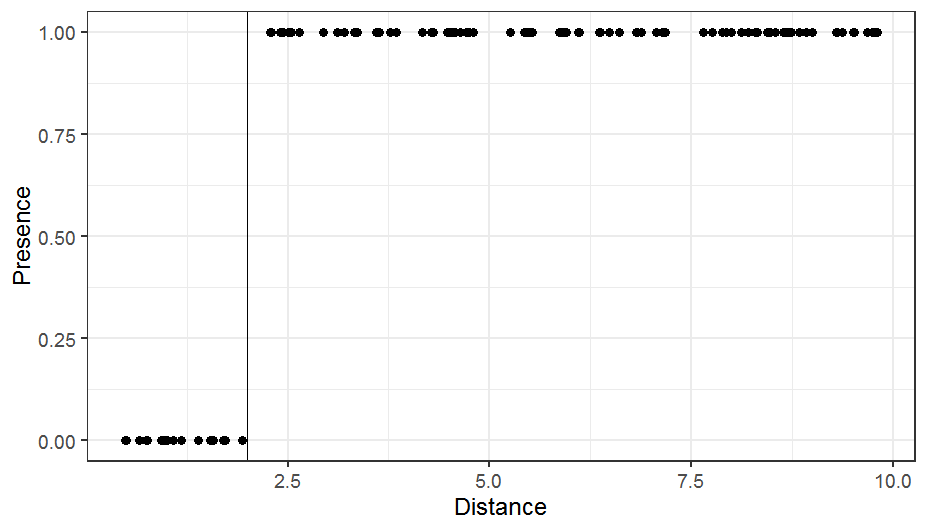
# Complete separation

*Complete separation*, also called perfect separation, is when the outcome is perfectly determined by a explanatory variable. Complete separation is a concept that is most often associated with binomial or binary data, since it is commonly a problem with models using those distributions. However, it can occur with other types of data (like counts, in particular).

In this week’s motivating example, the control group demonstrates complete separation for a categorical response variable. The outcome is perfectly determined if we know a tank is in the control group because the no fingerlings in that group got tumors and the outcome for all control tanks is 0.

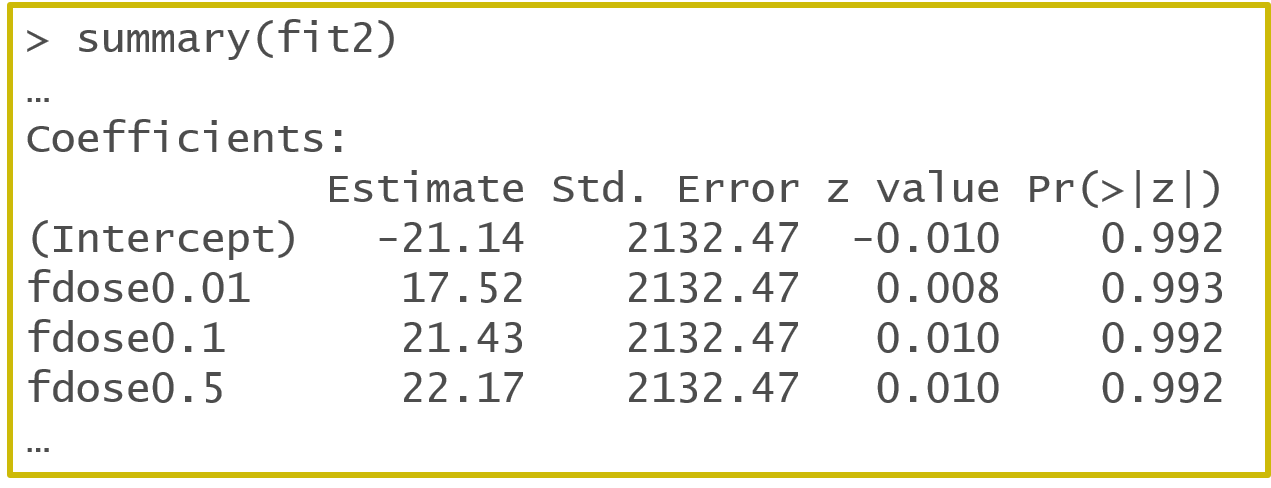


Complete separation can happen with continuous variables, as well. Below is a plot showing a binary response variable vs some continuous variable distance.



You can see that distance perfectly explains presence/absence in this case: all observations below 2 are 0 and all observations above 2 are 1. These data are *completely separated* at Distance = 2.

A symptom of complete separation will be extremely large standard errors along with estimated values of coefficients > 4. You can see this in the coefficients and standard errors from a model summary in R:



You also may see a warning message like this:



## Causes of and solutions to complete separation

Complete separation is caused by the model attempting to estimate log(0) or log(infinity), since at least one of the groups is all 0 or all 1. This is where it’s important to remember that it is the *mean* that is related to the linear predictor through a link function. Having 0 values is not a problem with a binomial GLM with a logit link in general. It’s only when, e.g., one group is made up of all 0 values so the *mean* of the variable is 0 that problems arise. Picturing the math may help you see why this is.

Here’s what the left-hand side of the model looks like in the mathematical notation we’ve been using for the probability model for a binomial GLM with a logit link, where p is the mean:

log(p/1-p)

You can see that if p is 0 this becomes log(0/1) = log(0). If p is 1 this becomes log(1/0) = log(infinity). Since log(0) is -infinity, the optimizer being used by the software to maximize the likelihood attempts and eventually fails to reach -infinity and returns the warning and the large standard errors. The same thing happens with log(infinity) = infinity.

A group of all 0 values is likely going to be extremely interesting scientifically. What we need to ask ourselves is if it interesting statistically. When a group is all 0, do we need to estimate differences from that group? This may be a case where the option of simply stating this very strong result is sufficient and statistical results won’t strengthen the evidence at all.

Having complete separation isn’t necessarily a problem. Tests based on the likelihood are still valid (these are discussed more below). Estimates and CI’s made from profiling the likelihood will be fine. Wald-based tests and confidence intervals, though, such as those used by package emmeans or output in the model summary in R, will not be valid.

## Alternatives when you have complete separation

Like so many decisions in statistics, if you are in a situation where complete separation needs to be addressed there are many options and the way you should approach modeling will take careful thought. Any choice needs to be justified and well described in your methods. While some solutions are fairly ad-hoc, penalized approaches are generally a good option. There are now Bayesian-based penalized approaches for more complicated models like GLMM’s that keep models constrained while working with the original data that are fairly attractive. Note that doing nothing can also be a fine option in many cases.

You can see a nice discussion and list of options to deal with complete separation on Cross Validated here, <https://stats.stackexchange.com/a/68917/29350>. The last option in this list is by far the least attractive option. I’ve included an additional link in “Additional resources” on Canvas from another source that makes a similar list.

## Quasi-complete separation

Quasi-complete separation can happen when the outcome variable separates a predictor variable or a combination of predictor variables *almost* completely. This means you may not be able to diagnose this in the raw data as we could in the two examples above. You will often only be able to tell you have issues with quasi-complete separation in complicated GLMM’s when the SE are extremely large compared to the estimate.

Options from working with quasi-complete separation in GLMM’s are more limited, and this is a case where using a Bayesian-based penalized model may come in handy. See the short discussion in the GLMM FAQ here, <https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#penalizationhandling-complete-separation>, for more background.

# Fitting a binomial GLM

Now that we’ve covered the complicated things that can happen with binomial GLM’s, let’s go through the process of how we would fit a binomial generalized linear model with a logit link and check that the model fits. This is what we will practice in R this week in lab.

## Response variable

When fitting a binomial GLM, the response variable is not just the observed proportion of successes for a study unit. You must also have information the number of trials. This is because 1., the number of trials is directly relevant to the variance and 2., the number of trials in a binomial GLM is a type of weighting variable. Observations with more trials are given more weight in analysis.

The response variable can therefore be stored in two ways.

Option 1:

Have one column for the number of successes and another for the number of failures or the total number of trials per observation.

Option 2:

Have one column for the proportion successes and another for the total number of trials.

If you have only the proportion and not the number of trials and the number of trials isn’t a fixed value you cannot use a binomial GLM.

## Check for overdispersion

Overdispersion is the first thing to check for when fitting a binomial GLM.

Since we know that overdispersion can by a symptom of a missing covariate, we will have anticipated this and collected and included covariates that we believe cause variation in our response in the model. We should be checking for overdispersion on the fullest reasonable model we can fit. Sometimes, as in the case of the motivating example, we have no additional information to use so the fullest model is fairly simple.

An estimate of the overall overdispersion is the sum of the squared Pearson residuals divided by the residual degrees of freedom. Values greater than 1 indicate overdispersion. There are no hard and fast rules for how much overdispersion is too much. An overdispersion value of 2 is definitely considered large. But a value of 1.1 is small and likely doesn’t matter too much. The researcher has to decide when the value is large enough to be problematic.

If we see evidence of overdispersion (or even if we don’t), we’d report that we checked for overdispersion and the estimated value for overdispersion. If we had to choose an alternative model to address the issue of overdispersion we’d need to then describe that model.

For binomial GLM we can often operate under the assumption that we probably have overdispersion. Since extra-binomial variation so common, it is likely OK to simply assume that you have overdispersion and start with a model to address the issue.

## Complete separation

We should already have a pretty good idea if we have *complete* separation because we’ll have already looked at our dataset prior to analysis. If we know this is an issue we will have already been thinking about what we want to do with a group that causes complete separation and have anticipated any modeling approaches that we might want to use.

If we have quasi-complete separation, we usually only become aware of it because of warning messages and giant SE in summary output. Again, this is not necessarily a problem, but quasi-complete separation can indicate overfitting (i.e., our model is too complex for our data) and we need to rethink our model. Alternatively, it may mean that some complex combinations of groups are all 0.

## Role of residuals

While we no longer have an error distribution to think about, we will still use plots of residuals to help check model fit. There are different types of residuals we can get, which we use for different things.

The Pearson residuals are what we use in the calculation for checking for overdispersion.

Deviance residuals are considered to be “well behaved”, and so are good for making residual plots. Most deviance residuals should fall between -2 and 2, allowing us to see any unusual points that the model doesn’t fit well. In addition, the way deviance residuals are calculated we expect to see reasonably similar variance among groups in any residual plots.

Note there is no assumption of normality of errors in a binomial GLM, so don’t check any assumptions of normality using residuals. While asymptotically the deviance residuals technically go in distribution to the normal distribution for large samples, we can’t expect this to be true for small samples. You can see an example where the deviance residuals will never be approximately normally distributed here, <https://github.com/florianhartig/DHARMa/blob/master/Code/DHARMaExamples/DevianceResiduals.md>. I find that checking for normality of the residuals is more confusing than helpful, since it makes people think that normality is some assumption of all GLM’s instead of only LM’s.

## Hypothesis tests and confidence intervals

Of course no hypothesis test or confidence interval should be calculated prior to addressing any issues you found in the model, such as overdispersion. Once you have a model that you are happy with you can get estimates with confidence intervals and any tests you want to report.

For any hypothesis tests, a good standard option is to use *drop in deviance* tests. These are also called *likelihood ratio* tests. Pick one of these terms and use it when describing these tests.

The “worst” tests we can use are the Wald z tests in the model summary output, so use those as a last resort. Among other issues, it is the Wald tests that are problematic when we have complete separation. We can also get Wald χ2 tests from the car package Anova() function, but those are worse than likelihood ratio tests. See a more complete discussion of hypothesis tests for GLM/LMM/GLMM models here, <https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#testing-hypotheses>.

If you decide to use a quasilikelihood model, it is standard to use an F-based likelihood ratio tests instead of a χ2-based likelihood ratio tests. F tests are more conservative. You will see an example of how to calculate these tests in R in lab this week.

The gold standard method to calculate confidence intervals from a GLM is by profiling the likelihood. This is true for any GLM and can be done with confint() on the object if the model was fit in R with the glm() function. Add-on packages will likely report Wald confidence intervals. These tend to be OK if there is no complete separation, in particular if we have a lot of data. Note if using quasilikelihood methods, any Wald confidence intervals should be based on the t-distribution and not the z-distribution.

## Results

The scale of interpretation for a binomial generalized linear model with a logit link is to odds. Estimated differences in groups are multiplicative and should be expressed as estimated odds ratios.

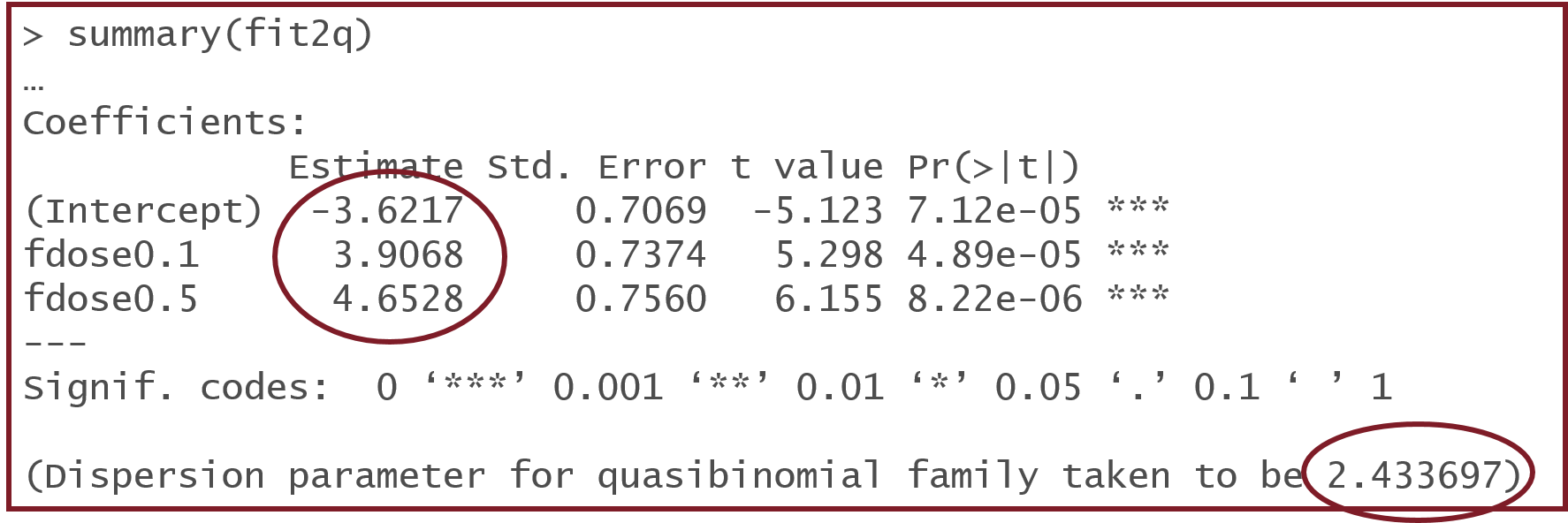
For categorical explanatory variables, reporting the estimated proportions/probabilities per group can help your reader interpret the odds ratios. This is especially useful when checking if a result is practically important or not. Make sure that you understand we can get a similar estimated odds ratio for very different estimated proportions. For example, do the math to compare the odds ratio of two groups where the first group has a probability of 0.75 and the second has a probability of 0.5 with another scenario where the first group has a probability of 0.03 and the second 0.01.

For continuous explanatory variables you will likely want to make a plot of the estimated relationship on the original scale for a similar reason.

# Output from R

I put together an excerpt of the results for a model fit to the lab example in Week 7 Handout 4, which I will hand out in class.

For a simple model like the one I fit to the motivating example this week, the summary of the model in R shows estimated differences in log odds among each group compared to the reference group.



The results above are on the model scale (log odds), so we would need to exponentiate the “Estimates” to get these estimated differences as odds ratios. If I wanted to show the estimated group proportions I would need to use the inverse link (plogis()) on the estimated log odds per group (not the *differences* in estimated log odds between groups).

You can see the overdispersion estimate (~2.4) is given at the bottom of the summary since I used the quasibinomial distribution for this analysis.

The overall test of an effect is called a drop-in-deviance or a likelihood ratio test. Be careful to never call this an “ANOVA”, because it’s not one. This is an F test if using a quasi-family distribution but would otherwise be a χ2 test.

