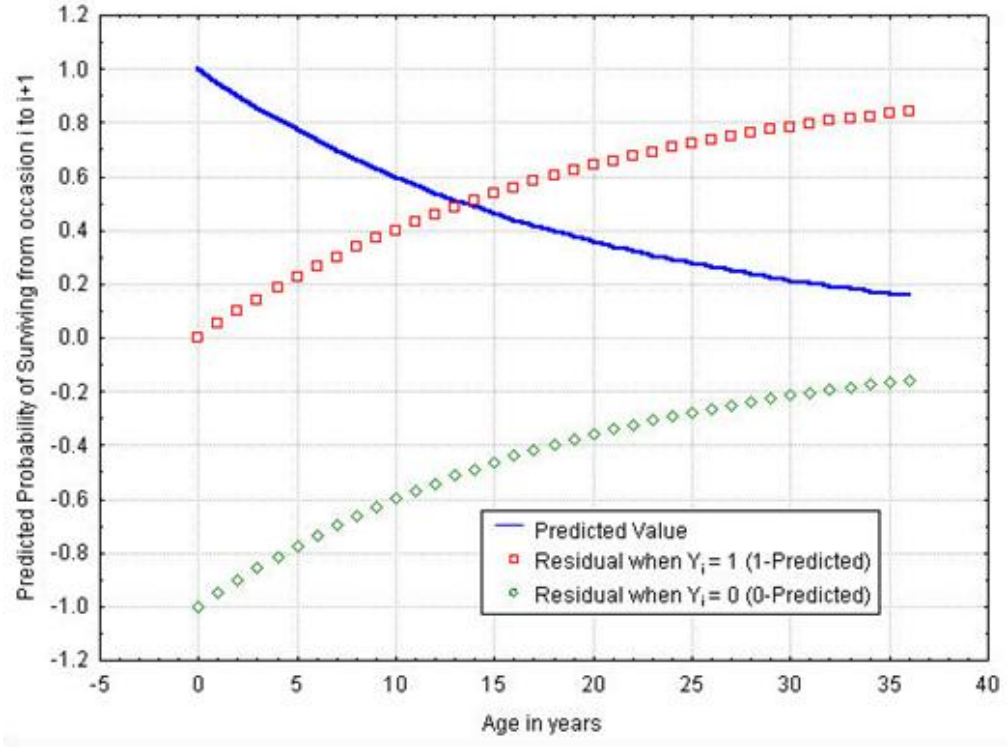
Testing GOF & Estimating Overdispersion

Your Most General Model Needs to Fit the Dataset

It is important that the most general (complicated) model in your candidate model list fits the data well. This model is a benchmark when evaluating other models. Thus, comparing the relative fit of simpler models to that of your most general model only makes sense IF the general model fits the data. That is, modeling only provides accurate description and inference for your data set if the general model fits the data set well. If your general model does not fit, you need to consider your data further and consider explanations for lack of fit. Such re-evaluation can improve your understanding of the problem at hand.

You evaluate whether the model fits the data via Goodness-of-Fit (GOF) testing. In typical regression problems with normally distributed errors, you can evaluate GOF by a variety of methods. For example, you can check R^2, plot the fitted model and the raw data against explanatory variables, plot the residuals, test for outliers and influential observations, etc. Some of these diagnostics are simply unavailable for binomial and multinomial response variables, e.g., R^2. Others are more difficult with binomial and multinomial response variables. For example, for a binomial response, analysis of residuals is difficult because each residual (ε_i) can take on only one of two values for any π_i (either 1-π_i or 0 - π_i). Thus, we don't expect the residuals to be normally distributed and don't know what distribution to expect them to follow under the assumption that the fitted model is appropriate. Consequently, plots of residuals against values of the explanatory variables are uninformative. So, we'll have to consider other diagnostics.



GOF is Evaluated with a Variety of Methods

Although GOF can be difficult for the data types we discuss in this class, there are a variety of methods that people have employed for testing GOF. To develop an understanding of how GOF works, it helps to start with a relatively simple GOF test first. Let's start with a simple comparison of observed and expected counts.

1. Observed vs. Expected Counts

An omnibus GOF test for CR models can be based on the m_{ij} array
Example with the European Dipper dataset – both sexes pooled

observed array

```
11    2    0    0    0    0
     24    1    0    0    0
          34    2    0    0
               45    1    2
                    51    0
                         52
```

expected (under CJS model, i.e., phi(t), p(t) given estimates and R_i)

```
11.0   1.9  0.1  0    0    0
      24.1  0.9  0.1  0    0
           34.0  1.8  0.1  0
                45.1  2.8  0.1
                     49.1  1.9
                          52.0
```

Many of the expected values are less than 2 (14 of 21), which is problematic for the test's performance. Also, the test gives only an overall idea of GOF. It does not take full advantage of the information that we have and inform us as to where problems exist, e.g., certain years or cohorts. Program RELEASE was developed to go further. Specifically, we can use Program RELEASE to examine whether animals behave the same:

1. *regardless of their past capture history (**Test 3**),* and
2. *regardless of whether they are captured on the current occassion (**Test 2**)*.

RELEASE partitions the data and analyzes it in pieces – RELEASE does this nicely for us and RELEASE can be run from inside Program MARK – it's under the TESTS button.
NOTE: *if you conduct a CJS-type analysis for serious work, you'll need to learn how to use Program RELEASE and Chapter 5 of C&W and references therein will help you to do that*.

For a logistic-regression analysis of known-fate data, you can compare the fitted model's predicted or expected counts of Y=1 and Y=0 against the observed counts for each level of the explanatory variable used in the study. The comparisons can be done using a Pearson X^2 or likelihood-ratio G^2 test statistic to test the null hypothesis that the model fits the data. For a fixed number of levels and when most expected counts are $\geq$~5, then these test statistics have approximate chi-squared distributions. The *df* for the test is equal to the number of levels of the explanatory variables minus the number of parameters estimated by the model.

$$\text{Pearson } X^2 = \sum_{i=1}^{\#levels} \frac{(Observed - Expected)^2}{Expected}$$

$$G^2 = 2 \cdot \sum_{i=1}^{\#levels} (Observed) \cdot \ln\left(\frac{Observed}{Expected}\right)$$

When the explanatory variables are continuous, it is difficult to analyze GOF without some form of grouping of the explanatory values. This is because the expected values for each observed value of the explanatory variable are usually small (<5) in this case. Thus, the test statistics do not have approximate chi-squared distributions. Let's consider an example for survival of cutthroat trout as a function of length. Length was measured as a continuous variable, 66 different lengths were obtained for 173 fish, and this variable was used to estimate S. The fitted model is: $\frac{e^{12.351+0.497 \cdot Length}}{1+e^{12.351+0.497 \cdot Length}}$. However, not many fish were the same length so the number of fish that are expected to live or die for any length is quite small for this model. The way around this is to group the fish into groups by length. Of course, decisions have to be made about how to group the data. It is common to partition the data so that all groups have ~equal sample size. Strategies for grouping and other forms of these tests (e.g., the Hosmer-Lemeshow test) are discussed well in books on the analysis of categorical data (e.g., Agresti 1996) and easily implemented in good logistic-regression software packages.

Length (cm)	Observed Survived	Expected Survived	$\frac{(Obs-Exp)^2}{Exp}$		Observed Died	Expected Died	$\frac{(Obs-Exp)^2}{Exp}$
<23.5	5	3.64	0.508		9	10.36	0.179
23.25-24.25	4	5.31	0.323		10	8.69	0.197
24.25-25.25	17	13.78	0.752		11	14.22	0.729
25.25-26.25	21	24.23	0.431		18	14.77	0.706
26.25-27.25	15	15.94	0.055		7	6.06	0.146
27.25-28.25	20	19.38	0.020		4	4.62	0.083
28.25-29.25	15	15.65	0.027		3	2.35	0.180
>29.25	14	13.08	0.065		0	0.92	0.920
		$Sum_{surv\ part}$	**2.181**			**$Sum_{died\ part}$**	**3.140**

There are 8 levels and the model contains 2 estimated parameters: an intercept (12.351) and slope term (0.497 for length) for predicting the survival of the fish as a function of length. Thus, *df* = 6. The 2 test statistic values are: Pearson X^2 = 5.321 (2.181+3.140) and G^2 = 6.2. Neither indicates lack of fit ($P > 0.4$).

As the number of explanatory variables increases, simultaneous grouping across many variables leads to a contingency table with many cells, and many will have small expected counts. Under these circumstances, you can group the observed and expected counts according to their predicted probabilities, 8 groups formed from low to high probability. If you are doing logistic regression, you might want to read Hosmer et al. (1997; A comparison of goodness-of-fit tests for the logistic regression model. Stat in Med 16:965–980) and see if anything newer has come out since then. The *rms* package in **R** implements some of the latest goodness-of-fit methods for logistic regression.

The Saturated Model & Model Deviance for Fitted Models

Because the deviance is used commonly in various diagnostics, it is worth working a bit to understand how deviance is estimated.

"The deviance of a fitted model is determined by comparing the log-likelihood of the fitted model to the log-likelihood of a model with *n* parameters that fits the *n* observations perfectly. Such a perfectly fitting model is called a *saturated model*." Neter et al. 1996. Applied linear statistical models. 4th Edition. Irwin.

It is rarely the case that the saturated model is in your candidate list. For example, consider CJS models for 1 group, 3 occasions, 2 releases (R1 & R2), and no covariates. Here, you *might* think that ϕ(t)p(t) would be the saturated model. But, … think about this a bit more. We said that we want a model with as many parameters as there are data points, and a truly saturated model would not just let ϕ and p vary by time. For this type of dataset we would have 6 data points: the number of animals with each of 6 different Encounter Histories: 111, 110, 101, 100, 011, and 010 . We could also envision parameters varying based on which Release or cohort the animal was part of. Thus, for CJS models, a reasonable way of calculating the saturated model's lnL is:

$$[Y_{111} \cdot \ln(Y_{111} / R_1) + Y_{110} \cdot \ln(Y_{110} / R_1) + Y_{101} \cdot \ln(Y_{101} / R_1) + Y_{100} \cdot \ln(Y_{100} / R_1) + Y_{011} \cdot \ln(Y_{011} / R_2) + Y_{010} \cdot \ln(Y_{010} / R_2)]$$

Think about this formula – this is definitely the lnL based on a set of probabilities that match the observed probabilities for each encounter history perfectly.

If you observed the following for a CJS study of 1 group over 3 occasions:

EH	111	110	101	100	011	010	Σ =6
Y_i	6	46	21	427	20	80	Σ =600

Here, R1 = 500, R2 = 100.

The lnL = 6 ln(6/500) + 46 ln(46/500) + 21 ln(21/500) + 427 ln(427/500) +
20 ln(20/100) + 80 ln(80/100)

Thus, the $\ln L_{\text{saturated model}}$ = -320.294, and the –2lnL = 640.589.

The saturated model fits the data perfectly assuming that the data are independent and that all animals with the same encounter history have parameter values that are identical (no overdispersion). If we don't know any covariate values for the animals we can't develop a more complex model of ϕ and p and we calculate the -2lnL as above. If individual covariates are included, the saturated model has a -2 log likelihood value of zero and the steps above aren't needed: the deviance for a model with individual covariates is just its -2 log likelihood value.

The saturated model gives us a baseline for evaluating the fits of other models. A fitted model with a small deviance fits the data almost as well as the saturated model, and a fitted model with a large deviance value does not fit very well. But, how do we decide when the deviance is too large?

Uses of Deviance in Analyses

Likelihood Ratio tests - LRTs compare deviances of nested models to conduct significance tests of various factors in the models. The difference between the deviance values (which is the same value as the difference between the –2lnL values) is distributed approximately as X^2 with the number of degrees of freedom equal to the difference in the number of parameters contained in the two models being tested.

Deviance GOF Test – Given the logic that a fitted model with a small deviance fits the data almost as well as the saturated model, and that a fitted model with a large deviance value does not fit very well, deviance itself is sometimes proposed as a GOF criterion. Formally, some people have proposed that deviance follows approximately a chi-squared distribution with degrees of freedom approximately equal to the difference in the degrees of freedom for the Saturated Model and the Fitted Model.

Example for CJS model: The Saturated Model for the CJS example estimates 4 probabilities. Okay, 6 probabilities are in the likelihood expression, but REMEMBER the probability of EH=100 is known once the probabilities of 111, 110, and 101 are estimated; and the probability of EH=010 is known once the probability of 011 is estimated (or vice versa).

The phi(t)p(t) model estimates 3 parameters. Remember: phi_2 and p_3 are confounded and estimated only as a product. Thus, when testing GOF for our most general model (ϕ(t)p(t)) using deviance, we have 1 degree of freedom. For our simple example, the
–2lnL for phi(t)p(t) = 642.41 and thus, the Deviance = 642.41 - 640.589 = 1.82
The probability of a chi-squared value $\geq$1.82 with 1 df = 0.823, i.e., it's a likely value to observe when the null hypothesis is true (H_0 = phi(t)p(t) fits the data as well as the saturated model).

Unfortunately, for CJS models it doesn't appear that, in most cases, deviance follows the chi-squared distribution well enough to provide a valid test. Ughh! Like LRTs, this is mentioned because you'll see this in many books and discussions of categorical data.

Deviance Residuals – GOF statistics provide summary indicators of overall fit. But, they do not inform us as to the nature of any lack of fit. Residuals are useful for this purpose. We're not going to discuss them at any great length, but you should know that there are several special kinds of residuals calculated for categorical response variables. One type is a deviance residual. As for regression of normally distributed Y values, there is a residual for each case in the study.

Overdispersion - ***The major use of deviance for our purposes will be to estimate the amount of over-dispersion in the data via c-hat.*** Overdispersion occurs when the variance of the response variable exceeds the nominal variance, e.g., for a binomial the nominal variance is *np(1-p)*. Overdispersion can be caused by:

1. lack of independence among animals – e.g., twins. You have less data than you think!
2. heterogeneity in the probabilities beyond that specified by the model, e.g., phi varies among animals within the same group during a single interval.

In ordinary regression models for a normally distributed random component, overdispersion due to heterogeneity is not a problem because the normal distribution describes variation using a separate parameter ($\hat{\sigma}^2$) than that used to describe the mean. This is not true for categorical response variables. For example, with the binomial distribution, the variance is a function of the mean ($\hat{p}$ (1- $\hat{p}$)/n).

Factors such as genetics, environmental conditions, etc. can cause overdispersion among animals. Apparent overdispersion occurs when the systematic component of the model (the regression string) is inadequate in some way. For example, you:

1. Omit important explanatory variables,
2. Fail to include sufficient interaction terms,
3. Assume a linear relationship between the transformed parameter (e.g., ln(S/(1-S)) and X β),
4. Have outliers in the dataset.

You should eliminate these possibilities before concluding that the data are overdispersed.

Let's think about overdispersion and how it relates to deviance and GOF. We said earlier that we wanted to know the deviance for our saturated model so that we'd have a baseline lnL value for evaluating other models. Hmmm, what if the saturated model is inadequate? How can this be? Well, take the simple example of estimating weekly survival for 4 weeks and 2 groups (let's say males and females). You might not know that within each sex, there are 2 sub-groups. Each of these sub-groups has different survival but you yet have no way of readily distinguishing these sub-groups by looking at them or measuring them! In this case, your saturated model might NOT fit the data and now your deviances are wrong.

As you can imagine, overdispersion, or at least apparent overdispersion, is common to many data sets. You need to concern yourself with it because **overdispersion causes you to underestimate the variance of parameter estimates and choose models that are too complex for the data at hand**.

To correctly estimate the variances, you need to inflate the nominal variance (that of the model) by the dispersion parameter, $\hat{c}$. When the data are not overdispersed, $\hat{c}$ =1.0. When the data are overdispersed, $\hat{c}$ is >1.0. Thus, $\hat{c}$ is often termed a *variance inflation factor*.

The good news is that the MLEs for the parameters are typically quite unaffected by overdispersion. That is, the estimates of the model's parameters are usually okay even when you have overdispersion.

Estimating c-hat - First, it is important that you estimate c-hat, the variance inflation factor, from your most general model (sometimes termed the *global model*). Your most general model is the most highly parameterized model in your set of candidate models, i.e., the model that you think explains the variation in your model quite thoroughly.

For some data types, you can use the deviance GOF test or Pearson GOF test to estimate c-hat. In either case the estimation is quite simple. For example, if the deviance is used, you simply take the deviance/df as an estimate of c-hat. Thus, in our earlier example for the CJS model, phi(t)p(t) is the most general model and would be used for estimating c-hat. Its deviance was

1.82 and there was 1 degree of freedom. Thus, the estimate of c-hat would be 1.82, which indicates some overdispersion. Unfortunately, this does NOT work well for open-population capture-recapture models such as CJS, and estimates of c-hat obtained this way tend to be biased high. Ugghh again!

We need to consider other ways of estimating c-hat. Fortunately, the median $\hat{c}$ procedure seems to work well for CJS models, and Lab 4 works through how to use the median $\hat{c}$ procedure. But, what do we do once we have $\hat{c}$?

Using c-hat – Improving Estimation and Model Selection for Overdispersed Data

Okay, let's say you've identified that there is some overdispersion in your dataset, what next? First you should be convinced that you don't have any missing covariates, interactions, etc., i.e., you've done your best at developing a good general model. You now need to complete 2 tasks: (1) adjust your AIC values so that your model-selection results are appropriate, and (2) inflate your variance estimates.

Model Selection based on QAIC$_c$ or AIC$_c$/ $\hat{c}$: If overdispersion has been identified, then model selection should be based on QAIC or QAICc. This is easily accomplished by simply dividing the –2ln(L) value used to calculate AIC or AICc values by c-hat, i.e.,

$$QAICc = \frac{-2\ln L}{\hat{c}} + 2k\frac{2k(k+1)}{n-k-1}$$

When $\hat{c}$ is >1, then the contribution to the QAICc value from the model likelihood declines relative to the penalty term for a given model. Thus, QAICc will tend to favor simpler models as $\hat{c}$ increases.

Variance Inflation - It turns out that it's easy to inflate your variance estimates. It's simply $\hat{c} \cdot \mathrm{var}(\hat{\theta})$. Similarly, you could multiply the estimated sampling standard error of each parameter estimate by the square root of $\hat{c}$: $\sqrt{\hat{c}} \cdot se(\hat{\theta})$. If $\hat{c}$ is close to 1, then there is little effect. ***Be sure to estimate $\hat{c}$ from your global model.***

Final thoughts: Your value of $\hat{c}$ may be >1 because the global model is structurally inadequate, i.e., there is no overdispersion. That is, in reality, had you obtained better covariates and better understood curvilinearities and interactions, you could have come up with a good model. With that model, the parameter values are constant within the constraints of the model. Further, you may truly have independent observations. Thus, in reality you could have a model in which no overdispersion exists. It will, of course, often be the case that you can't tell whether you estimate $\hat{c}$ >1 because of correctable model inadequacy (GOF problems), lack of independence, underlying heterogeneity in parameters (you can't measure covariates related to the heterogeneity and model it), or combinations of all of these. The good news is, that even in the face of not knowing why $\hat{c}$ is >1, incorporating $\hat{c}$ into model selection and variance inflation at least leads to conservative inference.