FES 524: Natural resources data analysis

Assumptions of linear models

2024-01-15

Contents

1 Checking model assumptions

Checking model assumptions is an important part of statistical analysis. Think back to lecture and the assumptions we had to make in order to derive the model of exponentially-distributed wait times between bird calls. We often prefer simple models (i.e., the principle of Occam's razor), but, in general, the simpler the model the stronger the assumptions. We need to have a way of checking those assumptions against the data to ensure our results are valid.

Since the assumptions for linear models are primarily about the model errors, we have to actually fit a model before we can check if the assumptions have been reasonably met. This can lead to a bad habit of looking at model results *prior* to checking model assumptions.

This is an approach I see a lot:

1. Fit model in R
2. `anova(model)`

3. `summary(model)`

In this approach, the analyst is looking at statistical results (using `anova()` and `summary()` prior to verifying the model is valid. This is poor statistical practice. In this class we will focus on using better statistical practice, where we make inference from models (i.e., look at results from models) only after making sure assumptions are reasonably met.

A better approach would be to:

1. Fit model in R
2. Check model assumptions
3. If assumptions not met, adjust the model
4. Once we have a model where assumptions are reasonably met, extract results to make statistical inference (e.g., `anova()`, `summary()`, `emmeans()`)

1.1 How to check model fit

We will be focusing on graphical checks of assumptions throughout the quarter. You will practice writing a succinct but thorough description of these graphical checks of assumptions and the results of those checks on assignments.

Using graphical checks for assumptions can feel subjective. This is because checking assumptions is subjective. Checking assumptions is always a judgment call, no matter how it is done.

Many people want to use statistical hypothesis tests to check assumptions because hypothesis tests have a veneer of objectivity. While we won't use such tests in this class, it is possible to use hypothesis tests in addition to graphical checks of assumptions. However, you should minimize their use and never rely solely on hypothesis tests to check assumptions.

Make sure you understand that, first, hypothesis tests don't return dichotomous results (will discuss this in detail in lecture). Second (and worse), hypothesis tests are extremely influenced by sample size. Minimal departures from an assumption will lead to tiny p-values if the sample size is large enough. Large departures from an assumption can still lead to large p-values if the sample size is small enough. Unless you have established a way to define a practically important departure for an assumption, the issue of sample size makes hypothesis tests fairly unhelpful for checking assumptions.

For assumptions about model errors, we use an estimate of the errors (assuming the model is an reasonable approximation to the data-generating process) for graphically checking assumptions. The residuals of a model (observed value minus model estimated value) are the estimates of the model errors.

2 Linear model assumptions

When we write a statistical model in mathematical notation, we can concisely state the assumptions of the model. Here is an example with two covariates, X_1 and X_2:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{1}$$

where $\epsilon_1, \epsilon_2, ..., \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. This model definition states that

1. The model is additive in the parameters (i.e., a linear model is appropriate)

2. The errors are independent
3. The errors are assumed to come from a distribution with a single overall variance (i.e., constant variance of errors / homoskedasticity)
4. The errors are assumed to come from a normal distribution

There is one more assumption that is often ignored since it is difficult or impossible to verify and is related to the assumption of proper model specification; that is the assumption of *strict exogeneity*, or the assumption that the explanatory variables are not correlated with the error term. In practice, this means that we are assuming there is no confounding variable that is missing from the model. For example, if you were to apply fertilizer to subsections of a field and measure plant response, but you just so happen to apply the high fertilizer treatment to the wettest parcel of the field, your treatment effect would be confounded with soil moisture and the estimates you get from the model will be biased. I bring this up because it is important to recognize that the estimates we get from regression models are often conditional on the model structure since the interpretation is "the estimated change in the mean response, holding all the other variables in the model fixed."

2.1 Additive in the parameters

You may hear this assumption described as "linear in the parameters" or "linear in the betas". Additivity in the mathematical representation of statistical model is captured with the addition signs: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$.

When a model is additive (i.e., linear), differences in means among groups are expressed with subtraction.

Example of results from a linear model:

> "The mean site growth increment for the heavy thinning group is estimated to be 0.8 inches higher than that of the light thinning group."

If this assumption is not met, the estimated means and standard errors will be biased and any hypothesis tests or confidence intervals will be incorrect.

2.1.1 Evaluating linearity

This assumption is primarily based on scientific expertise. There are variables that we expect should have multiplicative relationships, such as stream discharge or biomass. In other cases there may be interest in expressing results multiplicatively and so an additive model isn't useful.

The residuals versus fitted values plot will likely indicate a lack of fit if the additive model does not fit the data well. For example, a funnel shape in the residuals versus fitted values plot could be an indication that a multiplicative model is more appropriate (but not always).

An alternative model is a multiplicative model, which we will discuss later in the course. Figure 1 is a graphical example showing plots of the raw data for an additive versus multiplicative model:

2.2 Mutual independence of errors

The assumption of mutually independent errors (i.e., ϵ_i and $\epsilon_{i'}$ are independent for all $i \neq i'$) is the assumption of a linear model that is most important to valid inference (i.e., hypothesis test and confidence intervals). Note however, this assumption is not invoked in deriving the least squares estimates of the regression coefficients, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.

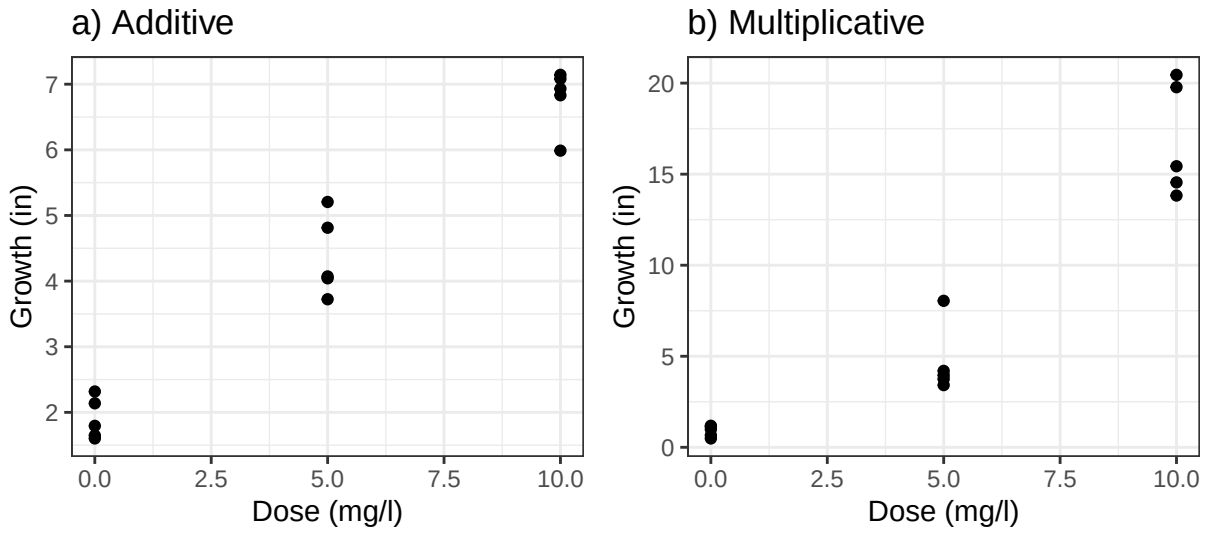

Figure 1: Comparison of data generated from a) an additive model and b) a multiplicative model of tree growth over doses of fertilizer.

The focus for checking this assumption is to check for a lack of correlation in the residuals.

Since the residuals are the observed values minus the value estimated by the model, all effects included in the model are subtracted out of the residuals. It might help some people to think of the residuals as showing "leftover" effects; it's everything leftover after subtracting out the effect of any variables in the model. If a variable that causes correlation in the observed values is in the model, the residuals will not be correlated. For example, if temporal autocorrelation in lake temperature can be explained by temporal autocorrelation in ambient temperature, then we may not even need a time series-specific model because the residuals will be independent.

This concept of model effects being subtracted out of the residuals is an important one. We certainly expect all observations within one group of a protocol to be related to each other if we expected the protocol to have an effect. This means the observed values should be correlated. If we didn't think the observed values were similar in some way within a protocol group, why would we have used that protocol at all? Since the protocol variable is in the model, though, the residuals are not correlated based on the protocol given. The effect of the protocol variable is subtracted out of the residuals.

Another indication that we might need to think about correlation is if the study design involves subsampling. For example, we usually expect measurements of trees in the same stand to be correlated because they all share a similar environment. The term pseudoreplication, often used for subsampling designs, is used to indicate that the subsamples are not independent observations of the response variable. Subsampling designs are a type of repeated measures, which we will discuss in detail in week 6.

If the assumption of independence of errors is not met, all estimated standard errors will be invalid and so all hypothesis tests and confidence intervals will also be invalid. Estimated standard errors are most often too small, so results can often be anticonservative (i.e., p-values too small and confidence intervals too narrow). However, the reverse can also occur.

2.2.1 Evaluating assumption of mutually independent errors

The assumption of independence of errors is most often justified based on a careful study design and scientific expertise. Generally this is done by justifying the independence of replicate study units, either by random sampling or random assignment of treatment conditions. We might expect study units close in space and/or time to be correlated, and so independence needs to be justified scientifically through, for example, professional judgement or past studies. In some cases, a graphical check of residuals can be performed.

For units measured through time or space, plot of the empirical autocorrelation function (ACF) or semivariograms of the residuals can be used to assess correlation. We will touch on this again in week 6.

Figure 2 shows a graphical example of an empirical ACF plot showing no correlation compared to correlation in the model residuals:

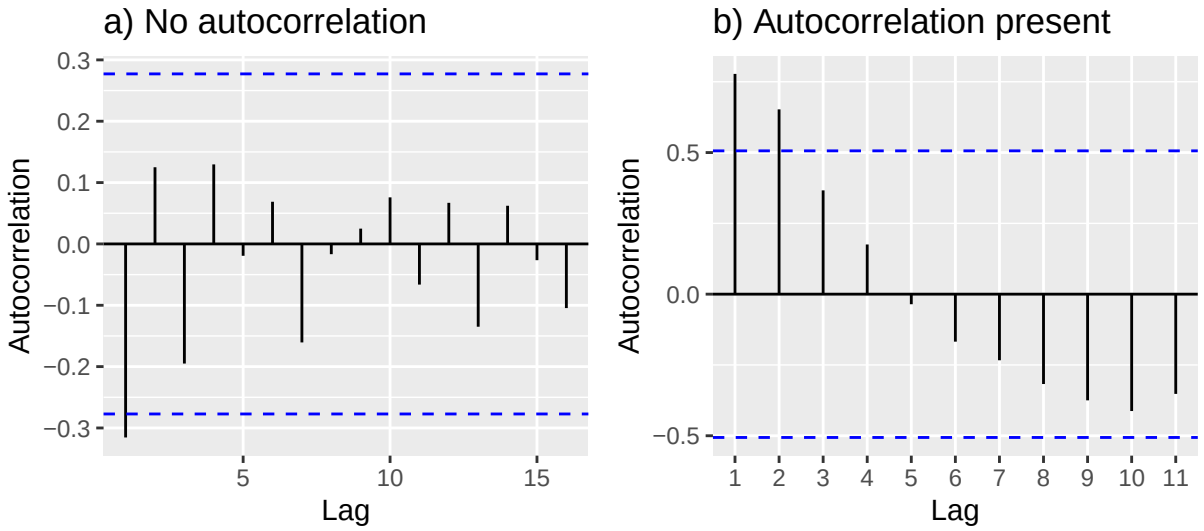

Figure 2: Empirical autocorrelation plots for a series with a) no correlation, and b) autocorrelation in time.

2.3 Constant error variance

This assumption comes from the definition of the errors as *identically distributed*, meaning every error term comes from a normal distribution with the same mean and variance. You may see this assumption described using the terms "homoscedasticity" (constant variance) and "heteroscedasticity" (nonconstant variance).

This assumption means that the variance of the errors are constant over all levels of a factor and across all values of continuous explanatory variables. Note that when we check the variances of the residuals, especially in small samples, we don't expect them to be identical among groups or across a variable; the variances need to be *reasonably* constant.

The term *reasonably* is, of course, "squishy". There is no hard and fast rule for how different the variances can be before it causes problems. *The Statistical Sleuth* book that you used in ST 511 and ST 512 has some discussion on how different the variances can be in Section 5.5. I often consider a doubling of the variance in one group compared to another as a red flag, but this can depend on other factors such as the sample size and expert knowledge about the response variable or study design.

If the assumption of constant variance of the errors is not met, the variances are too small for some comparisons and too large for others. P-values from hypothesis tests will therefore be invalid and prediction intervals will either be too wide or too narrow.

2.3.1 Evaluating homoskedasticity

This assumption can be evaluated using residuals versus fitted-values plots as well as residuals versus explanatory variable plots.

Figure 3 shows an example of constant variance and one with non-constant variance over an explanatory variable.

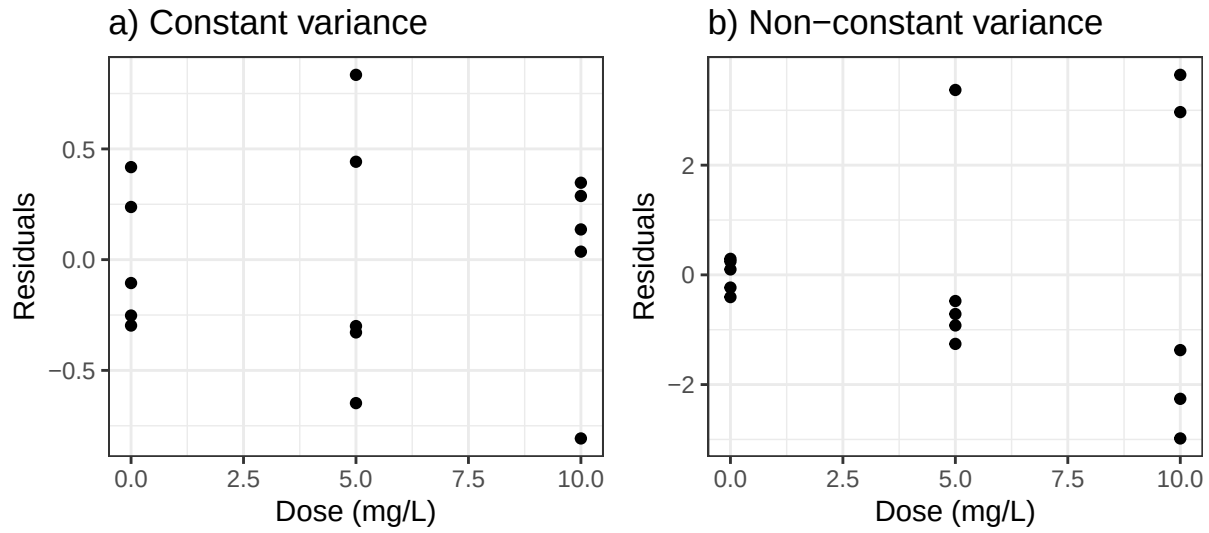

Figure 3: Example of residuals versus explanatory variable plot in a case with a) constant variance and b) one with non-constant variance.

2.4 Normality of the errors

This is the least important assumption of a linear model, even though it is often given considerable focus. A common mistake an analyst can make is to look for normality in the raw data. However, it is the *errors* that need to be approximately normally distributed. We see this in the statistical model: $\epsilon_1, ..., \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, not $y_1, ..., y_n$.

Normality is actually a very specific assumption with respect to the symmetry of the distribution (the bell-shaped curve) and spread (ratio of peak to tail probability, i.e. kurtosis = 0). Linear models are robust to this assumption, meaning they often perform well even when this assumption is not met.

Linear models are not robust to extreme skew, however. When we check the assumption of normality of errors, we are looking more for reasonably symmetric distributions of the residuals and not necessarily strict normality.

Note that assessing normality is very difficult when $n < 50$. This is ironic, since for large samples we may be able to rely on the central limit theorem to justify asymptotic normality of the errors, so deviations from normality may be less of a concern.

If the assumption of the normality of the errors is not met, estimates of means and standard errors are incorrect. P-values from hypothesis tests are incorrect and confidence intervals could contain impossible values (i.e., they could contain negative values even though the data are strictly positive).

2.4.1 Evaluating normality of errors

This assumption can be evaluated through plots of the residuals. We will use boxplots most often in this class, but you can also make histograms of the residuals or quantile-quantile normal plots (i.e., qq plots). We will discuss how to interpret qq plots later this term. For small samples like we have in this class, this can be a difficult assumption to assess.

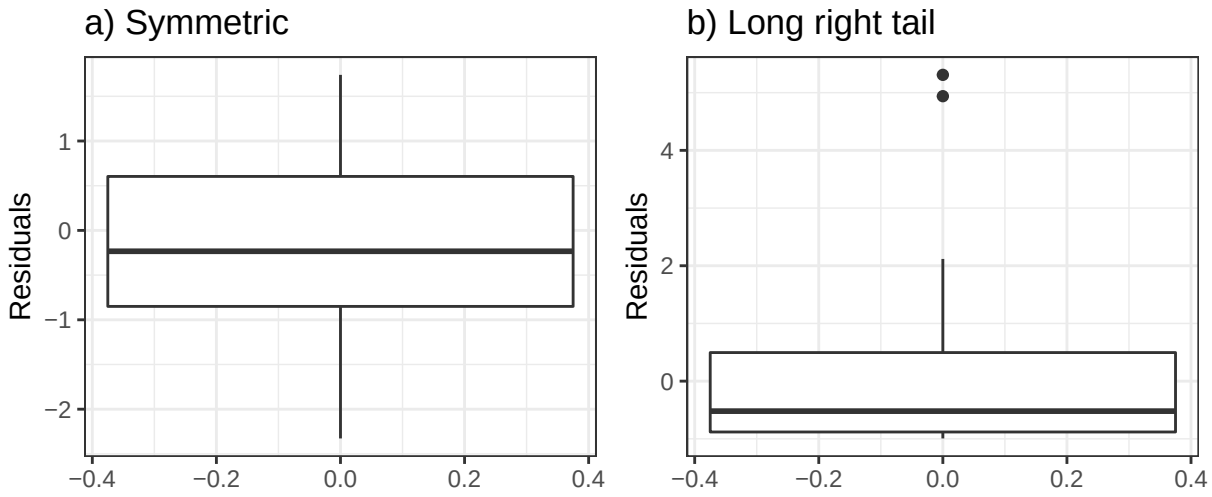

Figure 4: Example of a) symmetric and b) right-skewed residuals.

3 Reporting results

Once you have a model that you have deemed to have reasonably met all assumptions, you can report results from that model (note that we are assuming the "adjustments" we made to the model do not include variable selection steps, but rather adjustments to how the errors are modeled). These should include estimates and confidence intervals to answer all questions of interest. Results can also include p-values from hypothesis tests, although this is not required and should only be used in a thoughtful way (see for example Wasserstein and Lazar (2016), Wasserstein, Schirm, and Lazar (2019)).

3.1 Hypothesis tests

Overall tests of default null hypotheses are generally not particularly interesting. For example, remember that in a separate means model the overall null hypothesis is very coarse:

H_0: All group means are the same

H_A: At least one of the group means is different than the others

In addition, tests of hypotheses are done by assuming the null hypothesis is true. Many times we have designed a study because we think our protocol has an effect. We likely wouldn't waste time or money if we didn't think there was an effect *a priori*!

Another difficulty with hypothesis tests is how to interpret a large p-value. If you are using Fisherian hypothesis tests, you can never conclude that that the null hypothesis is true (note that you cannot conclude that the alternative hypothesis is true based on a null hypothesis test either). Supposedly, Sir Ronald Fisher, when asked how to interpret a large p-value, said "Get more data" (citation).

Issues with using p-values have been and continue to be much discussed throughout the statistics community and the topic of p-values is something we will revisit throughout the term. A good place to start in understanding p-values is the "ASA statement on statistical significance and p-values" (Wasserstein and Lazar 2016).

Throughout this course, we will practice using appropriate language for reporting statistics.

3.2 Estimates and confidence intervals

Plan on reporting estimates and confidence intervals for all comparisons or effects of interest. A common mistake is to only report confidence intervals for those comparisons that have small p-values, but this is not good statistical practice. See principles 3, 5, and 6 of the ASA's statement on p-values.

I generally subscribe to the idea of thinking of confidence intervals as compatibility intervals, which are interpreted as the set of values that could plausibly be the "true effect" given the data at hand and the modeling assumptions. This is a slightly inaccurate interpretation of a confidence interval *per se*, but is useful in practice. Check out this short blog post about this from one of the leading thinkers in statistics: https://statmodeling.stat.columbia.edu/2022/04/05/confidence-intervals-compatability-intervals-uncertainty-intervals/.

I would generally not recommend you report the confidence interval and then use it as a hypothesis test by checking if the null value is in the interval (despite the fact that many expressions for confidence intervals are derived by *inverting* a hypothesis test). If you want to do a hypothesis test, go ahead and report a p-value (along with the other relevant information of the test, e.g., degrees of freedom). After all, that's what null hypothesis testing was designed for. When you report confidence intervals, focus on the values in the confidence interval as a way to discuss if plausible values for the estimate are *practically* important.

3.3 Making conclusions

The primary focus when making conclusions based on the statistical results should be on practical importance. Too often investigators skip defining what value would indicate practical importance and instead focus on the null value (0 for linear models). Few differences are exactly 0, and results cannot be interpreted into scientific findings without some idea of what an important difference might be. You will define a practically important value to use in the conclusions from the analysis you do for your final project. Every assignment example has a practically important difference defined to use when making conclusions.

Some sort of graphic should be used to support the conclusions in most cases. This graphic should be "stand-alone", meaning everything in the graphic is defined in the caption.

3.4 Limitations

The limitations of a study involve things that may have gone wrong in the current study as well as recommendations on things to change for future studies. Limitations could involve, e.g., limitations on inference and concerns about making management recommendations on a single small study. For example, the scope of inference could be a limitation if it is very narrow.

We will practice identifying and discussing limitations throughout the course.

References

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. "The ASA Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129–33. https://doi.org/10.1080/00031305.2016.1154108.

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. "Moving to a World Beyond 'p < 0.05'." *The American Statistician* 73 (sup1): 1–19. https://doi.org/10.1080/00031305.2019.1583913.