Fundamental Modeling In R

Announcements

- Hope your spring break was rejuvenating!
- Final Project Proposal pushed back to April 1st

A quick goal setting anchor for today’s lecture

- We want to use statistics to better understand our data
- Most geography specific analytics are derived from traditional statistics
- Our data is a representation of real life processes
- Real life and the people measuring real life aren’t perfect, so much of statistics is determining how ‘wrong’ our data is
- How ‘wrong’ our data is can be quantified, which can help us adjust the interpretation of our data
- Our data is much more useful when we know how useful it really is!

1. Statistical Concepts

Statistical Concepts

- Sample vs. Population
- Error
  - What is Error?
  - Sums of Error (SE)
  - Sum of Squared Error (SSE)
- Variance
- Standard Deviation
- Probability Distributions
- Probability & Likelihood
- Confidence Intervals

Sample vs. Population

- **Population**: The entire set of individuals of interest, or the entire set of measurements obtained from all individuals of interest.
    - Represents the total set from which a statistical sample is drawn and to which any statistical analysis is ultimately aimed to be generalized.
- **Sample**: A subset of the population selected for measurement, observation, or analysis, to make inferences about the entire population. Samples are used because it is often impractical or impossible to collect data from every member of a population.

What is Error?

Error = $(x_i - \bar{x})$

xi = An observation we have

$\bar{x}$ = the average (mean) of our entire dataset

Error = Observed - Expected

- We would ideally expect to see the average of our dataset everywhere we go, so when we don't see our average, we have error.
- This is cool, but this is only for one observation, how is that useful?

Sums of Error (SE)

- Sums of errors: $\sum(x_i - \bar{x})$

- Adds up the error in our entire dataset (all observations)
- Does not consider their direction (positive or negative).
- A major issue with SE is that positive and negative errors can cancel each other out, leading to a misleading representation of the model's accuracy.
- This cancellation effect can make a model appear more accurate than it actually is because the errors in opposite directions neutralize each other.
- We can solve this by squaring!

Sum of Square Error (SSE)

- Sums of squared errors: $\sum(x_i - \bar{x})^2$

- Used to quantify error = the difference between an observation and the expected value.
- “The sum of the squared differences between each value and the mean”
- Squaring has some desirable properties
  - Converts negative values to positive
  - Penalizes high values
  - Large errors are very influential because the squaring function is nonlinear
- Problem: The amount of observations you have directly affects the SSE, making it not comparable to other datasets!

Normalization

- Normalizing by population size:
  - $\frac{1}{N}$ and $\frac{1}{N-1}$

- This allows us to compare results between different datasets
- The number of observations can change the value of a SSE, but by normalizing our data by how many observations we have, we can standardize the metric further.

Variance

- Variance: A measure of dispersion or spread caused by error
  - The average of the squared differences (observed - expected)
- It's like saying how much spread is there in a variable, with reference to itself.
  - That's why the (observed - mean) is squared
- **It's just the SSE but normalized** by the [adjusted] sample or population size.
- Note, the units are squared so they are somewhat uninterpretable.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ **Population Variance**

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ **Sample Variance**

u = mean

- Normalizing by population size:
  - $\frac{1}{N}$ and $\frac{1}{N-1}$

Covariance

- Covariance is variance, but crossed with another variable
- Measures the dispersion of one variable, x, in the context of the dispersion of a second variable, y!
- Covariance can be positive or negative - unlike variance which is strictly positive due the squaring!
- No Covariance
  - Above-average values of x are just as likely to occur with above or below average values of y.
  - The negative and positive terms cancel out: the overall sum should be near zero

$$\textit{Covariance}$$

$$s_{jk}^2 = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{n-1}$$

Standard Deviation

- Standard deviation: The square root of the variance. The standard deviation has special meaning for normally distributed variables as we'll get to…
- Note, **the standard deviation is in the same units as the measurement variable, which are therefore interpretable**.

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Sample Standard Deviation

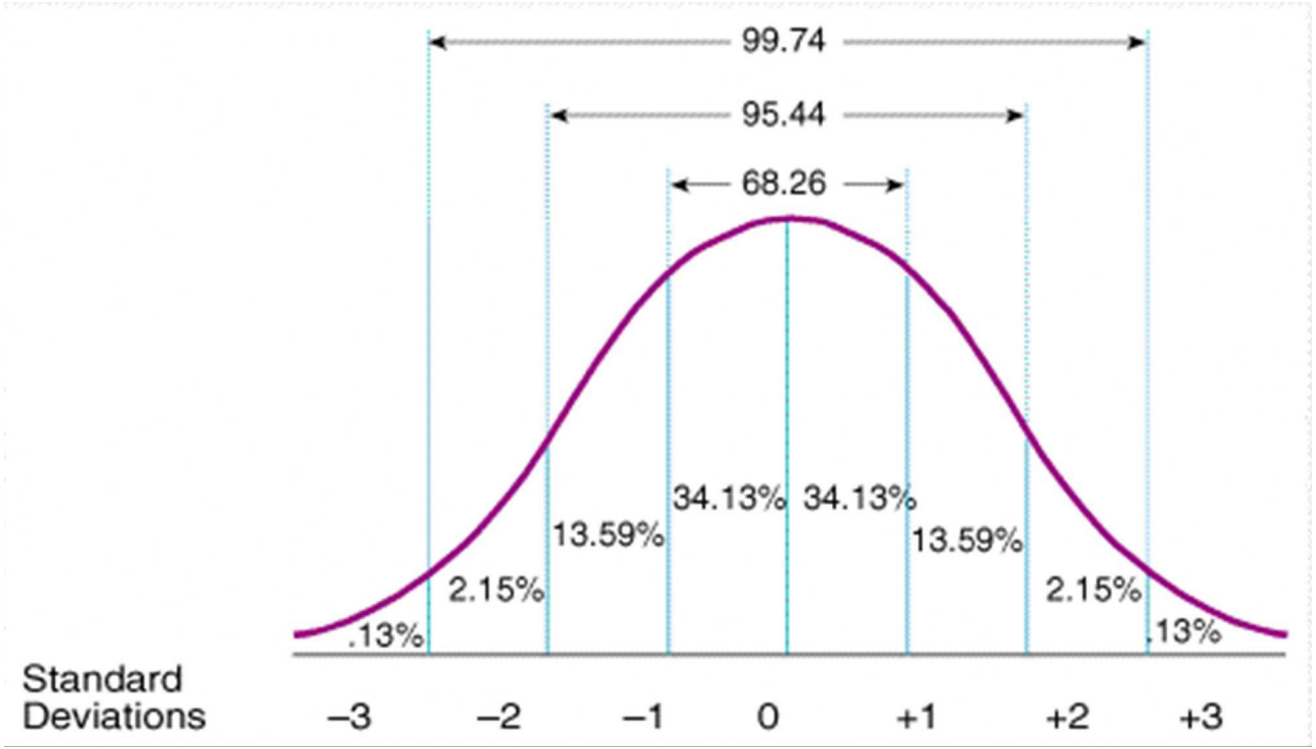
Basic Probability

- Probability measures the likelihood that an event will occur
- For example, in a coin flip there is close to a 50/50 chance you will get heads or tails. The probability of you getting tails is 50%.
- This is a single use case, but becomes more complicated once we look to describe the probability of events occurring on a continuous scale.
    - For example, the probability of temperature being Y at an elevation of X, where elevation X is a continuous set of values between 100 - 500 meters.
    - This is where probability distributions come in!

Probability Distributions

- When we move from simple events to looking at the distribution of outcomes across a continuous or discrete variable, we start working with Probability Distribution Functions (PDF).
- A PDF describes the likelihood of each possible outcome of a variable.

The Normal Distribution & Standard Deviation



Credit to SubjectCoach

Types of Distributions

- Uniform Distribution: All outcomes are equally likely.
- **Normal** Distribution (**Gaussian** Distribution): Describes data that clusters around a mean or average. It's symmetrical and has the well-known bell curve shape.
- **Poisson** Distribution: Models the number of times an event occurs in a fixed interval of time or space.
- **Binomial** Distribution: Represents the number of successes in a fixed number of binary (success/failure) experiments.
- Exponential Distribution
- Gamma Distribution…

Probability Distributions & Error

- Probability distributions describe how errors are distributed around the true values of our dependent response variable.
- By modeling the errors as coming from a specific distribution (ex: normal), we can quantify the uncertainty (error) in our predictions by it's deviation.
- This uncertainty reflects the natural variability in data and the limitations of the model to capture the true underlying process perfectly.
- Represents the random variability about the expected value (typically mean).
- **Probability distributions describe residual errors**
- **This is why we want our errors to be normally distributed!**

Likelihood

- Likelihood refers to **the probability of observing the given data under specific model parameters**. It's a measure of how well the model parameters explain the observed data.
- While probability predicts future outcomes based on known parameters, **likelihood assesses the plausibility of parameters given observed outcomes**.
- Example: Consider a coin toss with heads and tails!
  - Probability: What is the probability of flipping the coin twice and getting one head and one tail in any order?
  - Likelihood: Given some observed data (one head and one tail), how likely are different probabilities of landing a head (p) in explaining our observed data?

Confidence Intervals

- A confidence interval (CI) provides a range of values that is **likely** to contain a population parameter with a certain level of confidence.
- The **confidence level**, often set at 95% or 99%, indicates the probability that the calculated CI range of values contains the true population parameter.
- Example: A 95% confidence interval means that if we were to take 100 different samples and compute a CI for each, we would expect about 95 of those intervals to contain the true population parameter.

2. Modeling Concepts

Modeling Concepts

- Independent vs. Dependent Variable
- What is a model?
- Deterministic vs. Stochastic
- Parametric vs. Non-Parametric
- Parametric Model Assumptions

Independent vs. Dependent Variable

- **Independent Variables** (Predictors): These are the variables that (may) have an effect on another variable, which the dependent variable.
- **Dependent Variables** (Outcomes): These are the variables that respond to changes in the independent variables.
- The dependent variables ‘depends’ on the independent variable to change.
  - X = independent variable
  - Y = dependent variable
- Example: In looking to determine the relationship between elevation and temperature, elevation is our independent predictor variable while temperature is our dependent response variable!

What is a model?

- A statistical model represents the relationship between one or more independent variables (predictors) and a dependent variable (outcome).
- Statistical models are used both for prediction and inference.
- Models creates the best equation it can for your data, which is used to represent the relationship between the variables of interest.
- At the end of the day, a statistical model is an equation that we need to uncover by using our data.
    - $y = \mathbf{m}x + \mathbf{b}$
    - We are essentially solving for m and b with our data

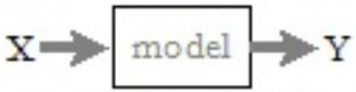
“All models are wrong, but some are useful”

Model Anatomy - Deterministic & Stochastic Models

- **Deterministic Model**: Represents the underlying ecological process, and estimating the parameters of this model is the focus of statistical modeling.
  - Example: y = **mx + b**
- **Stochastic Model**: Incorporates randomness or uncertainty into the model, which is typically thought of as error.
  - Example: y = mx + b **+ e**
    - e = error or stochastic component
- Variation from the deterministic model are errors also called "**residuals**" as they represent the residual variability not accounted for by the deterministic model.
  - **Residual = Error = Observed - Expected**

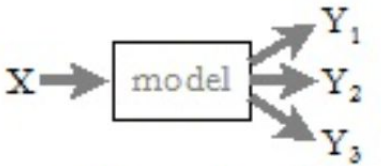
Deterministic vs. Stochastic Modeling

Deterministic model



- Given the input data, the model determines exactly the output; we always get the same result

Stochastic model



- Given the input data, the model gives variable output; we always get a different result due to randomness

A special thanks to Kevin McGarigal for this graphic.

Process Variability & Stochasticity

- **Measurement Error**: Error in measurement at a physical/manual level.
- **Process Variability**: Unlike measurement error, affects the future dynamics of the ecological system. This is important in more complicated models, where the process error has a chance to feed back on the dynamics.
- **Demographic stochasticity** is the innate variability in outcomes due to random processes even among otherwise identical units.
- **Environmental stochasticity** is variability imposed from “outside” the ecological system, such as climatic, seasonal, or topographic variation. We usually reserve environmental stochasticity for unpredictable variability.

Parametric Methods

- Parametric methods are any statistical techniques which operate under a set of fundamental assumptions.
- Parametric Assumptions
  - Normality
  - Homogeneity (Constant Variance)
  - Independence of observations
  - Fixed X (no measurement error in predictor variables)
- Limitations of Parametric Methods
  - Strict on assumptions, and in ecology these are frequently violated
  - Not great for describing complex, non-linear relationships
- **Non-Parametric Methods** are any models that violate these assumptions!

Parametric Assumptions - Normality

- Normality: Under repeated sampling, data would be normally distributed.
- This DOES NOT mean that values for each variable in a data set must be normally-distributed by themselves
- Normally distributed around each predicted value in the deterministic model
- **We really care about the normality of the residuals from a model** ***
    - This can be inspected by plotting a Q-Q Plot

Parametric Assumptions - Homogeneity of Variance

- we expect the variability of our observed values to be the same
- As x increases, the variance remains the same and does not change - which can be seen with an unchanging distribution of errors
- This is often very unrealistic!
- Ex: Plants that grow in poor soil have an avg. 2.0g biomass, while those grown in rich soils have an avg. 20.0g biomass - so the magnitude of variation does not make sense to be the same in each group

Parametric Assumptions - Independence of Observations

- Sampling is randomized
- Knowing something about observation x1 gives us no information about observation x2
- Non-independence can result from:
  - Proximity in space or time
    - Autocorrelation: the similarity between observations as a function of the time lag or between them
  - Hierarchical structure
    - Ex: if you are sampling invertebrates from two beaches, you can expect non-independence of invertebrates collected from the same beach - which is tricky to deal with

Parametric Assumptions - Fixed X

- Perfect accuracy and no measurement error in our predictor (explanatory) variables
- This assumption is frequently violated!
- It’s OK-ish if the noise in the predictor variables measurement is small relative to the noise in the response

Model Error

- Measurement and Process Variability co-occur with **Model Error**, which is the error in the modeled relationship introduced by the choice of a poorly specified or incorrect statistical model.
- Choosing the wrong **deterministic function** or **stochastic model** will exacerbate the apparent error, and thus the Model Error
    - If the model is improperly specified to reflect the process under investigation, then the signal-to-noise ratio will likely decrease

3. Statistical Modeling

Linear Models - Null vs. Alternative Hypothesis

- The **null hypothesis** in linear regression suggests that there is no relationship between the independent predictor variables and the dependent response variable.
- The **alternative hypothesis** suggests that there is a significant relationship between the independent variables and the dependent variable, which contradicts the null hypothesis.

Linear Models - Null Model Example

- lm.null = lm(y ~ 1)
  - y = dependent response variable
  - 1 = species no independent effect variable
- Includes no predictors, and essentially predicts the dependent variable, y, using only the mean of y (assuming a constant value across all observations).
- The null model serves as a baseline or reference model to compare against more complex models that include one or more predictors.
- It assesses the total variance in the dependent variable without accounting for any effects from independent variables.

Linear Models - Standard Example

- lm.reg = lm(y ~ x)
  - y = dependent response variable
  - x = independent effect variable
- A standard linear model in regression analysis includes at least one independent predictor, x, and examines how changes in x are associated with changes in y.
- summary(lm.reg)
  - Residuals
  - Standard errors
  - p-values for the coefficients
  - R2
  - F statistics

Linear Models - P-Values

- The probability of observing our data assuming the null hypothesis is true.
- In the context of linear models, a p-value is calculated for each predictor variable and tests the null hypothesis that the independent predictor variable has no effect on the dependent response variable.
- Small p-values (typically <0.05) suggest rejecting the null hypothesis, indicating a statistically significant relationship between the predictor and the outcome.
- If the probability of the null hypothesis being true is 5% (p=0.05), that is a pretty low chance in ecology so we can reject the null hypothesis!

Linear Models - F Statistic

- Is a measure used to evaluate the overall significance of the model.
- It tests the null hypothesis that all regression independent predictor variables are equal to zero (no effect) versus the alternative that at least one is not.
- A higher F-statistic (and corresponding small p-value) indicates that the model explains a significant portion of the variance in the dependent variable, suggesting that the model fits the data better than a model with no predictors (the null model).

Linear Models - R^2 (Coefficient of Determination)

- Represents the **proportion of variance in the dependent variable that is predictable from the independent variables**.
- Ranges from 0 to 1, where higher values indicate a better fit of the model to the data.
- In ecology, this is typically all over the place. I stray away from trusting this because you can have great data and significant values across the board, but a low R^2.

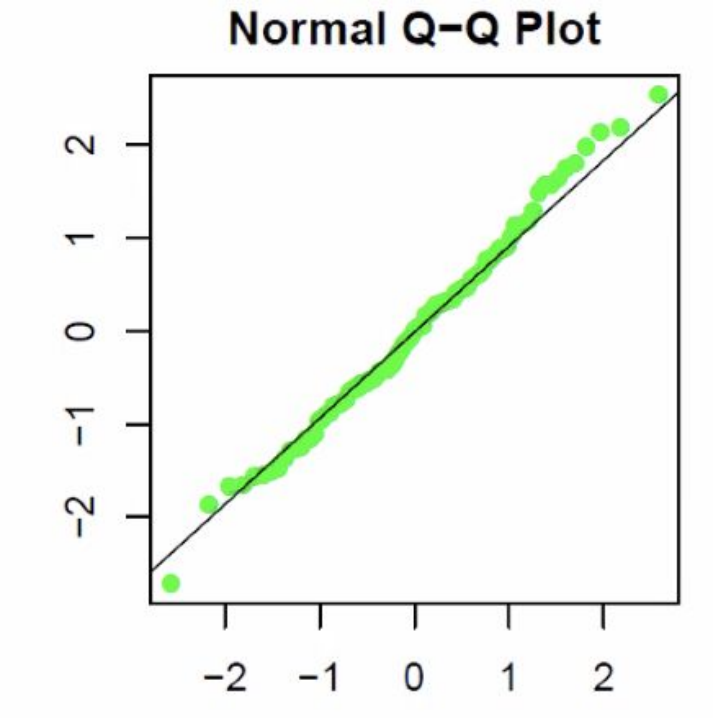
(QQ) Quantile-Quantile Plots

- Graphical tool used to determine whether the data follow a particular distribution
  - ex: Normal distribution
- If the resulting points lie roughly on a straight diagonal line, then the distribution of the data is considered to be the same as a normally distributed variable
- The qqnorm plot depicts the sample quantiles on the x axis against the theoretical quantiles from a normal distribution of the same sample size on the y axis.
- The straight line in the plot is typically obtained by connecting the 25th and 75th quartile points.

QQ Plot - Example

> plot(linear_model)

- Points are along the line, meaning that the data is normally distributed.
- The more the points deviate from the line, more less normally distributed the data is.
- QQ Plots are sensitive to deviations from normality, so you will know it when you see it!



4. Linear Modeling Demo

Linear Modeling Workflow

1. Check your data assumptions
    - Normality
    - Homogeneity (Constant Variance)
    - Independence of observations
    - Fixed X
2. Run linear model
    - linear_model = lm(response ~ predictor)
3. Plot your Q-Q Plot to ensure normality of residuals
    - plot(linear_model) # will give you a Q-Q Plot!
4. Check model statistics
    - summary(linear_model)

Linear Modeling - Example Dataset

- Bird Habitat & Diversity Dataset
- Simpson's Diversity Index is a measure of diversity which takes into account species richness and evenness.
- As species richness and evenness increase, so diversity increases.
- **b.sidi** = Simpson’s diversity index for breeding birds ***
- **s.sidi** = Simpson’s diversity index for vegetation cover types ***

Linear Modeling - Fitting Model Example

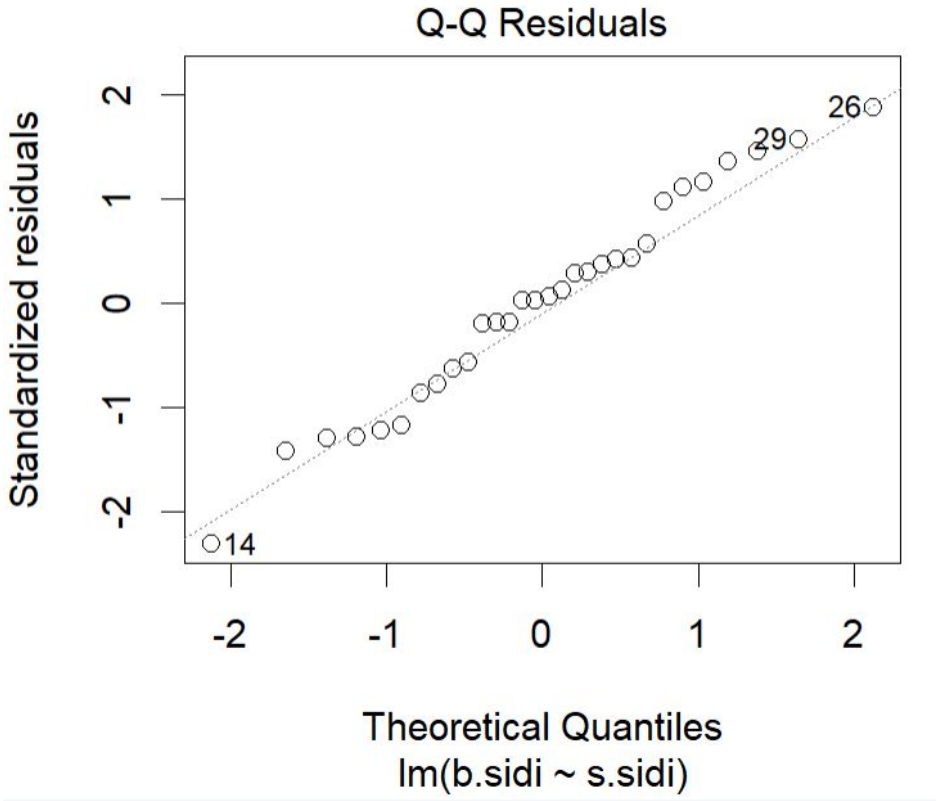
```
fit_1 = lm(b.sidi ~ s.sidi, data = dat_all) # s.sidi is the predictor
```

- We can use this model to determine the relationship between the diversity of birds and the diversity of the vegetation cover types they are in!
- s.sidi (veg diversity) is the independent predictor variable
- d.sidi (bird diversity) is our dependent response variable

Linear Modeling - Checking Residual Normality

> plot(fit_1)

- Points are close to line
- Looks normal to me!



Linear Modeling - Model Summary Example

```
> summary(fit_1)

Call:
lm(formula = b.sidi ~ s.sidi, data = dat_all)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0180574 -0.0055768  0.0003214  0.0041270  0.0146152

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.071170   0.003234  22.007  < 2e-16 ***
s.sidi      -0.024371   0.006418  -3.798 0.000721 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007995 on 28 degrees of freedom
Multiple R-squared:   0.34,     Adjusted R-squared:  0.3164
F-statistic: 14.42 on 1 and 28 DF,  p-value: 0.0007212
```

b.sidi → (Intercept)

← Less than 0.05 Significant!

high! → F-statistic: 14.42

← Less than 0.05 Significant!

Linear Modeling - Interpreting Variable Model Results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.071170   0.003234  22.007  < 2e-16 ***
s.sidi      -0.024371   0.006418  -3.798 0.000721 ***
```

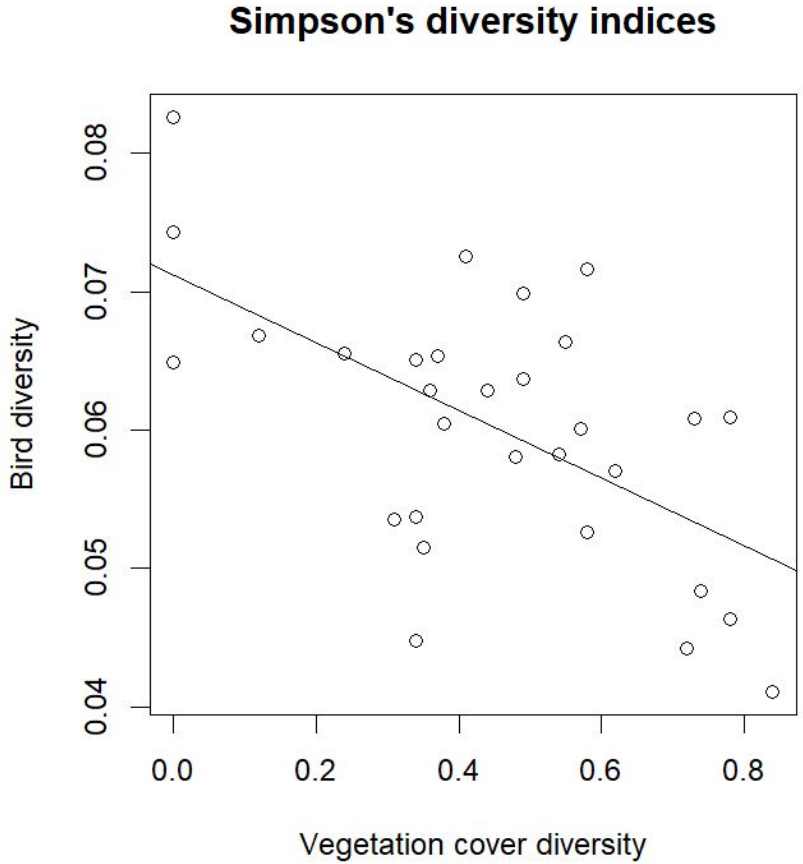
- The intercept in a linear regression model is the expected mean value of the dependent variable when all predictors are = 0.
- In this model, the intercept can be interpreted as the expected value of b.sidi when s.sidi is equal to zero.
- Estimate is the for the intercept is the “+b” in y=mx**+b**
- Estimate for s.sidi is the “m” in y=**mx**+b as it is the slope for our fit
- **Pr(>|t|)** is the p-value for the t-test, which determines if d.sidi and s.sidi are significantly different from each other.
    - Both values or statistically significant (<0.05) and thus we can reject the null hypothesis and now say that d.sidi & s.sidi have a statistically significant relationship that is different than 0!

Linear Modeling - Interpreting Model Results

- The **F-statistic is 14.42, which is high**
- The **p-value is 0.0007212, which is very low**
- High F + Low P = **the model is statistically significant!**
- Adjusted R^2 value is low, but because this ecology we don't mind because the F-statistic and p-values are both significant!

Linear Modeling - Plotting Example

```
{ # Plot with simple linear regression
plot(
  b.sidi ~ s.sidi, data = dat_all,
  main = "Simpson's diversity indices",
  xlab = "Vegetation cover diversity",
  ylab = "Bird diversity")
abline(fit_1)
}
```



What if we have more multiple predictor variables?

- This is what we call a **General Linear Model**, and these deserve an entire lecture of their own!
- We will go over GLMs in our advanced statistics in R lecture in a few weeks
- If you have taken or are currently enrolled in Applied Ecological Statistics, that is the core focus of that course!

Thanks!