Murphy's Law: The Biostatistician’s Version

Ashley E. Mullan, MS
September 11, 2026

https://ashleymullan.github.io/talks/

Anything that can go wrong will go wrong.

It all starts with a well-designed experiment.

Planning in advance is a superpower!

- A **well-scoped question** points you towards reasonable **measurements** on the appropriate **population.**
- Specifying your **unit of analysis** helps you plan a **sample size**.
- Understanding the **biological mechanism** of interest and **data collection protocol** move you towards an **analysis plan.**
- Thinking through **surprises** can suggest **sensitivity analysis**.
- Good decisions on the **front end** can save a headache on the **back end**!

Setting the Scene

(Variables and Notation)

- Let **X** be the variable we **manipulate**, like access to water.
- Let **Y** be the **outcome** we're interested in, like body temperature.
- Let **Z** be a **covariate** that we may want to **adjust** for, like bird weight.
- We'll tweak these a bit as we discuss **different models**. In general, some **function of the outcome** is modeled by a **function of the explanatory variables**.

$$f(Y) = g(X, Z)$$

Regression is a great (and very interpretable) starting place when thinking about your analysis!

Simple Linear Regression

E(Y) = mX + b **(think `lm()` in R)**

- Let's say we either give our birds **water** or **no water** and land on this **model output.**
- We'd expect a typical bird with **no access to water** to have a body temperature of about **90°F**.

Simple Linear Regression

term	estimate	standard error	t statistic	p value
(Intercept)	90.0	1.3718×10^{-4}	656,065.0	0.0000
x	−3.0	1.9400×10^{-4}	−15,462.5	8.4298×10^{-315}

- We'd expect birds with **access to water** to have lower body temperatures of about 3° **on average.**
- These interpretations can **generalize** outside of this experiment!

Simple Linear Regression

E(Y) = mX + b **(think `lm()` in R)**

- For a typical bird with **X = 0**, we **expect Y** to be about **b**.
- Between two birds with a **one unit difference in X**, we expect them to have an **m-unit difference in Y**.
- We can do a few **math tricks** to adjust those **interpretations**.
    - If we **center** the X variable before fitting the model, then **b** becomes the **expected Y** for a bird with the **average value of X**.
    - If we want to compare birds with a **K unit difference** in X, we expect their **difference in Y** to be about **Km**.

Multiple Linear Regression

$E(Y) = m_1X + m_2Z + b$ **(think `lm()` in R)**

- Let's suppose we think the **weight** of the bird (in ounces) might **impact its body temp** and model accordingly.
- Between birds an ounce apart, we expect the bigger bird to have a 0.2° higher temperature **on average** if they have the **same water access.**
- Among birds of the **same weight**, we expect those with **access to water** to have body temperatures **3° lower on average** than their **no-water** counterparts.
- Again, we can **generalize**!

Multiple Linear Regression

term	estimate	standard error	t statistic	p value
(Intercept)	90.0	4.9561×10^{-4}	181,591.1	0.0000
x	−3.0	1.8827×10^{-4}	−15,933.0	4.3960×10^{-313}
z	0.2	5.4755×10^{-5}	3,654.0	4.7584×10^{-251}

Multiple Linear Regression

$E(Y) = m_1X + m_2Z + b$ **(think `lm()` in R)**

- For a typical bird with **$X = 0$** and **$Z = 0$**, we **expect Y** to be about **b**.
- Between two birds with a **one unit difference in X** and **matching Z**, we expect them to have an **m_1-unit difference in Y**.
- Between two birds with a **one unit difference in Z** and **matching X**, we expect them to have an **m_2-unit difference in Y**.
- The **centering** and **scaling** tricks from simple linear regression **still apply**.
- In general, you can **plug in different inputs** and **subtract** their expected Y values to get that difference. This is a **contrast**.

Assumptions

(Simple and Multiple Linear Regression)

- Each model assumption is used for a **different purpose**.
- To look at **associations**, you'll need:
  - Your parameter estimates to be **approximately normal**
  - **Independent** observations
  - Approximately **equal variance** between groups
- There are a few ways you can **check** these and **correct** for them (if needed)!

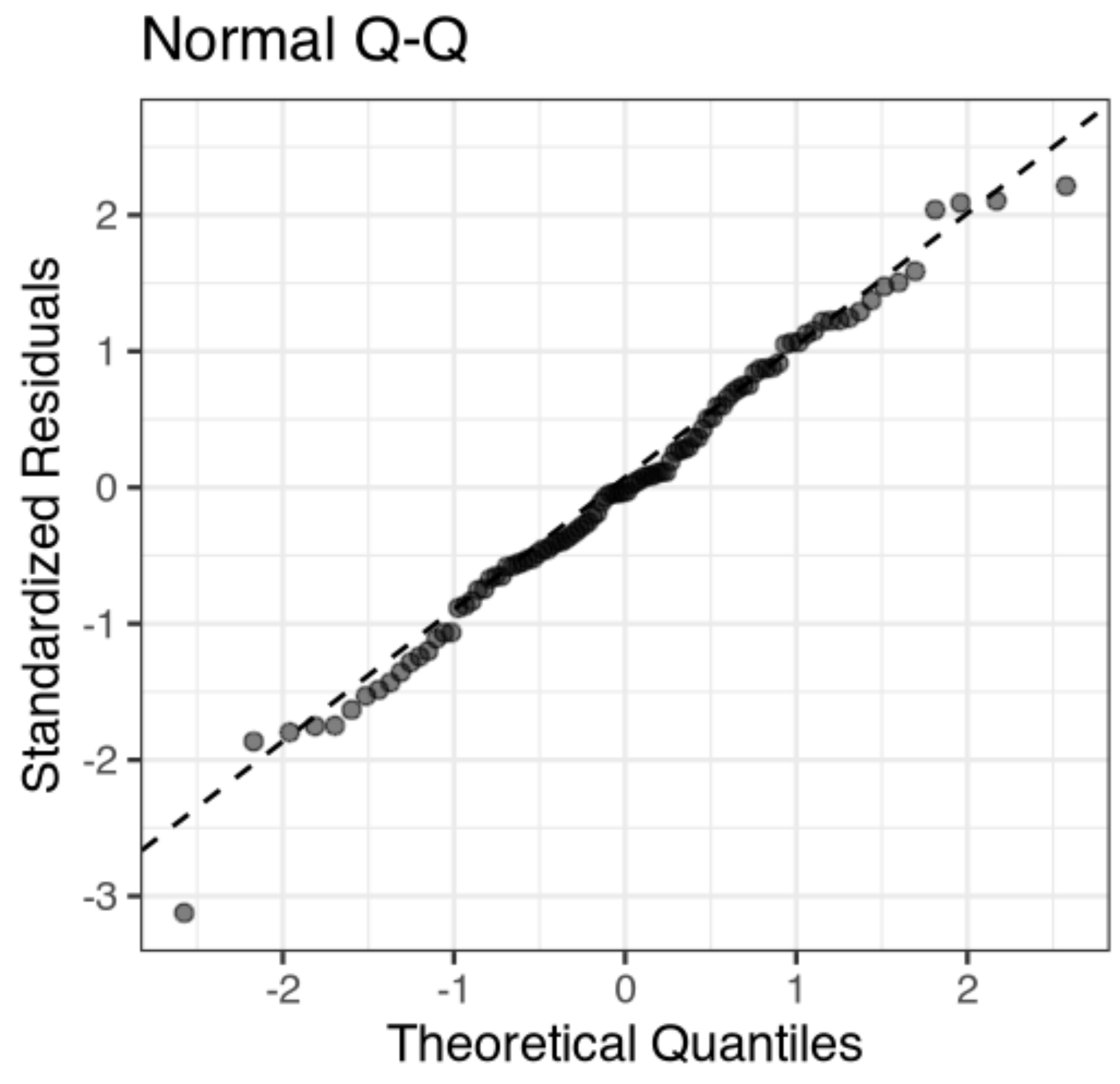
Assumptions (Linear Regression)

Approximately Normal Parameter Estimates

- An “approximately normal” **parameter estimate** (*see “estimate” column in previous examples*) does **NOT** mean your **data** have to be normally distributed!

- Either a normal **outcome** or a **large enough sample size** will do the trick. Watch your tails!

- This assumption is what allows you to use the default **t-statistic** and **p-value** estimates.



- What does “**large enough**” mean? Smaller than you’d think, check this out!

Assumptions (Linear Regression)

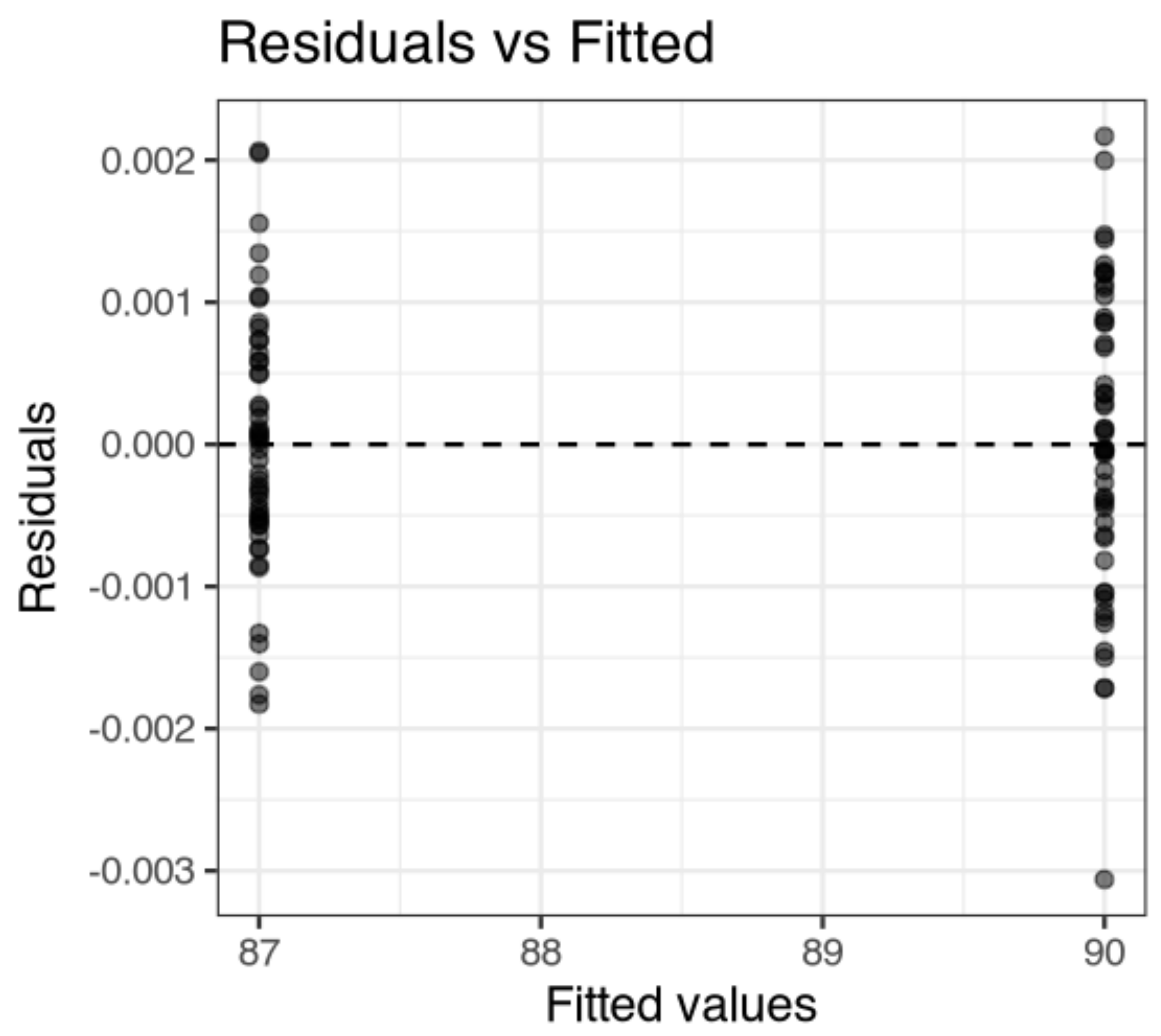
Independent Observations

- Unfortunately, you can't check for **independence** of observations.
- Luckily, **good experimental design** can save you here!
- If there is some sort of **clustering** going on, you'll need to either pivot to **another model** or use a **different standard error**. *(more on this later)*
- One example might be if you have large groups from three **species of birds** in your experiment, or if you collect **multiple measurements** on each bird.

Assumptions (Linear Regression)

Equal Variance Among Groups

- We check this assumption with the **residuals.**
- You're looking for **no particular pattern** with variation **centered around zero** of approximately the **same magnitude** as you **cross the x-axis.**
- If you don't believe this assumption holds, go for **robust standard errors** rather than model defaults.



Assumptions (Linear Regression)

Moving Beyond Associations

- If you want to **borrow information** across groups, you assume your model structure is correct, meaning you have the **right inputs** in the **right form**.
  - This assumption is **stronger** than the original three.
- If you want to learn about **individual observations** within a group, you have to assume normal distribution of errors among members of that group.
  - This is the **strongest assumption** yet! Luckily, these tasks are less common.

You might know something about your data structure that can point you to a different method of analysis!

Extending Past Linear Regression

Generalized Linear Models

- Instead of Y as **body temperature**, we might want to tweak our **outcome**.
- For example, we might be interested in whether a bird **displayed a specific thermoregulation behavior** within a given time frame.
- We can try a **generalized linear model (GLM)**, which **extends** linear regression!
- The **input (linear predictor)** is set up in the same way, but we model the **outcome** differently depending on **data type** and further **assumptions**.
- Try **logistic regression** when your outcome is **binary**, and **Poisson regression** is a good first try if your outcome is a **count**.

Logistic Regression

$\log(\frac{p}{1 - p}) = m_1X + m_2Z + b$, where $P(Y = 1) = p$ (think `glm()` in R)

- Let's go back to the **water** and **weight** inputs and record whether the bird **displays a behavior** by the end of the experiment.

Logistic Regression				
term	estimate	standard error	z statistic	p value
(Intercept)	1.43	1.2847	1.12	2.6414×10^{-1}
x	−2.47	5.0552×10^{-1}	−4.89	9.9989×10^{-7}
z	0.04	1.4129×10^{-1}	0.31	7.5599×10^{-1}

- We'd expect a typical **0 ounce** bird with **no access to water** to have **exp(1.43) odds** of displaying the behavior. (This is where **centering** might make sense!)

- Among birds of the same weight, we expect those with **access to water** to have **odds** of displaying the behavior **exp(-2.47)** times higher than their **no-water counterparts**.

- We might fiddle with the arithmetic to express it as **exp(2.47)** times lower.

Poisson Regression

$\log[E(Y)] = m_1X + m_2Z + b$ **(think `glm()` in R)**

Poisson Regression				
term	**estimate**	**standard error**	**z statistic**	**p value**
(Intercept)	4.96	2.6717×10^{-2}	185.72	0.0000
x	−3.01	2.2190×10^{-2}	−135.43	0.0000
z	0.20	2.8240×10^{-3}	72.20	0.0000

- Let's go back to the **water** and **weight** inputs and record a **count of behavior displays** by the end of the experiment.
- We'd expect a typical **0-ounce** bird with **no access to water** to have a **behavior rate** of **exp(4.96)** over the course of the experiment.
- Among birds of the same weight, we expect those with **access to water** to have a **behavior rate** of **exp(-3.01)** times higher than the **no-water birds**.
- We might fiddle with the arithmetic to express it as **exp(3.01)** times lower.

Assumptions

(Logistic and Poisson Regression)

- You'll again want to check for **independent** observations and think carefully about the **form of your linear predictor**.
- For **logistic regression**, getting your linear predictor right gets your variance right for free!
- For **Poisson regression**, your outcome **mean** and **variance** should be about the same. If they're not, models like **quasipoisson** or **negative binomial** might do the trick.
- Poisson regression requires you to count events over the same **period** or **place**, unless you use an **offset** term.

Extending Past Linear Regression

Nonlinear Least Squares (think `nls()` in R)

- What if we want to model a **biological mechanism** that has a relatively **well-understood data structure**?
- For example, certain kinds of **assays** and **ecological systems** often generate data that follow **consistent non-linear patterns.**
- The **non-linear least squares** model may be a good fit here, as its parameters often have a natural **interpretation.**
- For example, one of the parameters might tell you about a **population limit** or a **growth rate**.

Extending Past Linear Regression

Mixed Models and GEE

- Going back to the issue of **independence violations**, there are many models that can incorporate **dependence** in your data structure!

- Models such as **linear mixed models (LMM and GLMM)** or **generalized estimating equations (GEE)** incorporate a **cluster variable** into their setup.

- **Mixed models** are helpful for learning about **individual specific effects**, while **GEE**s can help you learn about a **population level effect**.

- Watch out for **missing data** here, you'll want to think a little on **why** different data points might be missing before choosing a method. GEE has a **stronger assumption** than mixed models.

Wait, I'm ~close~ to one of these setups but just a bit off!

If You Think	You Could Try	With Tool	And Advice
One particular variable in a linear model setup is not so linear.	Splines	library(splines)	Only use this for a variable you think belongs in your model but doesn't require a specific interpretation.

If You Think	You Could Try	With Tool	And Advice
One particular variable in a linear model setup is not so linear.	Splines	library(splines)	Only use this for a variable you think belongs in your model but doesn't require a specific interpretation.
One variable might be influencing the effect of another.	Interaction Terms	This usually works in a standard regression with the * setup.	Sometimes, combinations of coefficients are more helpful to interpret here.

If You Think	You Could Try	With Tool	And Advice
One particular variable in a linear model setup is not so linear.	Splines	library(splines)	Only use this for a variable you think belongs in your model but doesn't require a specific interpretation.
One variable might be influencing the effect of another.	Interaction Terms	This usually works in a standard regression with the * setup.	Sometimes, combinations of coefficients are more helpful to interpret here.
One particular variable in a linear model setup is not so linear.	A Term Transformation	I usually do this in a preprocessing step, often with dplyr.	This situation is where contrasts might be helpful.

If You Think	You Could Try	With Tool	And Advice
One particular variable in a linear model setup is not so linear.	Splines	library(splines)	Only use this for a variable you think belongs in your model but doesn't require a specific interpretation.
One variable might be influencing the effect of another.	Interaction Terms	This usually works in a standard regression with the * setup.	Sometimes, combinations of coefficients are more helpful to interpret here.
One particular variable in a linear model setup is not so linear.	A Term Transformation	I usually do this in a preprocessing step, often with dplyr.	This situation is where contrasts might be helpful.
You've got too many predictor variables and not all of them really belong.	Penalty Models (Ridge, Lasso, and Elastic Net are great examples!)	library(glmnet)	These are best used in a prediction setup; variable selection in this way comes at a cost of reliable inference.

If You Think	You Could Try	With Tool	And Advice
One particular variable in a linear model setup is not so linear.	Splines	library(splines)	Only use this for a variable you think belongs in your model but doesn’t require a specific interpretation.
One variable might be influencing the effect of another.	Interaction Terms	This usually works in a standard regression with the * setup.	Sometimes, combinations of coefficients are more helpful to interpret here.
One particular variable in a linear model setup is not so linear.	A Term Transformation	I usually do this in a preprocessing step, often with dplyr.	This situation is where contrasts might be helpful.
You’ve got too many predictor variables and not all of them really belong.	Penalty Models (Ridge, Lasso, and Elastic Net are great examples!)	library(glmnet)	These are best used in a prediction setup; variable selection in this way comes at a cost of reliable inference.
One of the assumptions about variance in your data is violated.	A Different Variance Estimator	library(sandwich)	Robust standard errors can fix different violations in different models, and there are different kinds.

If You Think	You Could Try	With Tool	And Advice
One particular variable in a linear model setup is not so linear.	Splines	library(splines)	Only use this for a variable you think belongs in your model but doesn't require a specific interpretation.
One variable might be influencing the effect of another.	Interaction Terms	This usually works in a standard regression with the * setup.	Sometimes, combinations of coefficients are more helpful to interpret here.
One particular variable in a linear model setup is not so linear.	A Term Transformation	I usually do this in a preprocessing step, often with dplyr.	This situation is where contrasts might be helpful.
You've got too many predictor variables and not all of them really belong.	Penalty Models (Ridge, Lasso, and Elastic Net are great examples!)	library(glmnet)	These are best used in a prediction setup; variable selection in this way comes at a cost of reliable inference.
One of the assumptions about variance in your data is violated.	A Different Variance Estimator	library(sandwich)	Robust standard errors can fix different violations in different models, and there are different kinds.
You have a large representative sample but distributional assumptions aren't quite right	Bootstrapping	library(boot)	This is often computationally intensive! If you have a complex sampling procedure, it may be easier write without a package.

Missingness in Data

- Missing completely at random **(MCAR)** is the rarest form of missingness, where each observation has an **equal chance of being missing**.
- Missing at random **(MAR)** means the reason for missingness is **captured by other variables**, like if you know male birds more often have missing data but you know the bird is male!
- Missing not at random **(MNAR)** occurs if we **can't justify** MCAR or MAR.
- Check the **default behavior** of your code for handling missing data!
- Different methods **allow different levels of missingness**. For example, mixed models work under MAR, while GEE requires MCAR.
- Options like **imputation** or **complete case analysis** might combat missingness.

Sensitivity Analysis

How robust are your conclusions?

- Sensitivity analysis allows you to explore **whether your conclusions hold** under slight tweaks in your assumptions.
- For example, fitting a model **with and without outliers** and seeing **matching conclusions** is a good sign that the outliers aren't driving the results.
- You could try **simplifying your model** and exploring how (or if) your conclusions change and to what extent.
- The **goal of your analysis** (predicting an outcome, drawing inference about an effect, etc) drives the **possible sensitivity analyses** you could conduct.

You’ve found a new tool in R, now what?

Questions to Ask Yourself

- Where did I find the **package** (and its **documentation**)?
- Is **source code** available and well commented?
- Are there **tutorials** out there?
- What happens when I change specific **arguments**?
- Are there any **companion packages** out there that I already know?
- What **subject area** do the authors work in?

Statistical Example

library(lme4)

- The package is on **CRAN** with source code available through **GitHub**.
- The source code has **comments** peppered throughout that label parts of each function.
- The **readme** file points you towards places where you can get **help**!
- The documentation on **CRAN** is well-developed.
- Douglas Bates is one of the lead **maintainers**, and he **publishes extensively** about linear mixed models.

Reporting Example

library(gt)

- The package lives on **CRAN** and has an entire **website manual** devoted to it.
- There are many **examples** available with its **built-in data**.
- The folks at Posit (formerly RStudio) were involved with its **development**, so it plays nicely with tidyverse functions!
- Many **tutorials** exist, and I particularly like Tom Mock's **cookbooks**.

Bonus Thoughts About Programming

(While I have you here, I might as well!) 😊

- **Comment your code** like you'll be returning to the project six years later and forgot everything about it!
- **Version control** can save you from yourself.
- Depending on the task, choose wisely between **scripts** and **notebooks**.
- **Functions** are your friends, especially if you often **repeat workflows**!
- Write **what works** first, and you can always **clean it up** later.
- Spend the time making your **figures and tables** easy to parse!

References

- Sensitivity analysis
- Missing data
- Penalized regression
- Mixed models and GEE
- Nonlinear regression
- Nonlinear regression (and more!)
- Varieties of variance (😊)
- Lots of applied biostatistical goodies

CAUTION ⚠️⚠️⚠️
This was a FLYOVER talk! I glossed over MANY details and caveats that I encourage you to read more about if any topic from today potentially aligns with your research.