Linear Mixed Effects Modeling

Jereme W Gaeta, PhD (CDFW and IEP)

3/23/2022

1 Recommended Reference Material

I recommend several books and book chapters to dig deeper into the underlying theory of mixed effects modeling and its application to ecological data. The most thorough book, in my opinion, is Gelman and Hill (2006); however, this book is not written with ecologists in mind, but is written for sociologists analyzing human data. Gelman and Hill (2006) do an amazing job of digging into the deep nuance of this analytical approach and provide a lot of advice on data preparation and method troubleshooting. One of my top go-to resources for advanced regression analyses with R code is Zuur et al. (2009). I highly recommend this book as an invaluable resource for anyone that regularly analyzes data as a part of their work duties.

We will only be covering two-level models in this micro-training. However, you may encounter a situation in which you need to use a three-level model. As a starting point, I suggest you see the model notation and analysis in Gaeta et al. (2011) in which my colleagues and I analyzed repeated measures of fish growth observations (first-level) nested within individuals (second-level), which, in turn, were nested in various lakes (third-level).

2 Background

Mixed effects modeling (AKA hierarcical or multilevel modeling; not to be confused with mixing models) is a form of statistics that allows users to account for the structure of our their data or sampling regime without violating one of the primary assumptions of regression: *independence* (i.e., any value of y_i given a value of x_i is NOT influenced by any other value of x_i; more on this in the next section). Many of us in the Interagency Ecological Program, for example, analyze monitoring data, most of which are not suitable for simple linear regression (e.g., equation 4) because they are time-series data, are repeated sampling events at fixed stations or in fixed regions, and/or contain multiple samples from a single sampling event, all of which violate the assumption of independence.

Take a situation in which we have a set of fixed sampling stations and we collect multiple samples from each station (Figure 1). Our samples are, therefore, "nested" within stations. This is a clear violation of the assumption of independence as environmental conditions at a given station likely make all data collected at that station more similar (or not independent) than data collected at different stations. One possible solution is to simply include a covariate in your model for each station. While this may work for a small project, many of our long-term monitoring programs have dozens of stations. This would greatly increase the number of parameters your model is fitting (decreasing your degrees of freedom drastically) and making your model output unintelligible. A mixed effects framework, however, allows us to "tell" the model, while we know we need to account for the grouping of station, we do not want station to be a primary parameter in our model. Let's get into it!

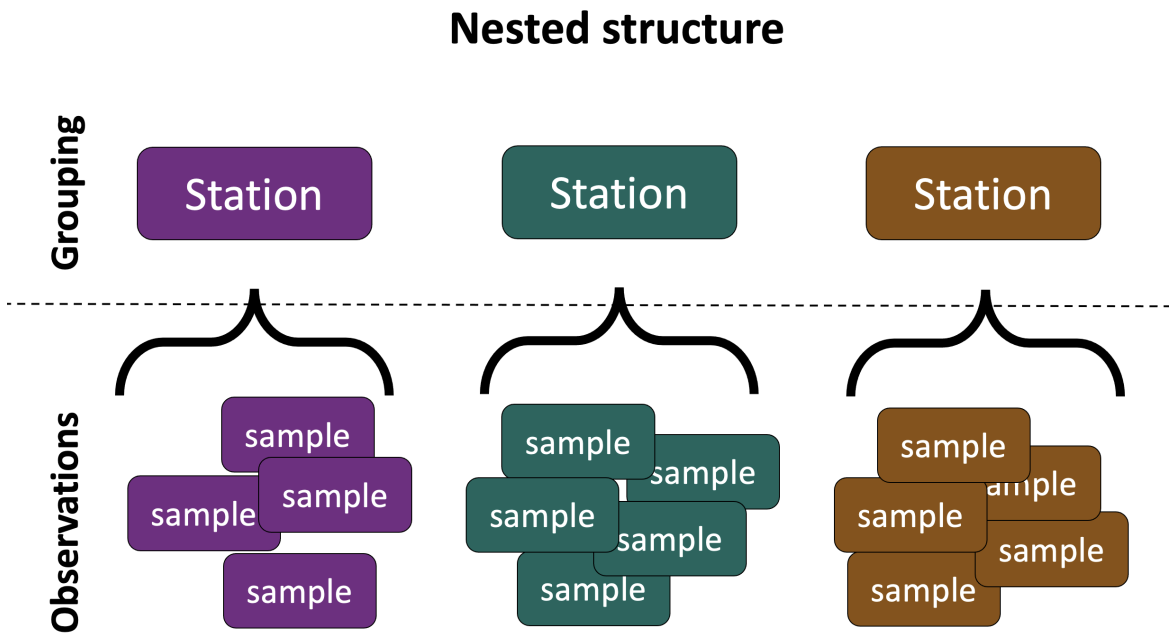

Figure 1: A conceptual model illustrating data nesting with samples at the observation-level nested within stations at the group-level.

2.1 The Assumptions (i.e., requirements) of Regression

Given modern computing power, open source software (e.g., R statistical software), and a plethora of peer-reviewed and non-peer-reviewed resources available on the Googles, we ecologists all too often throw data at an analysis without a deep understanding of the underlying statistical theory. While this micro-training session will not go too deep into the underpinnings of regression, I do want to briefly remind you of the four main assumptions (i.e., requirements) of regression: (1) normality, (2) homogeneity, (3) fixed x, and (4) independence (refer to Zuur et al. (2009) sections 2.3 for a more thorough discussion).

Disclaimer: It is important to remember that regression is most suited for highly controlled data such as those derived from a physics or engineering laboratory setting. We ecologists will *always* have more variability and uncertainty associated with our data. Therefore, while we should strive to always meet the assumptions of regression, it is understood by quantitative ecologists and peer-reviewers that ecologists will have to *flex* these assumptions.

2.1.1 Normality:

Regression users often believe this assumption means your response variable y must be normally distributed. While that is *part* of the assumption of normality, that is not the core its meaning. This assumption means that the replicates or sub-samples values y would be normally distributed at a given value of x, and this would hold true for **all** values of x. However, we ecologists rarely have numerous sub-samples at each value of x to test for normality. Therefore, our best option is to build a regression and evaluate the normality of the pooled residuals. This is why you should always include a histogram of model residuals when presenting model results.

2.1.2 Homogeneity:

This assumption is tightly linked with the assumption of normality. Specifically, this is the assumption that the distribution of your residuals hold constant across x. This is why you should include a plot of model residuals (y-axis) against observed values of your predictor variable (x-axis). The plot should look like a rectangle with the density of residuals increasing as you approach zero along the y-axis (see the green box below for a deeper discussion). A clear violation of this assumption would be your residuals changing shape as you move left to right (imagine a traffic cone on its side or visible curve).

2.1.3 Fixed X:

This assumption is really tricky for ecologists. Remember, regression is most suited for highly controlled data such as those derived from a physics or engineering laboratory setting. The assumption of fixed X has two components: (1) that you *a priori* know each value of x (e.g., you design an aquarium experiment and *a priori* decide the water temperatures for the experiment) and (2) you know the exact value of x (i.e., no measurement error). Neither of these components are ever *truly* met in field ecology. So, generally speaking, we ecologists slightly bend this assumption. You should really only be concerned if the measurement error around x is very large relative to the range of x in your study (e.g., if you are aging Delta Smelt that are usually $\leqslant 3$ years old and your measurement error is ± 2 years).

2.1.4 Independence:

The final assumption of regression is that your observation of y at a given value of x cannot be influenced by or related to another observation of y at a different value of x. This assumption is almost certainly violated by monitoring data. However, mixed effects modeling provides a framework to analyze these data without violating this assumption of regression.

2.2 Distinguishing Between Fixed and Random Effects

As with any analytical framework, mixed effects modeling has its own set of unique terminology; unfortunately, this terminology often differs among disciplines (e.g., sociologist, economists, and ecologist tend to use completely different terms with the same meaning; see Gelman and Hill (2006) pg 245-246, especially the footnotes, for a detailed discussion). I will, therefore, use one set of terminology with my own personal definitions that have served me well so far in my career, but please see Gelman and Hill (2006) and Zuur et al. (2009) to learn from the experts.

Simple linear regression has one response variable (your y-variable) and one or more predictor varaibles or covariables (your x-variable(s)). Mixed effects regression has one response variable (your y-variable), but has two kinds of "predictor" variables called "fixed effects" and "random effects":

- **Fixed Effects**:
  - Colloquially: this is the "stuff" you care about
  - Overly broad definition: the predictor variable(s) of interest; the variable(s) you mention in your ecological goal or question
- **Random Effects**:
  - Colloquially: this is the "stuff" you don't necessarily care about, but for which you must account
  - Overly broad definition: the aspects of your data structure or sampling regime that result in your data violating the assumption of independence

2.3 Random Effect Structures

Random effect structures come in three forms:

1) **Random intercept** (AKA varying intercept) models: here, you allow the model to derive a unique intercept for each factor in your grouping
   i. the model will derive a "grand mean" intercept coefficient and a single slope coefficient
   ii. Figure 2a
   iii. Equation 5
2) **Random slope** (AKA varying slope) models: here, you allow the model to derive a unique slope for each factor in your grouping
   i. the model will derive a "grand mean" slope coefficient and a single intercept coefficient
   ii. Figure 2c
   iii. Equation 7
3) **Random intercept and slope** (AKA varying intercept and slope) models: here, you allow the model to derive a unique intercept *and* a unique slope for each factor in your grouping
   i. the model will derive a "grand mean" intercept coefficient and a "grand mean" slope coefficient
   ii. Figure 2b
   iii. Equation 6

Generally speaking, most applications start with, if not stick to, a random intercept model. This allows users to account for the structure of their data and does not come with the headaches of the other model structures. For example, Figure 2c highlights a common shortcoming of random slope models: unusual and, often times, unrealistic patterns can occur when the y-intercept (i.e., where the model crosses $x = 0$) is outside of the data range. An exception to this rule is the rare condition in which you expect all groupings to have the same intercept. This may be suitable, for example, if you are modeling animal size over time ***if*** you are assuming that all individuals hatch or are born at the same size. More often, however, you can get around this challenge simply by centering your predictor variable data:

$$\text{centered}(x_i) = x_i - \mu_x \tag{1}$$

where μ_x is the mean value of x. Centering will shift your data such that the mean value of $x = 0$. The random slopes, therefore, will create an "X" pattern in your random effect regressions where the center of the "X" is at $x = 0$.

While random slope and intercept models logically seem like the most flexible and accurate option (see Figure 2b), they have their own set of challenges. Most obvious is that you are asking the model to double the number of parameters for which you are fitting, drastically reducing your degrees of freedom. Consequently, this random effect structure is really only suitable when your ratio of observations:groups is high. Even so, the model estimates the random slopes and intercepts simultaneously, resulting in an additional parameter: the correlation between random slopes and random intercepts (see equation 7; $\rho\sigma^2_\alpha\sigma^2_\beta$). Models often fail to converge or throw warnings if $\rho\sigma^2_\alpha\sigma^2_\beta$ approaches 1 or -1. A common solution to this problem is to z-score your response and predictor variable(s):

$$\text{z-scored}(x_i) = \frac{x_i - \mu_x}{\sigma_x} \tag{2}$$

where μ_x is the mean value of x and σ_x is the standard deviation of x. This is also referred to as "rescaling" your variables as z-scored data are 1) centered around zero and 2) scaled relative to their standard deviation. Therefore, if your variables are normally distributed (or approaching normally distributed), your observations will range from $\sim\geqslant -4$ to $\sim\leqslant 4$.

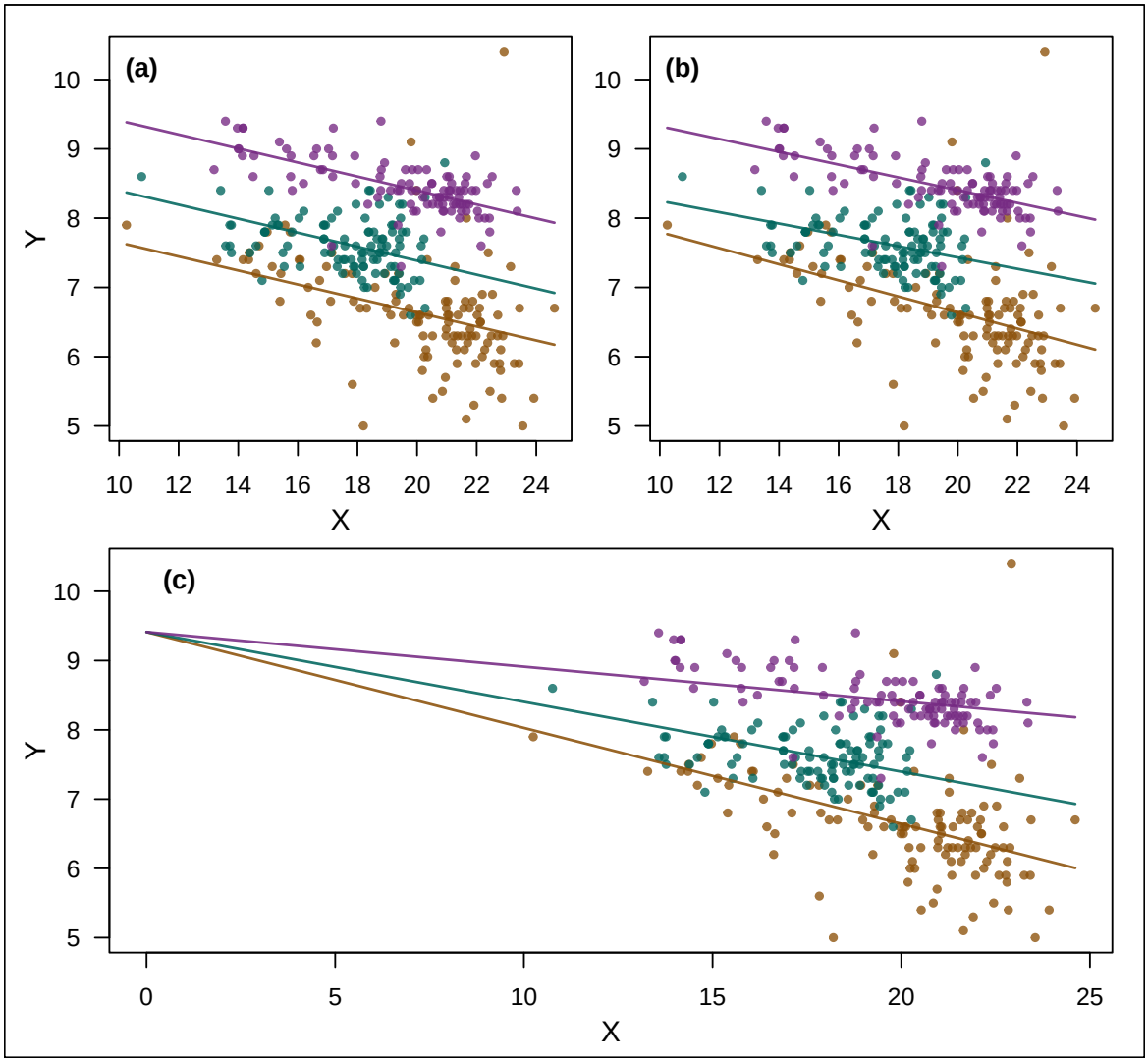

Figure 2: Visualization of random effects given (a) a random intercept model, (b) a random intercept and slope model, and (c) a random slope model. Points are color-coded according to random effect grouping (i.e., sampling station).

3 Model Notation

Note: feel free to skip this section on your first read and circle back after you explore model code and model outputs.

Personally, I believe you cannot fully understand a model without understanding the model notation. Do not run away yet, I do not mean you need to truly understand the math to the point of doing the analysis by hand on graph paper, I just mean you should be able to understand the underlying equations.

3.1 Simple Linear Regression

Take that old equation from grade school:

$$y = mx + b \tag{3}$$

where y is your response variable, x is your predictor variable, m is the slope (i.e., "rise-over-run"), and b is the intercept of a simple linear regression (are you having flashbacks of drawing a line on graph paper yet?). We can re-write equation 3 using the following standard mathematical notation:

$$\begin{gathered} y = \alpha + \beta x + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{gathered} \tag{4}$$

where y and x are your predictor and response variables, respectively, α and β are model (i.e., equation 4) coefficients (the same as b and m in equation 3, respectively), and ε is the remaining uncertainty not accounted for by your predictor variable x. Specifically, ε is the residual variance (i.e, the difference between the observations and the regression line created by the coefficients). The model uncertainty (ε) is described by the second line of the equation: the residual variance is normally distributed (N) with a mean of 0 and a variance of σ^2 (see the green box below for a deeper discussion).

3.2 Mixed Effects Linear Regression

Simple linear regression only allows us to account for one level of uncertainty: the observation level. This is annotated as $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ in equation 4 above. This essentially means the variability in our observations are either explained by our variable(s) of interest (i.e., predictor(s)) or remains unexplained and considered residual uncertainty.

A two-level mixed effects regression, on the other hand, allows us to account for **two** levels of uncertainty: 1) the variability in our data associated with our random effects groupings (i.e., data structure or sampling design) and, then, after accounting for this uncertainty, 2) the observation level. This essentially means the variability in our observations are explained by our variable(s) of interest (i.e., predictor(s)) or explained by our random effects groupings (i.e., data structure or sampling design); then the remaining unexplained variability is considered residual uncertainty.

3.2.1 Random Intercept

A random intercept mixed effects linear regression can be written using the following standard mathematical notation:

$$\begin{gathered} y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_{j[i]} \\ \alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha), \text{ for } j = 1, ..., J \\ \varepsilon_{j[i]} \sim N(0, \sigma^2_\varepsilon) \end{gathered} \tag{5}$$

where y_i is the i^{th} response variable observation given x_i, $\alpha_{j[i]}$ is the random intercept of the j^{th} group to which the i^{th} observation belongs, β is the coefficient (i.e., slope) of the response variable x, and ε is the remaining uncertainty not represented by the the predictor variable x and the random intercepts. The random intercepts, α_j, are normally distributed around the grand mean model intercept, μ_α, with a variance of σ^2_α. The remaining model uncertainty, ε, is normally distributed with a mean of zero and a variance of σ^2_ε.

3.2.2 Random Slope

A random slope mixed effects linear regression can be written using the following standard mathematical notation:

$$\begin{gathered} y_i = \alpha + \beta_{j[i]} x_i + \varepsilon_{j[i]} \\ \beta_j \sim N(\mu_\beta, \sigma_\beta^2), \text{ for } j = 1, ..., J \\ \varepsilon_{j[i]} \sim N(0, \sigma_\varepsilon^2) \end{gathered} \tag{6}$$

where y_i is the i^{th} response variable observation given x_i, α is the model intercept, $\beta_{j[i]}$ is the random slope of the j^{th} group to which the i^{th} observation belongs, and ε is the remaining uncertainty not represented by the the predictor variable x and the random slopes. The random slopes, β_j, are normally distributed around the grand mean model slope, μ_β, with a variance of σ_β^2. The remaining model uncertainty, ε, is normally distributed with a mean of zero and a variance of σ_ε^2.

3.2.3 Random Intercept and Random Slope

A random intercept and slope mixed effects linear regression can be written using the following standard mathematical notation:

$$\begin{gathered} y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \varepsilon_{j[i]} \\ \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha^2\sigma_\beta^2 \\ \rho\sigma_\alpha^2\sigma_\beta^2 & \sigma_\beta^2 \end{pmatrix} \right), \text{ for } j = 1, ..., J \\ \varepsilon_{j[i]} \sim N(0, \sigma_\varepsilon^2) \end{gathered} \tag{7}$$

where y_i is the i^{th} response variable observation given x_i, $\alpha_{j[i]}$ and $\beta_{j[i]}$ are the random intercept and slope, respectively, of the j^{th} group to which the i^{th} observation belongs. α_j and β_j are normally distributed around the grand mean model intercept, μ_α, and grand mean model slope, μ_β, with variances of σ_α^2 and σ_β^2, respectively, and a correlation of $\rho\sigma_\alpha^2\sigma_\beta^2$. The remaining model uncertainty, ε, is normally distributed with a mean of zero and a variance of σ_ε^2.

The $\rho\sigma_\alpha^2\sigma_\beta^2$ defines the correlation between the α_j and β_j estimates. A $\rho\sigma_\alpha^2\sigma_\beta^2$ approaching 1 or -1 should be viewed with caution. This often indicates your model is struggling to fit given your data, unless you have an ecological reason to believe this pattern. Nevertheless, I advise centering or z-scoring your data to determine whether this will push $\rho\sigma_\alpha^2\sigma_\beta^2$ toward zero.

3.2.4 A Final Note on Model Notation

The model notation I have provided above is just one way to annotate the models. Please see Gelman and Hill (2006) section 12.1 (titled "Notation") and section 12.1 (titled "Five ways to write the same model") for a much more thorough discussion. Also, refer to Gaeta et al. (2011) for three-level model notation, although the notation style differs from the style above shown for equations 4 through 7.

A Deeper Look at ε and the Assumption of Homogeneity

While ε in equation 4 may seem a little confusing, this is a simple, yet critical aspect of regression ecologists should understand. Specifically, this is related to the second assumption of regression: homogeneity. This assumption means the uncertainty, or noise, around the mean (i.e., where the mean is the linear regression model or prediction) must be constant across your predictor variable x. Consider the regression prediction (i.e., $\hat{y}$) at a given value of x, the observations of y must be normally distributed around $\hat{y}$ such that 68.3% of observations at that value of x must be $\hat{y} \pm 1 * \sigma^2$, 95.5% of observations at that value of x must be $\hat{y} \pm 2 * \sigma^2$, and 99.7% of observations at that value of x must be $\hat{y} \pm 3 * \sigma^2$. Because a linear regression model only produces one residual variance term (i.e., σ^2; the second part of equation 4), our observations must vary consistently around $\hat{y}$ for all values of x; that is, our data must have homogeneous variance (see Figure 3 for an illustration of this concept; refer to Zuur et al. (2009) sections 2.2 and 2.3.3 for a more thorough discussion).

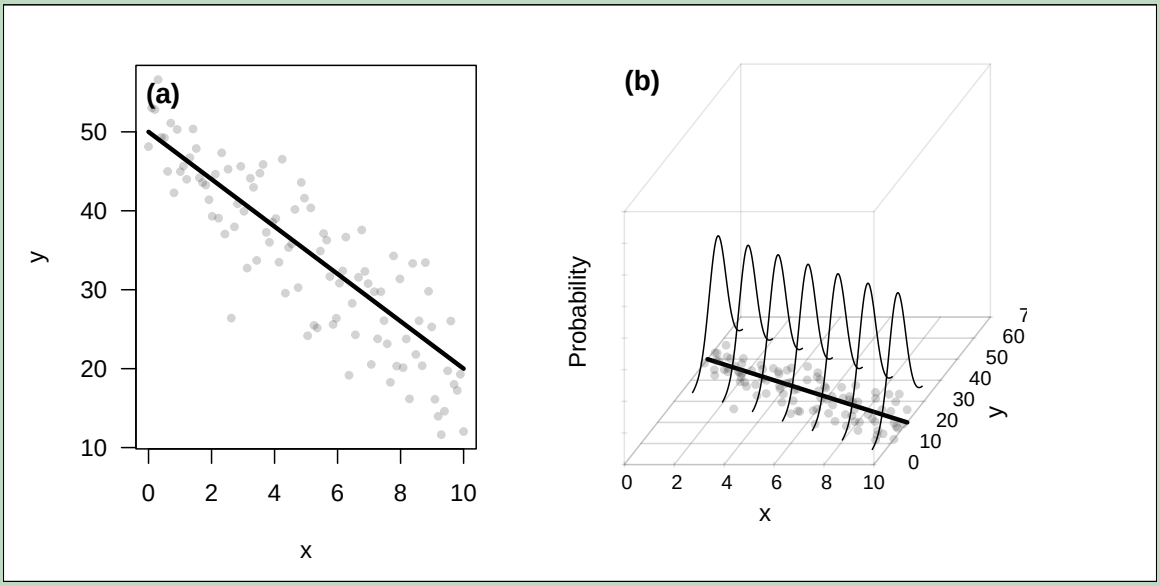

Figure 3: (a) A hypothetical dataset (points) shown with a linear regression model (black line). (b) The same hypothetical dataset (points) and linear regression model (black line), but shown with the several normal distributions across the range of x, each of which represents ε: a normal distribution around the regression line (i.e., a mean of zero) with a variance of σ^2.

4 Analysis

I recommend always taking a question-driven approach to statistical and mechanistic modeling. With a good question comes brainstorming about the best data to address your question and a plan for the ideal statistical or mechanistic modeling approach. Regardless of the question, statistical method, or mechanistic modeling, you should always follow the same overarching framework presented in Figure 4: explore your data; apply a model; evaluate, validate, and optimize your model; and, when validation and optimization are complete, focus on a clear, concise interpretation and effective presentation. While every step in this general framework may not be relevant for every project or every analysis, returning to this framework will ensure you produce defensible and publishable science.

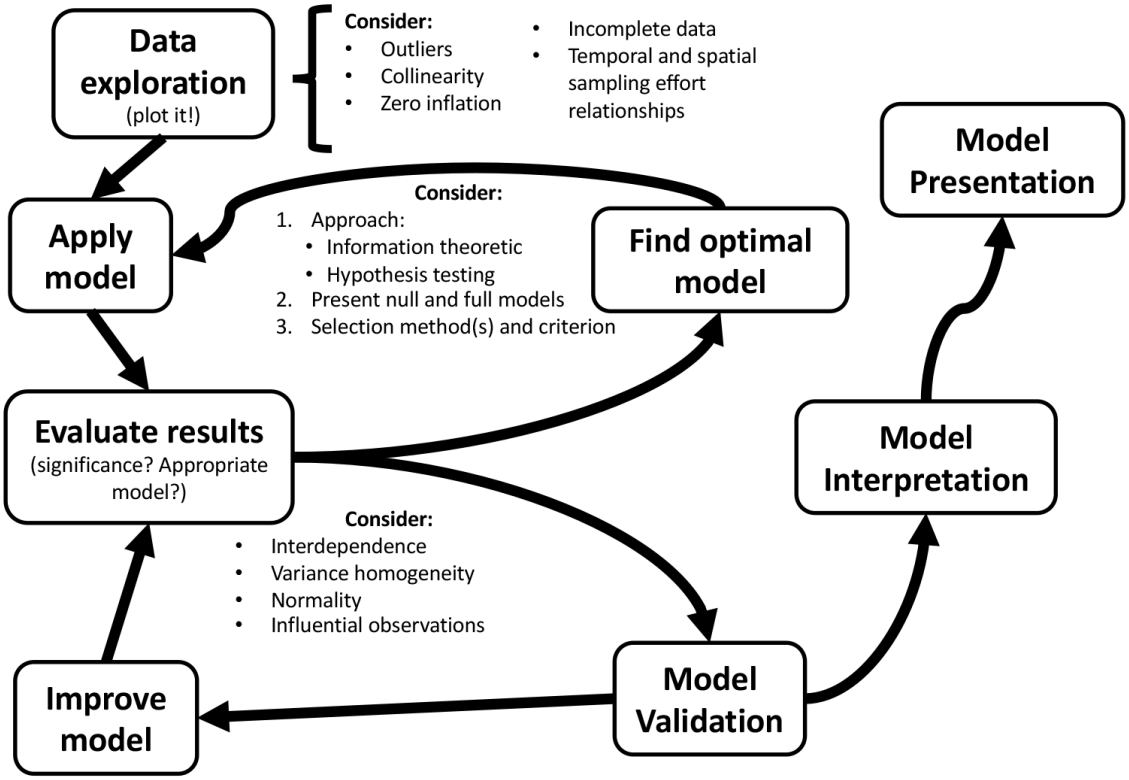

Figure 4: A general modeling framework I recommend following (at least loosely) for all analyses.

While, unfortunately, we will not have time during the workshop to walk through each step of the modeling framework presented in Figure 4, I will return to the figure and bring your attention to these concepts throughout this micro-training.

4.1 Load the R Packages, our Dataset, and Custom Functions

Prior to any analyses in R, you need to make sure you have the necessary packages loaded into the environment. We will be using the ***lme4*** package to develop mixed effects models and the ***effects*** package to visualize model uncertainty in this micro-training session.

```
#~ Load packages
library(lme4)
library(effects)

#~ A custom function to add a label to the same location on every panel
plot_label = function(lab="(a)", x_prop=0.08, y_prop=0.92,
                      font_type=2, fcex=1.15, usr=par('usr')){
  x_val = usr[1]+(usr[2]-usr[1])*x_prop
  y_val = usr[3]+(usr[4]-usr[3])*y_prop
  text(x = x_val, y = y_val, labels = lab, font = font_type, cex=fcex)
}

#~ Load the dataset
dat3 = read.csv(file = "USGS_SFBWQ_3stations.csv", header=T)
```

4.2 Case Study Data

I selected a publicly available dataset generated by the USGS's San Francisco Bay Project (SFBP), which is available via the United States Geological Survey (Schraga et al. 2020). The SFBP collects discrete water quality data throughout the San Francisco Estuary, including discrete dissolved oxygen and temperature observations. For this micro-training, I reduced the data to just surface data (2m depth) collected from June through November at three stations spanning the spatial extent of the sampling program (Figure 5). We will evaluate dissolved oxygen across water temperature.

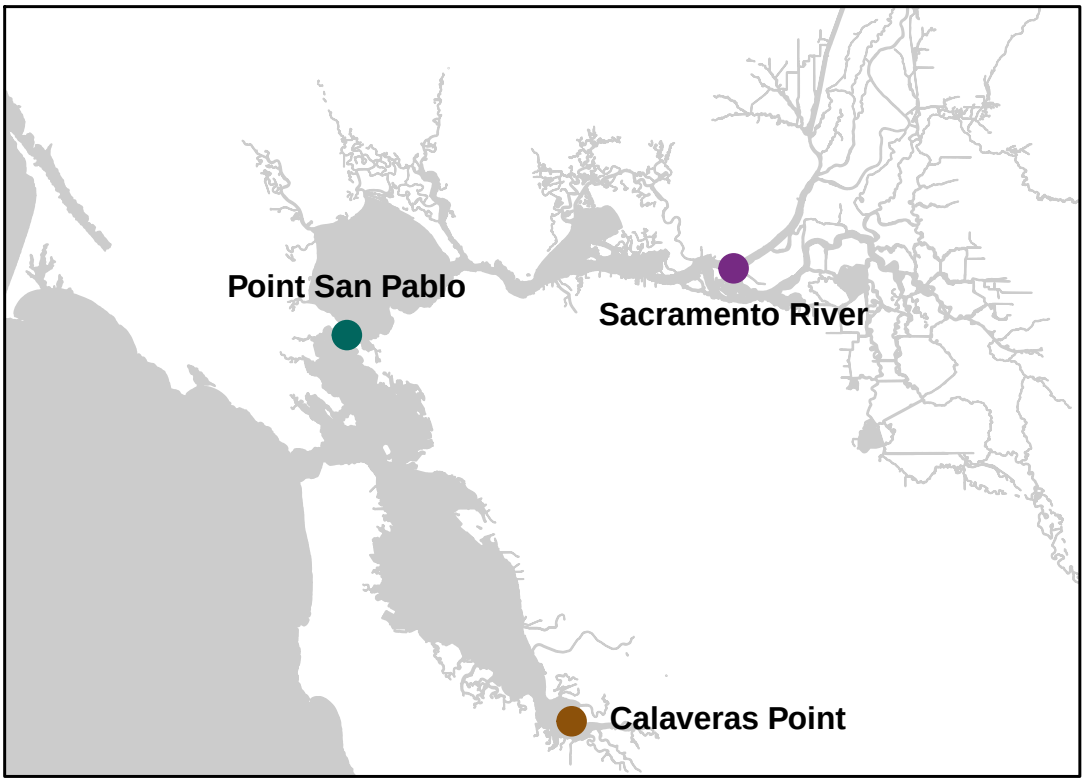

Figure 5: A map of the USGS's San Francisco Bay Project sampling stations included in this micro-training.

Let's take a look at the dataset:

```
head(dat3)
```

```
##      station_name Station_Number     date Year Discrete_Oxygen Temperature
## 1 Calaveras Point             36  11/9/77 1977             7.2        14.6
## 2 Point San Pablo             15 11/10/77 1977             7.6        14.6
## 3 Point San Pablo             15  7/13/78 1978             7.6        17.2
## 4 Calaveras Point             36  9/19/78 1978             5.0        18.2
## 5 Point San Pablo             15  9/20/78 1978             7.3        18.2
## 6 Point San Pablo             15  11/9/78 1978             7.8        14.9
```

Abbreviated metadata for the variables are as follows:

- *station_name*: station location
- *Station_Number*: station code
- *date*
- *Year*: calendar Year
- *Discrete_Oxygen*: concentration ($\mu g \cdot l^{-1}$) of oxygen dissolved in water analyzed in the laboratory using the Winkler method
- *Temperature*: water surface temperature (2m depth) in degrees Celsius

4.3 Study Goal

The study goal we will address in this micro-training is to ***"develop an Estuary-wide relationship between dissolved oxygen and temperature after accounting for sampling station"***. The overarching goal, therefore, of our analysis is to build a linear model to predict disolved oxygen (hereafter, "DO") across water temperature (hereafter, "temperature"). As per our modeling framework (Figure 4), let's start by exploring this relationship.

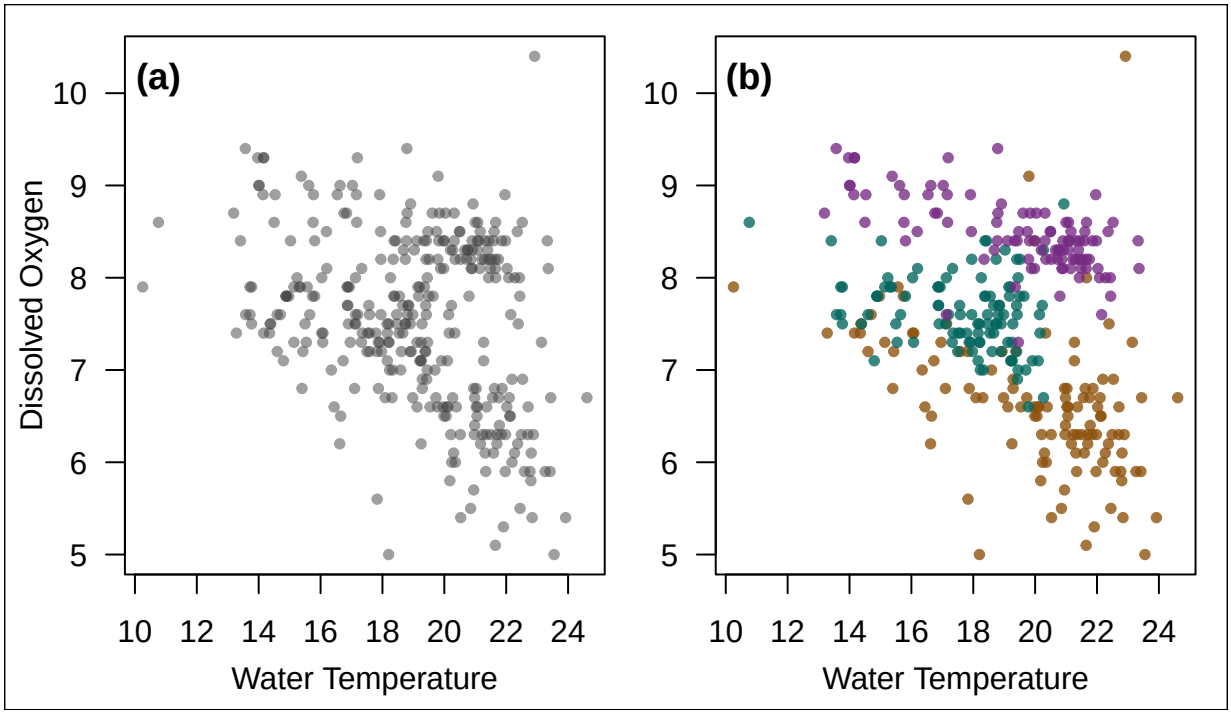

Figure 6: Disolved oxygen across water temperature in the San Francisco Estuary. Shown as (a) aggregated data and (b) color coded by sampling location. Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36.

We see a negative relationship between DO and temperature (Figure 6a), suggesting a linear regression is likely a suitable analytical method. However, the strength and variability of the relationship appears to vary by station (Figure 6b), indicating we should take sampling station into account in our analysis.

If we were analyzing these data for publication, I would ask data generators whether the DO observation >10 $\mu g \cdot l^{-1}$ and the temperature observations <12 °C are potential data entry errors. However, a few potential outliers are unlikely to inhibit our ability to model these data for this micro-training. Similarly, as per Figure 4, an assessment of the data suggests our data would be *slightly* more linear if $\log e$-transformed, particularly if we were to use the full dataset. However, we will forgo this transformation for ease of figure and result interpretation in this micro-training.

4.4 Building a Simple Linear Regression

Before we journey into the land of mixed effects, let's build our model in a simple linear framework:

```
mod = lm(Discrete_Oxygen ~ Temperature, data = dat3)
summary(mod)
```

```
##
## Call:
## lm(formula = Discrete_Oxygen ~ Temperature, data = dat3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6304 -0.6547 -0.1197  0.8393  3.3410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.83364    0.35917  27.379  < 2e-16 ***
## Temperature -0.12106    0.01867  -6.483 3.37e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8854 on 323 degrees of freedom
## Multiple R-squared:  0.1151, Adjusted R-squared:  0.1124
## F-statistic: 42.03 on 1 and 323 DF,  p-value: 3.367e-10
```

Note: the residual standard deviation (i.e., σ) is mislabeled as "Residual standard error" in the summary.lm() output

We can apply the model output to equation 4 to generate an annotated model equation suitable for publication in your results section:

$$y = 9.83 - 0.12x + \varepsilon$$
$$\varepsilon \sim N(0, 0.89^2) \tag{8}$$

4.5 Building a Random Intercept Mixed Effects Model

The ***lme4*** syntax is very similar to the *lm()* function. Building a mixed effects model, however, requires two additional arguments: 1) your random effect structure and 2) the method in which we are fitting our model. We will discuss the latter later in this micro-training; the former uses the following syntax:

```
(random effect structure | Random effect grouping)
```

The bar ("|") in the above syntax means "given." One important aspect of building your random effects model is that the random effect grouping ***must*** be a factor. In our case, however, the grouping factor, *Station_Number*, is a number. Let's convert this into a factor, build our model, and look at the model summary:

```
dat3$fStation = factor(dat3$Station_Number)
mem = lmer(Discrete_Oxygen ~ Temperature + (1|fStation),
           data = dat3, REML = TRUE)
summary(mem)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Discrete_Oxygen ~ Temperature + (1 | fStation)
##    Data: dat3
##
## REML criterion at convergence: 510.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5442 -0.4935 -0.0454  0.4666  7.9005
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  fStation (Intercept) 0.7837   0.8853
##  Residual             0.2637   0.5135
## Number of obs: 325, groups:  fStation, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  9.49399    0.55803  17.013
## Temperature -0.10098    0.01167  -8.651
##
## Correlation of Fixed Effects:
##             (Intr)
## Temperature -0.398
```

The *summary.lm()* output has a "Coefficients" section. The *summary.merMod()* output above, instead, has a "Random effects" section and a "Fixed effects" section. The random effects section tells us where the model uncertainty is allocated: the random effect (σ^2_α) or the residual (σ^2_ε). The fixed effects gives us our grand mean intercept (μ_α) and our slope parameter (β).

We can apply the mixed effects model output to equation 5 to generate an annotated model equation suitable for publication in your results section:

$$y_i = \alpha_{j[i]} - 0.1x_i + \varepsilon_{j[i]} \tag{9}$$
$$\alpha_j \sim N(9.49, 0.89^2), \text{ for } j = 1, ..., J$$
$$\varepsilon_{j[i]} \sim N(0, 0.51^2)$$

We should take note of a few differences in the *lm()* and *lmer()* model summary outputs:

1) While similar, the *lmer()* model fixed effects differ slightly from the *lm* coefficients. This suggests that the inclusion of the random effects slightly changes the relationship between DO and temperature.
2) The inclusion of the random effect of sampling station reduced our model uncertainty from 0.89 in the simple linear regression model to 0.51 in the mixed effects linear regression model. In other words, the inclusion of the random effect term reduced our uncertainty by 42%.

4.5.1 Model visualization

Let's start by visualizing the data, the random effect model predictions, and our grand mean model prediction with a 95% confidence interval. We will be using the ***effects*** package to generate the 95% confidence interval:

```r
# set up to for loop through the stations
plot_sta = unique(dat3$Station_Number)
col_mat = matrix(data = c(140,81,10, 1,102,94, 118,42,131),
                 nrow = 3, ncol = 3, byrow = TRUE)

# create a vector that spans the observed range of temperatures for simulations
temp_seq = seq(min(dat3$Temperature), max(dat3$Temperature),
               length.out=100)

# Use the effects package to calculate a 95% CI around the grand mean model
library(effects)
mem_eff=Effect(focal.predictors = "Temperature", mod=mem,
               xlevels=list(Temperature=temp_seq))

# Use the random effect coefficients to generate random effect predictions
# across the observed range of temperatures
y_15 = coef(mem)$fStation["15",1] + coef(mem)$fStation["15",2]*temp_seq
y_36 = coef(mem)$fStation["36",1] + coef(mem)$fStation["36",2]*temp_seq
y_649 = coef(mem)$fStation["649",1] + coef(mem)$fStation["649",2]*temp_seq

# plot it
par(mfrow=c(1,1), mar=c(4, 4.5, 1, 1)+0.1, oma=rep(0,4))
plot(dat3$Discrete_Oxygen ~ dat3$Temperature, type="n", las=1,
     ylab="Dissolved Oxygen", xlab = "Water Temperature")
for(i in 1:length(plot_sta)){
  sub = subset(dat3, dat3$Station_Number == plot_sta[i])
  points(sub$Discrete_Oxygen ~ sub$Temperature, pch=20,
         col=rgb(col_mat[i,1],col_mat[i,2],col_mat[i,3],
                 maxColorValue = 255, alpha = 200))
}
lines(temp_seq, y_36, type="l", lwd=1.5,
      col=rgb(col_mat[1,1],col_mat[1,2],col_mat[1,3],
              maxColorValue = 255, alpha = 225))
lines(temp_seq, y_15, type="l", lwd=1.5,
      col=rgb(col_mat[2,1],col_mat[2,2],col_mat[2,3],
              maxColorValue = 255, alpha = 225))
lines(temp_seq, y_649, type="l", lwd=1.5,
      col=rgb(col_mat[3,1],col_mat[3,2],col_mat[3,3],
              maxColorValue = 255, alpha = 225))

polygon(x = c(temp_seq, temp_seq[100], rev(temp_seq), temp_seq[1]),
        y=c(mem_eff$lower, mem_eff$upper[100],
            rev(mem_eff$upper), mem_eff$lower[1]),
        col=gray(0.25,0.25),
        border=NA)
lines(mem_eff$fit~temp_seq, lwd=3.5, col="black")
box(which="outer")
```

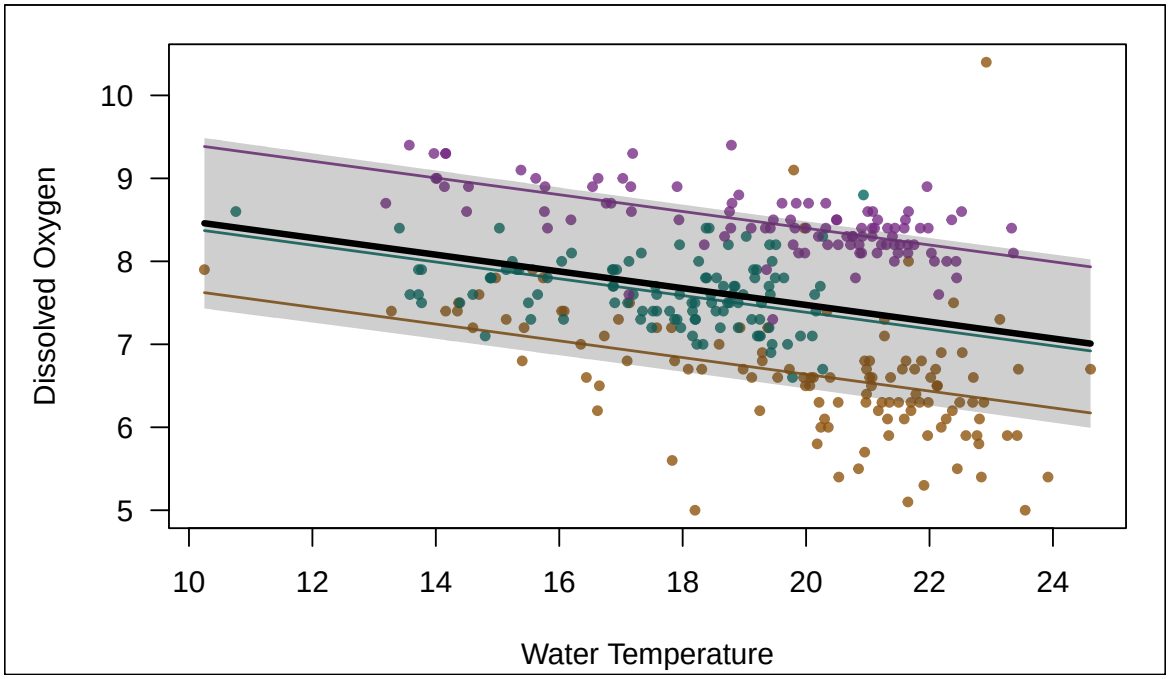

Figure 7: Disolved oxygen across water temperature in the San Francisco Estuary (points). Shown with grand mean random intercept mixed effects model prediction (black line)with uncertainty (95% confidence interval; gray transparent polygon) and random effect predictions (colored thin lines). Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36.

Figure 7 is suitable for publication with or without the random effect predictions (i.e., the thin colored lines). For the sake of a learning exercise, let's overlay the simple liner regression prediction with uncertainty:

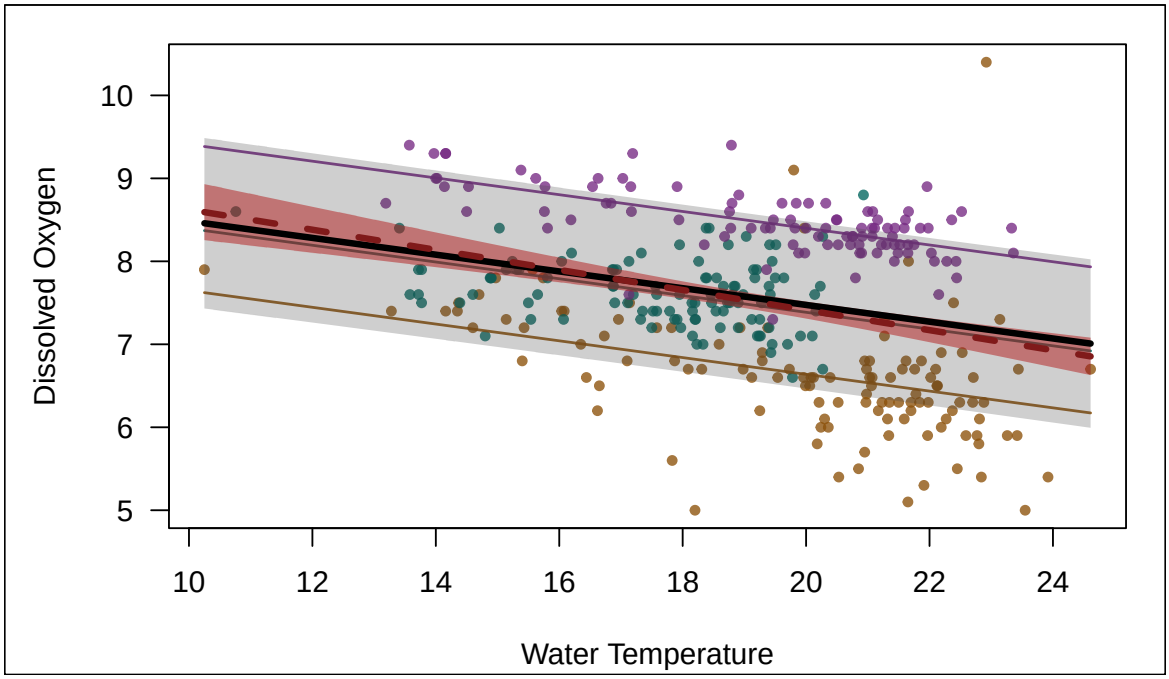

Figure 8: Disolved oxygen across water temperature in the San Francisco Estuary (points). Shown with grand mean random intercept mixed effects model prediction (black line) with uncertainty (95% confidence interval; gray transparent polygon) and random effect predictions (colored thin lines). A simple linear regression model (red dashed line) with uncertainty (95% confidence interval; red transparent polygon) is overlayed as a learning exercise. Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36.

The simple linear regression prediction and grand mean mixed effects prediction do not likely differ in any ecologically meaningful way. However, the difference in the 95% confidence interval between the two methods shown in Figure 7 highlights one of the consequences of ignoring the regression assumption of independence.

As per Figure 4, let's generate a few standard regression diagnostic plots:

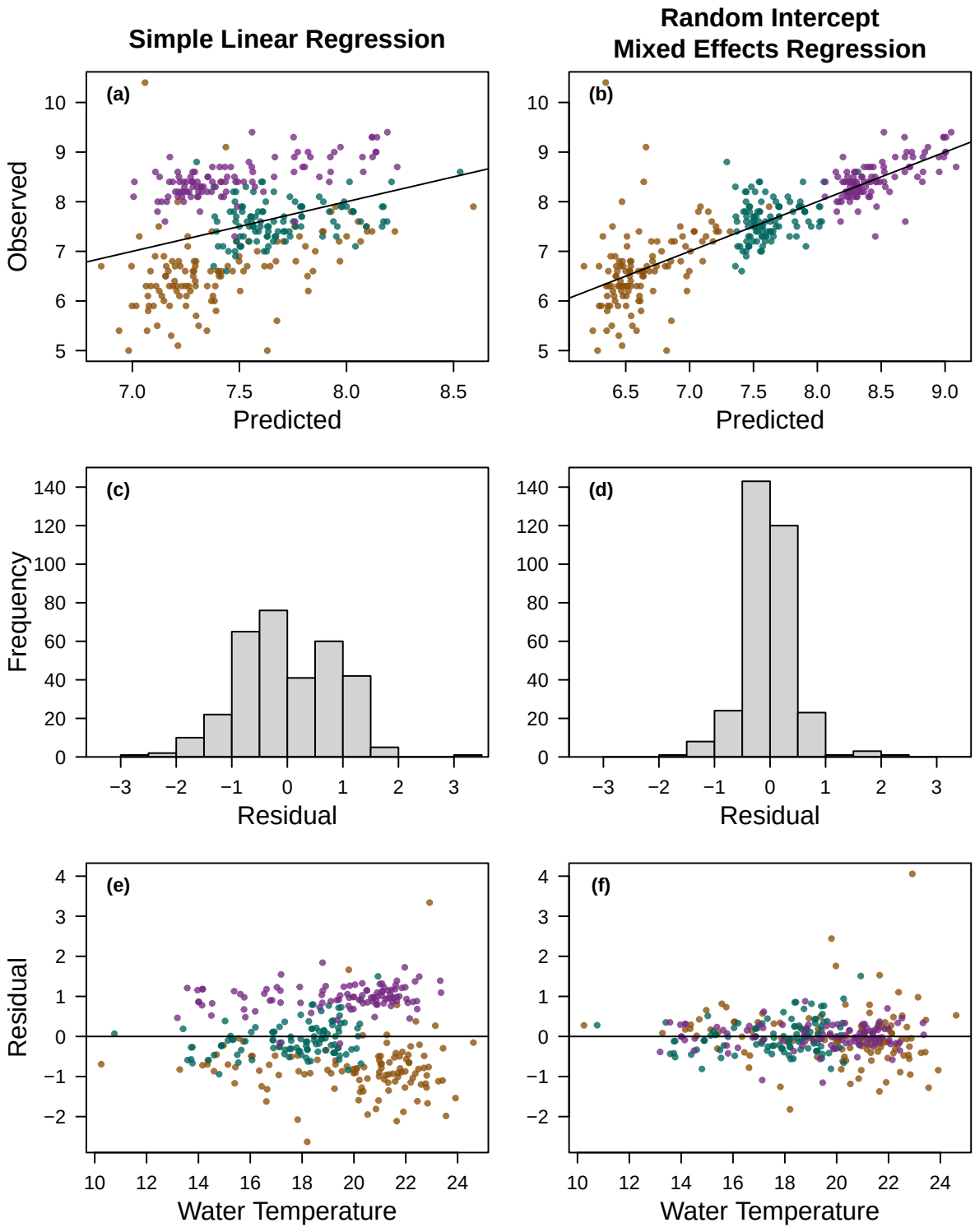

Figure 9: Model diagnostic plot for the simple linear regression model (left column) and the random intercept mixed effects regression model (right column) including (a & b) observed across predicted, (c & d) an assessment of residual normality, and (e & f) residuals across our predictor variable. Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36. Black lines in (a & b) are 1:1 lines; black lines in (e & f) indicate a residual of zero.

Finally, let's visualize the model estimated uncertainty:

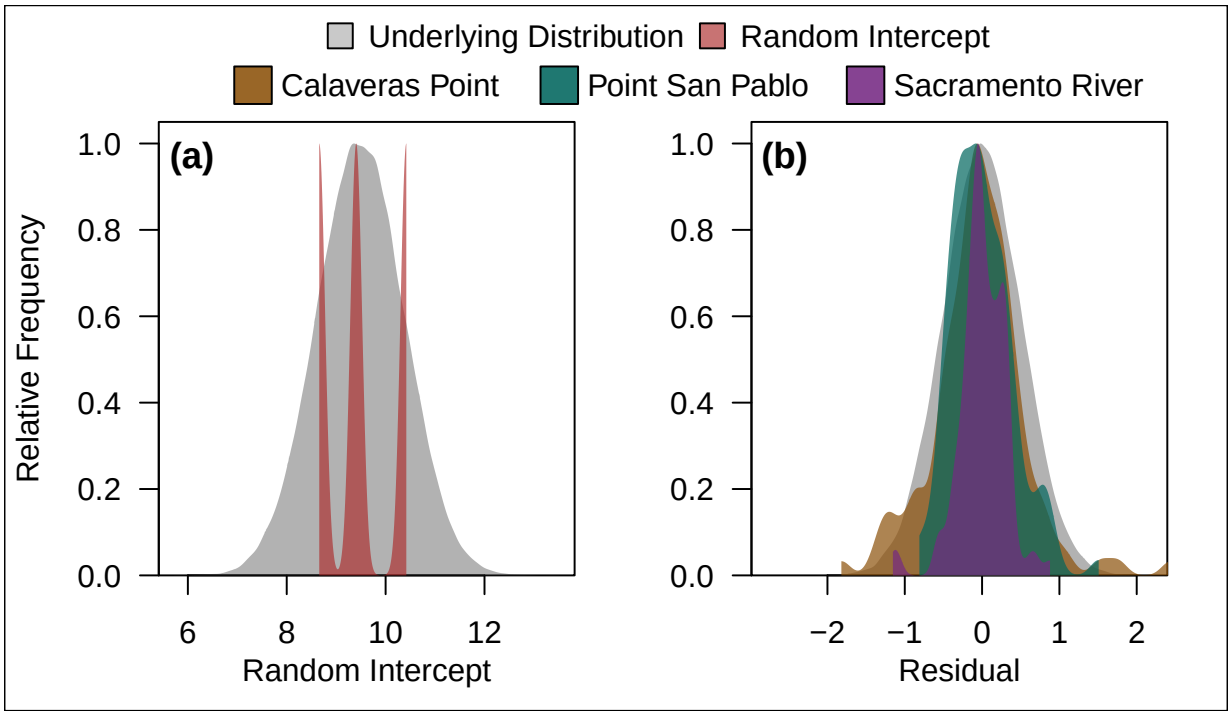

Figure 10: Model estimated underlying distributions of (a) α_j and (b) ε (gray normal distributions) as per equation 9 overlayed with (a) model predicted α_j values and (b) model residuals.

We see that the random intercept predictions (red polygon in Figure 10a) fall within the underlying distribution derived from equation 9. However, we cannot really tell whether the random intercept predictions (α_j) follow a normal distribution due to a sample size of 3. Conversely, the residuals associated with each of the random effect groups (i.e., stations) appear to be normally distributed. If I were publishing this model, however, I would further explore why the variance associated with the residuals of Sacramento River observations are substantially more narrow that the other stations.

4.6 Building a Random Slope Mixed Effects Model

```
mem = lmer(Discrete_Oxygen ~ Temperature + (Temperature - 1|fStation),
           data = dat3, REML = TRUE)
summary(mem)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Discrete_Oxygen ~ Temperature + (Temperature - 1 | fStation)
##    Data: dat3
##
## REML criterion at convergence: 516.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6544 -0.4914 -0.0630  0.4555  8.0324
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  fStation Temperature 0.001976 0.04446
##  Residual             0.268289 0.51797
## Number of obs: 325, groups:  fStation, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  9.41237    0.22863  41.168
## Temperature -0.09643    0.02831  -3.406
##
## Correlation of Fixed Effects:
##             (Intr)
## Temperature -0.419
```

Remember, the random effects section tells us where the model uncertainty is allocated: the random effect (σ^2_β) or the residual (σ^2_ε). The fixed effects gives us our intercept (α) and our grand mean slope parameter (μ_β).

We can apply the mixed effects model output to equation 6 to generate an annotated model equation suitable for publication in your results section:

$$y_i = 9.41 + \beta_{j[i]} x_i + \varepsilon_{j[i]} \tag{10}$$
$$\beta_j \sim N(-0.1, 0.04^2), \text{ for } j = 1, ..., J$$
$$\varepsilon_{j[i]} \sim N(0, 0.52^2)$$

We should take note of a few differences in the *lm()* and *lmer()* model summary outputs:

1) While similar, the *lmer()* model fixed effects differ slightly from the *lm* coefficients. This suggests that the inclusion of the random effects slightly changes the relationship between DO and temperature.
2) The inclusion of the random effect of sampling station reduced our model uncertainty from 0.89 in the simple linear regression model to 0.52 in the mixed effects linear regression model. In other words, the inclusion of the random effect term reduced our uncertainty by 41.5%.

4.6.1 Model visualization

Let's start by visualizing the data, the random effect model predictions, and our grand mean model prediction with a 95% confidence interval. We will be using the ***effects*** package to generate the 95% confidence interval:

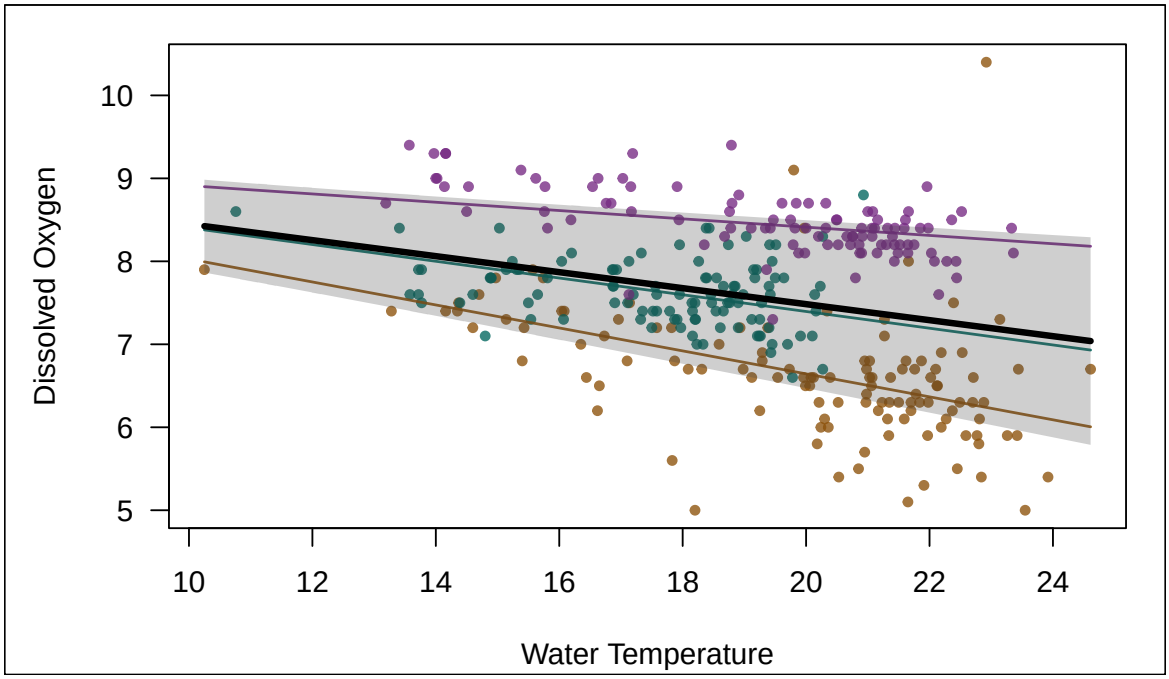

Figure 11: Disolved oxygen across water temperature in the San Francisco Estuary (points). Shown with grand mean random slope mixed effects model prediction (black line) with uncertainty (95% confidence interval; gray transparent polygon) and random effect predictions (colored thin lines). Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36.

Figure 11 is suitable for publication with or without the random effect predictions (i.e., the thin colored lines). However, the problem highlighted in Figure 2c is evident, suggesting this is an inappropriate random effect structure for this dataset (at least until we center the data).

For the sake of a learning exercise, let's overlay the simple liner regression prediction with uncertainty:

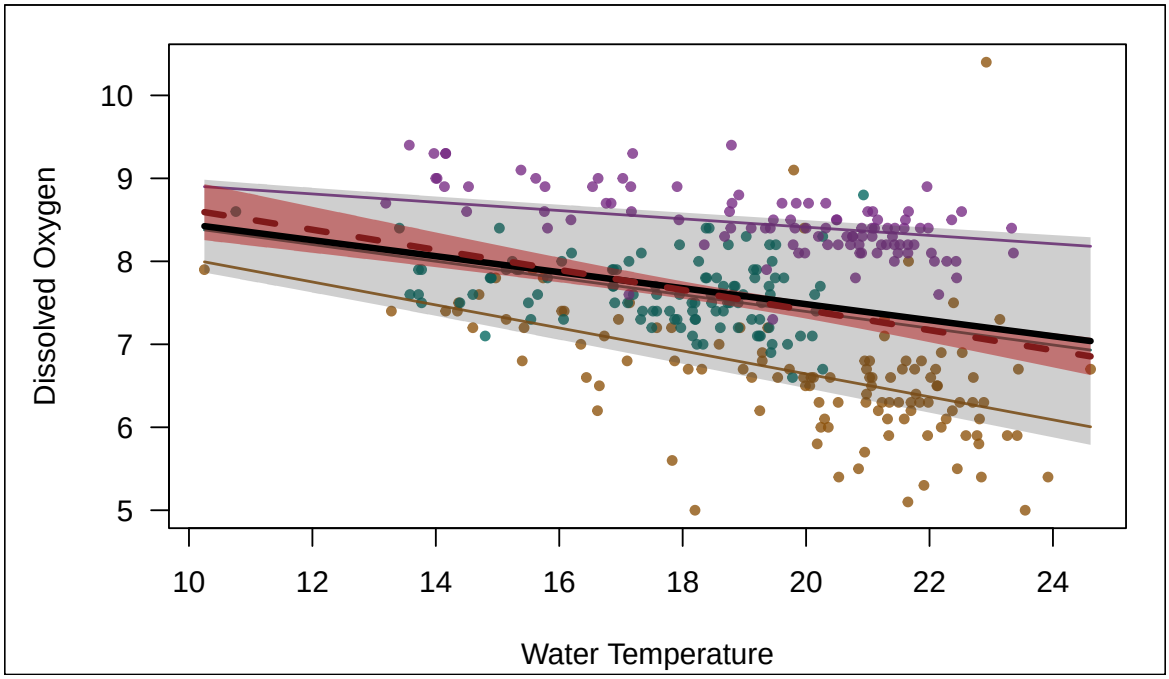

Figure 12: Disolved oxygen across water temperature in the San Francisco Estuary (points). Shown with grand mean random slope mixed effects model prediction (black line) with uncertainty (95% confidence interval; gray transparent polygon) and random effect predictions (colored thin lines). A simple linear regression model (red dashed line) with uncertainty (95% confidence interval; red transparent polygon) is overlayed as a learning exercise. Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36.

The simple linear regression prediction and grand mean mixed effects prediction do not likely differ in any ecologically meaningful way. However, the difference in the 95% confidence interval between the two methods shown in Figure 11 highlights one of the consequences of ignoring the regression assumption of independence.

As per Figure 4, let's generate a few standard regression diagnostic plots:

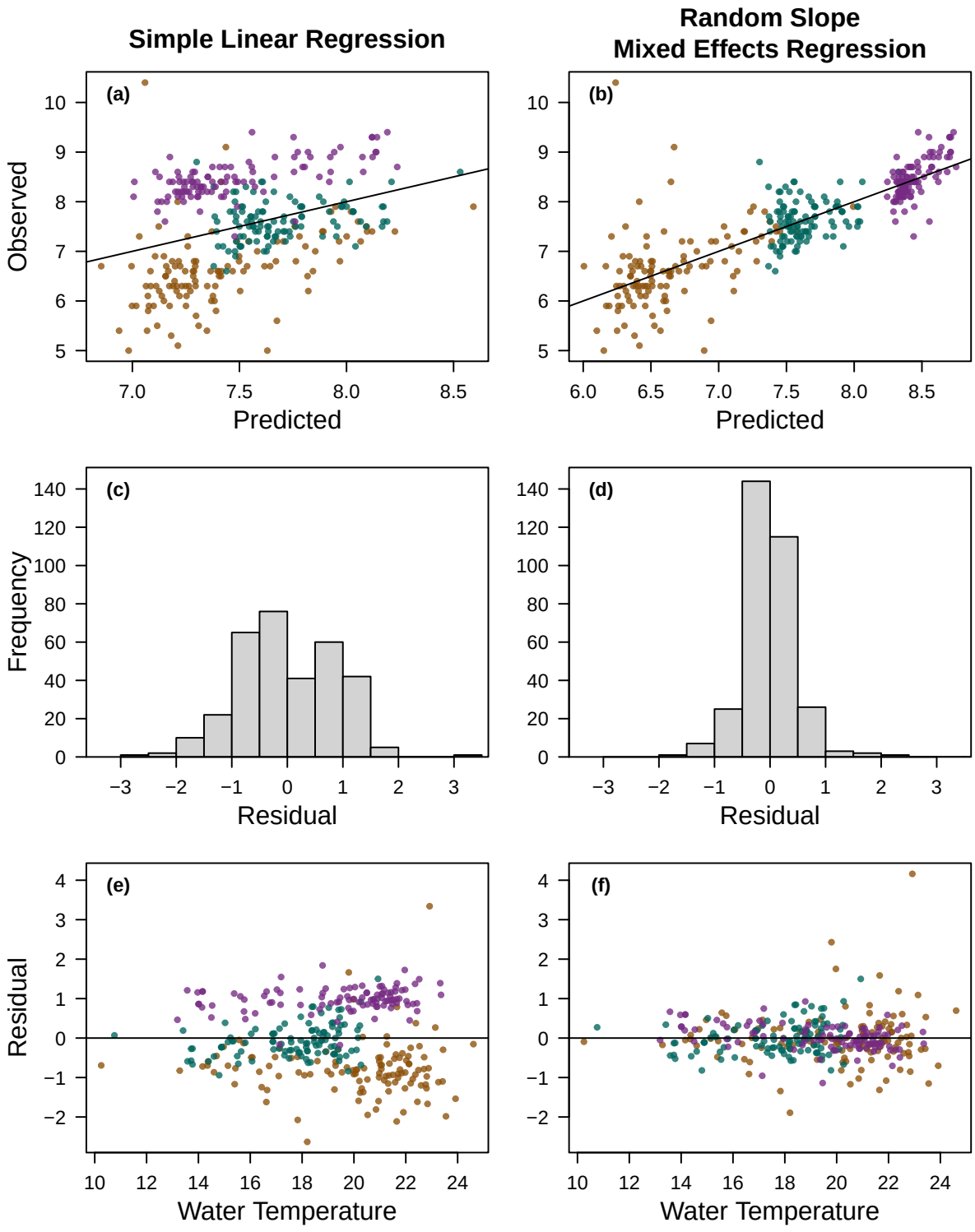

Figure 13: Model diagnostic plot for the simple linear regression model (left column) and the random slope mixed effects regression model (right column) including (a & b) observed across predicted, (c & d) an assessment of residual normality, and (e & f) residuals across our predictor variable. Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36. Black lines in (a & b) are 1:1 lines; black lines in (e & f) indicate a residual of zero.

Finally, let's visualize the model estimated uncertainty:

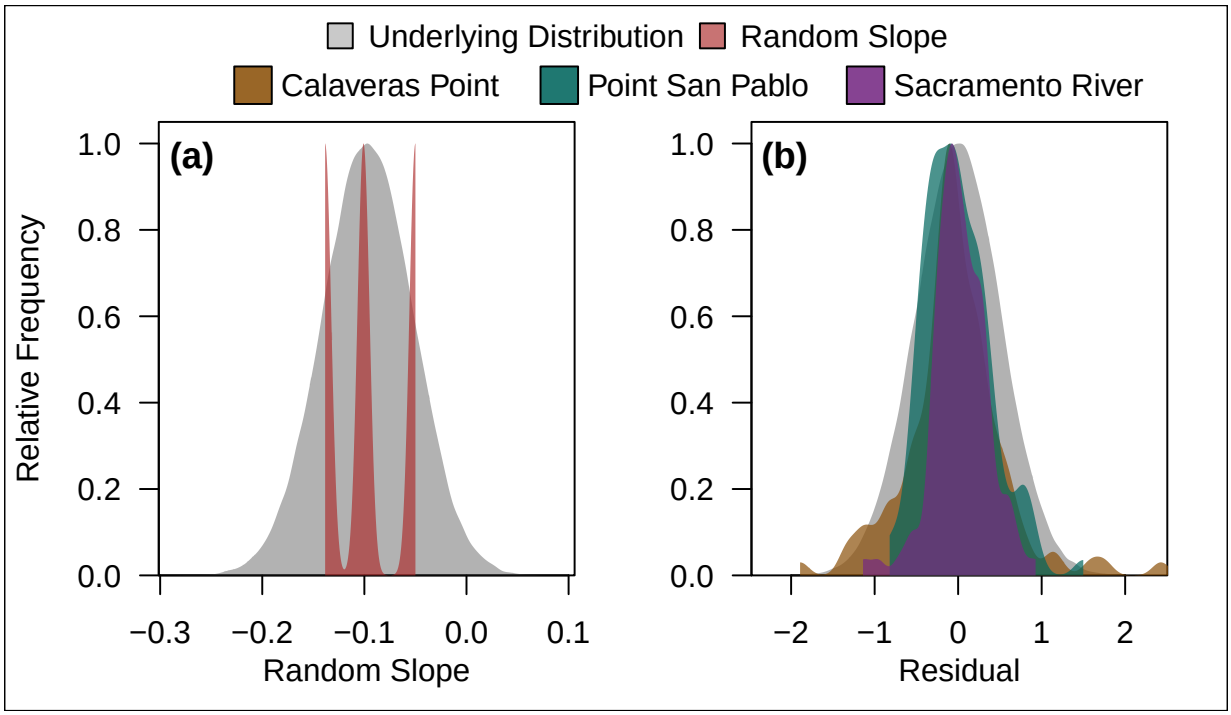

Figure 14: Model estimated underlying distributions of (a) β_j and (b) ε (gray normal distributions) as per equation 10 overlayed with (a) model predicted β_j values and (b) model residuals.

We see that the random slope predictions (red polygon in Figure 14a) fall within the underlying distribution derived from equation 10. However, we cannot really tell whether the random slope predictions (α_j) follow a normal distribution due to a sample size of 3. Conversely, the residuals associated with each of the random effect groups (i.e., stations) appear to be normally distributed. If I were publishing this model, however, I would further explore why the variance associated with the residuals of Sacramento River observations are substantially more narrow that the other stations.

4.7 Building a Random Intercept and Slope Mixed Effects Model

```
mem = lmer(Discrete_Oxygen ~ Temperature + (1 + Temperature|fStation),
           data = dat3, REML = TRUE)
summary(mem)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Discrete_Oxygen ~ Temperature + (1 + Temperature | fStation)
##    Data: dat3
##
## REML criterion at convergence: 509.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6091 -0.4692 -0.0549  0.4632  8.0176
##
```

```
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  fStation (Intercept) 0.5889051 0.76740
##           Temperature 0.0005195 0.02279  -0.02
##  Residual             0.2617046 0.51157
## Number of obs: 325, groups:  fStation, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  9.42467    0.49793  18.928
## Temperature -0.09669    0.01777  -5.442
##
## Correlation of Fixed Effects:
##             (Intr)
## Temperature -0.314
```

Remember, the random effects section tells us where the model uncertainty is allocated: the random effect (σ^2_β) or the residual (σ^2_ε). The fixed effects gives us our intercept (α) and our grand mean slope parameter (μ_β).

We can apply the mixed effects model output to equation 6 to generate an annotated model equation suitable for publication in your results section:

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \varepsilon_{j[i]} \tag{11}$$

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} 9.42 \\ -0.1 \end{pmatrix}, \begin{pmatrix} 0.77^2 & -0.02 \\ -0.016 & 0.02^2 \end{pmatrix} \right), \text{ for } j = 1, ..., J$$

$$\varepsilon_{j[i]} \sim N(0, 0.51^2)$$

We should take note of a few differences in the *lm()* and *lmer()* model summary outputs:

1) While similar, the *lmer()* model fixed effects differ slightly from the *lm* coefficients. This suggests that the inclusion of the random effects slightly changes the relationship between DO and temperature.
2) The inclusion of the random effect of sampling station reduced our model uncertainty from 0.89 in the simple linear regression model to 0.51 in the mixed effects linear regression model. In other words, the inclusion of the random effect term reduced our uncertainty by 42.22%.

4.7.1 Model visualization

Let's start by visualizing the data, the random effect model predictions, and our grand mean model prediction with a 95% confidence interval. We will be using the ***effects*** package to generate the 95% confidence interval:

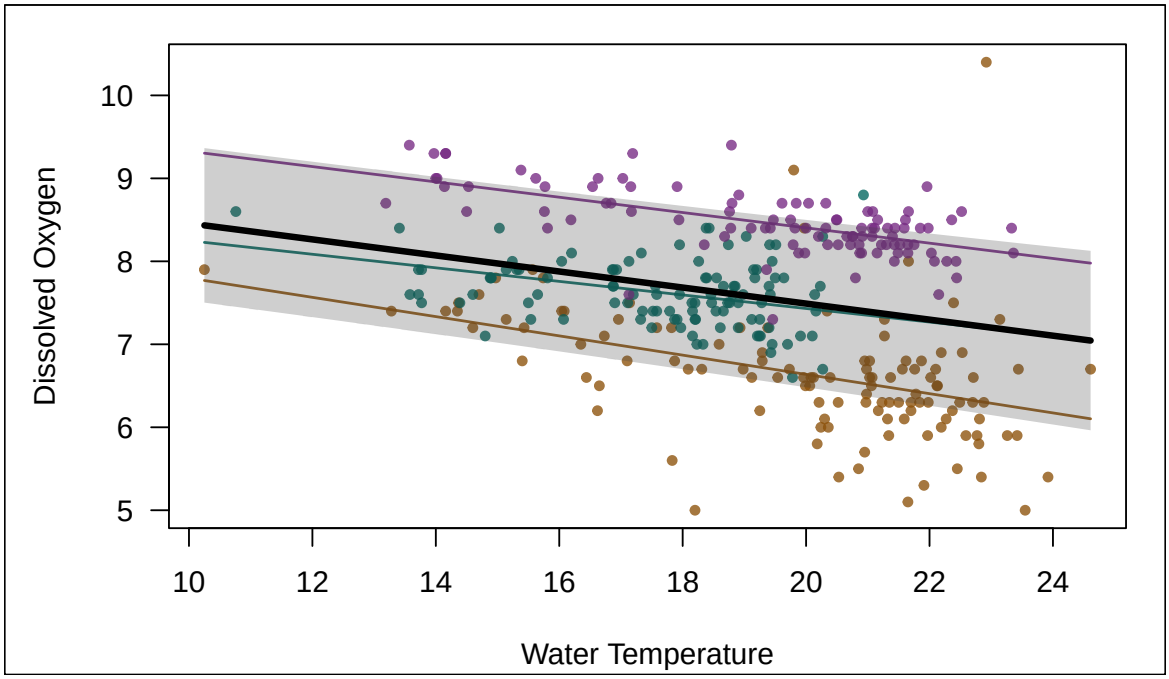

Figure 15: Disolved oxygen across water temperature in the San Francisco Estuary (points). Shown with grand mean random intercept and slope mixed effects model prediction (black line) with uncertainty (95% confidence interval; gray transparent polygon) and random effect predictions (colored thin lines). Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36.

Figure 15 is suitable for publication with or without the random effect predictions (i.e., the thin colored lines). For the sake of a learning exercise, let's overlay the simple liner regression prediction with uncertainty:

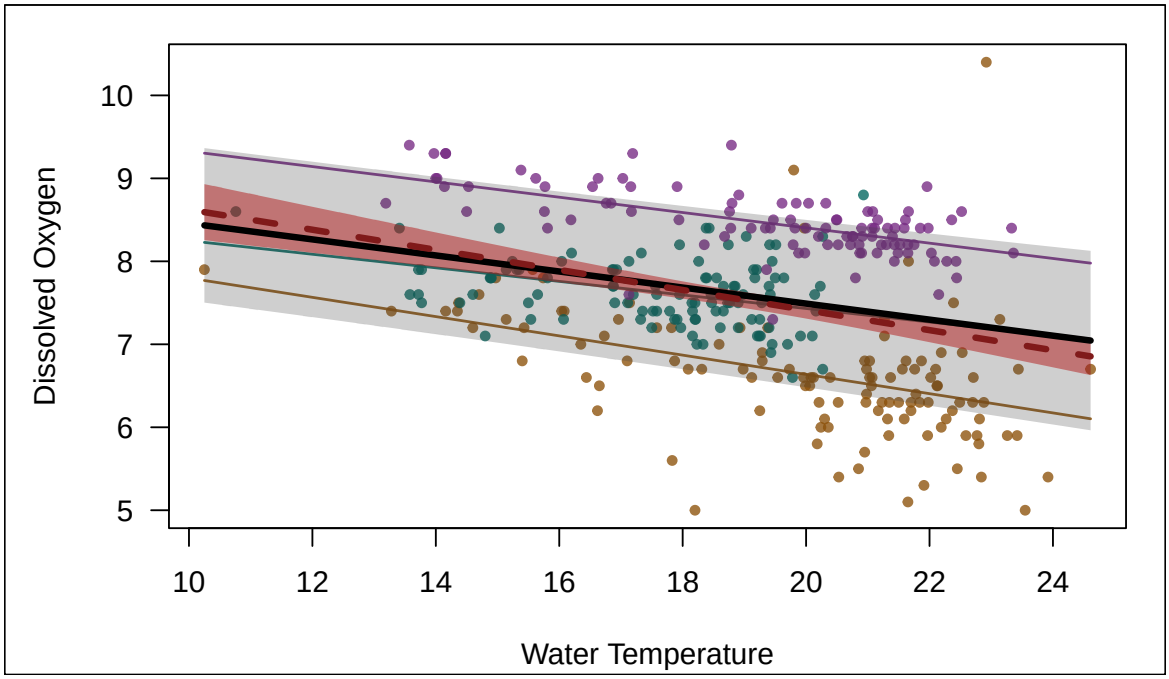

Figure 16: Disolved oxygen across water temperature in the San Francisco Estuary (points). Shown with grand mean random intercept and slope mixed effects model prediction (black line) with uncertainty (95% confidence interval; gray transparent polygon) and random effect predictions (colored thin lines). A simple linear regression model (red dashed line) with uncertainty (95% confidence interval; red transparent polygon) is overlayed as a learning exercise. Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36.

The simple linear regression prediction and grand mean mixed effects prediction do not likely differ in any ecologically meaningful way. However, the difference in the 95% confidence interval between the two methods shown in Figure 15 highlights one of the consequences of ignoring the regression assumption of independence.

As per Figure 4, let's generate a few standard regression diagnostic plots:

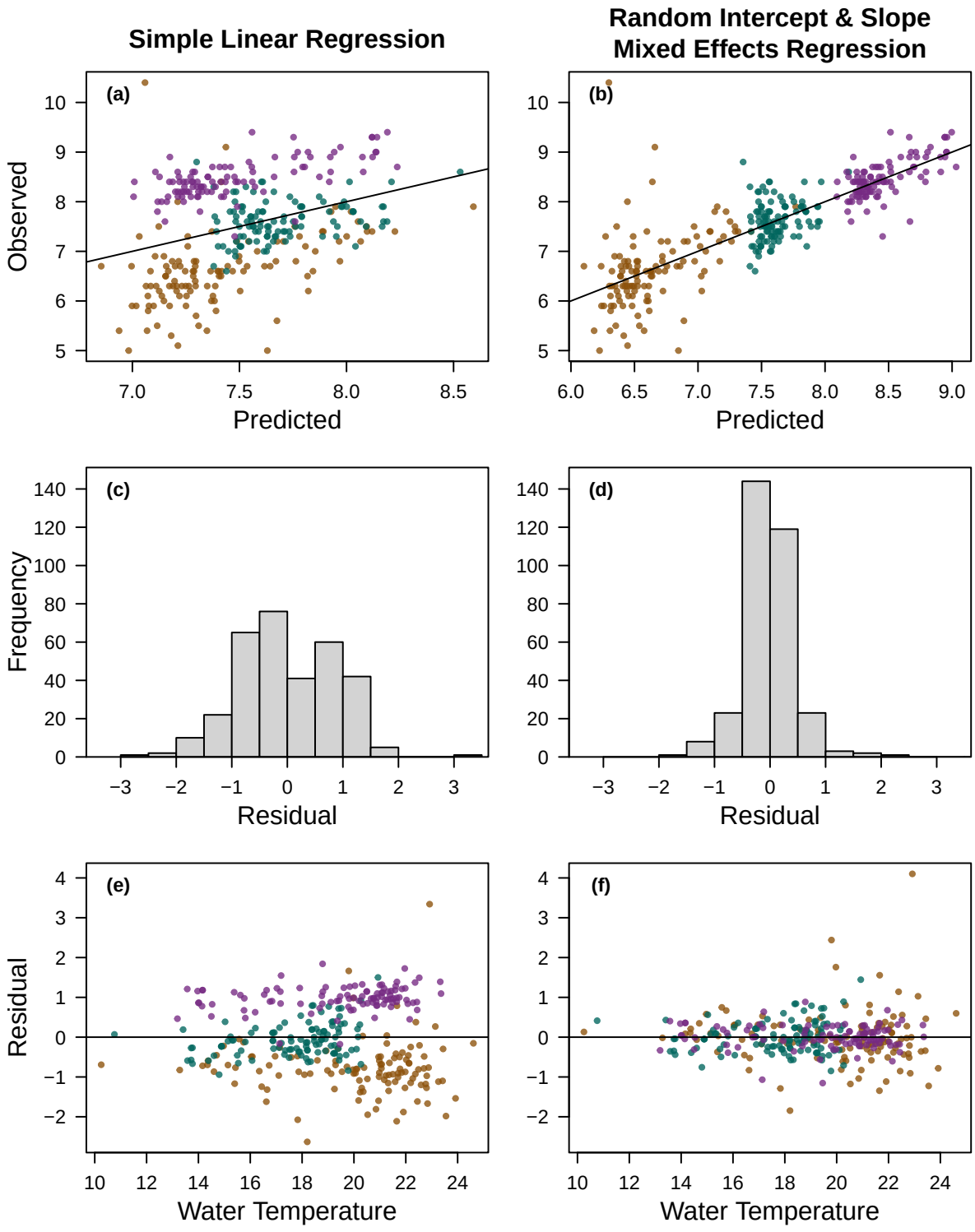

Figure 17: Model diagnostic plot for the simple linear regression model (left column) and the random intercept and slope mixed effects regression model (right column) including (a & b) observed across predicted, (c & d) an assessment of residual normality, and (e & f) residuals across our predictor variable. Purple denotes Sacramento River station 649, teal denotes Point San Pablo station 15, and brown denotes Calaveras Point station 36. Black lines in (a & b) are 1:1 lines; black lines in (e & f) indicate a residual of zero.

Finally, let's visualize the model estimated uncertainty:

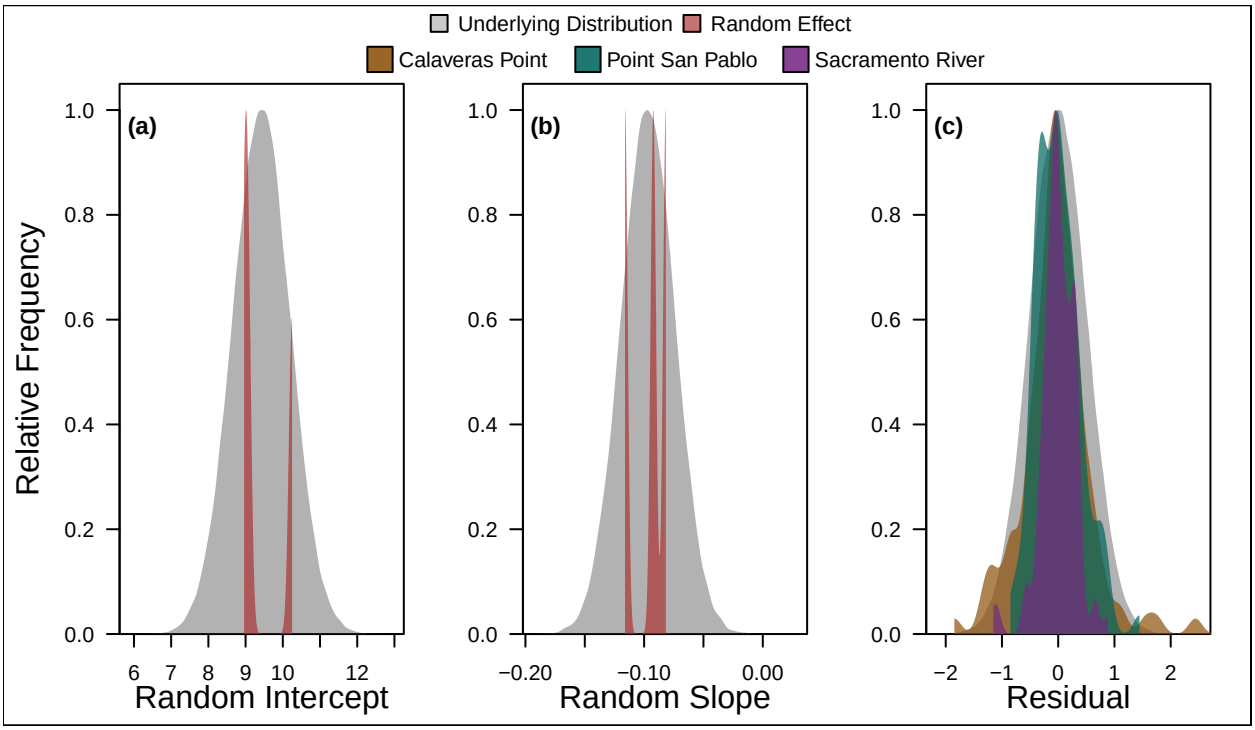

Figure 18: Model estimated underlying distributions of (a) α_j, (b) β_j, and (b) ε (gray normal distributions) as per equation 11 overlayed with (a) model predicted α_j values, (b) model predicted β_j values, and (b) model residuals.

We see that the random slope predictions (red polygon in Figure 18a) fall within the underlying distribution derived from equation 10. However, we cannot really tell whether the random intercepts or slopes predictions (α_j & β_j) follow a normal distribution due to a sample size of 3. Conversely, the residuals associated with each of the random effect groups (i.e., stations) appear to be normally distributed. If I were publishing this model, however, I would further explore why the variance associated with the residuals of Sacramento River observations are substantially more narrow that the other stations.

5 On REML, AIC, BIC, and Model Selection

The following assumes the reader is familiar with regression model selection procedures. If not, I advise a little Google time prior to continuing.

The following is an oversimplified explanation. Refer to Gelman and Hill (2006) and/or Zuur et al. (2009) for more information

5.1 REML

The *REML* argument in the *lmer()* function determines how the likelihood function is calculated. "REML" stands for restricted maximum likelihood, and you can set this argument to TRUE or FALSE. *REML=TRUE* tells the model to estimate parameters using the restricted maximum likelihood while *REML=FALSE* tells the model to estimate parameters using the maximum likelihood. For a deep dive into this subject, I recommend Zuur et al. (2009) section 5.6. Briefly, however, REML dictates *how* sample size is determined

and *how* you are penalized for adding parameters. This can slightly change your model estimated parameters (coefficients and error terms), although the effect increases with the number of parameters you are estimating.

Choosing *REML=TRUE* or *REML=FALSE* becomes critically important under two primary conditions: 1) when performing model selection and 2) when using an information criterion such as the Bayesian Information Criterion (BIC) or Deviance Information Criterion (DIC). To over simplify, the "number of parameters" we are estimating in a mixed effects regression is not straight forward because we are estimating coefficients and residual variance, but we have this middle set of parameters the random effect variance (see Gelman and Hill (2006) section 24.3 for more details). This becomes challenging when trying to evaluate whether or not to add or remove a covariate in a mixed effects model selection procedure.

5.2 AIC: Akaike Information Criterion

In a simple linear regression framework, we would identify the addition (or removal) of a covariate significant if the change was associated with a ΔAIC of ± 2, ± 4, or ± 7 (depending on how conservative you are being; see Burnham and Anderson (2002) for more details).

Let's take a look at how AIC is calculated:

$$AIC = -2log_e\hat{L}(\hat{\theta}|data, g) + 2K \tag{12}$$

where $-2log_e\hat{L}(\hat{\theta}|data, g)$ is two times the negative log-likelihood of the model parameters $\hat{\theta}$ *given* your data and the model g and K is the number of parameters being estimated. That is, regression models *maximize* the likelihood (or chance) that your model coefficients are correct given your data and model g. So, if our goal is to *maximize* the likelihood that the model coefficients are correct, a **more negative** AIC value is desirable. However, AIC penalizes us by a value of 2 for each additional parameter we are asking the model to estimate (equation 12). **But, AIC does not distinguish between random effects parameters and fixed effects parameters nor does it incorporate sample size.**

5.3 BIC: Bayesian Information Criterion

Let's take a look at how BIC is calculated:

$$BIC = -2log_e\hat{L}(\hat{\theta}|data, g) + log_e(n)K \tag{13}$$

$-2log_e\hat{L}(\hat{\theta}|data, g)$ is the same as for the calculation of AIC, but BIC includes a n term, where n is your sample size. *HOWEVER*, n changes whether *REML=TRUE* or *REML=FALSE*. When *REML=TRUE*, the n term is equal to your sample size *minus* the number of parameters you are estimating. When *REML=FALSE*, the n term is equal to your sample size.

5.4 REML and Model Selection

In general, I recommend the following:

- Use *REML = TRUE* to compare random effect structure
  - For example, comparing a random slope model to a random intercept model
- Use *REML = FALSE* to compare fixed effects structure
  - For example, when determining whether to add water temperature as a fixed effect
- Use *REML = TRUE* for final reporting
  - After determining your fixed and random effect structures, run your final model with *REML=TRUE* and report those coefficients in your publication or report

6 Final Thoughts

This micro-training handout is just a bare glimpse into the world of mixed effects modeling. I encourage you to explore these models and other mixed effects models including logistic regression and general additive modeling. I also encourage you to copy and paste any errors R may throw into google and see what the analytical community has to say.

6.1 nlme v lme4

While the code and syntax used in this micro-training is specific to the ***lme4*** package, the theory will hold whether you continue to use ***lme4*** or explore R's other common mixed effects package: ***nlme***. Here is what I view as the main distinction between the two (aside from syntax):

- ***lme4*** is your best option if you have complex random effect structures, such as a three level model. However, ***lme4*** cannot handle within-group correlation structures that are necessary to account for temporal or spatial autocorrelation
- ***nlme*** is your best option to handle within-group correlation structures that are necessary to account for temporal or spatial autocorrelation. However, ***nlme*** cannot handle complex random effect structures, such as a three level model.

Literature Cited

Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Inference: A Practical Information-Theoretic Approach. Second edition. Springer-Verlag, New York.

Gaeta, J. W., M. J. Guarascio, G. G. Sass, and S. R. Carpenter. 2011. Lakeshore residential development and growth of largemouth bass (Micropterus salmoides): A cross-lakes comparison. Ecology of Freshwater Fish 20:92–101.

Gelman, A., and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge.

Schraga, T., E. Nejad, C. Martin, and J. Cloern. 2020, March. USGS measurements of water quality in San Francisco Bay.

Zuur, A., E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer Science & Business Media.