Readings for MTH207

Derek H. Ogle

2020-12-20

Contents

Preface

This BOOK-IN-PROGRESS was last updated on: 20 Dec 2020.

Chapter 1

Model Types & Methods

During this course we will examine a variety of models called either *general linear* or *generalIZED linear* models. General linear models have a quantitative response variable and generally assume that the "errors" around the model follow a normal distribution. General linear models that we will discuss are a **One-Way ANOVA**[1], **Two-WAY ANOVA**, **Simple Linear Regression**, and **Indicator Variable Regression**. GeneralIZED linear models do not require a quantitative response variable nor "errors" that are normally distributed. Thus, generalIZED linear models are more flexible than general linear models. The only generalIZED linear model that we will encounter in this course is **Logistic Regression**, but the chi-square test from your introductory statistics course can also be cast as a generalIZED linear model.

Response Variable: The variable thought to depend upon, be explained by, or be predicted by other variables.

All models covered in this course will have **only one** response variable

Both general and generalIZED linear models can have a single explanatory variable that can be either quantitative or categorical, or multiple explanatory variables that can be all quantitative, all categorical, or a mixture of both quantitative and categorical. Ultimately, there can be several explanatory variables in a model, but we will only consider one or two explanatory variables in this course.

Explanatory Variable: A variable thought to explained or be able to predict the response variable.

[1] ANOVA is short for ANalysis Of VAriance

Table 1.1: Response and explanatory variables types for the linear models considered in this course.

Response	Explanatory	Linear.Model
Quantitative	Categorical (only one)	One-Way ANOVA
Quantitative	Categorical (two)	Two-Way ANOVA
Quantitative	Quantitative (only one)	Simple Linear Regression
Quantitative	Quantitative (one) & Categorical (one)	Indicator Variable Regression
Binomial	Quantitative (or Both)	(Binary) Logistic Regression

1.1 Distinguishing Methods

The five methods that will be covered in this course can be distinguished by considering only the type of response variable and the types and number of explanatory variables (Table 1.1). Thus, you will want to review variable types and definitions and distinctions of response and explanatory variables from your introductory statistics course.

1.2 Method Purposes

As seen above, each method uses different types of data. Not surprisingly then, each method is used to test different hypotheses or has a different analytical purpose. These purposes will be discussed in detail in subsequent modules. However, the major objective of each method is explained briefly below (in the order that we will cover them).

Each example uses a data set that contains data about mirex concentrations (`mirex`) for two species of salmon (`species`) captured in six years between 1977 and 1999 (`year`) in Lake Ontario. The weight of each fish (`weight`) and whether or not the mirex concentration exceeded the EPA limit of 0.1 mg/kg (`exceeds_limit`) were also recorded.

A **one-way ANOVA** is used to determine if the means of the quantitative response variable (`mirex`) differ among two or more groups defined by a single categorical variable (e.g., `year`).

A **two-way ANOVA** is used to determine if the means of the quantitative

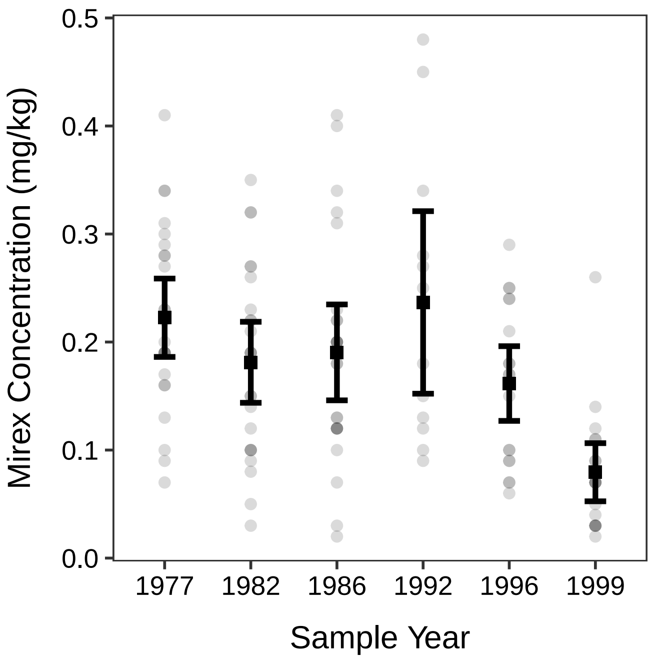

Figure 1.1: Mean mirex concentration by sample year. This is an example of a One-Way ANOVA.

response variable (`mirex`) differ among groups of one categorical variable (e.g., `year`), among groups of another categorical variable (e.g., `species`), or by the interaction between the two categorical variables.

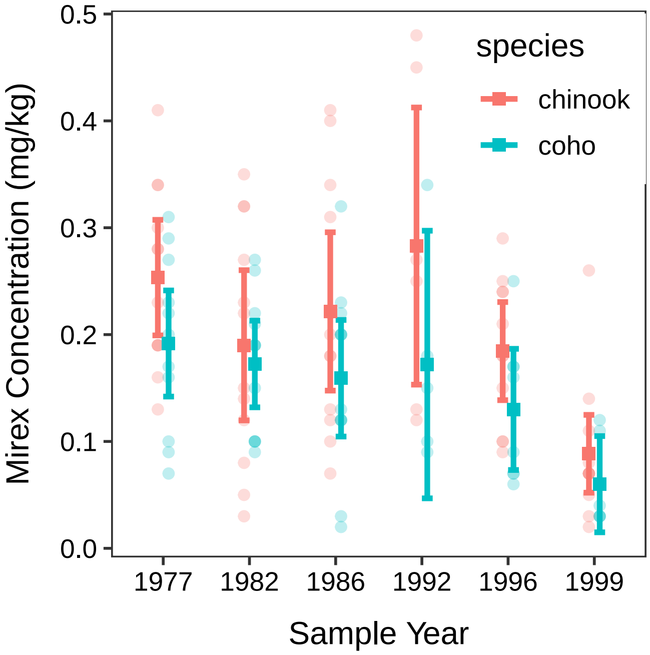

Figure 1.2: Mean mirex concentration by sample year and salmon species. This is an example of a Two-Way ANOVA.

A **simple linear regression** is used to determine if there is a relationship between the quantitative response variable (e.g., `mirex`) and a single quantitative explanatory variable (e.g., `weight`).

An **indicator variable regression** is used to determine if the relationship between a quantitative response (e.g., `mirex`) and a quantitative explanatory variable (e.g., `weight`) differs between two or more groups defined by a categorical explanatory variable (e.g., `species`). This will look like two (or more) simple linear regressions are being compared.

A **logistic regression** is used to determine if there is a relationship between the probability of "success" for a binary[2] categorical response variable (e.g.,

[2] Binary means there are only two categories – generically "success" and "failure".

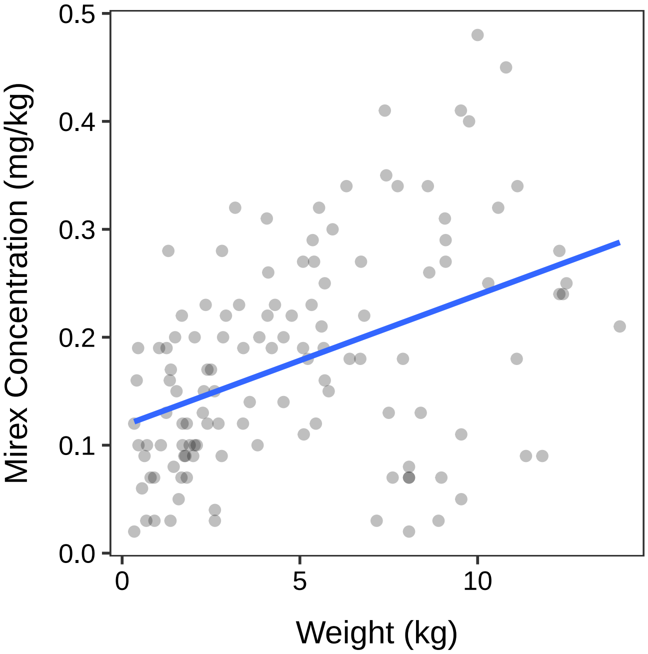

Figure 1.3: Mirex concentration by fish weight. This is an example of a Simple Linear Regression.

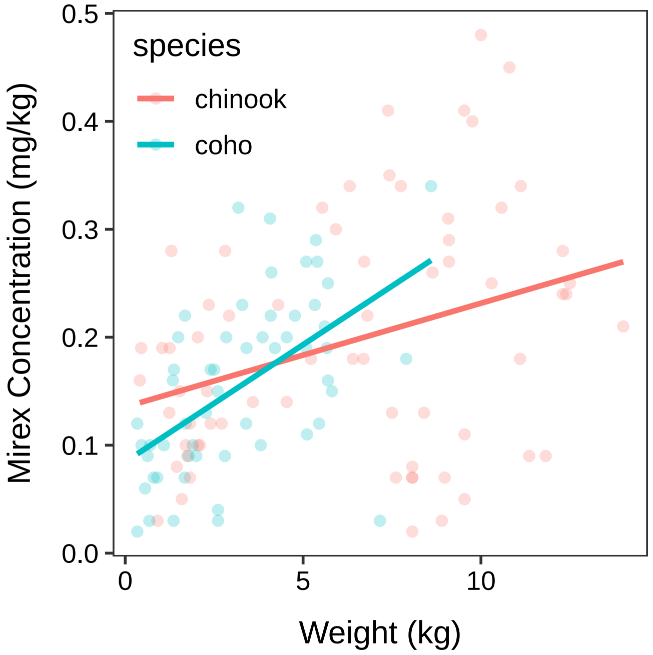

Figure 1.4: Mirex concentration by fish weight seprated by salmon species. This is an example of an Indicator Variable Regression.

`exceeds_limit`) and the quantitative explanatory variable (e.g., `weight`).

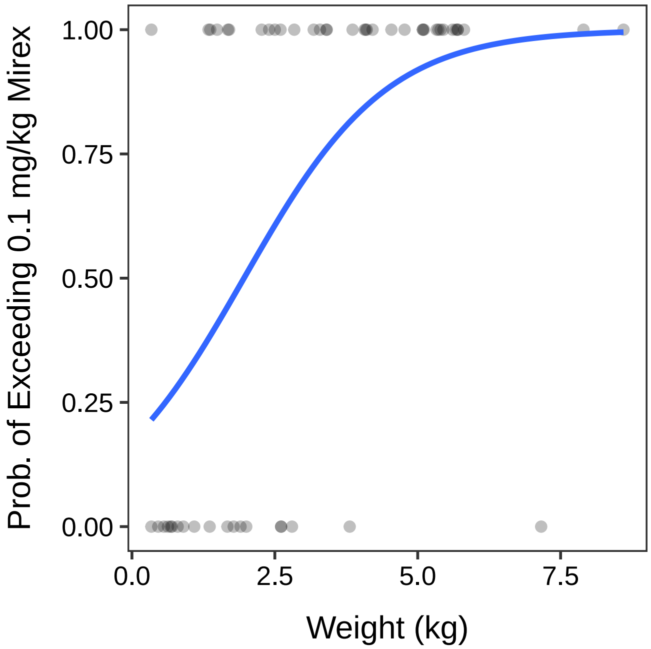

Figure 1.5: The probability that the mirex concentration exceed the 0.1 mg/kg threshold by fish weight. This is an example of a Logistic Regression.

From these examples it should be apparent that "ANOVAs" are about **comparing means** among groups and will look like means (usually with confidence intervals) plotted as points for each group. In contrast "regressions" are about exploring **relationships** and will look like a line or a curve when plotted.

ANOVAs compare means; regressions examine relationships.

Chapter 2

Model Concepts

2.1 What is a Model

A model is a representation of something or some phenomena. It is usually a simplification or an abstraction that helps our understanding of the more complex reality. A mathematical or statistical model is an equation or system of equations that is meant to characterize the general characteristics of observations. Statistical models do not represent every observation perfectly, rather they attempt to best represent the "central tendency" of the observations. Weather forecasts are based on mathematical and statistical models. You have observed at least two statistical models in your introductory statistics course – the mean and the regression line (Figure 2.1).

Models can predict an observation but generally not perfectly. For example, weather forecasters predict the temperature for tomorrow but will most likely be off by (hopefully only) a small amount.

An observed value of the response variable can be thought of as being equal to a value predicted from a model plus some deviation, or error, from that prediction; i.e.,

$$\text{Observed Response} = \text{Model Predicted Response} + \text{error}$$

For example, tomorrow's temperature may be 74oF, which is the predicted 76oF from the forecaster's model plus -2oF "error."

In statistics, one model for predicting the response variable for an individual in a group is to use the mean for the group. My best guess at the height of an unknown student is to guess that they are average for "their group." Obviously, most individuals are not truly average, so the specific individual will deviate

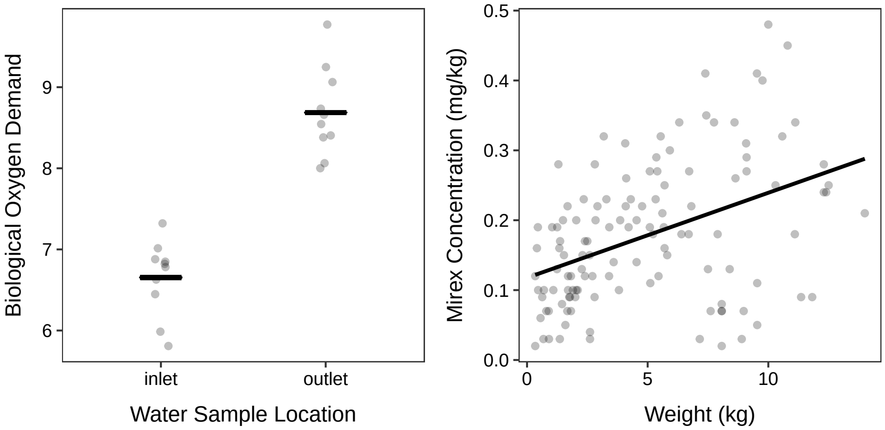

Figure 2.1: Two examples of models seen in your introductory statistics course – two means (Left) and regression line (Right).

from the mean. In Figure 2.2 an observation is shown as a red point, the predicted value for that individual is shown as a horizontal line at the mean for the individual's group, and the "error" from this prediction is shown as the vertical red line.

We always predict the **response** variable with a model.

In the context of a simple linear regression, the predicted value is obtained by plugging the observed value of the explanatory variable into the regression equation. Thus, the "error" is the vertical distance between an observed point and the corresponding point on the line (Figure 2.3).

Many hypothesis tests, including the two-sample t-test, can be cast in a framework of competing statistical models. Using this framework requires assessing the relative fit (to data) and complexity of a model. The remainder of this module is about measuring fit and complexity of models. We will discuss fit and formally compare two models to see which is "best" in the next module.

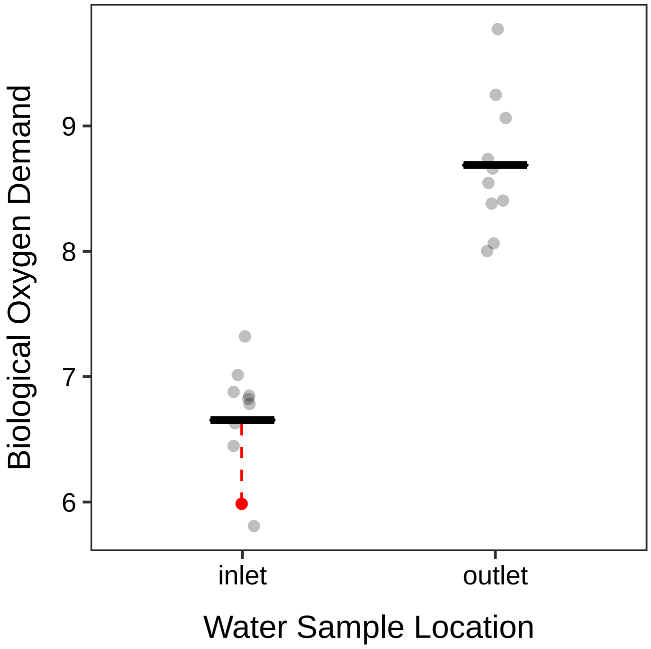

Figure 2.2: Biological oxygen demand versus sample location (points) with group means shown by horizontal segments. The residual from a model that uses a separate mean for both groups is shown.

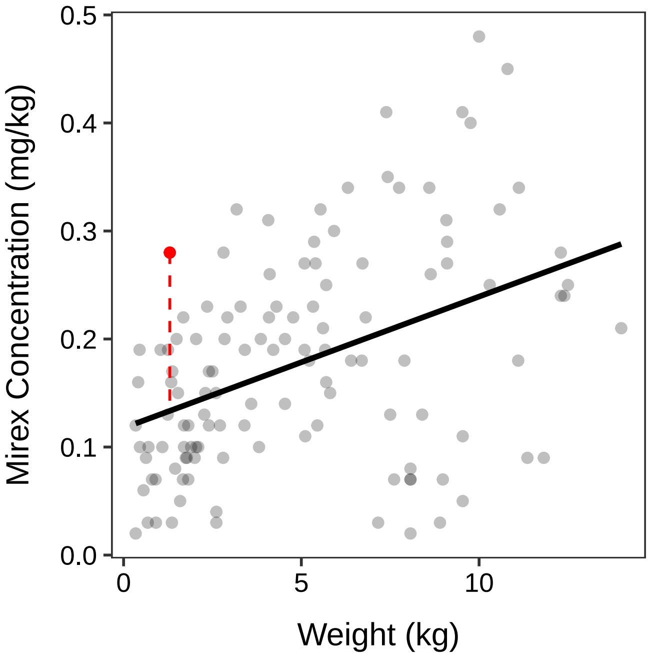

Figure 2.3: Mirex concentration versus fish weight with a simple linear regression line show. The residual from the regression line model is shown.

2.2 Assessing Fit (SS)

2.2.1 A Residual

A residual is an estimate of the “error” discussed in the previous section. If you rearrange the formula shown above and replace “error” with “residual” you see that

$$\text{residual} = \text{Observed Response} - \text{Model Predicted Response}$$

Visually a residual is the vertical distance between a point and the “model”, as shown by the vertical dashed lines above (or further below). Residuals are vertical distances because they are the difference between two values of the response variable, which is always plotted on the y-axis.

Residuals are *vertical* distances between an observation and the model.

Residuals are negative if the point is “below” the model prediction and positive if the point is “above” the model prediction. More importantly, the absolute value of the residual is a measure of how close the model prediction is to the point or how well the model fits the individual point. Large residuals (in an absolute value sense) mean that the point is far from the model prediction and, thus, the model does not represent that point very well. Points with small residuals, in contrast, are near the model prediction and are thus well-represented by the model. Figure 2.4 shows points with relatively large residuals in red and relatively small residuals in blue.

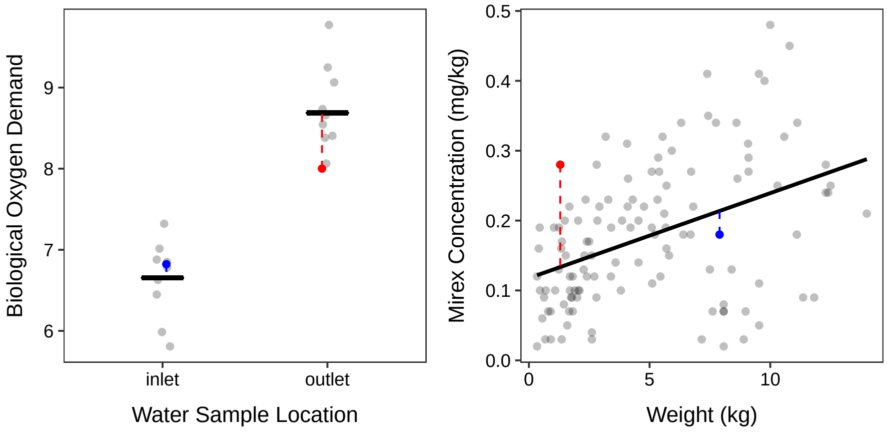

Figure 2.4: Same plots as previously but with a ”large” residual shown in red and a ”small” residual shown in blue.

2.2.2 Residual Sum-of-Squares

If a residual measures how closely a model comes to a point then it stands to reason that the sum of all of the residuals measures how closely a model comes to all of the points. Unfortunately, because residuals are negative and positive they always sum to 0.[1] Thus, the sum of all residuals is not a useful measure of the overall fit of a model.

Instead of summing residuals, statisticians sum squared residuals into a quantity called a **residual sum-of-squares (RSS)**.[2] Using the formula for a residual from above, an RSS for a given set of observed data and a model is computed with

$$\text{RSS} = \sum_{data} (\text{Observed Response} - \text{Model Predicted Response})^2$$

The RSS measures how closely the model comes to *all* of the observations.

The RSS is on an unfamiliar scale (squared residuals?) but it maintains the same conceptual idea that summing residuals would have. Mainly, the smaller the RSS the more closely the points are to the model. The full set of residuals required to compute an RSS are shown in Figure 2.5.

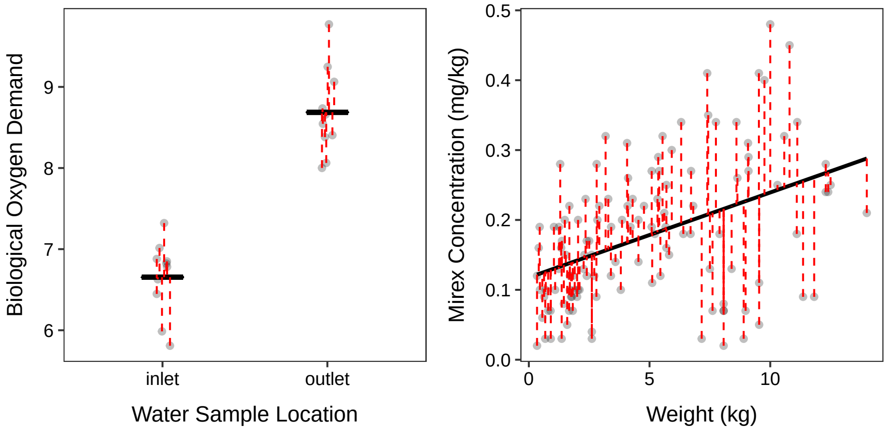

Figure 2.5: Same plots as previously but with all residuals shown.

As a value, the RSS is a measure of how *poorly* the model fits the data – i.e., small values are a good fit, large values are a poor fit. Thus, the RSS is often

[1] Under certain reasonable assumptions.

[2] Some statisticans call this an error sum-of-squares or a sum of squared errors (SSE)

called “a measure of lack-of-fit” of the model to the observations.

An RSS is a measure of the “lack-of-fit” of a model to the data.

Unfortunately, the magnitude of the RSS is only useful in comparison to other RSS computed for different models from the same data. We will discuss this further in the next module.

2.3 Residual Degrees-of-Freedom

You used degrees-of-freedom (df) with t-tests and chi-square tests in your introductory statistics course. However, you likely did not discuss what degrees-of-freedom mean and where they come from. I will discuss this briefly here, but we will use df more in the next module.

Residual degrees-of-freedom (Rdf) are the number of observations that are “free” to vary if the sample size (n) and number of parameters estimated is known. As a simple example, suppose that we know that $\bar{x}$=13 from n=4 observations. With just this information can I tell you the values for the four observations that went into $\bar{x}$? Clearly I cannot. If you give me one observation can I tell you the remaining three? No! If you tell me two? No! If you tell me three observations can I tell you the last observation? Yes, because the total of the four numbers must be 52 (=$n\bar{x}$=4×13); so the last number must be 52 minus the total of the three numbers you told me. In this case, three numbers were “free” to be any value before the last number was set. Thus, this case has three residual degrees-of-freedom.

Residual degrees-of-freedom are more complicated to explain in other situations, but generally

$$\text{Rdf} = \text{Number of Observations} - \text{Number of Model Parameters}$$

In the example above, there were four observations (n) and one model parameter – $\bar{x}$ – so df=4-1=3. In Figure 2.1-Left there are 20 observations and two parameters (i.e., two group means) so Rdf=20-2=18. In Figure 2.1-Right there are 122 observations and two parameters (i.e., the slope and intercept of the regression line) so Rdf=122-2=120.

As a general rule, parameter estimates are more precisely estimated with more residual degrees-of-freedom. Thus, models that “preserve” residual degrees-of-freedom (i.e., have fewer parameters) are preferred, all else being equal.

2.4 Mean-Squares

Sums-of-squares are useful measures of model fit, but they are largely uninterpretible on their own. However, if a sum-of-squares is divided by its corresponding degrees-of-freedom it is called a **Mean-Square (MS)**. Mean-squares are the **variance** (i.e., squared standard deviation) of individuals around a given model. Mean-squares have useful mathematical properties as you will see in future modules. However, visually the square root of a mean square loosely describes how far each point is from the model (i.e., the "errors"), on average. The mean-squares are thus a measure of the "noise" around each model.

MS are variances; thus, the square root of MS are standard deviations