Regression

ES/STT 7140: Statistical Modeling for Environmental Data

Brandon M. Greenwell

A good modelling tool

At a minimum:

- Universally applicable in classification and regression
- Unexcelled accuracy
- Capable of handling large data sets
- Handle missing values effectively

A good modelling tool

At a minimum:

- Universally applicable in classification and regression
- Unexcelled accuracy
- Capable of handling large data sets
- Handle missing values effectively

It would also be a plus to know:

- Which variables are important?
- How do variables interact?
- What is the shape of the data (i.e., how does it cluster?)
- How do the features separate classes?
- Are their novel cases and outliers?

A good modelling tool

At a minimum:

- Universally applicable in classification and regression
- Unexcelled accuracy
- Capable of handling large data sets
- Handle missing values effectively

It would also be a plus to know:

- Which variables are important?
- How do variables interact?
- What is the shape of the data (i.e., how does it cluster?)
- How do the features separate classes?
- Are their novel cases and outliers?

-Leo Breiman

Statistical models

Introduction

- In the one-sample t-test, we are interested in learning about the mean of a normal distribution/population

$$y_i \sim N\left(\mu, \sigma^2\right), \quad i = 1, 2, \ldots, n$$

- For example, y might represent the shell length of a randomly selected zebra mussel from a stream or lake in Michigan

- It is often convenient to think of the data y_i in terms of a statistical model:

$$y_i = \mu + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right)$$

Introduction

- In the one-sample t-test, we are interested in learning about the mean of a normal distribution/population

$$y_i \sim N\left(\mu, \sigma^2\right), \quad i = 1, 2, \ldots, n$$

- For example, y might represent the shell length of a randomly selected zebra mussel from a stream or lake in Michigan

- It is often convenient to think of the data y_i in terms of a statistical model:

$$y_i = \mu + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right)$$

  - *data* = *mean* + *remainder*

  - The above two expressions are mathematically equivalent

  - The remainder is the difference between the observed values and the mean, often refered to as the residuals

Introduction

- In the two-sample t-test problem, we are interested in the difference between the means of two populations (or groups):

$$y_{1i} \sim N\left(\mu_1, \sigma^2\right), \quad i = 1, 2, \ldots, n_1$$

$$y_{2j} \sim N\left(\mu_2, \sigma^2\right), \quad j = 1, 2, \ldots, n_2$$

$$\delta = \mu_2 - \mu_1$$

- As a linear model, we could use

$$y_k = \mu_1 + \delta g_k + \epsilon_k, \quad \sim N\left(0, \sigma^2\right), \quad k = 1, 2, \ldots, n_1 + n_2$$

  - Here, g_k is a *dummy variable* (one for `intake` group and zero for `discharge` group)

Introduction

- In the two-sample t-test problem, we are interested in the difference between the means of two populations (or groups):

$$y_{1i} \sim N\left(\mu_1, \sigma^2\right), \quad i = 1, 2, \ldots, n_1$$

$$y_{2j} \sim N\left(\mu_2, \sigma^2\right), \quad j = 1, 2, \ldots, n_2$$

$$\delta = \mu_2 - \mu_1$$

- As a linear model, we could use

$$y_k = \mu_1 + \delta g_k + \epsilon_k, \quad \sim N\left(0, \sigma^2\right), \quad k = 1, 2, \ldots, n_1 + n_2$$

  - Here, g_k is a *dummy variable* (one for `intake` group and zero for `discharge` group)

- Illustration: `clams.R`

- A similar approach can be used for ANOVA procedures as well

Introduction

- The primary focus of this chapter is to introduce more complicated models

Introduction

- The primary focus of this chapter is to introduce more complicated models
- In particular, we will focus on

Introduction

- The primary focus of this chapter is to introduce more complicated models
- In particular, we will focus on
  - *simple linear regression* (SLR) models (i.e., linear regression with a single predictor)

Introduction

- The primary focus of this chapter is to introduce more complicated models

- In particular, we will focus on

  - *simple linear regression* (SLR) models (i.e., linear regression with a single predictor)

  - *multiple linear regression* (MLR) models (i.e., linear regression with multiple predictors)

Introduction

- The primary focus of this chapter is to introduce more complicated models

- In particular, we will focus on

    - *simple linear regression* (SLR) models (i.e., linear regression with a single predictor)

    - *multiple linear regression* (MLR) models (i.e., linear regression with multiple predictors)

    - *nonlinear regression* (NLR) models

Introduction

- The primary focus of this chapter is to introduce more complicated models
- In particular, we will focus on
    - *simple linear regression* (SLR) models (i.e., linear regression with a single predictor)
    - *multiple linear regression* (MLR) models (i.e., linear regression with multiple predictors)
    - *nonlinear regression* (NLR) models
- In the next chapter, we will look at a more general class of regression models called *generalized linear models* (GzLMs)
    - GzLMs include both *logistic regression* and *Poisson regression* models

Simple linear regression

The SLR model

- A regression model is a formal means of expressing the two essential ingredients of a statistical model:

    1. A tenancy of the **response variable**, y, with a **predictor variable**, x, in some systematic fashion

    2. A scattering of points around the hypothesized curve of statistical relationship

- These two characteristics are embodied in a regression model by postulating that:

    1. There is a **probability distribution** of y for each level of x

    2. The means of these probability distributions vary in some systematic fashion with x

The SLR model

- There are many reasons for the vast popularity of regression models in statistical practice
- Regression models allow us to relate variables together in a mathematical form which can provide insight into the relationships between the variables of interest

The SLR model

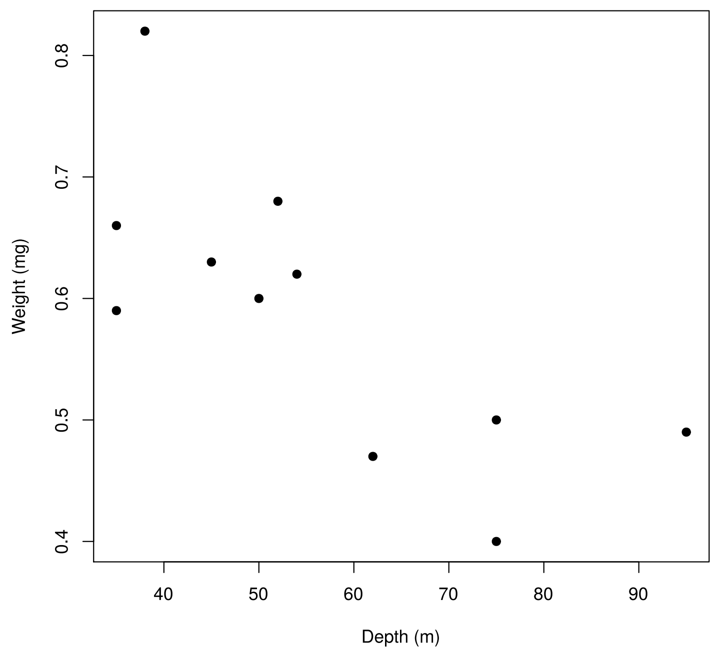
- There are many reasons for the vast popularity of regression models in statistical practice
- Regression models allow us to relate variables together in a mathematical form which can provide insight into the relationships between the variables of interest
- Regression models allow us to **determine statistically** if a *response variable* is related to one or more other *predictor variables*
    - For instance, we may want to determine the effect of increasing levels of DDT on eggshell thicknesses. How does increasing levels of DDT effect eggshell thickness?

The SLR model

- There are many reasons for the vast popularity of regression models in statistical practice
- Regression models allow us to relate variables together in a mathematical form which can provide insight into the relationships between the variables of interest
- Regression models allow us to **determine statistically** if a *response variable* is related to one or more other *predictor variables*
    - For instance, we may want to determine the effect of increasing levels of DDT on eggshell thicknesses. How does increasing levels of DDT effect eggshell thickness?
- Another common use of regression models is to **predict a response**
    - For instance, if water is contaminated with a certain level of toxin, can we predict the amount of accumulation of this toxin in a fish that lives in the water?

The SLR model

- Many statistical applications deal with modeling how a single response variable, denoted y, depends on a single predictor, denoted x
- Illustration: `shrimp.R`



The SLR model

- From the previous *scatterplot*, one can see a fairly strong relationship between between the weight of the *Diporeia* and the depth of water where the Diporeia are found

- The scatterplot suggests that a straight line relationship between weight of Diporeia (y) and water depth (x) may be a reasonable way to model the data:

$$weight = \beta_0 + \beta_1 depth$$

- Here, β_0 is the y-intercept of the regression line and β_1 is the slope (i.e., *rate of change*)

- Of course, the shrimp data do not lie perfectly on a straight line and a probabilistic model is needed to account for the variability of points about the line:

The SLR model

- From the previous *scatterplot*, one can see a fairly strong relationship between between the weight of the *Diporeia* and the depth of water where the Diporeia are found

- The scatterplot suggests that a straight line relationship between weight of Diporeia (y) and water depth (x) may be a reasonable way to model the data:

$$weight = \beta_0 + \beta_1 depth$$

- Here, β_0 is the y-intercept of the regression line and β_1 is the slope (i.e., *rate of change*)

- Of course, the shrimp data do not lie perfectly on a straight line and a probabilistic model is needed to account for the variability of points about the line:

$$weight_i = \beta_0 + \beta_1 depth_i + \epsilon_i, \quad i = 1, \ldots, 11$$

Assumptions of the SLR model

- The SLR model

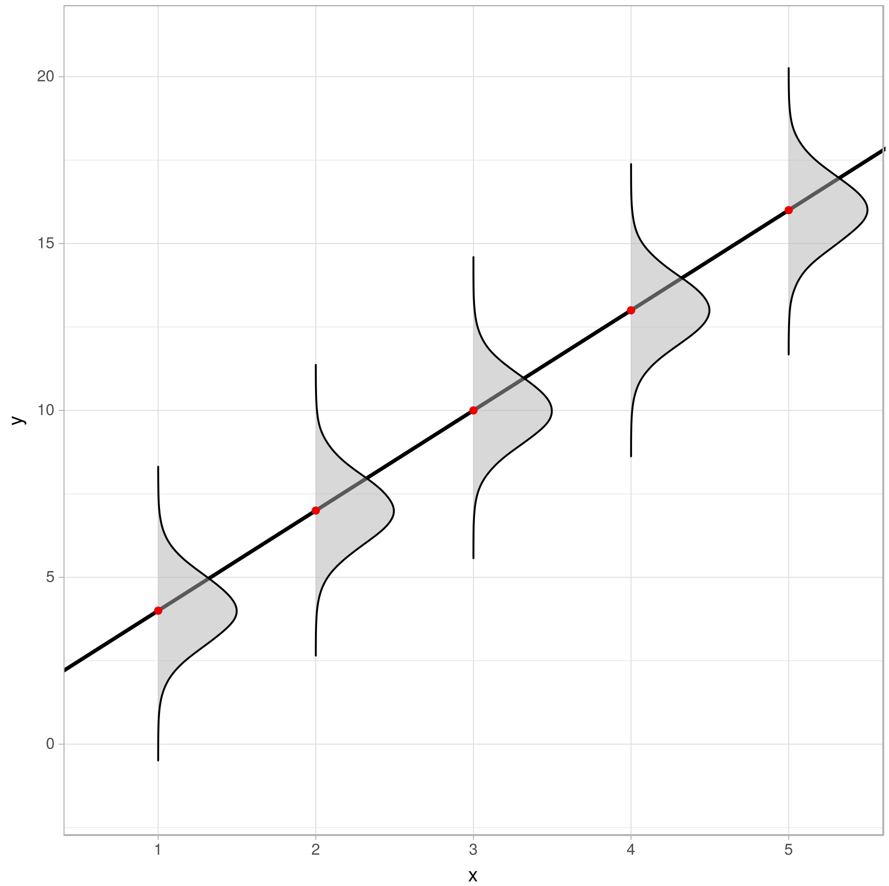
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

  assumes that

  1. Independent observations (i.e., the random errors are independent)
  2. The errors have constant variance (i.e., *homoscedasticity*)
  3. The errors are normally distributed (for statistical inference)

- If these assumptions are not met, then alternative methods need to be applied (e.g., *weighted least squares* or *mixed-effects models*)

Assumptions of the SLR model



Least squares estimation

- How do we estimate the model coefficients β_0 and β_1?
- There are an infinite number of lines passing through the data points $\{x_i, y_i\}_{i=1}^n$
- The *least squares* (LS) solution seeks to find β_0 and β_1 that minimize the *sum of squares*:

$$SS\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

- Hence, the LS line is the "best" fitting line in terms of minimizing $SS\left(\beta_0, \beta_1\right)$

Least squares estimation

- How do we estimate the model coefficients β_0 and β_1?

- There are an infinite number of lines passing through the data points $\{x_i, y_i\}_{i=1}^{n}$

- The *least squares* (LS) solution seeks to find β_0 and β_1 that minimize the *sum of squares*:

$$SS\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

- Hence, the LS line is the "best" fitting line in terms of minimizing $SS\left(\beta_0, \beta_1\right)$

- So, how do we minimize $SS\left(\beta_0, \beta_1\right)$?

Least squares estimation

- How do we estimate the model coefficients β_0 and β_1?
- There are an infinite number of lines passing through the data points $\{x_i, y_i\}_{i=1}^n$
- The *least squares* (LS) solution seeks to find β_0 and β_1 that minimize the *sum of squares*:

$$SS\left(\beta_0, \beta_1\right) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

- Hence, the LS line is the "best" fitting line in terms of minimizing $SS\left(\beta_0, \beta_1\right)$
- So, how do we minimize $SS\left(\beta_0, \beta_1\right)$?
    - **CALCULUS!!**

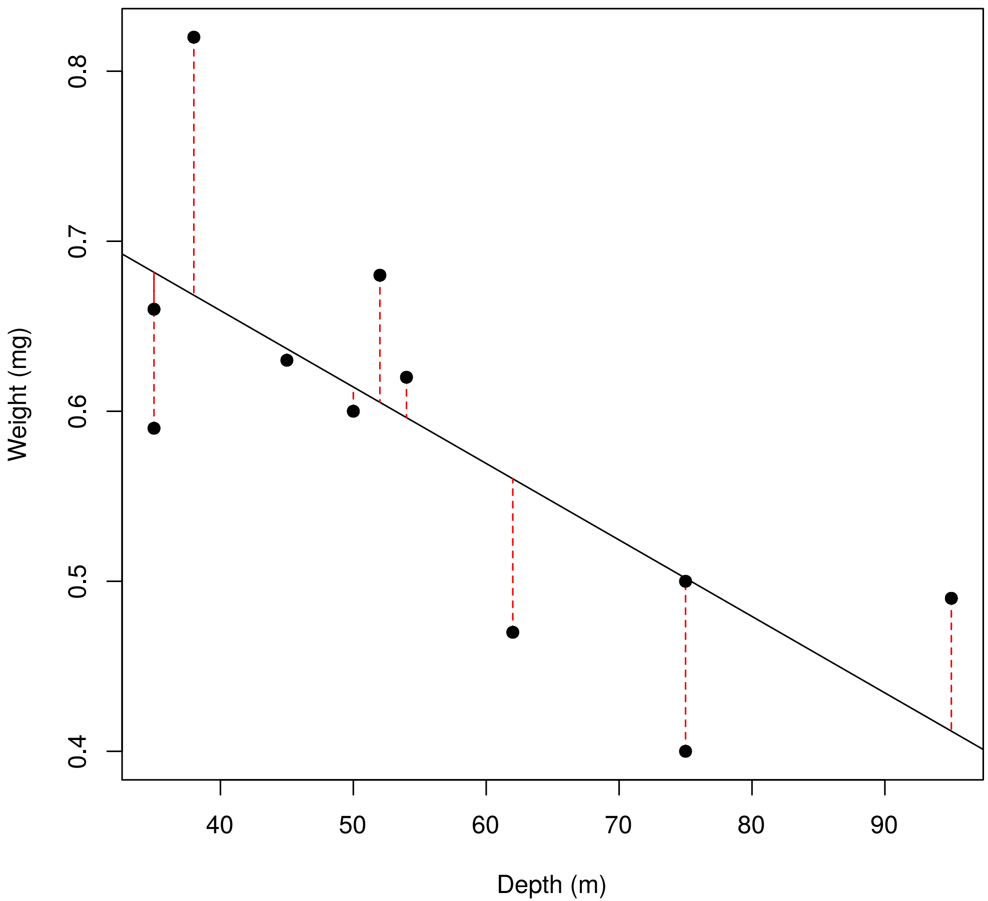
Least squares estimation

- The values of β_0 and β_1 that minimize $SS(\beta_0, \beta_1)$ are given by
    - $\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
    - $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$

Least squares estimation

- The values of β_0 and β_1 that minimize $SS(\beta_0, \beta_1)$ are given by
  - $\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
  - $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$
- These are called the LS estimators of β_0 and β_1
- Under the usual assumptions for the SLR model (normality not required), the LS estimators:
  - Are **unbiased** estimators of β_0 and β_1
  - Have **minimum variance** among all *linear* unbiased estimators of β_0 and β_1!
- How do we interpret $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for a fitted SLR model?

Least squares estimation



The SLR model

- The SLR model is belongs to a broad class of models called *linear models* (LMs)
    - In an LM, the response is a **linear function of the coefficients**
    - Later on we'll see how to deal with nonlinear models where the response is not linearly related to the model parameters
- The classic two-sample t-test and ANOVA are linear models where the predictors are indicators for the levels of the factors involved
- In the R software, the `lm()` function can be used to fit regression models
    - Illustration `shrimp.R`

The SLR model

- The SLR model is belongs to a broad class of models called *linear models* (LMs)
  - In an LM, the response is a **linear function of the coefficients**
  - Later on we'll see how to deal with nonlinear models where the response is not linearly related to the model parameters
- The classic two-sample t-test and ANOVA are linear models where the predictors are indicators for the levels of the factors involved
- In the R software, the `lm()` function can be used to fit regression models
  - Illustration `shrimp.R`
- For the shrimp example, we have $\hat{y} = 0.839135 - 0.004498x$, where $\hat{y}$ is the *predicted value* of y

Inference in the SLR model

- Typically, the parameter of primary interest in SLR is the slope, β_1
- The slope measures the average rate of change in y relative to x
- Occasionally, interest also lies in the y-intercept, β_0, but usually only in cases where x values are collected **near the origin**
- Otherwise, the y-intercept may not have any practical meaning
- For the shrimp example, the estimate slope is $\widehat{\beta}_1 = -0.0045$ (**how do we interpret this number?**)
- In SLR, it is natural to ask whether or not the **slope differs significantly from zero**.
    - If the slope equals zero and the model is correctly specified, then y will not depend on x (i.e., a horizontal regression line)
    - If the relation is quadratic, then one could fit a straight line and get an estimated slope near zero which could be very misleading (**always plot your data**)

Inference in the SLR model

- If $\epsilon \sim N\left(0, \sigma^2\right)$, then $\widehat{\beta}_1 \sim N\left(\beta, \sigma^2_{\beta_1}\right)$ **(why?)**
- The formulas for $\widehat{SE}\left(\widehat{\beta_0}\right)$ and $\widehat{SE}\left(\widehat{\beta_1}\right)$ are messy, but are provided my most statistical software
  - In R, these are located in the column labeled `Std. Error` after applying the `summary()` function (e.g., `summary(slr)`)

Inference in the SLR model

```
summary(slr)
```

```
##
## Call:
## lm(formula = weight ~ depth)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.101819 -0.056004 -0.006746  0.049235  0.151772
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.839135   0.081394  10.310 2.77e-06 ***
## depth       -0.004498   0.001382  -3.254  0.00993 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08363 on 9 degrees of freedom
## Multiple R-squared:  0.5406,    Adjusted R-squared:  0.4896
## F-statistic: 10.59 on 1 and 9 DF,  p-value: 0.009926
```

Inference in the SLR model

- In SLR, one can test the following hypothesis:

$$H_0 : \beta_1 = \beta_{10} \quad vs. \quad H_1 : \beta_1 \neq \beta_{10}$$

  - Typically, $\beta_{10} = 0$

- If the null hypothesis is true, then

$$t_{obs} = \frac{\widehat{\beta}_1 - \beta_{10}}{\widehat{SE}\left(\widehat{\beta_1}\right)}$$

  will have a t-distribution with $n-2$ degrees of freedom

- From this result,we can conduct various tests of hypotheses regarding β_1 and construct $100\,(1-\alpha)$% confidence intervals for β_1

Inference in the SLR model

- In SLR, one can test the following hypothesis:

$$H_0 : \beta_1 = \beta_{10} \quad vs. \quad H_1 : \beta_1 \neq \beta_{10}$$

  - Typically, $\beta_{10} = 0$

- If the null hypothesis is true, then

$$t_{obs} = \frac{\widehat{\beta}_1 - \beta_{10}}{\widehat{SE}\left(\widehat{\beta_1}\right)}$$

  will have a t-distribution with $n-2$ degrees of freedom

- From this result,we can conduct various tests of hypotheses regarding β_1 and construct $100\,(1-\alpha)$% confidence intervals for β_1

- Similar results hold for the y-intercept β_0

The coefficient of determination

- One of the most important statistics in regression analysis is the coefficient of determination, better known as the R^2 ("R-squared")

- In the SLR model, R^2 is just the square of the (Pearson) correlation between x and y

- In MLR (i.e., when we have more than one predictor), a more general definition of R^2 is required

- Illustration: `shrip.R`

- The value of R^2 is always between zero and one and **represents the proportion of variability in the response that is explained by the regression model**

- **Question:** Interpret R^2 from the SLR model fit to the shrimp data

Assessing the SLR fit

- TBD.