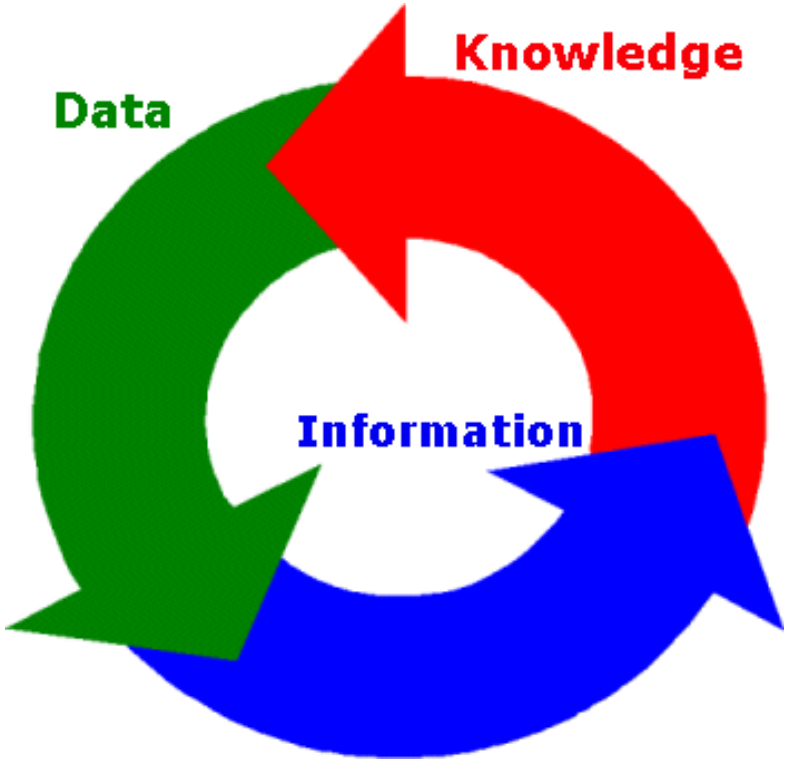
Linear Models

Developing and Deciphering

Kristan Reed, PhD
Assistant Professor of Dairy Cattle Nutrition
Cornell University

The Task

The Task



W Regression?

Describes the average, systematic response of the <u>dependent</u> variable to the <u>independent</u> variables *AND* the expected degree of variation of the dependent variable for a given set of independent variables

W Regression?

Describes the average systematic response of the <u>dependent</u> variable to the <u>independent</u> variables *AND* the expected degree of variation of the dependent variable for a given set of independent variables

(in)dependent = (In)puts → dependent = Outputs

Linear Models

What's the objective?

Before beginning, it's important to know how the model results will be used.

Linear Models

What's the objective?

Before beginning, it's important to know how the model results will be used.

- <u>Description</u>
  - How does the response vary w.r.t. the inputs?
- <u>Prediction</u>
  - Given a set of input conditions, what is the expected outcome? How much certainty is there in this prediction?

Linear Models

What's the objective?

Before beginning, it's important to know how the model results will be used.

Linear Models

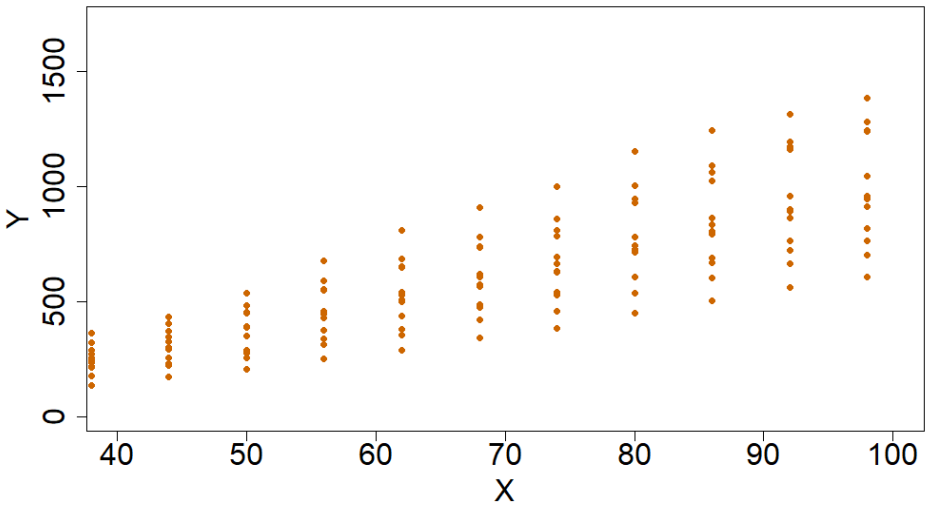
What's the Objective?

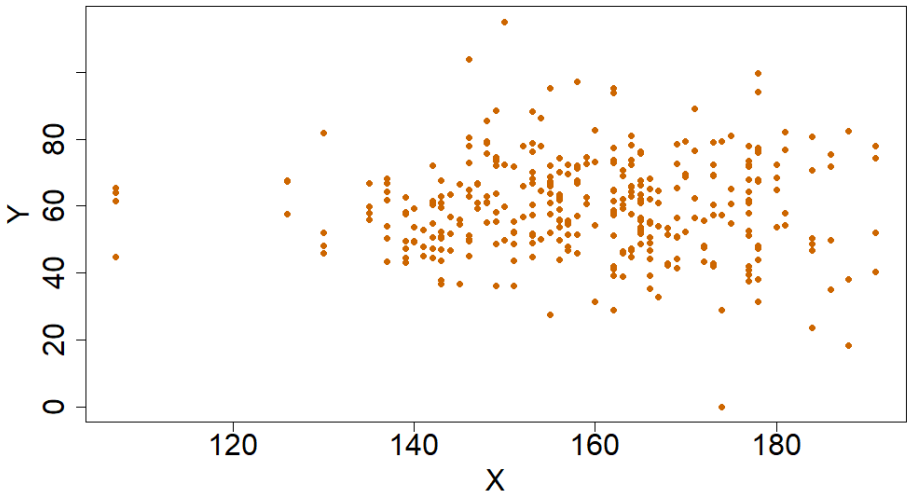
Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

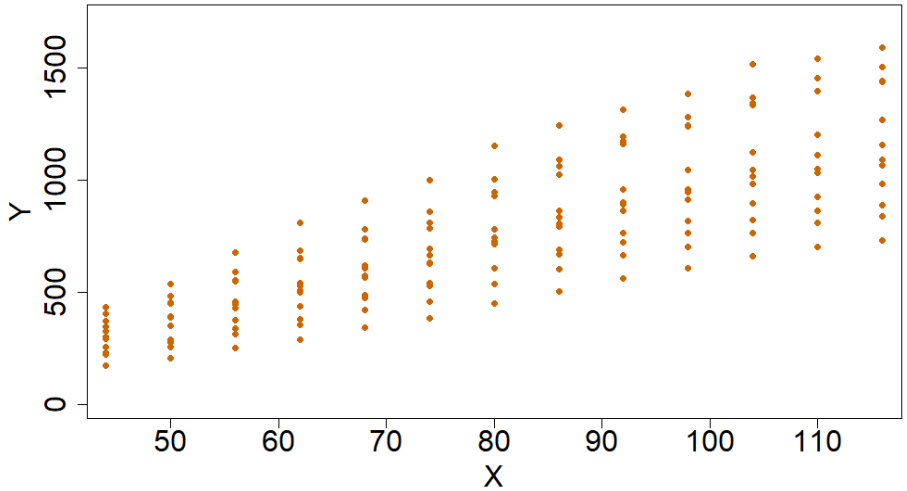
Data Exploration

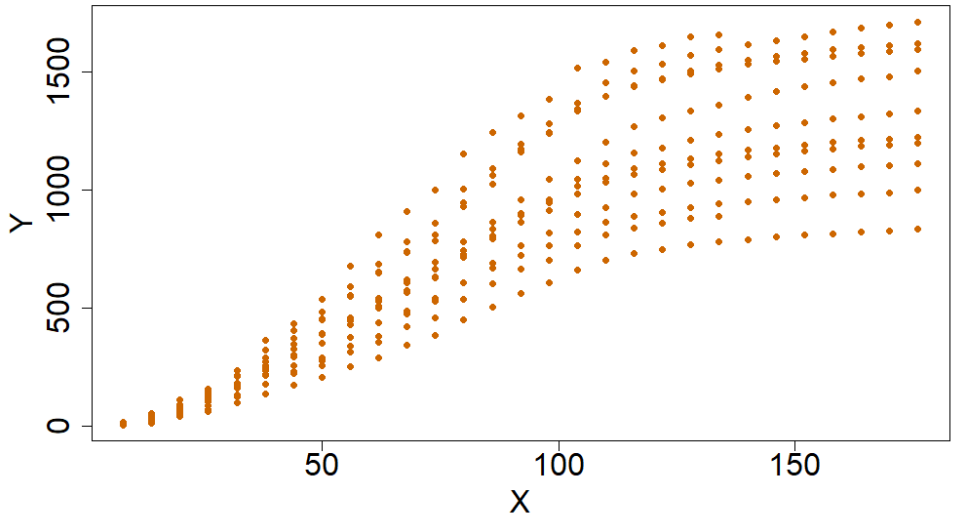




Visually assess plots of the dependent variable vs. independent variables

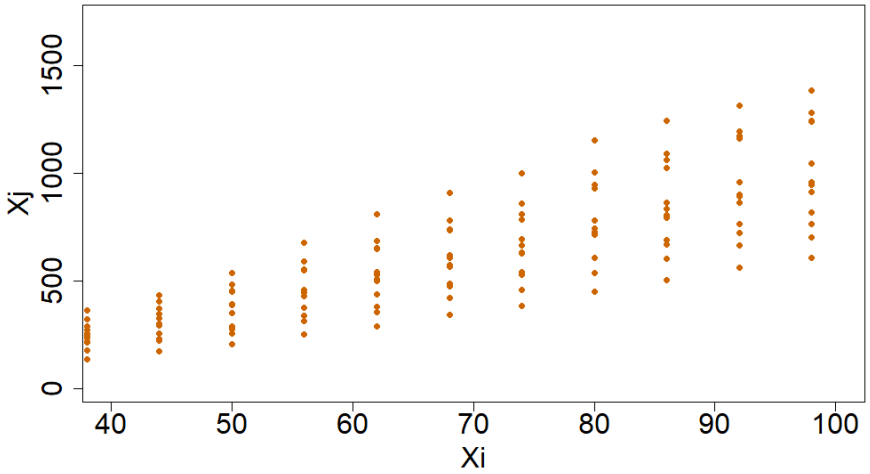
Data Exploration

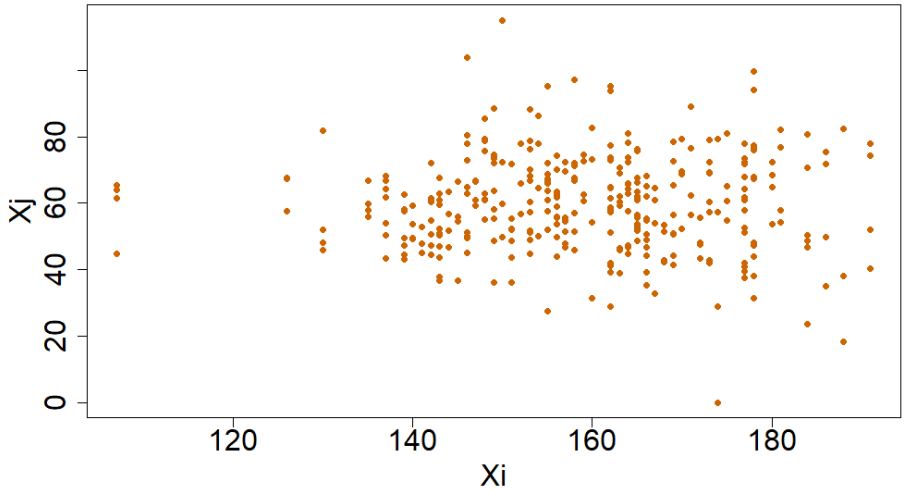




Be sure to assess the range of your independent variable and how it relates to the system studied and your desired inference

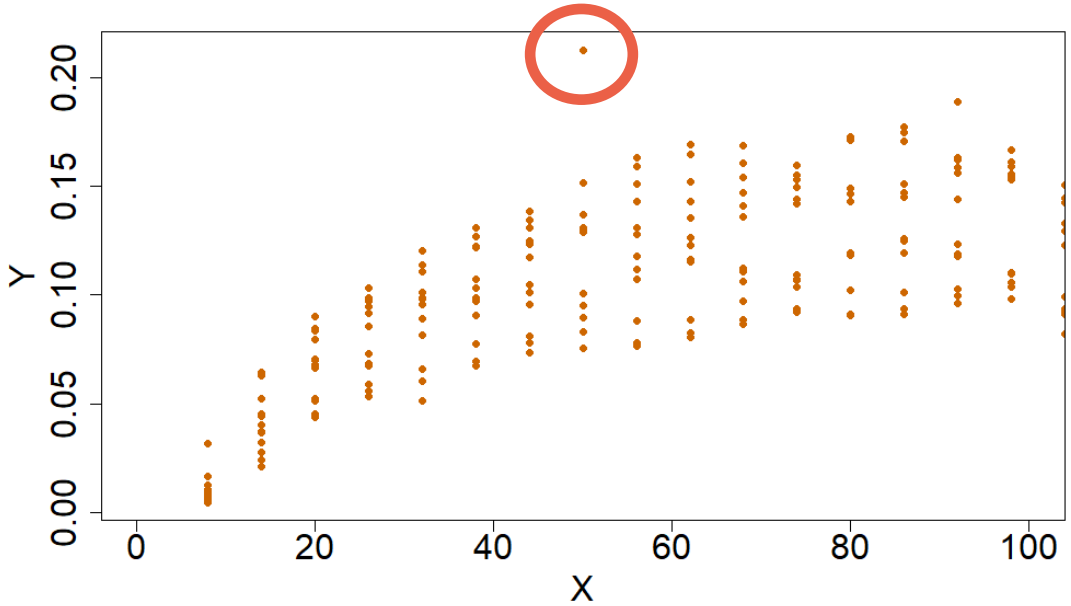
Data Exploration





Plots of dependent variables against each other are an important step in identifying highly correlated predictor variables

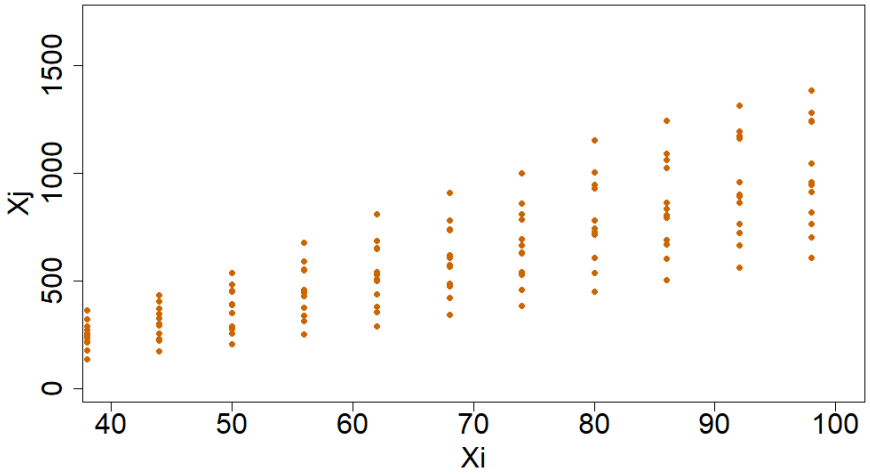
Data Exploration

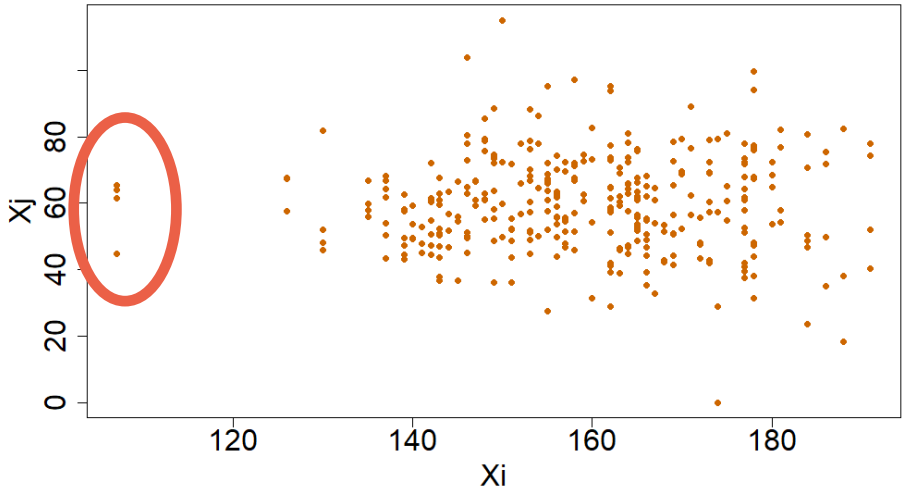


Visually look for outliers and remove if:

1) Meets statistical criteria (i.e. more than 2-3 SD from the mean)
2) A review of notes/procedure identifies an abnormality in the collection process

Data Exploration





- Plots of dependent variables against each other are an important step in identifying highly correlated predictor variables
- Extreme values in independent variables, like the plot on the left should not be removed

Linear Models

What's the Objective?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

- Plot, plot, plot!
  - visually assess the relationships between variables and to look for potential outliers.
- Does the plot reveal a well-defined relationship? Is this relationship linear?
- Are the ranges and precisions of the independent variables sufficient for your purposes?

Linear Models

What's the Objective?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

Model Fitting

Find the 'best' numerical relationship between the dependent and independent variables.

Linear Models (Simple Regression)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Y_i is the value of the response variable for the ith observation
- β_0 and β_1 are the parameters
- X_i is the known constant value of the predictor variable for the ith observation
- ε_i is the error for the ith observation

Situation

The Regression parameters are unknown

B_0 is the y-intercept
B_1 is the slope

Assumptions

- The expectation of Y_i is the mean of the regression equation
- The errors are uncorrelated and have a constant variance

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

$$E[\varepsilon_i] = 0$$

$$\sigma^2\{\varepsilon_i\} = \sigma^2$$

Objectives

To find the 'good' values that can serve as estimates for β_0, β_1, and ε_i

Linear Models (Multiple Regression)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Y_i is the value of the response variable for the ith record $(i = 1, ..., n)$
- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
- $X_{i1}, X_{i2}, \ldots, X_{i,p-1}$ are known predictor variables
- ε_i is the error for the ith observation

Situation

The Regression parameters are unknown

B_0 is the y-intercept
B_i are the slopes

Assumptions

-The expectation of Y_i is the mean of the regression equation

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

$$E[\varepsilon_i] = 0$$

-The errors are uncorrelated and have a constant variance

$$\sigma^2\{\varepsilon_i\} = \sigma^2$$

Objectives

To find the 'good' values that can serve as estimates for β_0, β_i, and ε_i

Linear Models (Matrix Notation)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Situation

The Regression parameters are unknown

B is the vector of parameter estimates

Assumptions

-The expectation of Y_i is the mean of the regression equation

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

$$E[\boldsymbol{\varepsilon}] = \mathbf{0}$$

-The errors are uncorrelated and have a constant variance

$$\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\} = \sigma^2\mathbf{I}_{n\times n}$$

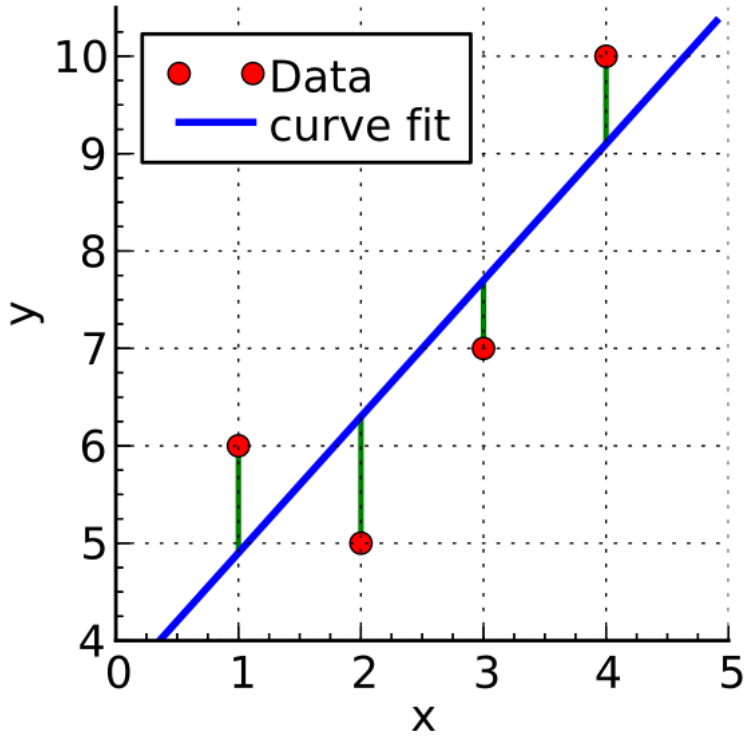
Objectives

To find 'good' values that can serve as estimates for β and ε

How do we find ‘good’ estimates?

Least Squares Estimation

- Minimizes the vertical distance of the observed responses to the regression line



Notation convention…

- β_0, β_1, β_2,… etc. are unknown
- Estimators from a sample are: b_0, b_1, b_2,… etc.

- ε_i and σ^2 are unknown
- Estimators from a sample are e_i and s^2

Least Squares Estimation

- Minimizes the vertical distance of the observed responses to the regression line

$$Q = \sum_{i}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

- To find the solution…
  - Take the derivative w.r.t. the regression parameters

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Least Squares Estimation

- Minimizes the vertical distance of the observed responses to the regression line

$$Q = \sum_{i}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

- To find the solution…
  - Take the derivative w.r.t. the regression parameters
  - Set equal to zero and solve for b_0 and b_1

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{Y}\bar{X}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i\right) = \bar{Y} - b_1\bar{X}$$

Least Squares Estimation

- The residuals are used to create an estimate of the variance s^2

$$s^2 = MSE = \frac{SSE}{n-p}$$

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$E[MSE] = \sigma^2$$

Least Squares Estimation - Properties

- Guarantees estimators are **unbiased with minimal variance**
  - (*does not require assumption of normality*)

$$E[b_0] = \beta_0 \text{ and } E[b_1] = \beta_1$$

$$E[MSE] = \sigma^2$$

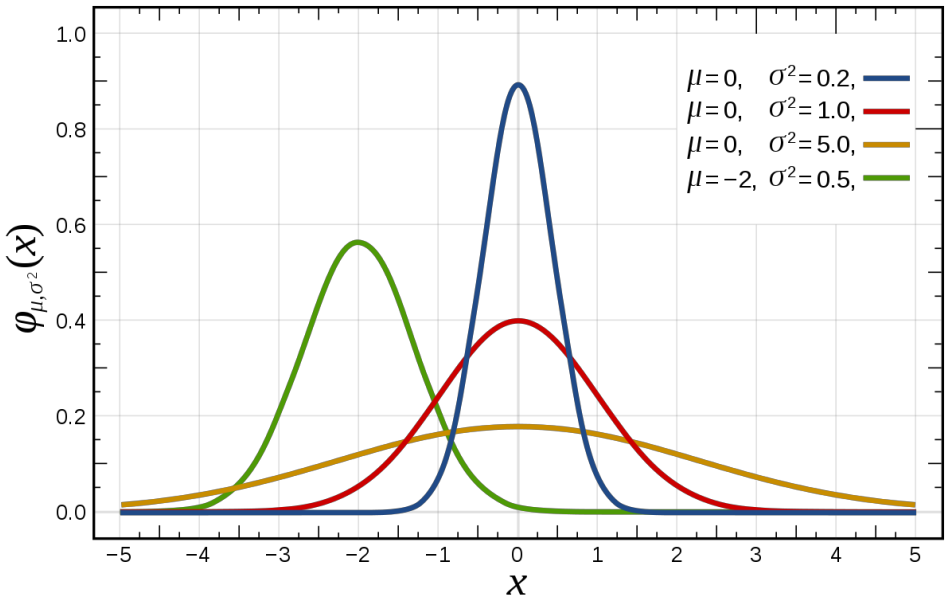
What if we want to make inferences about the significance of the estimates?

Or about the distribution of the residuals?

Normal Regression (Maximum Likelihood)

- The data are assumed to be a random selection from a normal population

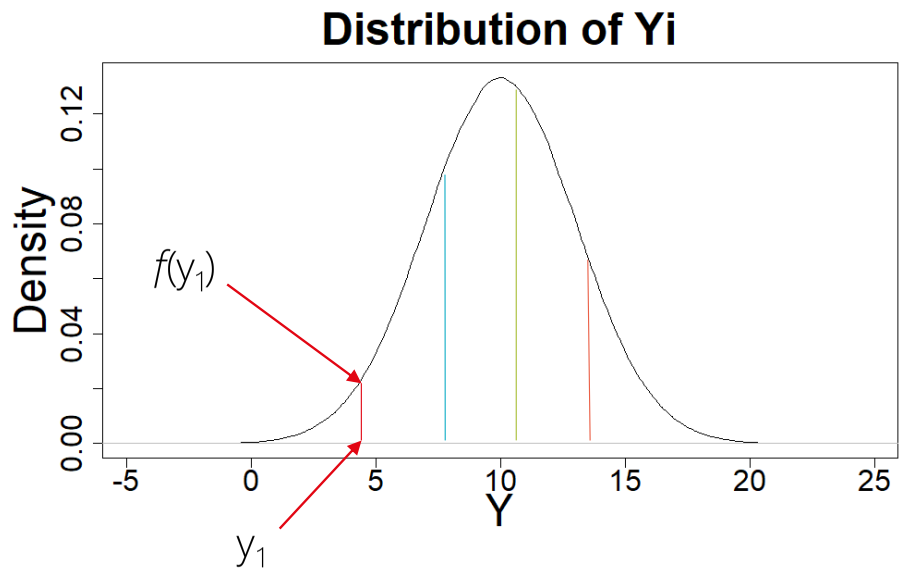
- The error terms are now assumed to be normally distributed
  - (i.e. that $\varepsilon_i \sim N(0, \sigma^2)$)



$$f(t) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]$$

Normal Regression (Maximum Likelihood)

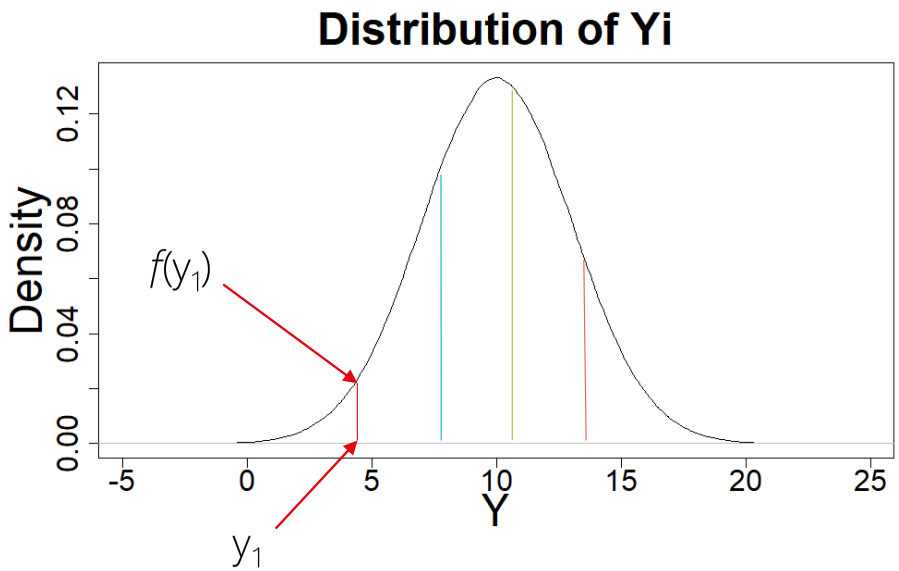
- The task of MLE estimation is to find estimators for β and σ^2 that maximize the normal density function for all y_i



$$f(y_i \mid \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(y_i - \boldsymbol{X\beta})^2}{2\sigma^2}\right]}$$

Normal Regression (Maximum Likelihood)

- The MLE estimators for β are the same as the least squares estimators
  - Note the similarity to Least Squares through inclusion of the negative squared distance in the exponential
- The MLE estimator for σ^2 is almost the same as the least squares standard error estimator
  - For large samples sizes they are essentially equivalent



$$f(y_i \mid \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(y_i - \boldsymbol{X\beta})^2}{2\sigma^2}\right]}$$

Normal Regression (Maximum Likelihood)

Assumptions - Errors

- Errors are uncorrelated and have a constant variance
- Errors are independent and distributed normally with a mean of "0" and a variance of σ^2

$$E[\varepsilon_i] = 0$$
$$\sigma^2\{\varepsilon_i\} = \sigma^2$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

Normal Regression (Maximum Likelihood)

Assumptions - Parameters

- The distribution assumption for the data carries over to the sampling distribution of the parameter estimates
- "The sampling distribution of b_1 refers to the different values of b_1 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample." ~ Kutner et al., 2004

$$b_0 \sim N\left(\beta_0,\ \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\right)$$

- $\dfrac{b_0 - \beta_0}{s\{b_0\}} \sim t_{n-2}$

$$b_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

- $\dfrac{b_1 - \beta_1}{s\{b_1\}} \sim t_{n-2}$

Normal Regression (Maximum Likelihood)

Assumptions - Parameters

- The distribution assumptions for the parameters enable hypothesis testing about parameter significance

$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

- $t^* = \dfrac{b_1}{s\{b_1\}}$
- If $|t^*| \leq t(1-\alpha/2; n-2) \Rightarrow$ conclude H_0
- If $|t^*| > t(1-\alpha/2; n-2) \Rightarrow$ conclude H_a

$$s^2\{b_1\} = \frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2}\right]$$

Beef Example

Description of Data

- 30 bull calves from 3 breeds were grazed on winter wheat or rye
- They were slaughtered around 18 mos and detailed carcass quality measurements were taken

Lots of data and measurements

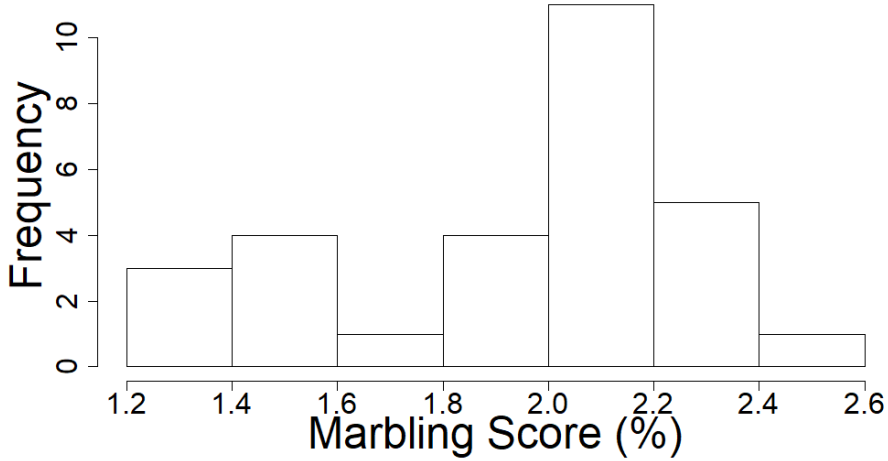
```
> names(beef)
 [1] "ID"              "SlaughterGrp"    "BirthDate"
 [4] "SlaughterDate"   "DaysAtSlaughter" "Group"
 [7] "Forage"          "Breed"           "Live.Wt"
[10] "Carcass.Wt"      "Dressing."       "Marbling.Score"
[13] "Backfat"         "Ribeye.Area"     "KPH"
[16] "Yield.Grade"     "Grade"           "BCS"
[19] "Protein"         "SFA"             "MFA"
[22] "CCPUFA"          "TFA"             "omega3"
[25] "omega6"          "X6_3_ratio"
```

Heins, et al. 2017. Organic Beef Data from Integration of Crops and Livestock Project.

Beef Example

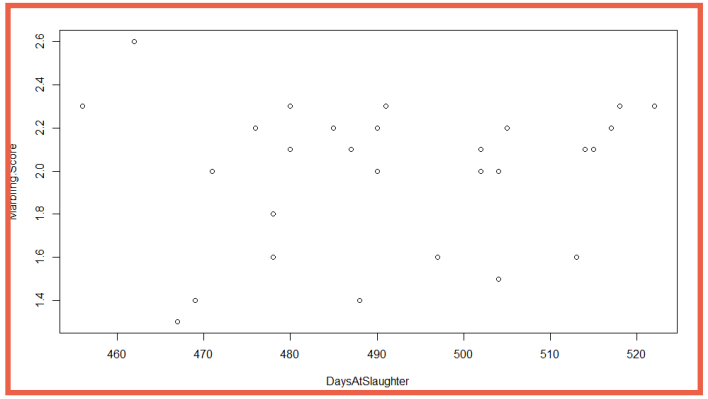
<u>Objective:</u>

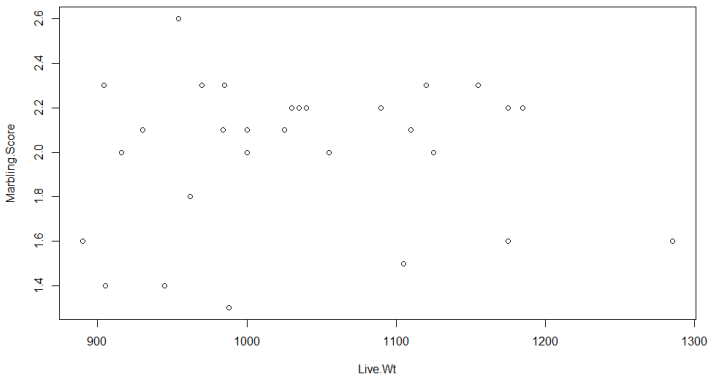
Develop a model to predict marbling score

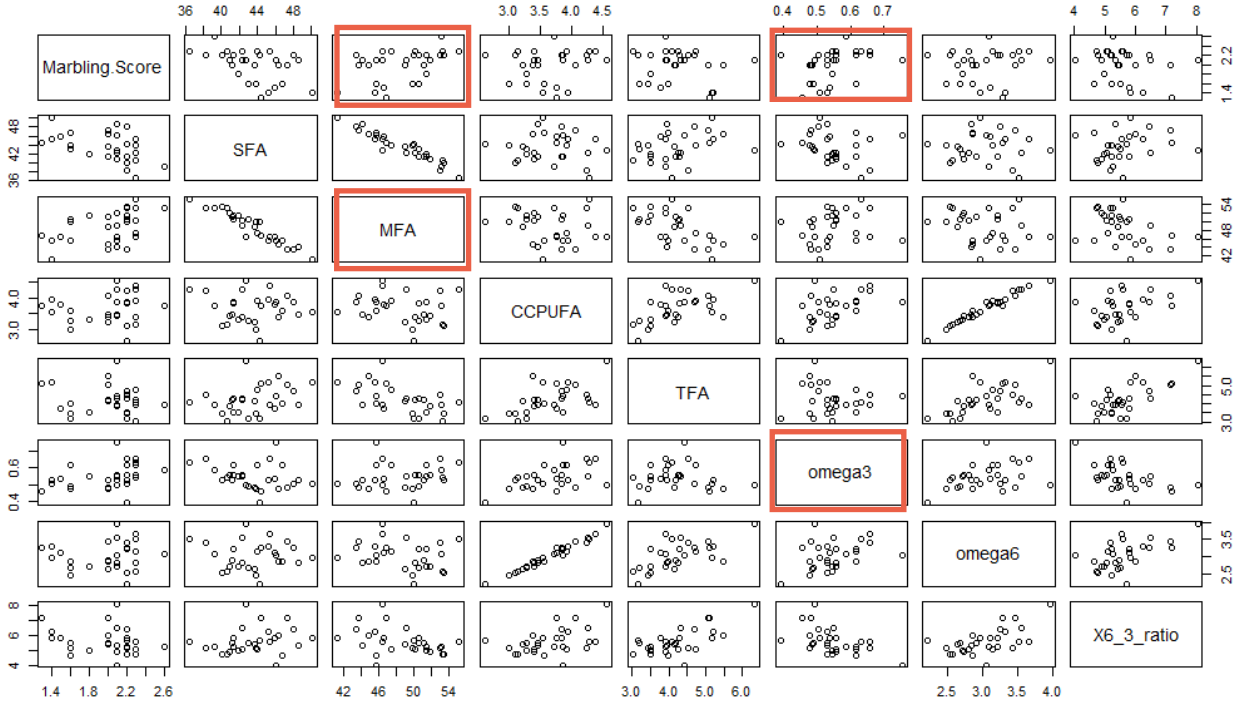


Heins, et al. 2017. Organic Beef Data from Integration of Crops and Livestock Project.

Beef Example








```
pairs(beef[, c(12,20:26)])
```

Heins, et al. 2017. Organic Beef Data from Integration of Crops and Livestock Project.

Beef Example

```
fit1 = lm(Marbling.Score ~ DaysAtSlaughter + MFA + omega3, data = beef)
```

Model Specification

```
> summary(fit1)

Call:
lm(formula = Marbling.Score ~ DaysAtSlaughter + MFA + omega3,
    data = beef)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47407 -0.23739  0.03711  0.23037  0.43549

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.347470   1.735404  -0.776   0.4448
DaysAtSlaughter  0.001775   0.003103   0.572   0.5724
MFA              0.036861   0.015931   2.314   0.0292 *
omega3           1.250162   0.767690   1.628   0.1160
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2968 on 25 degrees of freedom
Multiple R-squared:  0.285,     Adjusted R-squared:  0.1992
F-statistic: 3.321 on 3 and 25 DF,  p-value: 0.03597
```

Heins, et al. 2017. Organic Beef Data from Integration of Crops and Livestock Project.

Linear Models

What's the Objective?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

Model Fitting

Find the 'best' numerical relationship between the dependent and independent variables.

- For simple and multiple linear models, the important choice in model fitting is between Least Squares and MLE

Linear Models

What's the Objective?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

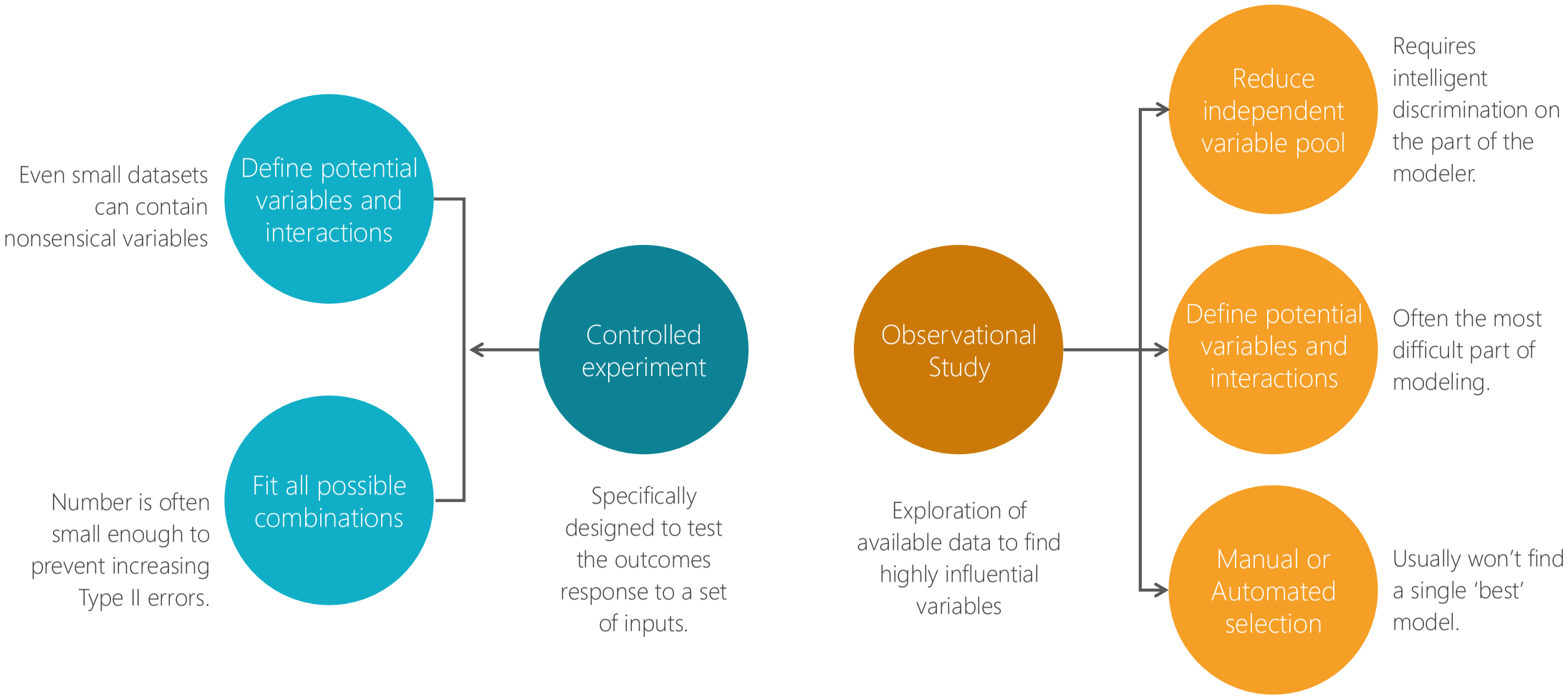
Model Fitting

Find the 'best' numerical relationship between the dependent and independent variables.

Model Selection

How do you know which dependent variables should remain in the model?

How was your data collected?



Criteria for selection

- R^2 or Adjusted R^2
  - Adding more predictor variables will always increase R^2
  - Adjusted R^2 corrects for the number of parameters w.r.t the number of observations

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

$$\text{Adjusted } R^2 : R_a^2 = \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}}$$

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Criteria for selection

- R^2 + Adjusted R^2
- AIC or BIC
  - Akaike Information Criterion
  - Bayesian Information Criterion
  - Estimates the 'amount of information lost' by summarizing the data with a model instead of using the exact data
  - Lower loss of information = Better
  - BIC has a stronger penalty for adding more parameters

$$\hat{L} = \prod_{i=1}^{n} f(y_i | \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(y_i - \boldsymbol{X\beta})^2}{2\sigma^2}\right]}$$

$$\mathrm{AIC} = 2\mathrm{p} - 2\ln(\hat{L})$$

$$\mathrm{BIC} = \ln(n)\mathrm{p} - 2\ln(\hat{L})$$

Criteria for selection

- R^2 + Adjusted R^2
- AIC or BIC
- MSPE – An assessment of prediction error.
  - This metric can also be decomposed to assess mean bias (ECT) and slope bias (ER) as described by Bibby and Toutenberg (1977).

$$\mathrm{MSPE} = \frac{\sum_{i=1}^{n}\left(\mathrm{O}_i - \widehat{\mathrm{P}_i}\right)^2}{n}$$

$$s_p^2 = \frac{\sum_{i=1}^{n}(\mathrm{P}_i - \overline{\mathrm{P}_i})^2}{n} \qquad s_a^2 = \frac{\sum_{i=1}^{n}(\mathrm{O}_i - \overline{\mathrm{O}_i})^2}{n}$$

$$r = \left(\frac{\sum_{i=1}^{n}(\mathrm{P}_i - \overline{\mathrm{P}_i})}{n}\right)\left(\frac{\sum_{i=1}^{n}(\mathrm{O}_i - \overline{\mathrm{O}_i})}{n}\right)\Big/\left(\sqrt{s_p^2}\sqrt{s_a^2}\right)$$

$$\mathrm{ECT} = \frac{(\overline{\mathrm{P}} - \overline{\mathrm{O}})^2}{\mathrm{MSPE}} \qquad \mathrm{ER} = \frac{\left(\sqrt{s_p^2} - r\sqrt{s_a^2}\right)^2}{\mathrm{MSPE}}$$

O = Observed; P = Predicted

Criteria for selection

- R^2 + Adjusted R^2
- AIC or BIC
- MSPE
- Mallow's CP- another metric with a strong correction for the number of parameters

$$C_p = \frac{SSE_p}{S^2} - n + 2p$$

$$SSE_p = \sum_{i=1}^{n} (y_i - \widehat{y.p}_i)^2$$

$$S^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y.full}_i\right)^2$$

Criteria for selection

- R^2 + Adjusted R^2
- AIC or BIC
- MSPE
- Mallow's CP
- LRT (likelihood ratio tests)
  - For testing nested models
  - $2\ln\left(\widehat{L_2}\,/\widehat{L_1}\right)$
  - Follows an F-distribution

```
lm(formula = Marbling.Score ~ DaysAtSlaughter + MFA + omega3,
    data = beef)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47407 -0.23739  0.03711  0.23037  0.43549

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.347470   1.735404  -0.776   0.4448
DaysAtSlaughter 0.001775   0.003103   0.572   0.5724
MFA             0.036861   0.015931   2.314   0.0292 *
omega3          1.250162   0.767690   1.628   0.1160
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2968 on 25 degrees of freedom
Multiple R-squared:  0.285,     Adjusted R-squared:  0.1992
F-statistic: 3.321 on 3 and 25 DF,  p-value: 0.03597
```

```
> anova(fit2, fit1)
Analysis of Variance Table

Model 1: Marbling.Score ~ DaysAtSlaughter + MFA
Model 2: Marbling.Score ~ DaysAtSlaughter + MFA + omega3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     26 2.4358
2     25 2.2022  1   0.23361 2.6519  0.116
```

Forward Stepwise Regression

- An algorithm to select models by starting with the smallest model and progressively adding independent variables until additional variables do not improve the model
  1. Add independent variable with the smallest P-value, or model with lowest AIC/BIC
  2. Add another independent variable with the next smallest P-value, or lowest AIC/BIC
  3. Test previously added variables to see if they are still needed based on a removal alpha criteria
  4. Repeat 2-3 until adding a new variable no longer improves the model

- A drawback of this method is that is identifies a single 'best' model but there is often not a truly 'best' model because different models will perform differently based on different criteria and diagnostics

Forward Stepwise Regression: Beef Example

Full Model:

Marbling Score ~ AgeatSlaughter + MFA + Omega3 + ForageType + Breed + Carcass.Wt + Dressing% + BCS + Backfat + RibeyeArea + TFA

R package – olsrr

- Can select on p-value or AIC

Selection with p-value

```
> ols_step_forward_p(fullfit)
Forward Selection Method
---------------------------

Candidate Terms:

1. DaysAtSlaughter
2. MFA
3. omega3
4. Forage
5. Breed
6. Carcass.Wt
7. Dressing.
8. BCS
9. Backfat
10. Ribeye.Area
11. TFA

We are selecting variables based on p value...

Variables Entered:

✔ Dressing.
✔ MFA
✔ omega3
✔ DaysAtSlaughter
✔ TFA
✔ Backfat
```

Forward Stepwise Regression: Beef Example

Selection with p-value

```
> ols_step_forward_p(fullfit)
Forward Selection Method
---------------------------

Candidate Terms:

1. DaysAtSlaughter
2. MFA
3. omega3
4. Forage
5. Breed
6. Carcass.Wt
7. Dressing.
8. BCS
9. Backfat
10. Ribeye.Area
11. TFA

We are selecting variables based on p value...

Variables Entered:

✔ Dressing.
✔ MFA
✔ omega3
✔ DaysAtSlaughter
✔ TFA
✔ Backfat
```

Forward Stepwise Regression: Beef Example

Selection with p-value

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Dressing.	0.2681	0.2410	3.1107	14.2173	0.2889
2	MFA	0.3681	0.3195	1.2718	11.9592	0.2736
3	omega3	0.4325	0.3644	0.7982	10.8420	0.2644
4	DaysAtSlaughter	0.4907	0.4058	0.5622	9.7033	0.2557
5	TFA	0.5190	0.4144	1.4765	10.0475	0.2538
6	Backfat	0.5454	0.4214	2.4615	10.4089	0.2523

Selection with AIC

Selection Summary

Variable	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
Dressing.	14.217	0.826	2.254	0.26813	0.24103
MFA	11.959	1.134	1.946	0.36808	0.31947
omega3	10.842	1.332	1.748	0.43248	0.36438
DaysAtSlaughter	9.703	1.511	1.569	0.49069	0.40581

Backward Stepwise Regression

- Similar to the forward selection
  1. Starts with full model with all of the variables in the model
  2. Removes a variable based on largest p-value or AIC
  3. Previously removed variables are tested to see if they should be added back in.
  4. Repeat steps 2-3

Model Selection

Backward Stepwise Regression

Full Model:

Marbling Score ~ AgeAtSlaughter + MFA + Omega3 + ForageType + Breed + Carcass.Wt + Dressing% + BCS + Backfat + RibeyeArea + TFA

R package – olsrr

- Can select on p-value or AIC

```
> ols_step_backward_p(fullfit)
Backward Elimination Method
---------------------------

Candidate Terms:

1 . DaysAtSlaughter
2 . MFA
3 . omega3
4 . Forage
5 . Breed
6 . Carcass.Wt
7 . Dressing.
8 . BCS
9 . Backfat
10 . Ribeye.Area
11 . TFA

We are eliminating variables based on p value...

Variables Removed:

✖ Breed
✖ Carcass.Wt
✖ omega3
✖ Forage
✖ Ribeye.Area
✖ BCS

No more variables satisfy the condition of p value = 0.3
```

Model Selection

Backward Stepwise Regression

Selection with p-value

```
> ols_step_backward_p(fullfit)
Backward Elimination Method
---------------------------

Candidate Terms:

1 . DaysAtSlaughter
2 . MFA
3 . omega3
4 . Forage
5 . Breed
6 . Carcass.Wt
7 . Dressing.
8 . BCS
9 . Backfat
10 . Ribeye.Area
11 . TFA

We are eliminating variables based on p value...

Variables Removed:

✖ Breed
✖ Carcass.Wt
✖ omega3
✖ Forage
✖ Ribeye.Area
✖ BCS

No more variables satisfy the condition of p value = 0.3
```

Selection with AIC

```
> ols_step_backward_aic(fullfit)
Backward Elimination Method
---------------------------

Candidate Terms:

1 . DaysAtSlaughter
2 . MFA
3 . omega3
4 . Forage
5 . Breed
6 . Carcass.Wt
7 . Dressing.
8 . BCS
9 . Backfat
10 . Ribeye.Area
11 . TFA


Variables Removed:

✖ Breed
✖ Carcass.Wt
✖ omega3
✖ Forage
✖ Ribeye.Area
✖ BCS

No more variables to be removed.
```

Beef Example: Forward and Backward

<u>Forward Selection</u>

Marbling Percent ~
- Age at Slaughter +
- MFA +
- Dressing percent +
- Backfat +
- TFA +
- Omega 3

<u>Backward Selection</u>

Marbling Percent ~
- Age at Slaughter +
- MFA +
- Dressing percent +
- Backfat +
- TFA

Beef Example: Forward and Backward

Forward Selection

```
Call:
lm(formula = Marbling.Score ~ Dressing. + MFA + omega3 + DaysAtSlaughter +
    TFA + Backfat, data = beef)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65154 -0.17202  0.09389  0.12840  0.39472

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8.224269   2.523414  -3.259  0.00359 **
Dressing.        9.276673   3.778174   2.455  0.02245 *
MFA              0.034901   0.016777   2.080  0.04936 *
omega3           0.571822   0.791551   0.722  0.47765
DaysAtSlaughter  0.005753   0.003141   1.832  0.08057 .
TFA              0.106287   0.081098   1.311  0.20351
Backfat          4.172663   3.689820   1.131  0.27029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2523 on 22 degrees of freedom
Multiple R-squared:  0.5454,    Adjusted R-squared:  0.4214
F-statistic: 4.399 on 6 and 22 DF,  p-value: 0.004565
```

Backward Selection

```
Call:
lm(formula = Marbling.Score ~ Dressing. + MFA + DaysAtSlaughter +
    TFA + Backfat, data = beef)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66721 -0.18319  0.05394  0.15671  0.40892

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8.204793   2.496905  -3.286  0.00324 **
Dressing.        8.732572   3.663656   2.384  0.02579 *
MFA              0.036189   0.016508   2.192  0.03875 *
DaysAtSlaughter  0.006363   0.002993   2.126  0.04447 *
TFA              0.112952   0.079729   1.417  0.16997
Backfat          5.670525   3.020264   1.877  0.07319 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2496 on 23 degrees of freedom
Multiple R-squared:  0.5346,    Adjusted R-squared:  0.4334
F-statistic: 5.284 on 5 and 23 DF,  p-value: 0.002243
```

Adding Omega3 as a predictor increased the SE of the other predictors and is not significant at a P-value < 0.1 , but it was included in both forward selection models. Why?

Independent Variable Correlations

```
> cor(model.matrix(forward.fit)[ , -1])
                 Dressing.        MFA      omega3 DaysAtSlaughter         TFA    Backfat
Dressing.        1.0000000  0.2738693  0.10353798     -0.32104997 -0.37066099  0.5596820
MFA              0.2738693  1.0000000  0.17791438     -0.01704450 -0.57886256  0.2743572
omega3           0.1035380  0.1779144  1.00000000      0.01456279 -0.07532196  0.4804054
DaysAtSlaughter -0.3210500 -0.0170445  0.01456279      1.00000000  0.19322647 -0.4574114
TFA             -0.3706610 -0.5788626 -0.07532196      0.19322647  1.00000000 -0.3559961
Backfat          0.5596820  0.2743572  0.48040541     -0.45741140 -0.35599611  1.0000000
```

Independent Variable Correlations

- When there is a high degree of correlation between independent variables, the coefficient estimates will have larger SEs and the estimates will vary more between samples
- Parameters may not be statistically significant even if there is a statistical relationship

Extreme case is a perfect linear relationship:

$$X_3 = 1.8 + 3.4\ X_2$$

The degree of correlation among independent variables in a model is known as *Multicollinearity*

Multicollinearity

- The <u>Variance Inflation Factor</u> is used to assess the degree of multicollinearity
  - General recommendations are to have a VIF less than 10

$$VIF(b_k) = (1 - R_k^2)^{-1}$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed on the $p - 2$ other predictor variables

```
> vif(fullfit)
                     GVIF Df GVIF^(1/(2*Df))
DaysAtSlaughter   3.505434  1        1.872280
MFA               2.209018  1        1.486277
omega3          100.503739  1       10.025155
Forage            2.104543  1        1.450704
Breed            13.270017  2        1.908613
Carcass.Wt        4.672553  1        2.161609
Dressing.         3.011468  1        1.735358
BCS               2.810502  1        1.676455
Backfat           4.438618  1        2.106803
Ribeye.Area       3.503715  1        1.871821
TFA               5.205457  1        2.281547
Protein           6.120376  1        2.473939
omega6           98.222727  1        9.910738
X6_3_ratio      138.124621  1       11.752643
```

```
> vif(forward.fit)
     Dressing.           MFA        omega3 DaysAtSlaughter           TFA       Backfat
      1.610185      1.585762      1.520016        1.419040      1.693856      2.544841
```

Model Selection Caveats

- Often there is not one “best” model for all criteria and so the ultimate selection is a choice made by the modeler by weighing the difference criteria
- If there is a set of qualitative predictors (i.e. a categorical variable with multiple values) it is best to include all values, even if only one or two is significant. Similarly, if there are interaction or second or terms, the first order term needs to be maintained in the model.

Linear Models

What's the Objective?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

Model Fitting

Find the 'best' numerical relationship between the dependent and independent variables.

Model Selection

How do you know which dependent variables should remain in the model?

All model selection methods still require intelligent direction, intervention, and evaluation

Linear Models

What's the Objective?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

Model Fitting

Find the 'best' numerical relationship between the dependent and independent variables.

Model Selection

How do you know which dependent variables should remain in the model?

Evaluation + Interpretation

How well does the model perform and what can you learn from it?

Prediction of sample data

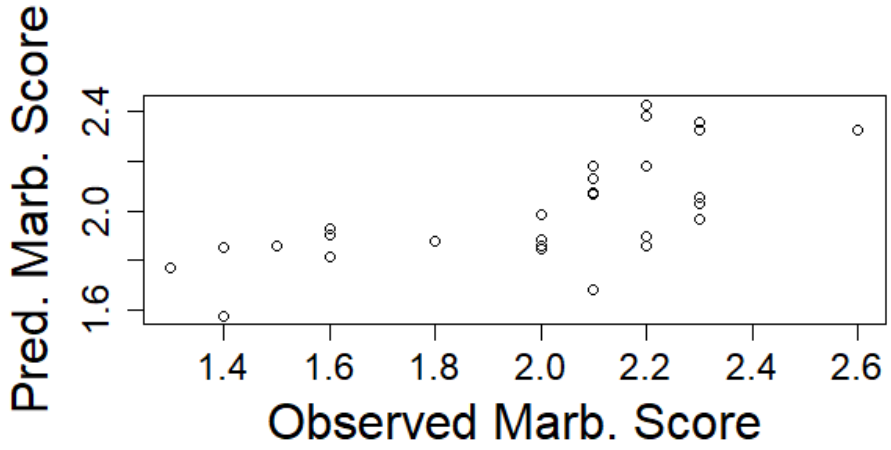
$$\hat{Y}_h = \mathbf{X}_h^{\mathsf{T}}\mathbf{b} = 1b_0 + X_{h1}b_1 + \cdots + X_{h,p-1}b_{p-1}$$

$$\mathbf{s}^2\{Y_h\} = MSE(\mathbf{X}_h^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}_h)$$

Confidence Interval: $\hat{Y}_h \pm t(1-\alpha/2, n-p)s\{\hat{Y}_h\}$

```
> pred = predict(fit3)
> plot(pred ~ beef$Marbling.Score, xlab = "Observed Marb. Score",
+      ylab = "Pred. Marb. Score", cex.axis = 1.5, cex.lab = 2)
```

```
> fit3 = lm(Marbling.Score ~ DaysAtSlaughter + MFA + Backfat + TFA, data=bee
f)
> summary(fit3)

Call:
lm(formula = Marbling.Score ~ DaysAtSlaughter + MFA + Backfat +
    TFA, data = beef)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47104 -0.18183  0.01278  0.24318  0.41460

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.985733   1.925195  -2.070  0.04935 *
DaysAtSlaughter  0.005717   0.003259   1.754  0.09213 .
MFA              0.038341   0.018019   2.128  0.04383 *
Backfat          8.711830   2.992610   2.911  0.00766 **
TFA              0.083155   0.086081   0.966  0.34367
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2729 on 24 degrees of freedom
Multiple R-squared:  0.4196,    Adjusted R-squared:  0.3229
F-statistic: 4.338 on 4 and 24 DF,  p-value: 0.008812
```



Model Evaluation

Objectives	Which Data to Use

• Check Assumptions!
  • Normality
  • Constant Variance
  • Uncorrelated
	• Sample data

• Fit
  • How well does the model describe the data?
  • Is the model biased?
	• Sample data

• Prediction Accuracy • How well does the model predict new data	• New Data • Data splitting • Bootstrapping

Metrics

Plots

- Residual Plots- Against predictor variables
  - Against predicted values
  - Against dependent and other, non-include variables (e.g. time, study, etc.)
  - Boxplots and normality plots of residuals
  - QQ plots
- Plots of observed vs. predicted
- Regression of observed vs. predicted

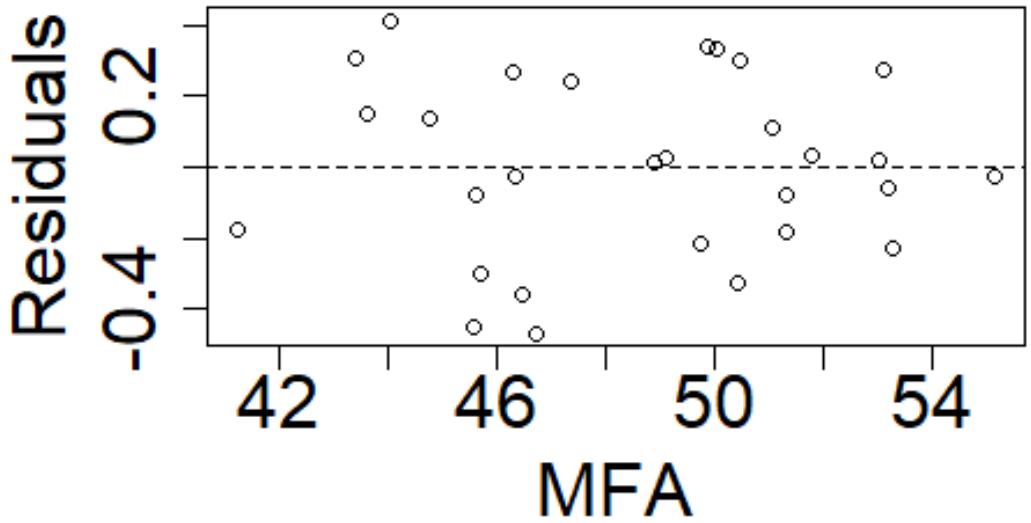
Formal Tests and Metrics

- Formal Tests
  - Shapiro-Wilkes test (normality)
  - Brown-Forsythe (Consistency in Variance)
  - Breusch-Pagan test (Consistency in Variance)
- MSPE/RMSPE
- CCC
- R^2

Beef Example

- Assumptions: Residual plots
  - Look for patterns and distribution between positive and negative values
  - Should be evenly and randomly distributed

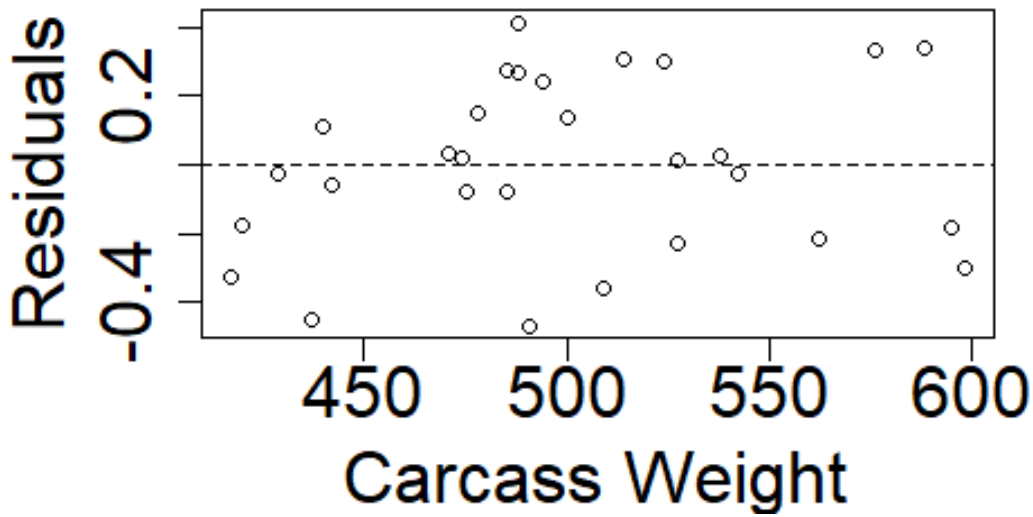
```
#Residuals vs. independent variables
plot(res ~ beef$MFA, ylab = "Residuals", xlab = "MFA", cex.axis = 2, cex.lab = 2)
abline (h = 0, lty = 2)
```



Beef Example

- Assumptions: Residual plots
  - Look for patterns and distribution between positive and negative values
  - Should be evenly and randomly distributed

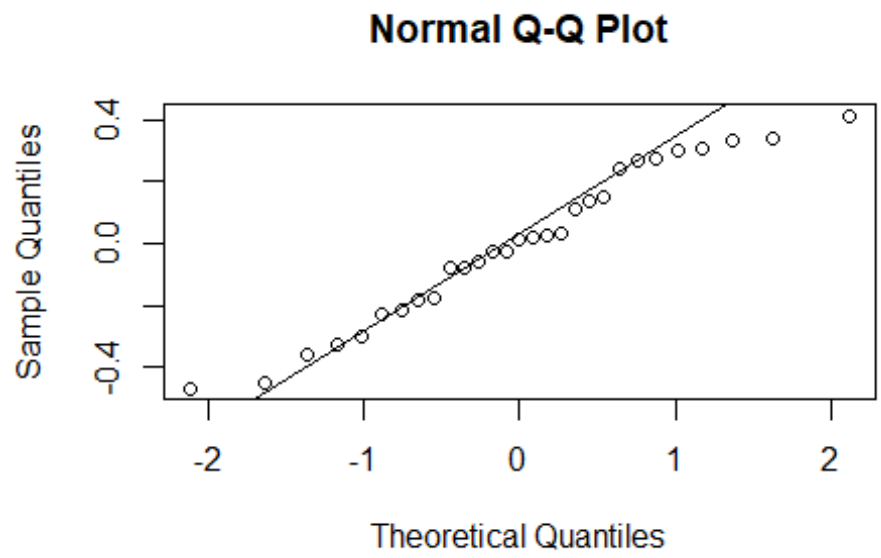
```
#Residuals vs. independent variables
plot(res ~ beef$Carcass.Wt, ylab = "Residuals", xlab = "Carcass Weight",
     cex.axis = 2, cex.lab = 2)
abline (h = 0, lty = 2)
```



Beef Example

- Assumptions: Quantile plots
  - Ordered residuals plotted against their quantile value if they followed a normal distribution exactly
  - Deviations from the line indicate departures from normality

```
## QQ plot
qqplot(qqnorm(res))
qqline(res)
```



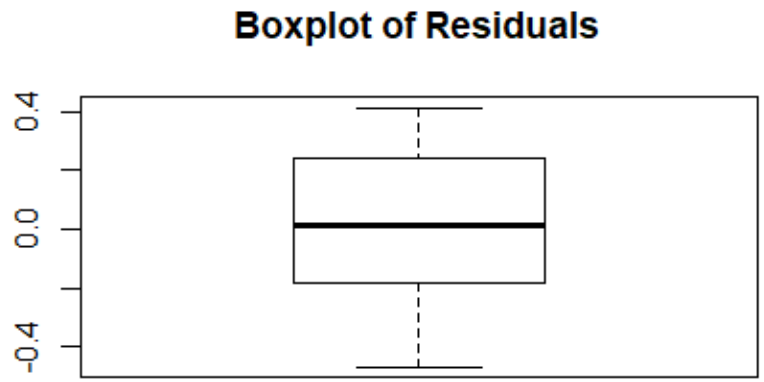
Beef Example

- Assumptions: Shapiro-Wilkes

H_0: Residuals ~ $N(0, \sigma^2)$

*If the null is rejected, it is not certain that the residuals are not normally distributed. Especially with large sample sizes, it is recommended to assess the degree of departure, e.g. with a QQ plot

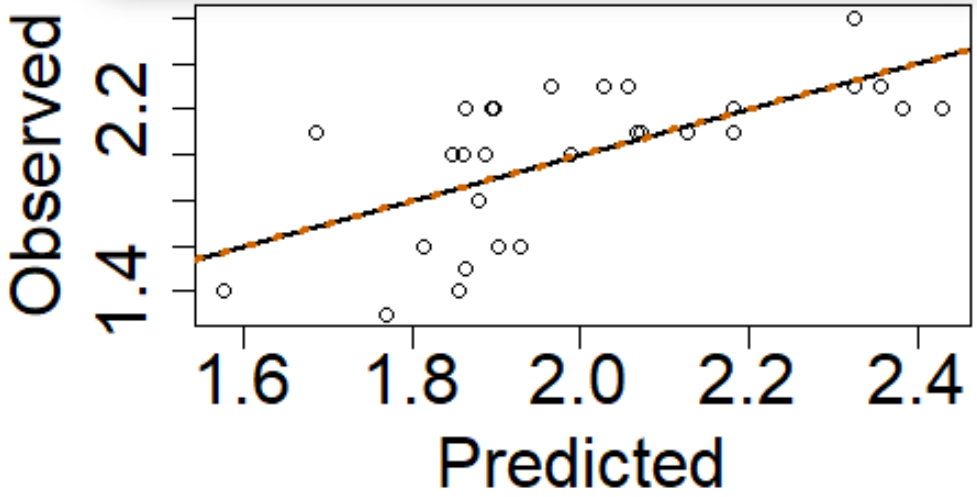
```
> #### Tests for normality
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.96138, p-value = 0.3553
```



Beef Example

- Model Fit: Plots and Regression of Observed Vs. Predicted
  - Points should fall along the line of unity
  - Regression of Obs ~ Pred should result in $\beta_0 = 0$, and $\beta_1 = 1$
  - Higher R^2 = better fit

```
> plot(beef$Marbling.Score ~ pred, xlab = "Predicted", ylab = "Observed", cex.axis = 2,
+      cex.lab = 2)
> ObsPredReg = lm (beef$Marbling.Score ~ pred)
> abline(a = 0 , b = 1, lwd = 2)
> abline(ObsPredReg, lty = 3, col = "darkorange3", lwd= 3)
> summary(ObsPredReg)

Call:
lm(formula = beef$Marbling.Score ~ pred)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47104 -0.18183  0.01278  0.24318  0.41460

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.639e-15  4.552e-01   0.000 1.000000
pred        1.000e+00  2.263e-01   4.418 0.000145 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2573 on 27 degrees of freedom
Multiple R-squared:  0.4196,    Adjusted R-squared:  0.3981
F-statistic: 19.52 on 1 and 27 DF,  p-value: 0.0001452
```



Beef Example

- Prediction Accuracy:
  - MSPE
    - Smaller values are better
    - Can take the square root to get the prediction error in units of observed value
  - CCC
    - Assesses degree to which the observed and predicted values agree – both precision and accuracy
    - $-1 < \rho < 1$
    - Values closer to 1 indicate a stronger concordance

```
> MSPE = sum((beef$Marbling.Score - pred)^2)/length(pred)
> MSPE
[1] 0.06163817
```

```
> epi.ccc(beef$Marbling.Score, pred)
$`rho.c`
        est     lower     upper
1 0.5911928 0.3402253 0.7635394

$s.shift
[1] 0.6477967

$l.shift
[1] -2.539602e-15

$c.b
[1] 0.9126207
```

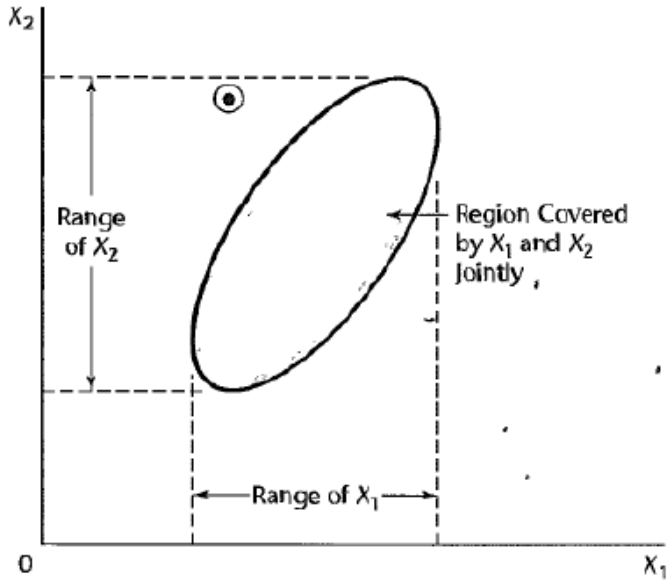
Prediction of new data

$$\hat{Y}_h = \mathbf{X}_h^\top \mathbf{b} = 1b_0 + X_{h1}b_1 + \cdots + X_{h,p-1}b_{p-1}$$

$$\mathbf{s}^2\{pred\} = MSE(1 + \mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h)$$

Prediction Interval: $\hat{Y}_{h(new)} \pm t(1-\alpha/2, n-p)s\{pred\}$

- Only use values that are within the multivariate region of the of the dependent variable samples



Kutner et al., 2014

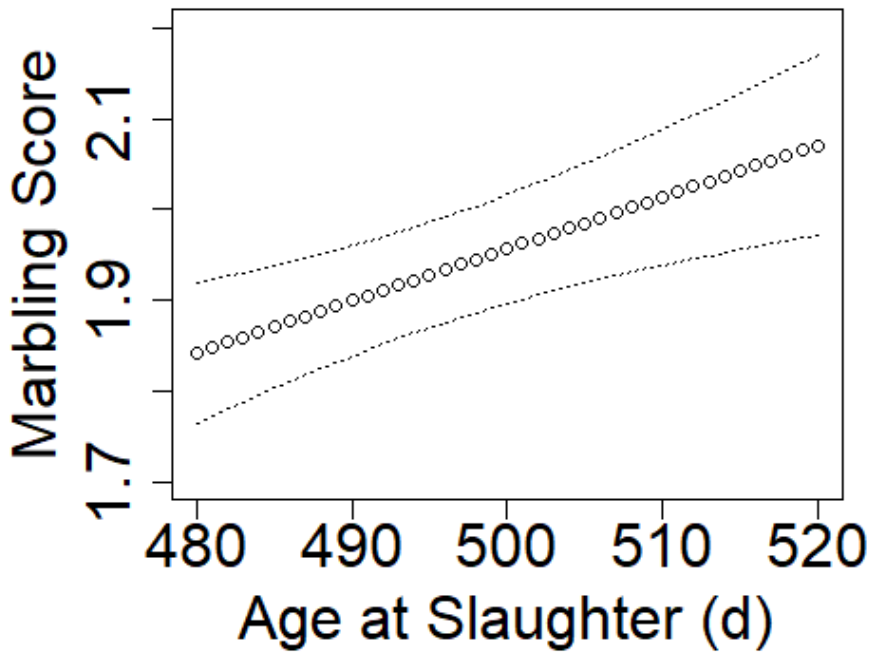
Prediction of new data

$$\hat{Y}_h = \mathbf{X}_h^{\mathsf{T}}\mathbf{b} = 1b_0 + X_{h1}b_1 + \cdots + X_{h,p-1}b_{p-1}$$

$$\mathbf{s}^2\{pred\} = MSE(1 + \mathbf{X}_h^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}_h)$$

Prediction Interval: $\hat{Y}_{h(new)} \pm t(1-\alpha/2, n-p)s\{pred\}$

```
#predict new data
summary(beef[ , c("DaysAtSlaughter", "MFA", "Backfat", "TFA")])

newData = data.frame(DaysAtSlaughter = 480:520, MFA = rep(48.6, 41),
            Backfat = rep(0.1, 41), TFA = rep(4.2, 41) )

new = predict(fit3, newdata = newData, se.fit = TRUE)

plot(new$fit ~ newData$DaysAtSlaughter, xlab = "Age at Slaughter (d)",
     ylab = "Marbling Score", ylim = c(1.7, 2.2), cex.axis = 2, cex.lab = 2)

ci.l = new$fit - new$se.fit
ci.u = new$fit + new$se.fit
lines(ci.l ~ newData$DaysAtSlaughter, lty = 3)
lines(ci.u ~ newData$DaysAtSlaughter, lty = 3)
```



Parameter Estimates

- How should the parameter mean and CI be interpreted?
  - Depends on…
    - Study design
    - Modeling objectives

- Remember that the parameter estimates are correlated

$$\text{Test statistic}: \quad t^* = \frac{b_k}{s\{b_k\}}$$

$\boldsymbol{\sigma}^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$

- $\mathbf{s}^2\{\mathbf{b}\} = MSE(\mathbf{X}^\top\mathbf{X})^{-1}$

$$\mathbf{s}^2\{\mathbf{b}\} = \begin{bmatrix} s^2\{b_0\} & s\{b_0, b_1\} & \cdots & s\{b_0, b_{p-1}\} \\ s\{b_1, b_0\} & s^2\{b_1\} & \cdots & s\{b_1, b_{p-1}\} \\ \vdots & \vdots & & \vdots \\ s\{b_{p-1}, b_0\} & s\{b_{p-1}, b_1\} & \cdots & s^2\{b_{p-1}\} \end{bmatrix}$$

Model Inference

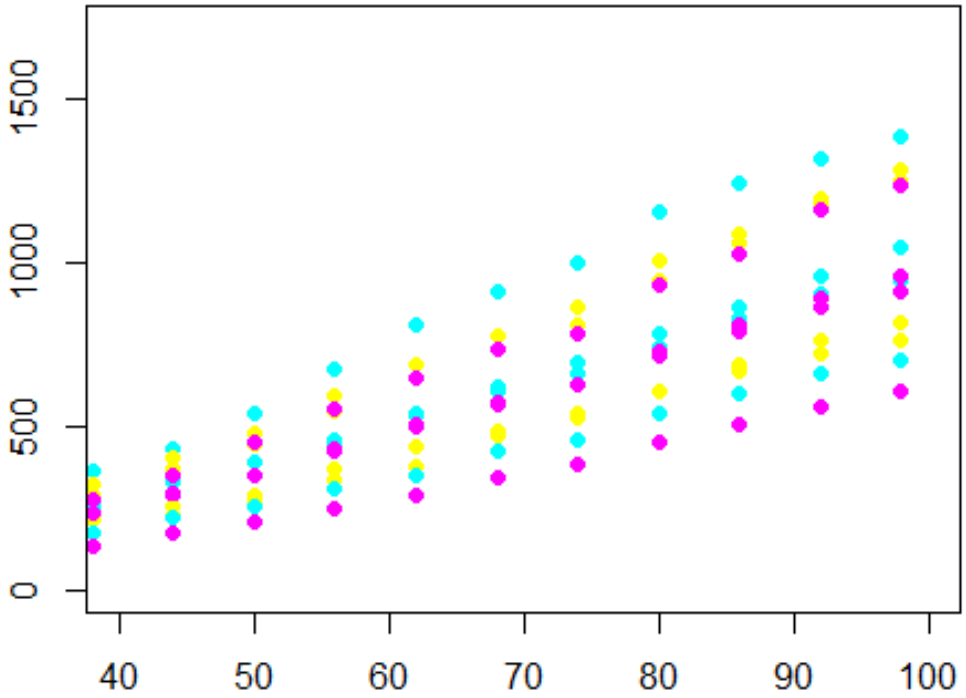
Parameter Estimates: Beef Example

```
Call:
lm(formula = Marbling.Score ~ DaysAtSlaughter + MFA + Backfat +
    TFA, data = beef)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47104 -0.18183  0.01278  0.24318  0.41460

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.985733   1.925195  -2.070  0.04935 *
DaysAtSlaughter  0.005717   0.003259   1.754  0.09213 .
MFA              0.038341   0.018019   2.128  0.04383 *
Backfat          8.711830   2.992610   2.911  0.00766 **
TFA              0.083155   0.086081   0.966  0.34367
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2729 on 24 degrees of freedom
Multiple R-squared:  0.4196,    Adjusted R-squared:  0.3229
F-statistic: 4.338 on 4 and 24 DF,  p-value: 0.008812
```

```
Call:
lm(formula = Marbling.Score ~ DaysAtSlaughter + MFA, data = beef)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5770 -0.2526  0.0351  0.1867  0.4672

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.932668   1.770301  -0.527   0.6028
DaysAtSlaughter  0.001865   0.003199   0.583   0.5649
MFA              0.041485   0.016167   2.566   0.0164 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3061 on 26 degrees of freedom
Multiple R-squared:  0.2091,    Adjusted R-squared:  0.1483
F-statistic: 3.438 on 2 and 26 DF,  p-value: 0.04735
```

Note that including additional parameters, can change the parameter estimates and their SE

What if groups of your data are related to each other?



Mixed Models

What is a
random effect?

Random Effects

- Random effects are used to represent correlations among observations that belong to the same group when the group characteristic is a random selection from the larger population.
- While a fixed effect describes an average response to a given factor or independent variable level, a random effect is assumed to have an average response of zero (it is random) and therefore influences the covariance structure rather than the mean or expected value of the response.

Mixed Models

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon$$

- $\mathbf{Y}$ is the $n \times 1$ vector of observations
- β is the $p \times 1$ vector of regression coefficients or fixed effects
- $\mathbf{X}$ is the $n \times p$ design matrix for the fixed effects
- $\mathbf{Z}$ is the $n \times m$ design matrix for the random effects
- $\mathbf{u}$ is the $m \times 1$ vector of random effects
- ε is the $n \times 1$ vector of errors

Situation

The Regression parameters are unknown

u is the vector of parameter estimates

Assumptions

-The expectation of Y_i and errors are the same as a simple linear model

-Have additional assumptions of the random effects

$$E[\mathbf{Y}] = \mathbf{X}\beta$$

$$E[\varepsilon] = \mathbf{0}$$

$$\boldsymbol{\sigma}^2\{\varepsilon\} = \sigma^2\mathbf{I}_{n\times n}$$

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}) \text{ and } \varepsilon \sim N(\mathbf{0}, \mathbf{R})$$

Objectives

To find 'good' values that can serve as estimates for ß, G, and R

Random Intercepts vs. Slopes

A random slope influences the *response* to the fixed effects AND this response can be separated from the effect of the random grouping on the overall mean response.

$$y_{ij} = \beta_0 + b_{0i} + \beta_1 x_{ij} + \varepsilon_{ij}$$

y_{ij} is the body weight on the jth time for the ith rat

β_0 and β_1 are the fixed intercept and slopes commom to all rats

b_{0i} is the random effect describing the deviation of rat i from the common intercept

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})x_{ij} + \varepsilon_{ij}$$

b_{1i} is the random effect describing the deviation of rat i from the common slope

ε_{ij} is the error

$i = 1, \ldots, 16$ and $j = 11$

$\varepsilon_{ij} \sim N(0, \sigma^2)$ and $\mathbf{b_i} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{int} & \sigma_{int,slope} \\ \sigma_{int,slope} & \sigma^2_{slope} \end{pmatrix}\right]$

Mixed Models

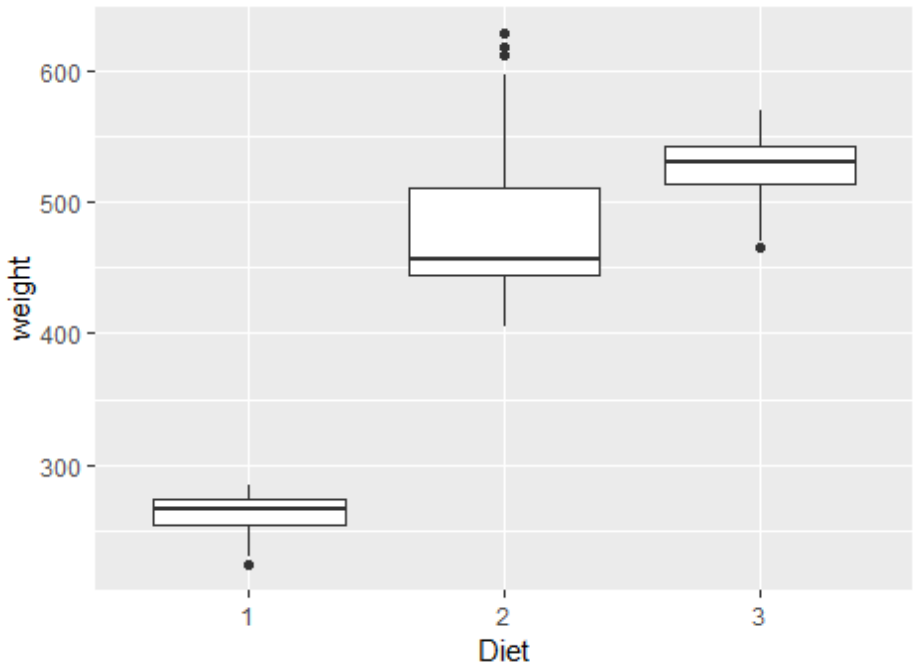
What is a random effect?

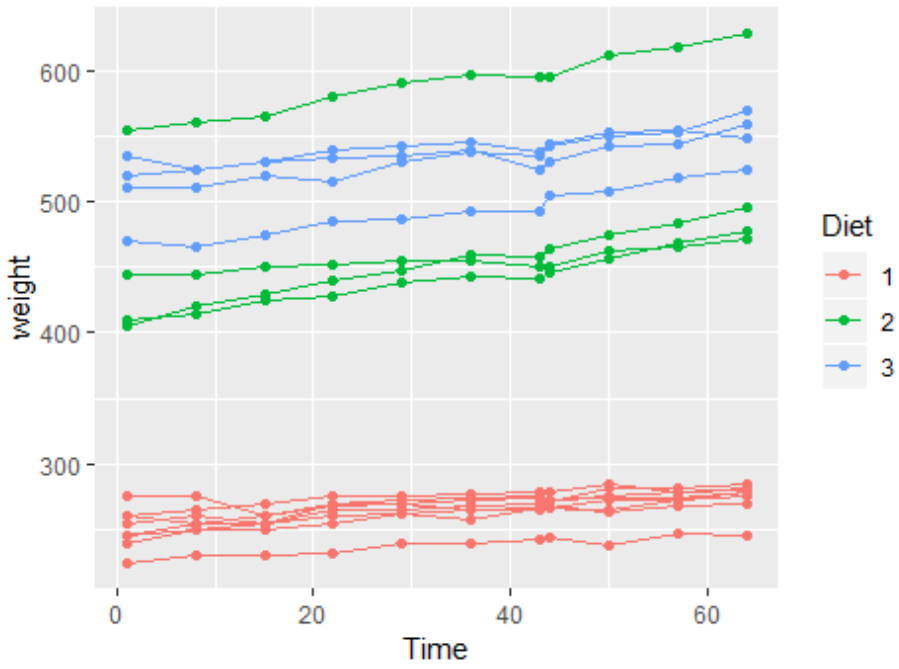
Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

Same data exploration in simple linear regression applies in mixed model regression except also want to separate by groups

Rat Bodyweight Example





Mixed Models

What is a random effect?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

Model Fitting

Find the 'best' numerical relationship between the dependent and independent variables.

Mixed Model Fitting

- MLE are the parameters that maximize the likelihood function of the data.
- Random effects increase the number of parameters that need to be estimated can make the optimization difficult.
- MLE variance estimates are biased.

MLE

Maximum Likelihood

REML

Restricted Maximum Likelihood

- Removes the fixed effects then estimates the variance components
- The estimated variance components act as constants in the second estimation procedure to find the MLE estimates of the fixed effects.
- REML estimates are less biased than MLE estimators, especially for balanced data.

Know your model fitting functions!

REML is often the default so if you need to fit in ML for model selection purposes, be sure to use the correct function inputs

Mixed Model Fitting

Note that the mean estimates are the same but their SE and the Random effect variances are different

ML

```
ML.fit = lmer(weight ~ Time + Diet + (1|Rat), data = BodyWeight, REML = FALSE)
summary(ML.fit)
```

```
> summary(ML.fit)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: weight ~ Time + Diet + (1 | Rat)
   Data: BodyWeight

     AIC      BIC   logLik deviance df.resid
  1333.2   1352.2   -660.6   1321.2      170

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5386 -0.5581 -0.0494  0.5693  3.0990

Random effects:
 Groups   Name        Variance Std.Dev.
 Rat      (Intercept) 1085.92  32.953
 Residual               66.44   8.151
Number of obs: 176, groups:  Rat, 16

Fixed effects:
             Estimate Std. Error t value
(Intercept) 244.06890   11.73107   20.80
Time          0.58568    0.03158   18.54
Diet2       220.98864   20.23577   10.92
Diet3       262.07955   20.23577   12.95

Correlation of Fixed Effects:
      (Intr) Time   Diet2
Time  -0.090
Diet2 -0.575  0.000
Diet3 -0.575  0.000  0.333
```

REML

```
REML.fit = lmer(weight ~ Time + Diet + (1|Rat), data = BodyWeight)
summary(REML.fit)
```

```
> REML.fit = lmer(weight ~ Time + Diet + (1|Rat), data = BodyWeight)
> summary(REML.fit)
Linear mixed model fit by REML ['lmerMod']
Formula: weight ~ Time + Diet + (1 | Rat)
   Data: BodyWeight

REML criterion at convergence: 1304.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5236 -0.5557 -0.0467  0.5667  3.0932

Random effects:
 Groups   Name        Variance Std.Dev.
 Rat      (Intercept) 1337.88  36.577
 Residual               66.85   8.176
Number of obs: 176, groups:  Rat, 16

Fixed effects:
             Estimate Std. Error t value
(Intercept) 244.06890   13.00477  18.768
Time          0.58568    0.03168  18.486
Diet2       220.98864   22.44958   9.844
Diet3       262.07955   22.44958  11.674

Correlation of Fixed Effects:
      (Intr) Time   Diet2
Time  -0.082
Diet2 -0.575  0.000
Diet3 -0.575  0.000  0.333
```

Mixed Models

What is a random effect?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

Model Fitting

Find the 'best' numerical relationship between the dependent and independent variables.

Model Selection

Have to select both fixed and random effects

Fixed effect selection

- Cannot use LRT with REML fitted models because the likelihoods are altered to remove the fixed effects so aren't comparable.
- Use ML fit models instead to select fixed effects in a mixed model

```
ML.fit = lmer(weight ~ Time + Diet + (1|Rat), data = BodyWeight, REML = FALSE)
ML.fit.red = lmer(weight ~ Time +  (1|Rat), data = BodyWeight, REML = FALSE)
anova(ML.fit, ML.fit.red)
```

```
Data: BodyWeight
Models:
ML.fit.red: weight ~ Time + (1 | Rat)
ML.fit: weight ~ Time + Diet + (1 | Rat)
           Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
ML.fit.red  4 1372.0 1384.7 -682.00   1364.0
ML.fit      6 1333.2 1352.2 -660.58   1321.2 42.841      2  4.979e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Mixed Models

What is a random effect?

Before beginning, it's important to know how the model results will be used.

Data Exploration

What data do you have to work with? What are their ranges? Patterns? Visual relationships?

Model Fitting

Find the 'best' numerical relationship between the dependent and independent variables.

Model Selection

How do you know which dependent variables should remain in the model?

Evaluation + Interpretation

The nature of the random effects and your objectives will determine how you can interpret them

Reporting Random effects

- Random effects are often not reproducible
  - Not valuable for inference on future populations
  - Rat #2 is on average 25.96 g below Rat #1 – but that estimate can only be used for inference on Rat #2

```
> random.effects(REML.fit)
$`Rat`
   (Intercept)
2   -25.961608
3    -3.518109
4     2.816750
1    -2.613129
8     1.730774
5     5.712685
6    10.961568
7    10.871070
11  -29.751211
9   -43.687900
10  -31.832664
12  105.271775
13  -32.737644
15   14.321305
14   10.429892
16    7.986447
```

Reporting Random effects

- Random effects are often not reproducible
  - Not valuable for inference on future populations
  - Rat #2 is on average 25.96 g below Rat #1 – but that estimate can only be used for inference on Rat #2
- So the interest usually lies in the degree of variation in the random grouping
  - Here the variance between rats 4x higher than the residual variance

```
> summary(REML.fit)
Linear mixed model fit by REML ['lmerMod']
Formula: weight ~ Time + Diet + (1 | Rat)
   Data: BodyWeight

REML criterion at convergence: 1304.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5236 -0.5557 -0.0467  0.5667  3.0932

Random effects:
 Groups   Name        Variance Std.Dev.
 Rat      (Intercept) 1337.88  36.577
 Residual               66.85   8.176
Number of obs: 176, groups:  Rat, 16
```

Residuals from Random Effects models

- The residuals() function gives the residuals from the fixed effect model

- The residuals object within the summary() of the model, gives the residuals from the BLUP (best linear unbiased predictor) that includes the random effects

Residuals calculated w/o random effects

```
> head(residuals(REML.fit))
        1         2         3         4         5         6 
-2.041451  3.858766  4.758983  5.659200  3.559417 -4.540366 
```

Residuals calculated w/ random effects

```
> head(summary(REML.fit)$residuals)
         1          2          3          4          5          6 
-0.2496745  0.4719366  0.5820353  0.6921341  0.4353255 -0.5552980 
```

* Estimates of prediction error should be based on the fixed effects model *without* the random effects!

Thank You

kfr3@cornell.edu