Introduction to Statistical Modeling

November 2 & 5, 2018
FANR 6750

Richard Chandler and Bob Cooper

Looking ahead

Linear models

Generalized linear models

Model selection and multi-model inference

Outline

Motivation

Why do we need this part of the course?

- We have been modeling all along
- Good experimental design + ANOVA is usually the most direct route to causal inference
- Often, however, it isn't possible (or even desirable) to control some aspects of the system being investigated
- When manipulative experiments aren't possible, observational studies and predictive models can be the next best option

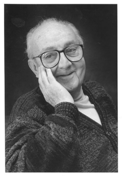
What is a model?

Definition
A model is an abstraction of reality used to describe the relationship between two or more variables

Types of models

- Conceptual
- Mathematical
- Statistical

Important point
"All models are wrong but some are useful" (George Box, 1976)



Statistical models

What are they useful for?

- Formalizing hypotheses using math and probability
- Evaulating hypotheses by confronting models with data
- Predicting future outcomes

Statistical models

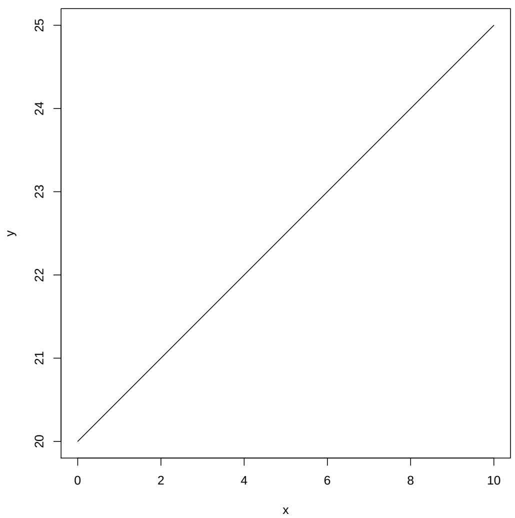
Two important pieces

(1) Deterministic component
- Equation for the expected value of the response variable

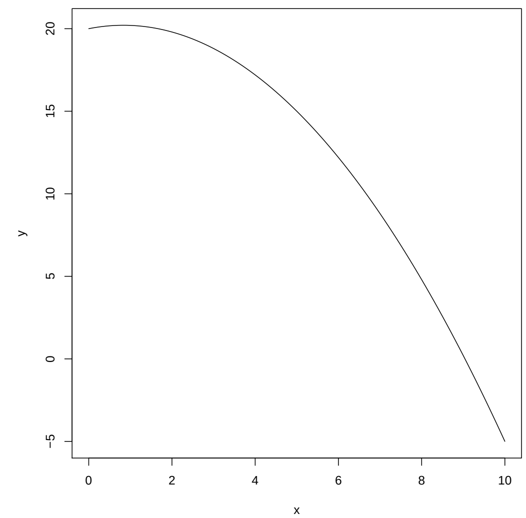
(2) Stochastic component
- Probability distribution describing the differences between the expected values and the observed values
- In parametric statistics, we assume we know the distribution, but not the parameters of the distribution

Outline

$$y = 20 + 0.5x$$



$$y = 20 + 0.5x - 0.3x^2$$



Linear model

A linear model is an equation of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

where the β's are coefficients, and the x values are predictor variables (or dummy variables for categorical predictors).

This equation is often expressed in matrix notation as:

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

where $\mathbf{X}$ is a design matrix and $\boldsymbol{\beta}$ is a vector of coefficients. More on matrix notation later...

Interpretting the β's

You must be able to interpret the β coefficients *for any model that you fit to your data*.

A linear model might have dozens of continuous and categorical predictors variables, with dozens of associated β coefficients.

Linear models can also include polynomial terms and interactions between continuous and categorical predictors

Interpretting the β's

The intercept β_0 is the expected value of y, when all x's are 0

If x is a **continuous** explanatory variable:

- β can usually be interpreted as a *slope* parameter.
- In this case, β is the change in y resulting from a 1 unit change in x (while holding the other predictors constant).

Interpretting β's for categorical explantory variables

Things are more complicated for **categorical** explantory variables (i.e., factors) because they must be converted to dummy variables

There are many ways of creating dummy variables

In **R**, the default method for creating dummy variables from unordered factors works like this:

- One level of the factor is treated as a reference level
- The reference level is associated with the intercept
- The β coefficients for the other levels of the factor are differences from the reference level.

The default method corresponds to:

```
options(contrasts=c("contr.treatment","contr.poly"))
```


Interpretting β's for categorical explantory variables

Another common method for creating dummy variables results in βs that can be interpretted as the α's from the additive models that we saw earlier in the class.

With this method:

- The β associated with each level of the factor is the difference from the intercept
- The intercept can be interpetted as the grand mean if the continuous variables have been centered
- One of the levels of the factor will not be displayed because it is redundant when the intercept is estimated

This method corresponds to:

```
options(contrasts=c("contr.sum","contr.poly"))
```

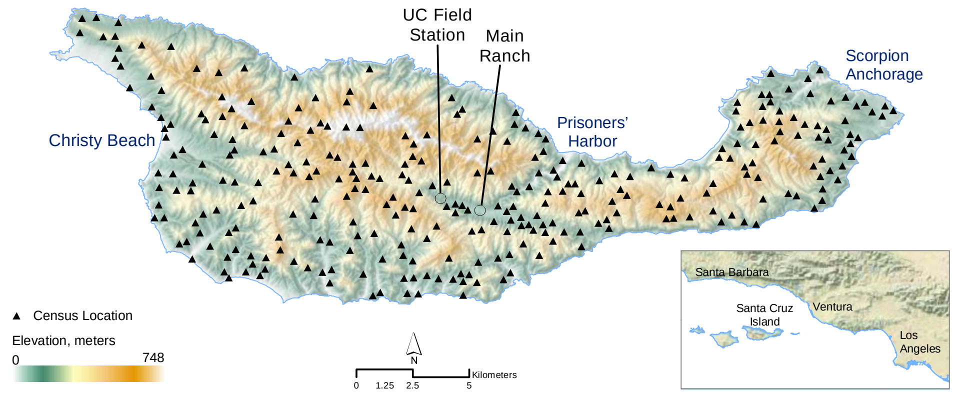

Outline

Example

The Island Scrub-Jay

Example

Santa Cruz Island

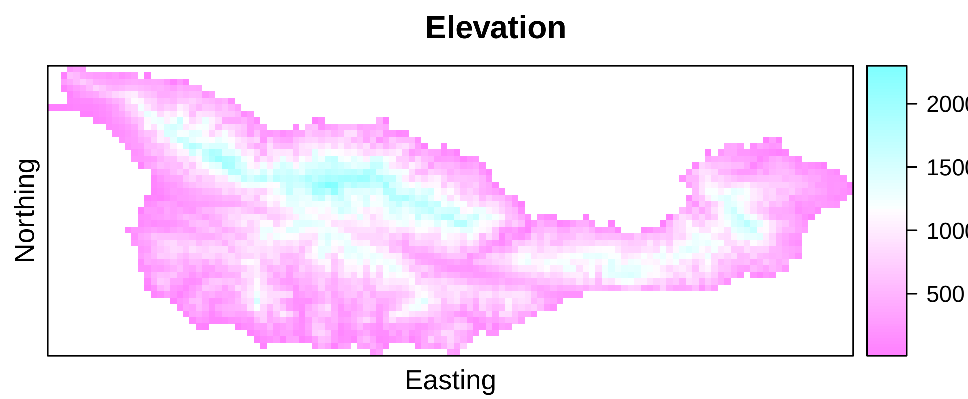


Santa Cruz Data

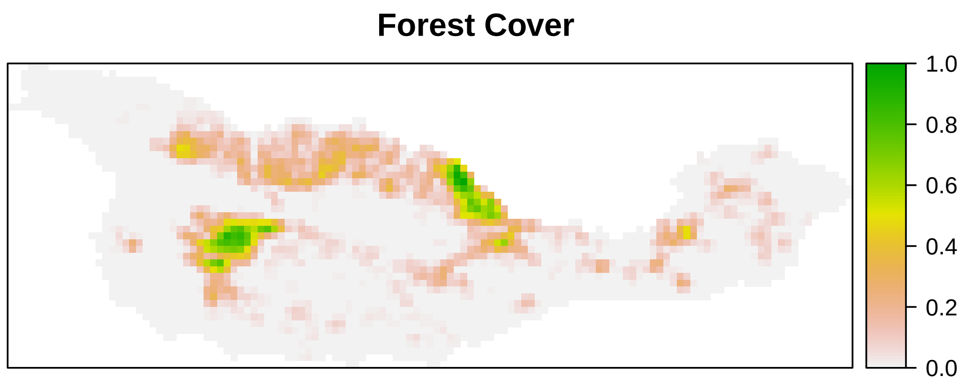
Habitat data for all 2787 grid cells covering the island

```
head(cruz2)

##          x       y elevation forest chaparral habitat seeds
## 1 230736.7 3774324       241      0         0     Oak   Low
## 2 231036.7 3774324       323      0         0    Pine   Med
## 3 231336.7 3774324       277      0         0    Pine  High
## 4 230436.7 3774024        13      0         0     Oak   Med
## 5 230736.7 3774024       590      0         0     Oak  High
## 6 231036.7 3774024       533      0         0     Oak   Low
```

Maps of predictor variables



Maps of predictor variables



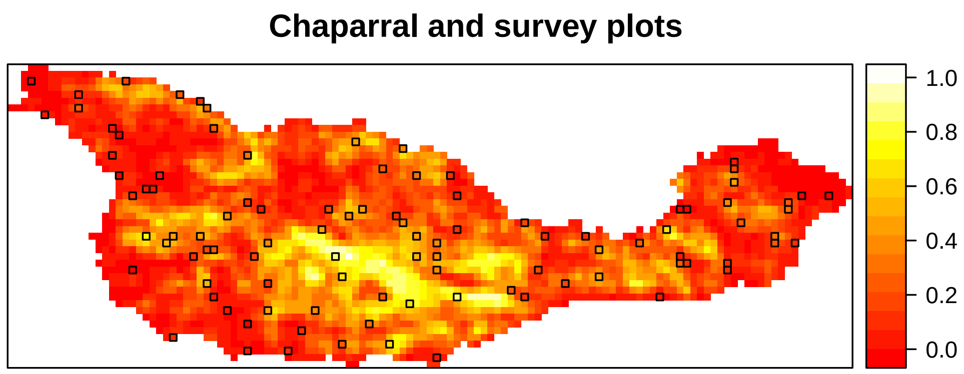
Questions

(1) How many jays are on the island?

(2) What environmental variables influence abundance?

(3) Can we predict consequences of environmental change?

Maps of predictor variables



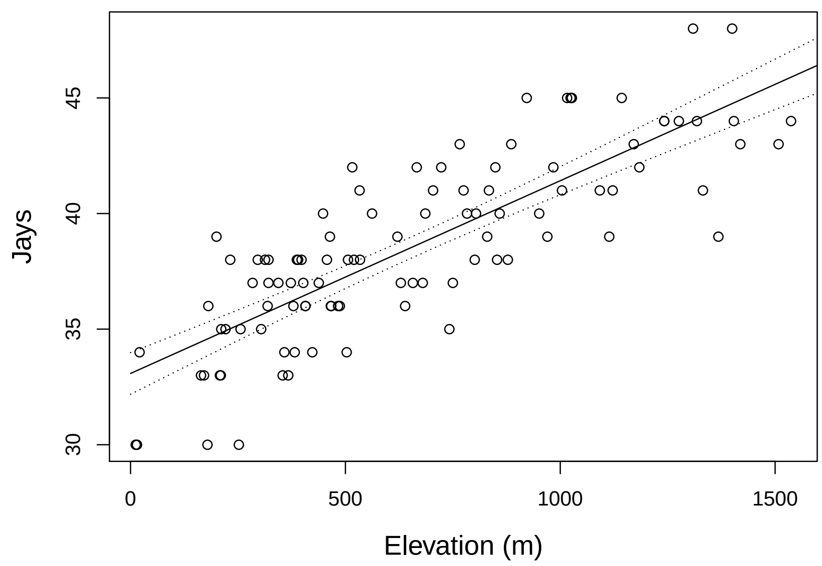
The (fake) jay data

```
head(jayData)

##             x       y elevation forest chaparral habitat seeds jays
## 2345 258636.7 3764124       423   0.00      0.02     Oak   Med   34
## 740  261936.7 3769224       506   0.10      0.45     Oak   Med   38
## 2304 246336.7 3764124       859   0.00      0.26     Oak  High   40
## 2433 239436.7 3763524      1508   0.02      0.03    Pine   Med   43
## 1104 239436.7 3767724       483   0.26      0.37     Oak   Med   36
## 607  236436.7 3769524       830   0.00      0.01     Oak   Low   39
```

Simple linear regression

```
fm1 <- lm(jays ~ elevation, data=jayData)
summary(fm1)

##
## Call:
## lm(formula = jays ~ elevation, data = jayData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4874 -1.7539  0.1566  1.6159  4.6155
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.082808   0.453997   72.87   <2e-16 ***
## elevation    0.008337   0.000595   14.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.285 on 98 degrees of freedom
## Multiple R-squared:  0.667, Adjusted R-squared:  0.6636
## F-statistic: 196.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

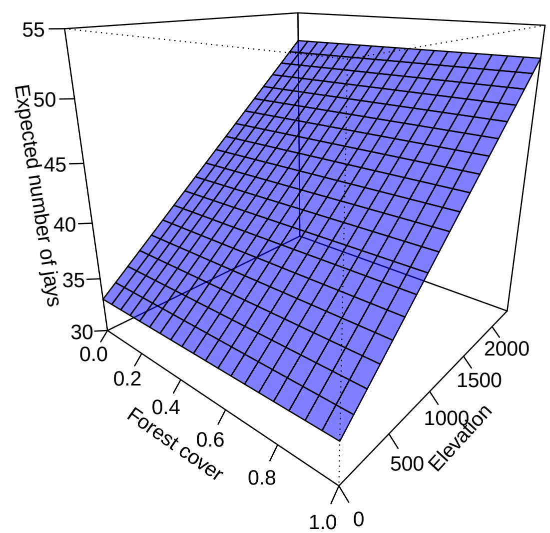

Simple linear regression



Multiple linear regression

```
fm2 <- lm(jays ~ elevation+forest, data=jayData)
summary(fm2)

##
## Call:
## lm(formula = jays ~ elevation + forest, data = jayData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4717 -1.7384  0.1552  1.5993  4.6319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.065994   0.467624  70.711   <2e-16 ***
## elevation    0.008337   0.000598  13.943   <2e-16 ***
## forest       0.294350   1.793079   0.164     0.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.296 on 97 degrees of freedom
## Multiple R-squared:  0.6671, Adjusted R-squared:  0.6603
## F-statistic: 97.21 on 2 and 97 DF,  p-value: < 2.2e-16
```

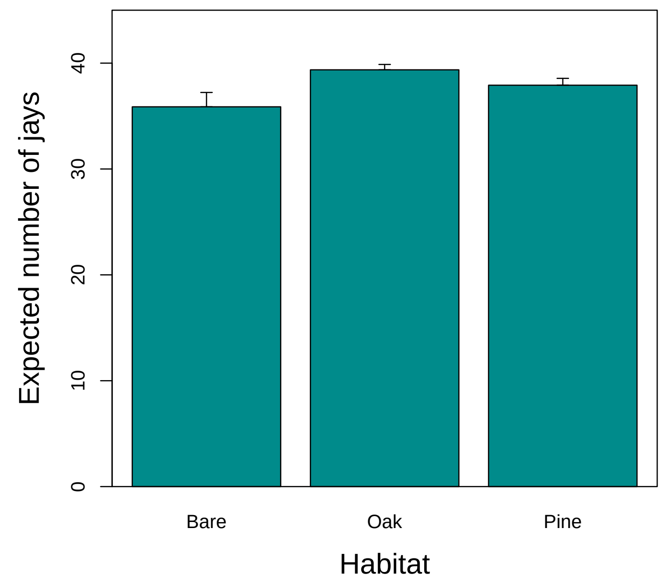

Multiple linear regression



One-way ANOVA

```
fm3 <- lm(jays ~ habitat, data=jayData)
summary(fm3)

##
## Call:
## lm(formula = jays ~ habitat, data = jayData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9143 -2.3684 -0.3684  3.0857  8.6316
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.875      1.356  26.456   <2e-16 ***
## habitatOak     3.493      1.448   2.413   0.0177 *
## habitatPine    2.039      1.503   1.357   0.1780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.835 on 97 degrees of freedom
## Multiple R-squared:  0.07126, Adjusted R-squared:  0.05211
## F-statistic: 3.721 on 2 and 97 DF,  p-value: 0.02773
```

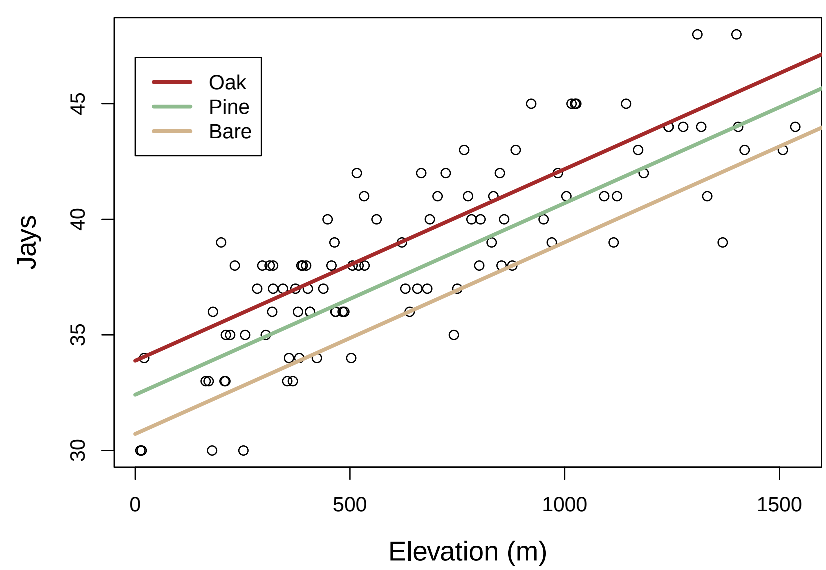

One-way ANOVA



ANCOVA

```
fm4 <- lm(jays ~ elevation+habitat, data=jayData)
summary(fm4)

##
## Call:
## lm(formula = jays ~ elevation + habitat, data = jayData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0327 -1.5356  0.0091  1.4686  4.2391
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.072e+01  8.084e-01  37.997  < 2e-16 ***
## elevation   8.289e-03  5.414e-04  15.308  < 2e-16 ***
## habitatOak  3.166e+00  7.850e-01   4.034  0.00011 ***
## habitatPine 1.695e+00  8.148e-01   2.081  0.04010 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.078 on 96 degrees of freedom
## Multiple R-squared:  0.7301, Adjusted R-squared:  0.7217
## F-statistic: 86.56 on 3 and 96 DF,  p-value: < 2.2e-16
```

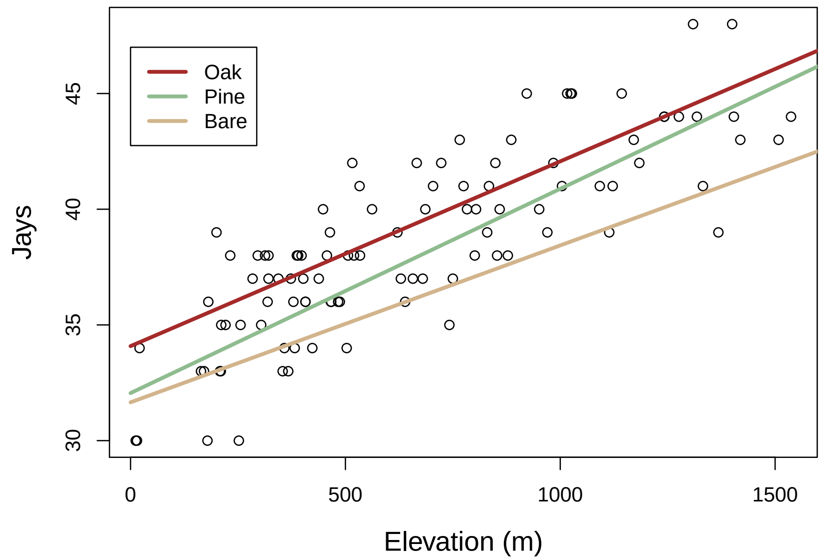

ANCOVA



Continuous-categorical interaction

```
fm5 <- lm(jays ~ elevation*habitat, data=jayData)
summary(fm5)

##
## Call:
## lm(formula = jays ~ elevation * habitat, data = jayData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.008 -1.581 -0.103  1.420  4.184
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          31.654383   1.446322  21.886  < 2e-16 ***
## elevation             0.006781   0.001999   3.393  0.00101 **
## habitatOak            2.428682   1.565227   1.552  0.12411
## habitatPine           0.399953   1.579874   0.253  0.80070
## elevation:habitatOak  0.001204   0.002153   0.559  0.57737
## elevation:habitatPine 0.002046   0.002151   0.951  0.34414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.087 on 94 degrees of freedom
## Multiple R-squared:  0.7334, Adjusted R-squared:  0.7192
```

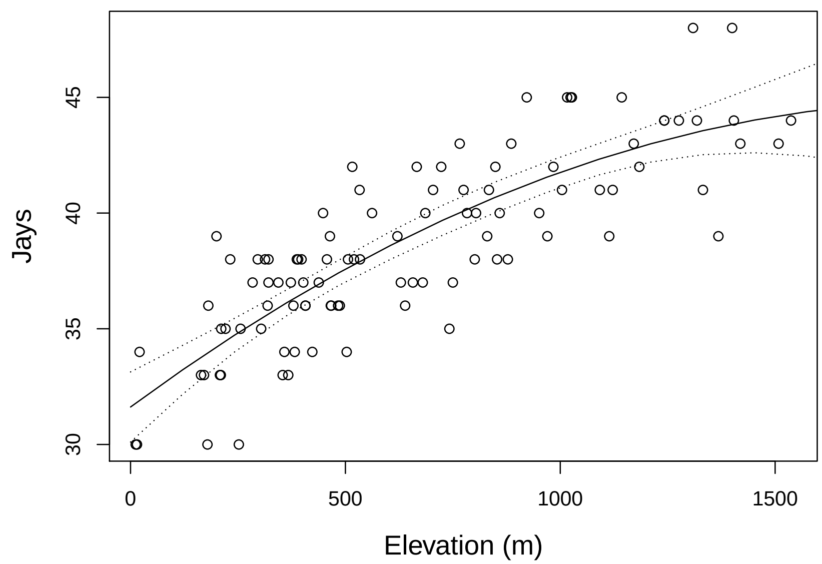

ANCOVA



Quadratic effect of elevation

```
fm6 <- lm(jays ~ elevation+I(elevation^2), data=jayData)
summary(fm6)

##
## Call:
## lm(formula = jays ~ elevation + I(elevation^2), data = jayData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8429 -1.4608  0.1304  1.5908  4.7854
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.162e+01  7.631e-01  41.434  < 2e-16 ***
## elevation      1.368e-02  2.342e-03   5.843 6.86e-08 ***
## I(elevation^2) -3.542e-06  1.503e-06  -2.357   0.0204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.233 on 97 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6786
## F-statistic: 105.5 on 2 and 97 DF,  p-value: < 2.2e-16
```


Quadratic effect of elevation



Interaction and quadratic effects

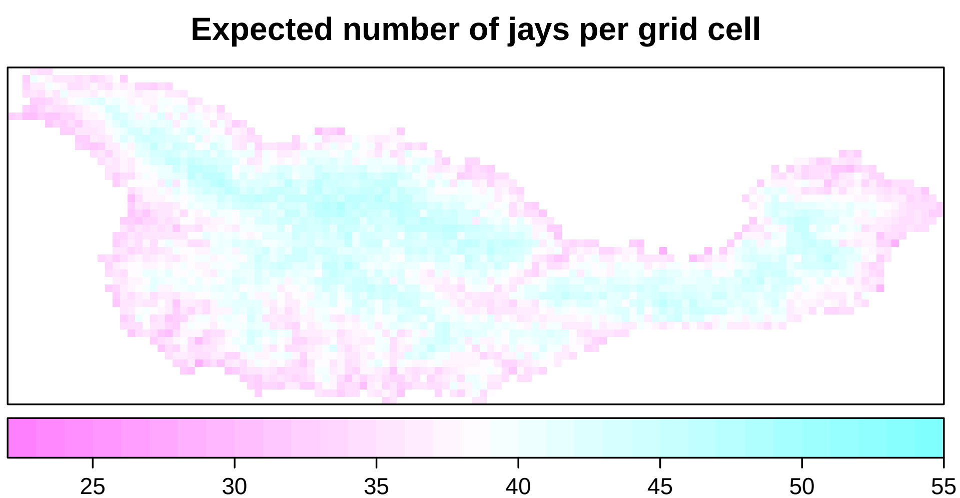
```
fm7 <- lm(jays ~ habitat * forest + elevation +
          I(elevation^2), data=jayData)
summary(fm7)

##
## Call:
## lm(formula = jays ~ habitat * forest + elevation + I(elevation^2),
##     data = jayData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2574 -1.4400  0.0487  1.4055  3.7924
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.920e+01  1.030e+00  28.338  < 2e-16 ***
## habitatOak           3.705e+00  8.433e-01   4.394 2.98e-05 ***
## habitatPine          2.216e+00  8.757e-01   2.531   0.0131 *
## forest               4.007e+01  2.780e+01   1.441   0.1529
## elevation            1.215e-02  2.300e-03   5.285 8.41e-07 ***
## I(elevation^2)      -2.554e-06  1.484e-06  -1.721   0.0886 .
## habitatOak:forest   -4.292e+01  2.785e+01  -1.541   0.1267
## habitatPine:forest  -3.918e+01  2.784e+01  -1.407   0.1627
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.044 on 92 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7307
## F-statistic: 39.37 on 7 and 92 DF,  p-value: < 2.2e-16
```


Predict jay abundance at each grid cell

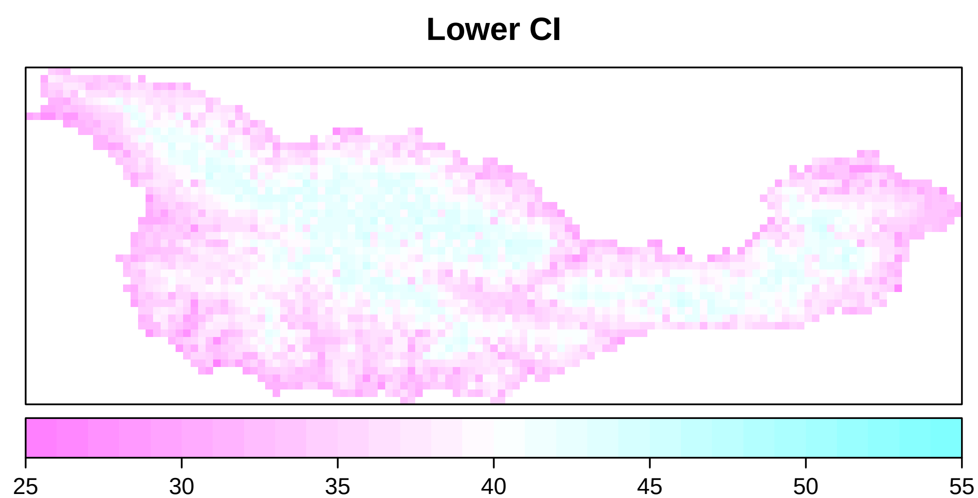
```
E7 <- predict(fm7, type="response", newdata=cruz2,
              interval="confidence")
```

```
E7 <- cbind(cruz2[,c("x","y")], E7)
head(E7)

##          x       y      fit      lwr      upr
## 1 230736.7 3774324 35.68349 34.86313 36.50386
## 2 231036.7 3774324 35.07284 34.22917 35.91652
## 3 231336.7 3774324 34.58427 33.72668 35.44186
## 4 230436.7 3774024 33.06042 31.55907 34.56177
## 5 230736.7 3774024 39.18440 38.49766 39.87113
## 6 231036.7 3774024 38.65512 37.98859 39.32165
```

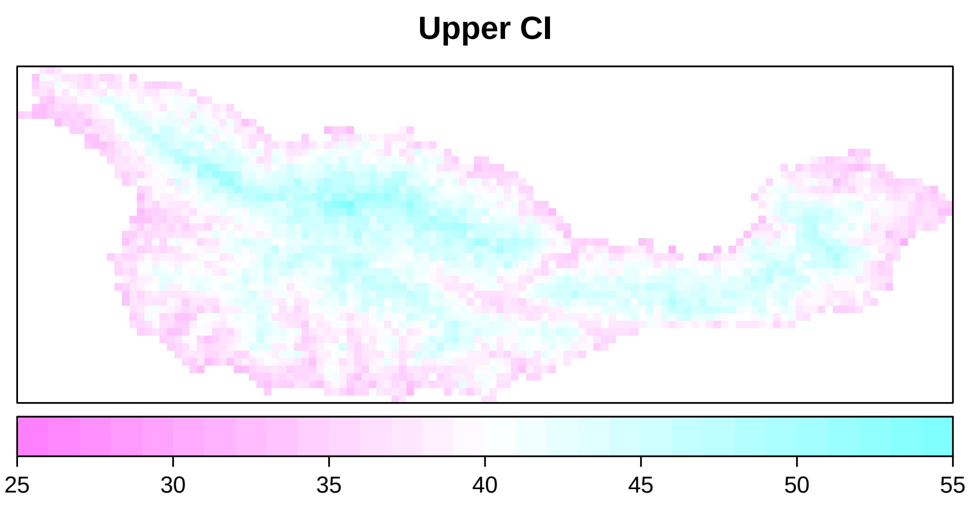

Map the predictions



Map the predictions

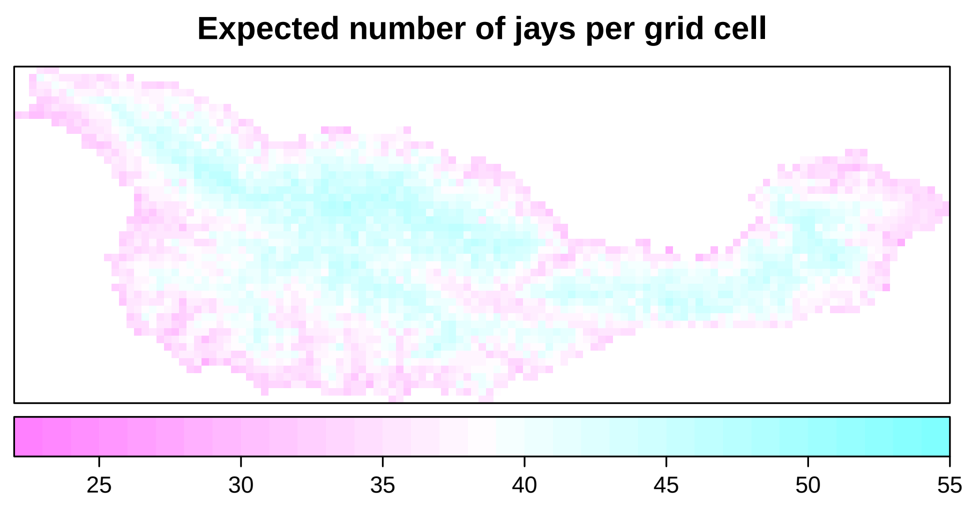


Map the predictions



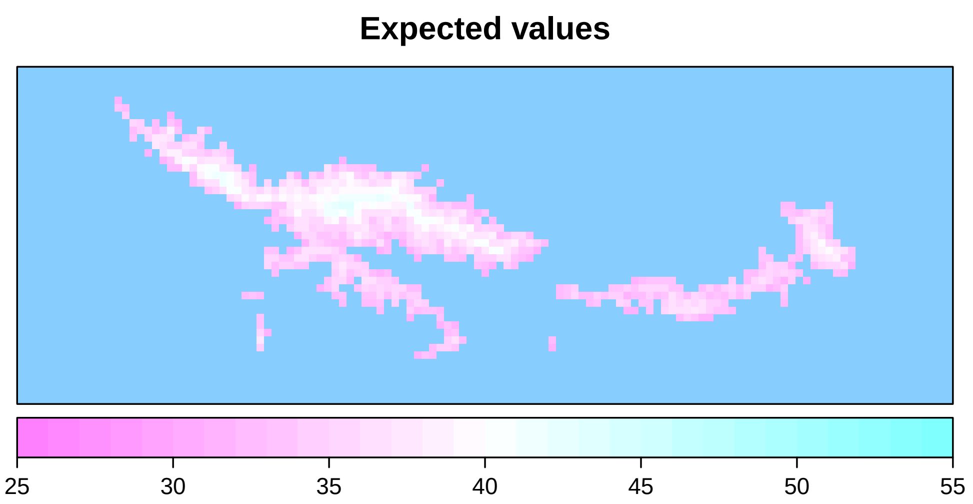
Future scenarios

What if pine and oak disapper?



Future scenarios

What if sea level rises?



Outline

Linear model

All of the fixed effects models that we have covered can be expressed this way:

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

where

$$\boldsymbol{\varepsilon} \sim \text{Normal}(0, \sigma^2)$$

Examples include

- Completely randomized ANOVA
- Randomized complete block designs with fixed block effects
- Factorial designs
- ANCOVA

Then how do they differ?

- The design matrices are different
- And so are the number of parameters (coefficients) to be estimated
- Important to understand how to construct design matrix that includes categorical variables

Design matrix

A design matrix has N rows and K columns, where N is the total sample size and K is the number of coefficients (parameters) to be estimated.

The first column contains just 1's. This column corresponds to the intercept (β_0)

Continuous predictor variables appear unchanged in the design matrix

Categorical predictor variables appear as dummy variables

In **R**, the design matrix is created internally based on the formula that you provide

The design matrix can be viewed using the `model.matrix` function

Design matrix for linear regression

Data

```
dietData <- read.csv("dietData.csv")
head(dietData, n=10)

##       weight    diet       age
## 1   23.83875 Control 11.622260
## 2   25.98799 Control 13.555397
## 3   30.29572 Control 15.357372
## 4   25.88463 Control  7.950214
## 5   18.48077 Control  5.493861
## 6   31.57542 Control 18.874970
## 7   23.79069 Control 12.811297
## 8   29.79574 Control 17.402436
## 9   21.66387 Control  7.379666
## 10  30.86618 Control 18.611817
```

Design matrix

```
X1 <- model.matrix(~age,
                   data=dietData)
head(X1, n=10)

##    (Intercept)       age
## 1            1 11.622260
## 2            1 13.555397
## 3            1 15.357372
## 4            1  7.950214
## 5            1  5.493861
## 6            1 18.874970
## 7            1 12.811297
## 8            1 17.402436
## 9            1  7.379666
## 10           1 18.611817
```

How do we multiply this design matrix ($\mathbf{X}$) by the vector of regression coefficients ($\boldsymbol{\beta}$)?

Matrix multiplication

$$\mathbb{E}(y) = \mathbf{X}\boldsymbol{\beta}$$

$$\begin{bmatrix} aw + bx + cy + dz \\ ew + fx + gy + hz \\ iw + jx + ky + lz \end{bmatrix} = \begin{bmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{bmatrix} \times \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix}$$

In this example

- The first matrix corresponds to the expected values of $\boldsymbol{y}$
- The second matrix corresponds to the design matrix $\mathbf{X}$
- The third matrix (a column vector) corresponds to $\boldsymbol{\beta}$

Matrix multiplication

The vector of coefficients

```
beta <- coef(lm(weight ~ age, dietData))
beta

## (Intercept)         age
##   21.325234    0.518067
```

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \text{ or } y_i = \beta_0 + \beta_1 x_i$$

```
Ey1 <- X1 %*% beta
head(Ey1, 5)

##       [,1]
## 1 27.34634
## 2 28.34784
## 3 29.28138
## 4 25.44398
## 5 24.17142
```


Summary

Linear models are the foundation of modern statistical modeling techniques

They can be used to model a wide array of biological processes, and they can be easily extended when their assumptions do not hold

One of the most important extensions is to cases where the residuals are not normally distributed. Generalized linear models address this issue.