Applied Ecological Statistics - ECO 636

Lab 4: Simple Linear Regression and a little Two-way ANOVA

Introduction

The purpose of this lab is to review how to interpret a two-way ANOVA and increase familiarity with applying simple linear regression to examine relationships between a single **continuous** response variable (y) and a single **continuous** explanatory variable (X). First, work through the 'worked' simple linear regression example together (or separate), where we will compare the *continuous* model with what we've seen in *categorical* predictors, and we will get more familiar with the `glm()` function. At the end of lab you will conduct a two-way ANOVA on your own as a review from last week and turn in a little report.

1.1 Simple linear regression background

Before diving in **Quiz yourself:** how would a linear model with a single *categorical* predictor (ANOVA) differ from a linear model with a single *continuous* predictor variable? What differences are there in the number of parameters the models estimate? What about the interpretation of those parameters? Okay ... onward!

Let's review the structure of a simple linear regression formulated as a linear model. As was the case with the one-way ANOVA, a simple linear regression has a single explanatory variable (X) but X is now continuous and the focus is on quantifying the **relationship** between y and X rather than estimating differences among expected group means. The model takes a familiar form:

$$y_i = \beta_0 + \beta_1 X_{1i} + e_i$$

Quiz yourself: This second model formula structure looks very familiar. Can you (together or individually) walk through what it means? In the ANOVA model, β_0 is the expected mean of the reference level, and β_1 is the contrast between expected means of different groups.

What do β_0 and β_1 represent here?

Think back to the basic formula for a line from high school algebra, $y = mx + b$, b was the intercept, and m was the slope. The same is true here! β_0 is the intercept, or, the value of y when X is 0, and β_1 is the slope, or the change in value of y with a one-unit increase in X. For example, if β_1=10, then the expected value of y increases by 10 units with a one unit increase in *X*. Similarly, if β_1=-10, then the expected value of y decreases by 10 units with a one unit increase in X. Notice that there is some conceptual similarity between interpreting β's when X is categorical and when X is continuous. They both represent changes in the expected value of y as you change the value of X. The only difference is whether those "values" represent categorical groups or real numbers.

Quiz yourself: What null hypothesis is being tested in the ANOVA? What about in the simple linear regression?

The null hypotheses for these two models is essentially the same, that there is no effect of X on y. When fitting a simple linear regression, we are implicitly asking how the expected value of y changes as a **linear** function of X. The simplest null hypothesis is that the expected value of y remains constant across all values of X, i.e., there is no change, which is equivalent to saying that β_1=0. Under this scenario, there is no effect of X ($\beta_1 = 0$) and so X drops out of the model and we are left with an intercept-only model where β_0 is the mean (i.e., expected value) of y.

Of course the assumptions of the linear model still hold through for continuous data as well (now we all pause while you *reflexivly* recite these - done? Okay, read on).

With categorical variables, observations were grouped and so residuals were assumed to normally distributed around the estimated group mean. For a continuous predictor, multiple observations aren't made at each value of X, but evaluating the homogeneity assumption is similarly straight-forward. We simply look for a lack of trend or pattern in the residuals when plotted against either X or the expected values of y. This lack of pattern would suggest that residuals are similarly distributed across all values of X.

1.2 Worked example

Now let's work through fitting and interpreting a simple linear regression model. The data for today's lab were collected from a radio telemetry study of prairie rattlesnakes, *Crotalus viridis*, from the northwestern Rocky Mountains. The primary goal of this study was to describe prairie rattlesnake movement patterns in a mountainous landscape, but researchers also looked at how movement metrics were influenced by body size. Some studies have reported that snake movement patterns vary by body size, specifically that larger bodied individuals move further than their smaller bodied counterparts. This may reflect age-specific differences in movement patterns or energetic requirements. Alternatively, in populations exhibiting male mate-searching, size-specific variation in movement patterns among males may reflect size-specific variation in mate-searching movements.

Figure 1: Prairie rattlesnake

We will use simple linear regression to examine the relationship between movement rate (distance moved per unit time) and body weight. We can therefore specify the statistical model as:

$$\text{Rate}_i = \beta_0 + \beta_{\text{Weight}}\text{Weight}_i + e_i$$

Quiz yourself: Write out the null and alternative hypotheses for this analysis? Can you write it out (or describe it) in words and in algebra?

Let's read in the data and have a look at it.

```r
setwd("<your directory>")
rattlers <- read.table("rattlesnake.rate.txt", header = TRUE)
head(rattlers)
str(rattlers)
```

Thinking back to the models we are testing here, the alternative hypothesis here is that the relationship between movement rate and body weight *is* different from zero, or that $\beta_{\text{Weight}} \neq 0$, the null is that there is *no* difference, or $\beta_{\text{Weight}} = 0$.

Now, take a look at the data, do this with `head(rattlers)` and `str(rattlers)`. The `rattlers` data consist of three columns: `rate`, the movement rate specified as meters moved per day, `weight`, snake body weight in grams, and `Sex`. Let's look at the data in more detail by using `summary()`, and tabulate male and female observations using `table()`.

```r
summary(rattlers)
table(rattlers$Sex)
```

Sample sizes are relatively small and made of up primarily of males. Because the mechanisms for size-specific differences in movement rate could vary by sex (e.g., due to sex-specific differences in mate-searching), let's remove the females from the data to avoid any confounding effects of sex.

```r
rattlers <- subset(rattlers,rattlers$Sex=="M",c('rate','weight','Sex'))
table(rattlers$Sex)
summary(rattlers)
```

As always, we should graphically explore the data before proceeding. Because the predictor is continuous, we will use a different set of plots than we have used in proceeding labs. Let's make *histograms* and *boxplots* of both rate and weight to examine the distribution of each variable.

Quiz yourself: Do you notice anything of note in these exploratory plots?

```r
par(mfrow=c(2,2))

#Histogram of rate
hist(rattlers$rate, breaks=seq(0,150,5), cex.axis=0.8,
     xlab="Movement rate (m/day)", main="")

#Boxplot of rate
boxplot(rattlers$rate, cex.axis=0.8,
        ylab="Movement rate (m/day)", main="")

#Histogram of weight
hist(rattlers$weight, breaks=seq(0,600,20),
     cex.axis=0.8, xlab="Body weight (g)", main="")

#Boxplot of weight
boxplot(rattlers$weight, cex.axis=0.8,
        ylab="Body weight (g)", main="")
```

These graphs communicate information about each variable's distribution independently, but we really want to visualize the relationship *between* these variables to visualize a **linear** relationship between movement rate and body weight. To do this let's use `scatter.smooth()` to create a *scatter plot* of rate and weight. We can also add a trend line to the plot using `abline()` and a fitted `glm()` object with a `gaussian` distribution. It is important to realize that we are **not** fitting the model at this point. This is data exploration, we merely want to see if there appears to be a **linear** relationship between movement rate and body weight. Does a *simple linear regression* seem like an appropriate model for this data?

```
#Scatter plot
par(mfrow=c(1,1))
grn <- adjustcolor("forestgreen",0.6)
scatter.smooth(rattlers$weight, rattlers$rate,
     ylim=c(0,125),xlim=c(150,500),
     ylab="Movement rate (m/day)",
     xlab="Body weight (g)",pch=21,
     bg=grn)
abline(glm(rate~weight,data=rattlers,family=gaussian),lwd=2,lty=2, col=4)
```

Quiz yourself: What does this plot tell you about the linear relationship between movement rate and body weight? What do you notice about the spread of the data points around the fitted line, that is, the variability in movement at different body weights? Could this possibly violate one of the model assumptions?

```
mod1 <- glm(rate~weight,data=rattlers, family = "gaussian")
par(mfrow=c(2,3))
plot(mod1,pch=16)
hist(resid(mod1),breaks=seq(-60,60,5),main="",xlab="Residuals")
plot(resid(mod1)~rattlers$weight,pch=16,
     ylab="Residuals",xlab="Body weight (g)")
```

Remember that for a linear regression model, plots of the residuals should not show a clear pattern or structure but rather should appear randomly distributed about 0.

Quiz yourself: Is there any pattern or structure to these residual plots that concerns you? Why, or why not?

There does appear to be a *bit* more residual variation at intermediate-upper body weights although this is arguably caused by only a single point (see that one way up at the top there). Does this pattern seen consistent with the scatter plot of movement rate and body weight we just looked at? Let's apply our old friend the log-transformation (seriously though, so - much - *logging*) to movement rate, refit the model, and re-evaluate the fit. Note that with glm, we have other ways of dealing with non-normal residuals, but for now we are going to stick with the familiar, we'll get to non-normal residuals in a few labs :).

```
rattlers$lograte <- log(rattlers$rate) # log transformation of rates
log.mod <- glm(lograte~weight,data=rattlers, family = "gaussian")
par(mfrow=c(2,3))
plot(log.mod,pch=16)
hist(resid(log.mod),breaks=seq(-2,2,0.1),main="",xlab="Residuals")
plot(resid(log.mod)~rattlers$weight,pch=16,
     ylab="Residuals",xlab="Body weight (g)")
```

After the log-transformation, the Q-Q plot indicates that the residuals are normally distributed. The residuals also have less evidence of structure or pattern. Since the model's assumptions now appear met, we can proceed with statistical inference. Let's also replot the *scatter plot* of log(movement rate) and body weight and add a new linear trend line using `abline()`.

```
summary(log.mod)
par(mfrow=c(1,1))
plot(lograte~weight, rattlers, pch=16,col = grn, ylim=c(2,5), xlim=c(150,500))
abline(log.mod,lwd=2,lty=2)
```

Quiz yourself: Let's break down this summary table. What do the estimates for `Intercept` and `weight` mean here?

Ok, so based on that, what can we say about the null hypothesis - do we reject? Well, the p-value is <0.05 and the $\hat{\beta}$ for `weight` is positive so we can conclude that there is a significant positive relationship between log(movement rate) and body weight. But how do we think about how well our model explains the variance of our data? With the `lm()` function we used the R^2 and an estimate of σ as the residual standard error, but with `glm()` we instead have measures of deviance. With `glm()` we often have non-normal residuals, so we can't make the same assumptions about our data and use R^2 or σ, instead we use the deviance. Remeber: the `Null deviance` refers to how well the reponse is predicted with nothing but an `intercept` (our null model), and the `Residual deviance` showes how well the response is predicted by the model with all explanatory variables included. A smaller deviance represents a better fit. For linear regression with normally distributed residuals (what we have!) the `Residual deviance` is equal to the residual sum of squares. For us, it is good to see that the `Residual deviance` is less than the `Null deviance` but the small difference indicates that the null model might explain most of the variation in the data (and there is not much gain in using our weight explanatory variable). If we want to approximate R^2 for `glm()` models, we can use the following equation:

$$R^2_{\text{approx}} = 1 - (\text{Residual deviance}/\text{Null deviance})$$

If we compute the approximate R^2, we find only ~24% of the variation in log(movement rate) is explained by body weight.

We have run the model, validated the model, and interpreted the output. Let's review. How can we interpret the $\hat{\beta}$'s for the `intercept` and `weight`? Remember that the intercept is the value of y when X=0. So, the log(movement rate) for a snake with a body weight of zero grams is approximately 3.05 (or approximately 21 meters/day). We can demonstrate this by re-plotting the data adding an `abline()` (the fitted line), and then adding a vertical line to represent X=0 and a horizontal line to represent the value of $\hat{\beta}_0$. Note that we have expanded the scale of the x-axis to include X=0. We can also demonstrate this using `predict()` to predict the expected log(movement rate) for a snake with weight = 0.

```
plot(lograte~weight, rattlers, pch=16, col = grn, ylim=c(2,5),xlim=c(-100,500))
abline(log.mod,lwd=2,lty=2)
abline(v=0,col="red")
abline(h=coef(log.mod)[1])
```

Let's work through the prediction of this model and it's confidence intervals in `R` using two new packages: `HH` and `ggplot2`, so load those into your workspace first.

```
# load libraries
library(HH)
library(ggplot2)


# set up new data frame for prediciton
```

```r
weight_new <- seq(from = min(rattlers$weight),
                to = max(rattlers$weight),
                length = 100)
df_preds <- data.frame(weight = weight_new)

# now we predict u sing the HH package
pred.1 <- predict(log.mod, newdata = df_preds, se.fit = TRUE)
pred.2 <- interval(log.mod,pred.1)
pred.int <- data.frame(df_preds,pred.2)
head(pred.int)

# Now lets plot using ggplot
ggplot(pred.int,aes(x=weight,y=fit))+
   geom_line()+
   geom_line(aes(x=weight,y=ci.low,color=2))+
   geom_line(aes(x=weight,y=ci.hi,color=2))+
   theme_minimal() + ylab("Log Movement Rate")+
   xlab("Weight")+
   theme(panel.border = element_rect(colour = "black",fill=NA))
```

However, we cannot have a snake with zero body weight so $\hat{\beta}_0$ is currently of no practical use. Is there some way to obtain an estimate of β_0 that is biologically meaningful? There is, using a method called *centering* which is accomplished by subtracting each value from its mean. This allows the mean of the *centered* variable to equal zero and $\hat{\beta}_0$ will then be interpreted as the expected value of y at the mean of the *centered* variable. In our example, if we re-fit the model using the *centered* body weight, $\hat{\beta}_0$ will then equal the expected value of log(movement rate) at the mean of `weight`. Let's see this in practice and show that $\hat{\beta}_0$ equals the mean of log(movement rate) calculated from the data.

Quiz yourself: will centering change the distribution of the data, or the model estimates?

```r
rattlers$weightC <- rattlers$weight-mean(rattlers$weight)
round(mean(rattlers$weightC),2)

log.modC <- glm(lograte~weightC, data = rattlers, family = "gaussian")
summary(log.modC)
mean(rattlers$lograte)
```

```r
plot(lograte~weightC, rattlers, pch=16, col=grn, ylim=c(2,5),xlim=c(-200,200))
abline(log.modC,lwd=2,lty=2)
abline(v=0,col="red")
abline(h=coef(log.modC)[1])

exp(coef(log.modC)[1]) # Back-transform to original scale
mean(rattlers$rate) # Compare with the observed mean
```

So how do we interpret $\hat{\beta}_{\text{weight}}$? Remember that the $\hat{\beta}$ for the slope is the change in the expected value of y per unit change of X. So log(movement rate) increases by approximately 0.0024 for every one gram increase in body weight. Using the centered body weight, we know that the expected log(movement rate) for a snake with average body weight (333 g) is approximately 3.83. So let's calculate the expected log(movement rate) for a snake with a **one unit** increase in body weight (i.e., 334 g).

```r
#b0 + 1*b1
coef(log.modC)[1]+coef(log.modC)[2]
```

The absolute value of $\hat{\beta}_{\text{weight}}$ is so small that the change in log(movement rate) over a one-unit change in body weight is barely perceptibly different from the log(movement rate) for a snake with average body weight. Remember that these are still in logged units. If we want to know what the expected movement rate is for every one unit change in body weight is in meters, we must back-transform the data.

```r
#b0 + 1*b1
exp(coef(log.modC)[1]+coef(log.modC)[2])
```

One alternative approach to look at how movement rates change with body weight when one unit changes are small is to multiply $\hat{\beta}_{\text{weight}}$ by some large value. For example, we may wish to know how the expected log(movement rate) changes with a **100** g increase in body weight. To calculate this change, we would simply multiple $\hat{\beta}_{\text{weight}}$ by 100, and add it to $\hat{\beta}_0$. Increasing body weight by 100 g corresponds to a movement rate of 58 meter/day, which is approximately 12 meters further than the expected movement rate for a snake with average body weight.

```r
#b0 + 100*b1
coef(log.modC)[1]+(100*coef(log.modC)[2])
exp(coef(log.modC)[1]+(100*coef(log.modC)[2]))
```

Alternatively, we could simply re-scale the explanatory variable. This may be particularly useful when the explanatory variable is a unit of length or weight (e.g., meters, grams). As an example, let's convert weight from grams to kilograms. The $\hat{\beta}_{\text{weight}}$ would then be interpreted as the change in log(movement rate) for a 1 kg change in body weight. Notice that the 1 kg change in body weight now corresponds to a movement rate of 486 meters/day. How do you think *scaling* will that change our model estimates?

```r
rattlers$weightCkg <- rattlers$weightC/1000
log.modCkg <- glm(lograte~weightCkg, data = rattlers, family = "gaussian")
summary(log.modCkg)

coef(log.modCkg)[1]+(coef(log.modCkg)[2])
exp(coef(log.modCkg)[1]+(coef(log.modCkg)[2]))
```

Now that we have explored the process of fitting a simple linear regression using `glm()`, let's plot the expected values (i.e., means) of movement rate as a function of body weight on the original scale of each variable using `log.mod`. We can also plot confidence and prediction intervals around the expected values. To do this, we will use `HH` and `ggplot2` again like above.

First, let's create a data frame with the the values of body weight that will be used to predict log(movement rate) (we can back-transform the predicted values later). Remember that this data frame must have the same column names as the data used to fit the original model. We will create a data frame with a single column (`weight`) containing a sequence of body weight values spanning the entire range of the observed data for body weight. Let's use `predict()` and `interval()` to get predicted values of the mean and corresponding confidence and prediction intervals, respectively. The default interval width for `predict()` is 95%. Finally, let's use `exp()` to back-transform the predicted values onto the original scale of movement rate (meters/day). We can then add these predicted values and their appropriate intervals to a scatter plot of movement rate and body weight using `polygon()`.

```r
# Plotting expected values, CI, and PI
# set up new data frame for prediciton
weight_new <- seq(from = min(rattlers$weight),
                  to = max(rattlers$weight),
                  length = nrow(rattlers))
df_preds <- data.frame(weight = weight_new)
# now we predict
```

```
pred.1 <- predict(log.mod, newdata = df_preds, se.fit = TRUE)
pred.2 <- interval(log.mod, pred.1)
pred.int <- data.frame(df_preds, pred.2)
head(pred.int)

# Now lets plot
   # this time we will add prediction intervals
ggplot(pred.int,aes(x=weight,y=rattlers$lograte))+
   geom_point(aes(alpha=0.5))+
   geom_line(aes(x=weight,y=fit))+
   geom_line(aes(x=weight,y=ci.low,color=5))+
   geom_line(aes(x=weight,y=ci.hi,color=5))+
   geom_line(aes(x=weight,y=pi.low,color=3))+
   geom_line(aes(x=weight,y=pi.hi,color=3))+
   theme_minimal()+ylab("Log Movement Rate")+xlab("Weight")+
   theme(panel.border = element_rect(colour = "black",fill=NA),
         legend.position = "none")
```

Why are the prediction intervals so much larger than the confidence intervals? Confidence intervals represent uncertainty in the expected values (the mean value of y at a given value of X). That is, we would expect the 95% confidence interval to contain the true expected values of movement rate 95% of the time. This is analogous to the standard error or CI for the mean. In contrast, the prediction interval represents the uncertainty in predicting a **single** value of y. This uncertainty must incorporate both the uncertainty in estimating the expected value and random error about that expected value. So we would be 95% sure that a future observation would lie within the prediction interval.

Notice that the spread of the simulated data falls approximately within the 95% prediction intervals.

Wahoo - made it to the end of the worked example! Let's move on to review a two-way ANOVA where you will work up the code yourself.

Take home assignement (DUE Monday night): *Interacting with Migrants*

A local ecologically minded non-profit is looking to begin a new project to re-forest areas in the northeast. They are interested in providing habitat for all organisms, but have a particular interest in migratory birds. The organization's project lead is curious whether she should purchase parcels in the midst of human presence or only in more "natural" settings. You and your team have been tasked with doing some research, analyzing data, and reporting back to the project lead. If the organization wants to provide habitat for as many *migrant* birds as possible should they purchase parcels in *anthro* or *natural* sites? Working in your groups or solo, and using the `migrants` data (on moodle), follow the modelling process (like we just did above, but now with categorical variables). You will write up a <u>brief</u> executive summary describing your findings and recommendations (no more than 2 pages of single spaced *writing*, not including code or figures). The summary should describe your process and results, including 1-2 supporting figures for model interpretation/prediction, and 1-2 figures for evaluating model assumptions. You should:

- Develop a set of hypotheses (null and alternative) based on the project goals
- Describe **all** relevant statistical model(s) in words and algebra
- Fit candidate models and evaluate using AIC to select best candidate model
- Evaluate and validate the top model(s)
- Interpret results, including description of all model parameters and what estimates mean including graphical and verbal summaries of the model predicitions
- Include an annotated R script or do this as an R markdown file