Why worry about assumptions?
Reproductive output of a rare plant

When we apply statistical models we entail a whole set of assumptions. For example, the use of linear regression models implies (i) normality, (ii) homogeneity of variance, (iii) independence, (iv) fixed measurement of the explanatory variable, and (v) correct model specification. In this session we will discuss how to verify these assumptions in a linear model and recognize the consequences of their violation. Quintana-Ascencio et al. (2003) collected demographic data of the rare and endemic plant *Hypericum cumulicola* in Archbold Biological Station, in Highlands, Florida. Reproduction is a critical vital rate contributing to determine population persistence so demographic studies of plants routinely characterize reproductive output of individuals with different sizes. Here, we will evaluate a linear regression model to predict number of reproductive structures of *Hypericum cumulicola* with different heights.

Figure 1. Vegetative and reproductive stages of *Hypericum cumulicola*

Enter the following command to load the *Hypericum cumulicola* data.

```
orig_data <- read.table("hypericum_data_94_07.txt", header=T)
```

This dataset contains, among other variables, height in cm of each plant (`ht_init`) and number of reproductive structures per individual (`rp_init`). For this demo we will focus on the regression of number of reproductive structures vs. height during 1994-1996. We will constrain the data to concentrate only on reproductive individuals collected during the first three years of study:

```
dt <- subset(orig_data, !is.na(ht_init) & rp_init > 0 & year < 1997)
height <- dt$ht_init
fruits <- dt$rp_init
id <- dt$tag
year <- dt$year
```

We can naively start analyzing these data evaluating a linear model of the relationship between height and fruits for reproductive individuals and hypothesizing that the number of fruits changes with height. This model, summarized below, indicates that number of fruits increases with height and explains 24% of the observed variation.

The underlying statistical model is:

$$\boxed{Number\ of\ fruits = \beta_1 + \beta_2 * height \qquad \epsilon \sim N(0, \sigma)}$$

The implementation and results in *R* are:

```
Call:
lm(formula = fruits ~ height)

Residuals:
   Min     1Q Median     3Q    Max 
-347.5  -81.4  -18.2   45.4 5620.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -269.3643    16.8864  -15.95   <2e-16 ***
height        11.2354     0.4416   25.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 240.7 on 2099 degrees of freedom
Multiple R-squared:  0.2357,    Adjusted R-squared:  0.2353 
F-statistic: 647.2 on 1 and 2099 DF,  p-value: < 2.2e-16
```

The plot of these two variables (Figure 2) for the first three years of the data shows that the number of fruits consistently increased with height through years, that the increase may be faster than a linear change, and that the uncertainty in the number of fruits augmented with height.

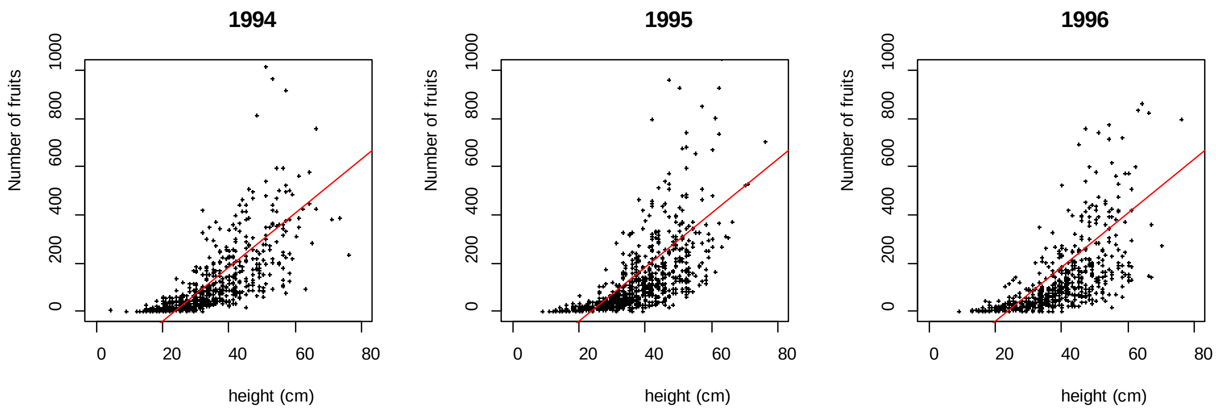

Figure 2. Plot of number of fruits as a function of plant height for *H. cumulicola* measured during 1994, 1995, and 1996 at Archbold Biological Station. The line in red is the linear model (y = -269.36 + 11.23x). The ordinate access was truncated at 1000.

A plot of the residuals of the linear model shows a clear pattern of increasing variance for the values with height (Figure 3).

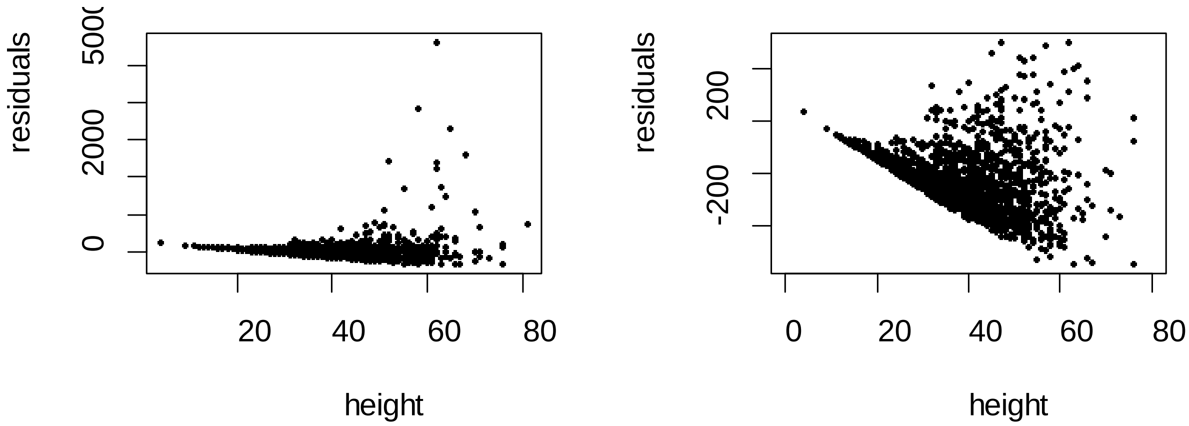

Figure 3. Plots of model residuals as a function of height. The plot at the right enlarges the area where the data is concentrated.

After inspecting the residuals in the previous plots we recognize the non-linear relationship between height and number of fruits and evaluate a power model. To accomplish this we implement a linear model of the natural logarithm of both variables.

The underlying statistical model is:

$$\log(\boldsymbol{Number\ of\ fruits}) = \boldsymbol{\beta_1} + \boldsymbol{\beta_2} * \log(\boldsymbol{height}) \qquad \boldsymbol{\epsilon \sim N(0, \sigma)}$$

The implementation and results in *R* are:

```
Call:
lm(formula = log(fruits) ~ log(height))

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2555 -0.4869  0.0289  0.5258  4.8065 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.35037    0.17916  -41.03   <2e-16 ***
log(height)  3.23867    0.05043   64.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8164 on 2099 degrees of freedom
Multiple R-squared:  0.6627,    Adjusted R-squared:  0.6625 
F-statistic:  4124 on 1 and 2099 DF,  p-value: < 2.2e-16
```

Notice the different values for the Betas between models.

This model characterizes the data much better (Figure 4 – red line) and the even distribution of the residuals indicates a more reliable model (Figure 5 – left). Notice that an outlier remains among the small plants. This plant has 4 cm and 7 fruits; an exceptional reproductive output for this size.

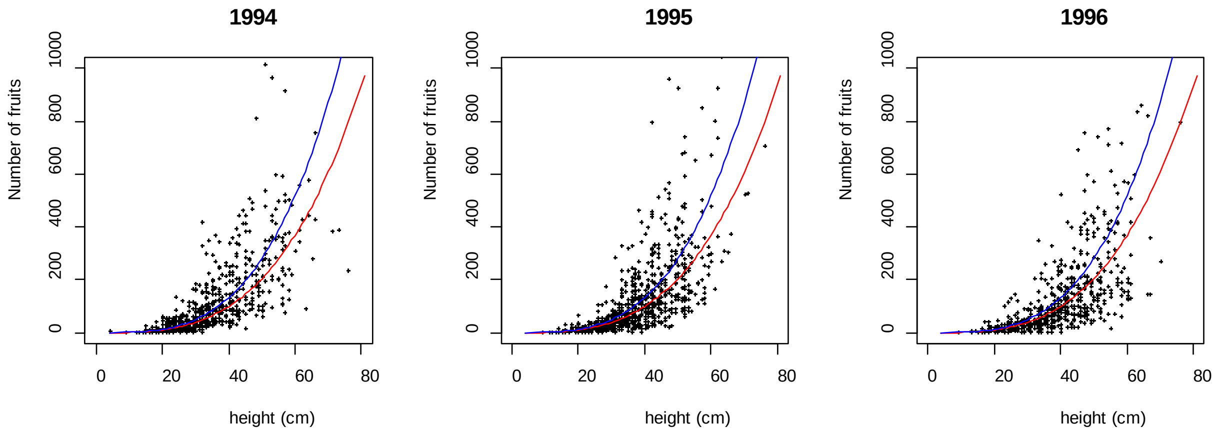

Figure 4. Plot of number of fruits as a function of plant height for *H. cumulicola* measured during 1994, 1995, and 1996 at Archbold Biological Station. The line in red is the power model using MLS ($y = -7.35x^{3.24}$). The line in blue is the power model using likelihood ($y = -7.60x^{3.37}$). The ordinate access was truncated at 1000.

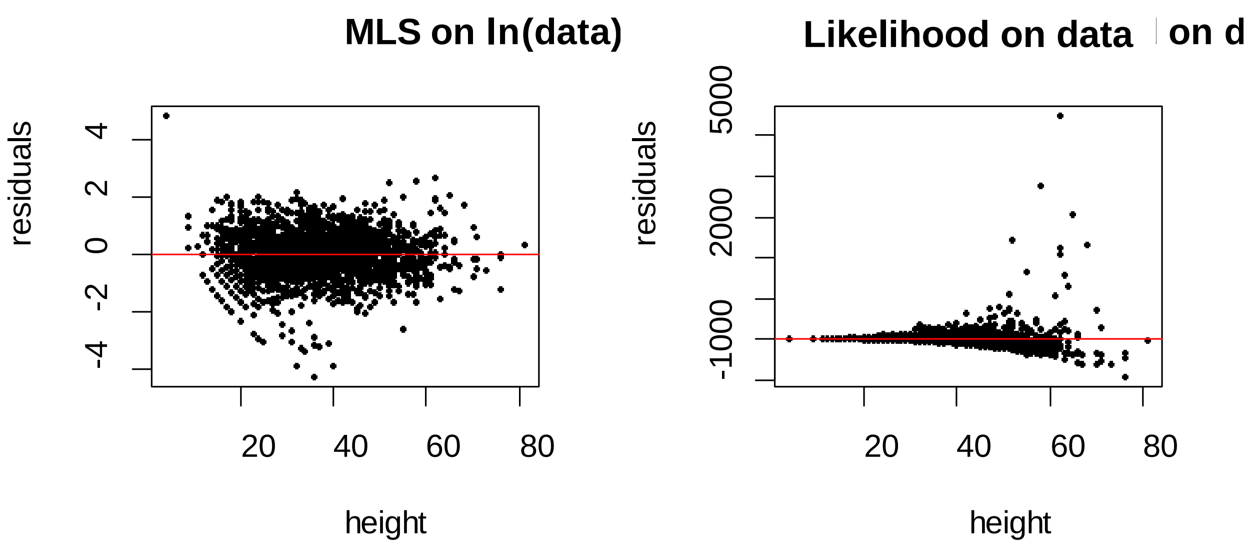

Figure 5. Plot of model residuals as a function of height for both power models.

When we transformed the data we changed the model that we hypothesized fits the relationship; in this case from a linear to a power model. The estimation of the parameters of the power model varies depending on whether we use **likelihood** or the **method of least squares** to approximate the solution. For the previous power model we used the method of least squares on the logarithm transformed data. Below we approximate the solution of this model using likelihood on the natural scale of the data with the function `nlsLM()`of the R package `minpack.lm`

The underlying statistical model is:

$$\boxed{y = \ a_1 * height^{a_2} \qquad \epsilon \sim N(0, \sigma)}$$

The implementation and results in *R* are:

```
library(minpack.lm)
Power <- nlsLM(fruits~a1*height^a2,start=list(a1=1,a2=1),data=dt,algorithm="LM")
summary(Power)
```

The output of this model is

```
Formula: fruits ~ a1 * height^a2

Parameters:
    Estimate Std. Error t value Pr(>|t|)
a1 0.0005229  0.0002615   1.999   0.0457 *
a2 3.3723063  0.1246992  27.044   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 232.2 on 2099 degrees of freedom

Number of iterations to convergence: 36
Achieved convergence tolerance: 1.49e-08
```

Notice that the parameters of the nonlinear model using likelihood (α1: ln(0.0005) = -7.56 and α2: 3.37) are commensurate, but different, from those of the MLS model (α1: -7.35 and α2: 3.24). Because the power model approximated with likelihood is evaluated on the same scale as the linear model we can compare their relative information using AIC. The Power model is more informative (AIC= 28857) than the linear one (AIC = 29008).

We plot both power models in Figure 4 (blue line) and their residuals in Figure 5 (right). Notice the more central location of the likelihood power model around larger heights. This model emphasizes the increasing variation of number of fruits with increasing height. To decide which model you prefer consider that the model estimated with likelihood does not change the distribution of the variance along the mean, as happens with the transformed data.

The plot of residuals of the MLS power model shows reduced variation for plants smaller than 20 cm and those larger than 60 cm. This may be a problem because these regions of the model do not meet homogeneity of variance, an assumption of the MLS method. In several R procedures, weights can be specified to indicate that different observations have different variances (with the values in weights being inversely proportional to the variances (R documentation). Estimates of the coefficients are adjusted accordingly. The model below was estimated with weights by height.

The underlying statistical model is:

$$\boxed{\log(Number\ of\ fruits) = \ \beta_1 + \ \beta_2 * \log(height) \qquad \epsilon \sim N(0, \sigma) \qquad \sigma \sim w * height}$$

The implementation and results in *R* are:

```
Call:
lm(formula = log(fruits) ~ log(height), weights = height)

Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-25.5362  -2.8483   0.1603   3.0937  21.0921 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.42995    0.19962  -37.22   <2e-16 ***
log(height)  3.26104    0.05452   59.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.796 on 2099 degrees of freedom
Multiple R-squared:  0.6302,    Adjusted R-squared:   0.63 
F-statistic:  3577 on 1 and 2099 DF,  p-value: < 2.2e-16
```

The coefficients of this model were very similar to the previous model with fixed variance (Figure 6) and we can use AIC to evaluate if fixed or variable variance are warranted. In this case the model with fixed variance has smaller AIC (5113.77) compared to the model with variable variance (5127.51) indicating that in this case the correction was not necessary.

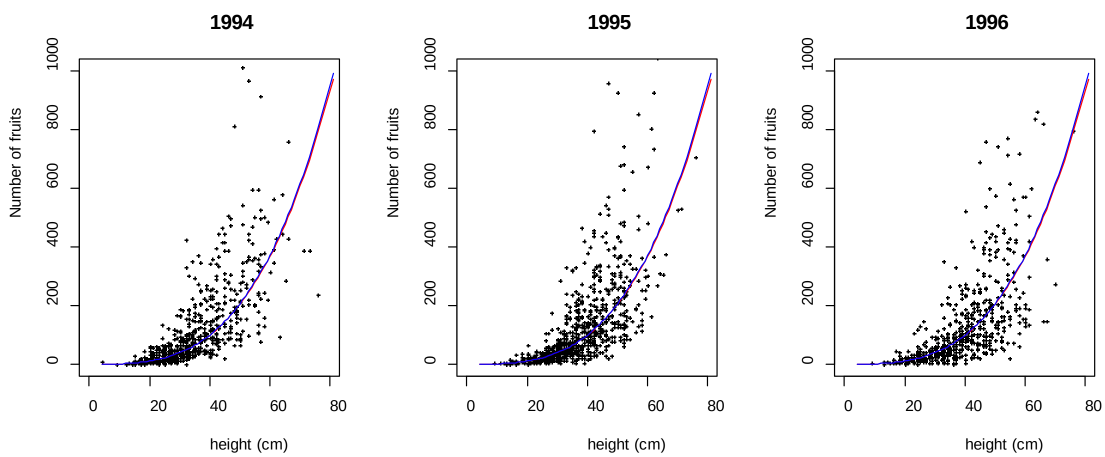

Figure 6. Plot of number of fruits as a function of plant height for *H. cumulicola* measured during 1994, 1995, and 1996 at Archbold Biological Station. The line in blue is the MLS power model with constant variances ($y = -7.35x^{3.24}$). The line in red is the MLS power model assuming variable variance ($y = -7.43x^{3.26}$). The ordinate access was truncated at 1000.

While the residual of the linear model were highly skewed to the left, the residuals for both the power models were normally distributed (Figure 7).

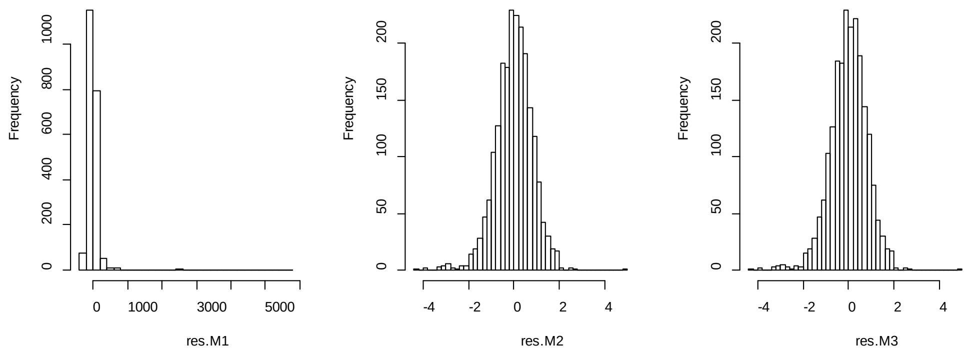

Figure 7. Histogram of the residuals of the general linear model, the MLS and the likelihood power models (with constant variance).

There are more aspects that we should inspect to validate the models. These data represent several years of work and there is a certain level of non-independence in the data. For example Population 63 had several individuals that were measured consecutively several times. The estimate of the coefficient of the slope of the power model based on those 32 individuals, results in biased estimates when compared to estimates based in random samples of the same size (Figure 8 and 9).

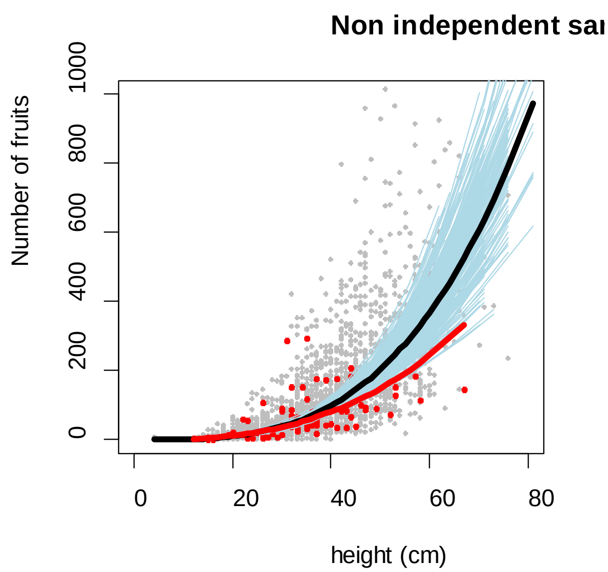

Figure 8. Power models based on 500 random samples of 32 individuals (in blue) compared to the overall estimate (in black) and the estimate of 32 individuals repeatedly sampled in Population 63 (in red).

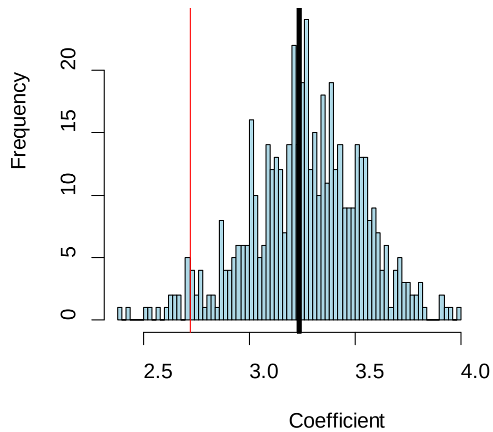

Figure 9. Histogram of the coefficients of the models with all data (central value in black) and with Population 63 (central value in red).

Finally, here, we assume a normal distribution for the errors, we could have evaluated alternative distributions for these data since number of fruits is a count. We will come back to this concern in another demo.

Data sources:

Quintana-Ascencio, P. F., E. S. Menges, and C. Weekley. 2003. A fire-explicit population viability analysis of *Hypericum cumulicola* in Florida rosemary scrub. Conservation Biology, 17: 433-449.