Why worry about assumptions?
Reproductive output of a rare plant

When we apply statistical models we consent to a whole set of assumptions. For example, the use of linear regression models implies assumptions on (i) the likelihood distribution, often assumed to be the normal distribution, (ii) variance heterogeneity, often assumed homogeneous, (iii) independence of the data, (iv) variation of the measurement of the explanatory variable, mostly assumed fixed, and (v) correct model specification. In this session we will discuss how to verify some of these assumptions in a linear model and recognize the consequences of their violation. Quintana-Ascencio et al. (2003, 2018, 2019) collected demographic data of the rare and endemic plant *Hypericum cumulicola* at Archbold Biological Station, in Highlands County, Florida. Reproduction is a critical vital rate contributing to determine population persistence, so demographic studies of plants routinely characterize the reproductive output for individuals of 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, the height in cm of each sampled plant (ht_init) and the number of reproductive structures per individual (rp_init). For this demo we will focus on the regression of number of reproductive structures as a function of height in 14 populations in Florida Rosemary Scrub patches at Archbold Biological Station during August 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$tag
bald <- dt$bald
year <- dt$year
```

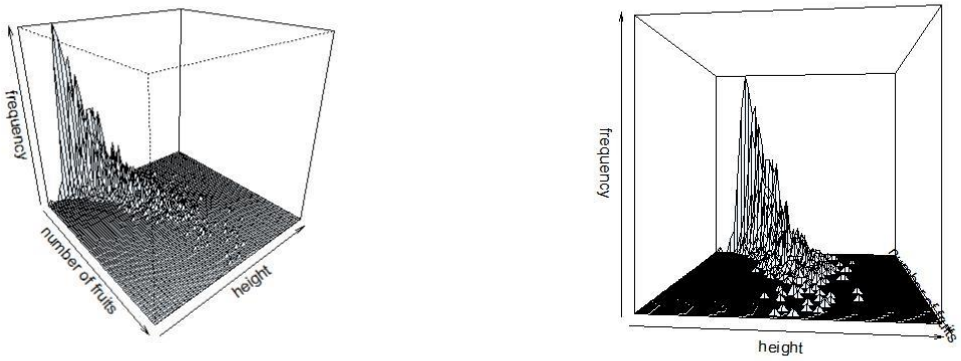

Figure 2. Number of number of fruits as a function of plant height for *H. cumulicola*

We can naively start analyzing these data by 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:

$$Number\ \ of fruits = \beta_1 + \beta_2 * height \quad Likelihood\ N(\mu, \sigma)$$

The implementation and results in *R* and *Stan* are:

```
model1 <- ulam(
  alist(
    rep_structures ~ dnorm(mu,sigma),
    mu <- a + b*height,
    a ~ dnorm(0,500),
    b ~ dnorm(0,100),
    sigma ~ dunif(0,400)
  ),
  data = dt,chains =3
)
```

```
precis(model1,digits=1)
         Mean     sd    5.5%   94.5% n_eff Rhat
a      -269.7   17.3  -296.4  -242.0   994    1
b        11.2    0.5    10.5    12.0   958    1
sigma   240.8    3.7   235.1   246.8  1602    1
```

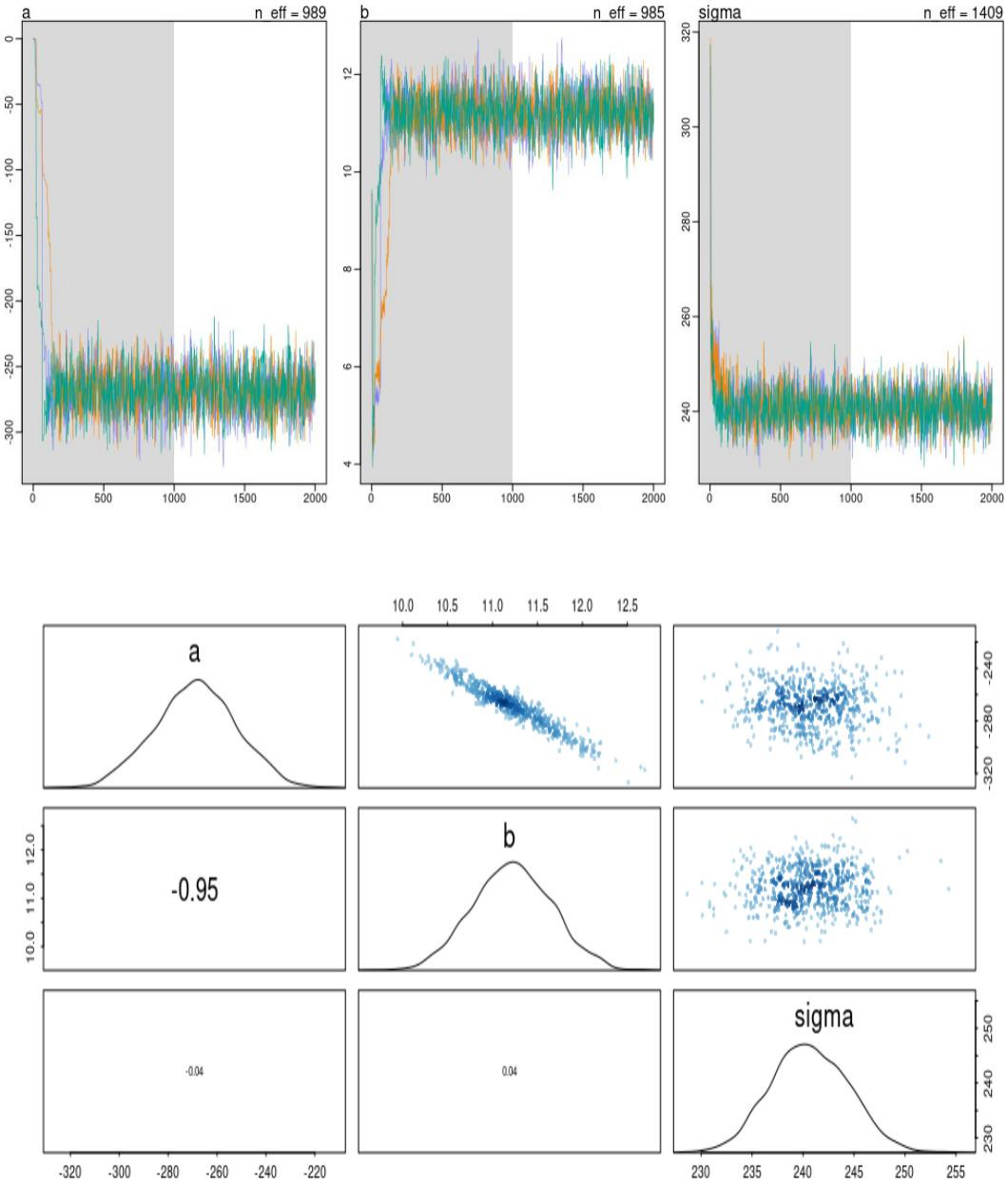

Figure 3. Plots of a model of number of fruits as a linear function of plant height for *H. cumulicola* measured at Archbold Biological Station. The plot on top shows the sequence of the generating chains and at the bottom the distribution and the correlation of the parameters of the model.

The plot of these two variables for these data shows that the number of fruits consistently increased with height, that the increase may be faster than a linear change, and that the uncertainty in the number of fruits increased with height.

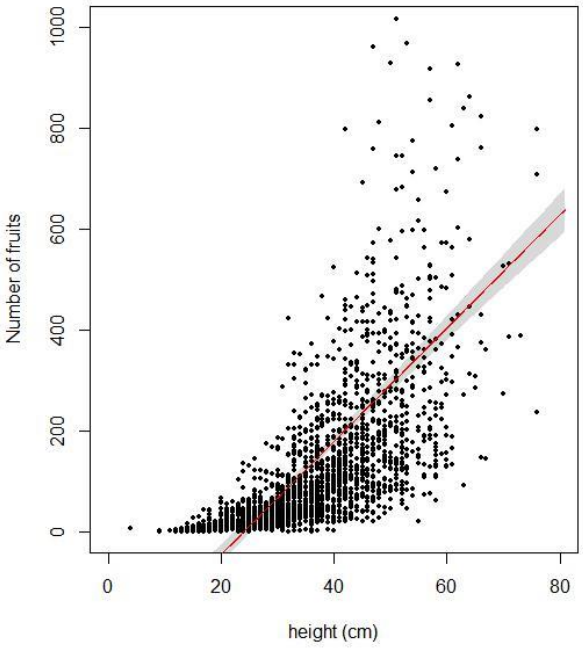

Figure 4. Plot of number of fruits as a function of plant height for *H. cumulicola* measured during 1994, 1995, and 1996 at Archbold Biological Station. The line in red is the linear model (y = -269 + 11x). The polygon is the 95 % credibility interval. The ordinate access was truncated at 1000.

After inspecting the residuals in the previous plot, we recognize the non-linear relationship between height and number of fruits. Particularly worrisome were the predicted negative number of fruits for plants smaller than 20 cm. We decide to evaluate a power model. To accomplish this, we implement a linear model of the natural logarithm of both variables.

The underlying statistical model is:

$$log(Numberoffruits) = \beta_1 + \beta_2 * log(height) Likelihood\ N(\mu, \sigma)$$

The implementation and results in *R* and *Stan* are:

```
dt$lnrep <- log(dt$rep_structures)
dt$lnht <-  log(dt$height)
sdlnfruits <- sd(dt$lnrep)

model2 <- ulam(
  alist(
    lnrep ~ dnorm(mu,sigma),
    mu <- a + b*lnht,
    a ~ dnorm(0,10),
    b ~ dnorm(0,10),
    sigma ~ dunif(0,10)
  ),
  data = dt,chains =3
)
```

```
> precis(model2,digits=1)

       Mean      sd     5.5%   94.5%   n_eff   Rhat
a      -7.3     0.2     -7.6    -7.1     784      1
b       3.2     0.0      3.1     3.3     786      1
sigma   0.8     0.0      0.8     0.8    1133      1
```

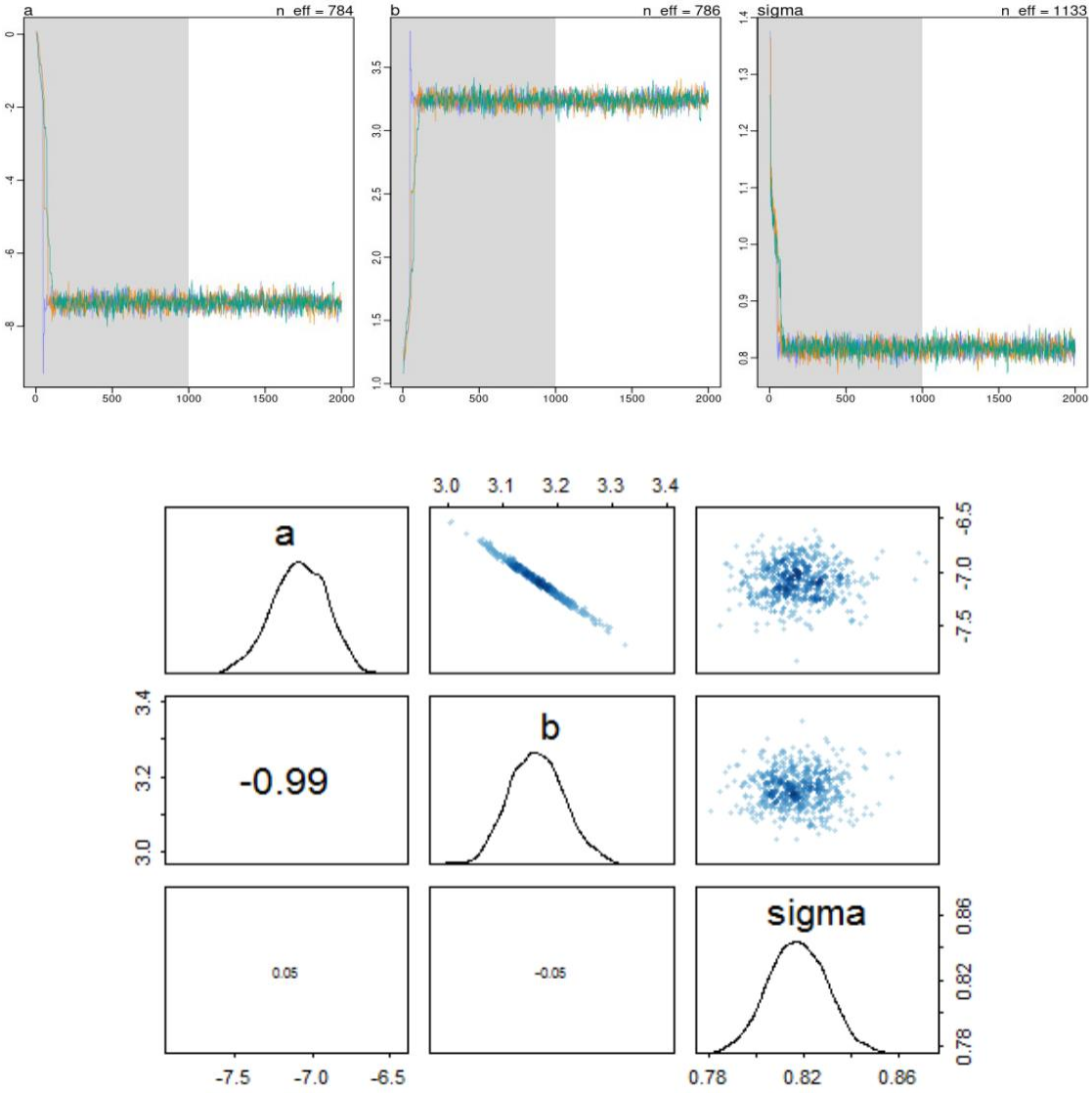

Figure 5. Plots of a model of the logarithm of number of fruits as a function of the logarithm of plant height for *H. cumulicola* measured at Archbold Biological Station. The plot on top shows the sequence of the generating chains and at the bottom the distribution of the posterior parameters and their correlations.

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

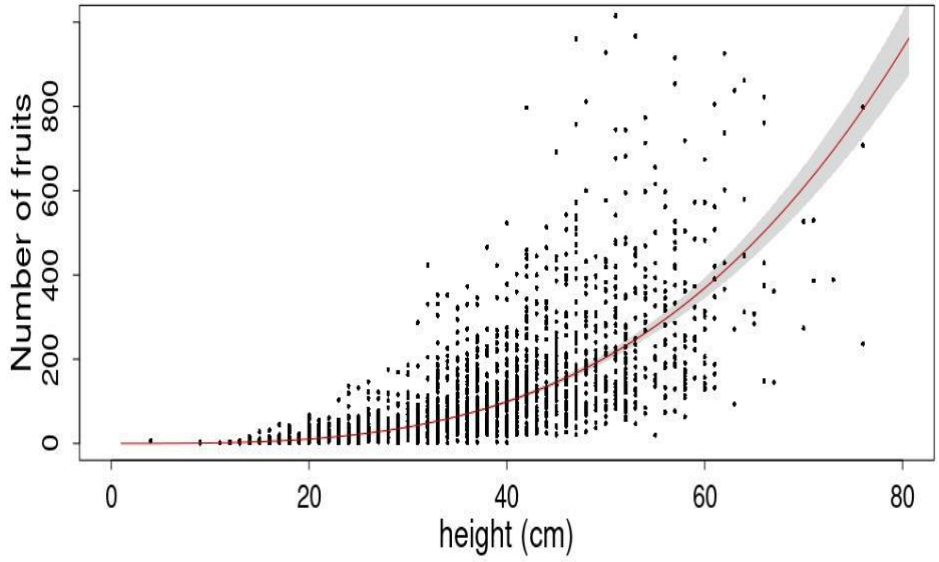

Figure 6. Plot of number of fruits as a function of plant height for *H. cumulicola* measured during 1994, 1995, and 1996 at Archbold Biological Station. The line in red is the power model using logarithm transformations ($y = \exp(-7.1) * x_{3.2}$).

When we transform the data, we actually change the model hypothesized to fit 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 **logarithms** or a **non-linear model** to approximate the solution. For the previous power model we used the method on the logarithmic transformed data. Below we approximate the solution of this model using a non-linear model on the natural scale of the data.

The underlying statistical model is:

$$y = a_1 * height^{a_2} \;\; Likelihood\; N(\mu, \sigma_i)$$

The implementation and results in *R* and *Stan* are:

```
model3 <- ulam(
  alist(
    rep_structures ~ dnorm(mu,sigma),
    mu <- exp(a)*height^exp(b),
    a ~ dnorm(0,10),
    b ~ lognormal(0,1),
    sigma ~ dunif(0,400)
  ),
  data = dt,chains =3
)
```

```
> precis(model3,digits=2)

          Mean      sd     5.5%    94.5%  n_eff  Rhat
a        -7.49    0.48    -8.29    -6.73    628     1
b         1.21    0.04     1.04     1.16    616     1
sigma   232.89    3.54   227.51   238.66   1162     1
```

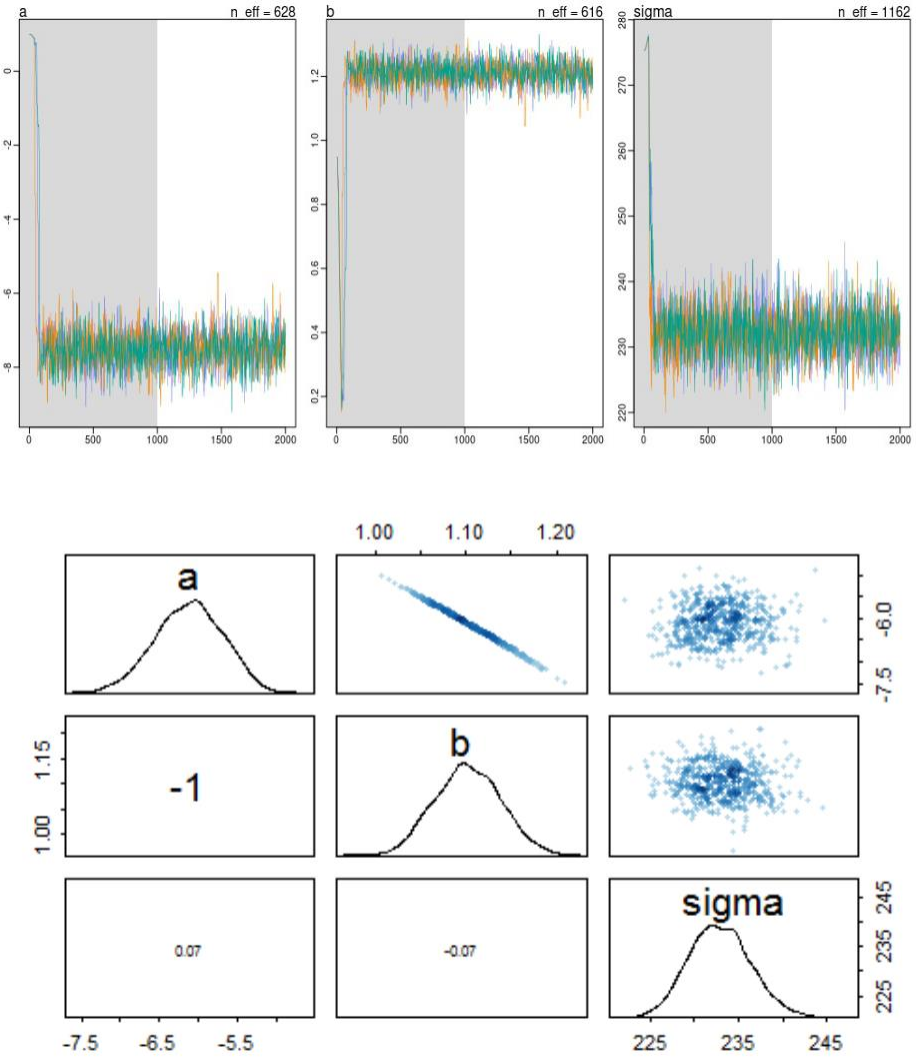

Figure 7. Plots of a non-linear model of number of fruits as a function of plant height for *H. cumulicola* measured at Archbold Biological Station. The plot on top shows the sequence of the generating chains and at the bottom the distribution and the correlation of the parameters of the model.

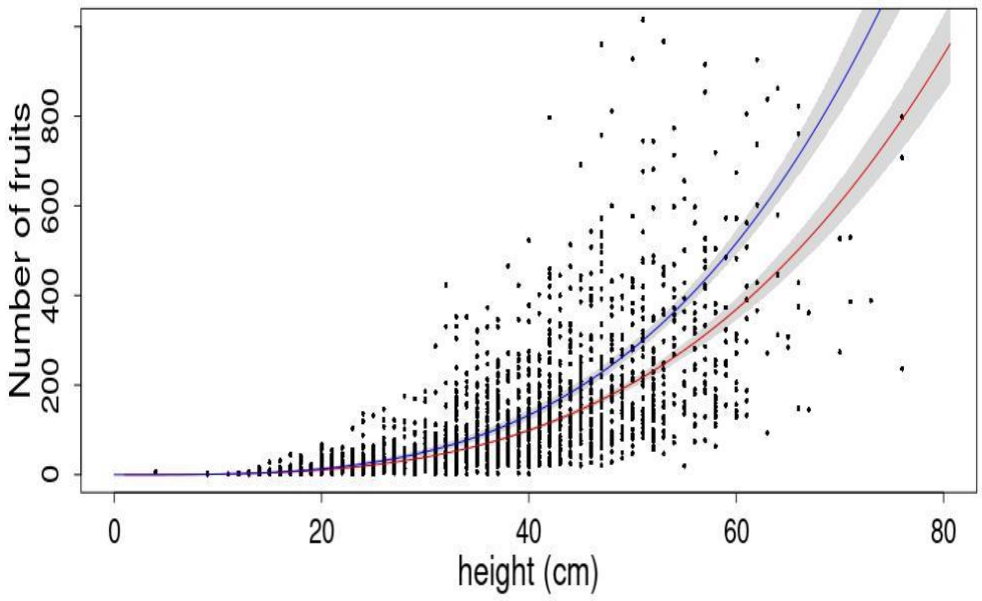

Figure 8. Plot of number of fruits as a function of plant height for *H. cumulicola* measured during 1994, 1995, and 1996 at Archbold Biological Station. The line in red is the power model using logarithm transformations ($y = \exp(-7.1)*x_{3.2}$). The line in blue is the power model directly on the data ($y = \exp(-6.1)*x_{1.1}$). The ordinate access was truncated at 1000.

Notice that the parameters of the model with logarithms (α1 = -7.3 and α2 = 3.2) are commensurate, but different, from those of the non-linear model (α1 = -7.49 and α2 = 1.21). Because the power model as non-linear model is evaluated on the same scale as the linear model we can compare their relative information. The Power model is definitively more informative than the linear one.

```
> compare(model1,model3)
           WAIC      SE  dWAIC    dSE pWAIC weight
model3  28994.0 763.59    0.0     NA 114.0      1
model1  29137.9 767.60  159.5 38.92 119.7      0
```

We plot both power models in Figure 8. Notice the more central location of the non-linear power model around larger heights. These models emphasize the increasing variation of number of fruits with increasing height. To decide which model do you prefer to consider for your inference, notice that the model estimated as non-linear model have a different distribution of the variance at each point in the x axes compared with the transformed data. Additionally, observe the distribution of number of reproductive structures in Figure 9.

```
hist(dt$rep_structures[dt$rep_structures<2000],100)
```

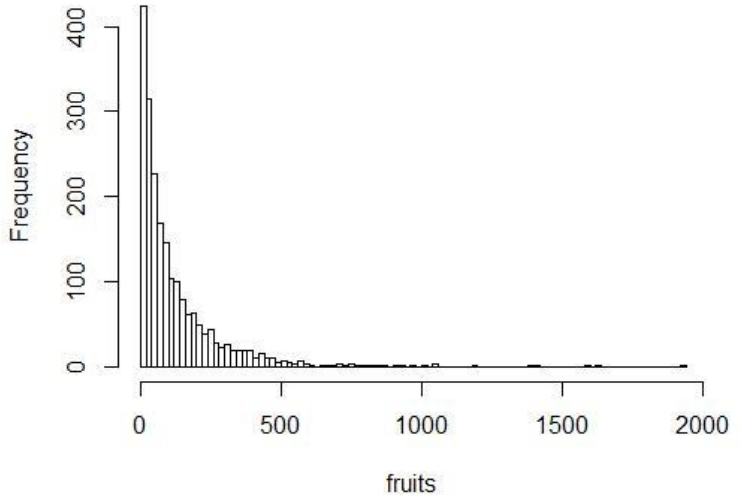

Figure 9. Histogram of the frequency of number of fruits per plant

Which one do you prefer and why?

There are more aspects that we should inspect to validate the models. These data represent several years of work in several populations, and there is a certain level of non-independence in the data because of these factors. In the next model we allow for random effects on the intercept by year (i) and population (j). We will explore more complicated mixed models later in the course.

The underlying statistical model is:

$$log(Numberoffruits) = \alpha_i + \gamma_j + \beta_1 + \beta_2 * log(height)$$
$$Likelihood\ N(\mu, \sigma)$$

The implementation and results in *R* and *Stan* are:

```
model4 <- map2stan(
  alist(
    lnrep ~ dnorm(mu,sigma),
    mu <- a  + y[yi] + p[pj] + b*lnht,
    a ~ dnorm(0,10),
    y[yi] ~ dnorm(0,sigmay),
    p[pj] ~ dnorm(0,sigmap),
    b ~ dnorm(0,10),
    sigmay ~ dcauchy(0,1),
    sigmap ~ dcauchy(0,1),
    sigma ~ dcauchy(0,1)
  ),
  data = dt,chains =3
)
```

```
> precis(model4,digits=1,depth=2)

        Mean      sd    5.5%  94.5%  n_eff Rhat
a       -8.0     0.4    -8.4   -7.5    228    1
y[1]     0.1     0.3    -0.2    0.4    170    1
y[2]     0.0     0.3    -0.3    0.3    168    1
y[3]    -0.1     0.3    -0.4    0.2    170    1
p[1]    -0.3     0.1    -0.5   -0.1   1831    1
p[2]     0.1     0.1    -0.1    0.2   1526    1
p[3]    -0.1     0.1    -0.3    0.0   1371    1
p[4]     0.1     0.1     0.0    0.3   1353    1
p[5]     0.0     0.1    -0.1    0.2   1363    1
p[6]    -0.1     0.1    -0.3    0.1   1599    1
p[7]     0.2     0.1     0.0    0.3   1391    1
p[8]     0.2     0.1     0.0    0.3   1461    1
p[9]    -0.3     0.1    -0.4   -0.1   1606    1
p[10]   -0.1     0.1    -0.3    0.0   1580    1
p[11]    0.7     0.1     0.5    0.9   1521    1
p[12]    0.0     0.1    -0.1    0.2   1459    1
p[13]   -0.2     0.1    -0.3    0.0   1342    1
p[14]   -0.1     0.1    -0.3    0.0   1426    1
b        3.4     0.1     3.3    3.5   3498    1
sigmay   0.3     0.4     0.1    0.8    279    1
sigmap   0.3     0.1     0.2    0.4   4170    1
sigma    0.8     0.0     0.8    0.8   5806    1
```

When we compare their relative information, the mixed model is definitively more informative than the non-mixed one. This means that the change in number of fruits as a function on plant size depends on the population evaluated.

```
> compare(model2,model4)
           WAIC     SE  dWAIC    dSE  pWAIC weight
model4   4931.8  99.48    0.0     NA   20.2      1
model2   5115.5  92.35  183.7  28.89    4.7      0
```

There are more considerations that we will explore in later sessions of this class.

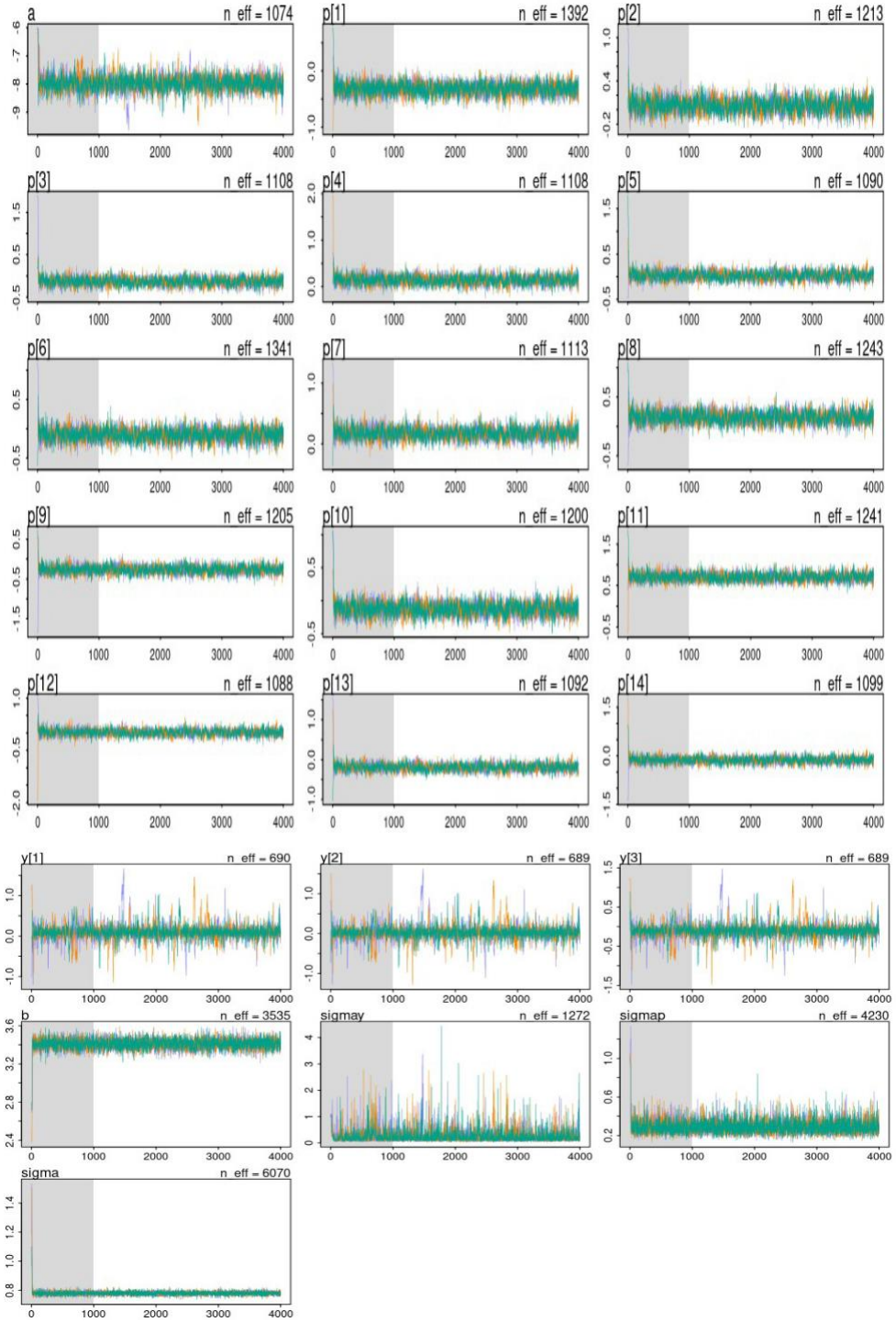

Figure 10. Plots of a mixed model of number of fruits as a function of plant height for *H. cumulicola* measured at Archbold Biological Station. The plot shows the sequence of the generating chains.

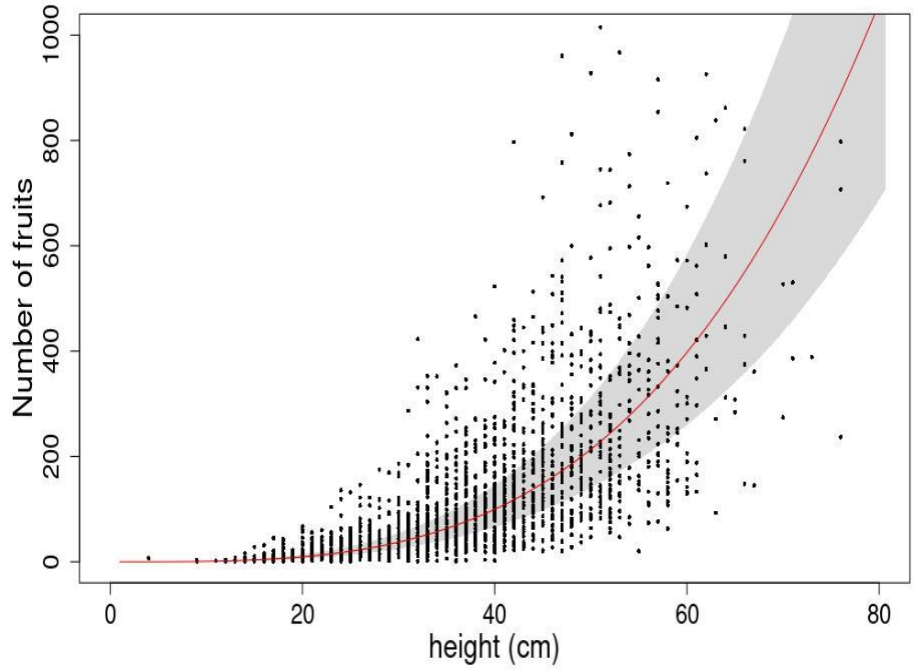

Figure 11. Plot of number of fruits as a function of plant height for *H. cumulicola* measured during 1994, 1995, and 1996 at Archbold Biological Station. The line in red is the average power mixed model using logarithm transformations with random variation by year and population. Notice the wider 95% credibility interval when we incorporate estimates of the variation by these factors.

NOTE: all the materials for this demo can be found at:

https://sciences.ucf.edu/biology/d4lab/methods-2/

References

McElreath, R.M. 2016. Statistical Rethinking: a Bayesian course with examples in R and Stan. Chapman and Hall.

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.

Quintana-Ascencio, P.F. Koontz, S., Smith, V., David, A., Sclater, V. L. & E. S Menges. 2018. Predicting landscape-level distribution and abundance: Integrating demography, fire, elevation, and landscape habitat configuration. Journal of Ecology, 106: 2395-2408

Quintana -Ascencio, P.F. Koontz, S.M., Ochocki, B., Sclater, V. L., López-Borghesi, F., Li, H. & E. S Menges. 2019. Assessing the roles of seed bank, seed dispersal and historical disturbances for metapopulation persistence of a pyrogenic herb. Journal of Ecology, 107: 2760-2771.