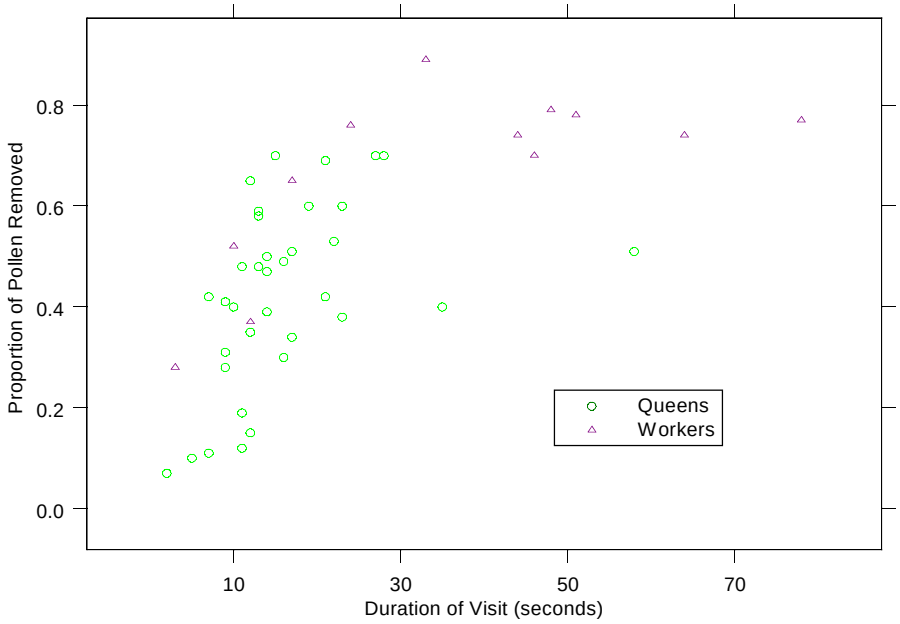
BusyBees

As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the the proportions of pollen removed by bumble bee queens and honeybee workers pollinating a species of lily (from page 75)

The data consist of the Proportion of Pollen removed (removed) the duration of visit in secs (duration) and a code for queen or worker (code = 1 = queen, code = 2 = worker)



Question:

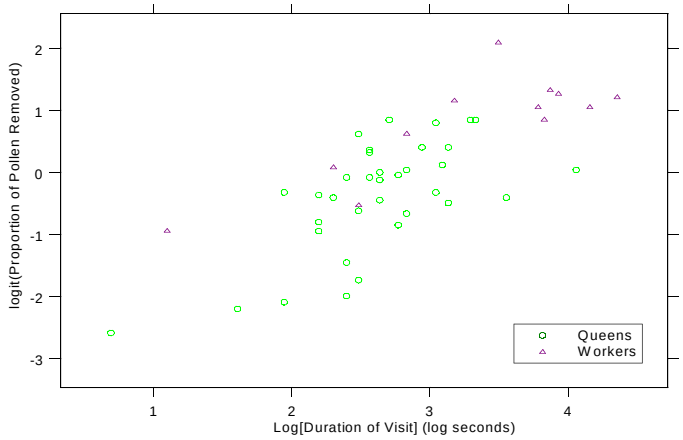
Is the relationship between duration and proportion of pollen removed the same for queen bumble bees and worker honey bees?

Clearly in both groups there is a nonlinear relationship between duration and proportion of pollen removed.

To answer this question it is useful to find out if there is some transformation so that the relationship is linear.

When measuring proportions, a useful transformation is the logit transformation, so that logit(P) = log(p/(1-p)). This is fine as long as P is never 0 or 1. The logit(P) can take on any value from - infinity to +infinity. This often makes the normal assumption more reasonable. A logit(.5) = log(1) = 0.

Use logit(removal) and trial and error to find that log(duration) makes the relationship between logit(removal) and log(duration) look close to linear for both groups. (an exercie for the energetic student :-)



Model Derivation:

If in both groups there is a linear relationship between log(duration) and logit(remove) with possibly different intercepts and slopes, then we have the model

$\mu(\text{logit}(R)_i \mid \text{duration, code = queens}) = \beta_0 + \beta_1 \log(\text{duration}) + \varepsilon_i$ for i = 1, ... 35

$\mu(\text{logit}(R)_i \mid \text{duration, code = workers}) = \alpha_2 + \alpha_3 \log(\text{duration}) + \varepsilon_i$ for i = 36, ... 47

The null hypothesis that there is no difference between workers and queens would imply that the slopes of the two lines are equal and that the intercepts of the two lines are the same. One way to fit this type of model is to introduce the idea of **dummy variables.**

We create a new variable I that is 1 for workers and 0 for queens, which is called a dummy variable We can fit one combined regression model using the dummy variable and an interaction, which is the product of the dummy variable I times the explanatory variable log(duration).

$$\mu(\text{logit}(R) \mid \text{duration, I}) = \beta_0 + \beta_1 \log(\text{duration}) + \beta_2 I + \beta_3 \log(\text{duration})*I + \varepsilon_i \text{ for i = 1, ... 47}$$

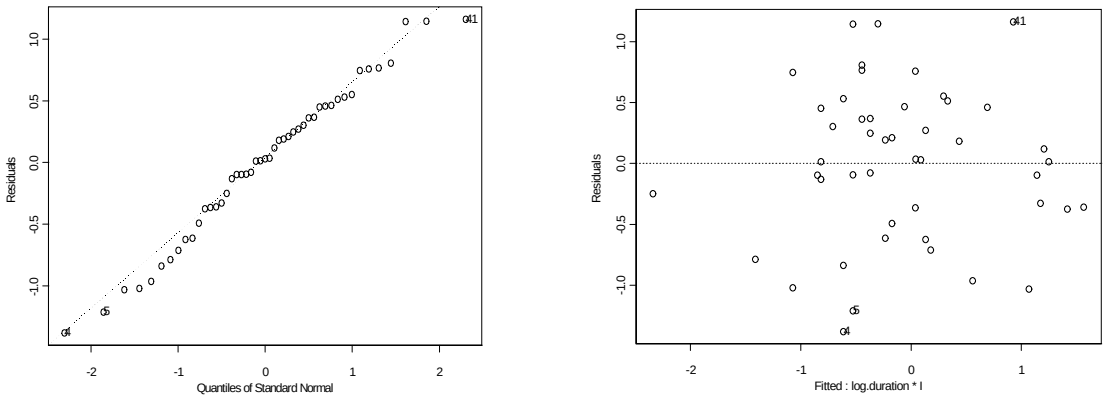
when I = 0 (queens) we have just the top row

when I = 1 (workers) we have that the intercept is $\beta_0 + \beta_2$ and that the slope is $\beta_1 + \beta_3$.

In this form the hypothesis of common regression lines coresponds to Ho: $\beta_2 = \beta_3 = 0$.

The next page contains the output from fitting the **multiple regression** model with log.duration, I, and I*log.duration (the ineraction in S-Plus is represented as I:log.duration).

Is model OK? Just as in Simple Linear Regression check residual plots:



The residual plots and normal probability plots look o.k., so at this point we can consider testing hypotheses.

As you will see there are potentially many hypotheses of interest and many F-stats and t-stats in the output. You should know how to interpret:

1) The **overall F-test** in the regression output is testing the null hypothesis that ALL regression coefficients (except the inercept) in the model are 0 against the alernative hypothesis that AT LEAST ONE regression coefficient is non-zero. This is the starting place for any analysis. If you do not reject the null, there is no point continuing. Period. The F-test is based on a F-statistic with p - 1 df, and n - p - 1 df where p = the number of variables in the model. If we reject, then we can look at individual t-tests

2) An **individual t-test** is testing the null hypothesis that an individual coefficient is 0 vs the alernative that the coefficient is non-zero. This assumes that the other variables are included in the model, so you can't use this for simplifing the model by dropping 2 variables at a time that both have large p-values. You can refit the model after dropping the variable with the most insignificant coefficient. Repeat as needed..

3) Once all variables are significant interpret the final model.

Multiple Regression Output

***** Linear Model 1 - Full Model all variables *****

```
Call: lm(formula = logit.remove ~ log.duration + I + I:log.duration, data =
      Ex0326, na.action = na.omit)
Residuals:
       Min         1Q     Median         3Q     Max
 -1.380351 -0.3698604 0.03070024 0.4551561 1.16113
```

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.0389525	0.5114996	-5.9412613	0.0000004
log.duration	1.0120846	0.1902043	5.3210408	0.0000035
I	1.3770009	0.8721766	1.5788096	0.1217089
I:log.duration	-0.2708987	0.2816798	-0.9617256	0.3415647

Residual standard error: 0.6525275 on 43 degrees of freedom
Multiple R-Squared: 0.6150824

F-statistic: 22.90407 on 3 and 43 degrees of freedom,
the p-value is 5.151446e-009

***** Linear Model 2 - Main Effects Only *****

```
Call: lm(formula = logit.remove ~ log.duration + I, data = Ex0326, na.action
       = na.omit)
Residuals:
       Min         1Q     Median         3Q      Max
 -1.408519 -0.4962697 0.08814885 0.4359778 1.155616
```

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.7145967	0.3842293	-7.0650433	0.0000000
log.duration	0.8885650	0.1401728	6.3390679	0.0000001
I	0.5696676	0.2364278	2.4094781	0.0202261

Residual standard error: 0.6519706 on 44 degrees of freedom
Multiple R-Squared: 0.6068029

F-statistic: 33.95159 on 2 and 44 degrees of freedom,
the p-value is 1.206218e-009