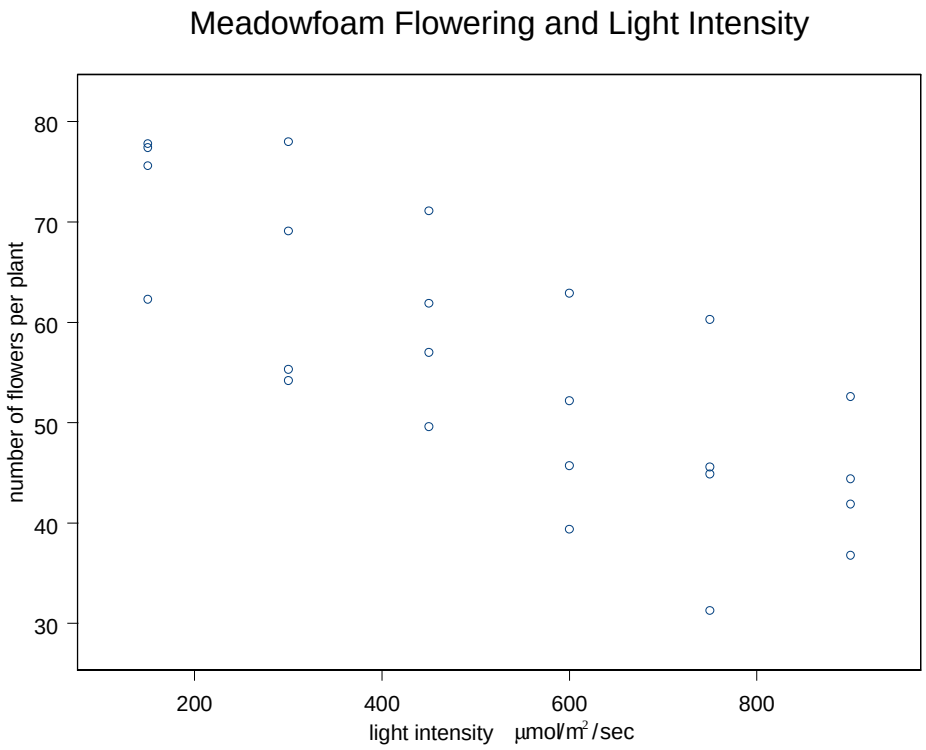
18. Simple Linear Regression

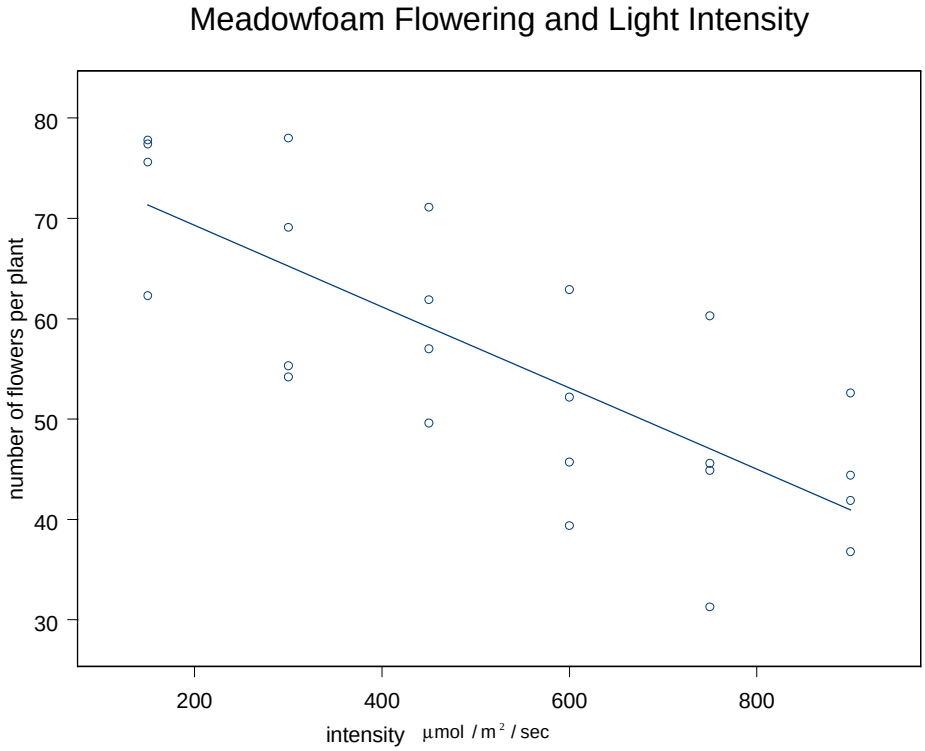


Is flowering affected by light intensity? Phrased another way, does the mean number of flowers appear to differ for different light intensities? Sounds like an ANOVA type of question, right? We discuss another method call regression which is similar to ANOVA, and, at some fundamental level, is no different. We continue to use some of the notation and terminology introduced in the last chapter on correlation.

18.1. See POB

18.2. The Model

18.2.1. The Population Regression Line



At each value of light intensity where we measure data, we can think of a population of flower counts per plant. In ANOVA we specified the means of the "y-variable" at each of the levels of the factor variable as

$$\mu_1, \mu_2, \ldots, \mu_k$$

if we had k levels of our factor. Here, intensity is our "factor", but we don't call it a factor in regression. We have various other names (explanatory, independent, predictor,…). We can use a different notation to indicate the mean number of flowers at a particular value of intensity:

$$\mu_{y|x}$$

("mean of y given the value x of the explanatory variable"). Since both our response and explanatory variables are numerical, we might specify a model for the mean as

$$\mu_{y|x} = \alpha + \beta x$$

This specifies the *mean* of our response variable, y, at any value of our explanatory variable, x. Our model for the response variable, y, may be written as

$$y = \mu_{y|x} + \varepsilon = \alpha + \beta x + \varepsilon$$

- $\mu_{y|x} = \alpha + \beta x$ is the (unknown) population regression line of y on x
- α is the y-intercept—the y value at which the line assumes when x=0.
- β is the slope of the line—the rate of change of y with x or the change in y given a 1 unit change in x
- ε is called the error—the distance y lies from the regression line
- simple linear regression

Some common objectives:

- Estimate the regression coefficients
- Determine the statistical significance of the regression coefficients (usually not interested in the intercept), i.e., test if a coefficient is, say, zero, or not
- Predict a mean value of y at a certain x value and give a confidence interval for the mean

- Predict a value of y at a certain x value and give a confidence interval

Some useful assumptions:

- x is measured without error
- at each x, y is normal with (population) mean $\mu_{y|x}$ and (population) standard deviation $\sigma_{y|x}$
- $\sigma_{y|x}$ does not change with x (i.e., constant variance across the value of x (like in the ANOVA case)
- the regression model is correct, i.e., $\mu_{y|x} = \alpha + \beta x$ is correctly specified
- the outcomes of different y values are independent

The above specifies a model for the (unknown) population mean and has some (unknown) parameters to estimate. How do we estimate the parameters?

18.2.2. The Method of Least Squares

Consider n independent observations of y at some x values, i.e., our data:

$$(y_1, x_1), \ldots, (y_n, x_n)$$

Note that we may have multiple observations of y at the same x value. Our model says:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

We want to fit a line to our data points in some "optimal" way. One criterion is given by the method of least squares which, in

a nutshell, is to minimize the sum of squared (vertical) distances from the line to each y_i:

$$\min_{\alpha,\beta}\left(\sum_{i=1}^{n}[y_i-(\alpha+\beta x_i)]^2\right)$$

After a bit of differentiating we find the least squares estimators of alpha and beta:

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\text{, or equivalently}$$

$$\hat{\beta}=r\frac{s_y}{s_x}\text{, and}$$

$$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$$

where r is the (sample) Pearson correlation coefficient and s_y and s_x are the sample standard deviations of the y_i values and the x_i values, respectively.

The estimated model is

$$y_i=\hat{\alpha}+\hat{\beta}x_i+e_i$$

The predicted value of y_i (the estimated mean $\hat{\mu}_{y|x_i}$) is given by

$$\hat{y}_i=\hat{\alpha}+\hat{\beta}x_i$$

Thus, we can write

$$e_i=y_i-\hat{y}_i\text{.}$$

e_i is called the i^{th} residual. Note the sum of squares of the residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

is at a minimum for all possible lines through the data; sometimes called the residual sum of squares or error sum of squares. Note that the residual sum of squares summarizes the variability of the y_i values about the line. We'll use it to estimate $\sigma_{y|x}$. Thus, we have estimates of means and estimates of variability. We're almost ready to do some inference.

TO BE CONTINUED…

18.2.3. In the next set of notes