STA303
Assignment 2
by Justin SJ Lee

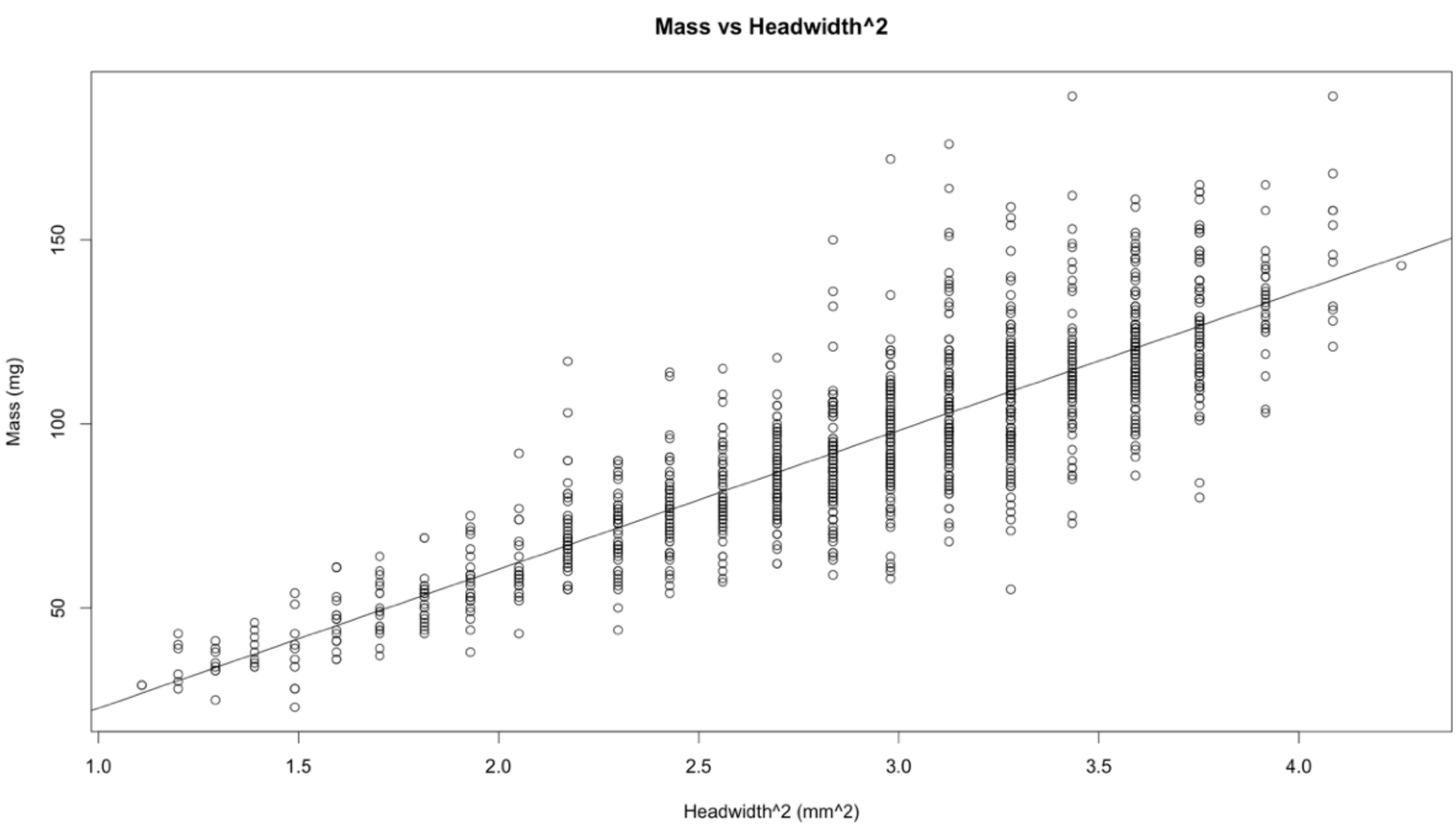
Let's further investigate the ants data collected by Peter Nonacs from UCLA. The dataset and a data dictionary and some relevant background information can be found here. You'll need to do some mild cleaning first. In particular, you should set appropriate levels for the size factor, and make Colony a factor as well. Distance you can leave as numeric. For this assignment, you may ignore the fact that ants from the same colony are likely not independent.

1. Let's try to work out the relationship between Mass (in mg) and Head Width (in mm), for thatched ants only.

(a) Make a plot of Mass vs. Headwidth.

(b) Since the relationship looks curvilinear, present a plot of Mass vs. Headwidth^2 and draw the line of best fit through this scatterplot. Look at the residual plots and note any problems you see on these plots.

From the Residual vs Fitted graph, we see that residuals are spread further apart as fitted value goes from 40 to 140. This means that the variance is increasing and is a sign of heteroscedasticity (non-constant variance). From the Normal QQ-Plot, we see that the distribution is tail heavy. These two aspects are problems for our purposes since the assumptions of simple linear regression is constant variance and a normal distribution.



(c) Fit a Weighted Least Squares regression model instead, using the (absolute) residuals to estimate the standard deviation function. Give the estimated fitted equation using English variable names.

Mass = -13.8509 + 37.3380(Headwidth^2) + Error

The above is a weighted least squares regression model for thatch ants. The units for mass and headwidth^2 are respectively mg and mm^2. For every unit increase in headwidth^2 the predicted mass increases by 37.3380mg. The intercept is not useful for us since ants with $0mm^2$ headwidth do not exist, in addition to the fact that mass can not be negative as it is in our model.

(d) Predict the mass of a thatched ant with 2.0 mm of head width, and give a 95% CI for this prediction. Compare this interval to the interval you would have obtained from an unweighted regression, and explain why it is larger or smaller.

With ordinary linear regression (OLS), the fitted value is 135.9767mg with 95 % prediction interval (106.9962mg, 164.9572mg)

With weighted linear regression (WLS), the fitted value is 135.5013mg with a 95% prediction interval (132.4004mg, 138.6021mg)

If we use weights

$$w_i \propto \frac{1}{\sigma_i^2},$$

the weighted least squares estimates have smaller standard errors than the ordinary least squares estimates for our case since weights are inversely correlated with variance and the variance for the intercept of our OLS model is greater than 1. Since the standard errors for the weighted least regression model is smaller, we can see how the prediction interval must also be smaller.

2. It's difficult to tell these ant species apart, as it requires a lab test.[1] The researchers are curious to see if you can predict the species from the mass of the ant, since they will be weighed anyway in the procedure. This would save an expensive lab test if it works. Do some analysis to determine if it is possible, and if so, quote some relevant numbers (with CI) to quantify what you've found.

In order to see if you can predict the species from the mass of the ant, we should use a logistics model since our response variable is binary. Let's check our assumptions for the logistics model.

Independence?
Observations are independent as each ant was measured separately and the measurements of one ant doesn't give information about another ant.

All relevant predictors?
No, there are many other predictors that we are leaving out such as colony, the distance the ant was found from the mound entrance as well as worker class.

Large samples?
Yes, we have 577 samples for seed ants and 1195 samples for thatch ants, thus we have satisfied the large sample assumption.

However, if our assumptions were indeed satisfied, the there would be a 7.90% decrease in the odds of being a seed ant for every unit increase in mass. We are 95% confident that for each additional increase in unit mass, the odds of being a seed ant decreases is by a number in the interval (7.19%, 8.65%).

There is 8.58% increase in the odds of being a thatch ant for every unit increase in mass.
We are 95% confident that for each additional increase in unit mass, the odds of being a thatch ant increase is by a number in the interval (7.75%, 9.47%)

These inferences however, are not valid since one of our assumptions is violated.