# Lecture 19

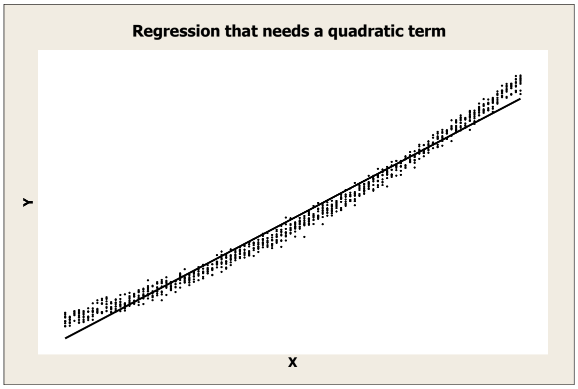
## Transformations and Regression

In this lecture we will discuss transformations of the predictor variable $X$ and the response variable $Y$ that have a natural interpretation in simple linear regression and we will discuss what those interpretations are.

### Review of the Interpretation for Linear Regression

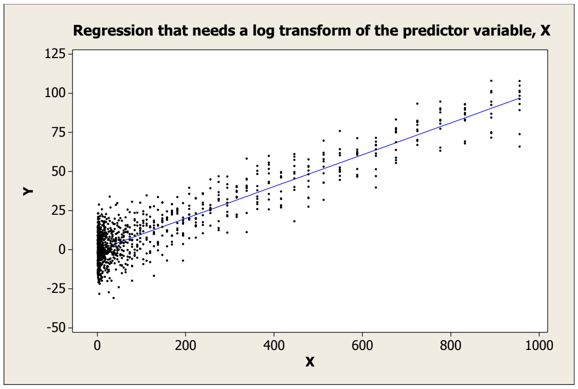
If $Y = a + bX$ then increasing $X$ by 1 increases $Y$ by $b$. That is, an additive change in $X$ corresponds to an additive change in $Y$. The change in $Y$ is proportional to the change in $X$ and the proportionality constant is the slope of the regression line which is the coefficient of $X$ in the equation above.

### Quadratic Terms



If you see curvature in the data, sometimes adding a quadratic term (the square of the $X$ variable) is appropriate. That just says your model should be quadratic rather than linear. Adding one to $X$ changes $Y$ by a constant and, additionally, by a portion that is proportional to the original value of $X$. That is, if the model is $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ then replacing $X$ with $X + 1$ replaces $Y$ with $Y + \beta_1 + \beta_2 + 2 * \beta_2 * X$. Another way of stating this is that adding 1 to $X$ causes $Y$ to increase by $\beta_1 + \beta_2 + 2 * \beta_2 * X$.

## Log Transform of the Predictor Variable

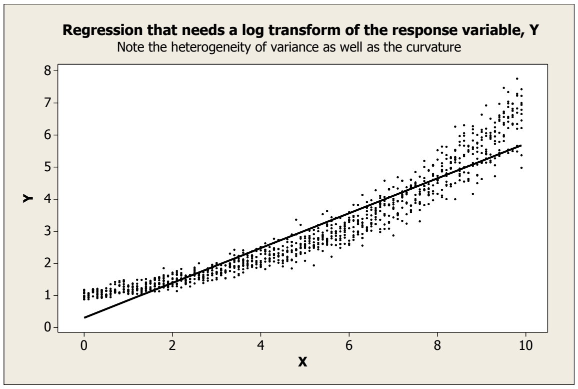


When it seems like the predictor variable, $X$, is on the multiplicative scale (for instance, when $X$ is positive and right skewed), and when there is curvature in the regression but no evidence of heterogeneity of variance, log transforming the $X$ variable may be appropriate. In this case the model becomes:

$$Y = \beta_0 + \beta_1 \log(X)$$

and the interpretation is that doubling $X$ causes a $\beta_1 \log(2)$ (additive) increase in the response variable, $Y$. If the logarithm was taken with a base of 2 to start with, then the additive increase in the response variable for doubling the predictor variable would just be the slope of the regression.

## Log Transform of the Response Variable

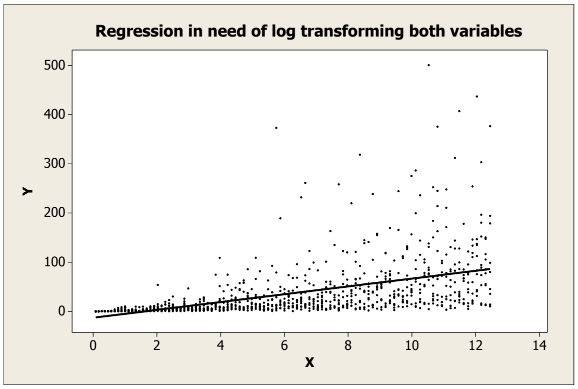


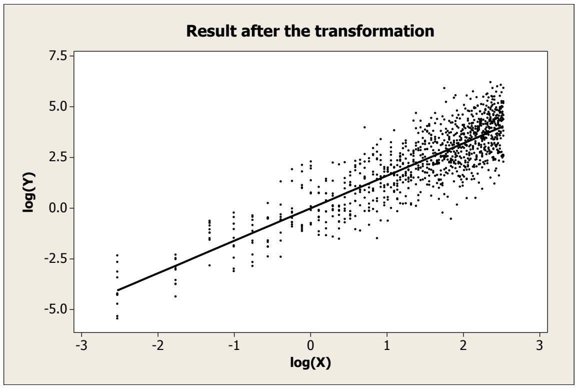
When it seems like there is heterogeneity of variance in the residuals and it is possible to log transform the response variable, $Y$ (if $Y$ is always positive), then you should try it to see if that cures the heterogeneity problem. The new model is

$$\log(Y) = \beta_0 + \beta_1 X \quad \text{or} \quad Y = e^{\beta_0 + \beta_1 X}$$

The interpretation is that increasing the predictor variable, $X$, by 1 causes an $e^{\beta_1}$-fold increase in the response variable, $Y$.

## Log Transforming Both Variables





When there is heterogeneity in the response variable and when you suspect an allometric model is the most appropriate for the data, log transforming both variables should be considered. The new model is

$$\log(Y) = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad Y = e^{\beta_0} X^{\beta_1}$$

and the interpretation is that doubling the predictor variable, $X$ causes a $2^{\beta_1}$-fold increase in the response variable, $Y$.

## Example

Consider the question of predicting heart rates from egg mass in various birds with data from Tazawa *et al.* 2001. Just log transforming the egg mass data still leaves a little curvature in the model, as the data subsetting test from Minitab indicates, whereas the allometric model shows no evidence of curvature and explains a little more of the variance in the heart rates.

The best fitting model has mathematical form

$$\text{Heart Rate} = \frac{365.75}{(\text{Egg Mass})^{0.09396}}$$

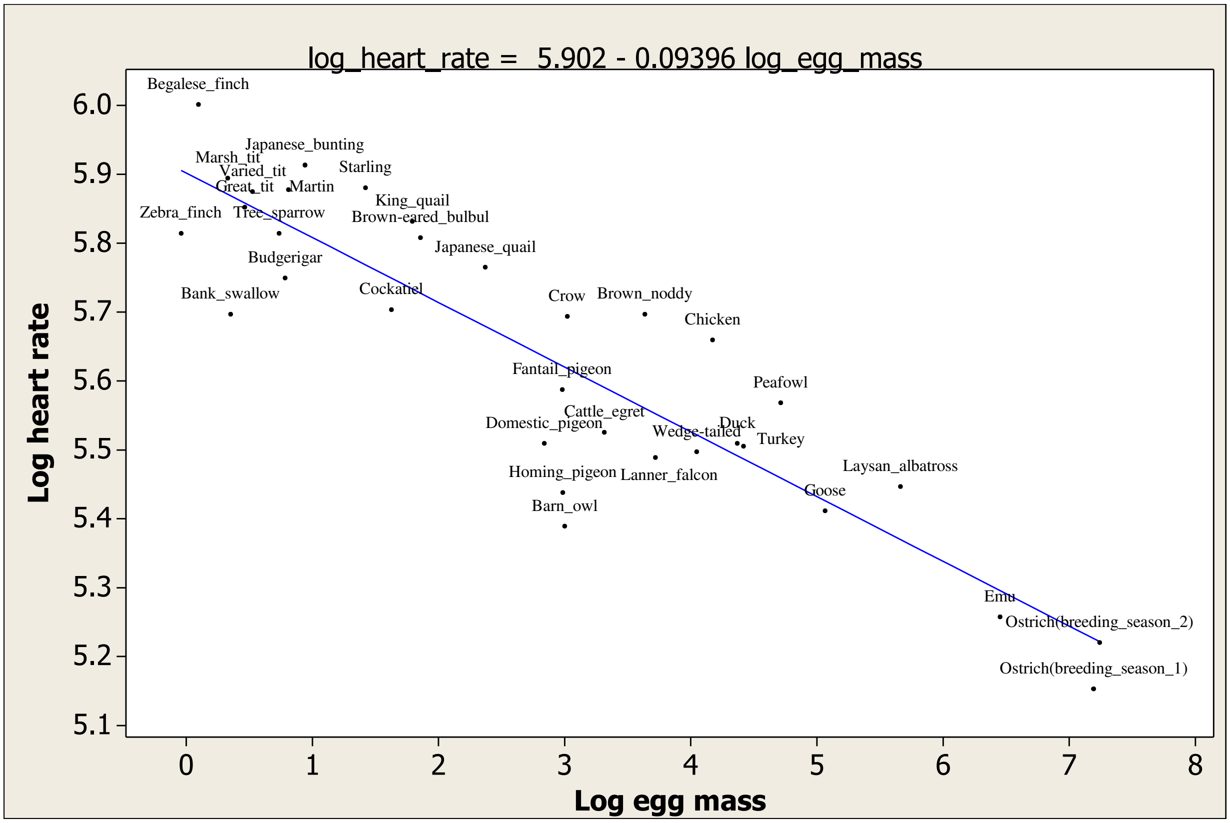


Figure 19.1: Predicting heart rate from egg mass in various birds. The allometric models fits best.

## References and Readings

[1] Hiroshi Tazawa, James T. Pearson, Takashi Komoro, and Amos Ar. Allometric relationships between embryonic heart rate and fresh egg mass in birds. *The Journal of Experimental Biology*, 204:165–174, 2001.

## Exercises for Lecture 19

1. –

2. –