# 16.0 Multiple and Nonlinear Regression

- Answer Questions
- Multiple Regression
- Nonlinear Regression
- Regression

# 16.1 Multiple Regression

Recall the regression assumptions:

**1.** Each point $(X_i, Y_i)$ in the scatterplot satisfies:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the $\epsilon_i$ have a normal distribution with mean zero and (usually) unknown standard deviation.

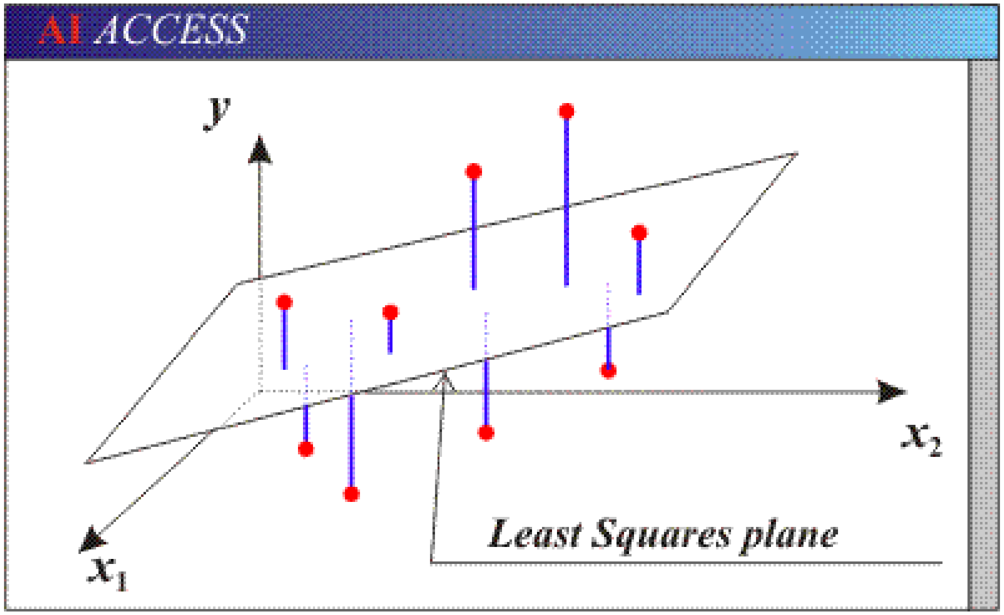
**2.** The errors $\epsilon_i$ have nothing to do with one another. A large error does not tend to be followed by another large error, for example.

**3.** The $X_i$ values are measured without error. (Thus all the error occurs in the vertical direction, and we do not need to minimize perpendicular distance to the line.)

In multiple regression, there is more than one explanatory variable. The model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_p X_{pi} + \epsilon_i.$$

Again, the $\epsilon_i$ are independent normal random variables with mean 0.



As an example, the Princeton economist and enophile Orley Ashenfelter built a model to predict the price of wine, along the following lines:

$$\begin{aligned}\textbf{price} &= \beta_0 + \beta_1(\textbf{avg. rainfall}) + \beta_2(\textbf{avg. temp.}) + \\ &\quad \beta_3(\textbf{calcium in soil}) + \beta_4(\textbf{soil pH}) + \epsilon.\end{aligned}$$

This general kind of model is used by wine speculators.

In building such a model, Ashenfelter considered many possible explanatory variables. He wanted to include only those that were relevant to viticulture (e.g., the shoe size of the vineyard owner is probably not helpful). If the model includes irrelevant explanatory variables, then it tends to give poor predictions.

To determine which variables to include and which to remove from his model, Ashenfelter did hypothesis tests to decide whether each estimated coefficient was significantly different from zero.

To make this test, the null and alternative hypotheses are:

$$\mathbf{H}_0 : \beta_i = 0 \ \textbf{vs.} \ \mathbf{H}_A : \beta_i \neq 0.$$

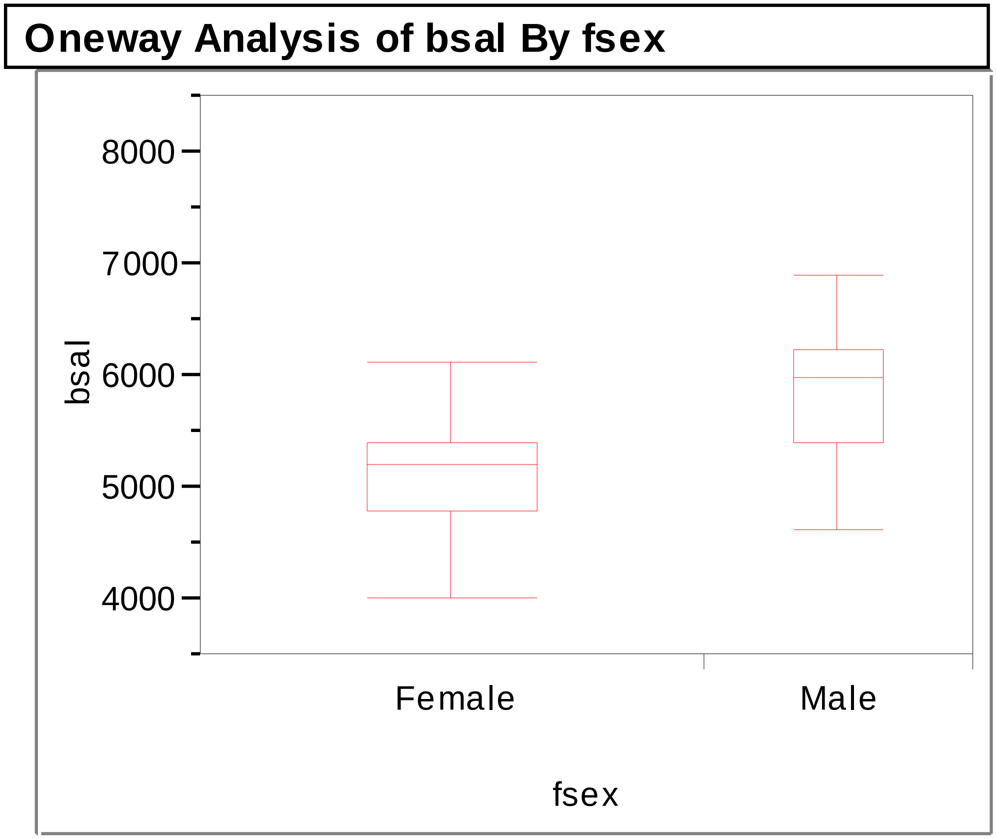
The test statistic is

$$ts = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\beta_i}}$$

where $\hat{\sigma}_{\beta_i}$ is the standard error of the estimate $\hat{\beta}_i$. It is a bit complicated, but can be found from the all standard statistics packages.

This $ts$ is compared to a $t$-distribution with $n - p - 1$ degrees of freedom. (Recall: we lose information equivalent to one observation for each estimate we make, and we had to estimate $\beta_0, \ldots, \beta_p$.) If $n - p - 1 > 30$, we can use the $z$-table.

In 1979, Harris Trust and Savings Bank was accused of gender discrimination in starting salaries. In particular, one main question was whether men in entry-level clerical jobs got higher salaries than women with similar credentials.



Harris Trust and Savings denied that they discriminated. They claimed that their starting salaries were based on many other factors, such as seniority, education, age and experience.

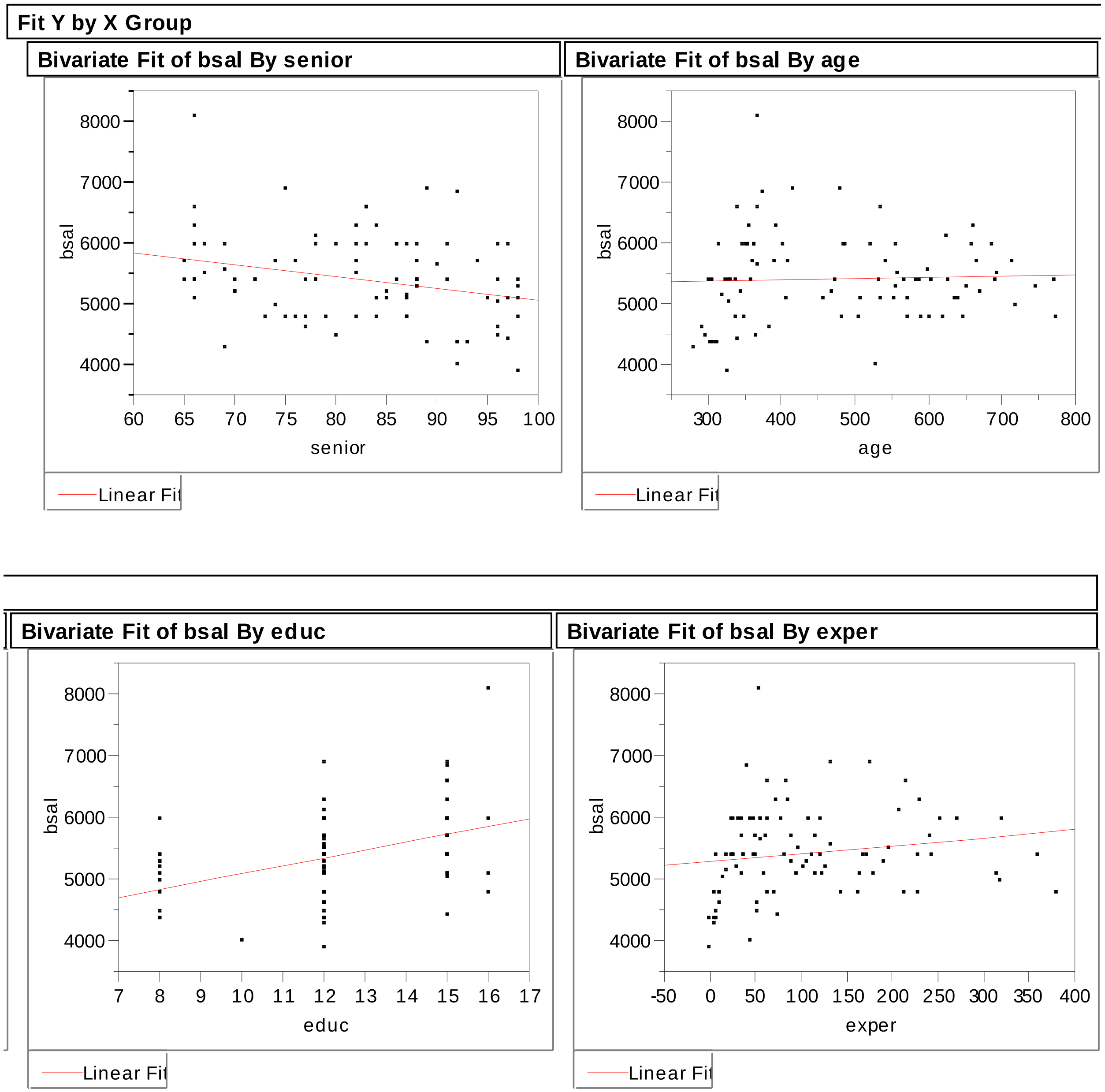
To assess that claim, the plaintiffs' lawyers used multiple regression:

$$\begin{aligned}\mathbf{salary} &= \beta_0 + \beta_1(\mathbf{sex}) + \beta_2(\mathbf{seniority}) + \beta_3(\mathbf{age}) + \beta_4(\mathbf{educ}) + \\ &\quad \beta_5(\mathbf{exper}) + \epsilon.\end{aligned}$$

Sex was recorded as 1 if the person was female, 0 for males.

Age, seniority, and experience were measured in months. Education was measured in years.

The legal question was whether the coefficient $\beta_1$ was significantly less than 0. If so, then the effect of gender was to lower the starting salary.



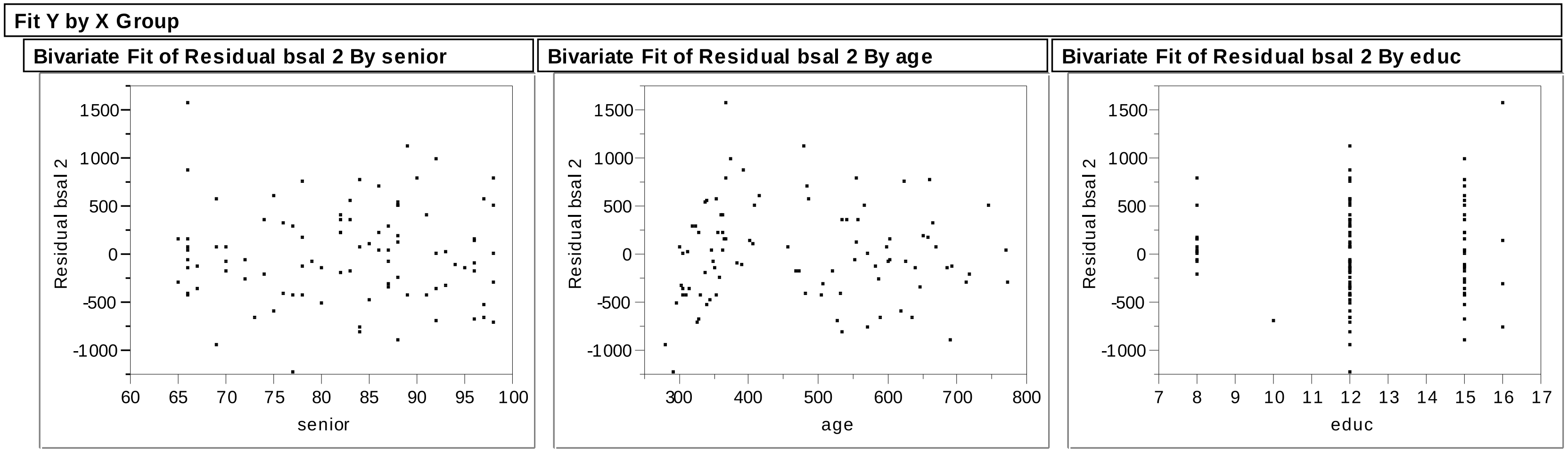
Fit Y by X Group

Bivariate Fit of bsal By senior

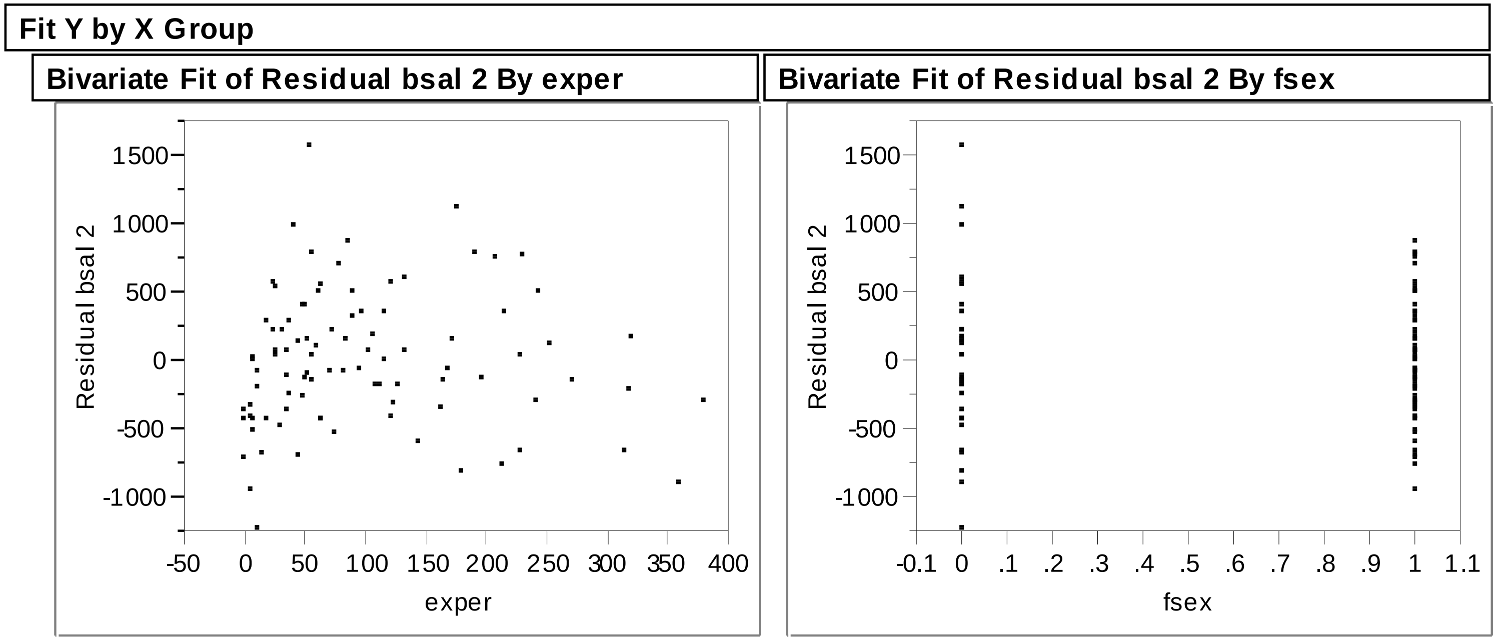
Bivariate Fit of bsal By age

Bivariate Fit of bsal By educ

Bivariate Fit of bsal By exper



These are some of the residual plots. The seniority plot looks pretty good, there is something at little odd for age at around 400 months (age 33), and the education scatterplot shows the striping associated with high school and college.


Fit Y by X Group

Bivariate Fit of Residual bsal 2 By exper

Bivariate Fit of Residual bsal 2 By fsex

These are more residual plots. Experience may show some patterning. Gender shows that there is more variance for men than for women.

One residual may be the boss's son?

Using the 93 available cases of entry-level clerical workers, the JMP statistical package found that the estimated model is

$$\begin{aligned}\textbf{salary} &= 6277.9 - 767.9(\textbf{sex}) - 22.6(\textbf{seniority}) + 0.63(\textbf{age}) + \\ &\quad 92.3(\textbf{educ}) + 50(\textbf{exper}) + \epsilon.\end{aligned}$$

The output showed that the standard error for the estimate of the coefficient on sex (i.e., the $\hat{\sigma}_{\beta_1}$) was 128.9.

We observe that the coefficient on sex is negative, which suggests that there may be discrimination against women. But we still need a significance test. We cannot interpret the size of the effect without one. Without a small significance probability, Harris Trust and Savings might argue in court that this result is only due to random chance.

The null and alternative hypotheses are:

$$\mathbf{H}_0 : b_1 \geq 0 \ \textbf{vs.} \ \mathbf{H}_A : b_1 < 0.$$

The test statistic is

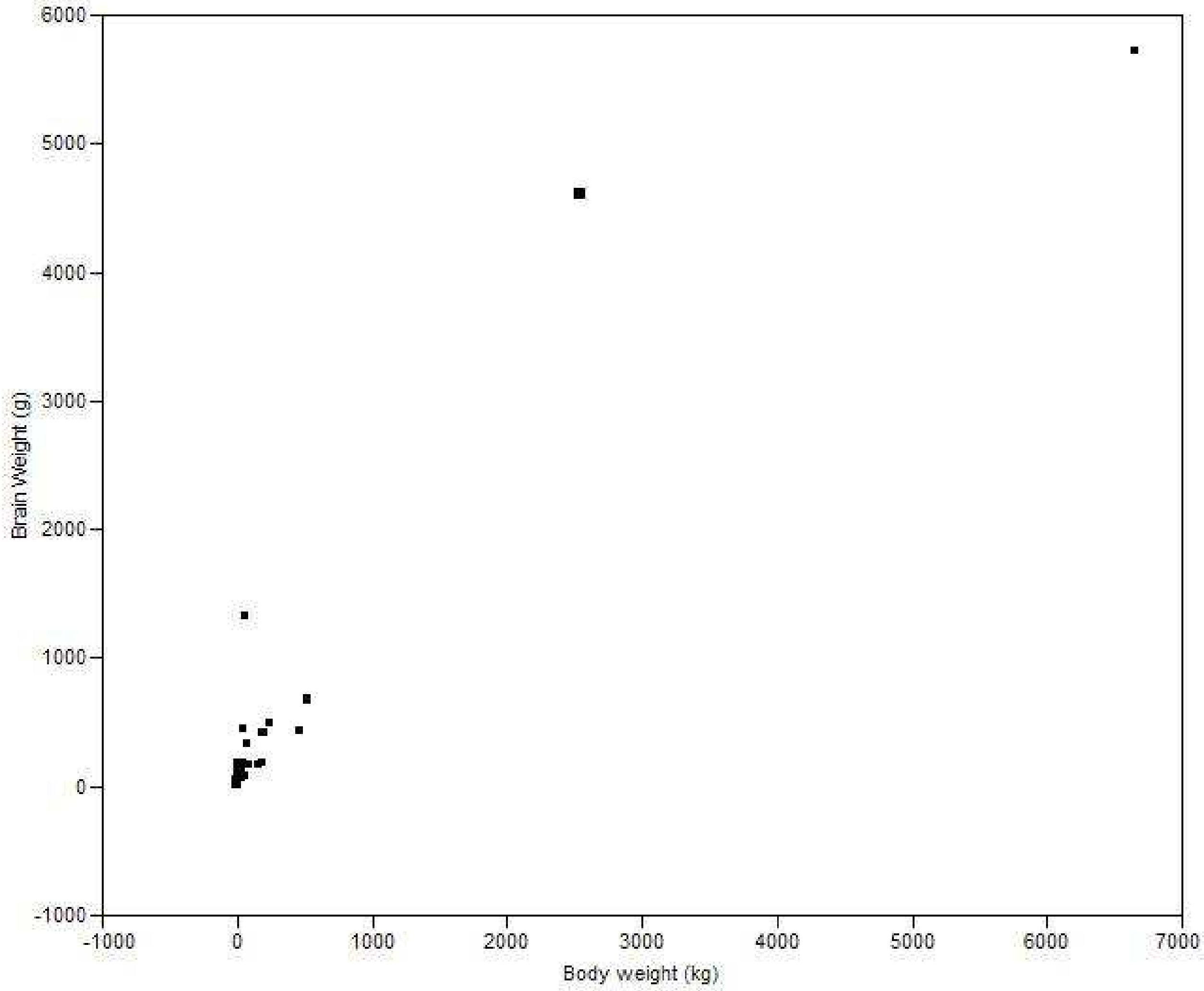
$$ts = \frac{\hat{b}_1 - 0}{se} = \frac{-767.9 - 0}{128.9} = -5.95.$$

This is compared to a $t$-distribution with $n - p - 1 = 93 - 5 - 1 = 87$ degrees of freedom. Since this is off our $t$-table scale, we use a $z$-table. The result is highly significant. Reject the null; there is evidence of discrimination.

# 16.2 Nonlinear Regression

A biologist wants to predict brain weight from body weight, based on a sample of 62 mammals. A portion of the data are shown below:

| | bodywt | brainwt | log(bodywt) | log(brainwt) |
|---|---:|---:|---:|---:|
| arctic fox | 3.385 | 44.5 | 0.529 | 1.648 |
| owl monkey | 0.48 | 15.5 | -0.318 | 1.190 |
| cow | 465 | 423 | 2.667 | 2.626 |
| grey wolf | 36.33 | 119.5 | 1.560 | 2.077 |
| roe deer | 14.83 | 98.2 | 1.171 | 1.992 |
| vervet | 4.19 | 58 | 0.622 | 1.763 |


6000
5000
4000
3000
2000
1000
0
-1000
Brain Weight (g)
-1000
0
1000
2000
3000
4000
5000
6000
7000
Body weight (kg)

The regression equation is

$$Y = 90.996 + 0.966X$$

The correlation is 0.9344, but it is heavily influenced by a few outliers (the Indian and African elephants). The standard deviation of the residuals is 334.721. This is the typical distance of a point to the line (in the vertical direction).

A 95% confidence interval on the brainweight of a mammal that weighed 100 kg would be

$$L, U = 90.996 + 0.966(100) \pm (334.721)(1.96)$$

so $U = 843.65$ and $L = -468.46$. This isn't very helpful.

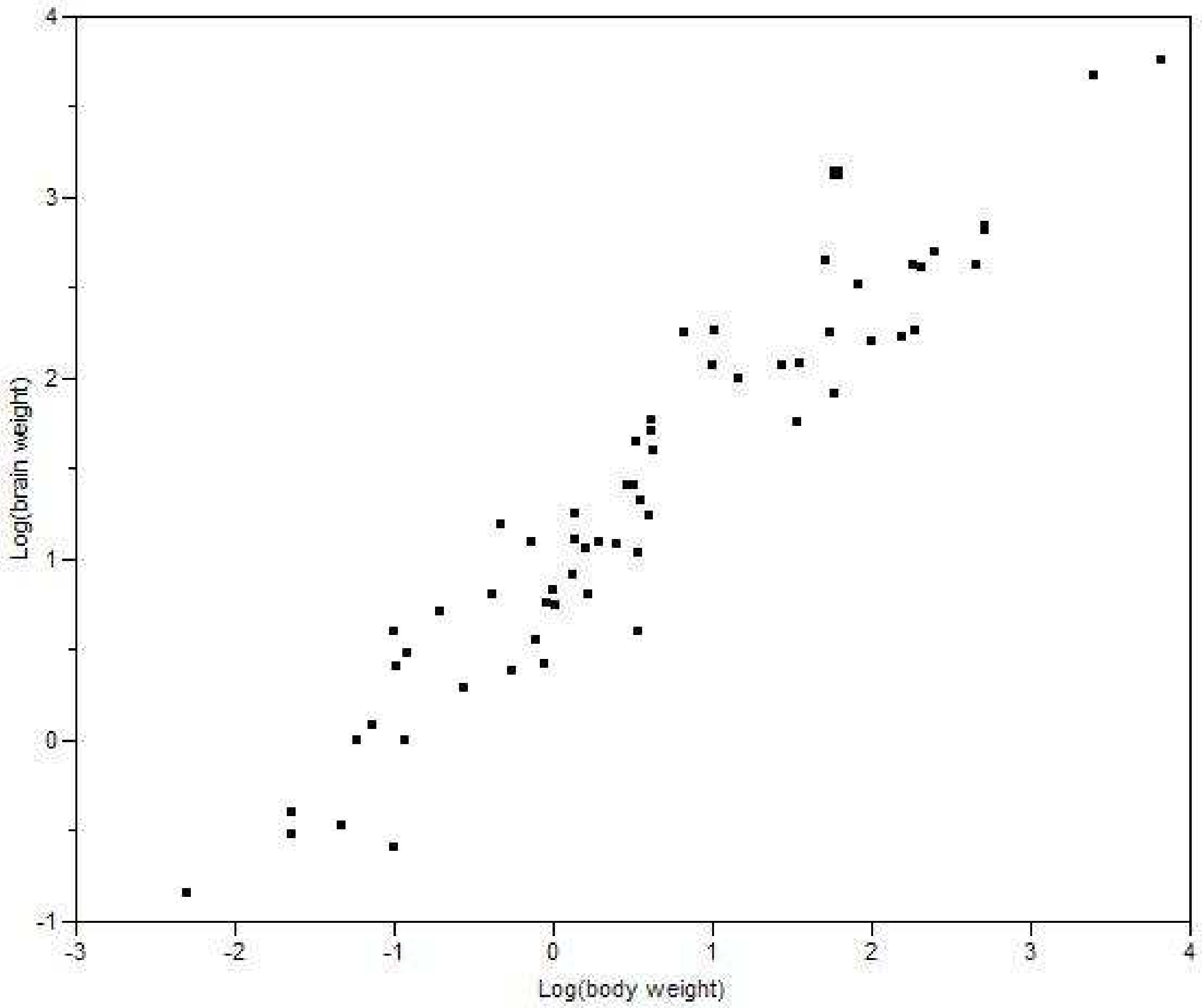
The scatterplot of the brainweight against bodyweight showed the the line was probably controlled by a few large values. These are sometimes called **influential points**.

Even worse, the scatterplot did not resemble the cigar-shaped point cloud that supports the regression assumptions listed before.

In cases like this, one can consider making a transformation of the response variable or the explanatory variable or both. It is hard to know what transformation to choose; usually this choice depends upon scientific knowledge or the judgment of a good statistician.

For this data, consider taking the logarithm (base 10) of the brainweight and body weight.

The following scatterplot is much better.


Log(brain weight)
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
4
Log(body weight)

Taking the log shows that the influential points are not surprising. The regression equation is now:

$$\log Y = 0.908 + 0.763 \log X$$

The coefficient of determination shows that 91.23% of the variation in log brain weight is explained by log body weight. Both the intercept and the slope are highly significant. The estimated standard deviation of $\epsilon$ is 0.317; this is the typical vertical distance between a point and the line.

Thus a 95% confidence interval on the log brain weight of a 100 kg mammal is

$$L, U = 0.908 + 0.763(\log 100) \pm (0.317)(1.96)$$

so $U = 3.06$ and $L = 1.81$.

Making transformations is an art. Here the analysis suggests that

$$Y = 10^{0.908} * X^{0.763} = 8.1 * X^{0.763}.$$

So there is a power-law relationship between brain mass and body mass.

**Note:** We are ignoring a technical issue about additivity of the errors.

Some standard transformations:

| **function** | **transformation** | **linear form** |
|---|---|---|
| $y = a \exp bx$ | $y^* = \ln y$ | $y^* = \ln a + bx$ |
| $y = ax^b$ | $y^* = \log y, x^* = \log x$ | $y^* = \log a + bx^*$ |
| $y = a + b/x$ | $x^* = 1/x$ | $y = a + bx^*$ |