STA 111: Probability & Statistical Inference

Lecture Nineteen – Multiple Linear Regression
D.S. Sections 11.3 & 11.5

Instructor: Olanrewaju Michael Akande

Department of Statistical Science, Duke University

Outline

- Questions from Last Lecture
- Multiple Regression
- Transformations
- Recap

Introduction

– In the last lecture we talked about simple linear regression and least squares method

– Today we will extend the regression idea to multiple explanatory variables.

– We often come across data that do not necessarily satisfy the normal distribution or linearity assumptions for the response variable and we will briefly talk about some common transformations that result in normality (or as least as close as possible).

Recap

Recall the simple linear 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 ϵ_i have a normal distribution with mean zero and (usually) unknown standard deviation.

2. The errors ϵ_i have nothing to do with one another (independence). 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.)

These weren't included in the last class but we also implicitly assume that:

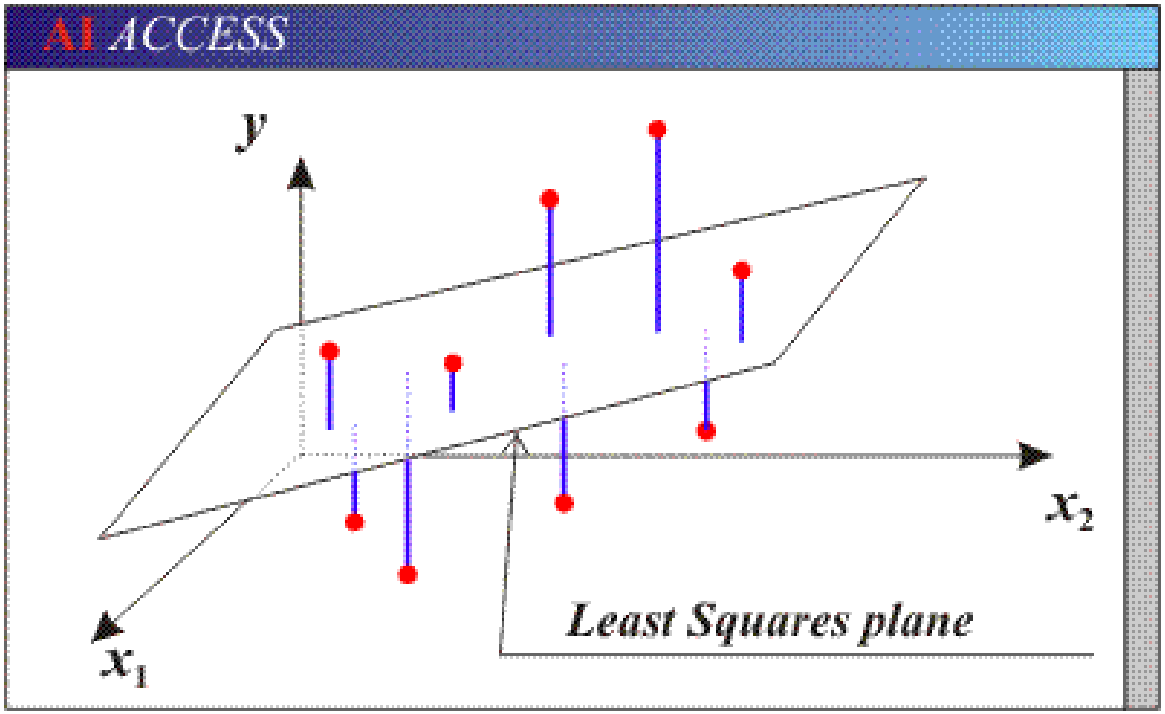
4. The errors ϵ_i are independent of the x_i's.
5. The relationship between y_i andx_i is linear.

Multiple Regression

In multiple regression, the set-up is still the same except that we now have 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 ϵ_i are independent normal random variables with mean 0 and all the x's are measured without error.



Multiple Regression

How then should we interpret the coefficients (that is, $\beta_0, \beta_1, \ldots, \beta_p$)?

Recall that for the simple linear regression, β_0 is usually the average of y when x is zero (and it is only meaningful when x can be zero) while for β_1, every unit increase in x corresponds to an increase in y by β_1.

The interpretation is quite similar here too except that for any β_p, every unit increase in x_p corresponds to an increase in y by β_p **when all the other x's are fixed or held constant.**

Variable Selection

In practice, we often have a long list of potential explanatory variables and need to make a decision on which variables to include.

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

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

This general kind of model is often used by wine speculators.

In building such a model, Ashenfelter considered many possible explanatory variables. He wanted to include only those that were relevant. 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.

Variable Selection

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 takes the form we are used to:

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

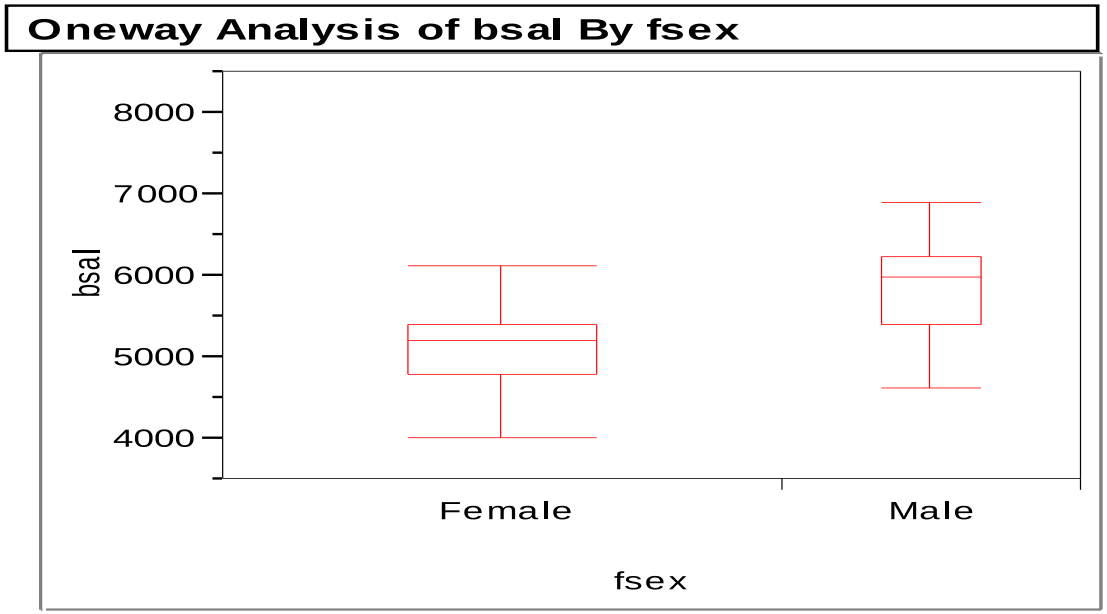
where $\hat{\sigma}_{\beta_i}$ is the standard error of the estimate $\hat{\beta}_i$. It is a bit complicated (remember that we found the variance of β_1 in the previous lecture), but can be found from the all standard statistics packages.

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

Examples

Why multiple regression? This follows intuitively from our discussion on confounders and the following example illustrates the point.

Example 1: 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. Exploratory box plots showed that the claim might be true.



Examples

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 (possible confounders).

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

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

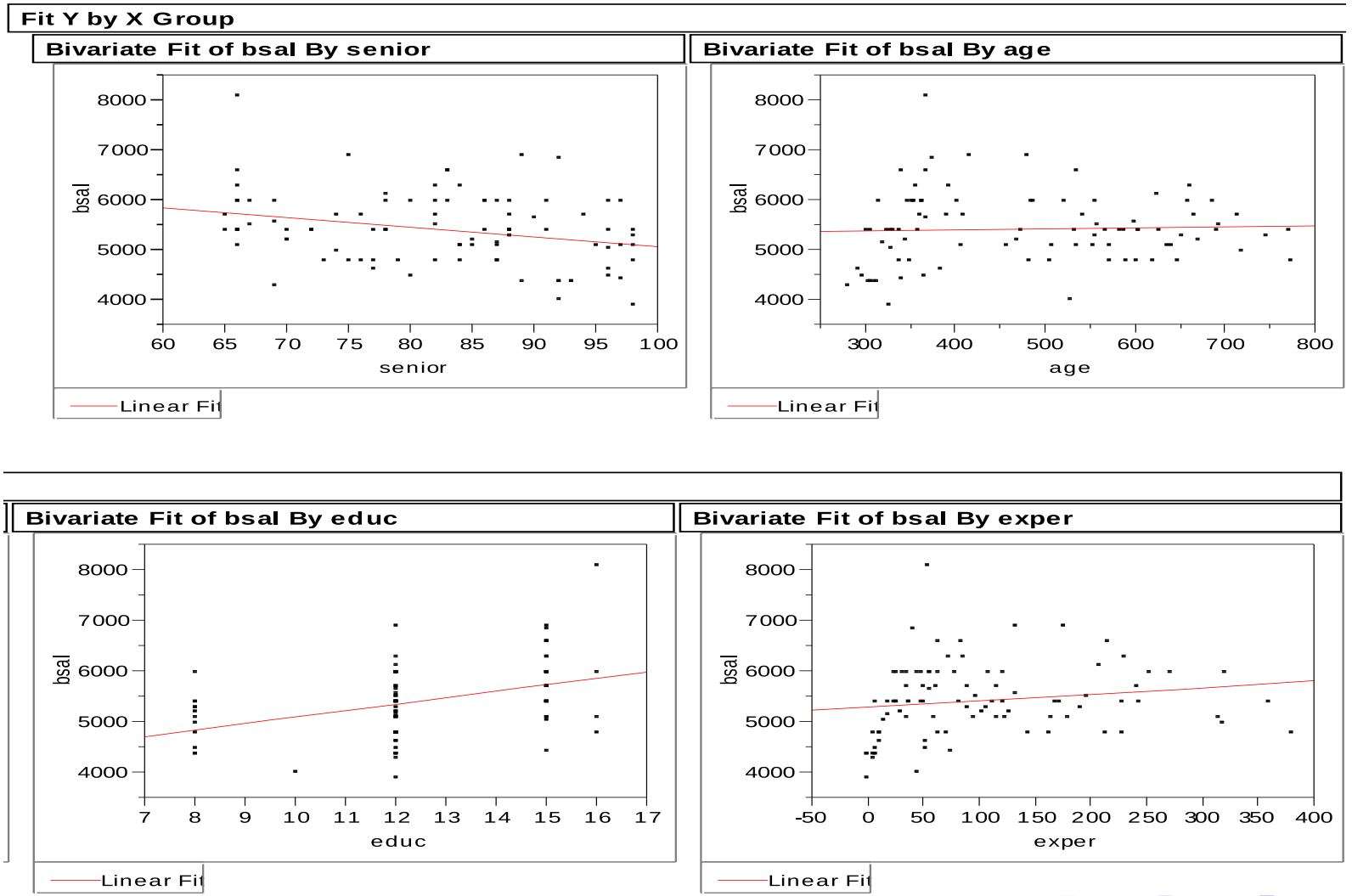
Sex was recorded as 1 if the person was female, 0 for males.

Age, seniority, and experience were measured in months. Education was measured in years (we are treating education as a numeric variable but that's not the only way to treat it; let's ignore that discussion here though).

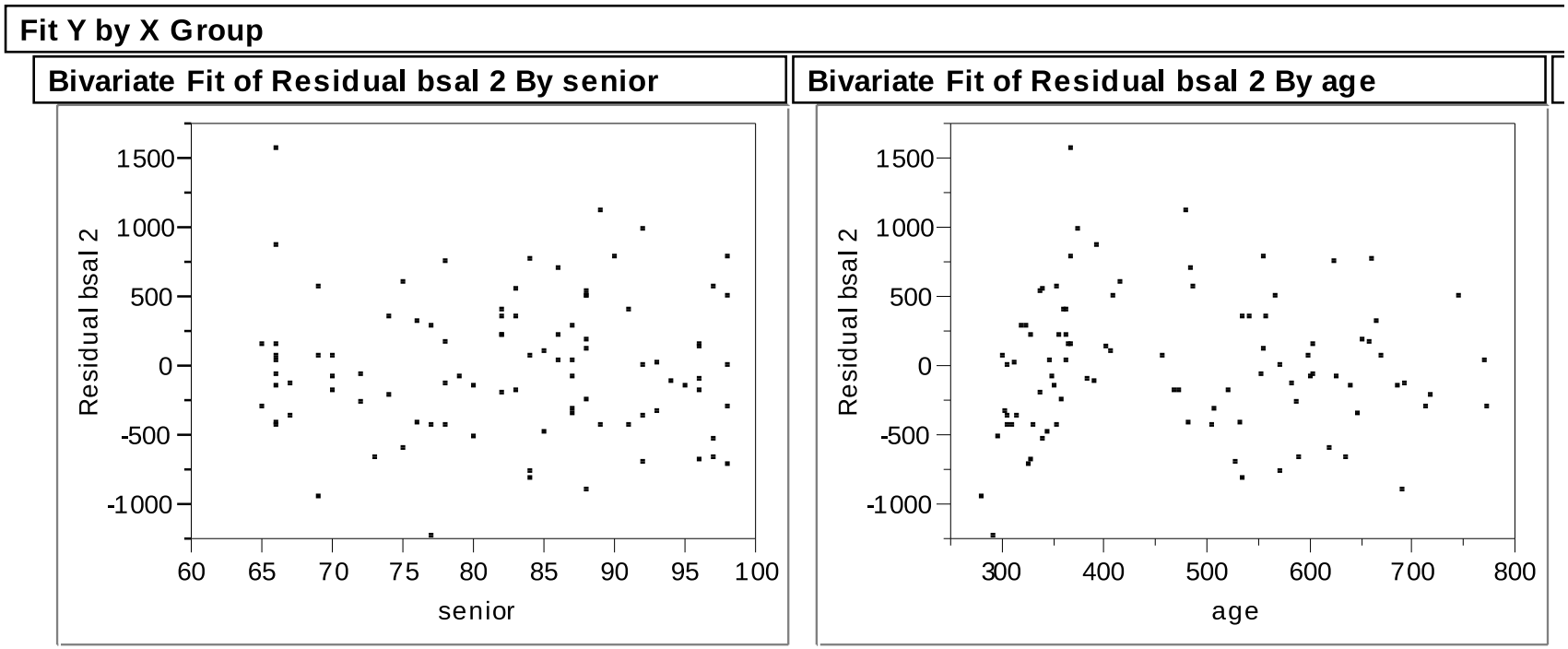
The legal question was whether the coefficient β_1 was significantly less than 0. If so, then the effect of gender was to lower the starting salary.

Examples

Let's look at the regression fit by each variable, holding the rest constant.

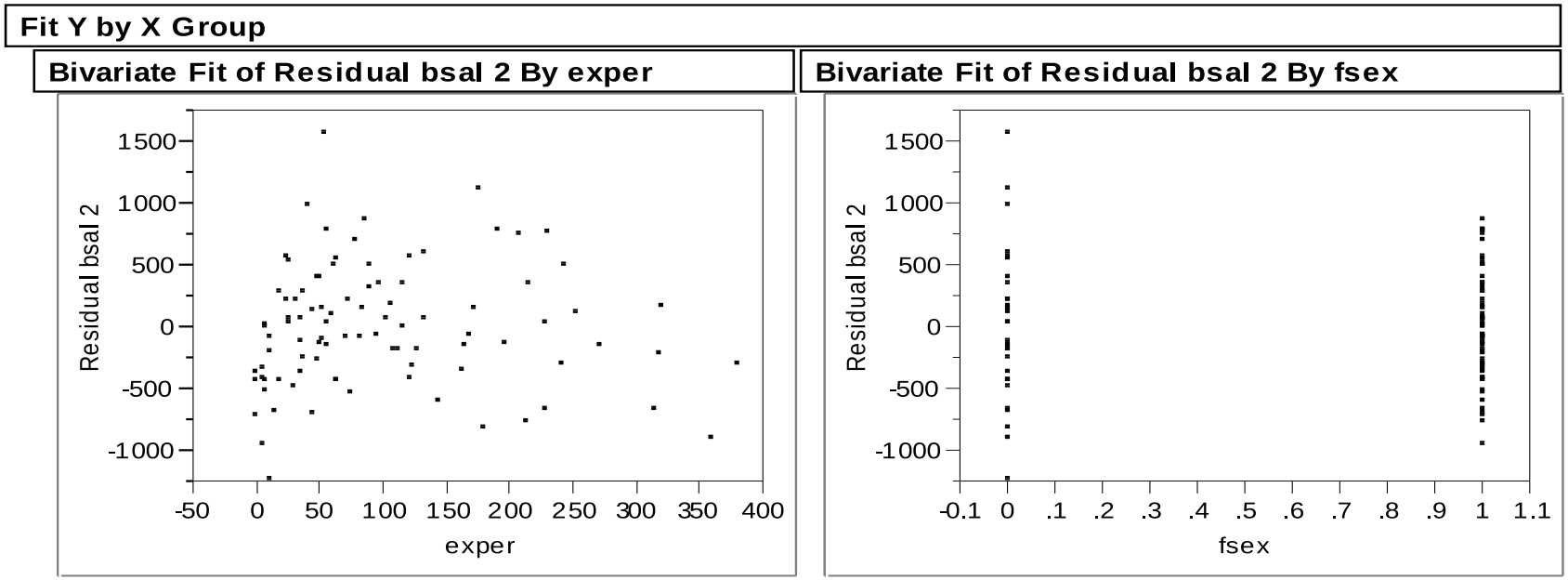


Examples



These are some of the residual plots, to examine our assumption of independence between the errors and each covariate. The seniority plot looks pretty good, there is something at little odd for age at around 400 months (age 33).

Examples



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?

Residual plots are really useful in examining our assumption of independence. They might help us figure out that we are still missing a function of one of the x's if we observe any pattern.

Examples

Back to the question we are interested in answering:

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

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

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 to be sure. We cannot interpret the size of the effect without one. Without a small p-value (below $\alpha = 0.05$ for example), Harris Trust and Savings might argue in court that this result is only due to random chance.

Examples

Because we care about a one-sided alternative hypothesis, 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.

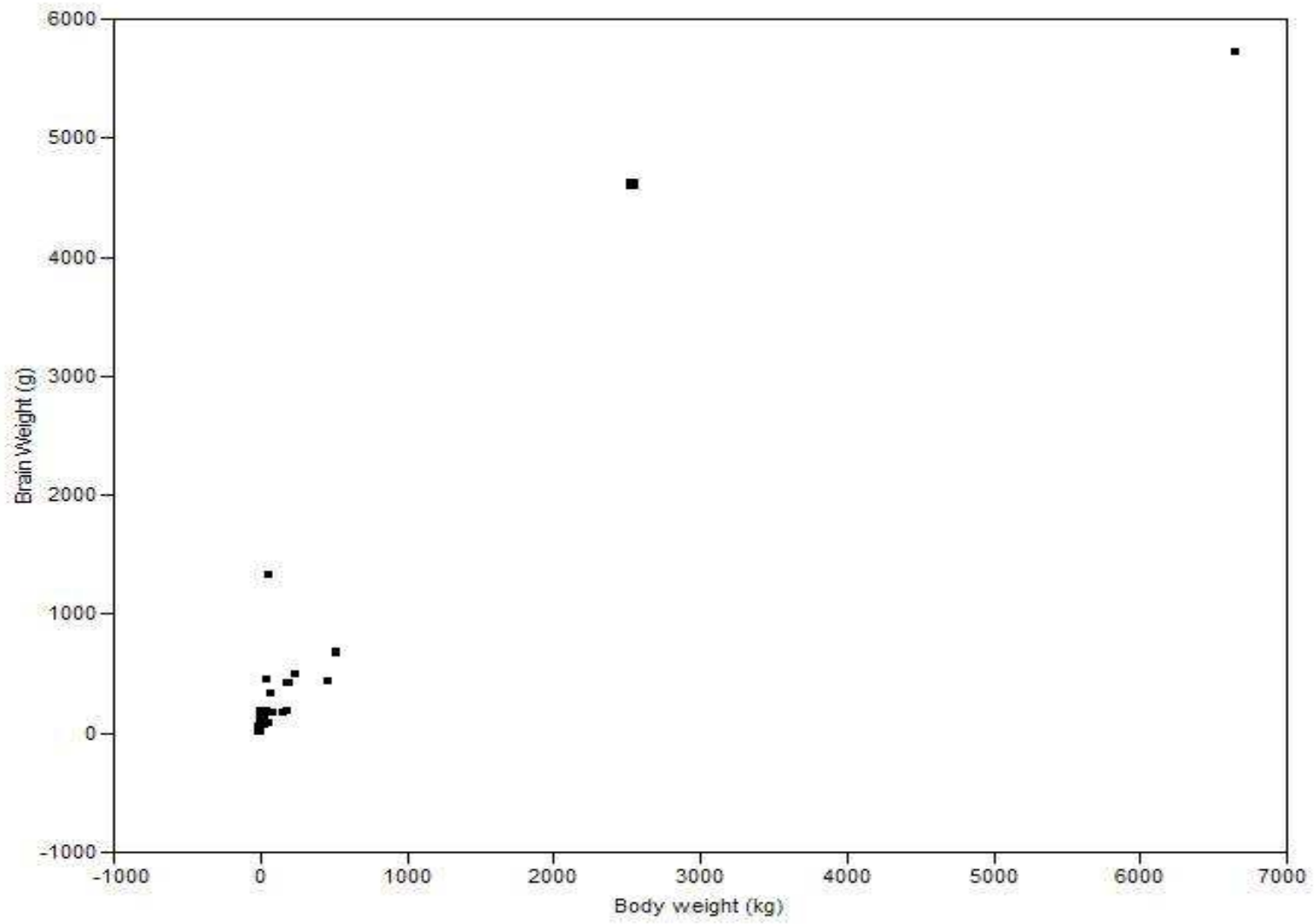
Transformations

We often come across data that are not necessarily normal or even linear and some transformations can be useful.

Example 2: 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

Transformations



Transformations

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.

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.

Transformations

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**.

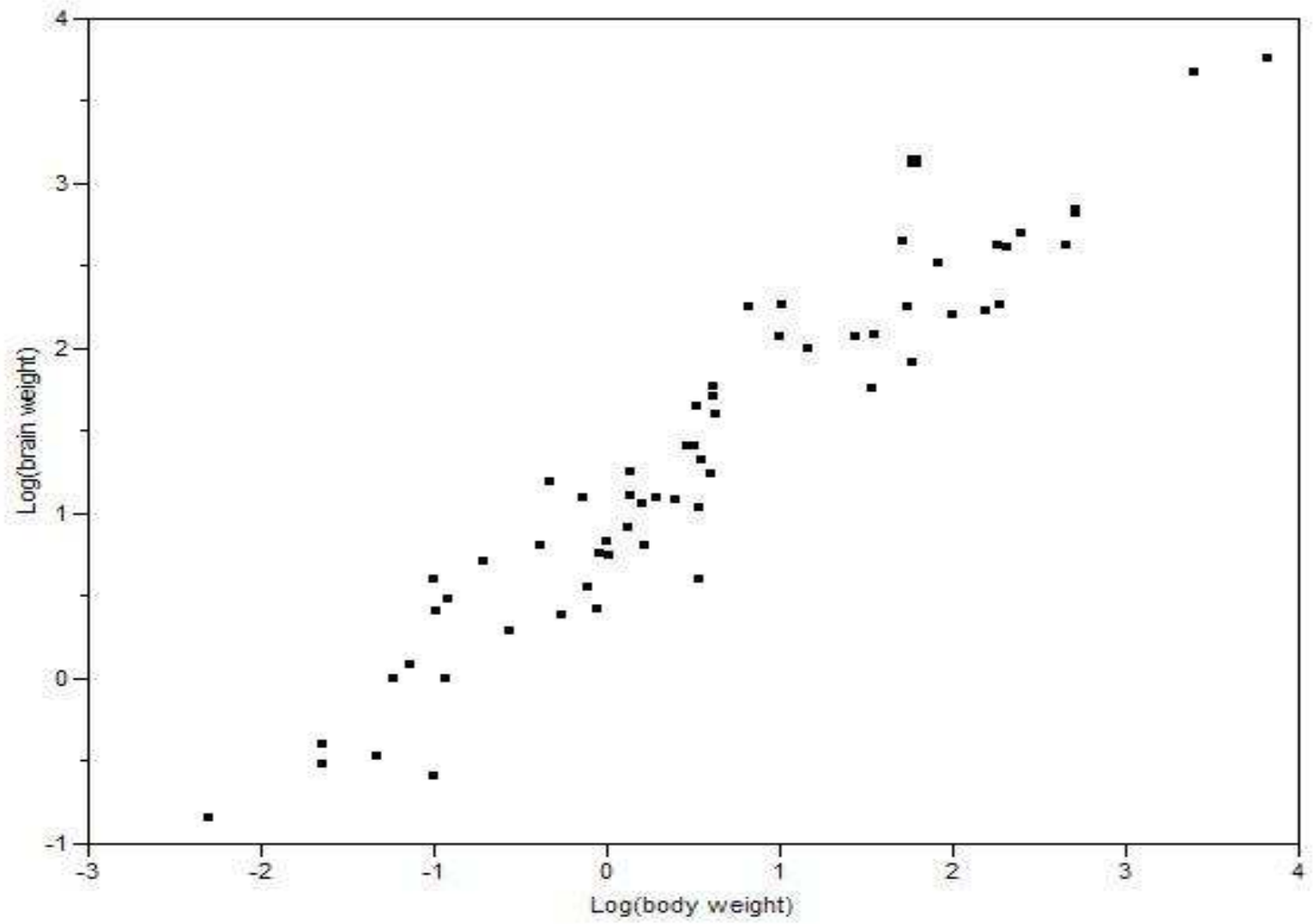
Even worse, the scatterplot did not resemble the linear trend that supports the regression assumptions listed before.

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.

Transformations



Transformations

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 ϵ is 0.317.

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 $L = 1.81$.

Transforming back to the original scale, $U = 10^{3.06}$ and $L = 10^{1.81}$, which is more helpful than before.

Transformations

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**
$y = ae^{bx}$	$y^* = \ln y$	$y^* = \ln a + bx$
$y = ax^b$	$y^* = \log y, x^* = \log x$	$y^* = \log a + bx^*$
$y = a + b/x$	$x^* = 1/x$	$y = a + bx^*$

Recap

Today we learned about extending the regression idea to multiple explanatory variables.

In the next lecture, we will talk about one specific type of regression known as analysis of variance (which doesn't have to be set-up as a classic regression problem exactly).