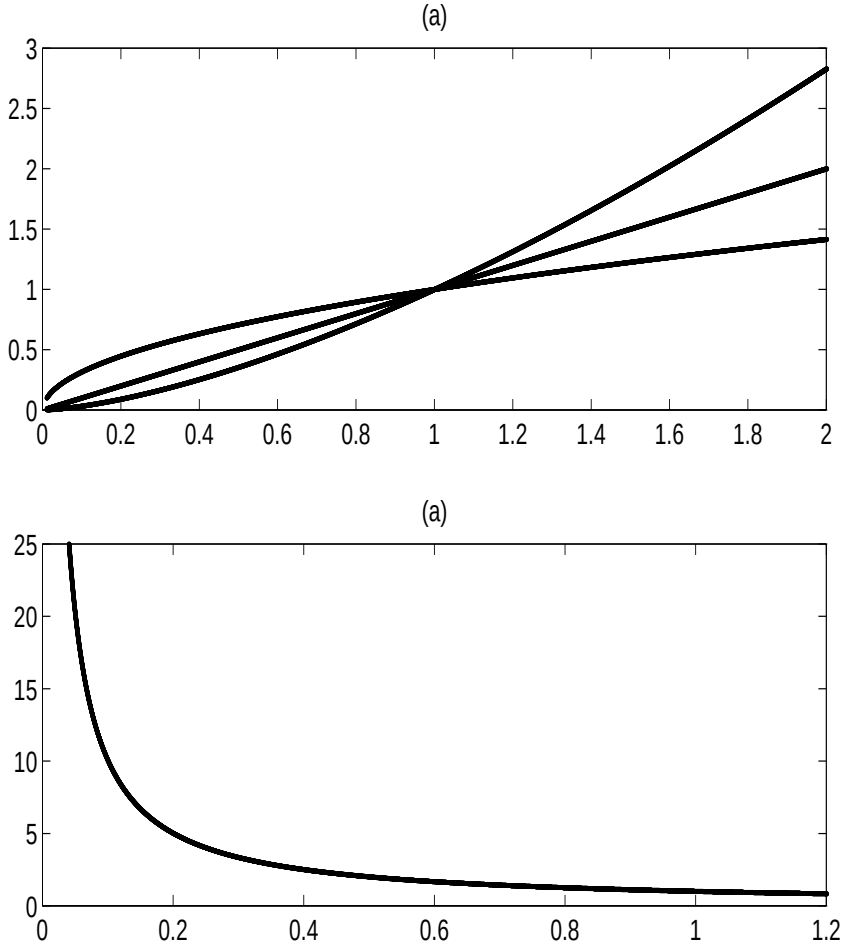
## Transformations in Regression

Simple linear regression is appropriate when the scatterplot of $Y$ against $X$ show a linear trend. In many problems, non-linear relationships are evident in data plots. Linear regression techniques can still be used to model the dependence between $Y$ and $X$, provided the data can be transformed to a scale where the relationship is roughly linear. In the ideal world, **theory** will suggest an appropriate transformation. In the absence of theory one usually resorts to empirical model building. Polynomial models, to be discussed later, are another method for handling nonlinear relationships.

I will suggest transformations that you can try if the **trend** in your scatterplot has one of the following functional forms. The responses are assumed to be non-negative in all cases.

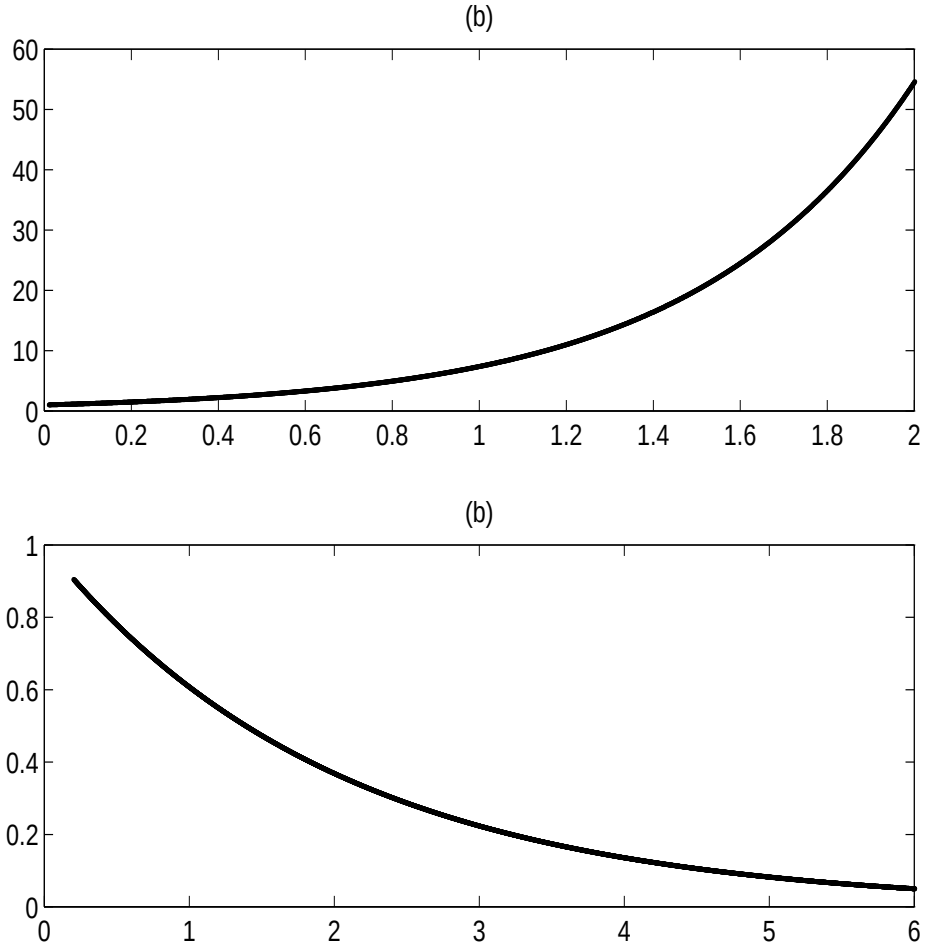



The functional relationship between $Y$ and $X$ in (a) is given by $Y = \beta_0 X^{\beta_1}$, that is $Y$ is related to a power of $X$, where the power is typically unknown. For the top plot, $\beta_1 > 0$

whereas $\beta_1 < 0$ for the plot on the bottom. For either situation, the logarithm of $Y$ is linearly related to the logarithm of $X$ (regardless of the base):

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

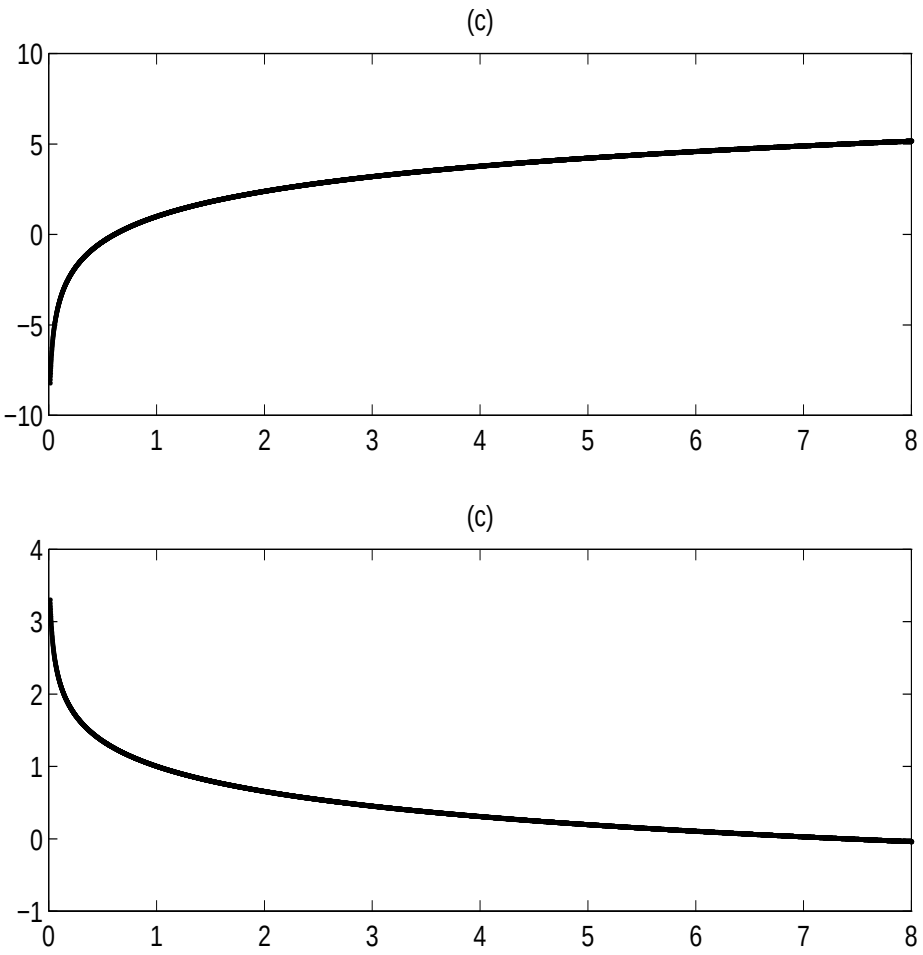
You should consider a simple linear regression of $Y' = \log(Y)$ on $X' = \log(X)$.




The functional relationship between $Y$ and $X$ in (b) is given by $Y = \beta_0 \exp(\beta_1 X)$, that is $Y$ is an exponential function of $X$. For the plot on top, $\beta_1 > 0$ whereas $\beta_1 < 0$ for the plot on the bottom. In each situation, the natural logarithm of $Y$ is linearly related to $X$:

$$\log_e(Y) = \log_e(\beta_0) + \beta_1 X.$$

You should consider a simple linear regression of $Y' = \log_e(Y)$ on $X$. Actually, the base of the logarithm is not important here either.

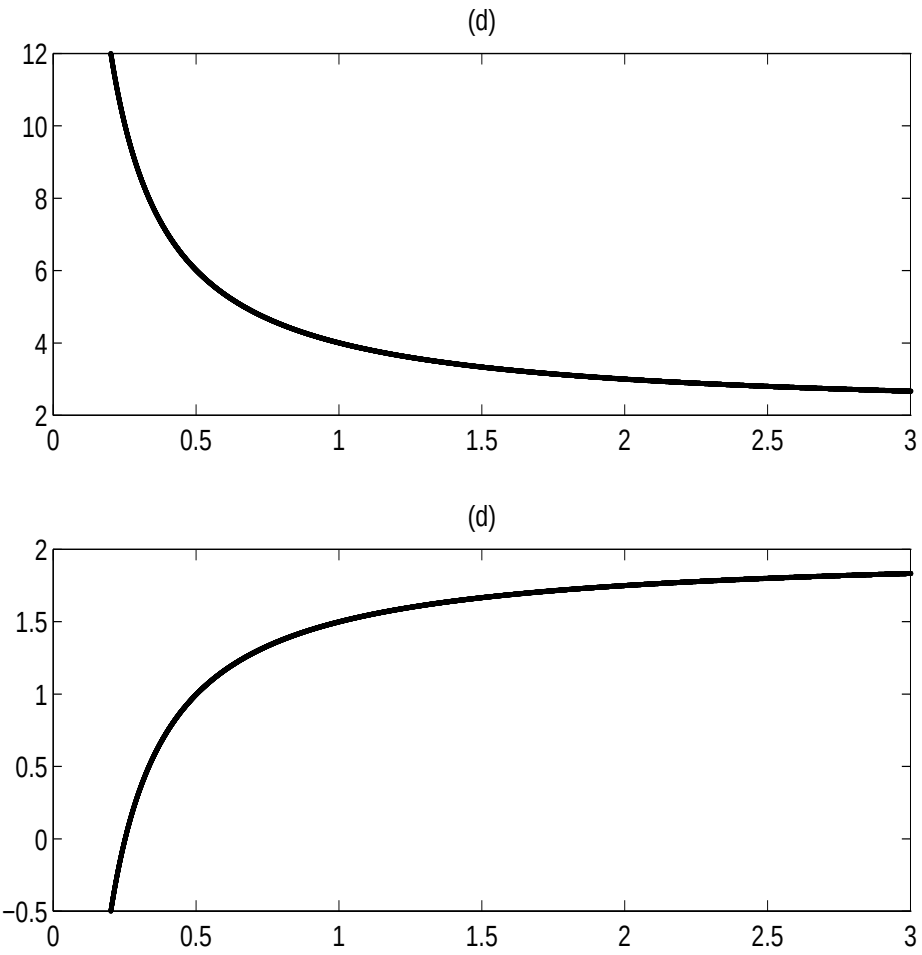



The functional relationship between $Y$ and $X$ in (c) is given by $Y = \beta_0 + \beta_1 \log(X)$, that is $Y$ is an logarithmic function of $X$. For the top plot, $\beta_1 > 0$ whereas $\beta_1 < 0$ for the bottom plot. In each situation, consider a simple linear regression of $Y$ on $X' = \log(X)$.

The functional relationship between $Y$ and $X$ in (d) is

$$Y = \beta_0 + \beta_1 \frac{1}{X}.$$

Hence, consider a simple linear regression of $Y$ on $X' = 1/X$. Note that each plot in (d) has a horizontal asymptote of $\beta_0$.




In most problems, the **trend or signal** will be buried in a considerable amount of **noise, or variability**, so the best transformation may not be apparent. If two or more transformations are suggested try all of them and see which is best - look at diagnostics from the various fits rather than (meaningless) summaries such as $R^2$. In situations where a logarithmic transformation is suggested, you might try a square root transformation as well.

The need to transform is sometimes more apparent in a plot of the studentized residuals against the predicted values from a **linear fit** of the **original data** because you tend not to perceive subtle deviations from linearity.

## Computing Predictions

Transforming the response to a new scale causes no difficulties if you wish to make predictions on the original scale. For example, suppose you fit a linear regression of $\log_e(Y)$ on $X$. The fitted values satisfy

$$\widehat{\log_e(Y)} = b_0 + b_1 X.$$

The predicted response $Y_p$ for an individual with $X = X_p$ is obtained by first getting the predicted value for $\log_e(Y_p)$:

$$\widehat{\log_e(Y_p)} = b_0 + b_1 X_p.$$

Our best guess for $Y_p$ is obtained by exponentiating our prediction for $\log_e(Y_p)$:

$$\hat{Y}_p = \exp(\widehat{\log_e(Y_p)}) = \exp(b_0 + b_1 X_p).$$

The same idea can be used to get prediction intervals for $Y_p$ from a prediction interval for $\log_e(Y_p)$.

Other transformations on $Y$ are handled analogously. For example, how do you predict $Y$ using a simple linear regression with $1/Y$ as the selected response?

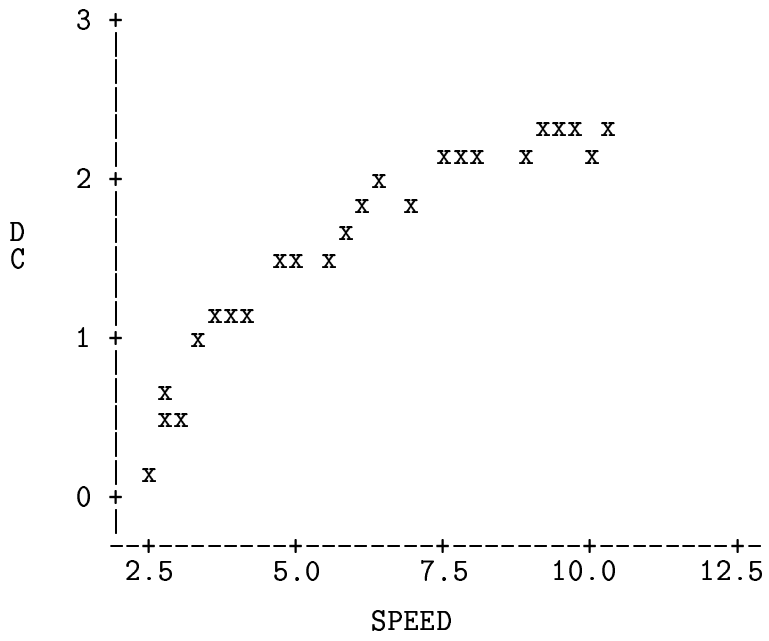
## Example of Transformations: Wind Speed Data

A research engineer is investigating the use of a windmill to generate electricity. She has collected data on the DC output from the windmill and the corresponding wind velocity. She wants to develop a model that explains the dependence of the DC output on wind velocity.

I will give you snipets of an analysis. The data are given below, followed by a plot.

| SPEED | DC |
|---|---|
| 5.00 | 1.582 |
| 6.00 | 1.822 |
| 3.40 | 1.057 |
| 2.70 | 0.500 |
| 10.00 | 2.236 |
| 9.70 | 2.386 |
| 9.55 | 2.294 |
| 3.05 | 0.558 |
| 8.15 | 2.166 |
| 6.20 | 1.866 |
| 2.90 | 0.653 |

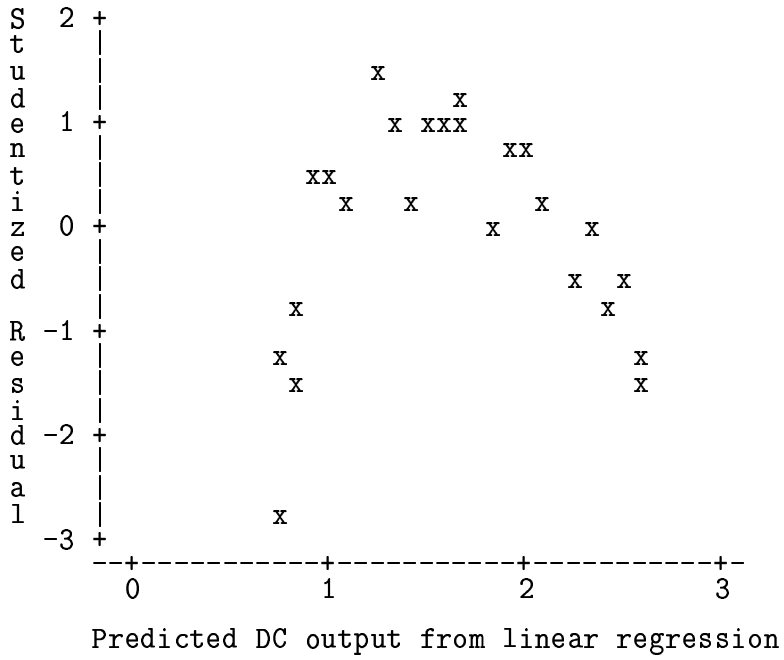
```
 6.35   1.930
 4.60   1.562
 5.80   1.737
 7.40   2.088
 3.60   1.137
 7.85   2.179
 8.80   2.112
 7.00   1.800
 5.45   1.501
 9.10   2.303
10.20   2.310
 4.10   1.194
 3.95   1.144
 2.45   0.123
```




The data plot shows a strong linear trend, but the relationship is nonlinear. If I ignore the nonlinearity and fit a simple linear regression model, I get

$$\text{Predicted DC Output} = .1809 + .2411 \text{ Wind Speed}.$$

Although the $R^2$ from this fit is high, $R^2 = .875$, I am unhappy with the fit of the model. The plot of the residuals against fitted values clearly points out the inadequacy. The plot shows that the linear regression systematically underestimates the DC output for wind speeds in the middle, and overestimates the DC output for low and high wind speeds. This model is not acceptable for making predictions - one can and should do better!



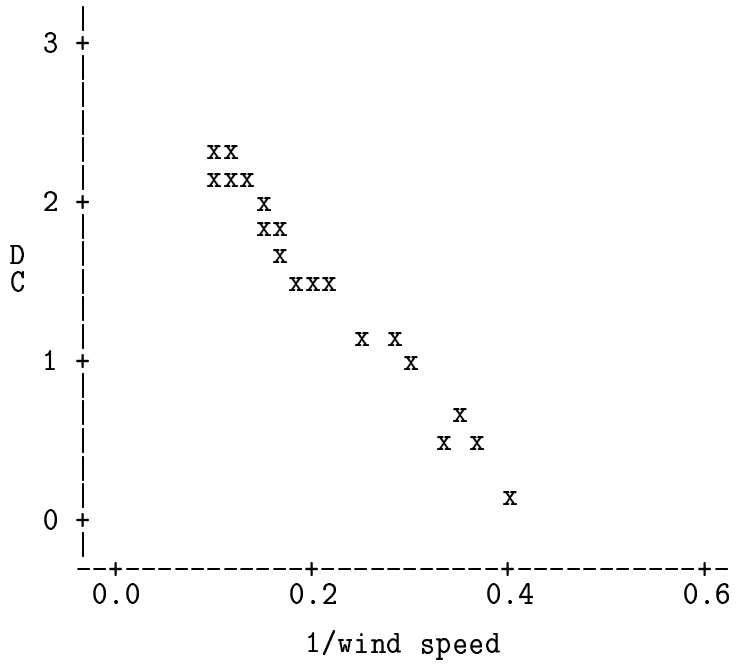

The original data plot indicates that DC output approaches an upper limit of about 2.5 amps as the wind speed increases. Given this fact, and the trend in the plot, I decided to use the inverse of wind speed as a predictor of DC output. Another reasonable first step would be a logarithmic transformation of wind speed but this function steadily increases without approaching a finite limit.

A plot of DC output against one over the wind speed (computed in the **MINITAB** calculator) is given below. The plot is fairly linear, suggesting that a simple linear regression fit on this scale is appropriate. Note that DC output is a decreasing function of one over the wind speed.

The LS regression line is

$$\text{Predicted DC output} = 2.9789 - 6.9345\frac{1}{\text{Wind speed}}.$$

A plot of the studentized residuals against the fitted values showed no apparent abnormalities. A more thorough influence analysis led us to conclude that the transformation worked well here.




## Brain Weights and Body Weights of Mammals

This is another old homework assignment that was too important to be excluded from class discussion.

The data below are the average brain weight (g) and body weights (kg) for 54 species of mammals. The brain weights for eight other species were omitted (given as $*$, the missing value code in **MINITAB**). These data were read into **MINITAB** from an **EXCEL** spreadsheet.

You are interested in developing a model for predicting brain weight from body weight. Based on a plot of the data, determine whether a transformation is needed to linearize the relationship between brain weight and body weight. Given the scale you select, perform a simple linear regression analysis of the data. Be sure to check for influential observations, outliers, and any deviations from assumptions. Do your best to correct any deficiencies with

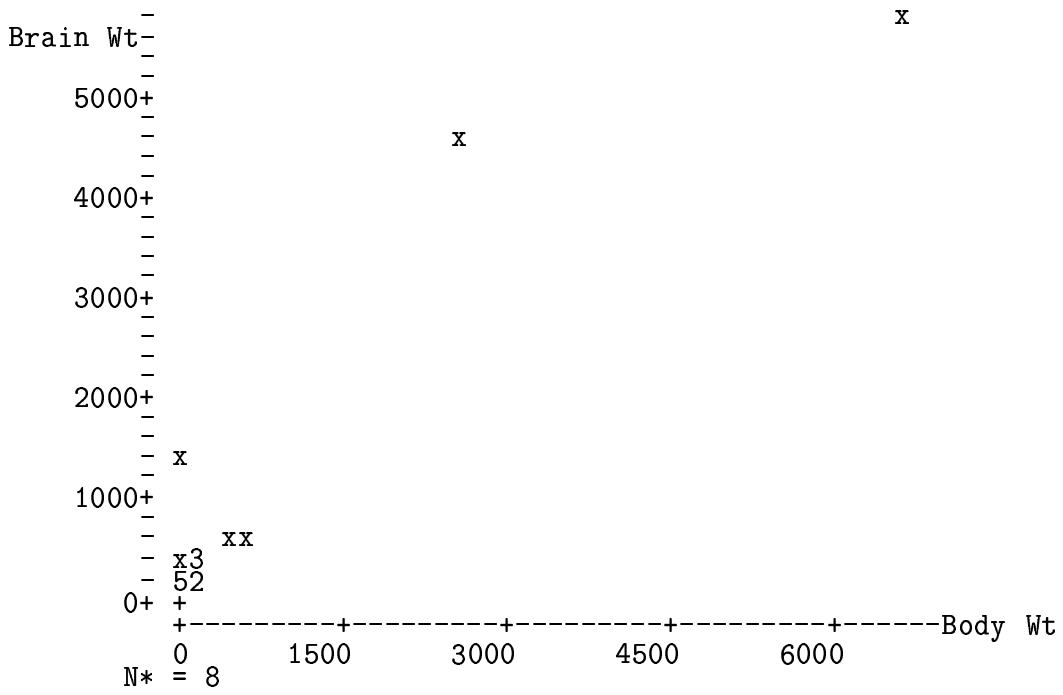
the model. Provide some justification for deleting any species from the analysis. Given the selected model, provide predictions and 95% prediction intervals for the brain weights of the 8 species with "missing" brain weights.

Data Display

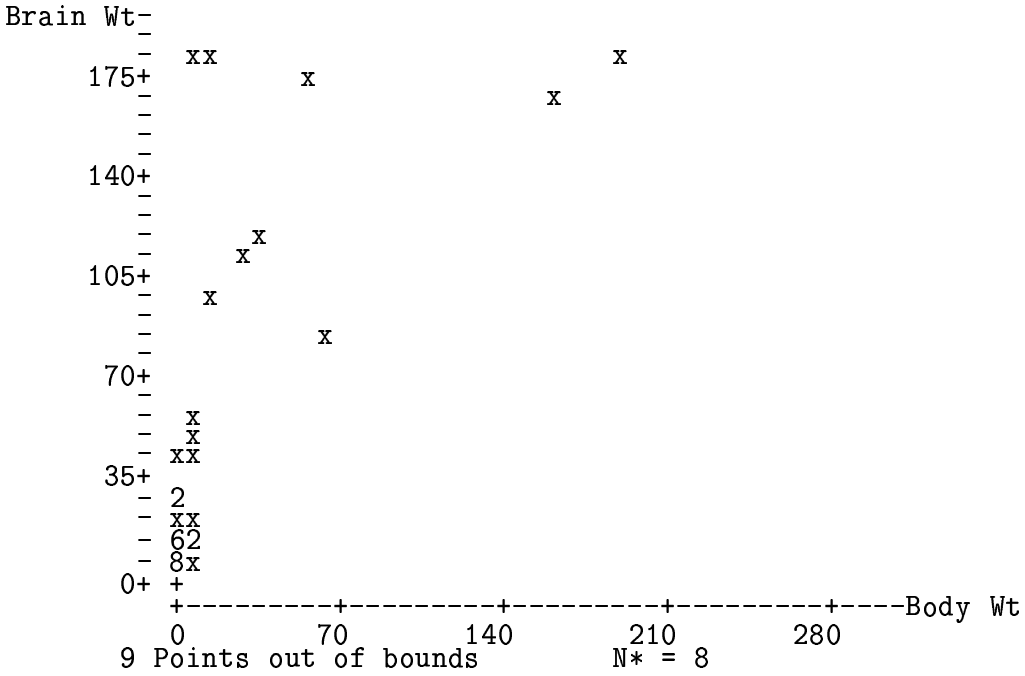
| Row | ID | Species | Body Wt | Brain Wt |
|---|---|---|---|---|
| 1 | 1 | Artic fox | 3.38 | 44.50 |
| 2 | 2 | Owl monkey | 0.48 | 15.50 |
| 3 | 3 | Mountain beaver | 1.35 | 8.10 |
| 4 | 4 | Cow | 465.00 | * |
| 5 | 5 | Gray wolf | 36.33 | 119.50 |
| 6 | 6 | Goat | 27.66 | 115.00 |
| 7 | 7 | Roe deer | 14.83 | 98.20 |
| 8 | 8 | Guinea pig | 1.04 | * |
| 9 | 9 | Vervet | 4.19 | 58.00 |
| 10 | 10 | Chinchilla | 0.42 | 6.40 |
| 11 | 11 | Ground squirrel | 0.10 | 4.00 |
| 12 | 12 | Arctic ground squirrel | 0.92 | 5.70 |
| 13 | 13 | Africa giant poached rat | 1.00 | 6.60 |
| 14 | 14 | Lesser short-tailed shre | .005 | 0.14 |
| 15 | 15 | Star-nosed mole | 0.06 | 1.00 |
| 16 | 16 | Nine-banded armadillo | 3.50 | 10.80 |
| 17 | 17 | Tree hyrax | 2.00 | 12.30 |
| 18 | 18 | N. American opussum | 1.70 | 6.30 |
| 19 | 19 | Asian elephant | 2547.00 | 4603.00 |
| 20 | 20 | Big brown bat | 0.02 | 0.30 |
| 21 | 21 | Donkey | 187.10 | 419.00 |
| 22 | 22 | Horse | 521.00 | 655.00 |
| 23 | 23 | European hedgehog | 0.79 | 3.50 |
| 24 | 24 | Patas monkey | 10.00 | * |
| 25 | 25 | Cat | 3.30 | 25.60 |
| 26 | 26 | Galago | 0.20 | 5.00 |
| 27 | 27 | Genet | 1.41 | 17.50 |
| 28 | 28 | Giraffe | 529.00 | 680.00 |
| 29 | 29 | Gorilla | 207.00 | 406.00 |
| 30 | 30 | Gray seal | 85.00 | 325.00 |
| 31 | 31 | Rock hyrax | 0.75 | 12.30 |
| 32 | 32 | Human | 62.00 | 1320.00 |
| 33 | 33 | African elephant | 6654.00 | 5712.00 |
| 34 | 34 | Water opussum | 3.50 | 3.90 |
| 35 | 35 | Rhesus monkey | 6.80 | 179.00 |
| 36 | 36 | Kangaroo | 35.00 | * |
| 37 | 37 | Yellow-bellied marmot | 4.05 | 17.00 |
| 38 | 38 | Golden hamster | 0.12 | 1.00 |
| 39 | 39 | Mouse | 0.02 | * |
| 40 | 40 | Little brown bat | 0.01 | 0.25 |
| 41 | 41 | Slow loris | 1.40 | 12.50 |
| 42 | 42 | Okapi | 250.01 | * |
| 43 | 43 | Rabbit | 2.50 | 12.10 |
| 44 | 44 | Sheep | 55.50 | 175.00 |
| 45 | 45 | Jaguar | 100.00 | * |
| 46 | 46 | Chimpanzee | 52.16 | 440.00 |
| 47 | 47 | Baboon | 10.55 | 179.50 |
| 48 | 48 | Desert hedgehog | 0.55 | 2.40 |
| 49 | 49 | Giant armadillo | 60.00 | 81.00 |
| 50 | 50 | Rock hyrax | 3.60 | 21.00 |

| | | | | |
|---|---|---|---|---|
| 51 | 51 | Raccoon | 4.29 | 39.20 |
| 52 | 52 | Rat | 0.28 | 1.90 |
| 53 | 53 | Eastern American mole | 0.07 | 1.20 |
| 54 | 54 | Mole rat | 0.12 | 3.00 |
| 55 | 55 | Musk shrew | 0.05 | 0.33 |
| 56 | 56 | Pig | 192.00 | 180.00 |
| 57 | 57 | Echidna | 3.00 | 25.00 |
| 58 | 58 | Brazilian tapir | 160.00 | 169.00 |
| 59 | 59 | Tenrec | 0.90 | 2.60 |
| 60 | 60 | Phalanger | 1.62 | 11.40 |
| 61 | 61 | Tree shrew | 0.10 | * |
| 62 | 62 | Red fox | 4.24 | 50.40 |

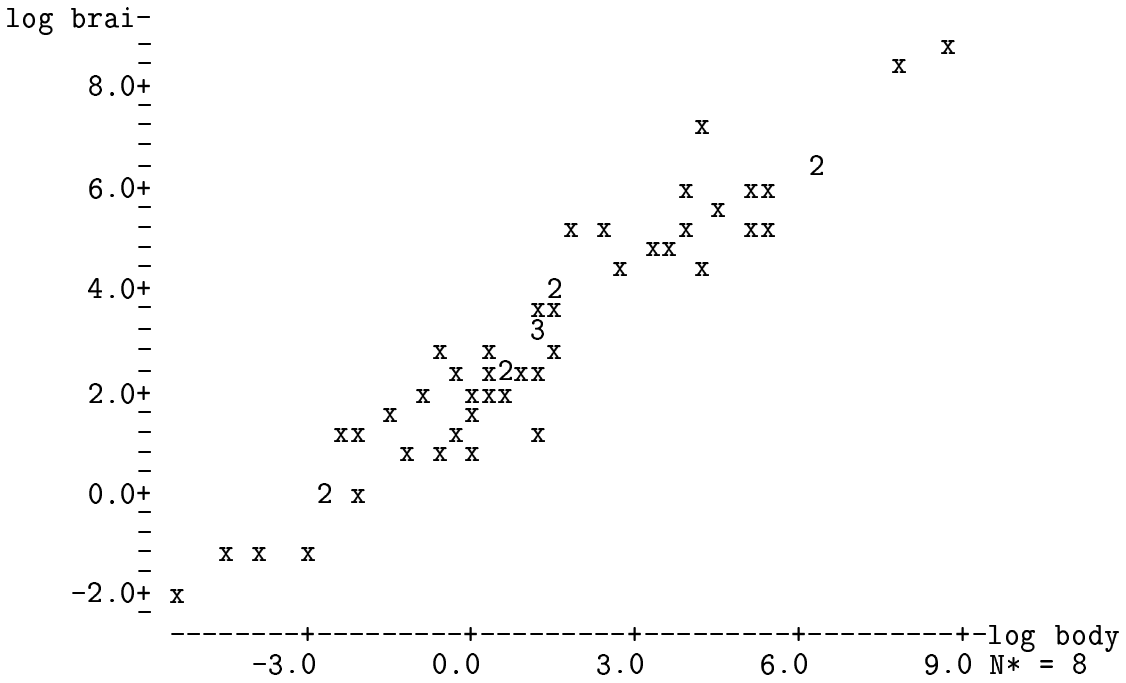
A plot of the brain weights against the body weights is non-informative because many species have very small brain weights and body weights compared to the elephants.




I get the following plot if I momentarily hold out the species with body weights exceeding $300kg$ or brain weights exceeding $200g$. This is easily done **MINITAB** by selecting the minimum and maximum values to be included in the plots. The plot shows that the brain weight of mammals typically increases with the body weight, but the relationship is nonlinear. The trend suggests transforming both variables to a logarithmic scale to linearize the relationship between brain weight and body weight. It does not matter which base logarithm you choose. The relationship is no more linear with one base than another. I will use natural logarithms.




I transformed the brain weights and body weights using the **MINITAB** calculator, and then replotted the data.



The plot of $\log_e$(brain weight) against $\log_e$(body weight) is fairly linear. At this point consider fitting the model:

$$\log_e(\text{brain weight}) = \beta_0 + \beta_1 \log_e(\text{body weight}) + \epsilon.$$

Summary information from fitting this model is given below.

```
Regression Analysis

The regression equation is: log brain = 2.14 + 0.765 log body

54 cases used 8 cases contain missing values

Predictor        Coef       StDev          T        P
Constant       2.1448      0.1042      20.59    0.000
log body      0.76498     0.03181      24.05    0.000

S = 0.7088      R-Sq = 91.8%     R-Sq(adj) = 91.6%

Analysis of Variance

Source             DF          SS          MS          F        P
Regression          1      290.58      290.58     578.38    0.000
Residual Error     52       26.13        0.50
Total              53      316.71

Unusual Observations
Obs   log body   log brai         Fit   StDev Fit    Residual    St Resid
 32       4.13     7.1854      5.3020      0.1332      1.8834        2.71R
 33       8.80     8.6503      8.8789      0.2593     -0.2286       -0.35 X
 34       1.25     1.3610      3.1032      0.0965     -1.7422       -2.48R
 35       1.92     5.1874      3.6112      0.0989      1.5762        2.25R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

The fitted relationship between $\log_e$(brain weight) and $\log_e$(body weight):

$$\text{Predicted}\log_e(\text{brain weight}) = 2.14 + .76\log_e(\text{body weight}),$$

explains about 92% of the variation in $\log_e$(brain weight). The $t-$test for $H_0 : \beta_1 = 0$ is highly significant ($p - value = .0001$). This summary information combined with the data plot indicates that there is a strong linear relationship between $\log_e$(brain weight) and $\log_e$(body weight), with the average $\log_e$(brain weight) increasing as $\log_e$(body weight) increases.

These conclusions are tentative, subject to a careful residual analysis. The residual summaries highlight three species with large $D_i$ or $r_i$ (where is case 33?):

| Case | Species | $r_i$ | $D_i$ |
|---|---|---|---|
| 32 | Man | 2.705 | .134 |
| 34 | Water opossum | -2.481 | .058 |
| 35 | Rhesus monkey | 2.246 | .050 |

These three species are easily spotted in the residual plot and the normal scores plot on the next page. Neither plot suggests any serious inadequacies with the model.

I decided to hold out the most influential species in the analysis, **Man**, and check the effects on the fitted model. Summary information for the fit with Man omitted is given below.

```
Regression Analysis

The regression equation is: log brain = 2.12 + 0.754 log body

53 cases used 8 cases contain missing values

Predictor         Coef       StDev           T         P
Constant       2.12273     0.09781       21.70     0.000
log body       0.75361     0.03003       25.09     0.000

S = 0.6634      R-Sq = 92.5%     R-Sq(adj) = 92.4%

Analysis of Variance

Source             DF          SS          MS          F         P
Regression          1      277.18      277.18     629.72     0.000
Residual Error     51       22.45        0.44
Total              52      299.63

Unusual Observations
Obs   log body   log brai         Fit   StDev Fit    Residual    St Resid
 32       8.80     8.6503      8.7568      0.2463     -0.1064       -0.17 X
 33       1.25     1.3610      3.0668      0.0912     -1.7059       -2.60R
 34       1.92     5.1874      3.5673      0.0938      1.6200        2.47R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

Deleting Man does not change our conclusion that there is a strong linear relationship between $\log_e$(brain weight) and $\log_e$(body weight). The following table shows that the deletion of Man has little effect on the summary information. In particular, the LS lines are similar and give relatively similar predicted brain weights.

| Feature | Full data | Omit Man |
|---|---|---|
| $b_0$ | 2.14 | 2.12 |
| $b_1$ | .76 | .75 |
| $SE(b_1)$ | .032 | .030 |
| $R^2$ | .918 | .925 |
| p-val for slope test | 0.0001 | 0.0001 |

Although including Man in the analysis hardly changes the fit of the model, there are some good reasons for deleting Man from the analysis. First, Man is a very **atypical** mammal with considerably greater brain weight than any other species with a similar body weight. This is why Man appears to be an outlier in the original analysis. Second, I would be uncomfortable using either model to predict Man's brain weight, for reasons just given. Thus, I will choose

$$\text{Predicted } \log_e(\text{brain weight}) = 2.12273 + .753612 \log_e(\text{body weight})$$

as the best model for summarizing the relationship for species other than Man. A residual analysis for this model shows no serious deficiencies. To predict brain weights, use the inverse transformation

$$\text{Predicted brain weight} = \exp\{\text{Predicted } \log_e(\text{brain weight})\}$$

or

$$\begin{aligned} \text{Predicted brain weight} &= \exp\{2.12273 + .753612 \log_e(\text{body weight})\} \\ &= \exp(2.12273) * \text{body weight}^{.753612} \\ &= 17.73 * \text{body weight}^{.75} \end{aligned}$$

to two decimal places.

**Remark:** If you used base 10 logarithms then

$$\text{Predicted brain weight} = 10^{\text{Predicted } \log_{10}(\text{brain weight})},$$

and the final prediction equation will be identical to that given above.

For the species with missing brain weights, I can get predicted brain weights and 95% prediction limits by following these steps:

1. Fit the model on the log-log scale, and save the fitted values and 95% prediction limits for the log brain weights of all species in the data. **MINITAB** will generate these summaries for all cases, including the 8 that were not used to fit the model!

2. Use the calculator to exponentiate the column of fitted values and the 95% prediction limits.

3. Print out these summaries for the selected cases.

Here is the result of that process when the model is fitted to the 53 species excluding man.

| ID | log(body wt) | pred log brainwt | pred brain wt | | |
|---|---|---|---|---|---|
| | | | | 95% prediction limits | |
| 4 | 6.14204 | 6.75145 | 855.29 | 215.76 | 3390.50 |
| 8 | 0.03922 | 2.15229 | 8.60 | 2.24 | 33.07 |
| 24 | 2.30259 | 3.85799 | 47.37 | 12.33 | 182.02 |
| 35 | 3.55535 | 4.80209 | 121.76 | 31.50 | 470.65 |
| 38 | -3.77226 | -0.72009 | 0.49 | 0.12 | 1.93 |
| 41 | 5.52150 | 6.28380 | 535.82 | 136.20 | 2107.91 |
| 44 | 4.60517 | 5.59324 | 268.61 | 68.93 | 1046.73 |
| 60 | -2.26336 | 0.41703 | 1.517 | 0.3893 | 5.91 |

For comparison, consider the predictions and prediction limits using the log-log model based on the ENTIRE data set. In the output I also included the actual brain weights for these species.

| ID | log(body wt) | true brwt | pred log brainwt | pred brain wt | | |
|---|---|---|---|---|---|---|
| | | | | | 95% prediction limits | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | 6.14204 | 423 | 6.84334 | 937.62 | 215.76 | 4074.61 |
| 8 | 0.03922 | 5.5 | 2.17483 | 8.80 | 2.09 | 37.05 |
| 24 | 2.30259 | 115 | 3.90625 | 49.71 | 11.81 | 209.20 |
| 36 | 3.55535 | 56 | 4.86459 | 129.62 | 30.62 | 548.75 |
| 39 | -3.77226 | .4 | -0.74086 | 0.48 | 0.11 | 2.07 |
| 42 | 5.52150 | 490 | 6.36864 | 583.27 | 135.28 | 2514.69 |
| 45 | 4.60517 | 157 | 5.66767 | 289.36 | 67.78 | 1235.37 |
| 61 | -2.26336 | 2.5 | 0.41341 | 1.512 | 0.3537 | 6.46 |

Given these summaries, which model appears to provide better predictions for the eight species?

I have illustrated a **cross-validation** method to assess the accuracy of predictions from a selected model. With a large data set, it is desirable to break the data into two parts (often halves). One part is used to build the model. Given the selected model, you compare the actual responses for observations in the second part with the predicted responses based on the model fit to the first part. This provides a means to assess how accurately the model will predict future responses.

**Class discussion?**

**An Important Final Point:**

The initial focus of a regression analysis should be modelling the trend correctly. If the trend is not modeled appropriately then the regression summaries (i.e. ANOVA table, p-values, CI, predictions, etc.) and assessments of poorly fitted and influential cases are not very informative.