Mathematica Laboratories for Mathematical Statistics

Emphasizing Simulation and Computer Intensive Methods

by Jenny A. Baglivo

ASA-SIAM Series on Statistics and Applied Probability

Mathematica Laboratories for Mathematical Statistics introduces an approach to incorporating technology in the mathematical statistics sequence, with an emphasis on simulation and computer intensive methods. The printed book is a concise introduction to the concepts of probability theory and mathematical statistics. The accompanying electronic materials are a series of in-class and take-home computer laboratory problems designed to reinforce the concepts, and to apply the techniques in real and realistic settings. The laboratory materials are written as *Mathematica* Version 5 notebooks (Wolfram Research, Inc.) and are designed so that students with little or no experience in *Mathematica* will be able to complete the work.

The materials are written to be used in the mathematical statistics sequence given at most colleges and universities (two courses of four semester hours each or three courses of three semester hours each). Multivariable calculus, and familiarity with the basics of set theory, vectors and matrices, and problem-solving using a computer are assumed. The order of topics generally follows that of a standard sequence. Chapters 1 through 5 cover concepts in probability. Chapters 6 through 10 cover introductory mathematical statistics. Chapters 11 and 12 are on permutation and bootstrap methods; in each case, problems are designed to expand on ideas from previous chapters so that instructors could choose to use some of the problems earlier in the course. Permutation and bootstrap methods also appear in the later chapters. Chapters 13, 14 and 15 are on multiple sample analysis, linear least squares and contingency tables, respectively. References for specialized topics in Chapters 10 through 15 are given at the beginning of each chapter.

Each chapter has a main laboratory notebook containing between five and seven problems, and a series of additional problem notebooks. The problems in the main laboratory notebook are for basic understanding, and can be used for in-class work or assigned for homework. The additional problem notebooks reinforce and/or expand the ideas from the main laboratory notebook and are generally longer and more involved.

This PDF File

This PDF file contains (I) the main laboratory notebook for Chapter 14 (linear least squares analysis), pages 2-11; (II) typical output from the examples in the notebook, pages 12-19; and (III) solutions to the problems in the notebook, pages 20-30.

Part I. Laboratory 14: Linear Least Squares

§1. Simple Linear Least Squares

Assume that the response random variable Y can be written as a linear function of the form

$$Y = \alpha + \beta X + \epsilon$$

where

- the predictor X and the error ϵ are independent random variables,
- The distribution of ϵ has mean 0 and standard deviation σ, and
- All parameters (α, β, σ) are unknown.

Then the conditional expectation of Y given $X = x$ is a linear function in x

$$E(Y \mid X = x) = \alpha + \beta x$$

and the standard deviation of the conditional distribution does not depend on x.

This section focuses on using least squares and permutation methods to estimate the parameters in the conditional mean formula using the Fit and SlopeCI functions. Please evaluate the following command before starting your work.

```
Needs["StatTools`Group1`"];
Needs["StatTools`Group2`"];
Needs["StatTools`Group3`"];
Needs["StatTools`Group4`"];
```

The form of the Fit function is as follows:

```
Fit[pairs,{1,x},x]
returns the estimated mean formula, where pairs is a list of pairs of real
numbers.
```

Note: The pairs are assumed to be the values of a random sample from the joint (X, Y) distribution or to have been generated independently from conditional distributions at fixed values of X.

Example 1:

To demonstrate the Fit function using simulation, let X be a uniform random variable on the interval [0,50], ϵ be a uniform random variable on the interval [–25,25] and

$$Y = 2 - 3X + \epsilon .$$

(1) Evaluate the following command to construct a list of 80 **pairs** from the joint (X, Y) distribution.

```
nn = 80;
xvals = RandomArray[UniformDistribution[0, 50], nn];
evals = RandomArray[UniformDistribution[-25, 25], nn];

yvals = 2 - 3 * xvals + evals;
pairs = Transpose[{xvals, yvals}];
```

The *y*-coordinates (**yvals**) are a linear function of the *x*-coordinates (**xvals**) and a random sample from the error distribution (**evals**).

(2) Evaluate the following command to construct a scatter plot of the observations and to display the sample correlation.

```
ScatterPlot[pairs,
 Correlation → True]
```

There is a strong negative association between *X* and *Y*.

(3) Evaluate the following command to define a function, f, whose value at *x* is the estimated conditional mean.

```
Clear[x]; Remove[f];
f[x_] = Fit[pairs, {1, x}, x]
```

Fit uses the pairs list to find least squares estimates of intercept and slope (the coefficients of 1 and *x*, respectively) as a function of *x*. The estimated formula is stored as the value of f(*x*). The graph of $y = f(x)$ is the gray line in the scatter plot in step (2).

(4) Repeat the commands in steps (1) through (3) several times to see different plots and estimates of the conditional expectation. Then repeat the simulation several times each assuming the error distribution is uniform on the interval [–5,+5] and assuming it is uniform on the interval [–100,+100]. In the first case, the correlations will be very close to –1 and the estimated formulas close to $2–3x$. In the second case, the correlations will be close to –0.60 and the estimated formulas will be much more variable.

■ Permutation confidence interval for β

Permutation methods can be used to construct $100(1–\gamma)\%$ confidence intervals for the slope parameter β. A value β_0 is in the confidence interval if the two sided test of

H_0 : The correlation between X and $Y - \beta_0 X$ equals zero

accepts H_0 at the γ significance level. The SlopeCI function uses simulation to approximate the permutation confidence interval:

```
SlopeCI[pairs,ConfidenceLevel→1-γ,RandomPermutations→r]
returns an approximate 100(1-γ)% permutation confidence interval for the
slope based on r random permutations, where pairs is a list of pairs of
numbers.
```

Note: If the options are omitted, then SlopeCI returns an approximate 95% permutation confidence interval based on 1000 random permutations.

Example 1, continued:

Evaluate the first command to initialize a list of 80 **pairs** from the joint (*X*, *Y*) distribution defined above. Evaluate the second command to construct an approximate 95% confidence interval for β using 2000 random permutations.

```
nn = 80;
xvals = RandomArray[UniformDistribution[0, 50], nn];
evals = RandomArray[UniformDistribution[-25, 25], nn];

yvals = 2 - 3 * xvals + evals;
pairs = Transpose[{xvals, yvals}];

SlopeCI[pairs, RandomPermutations → 2000]
```

Repeat the commands several times. The computed interval will contain $\beta = -3$ with probability approximately 0.95.

Problem 1: Assume that X and ϵ are independent uniform random variables, the range of X is [0, 80], and the range of ϵ is [–50,+50]. Let $Y = 2 - 3X + \epsilon$.

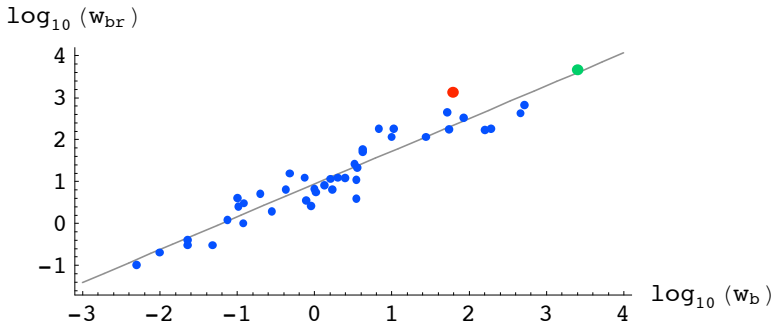
(a) Generate a random sample (pairs) of size 100 from the joint (X, Y) distribution. For these data,

- Compute the sample mean and sample standard deviation of the x- and y-coordinates.
- Construct a scatter plot of the pairs and display the sample correlation.
- Use Fit to estimate the conditional expectation.
- Use SlopeCI to construct an approximate 95% confidence interval for the slope based on 2000 random permutations. Is –3 in the interval?

(b) Compute E(X), SD(X), E(Y), SD(Y), and ρ=Corr(X,Y). Are the sample summaries (the sample means, standard deviations, and correlation) from part (a) close to these model summaries?

Example 2:

As part of a study on sleep in mammals, researchers collected information on the average brain and body weights for 43 different species. The graph below compares the common logarithms of average brain weight in grams (vertical axis) and average body weight in kilograms (horizontal axis) for the 43 species. The largest log-average brain weight (the green dot) corresponds to the Asian elephant; the second largest (the red dot) to man. The gray line is the least squares linear fit to the paired data. (Sources: Allison and Cicchetti, 1976; lib.stat.cmu.edu/DASL.)



Common logs of average brain (vertical axis) and body (horizontal axis) weights for 43 species of mammals.

The lists **species**, **wbody**, and **wbrain** give the species names (listed alphabetically) and corresponding body and brain weights.

species, **wbody**, **wbrain** are lists of length 43.

Body weights range from 0.005 kg (0.18 ounces) to 2547.0 kg (5,615.12 pounds). Brain weights range from 0.14 g (0.004 ounces) to 4603.0 g (10.15 pounds). To initialize the data, click on the rightmost bracket of the cell above and evaluate the command.

(1) Evaluate the following command to construct the list of pairs displayed above.

```
pairs = Transpose[{Log[10, wbody], Log[10, wbrain]}];
```

Note that the common logarithm is the logarithm function with base 10.

(2) Evaluate the following command to display a table of the paired data along with the species names.

```
TableForm[pairs,
 TableHeadings → {species, {"log10 (wb)", "log10 (wbr)"}}]
```

Problem 2:
(a) Using the paired (log-body weight, log-brain weight) data,

- Use Fit to find the least squares estimate of the conditional expectation of log-brain weight given log-body weight.
- Use SlopeCI to construct an approximate 95% confidence interval for the slope.
- Interpret the estimated slope in the context of the brain-body problem.

(b) One of our mammalian cousins, the gorilla, has been left off the list of species. The gorilla has an average body weight of 207.0 kg (456.35 pounds) and an average brain weight of 406.0 g (14.32 ounces).

Use the least squares formula from part (a) to estimate the gorilla's average brain weight from its average body weight. Is the estimated average brain weight close to the true average brain weight?

(c) Use the least squares formula from part (a) to define a function g whose input is an average body weight and whose output is an estimate of the average brain weight. Then evaluate the command below to produce a smoothed scatter plot of (w_b, w_{br}) pairs with the graph of $y = g(x)$ superimposed. Comment on the plot.

```
pairs2 = Transpose[{wbody, wbrain}];
SmoothPlot[{pairs2, g},
 AxesLabel → {"wb", "wbr"}]
```

Note: SmoothPlot generalizes ScatterPlot with the Correlation→True option. It is used to visualize non-linear relationships. Evaluate the command ?SmoothPlot to obtain information on this function.

§ 2. Linear Regression Analysis

Assume that the response random variable Y can be written as a linear function of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon$$

where

- The error random variable, ϵ, is independent of each predictor, X_i,
- ϵ is a normal random variable with mean 0 and standard deviation σ, and
- All p+1 parameters (the β_i's and σ) are unknown.

Let

$$\underline{X} = (1, X_1, X_2, \ldots, X_{p-1})$$

represent the list including the constant 1 and the p–1 predictors (the p basis functions).

Then the conditional expectation of Y given $\underline{X}=\underline{x}$ is a linear function:

$$E(Y \mid \underline{X} = \underline{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1}$$

and the conditional distribution is normal with standard deviation σ.

This section focuses on using least squares methods to estimate the parameters in the conditional mean formula using the Fit and Regress functions. The forms of the functions are as follows:

```
Fit[cases,functions,variables]
returns the estimated mean formula, where cases is the list of observations,
functions is the list of p basis functions, and variables is a single
variable or a list of variables needed in the formula.
```

```
Regress[cases,functions,variables,RegressionReport→options]
returns a list of rules for linear regression analysis.
```

Notes:

(1) The predictors, X_i, need not be mutually independent random variables. For numerical stability, the predictors should not be strongly correlated.

(2) If there are k variables, then each case must be a list of k+1 numbers (k variables plus response). The p basis functions must be functions of the k variables only.

(3) The following options list will be used in this section:

{ANOVATable, ParameterCITable, RSquared, EstimatedVariance, StandardizedResiduals}.

Additional options will be added in Section 3.

(4) See the Add-ons section of the Help Browser for other options and examples.

Example 3:

To demonstrate the analysis using simulation, assume that X is a uniform random variable on the interval [0, 20], ϵ is a normal random variable with mean 0 and standard deviation 8, and X and ϵ are independent. Let

$$Y = 40 - 15\,X + 0.60\,X^2 + \epsilon = -50 - 3\,(X - 10) + 0.60\,(X - 10)^2 + \epsilon.$$

(1) Evaluate the following command to construct a list of 100 **cases** from the joint (X, Y) distribution.

```
nn = 100;
xvals = RandomArray[UniformDistribution[0, 20], nn];
evals = RandomArray[NormalDistribution[0, 8], nn];

yvals = 40 - 15 * xvals + 0.60 * xvals ^ 2 + evals;
cases = Transpose[{xvals, yvals}];
```

The y-coordinates are generated using the first form of the linear equation.

(2) Evaluate the following command to use Fit to estimate the conditional expectation of Y given $X = x$ using the predictors for the second form of the linear equation.

```
Clear[x]; Remove[f];
f[x_] = Fit[cases, {1, x - 10, (x - 10) ^ 2}, x]
```

The estimated formula is stored as the value of f(x).

(3) Evaluate the following command to construct a scatter plot of the cases list with the estimated conditional expectation superimposed.

```
SmoothPlot[{cases, f}]
```

The least squares estimate is likely to approximate the observed pairs quite well. (If the standard deviation of ϵ was 28 instead of 8, for example, the fit might not so close.)

Problem 3: Let X be a uniform random variable on the interval [0, 20]. Compute

- The correlation between X and X^2.
- The correlation between $(X - 10)$ and $(X - 10)^2$.

■ Analysis of variance

Let N be the number of cases, p be the number of functions,

- $\underline{x}_i$ and Y_i $(i = 1, 2, \ldots, N)$ be the N observed predictor lists and responses, respectively,
- $f(\underline{x}_i)$ $(i = 1, 2, \ldots, N)$ be the estimated means (or predicted responses), and
- $\overline{Y}$ be the mean response.

A test of the null hypothesis that the regression effects are identically zero (or $\beta_j = 0$ for $j > 0$) has three sources of variation, as outlined in the following table:

	DF	SumOfSq	MeanSq
Model	p - 1	$SS_m = \sum_i (f(\underline{x}_i) - \overline{Y})^2$	$MS_m = SS_m / (p - 1)$
Error	N - p	$SS_e = \sum_i (Y_i - f(\underline{x}_i))^2$	$MS_e = SS_e / (N - p)$
Total	N - 1	$SS_t = \sum_i (Y_i - \overline{Y})^2$	

If the null hypothesis is true, then the ratio $F = MS_m / MS_e$ has an f ratio distribution with $p - 1$ and $N - p$ degrees of freedom. Large values of F provide evidence that the proposed predictors have some predictive value.

Example 3, continued

(1) Evaluate the following command to initialize a list of 100 **cases** from the joint (X, Y) distribution above.

```
nn = 100;
xvals = RandomArray[UniformDistribution[0, 20], nn];
evals = RandomArray[NormalDistribution[0, 8], nn];

yvals = 40 - 15 * xvals + 0.60 * xvals ^ 2 + evals;
cases = Transpose[{xvals, yvals}];
```

(2) Evaluate the following command to compute a list of **results** based on the cases from step (1) and the predictors for the second form of the linear equation.

```
results = Regress[cases, {1, (x - 10), (x - 10) ^ 2}, x,
   RegressionReport → {ANOVATable, ParameterCITable,
     RSquared, EstimatedVariance, StandardizedResiduals}];
```

(3) Evaluate the following command to retrieve the analysis of variance table.

```
ANOVATable /. results
```

The p value is likely to be virtually zero.

(4) Evaluate the following command to retrieve the coefficient of determination, R^2.

```
RSquared /. results
```

$R^2 = SS_m / SS_t$ is the proportion of the total variation explained by the proposed model. Approximately 90% of the total variation is explained by the model in this case. (Note that as σ increases, the value of R^2 generally decreases.)

■ Parameter estimates and standardized residuals

We next examine 95% confidence intervals for the β coefficients, compute the pooled estimate of the common standard deviation, and construct an enhanced normal probability plot of estimated standardized residuals.

Example 3, continued:

(1) Evaluate the following command to retrieve information about the β parameter estimates:

```
ParameterCITable /. results
```

The point estimates are likely to be close to –50, –3, and 0.60. The 95% confidence intervals are likely to indicate that each coefficient is significantly different from zero at the 5% significance level.

(2) Evaluate the following command to compute the pooled estimate of the common standard deviation.

```
Sqrt[EstimatedVariance] /. results
```

The value is likely to be close to 8.

(3) Evaluate the following command to retrieve the list of estimated **standardized** residuals and to construct an enhanced normal probability plot of standardized residuals.

```
standardized = StandardizedResiduals /. results;
ProbabilityPlot[NormalDistribution[0, 1], standardized,
 SimulationBands → True]
```

The points should approximate the line $y = x$.

Example 4:

In the study on sleep in mammals (Example 2 and Problem 2), researchers examined the relationship between non-dreaming or slow-wave sleep (SWS) and two variables: the average body weight (w_b) and an overall danger index. The danger index is a five-point scale where 1 indicates the least danger (from predation, exposure to the elements, etc.) and 5 indicates the most danger. The indices for the 43 species were as follows:

Danger = 1	Danger = 2	Danger = 3	Danger = 4	Danger = 5
Big brown bat	European hedgehog	African giant rat	Asian elephant	Cow
Cat	Galago	Ground squirrel	Baboon	Goat
Chimpanzee	Golden hamster	Mountain beaver	Brazilian tapir	Horse
E. Amer. mole	Owl monkey	Mouse	Chinchilla	Rabbit
Gray seal	Phanlanger	Musk shrew	Guinea pig	Sheep
Little brown bat	Rhesus monkey	Rat	Short tail shrew	
Man	Rock hyrax (h.b.)	Rock hyrax (p.h.)	Patas monkey	
Mole rat	Tenrec	Tree hyrax	Pig	
N.Amer. opossum	Tree shrew		Vervet	
Nine banded armadillo				
Red fox				
Water opossum				

The researchers determined that a model of the form

$$\log_{10}(\text{SWS}) = \beta_0 + \beta_1 \log_{10}(w_b) + \beta_2 \,\text{danger} + \epsilon,$$

where ϵ is a normal random variable with mean 0, approximated the data reasonably well.

The lists **danger** and **sws** give the danger indices and the values of slow-wave sleep in hours for the 43 species (in alphabetical order).

danger, **sws** are lists of length 43.

To initialize the data, click on the rightmost bracket of the cell above and evaluate the command. If necessary, re-initialize the data in Example 2.

(1) Evaluate the first command to initialize the list of 43 cases. Evaluate the second command to use Fit to compute the estimated conditional expectation.

```
cases = Transpose[{Log[10, wbody], danger, Log[10, sws]}];

Clear[x1, x2]; Remove[f];
f[x1_, x2_] = Fit[cases, {1, x1, x2}, {x1, x2}]
```

Note that each element of the cases list is of the form $\{x_1, x_2, y\}$, where x_1 corresponds to log-average body weight, x_2 corresponds to the danger index, and y corresponds to log-SWS.

(2) Evaluate the following command to view the relationship of

- log-SWS adjusted for the effect of danger (vertical axis) against
- log-w_b (the first variable) adjusted for the effect of danger (horizontal axis)

and to display the partial regression line.

```
PartialPlot[cases, 1]
```

The slope of the partial regression line is the estimate of β_1 from step (1). Repeat the PartialPlot command using 2 instead of 1 as the second argument to view the partial relationship between log-SWS (adjusted for log-w_b) and danger (adjusted for log-w_b).

Note: PartialPlot generalizes ScatterPlot with the Correlation→True option. Evaluate the command ?PartialPlot to obtain more information on this function.

Problem 4:
(a) Use Regress to analyze the SWS cases data. Report

- the p value from the analysis of variance f test,
- the coefficient of determination,
- 95% confidence intervals for the β parameters, and
- the estimated standard deviation of the error distribution.

In addition, construct an enhanced normal probability plot of standardized residuals. Comment on the computations.

(b) Use the least squares fitted formula from step (1) of the example to construct five lists of pairs (pairs1 for animals with danger score 1, pairs2 for animals with danger score 2, etc.) of elements of the form

$\{x, \text{sws}_x\}, \; x = -1, 0, 1, 2, 3$

where sws_x is an estimate of the number of hours of SWS sleep for an animal with body weight 10^x kg. Then evaluate the command below to plot the 5 pairs lists. Comment on the plot.

```
ScatterPlot[pairs1, pairs2, pairs3, pairs4, pairs5,
 PlotJoined → True,
 AxesLabel → {"log10(wb)", "SWS"}]
```

§ 3. Regression Diagnostics

This section focuses on additional methods for interpreting the results of a linear regression analysis. Specifically, scatter plots of

- (estimated mean, estimated error) pairs, and
- (case number, estimated standardized influence) pairs

will be constructed and interpreted.

Note that the estimated means, errors, and standardized influences are returned with the regression results by adding

PredictedResponse, FitResiduals, PredictedResponseDelta

to the regression report options list.

Example 5:

To demonstrate the plots using simulation, assume that X is a uniform random variable on the interval [0, 80], ϵ is a normal random variable with mean 0 and standard deviation 10, and X and ϵ are independent. Let

$$Y = 2 + 3X + \epsilon.$$

(1) Evaluate the following command to construct a random sample (**pairs**) of size 50 from the joint (X, Y) distribution.

```
nn = 50;
xvals = RandomArray[UniformDistribution[0, 80], nn];
evals = RandomArray[NormalDistribution[0, 10], nn];
yvals = 2 + 3 * xvals + evals;
pairs = Transpose[{xvals, yvals}];
```

(2) Evaluate the following command to replace the first observed pair with the pair (40, 202) and to construct a scatterplot of the altered pairs list.

```
pairs[[1]] = {40, 202};
ScatterPlot[pairs, Correlation → True]
```

The plot shows a strong positive association, but there is a point with an unusually large y-coordinate.

(3) Evaluate the following command to construct a list of regression **results**, and retrieve the estimated errors (**residuals**), **means**, and standardized influences (**deltas**).

```
Clear[x];
results = Regress[pairs, {1, x}, x,
   RegressionReport → {ANOVATable, ParameterCITable,
     RSquared, EstimatedVariance, StandardizedResiduals,
     PredictedResponse, FitResiduals, PredictedResponseDelta}];

residuals = FitResiduals /. results;
means = PredictedResponse /. results;
deltas = PredictedResponseDelta /. results;
```

(4) Evaluate the following command to compare estimated means and errors:

```
rpairs = Transpose[{means, residuals}];
ScatterPlot[rpairs,
 AxesLabel → {"m", "e"},
 Correlation → True]
```

The apparent large outlier should be the only pattern visible in the scatter plot.

(5) Evaluate the first command to construct a plot of standardized influences. Evaluate the second command to construct a list of case-delta pairs with standardized influence outside the interval

$$\left[-2\sqrt{p/N},\ +2\sqrt{p/N}\right] = [-0.40,\ +0.40].$$

```
dpairs = Transpose[{Range[1, nn], deltas}];
ScatterPlot[dpairs,
 AxesLabel → {"i", "δ"}]

p = 2; nn = 50; cutoff = N[2 * Sqrt[p / nn]];
Select[dpairs, (Abs[Last[#]] > cutoff) &]
```

The plot shows a single point (the first) whose y-coordinate is much larger than the others, indicating that the prediction for $x = 40$ is much larger when all 50 points are included in the regression than when only the 49 remaining points are included. Case 1 is highly influential. The y-coordinates of the remaining points generally fall in the interval [–0.40, +0.40].

(6) Repeat the simulation several times each using {40, 202} and using {40, 42} as the first point to see different diagnostic plots. To see more unusual plots, try changing the first two points.

Problem 4, continued:
(c) Construct and interpret an (estimated mean, estimated error)-pairs plot using the SWS cases data.
(d) Construct and interpret a (case number, standardized influence)-pairs plot using the SWS cases data.

Example 6:

The sleep researchers also compared dreaming or paradoxical sleep (PS) in hours to other ecological and environmental factors, including the average gestation time (t_g) in days for the species and the danger index. They determined that a model of the form

$$\log_{10}(\text{PS}) = \beta_0 + \beta_1 \log_{10}(t_g) + \beta_2\ \text{danger} + \epsilon,$$

where ϵ is a normal random variable with mean 0, approximated the data reasonably well.

The lists **ps** and **tgestation** give the PS values (in hours) and the t_g values (in days) for the 43 species (in alphabetical order).

ps, **tgestation** are lists of length 43.

Click on the rightmost bracket of the cell above and evaluate the command to initialize the data. Re-initialize the data in Examples 2 and 4, if necessary.

Problem 5:
(a) Construct a PS cases list where x_1 corresponds to $\log_{10}(t_g)$, x_2 corresponds to danger, and y corresponds to $\log_{10}(\text{PS})$. Use Fit to determine estimates of the β parameters in the formula above. Use PartialPlot to examine the partial regression plots. Comment on the computations.

(b) Repeat Problem 4(a), 4(c), and 4(d) using the PS cases list.

(c) Use the least squares estimated formula from part (a) to construct five lists of pairs (pairs1 for animals with danger score 1, pairs2 for animals with danger score 2, etc.) of elements of the form

$\{x, \text{ps}_x\}$, $x = 20, 80, 140, \ldots, 620$

where ps_x is an estimate of the number of hours of PS sleep for a species with average gestation period equal to x days. Then evaluate the command below to plot the 5 pairs lists. Comment on the plot.

```
ScatterPlot[pairs1, pairs2, pairs3, pairs4, pairs5,
 PlotJoined → True,
 AxesLabel → {"tg", "PS"}]
```

Part II. Laboratory 14 Examples

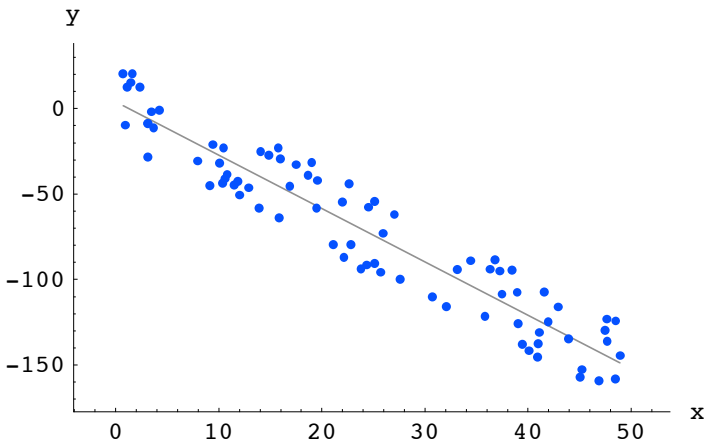
Example 1:

X is a uniform random variable on [0,50], ϵ is a uniform random variable on [–25,25] and

$$Y = 2 - 3X + \epsilon .$$

Students are asked to generate and work with 80 **pairs** from the joint (X, Y) distribution:

```
nn = 80;
xvals = RandomArray[UniformDistribution[0, 50], nn];
evals = RandomArray[UniformDistribution[-25, 25], nn];

yvals = 2 - 3 * xvals + evals;
pairs = Transpose[{xvals, yvals}];

ScatterPlot[pairs,
 Correlation → True]
```




```
Correlation:       -0.9501
```

```
Clear[x]; Remove[f];
f[x_] = Fit[pairs, {1, x}, x]
```

```
3.91097 - 3.1216 x
```

Students are asked to repeat the simulation several times each using uniform error distributions on

$$[-25, 25], [-5, 5], \quad [-100, 100],$$

and to compare simulation results.

Example 1, continued:

Students are asked to construct an approximate 95% confidence interval for β using 2000 random permutations.

```
SlopeCI[pairs, RandomPermutations → 2000]
```

```
{-3.35288, -2.88548}
```

Example 2:

Students are introduced to a study and asked to setup the paired data needed in Problem 2. They can also view the pairs, along with the species names, using a TableForm command. (The size has been reduced so that the output fits on this page.)

```
pairs = Transpose[{Log[10, wbody], Log[10, wbrain]}];

TableForm[pairs,
 TableHeadings → {species, {"log10(wb)", "log10(wbr)"}}]
```

	$\log_{10}(w_b)$	$\log_{10}(w_{br})$
African giant rat	0.	0.819544
Asian elephant	3.40603	3.66304
Baboon	1.02325	2.25406
Big brown bat	-1.63827	-0.522879
Brazilian tapir	2.20412	2.22789
Cat	0.518514	1.40824
Chimpanzee	1.71734	2.64345
Chinchilla	-0.371611	0.80618
Cow	2.66745	2.62634
E. Amer. mole	-1.12494	0.0791812
European hedgehog	-0.10513	0.544068
Galago	-0.69897	0.69897
Goat	1.44185	2.0607
Golden hamster	-0.920819	0.
Gray seal	1.92942	2.51188
Ground squirrel	-0.995679	0.60206
Guinea pig	0.0170333	0.740363
Horse	2.71684	2.81624
Short tail shrew	-2.30103	-1.
Little brown bat	-2.	-0.69897
Man	1.79239	3.12057
Mole rat	-0.91364	0.477121
Mountain beaver	0.130334	0.908485
Mouse	-1.63827	-0.39794
Musk shrew	-1.31876	-0.522879
N. Amer. opossum	0.230449	0.799341
Nine banded armadillo	0.544068	1.03342
Owl monkey	-0.318759	1.19033
Patas monkey	1.	2.0607
Phanlanger	0.209515	1.0569
Pig	2.2833	2.25527
Rabbit	0.39794	1.08279
Rat	-0.552842	0.278754
Red fox	0.626853	1.70243
Rhesus monkey	0.832509	2.25285
Rock hyrax (h.b.)	-0.124939	1.08991
Rock hyrax (p.h.)	0.556303	1.32222
Sheep	1.74429	2.24304
Tenrec	-0.0457575	0.414973
Tree hyrax	0.30103	1.08991
Tree shrew	-0.982967	0.39794
Vervet	0.622214	1.76343
Water opossum	0.544068	0.591065

Example 3:

X is a uniform random variable on [0, 20], ϵ is a normal random variable with $\mu = 0$ and $\sigma = 8$, X and ϵ are independent, and

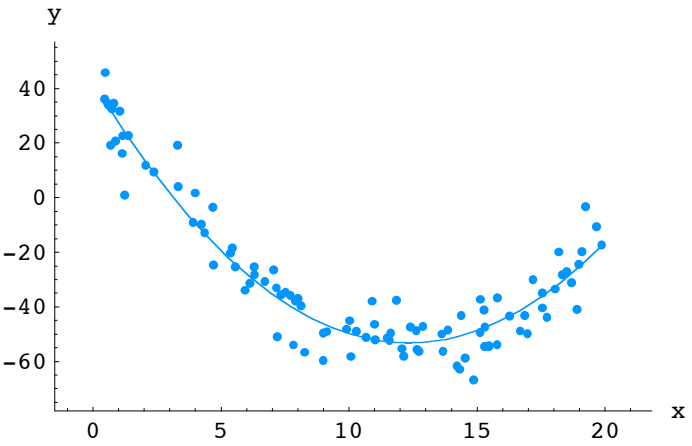
$$Y = 40 - 15\,X + 0.60\,X^2 + \epsilon = -50 - 3\,(X - 10) + 0.60\,(X - 10)^2 + \epsilon.$$

Students are asked to generate and work with 100 **cases** from the joint (X, Y) distribution.

```
nn = 100;
xvals = RandomArray[UniformDistribution[0, 20], nn];
evals = RandomArray[NormalDistribution[0, 8], nn];

yvals = 40 - 15 * xvals + 0.60 * xvals ^ 2 + evals;
cases = Transpose[{xvals, yvals}];

Clear[x]; Remove[f];
f[x_] = Fit[cases, {1, x - 10, (x - 10) ^ 2}, x]
```

```
-49.8745 - 2.94747 (-10 + x) + 0.628676 (-10 + x)^2
```

```
SmoothPlot[{cases, f}]
```



Students can then compare these simulation results with results obtained using $\sigma = 28$ instead of $\sigma = 8$.

Example 3, continued

The 100 **cases** from the joint (X, Y) distribution above can be analyzed using Regress.

```
results = Regress[cases, {1, (x - 10), (x - 10) ^ 2}, x,
   RegressionReport → {ANOVATable, ParameterCITable,
     RSquared, EstimatedVariance, StandardizedResiduals}];

ANOVATable /. results
```

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	71792.	35896.	574.567	0.
Error	97	6060.07	62.4749		
Total	99	77852.1			

```
RSquared /. results
```

```
0.922159
```

Example 3, continued:

Continuing with the regression analysis of the 100 **cases** from the joint (*X*, *Y*) distribution above:

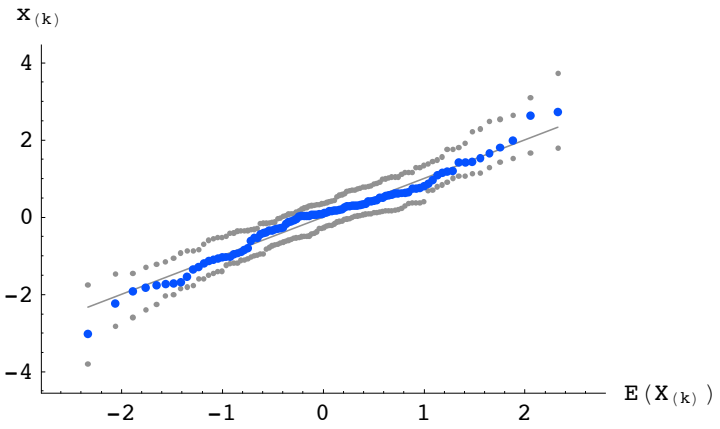
```
ParameterCITable /. results
```

	Estimate	SE	CI
1	-49.8745	1.18371	{-52.2238, -47.5251}
-10 + x	-2.94747	0.135425	{-3.21625, -2.67869}
$(-10 + x)^2$	0.628676	0.025648	{0.577772, 0.67958}

```
Sqrt[EstimatedVariance] /. results
```

```
7.90411
```

```
standardized = StandardizedResiduals /. results;
ProbabilityPlot[NormalDistribution[0, 1], standardized,
 SimulationBands → True]
```



Students can then compare these simulation results with results obtained larger values of σ.

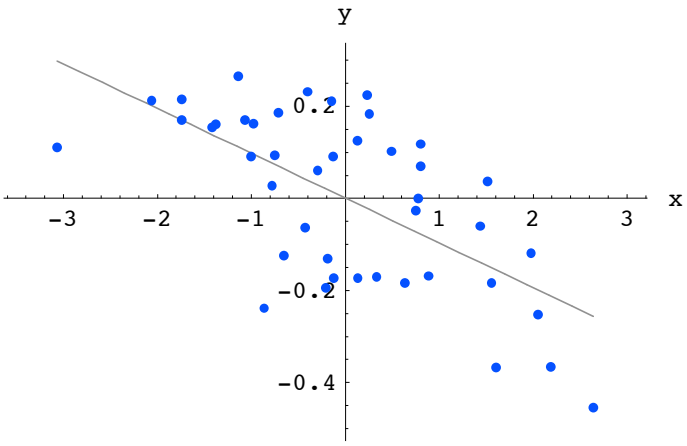
Example 4:

Students are introduced to a study and asked to setup a list of cases for analysis here and in Problem 4.

```
cases = Transpose[{Log[10, wbody], danger, Log[10, sws]}];

Clear[x1, x2]; Remove[f];
f[x1_, x2_] = Fit[cases, {1, x1, x2}, {x1, x2}]
```

```
1.06671 - 0.097165 x1 - 0.0539304 x2
```

```
PartialPlot[cases, 1]
```




```
Equation of line:  y = -0.097165 x
```

Students construct a partial regression plot for the second predictor as well.

Example 5:

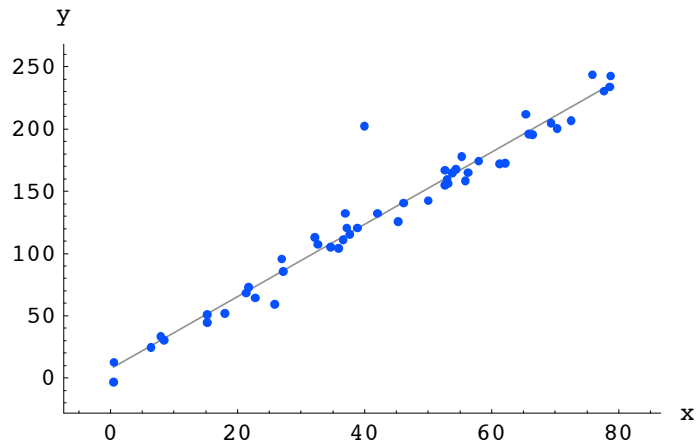
X is a uniform random variable on [0, 80], ϵ is a normal random variable with $\mu = 0$ and $\sigma = 10$, X and ϵ are independent and

$$Y = 2 + 3X + \epsilon.$$

Students construct 50 random **pairs** from the joint (X, Y) distribution and change the first pair.

```
nn = 50;
xvals = RandomArray[UniformDistribution[0, 80], nn];
evals = RandomArray[NormalDistribution[0, 10], nn];

yvals = 2 + 3 * xvals + evals;
pairs = Transpose[{xvals, yvals}];
```

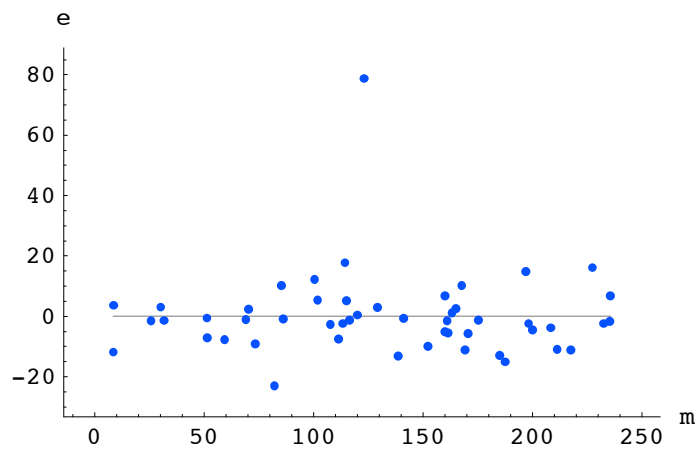
```
pairs[[1]] = {40, 202};
ScatterPlot[pairs, Correlation → True]
```




```
Correlation:        0.9759
```

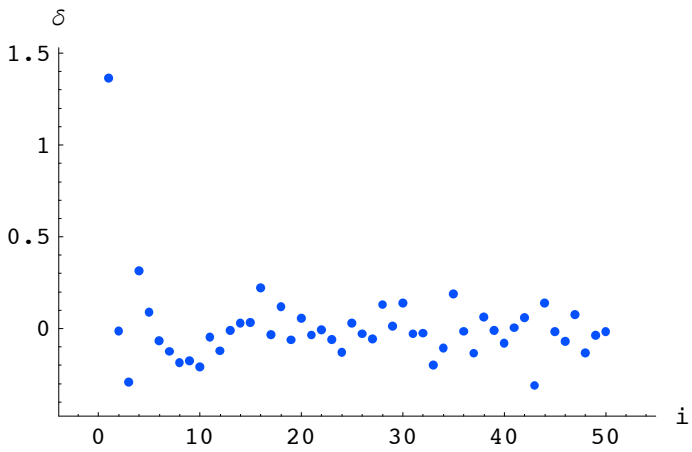
Output of the Regress function is used to construct (1) a plot of estimated errors versus estimated means and (2) a plot of standardized influences versus index number. Cases that are highly influential are selected.

```
Clear[x];
results = Regress[pairs, {1, x}, x,
   RegressionReport → {ANOVATable, ParameterCITable,
     RSquared, EstimatedVariance, StandardizedResiduals,
     PredictedResponse, FitResiduals, PredictedResponseDelta}];

residuals = FitResiduals /. results;
means = PredictedResponse /. results;
deltas = PredictedResponseDelta /. results;

rpairs = Transpose[{means, residuals}];
ScatterPlot[rpairs,
 AxesLabel → {"m", "e"},
 Correlation → True]
```




```
Correlation:        7.034 × 10^-16
```

```
dpairs = Transpose[{Range[1, nn], deltas}];
ScatterPlot[dpairs,
 AxesLabel → {"i", "δ"}]
```




```
p = 2; nn = 50; cutoff = N[2 * Sqrt[p / nn]];
Select[dpairs, (Abs[Last[#]] > cutoff) &]
```

```
{{1, 1.36365}}
```

Students are asked to repeat the simulation several times each using

(40, 202), (40, 42)

as first pair, and to compare the simulation results.

Example 6:

Students are introduced to a study and asked to initialize the data needed in Problem 5.

Part III. Laboratory 14 Solutions

Problem 1: $Y = 2 - 3X + \epsilon$ where X is uniform on [0,80] and ϵ is uniform on [–50,50].

(a) 100 Simulated pairs:

```
nn = 100;
xvals = RandomArray[UniformDistribution[0, 80], nn];
evals = RandomArray[UniformDistribution[-50, 50], nn];

yvals = 2 - 3 * xvals + evals;
pairs = Transpose[{xvals, yvals}];
```

Sample summaries for the X and Y samples are as follows:

```
{mx, sdx} = {Mean[xvals], StandardDeviation[xvals]}
```

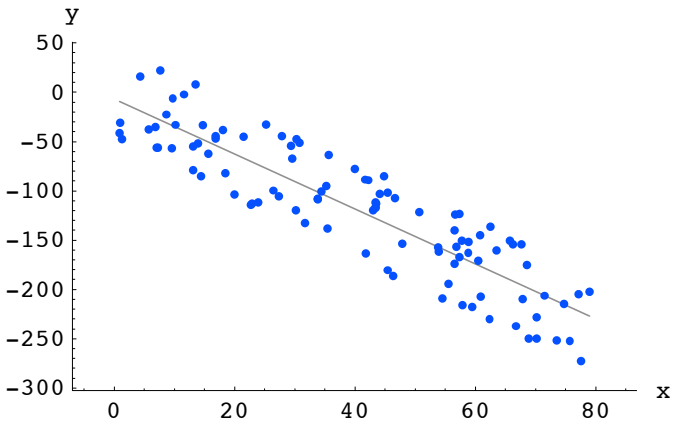
```
{39.9132, 22.2644}
```

```
{my, sdy} = {Mean[yvals], StandardDeviation[yvals]}
```

```
{-118.564, 69.4279}
```

A scatter plot of pairs and the least squares formula for the conditional mean are shown below:

```
ScatterPlot[pairs, Correlation → True]
```




```
Correlation:       -0.8931
```

```
Clear[x]; Remove[f]
f[x_] = Fit[pairs, {1, x}, x]
```

```
-7.40173 - 2.7851 x
```

The approximate 95% permutation confidence interval (shown below) contains –3.

```
SlopeCI[pairs, RandomPermutations → 2000]
```

```
{-3.06638, -2.49856}
```

(b) Model summaries:

```
model1 = UniformDistribution[0, 80];
model2 = UniformDistribution[-50, 50];
```

(1) The mean and standard deviation of X are as follows:

```
{μx, σx} = N[{Mean[model1], StandardDeviation[model1]}]
```

```
{40., 23.094}
```

(2) The mean and standard deviation of $Y = 2 - 3X + \epsilon$ are

$$E(Y) = 2 - 3E(X) + E(\epsilon) = -118 \text{ and } \mathrm{SD}(Y) = \sqrt{9\,\mathrm{Var}(X) + \mathrm{Var}(\epsilon)} \approx 75.06$$

as demonstrated below:

```
μy = 2 - 3 * Mean[model1] + Mean[model2]
```

```
-118
```

```
σy = N[Sqrt[9 * Variance[model1] + Variance[model2]]]
```

```
75.0555
```

(3) To compute the correlation, first note that

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, 2 - 3X + \epsilon) = -3\,\mathrm{Var}(X).$$

Thus, the correlation is as follows:

```
ρ = (-3 Variance[model1]) / (σx * σy)
```

```
-0.923077
```

(4) Comparison of estimates with model values:

```
{mx, sdx, my, sdy, Correlation[pairs]}
```

```
{39.9132, 22.2644, -118.564, 69.4279, -0.893136}
```

```
{μx, σx, μy, σy, ρ}
```

```
{40., 23.094, -118, 75.0555, -0.923077}
```

In each case, the estimate is close to the model summary.

Problem 2: Analysis of brain-body data.

species, **wbody**, **wbrain** are lists of length 43.

(a) The least squares linear fit formula, and approximate 95% confidence interval for the slope are shown below:

```
pairs = Transpose[{Log[10, wbody], Log[10, wbrain]}];
Clear[x]; Remove[f];
f[x_] = Fit[pairs, {1, x}, x]
```

```
0.930348 + 0.782263 x
```

```
SlopeCI[pairs, RandomPermutations → 2000]
```

```
{0.701258, 0.862863}
```

Since the 95% confidence interval does not contain zero, we know that there is a significant regression effect.

Since slope corresponds to the rate of change of *y* with respect to a unit change in *x* and we are working on the log scale, the interpretation is as follows: If average body weight increases by a factor of 10, then average brain weight will increase by a factor of about 6.06, as demonstrated below.

```
10^0.782263
```

```
6.05708
```

(b) The gorilla's predicted response is 2.74205 log-grams (552.136 grams).

```
lwb = Log[10, 207.0];
{f[lwb], 10^f[lwb]}
```

```
{2.74205, 552.136}
```

The actual response is 2.60853 log-grams (406 grams).

```
lwbr = Log[10, 406.0]
```

```
2.60853
```

The values are reasonably close on the log-scale (error is approximately 5.1% of actual log-value),

```
(f[lwb] - lwbr) / lwbr
```

```
0.0511859
```

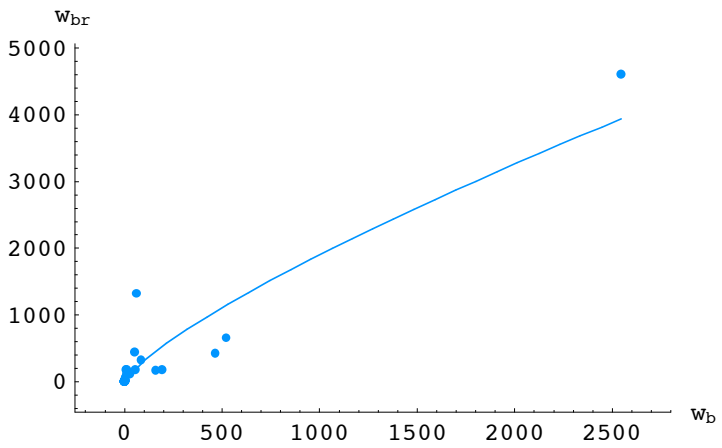
On the original scale, however, the values are not very close (the error is approximately 36% of the actual value).

(c) The definition of g and smoothed scatter plot are shown below.

```
Clear[x]; Remove[g];
g[x_] = Simplify[10^f[Log[10, x]]]
```

$$8.51821\, x^{0.782263}$$

```
pairs2 = Transpose[{wbody, wbrain}];
SmoothPlot[{pairs2, g},
 AxesLabel → {"wb", "wbr"}]
```



Brain weight increases as body weight increases, although the rate of increase decreases with body weight. Body and brain weights for the Asian elephant are very different from those of other species considered in this problem.

Problem 3: *X* is a uniform random variable on the interval [0,20].

(1) The correlation between X and X^2 is approximately 0.97, as demonstrated below:

- Mean and standard deviation of X :

```
model = UniformDistribution[0, 20];
{μ1, σ1} = {Mean[model], StandardDeviation[model]}
```

$$\left\{10, \frac{10}{\sqrt{3}}\right\}$$

- Mean and standard deviation of X^2:

```
μ2 = Integrate[x^2 / 20, {x, 0, 20}]
```

$$\frac{400}{3}$$

```
σ2 = Sqrt[Integrate[(x^2 - μ2)^2 / 20, {x, 0, 20}]]
```

$$\frac{160\sqrt{5}}{3}$$

- Covariance between X and X^2:

```
σ12 = Integrate[x^3/20, {x, 0, 20}] - μ1*μ2
```

$$\frac{2000}{3}$$

- Correlation between X and X^2:

```
N[σ12/(σ1*σ2)]
```

```
0.968246
```

(2) Since

$$\text{Cov}(X-10,(X-10)^2)=\text{Cov}(X-10,X^2-20X+100)=\text{Cov}(X,X^2)-20\,\text{Var}(X)=0$$

```
σ12 - 20 σ1^2
```

```
0
```

the correlation between X–10 and $(X$–$10)^2$ is also zero.

Problem 4: Analysis of the SWS cases data.

danger, **sws** are lists of length 43.

```
cases = Transpose[{Log[10, wbody], danger, Log[10, sws]}];
Clear[x1, x2]; Remove[f];
f[x1_, x2_] = Fit[cases, {1, x1, x2}, {x1, x2}]
```

```
1.06671 - 0.097165 x1 - 0.0539304 x2
```

(a) Regression analysis of the SWS cases list:

```
Clear[x1, x2];
results = Regress[cases, {1, x1, x2}, {x1, x2},
   RegressionReport → {ANOVATable, ParameterCITable,
     RSquared, EstimatedVariance, StandardizedResiduals,
     PredictedResponse, FitResiduals, PredictedResponseDelta}];
```

(1) Analysis of variance:

```
ANOVATable /. results
```

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	1.23174	0.615872	28.6667	1.8878×10^{-8}
Error	40	0.859355	0.0214839		
Total	42	2.0911			

Since the p value is virtually zero, the regression effects are not all equal to zero.

(2) The model explains approximately 58.9% of the variation in the data:

```
RSquared /. results
```

```
0.589042
```

(3) 95% confidence intervals for the β-parameters are as follows:

```
ParameterCITable /. results
```

	Estimate	SE	CI
1	1.06671	0.0500724	{0.965506, 1.16791}
x1	-0.097165	0.0181983	{-0.133945, -0.060385}
x2	-0.0539304	0.0173961	{-0.0890892, -0.0187715}

Note that each β coefficient is significantly different from zero.

(4) The estimated standard deviation of the error distribution is as follows:

```
Sqrt[EstimatedVariance] /. results
```

```
0.146574
```

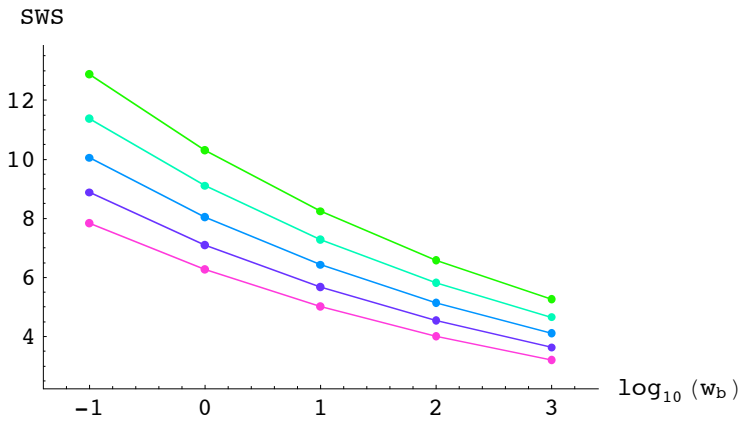
(5) An enhanced normal probability plot of standardized residuals (not displayed) indicates that normal theory methods are reasonable in this case.

(6) Comments: The analyses suggest that the amount of SWS sleep decreases as body weight increases and as the danger index increases and that each predictor contributes significantly to the model. Normal theory methods seem reasonable in this case.

(b) Smoothed plot of pairs:

The five lists of pairs are the five elements of **pairslist**.

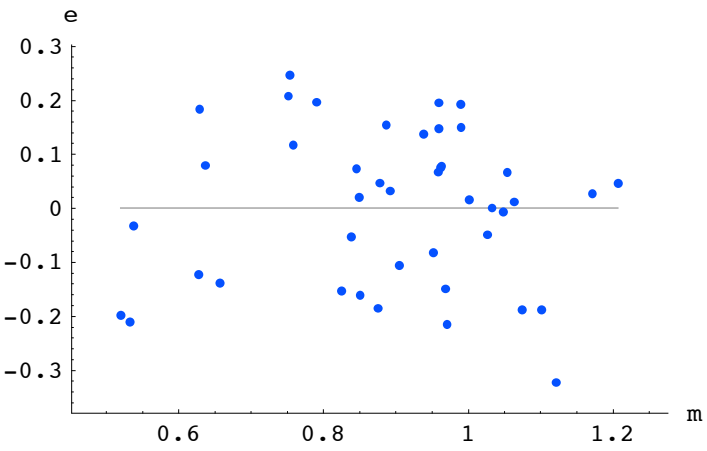
```
pairslist = Table[
   Table[{x, 10^f[x, j]}, {x, -1, 3}], {j, 1, 5}];
ScatterPlot[pairslist,
 PlotJoined → True,
 AxesLabel → {"log10 (wb)", "SWS"}]
```



The plot suggests that the amount of SWS sleep decreases with increasing body weight and with the danger index. Danger index 1 (green) corresponds to the curve with the largest values of SWS sleep; danger index 5 (light purple) corresponds to the curve with the smallest values of SWS sleep.

(c) Diagnostic plot of (estimated mean, estimated error) pairs:

```
residuals = FitResiduals /. results;
means = PredictedResponse /. results;
rpairs = Transpose[{means, residuals}];
ScatterPlot[rpairs,
 AxesLabel → {"m", "e"},
 Correlation → True]
```

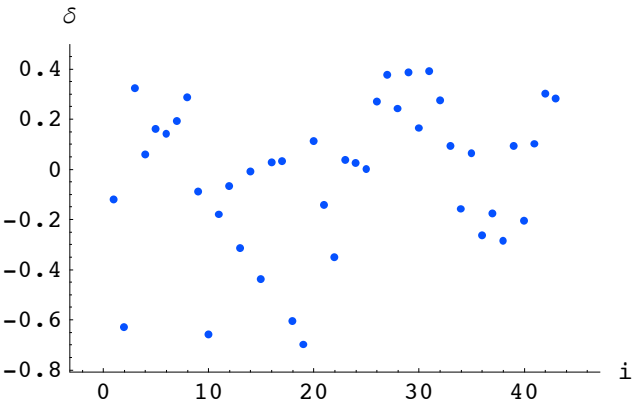


Correlation: 6.697×10^{-17}

There is no apparent relationship between residuals and predicted responses.

(d) Diagnostic plot of (case number, standardized influence) pairs:

```
nn = 43;
deltas = PredictedResponseDelta /. results;
dpairs = Transpose[{Range[1, nn], deltas}];
ScatterPlot[dpairs,
 AxesLabel → {"i", "δ"}]
```




```
p = 3; nn = 43;
cutoff = N[2 * Sqrt[p / nn]]
```

```
0.528271
```

```
Select[dpairs, (Abs[Last[#]] > cutoff) &]
```

```
{{2, -0.630684}, {10, -0.658695}, {18, -0.605661}, {19, -0.699289}}
```

```
species[[{2, 10, 18, 19}]]
```

```
{Asian elephant, E. Amer. mole, Horse, Short tail shrew}
```

```
residuals[[{2, 10, 18, 19}]]
```

```
{-0.197819, -0.32274, -0.210854, -0.188074}
```

Although 4 species have values outside the interval [–0.528, +0.528], none of the values are very far from the interval. In each case, the observed response was less than the predicted response.

Problem 5: Analysis of PS cases data.

ps, **tgestation** are lists of length 43.

(a) Initial analyses:

```
cases = Transpose[{Log[10, tgestation], danger, Log[10, ps]}];
Clear[x1, x2]; Remove[f]
f[x1_, x2_] = Fit[cases, {1, x1, x2}, {x1, x2}]
```

```
1.06246 - 0.300001 x1 - 0.108528 x2
```

Partial plots (not shown) indicate that adjusted log-PS values are negatively associated with adjusted x_i values. In each case, there was one point (the Asian elephant) with an unusually high y-coordinate.

(b) Regression analysis:

```
Clear[x1, x2];
results = Regress[cases, {1, x1, x2}, {x1, x2},
   RegressionReport → {ANOVATable, ParameterCITable,
     RSquared, EstimatedVariance, StandardizedResiduals,
     PredictedResponse, FitResiduals, PredictedResponseDelta}];
```

(1) Analysis of variance:

```
ANOVATable /. results
```

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	2.29653	1.14827	36.5731	9.29817×10^{-10}
Error	40	1.25586	0.0313965		
Total	42	3.55239			

Since the p value is virtually zero, the regression effects are not all equal to zero.

(2) The model explains approximately 64.6% of the variation in the data:

```
RSquared /. results
```

```
0.646475
```

(3) 95% confidence intervals for the β parameters are as follows:

```
ParameterCITable /. results
```

	Estimate	SE	CI
1	1.06246	0.118072	{0.823826, 1.30109}
x1	-0.300001	0.0632294	{-0.427792, -0.17221}
x2	-0.108528	0.0207053	{-0.150375, -0.0666814}

Each β coefficient is significantly different from zero.

(4) The estimated standard deviation of the error distribution is as follows:

```
Sqrt[EstimatedVariance] /. results
```

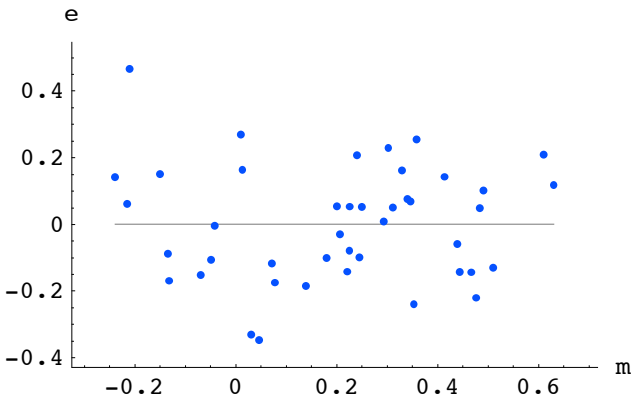
```
0.177191
```

(5) An enhanced normal probability plot (not shown) suggests that normal theory methods are reasonable in this case.

(6) Comments: The analyses suggest that the amount of PS sleep decreases as the average gestational time and danger indices increase and that each predictor contributes significantly to the model. Normal theory methods seem reasonable in this case.

(7) Diagnostic plot of (estimated mean, estimated error) pairs:

```
residuals = FitResiduals /. results;
means = PredictedResponse /. results;
rpairs = Transpose[{means, residuals}];
ScatterPlot[rpairs,
 AxesLabel → {"m", "e"},
 Correlation → True]
```




```
Correlation:        -1.677 × 10^-16
```

```
Last[Sort[Transpose[{residuals, Range[1, 43]}]]]
```

```
{0.465486, 2}
```

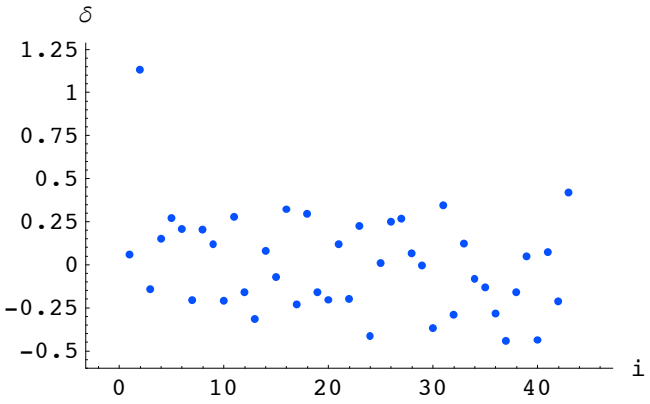
```
species[[2]]
```

```
Asian elephant
```

The plot shows no apparent relationship between estimated means and standardized residuals. Note, however, that there is one large outlier, corresponding to the Asian elephant.

(8) Diagnostic plot of (case number, standardized influence) pairs:

```
nn = 43;
deltas = PredictedResponseDelta /. results;
dpairs = Transpose[{Range[1, nn], deltas}];
ScatterPlot[dpairs,
 AxesLabel → {"i", "δ"}]
```




```
p = 3; nn = 43;
cutoff = N[2 * Sqrt[p / nn]]
```

```
0.528271
```

```
Select[dpairs, (Abs[Last[#]] > cutoff) &]
```

```
{{2, 1.13166}}
```

```
species[[2]]
```

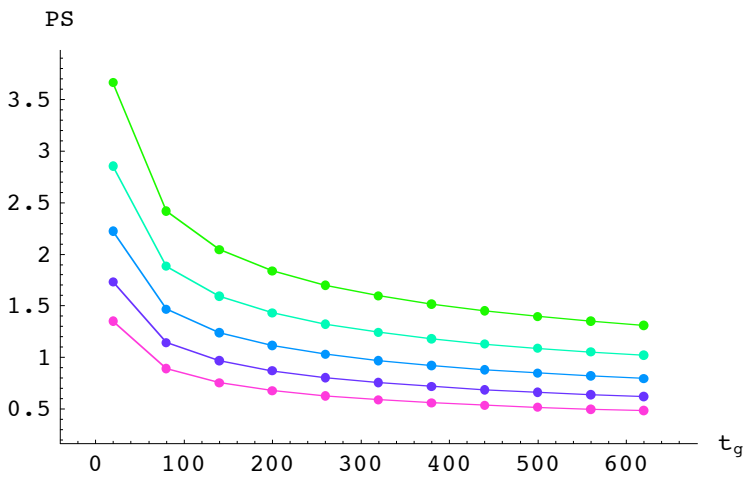
```
Asian elephant
```

The standardized influence for the Asian elephant (1.132) is unusually high.

(c) Plot of pairs lists:

The five lists of pairs are the five elements of **pairslist**.

```
pairslist = Table[
   Table[{x, 10^f[Log[10, x], j]}, {x, 20, 620, 60}], {j, 1, 5}];
ScatterPlot[pairslist,
 PlotJoined → True,
 AxesLabel → {"t_g", "PS"}]
```



The plot suggests that the amount of PS sleep decreases with increasing gestational time and with the danger index. Danger index 1 (green) corresponds to the curve with the largest values of PS sleep; danger index 5 (light purple) corresponds to the curve with the smallest values of PS sleep.