# Statistical Science.
## Part II. Quantifying Uncertainty and Evidence.
## Chapter 5   Data Equations

ReCap.        Part I (Chapters 1,2,3,4)
5.1  Introduction
5.2  Data
5.3  Deviations from a Single value model
5.4  Deviations from the Mean
5.5  Deviations from a Linear Trend
5.6  Comparison of Models – ANOVA and  ANODEV
5.7  Deviations from the Means of Two or More Groups
5.8  Review questions.

Red chalk for residuals
Yellow chalk for model
White chalk for data

on chalk board

**Recap** Part I
Ch1 Quantitative reasoning: Example of scallops, which combined stats and models
Ch2 Quantities: Five part definition
Ch3 Rescaling
Ch4 Equations express an idea or concept about the relation of one quantity to another

Today-- Data Equations
 Defined, then an example to demonstrate the idea
   and its application in statistical analysis.

Wrap-up.  Data equations summarize pattern in data.
        Data equations  have 3 parts: the data, the model, and the residuals.
        The sum of the residuals is a measure of bias in fit.
        The sum of the squared residuals
                measures the goodness of fit.
                allows us to quantify the improvement in fit.
        The likelihood ratio is calculated from the sum of squared residuals
                It measures which of two models is more likely than the other.
        Data equations apply to regression lines (ratio scale explanatory variable)
                and to contrasting means among groups (nominal scale explanatory variable)
        We will use likelihood ratios to make statistical tests of  regression lines and
                contrasting means  (t-tests, F-tests, chisquare tests, *etc*).

## 5.1 Introduction

Statistics are often presented in the following sequence

Here are some statistics and how to compute them (means, variances)

Here is a little bit of probability theory (How many blue marbles do you expect to draw in 5 tries, from an urn with 5000 blue marbles and 5000 red marbles ?)

Here are some more probability distributions--normal, $X^2$, etc.

Here is how to do a t-test to compare two means.

Here is how to do a regression, ANOVA, *etc.*

In this course we will take a different approach, one that will allow you to set up an appropriate analysis of data, without having to search for the 'right' test. This approach is readily learned. It allows you to accomplish far more than is possible by learning a series of tests. This model based approach begins with the concept of a set of data equations. We will use data equations to compare models to data.

$$\text{Data} = \text{Model} + \text{Residual}$$

This allows us to set up simple symbolic expressions to undertake statistical analysis of our data. In order to use these simple symbolic expressions with confidence, we will present them with verbal and graphical expressions of the same idea.
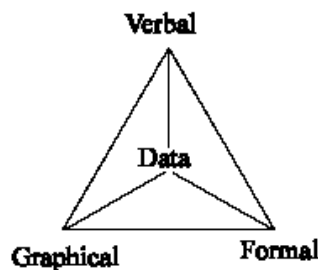


We will use data equations to measure how well a model fits the data. We will use data equations to compare competing models. Later, we will use data equations to calculate a measure of evidence, measures of uncertainty, and the error rate if we assume the model to be true (hypothesis testing).

The goal of the equations lab (Lab 2) is to gain experience in working with models that relate one variable to another. The models in this lab expressed:

-metabolic rate as an allometric function of body mass (Kleiber's Law)

-hormone levels in mothers as a function of time before birth

-femur length as allometric function of body length

-survival rate of snails that host human parasites as a function of temperature

These equations provide expected values of one quantity, as a function of some other quantity. The model values will not match observed values perfectly. But the match should be close enough to be convincing. Data equations allow us to quantify the fit of data to models that express important concepts in biology.

**Definition**: A data equation has three parts: observed values *Y*, expected values *E(H)*, and the residual $\varepsilon$.

$$\text{Data} = \text{Model} + \text{Residual}$$

$$\text{Data} = \text{Expected} + \text{Residual}$$
$$Y = E(H) + \varepsilon$$

$$\text{Data} = \text{Fitted values} + \text{Residual}$$
$$Y = \hat{Y} + \varepsilon$$

The data are a set of measured values of a quantity of scientific interest.

Expected values are calculated from a functional expression that expresses science-based relation of variables. Lab 2 (Equations) shows examples – hormone levels in tamarin mothers, insect allometry, and parasite demography.

The residuals are the difference between the data and fitted value.

$$\text{Residuals} = \text{Data} - \text{Fitted}$$

```
In following examples, keep the data, the graphics,
and the stack of data equations coordinated.

In stack of Data Equations:
Keep data on board (white chalk)
Erase and replace the model (yellow)
Erase and replace residuals (in red)

Graphs:
Keep axes and data points on board (white)
Erase and replace model (yellow lines)
Erase and replace residuals (vertical lines in red)
```

Here is an example, based on the expected value from an equation: metabolic rate of a non-passerine bird as a function of body mass (Lasiewski and Dawson 1967).
$$\dot{E} = 78.3 \, M^{0.723}$$

$$\frac{kcal}{day} = 78.3 kg^{0.723}$$

Here is an observed value for a single ostrich on a single day.
     2.7 m tall.   Weight = 115 kg.  Metabolic rate $E$ observed = 2300 kcal/day
Here is the expected value, as estimated from the function. .

$$\hat{\dot{E}} = 78.3 \, (115 \, kg)^{0.783} = 2420 \text{ kcal/day}$$

The data equation is $\dot{E}_{observed} = \hat{\dot{E}} + error$

$$2300 = 2420 - 120$$

## 5.2  Example:  Fruit Fly Heterozygosity.  Data from Brussard (1984)

To become familiar with data equations we will use data collected by Theodosius Dobzhansky, who was one of the pioneers of populations genetics.  Dobzhansky established the fact that natural populations had unexpectedly high levels of genetic variation.  Dobzhansky devoted his professional life to one of the central questions in population biology: the origin and maintenance of this genetic variability.  Dobzhansky worked on fruit flies in laboratory populations, but during the summer he would leave the lab and study wild populations of fruit flies. With colleagues he investigated the ecological processes that generate or erode genetic variability in nature. Harsher environments at higher altitudes are expected to select for a narrower range of phenotypes, hence reduce genetic variability.  The research question was:

> Does genetic variability decrease at higher altitude, as expected from stronger selection in extreme environments ?

To address this question Dobzhansky collected flies at different altitudes in Yosemite Park in the Sierra Nevada range, a spectacularly scenic location that offered pleasant relief from the hot summer conditions of urban Los Angeles. Dobzhansky used the best technique at the time, called inversion heterozygosity, to measure genetic variability. Here is Dobzhansky's data on inversion heterozygosity (assuming Hardy Weinberg equilibrium) of 3rd chromosome inversions from the fruit fly *Drosophila persimilis*.

The data were reported by P.F. Brussard 1984.  Geographic patterns and environmental gradients: The central-marginal model in Drosophila revisited.
*Annual Review of Ecology and Systematics* 15: 25-64.

```
   850   0.59   0.26
  3000   0.37   0.91
  4600   0.41   1.40
  6200   0.40   1.89
  8000   0.31   2.44
  8600   0.18   2.62
 10000   0.20   3.05


 Elev     H(%)   Elev(km)
```

H = heterozygosity.
Elev = elevation in feet, converted to km

Under Hardy-Weinberg equilibrium, heterozygosity (2pq) reaches a maximum at an inversion frequency of p = 50%, where 2pq = 2(0.5)(1-0.5). There is clearly a reduction in heterozygosity in this data, reduction in inversions with altitude, and hence increase in a predominant form.  With this data we can address two questions.

> How do these data compare to prior estimates of heterozygosity?
> How good is the evidence for a decrease in heterozygosity with elevation?

4

## 5.3 Deviations from a Single Value Model.    Model: $E(H) = \beta_{ext}$

In this case we do not have a theory based value of heterozygosity. If we did, we could use what R.A. Fisher called "Bayesian inference" in which a prior probability known from theory is updated by evidence.

While we do not have a theoretical value to test, we can compare the data to estimates of heterozygosity in this species. Lewontin and Hubby (1966) showed that heterozygosity over many loci ran at an average of 30% in *D. persimilis*.

Our model is $E(H) = \beta_{ext}$ where $\beta_{ext}$ has a single value, 30%.

$$\text{Data} = \text{Model} + \text{Residual}$$
$$H = E(H) + \varepsilon$$
$$H = \beta_{ext} + \varepsilon$$
$$H = 0.3 + \varepsilon$$

Using this simple model we form 7 data equations, one for each observed value.

| Data | = | Model | residual | | res$^2$ |
|------|---|-------|------|---|------|
| H | = | $\hat{H}$ | res | | res$^2$ |
| 0.59 | = | 0.3 | 0.29 | | 0.0841 |
| 0.37 | = | 0.3 | 0.07 | | 0.0049 |
| 0.41 | = | 0.3 | 0.11 | | 0.0121 |
| 0.40 | = | 0.3 | 0.10 | | 0.01 |
| 0.31 | = | 0.3 | 0.01 | | 0.0001 |
| 0.18 | = | 0.3 | -0.12 | | 0.0144 |
| 0.20 | = | 0.3 | -0.10 | | 0.01 |
| | | | 0.36 | | 0.13560 |

$\Sigma\text{res} = \underline{\ 0.36\ }$

$\Sigma\text{res}^2 = \underline{\ 0.1356\ }$

$L(\beta_{ext} \mid H) = 0.1356^{-n/2}$

$n = 7$

The values tend to be more than 0.30 and hence the residuals sum to a positive value.

The summed residuals measure the deviation from fit to 30% heterozygosity.
The sum of the squared residuals measures the goodness of fit.
$L(\beta_{ext} \mid H)$ is the likelihood of the parameter $\beta_{ext}$ given the data $H$. We will use it to compare the weight of evidence for one model versus another.

## 5.4 Deviations from the Mean. Model: $E(H) = \overline{H}$ estimated from data

Often we have no prior parameter value, as in Fisher's "Bayesian" inference or in Lewontin and Hubby's compilation of estimates.
In such cases we estimate the parameter from the data we have.
We use "frequentist" inference to make the "best" estimate.

What does "best" mean ? There are several criteria. We will start with two statistical criteria: that the residuals sum to zero (the estimate is unbiased) and that the residuals be as small as possible. The number that meets both these criteria is the mean value.

$$\text{Mean}(H) = \overline{H} = n^{-1} \Sigma H = 7^{-1} \cdot 2.46 = 0.3514$$
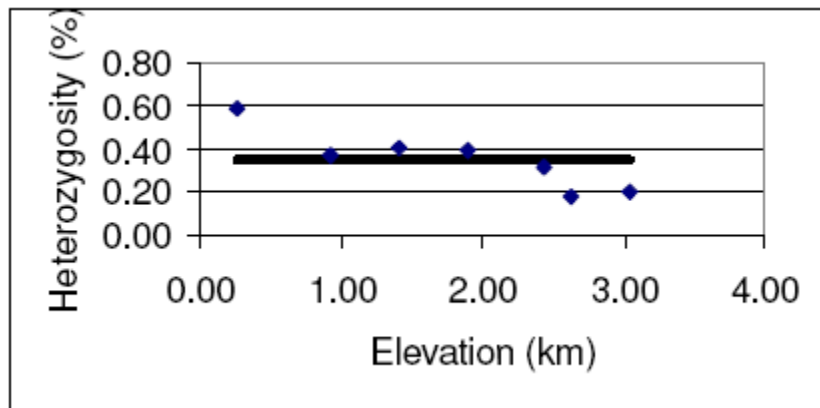


Add the model value 0.3514 to the graph as a horizontal line.

Run a perpendicular from the line to each point. These are the residuals.

White chalk for data, yellow chalk for model = straight line at 0.351

Our model is $E(H) = \overline{H}$ where $\overline{H}$ has a single value, 35.14%.

$$\text{Data} = \text{Model} + \text{Residual}$$
$$H = E(H) + \varepsilon$$
$$H = \beta_o + \varepsilon$$
$$H = 0.3514 + \varepsilon$$

Here are the data equations when the model value is the mean

| Data = | Model | + | Residual | |
|---|---|---|---|---|
| $H$ = | | + | res | res$^2$ |
| 0.59 = | 0.3514 | + | 0.2386 | 0.0569 |
| 0.37 = | 0.3514 | + | 0.0186 | 0.0003 |
| 0.41 = | 0.3514 | + | 0.0586 | 0.0034 |
| 0.40 = | 0.3514 | + | 0.0486 | 0.0024 |
| 0.31 = | 0.3514 | − | 0.0414 | 0.0017 |
| 0.18 = | 0.3514 | − | 0.1714 | 0.0294 |

$\Sigma\text{res} = \underline{\ 0\ }$

$\Sigma\text{res}^2 = \underline{\ 0.1171\ }$

$L(\beta_o \mid H) = 0.1171^{-n/2}$

$n = 7$

The residuals sum to zero, meeting the criterion of no bias. It can be shown mathematically that the squared deviations sum to the smallest possible value.

This model is an improvement over our previous model ($H = 0.30 + \text{residual}$) because it is unbiased and has the minimum deviation of the data

Keep $\Sigma\text{res}^2$ on board for each successive model

6

from the model. It is not necessarily an improvement from the point of view of the underlying biology. We have substituted a statistical criterion (best fit) for a biological criterion (prior knowledge).

## 5.5 Deviations from a Linear Trend: $\hat{H} = \hat{\beta}_E E + \hat{\beta}_I$ estimated from large scale data.

Our single value model of the data fails to capture the trend in heterozygosity relative to elevation. This shows up in the residuals when listed from low to high elevation: positive residuals early in the series, negative residuals later in the series. So we investigate a model that describes a decrease in heterozygosity with increase in elevation. We start with the simplest possible model, a straight line increase with altitude.

$$H = \hat{H} + \varepsilon$$
$$H = \hat{\beta}_E E + \hat{\beta}_I + \varepsilon$$

The parameter $\beta_E$ in this equation (model) is the heterozygosity gradient. It is the change in heterozygosity for each unit change in elevation. The gradient $\beta_E$ has units of % /km . The product $\beta_E \cdot E$ has units of (%/km)(km) = %. Thus it has the same units as the response variable $H$, as it must for the equation to be dimensionally homogeneous. $\beta_I$ is the $Y$ intercept.

We will start with the large scale heterozygosity gradient $\beta_E$. The large scale gradient is the difference in heterozyzosity from the lowest to highest elevation.

$$\hat{\beta}_E = \frac{(20 - 59) \%}{(3.05 - 0.26) \text{ km}} = -13.98 \%/ \text{ km}$$

$\hat{\beta}_E$ is an estimate of the slope of the line in the plot of heterozygosity versus elevation. The estimate is $-13.98\%$ per km increase in elevation. The hat over the symbol signifies that this is an estimate of the parameter.

Next we calculate an offset so that the line runs through the value at 0.26 km.

$$\hat{H} \quad = \hat{\beta}_E \cdot E \qquad\qquad\qquad + \text{offset}$$
$$0.59 \quad = (-13.98 \%/ \text{ km})(0.26 \text{ km}) \quad + \text{offset}$$

$$\text{offset} = 0.59 - (-13.98 \%/ \text{ km})(0.26) = 0.6262$$

$$\hat{H} = -0.1398 \cdot E + 0.626$$

Here is the gradient model, with estimates from the highest and lowest elevation.

$$H = \hat{H} + \varepsilon$$
$$H = \beta_E \cdot E + \text{offset} + \text{res}$$
$$H = -0.1398 \cdot E + 0.626 + \text{res}$$

## 5.4  Deviations from a Linear Trend: $\hat{H} = \hat{\beta}_E E + \hat{\beta}_I$ estimated from data

With this model we can calculate an expected value for each observation of heterozygosity, based on the elevation.

| Data | = | Model | + | Residual | |
|---|---|---|---|---|---|
| $H$ | = | $\hat{\beta}_E \cdot E + \textit{Offset}$ | + | res | Res$^2$ |
| 0.59 | = | $-0.1398 \cdot 0.26 + 0.63$ | + | 0.0000 | 0.000000 |
| 0.37 | = | $-0.1398 \cdot 0.91 + 0.63$ | + | $-0.1284$ | 0.016476 |
| 0.41 | = | $-0.1398 \cdot 1.40 + 0.63$ | + | $-0.0202$ | 0.000407 |
| 0.40 | = | $-0.1398 \cdot 1.89 + 0.63$ | + | $+0.0380$ | 0.001446 |
| 0.31 | = | $-0.1398 \cdot 2.44 + 0.63$ | + | $+0.0248$ | 0.000613 |
| 0.18 | = | $-0.1398 \cdot 2.62 + 0.63$ | + | $-0.0797$ | 0.006348 |
| 0.20 | = | $-0.1398 \cdot 3.05 + 0.63$ | + | 0.0000 | 0.000000 |

Note bias:  $\Sigma \text{res} = \underline{-0.17}$  $\Sigma \text{res}^2 = \underline{0.0253}$

$$L(\beta_e \,|H) = 0.0253^{-7/2}$$

The model exactly predicts the heterozygosity at the highest and lowest elevation. We examine the deviations at each elevation to determine whether this model is acceptable.

    high and low residuals are interspersed.

    there are about as many negative as positive residuals

    the sum of the residuals is somewhat biased (negative):  $\Sigma \text{res} = -0.17$

The model is satisfactory in that there are no patterns in the residuals. The model is somewhat biased, leading to an estimate of the heterozygosity gradient that is more negative than an unbiased estimate.

A better estimate of the gradient $\beta_E$ can be obtained from the formula for the least squares estimate.

$$\hat{\beta}_E = \frac{\Sigma \, (H - \overline{H})(E - \overline{E})}{\Sigma \, (E - \overline{E})^2}$$

We have already met the quantity $H - \overline{H}$. It is the deviation of the observed value from the average value of heterozygosity.

Similarly, the quantity $(E - \overline{E})$ represents the deviations from the average elevation $\overline{E}$.

And $(E - \overline{E})^2$ represents the squared deviations from average elevation.

## 5.4 Deviations from a Linear Trend.

The formula produces an estimate of $\beta_E$ that minimizes $(H - \bar{H})^2$

With this formula, the estimate of the heterozygosity gradient comes to $\hat{\beta}_E = -12.73$ %/km, a value that is not quite as negative (as steep) as our crude estimate of $-13.98$%/km.

As before we compute an offset. This offset runs through the mean values (0.3514,1.796), which we use as our reference point because we always have more information about this point than we do about the $Y$-intercept, where $E = 0$).

$$\bar{H} = \hat{\beta}_E \cdot \bar{E} \qquad + \text{offset}$$
$$0.3514 = (-12.73 \text{ %/ km})(1.796) \qquad + \text{offset}$$

$$\text{offset} = 0.3514 - (-12.73 \text{ %/ km})(1.796) = 0.58$$

$$\hat{H} = -0.1273 \cdot E + 0.58$$

Here is the gradient model, with unbiased estimates from the data.
$$H = -0.1273 \cdot E + 0.58 + \text{res}$$

The data equations based on this new estimate:

| Data | = | Model | + | Residual | |
|---|---|---|---|---|---|
| $H$ | = | $\beta \cdot E + \text{Offset}$ | + | res | res$^2$ |
| 0.59 | = | $-0.1273 \cdot 0.26 + 0.58$ | + | +0.0429 | 0.001842 |
| 0.37 | = | $-0.1273 \cdot 0.91 + 0.58$ | + | −0.0937 | 0.008773 |
| 0.41 | = | $-0.1273 \cdot 1.40 + 0.58$ | + | +0.0084 | 0.000071 |
| 0.40 | = | $-0.1273 \cdot 1.89 + 0.58$ | + | +0.0605 | 0.003659 |
| 0.31 | = | $-0.1273 \cdot 2.44 + 0.58$ | + | +0.0403 | 0.001626 |
| 0.18 | = | $-0.1273 \cdot 2.62 + 0.58$ | + | −0.0664 | 0.004408 |
| 0.20 | = | $-0.1273 \cdot 3.05 + 0.58$ | + | +0.0079 | 0.000063 |

$\Sigma$res = 0.0000    No bias

$\Sigma$res$^2$ = 0.0204    Much less than 0.1171, the value of the previous no bias model

$L(\beta_E \,|H) = 0.0204^{-7/2}$ More likely than previous no bias model based on $\beta_o$ because a smaller fraction raised to the same power.

We examine the residuals to evaluate this model.
 -residuals are all approximately the same magnitude
 -about the same number of positive and negative residuals
 -no pattern of low residuals followed by high residuals
 -sum of residuals is zero

The fact that the residuals sum to zero is no accident, The least squares estimation method guarantees this. The estimation also produces the smallest possible value of the sum of squared residuals $\Sigma$res$^2$.

9

## 5.5   Comparison of Models - Analysis of Variance and Analysis of Deviance

We have considered several models of the data. We use three measures to compare them.

<u>First:</u>  The sum of the residuals.  This measures bias.
We looked at several models. We find that some have biased estimates, two do not.

$\beta$ from previous information.                                        $\Sigma res = \underline{-0.34}$   -0.36
$\hat{\beta}_o = mean(H)$ estimated from data (least squares)             $\Sigma res = \underline{0.00}$
$\hat{\beta}_E$ estimated from data (using two data points)              $\Sigma res = \underline{-0.17}$
$\hat{\beta}_E$ estimated from data (least squares)                      $\Sigma res = \underline{0.00}$

Those having unbiased estimates are preferable.

<u>Second:</u> The sum of the squared residuals.  This measures goodness of fit.

$H_o:$  $H = \hat{\beta}_o + \epsilon$                          $SS_{tot}$ $\Sigma res^2 = \underline{0.1171}$
$H_A:$  $H = \hat{\beta}_E \cdot E + Offset + \epsilon$                 $\Sigma res^2 = \underline{0.0204}$

The reduction in squared deviation is:                 $SS_{model}$ $\Sigma res^2 = \underline{0.0966}$

We conclude that the model that includes the gradient in heterozygosity $\beta_E$ fits the data better than the model without the gradient.

This comparison is displayed in either of two forms: the analysis of variance (ANOVA) table or the analysis of deviance (ANODEV) table.

```
            Df   Sum of Sq
  ElevKm    1  0.09648707
Residuals   5  0.02059864
  Total     6  0.11708571
```
Analysis of Variance (ANOVA)
format

```
           Df   Deviance Resid. Df Resid. Dev
  NULL                            6   0.1170857
ElevKm   1       0.09648707       5   0.0205986
```
Analysis of Deviance (ANODEV) format

<u>Third:</u>  Likelihood.     Which model is more likely?
The evidential support for the full (null) model of no gradient is $SS_{tot} = 0.1174^{-7/2}$

This is called the full model because it encompasses all of the variation in the data, relative to the simplest model, that of single value, in this case the mean.

The support for the reduced (alternative) model of a gradient is $SS_{tot} = 0.0204^{-7/2}$
This is called the reduced model because it encompasses the information in the data, after reducing the variance by including an explanatory variable, in this case a gradient.
The likelihood of the reduced relative to the full model is  $LR = (0.0204/0.1174)^{-7/2} = 453$

The gradient model is 450 times more likely than the no gradient model.

## 5.5   Comparison of Models - Analysis of Variance and Analysis of Deviance

   The evidence is stronger for the gradient model than the no-gradient model.
   We have a measure of strength of evidence.  $\text{Log}_{10}(453) = 2.66$ hartleys
      The evidence is better than 100:1 odds ($\log_{10}(100) = 2$
      The evidence is not as strong as 1000:1 odds ($\log_{10}(1000) = 3$
   Later we will use the *LR* to develop a statistical test of whether the no-gradient model
   could have arisen by chance.


Statistical results report degrees of freedom (df).What is a degree of freedom ?
We started with 7 observations.  We have 7 degrees of freedom.
      If we know 6 values, the 7th is still free to vary.
We lose 1 degree of freedom when we estimate the mean from the data.
      If we know 6 values and the mean, the 7th is not free to vary. We only have 6 df
      Hence the total sum of squares (ANOVA) and null model (ANODEV) has 6 df.
We lose another degree of freedom when we estimate the slope parameter.
      We assign this degree of freedom to the regression term in the model
      Hence the residual term has $7-1-1 = 5$ df

We will use the ANOVA table for the general linear model, which assumes normal
errors.  We use the ANODEV table for the generalized linear model (any error structure)

Lewontin, R. C.; Hubby, J. L. (1966). "A molecular approach to the study of genic
heterozygosity in natural populations. II. Amount of variation and degree of
heterozygosity in natural populations of Drosophila pseudoobscura". Genetics. 54 (2):
595–609.


   \* \* \* \*

Next, we will look at the data equations for another common situation in biology, where
we wish to compare mean values in two or more groups.

**5.7    Deviations from the Means of Two or More Groups    $Y = \beta_0 + \beta_x \cdot X$**
**Oat Yield data from Steel and Torrie (1960)**

To illustrate the idea of data equations for two or more groups we will use data from
Steel and Torrie 1960 (p237) who reported yield
(bushels/acre) of oats (Vicland variety, infected with *H. victoriae*) with or without a chemical seed treatment
(Panogen).

$Y$ = Yield of oats (bushels / acre)
$X$ = group (−1 = control, +1 = treated)

Null Model. $\hat{Y} = \hat{\beta}_o$ estimated from data

```
−1  42.9   untreated
−1  41.6   untreated
−1  28.9   untreated
−1  30.8   untreated
+1  49.5   treated
+1  53.8   treated
+1  40.7   treated
+1  39.4   treated

Grp   Y
```

As is often the case we do not have a parameter value
from prior evidence. We estimate the parameter from the data we have.  The 'best'
estimate is the overall mean, which is unbiased (residuals sum to zero) with smallest
deviation (smallest sum of squared residuals).

$$\text{Mean}(Y) = \overline{Y} = n^{-1} \Sigma Y = 8^{-1} \cdot 327.6 = 40.95$$
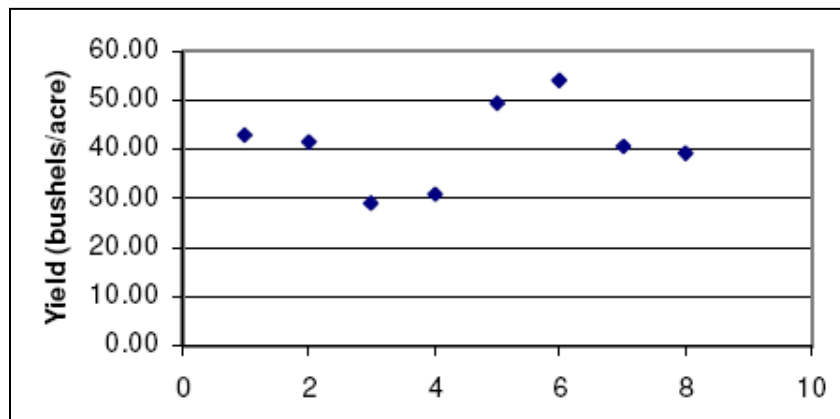


Draw in, as
horizontal line, the
model value of 40.95.

Connect each dot   to
the line with a
perpendicular.   These
are the residuals.

White chalk for data,
yellow chalk for model:
horizontal line at 40.95

Null Model (continued)    $\hat{Y} = \hat{\beta}_o$ from data

Our model is $\hat{Y} = \hat{\beta}_o$ where $\hat{\beta}_o$ is an estimate, 40.95 bushels / acre

$$\begin{array}{ccc} \text{Data} & = & \text{Model} + \text{Residual} \\ Y & = & Y \quad + \epsilon \\ Y & = & \hat{\beta}_o \quad + \epsilon \\ Y & = & 40.95 \quad + \epsilon \end{array}$$

Here are the data equations when the model value is the mean.

| Data | = | Model | + | Residual | | |
|------|---|-------|---|------|------|---|
| $Y$ | = | $\overline{Y}$ | + | res | res$^2$ | |
| 42.90 | = | 40.95 | + | 1.95 | 3.8025 | |
| 41.60 | = | 40.95 | + | 0.65 | 0.4225 | |
| 28.90 | = | 40.95 | + | -12.05 | 145.2025 | |
| 30.80 | = | 40.95 | + | -10.15 | 103.0225 | |
| 49.50 | = | 40.95 | + | 8.55 | 73.1025 | |
| 53.80 | = | 40.95 | + | 12.85 | 165.1225 | |
| 40.70 | = | 40.95 | + | -0.25 | 0.0625 | |
| 39.40 | = | 40.95 | + | -1.55 | 2.4025 | |

$\Sigma\text{res} = \underline{\ 0\ }$
$\Sigma\text{res}^2 = \underline{\ 493.14\ }$

$L(\hat{\beta}_o|Y) = (493.14)^{-8/2}$

Keep $\Sigma\text{res}^2$ on board for each successive model

Alternative Model   $Y = \beta_0 + \beta_x \cdot X$
          where $X$ = two groups

We have no prior information to gove us an expected value for treated and untreated oats, so we use means estimated from the data in each group.

For untreated oats the mean value is    $\hat{\beta}_o + \hat{\beta}_x(-1) = 40.95 - 4.90 = 36.05,$
For treated oats  the mean value is    $\hat{\beta}_o + \hat{\beta}_x(+1) = 40.95 + 4.90 = 45.85$

| Data = | | Model | | | | | + | Residual |
|--------|------|-----|------|------|---|-----|---|------|
| $Y$ = | $\hat{\beta}_o$ | + | $\hat{\beta}_x$ | $\cdot$ | $X$ | + | res | |
| 42.90 = | 40.95 | + | 4.90 | $\cdot$ | $-1$ | + | +6.85 | |
| 41.60 = | 40.95 | + | 4.90 | $\cdot$ | $-1$ | + | +5.55 | |
| 28.90 = | 40.95 | + | 4.90 | $\cdot$ | $-1$ | + | -7.15 | |
| 30.80 = | 40.95 | + | 4.90 | $\cdot$ | $-1$ | + | -5.25 | |
| 49.50 = | 40.95 | + | 4.90 | $\cdot$ | $+1$ | + | +3.65 | |
| 53.80 = | 40.95 | + | 4.90 | $\cdot$ | $+1$ | + | +7.95 | |
| 40.70 = | 40.95 | + | 4.90 | $\cdot$ | $+1$ | + | -5.15 | |
| 39.40 = | 40.95 | + | 4.90 | $\cdot$ | $+1$ | + | -6.45 | |

$\Sigma\text{res} = \underline{\ 0\ }$
$\Sigma\text{res}^2 = \underline{\ 301.06\ }$

$L(\hat{\beta}_x|Y) = (301.06)^{-8/2}$

**Data Equations.**   $Y = \beta_0 + \beta_x \cdot X$   two groups

The data equations are written using the overall mean $\hat{\beta}_o = 40.95$ bushels/acre, together with the group means expressed as deviations $\hat{\beta}_x$ from the grand mean.

Comparing models

How does our two group model  (treated versus control) compare to the  single value  (no effect) model?

To draw a conclusion about the effects of panogen on oat yields, we compare the model with the group means ($Y = \beta_o + \beta_x X$) to the model without the group means ($Y = \beta_o$).

That is, we compare the fit of the model having the explanatory term ($X$ = presence or absence of panogen) to the simpler model, in which the explanatory term $X$ is absent. The no effect model -- the one with no explanatory term --  is called the null model.  We have already met the null model $H_o$.  It is the model based on the grand mean $\beta_o$, for which the estimate was $\hat{\beta}_o = 40.95$ bushels/acre.  The alternative model $H_A$ includes the explanatory variable $X$, as shown in the above data equations.

We apply three criteria.

First:  Sum of the residuals.   $\Sigma$res   $= 0$ for both models.  Both are unbiased.

Second:   Sum of the squared residuals.   This measures goodness of fit.

| | | | |
|---|---|---|---|
| $H_o$: | $Y = \beta_o$ | $\Sigma\text{res}^2 = \underline{493.14}$ | $= \text{SStotal}$ |
| $H_A$: | $Y = \beta_o + \beta_x X$ | $\Sigma\text{res}^2 = \underline{301.06}$ | |

$H_A$ is a better fit than the null model, $H_o$ .
The improvement in fit (reduction in squared deviation) is:
      $\Sigma\text{res}^2 = 493.14 - 301.06 = 192.08$     $= \text{SSmodel}$
The improvement in fit as a percentage is $192.08/493.14 = 32\%$

Third:  Likelihood.   Which model is more likely, given the data ?

The likelihood ratio is $LR = (301.06/493.14)^{(-8/2)} = 7.2$
The model with treatment effects is only 7 times more likely than the no effect model. The weight of evidence ($LR < 20$) is not in favor of the treatment effect model.

Could this improvement in fit and in likelihood have been due to chance variation in measurements ?
Later, we will use the likelihood to develop a test for deciding whether the improvement in fit and is attributable to chance.


See worksheet    ( ST237.xls)

## 5.8    Review Questions

1.    Here are the 7 values of fly heterozygosity assigned randomly to the 7 values of altitude .

Calculate the sum of the residuals and the sum of the squared residuals for the null and alternative model.

How does the random improvement in fit compare to that for the unrandomized data?

Calculate the likelihood ratio for the gradient model compared to the no gradient model (no change with increasing altitude) for the random data.

| ft | km | Random sample of H |
|---|---|---|
| 850 | 0.25908 | 0.37 |
| 3000 | 0.91440 | 0.59 |
| 4600 | 1.40208 | 0.18 |
| 6200 | 1.88976 | 0.20 |
| 8000 | 2.43840 | 0.40 |
| 8600 | 2.62128 | 0.31 |
| 10000 | 3.04800 | 0.41 |

How does the likelihood ratio for the random data compare to that before randomization?  (take LR for unrandomized compared to randomized).

2.  Here are the heterozygosity values for *Drosophila pseudoobscura*, measured at the same locations as *D. persimilis*.

Calculate the sum of the residuals and the sum of the squared residuals  for the null and alternative model.

Calculate the likelihood ratio for the gradient model compared to the no gradient model (no change with increasing altitude)

How does the likelihood ratio for this species compare to that for *D. persimilis*?  (take LR for *D. persimilis* compared to *D. pseudoobscura*).

| ft | km | H |
|---|---|---|
| 850 | 0.25908 | 0.70 |
| 3000 | 0.91440 | 0.69 |
| 4600 | 1.40208 | 0.71 |
| 6200 | 1.88976 | 0.70 |
| 8000 | 2.43840 | 0.70 |
| 8600 | 2.62128 | 0.62 |
| 10000 | 3.04800 | 0.68 |

```
        *** Linear Model ***

Call: lm(formula = H ~ Alt.km., data = Brussard, subset = 1:7, na.action =
      na.exclude)
Residuals:
        1        2        3        4        5        6        7
 0.04292 -0.09367 0.008411 0.06049 0.04032 -0.0664 0.007921

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  0.5801    0.0529   10.9711   0.0001
   Alt.km.  -0.1273    0.0262   -4.8619   0.0046

Residual standard error: 0.06394 on 5 degrees of freedom
Multiple R-Squared: 0.8254
F-statistic: 23.64 on 1 and 5 degrees of freedom, the p-value is 0.004625

Analysis of Variance Table
Response: H
Terms added sequentially (first to last)
         Df  Sum of Sq    Mean Sq  F Value        Pr(F)
  Alt.km.  1 0.09664358 0.09664358 23.63833 0.004625356
Residuals  5 0.02044213 0.00408843


        *** Generalized Linear Model ***

Call: glm(formula = H ~ Alt.km., family = gaussian, data = Brussard, subset = 1:7, na.action
= na.exclude, control = list(epsilon = 0.0001, maxit = 50,
trace = F))
Deviance Residuals:
          1          2          3          4          5          6
 0.04291757 -0.09366627 0.008410879 0.06048802 0.04032481 -0.06639626


          7
 0.007921242

Coefficients:
               Value Std. Error     t value
(Intercept)  0.5800609 0.05287169   10.971105
   Alt.km. -0.1272907 0.02618113   -4.861927

(Dispersion Parameter for Gaussian family taken to be 0.0040884 )

    Null Deviance: 0.1170857 on 6 degrees of freedom
Residual Deviance: 0.0204421 on 5 degrees of freedom

Number of Fisher Scoring Iterations: 1
Analysis of Deviance Table
Gaussian model
Response: H

Terms added sequentially (first to last)
        Df   Deviance Resid. Df Resid. Dev
   NULL                       6  0.1170857
Alt.km.  1 0.09664358         5  0.0204421
```