Chapter 1

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Functional versus stochastic relations

Model: a mathematical approximation of the relationship among real quantities (equation & assumptions about terms).

- We have seen several models for an outcome variable from either one or two populations.
- Now we'll consider models that relate an outcome to one or more continuous predictors.

- **Functional relationships** are perfect. Realizations (X_i, Y_i) solve the relation $Y = f(X)$.
- A **statistical relationship** is not perfect. There is a trend plus error. Signal plus noise.

Section 1.1: relationships between variables

- A **functional relationship** between two variables is deterministic, e.g. $Y = \cos(2.1x) + 4.7$. Although often an approximation to reality (e.g. the solution to a differential equation under simplifying assumptions), the relation itself is "perfect." (e.g. page 3)
- A **statistical** or **stochastic** relationship introduces some "error" in seeing Y, typically a **functional relationship** plus **noise**. (e.g. Figures 1.1, 1.2, and 1.3; pp. 4–5).

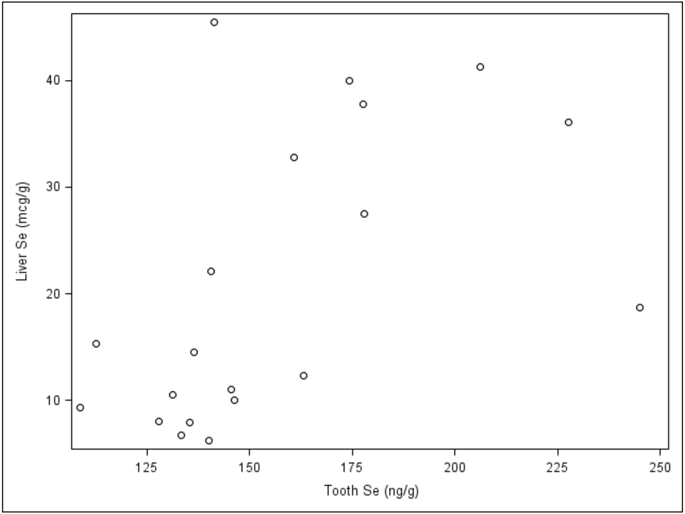
Statistical relationship: not a perfect line or curve, but a general tendency plus slop.

Whale Selenium

- Selenium protects marine animals against mercury poisoning.
- $n = 20$ Beluga whales were sampled during a traditional Eskimo hunt; tooth Selenium (Se) and liver Se were measured.
- Would be useful to be able to use tooth Selenium as a proxy for liver Selenium (easier to get).
- Same idea with "biomarkers" in biostatistics.

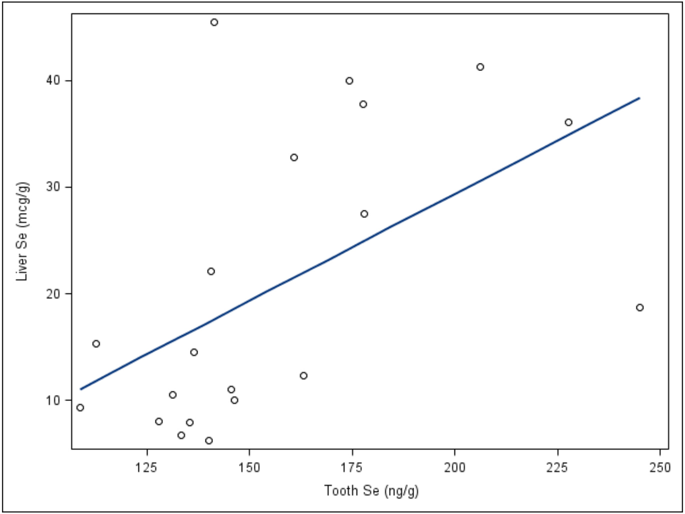
```
data whale;
input liver tooth @@;
label liver="Liver Se (mcg/g)"; label tooth="Tooth Se (ng/g)";
datalines;
 6.23 140.16  6.79 133.32  7.92 135.34  8.02 127.82  9.34 108.67
10.00 146.22 10.57 131.18 11.04 145.51 12.36 163.24 14.53 136.55
15.28 112.63 18.68 245.07 22.08 140.48 27.55 177.93 32.83 160.73
36.04 227.60 37.74 177.69 40.00 174.23 41.23 206.30 45.47 141.31
;
proc sgscatter; plot liver*tooth / reg; * or pbspline or nothing;
```

Whale Selenium



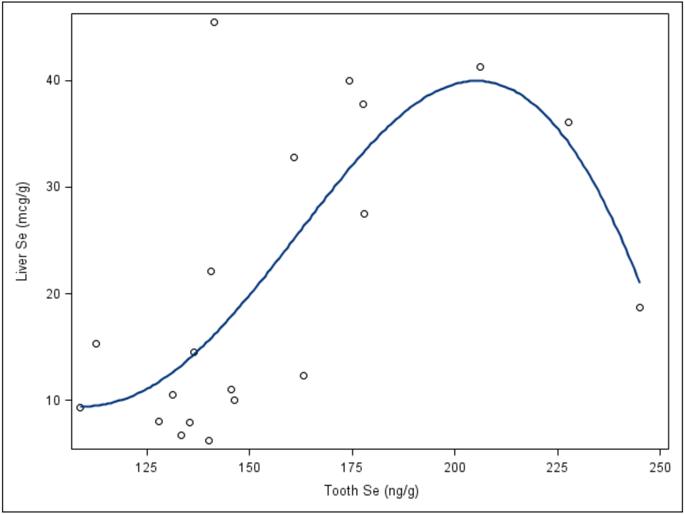
Must decide what is the proper *functional form* for the trend in this relationship, e.g. linear, curved, piecewise continuous, cosine?

Whale Selenium



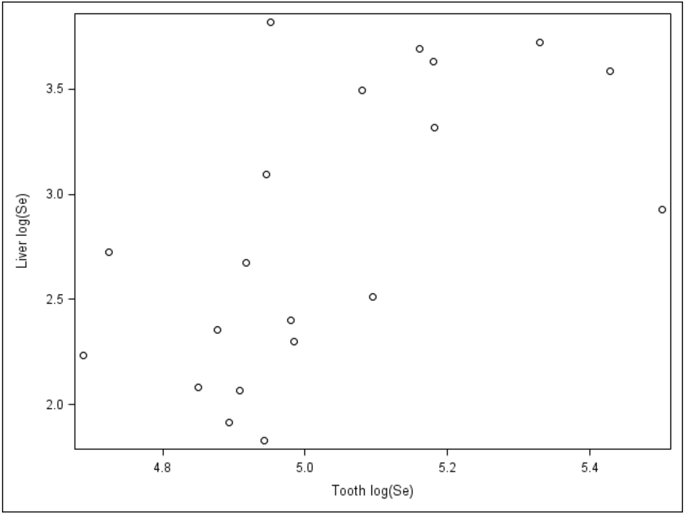
Is a line "correct?"

Whale Selenium



How about a curve?

Whale Selenium



Taking log of both variables.

For a sample of n pairs $\{(x_i, Y_i)\}_{i=1}^{n}$, let

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \ i = 1, \dots, n,$$

where

- $Y_1, \dots, Y_n$ are realizations of the response variable,
- $x_1, \dots, x_n$ are the associated predictor variables,
- β_0 is the intercept of the regression line,
- β_1 is the slope of the regression line, and
- $\epsilon_1, \dots, \epsilon_n$ are unobserved, uncorrelated random errors.

This model assumes that x and Y are *linearly* related, i.e. the mean of Y changes linearly with x.

Assumptions about the random errors

We assume that $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2$, and $\text{corr}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$: mean zero, constant variance, uncorrelated.

- $\beta_0 + \beta_1 x_i$ is the *deterministic* part of the model. It is fixed but unknown.
- ϵ_i represents the random part of the model.

The goal of statistics is often to separate signal from noise; which is which here?

Mean and variance of each Y_i

Note that

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + E(\epsilon_i) = \beta_0 + \beta_1 x_i,$$

and similarly

$$\text{var}(Y_i) = \text{var}(\beta_0 + \beta_1 x_i + \epsilon_i) = \text{var}(\epsilon_i) = \sigma^2.$$

Also, $\text{corr}(Y_i, Y_j) = 0$ for $i \neq j$.

These use results from A.3.

Example: Pages 10–11

- Consultant studies relationship between number of bids requested by construction contractors for lighting equipment over a week x_i (i denotes which week) and the time required to prepare the bids Y_i. Suppose *we know*

$$Y_i = 9.5 + 2.1x_i + \epsilon_i.$$

- If we see $(x_3, Y_3) = (45, 108)$ then $\epsilon_3 = 108 - [9.5 + 2.1(45)] = 4$. See Fig. 1.6.

Example: Pages 10–11

- The mean time given x is $E(Y) = 9.5 + 2.1x$. When $x = 45$, our expected y-value is 104, but we will actually *observe* a value somewhere around 104.
- What does 9.5 represent here? Is it sensible/interpretable?
- How is 2.1 interpreted here?
- In general, β_1 represents how the mean response changes when x is increased one unit.

Simple linear regression using matrices

Note the simple linear regression model can be written in matrix terms as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

or equivalently

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon},$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

This will be useful later on.

Section 1.6: Estimation of (β_0, β_1)

- β_0 and β_1 are unknown parameters to be estimated from the data: $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$.
- They completely determine the unknown mean at each value of x:

$$E(Y) = \beta_0 + \beta_1 x.$$

- Since we expect the various Y_i to be both above and below $\beta_0 + \beta_1 x_i$ roughly the same amount $(E(\epsilon_i) = 0)$, a good-fitting line $b_0 + b_1 x$ will go through the "heart" of the data points in a scatterplot.
- The method of least-squares formalizes this idea by minimizing the sum of the squared deviations of the observed y_i from the line $b_0 + b_1 x_i$.

Least squares method for estimating (β_0, β_1)

The sum of squared deviations about the line is

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 x_i)\right]^2 .$$

Least squares minimizes $Q(\beta_0, \beta_1)$ over all (β_0, β_1).
Calculus shows that the least squares estimators are

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b_0 &= \bar{Y} - b_1 \bar{x} \end{aligned}$$

Proof:

$$\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 x_i)(-x_i) = -2\left[\sum_{i=1}^{n} x_i Y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2\right],$$

$$\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 x_i)(-1) = -2\left[\sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i\right].$$

Two equations in two unknowns

Setting these equal to zero, and dropping indexes on the summations, we have

$$\left\{ \begin{array}{rcl} \sum x_i Y_i & = & b_0 \sum x_i + b_1 \sum x_i^2 \\ \sum Y_i & = & n b_0 + b_1 \sum x_i \end{array} \right\} \Leftarrow \text{“normal” equations}$$

Multiply the first by n and multiply the second by $\sum x_i$ and subtract yielding

$$n \sum x_i Y_i - \sum x_i \sum Y_i = b_1 \left[n \sum x_i^2 - \left(\sum x_i \right)^2 \right].$$

Solving for b_1 we get

$$b_1 = \frac{n \sum x_i Y_i - \sum x_i \sum Y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i Y_i - n \bar{Y} \bar{x}}{\sum x_i^2 - n \bar{x}^2}.$$

Not quite as nice as (1.10) p. 17)

The second normal equation immediately gives

$$b_0 = \bar{Y} - b_1\bar{x}.$$

Our solution for b_1 is correct but not as aesthetically pleasing as the purported solution

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Show

$$\begin{aligned} \sum(x_i - \bar{x})(Y_i - \bar{Y}) &= \sum x_i Y_i - n\bar{Y}\bar{x} \\ \sum(x_i - \bar{x})^2 &= \sum x_i^2 - n\bar{x}^2 \end{aligned}$$

Properties of least squares estimators

The line $\hat{Y} = b_0 + b_1 x$ is called the *least squares* estimated regression line. Why are the least squares estimates (b_0, b_1) "good?"

- They are unbiased: $E(b_0) = \beta_0$ and $E(b_1) = \beta_1$.
- Among all linear unbiased estimators, they have the smallest variance. They are **b**est **l**inear **u**nbiased **e**stimators, BLUEs.

We will show the first property next. The second property is formally called the "Gauss-Markov" theorem (1.11) and is proved in linear models (page 18).

b_0 and b_1 are unbiased (Section 2.1, p. 42) Recall that least-squares estimators (b_0, b_1) are given by:

$$b_1 = \frac{n\sum x_i Y_i - \sum x_i \sum Y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i Y_i - n\bar{Y}\bar{x}}{\sum x_i^2 - n\bar{x}^2},$$

and

$$b_0 = \bar{Y} - b_1\bar{x}.$$

Note that the numerator of b_1 can be written

$$\sum x_i Y_i - n\bar{Y}\bar{x} = \sum x_i Y_i - \bar{x}\sum Y_i = \sum (x_i - \bar{x})Y_i.$$

Keep going...

Then the expectation of b_1's numerator is

$$\begin{aligned} E\left\{\sum(x_i - \bar{x})Y_i\right\} &= \sum(x_i - \bar{x})E(Y_i) \\ &= \sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum x_i - n\bar{x}\beta_0 + \beta_1 \sum x_i^2 - n\bar{x}^2\beta_1 \\ &= \beta_1\left(\sum x_i^2 - n\bar{x}^2\right) \end{aligned}$$

Finally,

$$\begin{aligned} E(b_1) &= \frac{E\left\{\sum(x_i - \bar{x})Y_i\right\}}{\sum x_i^2 - n\bar{x}^2} \\ &= \frac{\beta_1\left(\sum x_i^2 - n\bar{x}^2\right)}{\sum x_i^2 - n\bar{x}^2} \\ &= \beta_1. \end{aligned}$$

$E(b_0) = \beta_0$

Also,

$$
\begin{aligned}
E(b_0) &= E(\bar{Y} - b_1\bar{x}) \\
&= \frac{1}{n}\sum E(Y_i) - E(b_1)\bar{x} \\
&= \frac{1}{n}\sum[\beta_0 + \beta_1 x_i] - \beta_1\bar{x} \\
&= \frac{1}{n}[n\beta_0 + n\beta_1\bar{x}] - \beta_1\bar{x} \\
&= \beta_0.
\end{aligned}
$$

As promised, b_1 is unbiased for β_1 and b_0 is unbiased for β_0.

Whale Selenium, SAS code

- `proc reg` and `proc glm` fit regression models.
- Both include a `model` statement that tells SAS what the explanatory variable(s) are (on the right of = separated by spaces) and the response (on the left).

```
data whale;
input liver tooth @@;
label liver="Liver Se (mcg/g)"; label tooth="Tooth Se (ng/g)";
datalines;
 6.23 140.16  6.79 133.32  7.92 135.34  8.02 127.82  9.34 108.67
10.00 146.22 10.57 131.18 11.04 145.51 12.36 163.24 14.53 136.55
15.28 112.63 18.68 245.07 22.08 140.48 27.55 177.93 32.83 160.73
36.04 227.60 37.74 177.69 40.00 174.23 41.23 206.30 45.47 141.31
;
proc reg;
 model liver=tooth;
```

Whale Selenium, SAS output

```
                         The REG Procedure
                          Model: MODEL1
              Dependent Variable: liver Liver Se (mcg/g)

               Number of Observations Read          20
               Number of Observations Used          20

                        Analysis of Variance

                                   Sum of          Mean
Source                   DF       Squares        Square    F Value    Pr > F
Model                     1     992.10974     992.10974       7.31    0.0146
Error                    18    2444.58376     135.81021
Corrected Total          19    3436.69350

             Root MSE             11.65376    R-Square     0.2887
             Dependent Mean       20.68500    Adj R-Sq     0.2492
             Coeff Var            56.33920

                        Parameter Estimates

                                       Parameter       Standard
Variable     Label              DF      Estimate          Error    t Value    Pr > |t|
Intercept    Intercept           1     -10.69641       11.89954      -0.90      0.3806
tooth        Tooth Se (ng/g)     1       0.20039        0.07414       2.70      0.0146
```

From this, $b_0 = -10.69$, $b_1 = 0.2004$, and $\hat{\sigma} = 11.65$.
Interpretation of each?