History of Linear Regression

Early Ideas and Methodology

- Early 1800's, Legendre, Laplace, Gauss (1822) fully established key properties of the method of *least squares* for estimating linear relationships to paired observations.
  - Used in various fields - astronomy, physics, for example
  - Method was intuitive - fit linear function through points using some objective criterion for 'fit'
  - This work predated by many decades the beginning development of modern statistics

Gauss and Estimation of Planetary Distance

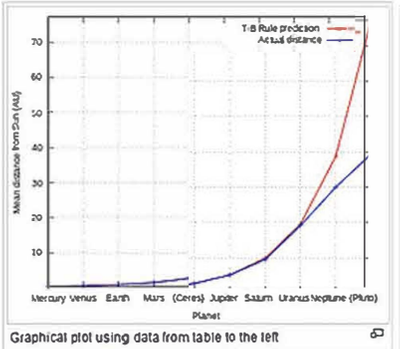
The 'missing' planet?

On 1 January 1801, the planetoid Ceres was discovered and tracked for 40 days before being lost in the glare of the Sun. This was an exciting discovery, as the Titius-Bode Law predicted a 'missing' planet between Mars and Jupiter. This law gave an approximate relationship between the size of the orbits of each planet to follow this simple formula:

$$a = 0.4 + 0.3 \times 2^m$$
$$\text{for } m = -\infty, 0, 1, 2, 3, 4, 5, 6, 7, 8$$

Where 'a' is the semi-major axis of each planet in terms of astronomical units (AU) i.e. distance from Earth to the Sun. For the outer planets (Jupiter onwards), the formula implies that the next planet is roughly twice as far from the Sun as the previous planet, as presented in the following table:

Planet	k	T–B rule distance (AU)	Real distance (AU)	% error (using real distance as the accepted value)
Mercury	0	0.4	0.39	2.56%
Venus	1	0.7	0.72	2.78%
Earth	2	1.0	1.00	0.00%
Mars	4	1.6	1.52	5.26%
Ceres[1]	8	2.8	2.77	1.08%
Jupiter	16	5.2	5.20	0.00%
Saturn	32	10.0	9.54	4.82%
Uranus	64	19.6	19.2	2.08%
Neptune	128	38.8	30.06	29.08%
Pluto[2]	256	77.2[2]	39.44	95.75%



Graphical plot using data from table to the left

Figure 2: *Actual vs Predicted with the Titius-Bode Law for the planets, which is fairly accurate from Mercury to Uranus The accuracy of this simple rule of thumb is mostly due to coincidence.*
Source: *Wikipedia*

From article by Milton Lim: "Gauss, Least Squares, and the Missing Planet"

Gauss and Estimation of Planetary Distance

Estimating the equation from the actual data (up to Neptune), does it work (i.e. confirm the rule)?

```
. regress distance kfactor if kfactor < 128
```

distance	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
kfactor	.2925629	.0018878	154.98	0.000	.2879436	.2971822
_cons	.3980644	.0493227	8.07	0.000	.2773762	.5187527

So equation (based on actual data would be :

distance $= 0.3981 + .2925 * k$ ($k = 2^m$ for $m = -\infty, 0, 1, 2, 3, 4, 5, 6, 7, 8$

Looks good:

constant (β_0 or intercept of line) equals about .40

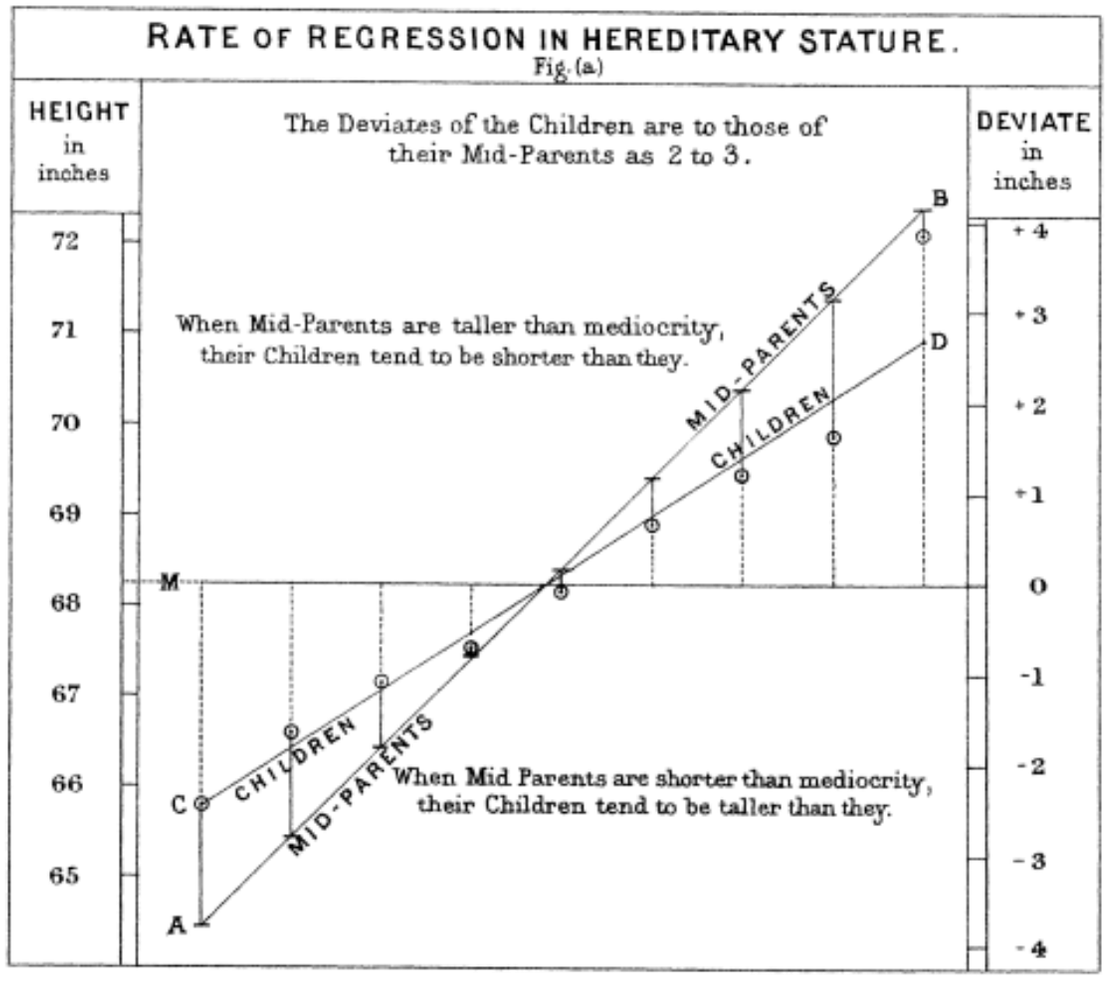
multiplier on k-factor is about 0.30

Regression: name origin and motivating application

- Developed in the 1870s by Sir Francis Galton (cousin of Darwin) for quantifying genetic inheritance, i.e. how strongly parents' characteristics influence those of their offspring.
  - Experiment on peas, weighing a sample of mother and daughter seeds, plotting the weights of the seed pairs.
  - The median weights of daughter seeds from a particular size of mother seed approximately followed a straight line with positive slope, but less than 1.0.
  - Galton called this *regression* to the mean.
    This is the origin of the name *regression analysis*
  - Later did a similar analysis of the heights of a large sample of (human) fathers and sons, and found a similar result. He wrote down the simple linear regression model, but did not propose a systematic way of estimation

Galton's Regression to the Mean

Plate IX.



J.P. & W.R. Emslie, lith.

Modern Regression

- In 1896, Karl Pearson further developed, formalized the Ordinary Least Squares (OLS) method of estimation. At this time, much of 'empirical' statistical methods were becoming formalized mathematically
  - In and of itself, least squares is an estimation method, but is thought of as synonymous with regression modeling
  - More general development of *the general linear model* a comprehensive mathematical/statistical framework that subsumes linear regression as a special case. (C.R. Rao, S. Searle)
  - Generalized linear models (McCullagh and Nelder), extend the capabilities to many response variable types not accommodated by linear model methods
  - Longitudinal models, other extensions and generalizations also extensively developed

A Relationship via Simple Linear Regression

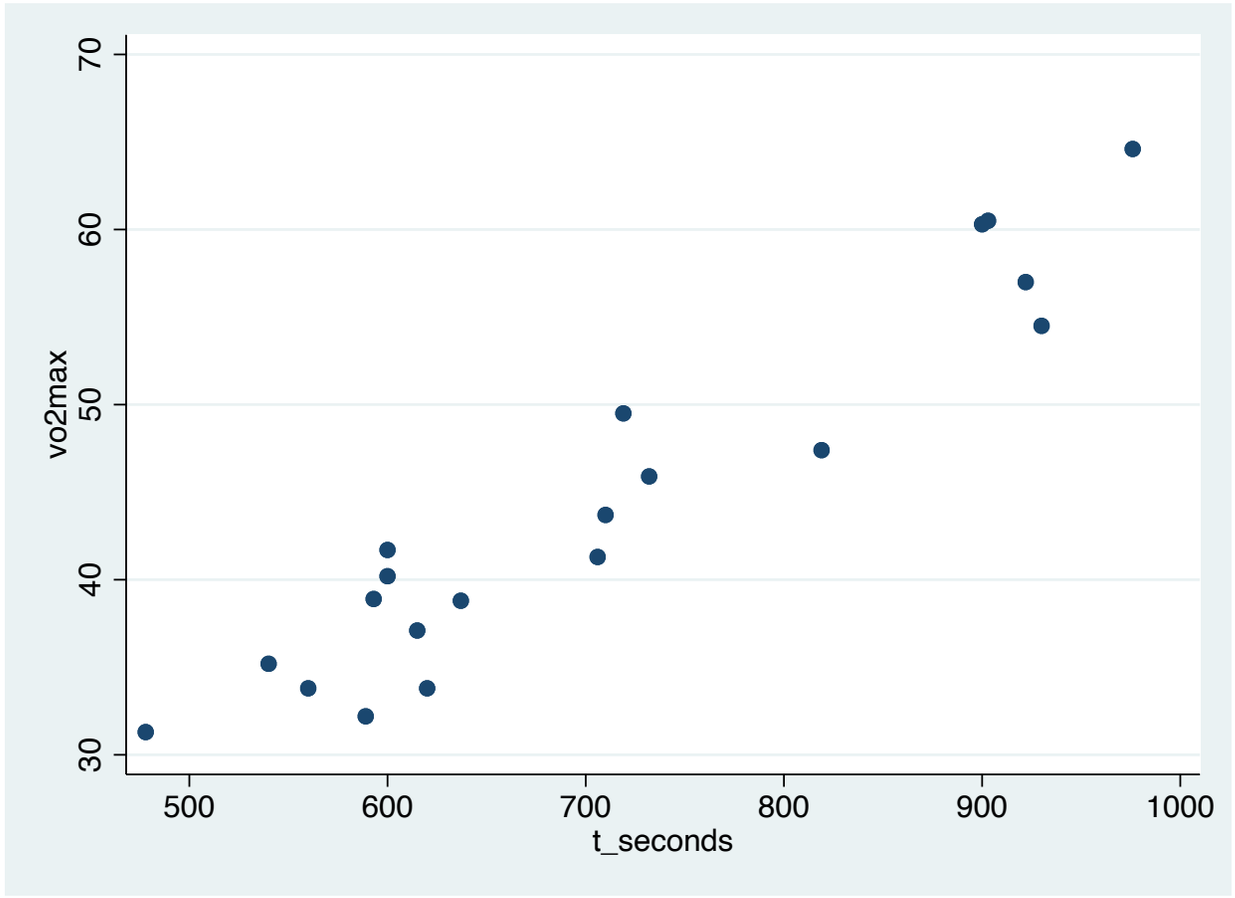
Example: (From Fisher and Van Belle biostatistics text, 1993)
VO_2 MAX is a measure of the body's capacity to utilize oxygen, with higher values indicating greater capacity. Using a treadmill test, recording the time to VO_2 MAX, one can evaluate the relationship between oxygen uptake and endurance during strenuous exercise

- We collect observations (X_i, Y_i) observations. (where i indicates the subject number) and plot X versus Y (called a scatterplot) and speculate on f by looking at what the relationship might be.

A Relationship via Simple Linear Regression

– In Stata, use the following code to load the data and make a scatter plot

```
. use  vo2data , clear
. scatter t_seconds vo2max
```



– If the shape looks roughly linear we may start simply by guessing that we have a straight line relationship, observed imperfectly:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where ϵ_i (the "error term") denotes deviation from a perfect line

Correlation and Regression

Taking a step back from regression, after graphing the data, we could look at the *correlation* of the two variables, $\mathrm{cor}(X, Y)$.

What is correlation?

Correlation

SPRM 2.11, also (oddly), Chapter 6

- The relationship between the two variables in the scatterplot may be expressed as a single summary measure: correlation
- Terms: "correlation", "Pearson correlation", "product-moment correlation", "correlation coefficient" - not necessarily interchangeable - need to check what was computed - there are other correlation measures
- Notation: correlation coefficient between x and y is $r = \mathrm{corr}(x, y)$
- ρ (estimator denoted by r) is a *measure of association* for two continuous variables:
  - strength of association – how far r is from zero - r takes on values from -1 to $+1$
  - direction of association – r positive or negative

Computing Correlation

- Recall: variance of x is

$$\text{var}(x) = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})(x_i - \bar{x})$$

- Define: **covariance** of x and y:

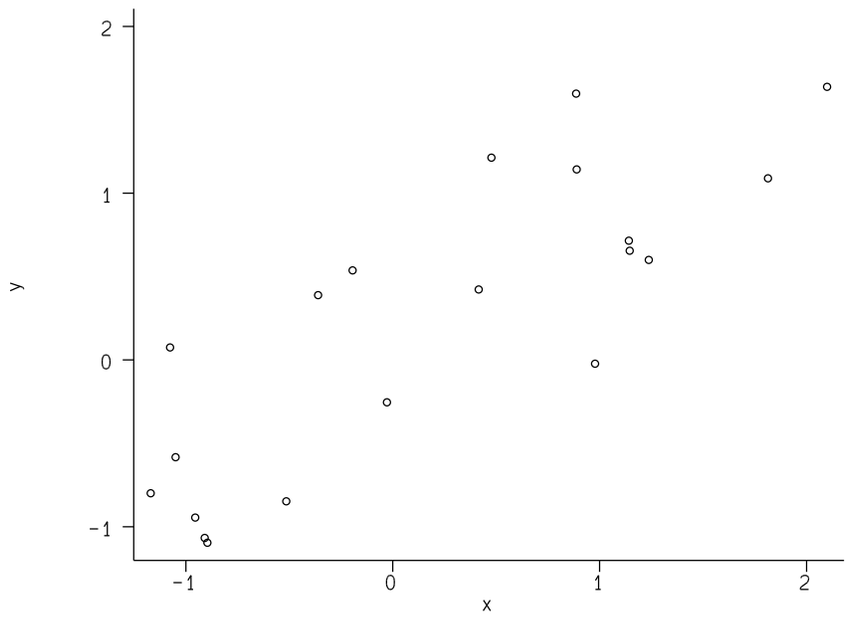
$$\text{cov}(x, y) = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- Define: **correlation** of x and y (equivalent to SPRM eqn. 2.21):

$$\text{corr}(x, y) = \frac{\text{cov(x, y)}}{\sqrt{\text{var}(x)\text{var}(y)}}$$

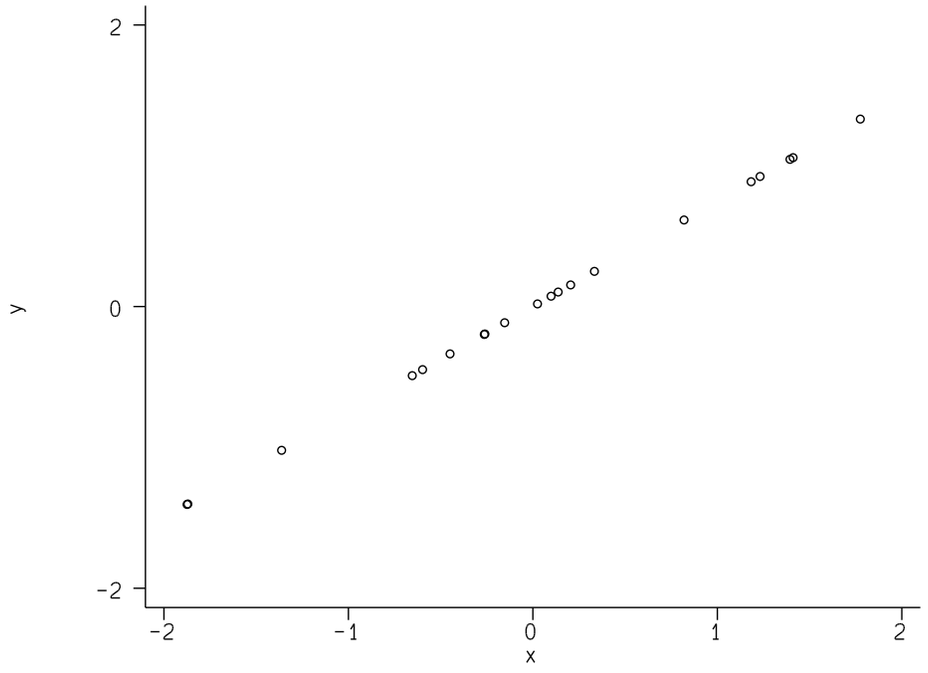
Correlation
Examples and Interpretation

$$r = \text{corr}(x, y) = 0.82$$



$\Longrightarrow$ strong positive or direct correlation

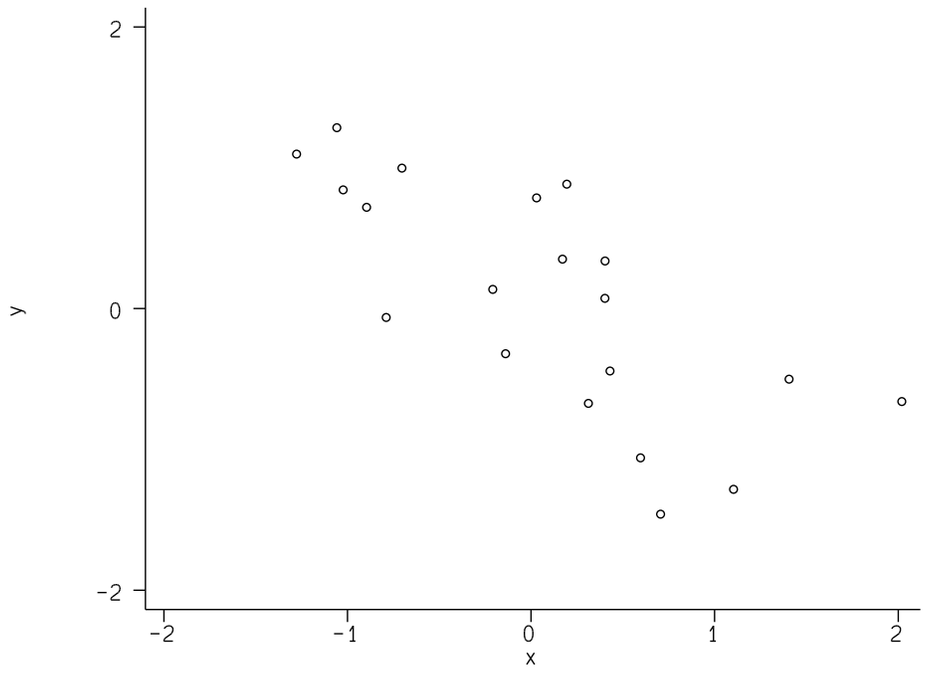
$$r = \text{corr}(x, y) = 1.00$$



$\Longrightarrow$ perfect positive correlation — if we know x, we can perfectly predict y

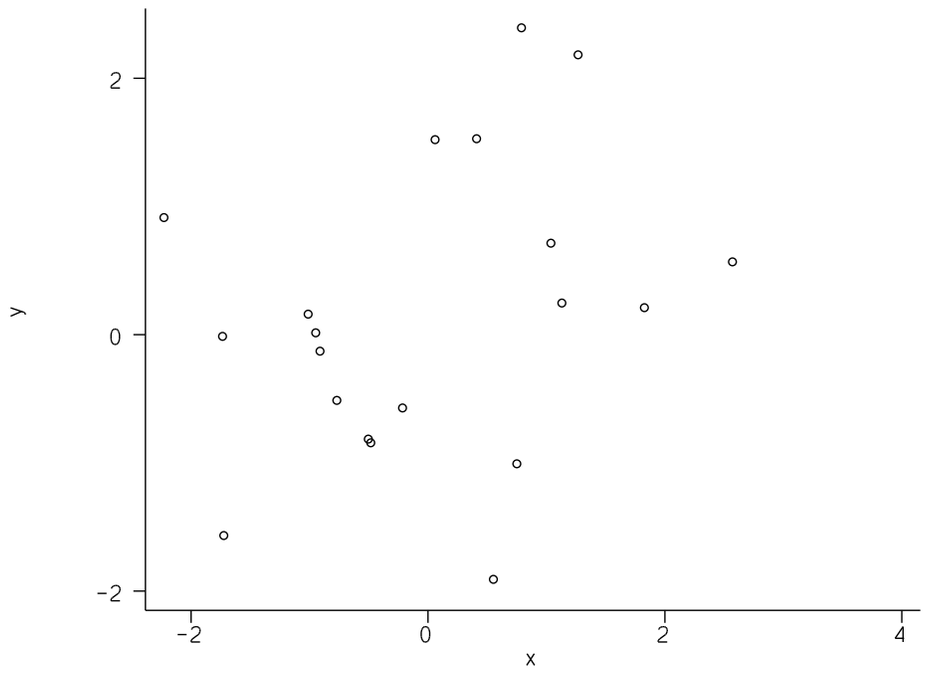
Correlation Examples (cont.)

$$r = \text{corr}(x, y) = -0.74$$



$\Longrightarrow$ strong negative or inverse correlation

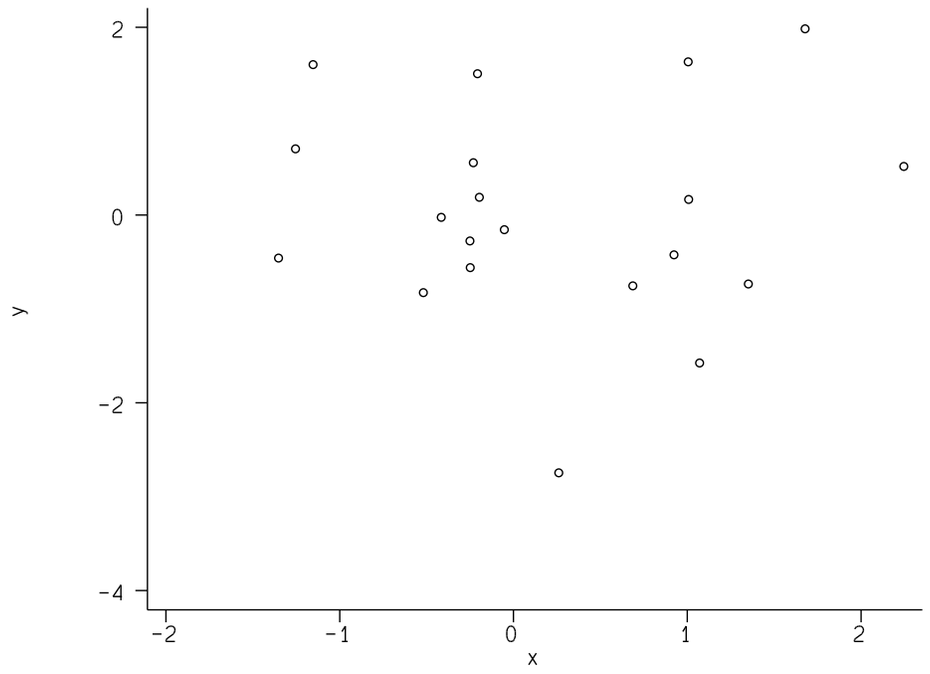
$$r = \text{corr}(x, y) = 0.30$$



$\Longrightarrow$ fairly weak, but positive association between y and x

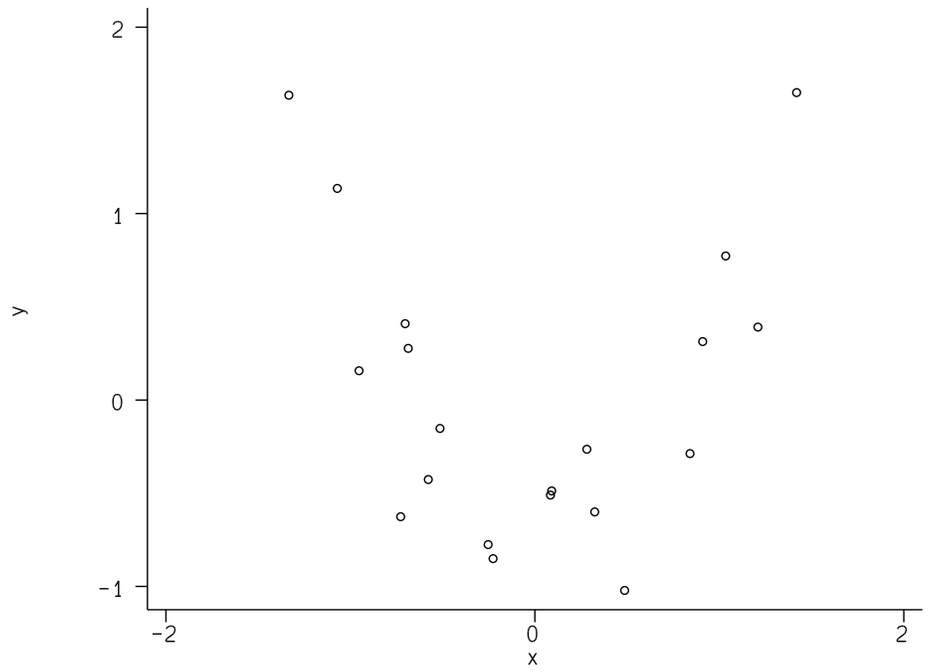
Correlation Examples (cont.)

$$r = \text{corr}(x, y) = -0.0006$$



$\Longrightarrow$ no association between x and y

$$r = \text{corr}(x, y) = -0.003$$

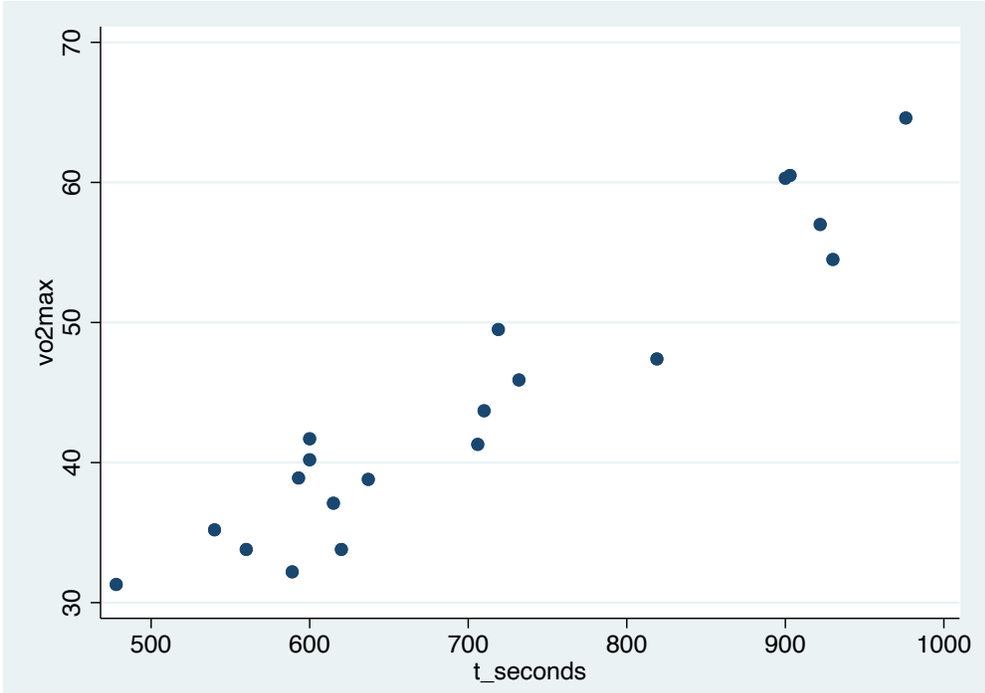


$\Longrightarrow$ here, there is clearly a strong *curvilinear* relationship between y and x. This relationship is *not* captured by r because r assesses the degree of *linear relationship* between y and x. A line through the points would be flat(e.g., parallel to the x-axis)

Correlation Coefficient

- **Additional notes:**
  - r is symmetric: $r = \text{corr}(x, y) = \text{corr}(y, x)$
  - r assesses **linear** (part of) relationship between x and y.

  Back to the VO_2 MAX example:



$\Longrightarrow$ Strong positive correlation is apparent

Correlation Coefficient

computing the correlation in Stata:

```
.. correlate t_seconds vo2max
(obs=20)

             | t_seco~s   vo2max
-------------+------------------
   t_seconds |   1.0000
      vo2max |   0.9533   1.0000
```

computing the covariance in Stata:

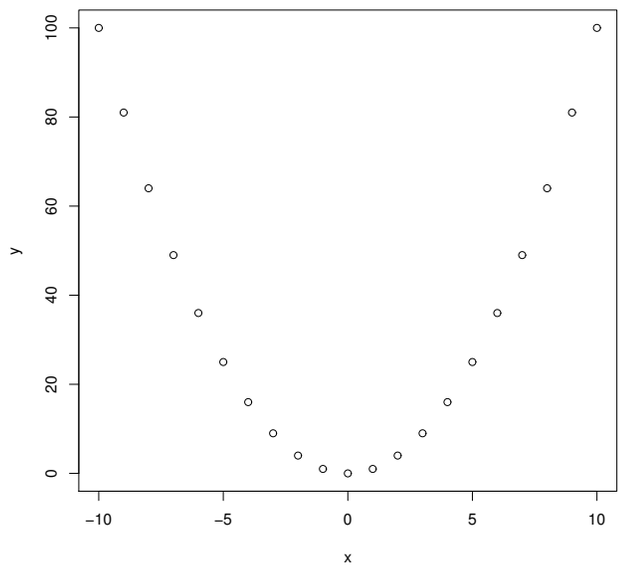
```
. correlate t_seconds vo2max, cov
(obs=20)

             | t_seco~s   vo2max
-------------+------------------
   t_seconds |    22631
      vo2max |  1470.55  105.149
```

There is an extremely strong positive correlation, as the plot suggested.

Correlation and Dependence

Again, $\mathrm{cor}(X, Y) = 0$ does not imply X and Y are independent.

An example, $Y = X^2$, where (Pearson) correlation is 0:




```
. cor x y
             |      X       Y
-------------+------------------
           X |   1.0000
           Y |   0.0000   1.0000
```

Correlation and Dependence

(Pearson) correlation measures the *linear association* between variables. If two variables are correlated, they are dependent. But two variables can be dependent while still having a correlation of zero, as shown in the example. Only when both X and Y are normally distributed, $\text{cor}(X, Y) = 0$ implies X and Y are independent.

Correlation vs Causation

Is high correlation evidence for causation? (ex/ DeVogli et al).

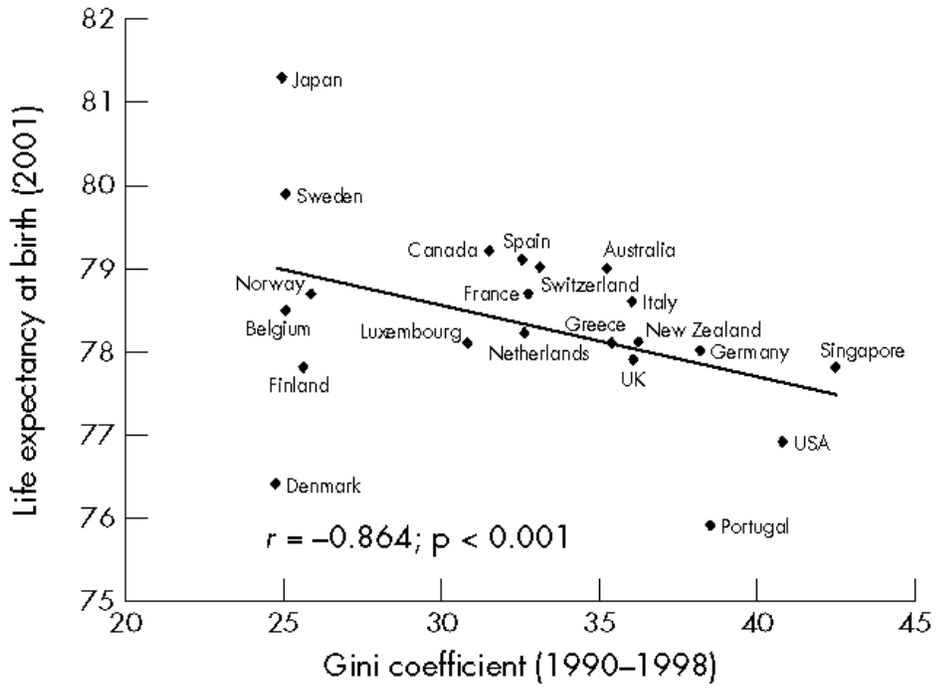


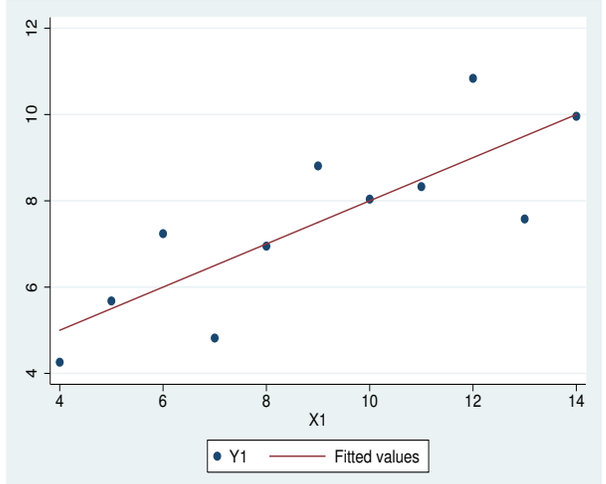
Figure 2 Income inequality and life expectancy at birth among industrialised countries (n = 21). Data are from the human development indicators 2003. The correlation presented in the figure is weighted by population size and adjusted for per capita gross domestic product

Correlation is symmetric while causation is not, implies a directional relationship (and additional conditions).

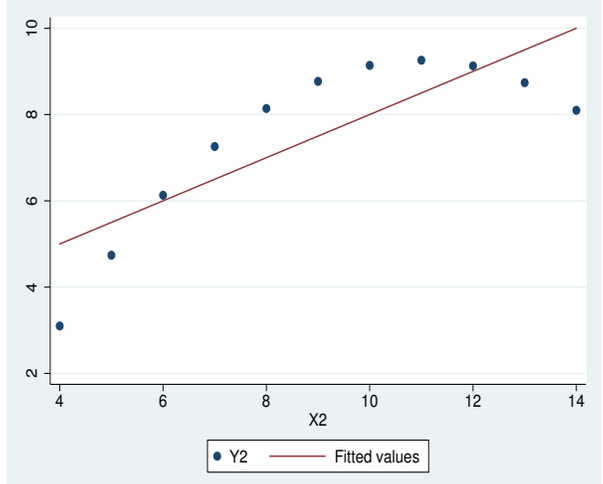
Pitfalls, Misuses of Linear Regression

A classic example was constructed by the statistician F.J. Anscombe to demonstrate both the importance of graphing data to assess the appropriateness of a linear model relationship. Each dataset consists of eleven (x,y) points.

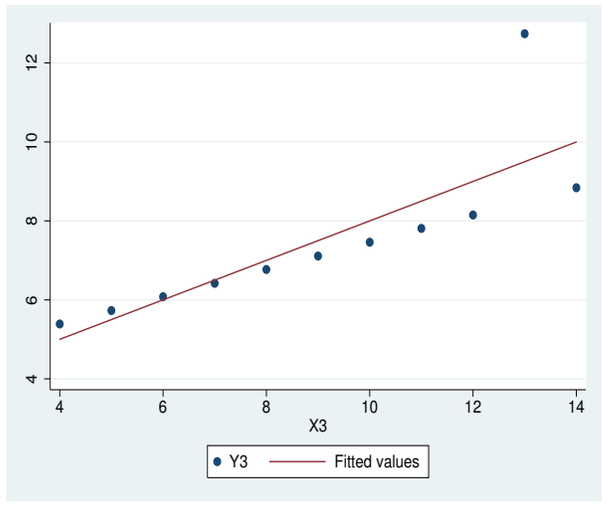
We will come back to this example later in the homework, seeing that these data share same correlation coefficient, slope of regression line (shown), inferential statistics and conclusions, etc



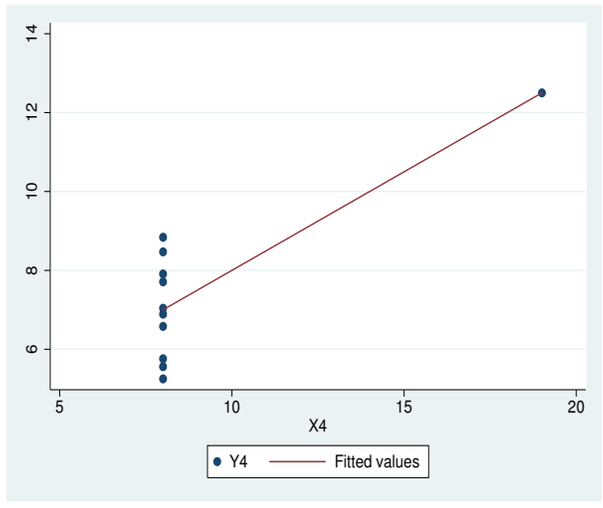
(a) fitted



(b) non-linear



(c) outlier



(d) influence

Linear Regression

Going beyond correlation, we want to characterize the relationship with a function f, possibly to use for prediction, etc. Simple linear regression provides this. The VO_2Max example:

```
. regress vo2max t_seconds

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =    179.29
       Model |  1815.55337         1  1815.55337   Prob > F        =    0.0000
    Residual |  182.272082        18  10.1262268   R-squared       =    0.9088
-------------+----------------------------------   Adj R-squared   =    0.9037
       Total |  1997.82545        19  105.148708   Root MSE        =    3.1822

------------------------------------------------------------------------------
      vo2max |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   t_seconds |   .0649794   .0048528    13.39   0.000      .054784    .0751748
       _cons |  -1.584694   3.506099    -0.45   0.657    -8.950735    5.781347
------------------------------------------------------------------------------
```

Note: In Stata, $\beta_0 =$ _cons, $\beta_1 =$ t_ seconds (name of predictor var)

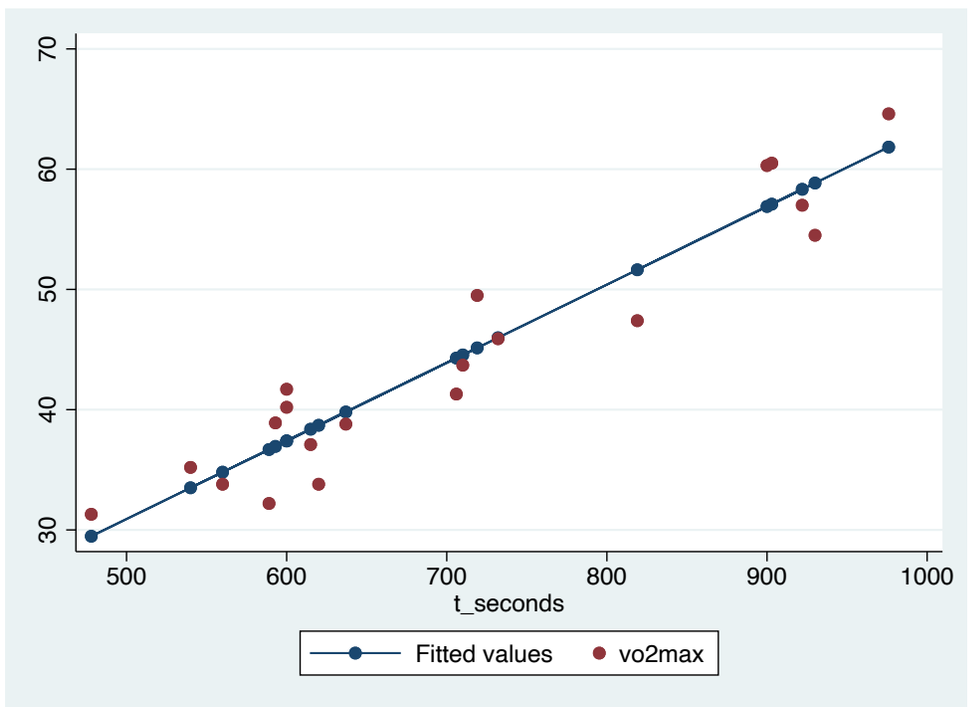
Linear Regression

Drawing the resultant linear relationship (that the model imposes):

```
. predict yhat
(option xb assumed; fitted values)
. scatter yhat vo2max t_seconds, c(l)
```

The model is

$$\text{predicted VO2MAX} = -1.585 + .0650 \times \text{seconds}$$



Linear Regression - Model in R

```
> library(foreign)
> v02 <- read.dta("v02.dta")
> v02
   t_seconds vo2max
1        706   41.3
2        732   45.9
3        930   54.5
.
.
> vo2model <-lm(v02$vo2max ~ v02$t_seconds)
> summary(vo2model)
Call:
lm(formula = v02$vo2max ~ v02$t_seconds)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.584694   3.506099  -0.452    0.657
v02$t_seconds  0.064979   0.004853  13.390 8.48e-11 ***
---

Residual standard error: 3.182 on 18 degrees of freedom
Multiple R-squared:  0.9088,Adjusted R-squared:  0.9037
F-statistic: 179.3 on 1 and 18 DF,  p-value: 8.479e-11
```

Formal Statement of Model

From SPRM Section 2.3

- Consider a statistical model where there is only one predictor (X) variable and the relationship is linear to some response variable (Y). We frame the model around predicting the mean of Y at each X. The theoretical model can be stated as follows:

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

where:

- X is the value of the predictor variable
- $\mu_{Y|X}$ is the mean value (also called Expected value, $E(Y|X)$) of the response variable at the value of X
- β_0 is the intercept term, equaling the mean of Y when $X = 0$
- β_1 is the slope, equaling the rate of change of Y per unit change in X

Formal Statement of Model

From SPRM Section 2.3

- The realized regression model is expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where:

- Y_i is the value of the response variable in the ith subject or unit for $i = 1, \ldots, n$
- β_0 is the intercept term, and β_1 is the slope
- X_i is the value of the predictor variable in the ith unit
- ϵ_i is a (random) error term with mean $\mathrm{E}(\epsilon_i) = 0$ and variance $\sigma^2_{\epsilon_i} = \sigma^2$; ϵ_i and ϵ_j are uncorrelated so that their covariance is zero for all i, j $(i \neq j)$,$i = 1, \ldots, n$. $\Rightarrow$ In a random sample, individual i and j are uncorrelated.

Formal Statement of Model

- This regression model is said to be *simple* because there is one predictor. It is *linear in the parameters* because no parameter appears as an exponent or is multiplied or divided by another parameter. This model is said to be *linear in the predictor variable*, because of the functional form of Y in relation to X.
- Important Features of Model
  - The response Y_i in the ith unit is the sum of two components: (1) $\beta_0 + \beta_1 X_i$ and (2) the random term ϵ_i. Hence, Y_i is a random variable.
  - Since $\mathrm{E}(\epsilon_i) = 0$,

    $$\mathrm{E}(Y_i|X_i) = \mathrm{E}(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i + \mathrm{E}(\epsilon_i) = \beta_0 + \beta_1 X_i$$

    This is also called the expected mean model. The expected mean model is the theoretical form and has no error term.

Regression Model Conceptually

Note that $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ implies that the true observations Y are still random, even after learning what the relationship is. This is because we have a statistical model and not a physical one. This randomness is characterized by ϵ_i and their distribution.

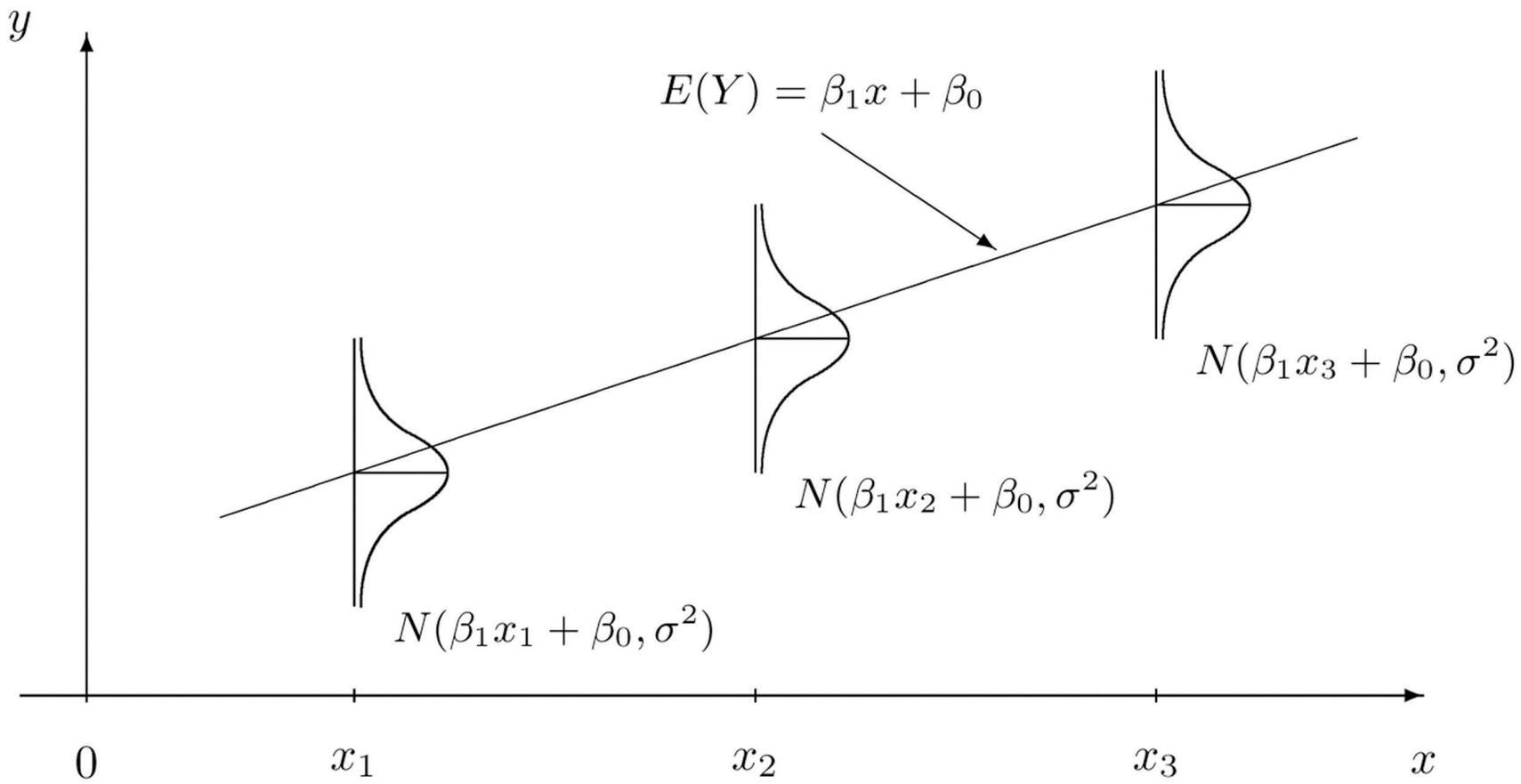
If the errors ϵ_i are normally distributed, say from $N(0, \sigma^2)$, then the Y_i are also normally distributed and come from distributions

$$N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Because the mean of this distribution varies in a deterministic way with X, we are able to make statements about how much the mean (or average, or expected value) of Y changes as X changes.

Another way to think about it is to picture a whole population of Ys associated with one particular X, and the mean of this population will be $\beta_0 + \beta_1 X$, and its variance will be σ^2.

Regression Model Conceptually



Interpreting the Model - Coefficients

SPRM 2.3

In simple linear regression, there are two *coefficients*.

1. β_0, the intercept, equaling to the expected response Y when $X = 0$.
   $E(Y|X = 0) = \beta_0 + \beta_1 X = \beta_0 + \beta_1 \times 0 = \beta_0$

2. β_1: the slope. It corresponds to the average (or mean) change in Y (note: this is the same as the change in average or mean of Y) associated with 1 unit increase in X.
   $E(Y|X = x + 1) - E(Y|X = x)$
   $= (\beta_0 + \beta_1 \times (x + 1)) - (\beta_0 + \beta_1 \times x) = \beta_1$

Obtaining the Coefficients - Ordinary Least Squares (OLS)

SPRM Section 2.7

- How to estimate β_0 and β_1 based on the data Y and X? In 1896, Karl Pearson developed the Ordinary Least Squares (OLS) method
- The OLS method finds the parameter (here β_0, β_1) estimates that can minimize the Sum of Squared Errors (SSE).

$$\text{SSE} = \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$$

- First, we introduce "residual" e_i as $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \times X_i$ is the "fitted" value predicted by our model
  - the e_i's are vertical distances between where we predict and the observed value of Y_i.
  - Why squared distances? Intuitively, both positive and negative distances count. Also, large differences weighted more so than small differences. Other reasons are more mathematical.

Ordinary Least Squares (OLS) Estimation

- We obtain the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the sum of squared errors. find β's that lead to:

$$\min \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

We do by differentiating the SSE with respect to β_1 and β_0, setting the equations equal to zero, and solving for the β's:

$$\text{w.r.t } \beta_0 \Rightarrow -2 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0, \quad (1)$$

$$\text{w.r.t } \beta_1 \Rightarrow -2 \sum X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0. \quad (2)$$

From equation (1):

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Plug in equation (2):

$$\sum X_i (Y_i - \bar{Y} - \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i) = 0$$

$$\Rightarrow \quad \sum X_i(Y_i - \bar{Y}) = \hat{\beta}_1 \sum X_i(X_i - \bar{X})$$

$$\Rightarrow \quad \hat{\beta}_1 = \frac{\sum X_i(Y_i - \bar{Y})}{\sum X_i(X_i - \bar{X})} \tag{3}$$

Note that given the data, the solution (i.e., estimates) is unique.

Doing some algebra, another way to write the estimator is

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

So, we can use either of these and the data to get β_1, the slope of the line.

Assumptions - SPRM Section 2.4

So far we only assumed linearity to obtain the estimates (by writing down the linear function and doing the calculus).

- We add 3 basic assumptions and one more about the variability of the error term or LS estimation to work well for us analytically:

  1. All observations are uncorrelated:

$$\text{cor}(Y_i, Y_j) = \text{cor}(\epsilon_i, \epsilon_j) = 0, \ \forall i \neq j$$

  2. All observations are equally informative:

$$\text{Var}(Y_i|X_i) = \text{Var}(\epsilon_i) = \sigma^2, \ \forall i$$

  3. There is no systematic bias in our model:

$$\text{E}(\epsilon_i) = 0, \ \forall i$$

  4. For testing and inference, we also add the normality assumption: all ϵ_i's are normally distributed.

Assumptions Lead to Properties of the Estimators

SPRM Section 2.8

Having the normality assumption on the error term helps us derive the distribution of these estimators:
We will focus on the distribution of $\hat{\beta}_1$ below, and $\hat{\beta}_0$ can be very similarly derived.

Recall E$(Y_i|X_i) = \beta_0 + \beta_1 X_i$, and Var$(Y_i|X_i)$ =Var$(\epsilon_i) = \sigma^2$.

Therefore, for $\hat{\beta}_1$,

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= \mathrm{E}(\sum k_i Y_i) = \sum k_i \mathrm{E}(Y_i) = \sum k_i(\beta_0 + \beta_1 X_i) \\
&= \beta_0 \sum k_i + \beta_1 \sum k_i X_i = \beta_1 \\
\mathrm{var}(\hat{\beta}_1) &= \mathrm{var}(\sum k_i Y_i) = \sum k_i^2 \mathrm{var}(Y_i) = \sigma^2 \sum k_i^2 \\
&= \frac{\sigma^2}{\sum (X_i - \bar{X})^2}
\end{aligned}$$

Assumptions Lead to Properties of the Estimators

Also, $\mathrm{E}(\hat{\beta}_0) = \mathrm{E}(\bar{Y} - \hat{\beta}_1\bar{X}) = (\beta_0 + \beta_1\bar{X}) - \beta_1\bar{X} = \beta_0$ and $\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$.

If $\epsilon_i \sim N(0, \sigma^2)$ and σ^2 is known, we can do inference about β_0 and β_1 with the normal distribution. However, σ^2 is not known and needs to be estimated from the data.

Variance Estimators for Coefficients

The variance expression we use in practice is

$$\text{var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}$$

where

$$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}.$$

The square root of $\hat{\sigma}^2$ (i.e., $\hat{\sigma}$) is also called root MSE is a critical quantity (we will discuss more later), as it reflects the general fit quality of the model as a 'average' deviation between actual Y's and predicted Y's. It is produced during OLS estimation and used to determine the variance estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$.

For inference, we use the t_{n-2}, a t-distribution with $n-2$ degrees of freedom (SPRM Section 2.9)

Summary So Far

From the previous we have:

- Least squares estimators are readily computed from X and Y

$$\Rightarrow \hat{\beta}_1 = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{\sum(X_i - \bar{X})^2},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- Estimators are unbiased (expected value of estimate equals population value), and 'best' in the sense of minimizing the criterion we are using (minimize squared distances from the line)
- Variability expressions readily available, can use standard methods (t tests) to evaluate hypotheses about β's

The Estimated Regression Model

Back to the VO_2MAX example, let's identify these key quantities:

```
regress vo2max t_seconds
```

```
      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =    179.29
       Model |  1815.55337         1  1815.55337   Prob > F        =    0.0000
    Residual |  182.272082        18  10.1262268   R-squared       =    0.9088
-------------+----------------------------------   Adj R-squared   =    0.9037
       Total |  1997.82545        19  105.148708   Root MSE        =    3.1822

------------------------------------------------------------------------------
      vo2max |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   t_seconds |   .0649794   .0048528    13.39   0.000      .054784    .0751748
       _cons |  -1.584694   3.506099    -0.45   0.657    -8.950735    5.781347
------------------------------------------------------------------------------
```

Inference: testing the importance of the predictor

SPRM 2.9

The fitted regression line, i.e., the fitted model, looks as follows:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$= -1.584694 + .0649794 * t_{seconds}$$

It is a convention to denote the predicted Y value by $\hat{Y}$, or "Y-hat"

The sampling distributions of the coefficients β_0 and β_1 allow us test hypotheses about these parameters.

The slope captures the relationship between Y and X. Testing whether **the slope equals to zero** is equivalent to testing whether there is a linear relationship between Y and X. This test is typically provided by default in any computer package.

Inference: testing the importance of the predictor

One can test more generally, for any specified β_{1_0}

$$H_0 : \beta_1 = \beta_{1_0}$$

$$H_1 : \beta_1 \neq \beta_{1_0}$$

at a given significance level α by the statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1_0}}{s.e.(\hat{\beta}_1)}$$

where $s.e.(\hat{\beta}_1)$ is the square root of the quantity we obtained earlier for $\mathrm{var}(\hat{\beta}_1)$ (above and eqn in SPRM, but computer output already tabulates this for us)

Inference: testing the importance of the predictor

The default test is whether there is zero slope, so this is the result for treadmill time as a predictor of VO_2MAX:

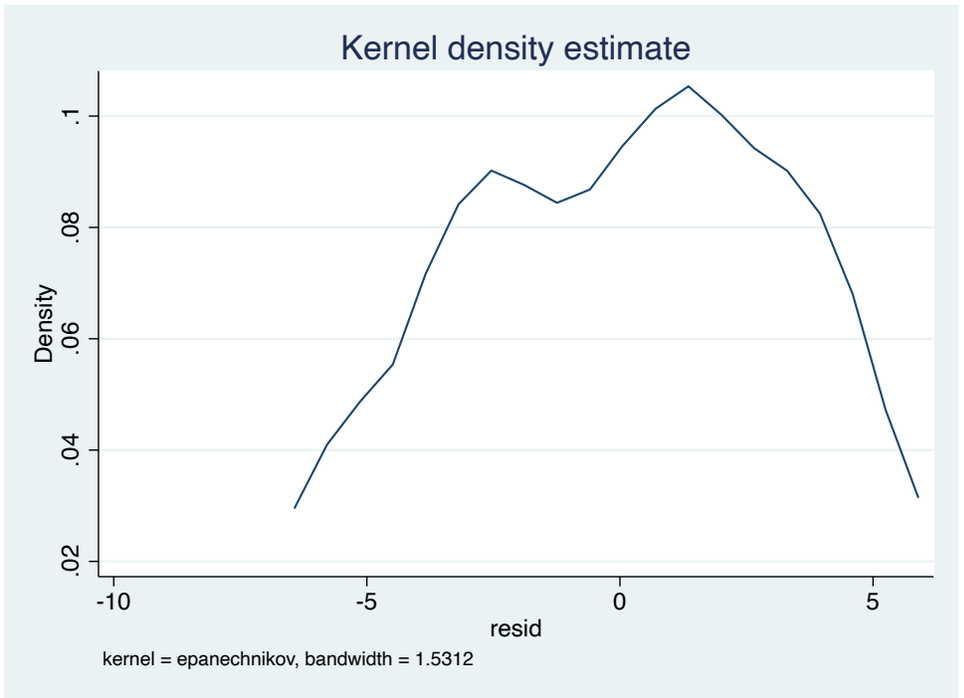
$$t = \frac{0.0649794 - 0}{0.0048528} = 13.39$$

We look at our t in relation to a $T_{n-2,\alpha/2}$ to see how 'extreme' it is, indicating whether the data support a slope of zero or some nonzero value. Here, the slope appears clearly different from zero

Checking Normality Assumption for the model

$Y_i - \hat{Y}_i$, called the residuals, should be normally distributed, mean zero - looks pretty good!

```
. gen resid = vo2max - yhat
. kdensity resid
```



More Simple Linear Regression (Chapter 2)

An Environmental Problem: Mercury levels in fish tissue for largemouth bass in the Wacamaw and Lumber Rivers

Rivers in North Carolina contain small concentrations of mercury, which can accumulate in fish over their lifetimes because mercury cannot be excreted.

Directly measuring the mercury concentration in the water is impossible since it is almost always below detectable limits when dispersed. However, the concentration of mercury in fish tissue can be obtained by catching fish and conducting laboratory analysis.

Mercury levels in largemouth bass

A study was conducted to investigate mercury levels in tissues of largemouth bass. At several stations along each river, a sample of fish were caught, weighed, and measured. In addition, a tissue sample was analyzed for concentration of mercury (parts per million).

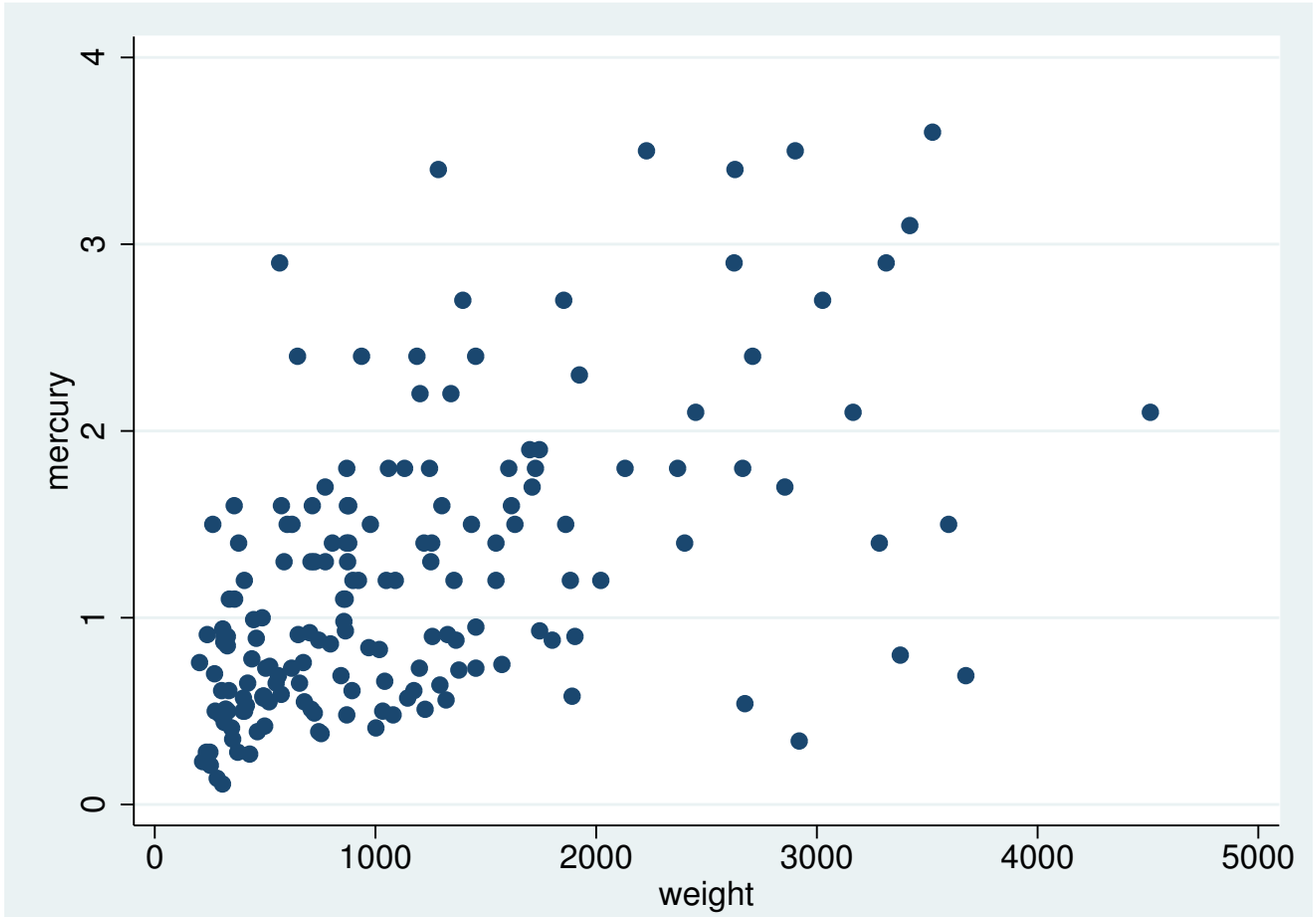
Questions:

1. Is there a relationship between mercury concentration and size (weight and/or length) of a fish?
2. A concentration over 1 part per million is considered unsafe for human consumption. In light of this, what recommendations can you make for fish caught from these rivers?

(data is on Canvas, under files - fish.dta)

To address question 1, we first look at the data. Here is a scatterplot of mercury concentration versus weight (grams).

```
. use fish.dta
. scatter mercury weight
```



Taking a step back to look at the raw data . . .

```
. sum weight, detail
                            weight
-------------------------------------------------------------
      Percentiles      Smallest
 1%           217            203
 5%           274            217
10%           320            234      Obs                 171
25%           491            238      Sum of Wgt.         171

50%           873                     Mean           1147.912
                        Largest       Std. Dev.      875.5318
75%          1455           3524
90%          2625           3597      Variance       766555.9
95%          3164           3675      Skewness       1.390022
99%          3675           4511      Kurtosis       4.518831
```

```
. sum mercury, detail
                            mercury
-------------------------------------------------------------
      Percentiles      Smallest
 1%           .14            .11
 5%           .34            .14
10%           .48            .21       Obs                 171
25%           .59            .23       Sum of Wgt.         171

50%           .93                      Mean           1.191754
                        Largest        Std. Dev.      .7616633
75%           1.6            3.4
90%           2.3            3.5       Variance       .580131
95%           2.9            3.5       Skewness       1.169084
99%           3.5            3.6       Kurtosis       4.01653
```

Checking the correlation between the two variables, as a measure of whether two variables are linearly related, and if so the direction and strength of the relationship.

```
. correlate mercury weight
(obs=171)
             |  mercury   weight
-------------+------------------
     mercury |   1.0000
      weight |   0.5538   1.0000

. corr mercury weight, cov
(obs=171)

             |  mercury   weight
-------------+------------------
     mercury |   .580131
      weight |  369.333   766556
```

Mercury levels in largemouth bass

The regression analysis

```
. regress mercury weight
```

```
      Source |       SS       df       MS              Number of obs =     171
-------------+------------------------------           F(  1,   169) =   74.77
       Model |  30.2510497     1  30.2510497            Prob > F      =  0.0000
    Residual |  68.3712259   169  .404563467           R-squared     =  0.3067
-------------+------------------------------           Adj R-squared =  0.3026
       Total |  98.6222756   170  .580131033           Root MSE      =  .63605

------------------------------------------------------------------------------
     mercury |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0004818   .0000557     8.65   0.000     .0003718    .0005918
       _cons |   .6386813   .0803536     7.95   0.000     .4800552    .7973073
------------------------------------------------------------------------------
```

With the regression model, we could predict the expected value (fitted) of mercury for each observation.

```
. predict yhat
(option xb assumed; fitted values)
```

```
. twoway (scatter mercury weight) (line yhat weight, sort clcolor(red))
```

This line results from the estimated intercept $\beta_0 = 0.6387$ and slope $\beta_1 = 0.00048$.

Mercury levels in largemouth bass

- Intercept denotes the average mercury level for a fish that has weight of 0. It is meaningless in many situations as you can see.

- Slope denotes the average change in mercury when weight goes up by 1 unit (in this case gram). So for two fish, one weighing x and the other weighing $(x + 1)$ grams, you would expect their mercury levels to differ by 0.00048 ppm.

So, the expectation (model line) is:

$$\mathrm{E}(\mathrm{mercury}|\mathrm{weight}) = 0.6387 + 0.00048 \times \mathrm{weight}$$

This model lets us predict how much mercury a fish of any given weight will have on average (i.e., among fish of that given weight, the average observed mercury level will be equal to the fitted value from the model above).

Hypothesis Testing

Is the positive linear relationship between these two variables significantly different from zero (say, at $p < .05$)? This is the default test carried out by Stata or R. From the output, the t-statistic is

$$t = \frac{\hat{\beta}_1 - \beta_{1_{H0}}}{s.e.(\hat{\beta}_1)} = \frac{(0.00048 - 0)}{.0000557} = 8.65$$

The associated p-value (< 0.0005 indicates weight is a statistically significant predictor. Recall that this p-value is from the t-test with $n - 2 = 169$ degrees of freedom. Note: A t-distribution with df this large is essentially N(0,1).

Alternatively, Examine the 95% confidence interval: it does not include 0, which tells us that at 5% significance level we would reject the null hypothesis of no relationship between mercury and weight. The mercury-weight relationship is significant at 5%

Hypothesis Testing
Confidence Intervals on β - Not mentioned (?) in SPRM

As $\hat{\beta}$ comes from a $t-$distribution . . .

The $1-\alpha/2$ % confidence interval on β_k (k = 0 or 1) is defined as

$$(\hat{\beta}_k - t_{n-2,\alpha/2} \times se(\hat{\beta}_k), \hat{\beta}_k + t_{n-2,\alpha/2} \times se(\hat{\beta}_k))$$

Stata and R provide the 95% interval by default. Standard errors of the estimates are given or refer to the formulas shown earlier. For n as large as we have here, the t critical value is 1.974, so this interval, based on the normal (N(0,1)) distribution, is nearly identical. For β_1:

$$(\hat{\beta}_k - 1.96 \times se(\hat{\beta}_k), \hat{\beta}_k + 1.96 \times se(\hat{\beta}_k))$$

$$= (.00037263, .00059097)$$

Hypothesis Testing

An alternative test that is useful for testing hypotheses about multiple β's jointly (more useful later) is provided by the F-test
We can test a single hypothesis $H_0 : \beta_1 = 0$ in STATA:

```
. test weight
 ( 1)  weight = 0
       F(  1,   169) =   74.77
            Prob > F =    0.0000
```

NOTE: **only in SLR**, the F-test is equivalent to the t-test we just did, and this F-test is also the overall F test in the upper right corner of the regression table. The overall F-test tests all predictor β's (not the intercept) equalling zero in the model. In fact mathematically, the F-statistic is the square of t-statistic for β_1 in the SLR above. SPRM mention the F-test in 2.9.2 in relation to decomposing sources of variation in Y, and we will return to it soon.

Hypothesis Testing

For now, this test provides a handy 'calculator' for testing some hypothesis we are interested in about β_1 and/or β_0 (we could of course just do the t-test by hand, writing out the statistic and looking up the p-value)

For example, we want to test whether the incremental increase in mercury per gram is equal to a specific value, for example, $H_0 : \beta_1 = 0.0003$, which might be a threshold value for material risk.

```
. test weight = .0003
 ( 1)  weight = .0003
       F(  1,   169) =   10.65
            Prob > F =    0.0013
```

NOTE: we could test this hypothesis (at $\alpha = .05$) using the CI for the weight coefficient in the table as well. This value is outside the CI, smaller than the loer bound, and so exceeds the threshold

Hypothesis Testing

We can also use R as a 'calculator' to compute any t-test we need

```
> bassreg = lm(bass$mercury ~ bass$weight)
> summary(bassreg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71933 -0.41388 -0.09201  0.33615  2.14220

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.387e-01  8.035e-02   7.948 2.56e-13 ***
bass$weight 4.818e-04  5.572e-05   8.647 3.93e-15 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.6361 on 169 degrees of freedom
Multiple R-squared: 0.3067,Adjusted R-squared: 0.3026
F-statistic: 74.77 on 1 and 169 DF,  p-value: 3.929e-15
```

```
> # test a specific hypothesis - whether estimate is consistent w/ 0.0003
> t_threshold = (4.818e-04 - .0003)/5.572e-05
>
> pt(t_threshold,169,lower.tail=F)
[1] 0.0006675683
 2*0.0006675683
[1] 0.001335137
> # same p-value as Stata
>
```

Actually, the one-sided p-value is perhaps more relevant here. We are interested in whether mercury concentration is *increased* per weight unit above this threshold

Note that one-side test at the 5% significance level has a same critical value as a two-sided test at the 10% significance level.

Hypothesis Testing - Intercept

We could perform similar specific tests on β_0 or even both coefficients concurrently. To test $H_0 : \beta_0 = 0.50$:

```
. test   _cons = .5

 ( 1)  _cons = .5

       F(  1,   169) =    2.98
            Prob > F =    0.0862
```

We could conclude from this that for very small fish (weight 'zero'), the level is not above 0.50. But since this is already one-half the EPA threshold, many fish of any catchable size will likely have expected mercury value above the threshold

Measuring the Quality of Fit

We have fit a line to the data and generated predicted values. How well did we do?

- We have $\hat{Y}$, and can just compare to corresponding Y or $Y - \hat{Y}$ (say, in a scatterplot) to note the discrepancies. Note in SLR, since $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, the plots of X being the predictor or $\hat{Y}$ being the predictor will look about the same.
- Clearly, $Y - \hat{Y}$ being 'small' is much better than large, and the sum $\sum(\hat{Y}_i - Y_i)^2$, has been minimized by the method of estimation, so is the smallest it can be for these X and Y values.
- We do have a measure of 'fit' in terms of the slope of the line and its statistical significance (difference from zero or no predictive value). We also have the correlation coefficient.

Coefficient of Determination or R^2

SPRM 2.92, 2.10

A key summary quantity in linear regression models is the *coefficient of determination*, or more commonly, just R^2

Going back to regression function estimation, there are three quantities that appear in the 'ANOVA Table' at the top of the analysis run via computer. One of these was computed in formulating the β's. The other two are functions of $\hat{Y}$ and Y:

Source	SS	df	MS
Model	30.2510497	1	30.2510497
Residual	68.3712259	169	.404563467
Total	98.6222756	170	.580131033

Coefficient of Determination or R^2

SPRM 2.9.2

These are:

1. Sum of Squares for Regression (or Model): $SSR = \sum(\hat{Y}_i - \bar{Y})^2$ - variation in a predicted Y (which is the mean of Y and a given X) around the overall mean of Y
2. Sum of Squares for Error (or Residual): $SSE = \sum(\hat{Y}_i - Y_i)^2$ - variation between predicted and observed Y_i's
3. Sum of Squares Total: $SST = \sum(Y_i - \bar{Y})^2$ - variation of individual Y_i's around the overall mean of Y

Then note that

$$SST = SSR + SSE$$

The R^2

The R^2 measures the fraction of SSR or 'model' variation relative to total variation (SST or simple variation of Y around its mean, ignoring X).

$$R^2 = \frac{SSR}{SST}$$

sometimes also written as

$$R^2 = 1 - \frac{SSE}{SST}$$

Heuristically, if X is a good predictor, $\hat{Y}_i$'s *should* differ a lot from the overall mean, because there are actually different $\bar{Y}$'s in effect at each X. If X is not a good predictor, then $\bar{Y}$ at X is about the same as $\bar{Y}$ overall.

The R^2

In the mercury data:

```
. regress mercury weight

      Source |       SS       df       MS              Number of obs =     171
-------------+------------------------------           F(  1,   169) =   74.77
       Model |  30.2510497     1  30.2510497            Prob > F      =  0.0000
    Residual |  68.3712259   169  .404563467           R-squared     =  0.3067
-------------+------------------------------           Adj R-squared =  0.3026
       Total |  98.6222756   170  .580131033           Root MSE      =  .63605
```

SSR $=$ 30.2510497

SSE $=$ 68.3712259

SST $=$ 98.6222756

From the regression table, $R^2 = 30.25/98.62 = .3067$, as given on the right.

Coefficient of Determination

Notes:

- In this analysis, $R^2 = 0.3067$, so we say that 30% of the variation in mercury level is explained by variation in fish weight

- In any SLR (one predictor variable), R^2 is the square of the correlation coefficient between X and Y,
  $R^2 = \text{cor}^2(Y, \hat{Y}) = \text{cor}^2(Y, \hat{\beta}_0 + \hat{\beta}_1 X) = \text{cor}^2(Y, X) = 0.5538^2$.

- R^2 has range [0, 1]. A 'good' R^2 is rather context-specific. We typically do not perform hypothesis tests on R^2 directly. It will be small when r is near zero, but may also be small even in the presence of a statistically significant β

Using the Model for Predictions

SPRM Section 2.12 and 2.14

- Q2: A concentration over 1 part per million is considered unsafe for human consumption. In light of this, what recommendations can you make for fish caught from these rivers?
- Given the model, we can predict the mean mercury level for a fish of a given weight, for example 1kg. We place a confidence interval on this prediction

We can also say something about the actual mercury level of a fish of 1kg based on our model – that is, we can find not only the expected or mean value, but also its forecast interval:

A forecast interval can be thought of as the middle 95% of all mercury weights for fish of 1kg. This interval takes into account the uncertainty in the prediction of the mean as well as the actual randomness of mercury weight around the mean at that weight.

Using the Model for Predictions

For predicting the mean at some X_0, we have as before

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

The standard error for this predicted mean response is estimated by (see also eqn 2.24)

$$s.e.(\hat{\mu}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

SPRM refers to $\hat{\sigma}$, the root MSE of the model, as $S_{Y|X}$

Using the Model for Forecasting

For "predicting" the actual value at x_0, i.e., forecasting the value at x_0, we have

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Looks the same (?). It is, because the regression model always predicts the mean of Y given X and the expectation of the error term is zero. The difference is in the standard error, i.e., the width of the interval. The standard error for forecasting (i.e., prediction of one observation) is estimated by (see also eqn 2.25)

$$s.e.(\hat{y}_0|x_0) = \hat{\sigma}\sqrt{\mathbf{1}+\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Since $y_0 = E(y_0)+\epsilon$, this standard error estimate for forecasting is larger, and even when n gets very large, it will not shrink much.

Confidence intervals on Predictions

Intervals are then constructed in the usual manner:

on the mean (eqn 2.22)

$$\hat{\mu}_0 \pm t_{n-2,\alpha/2} \times s.e.(\hat{\mu}_0),$$

OR

on an individual value (eqn 2.23)

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \times s.e.(\hat{y}_0),$$

The standard errors can be computed in Stata or R, either by obtaining the necessary components and computing, or via shortcuts that the programs provide. The predictions with confidence intervals are then obtained.

Confusing Nomenclature

- SPRM refer to the interval on the mean as simply that, or the standard error/CI on the conditional mean of y at a given x. The refer to the interval on an individual y value as the *prediction interval*. They refer to it as a confidence interval on the regression line, which it indeed is.

- Other texts (and `STATA`) refer to the interval on the mean as the prediction interval, and the interval on the individual values as the *forecast* interval. This name may reflect the fact that out of sample (new observation) data may be used to forecast an outcome.

Using the Model for Predictions/Forecasting

In Stata, these intervals are available after running *regress* . Refer to `help regress postestimation` for a complete description of pre-programmed shortcut for many post-estimation predictions in Stata.

stdp calculates the standard error of the prediction, $s.e.(\hat{\mu}|x_j)$, running over the range of X values in the data. Presumably, if you were to predict with a novel x_0, it would be in this range.

stdf calculates the standard error of the forecast, $s.e.(\hat{y}_j|x_j)$, which is the standard error of the point prediction for 1 observation.

Predictions/Forecasting

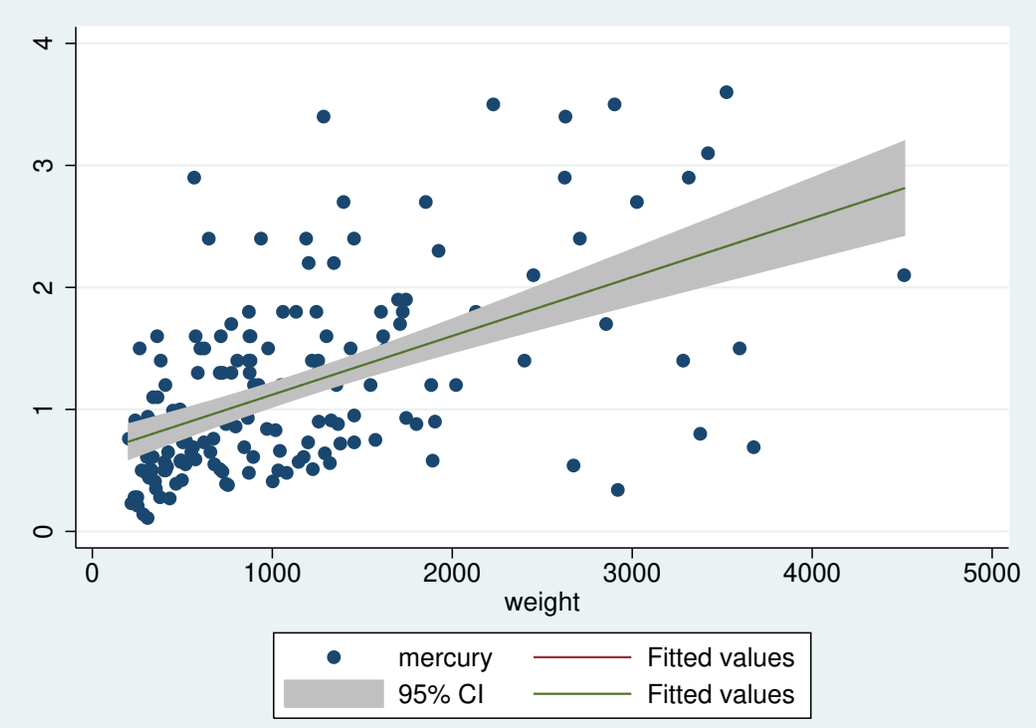
```
. quiet regress mercury weight
. predict yhat
. * ask for standard errors for prediction and forecast, name them to
. * new variables (stdp and stdf)
. predict spred, stdp
. predict sfor, stdf
.* listing observation 1-5
. list spred sfor in 1/5
     +---------------------+
     |    spred       sfor |
     |---------------------|
  1. | .0551914   .6384431 |
  2. | .0628405   .6391497 |
  3. | .1068311   .6449623 |
  4. | .0487234   .6379165 |
  5. | .0495763   .6379822 |

* get Prediction intervals for each set of X (from each observation)
. gen PI_lower = yhat - invttail(171-2, 0.025)*spred
. gen PI_upper = yhat + invttail(171-2, 0.025)*spred
 list weight yhat PI_lower PI_upper in 1/10, clean
```

	weight	yhat	PI_lower	PI_upper
1.	1616	1.417283	1.308339	1.526227
2.	1862	1.535807	1.411765	1.65985
3.	2855	2.014243	1.803365	2.22512
4.	1199	1.216369	1.120192	1.312546
5.	1320	1.274668	1.176807	1.372528
6.	1225	1.228896	1.13251	1.325282
7.	870	1.057854	.9570935	1.158615
8.	1455	1.339712	1.237932	1.441491
9.	1220	1.226487	1.130147	1.322826
10.	1033	1.136389	1.039548	1.23323

Predictions/Forecasting

In above, we computed the bounds by 'hand'. Stata does this for you, using plot commands (see menu options under Graphics, two-way plots) Fitted values with confidence interval on the mean:

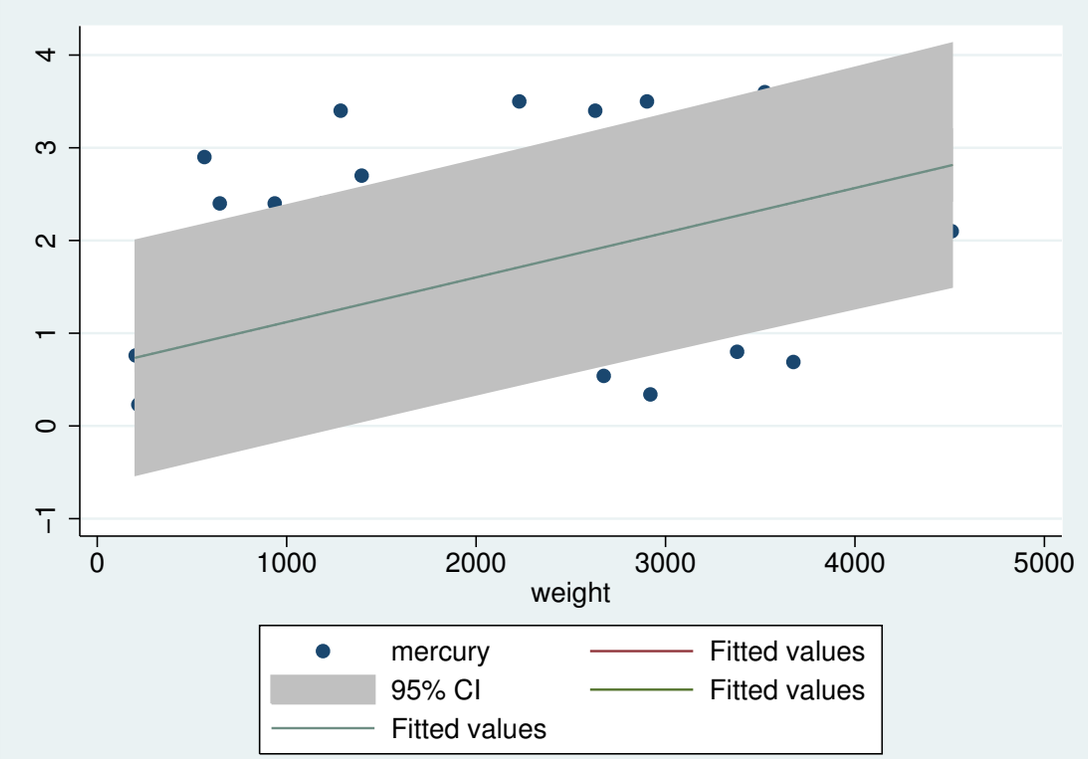
```
. twoway (scatter mercury weight) (lfit mercury weight)
 (lfitci mercury weight)
```



Predictions/Forecasting

The fitted value and the 95% highest density interval individual predicted values (forecast interval)

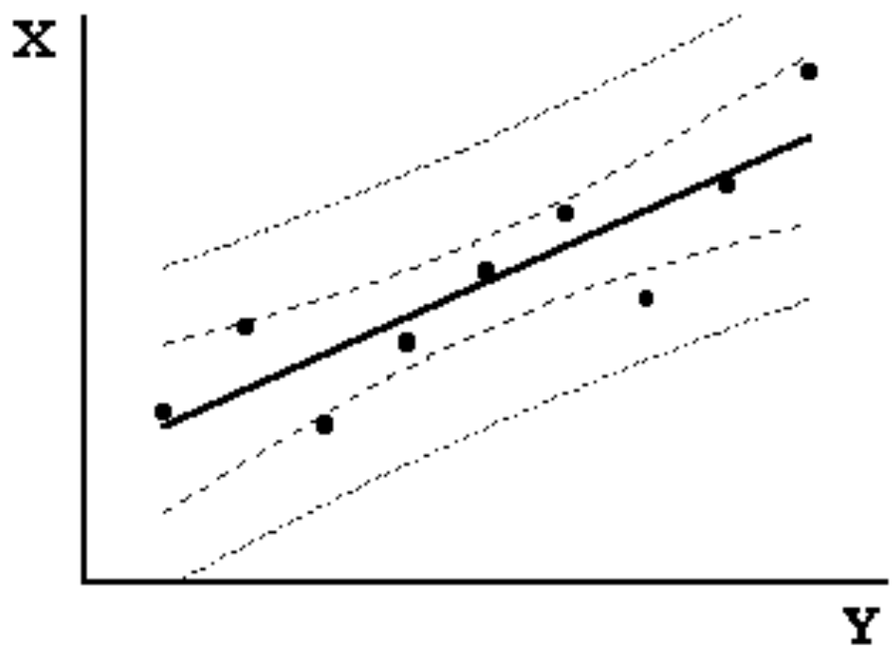
```
. twoway (scatter mercury weight) (lfit mercury weight)
(lfitci mercury weight, stdf)
```



Predictions/Forecasting

Notes:

- Forecast interval is technically NOT a confidence interval (not for testing purposes). This is a 95% highest probability density interval for the new forecast variable. For a new observation with weight $X = 3000$ (or some other number), this figure plots the interval that with 95% probability the mercury level for the new observation will fall in.
- Note how much wider the forecast interval is than the confidence interval on the predicted mean. There is much less precision in individual predictions
- The confidence interval on the mean is at a minimum at $(\bar{Y}, \bar{X})$ and 'flairs' as one moves away from these values. This pattern and is a function of how the intervals are formulated and holds in general



- Neither interval is recommended for predicting outside the observed range of X

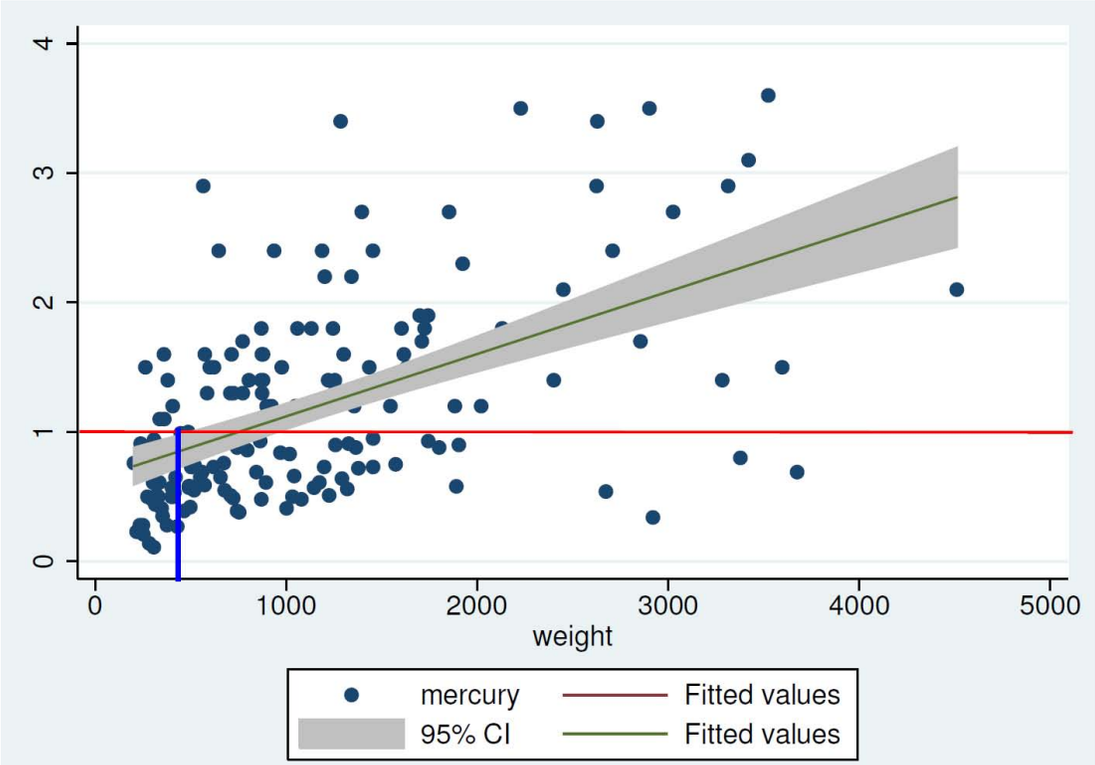
See the text for discussion of these intervals in relation to the triglyceride data

Now back to our original questions of interest.

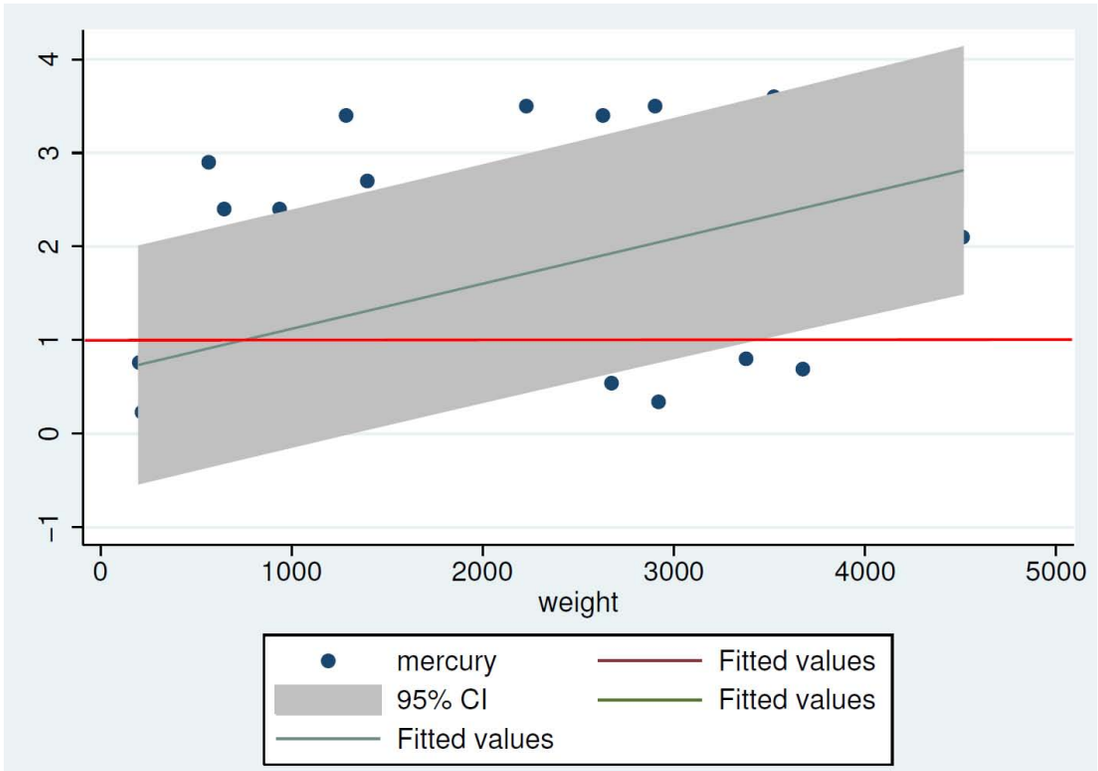
Questions:

1. Is there a relationship between mercury concentration and size (weight and/or length) of a fish?
2. A concentration over 1 part per million is considered unsafe for human consumption. In light of this, what recommendations can you make for fish caught from these rivers?

For a repeated consumer or a group of consumers (or both) concerned about the average mercury level:



For a one-time consumer concerned with the consumption of one fish:



Know the Bass of Illinois

(e) largemouth

(f) smallmouth

Summary: Simple Linear Regression

We have discussed simple linear regression, highlighting the following key components:

1. Relationship to correlation and empirical plot (scatterplot)
2. How β's are obtained and what additional properties we assume to make things work nicely
3. Inference on β's, confidence intervals on β's,
4. R^2 as a measure of explained variation
5. Prediction, interpretation of model results

All of this will naturally extend to multiple linear regression (Ch. 3)