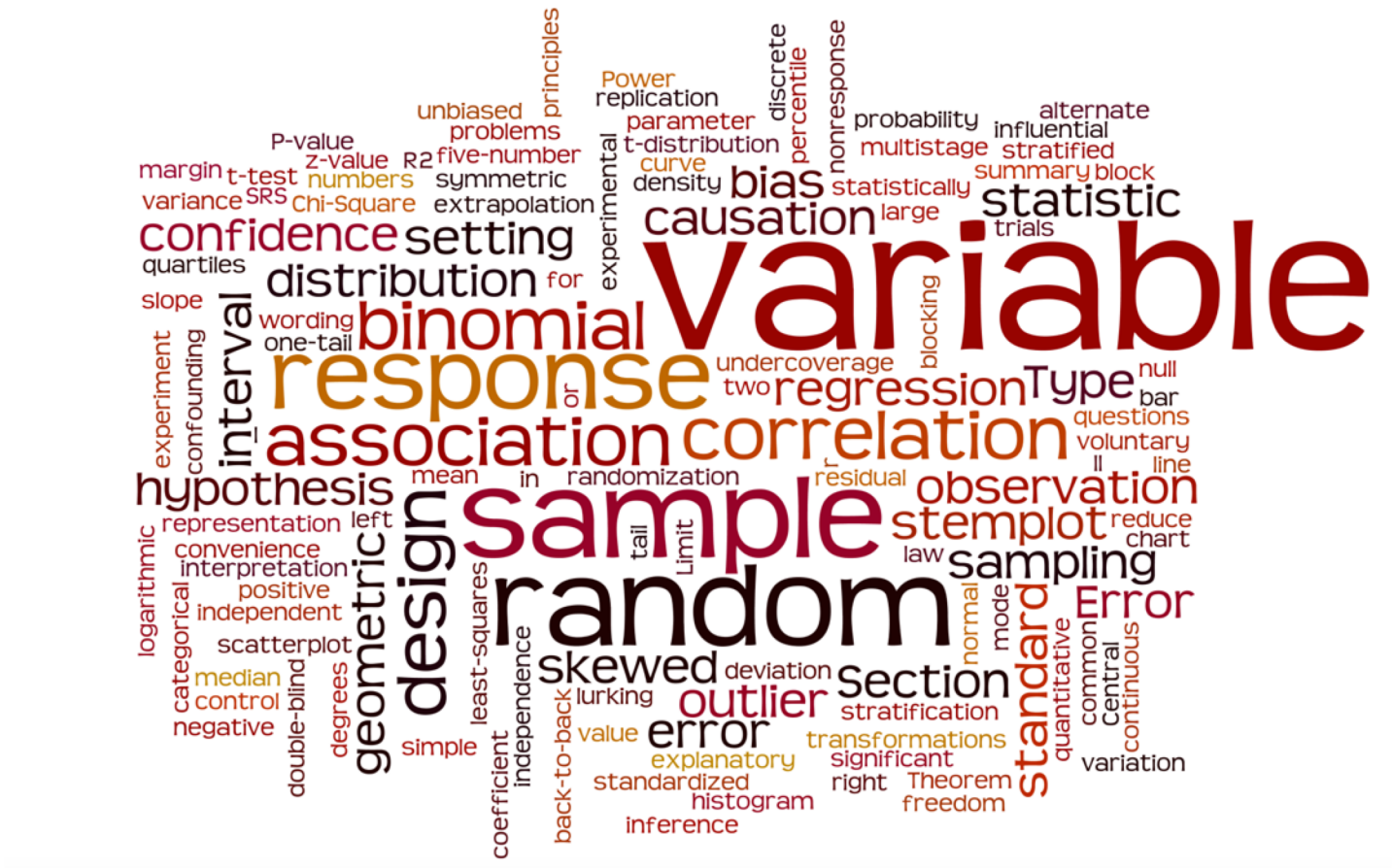
Introduction to Quantitative Methods



Kristin Eccles

Outline

- Review: Covariance/ Correlation
- Regression models
- Equation of a line (review)
- Fitting a line
- Residuals
- Testing slope
- Goodness of fit
- Testing model
- Assumptions
- Residuals – analysis

Can we trust the numbers? (January 26th, 2018)

https://www.npr.org/programs/ted-radio-hour/archive?date=1-31-2018

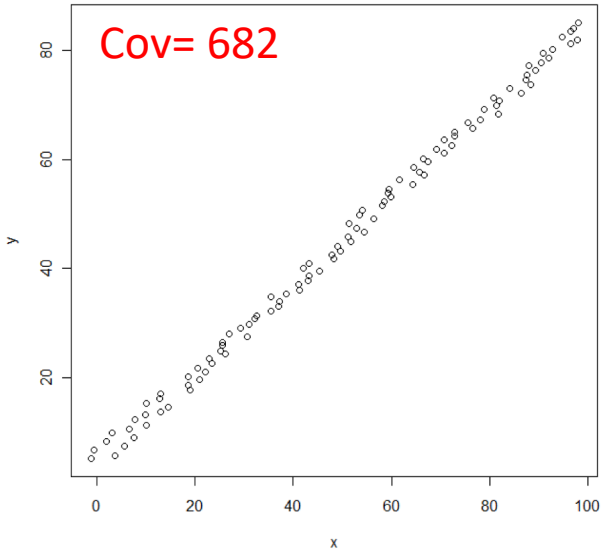
"The people who have access to the data and can decide how that data is presented, have a tremendous amount of power"

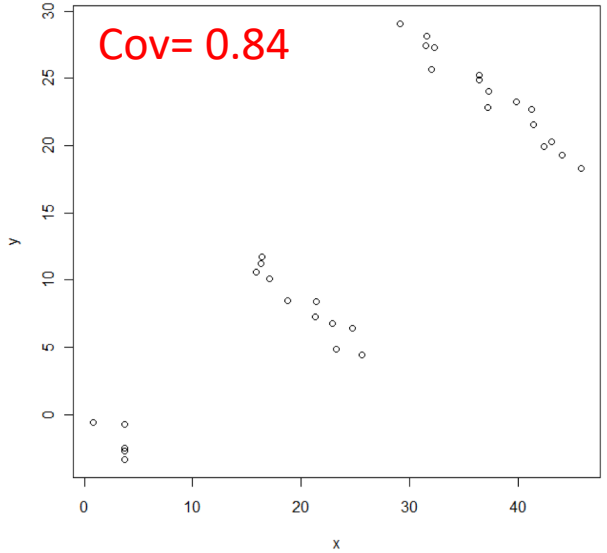
Review: Covariance

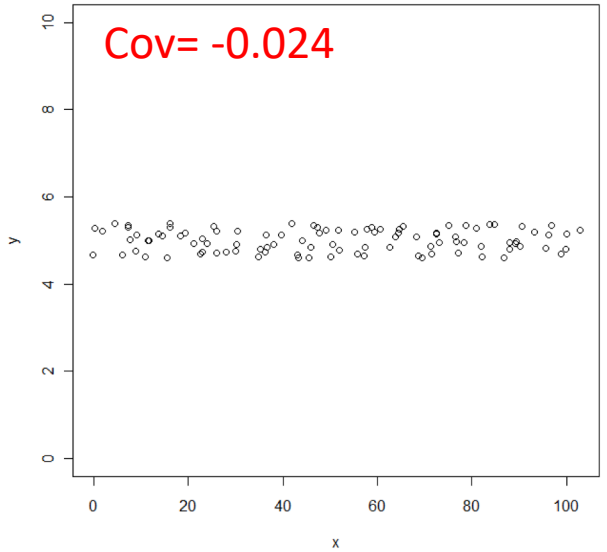
- How much to the variables co-vary?
- The higher the covariance, the more this is the case
- Covariance changes with the magnitude of x and y
  - Just like variance
- Need a way to standardize this measure of association

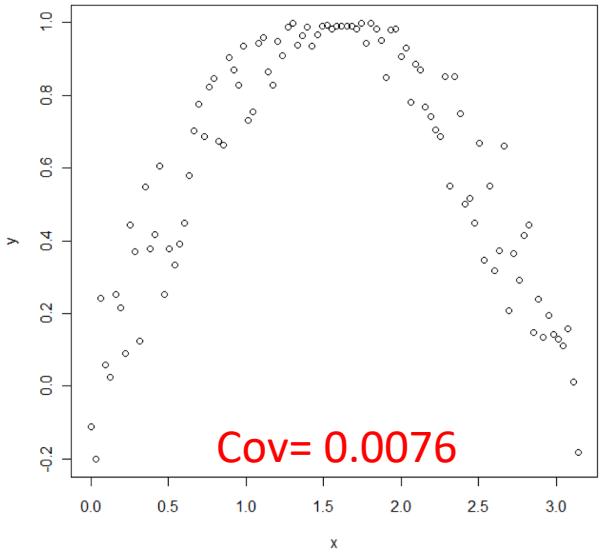
$$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

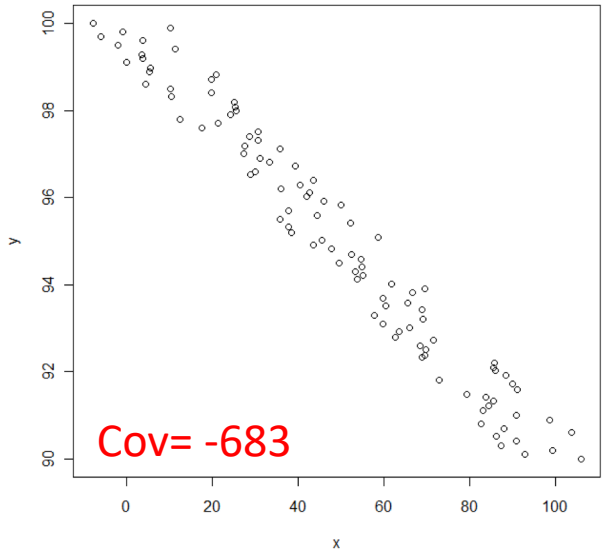
Relationship: Direction, form, strength?

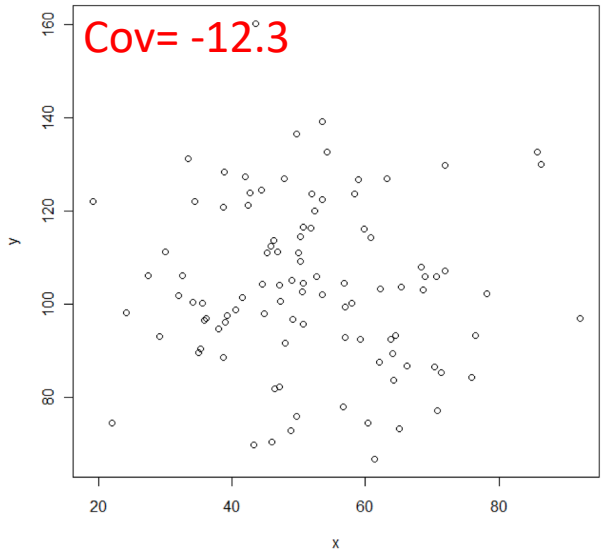












Reviews: Pearson's correlation coefficient

- AKA Pearson's r or Pearson's product moment correlation coefficient
- Measures the amount of correlation between 2 variables
- Large r (- or +) means a strong linear association
- r is a number from -1 to 1
  - -1: perfect negative correlation
  - 0: no correlation
  - 1: perfect positive correlation
- Formula: same as the covariance, but use z-scores for x and y

$$r = \frac{\sum_{i=1}^{n} z_{xi} \cdot z_{yi}}{n-1}$$

$$z = \frac{X - \bar{X}}{s}$$

- Note that correlation is only one type of association
  - it is possible to have r=0 and yet have a strong relationship

Relational statistics

- Explore the relationship between variables
- We know how to tell if two variables are associated
  - Correlation
- Does one explain the other?
- If so, how much (is it significant)?
- Can we predict y, if we know x?

- Easy to examine this, if you have a linear relationship
- Known as simple linear regression

Models

- Statistical models can capture the essence of a relationship in order to make **predictions**
- Remember:
  - you are using data from the real world (i.e., uncertainty, error, randomness, etc.)
  - **you have a sample, but you are predicting for a population**
- Models combine parameters (invariant) and variables (which vary) to make predictions
- 'All models are wrong, but some are useful' (George Box)

How to represent a relationship

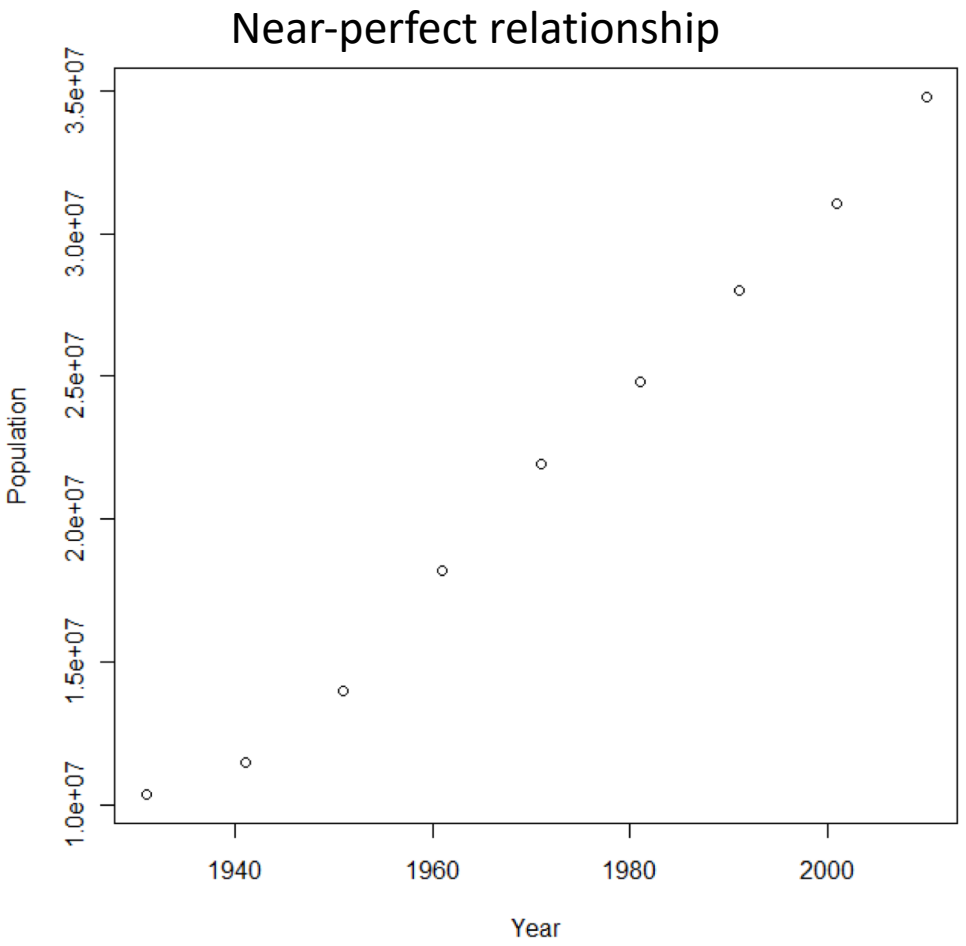


Fig 1. Canada's population from 1921 to present. Source: Statistics Canada.

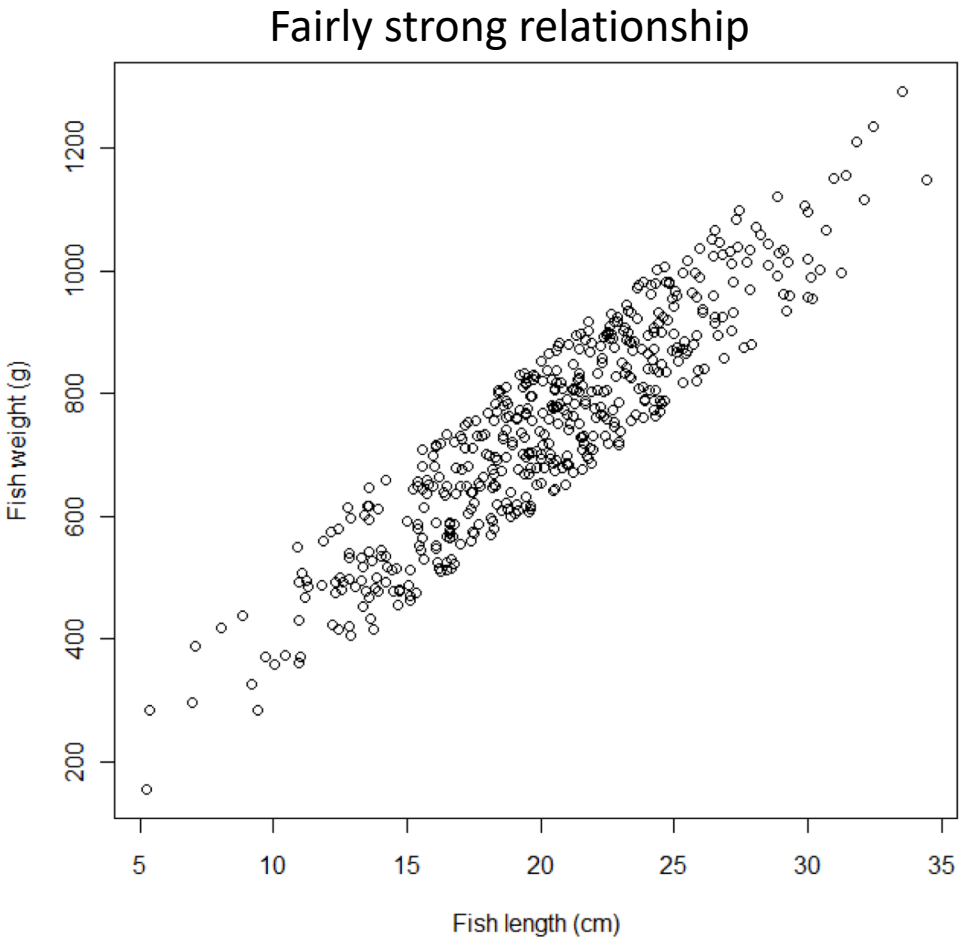


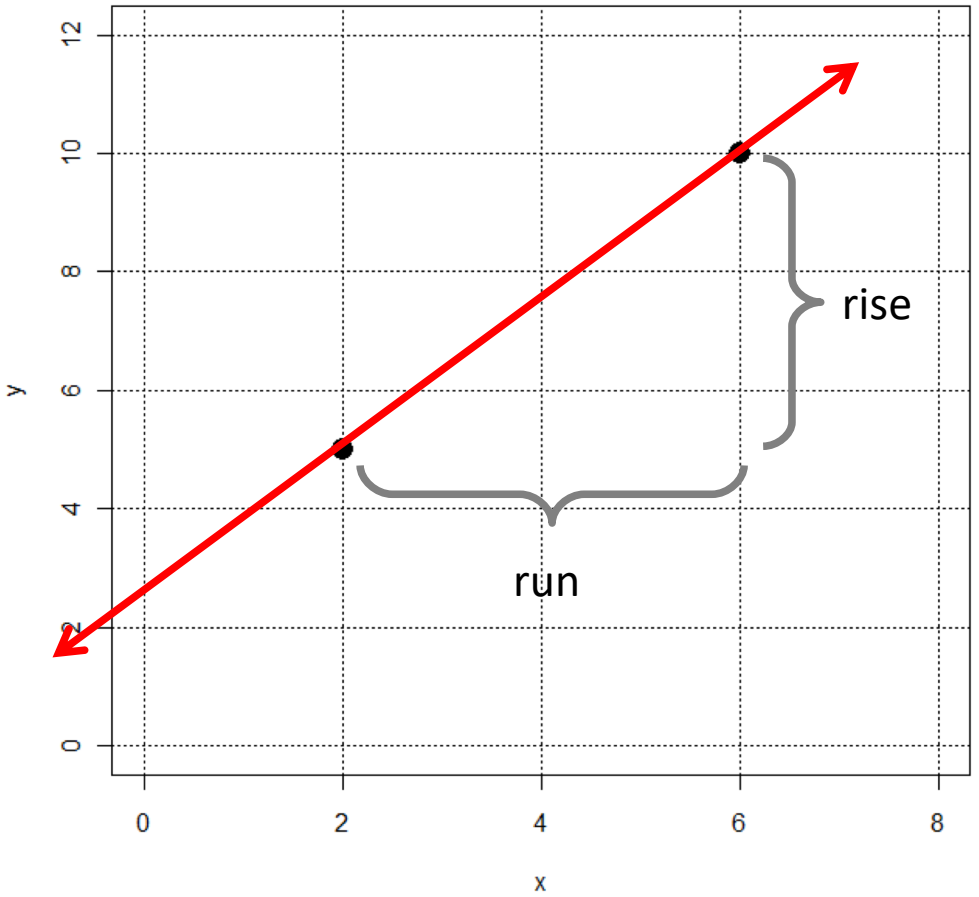
Fig 2. Fish length vs. weight for Meech Lake.

Algebra – equation of a line

You can describe a straight line on a 2-dimensional (x,y) graph by using the equation:

$$y = mx + b$$

- **m= slope** (aka gradient) is *rise* over *run*
  - rise = $(y_2-y_1) = 10-5$
  - run = $(x_2-x_1) = 6-2$
  - rise/run = 5/4 = 1.25
- **b= intercept** (aka y-intercept) is the value of y when x is 0
  - $(y-y_1) = m(x-x_1)$
  - $y-5 = 1.25(0-2)$
  - $y = 2.5$



Slope

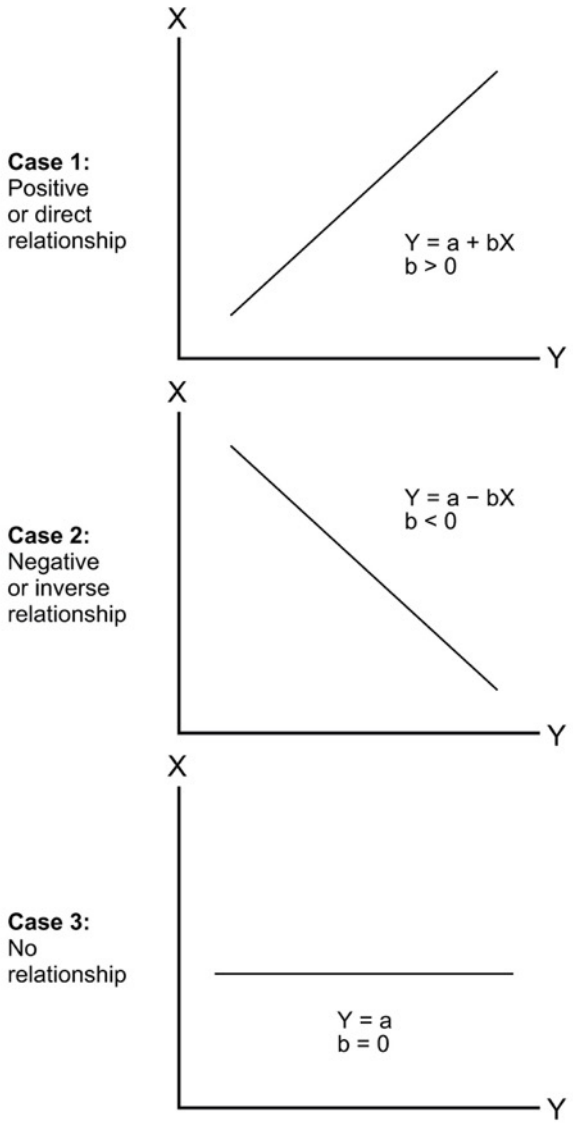
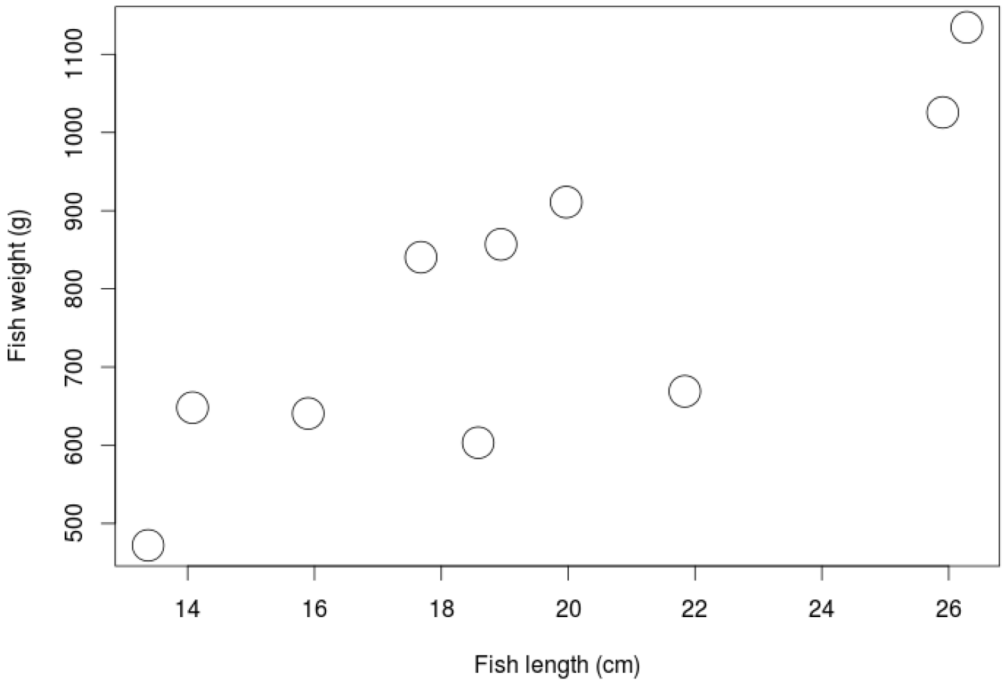


FIGURE 17.3
Interpretation of Slope in Simple Linear Regression

Fit a line

- Make a model by fitting a line to the data that captures the overall pattern:
  - $y = b_0 + b_1x$

- 2 variables:
  - x - independent / predictor
  - y - dependent / response
- 2 parameters:
  - b_0 = y-intercept
  - b_1 = slope

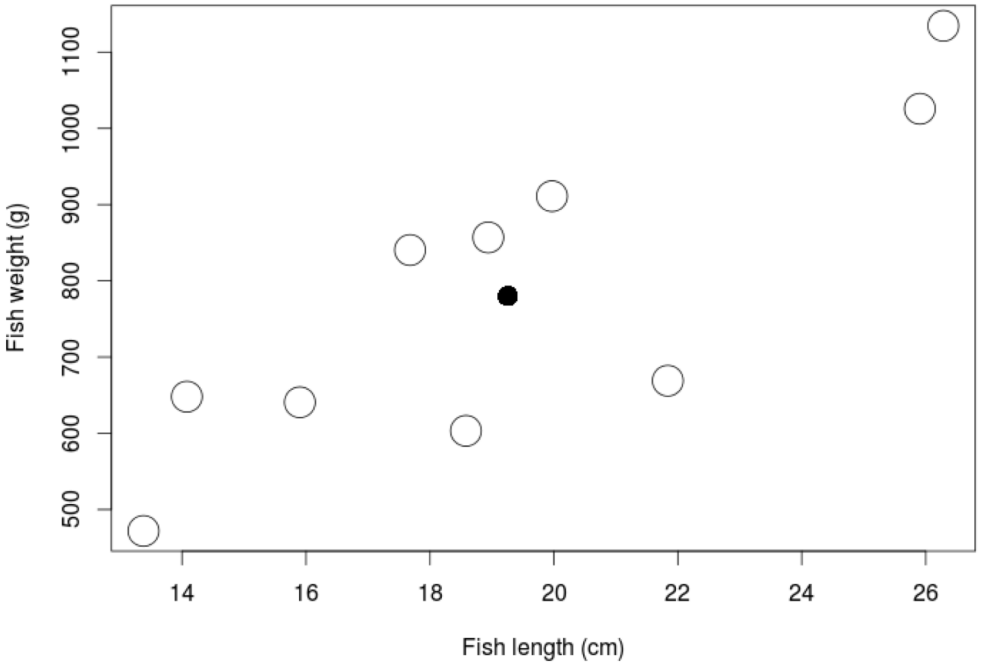


Residuals

- Residuals are the difference between the *observed* (y) and the *predicted* ($\hat{y}$)
  - note the hat means a predicted variable
  - $y_i - \hat{y}_i$
- Can be positive (under-predicted) or negative (over-predicted)
- Can sum all the residuals
  - $\Sigma(y_i - \hat{y}_i)$
- Residuals are also known as **error**
  - But not error as in a mistake – rather, this is *unexplained variance*
  - Models that are perfect will explain how y varies perfectly
  - Models that are ok will explain most of how y varies
  - Models that are bad, will explain none of the variance in y

Minimizing residuals

- Want the residuals to:
  - add to zero
  - be as small as possible
- Place the line so it passes through the point $(\bar{x}, \bar{y})$
  - residuals will now add to zero!
- Rotate until you get the smallest residuals overall
  - need to square the residuals to calculate the sum of squares of the error (SSE)
  - minimize SSE!



$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

More sums of squares

SSY – *represents* the variance of y

$$SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

SSX – *represents* the variance of x

$$SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

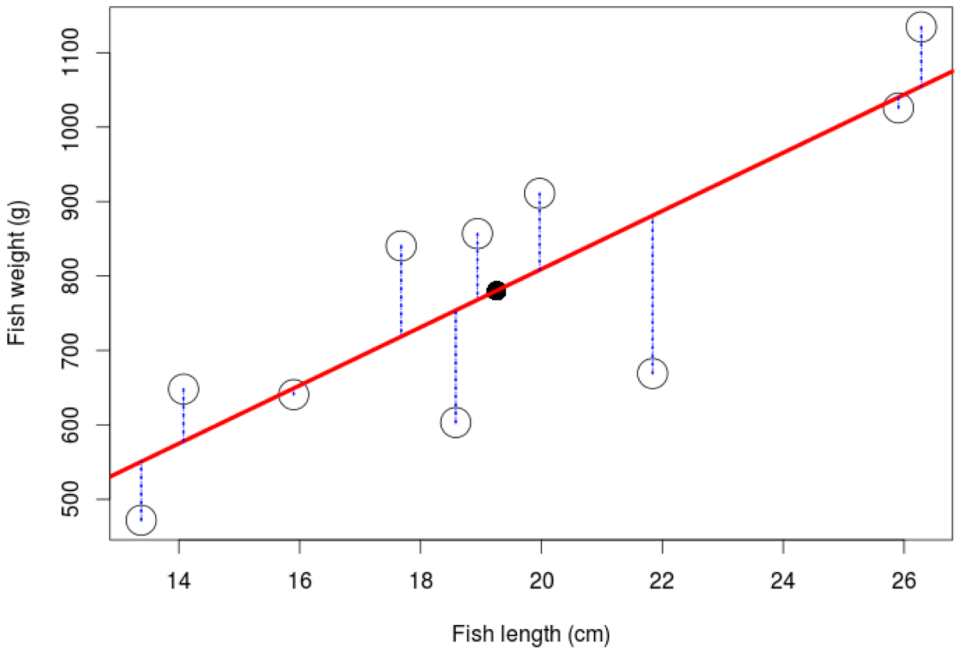
SSXY – *represents* the co-variance of x and y

$$SSXY = \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$$

Minimizing residuals

- The slope that minimizes the residuals gives the least squares (of the error)
- Linear regression (aka ordinary least squares regression)
- Slope of the regression line is:

$$b_1 = \frac{SSXY}{SSX}$$



- Solve for the intercept b_0 by using the equation and a point on the line:

$$(\bar{y} = b_0 + b_1\bar{x})$$

OLS Objective

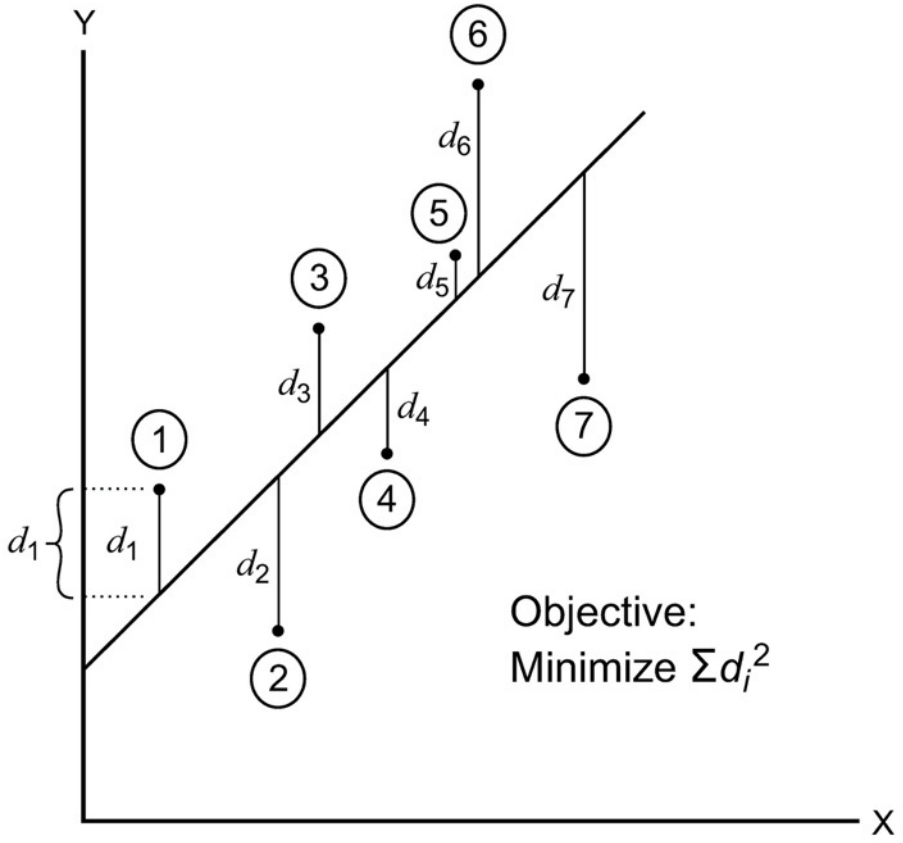


FIGURE 17.1
The Objective of Least-squares Regression

Is it significant?

- Now have a **line of best fit** calculated from a *sample*
- $y = b_0 + b_1x$

- Can you model the *population*?
  - $y = \beta_0 + \beta_1x + \varepsilon$
  - where β_0 is the intercept, β_1 is the slope and ε (Greek letter epsilon) is the error that we assume will exist (would be zero in a perfect world)

- Check see if your estimate of the 'true' slope is not zero (in other words that X does have *an effect** on Y)
  - H_0: $\beta_1 = 0$
  - H_1: $\beta_1 \neq 0$

*Fine Print: In the statistical sense, does not imply causation

Is it significant?

- Use the *t* statistic
- This considers the estimate of the slope along with the standard error of the slope
- Standard error of the slope is:
  - The standard deviation of the residuals
  - Error in prediction (SSE) degrees of freedom
  - The variation in x (SSX)
- Why *n-2* degrees of freedom?
  - To predict y from x you need to know 2 things (slope and intercept)

$$t = \frac{b_1}{se_{b_1}}$$

$$se_{b_1} = \sqrt{\frac{SSE/(n-2)}{SSX}}$$

Goodness of fit

- If you know x, how well can you explain y? How well does the model fit the data?
- Divide up all the variation (variance) in *y*:
  - SSY – all the variation to be explained
  - SSE – what can't be explained
  - SSR – what is explained by the regression, the regression sum of squares:
    - SSY = SSR + SSE

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Proportion of variance explained is:
  - R^2 = SSR/SSY
- R^2 is the **coefficient of determination** (*don't confuse this with r*) between 0 and 1 (or 0 and 100%)
- **Adjusted R^2** accounts for small sample sizes and is more conservative than R^2 (use this if available)

Is the overall model significant?

- Is the variation that the model explains significant relative to the total variation in *y*?
- *F* statistic
  - H_0: model is not significant
  - H_1: model is significant

$$F_{(1,n-2)} = \frac{SSR/1}{SSE/(n-2)}$$

- Compare to the *F*-distribution with degrees of freedom = 1 and *n*-2
  - SSR has 1 d.f. (you need x to get $\hat{y}$)
  - SSE has *n*-2 d.f.
- If the slope is not significant, the model (overall) will not likely be significant!

Regression in R (part 1/3)

```
Call:
lm(formula = y ~ x, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-212.04  -62.52   30.87   86.77  121.88

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.149    180.756   0.156  0.88010
x             39.057      9.172   4.258  0.00277 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121.8 on 8 deg. of freedom
Multiple R-squared: 0.69, Adjusted R-squared: 0.65
F-statistic: 18.13 on 1 and 8 DF,  p-value: 0.002767
```

- What you asked for:
  - the formula here says predict 'y' from 'x' with data from 'dataset'
- Residuals:
  - 5 number summary of the residuals

Regression in R (part 2/3)

```
Call:

lm(formula = y ~ x, data = dataset)

Residuals:

    Min      1Q  Median      3Q     Max

-212.04  -62.52   30.87   86.77  121.88

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   28.149    180.756   0.156  0.88010

x             39.057      9.172   4.258  0.00277 **

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121.8 on 8 deg. of freedom

Multiple R-squared: 0.69, Adjusted R-squared: 0.65

F-statistic: 18.13 on 1 and 8 DF,  p-value: 0.002767
```

Model coefficients (aka parameters)

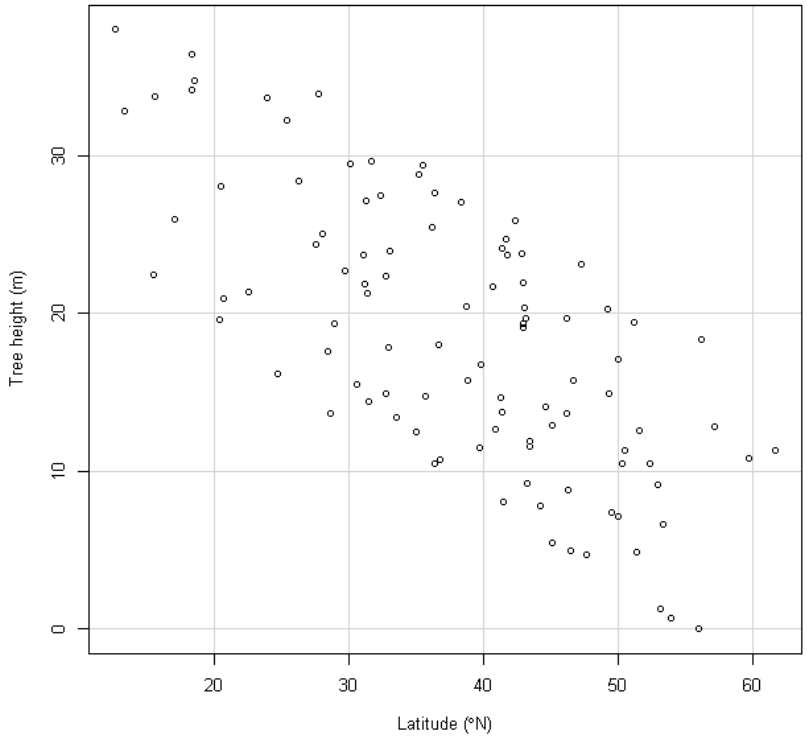
- cols are the coefficient **estimates**, the **standard error** of the coefficient, the ***t* value** for the coefficient & the ***p*-value** for this *t* value
- y-intercept on this row
- slope on this row
- legend explaining level of significance codes

Regression in R (part 3/3)

```
Call:

lm(formula = y ~ x, data = dataset)

Residuals:

    Min      1Q  Median      3Q     Max

-212.04  -62.52   30.87   86.77  121.88

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   28.149    180.756   0.156  0.88010

x             39.057      9.172   4.258  0.00277 **

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121.8 on 8 deg. of freedom

Multiple R-squared: 0.69, Adjusted R-squared: 0.65

F-statistic: 18.13 on 1 and 8 DF,  p-value: 0.002767
```

- Overall model performance
- standard error of the residuals (SSE)
- R2 and the adjusted R2
  - Use the adj. R2 since it is more conservative
- F statistic and the p-value (prob. of getting a bigger F)

Tree vs. Latitude Example Interpreting Outputs




```
> lm1=lm(tree~lat)
> summary(lm1)

Call:
lm(formula = tree ~ lat)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7257  -4.6574   0.3625   4.7728  10.3758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.12757    2.09226  12.010  < 2e-16 ***
lat         -0.30583    0.05247  -5.828 7.17e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.705 on 98 degrees of freedom
Multiple R-squared:  0.2574,    Adjusted R-squared:  0.2498
F-statistic: 33.97 on 1 and 98 DF,  p-value: 7.167e-08
```

Testing Assumptions

- Interval ratio level data
- Random/ probabilistic sample
- Linearity
  - Linear relationship (or straight enough) between continuous paired variables
- Independence (no autocorrelation)
  - Errors (residuals) are statistically independent
- Equal variance (homoscedastic)
  - For every x value (independent variable), the variance of residual error is equal
- Normality
- Raw data should have a bivariate normal distribution
- This means the residuals should be normally distributed

Testing Assumptions

- Linearity
  - Check the scatterplot
- Independence (serial auto correlation)
  - Check a plot of x vs. residuals to look for patterns (random scatter is what you want)
- Equal variance (homoscedastic)
  - Check a plot of x vs. residuals to see if 'the plot thickens'?
  - Do a Leverne's test
- Normality
  - Look at the histogram of the raw data and the residuals

Analysis of the residuals

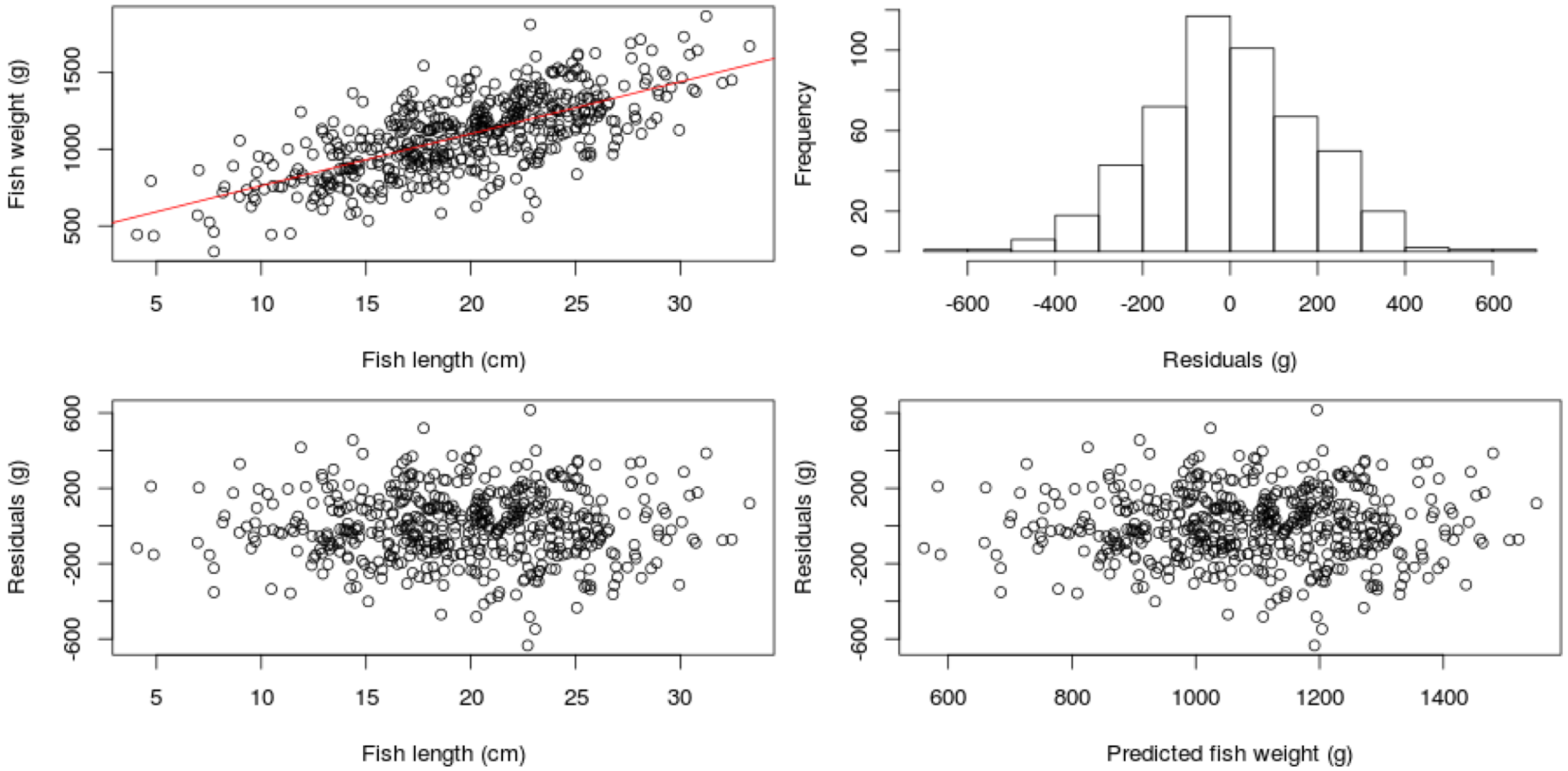
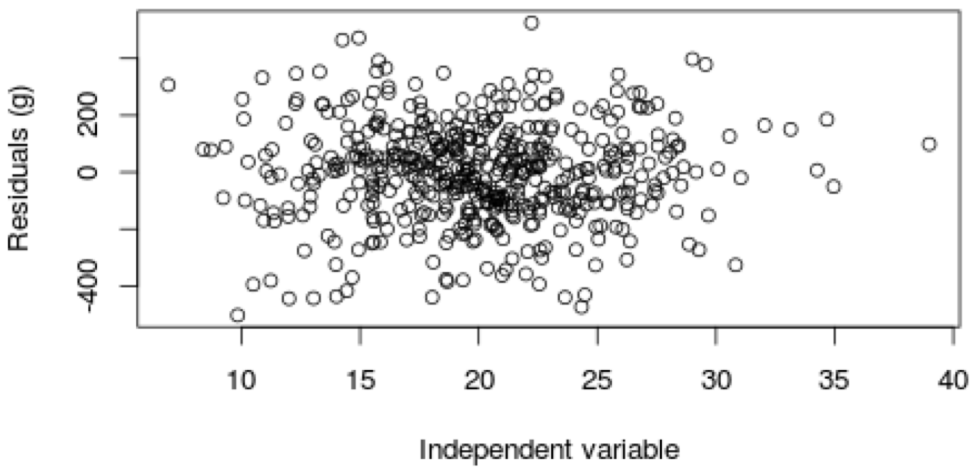
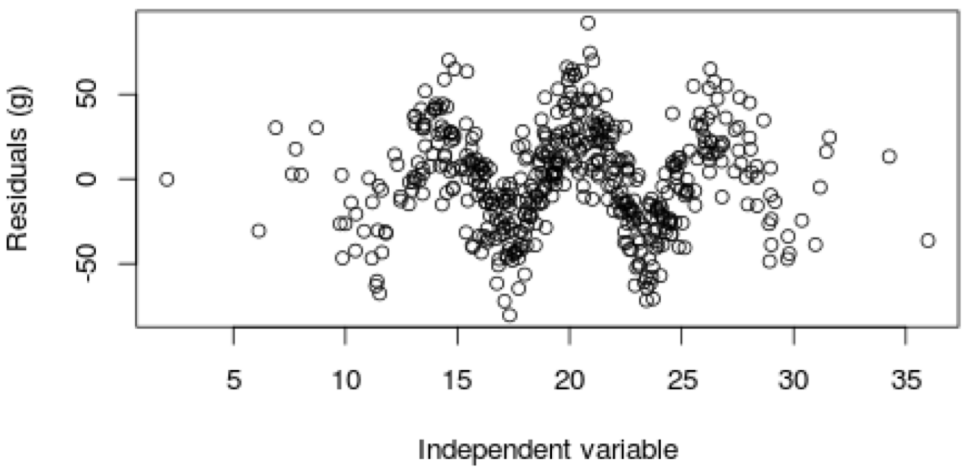


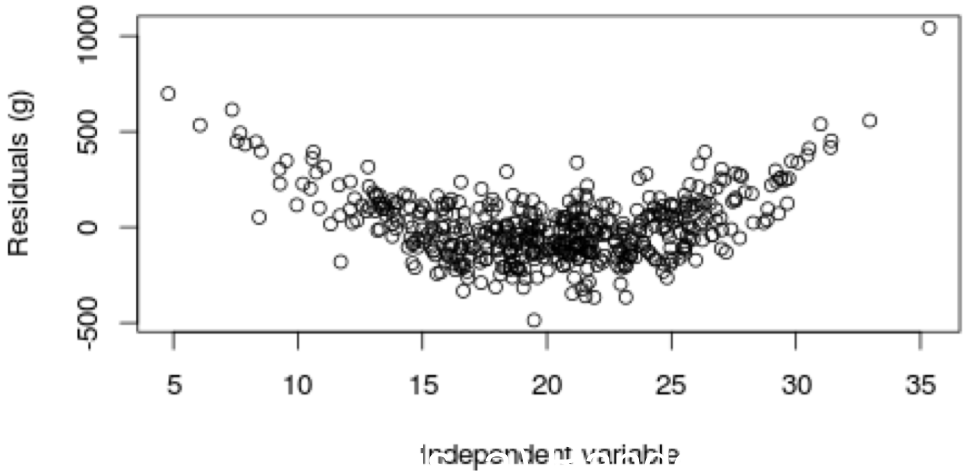
Fig.1 a) Scatterplot and regression line. b) Histogram of residuals. c) Independent variable vs residuals. There is no systematic pattern that would indicate non-random or dependence. d) Predicted values vs residuals. There is no evidence of thickening in this plot (i.e., the dispersion in y does not change systematically with x, which would give a wedge shape).

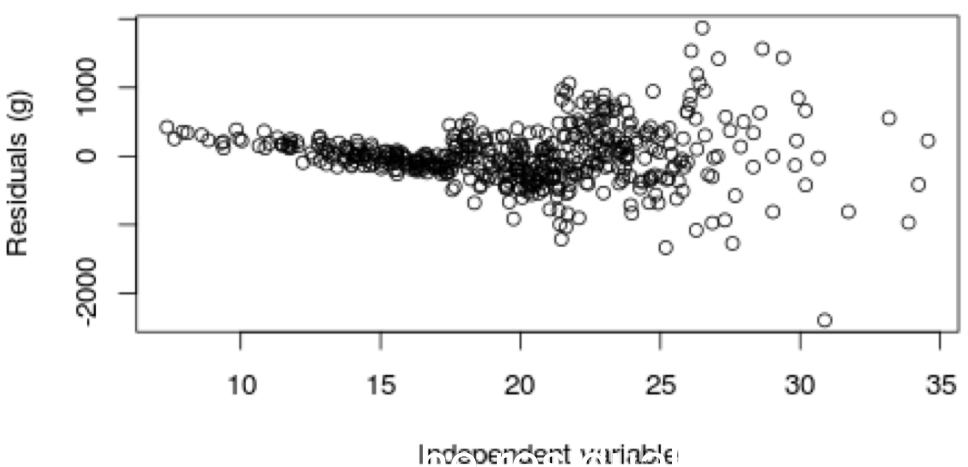
Note that c and d are similar to each other, but the scale of the x axis differs.

What do patterns in Residuals look like….





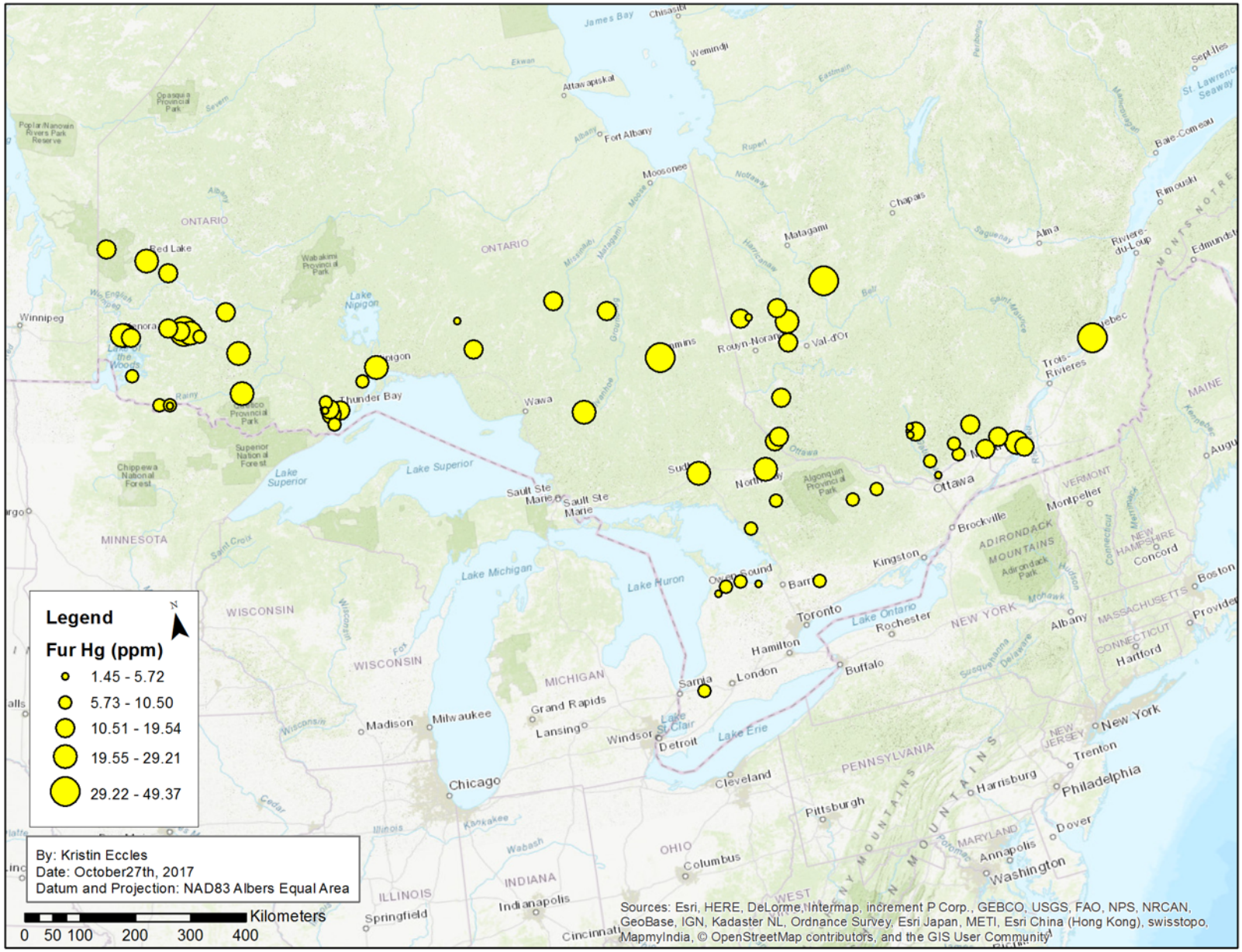




Outliers

- You may notice that certain residuals are really very large compared to the others
- These may be **outliers** or values that are abnormally high or low
- Outliers may:
  - Be caused by a mistake (instrument, transcription, lying respondent on a survey, contamination, etc.),
  - Arise due to chance alone or
  - Signal a problem with your understanding of the system (they could be the 'black swan')
- Outliers may unduly influence model parameters (slope) and so you should consider removing them, *if you have a good reason* to do so
  - Then re-run the analysis

Fur Mercury Example

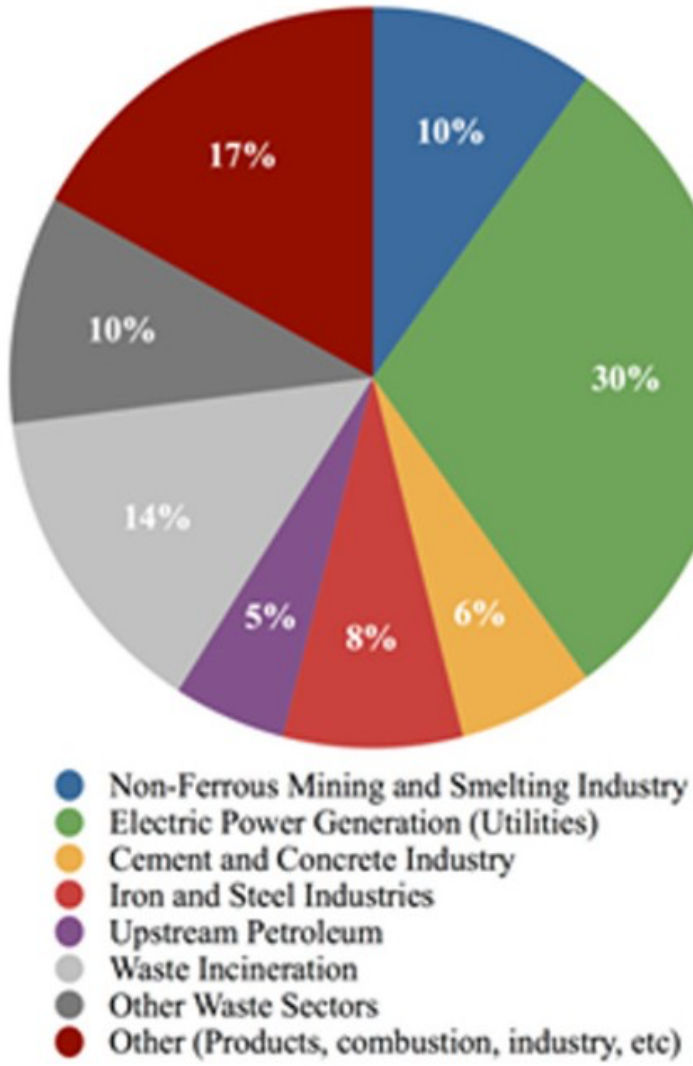


Methods

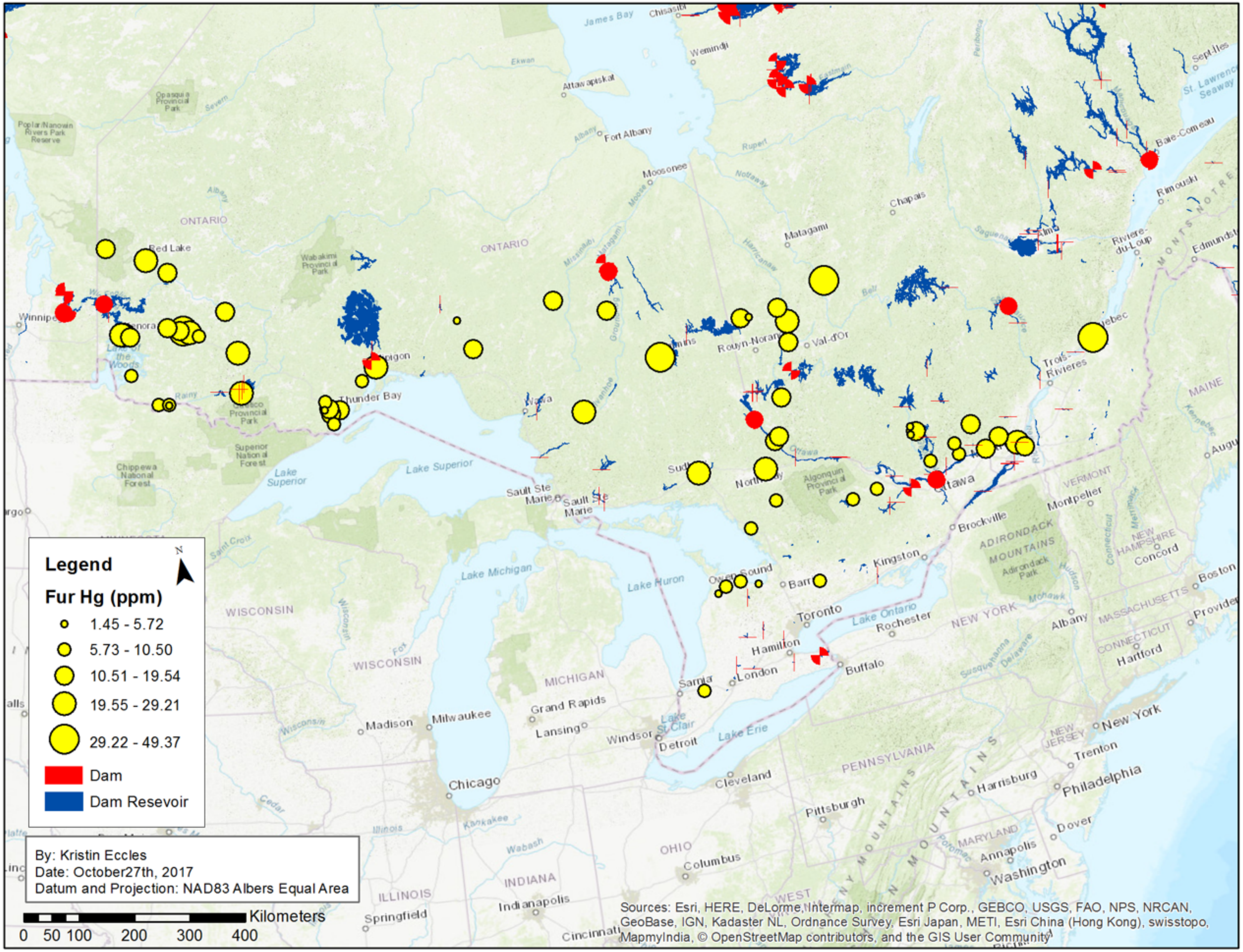
Mercury Sources

Data:

- National pollution release inventory (NPRI)
- Contaminated Sties
- Dams

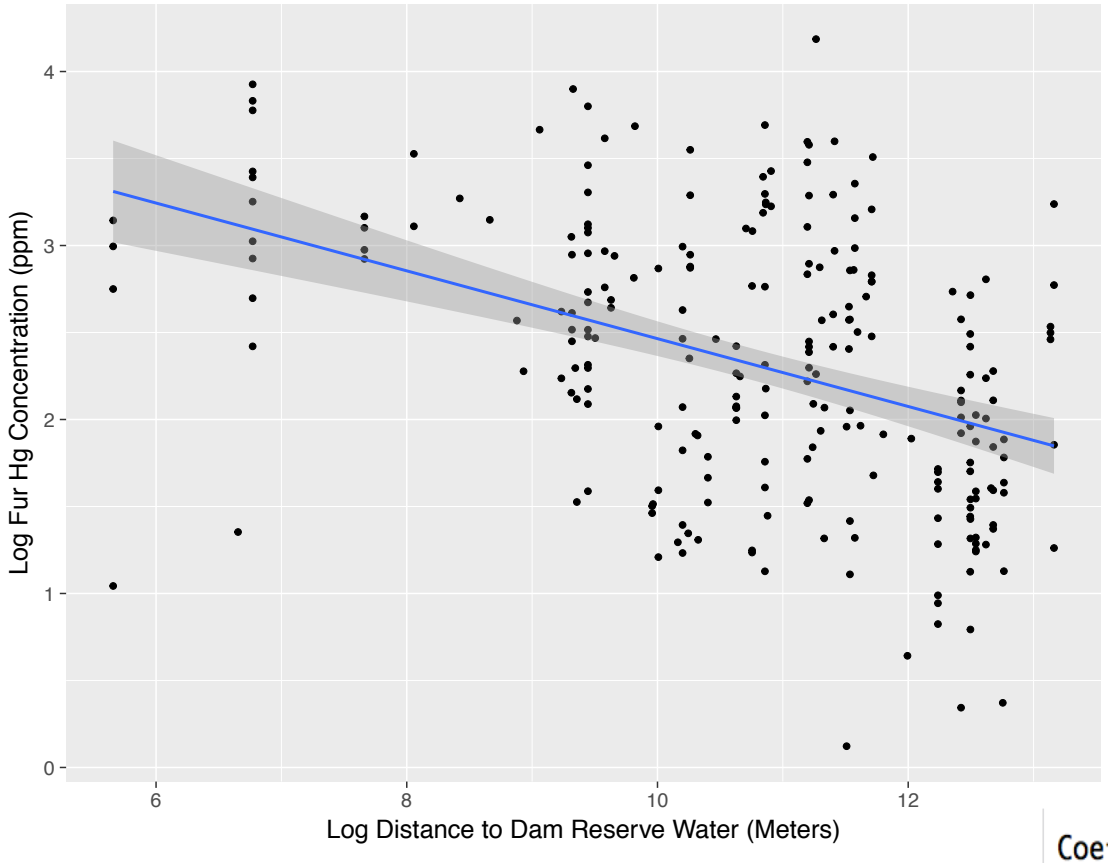


Canadian Mercury Science Assessment , 2016



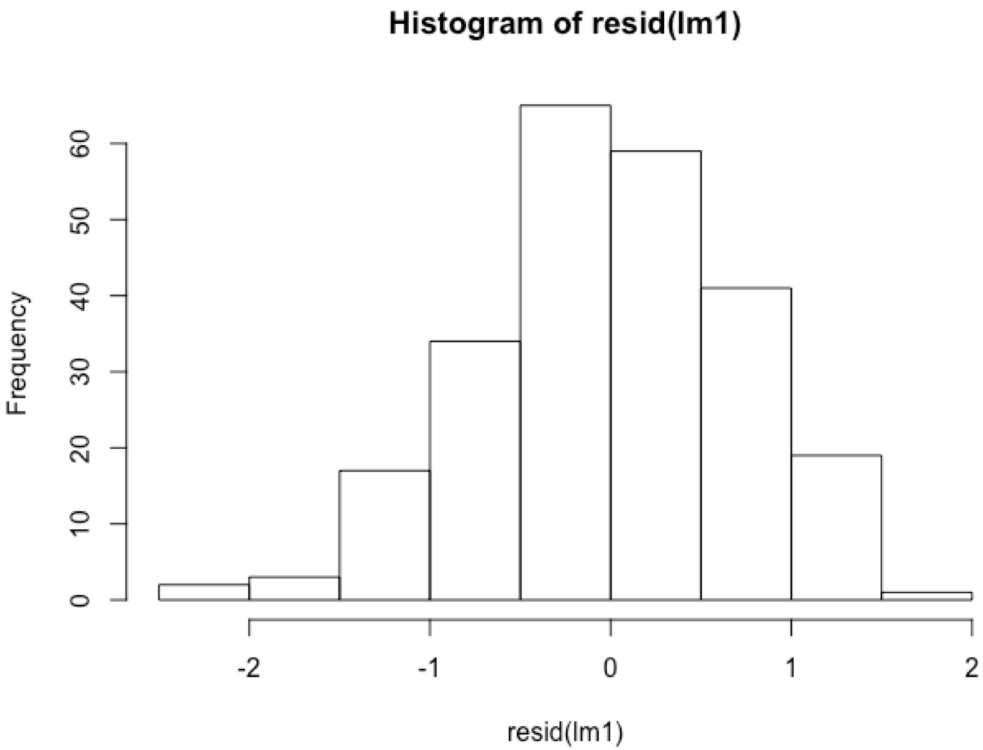
Legend
Fur Hg (ppm)
1.45 - 5.72
5.73 - 10.50
10.51 - 19.54
19.55 - 29.21
29.22 - 49.37
Dam
Dam Resevoir
By: Kristin Eccles
Date: October27th, 2017
Datum and Projection: NAD83 Albers Equal Area
Kilometers
0 50 100 200 300 400
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

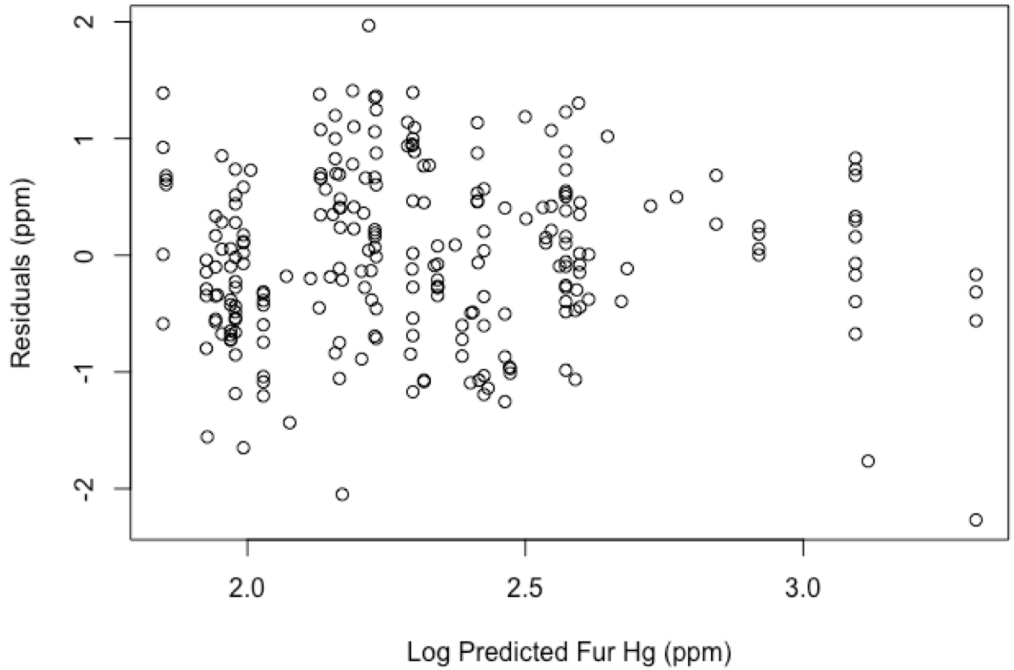
Results



← ME

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.91667    0.13091  14.641  < 2e-16 ***
logres      -0.19487    0.02771  -7.033 2.11e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3112 on 239 degrees of freedom
Multiple R-squared:  0.1715,    Adjusted R-squared:  0.168
F-statistic: 49.46 on 1 and 239 DF,  p-value: 2.111e-11
```


Histogram of resid(lm1)
Frequency
0
10
20
30
40
50
60
-2
-1
0
1
2
resid(lm1)


Residuals (ppm)
-2
-1
0
1
2
2.0
2.5
3.0
Log Predicted Fur Hg (ppm)

Regression vs. Correlation

- Both describe relationships between two (continuous) interval or ratio variables.
- Correlation focuses on the strength of the association (linear relationship)
- Regression gives a quantitative summary (equation) that describes the relationship
- In most cases, it is regression that you will want to do.
  - Your data will not allow inference about the correlation
  - Regression provides a more complete description of your data.

Regression vs. Correlation

- **Correlation:** Measures strength of the association (i.e. linear relationship) between X and Y. No distinction between dependent and independent variable.
  - Correlation of x with y is the SAME as that of y with x.
- **Regression:** predicts vales of the dependent variable from values of the independent variable. Hence there is a distinction between dependent and independent variables
  - Regression of x on y is NOT the same as that of y on x.

Summary

- Modelling can help understand a statistical relationship and can be used for prediction
- A linear regression requires estimating 2 parameters using the least squares technique
- The parameters (slope especially) can be tested for significance (*t*-test)
- The variance in y that the model explains is given by R^2 – this can be tested for significance (*F*-test)
  - **Do not confuse R^2 with r**
- Be aware of the assumptions of raw data and residuals
- Consider the influence of outliers
- Do not assume that a strong relationship *confirms* causality