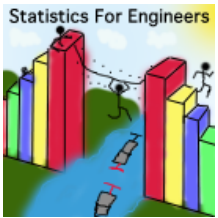
Statistics for Engineers



© Kevin Dunn, 2016
kevin.dunn@mcmaster.ca
http://learnche.mcmaster.ca/

Least Squares Modelling

Copyright, sharing, and attribution notice

We appreciate:

- if you let us know about **any errors** in the slides
- **any suggestions to improve the notes**

All of the above can be done by writing to

kevin.dunn@mcmaster.ca

or anonymous messages can be sent to Kevin Dunn at

http://learnche.mcmaster.ca/feedback-questions

If reporting errors/updates, please quote the current revision number:

Least squares models in context ...

1. Data visualization
2. Univariate data analysis
3. Least squares modelling
4. Design and analysis of experiments
5. Process monitoring

- We will use confidence intervals and visualization heavily though
- Some of the most important classes are this section: we develop the base for DOE and almost all modelling tools you will see in your career.

Usage examples

- **Quantify** relationship between 2 variables:
  - *Manager*: How does yield from the lactic acid batch fermentation relate to the purity of sucrose?
  - *Engineer*: The yield can be predicted from sucrose purity with an error of plus/minus 8%
  - *Manager*: And how about the relationship between yield and glucose purity?
  - *Engineer*: Over the range of our historical data, there is no discernible relationship.

Usage examples

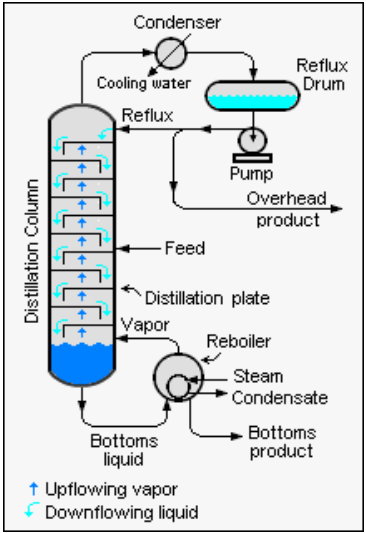
- **Assess** relationship between 2 variables:
  - *Engineer 1*: the theoretical equation for the melt index is non-linearly related to the viscosity
  - *Engineer 2*: the linear model does not show any evidence of that, but the model's prediction ability does improve slightly when we use a non-linear transformation in the least-squares model.

You've seen the least squares model hundreds of times

$$\hat{y} = \beta_0 + \beta_1 x$$

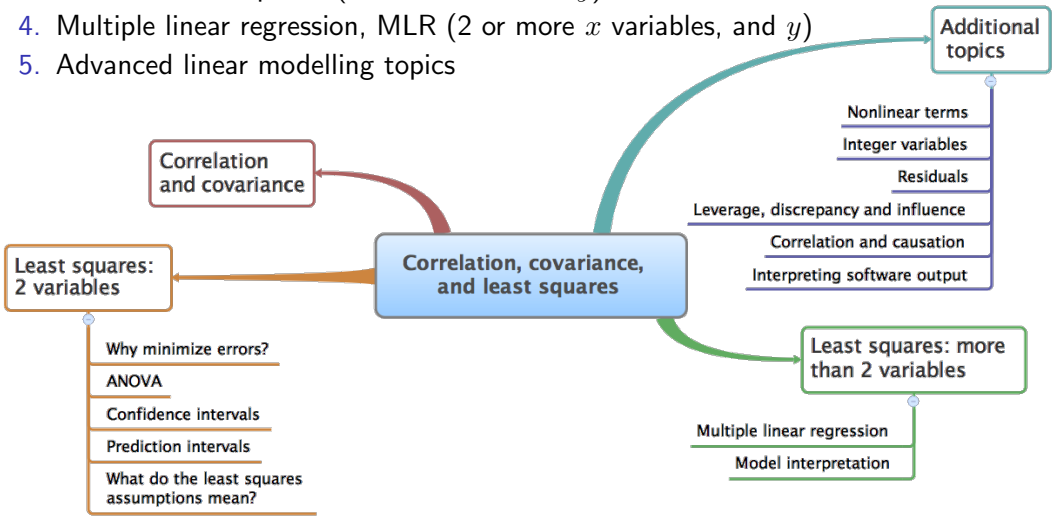
- What is a good predictor, x?
- What about confidence intervals in model building to tell if there is a relation?
- Most least squares models these days are built and never even seen by humans.
- What does R^2 really mean?
- How to judge predictions from the model? $\hat{y} \pm$ ___

A real example: predict the vapour pressure of the overhead product



[Both figures from Wikipedia]

1. Covariance
2. Correlation
3. Bivariate least squares (2 variables: x and y)
4. Multiple linear regression, MLR (2 or more x variables, and y)
5. Advanced linear modelling topics



References and readings

- **Recommended**: John Fox, *Applied Regression Analysis and Generalized Linear Models*
- **Recommended**: Draper and Smith, *Applied Regression Analysis*
- Box, Hunter and Hunter, *Statistics for Experimenters*, portions of Chapter 10 (2nd edition)
- Montgomery and Runger, *Applied Statistics and Probability for Engineers*

Learning about the covariance, with an example

- Consider measurements from a gas cylinder: temperature (K) and pressure (kPa).
- Ideal gas law applies under moderate condition: $pV = nRT$
  - Fixed volume, $V = 20 \times 10^{-3}\text{m}^3 = 20$ L
  - Moles of gas, $n = 14.1$ mols of chlorine gas, (1 kg gas)
  - Gas constant, $R = 8.314$ J/(mol.K)
- Simplify the ideal gas law to: $p = \beta_1 T$, where $\beta_1 = \dfrac{nR}{V} > 0$

Learning about the covariance, with an example

Raw data measured:

	Cylinder temperature (K)	Cylinder pressure (kPa)	Room humidity (%)
	273	1600	42
	285	1670	48
	297	1730	45
	309	1830	49
	321	1880	41
	333	1920	46
	345	2000	48
	357	2100	48
	369	2170	45
	381	2200	49
Mean	327	1910	46.1
Variance	1188	38940	7.0

Learning about the covariance, with an example

- Formal definition:

$$\text{Cov}\{x, y\} = \mathcal{E}\{(x - \overline{x})(y - \overline{y})\} \qquad \text{where} \qquad \mathcal{E}\{z\} = \overline{z}$$

1. Calculate deviation variables: $T - \overline{T}$ and $p - \overline{p}$
   - Subtracting off mean centers the vector at zero
2. Multiply the centered vectors: $(T - \overline{T})(p - \overline{p})$
   - `16740 10080 5400 1440 180 60 1620 5700 10920 15660`
3. Calculate the expected value (mean): 6780
4. Covariance has units: 6780 [K.kPa]
   - Product of the individual units
   - *Awkward units*!

- Covariance between pressure and humidity is: 202 [kPa.%]
  - *Awkward units*!

Some properties about the covariance

- Covariance is symmetrical:

$$\mathcal{E}\left\{(x-\overline{x})(y-\overline{y})\right\} = \mathrm{Cov}\left\{x,y\right\} = \mathrm{Cov}\left\{y,x\right\}$$

- (Co)variance of a centered vector $=$ (co)variance of the uncentered vector
- Covariance with itself is the variance:

$$\mathrm{Cov}\left\{x,x\right\} = \mathcal{E}\left\{(x-\overline{x})(x-\overline{x})\right\} \triangleq \mathcal{V}(x)$$

- What does $x^T x$ represent if x is already centered? *Hint*: compare it to

$$N \cdot \mathcal{E}\left\{(x-\overline{x})(x-\overline{x})\right\}$$

- And $x^T y$? *Hint*: compare it to

$$N \cdot \mathcal{E}\left\{(x-\overline{x})(y-\overline{y})\right\}$$

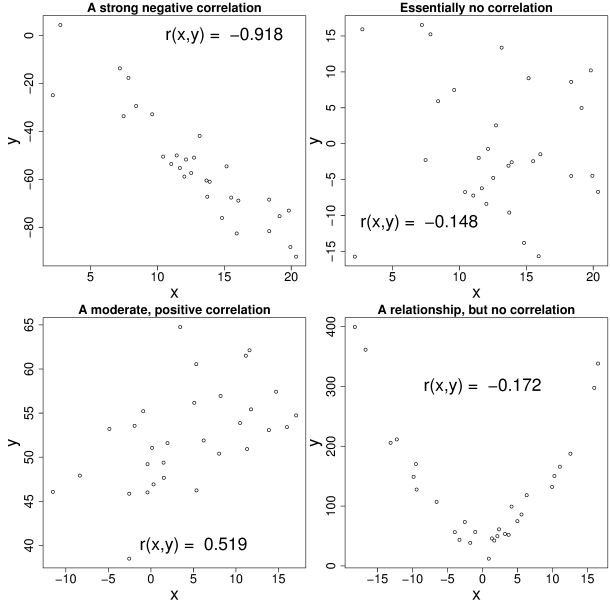
Correlation: measures "co-relationship"

- (Co)variance depends on units: e.g. different covariance for grams vs kilograms
- Correlation removes the scaling effect:

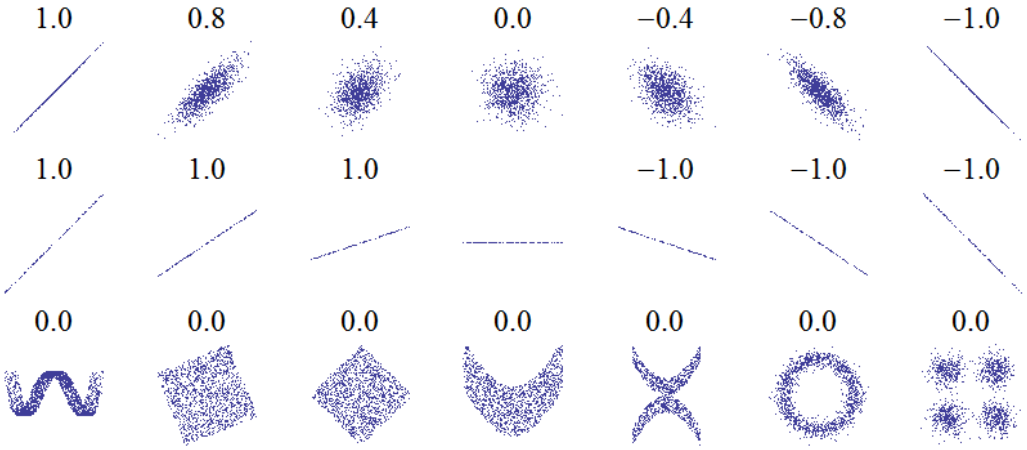
$$\text{correlation} = r(x,y) = \frac{\mathcal{E}\left\{(x-\overline{x})(y-\overline{y})\right\}}{\sqrt{\mathcal{V}\left\{x\right\}\mathcal{V}\left\{y\right\}}} = \frac{\text{Cov}\left\{x,y\right\}}{\sqrt{\mathcal{V}\left\{x\right\}\mathcal{V}\left\{y\right\}}}$$

- Divides by the units of x and y: dimensionless result
- $-1 \leq r(x,y) \leq 1$. you get this by calculating $r(x,x)$ and $r(x,-x)$
- $r = 0$? calculate the correlation between two random, unrelated vectors in R
- Please verify these from the gas cylinder example.
  - r(temperature, pressure) $= 0.997$
  - r(pressure, temperature) $= 0.997$
  - r(pressure, humidity) $= 0.380$
- Do these agree with the interpretation we just learned?

Correlation examples



Learning to interpret the correlation values more carefully



[Wikipedia: Pearson product-moment correlation coefficient]

Formal definitions for covariance and correlation (you've seen this before)

- $\mathcal{E}\{x\} = \overline{x}$
- $\mathcal{E}\{x + y\} = \mathcal{E}\{x\} + \mathcal{E}\{y\} = \overline{x} + \overline{y}$
- $\mathcal{V}\{x\} = \mathcal{E}\{(x - \overline{x})^2\}$
- $\mathcal{V}\{cx\} = c^2\mathcal{V}\{x\}$
- $\mathcal{V}\{x + y\} \neq \mathcal{V}\{x\} + \mathcal{V}\{y\}$, in general
- $\mathcal{V}\{x + y\} = \mathcal{V}\{x\} + \mathcal{V}\{y\}$, only if x and y are independent

Go through the other definitions on your own. They build on each other. The last one was used to derive the confidence interval in the prior module.

Some interesting examples to help test your knowledge

1. x=hours worked per week and y=take home pay
2. x=age of married partner 1 and y=age of married partner 2
3. x=cigarettes smoked per month and y=age at death
4. x=temperature on top tray of distillation column and y=top product purity
5. x=temperature on feed tray of distillation column and y=top product purity

For each one:

- draw a scatter plot of what you think it will look like
- give a rough value for the correlation number
- add an outlier to your scatter plot: what's the practical interpretation of it?
  - In statistics, outliers are our most interesting data points!

Introduction to least squares: relating 2 variables

- It is the basis for a number of algorithms in data analysis
  - Designed experiments
  - Used extensively for latent variable methods
  - Almost all data mining tools embed least squares is some (automatic) way
- We consider only 2 variables for now: x and y.

Other names you will see for the x and y variables:

x	y
input	**output**
predictor	prediction (or "predicted")
endogenous	exogenous
explanatory variable	explained variables
independent	dependent

Model definition

We have 2 vectors of data, x and y. Presume the relationship between them:

$$\begin{aligned} \mathcal{E}\{\mathrm{y}\} &= \beta_0 + \beta_1 \mathrm{x} \\ \mathrm{y} &= \beta_0 + \beta_1 \mathrm{x} + \varepsilon \end{aligned}$$

- β_0, β_1 and ε: are *population* parameters,
- ε term:
  - unmodelled components of the linear model
  - measurement error
  - other random variation

Important: error is from y, not from x, because the model is:

$$\mathrm{y} = \underbrace{\beta_0 + \beta_1 \mathrm{x}}_{\text{prediction component}} + \underbrace{\varepsilon}_{\text{error}}$$

The model does not assume error for x: (the x's are known exactly)

$$\mathrm{y} = \beta_0 + \beta_1 (\mathrm{x} + \epsilon)$$

Linear model definition: $\mathrm{y} = \beta_0 + \beta_1 \mathrm{x} + \varepsilon$

Use **any method** to obtain an estimate of these parameters:

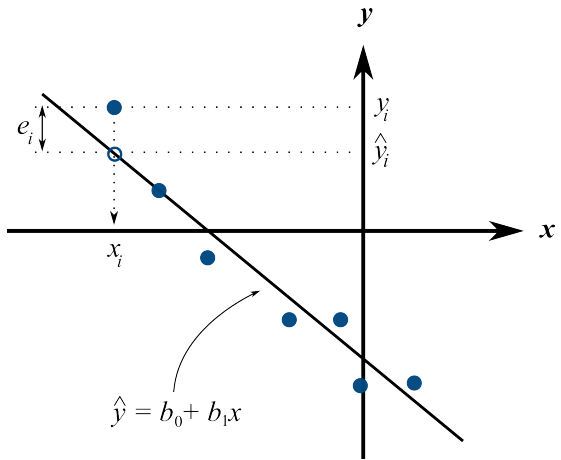
- $b_0 = \hat{\beta}_0$
- $b_1 = \hat{\beta}_1$
- $e = \hat{\varepsilon}$

So we can write:

$$
\begin{aligned}
y_i &= b_0 + b_1 x_i + e_i && \longleftarrow \text{model-building form} \\
\hat{y}_i &= b_0 + b_1 x_i && \longleftarrow \text{predictive form}
\end{aligned}
$$

Model definition

- Calculate the estimates b_0 and b_1
- We want: $\mathcal{E}\{e\} = 0$
- Then for a new x-observation, x_i, we can predict $\hat{y}_i = b_0 + b_1 x_i$



Minimizing errors

How do we calculate b_0 and b_1?
Given: that we have n pairs of data collected: (x_i, y_i)
Aim: make the e_i values small in some way and hopefully also have $\mathcal{E}\{e\} = 0$

Some options we could use:

1. $\sum_{i=1}^{n} (e_i)^2$: the standard least squares objective function
2. $\sum_{i=1}^{n} (e_i)^4$
3. sum of perpendicular distances to the model's line
4. $\sum_{i=1}^{n} \|e_i\|$ (aka least absolute deviations, ℓ-1 norm problem)
5. median $\{e_i^2\}$: least median of squared errors model (one type of robust least squares model)

Why minimize the sum of squares ?

The least squares model:

- has the lowest possible variance for b_0 and b_1 when certain assumptions are met (more later)
- computationally tractable by hand
- very fast on computers
- easy to prove various mathematical properties
- intuitive: penalize deviations quadratically

Other forms: multiple solutions, unstable, high variance solutions, mathematical proofs are difficult

How do we calculate b_0 and b_1 for the least squares model

Given: that we have n pairs of data collected: (x_i, y_i)
Aim: make the e_i values small in some way and hopefully also have $\mathcal{E}\{e\} = 0$

Some options we could use:

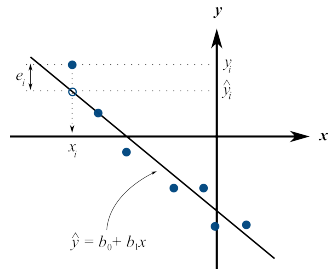
1. $\sum_{i=1}^{n} (e_i)^2$: the standard least squares objective function
2. $\sum_{i=1}^{n} (e_i)^4$
3. sum of perpendicular distances to the model's line
4. $\sum_{i=1}^{n} \|e_i\|$ (aka least absolute deviations, ℓ-1 norm problem)
5. median $\left\{e_i^2\right\}$: least median of squared errors model

(one type of robust least squares model)

Why minimize the sum of squares for a linear model $y_i = b_0 + b_1 x_i + e_i$?

The least squares model:

- is very fast on computers
- has a differentiable objective function
- easy to prove various mathematical properties
- intuitive: penalize deviations quadratically
- Other forms of solving for b_0 and b_1 have multiple solutions, are unstable, have high variance solutions, and mathematical proofs are difficult
- has the lowest possible variance for $b_0 \leftarrow \beta_0$ and $b_1 \leftarrow \beta_1$ when certain assumptions are met (more later)



Least squares is computationally tractable by hand

Solving the least squares model to find values for b_0 and b_1

Has to be an optimization problem: **minimizing** the sum of squared errors

- Easy to solve! Unconstrained optimization problem (know this from an undergrad optimization course)

$$\begin{aligned} \min_{b_0, b_1} \; f(b_0, b_1) &= \sum_{i=1}^{n} (e_i)^2 \\ &= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \end{aligned}$$

Sub in that $e_i = y_i - (\hat{y}_i) = y_i - (b_0 + b_1 x_i)$

Solving the least squares model to find values for b_0 and b_1: grid search

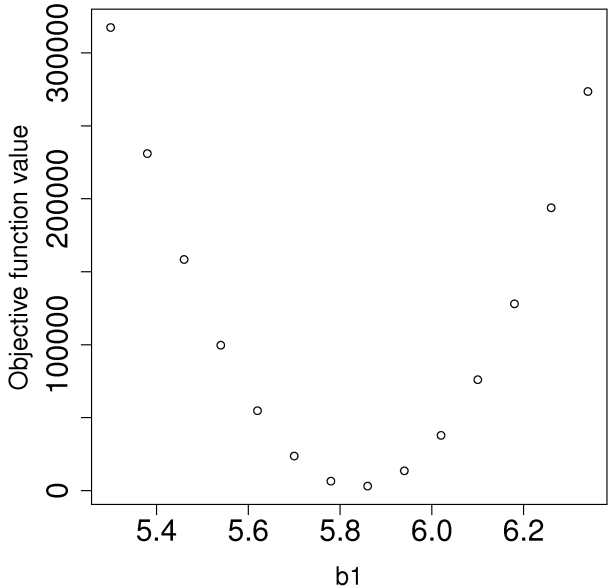
Let's come back to the gas cylinder example: $p = \beta_0 + \beta_1 T$

- We know that $\beta_0 = 0$ from theoretical principles
- Solve for β_1 by trial and error. Initial guess?
- $b_1 \leftarrow \beta_1 = \dfrac{nR}{V} = \dfrac{(14.1 \text{ mol})(8.314 \text{ J/(mol.K)})}{20 \times 10^{-3} \text{ m}^3} = 5.861 \text{ kPa/K}$

$$\text{objective function} = \min_{\cancel{b_0}, b_1} f(\cancel{b_0}, b_1) \quad = \quad \sum_{i=1}^{n} (y_i - \cancel{b_0} - b_1 x_i)^2 = \sum_{i=1}^{n} (y_i - b_1 x_i)^2$$

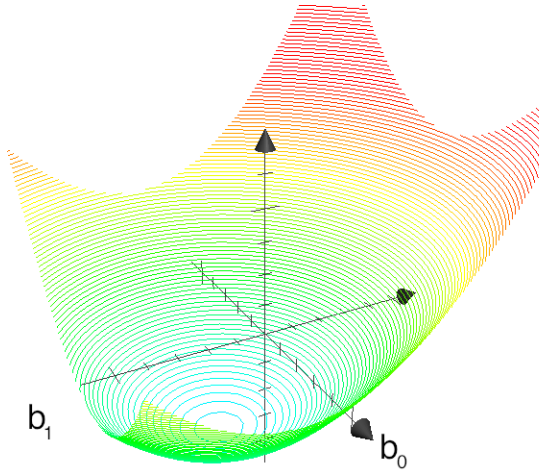
1. Construct equally spaced points between 5.0 and 6.5,
2. Set $b_1 =$ guessed value
3. Calculate the objective function
4. Plot b_1 value vs objective function

Solving the least squares model to find values for b_0 and b_1: grid search



Solving the least squares model to find values for b_0 and b_1: grid search in 2 variables

- objective function shape is a bowl
- a unique minimum can always be found
- because the objective function is convex
- guarantees it is a global, unique optimum



Solving the least squares model to find values for b_0 and b_1: analytically

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

At the optimal point we know that:

- The partial derivatives wrt b_0 and b_1 are zero

$$\begin{aligned} \frac{\partial f(b_0, b_1)}{\partial b_0} &= -2 \textstyle\sum_i^n (y_i - b_0 - b_1 x_i) = 0 \\ \frac{\partial f(b_0, b_1)}{\partial b_1} &= -2 \textstyle\sum_i^n (x_i)(y_i - b_0 - b_1 x_i) = 0 \end{aligned}$$

- This represents 2 linear equations in 2 unknowns

Solving the least squares model to find values for b_0 and b_1: analytically

$$\begin{aligned}
\frac{\partial f(b_0, b_1)}{\partial b_0} &= -2\sum_i^n (y_i - b_0 - b_1 x_i) = 0 \\
\frac{\partial f(b_0, b_1)}{\partial b_1} &= -2\sum_i^n (x_i)(y_i - b_0 - b_1 x_i) = 0
\end{aligned}$$

The least squares estimates are:

$$\begin{aligned}
b_0 &= \overline{\mathrm{y}} - b_1 \overline{\mathrm{x}} \\
b_1 &= \frac{\sum_i (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{\sum_i (x_i - \overline{\mathrm{x}})^2}
\end{aligned}$$

Solving the least squares model to find values for b_0 and b_1: analytically

$$
\begin{aligned}
b_0 &= \overline{\mathrm{y}} - b_1 \overline{\mathrm{x}} \\
b_1 &= \frac{\sum_i \left(x_i - \overline{\mathrm{x}}\right)\left(y_i - \overline{\mathrm{y}}\right)}{\sum_i \left(x_i - \overline{\mathrm{x}}\right)^2} = \frac{\left(\mathrm{x} - \overline{\mathrm{x}}\right)'\left(\mathrm{y} - \overline{\mathrm{y}}\right)}{\left(\mathrm{x} - \overline{\mathrm{x}}\right)'\left(\mathrm{x} - \overline{\mathrm{x}}\right)}
\end{aligned}
$$

The least squares solutions has these properties:

1. The units for b_1 are $\frac{\text{units of "y"}}{\text{units of "x"}}$
2. $b_0 = \overline{\mathrm{y}} - b_1 \overline{\mathrm{x}}$ can be written as $\overline{\mathrm{y}} = b_0 + b_1 \overline{\mathrm{x}}$
   - indicates the straight line equation passes through $(\overline{\mathrm{x}}, \overline{\mathrm{y}})$ without error

Solving the least squares model to find values for b_0 and b_1: analytically

$$
\begin{aligned}
b_0 &= \overline{\mathrm{y}} - b_1 \overline{\mathrm{x}} \\
b_1 &= \frac{\sum_i (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{\sum_i (x_i - \overline{\mathrm{x}})^2} = \frac{(\mathrm{x} - \overline{\mathrm{x}})'(\mathrm{y} - \overline{\mathrm{y}})}{(\mathrm{x} - \overline{\mathrm{x}})'(\mathrm{x} - \overline{\mathrm{x}})}
\end{aligned}
$$

The least squares solutions has these properties:

3. We don't usually examine the objective function value
4. This value is useful though: *average error*

$$
\begin{aligned}
\frac{\sum_i^n e_i}{n} &= \frac{\sum_i^n (y_i - b_0 - b_1 x_i)}{n} \\
&= \frac{\sum_i^n (y_i)}{n} - b_0 \frac{\sum_i^n (1)}{n} - b_1 \frac{\sum_i^n (x_i)}{n} \\
&= \overline{y} - b_0 \frac{\sum_i^n 1}{n} - b_1 \frac{\sum_i^n x_i}{n} \\
&= \overline{y} - b_0 - b_1 \overline{x} \\
&= 0
\end{aligned}
$$

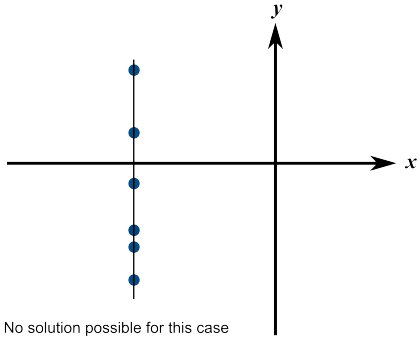
5. Estimate of b_0 depends on b_1: the estimates are correlated

Solving the least squares model to find values for b_0 and b_1: analytically

$$
\begin{aligned}
b_0 &= \overline{\mathrm{y}} - b_1 \overline{\mathrm{x}} \\
b_1 &= \frac{\sum_i \left(x_i - \overline{\mathrm{x}}\right)\left(y_i - \overline{\mathrm{y}}\right)}{\sum_i \left(x_i - \overline{\mathrm{x}}\right)^2} = \frac{\left(\mathrm{x} - \overline{\mathrm{x}}\right)'\left(\mathrm{y} - \overline{\mathrm{y}}\right)}{\left(\mathrm{x} - \overline{\mathrm{x}}\right)'\left(\mathrm{x} - \overline{\mathrm{x}}\right)}
\end{aligned}
$$

The least squares solutions has these properties:

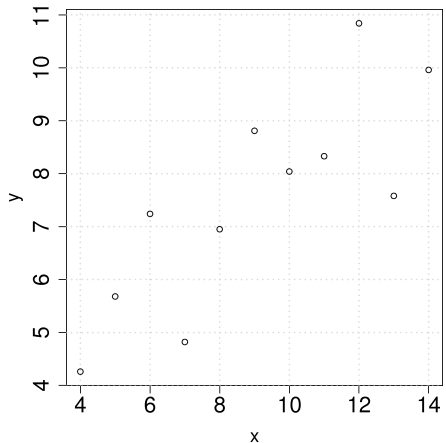
6. We will always get a solution, except in 1 trivial case (*hint*: look at the denominator)



No solution possible for this case

Example problem

1. Calculate model parameter estimates: $y = b_0 + b_1 x$ from the given data
2. Calculate predicted value $\hat{y}_i$ when $x_i = 5.5$



x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

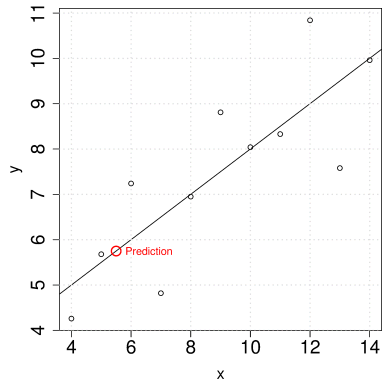
- $\bar{x} = 9.0$
- $\bar{y} = 7.5$
- $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 55.0$
- $\sum_i (x_i - \bar{x})^2 = 110$

Example problem: solution

1. $b_1 = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \dfrac{55}{110} = 0.5$

   $b_0 = \bar{y} - b_1 \bar{x} = 7.5 - (0.5)(9.0) = 3.0$

2. $\hat{y}_i = b_0 + b_1 x_{i,\text{new}} = 3.0 + (0.5)(5.5) = 5.75$



x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

- $\bar{x} = 9.0$
- $\bar{y} = 7.5$
- $\sum_i (x_i - \bar{x})(y_i - \bar{y}) = 55.0$
- $\sum_i (x_i - \bar{x})^2 = 110$

Least squares model analysis after we have calculated values for b_0 and b_1

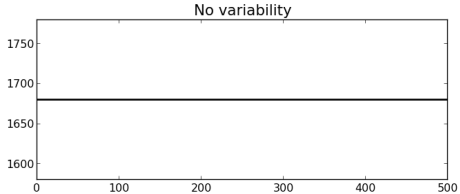
1. How well does the model perform?
2. What part of the data is error
   - noise
3. What part of the data is systematic
   - signal
4. Confidence interval for the model coefficients: b_0 and b_1

   For example:
   - intercept: $-2.3 \leq \beta_0 \leq +3.2$
   - slope = rate constant (1/seconds): $0.55 \leq \beta_1 \leq 1.07$
5. Prediction interval for the y-variable
   - e.g. the predicted yield is 8 ± 1.7 kg

But, we need to **make assumptions** about the data first. We will get to points 4 and 5 soon. Let's look at points 1, 2, and 3 next.

The variance breakdown for our model: $\mathrm{y} = b_0 + b_1\mathrm{x} + \mathrm{e}$

Recall: **life is pretty boring without variability**



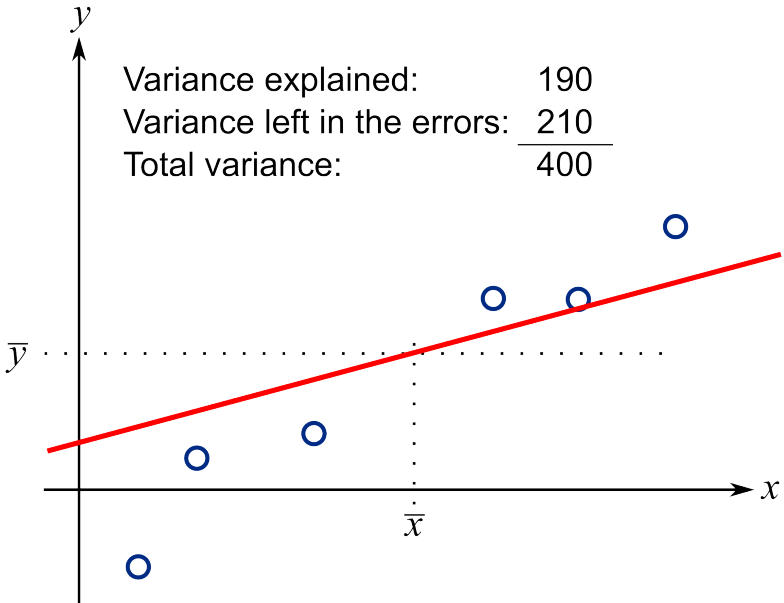
Analysis of variance (ANOVA) - just a tool to show the breakdown of variability in y

1. doing nothing, no model: implies $\hat{y} = b_0 = \overline{y}$
   - recall from the prior video: the model must pass through $\overline{x}$ and $\overline{y}$
2. now add a slope term to this base case: $\hat{y}_i = b_0 + b_1 x_i$ (intercept *plus* slope)
3. then, how much variance is left over in the errors $y_i = b_0 + b_1 x_i + e_i$?

These add up to the total variance.

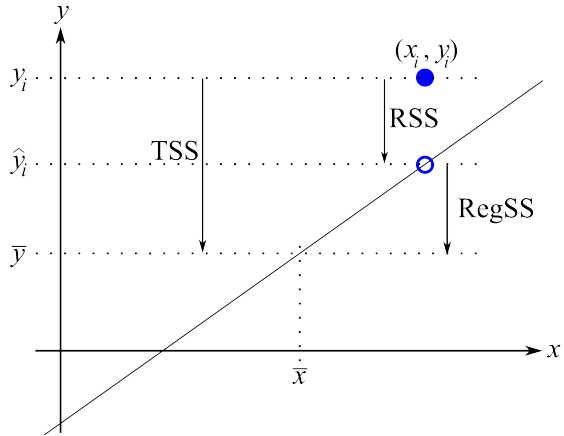
- Total variance of y = model's variance (signal) + error variance (noise)
- variance is quantified as a deviation from the mean

Adjust values of b_0 and b_1 to get the highest amount of variance explained



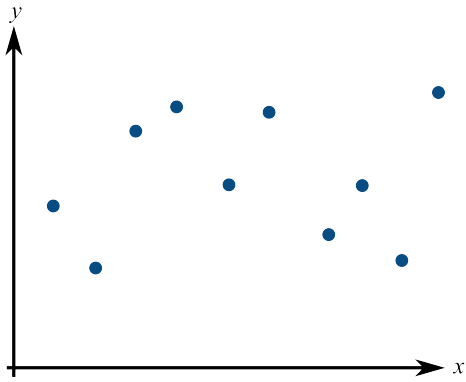
The analysis of variance: algebraically and geometrically

$$
\begin{aligned}
(y_i - \overline{y}) &= (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i) \\
(y_i - \overline{y})^2 &= (\hat{y}_i - \overline{y})^2 + 2(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) + (y_i - \hat{y}_i)^2 \\
\sum (y_i - \overline{y})^2 &= \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y}_i)^2 \\
\text{Total SS (TSS)} &= \text{Regression SS (RegSS)} + \text{Residual SS (RSS)}
\end{aligned}
$$

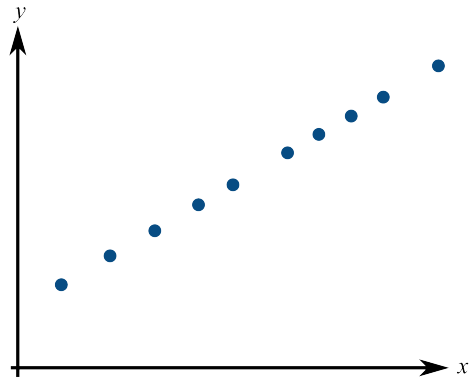


This geometric construction is for any value of x_i and y_i (above or below the line; *prove it*).

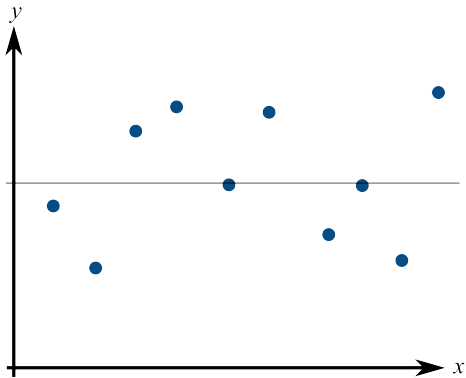
Worst case situation



Best case situation



Worst case situation



RegSS = zero $= \sum(\hat{y}_i - \overline{y})^2$

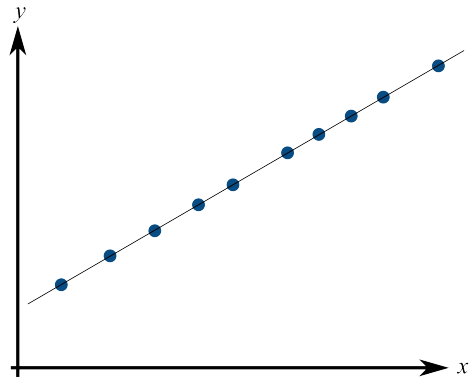
RSS = a large value $= \sum(y_i - \hat{y}_i)^2$

TSS = a large value $= \sum(y_i - \overline{y})^2$

Key point: $\hat{y}_i = \overline{y}$ or $\hat{y}_i - \overline{y} = 0$

$\frac{\text{RegSS}}{\text{TSS}} =$ 0.0 $= R^2$

Best case situation



RegSS = a large value $= \sum(\hat{y}_i - \overline{y})^2$

RSS = zero $= \sum(y_i - \hat{y}_i)^2$

TSS = a large value $= \sum(y_i - \overline{y})^2$

Key point: $\hat{y}_i = y_i$ or $\hat{y}_i - y_i = 0$

$\frac{\text{RegSS}}{\text{TSS}} =$ 1.0 $= R^2$

Confidence intervals for the least squares model: β_0 and β_1

$$b_0 - c_t S_E(b_0) \leq \beta_0 \leq b_0 + c_t S_E(b_0)$$

$$b_1 - c_t S_E(b_1) \leq \beta_1 \leq b_1 + c_t S_E(b_1)$$

- No assumptions required about the data in order to calculate the b_0 and b_1 in the least squares model
- No assumptions required to make model predictions either.

But, 6 assumptions are required to derive and interpret the confidence intervals above.

Assumption 1: the model has a linear structure

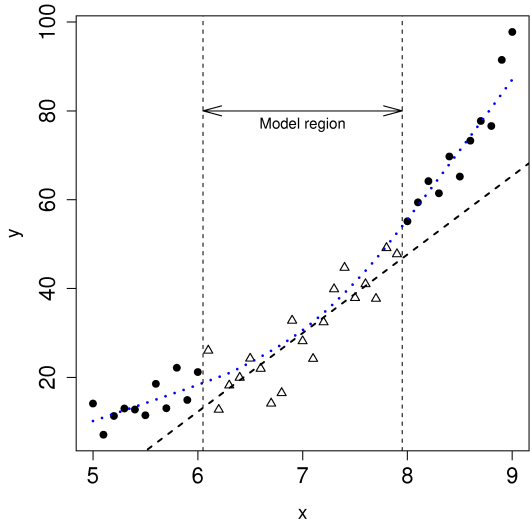
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- $b_0 \longleftarrow \hat{\beta}_0$ and $b_1 \longleftarrow \hat{\beta}_1$
- i.e. we will estimate the fixed population values (that is what the "hats" above $\hat{\beta}_0$ and $\hat{\beta}_1$ show)

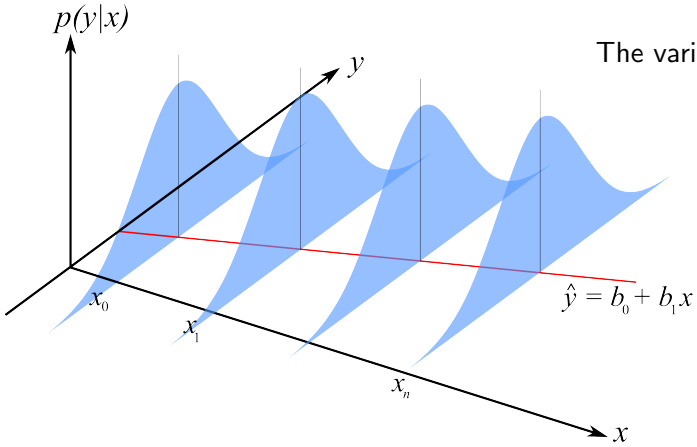
- Implies ε is the error of y, since $\beta_0 + \beta_1 \mathrm{x}$ terms are fixed.
- Means your x variable has much less uncertainty than the y variable

Assumption 1: the model has a linear structure

Systems are known to be non-linear; but a linear model structure might be good enough (depends on the model's purpose)



Assumption 2: the constant error variance assumption



The variance of y is assumed constant at all values of x.

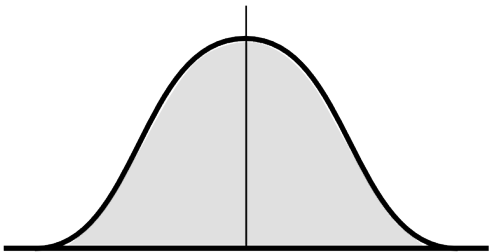
In practice: the variability of y is often non-constant

- measurement accuracy deteriorates at extreme conditions of x
- at high levels of the input variable (e.g. x = temperature, or x = pressure), the output can be more variable.

Assumption 3: Errors are normally distributed: $e_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$

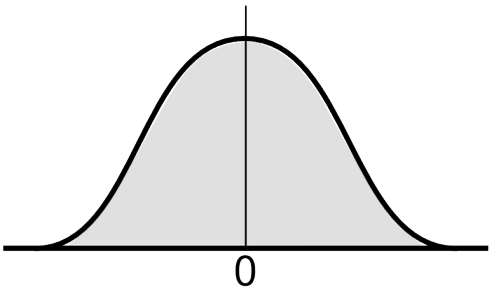
The y_i values

- mean $= \beta_0 + \beta_1 x_i + \varepsilon$
- variance $= \sigma_\varepsilon^2$



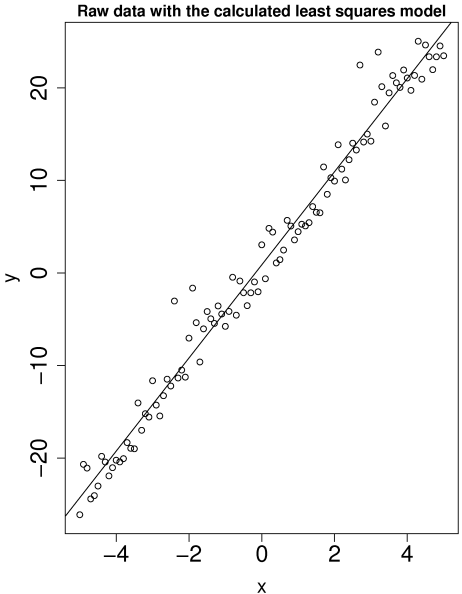
The e_i residual (error) values

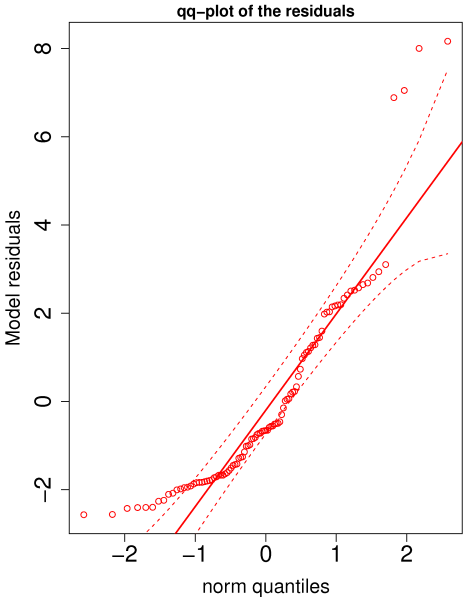
- mean $= 0$
- variance $= \sigma_\varepsilon^2$



- Implies that $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma_\varepsilon^2)$, from first linearity assumption
- We cannot always certain the residuals, e_i, are normally distributed
- But at least it is easy to test with a q-q plot

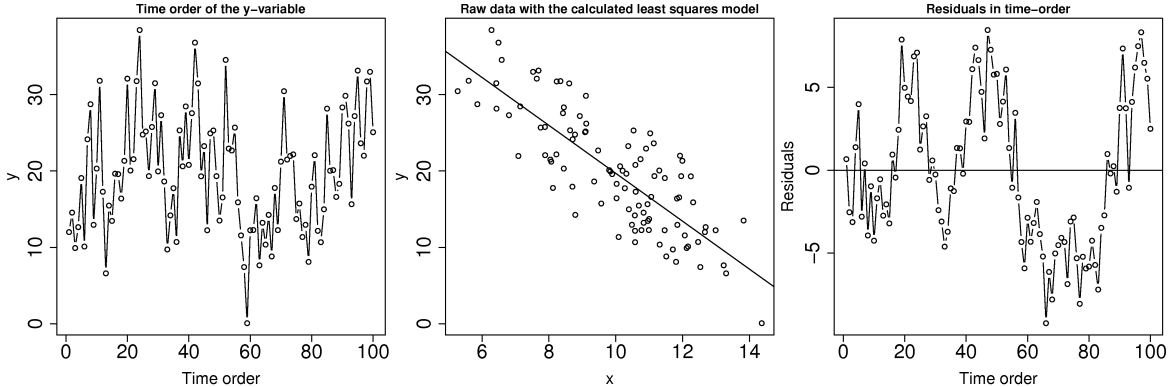
Assumption 3: e_i values are normally distributed: *never believe your eyes!*





Assumption 4: Each error, e_i, is independent of the other

- Often violated in practice
- Observations taken closely together in time on a slow processes
- E.g. if you have a positive error now, your next sample is also likely to have a positive error (called "*autocorrelation*")



Assumption 5: Assume x_i values are fixed and independent of the error

- Closely tied to first assumption: the model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Most engineering cases we sample (measure) the x values – with error!

- Being "independent of the error" means the error are about the same at all values of x: there is no relationship between the x_i values and the e_i values.

Assumption 6: All y_i values are independent of each other

- Closely tied to 4th assumption: the errors, e_i, are independent of the other
- This is the same as assumption 4, if the x_i values are constant
- Violated in cases where data are collected in time order and y_i are *autocorrelated*

Want to work ahead?

- look up the "autocorrelation" article on Wikipedia
- read the help for the `acf(...)` function in R

Why do we need all these assumptions?

1. the model has a linear structure
2. the errors have constant variance at all levels of x
3. e_i values are normally distributed
4. each error, e_i, is independent of the other
5. x_i values are fixed and independent of the error
6. all y_i values are independent of each other

They are all used in the lengthly mathematical derivation for the confidence intervals of β_0 and β_1

- mild violations of the 6 assumptions can be tolerated

Assumptions required for the least squares confidence intervals

1. the model has a linear structure
2. the errors have constant variance at all levels of x
3. e_i values are normally distributed
4. each error, e_i, is independent of the other
5. x_i values are fixed and independent of the error
6. all y_i values are independent of each other

Fortunately: least squares is pretty "robust" to violations in these assumptions.

Confidence intervals for the least squares model: β_0 and β_1

β_0

$$z = \frac{b_0 - \beta_0}{S_E(\beta_0)}$$

$$-c_t \leq \frac{b_0 - \beta_0}{S_E(\beta_0)} \leq +c_t$$

$$b_0 - c_t S_E(\beta_0) \leq \beta_0 \leq b_0 + c_t S_E(\beta_0)$$

$$b_0 \sim \mathcal{N}(\beta_0, \mathcal{V}\{\beta_0\})$$

β_1

$$z = \frac{b_1 - \beta_1}{S_E(\beta_1)}$$

$$-c_t \leq \frac{b_1 - \beta_1}{S_E(\beta_1)} \leq +c_t$$

$$b_1 - c_t S_E(\beta_1) \leq \beta_1 \leq b_1 + c_t S_E(\beta_1)$$

$$b_1 \sim \mathcal{N}(\beta_1, \mathcal{V}\{\beta_1\})$$

Variance calculation for $b_1 \leftarrow \hat{\beta}_1$

The details are in the course notes (PID book, chapter 4)

$$b_1 = \frac{\sum_i (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{\sum_j (x_j - \overline{\mathrm{x}})^2}$$

$$b_1 = m_1 y_1 + m_2 y_2 + \ldots + m_i y_i + \ldots m_N y_N \qquad \text{where} \qquad m_i = \frac{x_i - \overline{\mathrm{x}}}{\sum_j (x_j - \overline{\mathrm{x}})^2}$$

$$\mathcal{E}\{b_1\} = \mathcal{E}\{m_1 y_1\} + \mathcal{E}\{m_2 y_2\} + \ldots + \mathcal{E}\{m_N y_N\}$$

$$\mathcal{V}\{b_1\} = m_1^2 \mathcal{V}\{y_1\} + m_2^2 \mathcal{V}\{y_2\} + \ldots + m_N^2 \mathcal{V}\{y_N\}$$

$$\mathcal{V}\{b_1\} = \sum_i \left(\frac{x_i - \overline{\mathrm{x}}}{\sum_j (x_j - \overline{\mathrm{x}})^2} \right)^2 \mathcal{V}\{y_i\} = \frac{\mathcal{V}\{y_i\}}{\sum_j (x_j - \overline{\mathrm{x}})^2} = \frac{S_E^2}{\sum_j (x_j - \overline{\mathrm{x}})^2}$$

Variance calculation for $b_1 \leftarrow \hat{\beta}_1$

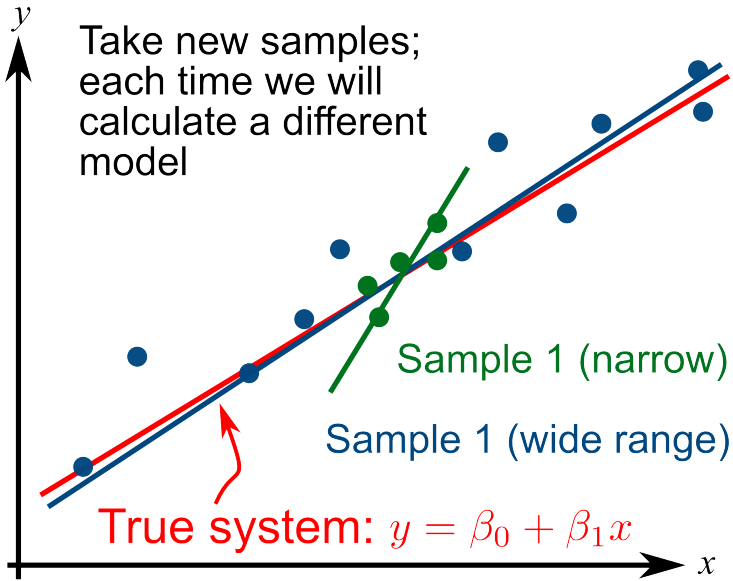
$$\mathcal{V}\{b_1\} = \frac{S_E^2}{\sum_j (x_j - \overline{\mathrm{x}})^2}$$

- Interpretation of the numerator: it is the standard error S_E^2.
- Denominator's interpretation:
  - use samples that are far from the mean of the x-data
  - use more samples (more terms in the summation)

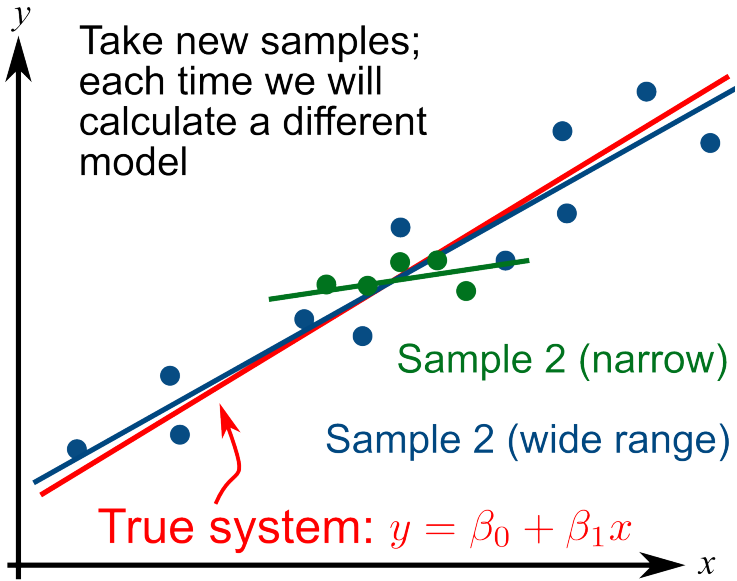
Aside: What is the numerator in terms of the standard error?

- Recall the assumptions had $\mathcal{V}\{y_i\} = \mathcal{V}\{e_i\} = S_E^2$
- and $S_E^2 = \dfrac{\sum e_i^2}{n-k}$ or $S_E = \sqrt{\dfrac{\sum e_i^2}{n-k}}$

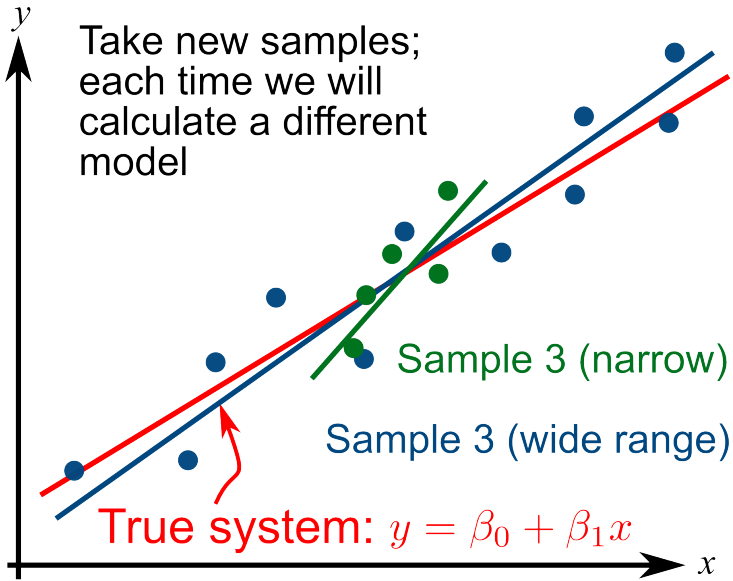
Why are small variances important for us?



Why are small variances important for us?



Why are small variances important for us?



Why are small variances important for us? (i.e. why having a small S_E is important)

It gives an indication how variable (*how stable*) the model is.

If we take second sample,

- a *small variance* implies we get a similar estimate to before (desirable; more stable)
- a *large variance* indicates different estimates every time (undesirable; not stable).

Variance calculation for $b_0 \leftarrow \hat{\beta}_0$

Recall: $b_0 = \overline{y} - b_1 \overline{x}$

$$\mathcal{V}\{b_0\} \quad = \quad \frac{S_E^2}{N} + \frac{S_E^2}{\sum_j (x_j - \overline{\mathrm{x}})^2} \cdot \overline{\mathrm{x}}^2 \quad = \quad S_E^2 \left(\frac{1}{N} + \frac{\overline{\mathrm{x}}^2}{\sum_j (x_j - \overline{\mathrm{x}})^2} \right)$$

Summary of important equations for variances

1. $\mathcal{V}\{\beta_0\}$ is approximated by $\mathcal{V}\{b_0\} = S_E^2(b_0) = S_E^2\left(\frac{1}{N} + \frac{\overline{\mathrm{x}}^2}{\sum_j (x_j - \overline{\mathrm{x}})^2}\right)$
2. $\mathcal{V}\{\beta_1\}$ is approximated by $\mathcal{V}\{b_1\} = S_E^2(b_1) = \frac{S_E^2}{\sum_j (x_j - \overline{\mathrm{x}})^2}$
3. $\mathcal{V}\{y_i\} = \mathcal{V}\{\varepsilon_i\}$ is approximated by $S_E^2 = \frac{\sum e_i^2}{n-k}$

We will give shorter names to the terms above:

1. $S_E(b_0)$ standard deviation of b_0
2. $S_E(b_1)$ standard deviation of b_1
3. S_E standard deviation of the error (residuals)

Confidence intervals for the least squares model: β_0 and β_1

β_0

$$z = \frac{b_0 - \beta_0}{S_E(\beta_0)}$$

$$-c_t \leq \frac{b_0 - \beta_0}{S_E(\beta_0)} \leq +c_t$$

$$S_E(\beta_0) \approx S_E(b_0)$$

$$z = \frac{b_0 - \beta_0}{S_E(b_0)} \sim t(\nu = n - k)$$

$$b_0 - c_t S_E(b_0) \leq \beta_0 \leq b_0 + c_t S_E(b_0)$$

β_1

$$z = \frac{b_1 - \beta_1}{S_E(\beta_1)}$$

$$-c_t \leq \frac{b_1 - \beta_1}{S_E(\beta_1)} \leq +c_t$$

$$S_E(\beta_1) \approx S_E(b_1)$$

$$z = \frac{b_1 - \beta_1}{S_E(b_1)} \sim t(\nu = n - k)$$

$$b_1 - c_t S_E(b_1) \leq \beta_1 \leq b_1 + c_t S_E(b_1)$$

An example: to apply the calculations for confidence intervals

From a previous example we had:

- $b_1 = \dfrac{\sum_i (x_i - \overline{\mathrm{x}})(y_i - \overline{\mathrm{y}})}{\sum_i (x_i - \overline{\mathrm{x}})^2} = \dfrac{55}{110} = 0.5$
- $b_0 = \overline{\mathrm{y}} - b_1 \overline{\mathrm{x}} = 7.5 - (0.5)(9.0) = 3.0$
- $S_E = 1.237$
- $n = 11$

Pay careful attention to the above values. Use them to calculate 95% confidence intervals for β_0 and β_1.

An example: to apply the calculations for confidence intervals – **solution**

First calculate all the variances that we require:

- $S_E^2(b_1) = \dfrac{S_E^2}{\sum_j (x_j - \overline{\mathrm{x}})^2} = \dfrac{1.237^2}{110} = 0.0139$

  $S_E(b_1) = 0.1179$

- $S_E^2(b_0) = S_E^2 \left(\dfrac{1}{N} + \dfrac{\overline{\mathrm{x}}^2}{\sum_j (x_j - \overline{\mathrm{x}})^2} \right) = \left(\dfrac{1}{11} + \dfrac{9^2}{110} \right) 1.237^2 = 1.266$

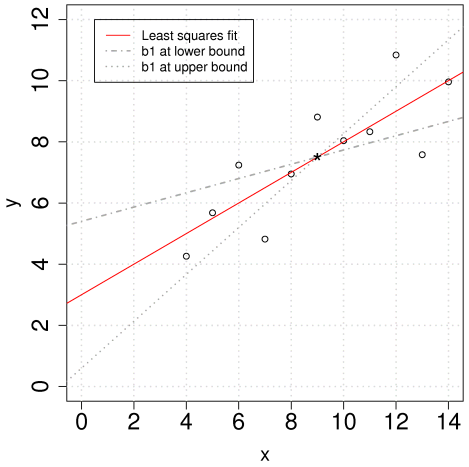
  $S_E(b_0) = 1.125$

- The critical t-value: $c_t = \pm 2.26$ at 95% confidence, using $11 - 2 = 9$ degrees of freedom

An example: to apply the calculations for confidence intervals – **solution**

$$
\begin{array}{rcccl}
 & & -c_t & \leq & \dfrac{b_1 - \beta_1}{S_E(b_1)} \leq +c_t \\
0.5 - 2.26 \times 0.1179 & \leq & \beta_1 & \leq & 0.5 + 2.26 \times 0.1179 \\
0.23 & \leq & \beta_1 & \leq & 0.77
\end{array}
$$

$$
\begin{array}{rcccl}
 & & -c_t & \leq & \dfrac{b_0 - \beta_0}{S_E(b_0)} \leq +c_t \\
3.0 - 2.26 \times \sqrt{1.266} & \leq & \beta_0 & \leq & 3.0 + 2.26 \times \sqrt{1.266} \\
0.457 & \leq & \beta_0 & \leq & 5.54
\end{array}
$$

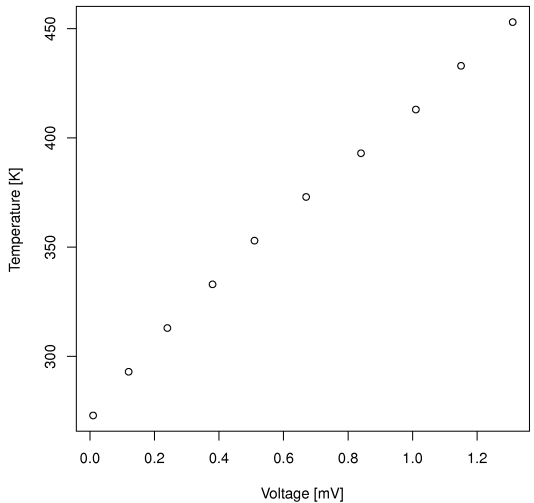


Interpret the software output: $y = 15.7 - (0.70)\text{cameras} = 15.7 - 0.70x$

- $9.61 \leq \beta_0 \leq 21.8$
- $-1.8 \leq \beta_1 \leq 0.36$
- $S_E = 10.9$
- $R^2 = 0.075$

Thermocouple example: predicting temperature using the recorded voltage

Thermocouples produce a mostly linear voltage [mV] response at different temperatures. They often record to an accuracy of ± 0.5K with cheap thermocouples.



1. The calculated model is $\hat{T} = 278.6 + 135.3V$
   - Interpretation of slope?
2. Can you interpret these confidence intervals to your colleague?

$$273 \leq \beta_0 \leq 284$$
$$128 \leq \beta_1 \leq 142$$

3. The $R^2 = 0.996$. What can you say about the model?
4. The $S_E = 3.9K$. Satisfied with the model's prediction ability?

Confidence intervals for the least squares model: β_0 and β_1

β_0

$$z = \frac{b_0 - \beta_0}{S_E(\beta_0)}$$

$$-c_t \leq \frac{b_0 - \beta_0}{S_E(\beta_0)} \leq +c_t$$

$$S_E(\beta_0) \approx S_E(b_0)$$

$$z = \frac{b_0 - \beta_0}{S_E(b_0)} \sim t(\nu = n - k)$$

$$b_0 - c_t S_E(b_0) \leq \beta_0 \leq b_0 + c_t S_E(b_0)$$

β_1

$$z = \frac{b_1 - \beta_1}{S_E(\beta_1)}$$

$$-c_t \leq \frac{b_1 - \beta_1}{S_E(\beta_1)} \leq +c_t$$

$$S_E(\beta_1) \approx S_E(b_1)$$

$$z = \frac{b_1 - \beta_1}{S_E(b_1)} \sim t(\nu = n - k)$$

$$b_1 - c_t S_E(b_1) \leq \beta_1 \leq b_1 + c_t S_E(b_1)$$

Using least squares models for two main reasons

The true model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

1. **Prediction**
   - Our predicted value is $\hat{y}_i = b_0 + b_1 x_i$
   - and a prediction interval is $\hat{y}_i \pm$ 
2. **Learn from the model**
   - examine and interpret the b_0 and b_1 values

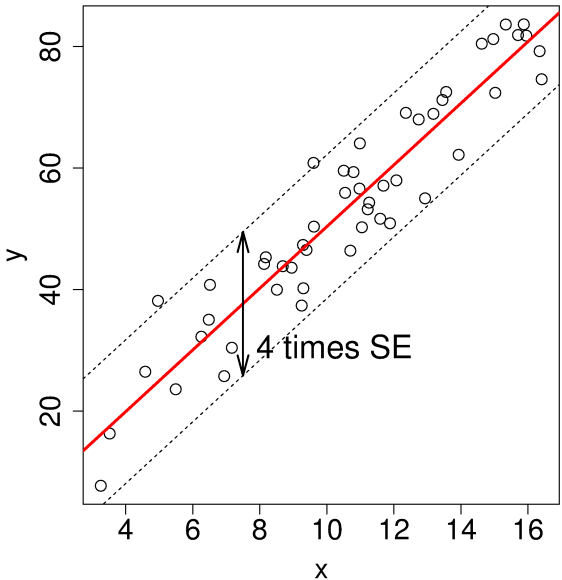
Finding an approximate prediction interval for y

Assuming:

- we want a 95% prediction interval
- the residuals are normally distributed

$$\text{Prediction} = \hat{y}_i \pm 2\,S_E$$

Use this when you are in a hurry:
you're going to get a good estimate!



Making the approximate prediction interval more accurate

- Approximation assumes: $\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Actual model is: $\hat{y}_i = b_0 + b_1 x_i + e_i$

The approximation ignores that b_0 and b_1 have error that is propagated into $\hat{y}_i$.

Taking that error into account now:

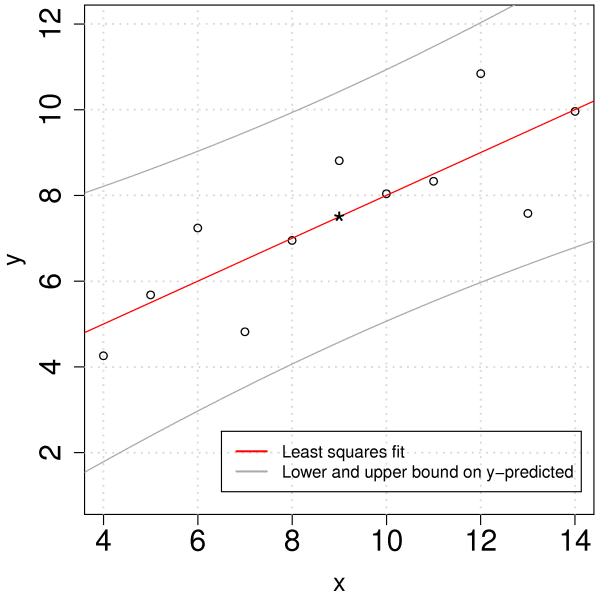
$$\hat{y}_i = b_0 + b_1 x_i + e_i$$

$$\mathcal{V}\{\hat{y}_i\} = \left(\frac{1}{n} + \frac{(x_i - \overline{\mathrm{x}})^2}{\sum_j (x_j - \overline{\mathrm{x}})^2} + 1 \right) S_E^2$$

$\hat{y}_i \pm c_t \sqrt{\mathcal{V}\{\hat{y}_i\}}$ is the actual prediction interval, with $n - k$ degrees of freedom

Above you can see another reason why you want S_E as small as possible.

Prediction interval for y: *it has curvature*



- General shape of the equation: it has a quadratic appearance
- Smallest prediction error at the model center
- Prediction error expands progressively wider as one moves away from the model center

The 6 assumptions required for least squares model confidence intervals

1. the model has a linear structure
2. the errors have constant variance at all levels of x
3. e_i values are normally distributed
4. each error, e_i, is independent of the other
5. x_i values are fixed and independent of the error
6. all y_i values are independent of each other

They are used in the derivations for the confidence intervals of β_0 and β_1

- mild violations of the 6 assumptions can be tolerated

Outliers are the more interesting data (usually). Many assumptions are violated because of outliers. And: **residuals contain the clues to problems with the model**.

Dealing with the 4 major least squares assumptions

The assumptions that we will investigate:

1. normally distributed residuals
2. non-constant error variance
3. independence in the data
4. model specification (linearity)

The procedure we will follow

- How to detect a violation of the assumption
  - "**Detecting it**"
- Dealing with the problem
  - "**Dealing with it**"
- How to know we have successfully resolved it
  - "**Everything OK when**"

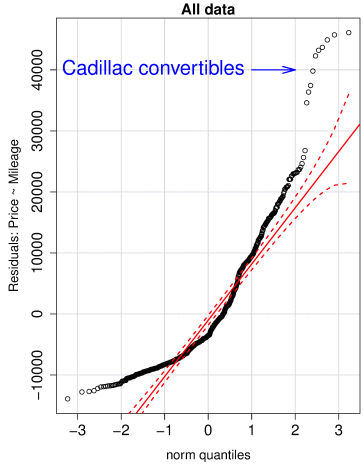
The assumption of "normally distributed residuals"

If non-normal, the S_E and other variances such as $\mathcal{V}(b_1)$ are too large or too small. This impacts the confidence interval interpretation.

Detecting it:

- Use a q-q plot: should match 45 degree line
  - Most accurate way
- A histogram is somewhat helpful
  - Human eye not good at picking up heavy tails
- Do *not* plot the residuals in time-order to try detect if normally distributed
  - Human eye not good at picking up heavy tails

The assumption of "normally distributed residuals"



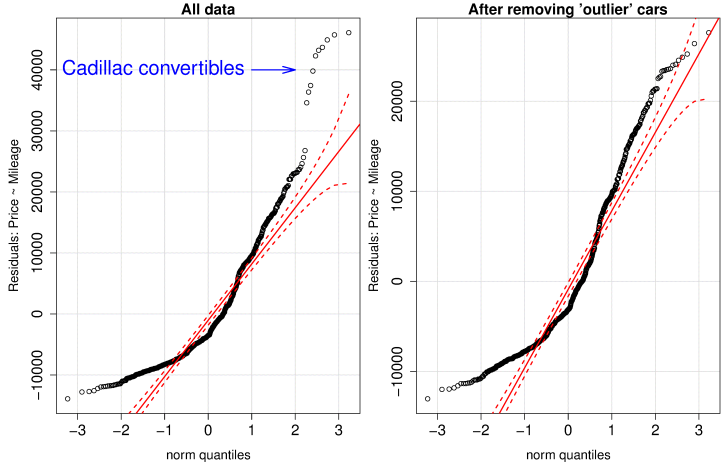
Before : $b_1 = -0.173$ $\quad -0.255 \leq \beta_1 \leq -0.0898$

$S_E = 9789$

After : $b_1 = -0.155$ $\quad -0.230 \leq \beta_1 \leq -0.0807$

$S_E = 8655$ $\quad \longleftarrow$ it has been reduced

The assumption of "normally distributed residuals"



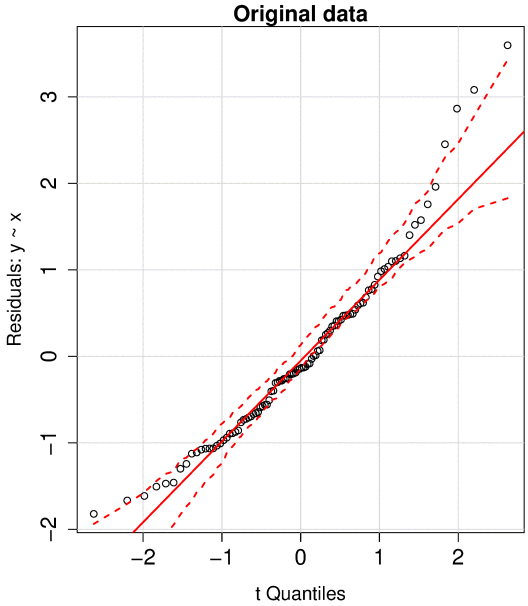
Before : $b_1 = -0.173$ $\quad -0.255 \leq \beta_1 \leq -0.0898$

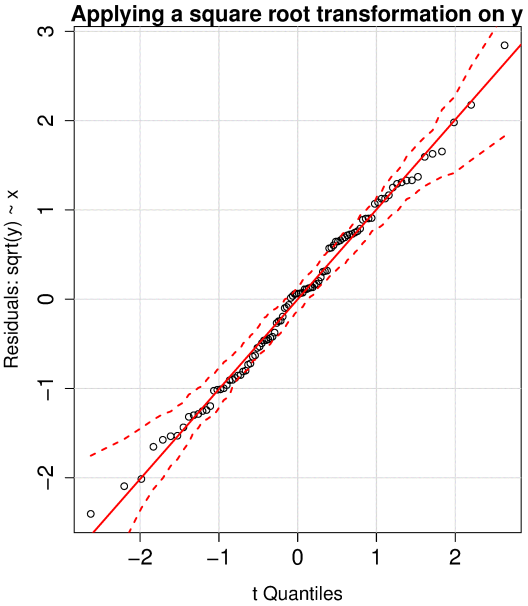
$S_E = 9789$

After : $b_1 = -0.155$ $\quad -0.230 \leq \beta_1 \leq -0.0807$

$S_E = 8655$ $\quad \longleftarrow$ it has been reduced

The assumption of "normally distributed residuals"





The assumption of "normally distributed residuals"

Dealing with it:

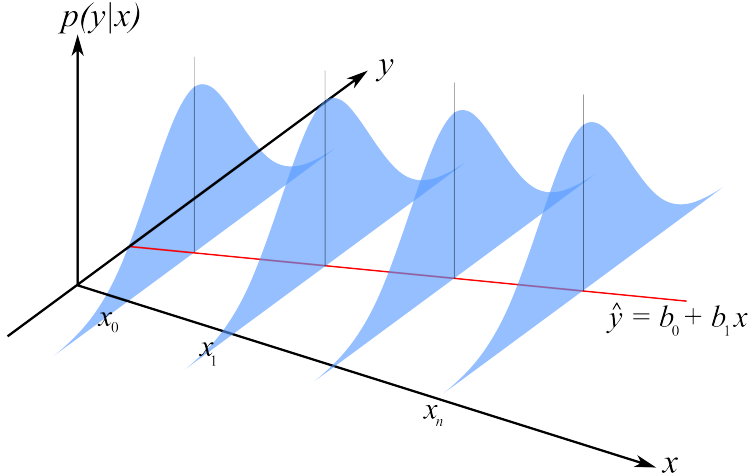
- Remove any outlying observation(s) (after investigation)
- Transform the y-variable

We will see in the section on multiple linear regression that non-normal residuals means you have perhaps forgot a term in the model: should you maybe add an extra variable in the equation?

Everything OK when:

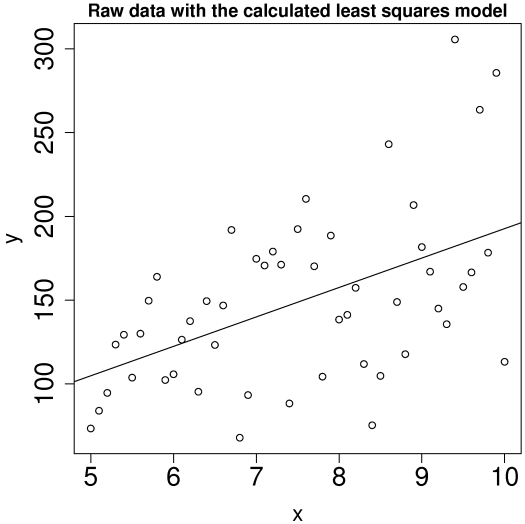
- q-q plot has normally-distributed residuals

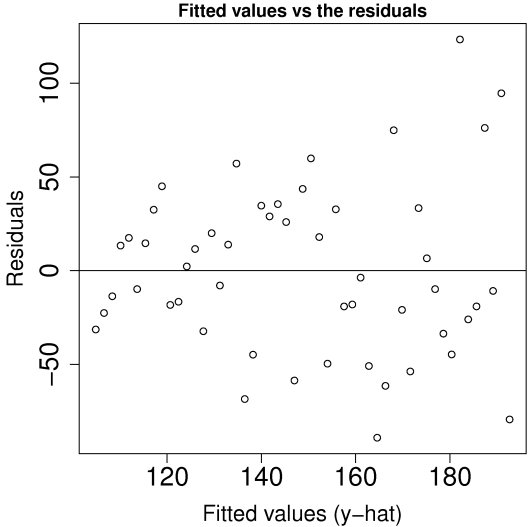
The assumption of 'Non-constant error variance"



Variability in y should be constant at all levels of x. If not, it increases S_E, undermining confidence intervals.

There is non-constant error variance in this example

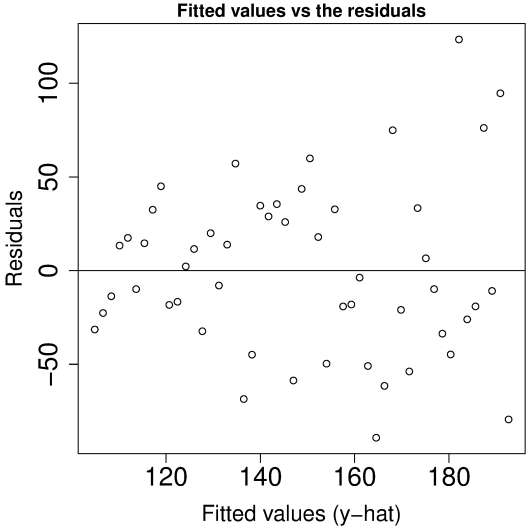


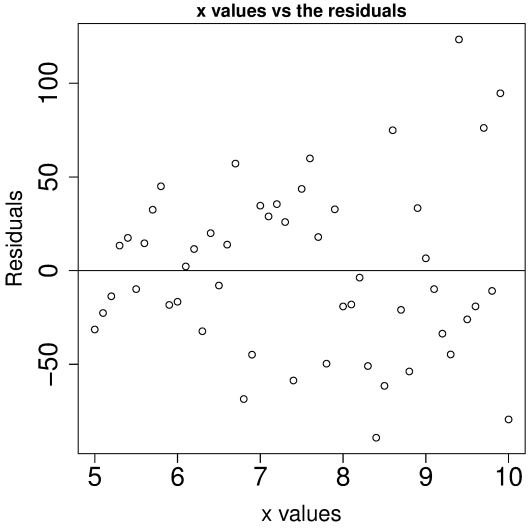


Detecting it:

- plot $\hat{y}$ (on the horizontal axis) against residuals (on the vertical axis)

There is non-constant error variance in this example





Detecting it:

- plot $\hat{y}$ or x (on the horizontal axis) against residuals (on the vertical axis)

The assumption of 'Non-constant error variance"

Detecting it:

- plot $\hat{y}$ (on the horizontal axis) against residuals (on the vertical axis)
- plot x (on the horizontal axis) against residuals (on the vertical axis)
- look for fan-shapes, and regions of non-constant variance

Dealing with it:

- Use weighted least squares: $f(\mathrm{b}) = \sum_i^n (w_i e_i)^2$
- Weights inversely proportional to the variance
- See: *Draper and Smith* (p 224 to 229, 3rd edition)
- Not usually too problematic if this assumption is violated (LS is pretty robust)

Everything OK when:

- no visible structure in the above plots

Dealing with the 4 major least squares assumptions

The assumptions that we will investigate:

1. normally distributed residuals
2. non-constant error variance
3. independence in the data
4. model specification (linearity)

The procedure we will follow

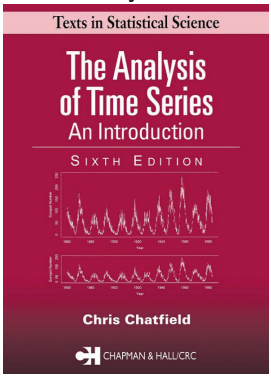
- How to detect a violation of the assumption
  - "**Detecting it**"
- Dealing with the problem
  - "**Dealing with it**"
- How to know we have successfully resolved it
  - "**Everything OK when**"

The assumption of "Independence in the data"

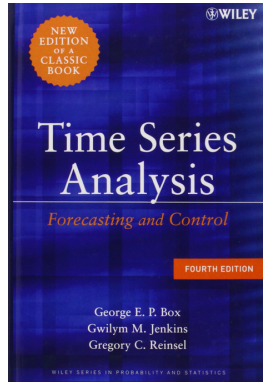
- it is often violated in data sampled from engineering systems; especially on a slow-moving process.
- it inflates/deflates the variances estimates used by the Central limit theorem

Two useful and highly recommended resources on this topic:

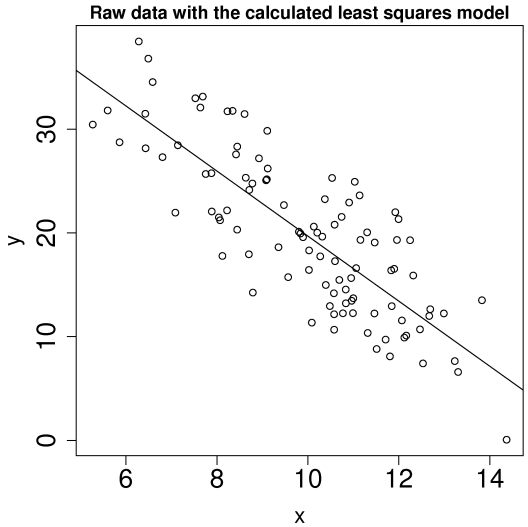
Chatfield: "The Analysis of Time Series"

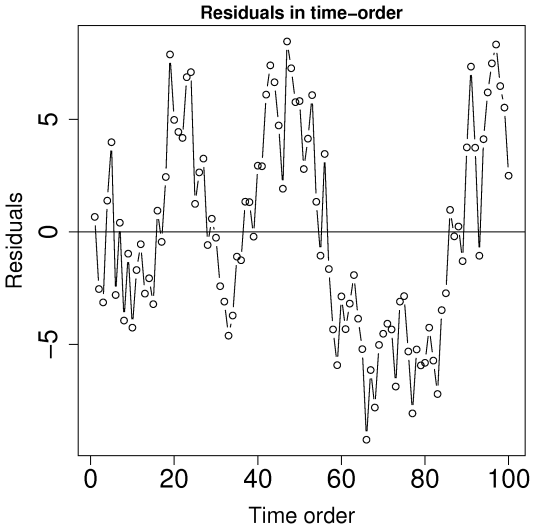


Box and Jenkins: "Time Series Analysis"



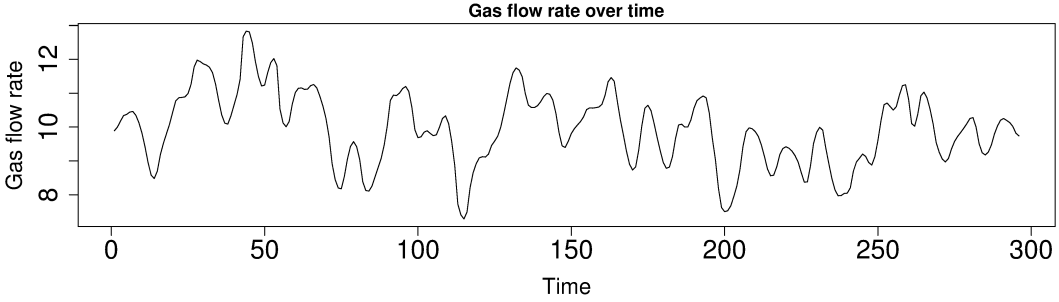
How to detect lack of independence in the data





How to detect lack of independence in the data

Look for any patterns in the data:



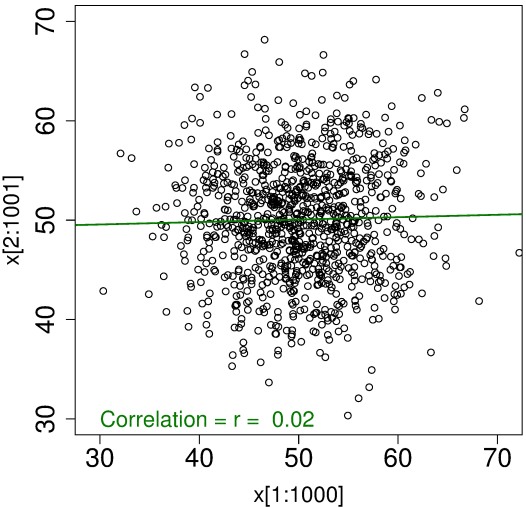
- cycling (as shown above)
- rapid alternation (crisscrossing)
- slow drifts

We cannot tell if data are independent by looking at a plot, but we can tell if they are **not independent** if we see these patterns.

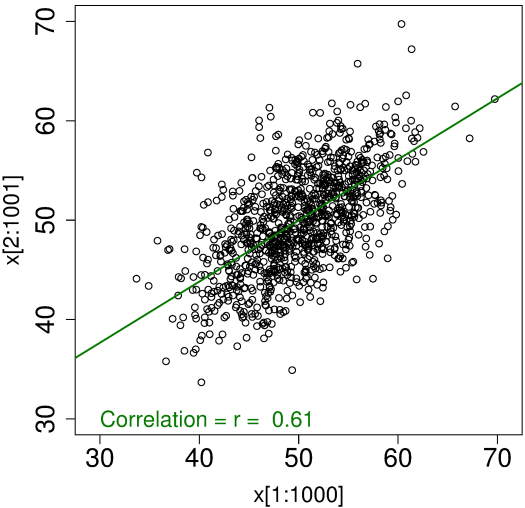
Understanding lack of independence: introducing the "autocorrelation" concept

- detect if samples are related to each other
- uses scatterplots to do this

Understanding and interpreting the autocorrelation: 1 sample apart

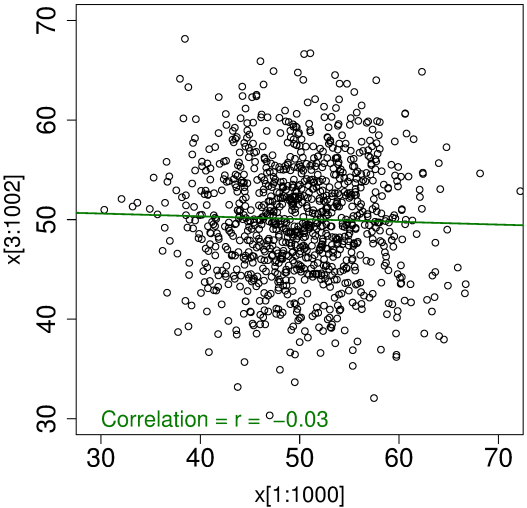



```
x <- rnorm(1004, 50, 6)
plot(x[1:1000], x[2:1001])
```

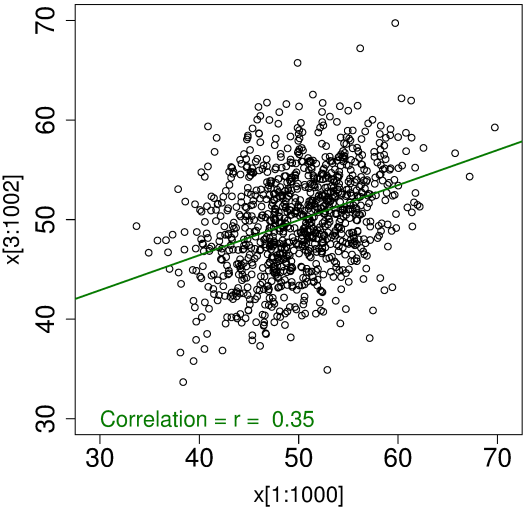



```
x <- read.csv('acf-data.csv')
plot(x[1:1000], x[2:1001])
```

Understanding and interpreting the autocorrelation: 2 samples apart

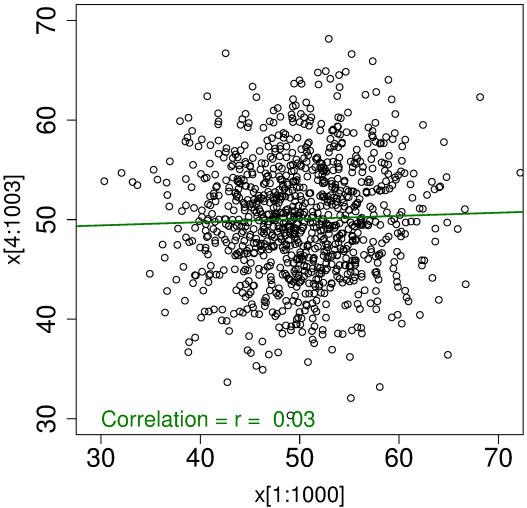



```
plot(x[1:1000], x[3:1002])
```

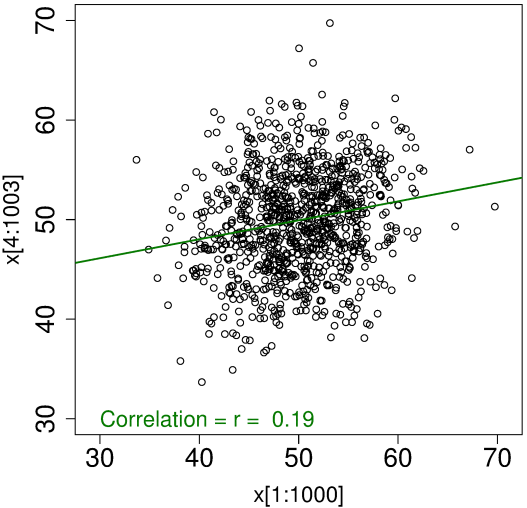



```
plot(x[1:1000], x[3:1002])
```

Understanding and interpreting the autocorrelation: 3 samples apart

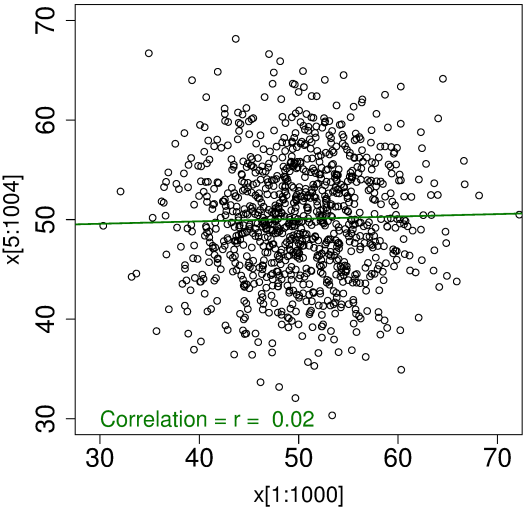



```
plot(x[1:1000], x[4:1003])
```

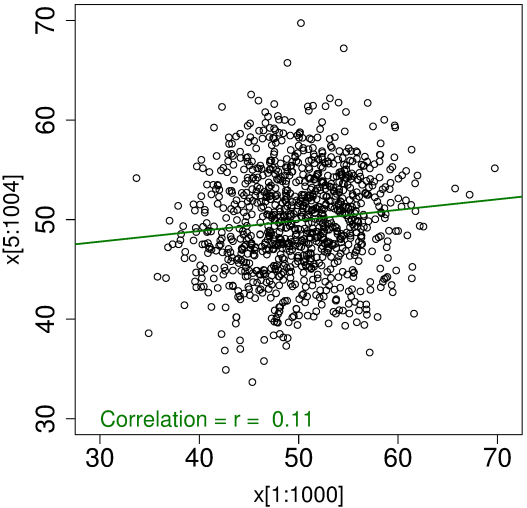



```
plot(x[1:1000], x[4:1003])
```

Understanding and interpreting the autocorrelation: 4 samples apart

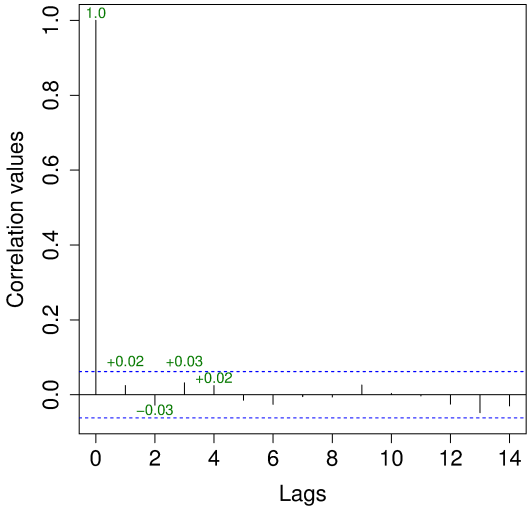



```
plot(x[1:1000], x[5:1004])
```

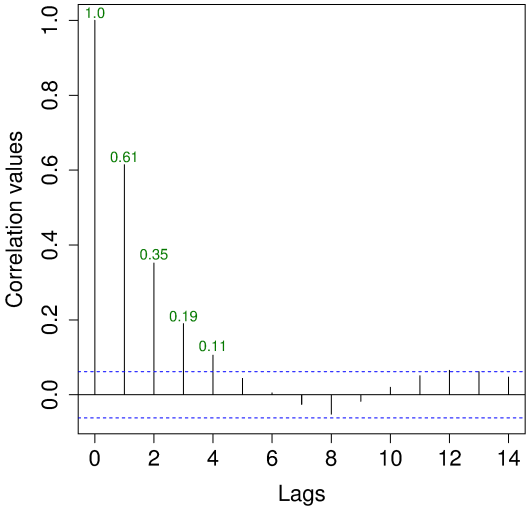



```
plot(x[1:1000], x[5:1004])
```

Understanding and interpreting the autocorrelation plot

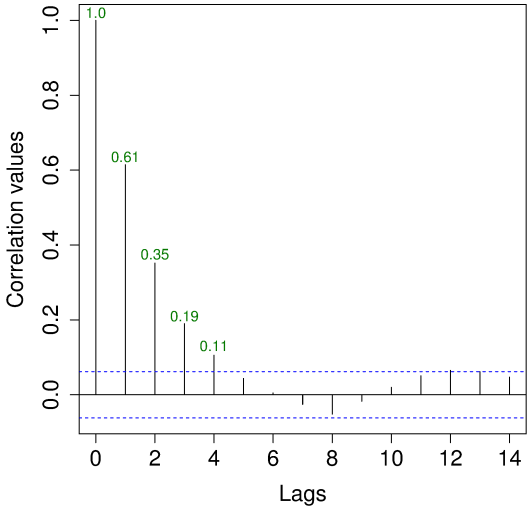


`acf(x)`

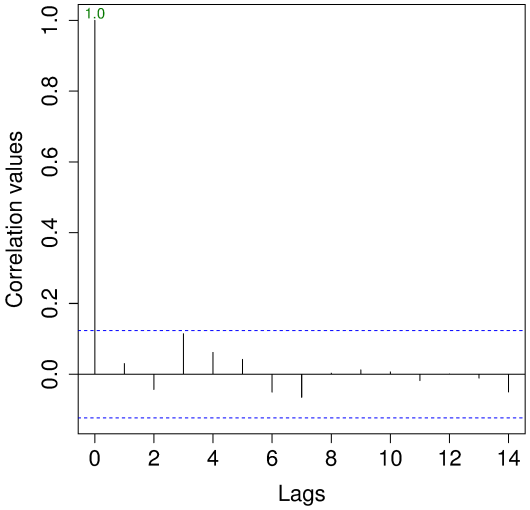


`acf(x)`

Using the autocorrelation function to sub-sample the data




```
acf(x)
```




```
x.subsample <- x[seq(1, length(x), 4)]
acf(x.subsample)
```

The assumption of "Independence in the data"

Detecting it:

- by observing the raw data plots
- use the autocorrelation plot
- advanced students: investigate the Durbin-Watson test (*Draper and Smith*, chapter 7)

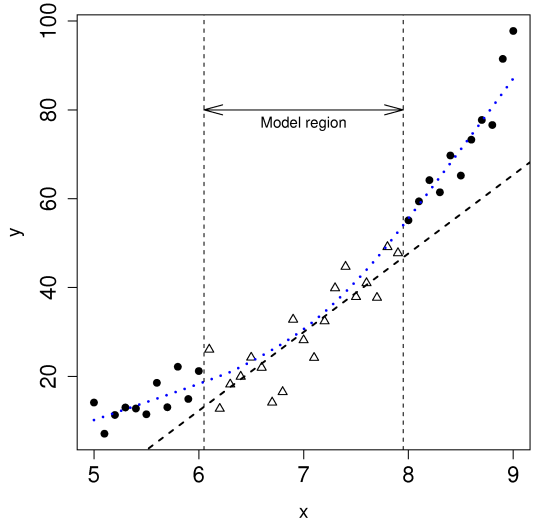
Dealing with it: sub-sample the data, every k^{th} sample

Everything OK when:

- `acf(y)` and `acf(e)` show no lags beyond lag 0
- where `e` is the model residuals

The assumption of "Correct model specification"

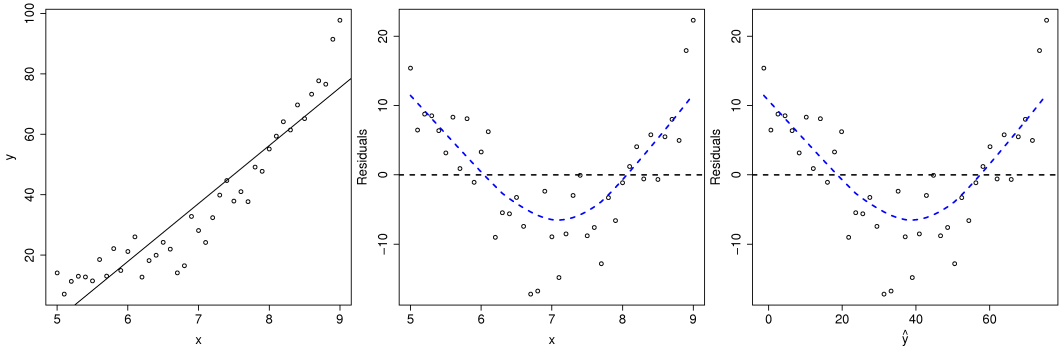
Systems are known to be non-linear; linear might be good enough (depends on model's purpose)



The assumption of "Correct model specification"

Detecting it:

- plot $\hat{y}$ against the residuals (y-axis)
- plot x against residuals
- q-q plot might show it also



Use the `lowess(...)` function to get the dashed blue line

The assumption of "Correct model specification"

Dealing with it:

1. Non-linear least squares:
   - *Example*: $f(x) = y = b_1\left(1 - e^{-b_2 x}\right)$
   - Objective function $= \sum e_i^2 = \sum \left(y_i - b_1\left(1 - e^{-b_2 x_i}\right) \right)^2$
   - Differentiating this and solving it is dependent on the function $f(x)$
   - Investigate the `nls(...)` function in R
   - This topic is a standalone topic of study [too broad for this course]

The assumption of "Correct model specification"

Dealing with it:

2. Transform the x or y variable; then use linear model
   - $x_{\text{transformed}} \leftarrow x^{p}_{\text{original}}$
   - *Example*: use $x_{\text{transformed}} \leftarrow \sqrt{x_{\text{original}}}$

Don't use trial-and-error; there is a process you should follow.

$$x_{\text{transformed}} \leftarrow x^{p}_{\text{original}}$$

- Base case: $p = 1$
- Stepping up the ladder p: $1, \quad 1.5, \quad 1.75, \quad 2.0, \quad$ *etc*
- Stepping down the ladder p: $1, \quad 0.5, \quad -0.5, \quad -1.0, \quad -1.5, \quad -2.0, \quad$ *etc*
- $\log(x)$: approximates $p = 0$ in terms of severity

The assumption of "Correct model specification"

3. Rearrange first-principles equations
   - Distillation: T inversely proportional to log(VP). $y = b_0 + b_1 x$:
     - $x \leftarrow 1/T$
     - $y \leftarrow \log(P)$.
     - The slope coefficient $=$
   - $y = p \times q^x$; take logs so that $\log(y) = \log(p) + x \log(q)$,
     - Slope coefficient $= \log(q)$
   - $y = \dfrac{1}{p + qx}$, invert to get: $y = b_0 + b_1 x$:
     - $b_0 \leftarrow p$
     - $b_1 \leftarrow q$
     - $y \leftarrow 1/y$

Summary: The assumption of "Correct model specification"

Everything OK when:

- No more structure in the detection plots

We can try using some/combination of these tools to help:

1. Use nonlinear least squares
2. Apply transformations *systematically*
3. Linearize the equation

Residuals (errors) play a crucial role in the linear model. They are initially the most interesting feature of a model.

Multiple linear regression (MLR)

AIM: We want to include more than one x input variable in the model

- most real systems have more than 1 factor affecting the output, y

Why do we build regression models?

1. **Improve our understanding of systems:**
   - $x_1 =$ reactant concentration
   - $x_2 =$ temperature
   - $y =$ reaction rate, where $y = b_0 + b_1 x_1 + b_2 x_2$
   - Used to understand the effect of concentration: b_1
   - Used to understand the effect of temperature: b_2
2. **Improve our model's predictions:**
   - $x_1 =$ temperature
   - $x_2 =$ feed flow rate
   - $y =$ melt index, where $y = b_0 + b_1 x_1 + b_2 x_2$
   - Less accurate with $y = b_0 + b_1 x_1$ ⟵ single-variable least squares
   - Less accurate with $y = b_0 + b_2 x_2$ ⟵ single-variable least squares
   - Better predictions with $y = b_0 + b_1 x_1 + b_2 x_2$ ⟵ multiple linear regression

Integer variables with 3 or more levels will automatically lead to a multiple linear regression model

Examples of integer variables at more than 2 levels:

- 3 mixing tanks: for example `TK-104`, `TK-107` and `TK-108`
- 3 or more operators
- multiple suppliers (we will see an example later on)

What lies ahead, and how it is similar to what you have done

We will see this most important equation:

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Similarities with MLR (*multiple linear regression*) and OLS (*ordinary least squares*):

- Objective function
- How it is solved
- Interpretation of the slope coefficients
- Confidence intervals

Introducing matrix notation for multiple linear regression (MLR)

We will remove the intercept by centering the data:

$$\begin{aligned} y_i &= b_0 + b_1 x_i \\ \overline{y} &= b_0 + b_1 \overline{x} \\ y_i - \overline{y} &= 0 + b_1 (x_i - \overline{x}) \end{aligned}$$

by subtracting previous lines

- Let $x = x_{\text{original}} - \text{mean}\,(x_{\text{original}})$
- Let $y = y_{\text{original}} - \text{mean}\,(y_{\text{original}})$
- Model is still the same, except intercept term is forced to zero: $b_0 = 0$

- Intercept can always be recovered afterwards, if required
- Using these deviation variables is optional, and not always done

- Centered data is more interpretable:
  - $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$ ⟵ these are "variances" and "covariances"

Matrix notation for multiple linear regression (MLR): 1 observation

So now our general linear model (without intercept) is given by:

$$
\begin{aligned}
y_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k} + \varepsilon_i \\
y_i &= [x_{i,1},\ x_{i,2},\ \ldots,\ x_{i,k}] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon_i \\
y_i &= \underbrace{x_i^T}_{(1\times k)} \underbrace{\beta}_{(k\times 1)} + \varepsilon_i
\end{aligned}
$$

Matrix notation for multiple linear regression (MLR): n observations

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,k} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

- $\mathbf{y}$: $n \times 1$
- $\mathbf{X}$: $n \times k$
- $\mathbf{b}$: $k \times 1$
- $\mathbf{e}$: $n \times 1$

Estimating the model parameters via optimization

Objective function: minimize sum of squares of the errors

$$\begin{aligned} f(\mathbf{b}) &= \mathbf{e}^T\mathbf{e} \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{b} + \mathbf{b}\mathbf{X}^T\mathbf{X}\mathbf{b} \end{aligned}$$

- Solved by setting $\frac{f(\mathbf{b})}{\partial \mathbf{b}} = 0$
- this is k equations in k unknowns: $[b_1, b_2, \ldots, b_k]$
- k is the number of parameters estimated in the model

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Let's look at an example

Original variables

$x_{1,\text{original}} = [1, \quad 3, \quad 4, \quad 7, \quad 9, \quad 9]$

$x_{2,\text{original}} = [9, \quad 9, \quad 6, \quad 3, \quad 1, \quad 2]$

$y_{\text{original}} = [3, \quad 5, \quad 6, \quad 8, \quad 7, \quad 10]$

Centered variables

$x_1 = [-4.5, \ -2.5, \ -1.5, \ 1.5, \ 3.5, \ 3.5]$

$x_2 = [4, \ 4, \ 1, \ -2, \ -4, \ -3]$

$y = [-3.5, \ -1.5, \ -0.5, \ 1.5, \ 0.5, \ 3.5]$

$$\mathbf{X} = \begin{bmatrix} -4.5 & 4 \\ -2.5 & 4 \\ -1.5 & 1 \\ 1.5 & -2 \\ 3.5 & -4 \\ 3.5 & -3 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} -3.5 \\ -1.5 \\ -0.5 \\ 1.5 \\ 0.5 \\ 3.5 \end{bmatrix}$$

- $\mathbf{y}$: $n \times 1$
- $\mathbf{X}$: $n \times k$
- $\mathbf{b}$: $k \times 1$
- $\mathbf{e}$: $n \times 1$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 55.5 & -57.0 \\ -57.0 & 62 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 36.5 \\ -36.0 \end{bmatrix} \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.323 & 0.297 \\ 0.297 & 0.289 \end{bmatrix}$$

Learning from our example

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 55.5 & -57.0 \\ -57.0 & 62 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 36.5 \\ -36.0 \end{bmatrix}$$

- $\mathbf{X}^T\mathbf{X}$: is a $k \times k$ matrix
- scaled version of the covariance matrix of $\mathbf{X}$
- Off-diagonal entries: symmetrical, strength of relationship between variables
- Diagonal entries: (co)variance, always positive
- What does $\mathbf{X}^T\mathbf{X}$ look like for uncorrelated variables?
- What does $\mathbf{X}^T\mathbf{y}$ represent?
- Real data sets can cause a problem when calculating $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$
  - Use the QR decomposition instead

Back to the standard error again

Objective function:

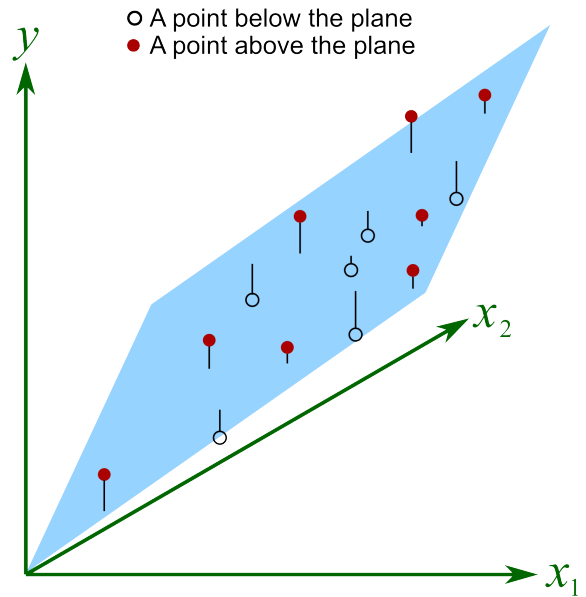
$$f(\mathbf{b}) = \mathbf{e}^T\mathbf{e}$$

The standard error $= S_E = \sqrt{\mathcal{V}\{\mathbf{e}\}} = \sqrt{\dfrac{\mathbf{e}^T\mathbf{e}}{n-k}}$

- where $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$
- since the mean of the errors $= 0$
- and degrees of freedom $= n - k$

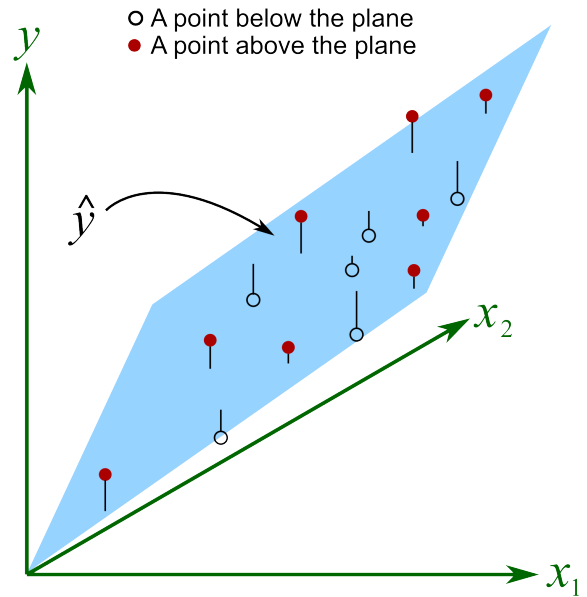
Interpretation of the model coefficients with the 2-variable example:

$y = b_1 x_1 + b_2 x_2$



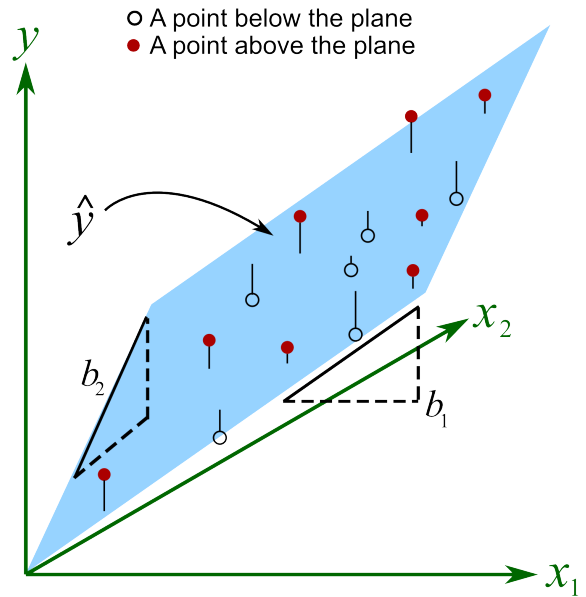
Interpretation of the model coefficients with the 2-variable example:

$y = b_1 x_1 + b_2 x_2$



Interpretation of the model coefficients with the 2-variable example:

$y = b_1 x_1 + b_2 x_2$



Interpretation of the model coefficients with the 2-variable example: $y = b_1 x_1 + b_2 x_2$

General interpretation

Coefficient b_1 is the average change in $\mathbf{y}$ for a one unit change in x_1 **provided we hold x_2 fixed**

Example

$$y = b_V V + b_T T$$

$$y = -0.5V + 4.2T$$

- V = reactor tank volume, measured in L
- T = reactor temperature, measured in Kelvin
- y = yield in μg

Interpretation of $b_V = -0.5$? "0.5 μg decrease in yield, on average, for every 1 L increase in volume, holding the temperature fixed"

Interpretation of the confidence intervals in the model

$$y = b_V V + b_T T = -0.5V + 4.2T$$

Interpreting the volume effect,
$b_V = -0.5$

Confidence interval for b_V spans zero: the effect of the volume, controlling for temperature, is not significant.

Interpreting the temperature effect,
$b_T = +4.2$

Confidence interval for b_T does not span zero: the effect of the temperature, controlling for volume, is to increase the yield by 4.2 μg, on average, for every 1 K increase in temperature, T.

The "controlling for" indicates the controlled variable was used in the model.

Examples of integer variables in a least squares regression model

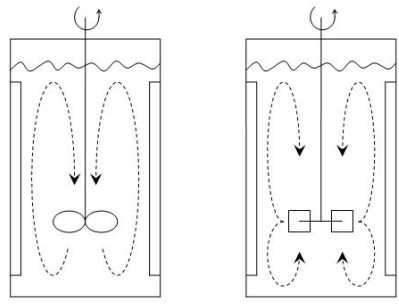
We regularly come across this problem.

- method A or method B used to process raw materials
  - it is a categorical variable
  - it cannot be in between
- indicates a particular feature:
  - convertible car, *or*
  - regular car
- we know that a particular characteristic affects the output, y, differently:
  - morning shift is more productive
  - than the evening shift

The integer variable example that we will use in this video

Yield $= y = f$(temperature, impeller type)

"impeller type" is a categorical variable, as shown on the right



[From Wikipedia]

Build two models: one for radial, one for axial

- Not an efficient use of the data
- Temperature effect is the same in both cases (independent of impeller)
- Fewer degrees of freedom in each separate model
- Increases S_E, and therefore we get larger confidence intervals

Conclusion: a unified model is more desirable.

The integer variable example that we will use in this video

$$
\begin{aligned}
y &= \beta_0 + \beta_1 T + \gamma d + \varepsilon \quad &\longleftarrow \text{population model} \\
y &= b_0 + b_1 T + g d_i + e_i \quad &\longleftarrow \text{statistical model}
\end{aligned}
$$

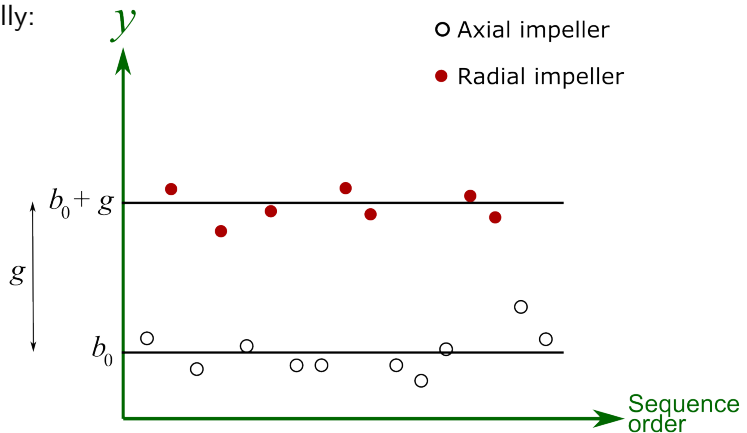
- Let: $d_i = 0$ for axial impeller
- Let: $d_i = 1$ for radial impeller

Interpretation of the integer variable in the least squares model

Assume $\beta_1 = 0$ (temperature has no effect) for now.

$$y = b_0 + \cancel{b_1 T} + g d_i \qquad = \qquad b_0 + g d_i$$

Geometrically:



Axial impeller:	$d_i = 0$	$y = b_0 + 0$	$y = b_0$
Radial impeller:	$d_i = 1$	$y = b_0 + g d_i$	$y = b_0 + g$

Interpretation of the integer variable in the least squares model

Now let $\beta_1 \neq 0$:

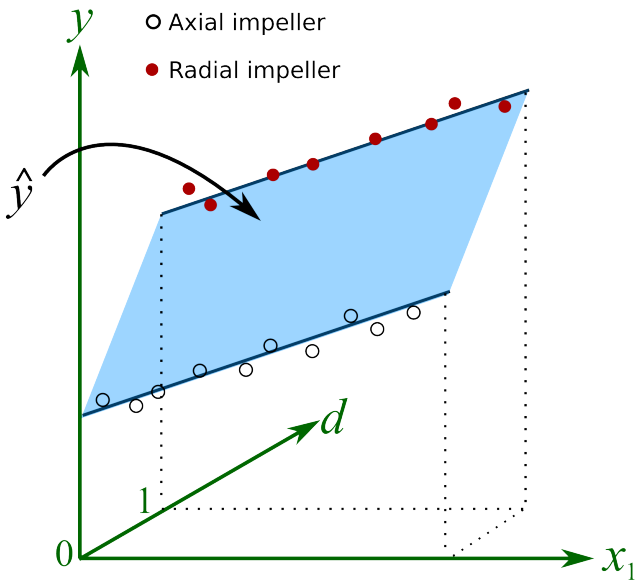
$$
\begin{aligned}
y &= \beta_0 + \beta_1 T + \gamma d + \varepsilon \quad \longleftarrow \text{population model} \\
y &= b_0 + b_1 T + g d_i + e_i \quad \longleftarrow \text{statistical model}
\end{aligned}
$$

Axial impellers: $y = b_0 + b_1 T + 0$

Radial impellers: $y = b_0 + b_1 T + g$

- If $g = -56\mu$g: the decrease in yield is expected to be 56 μg, on average, when changing from an axial to a radial impeller, controlling for temperature.

Geometric interpretation of the integer variable in the least squares model



$$y = \beta_0 + b_1 T + g d_i$$

- Slope coefficient: like the regular interpretation except it is the "incremental effect"
- Confidence interval for integer variables: no different to other variables

Interpretation of integer variables in the least squares model

Raw material from Spain, India, or Vietnam [3 levels]

2 integer variables: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \boldsymbol{\gamma_1 \mathbf{d_1}} + \boldsymbol{\gamma_2 \mathbf{d_2}} + \varepsilon$

- $d_{i1} = 0$ and $d_{i2} = 0$ for Spain
- $d_{i1} = 1$ and $d_{i2} = 0$ for India
- $d_{i1} = 0$ and $d_{i2} = 1$ for Vietnam
- Interpret coefficients relative to the (0,0) baseline for Spain

Summary of integer variables in a linear model

- *Slope coefficient*: like the regular least squares interpretation except it is the "incremental effect"
- Confidence interval for integer variables: interpreted no different to other variables