Statistics for Engineers

ChE 4C3 and 6C3

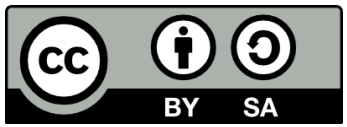
© Kevin Dunn, 2014

kevin.dunn@mcmaster.ca
http://learnche.mcmaster.ca/4C3

Overall revision number: 119 (January 2014)

Copyright, sharing, and attribution notice

This work is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License. To view a copy of this license, please visit http://creativecommons.org/licenses/by-sa/3.0/



This license allows you:

- **to share** - to copy, distribute and transmit the work, including print it
- **to adapt** - but you must distribute the new result under the same or similar license to this one
- **commercialize** - you *are allowed* to use this work for commercial purposes
- **attribution** - but you must attribute the work as follows:
  - "Portions of this work are the copyright of Kevin Dunn", *or*
  - "This work is the copyright of Kevin Dunn"

  (when used without modification)

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: 119

Please note that all material is provided "as-is" and no liability will be accepted for your usage of the material.

Least squares models in context ...

- A new development: we start looking at more than one variable at a time.
- 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.

Where are we headed

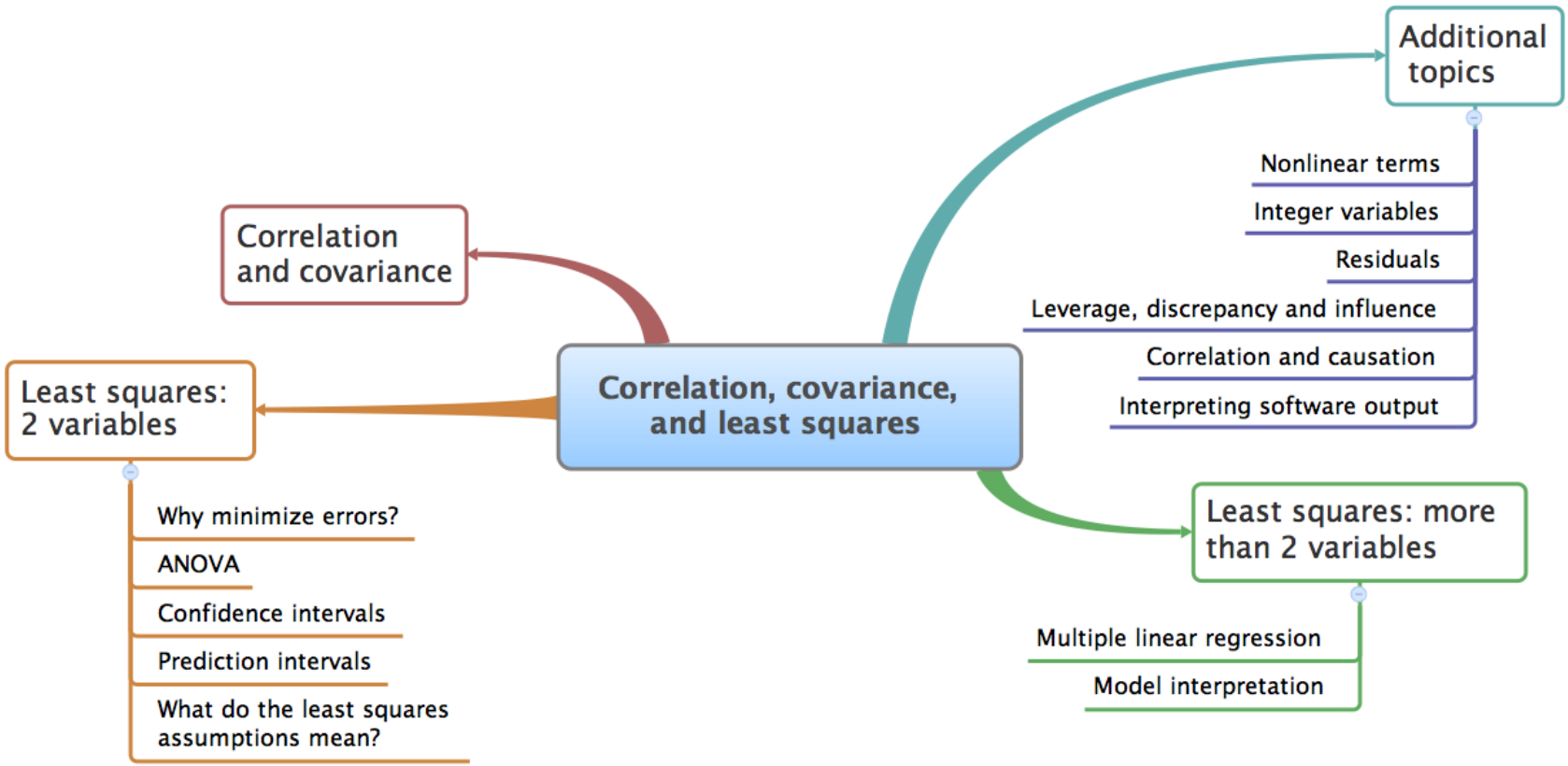
- You've seen the least squares model hundreds of times:

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x} + \varepsilon$$

  - What are confidence intervals for β_0 and β_1?
  - Why do we even care about these confidence intervals?
  - What does R^2 really mean?
  - How to judge predictions from the model? $\hat{y} \pm$ ___

What we will cover

1. Covariance
2. Correlation
3. Bivariate least squares (2 variables)
4. Multiple linear regression (2+ variables)
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*

Covariance: by example

What does the term covariance mean to you?

- 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 $\beta_1 = \dfrac{nR}{V} > 0$

Covariance

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	1320	43267	8.1

Covariance

- Formal definition:

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

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 temperature and humidity is: 202 [K.%]
  - *Awkward units*!

Covariance

- Covariance with itself is the variance:

$$\text{Cov}\{x, x\} = \mathcal{V}(x) = \mathcal{E}\{(x - \overline{x})(x - \overline{x})\}$$

- (Co)variance of a centered vector = (co)variance of the uncentered vector
- What does $x^T x$ represent?
- And $x^T y$?

- Go through the examples in the notes
  - x=hours worked per week and y=take home pay
  - x=age of married partner 1 and y=age of married partner 2
  - x=cigarettes smoked per month and y=age at death
  - x=temperature on top tray of distillation column and y=top product purity

- Covariance and correlation describe overall tendency, outliers don't follow the expected pattern.

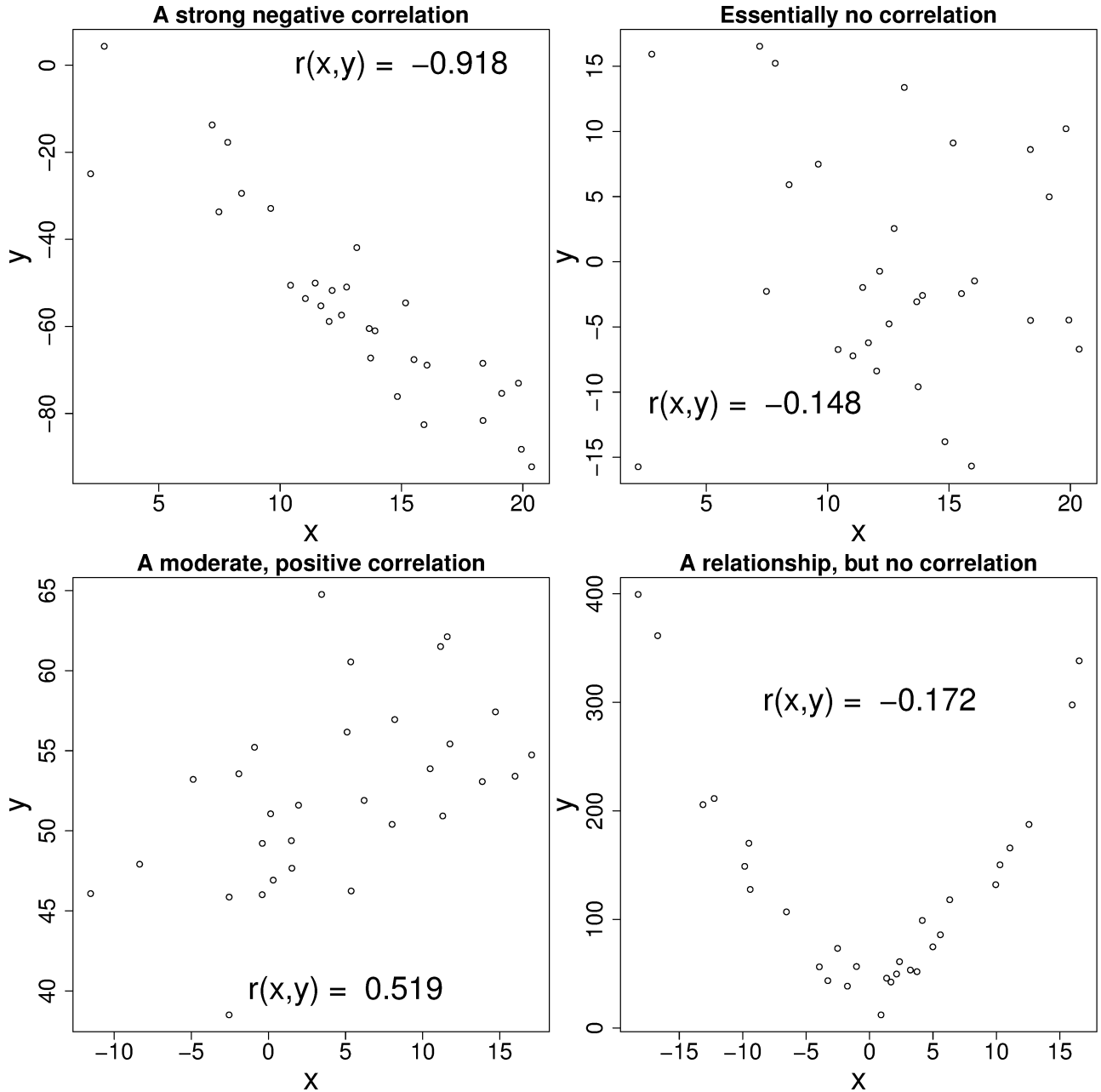
Correlation: measures "co-relationship"

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

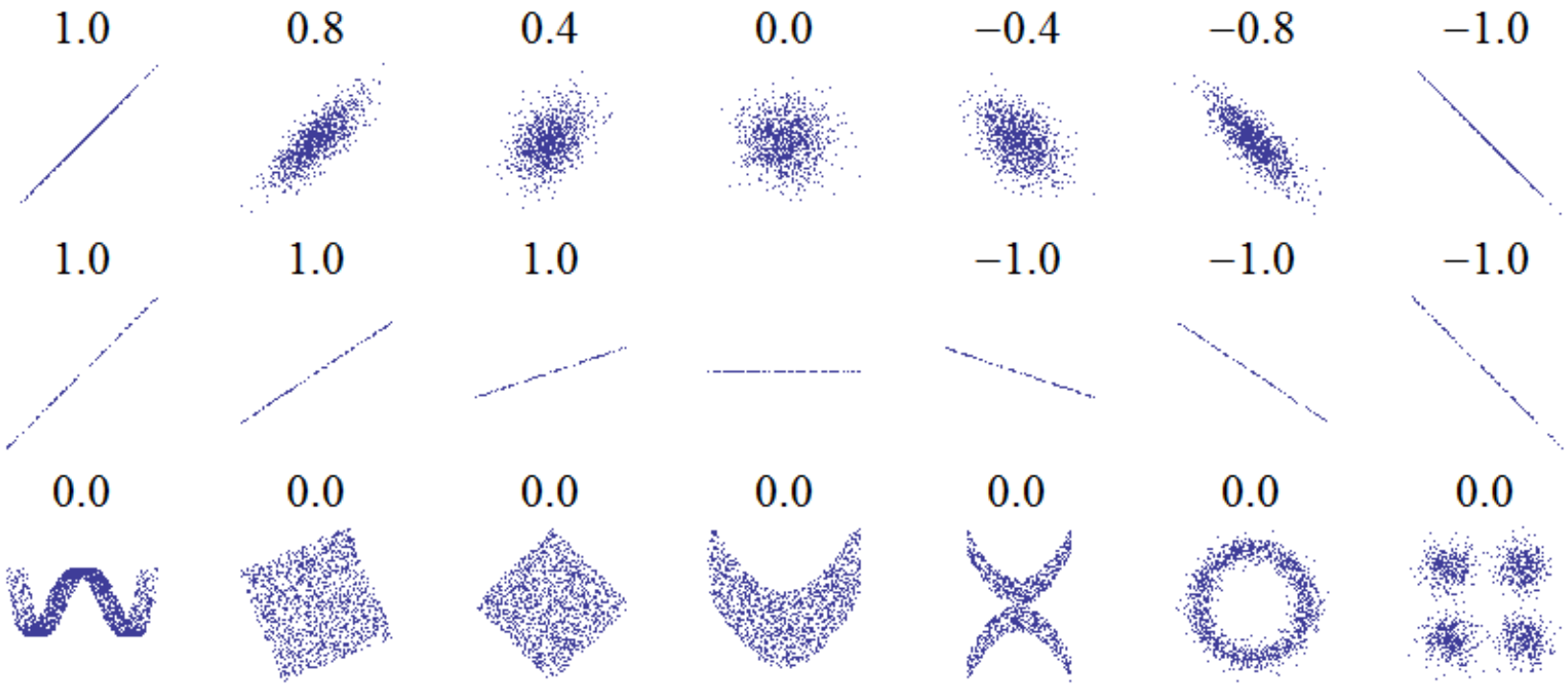
$$r(x,y) = \frac{\mathcal{E}\left\{(x-\overline{x})(y-\overline{y})\right\}}{\sqrt{\mathcal{V}\left\{x\right\}\mathcal{V}\left\{y\right\}}} = \frac{\mathrm{Cov}\left\{x,y\right\}}{\sqrt{\mathcal{V}\left\{x\right\}\mathcal{V}\left\{y\right\}}}$$

- Divides by the units of x and y: dimensionless result
- $-1 \leq r(x,y) \leq 1$
- Gas cylinder example:
  - $r(\text{temperature, pressure}) = 0.997$
  - $r(\text{pressure, temperature}) = 0.997$
  - $r(\text{temperature, humidity}) = 0.380$

Correlation examples



Correlation examples



[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 these.

Introduction to least squares: relating 2 variables

- Linear least squares model is a useful tool
- It is the basis for a number of algorithms in data analysis
  - DOE
  - Latent variable methods
  - Almost all other advanced methods build on its concepts

- We consider only 2 variables: x and y.

1. Model definition
2. Building (fitting) the model
3. Interpret model parameters and model outputs
4. Software

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.

Model definition

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

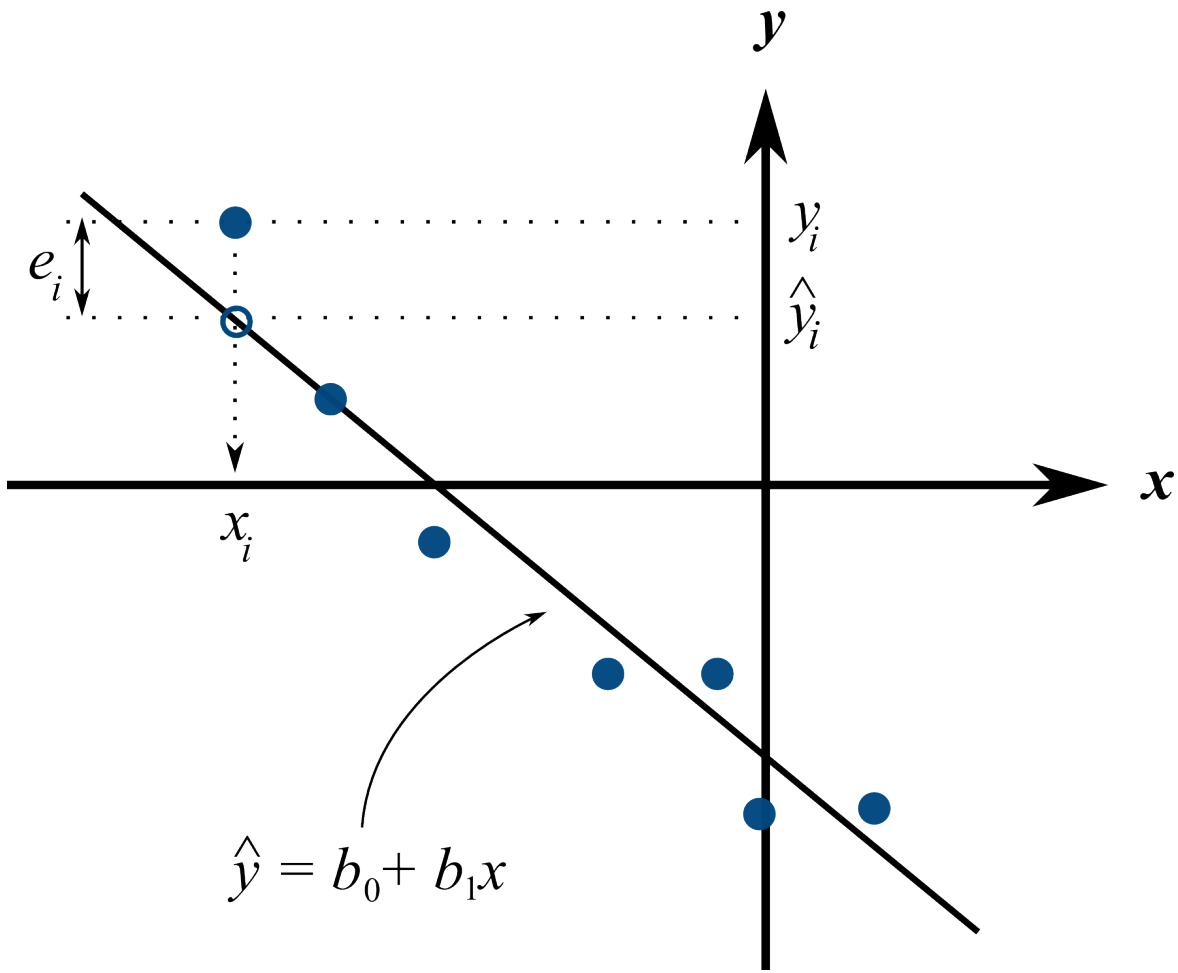
- $b_0 = \hat{\beta}_0$
- $b_1 = \hat{\beta}_1$
- $e = \hat{\varepsilon}$

So we can write:

$$\begin{aligned} y_i &= b_0 + b_1 x_i + e_i \\ \hat{y}_i &= b_0 + b_1 x_i \end{aligned}$$

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 $\left\{e_i^2\right\}$: 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

Solving the model

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)

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

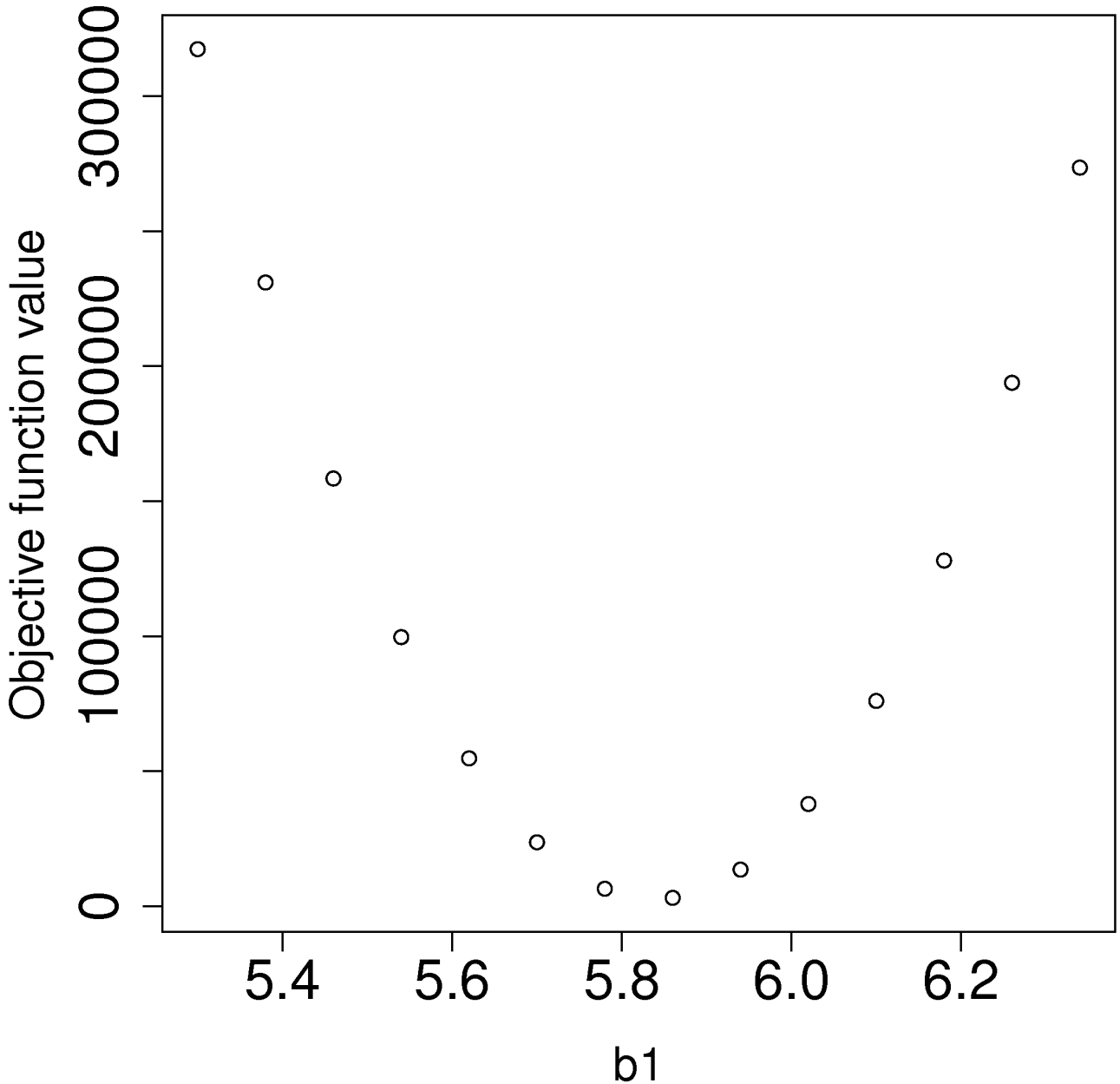
Solving the model: grid search

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

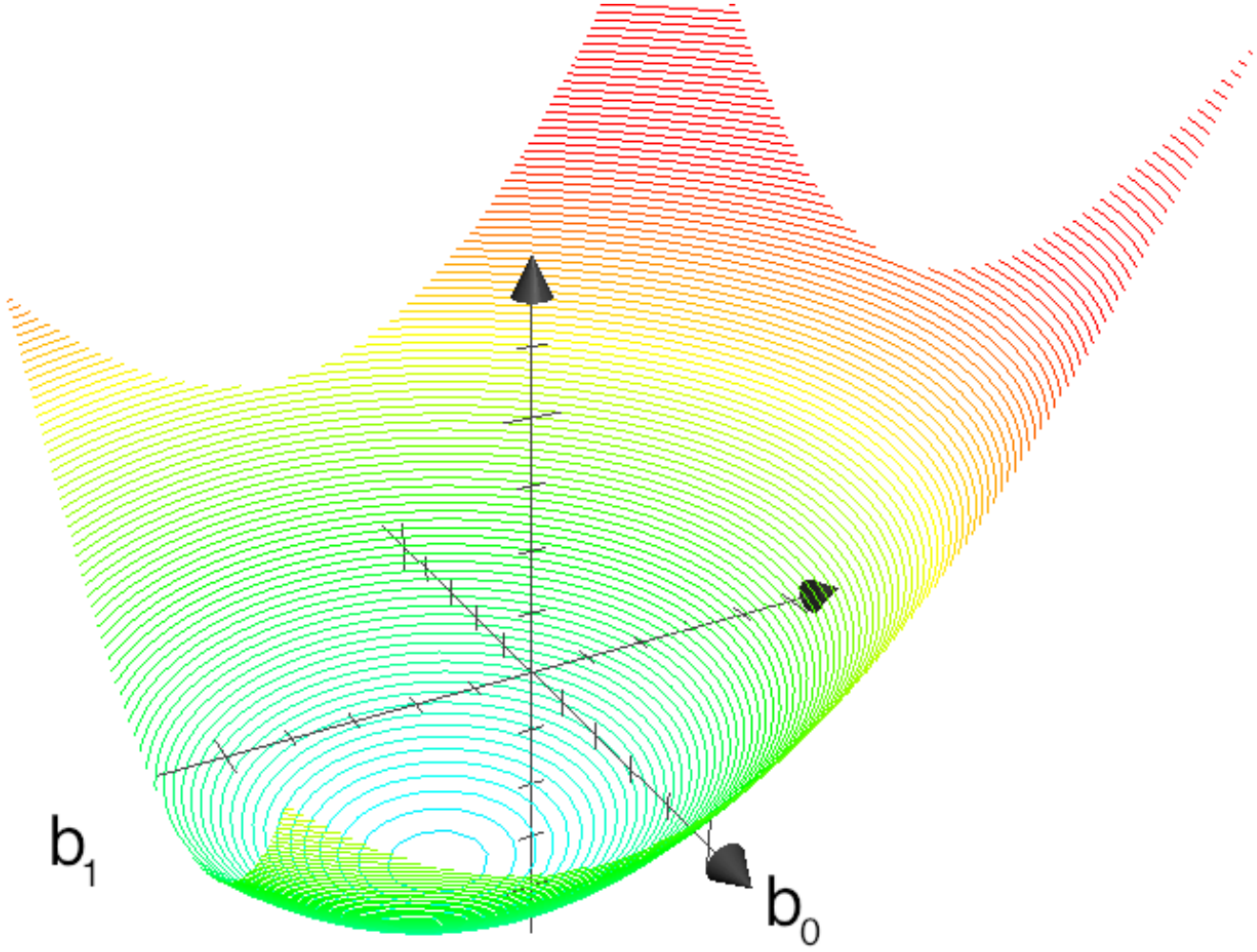
- We know that $\beta_0 = 0$ from theoretical principles
- Solve for β_1 by trial and error. Initial guess?
- $b_1 \approx \beta_1 = \dfrac{nR}{V} = \dfrac{(14.1 \text{ mol})(8.314 \text{ J/(mol.K)})}{20 \times 10^{-3} \text{ m}^3} = 5.861 \text{ kPa/K}$

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 model: grid search



Solving the model: grid search in 2 variables



- objective function shape is a bowl
- a unique minimum can always be found
- because the objective function is convex

Solving the model: analytically

Least squares objective function

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

At the optimal point we know that:

- The partial derivatives wrt b_0 and b_1 are zero
- This represents 2 equations in 2 unknowns
- Take the 2nd derivative to confirm that the optimum is indeed a minimum $\left(\frac{\partial^2 f}{\partial^2 b_0} > 0 \quad \text{and} \quad \frac{\partial^2 f}{\partial^2 b_1} > 0 \right)$

Solving the model: analytically

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

Least squares optimum for $f(b_0, b_1)$:

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

Remarks

Least squares solution

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

1. $\sum_i e_i = 0$
2. The straight line equation passes through $(\overline{\mathrm{x}}, \overline{\mathrm{y}})$ without error
3. Prove to yourself that $\sum_i (x_i e_i) = \mathrm{x}^T \mathrm{e} = 0$
   - The residuals are uncorrelated with the input variables, x
   - There is no information in the residuals that is in the x's
4. Prove and interpret that $\sum_i (\hat{y}_i e_i) = \hat{\mathrm{y}}^T \mathrm{e} = 0$
   - The fitted values are uncorrelated with the residuals
5. Estimate of b_0 depends on b_1: the estimates are correlated

Questions

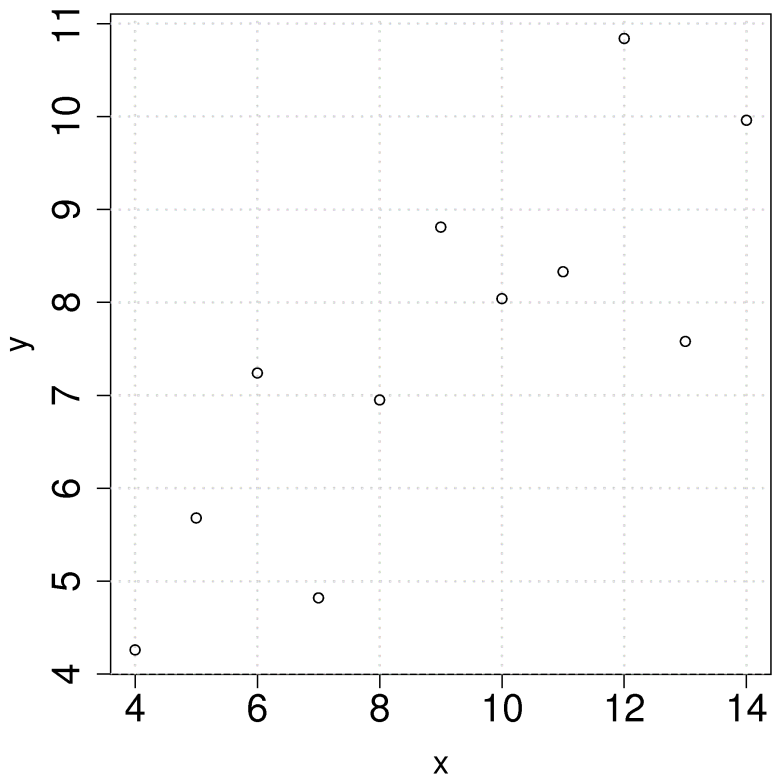
Least squares solution

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

1. Units of b_1 ?
   - The units of y divided by the units of x
2. Gas cylinder example. Let $\hat{p}_i = b_0 + b_1 T_i$:
   - What is the interpretation of coefficient b_1?
   - What is the interpretation of coefficient b_0?
3. How could the denominator term for b_1 equal zero? And what would that mean?
   - As long as there is variation in the x-data that we will obtain a solution

Example

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

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

Least squares model analysis

Assume we have calculated values for b_0 and b_1 from the data.
Now:

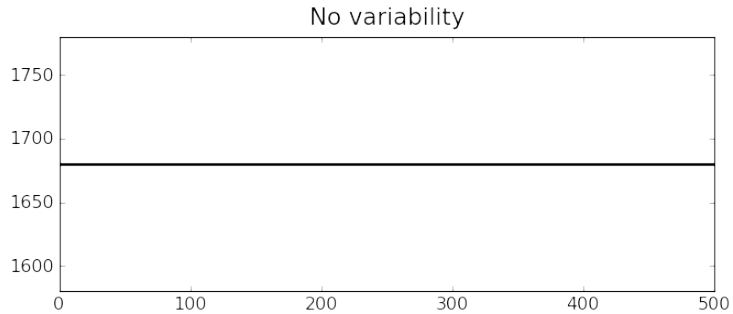
1. How well does the model perform?
2. What part of the data is error
   - noise
3. What part of the data is systematic
   - signal
4. Confidence interval for the model coefficients: b_0 and b_1

   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

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



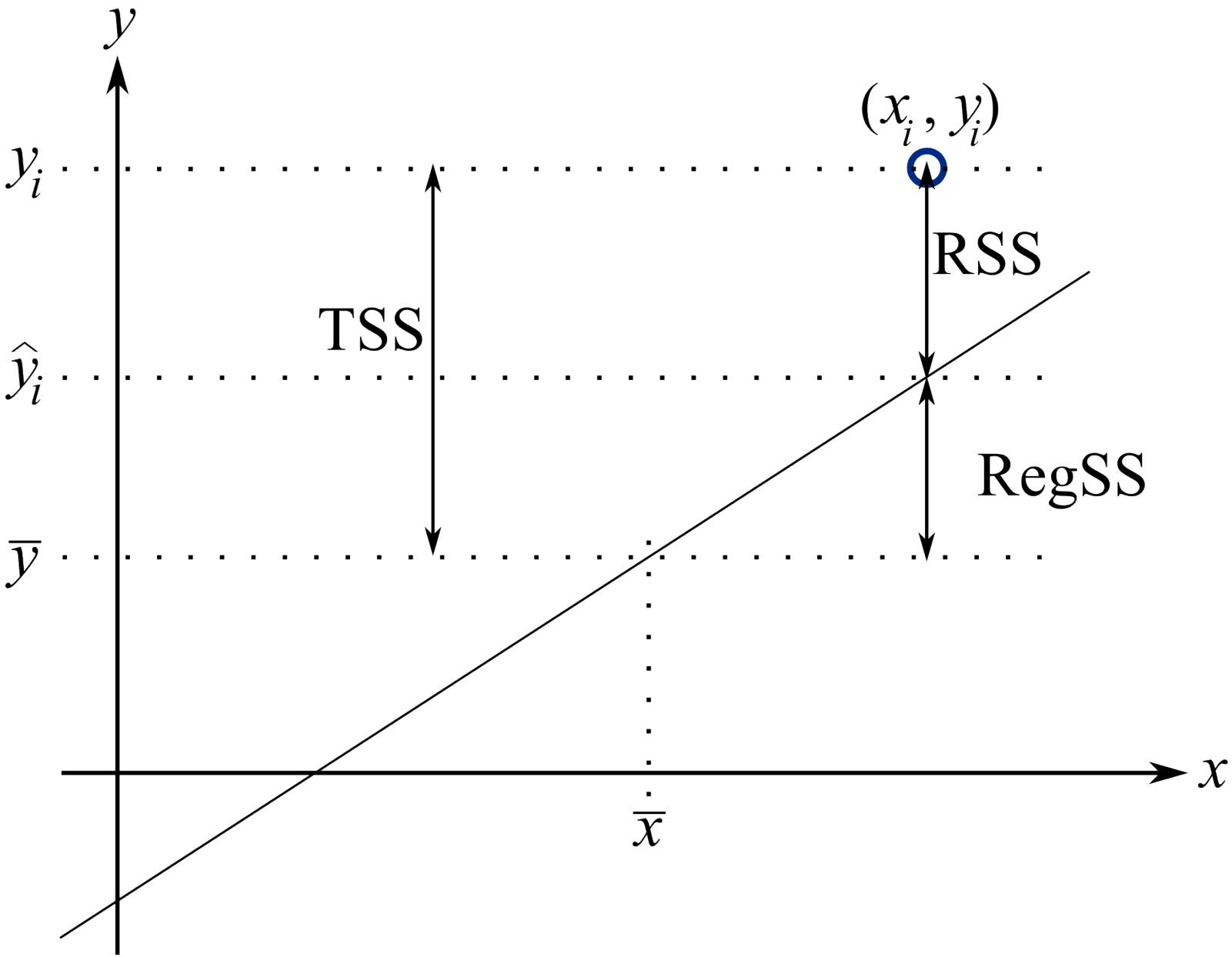
Analysis of variance (ANOVA) - just a tool to show the breakdown of variability in the y vector: $\mathrm{y} = b_0 + b_1\mathrm{x} + \mathrm{e}$

1. doing nothing, no model: implies $\hat{y} = b_0 = \overline{y}$ (intercept only)
2. the model: $\hat{y}_i = b_0 + b_1 x_i$ (intercept *plus* slope)
3. then, how much variance is left over in the errors e_i?

All 3 add up to the total variance. Note that:

- variance is quantified as deviation from the mean
- Variance for first portion is 0, by definition
- Total variance of y = model's variance + error variance
- This is why it was important that $\sum_i (\hat{y}_i e_i) = 0$ (so we can sum them independently)

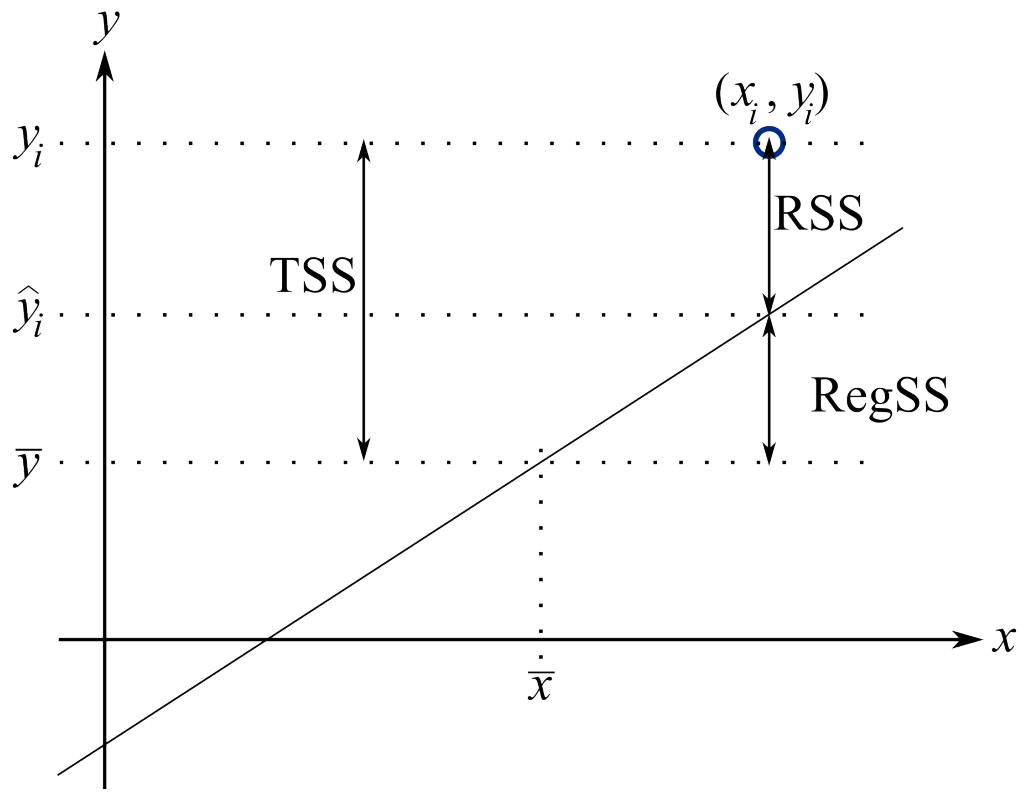
The variance breakdown: geometrically



The variance breakdown: geometrically

For a fixed value of x_i, and for any y_i (above or below):

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



The variance breakdown: table form

Variance of	**Distance**	**DOF**	**SSQ**	**Mean square***
Regression	$\hat{y}_i - \overline{y}$	k (k=2)	RegSS	RegSS/k
Residuals	$y_i - \hat{y}_i$	$n - k$	RSS	RSS/$(n - k)$
Total	$y_i - \overline{y}$	n	TSS	TSS/n

* mean square = "variance"

- Regression SS (RegSS) = $\sum (\hat{y}_i - \overline{y})^2$
  - from the mean to the prediction
- Residual SS (RSS) = $\sum (y_i - \hat{y}_i)^2$
  - from the prediction to the data point
- Total SS (TSS) = $\sum (y_i - \overline{y})^2$
  - from the data point to the mean

Exercise

For each of these cases:

1. $y_i = b_0 + e_i$, i.e. where $b_0 = \overline{y}$ and $b_1 = 0$
2. $y_i = b_0 + b_1 x_i + e_i$, for any values of b_0 and b_1, and the models fits the data without error

Do the following:

- draw a generic plot
- create an ANOVA table with fake values
- write down the value of the ratio $\dfrac{\text{RegSS}}{\text{TSS}}$

Exercise: what did we learn

- No linear model (just noise around a baseline):

$$y_i = \overline{y} + e_i = b_0 + e_i$$

- ratio $= \dfrac{\text{RegSS}}{\text{TSS}} = 0$
- Perfect predictions: $\dfrac{\text{RegSS}}{\text{TSS}} = 1$
- $\dfrac{\text{RegSS}}{\text{TSS}} = R^2$

Derivation of R^2

- $R^2 = \frac{\text{RegSS}}{\text{TSS}} = \frac{\sum_i (\hat{y}_i - \bar{\text{y}})^2}{\sum_i (y_i - \bar{\text{y}})^2}$
- RegSS = variance we can explain with the model
- TSS = total variance we started off with
- TSS = RegSS + RSS, so $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$ also

If

- $R^2 = 0$: requires $\hat{y}_i = \bar{\text{y}}$: predicting a flat line $= \text{y}$ data
- $R^2 = 1$: requires $\hat{y}_i = y_i$: no error in predictions

Meaning of R^2

R^2 name: square of the correlation between x and y.

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

- $-1 \leq r(x,y) \leq +1$
- $R^2 = [r(x,y)]^2$
- $0 \leq R^2 \leq +1$

Just tells us how far we are between predicting from **no variation** to **predictions without error**

Meaning of R^2

- R^2 is computed from the data used to build the model
- Note: R^2 can be calculated without even fitting the model (i.e. without finding b_0 and b_1)
- **No guarantee at all** as to future performance
- It is widely **abused** as a way to measure "*how good is my model*"
  1. "the R^2 value is really high, 90%, this is a good model"
  2. "Wow, that's a really low R^2, this model can't be right - it's no good"

Meaning of R^2

How **good** a model is *for a particular purpose* is almost never related to the R^2 value.

Goodness of a model is better judged by:

- your engineering judgment: does the interpretation of model parameters make sense?
- use the standard error, S_E (next section)
- use testing data to verify the model's predictive performance,
- using cross-validation/randomization tools (more later)

R^2 can be arbitrarily inflated by adding terms to the model. Use adjusted R^2 to account for this:

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}$$

where $k = 2$ for the case of estimating a model $y_i = b_0 + b_1 x_i$

Judging the standard error, S_E

A better way to judge model's predictive performance:

- mean square error $= \text{RSS}/(n-k)$
- standard error $= S_E = \sqrt{\text{mean square error}} = \sqrt{\text{RSS}/(n-k)}$
- The standard deviation of the errors, accounting for degrees of freedom
- $S_E^2 = \dfrac{\text{RSS}}{n-k} = \dfrac{e^T e}{n-k}$
- This is related to the model's objective function: minimizing RSS

Assume our model predicts

- $\hat{y}_i =$ batch yield [kg]
- $x_i =$ raw material purity
- What does a standard error of 3.4 kg mean?
  - If errors are normally distributed:
    - About 2/3 of predictions lie within ± 3.4 kg
    - About 95% of predictions lie within $\pm 2 \times 3.4$ kg

Confidence intervals for b_0 and b_1

- We have made no assumptions about the data
- We can calculate b_0 and b_1
- We can calculate model predictions

Now we want to know more. Let's learn from the data.

Assumptions required

- Model structure: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$

Assumption 1

Model is linear

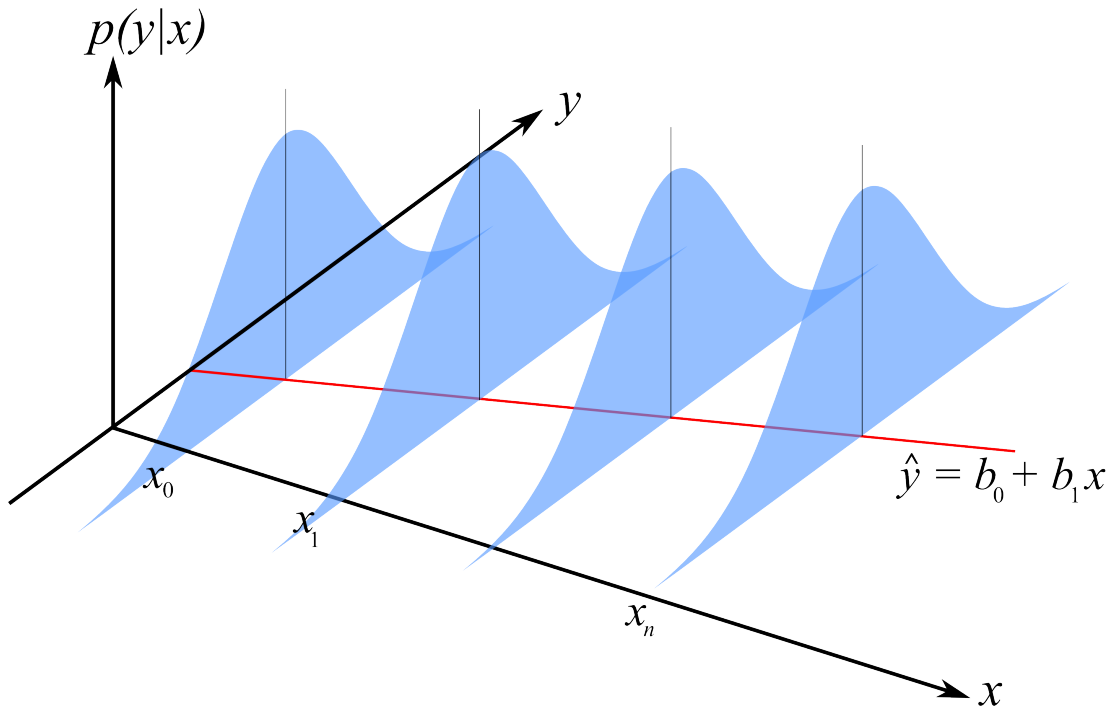
- Implies ε is the error of y ($\beta_0 + \beta_1 \mathrm{x}$ terms are fixed)
- Means your x variable has much less uncertainty than the y variable

Assumptions required

Assumption 2

Variance of y is constant at all values of x

- Constant error variance assumption
- Practice: variability of y often non-constant
  - measurement accuracy deteriorates at extreme conditions of x and y



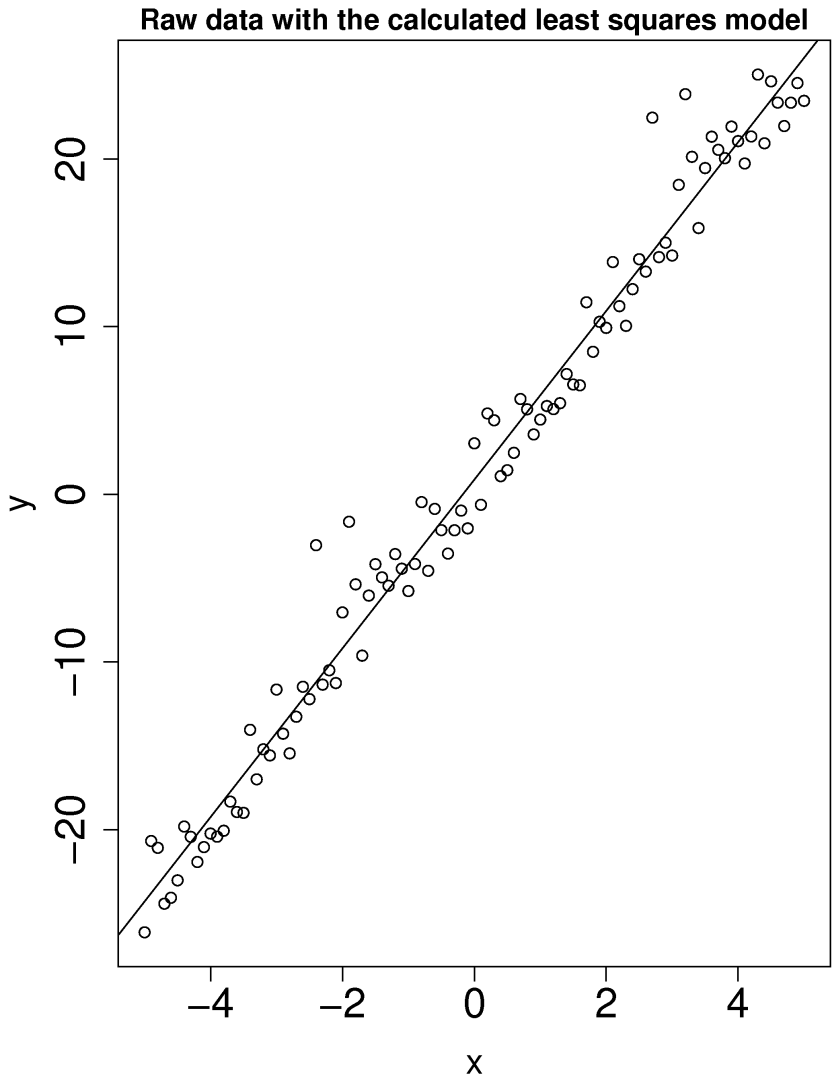
Assumptions required

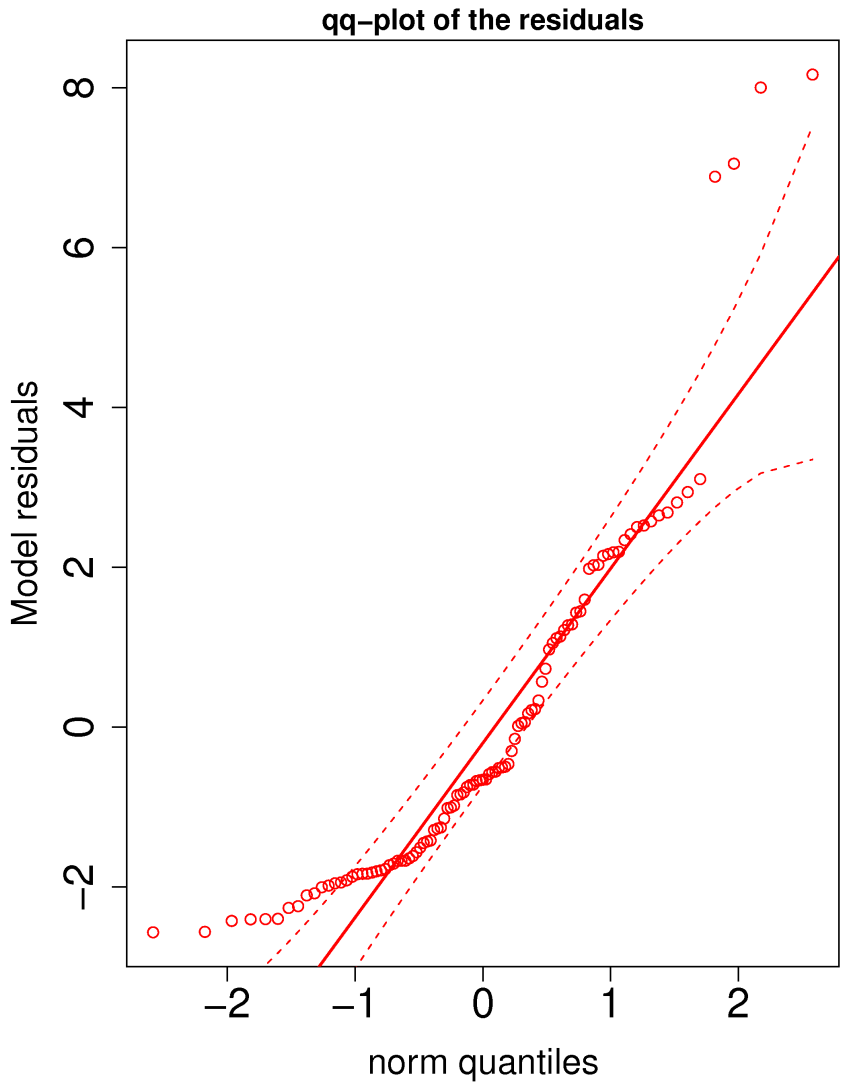
Assumption 3

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

- Implies that $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma_\varepsilon^2)$ (from first linearity assumption)
- Practice: not always certain
- But at least it is easy to test with a q-q plot

Assumptions required





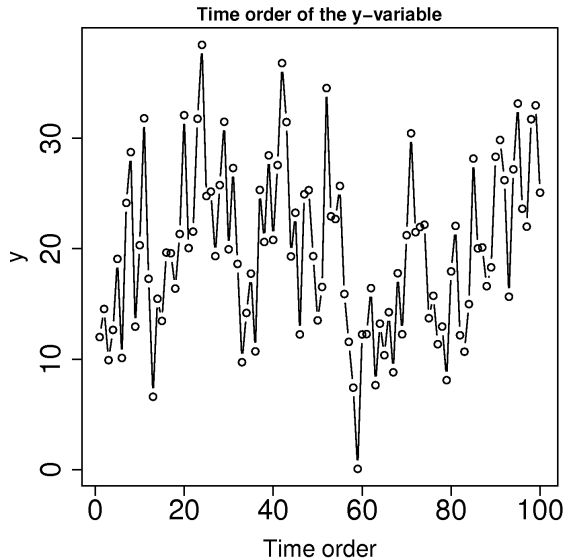
Assumptions required

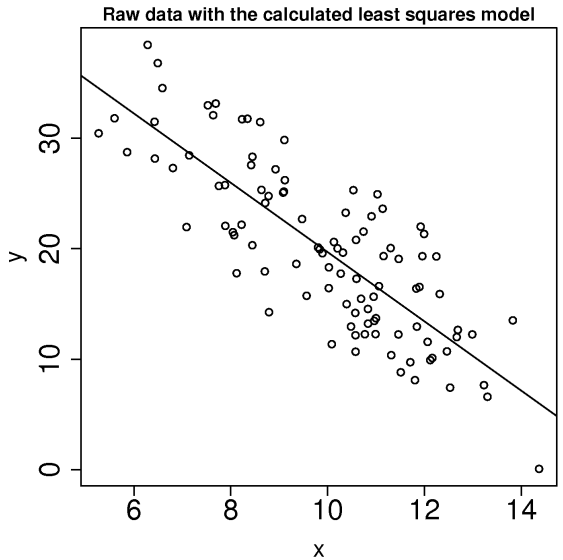
Assumption 4

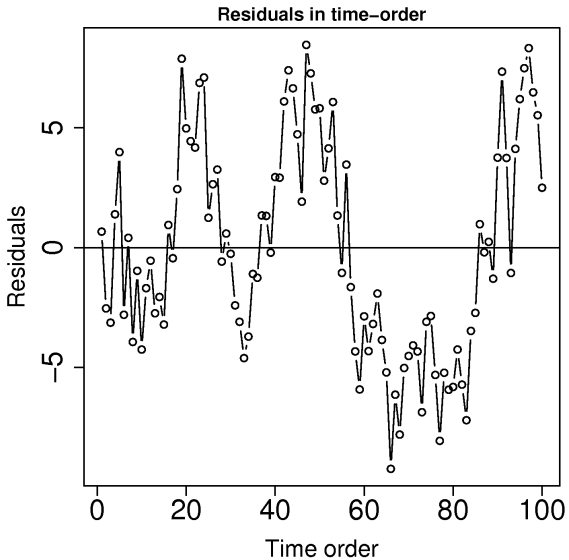
Each error is independent of the other

- Often violated in practice
- Observations taken 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 (autocorrelation)

Assumptions required







Assumptions required

Assumption 5

Assume x are fixed and independent of the error

- Closely tied to first assumption
- Most engineering cases we sample the x and y
- This assumption requires that our error variance does not change as we sample x

Assumptions required

Assumption 6

All y_i values are independent of each other.

- Closely tied to 4th assumption
- Is the same as assumption 4 if the x are constant
- Violated in cases where data are collected in time order and y_i are autocorrelated

Confidence intervals for b_0 and b_1

To construct a confidence interval, we must know the distribution

$$b_0 \sim \mathcal{N}(\beta_0, \mathcal{V}\{\beta_0\}) \qquad \text{and} \quad b_1 \sim \mathcal{N}(\beta_1, \mathcal{V}\{\beta_1\})$$

1. Create z-value for b_0 and b_1
2. Calculate the confidence interval for β_0 and β_1

Objective: Calculate a suitable $\mathcal{V}\{\beta_0\}$ and $\mathcal{V}\{\beta_1\}$

Variance of b_1

Details are in your notes. Strongly encouraged to go through the derivation.

- Recall the standard error:

$$S_E^2 = \mathcal{V}\{e_i\} = \mathcal{V}\{y_i\} = \frac{\sum e_i^2}{n-k} \qquad \text{or} \qquad S_E = \sqrt{\frac{\sum e_i^2}{n-k}}$$

- The variance of $b_1 = \mathcal{V}\{b_1\}$:

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

- Apart from the numerator, how could you decrease the variance your model's b_1 coefficient?
  - Use samples that are far from the mean of the x-data
  - Use more samples
- The numerator term: this is the standard error S_E^2 term we saw earlier.

Variance of b_0

Variance of b_0:

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

Summary of important equations:

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

- S_E standard deviation of the error (residuals)
- $S_E(b_0)$ standard deviation of b_0
- $S_E(b_1)$ standard deviation of b_1

Confidence intervals for b_0 and b_1

Finally! Construct confidence intervals for the least squares model parameters:

$$\begin{array}{rcccl} -c_t & \leq & \dfrac{b_0 - \beta_0}{S_E(b_0)} & \leq & +c_t \\ b_0 - c_t S_E(b_0) & \leq & \beta_0 & \leq & b_0 + c_t S_E(b_0) \end{array}$$

$$\begin{array}{rcccl} -c_t & \leq & \dfrac{b_1 - \beta_1}{S_E(b_1)} & \leq & +c_t \\ b_1 - c_t S_E(b_1) & \leq & \beta_1 & \leq & b_1 + c_t S_E(b_1) \end{array}$$

- Use t-distribution: we used estimates of the variance

Example

Confidence interval for β_0 and β_1 from previous example. Already calculated:

- $b_0 = 3.0$
- $b_1 = 0.5$
- $S_E = 1.237$
  - `predictions = b0 + x*b1`
  - `8.001 7.000 9.501 7.501 8.501 10.001 6.00 5.000 9.001 6.500 5.501`
  - `error = y - predictions`
  - `SE = sqrt(sum(error)**2 / (N-2))`
- $c_t = 2.26$ at 95% confidence
  - `qt(0.975, df=(11-2))`

Example

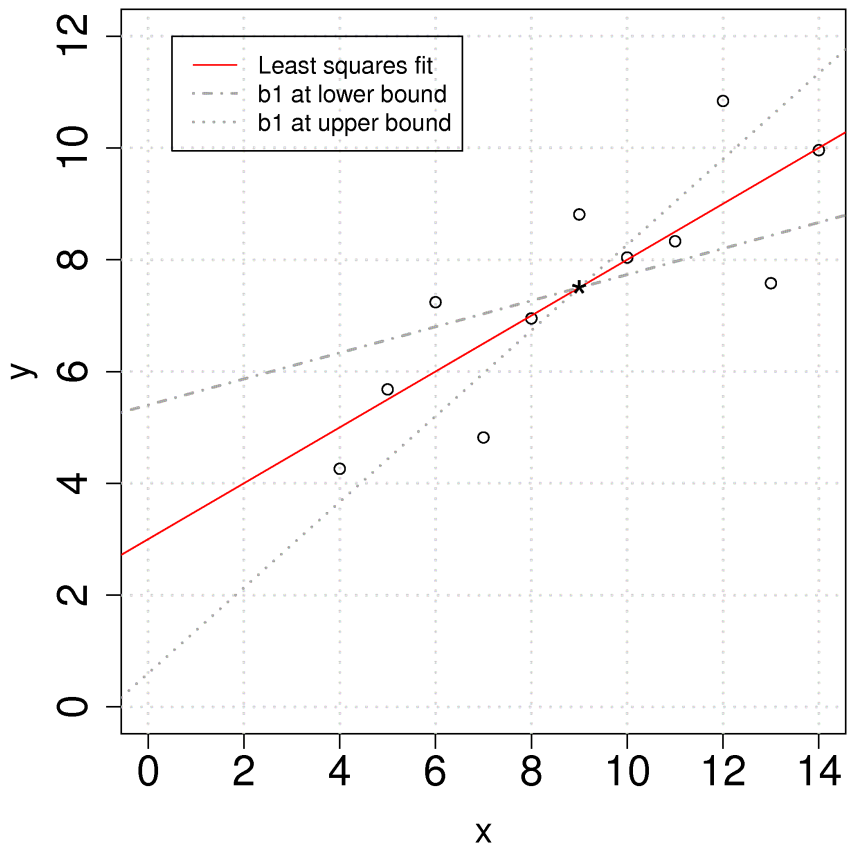
- $S_E^2(b_1) = \dfrac{S_E^2}{\sum_j (x_j - \bar{x})^2} = \dfrac{1.237^2}{110} = 0.0139$; $S_E(b_1) = 0.1179$
- $S_E^2(b_0) = S_E^2 \left(\dfrac{1}{N} + \dfrac{\bar{x}^2}{\sum_j (x_j - \bar{x})^2} \right) = \left(\dfrac{1}{11} + \dfrac{9^2}{110} \right) 1.237^2 = 1.266$; $S_E(b_0) = 1.125$
- For β_0:

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

Example

- For β_1:

$$\begin{array}{rcccl} -c_t & \leq & \dfrac{b_1 - \beta_1}{S_E(b_1)} & \leq & +c_t \\ 0.5 - 2.26 \times \sqrt{0.0139} & \leq & \beta_1 & \leq & 0.5 + 2.26 \times \sqrt{0.0139} \\ 0.233 & \leq & \beta_1 & \leq & 0.767 \end{array}$$



Software output

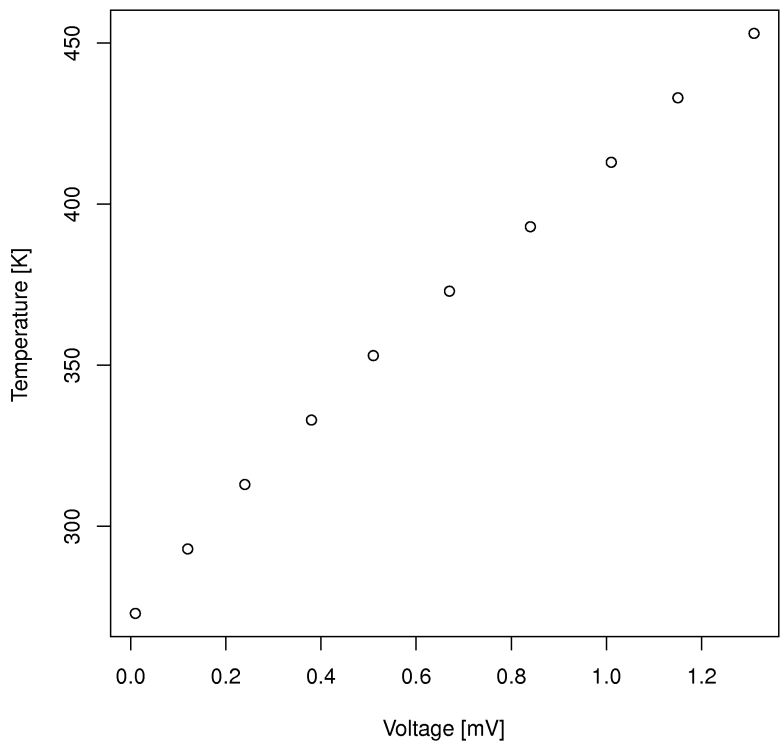
```
> x <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
> y <- c(8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68)
> model <- lm(y ~ x)    # "The linear model, where y is described by x"
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x             0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665,     Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.002170
```

Example

- What is the causal direction (line of reasoning) that the plot's author is wanting you to follow?
- Interpret the software output: $y = 15.7 - (0.70)\text{cameras}$
  - $9.61 \leq \beta_0 \leq 21.8$
  - $-1.8 \leq \beta_1 \leq 0.36$
  - $S_E = 10.9$
  - $R^2 = 0.075$

Example

Thermocouples produce an almost linear voltage (millivolt) response at different temperatures. (Data are on the course website). Temperatures usually recorded to an accuracy of ± 0.5K with inexpensive thermocouples.



1. Calculate and interpret model parameter estimates, $y = b_0 + b_1 x$, from the given data
   - Model calculated is $\hat{T} = 278.6 + 135.3V$. Interpretation?
2. Temperature prediction at 1.00 mV?
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?

Review so far

$$\begin{aligned} \text{Population model} : \mathrm{y} &= \beta_0 + \beta_1 \mathrm{x} + \varepsilon \\ \text{Estimated model} : y_i &= b_0 + b_1 x_i + e_i \end{aligned}$$

Solve an optimization problem:

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

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

$$S_E = \sqrt{\frac{\mathrm{RSS}}{n-k}} = \sqrt{\frac{e^T e}{n-k}} \qquad \sum e_i = 0$$

Review of assumptions

1. Model is linear
2. Variance of y is constant at all values of x
3. Errors are normally distributed
4. Each error is independent of the other
5. Assume x 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.

Analysis

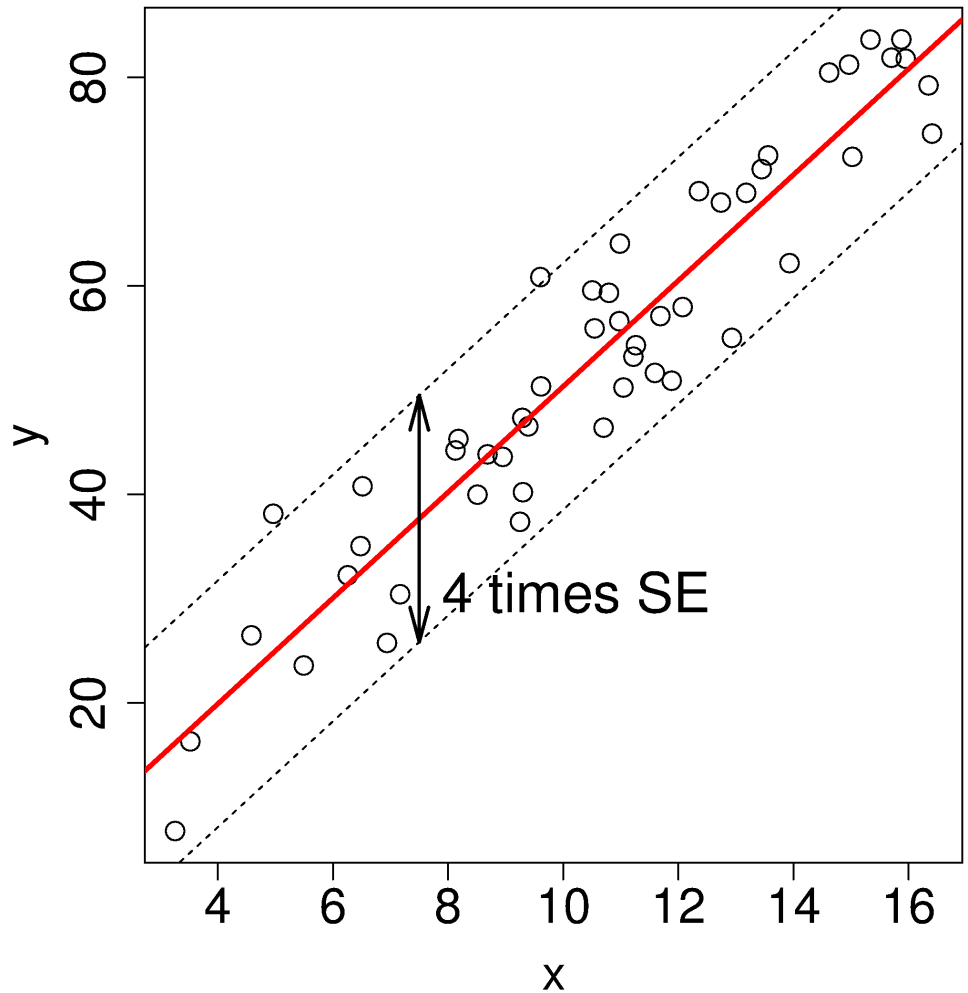
If those above assumptions hold then:

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

and then these are accurate estimates of the variance.
We then go on to use those variances to calculate confidence intervals.

An approximate prediction interval for y

- True model: $y = \beta_0 + \beta_1 x + \varepsilon \quad \Longrightarrow \quad \hat{y} = \beta_0 + \beta_1 x$
- Predict a new y-value: $\hat{y}_i = b_0 + b_1 x_i$
- We could assume $\hat{y}_i \pm 2S_E$



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

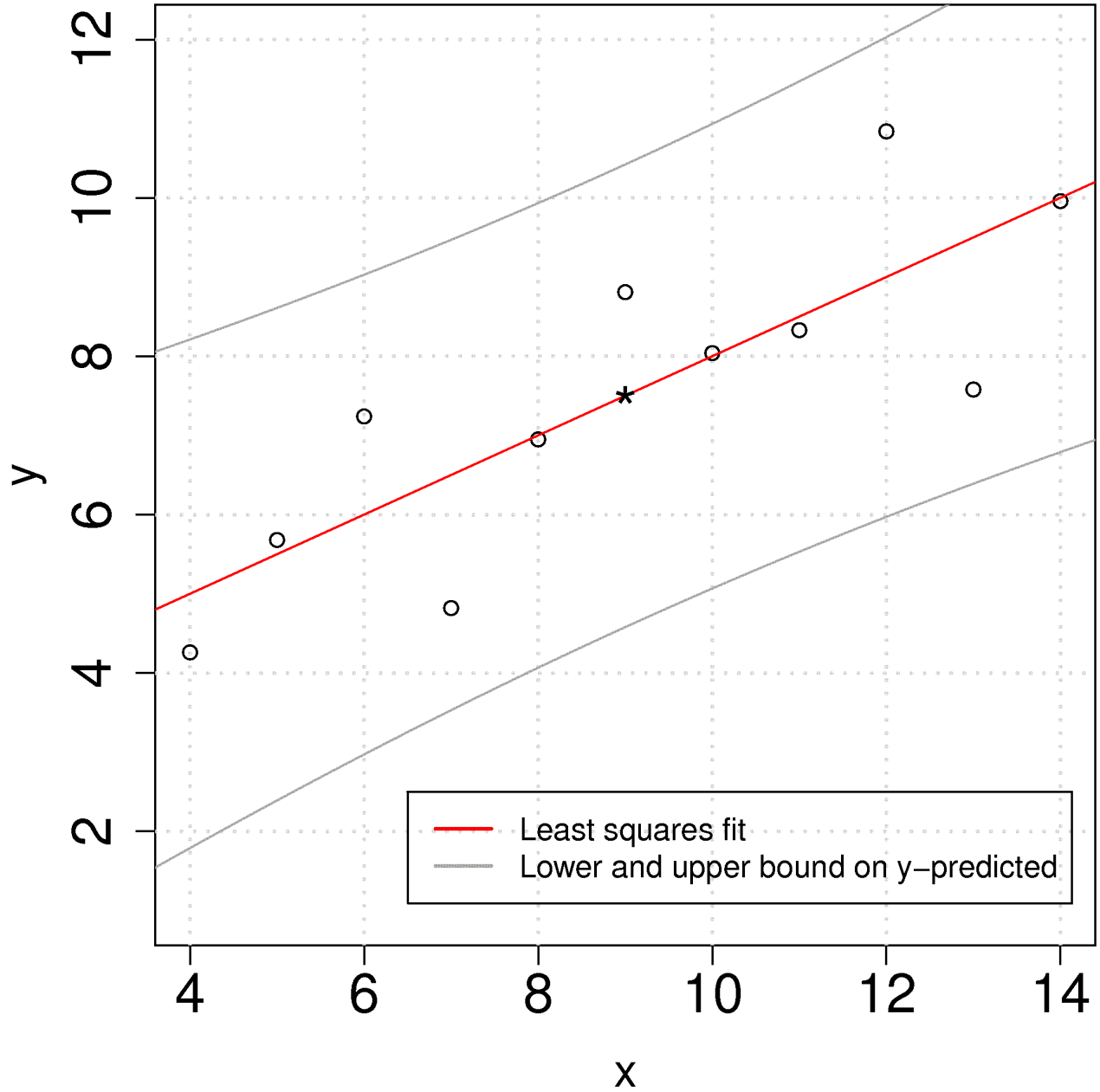
More accurate prediction interval for y

But, the previous slide ignores that b_0 and b_1 have error: that gets propagated into $\hat{y}_i$. True prediction error:

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

- Example in the notes
- General shape: quadratic shape
- Smallest prediction error at the model center
- Prediction error expands progressively wider as one moves away from the model center
- Makes intuitive sense

Prediction interval for y: *it has curvature*



Investigating a linear model

- Transformations of data
- Leverage, outliers, influence and discrepancy
- Multiple linear regression, MLR
- Training and testing data

The residuals contain the clues to problems with the model.

Normally distributed errors

Aim

Ensure the residuals are normally distributed

If non-normal: S_E and other variances such as $\mathcal{V}(b_1)$ are inflated.

Detecting it:

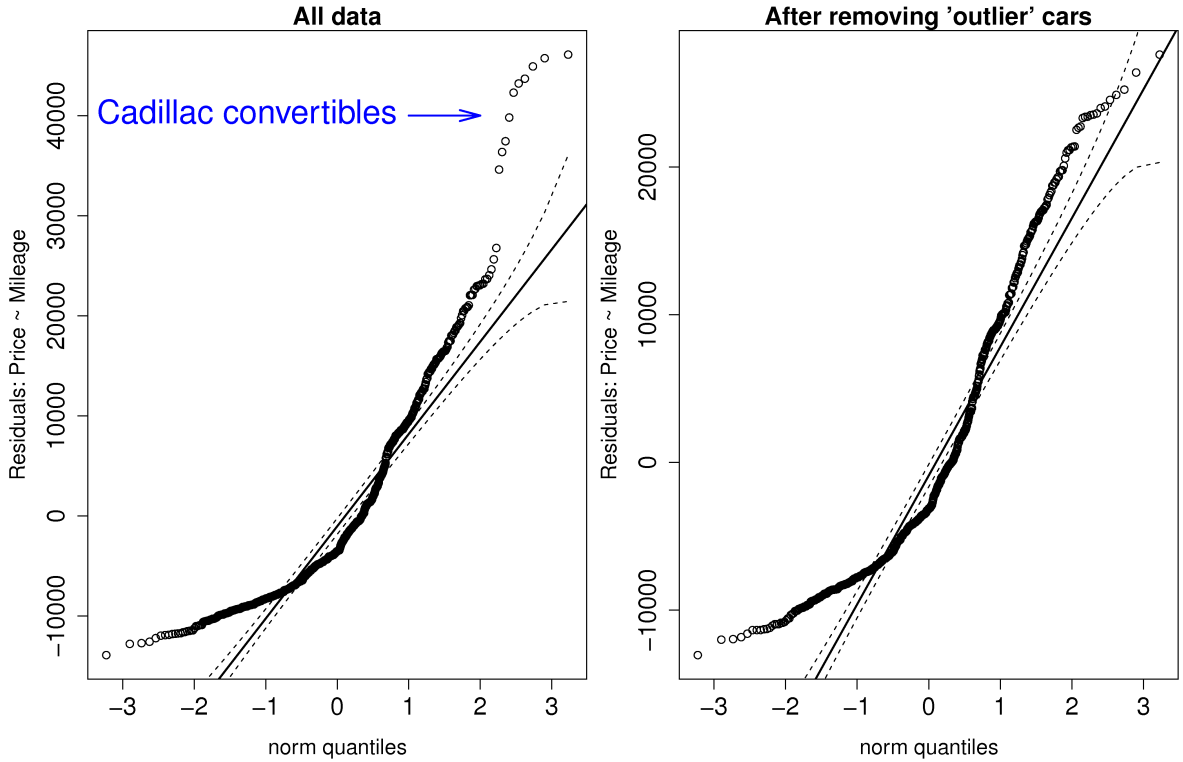
- Use a q-q plot: should match 45 degree line
  - Most accurate way
- A histogram is somewhat helpful
  - Human eye not good at picking up heavy tails
- Do *not* plot the residuals in time-order to try detect if normally distributed
  - Human eye not good at picking up heavy tails

Normally distributed errors

Dealing with it:

- Remove any outlying observation(s) (after investigation)
- Transform the y-variable
- Add additional explanatory terms to the model (see this in MLR later)

Normally distributed errors



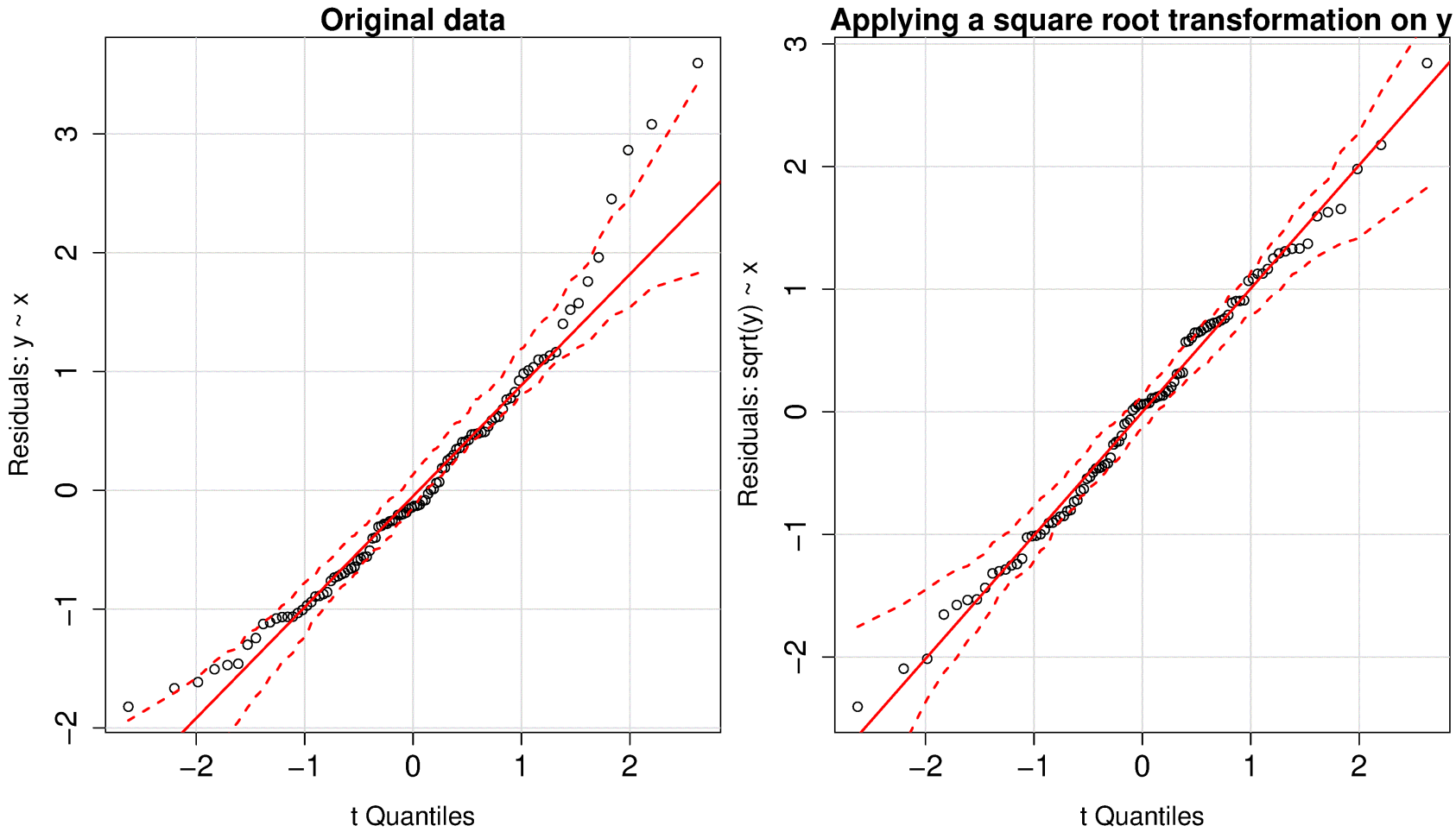
Before : $b_1 = -0.173$ $\quad -0.255 \leq \beta_1 \leq -0.0898$

$S_E = 9789$

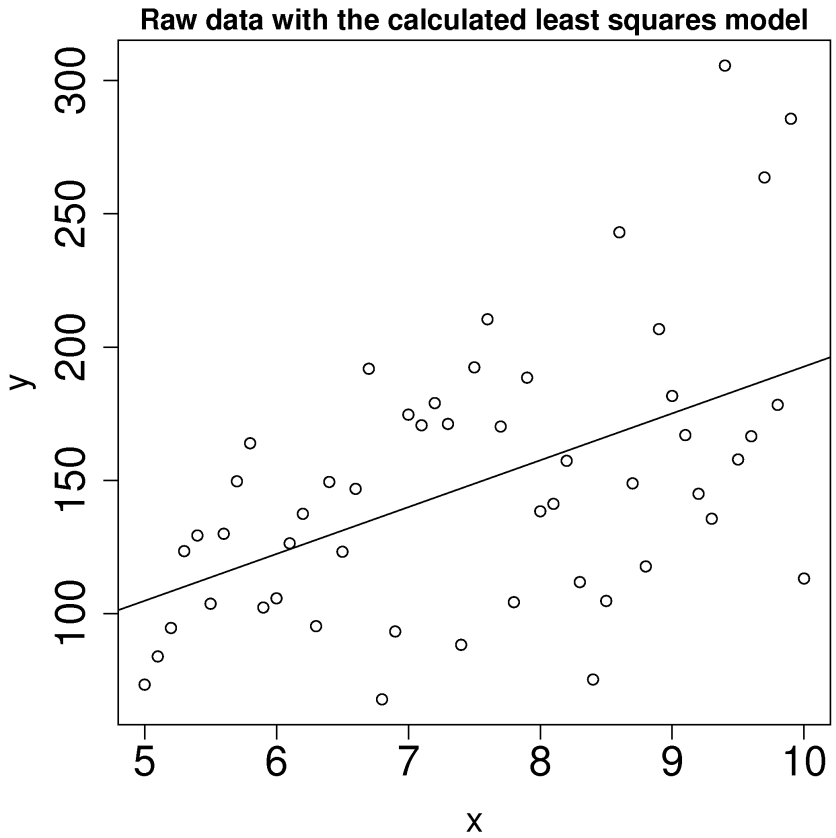
After : $b_1 = -0.155$ $\quad -0.230 \leq \beta_1 \leq -0.0807$

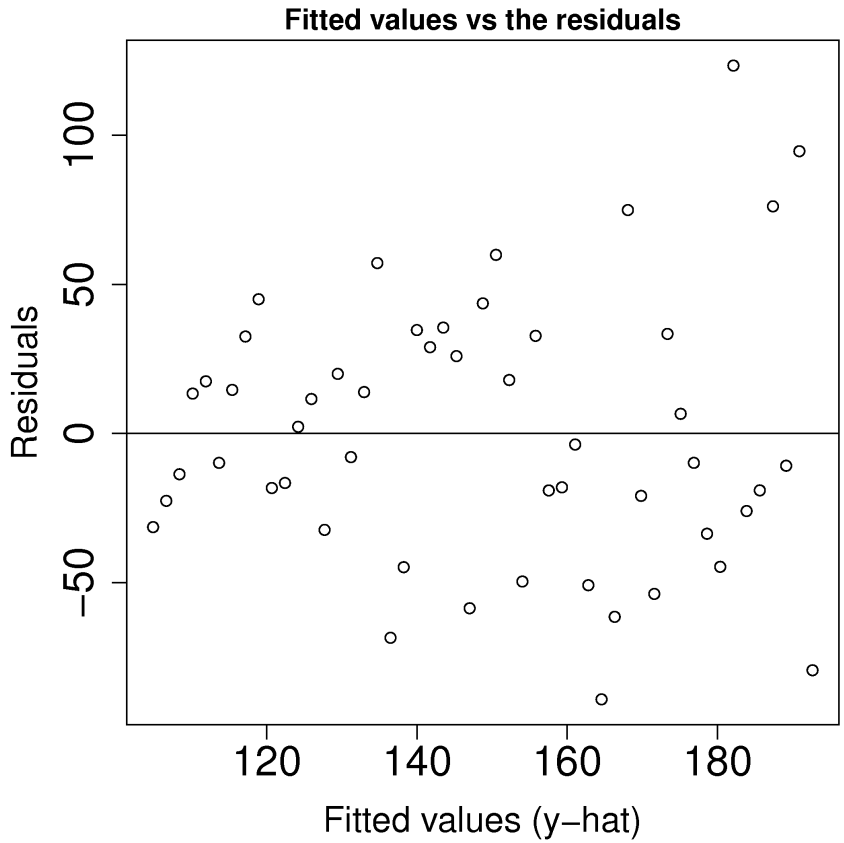
$S_E = 8655$ $\quad \longleftarrow$ it has been reduced

Normally distributed errors



Non-constant error variance





- Variability in y is non-constant as x or y change.
- Increases S_E, undermining confidence intervals
- Not usually too problematic (LS is pretty robust)

Non-constant error variance

Detecting it:

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

Everything OK when:

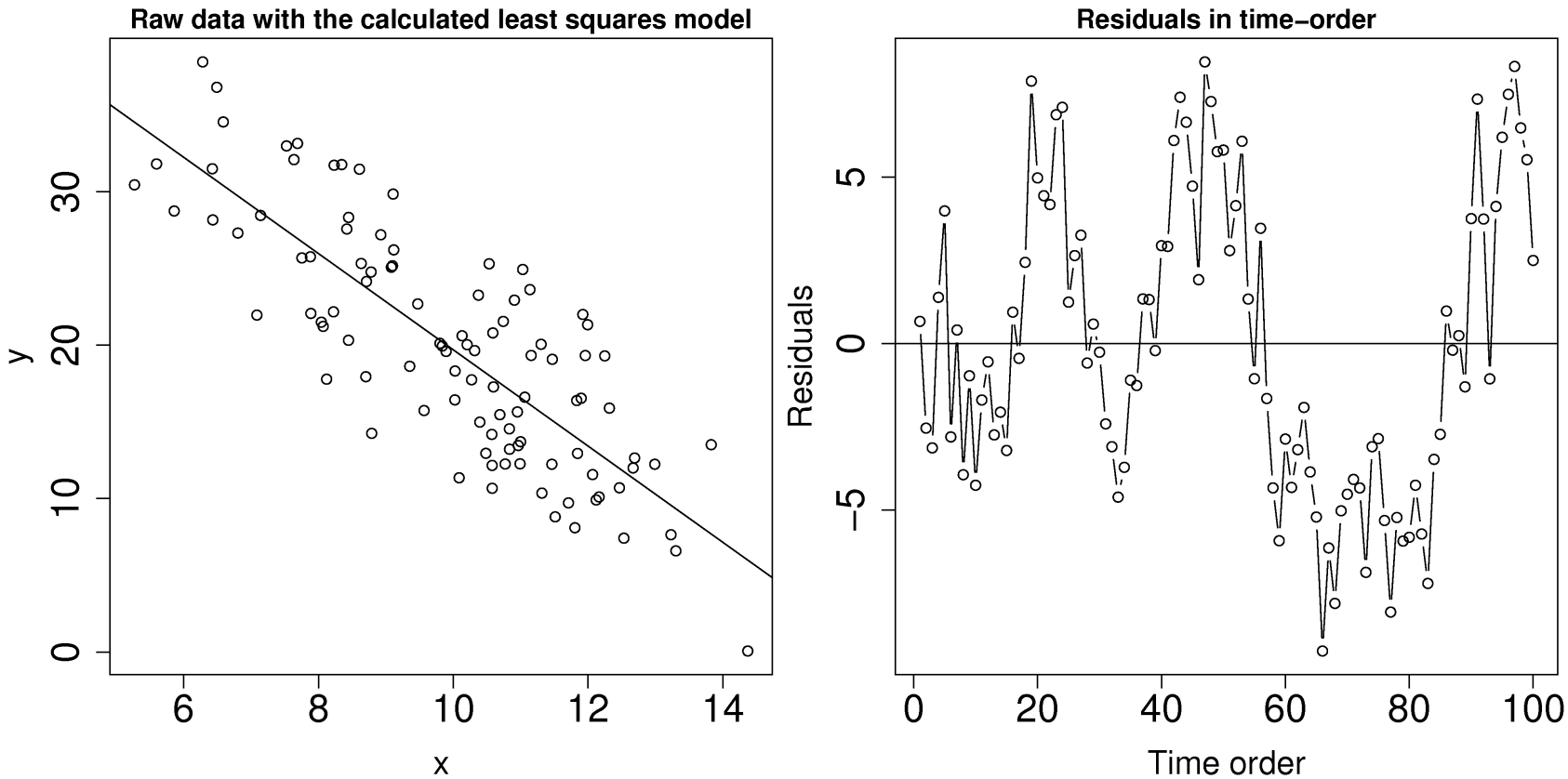
- no visible structure in the above plots

Lack of independence in the data

- **Serious issue**:
  - independence required to use CLT
  - CLT is used for confidence intervals of β_0 and β_1
- Not-independent: *autocorrelated*, e.g. a slow-moving process.
- Dealing with it: topic is called time-series analysis; many good textbooks, e.g.
  - Chatfield: The Analysis of Time Series ⟵highly recommended
  - Box and Jenkins: Time Series Analysis ⟵comprehensive
- We will use a very simple method here

Lack of independence in the data

Detecting it: time-ordered plot of residuals

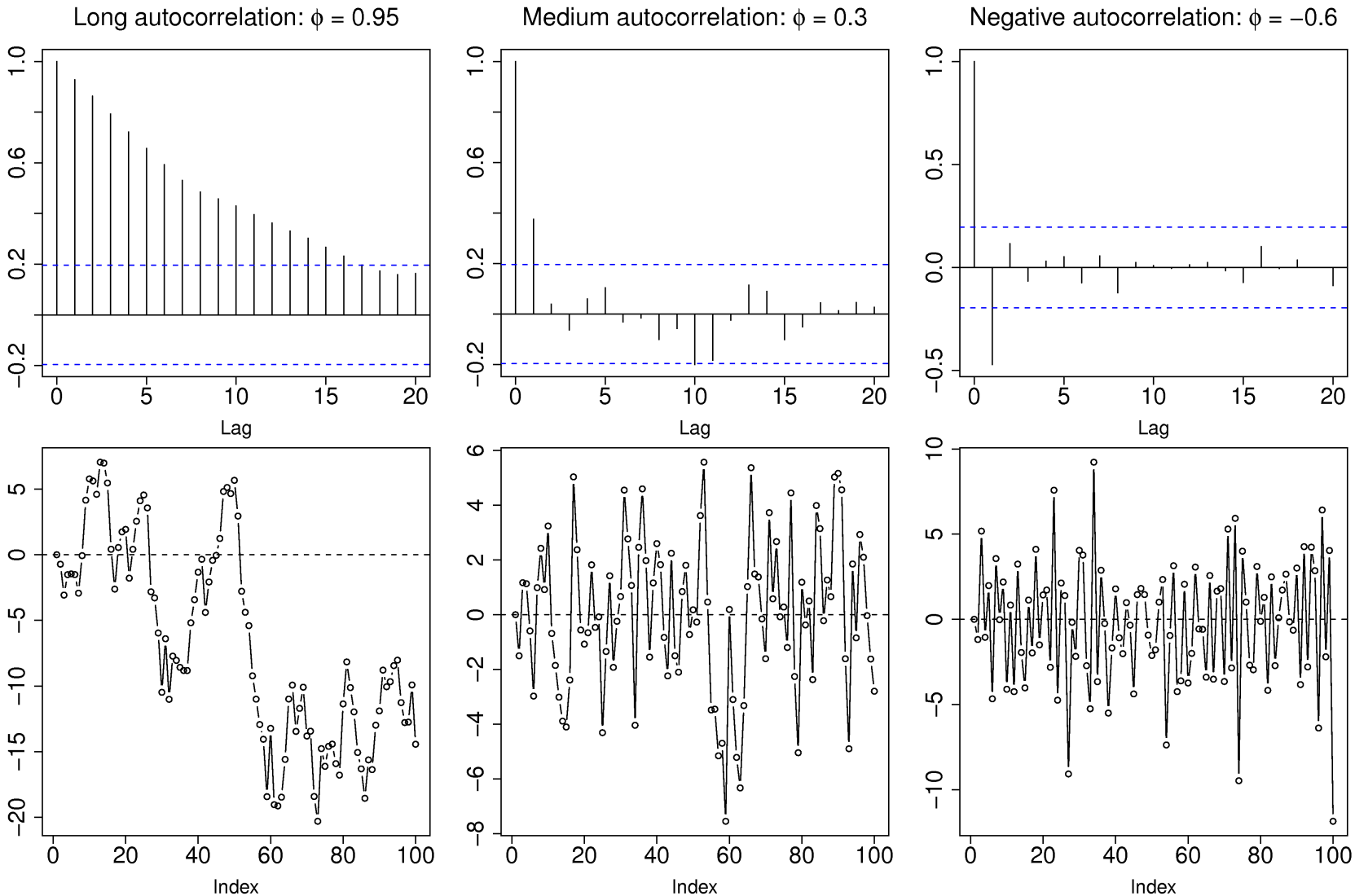


Look for patterns:

- slow drifts
- rapid criss-crossing of the zero axis
- cycles (e.g. daily cycle)

Lack of independence in the data

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



Lack of independence in the data

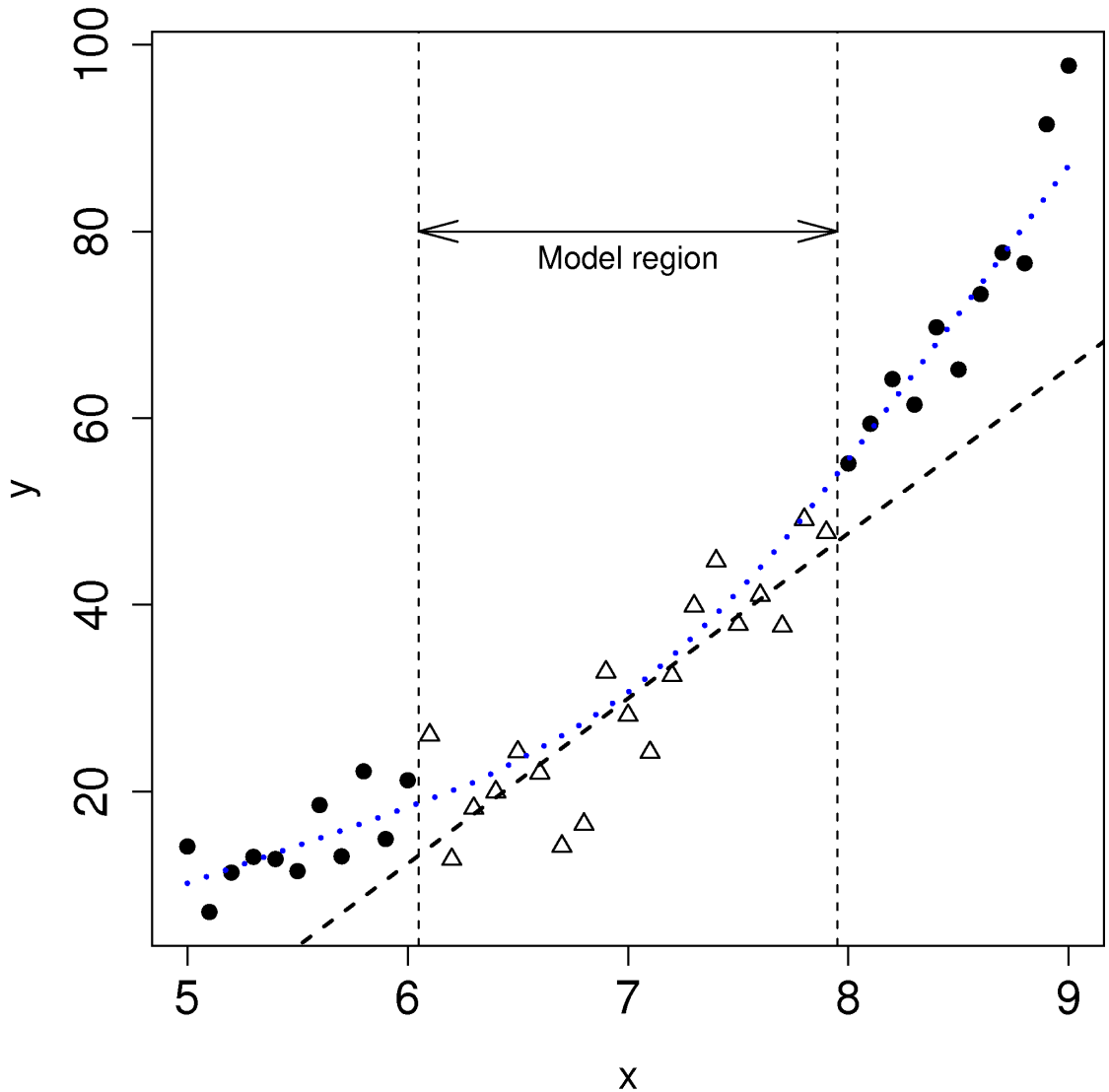
- Another test: Durbin-Watson test: *Draper and Smith* (chapter 7)

Everything OK when:

- `acf(y)` and `acf(e)` show no lags beyond lag 0

Linear model (incorrect model specification)

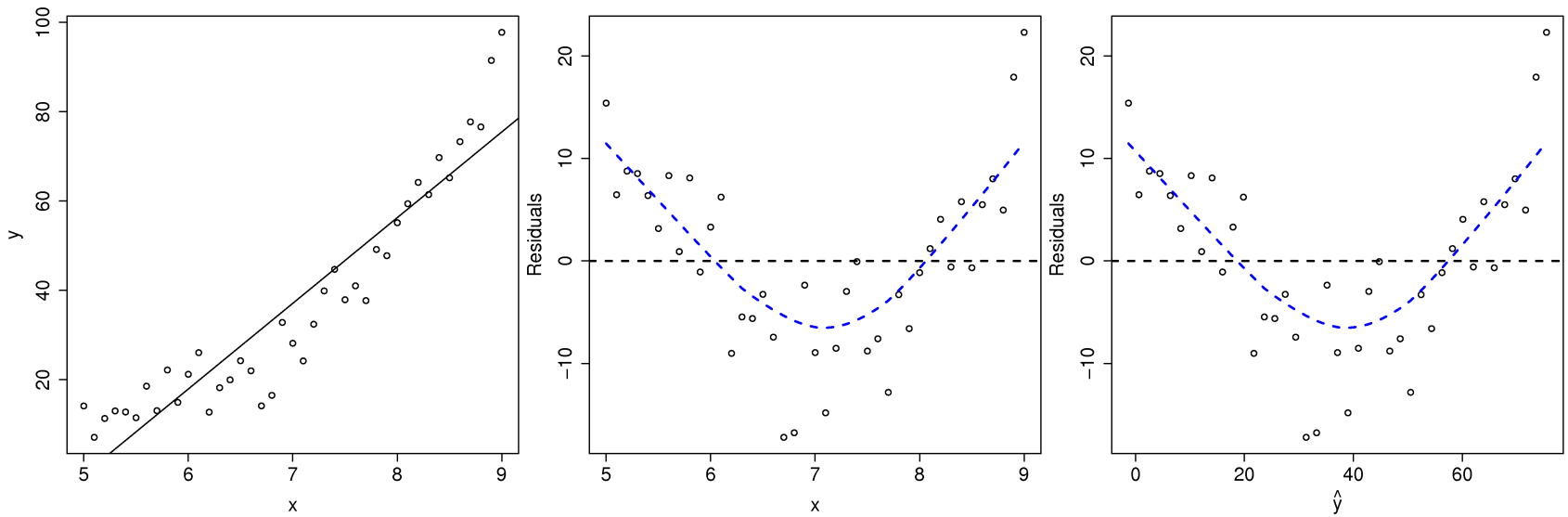
- Systems are known to be non-linear; linear might be good enough (depends on model's purpose)



Linear model (incorrect model specification)

Detecting it:

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



Linear model (incorrect model specification)

Dealing with it:

1. Non-linear least squares
   - Too many details to go into for this course
   - *Example*: $y = \beta_1 \left(1 - e^{-\beta_2 x}\right)$
     - $\beta_1 = [10, 20, 30, 40, 50]$ (e.g. picked 5 values that span β_1 range)
     - $\beta_2 = [0.0, 0.2, 0.4, 0.8]$ (e.g. and here I picked 4 values for β_2)
   - Evaluate sum of squares of errors (SSE $= \sum e_i^2$) objective function at all $4 \times 5 = 20$ combinations of β_1 and β_2

Linear model (incorrect model specification)

Dealing with it:

2. Transform the x or y variable; then use linear model
   - $x_{\text{transformed}} \leftarrow x^{p}_{\text{original}}$
   - p: 1 up to 1.5, 1.75, 2.0, *etc*
   - p: 1 down to 0.5 ($\sqrt{x}$), 0.25, -0.5, -1.0 ($1/x$), -1.5, -2.0, *etc*
   - $\log(x)$: approximates $p = 0$ ito severity

Linear model (incorrect model specification)

3. Rearrange first-principles equations
   - Distillation: T inversely proportional to log(VP). $y = b_0 + b_1 x$:
     - $x \leftarrow 1/T$
     - $y \leftarrow P$.
     - 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: $y = b_0 + b_1 x$:
     - $b_0 \leftarrow p$
     - $b_1 \leftarrow q$
     - $y \leftarrow 1/y$

Everything OK when:

- No more structure in the detection plots

Summary of steps to build and investigate a linear model

1. Plot the data
2. Fit model and examine text:

```
model <- lm(y ~ x)      # fit the model: "y as described by variable x"
summary(model)
confint(model)
```

   - Model's standard error, compare it to y
   - Interpret confidence interval

3. Visualize predictions:

```
plot(x, y)
lines(lowess(x,y))          # show the smoother
abline(model, col="red")    # and show the least squares model
```

Summary of steps to build and investigate a linear model

4. q-q plot of the residuals. Are they normally distributed?
5. Residuals against the x-values.
   - Expect to see no particular structure
6. Residuals in time (sequence) order
   - Also try `acf(model$residuals))`
7. Residuals against the fitted-values.
8. Plot the $\hat{y}$ against actual values of y

Multiple linear regression (MLR)

AIM: Include more than one x-variable in the model.

1. introduce some matrix notation
2. redo the optimization problem
3. interpret the model coefficients
4. use integer variables

Multiple linear regression (MLR)

1. Improve our understanding of systems:
   - x_1 = reactant concentration
   - x_2 = temperature
   - y = reaction rate
   - Understand the effect of temperature also:
     $y = b_0 + b_1 x_1 + b_2 x_2$
2. Improve predictions:
   - x_1 = temperature
   - x_2 = feed flowrate
   - y = melt index
   - Better predictions with $y = b_0 + b_1 x_1 + b_2 x_2$ than with $y = b_0 + b_1 x_1$
3. Integer variables at more than 2 levels:
   - mixing tank
   - operator
   - supplier
   - day of week

Notation for MLR

We will first center the data:

$$\begin{aligned} y_i &= b_0 + b_1 x_i \\ \overline{y} &= b_0 + b_1 \overline{x} \\ y_i - \overline{y} &= 0 + b_1 (x_i - \overline{x}) \qquad \text{by subtracting previous lines} \end{aligned}$$

- 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 zero: $b_0 = 0$
- If you work with *uncentered* data, but add a $b_0 = 0$ term to the model, you will get an **identical** MLR model
  - Centered data is more interpretable: $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$
  - Intercept can always be recovered afterwards, if required

Notation for MLR

The general linear model is given by:

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

Notation for MLR

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

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

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

Estimating the model parameters via optimization

Objective function: minimize sum of squares of the errors

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

- Solving $\frac{f(\mathbf{b})}{\partial \mathbf{b}} = 0$ gives us that $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$.

Estimating the model parameters via optimization

1. $\mathcal{E}\{\mathbf{b}\} = \beta$
2. $\mathcal{V}\{\mathbf{b}\} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} S_E^2$
3. Standard error $= \sigma_e \approx S_E = \sqrt{\dfrac{\mathbf{e}^T\mathbf{e}}{n-k}}$,

- k is the number of parameters estimated in the model
- n is the number of observations

Earlier we had: $\mathcal{V}\{b_1\} = \dfrac{S_E^2}{\sum_j (x_j - \overline{\mathrm{x}})^2}$

- Matrix form is identical
- Decrease $\mathcal{V}\{\mathbf{b}\}$ by:
  - taking more samples (increases denominator size)
  - including observations away from the center of the model

Example

- $x_{1,\text{raw}} = [1, 3, 4, 7, 9, 9]$
  - $x_1 = [-4.5, -2.5, -1.5, 1.5, 3.5, 3.5]$,
- $x_{2,\text{raw}} = [9, 9, 6, 3, 1, 2]$
  - $x_2 = [4, 4, 1, -2, -4, -3]$
- $y_{\text{raw}} = [3, 5, 6, 8, 7, 10]$
  - $y = [-3.5, -1.5, -0.5, 1.5, 0.5, 3.5]$

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

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

Example

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

- $\mathbf{X}^T\mathbf{X}$: scaled version of the covariance matrix of $\mathbf{X}$.
- Diagonal entries: variance, always positive
- Off-diagonal entries: symmetrical, strength of relationship between variables
- What does $\mathbf{X}^T\mathbf{X}$ look like for uncorrelated variables?

Let's take a look at building the model in R.

- `model <- lm(y ~ x1 + x2)`
- `summary(model)`

Example: 600-level students self-study

- Standard error $= \mathcal{V}\{\mathbf{b}\} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} S_E^2$

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} 0.323 & 0.297 \\ 0.297 & 0.289 \end{bmatrix}$$

- Variances are not independent of each other! CI are not independent either.
- For 2 variables:

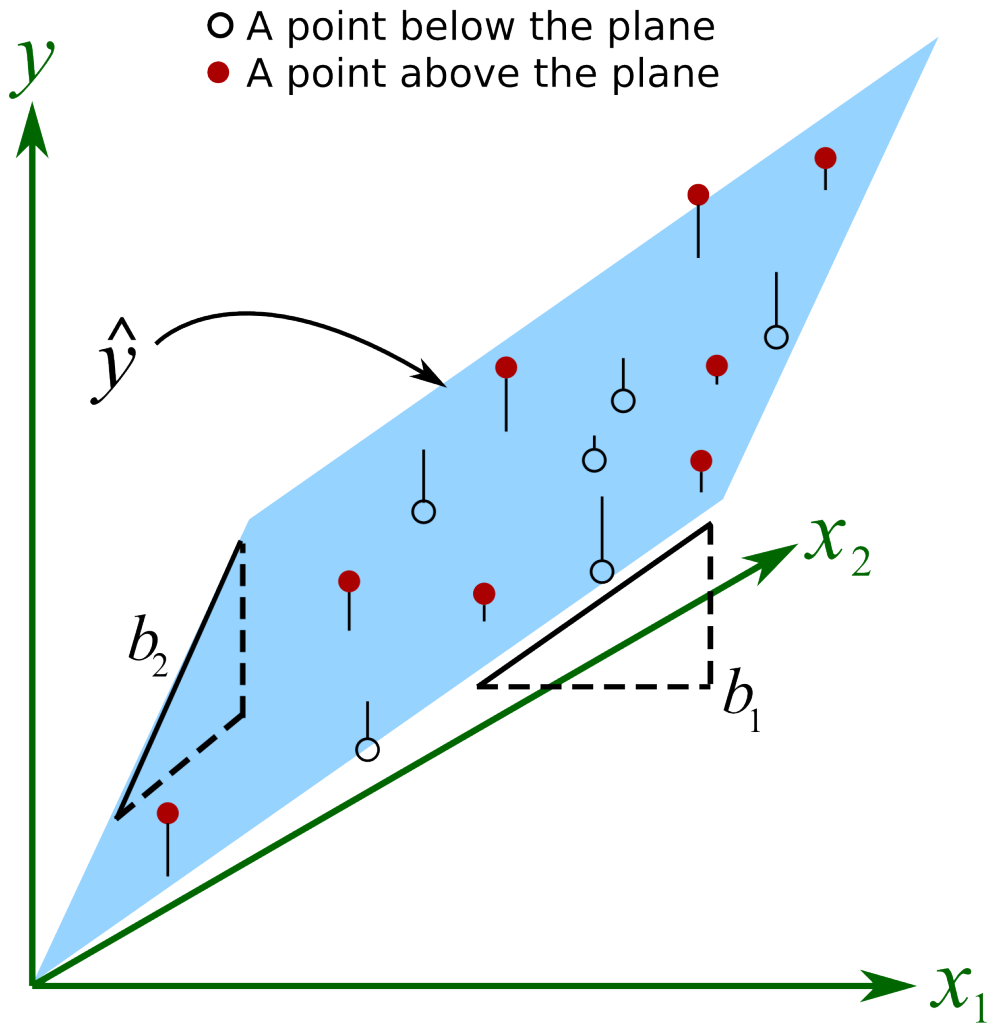
$$\begin{aligned} \mathcal{V}(b_1) &= \frac{1}{1 - r_{12}^2} \times \frac{S_E^2}{\sum x_1^2} \\ \mathcal{V}(b_2) &= \frac{1}{1 - r_{12}^2} \times \frac{S_E^2}{\sum x_2^2} \end{aligned}$$

- $r_{12}^2 =$ correlation between x_1 and x_2
- What happens as the correlation between the two variables increases?
- Each $\mathcal{V}(b_k)$ variance has $n - k - 1$ degrees of freedom

Interpretation of the model coefficients

Model

$$y = b_1 x_1 + b_2 x_2$$



Interpretation of the model coefficients

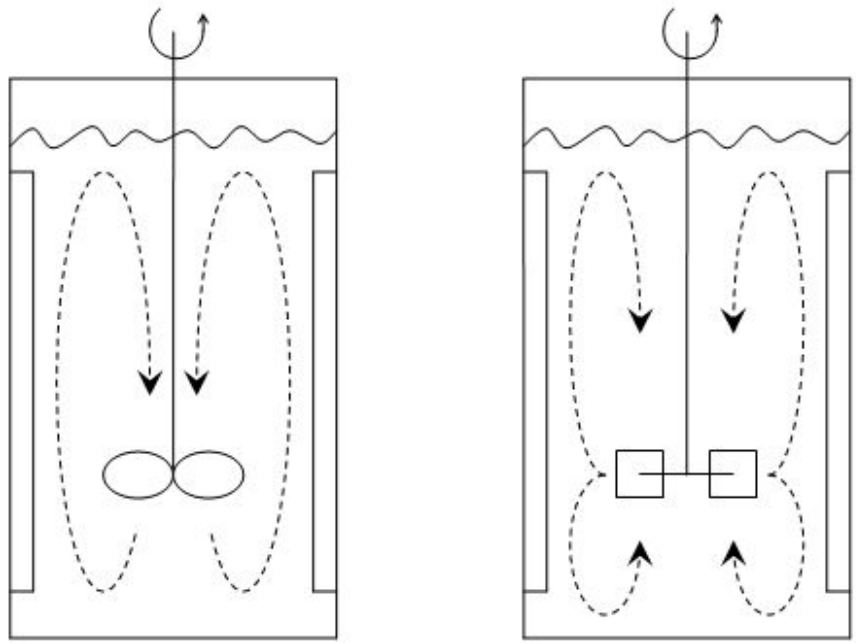
- $y = b_1 x_1 + b_2 x_2$
- Coefficient b_1 is the average change in **y** for a one unit change in x_1 **provided we hold** x_2 **fixed**
- *Example*: $y = b_T T + b_S S = -0.52T + 3.2S$
  - T = reactor temperature in Kelvin,
  - S = substrate concentration in g/L
  - y = yield in μg
  - $b_T = -0.52 \mu$g/K: decrease in yield for every 1 Kelvin increase in temperature, holding the substrate concentration fixed

Interpretation of the model coefficients

$$y = b_T T + b_S S = -0.52T + 3.2S$$

- CI for b_T spans zero: "*the effect of the temperature, controlling for concentration, is not significant*"
- The "controlling for" indicates the controlled variable was used in the model
- CI for b_S does not span zero: "*the effect of concentration, controlling for temperature, is to increase the yield by 3.2 μg for every 1g/L increase in concentration, S*"

Integer variables in the model



[From Wikipedia]

Yield = f(temperature, impeller type); impeller = [radial, axial]

- Build two models: one for radial, one for axial
  - Not efficient use of the data
  - Temperature effect is the same, independent of impeller
  - Fewer degrees of freedom in each model
  - Increases S_E

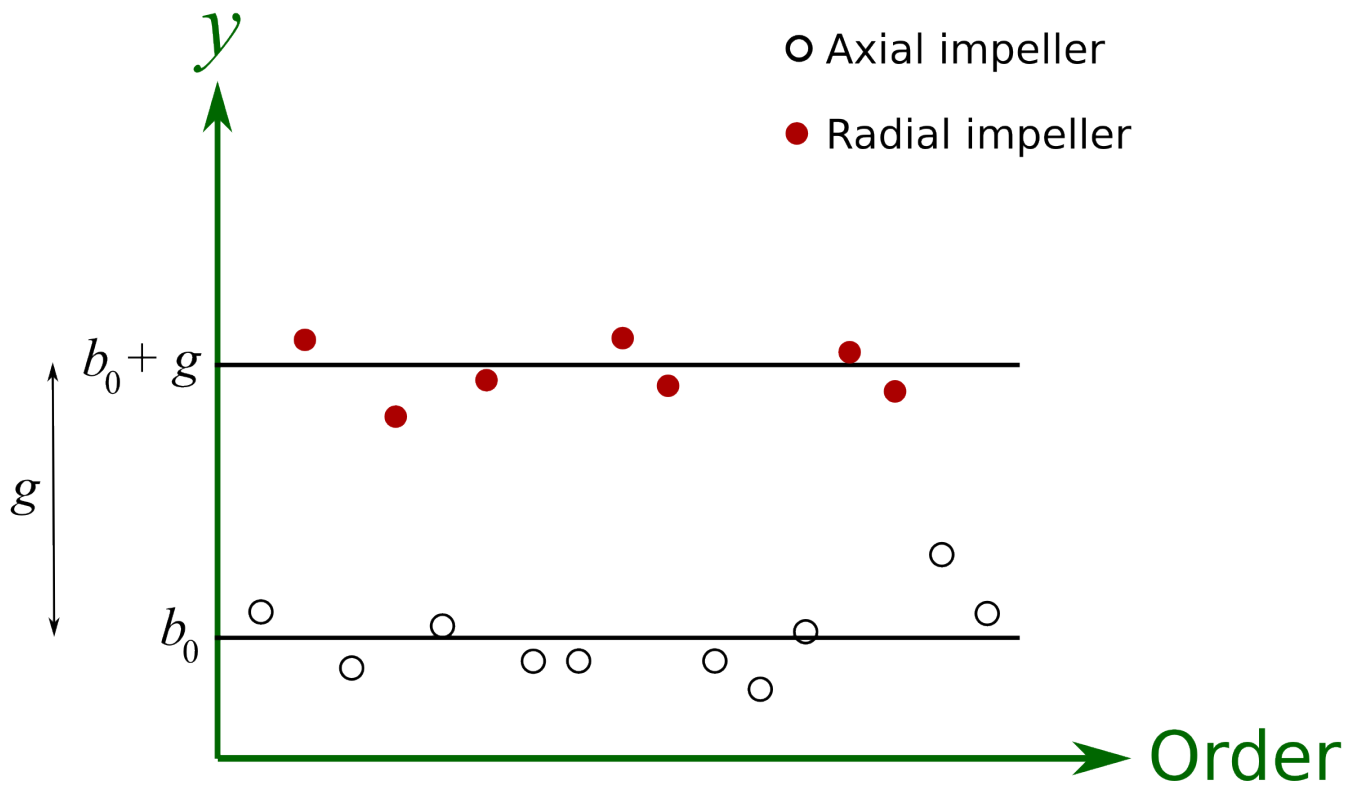
Integer variables in the model

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \gamma d + \varepsilon \\ y &= b_0 + b_1 T + g d_i + e_i \end{aligned}$$

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

Integer variables in the model

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



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

- g = difference in yield when changing the impeller from axial to radial

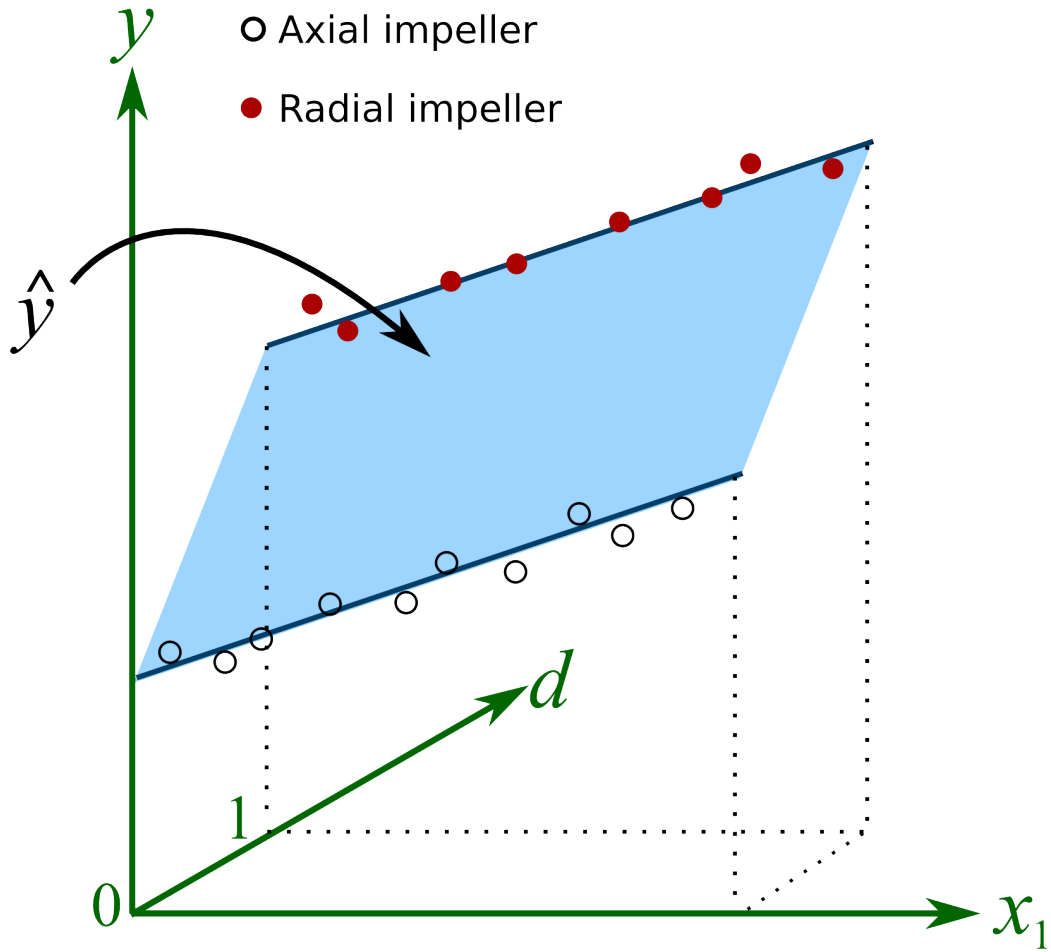
Integer variables in the model

Now let $\beta_1 \neq 0$:

$$\begin{aligned} &\text{Axial impellers:} && y = b_0 + b_1 T + 0 \\ &\text{Radial impellers:} && y = b_0 + b_1 T + g \end{aligned}$$

- The lines are still parallel
- If $g = -56\mu$g: the decrease in yield is expected to be 56 μg when changing from an axial to a radial impeller

Integer variables in the model



- Example: assume the 95% confidence interval for impeller parameter was: $-32\mu g \leq \gamma \leq 21\mu g$; i.e. no significant effect on yield
- CI for integer variables: no different to other variables

Integer variables in the model

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

- 2 integer variables:
  $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \gamma_1 \mathbf{d_1} + \gamma_2 \mathbf{d_2} + \varepsilon$
  - $d_{i1} = 0$ 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
- 3 integer variables:
  $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \gamma_1 \mathbf{d_1} + \gamma_2 \mathbf{d_2} + \gamma_3 \mathbf{d_3} + \varepsilon$
  - $d_{i1} = 1$ and $d_{i2} = 0$ and $d_{i3} = 0$ for Spain
  - $d_{i1} = 0$ and $d_{i2} = 1$ and $d_{i3} = 0$ for India
  - $d_{i1} = 0$ and $d_{i2} = 0$ and $d_{i3} = 1$ for Vietnam
  - These coefficients are more easily interpretable
  - Loose a single degree of freedom
  - IMPOSSIBLE to calculate model: the $\mathbf{X}^T\mathbf{X}$ is perfectly collinear, and cannot be inverted.

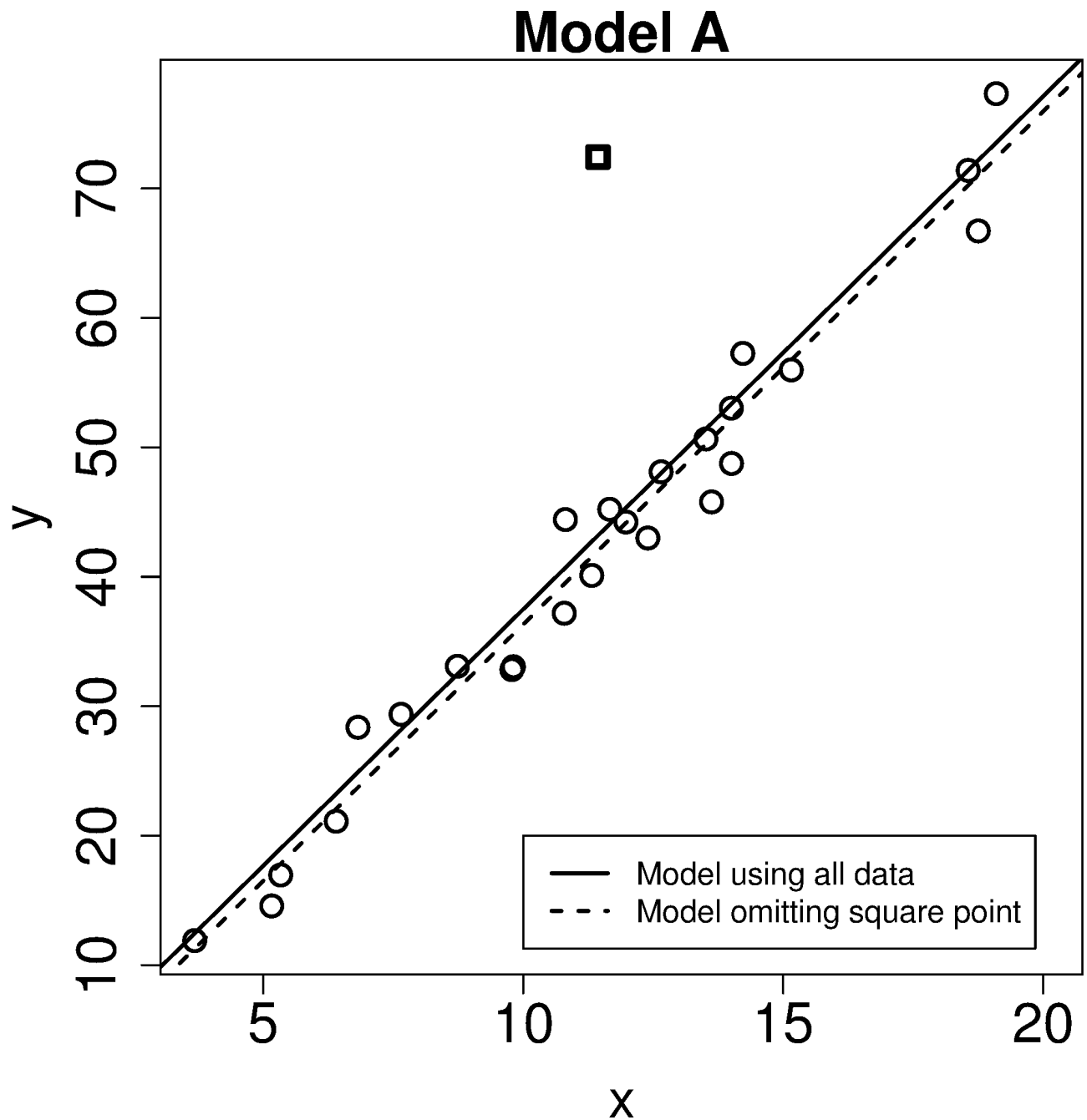
Outliers: discrepancy, leverage, and influence

- Unusual observations influence the model parameters and our interpretation
- Outliers are the most interesting data:
  - Do not delete without investigating them
  - If outlier is correct it could be very valuable information
  - Might have to add a new variable in model (avoid overfitting)

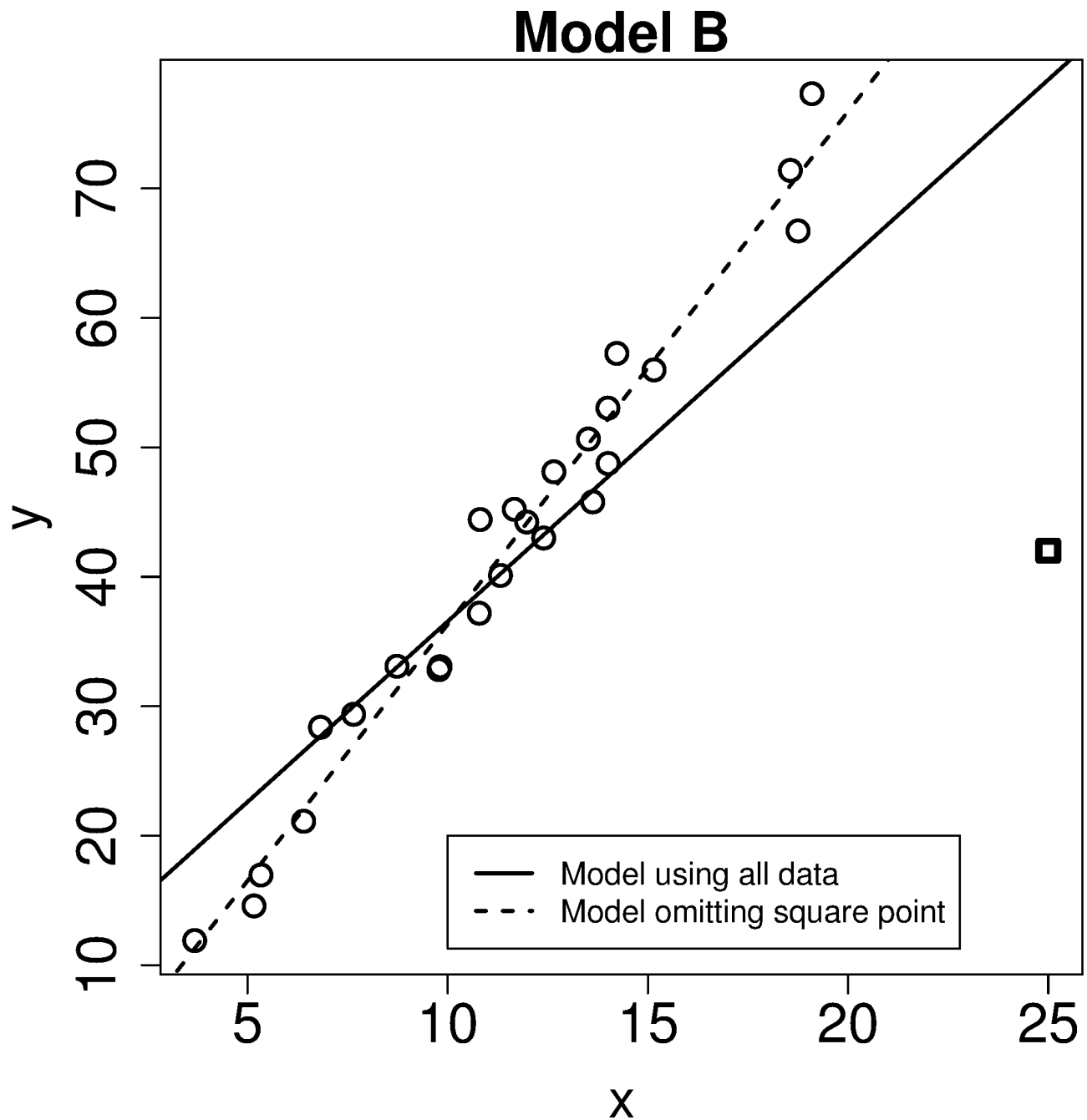
For unusual data:

$$\text{Leverage} \times \text{Discrepancy} = \text{Influence on the model}$$

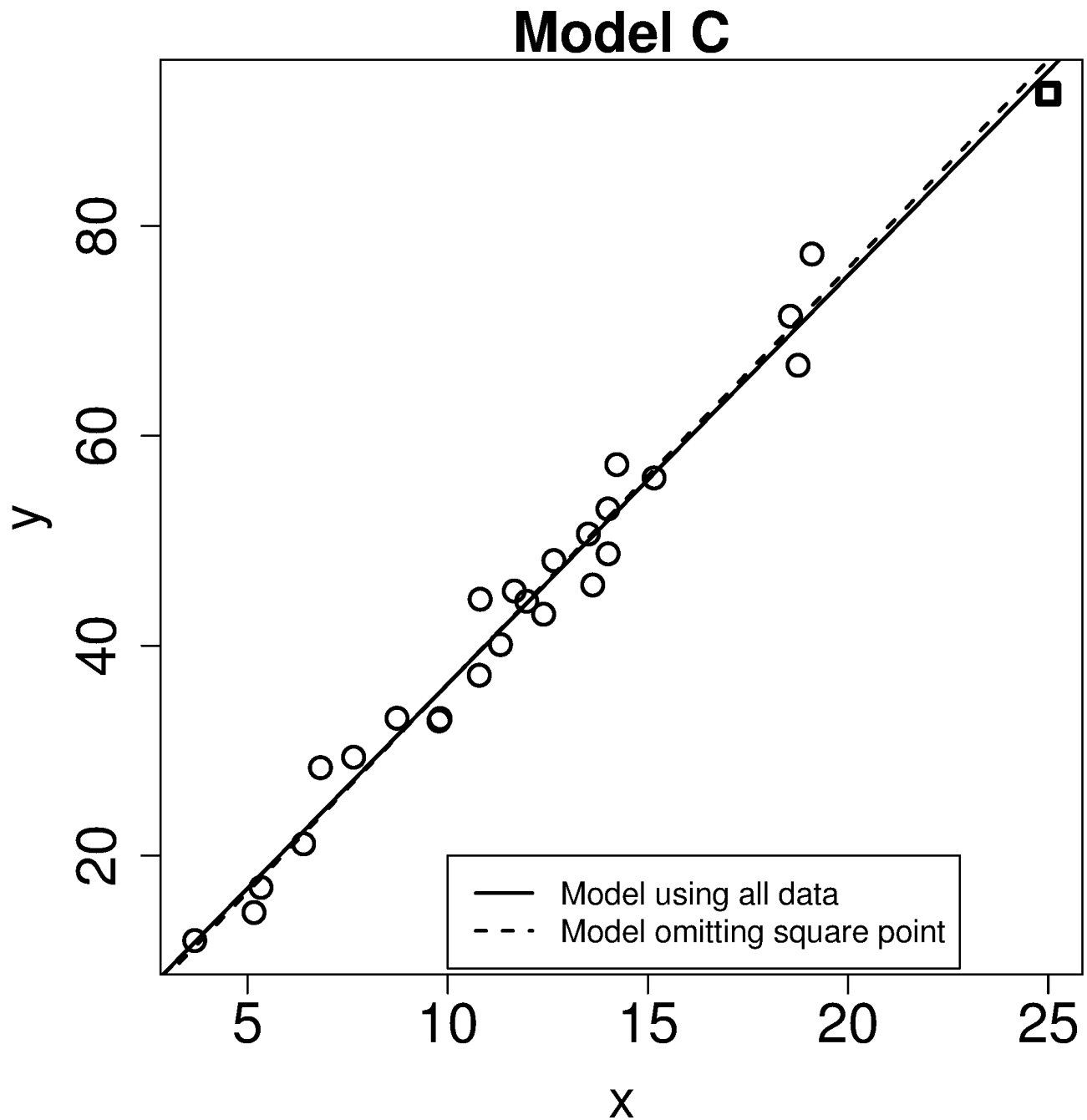
Model A


Model A
y
x
10
20
30
40
50
60
70
5
10
15
20
Model using all data
Model omitting square point

Model B

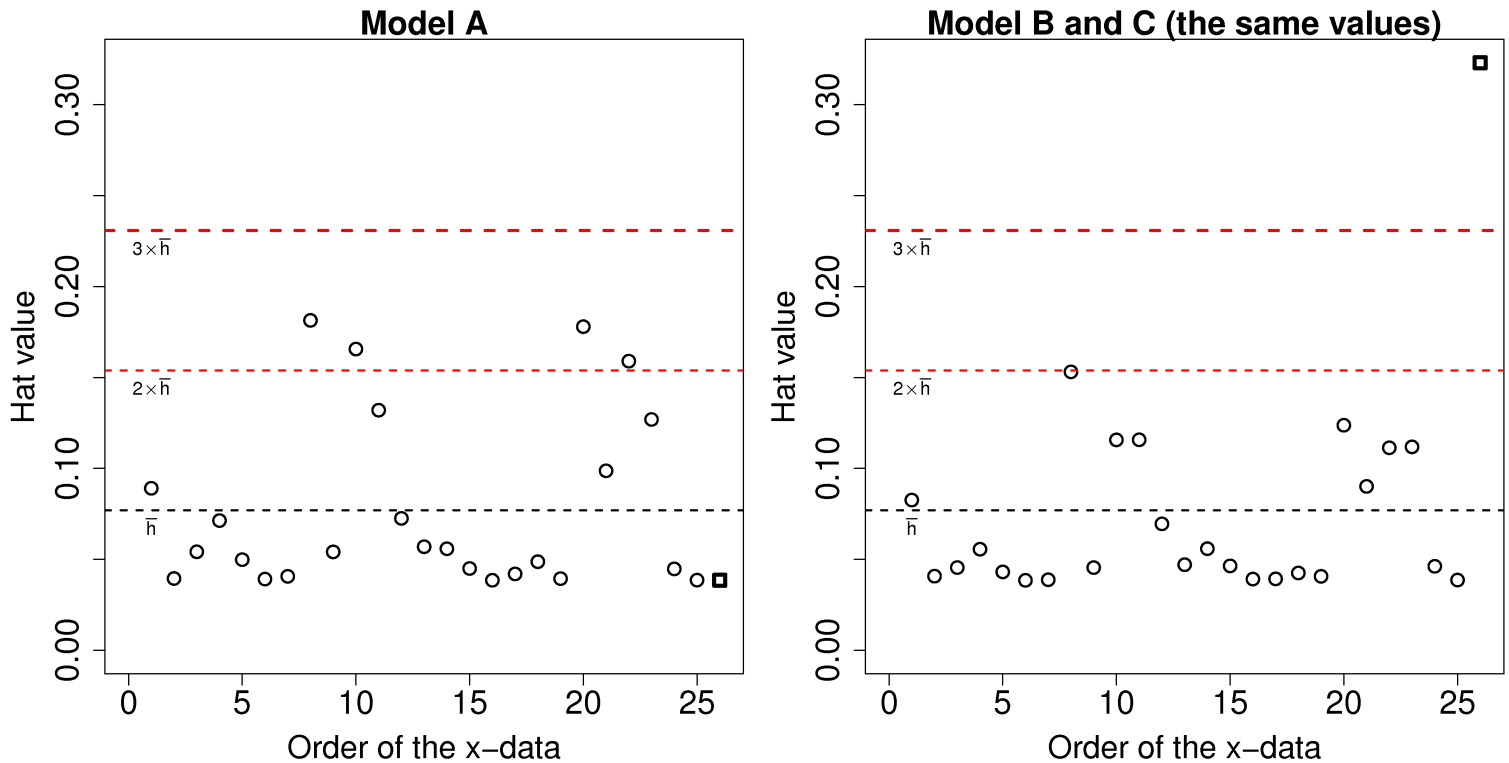

Model B
y
x
10
20
30
40
50
60
70
5
10
15
20
25
Model using all data
Model omitting square point

Model C


Model C
y
x
20
40
60
80
5
10
15
20
25
Model using all data
Model omitting square point

Leverage

- How much each observation contributes to $\hat{y}_i$; called hat value, h_i
- Measures how far the data point is from the center of the model
- $h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n} (x_j - \overline{x})^2}$ $\qquad \overline{h} = \frac{k}{n}$ $\qquad \frac{1}{n} \leq h_i \leq 1.0$
- Show limits at 2 and 3 times average hat value



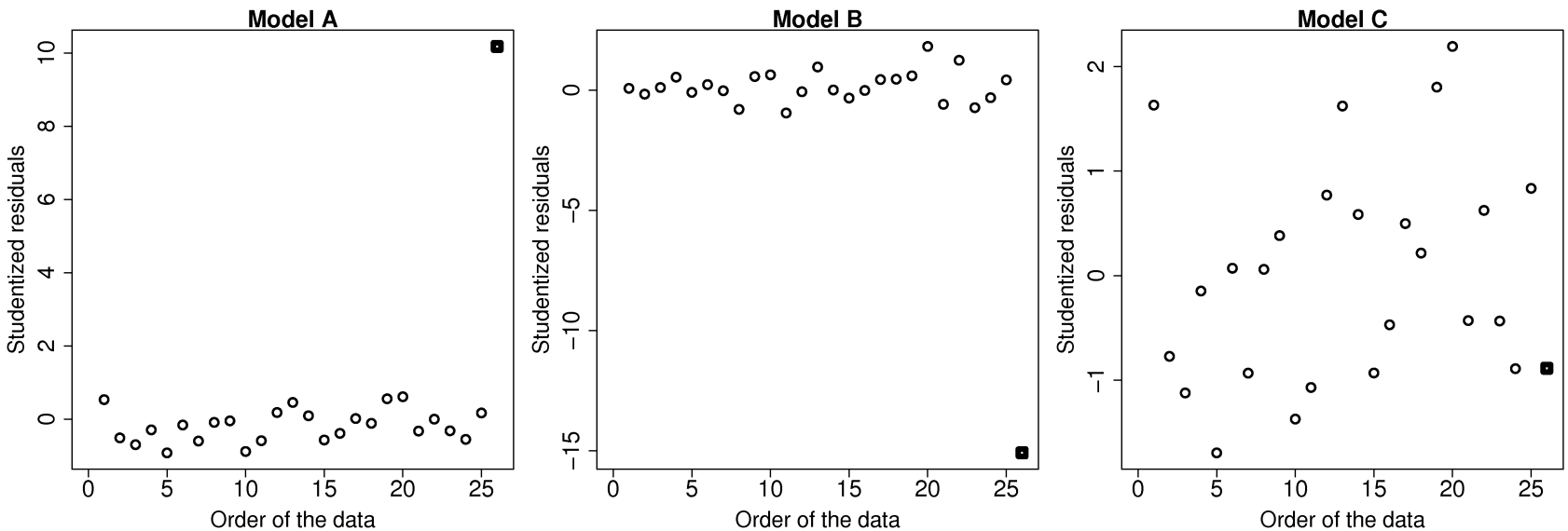
Discrepancy

Discrepancy: unusual data pair *in the context of the least squares model*

- Measured as residual distance, accounting for leverage: *studentized residuals*

$$e_i^* = \frac{e_i}{S_{E(-i)}\sqrt{1-h_i}} = \frac{e_i - 0}{\text{“corrected” standard deviation } \{e_i\}}$$

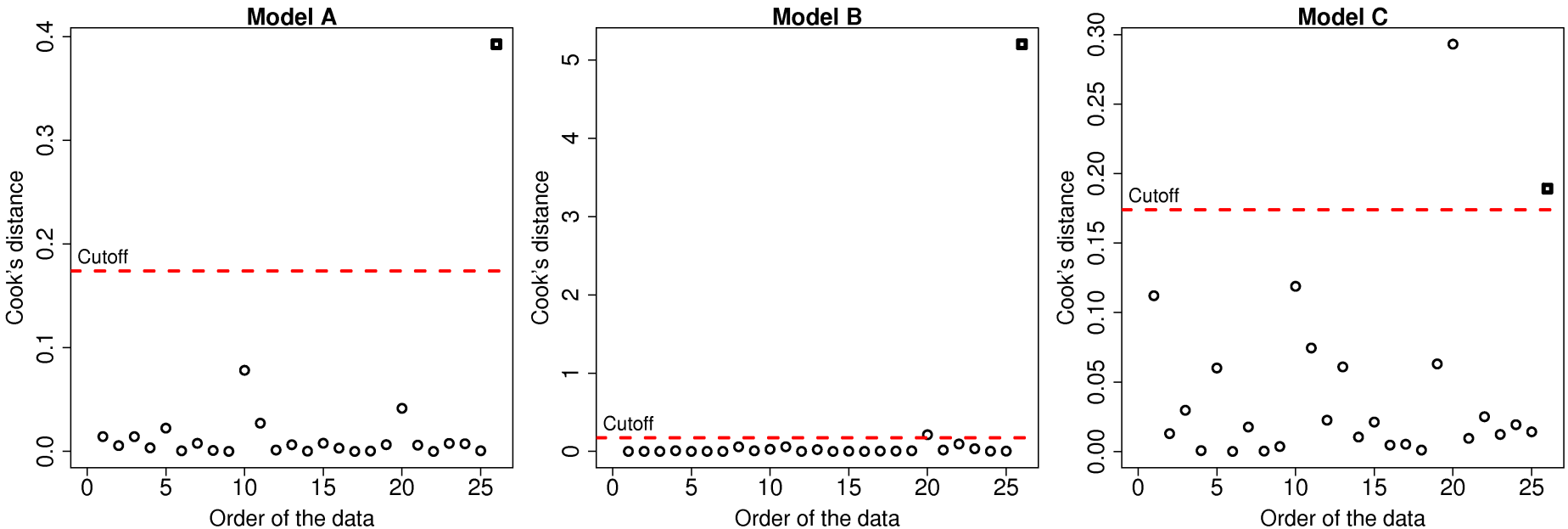
- $S_{E(-i)}$ = standard error without i^{th} point in model



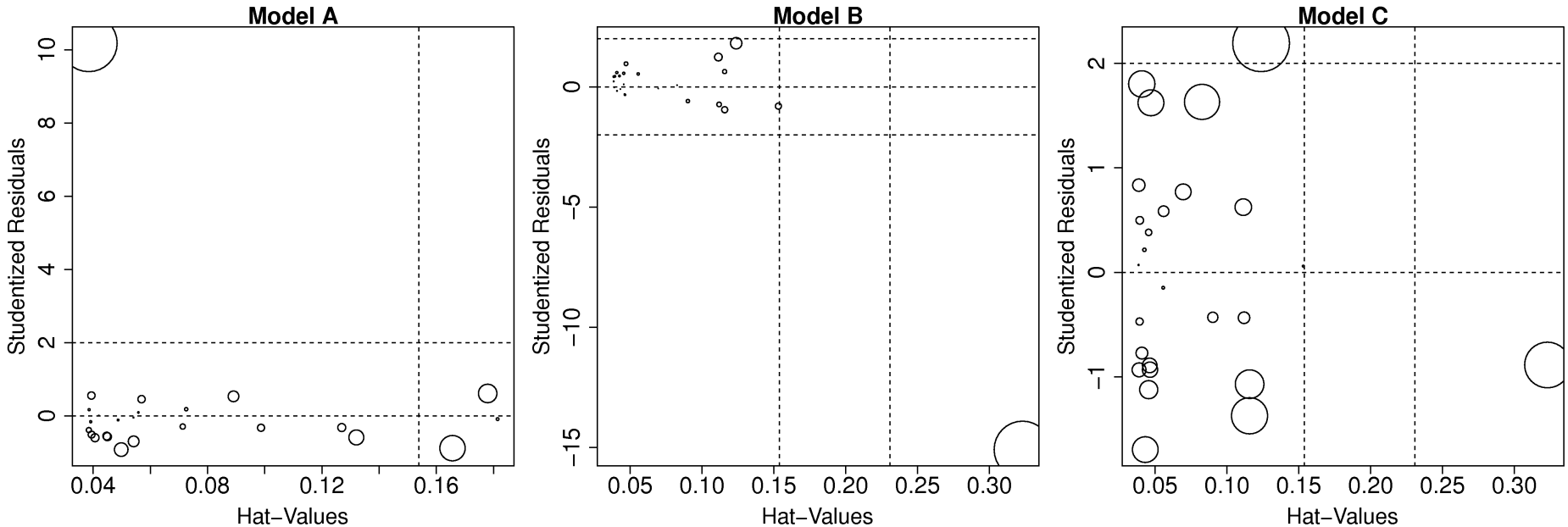
Influence

- *How much does the model change when omitting a data point*
- Influence on the model = Discrepancy × Leverage
- *Cook's statistic*, called $D_i = \dfrac{e_i^2}{k \times \frac{1}{n}\sum e_i^2} \times \dfrac{h_i}{1-h_i}$
- `cooks.distance(model)` in R
- Cut-off value: $D_i > \dfrac{4}{n-k}$

Influence


Model A
Cook's distance
Order of the data
Cutoff
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
Model B
Cook's distance
Order of the data
Cutoff
0
1
2
3
4
5
Model C
Cook's distance
Order of the data
Cutoff
0.00
0.05
0.10
0.15
0.20
0.25
0.30

Combining all measures of influence



- `library(car)`
- `model <- lm(y ~ x)`
- `influencePlot(model)`
- `influenceIndexPlot(model)`

Enrichment Topics

- Enrichment: *robust least squares models*
  - Useful for automated model building
  - Human reviewer not trained in understanding plots

- Enrichment: *Logistic modelling*
  - *y*-variable is discrete (yes/no or pass/fail)
  - Model then predicts probability of success/failure from various x's

Enrichment: Testing of least-squares models

1. Objective: learn more about our system
   - effect of one variable on another,
   - significant effect (use the CI)
2. Objective: prediction
   - How well does it work?

Gold standard: use an independent testing data set

- $\text{RMSEP} = \sqrt{\frac{1}{n}\sum_i^n \left(y_{\text{new},i} - \hat{y}_{\text{new},i}\right)^2}$
- RMSEE is the same as RMSEP, but for building the model
- The RMSEE $\approx S_E$ = standard error
- Or, use some other measure of "closeness"

Testing of least-squares models

Example:

- Need to build a predictive model for product viscosity using 3 x-variables
- Observations: one per day, from 2006 and 2007 (730 observations)

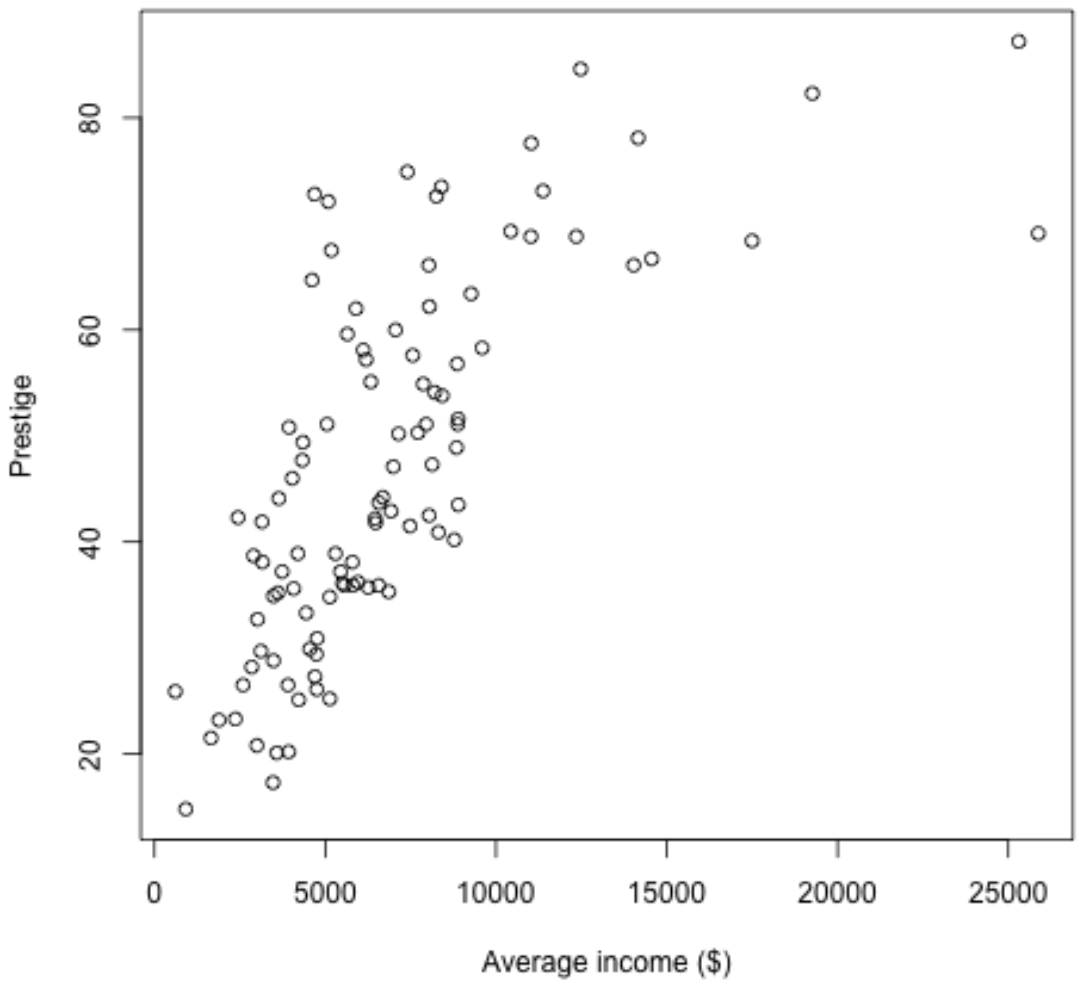
Which situation is better?

- **A**
  - Observations: 1, 3, 5, 7, ... 729 to build
  - Observations: 2, 4, 6, 8, ... 730 to test
- or **B**
  - Observations: 1 to 365 (2006 data) to build
  - Observations: 366 to 730 (2007 data) to test

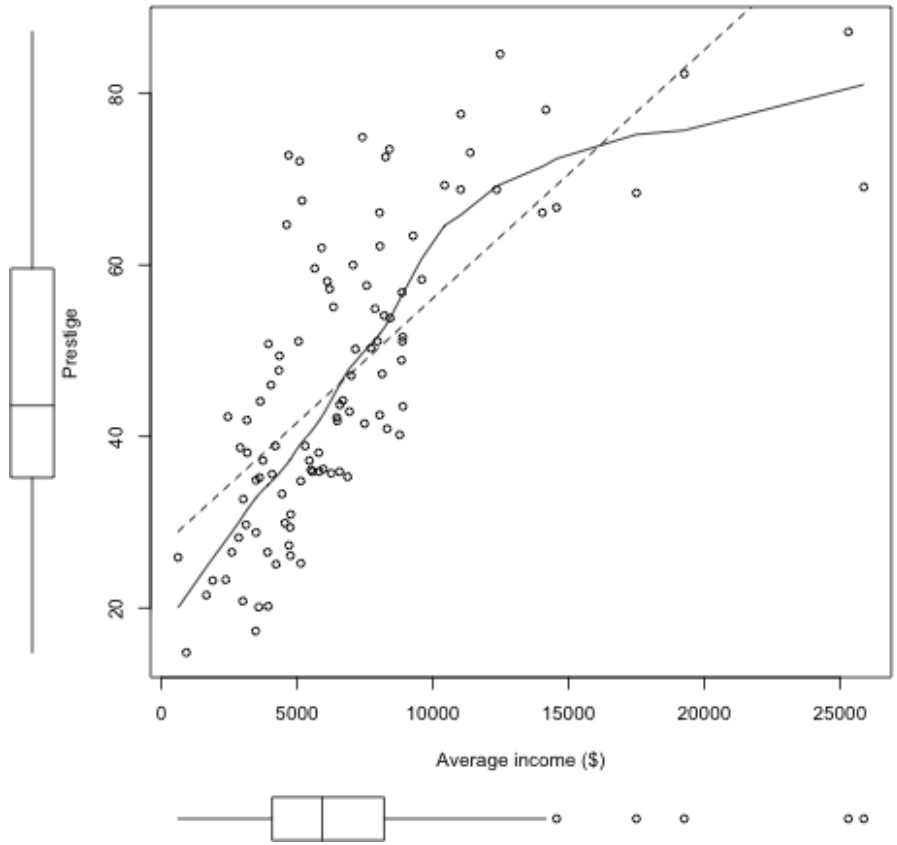
Enrichment: Nonparametric modelling

Relationship between x and y is of the form: $y = f(x) + \varepsilon$,

- The function (model), $f(x)$, is unspecified
- The model is usually a smooth function



Enrichment: Nonparametric modelling



- The smoothed line is the nonparametric function, $f(x)$
- See the Cleveland reference *Robust locally weighted regression and smoothing scatterplots*
- See the `lowess` function in R

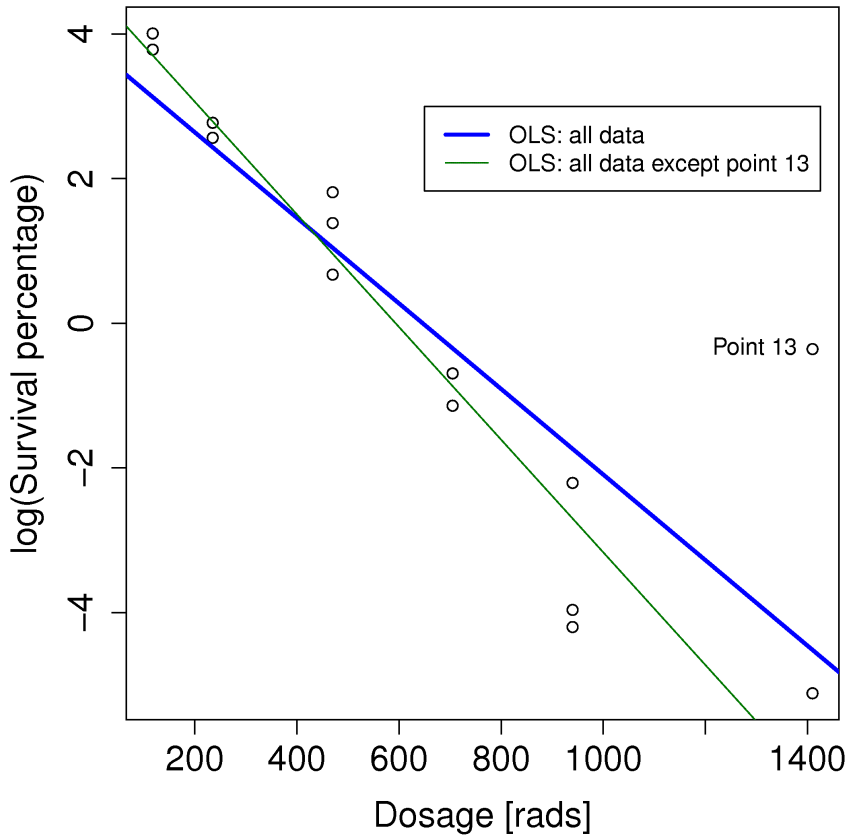
Enrichment: Bootstrapping

- Enrichment topic
- *Example*: bootstrapping strictly not required, but useful.
  - Model: $y = \beta_0 + \beta_1 x$
  - Interested in the slope β_1
  - Uncertainty associated with it? How much does it depend on the data?
  - The CI is:

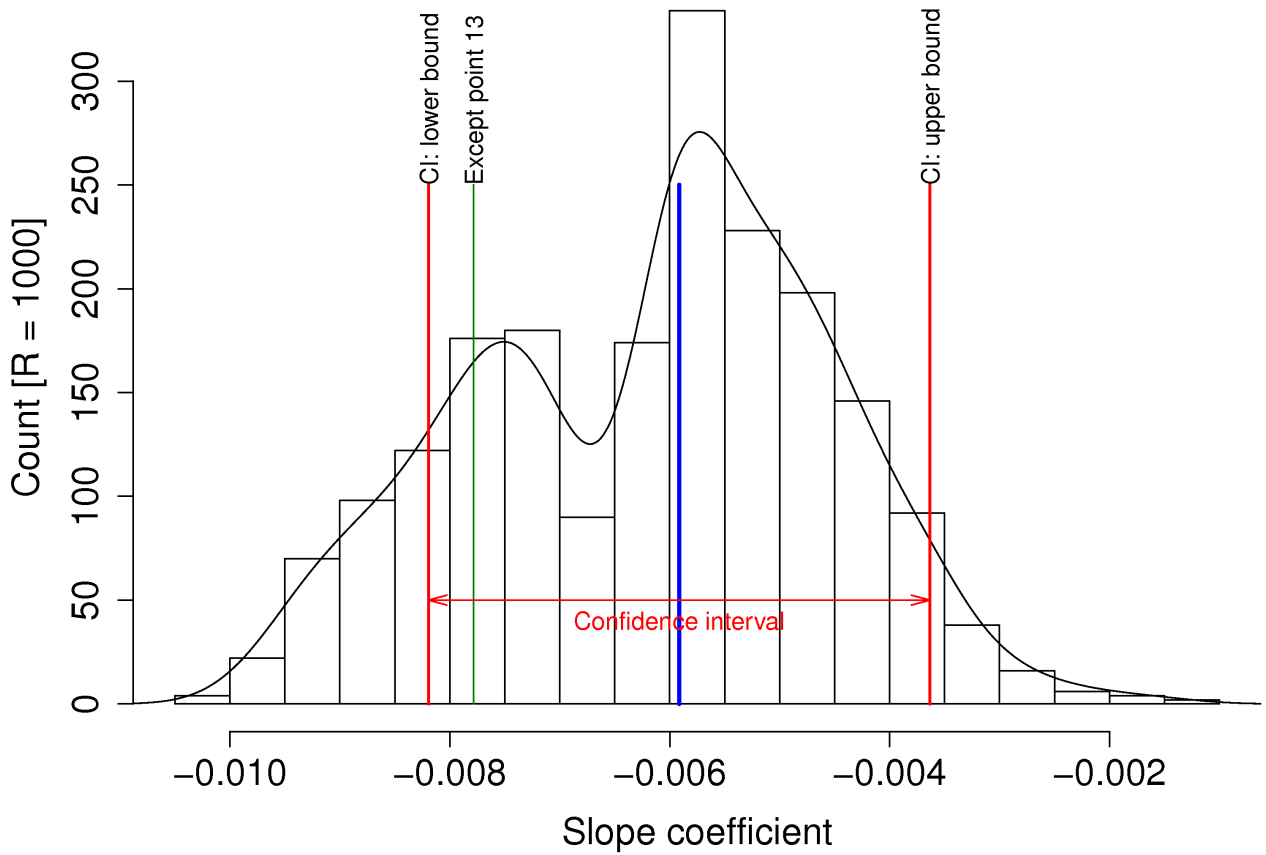
$$\begin{array}{rcccl} -c_t & \leq & \dfrac{b_1 - \beta_1}{S_E(b_1)} & \leq & +c_t \\ b_1 - c_t S_E(b_1) & \leq & \beta_1 & \leq & b_1 + c_t S_E(b_1) \end{array}$$

Bootstrapping example

- x = dose of radiation administered
- y = survival percentage



- Point 13 is influential
  - slope = −0.0059 with point 13
  - slope = −0.0078 without it

Bootstrapping example

- Randomly select n=14 rows from the raw data
  - [4] 9 1 6 5 5 5 4 11 8 2 9 13 8 13
  - [5] 4 7 12 10 13 12 12 3 8 4 14 1 6 4
  - [6] 3 4 10 14 14 9 10 10 8 11 1 1 4 14
- Calculate slope for each round
- CI: $-0.0082 \leq \beta_1 \leq -0.0036$