Lecture 10: Simple Linear Regression

STA261 – Probability & Statistics II

Ofir Harari

Department of Statistical Sciences

University of Toronto

UNIVERSITY OF TORONTO

Outline

UNIVERSITY OF TORONTO

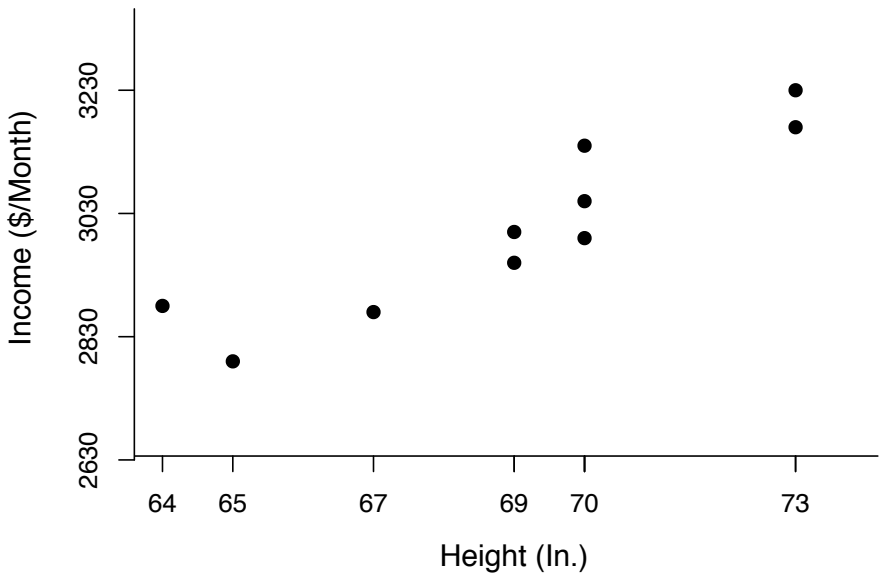
A motivating example: Income vs. Height

In a University of Pittsburgh study (reported in the Wall Street Journal , December 30, 1986), 250 MBA graduates, all about 30 years old, were polled and asked to report their height (in inches) and their monthly income (in USD). Here we shall focus on a subset of size 10 of those records.

joint distribution of 2 RVs.

	Height (X)	Income (Y)
1	70	2990
2	67	2870
3	69	2950
4	70	3140
5	65	2790
6	73	3230
7	64	2880
8	70	3050
9	69	3000
10	73	3170

Example (cont.)



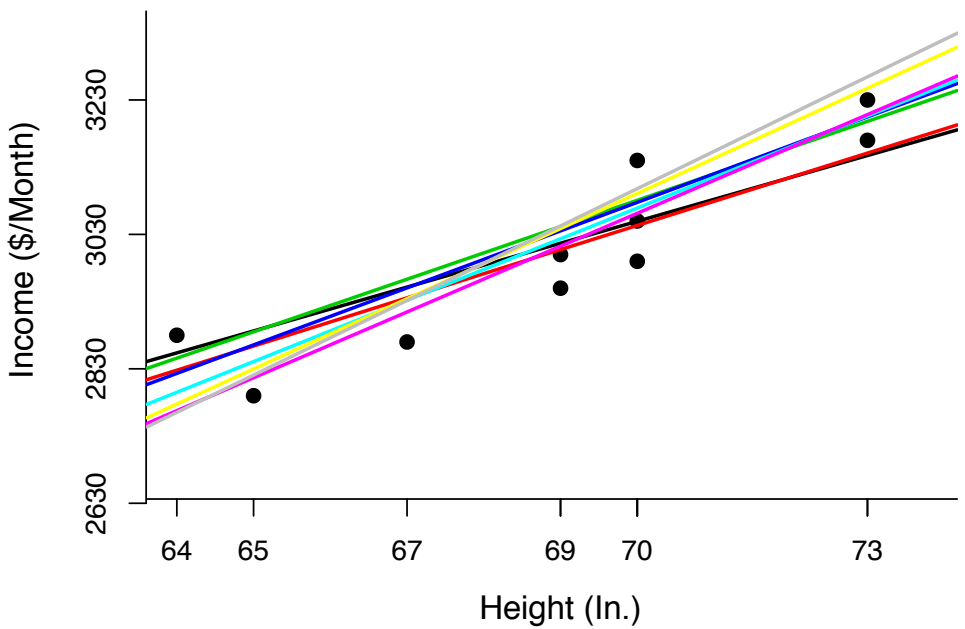
- Any noticeable linear trend here? Is it just an artifact of the small sample?
- Can we successfully predict the income of any person within this range? Can we quantify the uncertainty about our prediction?
- Does the change in height explain the change in income well? How can that be measured?

Example (cont.)

- We may consider fitting the simplest of models to the data -

$$\underbrace{Y}_{\text{Income,}} = \underbrace{\beta_0}_{\textit{intercept}} + \underbrace{\beta_1}_{\textit{slope}} \underbrace{X}_{\text{Height,}}$$

dependent variable | *independent variable*



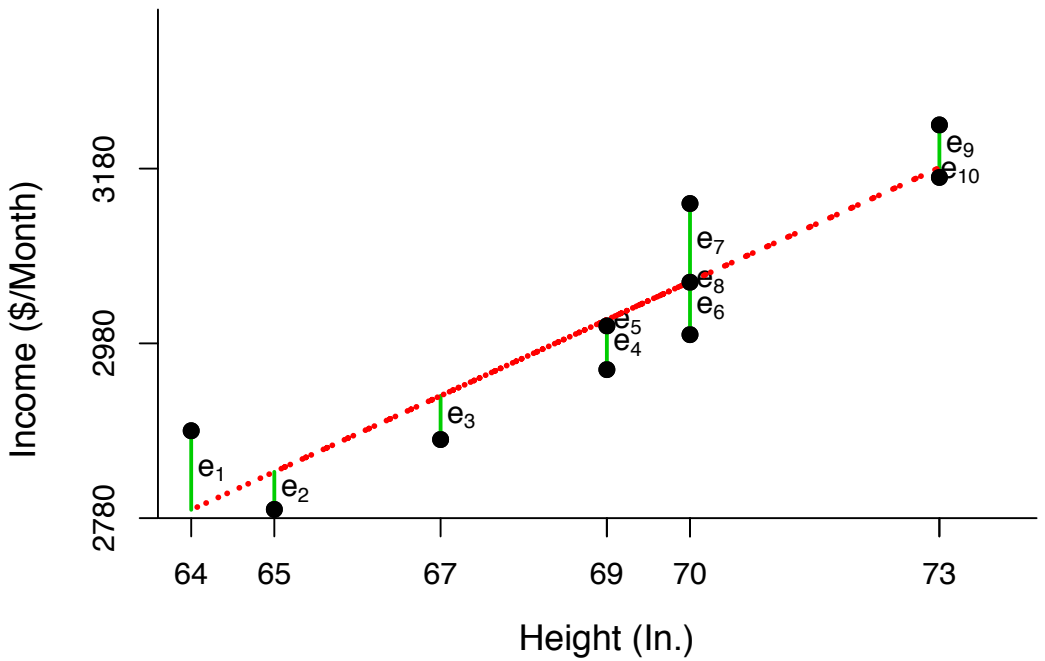
- Infinitely many possibilities - how do we choose the "best" one?

The method of Least Squares

- Obviously, no "perfect" linear model exists for the data. For example, three different persons, all 70" tall, earn three different salaries. This means that any linear fit would miss out on (at least) some of the observations.

- For fixed β_0 and β_1, denote

$$e_i = y_i - \beta_0 - \beta_1 x_i - \text{the } i\text{th } \textit{residual}$$



UNIVERSITY OF TORONTO

Least Squares (cont.)

- Quite clearly, we would like the residuals to be as small as possible. This should be reflected in the choice of β_0 and β_1.
- An obvious choice: choose a linear fit $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$, such that

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} |e_i| = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

  - Doable, but requires numerical optimization (linear programming).
  - No closed form for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ – makes studying their properties more challenging.

Definition

The *least squares* estimators of β_0 and β_1 are the minimizers of the *residual sum of squares*

$$\mathrm{RSS}(\beta_0, \beta_1) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

UNIVERSITY OF TORONTO

The Least Squares estimators

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Proposition

The least squares estimators of β_0 and β_1 are given by

$$\begin{cases} \widehat{\beta}_1 &= \dfrac{S_{XY}}{S_X^2} \text{ and} \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x}, \end{cases}$$

where $S_{XY} = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ is the *sample covariance* of X and Y, and S_X^2 is the sample variance of X.

consistent estimator of true covariance

The Least Squares estimators (cont.)

Proof:

- It will be useful from now on to use the (easily verifiable) identities –

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2, \text{ and}$$

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

- Solve finding minimum of sum of residue by taking partial derivatives

$$\frac{\partial \text{RSS}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \Longrightarrow \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}, \text{ and}$$

$$\frac{\partial \text{RSS}}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i\,(y_i - \beta_0 - \beta_1 x_i) = 0 \Longrightarrow \sum_{i=1}^{n} x_i y_i = \widehat{\beta}_0\sum_{i=1}^{n} x_i + \widehat{\beta}_1\sum_{i=1}^{n} x_i^2$$

The Least Squares estimators (cont.)

Proof (cont.):

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$

$$\sum_{i=1}^{n} x_i y_i = \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \overline{y} \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$= n\overline{x}\,\overline{y} + \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right) = n\overline{x}\,\overline{y} + \widehat{\beta}_1 \sum_{i=1}^{n} (x_i - \overline{x})^2$$

prop. 1

prop. 2

$$\Longrightarrow \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{j=1}^{n} (x_j - \overline{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} = \frac{S_{XY}}{S_X^2}.$$

multiply by 1/(n-1) on num and denom

- It is easy to verify that the Hessian (second derivatives) matrix is positive definite, hence the least squares estimators minimize the RSS.
- Since $\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}$, the least squares fit always passes through the averages point $(\overline{x}, \overline{y})$.

Back to the Income vs. Height example

	Height (X)	Income (Y)	
1	70	2990	
2	67	2870	
3	69	2950	
4	70	3140	
5	65	2790	
6	73	3230	
7	64	2880	
8	70	3050	
9	69	3000	
10	73	3170	
	$\sum x_i = 690$	$\sum x_i^2 = 47{,}690$	
	$\sum y_i = 30{,}070$	$\sum y_i^2 = 90{,}601{,}900$	$\sum x_i y_i = 2{,}078{,}310$

formula facilitates computation

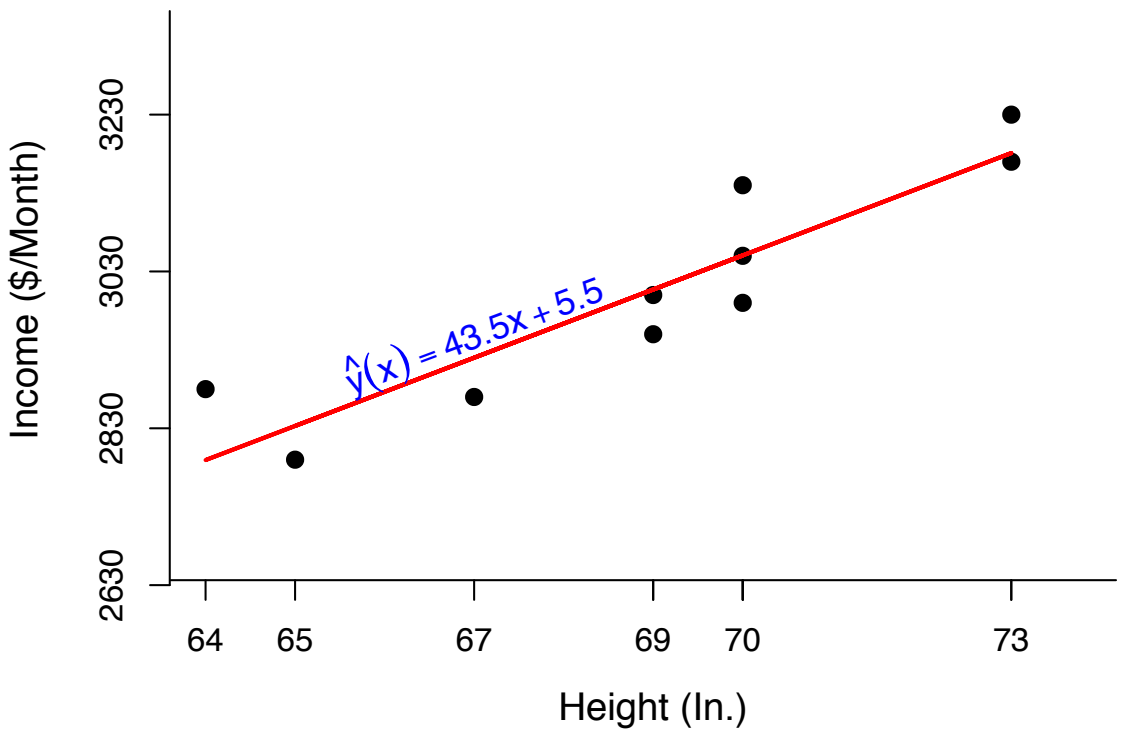
$$S_{XY} = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{n-1} = \frac{2078310 - 10 \times 69 \times 3007}{10-1} = 386.67,$$

$$S_X^2 = \frac{\sum x_i^2 - n\overline{x}^2}{n-1} = \frac{47690 - 10 \times 69^2}{10-1} = 8.89$$

The Income vs. Height example (cont.)

$$\overline{x} = 69 \ , \ \overline{y} = 3007 \ , \ S_{XY} = 386.67 \ , \ S_X^2 = 8.89$$

$$\begin{cases} \widehat{\beta}_1 &= \dfrac{S_{XY}}{S_X^2} = \dfrac{386.67}{8.89} = 43.5, \\ \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} = 3007 - 43.5 \times 69 = 5.5, \end{cases}$$



UNIVERSITY OF TORONTO

The normal equations

- Denote $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \quad i = 1, \ldots, n$
- The residuals, as before, are $e_i = y_i - \widehat{y}_i, \quad i = 1 \ldots, n$
- When deriving $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we showed that the residuals of the least square fit satisfy the following equations – i.e. first order partial w.r.t. beta_0 and beta_1

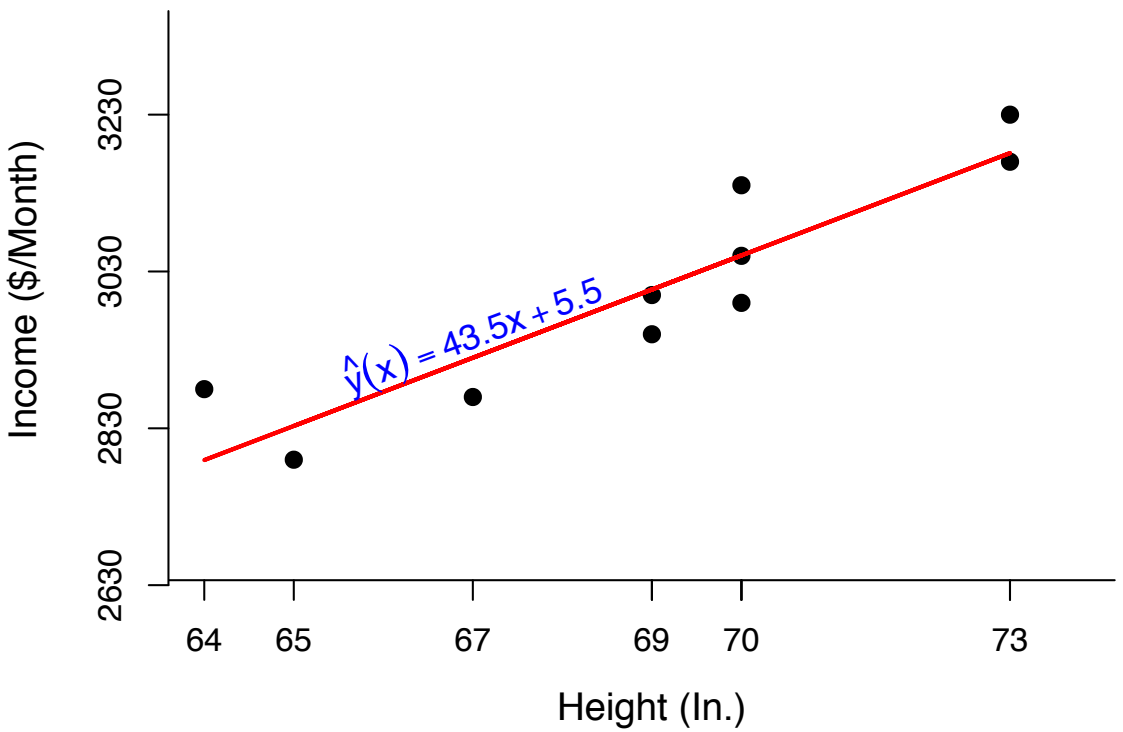
$$0 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i \right) = \sum_{i=1}^{n} e_i, \tag{1}$$

$$0 = \sum_{i=1}^{n} x_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) = \sum_{i=1}^{n} x_i \left(y_i - \widehat{y}_i \right) = \sum_{i=1}^{n} x_i e_i, \tag{2}$$

famously known as the *normal equations*. These two equations will come handy later on.

A statistical model

- So far what we have done has been purely numerical analysis – and nothing to do with statistics.
- A question that must be asked: if we believe that Y is indeed linear in X, why aren't all the observations lying exactly on the straight line?



A statistical model (cont.)

- Statistically, we may explain these fluctuations of the observations about the linear trend by the existence of "random noise" in the model, e.g.

$$y(x) = \beta_0 + \beta_1 x + \varepsilon(x),$$

where x (the independent variable) is not thought of as a random variable (the *fixed X* assumption), and $y(x)$ is random (or *stochastic*) through the inclusion of the random noise, $\varepsilon(x)$.

- Further, the standard model assumes a "white noise": by Normal Equation
  1. $\mathbb{E}[\varepsilon(x)] = 0 \quad \forall x$ (why is this assumption an obvious one?), and
  2. $\mathrm{Cov}\big(\varepsilon(x), \varepsilon(x')\big) = \begin{cases} \sigma^2 & , \quad x = x' \\ 0 & , \quad x \neq x' \end{cases}$ y at different position are uncorrelated

- When restricted to the observed data, the model can be written as

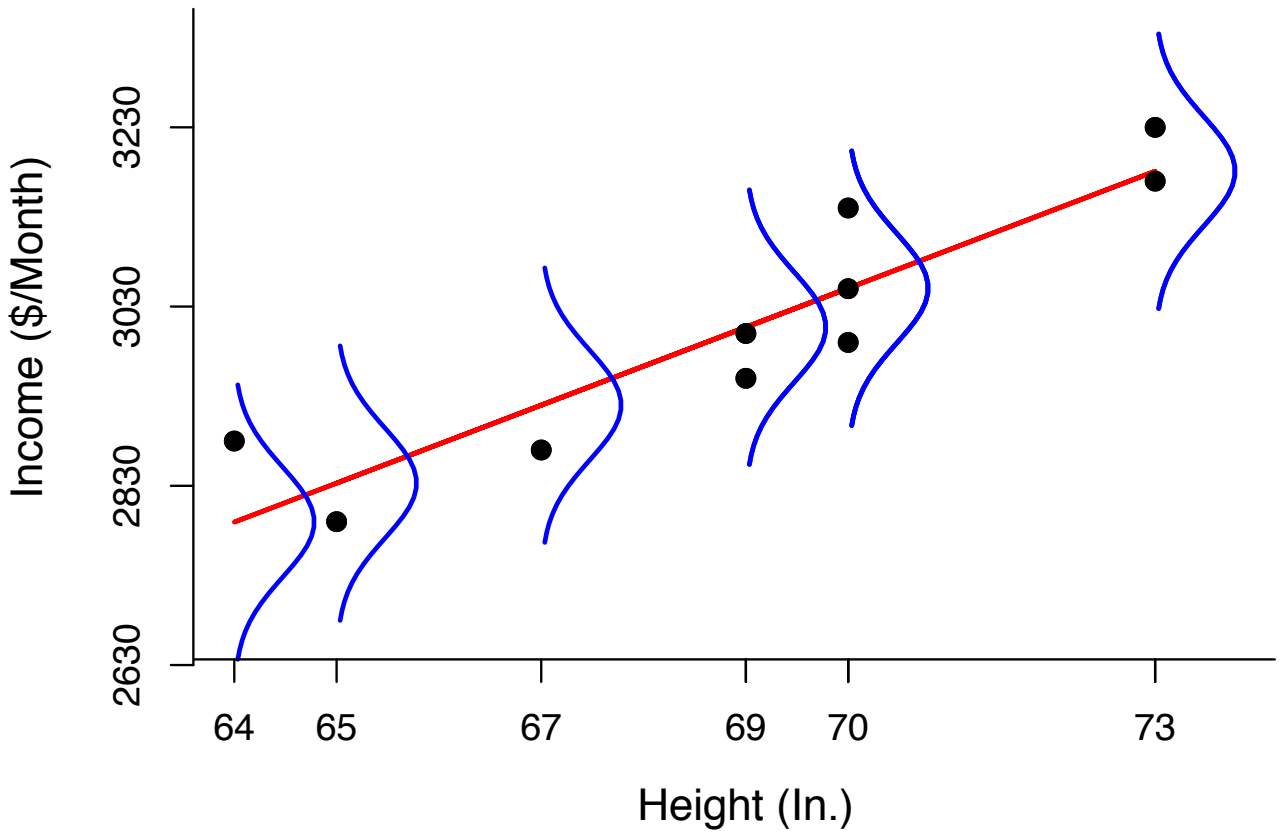
COV(X,Y) = E(XY) - E(X)E(Y)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

with $\mathbb{E}[\varepsilon_i] = 0 \ \forall i$, $\mathrm{Var}[\varepsilon_i] = \sigma^2 \ \forall i$ and $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$.

note x is not RV and y is RV two different error term uncorrelated

A statistical model (cont.)

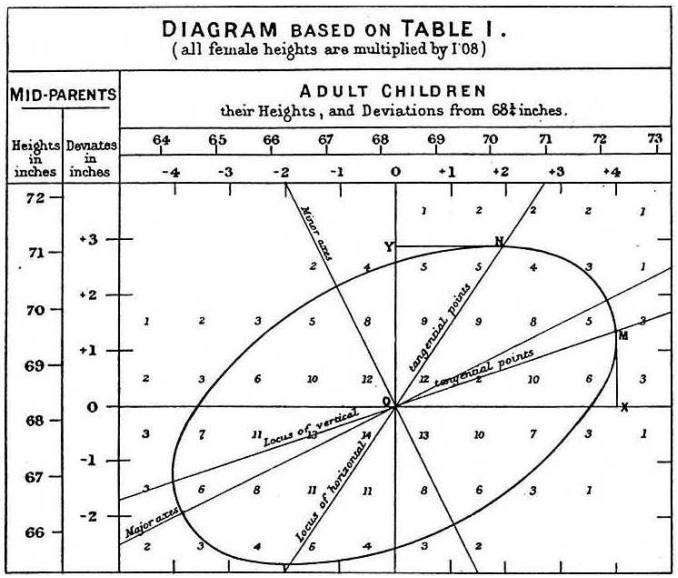


Var(epsilon) = 0 for all epsilon

⋆ The assumption that the variance around the regression line is the same for all values of X is called *homoscedasticity*, and is crucial for all of our following analyses.

a very important assumption

Historical perspective: why "Regression"?



In his 1885 work, Sir Francis Galton related the heights of children to the average height of their parents. The resultant least squares fit had a slope < 1, and thus the mean height of children of taller (shorter) than average parents was closer to the mean height of all children than the mean height of their parents was to the mean height of all parents. Galton called this phenomenon "regression towards mediocrity" – and the process of fitting such lines became known as "regression".

The LS estimators as linear estimators

Definition

Let $y_1, \ldots, y_n \sim f_\theta$. Any estimator of θ of the form

$$\widehat{\theta} = \sum_{i=1}^{n} c_i y_i$$

is called a *linear estimator*.

linear observation of observations

sample mean is linear estimator

- Let us show now that $\widehat{\beta}_1$ is a linear estimator of β_1:

$$\widehat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_j (x_j - \overline{x})^2} = \frac{\sum_i (x_i - \overline{x}) y_i}{\sum_j (x_j - \overline{x})^2} - \boxed{\frac{\overline{y} \sum_i (x_i - \overline{x})}{\sum_j (x_j - \overline{x})^2}}$$

= 0 because num = 0

$$= \frac{\sum_i (x_i - \overline{x}) y_i}{\sum_j (x_j - \overline{x})^2} = \sum_i a_i y_i,$$

for $a_i = \boxed{\dfrac{x_i - \overline{x}}{\sum_j (x_j - \overline{x})^2}}$.

The LS estimators as linear estimators (cont.)

- Similarly,

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = \frac{1}{n}\sum_i y_i - \overline{x}\frac{\sum_i (x_i - \overline{x}) y_i}{\sum_j (x_j - \overline{x})^2}$$

$$= \sum_i \left[\frac{1}{n} - \frac{\overline{x}(x_i - \overline{x})}{\sum_j (x_j - \overline{x})^2} \right] y_i = \sum_i b_i y_i,$$

proof that `

for $b_i = \frac{1}{n} - \frac{\overline{x}(x_i - \overline{x})}{\sum_j (x_j - \overline{x})^2}$, hence $\widehat{\beta}_0$ is a linear estimator of β_0.

- We are now ready to calculate the mean and the variance of the LS estimators –

$$\mathbb{E}[\widehat{\beta}_1] = \frac{\sum_i (x_i - \overline{x})\mathbb{E}[y_i]}{\sum_j (x_j - \overline{x})^2} = \frac{\sum_i (x_i - \overline{x})(\beta_0 + \beta_1 x_i)}{\sum_j (x_j - \overline{x})^2}$$

$$= \beta_0 \underbrace{\frac{\sum_i (x_i - \overline{x})}{\sum_j (x_j - \overline{x})^2}}_{0} + \beta_1 \frac{\sum_i (x_i - \overline{x}) x_i}{\sum_j (x_j - \overline{x})^2} = \beta_1 \frac{\sum_i (x_i - \overline{x}) x_i}{\sum_j (x_j - \overline{x})^2}$$

numerator = 0

The LS estimators as linear estimators (cont.)

$$\mathbb{E}[\widehat{\beta}_1] = \beta_1 \frac{\sum_i (x_i - \overline{x}) x_i}{\sum_j (x_j - \overline{x})^2} = \beta_1 \left[\frac{\sum_i (x_i - \overline{x}) x_i}{\sum_j (x_j - \overline{x})^2} - \underbrace{\frac{\overline{x} \sum_i (x_i - \overline{x})}{\sum_j (x_j - \overline{x})^2}}_{0} \right]$$

$$= \beta_1 \frac{\sum_i (x_i - \overline{x})^2}{\sum_j (x_j - \overline{x})^2} = \beta_1,$$

hence $\widehat{\beta}_1$ is an unbiased linear estimator of β_1.

- Showing that $\widehat{\beta}_0$ is unbiased is similar.
- The variance of the slope LS estimator is given by

y_i are uncorrelated; since only epsilon are uncorrelated

$$\mathrm{Var}[\widehat{\beta}_1] = \mathrm{Var}\left[\sum_i a_i y_i\right] = \sum_i a_i^2 \mathrm{Var}[y_i] = \sum_i a_i^2 \mathrm{Var}[\beta_0 + \beta_1 x_i + \varepsilon_i]$$

$$= \sum_i a_i^2 \mathrm{Var}[\varepsilon_i] = \sigma^2 \sum_i a_i^2 = \sigma^2 \sum_i \frac{(x_i - \overline{x})^2}{\left\{\sum_j (x_i - \overline{x})^2\right\}^2} = \frac{\sigma^2}{\sum_j (x_i - \overline{x})^2}.$$

beta's and x is constant and Var{epsilon} = sigma^2

The LS estimators as linear estimators (cont.)

- Some more work is required to show that

$$\mathrm{Var}[\widehat{\beta}_0] = \sigma^2 \left\{ \frac{1}{n} + \frac{\overline{x}^2}{\sum_j (x_j - \overline{x})^2} \right\} \quad \text{and} \quad \mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\sigma^2 \overline{\mathrm{x}}}{\sum_{\mathrm{j}} (\mathrm{x_j} - \overline{\mathrm{x}})^2}.$$

- The following result, brought here without a proof, tells us that the LS estimators are the Best Linear Unbiased Estimators (BLUE) of the linear regression coefficients.

The Gauss–Markov Theorem

Under the standard model assumptions, no linear unbiased estimator of β_0 (β_1) has a smaller variance than the least squares estimator $\widehat{\beta}_0$ ($\widehat{\beta}_1$).

UNIVERSITY OF TORONTO

The correlation coefficient

Reminder invariance under measurement system transformation

The *correlation coefficient* of random variables X and Y is

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

an unknown RV

where σ_X and σ_Y are the standard deviations of X and Y, respectively.

- Recall that $|\rho_{XY}| \leq 1$, where equality holds $\iff$ X and Y are perfect linear functions of one another. Hence $|\rho_{XY}|$ measures the strength of the linear relationship between X and Y.

- Similarly, for a sample of pairs $\{(x_i, y_i)\}$, we can define the *sample correlation coefficient* of X and Y ~~+~~used for estimating true rho, which happens to be consistent

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}.$$

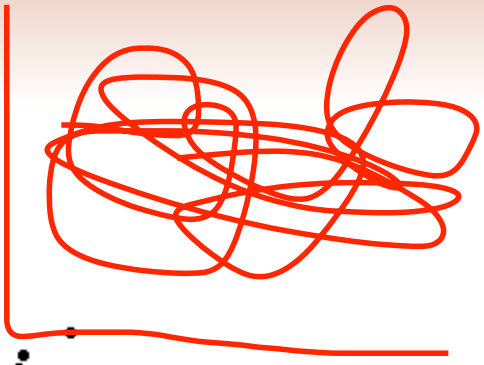
This is an estimator of ρ_{XY}, that has similar properties.

| r | bounded by 1 and is perfectly correlated at -1 and 1

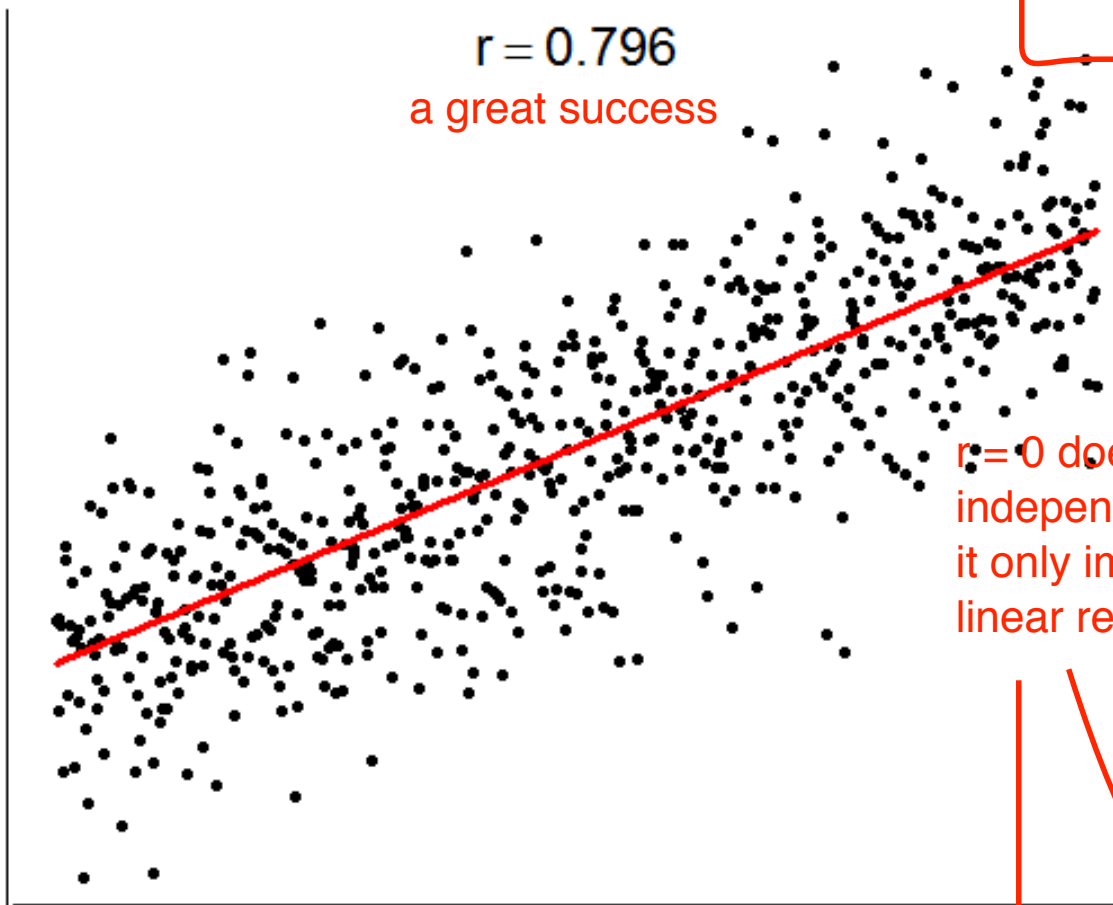
UNIVERSITY OF TORONTO

r = 0; but in this example two RV is independent

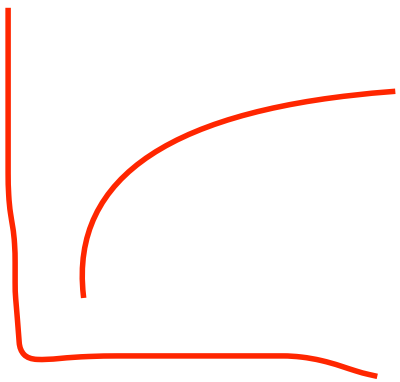
Examples of linear association



r = 0

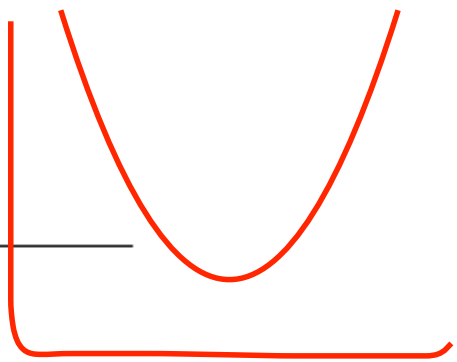


r = 0 does not mean independent in this example; it only implies there is no linear relationship;



r = 0.85

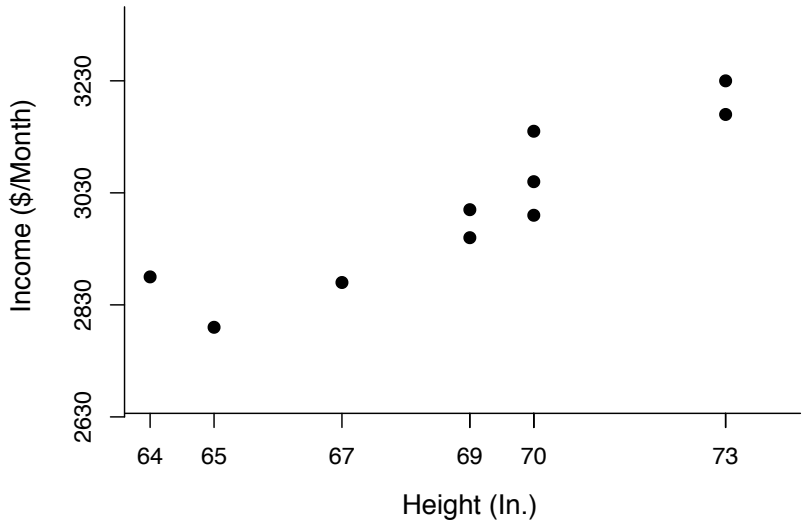
Text



r = -0.01

Back to the Income vs. Height example

What is your assessment of the strength of the linear association here?



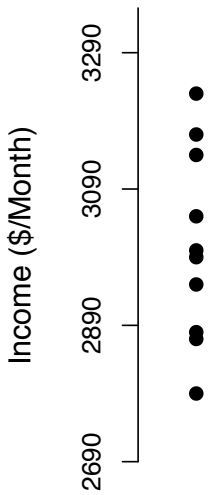
$$S_{XY} = 386.67 \ , \ S_X^2 = 8.89 \ , \ \overline{y} = 3{,}007 \ , \ \sum_{i=1}^{10} y_i^2 = 90{,}601{,}900$$

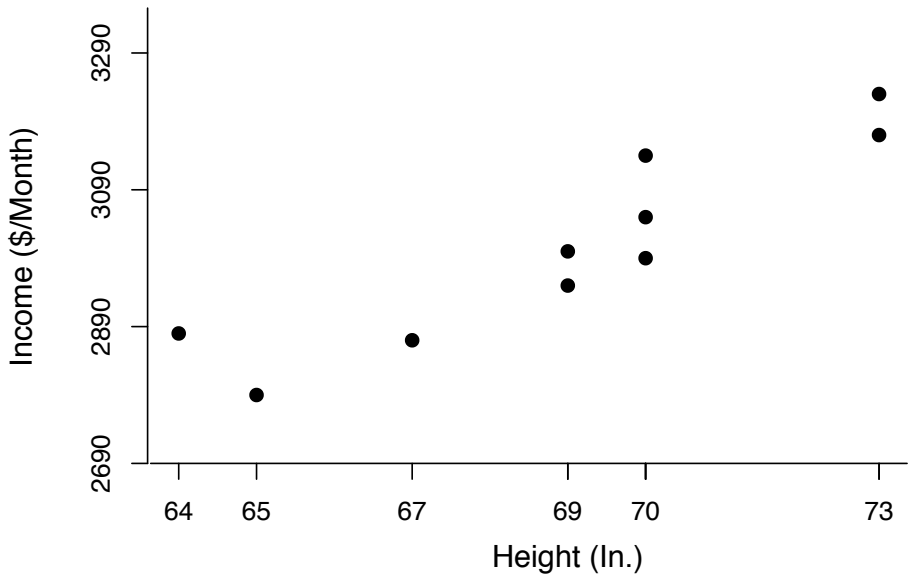
$$S_Y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}{n-1} = \frac{90601900 - 10 \times (3007)^2}{10-1} = 20156.67,$$

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{389.67}{\sqrt{8.89}\sqrt{20156.67}} = 0.9135$$

quite convincing..
strong linear relationship
n small.. so CI quite wide

Explained variation





What appears to be complete randomness when plotting 'Income' alone, suddenly follows a clear pattern when plotted vs. 'Height'.

- It seems like a lot of the variability of 'Income' can be attributed to the variability of 'Height' – but how can that be quantified?

Explained variation (cont.)

- One measure of the variation in the values of Y is the Total Sum of Squares –

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

variability from the mean

- Writing

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2$$

$$= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + 2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}),$$

and, noting that (using the normal equations)

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = \sum_{i=1}^{n}(\underbrace{y_i - \widehat{y}_i}_{e_i})(\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \overline{y})$$

$$= (\widehat{\beta}_0 - \overline{y})\underbrace{\sum_{i=1}^{n} e_i}_{0} + \widehat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i e_i}_{0} = 0,$$

Explained variation (cont.)

we are left with

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{\text{TSS}}=\underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}_{\text{RSS}}+\underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}_{\text{ESS}},$$

where

- **ESS** is the *explained sum of squares*. It is easy to verify that $\overline{\widehat{y}}=\overline{y}$, hence it is a measure of the variability $\widehat{Y}$ – which is just a reflection of the variability of X.

proportion of y explained by x

- **RSS** is the good old Residual Sum of Squares – it is a measure of the variability of Y that the linear regression model cannot explain.

everything not explained by x

- We define the *proportion of explained variation* to be

$$R^2=\frac{\text{ESS}}{\text{TSS}}=\frac{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}.$$

used to examine goodness-of-fit for linear model, we look at R^2

Explained variation (cont.)

- The independent variable X explains the dependent variable Y well, if the RSS is small, or alternatively – if the ESS accounts for most of the TSS. A large (close to 1) R^2 value is thus an indication of a good linear fit.

Proposition

$$R^2 = r_{XY}^2$$

Proof:

$$R^2 = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = \frac{\sum_{i=1}^n (\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \overline{y})^2}{(n-1)S_Y^2} = \frac{\sum_{i=1}^n (\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i - \overline{y})^2}{(n-1)S_Y^2}$$

$$= \widehat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{(n-1)S_Y^2} = \widehat{\beta}_1^2 \frac{(n-1)S_X^2}{(n-1)S_Y^2} = \left(\frac{S_{XY}}{S_X^2}\right)^2 \cdot \frac{S_X^2}{S_Y^2} = \frac{S_{XY}^2}{S_X^2 S_Y^2} = r_{XY}^2 .$$

UNIVERSITY OF TORONTO

Explained variation (cont.)

- In the Income vs. Height example, we calculated $r_{XY} = 0.9135$, thus the proportion of the variation in income that is explained by height is

$$R^2 = r_{XY}^2 = (0.9135)^2 = 83.45\%.$$

- What about the remaining 16.55%?
  - Other influential factors that were omitted from the model (trifles like talent, number of hours at work etc...).
  - Misspecification of the model (it is probably more complicated than linear).
  - A degree of randomness that cannot be overcome (an "irreducible error")?

UNIVERSITY OF TORONTO

Adding the Normality assumption

- So far, we have made the following modelling assumptions:

  1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$
  2. $\mathbb{E}[\varepsilon_i] = 0, \quad i = 1, \ldots, n$
  3. $\text{Var}[\varepsilon_i] = \sigma^2, \quad i = 1, \ldots, n \text{ and } \mathbb{E}[\varepsilon_i \varepsilon_j] = 0 \text{ for } i \neq j$

- We are now adding a 4th assumption –

  4. The distribution of ε_i is normal, $i = 1, \ldots, n$

- The additional Normality assumption will allow us to do *statistical inference*: perform hypothesis testing, calculate confidence interval etc. Granted, it must be validated through careful diagnostics, as we shall later show.

addition of 4th assumption => errors are independent, not just uncorrelated

⋆ Because Normal random variables are uncorrelated $\iff$ they are independent, one implication of the 4^{th} assumption is that the random errors ε_i are not only uncorrelated (assumption 3), but independent.

UNIVERSITY OF TORONTO

Estimating the noise variance

- Recall that the residuals of the linear fit where of the form

$$e_i = y_i - \widehat{y}_i \quad , \quad i = 1, \ldots, n$$

- Simple, yet tedious calculations can show that

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

is, under the standard model assumptions, an unbiased estimator of the noise variance σ^2 (even without the Normality assumption).

- With the additional Normality assumption, it can be shown (but that requires some matrix algebra and knowledge of multivariate Normal distributions) that

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

The "loss" of 2 degrees of freedom can be attributed to the need of the residuals to satisfy the normal equations 2 constraints

$$\sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i\, e_i = 0.$$

UNIVERSITY OF TORONTO

Inference on the regression coefficients

- Remember that both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ were shown to be linear estimators, i.e., they are both of the form $\widehat{\beta} = \sum_i c_i y_i$.
- Now, with the additional Normality assumption, we know that

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2) \quad , \quad i = 1, \ldots, n$$

note x_i is not RV, so does not contribute to variance

- This means that each of the LS estimators is now a linear combination of independent (why?) Normals, and is thus Normal.
- For example, $\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_j (x_j - \overline{x})^2}\right)$.
- The (true) slope, β_1, is **the mean change in Y for an increase of one unit in X**. Also, if it is (close to) zero, that means that there is no linear association between X and Y. For these reasons it is important to make inferences about it.

so beta_0 is less interesting

- The intercept is generally of no interest other than prediction.

beta_1 is of more inferential value than beta_0

Hypothesis tests on the slope

- Under Normality,

$$\frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum_j (x_j - \overline{x})^2}}} \sim \mathcal{N}(0,1).$$

- As always, replacing the σ^2 with its unbiased estimator results in the t distribution, namely –

$$\frac{\widehat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{\sum_j (x_j - \overline{x})^2}}} \sim t_{n-2}.$$

Testing $\mathcal{H}_0 : \beta_1 = 0$ vs. $\mathcal{H}_1 : \beta_1 \neq 0$, then, will be based on the fact that

$$\mathcal{T} = \frac{\widehat{\beta}_1}{\frac{S}{\sqrt{\sum_j (x_j - \overline{x})^2}}} \overset{\mathcal{H}_0}{\sim} t_{n-2},$$

and a $100(1-\alpha)\%$ confidence interval for β_1 will be given by

$$\widehat{\beta}_1 \pm t_{n-2,1-\alpha/2} \frac{S}{\sqrt{\sum_j (x_j - \overline{x})^2}}$$

Example: Income vs. Height

	Height (X)	Income (Y)	Fitted ($\widehat{Y}$)	Residual (e)
1	70	2990	3050.5	−60.5
2	67	2870	2920.0	−50.0
3	69	2950	3007.0	−57.0
4	70	3140	3050.5	89.5
5	65	2790	2833.0	−43.0
6	73	3230	3181.0	49.0
7	64	2880	2789.5	90.5
8	70	3050	3050.5	−0.5
9	69	3000	3007.0	−7.0
10	73	3170	3181	−11.0
				$\sum e_i^2 = 30{,}030$

$$S^2 = \frac{1}{n-2}\sum e_i^2 = \frac{30030}{8} = 3753.75$$

(a standard deviation of \$61.27).

hypothesis testing on if there exists a correlation between income and height. equivalent to testing if beta_1 = 0
1. p-value from test statistic for beta_1
2. CI does not cover 0

UNIVERSITY OF TORONTO

Example: Income vs. Height (cont.)

$\widehat{\beta}_1 = 43.5$, $S_X^2 = 8.89$, $S^2 = 3753.75$, $\sum_j (x_j - \overline{x})^2 = (n-1)S_X^2 = 80$

calculate beta_1 from LS estimator S_XY / S_X^2

- To test $\mathcal{H}_0 : \beta_1 = 0$ vs. $\mathcal{H}_1 : \beta_1 \neq 0$, we calculate

$$\mathcal{T} = \frac{\widehat{\beta}_1}{\frac{S}{\sqrt{\sum_j (x_j - \overline{x})^2}}} = \frac{43.5}{\frac{\sqrt{3753.75}}{\sqrt{80}}} = 6.35,$$

6 std away from 0, very large

d.f. = 10 - 2 = 8

and

$$\text{p-value} = \mathbb{P}\left(|\mathcal{T}| \geq 6.35 \middle| \mathcal{H}_0\right) = 2\mathbb{P}(t_8 \geq 6.35) = 0.0002,$$

thus, at any reasonable significance level, we conclude that a linear association exists between height and income.

- A 95% confidence interval for the slope would be

$$\widehat{\beta}_1 \pm t_{n-2,1-\alpha/2} \frac{S}{\sqrt{\sum_j (x_j - \overline{x})^2}} = 43.5 \pm \underbrace{t_{8,0.975}}_{2.306} \frac{\sqrt{3753.75}}{\sqrt{80}}$$

$$= 43.5 \pm 15.8 = [\$27.7, \$59.3].$$

does not include 0, correspond to hypothesis testing result

UNIVERSITY OF TORONTO

Confidence interval for the mean response

- Recall that the model is

$$y(x) = \underbrace{\beta_0 + \beta_1 x}_{\textit{mean response}} + \underbrace{\varepsilon(x)}_{\text{random noise}} .$$

- Our next goal is to estimate the mean response,

$$\mu(x_0) := \mathbb{E}[y(x_0)] = \beta_0 + \beta_1 x_0 ,$$

since expected value of error = 0

at a new site x_0 (that was not part of the original sample), as well as to find a $100(1-\alpha)\%$ confidence interval.

- An obvious choice of a point estimator would be the *prediction* at x_0,

$$\widehat{y}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 .$$

It is unbiased, since since its a linear combination of unbiased estimators

$$\mathbb{E}[\widehat{y}(x_0)] = \mathbb{E}[\widehat{\beta}_0 + \widehat{\beta}_1 x_0] = \mathbb{E}[\widehat{\beta}_0] + \mathbb{E}[\widehat{\beta}_1] x_0 = \beta_0 + \beta_1 x_0 = \mu(x_0).$$

CI for the mean response (cont.)

- In addition,

bi-linearity of covariance; takes x_0 out

$$\begin{aligned}\mathrm{Var}[\widehat{y}(x_0)] &= \mathrm{Var}[\widehat{\beta}_0 + \widehat{\beta}_1 x_0] = \mathrm{Var}[\widehat{\beta}_0] + x_0^2 \mathrm{Var}[\widehat{\beta}_1] + 2x_0 \,\mathrm{Cov}\big(\widehat{\beta}_0, \widehat{\beta}_1\big) \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{\overline{x}^2}{\sum_j (x_j - \overline{x})^2} \right\} + \frac{\sigma^2 x_0^2}{\sum_j (x_i - \overline{x})^2} - \frac{2\sigma^2 x_0 \overline{x}}{\sum_j (x_j - \overline{x})^2} \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_j (x_j - \overline{x})^2} \right\}.\end{aligned}$$

- Furthermore, since the LS estimators are both linear estimators, we have

$$\widehat{y}(x_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 = \sum_i b_i y_i + x_0 \sum_i a_i y_i = \sum_i (b_i + x_0 a_i) y_i,$$

  hence $\widehat{y}(x_0)$ is Normal (as a linear combination of independent Normals).

- The bottom line is that

$$\widehat{y}(x_0) \sim \mathcal{N}\left(\mu(x_0), \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_j (x_j - \overline{x})^2} \right\} \right).$$

UNIVERSITY OF TORONTO

CI for the mean response (cont.)

- Replacing σ with S^2 would again give rise to the t distribution, and so

$$\widehat{y}(x_0) \pm t_{n-2,1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_j (x_j - \overline{x})^2}}$$

is a $100(1-\alpha)\%$ confidence interval for the mean response $\mathbb{E}[y(x_0)]$.

- When we let x_0 vary, we obtain *confidence bands* centered at the least squares fit, that get narrower as x_0 draws closer to $\overline{x}$.

more points in center: higher confidence
narrower band

- In the Income vs. Height example we had $n = 10$, $\widehat{\beta}_0 = 5.5$, $\widehat{\beta}_1 = 43.5$ $\overline{x} = 69$, $S^2 = 3753.75$ and $\sum_j (x_j - \overline{x})^2 = 80$.

- A point estimate of the mean monthly income of a 72" tall person will be

$$\widehat{y}(72) = 5.5 + 43.5 \times 72 = \$3137.5.$$

(point estimation) = beta_0 + beta_1 x

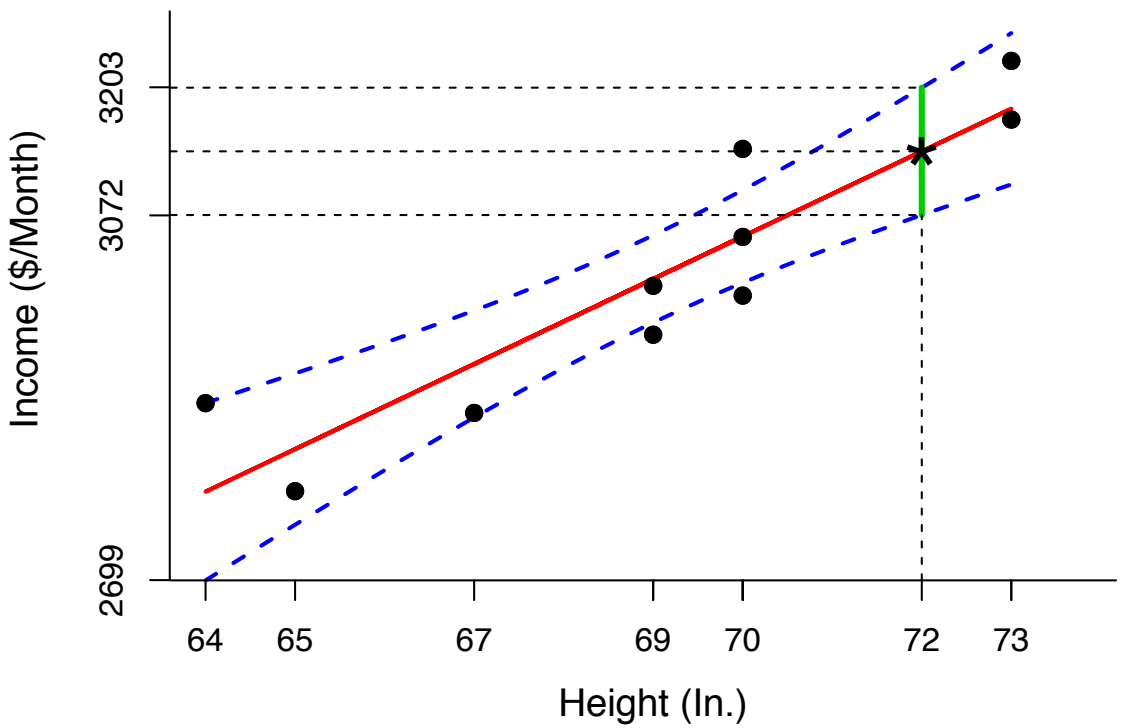
CI for the mean response (cont.)

finding confidence interval when x_0 = 72 and alpha=0.95

- A 95% confidence interval for that person's monthly income will then be

$$3137.5 \pm \underbrace{t_{8,0.975}}_{2.306} \times \sqrt{3753.75}\sqrt{\frac{1}{10} + \frac{(72-69)^2}{80}} = 3137.5 \pm 65.13$$

$$= [\$3072.37, \$3202.63].$$



The Least Squares estimators as MLEs

Proposition

Under the additional assumption that the random noise is Gaussian, the least squares estimators $\widehat{\beta}_0^{\text{LS}}$ and $\widehat{\beta}_1^{\text{LS}}$ are the maximum likelihood estimators of β_0 and β_1, respectively.

Proof: LS estimators happens to be MLE if noise is Gaussian

Since $y_i \overset{\text{ind.}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \ldots, n$, we have

$$\left(\widehat{\beta}_0^{\text{MLE}}, \widehat{\beta}_1^{\text{MLE}}\right) = \arg\max_{\beta_0, \beta_1} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$= \arg\min_{\beta_0, \beta_1} \sum_{i=1}(y_i - \beta_0 - \beta_1 x_i)^2 = \left(\widehat{\beta}_0^{\text{LS}}, \widehat{\beta}_1^{\text{LS}}\right).$$

UNIVERSITY OF TORONTO

Diagnostic plots

- Any inference drawn from the linear regression model, relies heavily on the assumptions made about the random errors, $\{\varepsilon_i\}$. It is not surprising then, that validating these assumptions involves their "proxy", the residuals $\{e_i\}$. Failure to show evidence in support of the basic assumptions, deems any such inference invalid.

- Under the Normality assumption, the residuals (like the unobserved errors) are draws from a Normal distribution. They are in fact correlated, but if the sample size is reasonably large, their correlation becomes negligible.

error has to be normal, independent and 0 mean

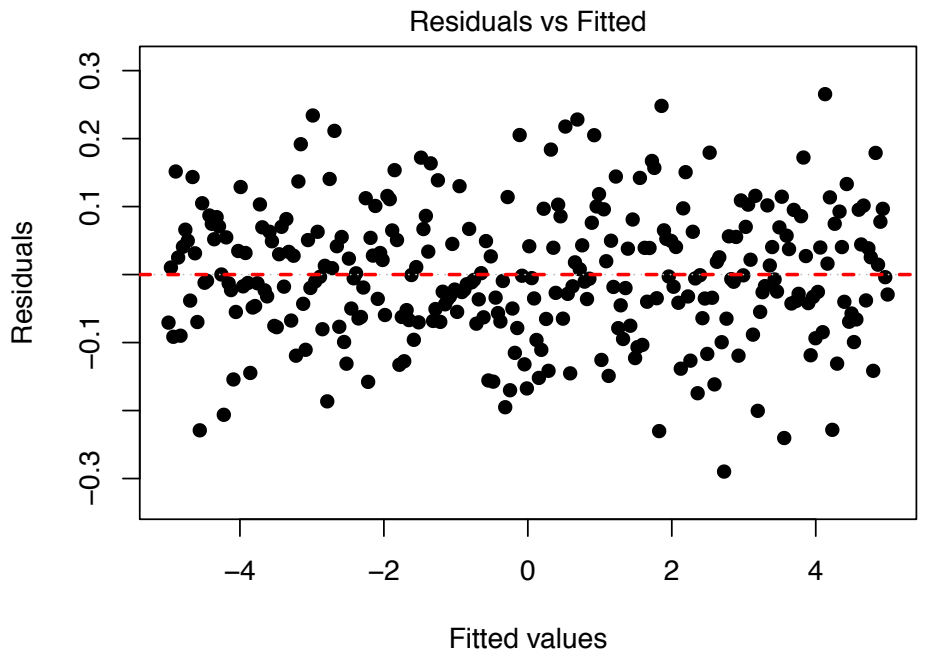
- Moreover, the residuals inherit the homoscedasticity (homogeneity of the variance) property of the ε_i's.

- The distribution of the *standardized residuals* $\dfrac{e_i}{S}$ should be nearly $\mathcal{N}(0,1)$.

- These principles are the basis for many *diagnostic plots*. We shall focus on two simple and highly informative plots: the Residuals vs. Fitted values and Quantile–Quantile plots.

The Residuals vs. Fitted values plot

- Simply plotting the e_i's vs. the $\widehat{y}_i$'s.
- This is what the ideal plot looks like –



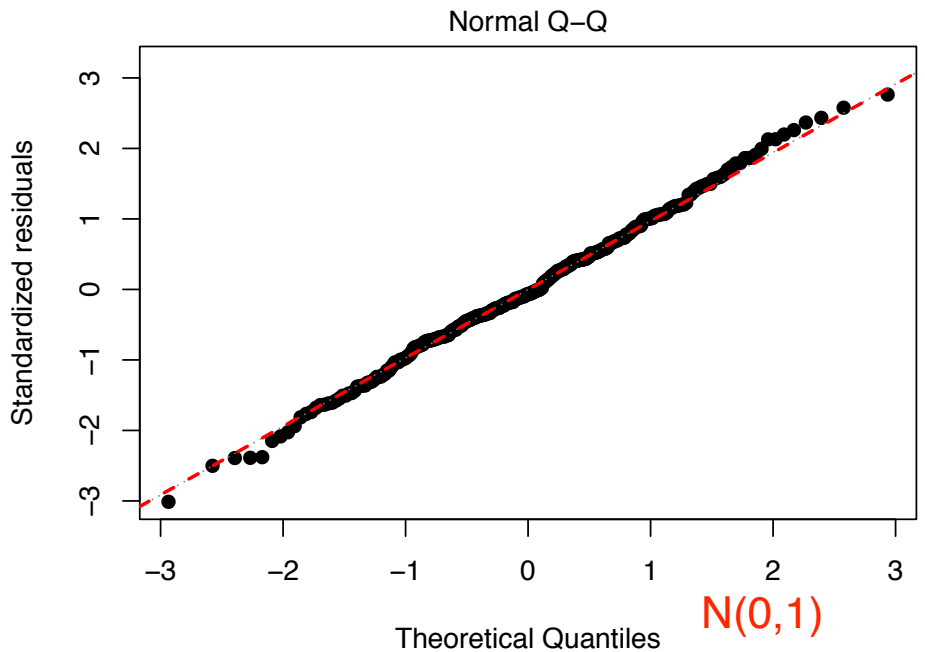
Symmetry about 0, homogeneity of the variance, no trends or pattern – typical white noise.

UNIVERSITY OF TORONTO

The Quantile–Quantile plot

- Plotting (some variant of) the sample quantiles (percentiles) of the standardized residuals vs. the theoretical quantiles of the standard Normal distribution.

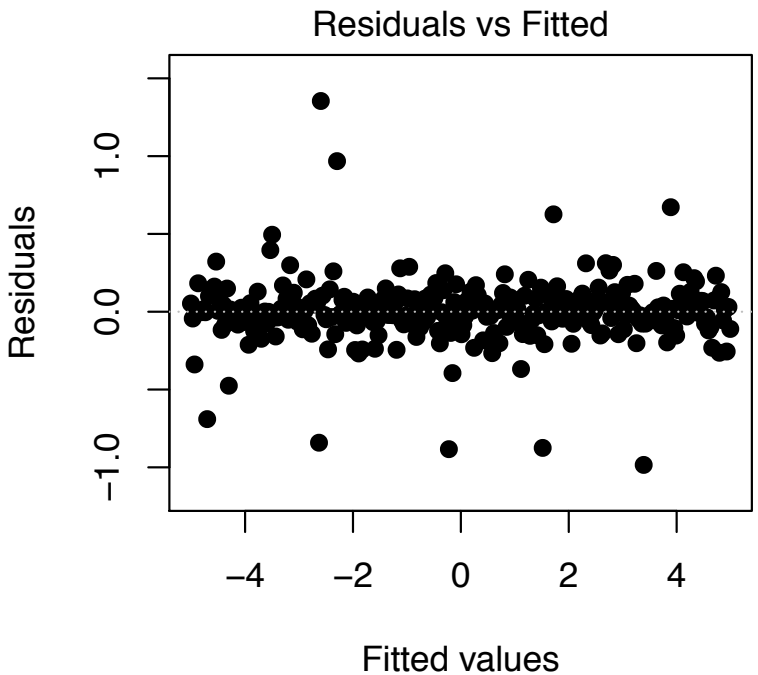
- This is what the ideal plot looks like –



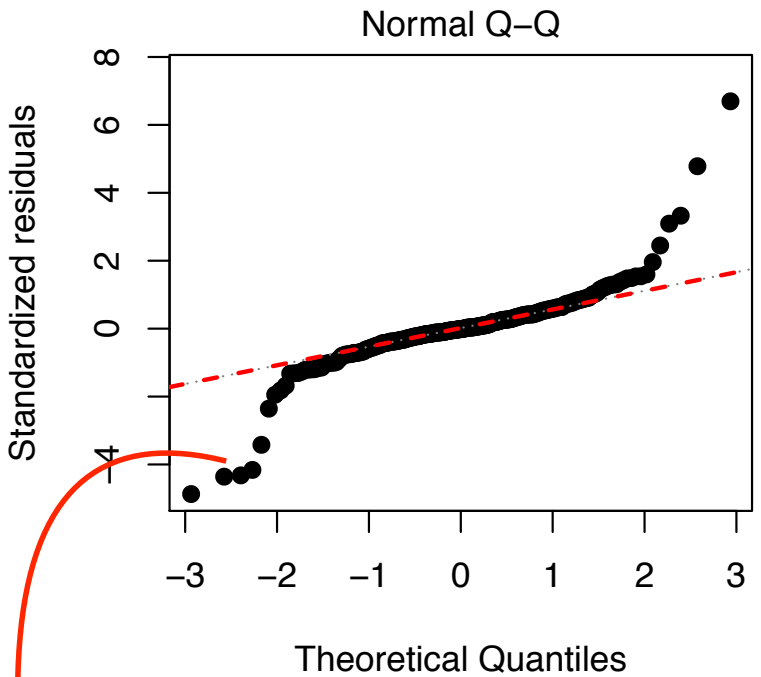
the points in the scatterplot align along the $y = x$ line, implying (nearly) perfect match.

Violation of the model assumptions: examples

looks fine here



normality assumption violated:

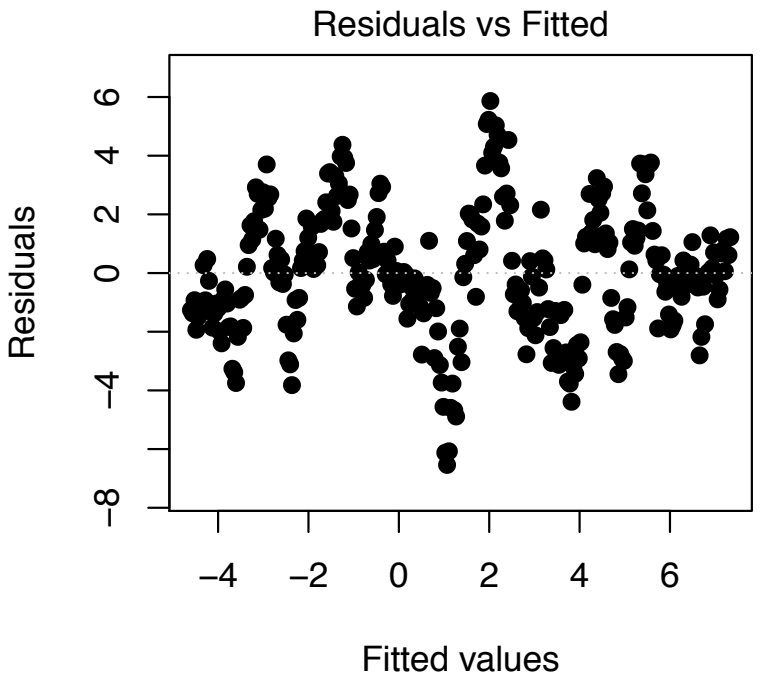


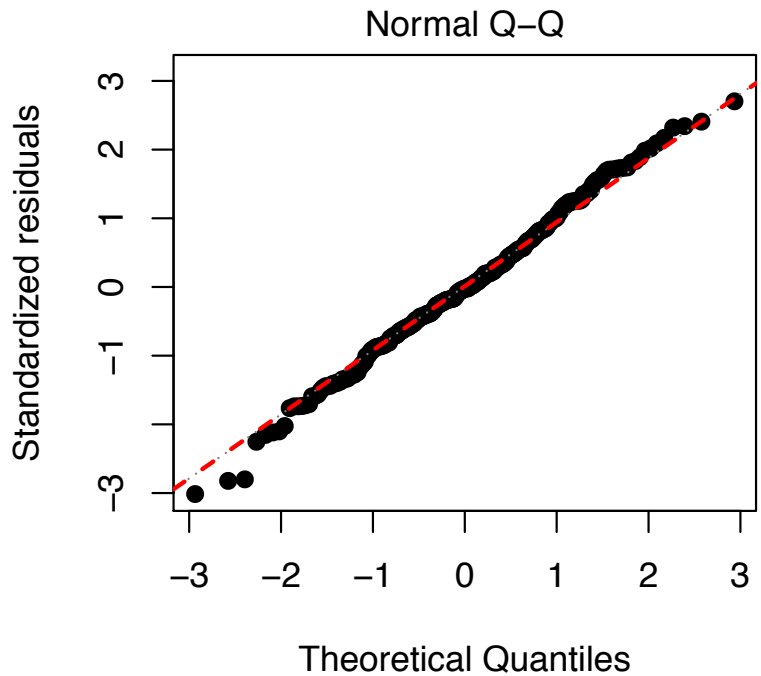
seems to be t-distr: tail heavy.

same quantile, observed standardized residual is smaller here.

- Lower quantiles are too small and upper quantiles too large – a "heavy-tailed" noise.

Violation of the assumptions: examples (cont.)





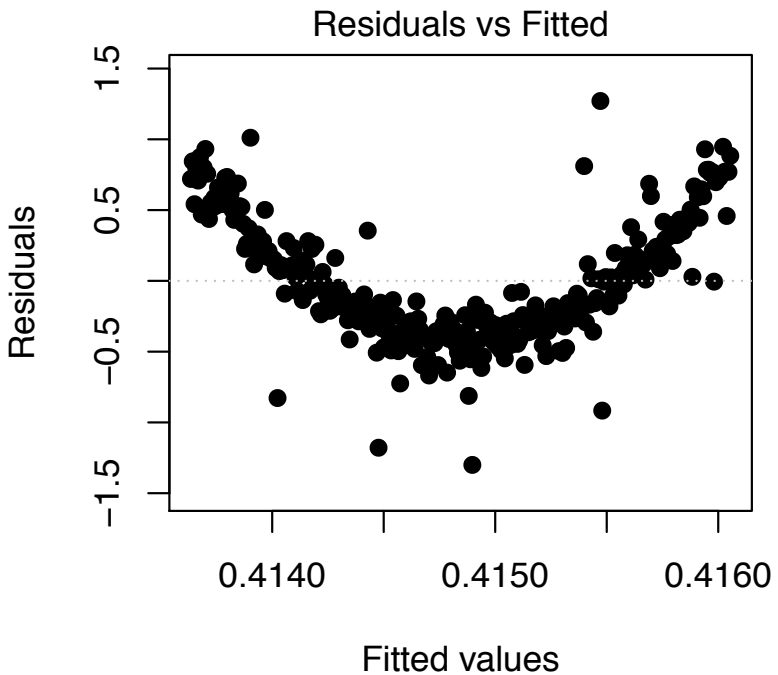
temp change is usual graduate, so observation is correlated

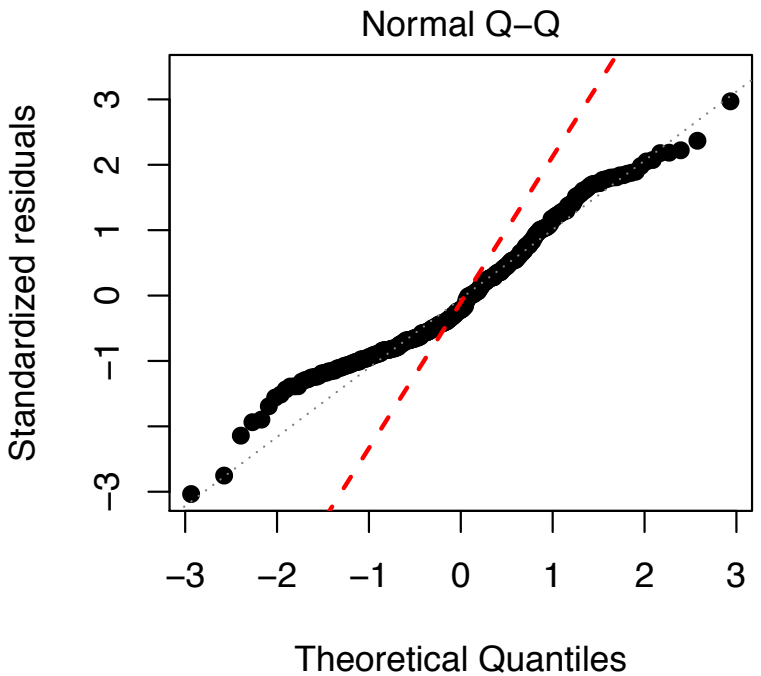
- Streaks of positive/negative residuals – *autocorrelated* noise.

- Think daily maximum temperature vs. day.
  usually when dependent variable is time

Diagnostic plots (cont.)

assumption of linearity is off: probably
fit 2nd order polynomial instead

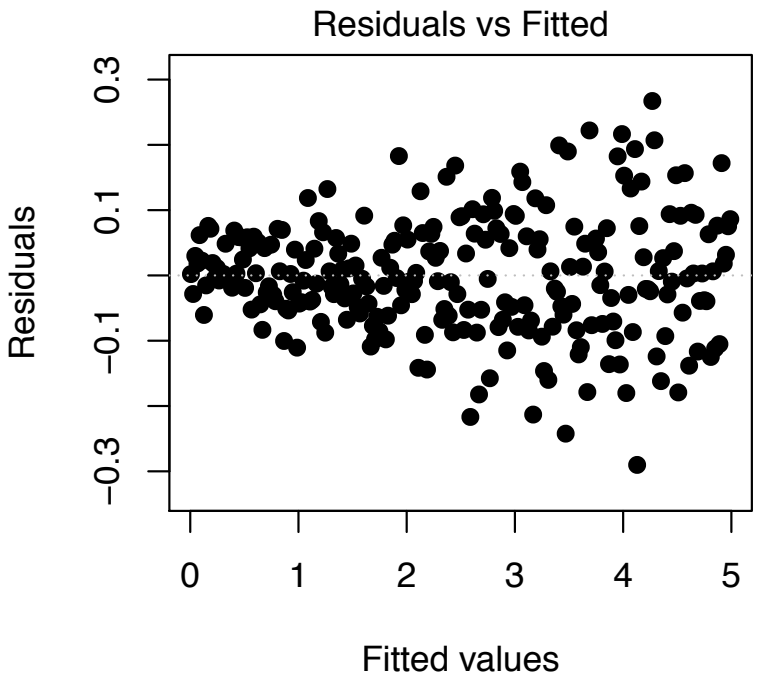


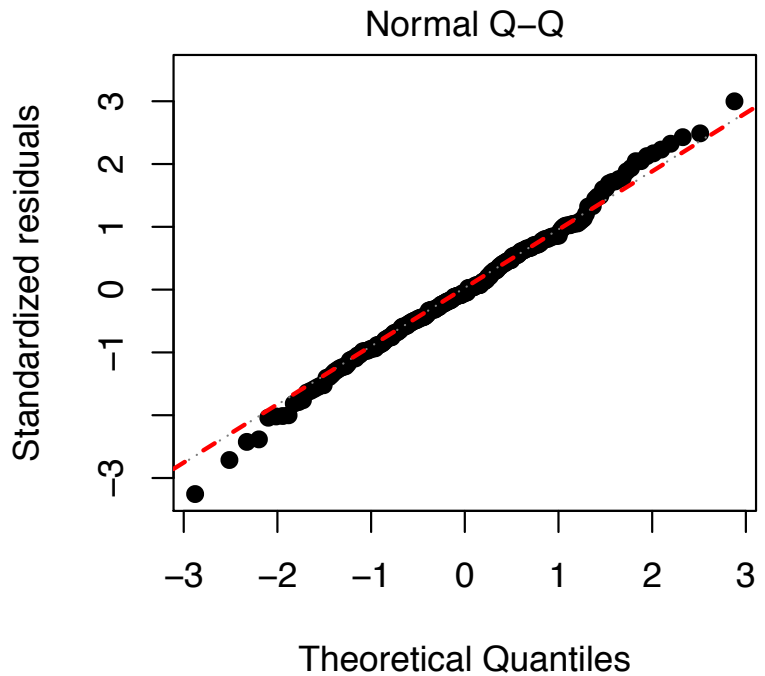


- Model misspecification – should be quadratic, not linear.

Diagnostic plots (cont.)

variance not homogeneous..





- Variance increases along with the predicted values – no homoscedasticity.