7.2 Linear Correlation and Regression

POPULATION

Random Variables X, Y: ***numerical***

Definition: **Population Linear Correlation Coefficient of *X*, *Y***

$$\rho = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y}$$

FACT: $-1 \le \rho \le +1$

$\downarrow$

SAMPLE, size *n*

Definition: **Sample Linear Correlation Coefficient of *X*, *Y***

$$\hat{\rho} = \boldsymbol{r} = \frac{s_{xy}}{s_x \, s_y}$$

Example: $\boldsymbol{r} = \dfrac{600}{\sqrt{250}\,\sqrt{1750}} = \mathbf{0.907}$ strong, positive linear correlation

FACT:

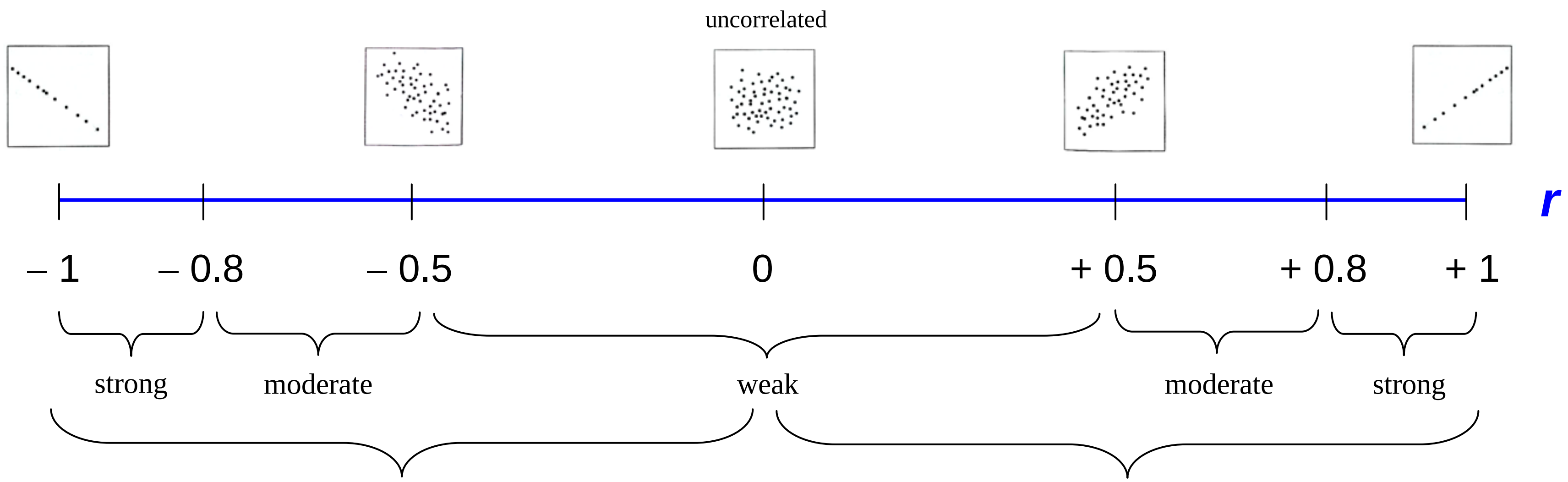
$$\boxed{-1 \le r \le +1}$$

Any set of data points (x_i, y_i), $i = 1, 2, \ldots, n$, having $r > 0$ (likewise, $r < 0$) is said to have a **positive linear correlation** (likewise, **negative linear correlation**). The linear correlation can be **strong**, **moderate**, or **weak,** depending on the magnitude. The closer r is to $+1$ (likewise, -1), the more strongly the points follow a straight line having *some positive* (likewise, *negative*) *slope*. The closer r is to 0, the weaker the linear correlation; if $r = 0$, then EITHER the points are *uncorrelated* (see 7.1), OR they are correlated, but *nonlinearly* (e.g., $Y = X^2$).

Exercise: Draw a scatterplot of the following $n = 7$ data points, and compute r.

$$(-3, 9),\ (-2, 4),\ (-1, 1),\ (0, 0),\ (1, 1),\ (2, 4),\ (3, 9)$$

(Pearson's) Sample Linear Correlation Coefficient $r = \dfrac{s_{xy}}{s_x\, s_y}$



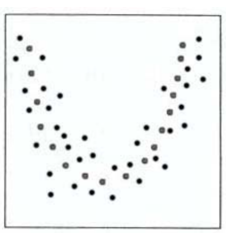
negative linear correlation

As X increases, Y decreases.
As X decreases, Y increases.

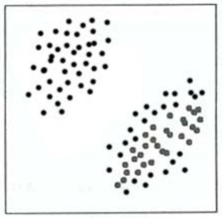
positive linear correlation

As X increases, Y increases.
As X decreases, Y decreases.

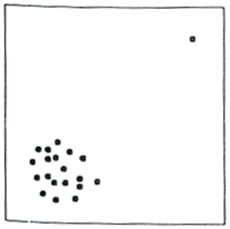
➢ Some important exceptions to the "typical" cases above:



$r = 0$, but X and Y are correlated, *non*linearly



$r > 0$ in each of the two individual subgroups, but $r < 0$ when combined



$r > 0$, only due to the effect of one *influential* outlier; if removed, then data are uncorrelated ($r = 0$)

Statistical Inference for ρ

Suppose we now wish to conduct a formal test of…

Hypothesis H_0: $\rho = 0$ $\Leftrightarrow$ **"There is <u>no</u> linear correlation between X and Y."**
vs.
Alternative Hyp. H_A: $\rho \neq 0$ $\Leftrightarrow$ **"There is a linear correlation between X and Y."**

Test Statistic

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

<u>Example</u>: ***p*-value** $= 2\,P\left(T_3 \geq \frac{.907\sqrt{3}}{\sqrt{1-(.907)^2}}\right) = 2\,P(T_3 \geq \mathbf{3.733}) = 2(.017) = \mathbf{.034}$

As $p < \alpha = .05$, the null hypothesis of no linear correlation can be rejected at this level.

<u>Comments</u>:

- Defining the numerator "sums of squares" $S_{xx} = (n-1)\,s_x^2$, $S_{yy} = (n-1)\,s_y^2$, and $S_{xy} = (n-1)\,s_{xy}$, the correlation coefficient can also be written as $r = \dfrac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$.

- The general null hypothesis H_0: $\rho = \rho_0$ requires a more complicated Z-test, which first applies the so-called **Fisher transformation**, and will not be presented here.

- The assumption on X and Y is that their *joint distribution* is **bivariate normal**, which is difficult to check fully in practice. However, a consequence of this assumption is that X and Y are linearly uncorrelated (i.e., $\rho = 0$) *if and only if* X and Y are independent. That is, it overlooks the possibility that X and Y might have a <u>nonlinear</u> correlation. The moral: ρ – and therefore the **Pearson sample linear correlation coefficient** r calculated above – only captures the strength of <u>linear</u> correlation. A more sophisticated measure, the **multiple correlation coefficient**, can detect nonlinear correlation, or correlation in several variables. Also, the *nonparametric* **Spearman rank-correlation coefficient** can be used as a substitute.

- <u>Correlation does not imply causation!</u> (E.g., X = "children's foot size" is indeed positively correlated with Y = "IQ score," but is this really cause-and-effect????) The ideal way to establish causality is via a well-designed randomized clinical trial, but this is not always possible, or even desirable. (E.g., X = smoking vs. Y = lung cancer)

Simple Linear Regression and the Method of Least Squares

$\boldsymbol{k = 2}$ parameters,
"regression coefficients"

Predictor Variable,
Explanatory Variable

$$Y = \underbrace{\beta_0 + \beta_1 X}\ + \ \varepsilon$$

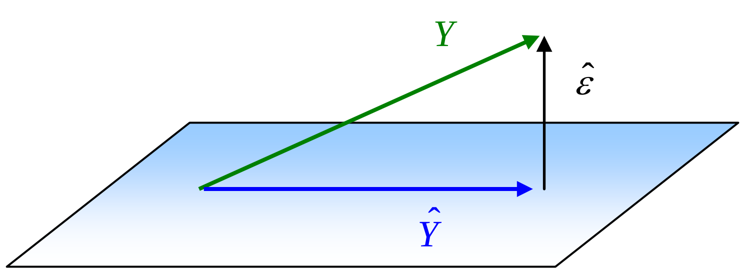
"Response = (Linear) Model + Error"

If a linear association exists between variables X and Y, then it can be written as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

intercept = b_0 b_1 = slope

Sample-based estimator of response

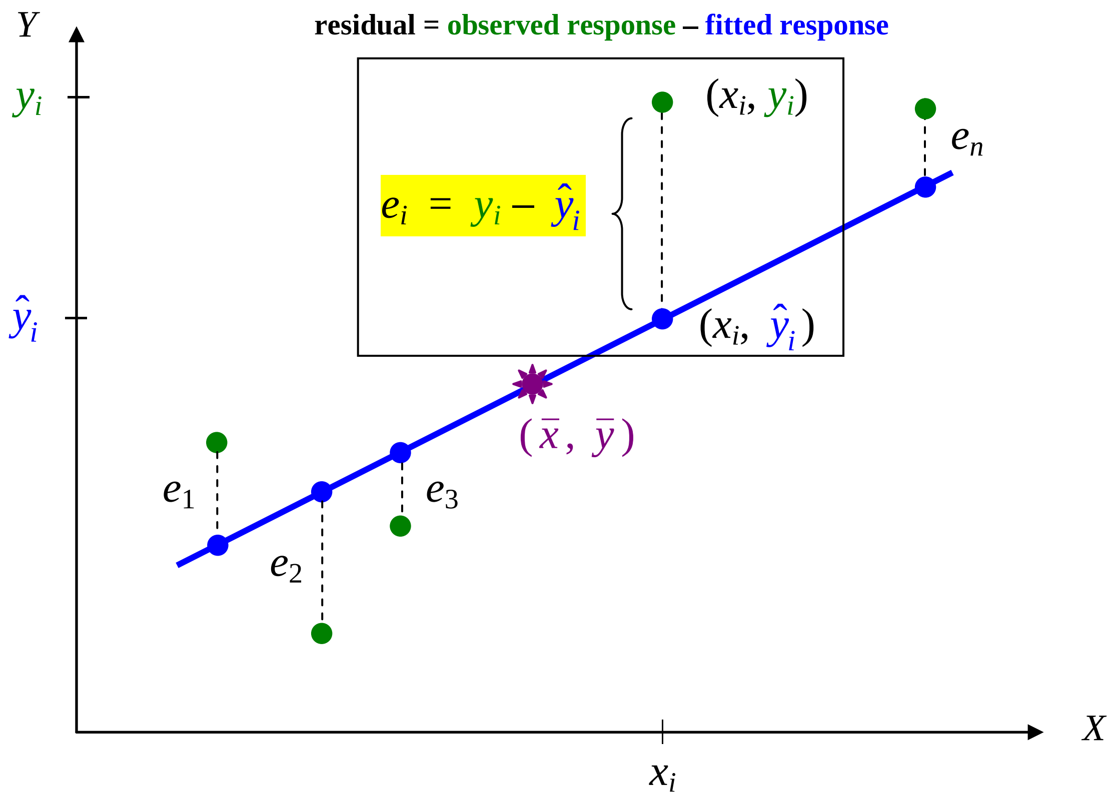


That is, given the "response vector" Y, we wish to find the linear estimate $\hat{Y}$ that makes the magnitude of the difference $\hat{\varepsilon} = Y - \hat{Y}$ <u>as small as possible</u>.

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \Rightarrow \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

How should we define the line that "best" fits the data, and obtain its coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$?

For *any* line, **errors** ε_i, $i = 1, 2, \ldots, n$, can be estimated by the **residuals** $\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$.



The **least squares regression line** is the *unique* line that minimizes the **Error** (or **Residual**) **Sum of Squares** $\mathbf{SS_{Error}} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

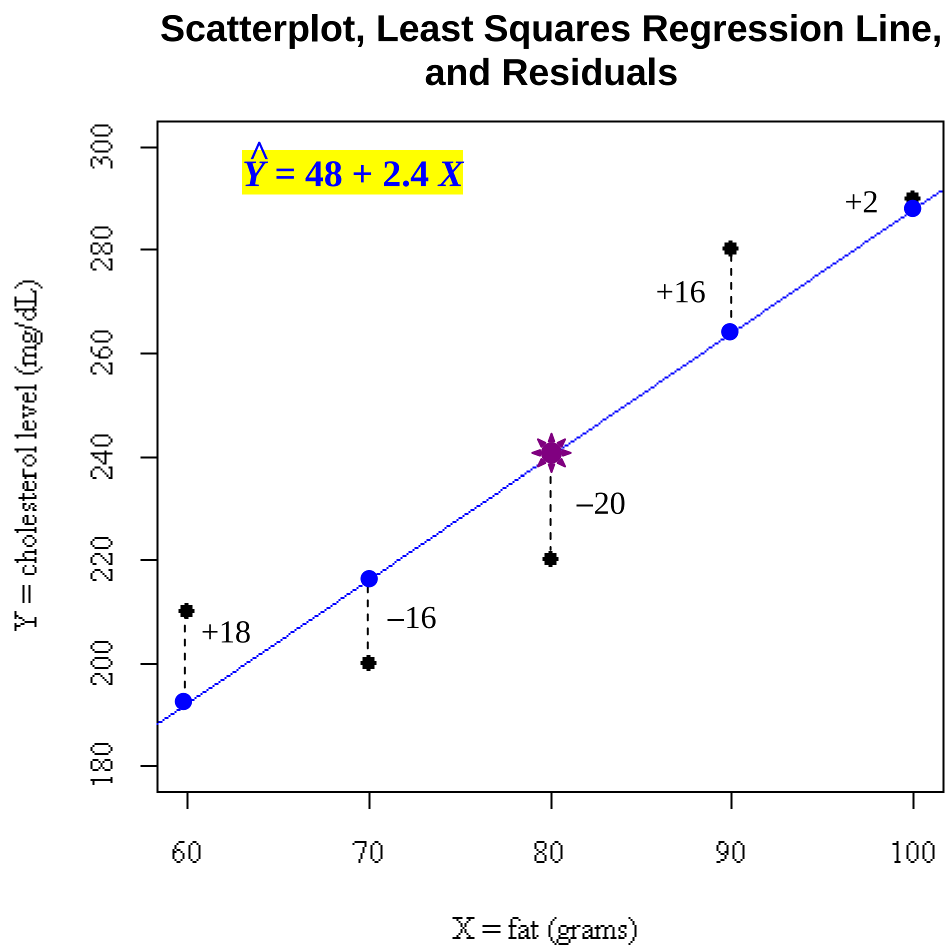
Slope: $\hat{\beta}_1 = b_1 = \dfrac{s_{xy}}{s_x^{\,2}}$

Intercept: $\hat{\beta}_0 = b_0 = \bar{y} - b_1\,\bar{x}$

$$\hat{\mathbf{Y}} = \mathbf{b_0} + \mathbf{b_1}\,\mathbf{X}$$

Example (cont'd): Slope $b_1 = \dfrac{600}{250} = \mathbf{2.4}$ Intercept $b_0 = 240 - (\mathbf{2.4})(80) = \mathbf{48}$

Therefore, the **least squares regression line** is given by the equation $\hat{\mathbf{Y}} = \mathbf{48 + 2.4\,X}$.



predictor values	x_i	60	70	80	90	100
observed responses	y_i	**210**	**200**	**220**	**280**	**290**
fitted responses, predicted responses	$\hat{y}_i$	**192**	**216**	**240**	**264**	**288**
residuals	$e_i = y_i - \hat{y}_i$	+18	–16	–20	+16	+2

Note that the sum of the residuals is equal to zero. But the sum of their squares,

$$\left\| \hat{\varepsilon} \right\|^2 = SS_{\text{Error}} = (+18)^2 + (-16)^2 + (-20)^2 + (+16)^2 + (+2)^2 = 1240$$

is, by construction, the <u>smallest</u> such value of all possible regression lines that could have been used to estimate the data. Note also that the **center of mass** (80, 240) lies on the least squares regression line.

<u>Example</u>: The population cholesterol level corresponding to x^* = 75 fat grams is estimated by $\hat{y} = 48 + 2.4(75)$ = 228 mg/dL. But how *precise* is this value? (Later...)

Statistical Inference for β_0 and β_1

It is possible to test for significance of the intercept parameter β_0 and slope parameter β_1 of the least squares regression line, using the following:

$(1 - \alpha) \times 100\%$ Confidence Limits

For β_0: $$b_0 \pm t_{n-2,\,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

For β_1: $$b_1 \pm t_{n-2,\,\alpha/2} \cdot s_e \frac{1}{\sqrt{S_{xx}}}$$

Test Statistic

For β_0: $$T = \left(\frac{b_0 - \beta_0}{s_e}\right)\sqrt{\frac{n\,S_{xx}}{S_{xx} + n\,(\bar{x})^2}} \sim t_{n-2}$$

For β_1: $$T = \left(\frac{b_1 - \beta_1}{s_e}\right)\sqrt{S_{xx}} \sim t_{n-2}$$

where $s_e^{\,2} = \dfrac{\text{SS}_{\text{Error}}}{n-2}$ is the so-called **standard error of estimate**, and $S_{xx} = (n-1)\,s_x^{\,2}$. (Note: $s_e^{\,2}$ is also written as MSE or MS_{Error}, the "mean square error" of the regression; see ANOVA below.)

Example: Calculate the p-value of the slope parameter β_1, under…

Null Hypothesis H_0: $\beta_1 = 0$ $\Leftrightarrow$ **"There is no linear association between X and Y."**

vs.

Alternative Hyp. H_A: $\beta_1 \neq 0$ $\Leftrightarrow$ **"There is a linear association between X and Y."**

First, $s_e^{\,2} = \dfrac{1240}{3} = 413.333$, so $s_e = 20.331$. And $S_{xx} = (4)(250) = 1000$. So…

p-value $= 2\,P\left(T_3 \geq \left(\dfrac{2.4 - 0}{20.331}\right)\sqrt{1000}\right) = 2\,P(T_3 \geq \mathbf{3.733}) = 2\,(.017) = \mathbf{.034}$

As $p < \alpha = .05$, the null hypothesis of no linear association can be rejected at this level.

Note that the T-statistic (**3.733**), and hence the resulting p-value (**.034**), is *identical* to the test of significance of the linear correlation coefficient H_0: $\rho = 0$ conducted above!

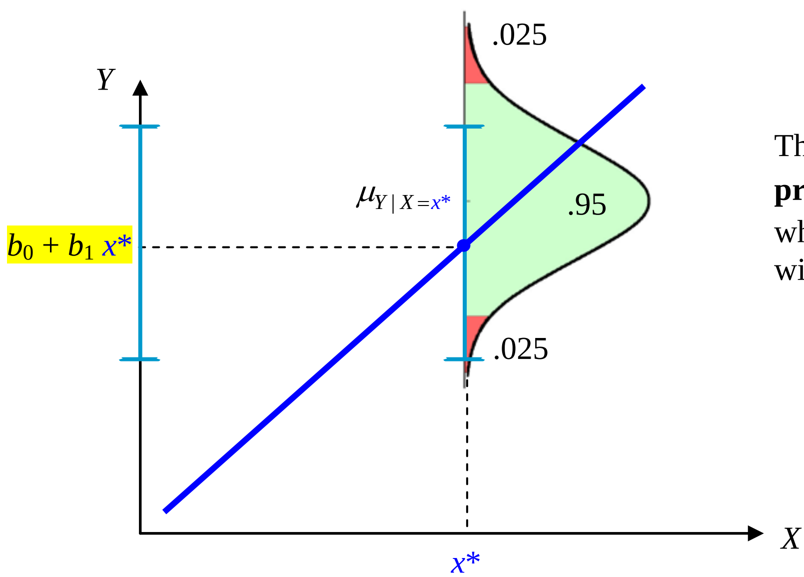
Exercise: Calculate the 95% confidence interval for β_1, and use it to test H_0: $\beta_1 = 0$.

Confidence and Prediction Intervals

Recall that, from the discussion in the previous section, a regression problem such as this may be viewed in the formal context of starting with n normally-distributed populations, each having a conditional mean $\mu_{Y\,|\,X=x_i}$, i = 1, 2, ..., n. From this, we then obtain a linear model that allows us to derive an estimate of the response variable via $\hat{Y} = b_0 + b_1 X$, for *any* value $X = x^*$ (with certain restrictions to be discussed later), i.e., $\hat{y} = b_0 + b_1 x^*$. There are two standard possible interpretations for this fitted value. First, $\hat{y}$ can be regarded simply as a "predicted value" of the response variable Y, for a randomly selected individual from the specific normally-distributed population corresponding to $X = x^*$, and can be improved via a so-called **prediction interval**.

(1 − α) × 100% Prediction Limits for *Y* at *X* = *x**

$$(b_0 + b_1\, x^*) \;\pm\; t_{n-2,\,\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$



This diagram illustrates the associated 95% **prediction interval** around $\hat{y}$ = $b_0 + b_1\, x^*$, which contains the true *response value* Y with 95% probability.

Exercise: Confirm that the 95% prediction interval for $\hat{y}$ = 228 (when x^* = 75) is (156.3977, 299.6023).

Example (α = .05):

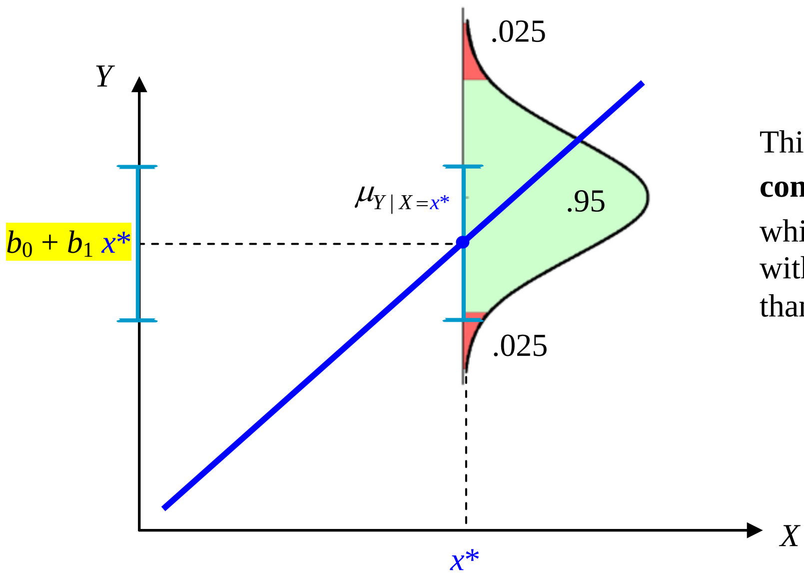
95% Prediction Bounds

X	fit	Lower	Upper
60	**192**	110.1589	273.8411
70	**216**	142.2294	289.7706
80	**240**	169.1235	310.8765
90	**264**	190.2294	337.7706
100	**288**	206.1589	369.8411

The second interpretation is that $\hat{y}$ can be regarded as a point estimate of the conditional mean $\mu_{Y\,|\,X=x^*}$ of this population, and can be improved via a **confidence interval**.

(1 − α) × 100% Confidence Limits for $\mu_{Y\,|\,X\,=\,x^*}$

$$(b_0 + b_1\,x^*) \;\pm\; t_{n-2,\,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$



This diagram illustrates the associated 95% **confidence interval** around $\hat{y}$ = $b_0 + b_1\,x^*$, which contains the true *conditional mean* $\mu_{Y\,|\,X=x^*}$ with 95% probability. Note that it is narrower than the corresponding prediction interval above.

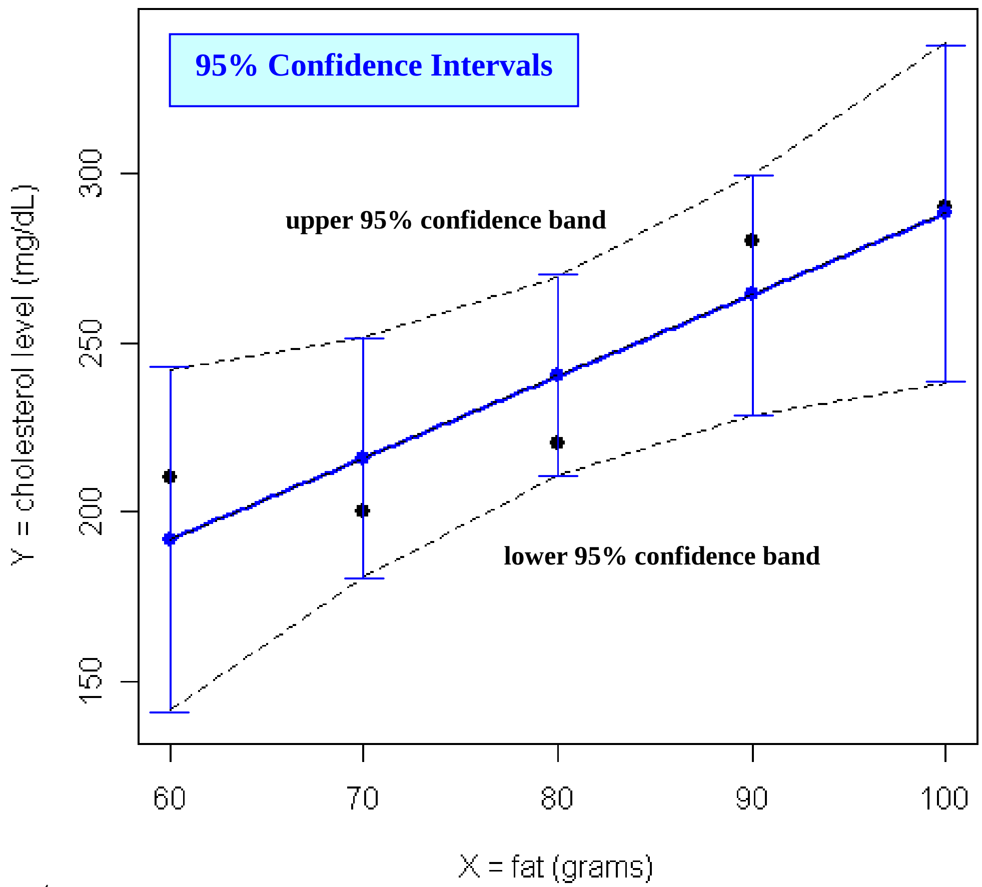
Exercise: Confirm that the 95% confidence interval for $\hat{y}$ = 228 (when x^* = 75) is (197.2133, 258.6867).

Note: Both approaches are based on the fact that there is, in principle, variability in the coefficients b_0 and b_1 themselves, from one sample of n data points to another. Thus, for *fixed* x^*, the object $\hat{y} = b_0 + b_1\,x^*$ can actually be treated as a random variable in its own right, with a computable sampling distribution.

Also, we define the *general* **conditional mean** $\mu_{Y\,|\,X}$ – i.e., **conditional expectation** $E[Y\,|\,X]$ – as $\mu_{Y\,|\,X=x^*}$ – i.e., $E[Y\,|\,X = x^*]$ – for *all* appropriate x^*, rather than a specific one.

Example ($\alpha = .05$): **95% Confidence Bounds**

X	fit	Lower	Upper
60	**192**	141.8827	242.1173
70	**216**	180.5617	251.4383
80	**240**	211.0648	268.9352
90	**264**	228.5617	299.4383
100	**288**	237.8827	338.1173



Comments:

- Note that, because individual responses have greater variability than mean responses (recall the Central Limit Theorem, for example), we expect prediction intervals to be wider than the corresponding confidence intervals, and indeed, this is the case. The two formulas differ by a term of "1 +" in the standard error of the former, resulting in a larger margin of error.

- Note also from the formulas that both types of interval are narrowest when $x^* = \bar{x}$, and grow steadily wider as x^* moves farther away from $\bar{x}$. (This is evident in the graph of the 95% confidence intervals above.) *Great care should be taken if x^* is outside the domain of sample values!* For example, when fat grams $x = 0$, the linear model predicts an unrealistic cholesterol level of $\hat{y} = 48$, and the margin of error is uselessly large. The linear model is not a good predictor there.

ANOVA Formulation

As with comparison of multiple treatment means (§6.3.3), regression can also be interpreted in the general context of **analysis of variance**. That is, because

Response = Model + Error,

it follows that the total variation in the original response data can be **partitioned** into a source of variation due to the model, plus a source of variation for whatever remains. We now calculate the three **"Sums of Squares (SS)"** that measure the variation of the system and its two component sources, and their associated **degrees of freedom (df)**.

1. **Total Sum of Squares** = sum of the squared deviations of *each observed response value* y_i from the *mean response value* $\bar{y}$.

$$\mathbf{SS_{Total}} = (210 - 240)^2 + (200 - 240)^2 + (220 - 240)^2 + (280 - 240)^2 + (290 - 240)^2 = \mathbf{7000}$$

$\mathbf{df_{Total}} = 5 - 1 = \mathbf{4}$ *Reason*: n data values – 1

Note that, by definition, $s_y^{\,2} = \dfrac{\text{SS}_{\text{Total}}}{\text{df}_{\text{Total}}} = \dfrac{7000}{4} = 1750$, as given in the beginning of this example in 7.1.

2. **Regression Sum of Squares** = sum of the squared deviations of *each fitted response value* $\hat{y}_i$ from the *mean response value* $\bar{y}$.

$$\mathbf{SS_{Reg}} = (192 - 240)^2 + (216 - 240)^2 + (240 - 240)^2 + (264 - 240)^2 + (288 - 240)^2 = \mathbf{5760}$$

$\mathbf{df_{Reg} = 1}$ *Reason*: As the regression model is linear, its degrees of freedom = one less than the $k = 2$ parameters we are trying to estimate (β_0 and β_1).

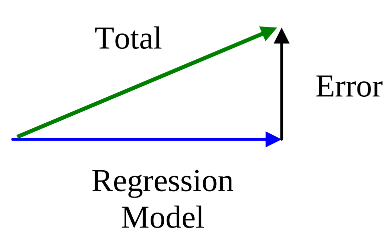
3. **Error Sum of Squares** = sum of the squared deviations of *each observed response* y_i from its corresponding *fitted response* $\hat{y}_i$ (i.e., the sum of the squared **residuals**).

$$\mathbf{SS_{Error}} = (210 - 192)^2 + (200 - 216)^2 + (220 - 240)^2 + (280 - 264)^2 + (290 - 288)^2 = \mathbf{1240}$$

$\mathbf{df_{Error}} = 5 - 2 = \mathbf{3}$ *Reason*: n data values – k regression parameters in model

$$\mathbf{SS_{Total} = SS_{Reg} + SS_{Error}}$$

$$\mathbf{df_{Total} = df_{Reg} + df_{Error}}$$



ANOVA Table

Source	df	SS "Sum of Squares"	"Mean Squares" $MS = \frac{SS}{df}$	Test Statistic ($F_{1,3}$ distribution) $F = \frac{MS_{Reg}}{MS_{Err}}$	p-value
Regression	1	5760	5760	13.94	.034
Error	3	1240	413.333		
Total	4	7000	–		

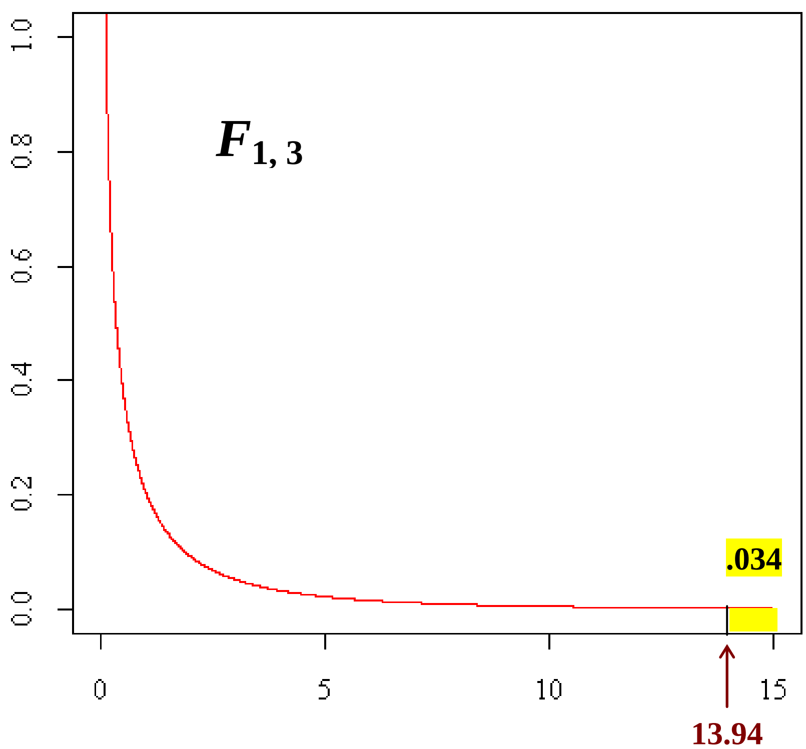
According to this F-test, we can reject…

Null Hypothesis $H_0: \beta_1 = 0$ $\Leftrightarrow$ **"There is no linear association between X and Y."**

vs.

Alternative Hyp. $H_A: \beta_1 \neq 0$ $\Leftrightarrow$ **"There is a linear association between X and Y."**

at the $\alpha = .05$ significance level, which is consistent with our earlier findings.



Comment: Again, note that **13.94 = (± 3.733)^2**, i.e., $F_{1,3} = t_3^{\,2}$ $\Rightarrow$ equivalent tests.

How well does the model fit? Out of a total response variation of 7000, the linear regression model accounts for 5760, with the remaining 1240 unaccounted for (perhaps explainable by a better model, or simply due to random chance). We can therefore assess how well the model fits the data by calculating the ratio $\dfrac{SS_{Reg}}{SS_{Total}} = \dfrac{5760}{7000} = 0.823$. That is, *82.3% of the total response variation is due to the linear association between the variables*, as determined by the least squares regression line, with the remaining 17.7% unaccounted for. (**Note:** This does NOT mean that 82.3% of the original data points lie on the line. This is clearly false; from the scatterplot, it is clear that *none* of the points lies on the regression line!)

Moreover, note that $0.823 = (0.907)^2 = r^2$, the *square* of the correlation coefficient calculated before! This relation is true in general…

Coefficient of Determination

$$r^2 = \frac{SS_{Reg}}{SS_{Total}} = 1 - \frac{SS_{Err}}{SS_{Total}}$$

This value (always between 0 and 1) indicates the proportion of total response variation that is accounted for by the least squares regression model.

Comment: In practice, it is tempting to over-rely on the coefficient of determination as the sole indicator of linear fit to a data set. As with the correlation coefficient r itself, a reasonably high r^2 value is suggestive of a linear trend, or a strong linear component, but should not be used as the definitive measure.

Exercise: Sketch the $n = 5$ data points (X, Y)

$$(0, 0),\ (1, 1),\ (2, 4),\ (3, 9),\ (4, 16)$$

in a scatterplot, and calculate the **coefficient of determination** r^2 in two ways:

1. By squaring the **linear correlation coefficient** r.
2. By explicitly calculating the ratio $\dfrac{SS_{Reg}}{SS_{Total}}$ from the regression line.

Show agreement of your answers, and that, despite a value of r^2 very close to 1, the *exact* association between X and Y is actually a nonlinear one. Compare the linear estimate of Y when $X = 5$, with its exact value.

Also see **Appendix > Geometric Viewpoint > Least Squares Approximation**.

Regression Diagnostics – *Checking the Assumptions*

	Response = $\overbrace{\text{Model}}$ + Error			
True Responses:	$Y = \overbrace{\beta_0 + \beta_1 X} + \varepsilon$	$\Leftrightarrow$	$\boxed{y_i = \beta_0 + \beta_1 x_i + \boldsymbol{\varepsilon_i}}$,	$i = 1, 2, ..., n$
Fitted Responses:	$\hat{Y} = b_0 + b_1 X$	$\Leftrightarrow$	$\hat{y}_i = b_0 + b_1 x_i$,	$i = 1, 2, ..., n$
Residuals:	$\hat{\varepsilon} = Y - \hat{Y}$	$\Leftrightarrow$	$\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$,	$i = 1, 2, ..., n$

1. The model is "correct."

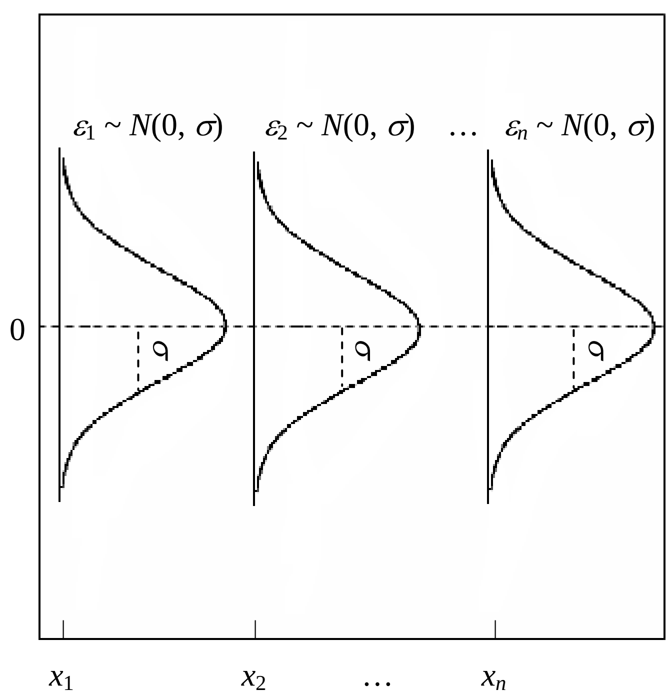
Perhaps a better word is "useful," since correctness is difficult to establish without a theoretical justification, based on known mathematical and scientific principles.

Check: Scatterplot(s) for general behavior, $r^2 \approx 1$, overall balance of simplicity vs. complexity of model, and robustness of response variable explanation.

2. Errors ε_i are independent of each other, $i = 1, 2, \ldots, n$.

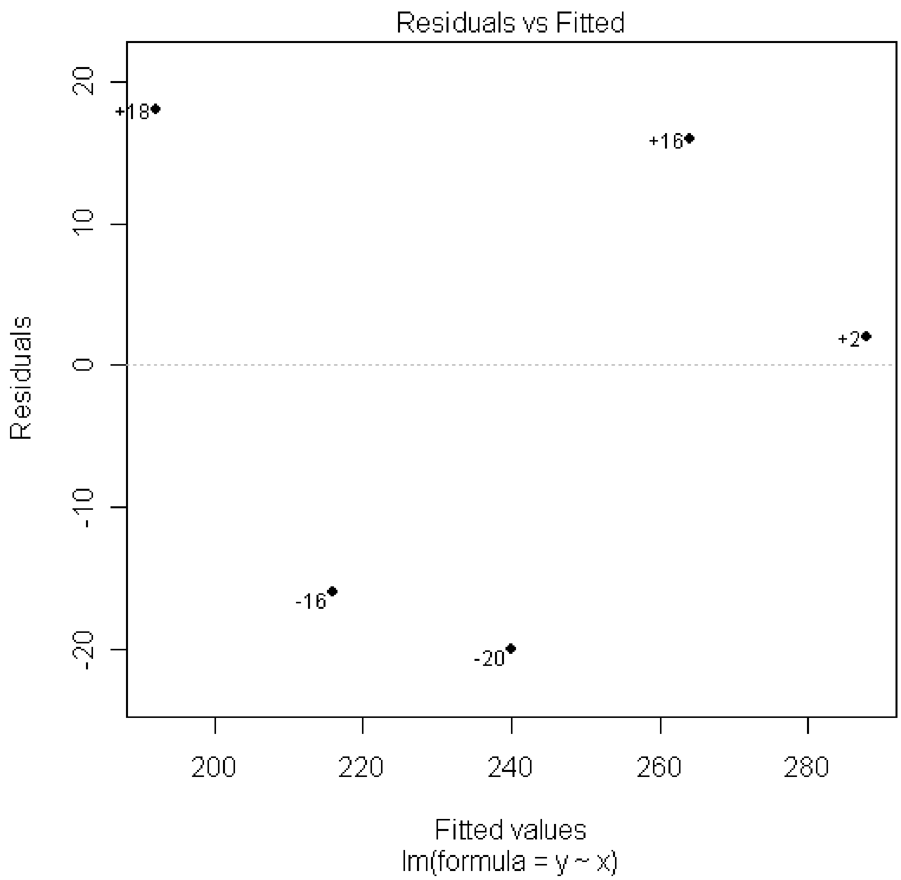
This condition is equivalent to the assumption that the *responses* y_i are independent of one other. Alas, it is somewhat problematic to check in practice; formal statistical tests are limited. Often, but not always, it is implicit in the design of the experiment. Other times, errors (and hence, responses) may be **autocorrelated** with each other. Example: Y = "systolic blood pressure (mm Hg)" at times $t = 0$ and $t = 1$ minute later. Specialized **time-series** techniques exist for these cases, but are not pursued here.

3. Errors ε_i are normally distributed with mean 0, and equal variances $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$ $(= \sigma^2)$, i.e., $\varepsilon_i \sim N(0, \sigma)$, $i = 1, 2, \ldots, n$.

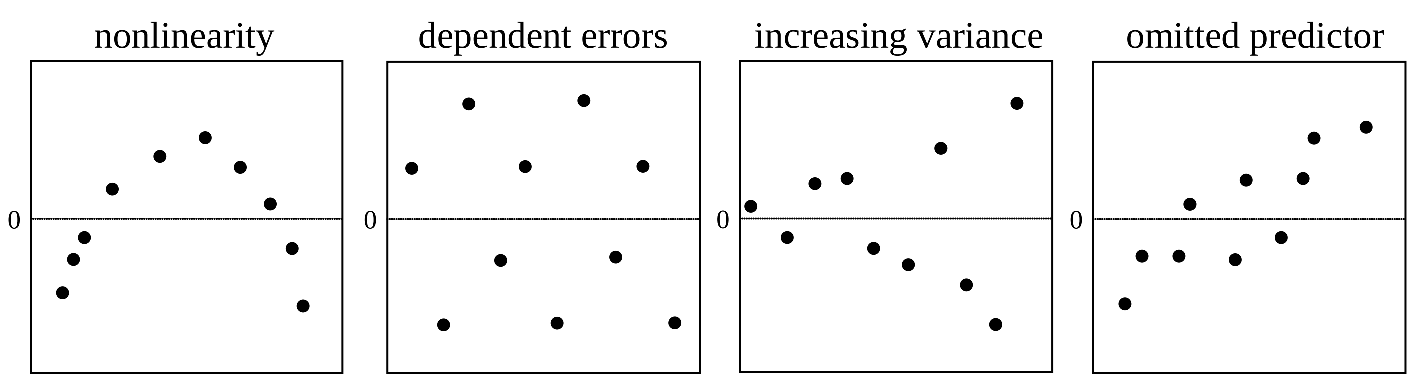


This condition is equivalent to the original normality assumption on the *responses* y_i. Informally, if for each fixed x_i, the true response y_i is normally distributed with mean $\mu_{Y\,|\,X=x_i}$ and variance σ^2 – i.e, $y_i \sim N(\mu_{Y\,|\,X=x_i}, \sigma)$ – then the error ε_i that remains upon "subtracting out" the true model value $\beta_0 + \beta_1 x_i$ (see boxed equation above) turns out also to be normally distributed, with mean 0 and the same variance σ^2 – i.e., $\varepsilon_i \sim N(0, \sigma)$. Formal details are left to the mathematically brave to complete.

<u>*Check*</u>: **Residual plot** (residuals e_i vs. fitted values $\hat{y}_i$) for a general random appearance, evenly distributed about zero. (Can also check the **normal probability plot**.)



Typical residual plots that violate Assumptions 1-3:



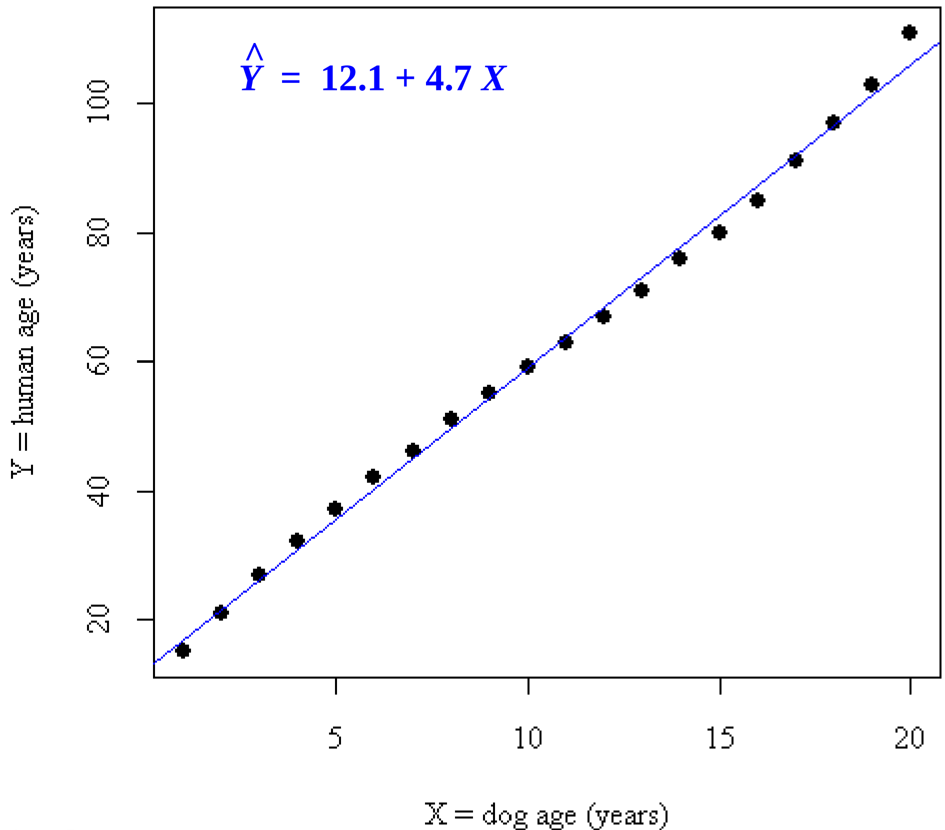
Nonlinear trend can often be described with a **polynomial regression** model, e.g., $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$. If a residual plot resembles the last figure, this is a possible indication that more than one predictor variable may be necessary to explain the response, e.g., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, **multiple linear regression**. Nonconstant variance can be handled by **Weighted Least Squares (WLS)** – versus **Ordinary Least Squares (OLS)** above – or by using a **transformation** of the data, which can also alleviate nonlinearity, *as well as violations of the third assumption that the errors are normally distributed.*

Example: Regress Y = "human age (years)" on X = "dog age (years)," based on the following n = 20 data points, for adult dogs 23-34 lbs.:

X	1	2	3	4	5	6	7	8	9	10
Y	15	21	27	32	37	42	46	51	55	59

11	12	13	14	15	16	17	18	19	20
63	67	71	76	80	85	91	97	103	111

Sadie

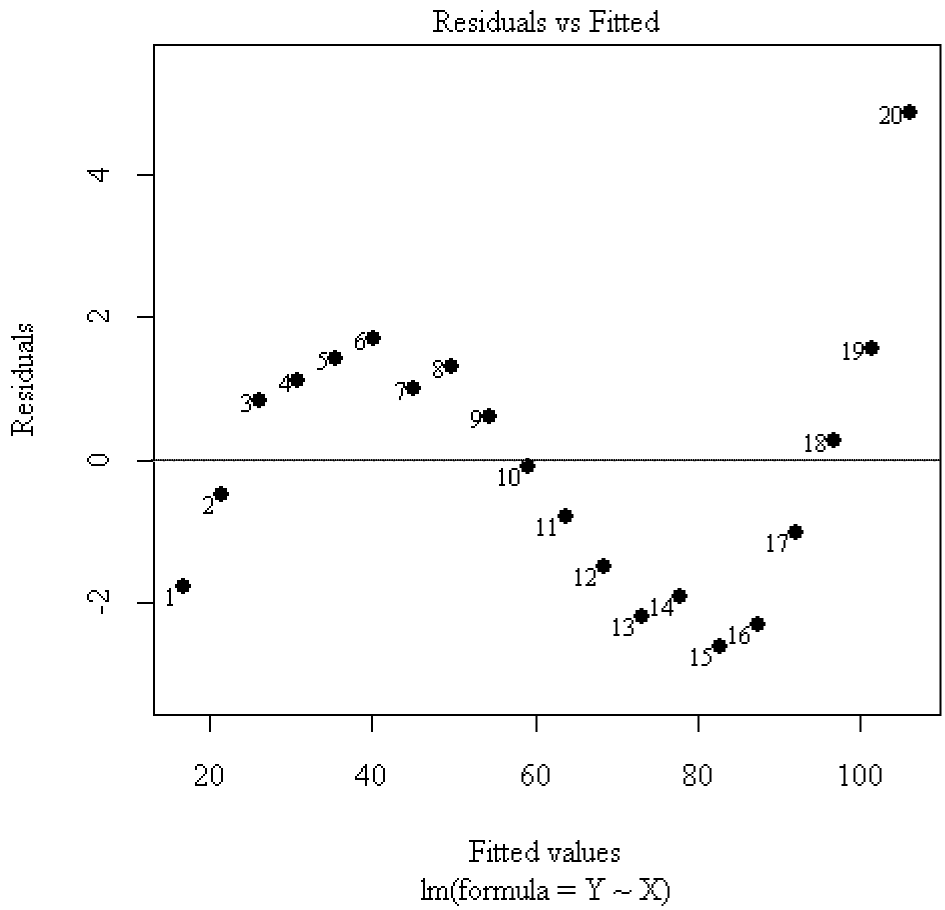



```
Residuals:
    Min       1Q     Median     3Q       Max
-2.61353 -1.57124  0.08947  1.16654  4.87143

Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
(Intercept) 12.06842    0.87794   13.75  5.5e-11 ***
X            4.70301    0.07329   64.17  < 2e-16 ***

Multiple R-Squared: 0.9956,     Adjusted R-squared: 0.9954
F-statistic:  4118 on 1 and 18 degrees of freedom,
p-value:     0
```

The residual plot exhibits a clear nonlinear trend, despite the excellent fit of the linear model. It is possible to take this into account using, say, a cubic (i.e., third-degree) polynomial, but this then begs the question: *How complicated should we make the regression model?*



My assistant and I, thinking hard about regression models.