

Index of 11-L

Page	Title
1	Practical information
2	Scatterplots and data example
3	Linear regression: Data + problem
4	Linear relation
5	Supplementary exercises 2.31 and 2.32
6	Linear regression model
7	Least squares estimation
8	Parameter estimates
9	Tests and confidence intervals
10	ANOVA Table for Linear Regression
11	Prediction
12	Standard errors in linear regression
13	How to report statistics in scientific papers?
14	Statistical reporting guidelines I
15	Statistical reporting guidelines II

PRACTICAL INFORMATION

Schedule:

- third home assignment and optional project description due today,
- home assignment IV on the website sometime next week,
- next lab session: Tuesday afternoon, 1-4pm.

Today's lecture:

- ANOVAs revisited: based on summary worksheet (11.3),
- main topic of next two lectures: correlation and regression,
 - * today regression, including prediction,¹
 - * next week correlation + extra regression,²
 - * postpone references to r^2 ; entirely skip extra topics (IPS: scatterplot smoothers, nonlinear regression),
- additional topic:
 - * guidelines on how to report statistics in papers/theses,
- another “misconceptions clinic” after the break.

Updates of course information:

- material on reporting guidelines (webpage + Moodle),
- “after the ANOVA”: schematic + nonparam. (media page).

¹ PLS 4e: Chapters 4, 23; S: Sections 10.1-2; IPS 7e: Sections 2.1, 2.3, 10.1-2.

² PLS 4e: Chapters 3, 23; S: Section 10.1; IPS 7e: Sections 2.2, 2.4, 10.2.

SCATTERPLOTS AND DATA EXAMPLE

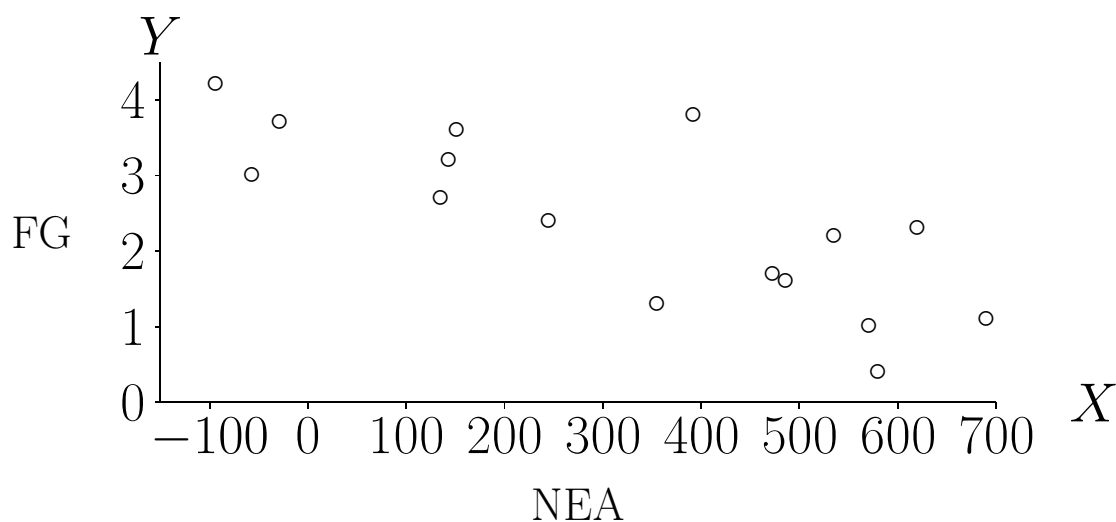
Scatterplot: a plot of two variables against each other:

- explanatory variable (if any) goes on the horizontal axis,
- one point per observation pair (Y, X) .

Data example: non-exercise activity (NEA) and fat gain (FG) in humans,³

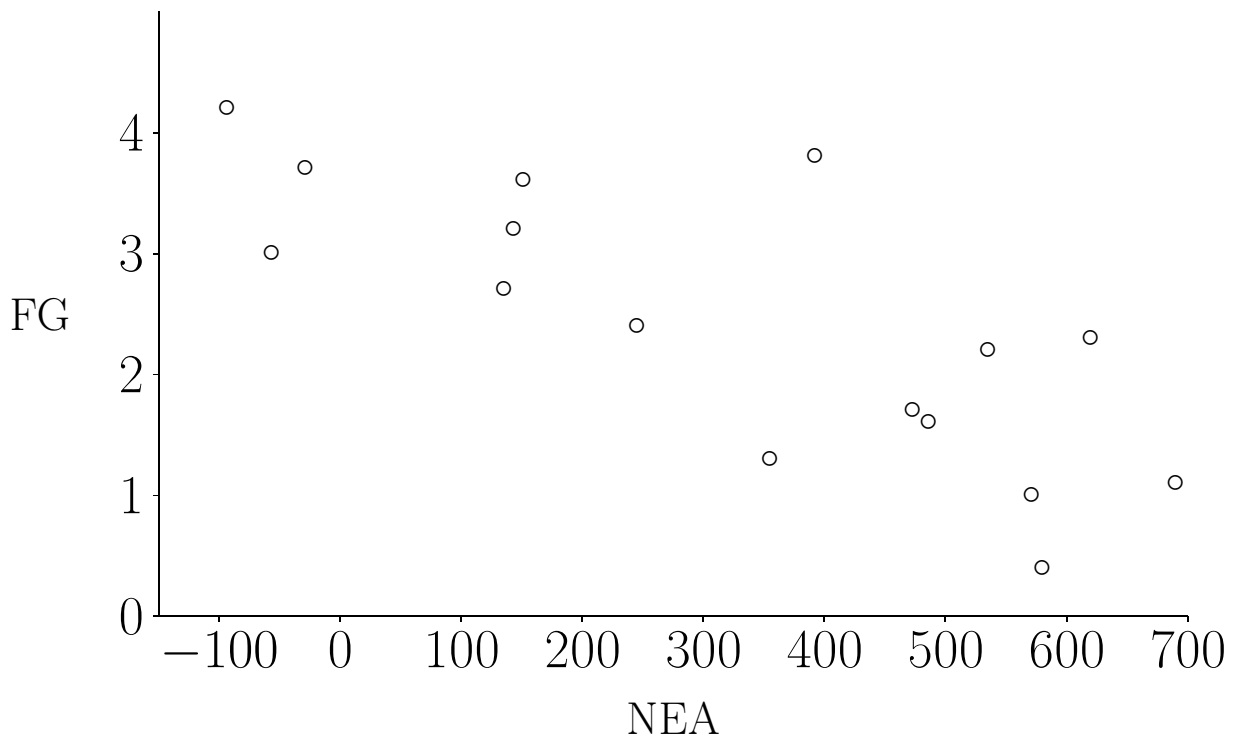
- for 16 young adults that were overfed for 8 weeks, measures of
 - * increase in NEA⁴, measured in calories,
 - * fat gain (FG), measured in kilograms,
- interest is in predicting fat gain from NEA,

- | | | | | | | | |
|----------|-----|-----|-----|-----|-----|-----|-----|
| fat gain | Y | 4.2 | 3.0 | 3.7 | 2.7 | ... | 1.1 |
| NEA | X | -94 | -57 | -29 | 135 | ... | 690 |



³ IPS 7e Example 2.18, data from Levine et al. (1999), *Science* **283**, 212-214.
⁴ NEA = any activity other than deliberate exercise, such as fidgeting, daily living, etc.; fidget(v): to make continuous small movements that annoy other people.

LINEAR REGRESSION: DATA + PROBLEM



Data:

$$\left. \begin{array}{l} Y_i = \text{fat gain} \\ X_i = \text{NEA} \end{array} \right\} \text{ for subject } i, i = 1, \dots, 16 = n.$$

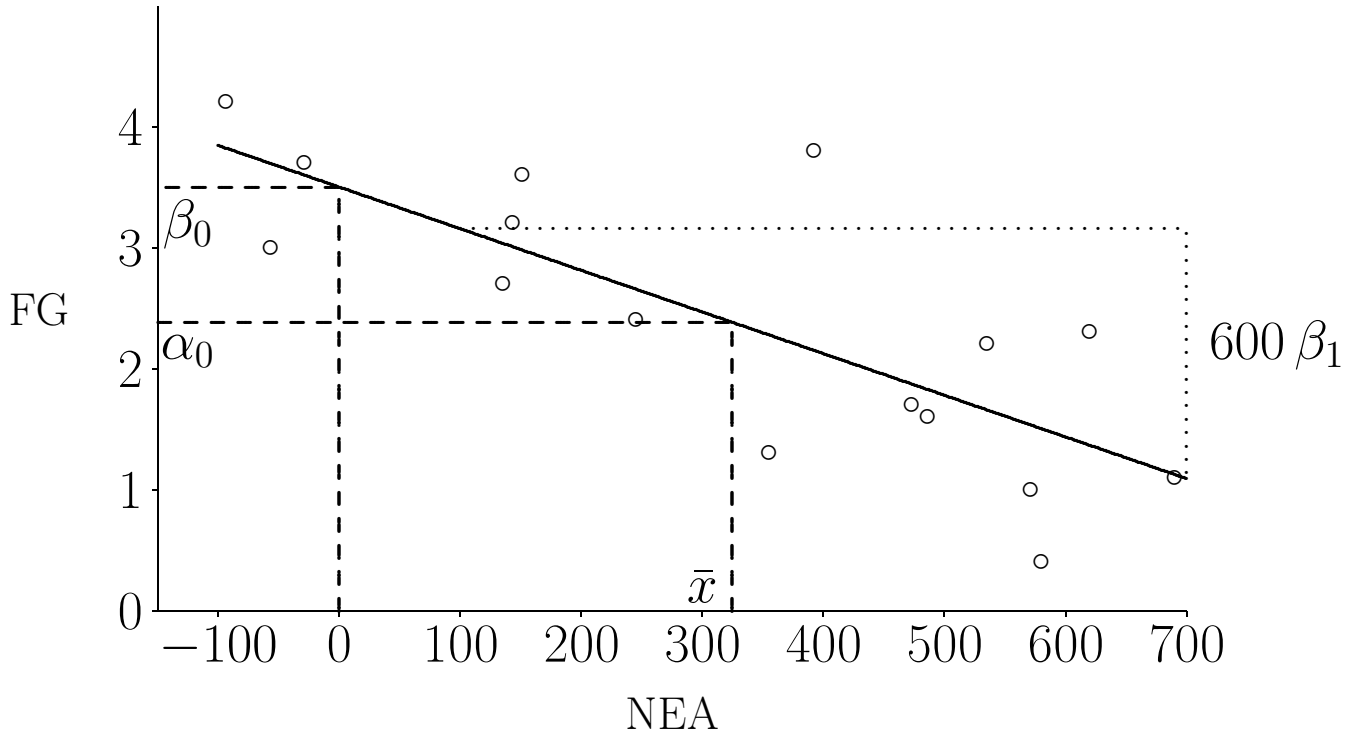
Problem: seek description of relationship between Y and X , in particular as: $y = f(x)$,

Why y as a function of x ?⁵

- causal relation? (if x controllable, we hope to impact y),
- interest in predicting y from x ?
(for prediction, x 's would be taken as fixed),
- X is not a random variable (\Rightarrow explanatory).

⁵ Commonly used (but somewhat imprecise) terminology to reflect this: $y =$ dependent variable, $x =$ independent variable.

LINEAR RELATION



Linear relation: $y = \beta_0 + \beta_1 \cdot x$

(or $y = a + bx$, as in Chapter 5 of PSLS, Chapter 2 of IPS):

- β_1 (or b) = slope of the line,
- β_0 (or a) = intercept of line with vertical axis ($x=0$),
- interpretation of slope: one unit increase in x implies a β_1 units change (increase or decrease) in y .

Alternative writing of the same line: $y = \alpha_0 + \beta_1(x - \bar{x})$:

- $\alpha_0 = y$ -value corresponding to $x = \bar{x}$,
- “centering” of x to avoid parameter (β_0) out of x ’s range,
- $\beta_0 = \alpha_0 - \beta_1 \bar{x}$, or $\alpha_0 = \beta_0 + \beta_1 \bar{x}$.

SUPPLEMENTARY EXERCISES 2.31 AND 2.32

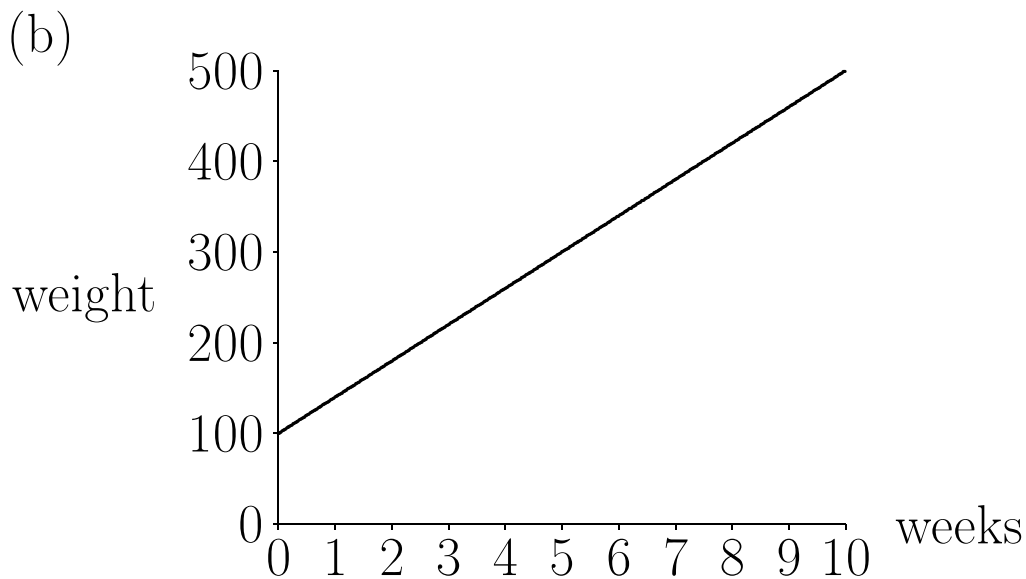
Exercise 2.31:

If x = number of seconds since splash and y = distance in meters, the equation is

$$y = 1500 \text{ (m/s)} \cdot x \text{ (s)}.$$

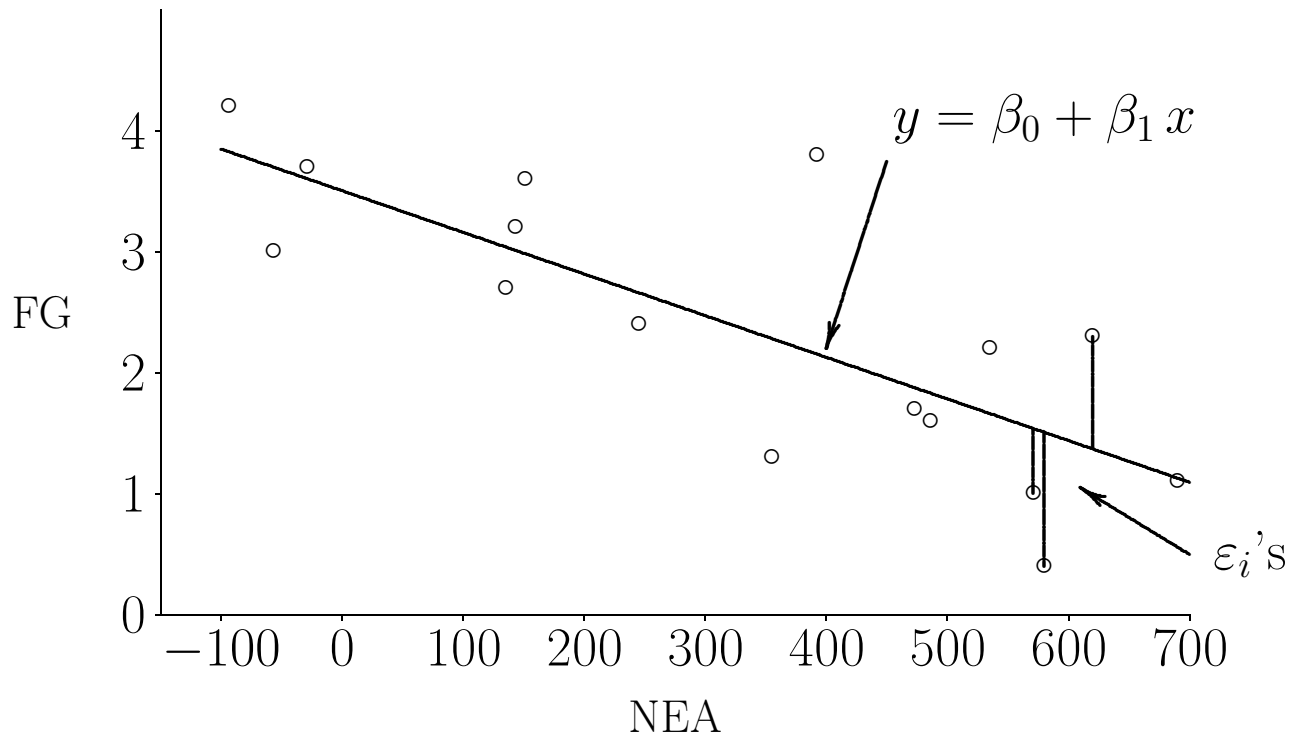
Exercise 2.32:

(a) equation: $\text{weight} = 100 + 40 \cdot \text{weeks}$,
and the slope of the line is 40 (g/week) .



(c) to use the linear equation for 2 years (104 weeks) would be an extreme case of *extrapolation* and clearly invalid, because rats do not continue to grow at a linear rate. Predicted value = $100 + 40 \cdot 104 = 4260 \text{ g}$.

LINEAR REGRESSION MODEL



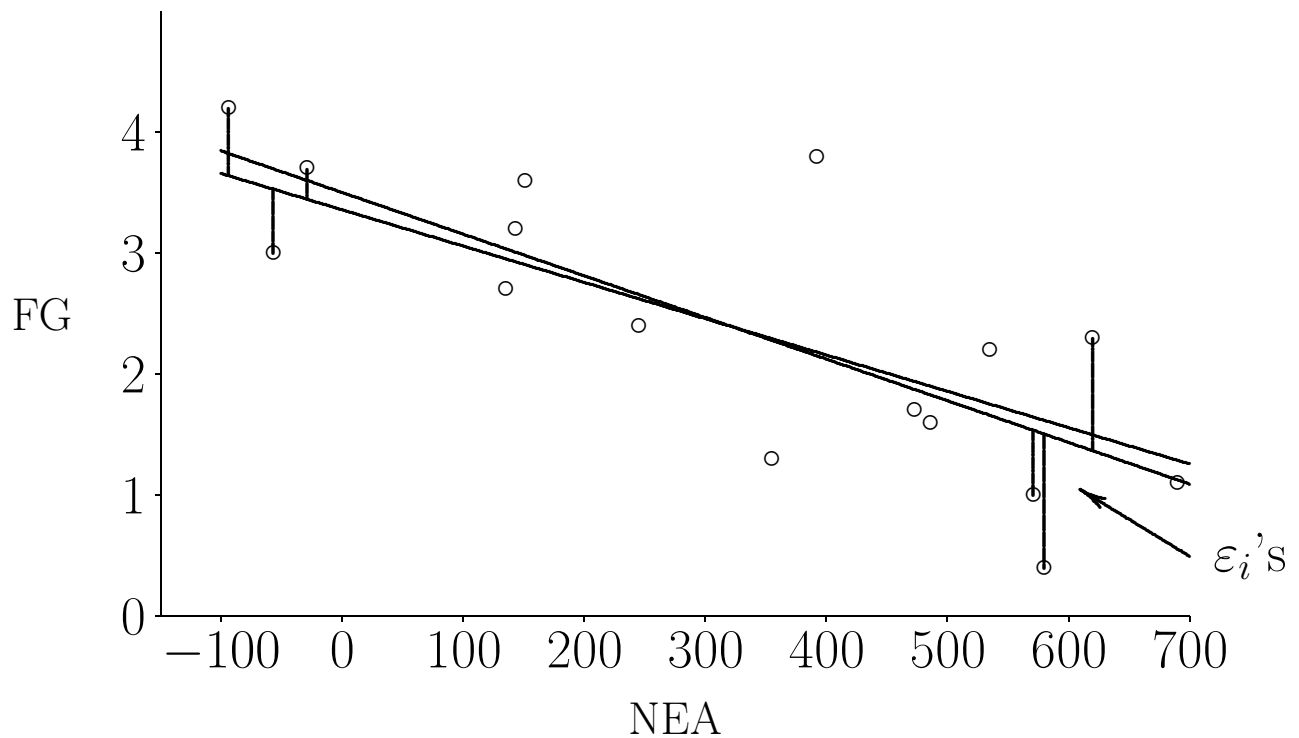
Statistical model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (= \alpha_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i),$$

where the (*vertical*) errors $\varepsilon_1, \dots, \varepsilon_{16}$ are i.i.d. and $\sim N(0, \sigma)$,

- parameters: β_1, β_0 (or α_0) and σ ,
- x 's considered fixed — thus no capitals,
- assumptions:
 - * the linear relation: $EY_i = \beta_0 + \beta_1 x_i$,
 - * normal distribution of errors ε_i ,
 - * same stand.dev. of all observations (homogeneity),
 - * independence of errors (and of observations).

LEAST SQUARES ESTIMATION



How to choose the regression line (estimate β_0, β_1)?

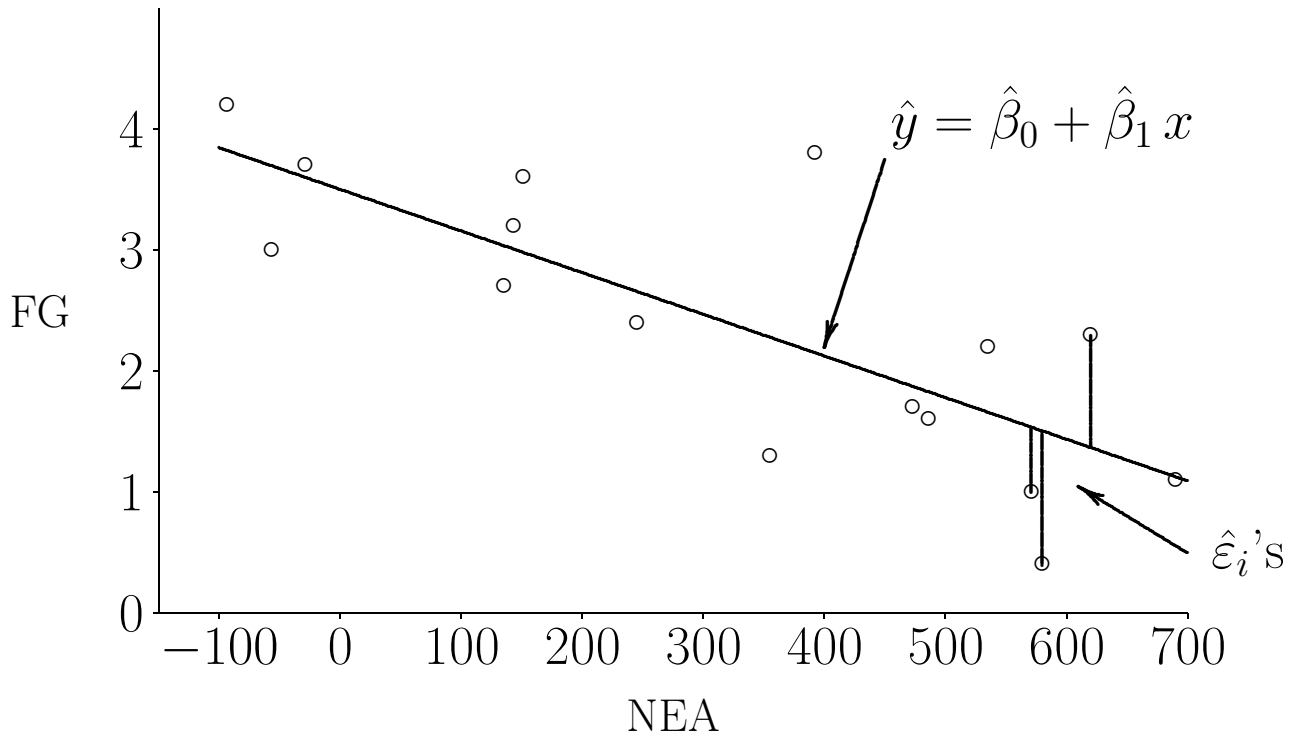
Idea: “best” line minimizes the sum of squared errors

$$\sum_i \varepsilon_i^2 = \sum_i (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Motivations:

- intuitive (minimizes squared vertical deviations),
- easy to calculate (solutions have closed formulae),
- resulting estimates have good theoretical properties (unbiased and optimal for present model).

PARAMETER ESTIMATES

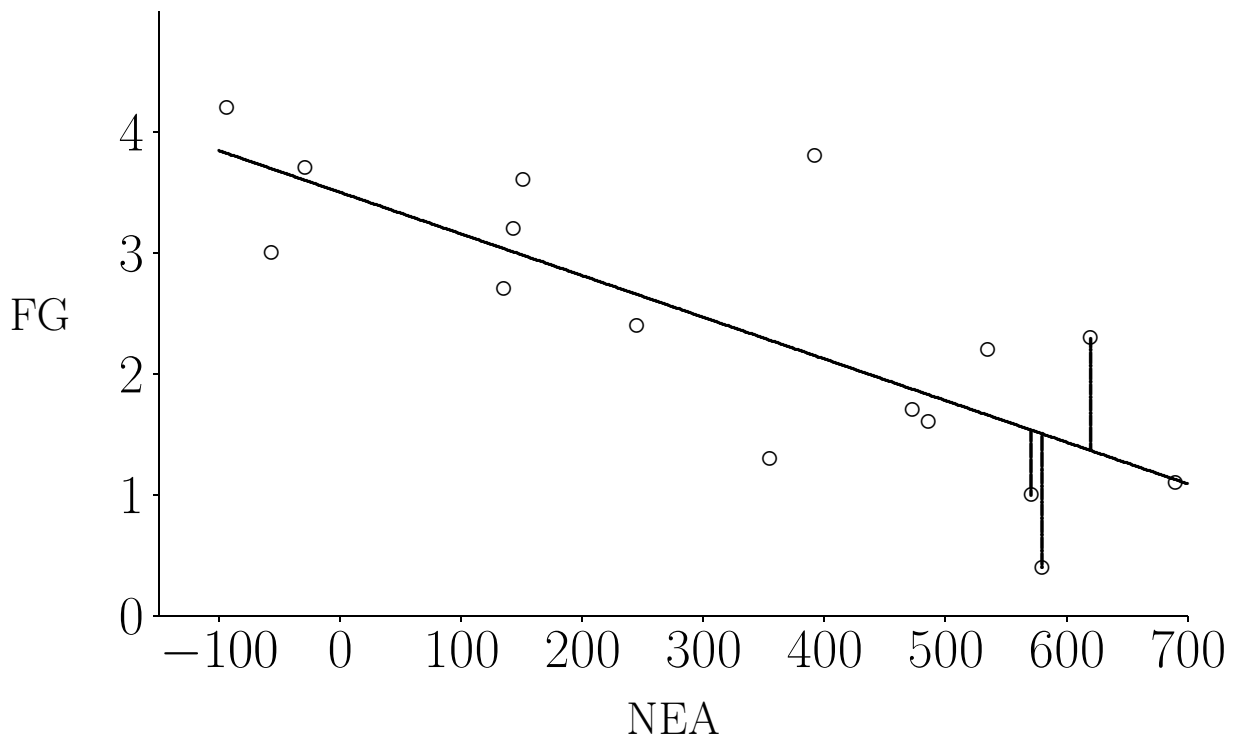


Parameter estimates:

- slope: $\hat{\beta}_1 = \frac{\sum_i (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = r s_y / s_x$, ($r = \text{correlation}$),
- $\hat{\alpha}_0 = \bar{Y}$ (\Rightarrow estimated line passes through (\bar{x}, \bar{Y})),
- intercept: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$,
- estimated line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$,
- residual: $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, (“observed – predicted”)
- $\hat{\sigma}^2 = s^2 = \frac{1}{n-2} \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n-2} \sum_i \hat{\varepsilon}_i^2$.

Minitab/Stata/R give the estimates and associated standard errors for mean parameters.

TESTS AND CONFIDENCE INTERVALS

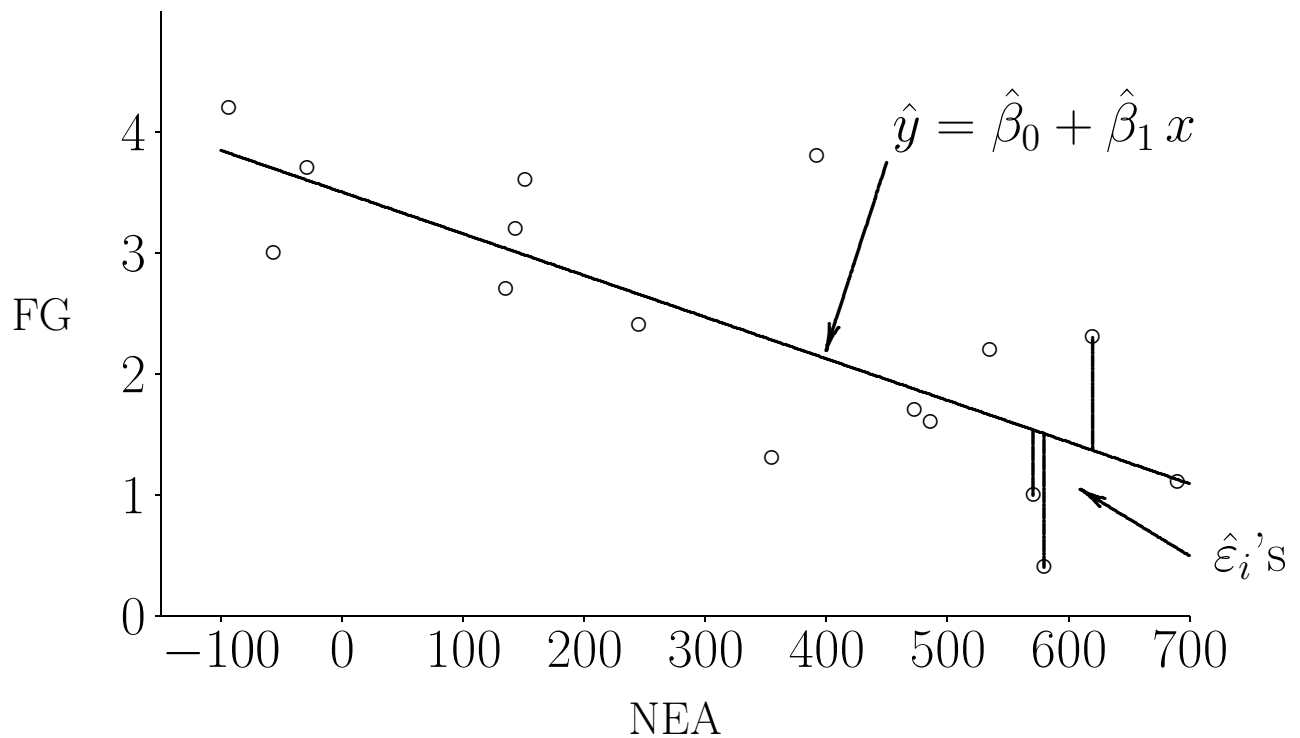


Statistical hypotheses about the parameters of the regression line are tested the “usual way”, using estimates and their standard errors:

- degrees of freedom for s^2 : $DFE = n - 2$,
- example: test of slope equal to known value:
 - * $H_0: \beta_1 = b$, (b known, fixed value),
 - * $H_a: \beta_1 \neq b$ (two-sided alternative),
 - * test: $t = (\hat{\beta}_1 - b) / SE(\hat{\beta}_1) \sim t(DFE)$ -dist. under H_0 ,
 most common example is $b = 0$ (horizontal line \sim no linear relation between x and y).

Confidence intervals: also the “usual way”, for example,
 95% CI: estimate $\pm t^* \cdot SE(\text{estimate})$, $t^* = t_{.975}(DFE)$.

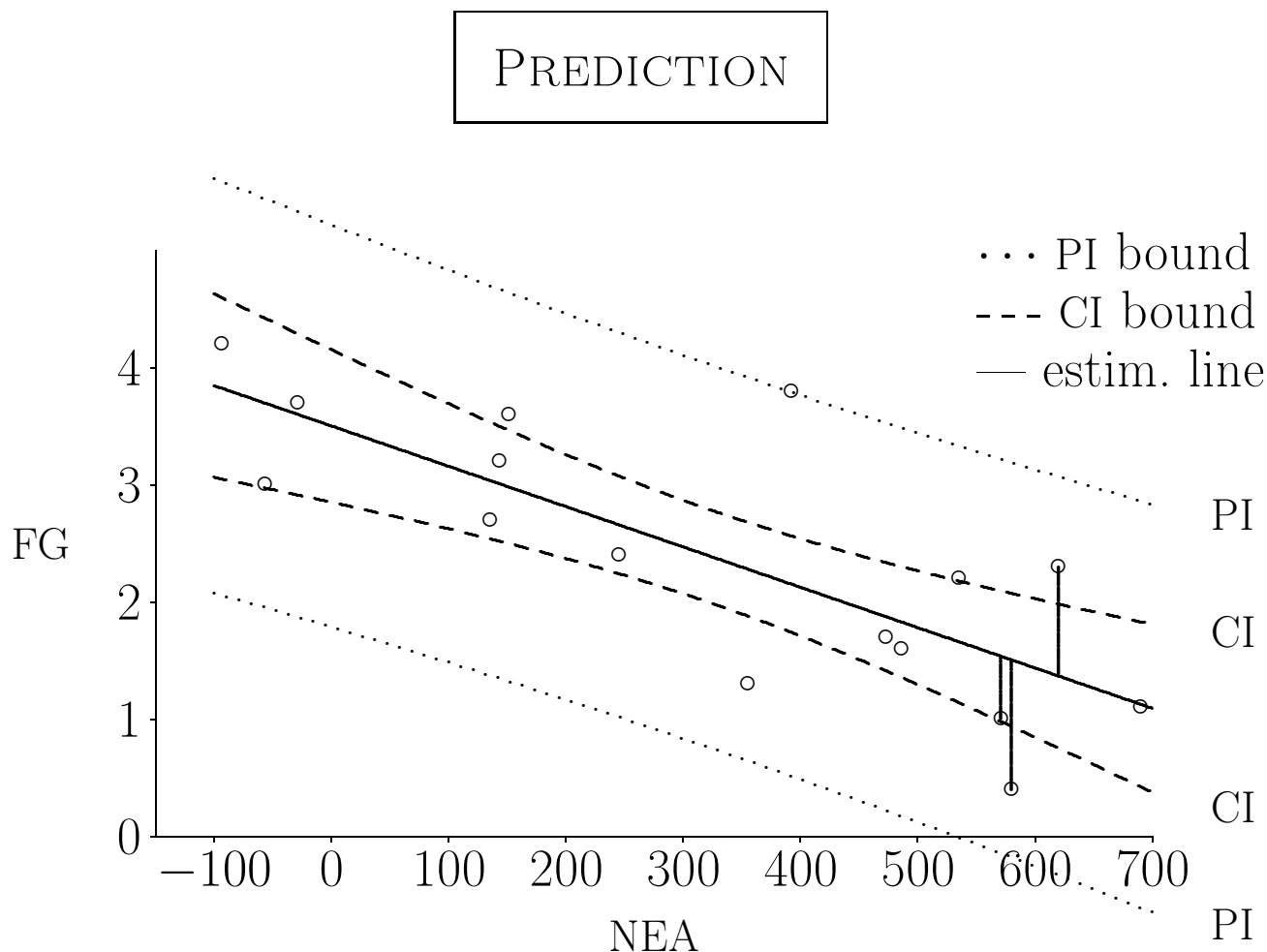
ANOVA TABLE FOR LINEAR REGRESSION



Hypothesis $H_0: \beta_1 = 0$ can also be tested by ANOVA:

Source of variation	Degrees of freedom	Sum of squares	Mean square	F
Regression Model	DFM = 1	SSM	MSM = SSM/DFM	MSM/MSE
Error	DFE = $n - 2$	SSE = $\sum_i \hat{\epsilon}_i^2$	MSE = SSE/DFE	
Total	DFT = $n - 1$	SST		

- estimated error variance = $s^2 = \text{MSE}$, as usual,
- F -test equivalent to t -test, because $F = t^2$, (same P)
- ANOVA table not really needed for simple linear regression (but for models with more x -variables).

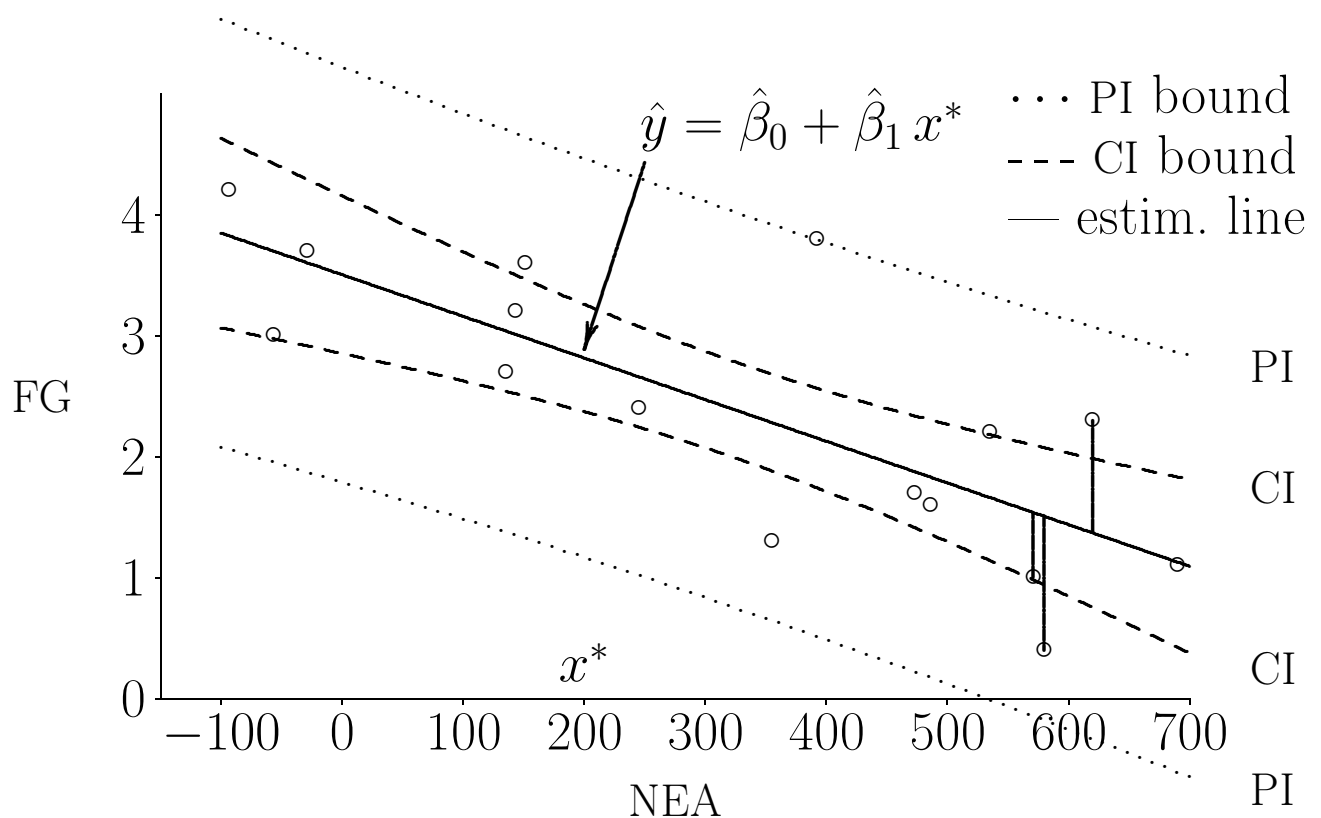


Prediction / Estimation:⁶

- 2 situations / purposes:
 - (CI) estimation of the regression line for given x , and CI to indicate the precision of the estimation,
 - (PI) prediction of a new observation for given x , and prediction interval (PI) to indicate *both* precision of the line (mean) and the dispersion around it,
- same estimated/predicted value: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$,
- CI for line more narrow than PI for new observation.

⁶ Stata terminology: prediction \sim estimation, forecasting \sim prediction.

STANDARD ERRORS IN LINEAR REGRESSION



Formulae for standard errors⁷ in linear regression:

$$\begin{aligned} \text{slope : } SE(\hat{\beta}_1) &= s / \sqrt{\sum_i (x_i - \bar{x})^2}, \\ \text{intercept : } SE(\hat{\beta}_0) &= s \sqrt{1/n + \bar{x}^2 / \sum_i (x_i - \bar{x})^2}, \\ \text{CI}(x^*) : SE(\hat{\mu}) &= s \sqrt{1/n + (x^* - \bar{x})^2 / \sum_i (x_i - \bar{x})^2}, \\ \text{PI}(x^*) : SE(\hat{y}) &= s \sqrt{1 + 1/n + (x^* - \bar{x})^2 / \sum_i (x_i - \bar{x})^2}. \end{aligned}$$

Notes:

- we usually don't compute by hand (calculator)!
- the squared variation of x 's determines accuracy,
- added "1" in prediction formula compared to estimation.

⁷ Strictly speaking, the error involved in a prediction is not a standard error but a prediction error; however, we stick here to the textbook notation/terminology.

HOW TO REPORT STATISTICS IN SCIENTIFIC PAPERS?

The last two decades has seen much stronger focus on appropriate conduct and reporting of statistical analysis in the published (peer-reviewed) literature, due to

- greater awareness of problems/issues, in particular in (human) health sciences,
- much larger variety of statistical methods becoming accessible through (variety of) statistical software,
- development of systematic review and meta-analysis (studies reported inappropriately cannot be included),
- debate on philosophical issues around statistical analysis (e.g., P -values; 6L–8),
- general interest in establishing more strict guidelines for scientific research.

Many guidelines for specific study types exist⁸, covering planning, execution, analysis and interpretation, in particular through the EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network. Scientific journals increasingly require compliance with relevant guideline(s).

Statistical reporting guidelines

- “original” (old, still useful): Bailar & Mosteller, 1988 (I),
- more recent: Lang & Altman, 2013 (II).

⁸ See links at the VHM 801 homepage, people.upei.ca/hstryhn/vhm801.

STATISTICAL REPORTING GUIDELINES I

Developed by the “International Committee of Medical Journal Editors”⁹:

1. Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results.
2. When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals).
3. Avoid sole reliance on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information.
4. Discuss eligibility of experimental subjects.
5. Give details about randomization.
6. Describe the methods for, and success of, any blinding of observations.
7. Report treatment complications.
8. Give numbers of observations (and give the experimental unit).
9. Report losses to observation (such as dropouts from a clinical trial).
10. References for study design and statistical methods should be to standard works (with pages stated) when possible rather than to papers where designs or methods were originally reported.
11. Specify any general-use computer programs used.
12. Put general descriptions of statistical methods in the Methods section. When data are summarized in the Results section, specify the statistical methods used to analyze them.
13. Restrict tables and figures to those needed to explain the argument of the paper and to assess its support. Use graphs as an alternative to tables with many entries; do not duplicate data in graphs and tables.
14. Avoid non-technical uses of technical terms in statistics, such as "random" (which implies a randomizing device), "normal", "significant", "correlation", and "sample".
15. Define statistical terms, abbreviations, and most symbols.

⁹ Bailar, J.C. & Mosteller, F. (1988), *Ann. Intern Med.* **108**, 266–73.

STATISTICAL REPORTING GUIDELINES II

Developed by two prominent statistics authors, but without formal consultation process (as for other guidelines)¹⁰, and aims to cover “basic” statistical analyses and methods.

- First guiding principle: statements (1, 2, 3, 10, 11, 15) from Bailar & Mosteller.
- Second guiding principle: provide enough detail that the results can be incorporated into other analyses.
- Reporting statistical methods — split into: preliminary, primary and secondary analyses (\Rightarrow researchers need to categorize their analyses and in particular identify their primary analyses, for which (among other things) a purpose must be described).
- Reporting statistical results — split into different types of outcomes and analyses, e.g. for hypothesis testing (selected topics/items):
 - * state the hypothesis being tested; report whether the test was one- or two-tailed (justify the use of one-tailed tests),
 - * identify the variables in the analysis and the name of the test used,
 - * if possible, identify the minimum difference considered to be clinically important; for equivalence and non-inferiority studies, report the largest difference between groups that will still be accepted as indicating biological equivalence,
 - * confirm that the assumptions of the test were met by the data,
 - * report the alpha level that defines statistical significance,
 - * at least for primary outcomes, report a measure of precision, such as the 95% confidence interval (do NOT use the standard error to indicate the precision of an estimate),
 - * although not preferred to confidence intervals, if desired, P values should be reported as equalities when possible and to one or two decimal places; the smallest P value that need be reported is $P < 0.001$, save in studies of genetic associations,
 - * report whether and how many adjustments were made for multiple statistical comparisons.

¹⁰ Lang T, Altman D (2013). Statistical Analyses and Methods in the Published Literature: the SAMPL guidelines, <http://www.equator-network.org/reporting-guidelines/sampl/>.