Biostats VHM 801 Course
Fall 2018, Atlantic Veterinary College, PEI
Henrik Stryhn

Index of 11-L

Practical Information

Schedule:

- third home assignment and optional project description due today,
- home assignment IV on the website sometime next week,
- lab session tomorrow afternoon.

Today's lecture:

- ANOVAs revisited: brief review with extra idea, based on summary worksheet (Chapter 11),
- 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, and regression vs. correlation until next week,
  - entirely skip extra topics (IPS: scatterplot smoothers, nonlinear regression),
- additional topic:
  - guidelines on how to report statistics in papers/theses,

Updates of course information:

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

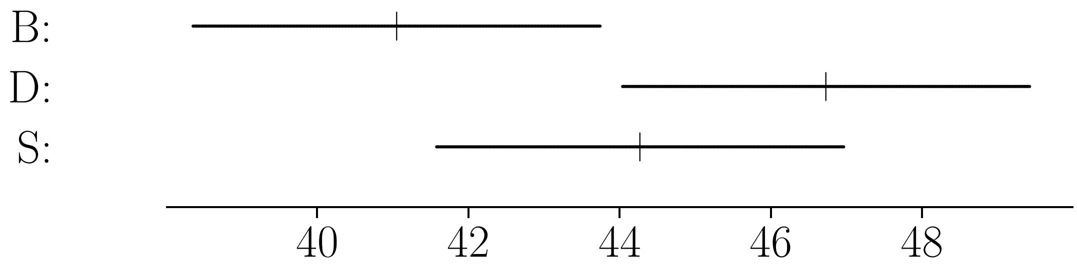
[1] PSLS 3e: Chapters 4, 23; S: Sections 10.1-2; IPS 7e: Sections 2.1, 2.3, 10.1-2.
[2] PSLS 3e: Chapters 3, 23; S: Section 10.1; IPS 7e: Sections 2.2, 2.4, 10.2.

Tips for Comparing and Presenting Groups

Group comparisons based on confidence intervals:

- based on test/CI for difference between parameters (e.g., $\mu_1 - \mu_2$),
- but conclusions available from group CIs in 2/3 cases (see figure):

  Reading data example: 95% CIs for `post3`:



  * B vs. D: disjoint (non-overlapping) CIs ⇒ signif. ($P < 0.05$),
  * D vs. S: estimate in another CI ⇒ no signif. ($P > 0.05$),
  * B vs. S: need CI for difference ($\mu_B - \mu_S$) to assess signif.

  assumes independent estimates, unadjusted for multiple testing.

Significance letter coding[3]: from software or constructed manually,

- order group means from lowest to highest,
- designate letter a to highest group + all groups not significantly different from it,
- designate letter b to next group in the same way (but drop if same pattern as for a),
- continue through all groups,
- Reading data: (uncorrected 5% error) coding: $\mathrm{B}^b\ \mathrm{S}^{ab}\ \mathrm{D}^a$,
- Ex. 12.55 (vit A): (Bonferroni corrected) coding: $7^b\ 3^{ab}\ 1^{ab}\ 5^a\ 0^a$.

[3] Meaning of letter codes: groups with same letter are *not* significantly different.

Scatterplots and Data Example

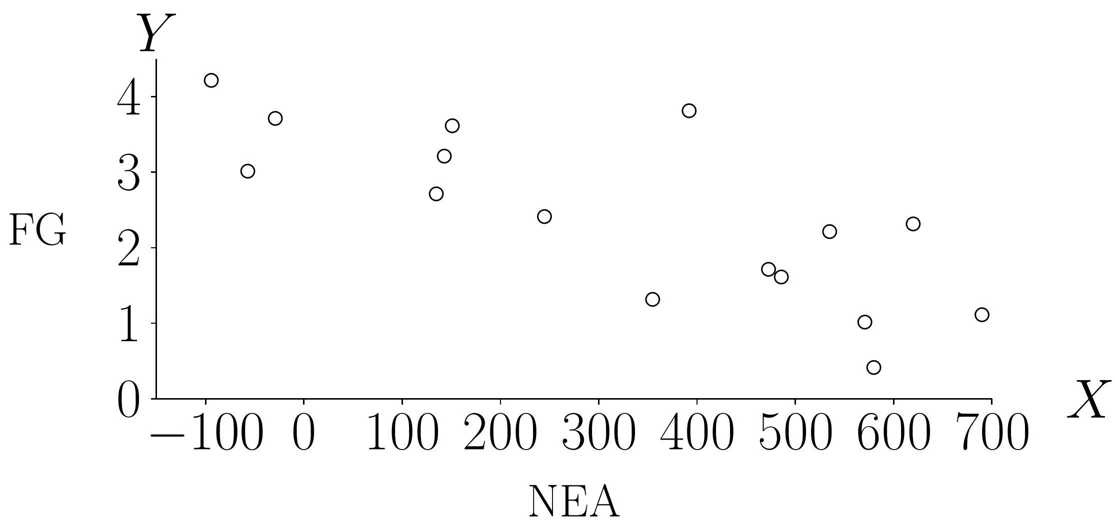
Scatterplot: a plot of two variables against each other:

- 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,[4]

- for 16 young adults that were overfed for 8 weeks, measures of
  - ∗ increase in NEA[5], measured in calories,
  - ∗ fat gain (FG), measured in kilograms,
- interest is in predicting fat gain from NEA,
-

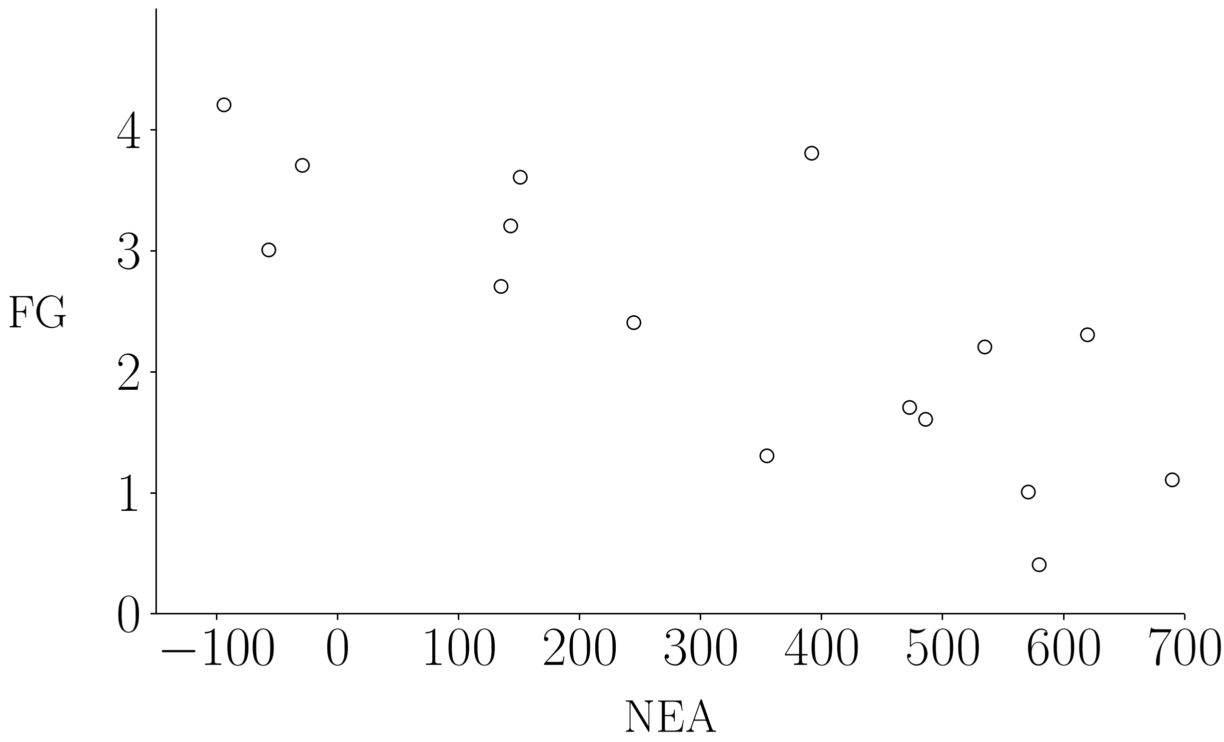
fat gain Y	4.2	3.0	3.7	2.7	...	1.1
NEA X	-94	-57	-29	135	...	690



[4] IPS 7e Example 2.18, data from Levine et al. (1999), *Science* **283**, 212-214.

[5] 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:

$$\left.\begin{aligned} Y_i &= \text{fat gain} \\ X_i &= \text{NEA} \end{aligned}\right\} \quad \text{for subject } i,\ i = 1, \ldots, 16 = n.$$

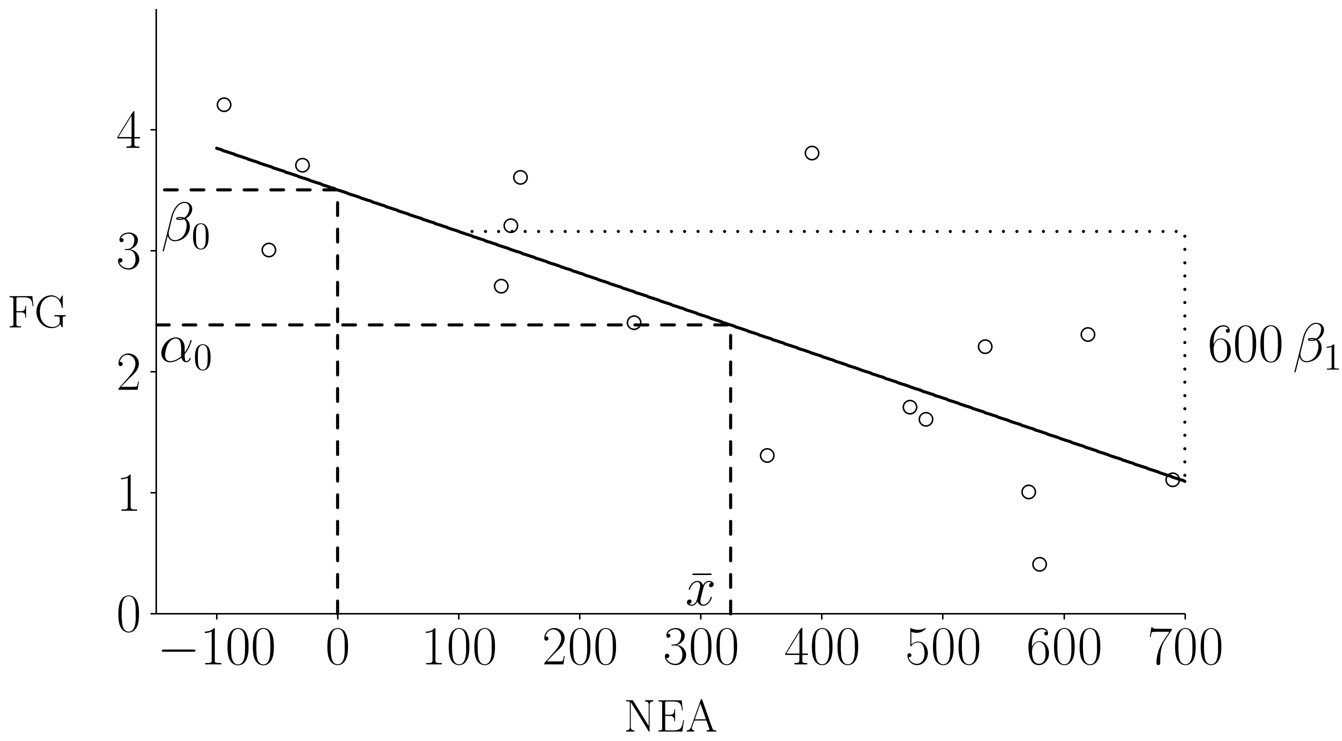
Problem: seek description of relationship between Y and X, in particular as: $y = f(x)$,

Why y as a function of x?[6]

- 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).

[6] 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: $y = \alpha_0 + \beta_1(x - \bar{x})$:

- α_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:

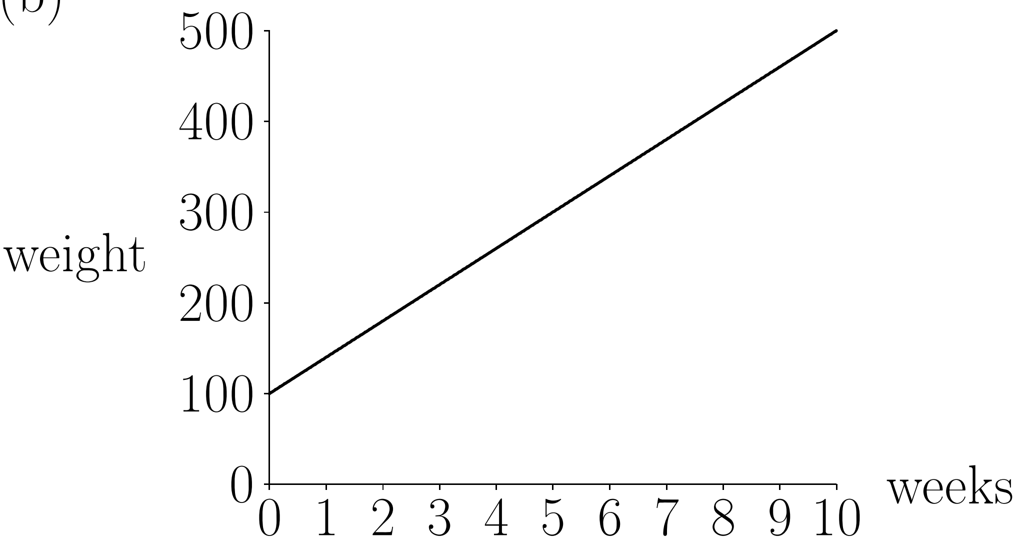
If x = number of seconds since splash and y = distance in meters, the equation is

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

Exercise 2.32:

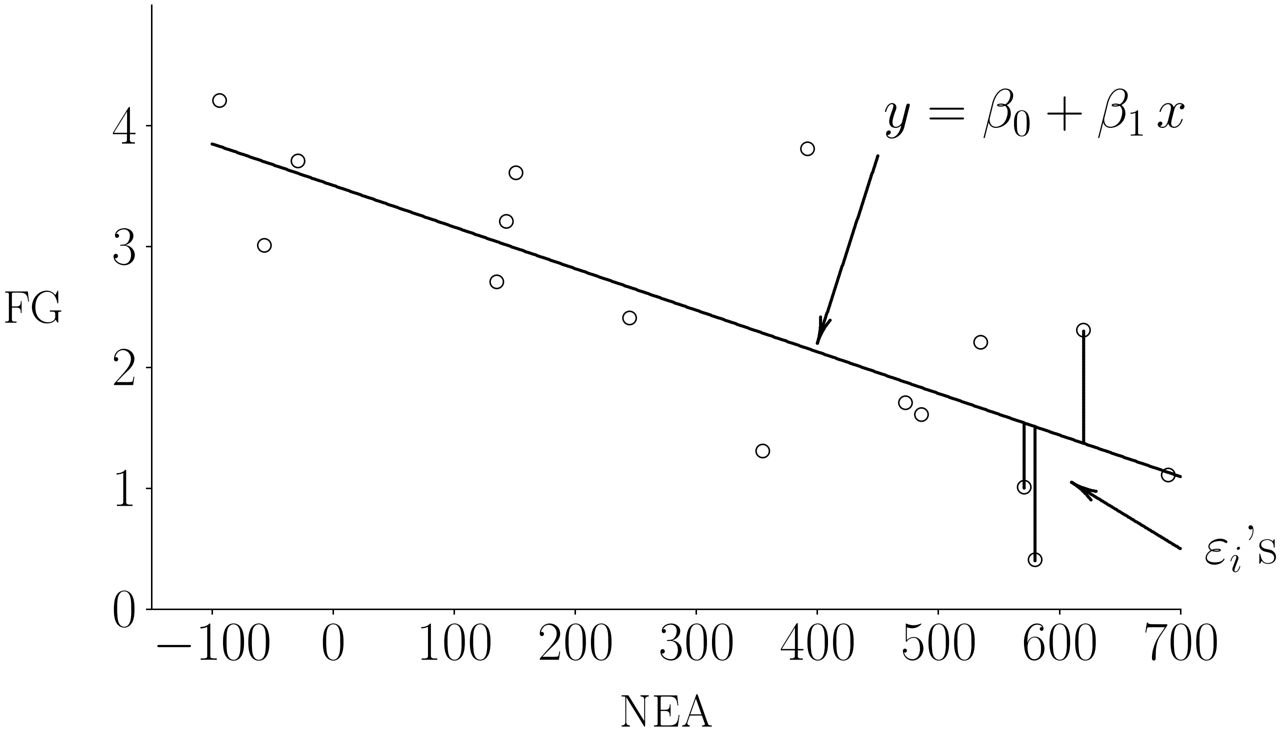
(a) equation: weight $= 100 + 40 \cdot$ weeks,
and the slope of the line is 40 (g/week).

(b)



(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 $= 100 + 40 \cdot 104 = 4260\, g$.

Linear Regression Model



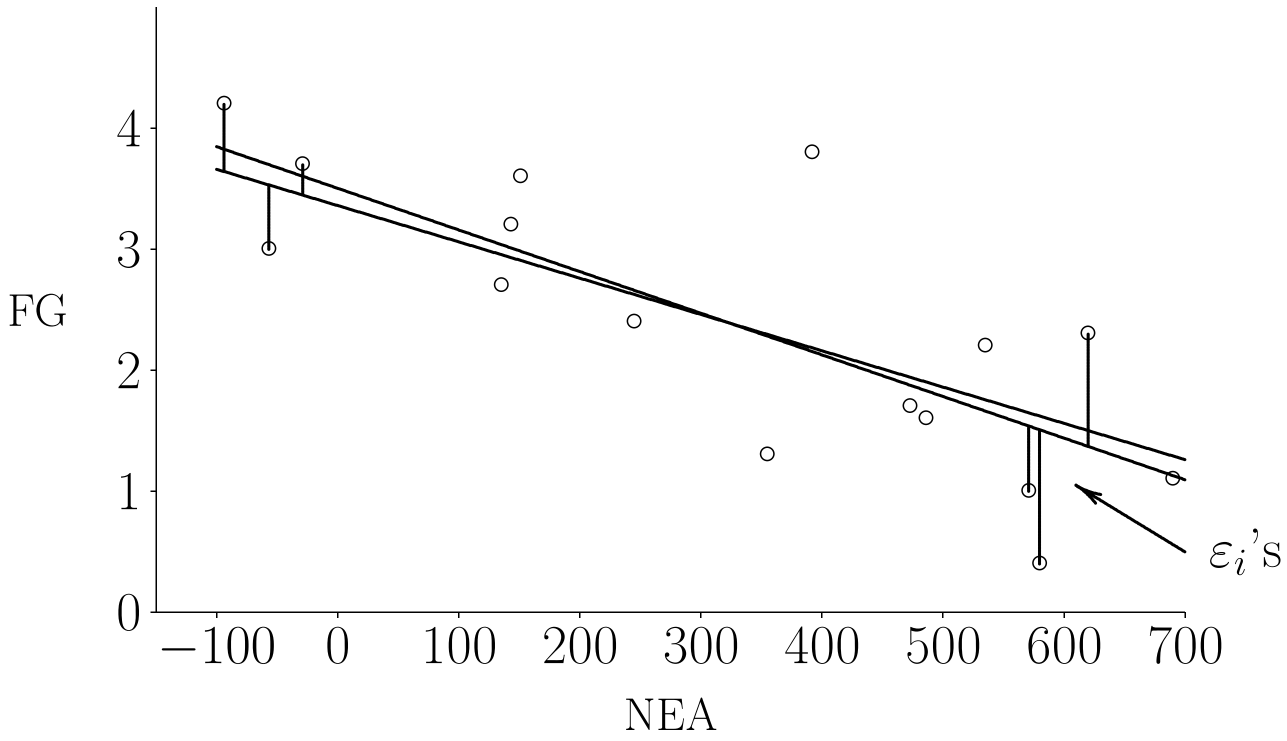
Statistical model:

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

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

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

Least Squares Estimation



<u>How to choose</u> the regression line (estimate β_0, β_1)?

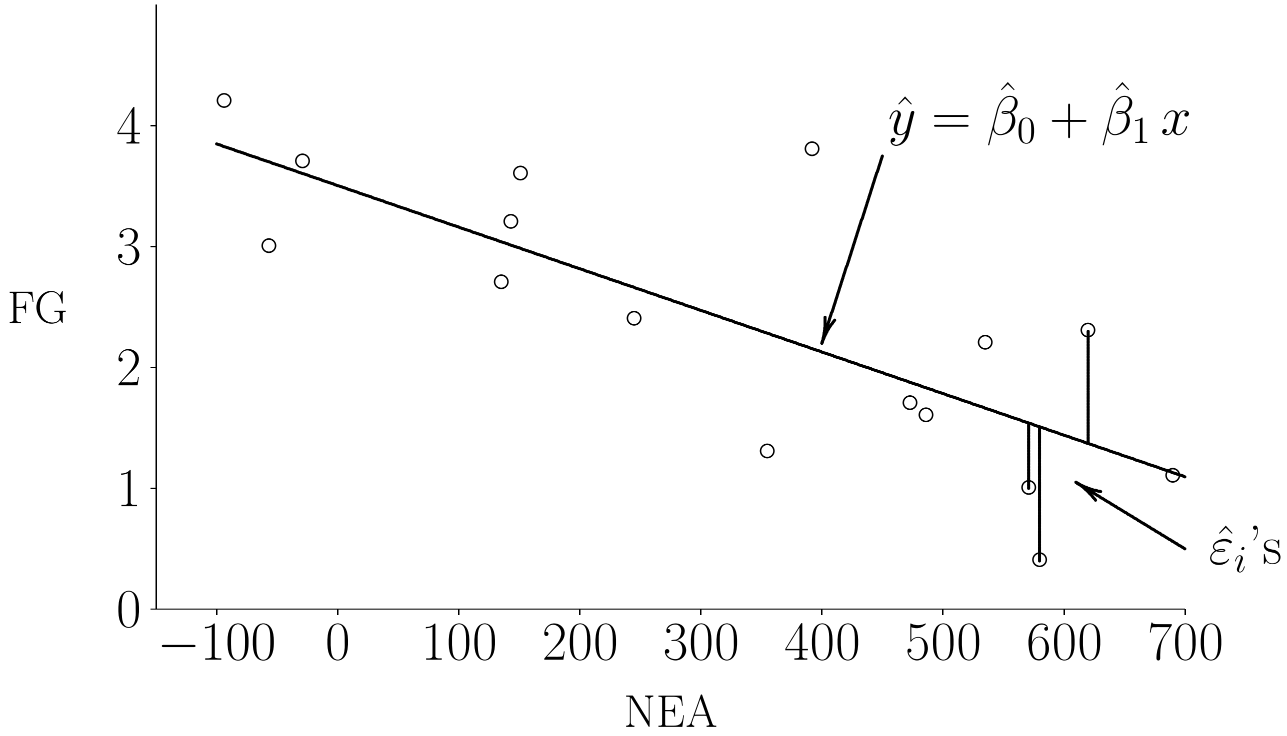
<u>Idea</u>: "best" line minimizes the sum of squared errors

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

<u>Motivations</u>:

- 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

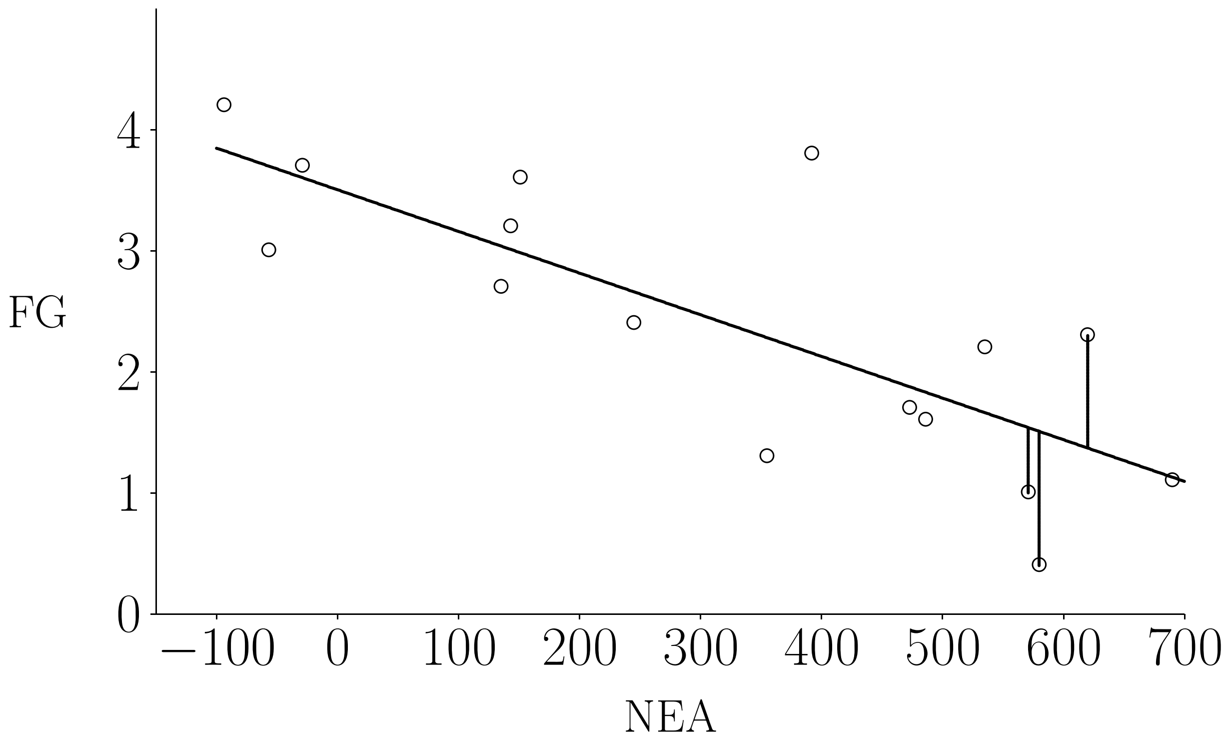


Parameter estimates:

- slope: $\hat{\beta}_1 = \frac{\sum_i (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = r s_y / s_x$, ($r$ = 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:

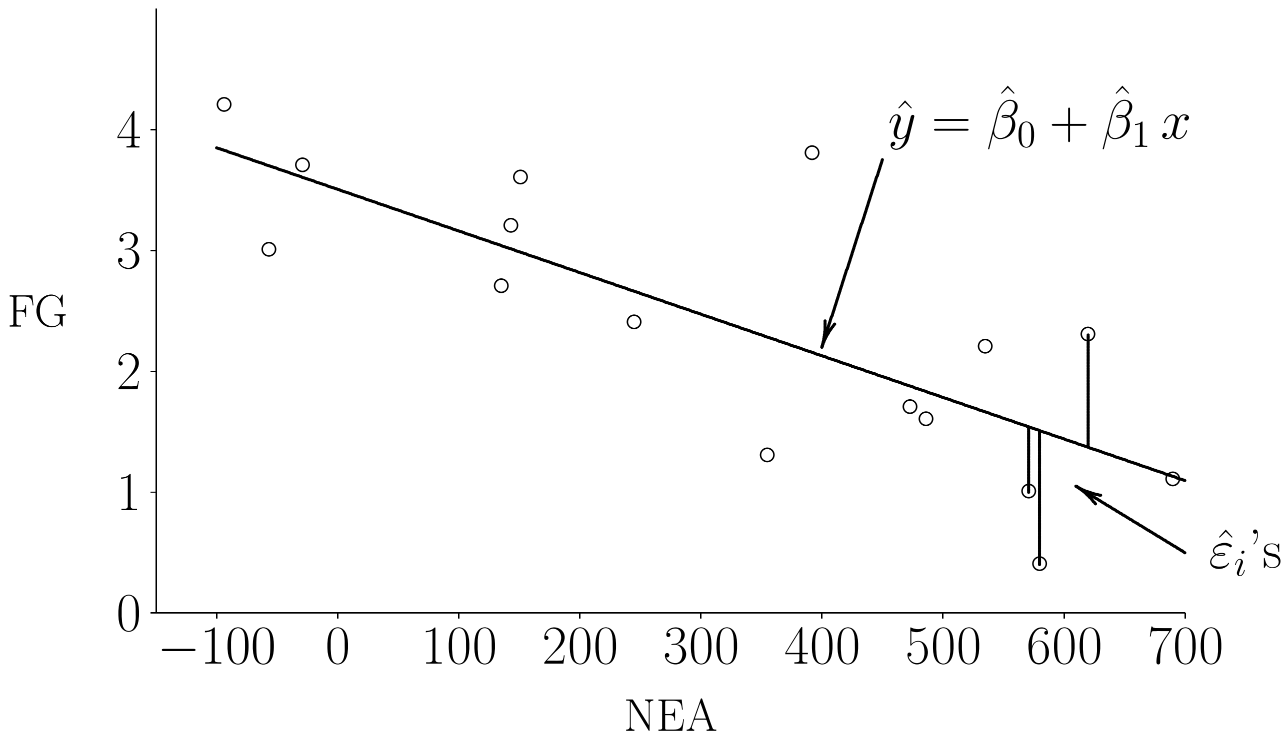
- degrees of freedom for s^2: $\text{DFE} = n-2$,
- example: test of slope equal to known value:
  - ∗ H_0: $\beta_1 = b$, (b known, fixed value),
  - ∗ H_a: $\beta_1 \neq b$ (two-sided alternative),
  - ∗ test: $t = (\hat{\beta}_1 - b)/\text{SE}(\hat{\beta}_1) \sim t(\text{DFE})$-dist. under H_0,

  most common example is $b = 0$ (horizontal line ∼ no linear relation between x and y).

Confidence intervals: also the "usual way", for example,

95% CI: estimate $\pm\, t^* \cdot \text{SE}(\text{estimate})$, $t^* = t_{.975}(\text{DFE})$.

ANOVA Table for Linear Regression

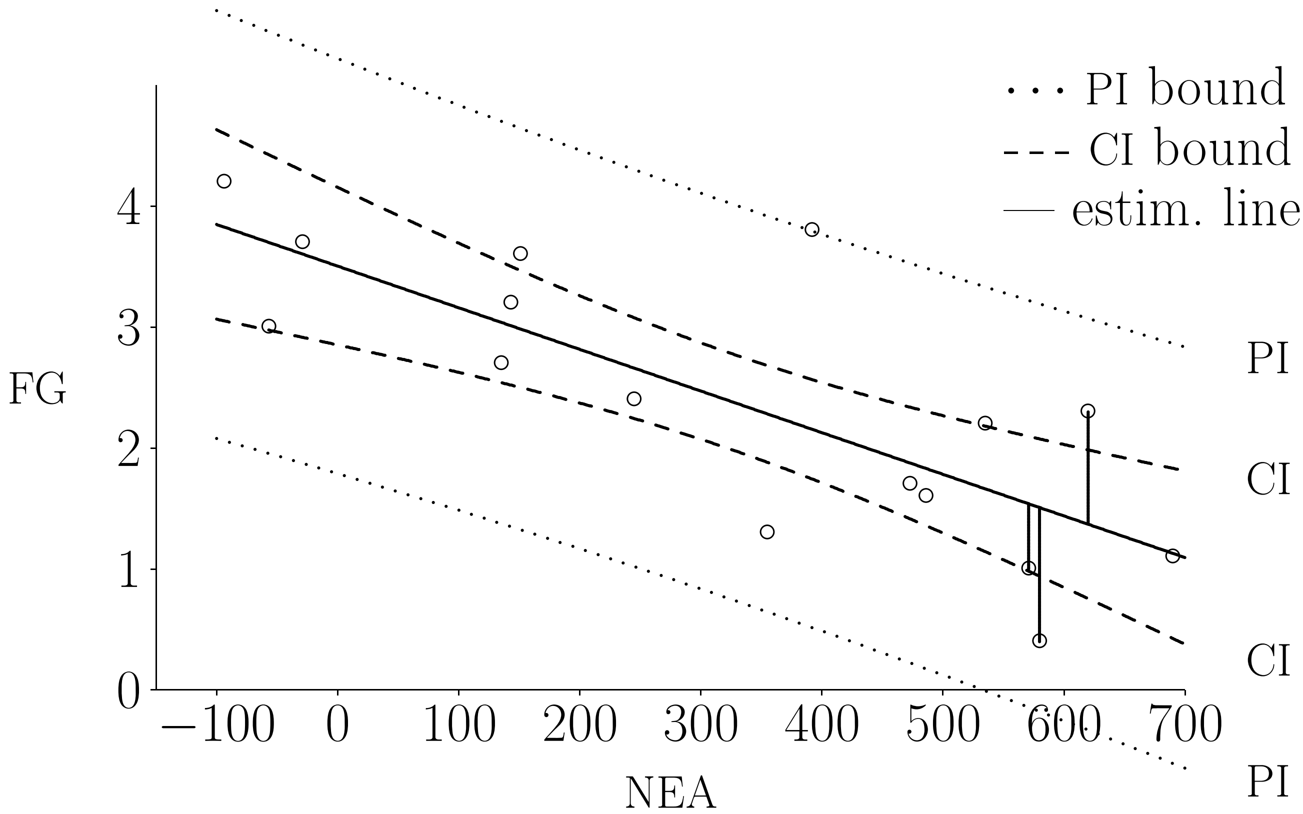


Hypothesis H_0: $\beta_1 = 0$ can also be tested by ANOVA:

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 = $\sum_i \hat{\varepsilon}_i^2$	MSE = SSE/DFE	
Total	DFT = $n-1$	SST		

- estimated error variance = s^2 = 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

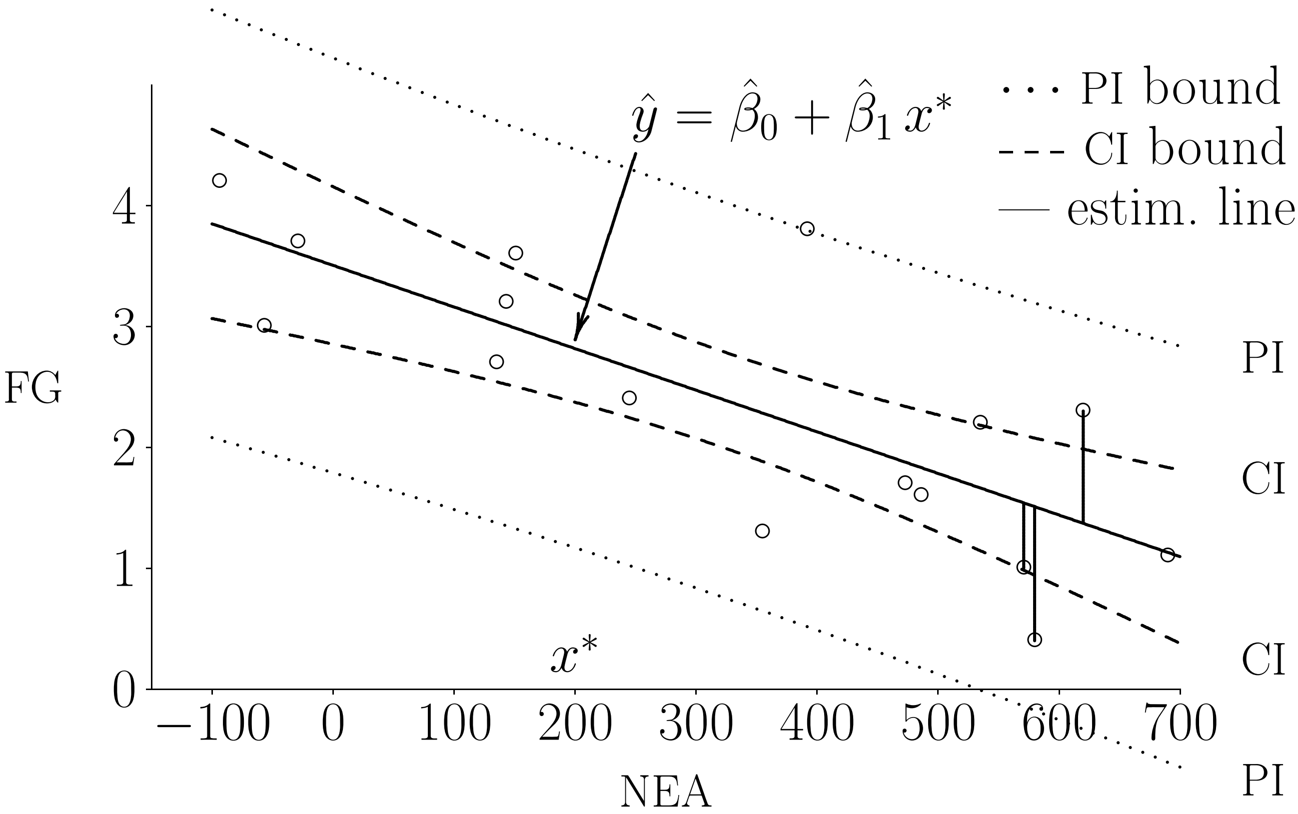


Prediction / Estimation:[7]

- 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.

[7] Stata terminology: prediction ∼ estimation, forecasting ∼ prediction.

Standard Errors in Linear Regression



Formulae for standard errors[8] in linear regression:

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

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.

[8] 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 exist[9], 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

- recent: Lang & Altman, 2013 (at EQUATOR network),
- older (still useful): Bailar & Mostellar, 1988 (next page).

[9] See links at the VHM 801 homepage, `people.upei.ca/hstryhn/vhm801`.

Sample Reporting Guidelines

Older guidelines for medical journals[10]:

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.

[10] Bailar, J.C. & Mostellar, F. (1988), *Ann. Intern Med.* **108**, 266−73.