

Index of 10-L

| Page | Title |
|------|--|
| 1 | Practical information |
| 2 | Scatterplots and data example |
| 3 | Linear regression: Data + problem |
| 4 | Linear relation |
| 5 | Exercises 2.31 and 2.32 |
| 6 | Linear regression model |
| 7 | Least squares estimation |
| 8 | Parameter estimates |
| 9 | Confidence intervals and tests |
| 10 | Prediction |
| 11 | Tools for model checking: residuals |
| 12 | Model checking in regression/ANOVA |
| 13 | Correlation |
| 14 | Statistical inference for correlation |
| 15 | Correlation vs. regression |
| 16 | Correlation II |
| 17 | Last comments about correlation and regression |
| 18 | Summary notes |
| 19 | Appendix: Schematic residual plots (versus fitted) |

PRACTICAL INFORMATION

Schedule:

- third **home assignment** ongoing (until Sunday), and **last home assignment** will be on the website by Monday,
- links for skipped lecture 11 added to Moodle site,
- **note** follow-up from last week: added **flowchart** for steps after the ANOVA test at the media page.

Today's lecture — **regression** and **correlation**:

- simple linear **regression**, including prediction and residuals,¹
 - * detailed discussion (beyond textbook coverage) of model checking using residuals based on Minitab demonstrations,
 - * “warnings” and extensions (supplementary notes: Moodle), but **entirely** skip extra topics (IPS: scatterplot smoothers, nonlinear regression),
- correlation coefficient and statistical inference for correlation,²
- links between models/procedures for correlation and regression.

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

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

SCATTERPLOTS AND DATA EXAMPLE

Recap — **scatterplot**: a plot of two variables (on the same units) against each other:

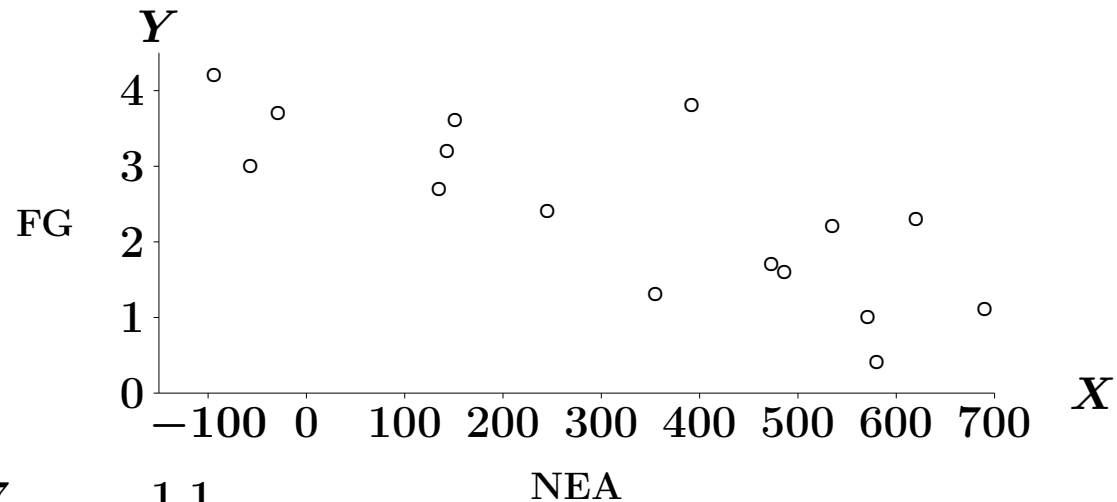
- explanatory variable (if any) goes on the horizontal axis,
- one point per observation pair (Y, X) .

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

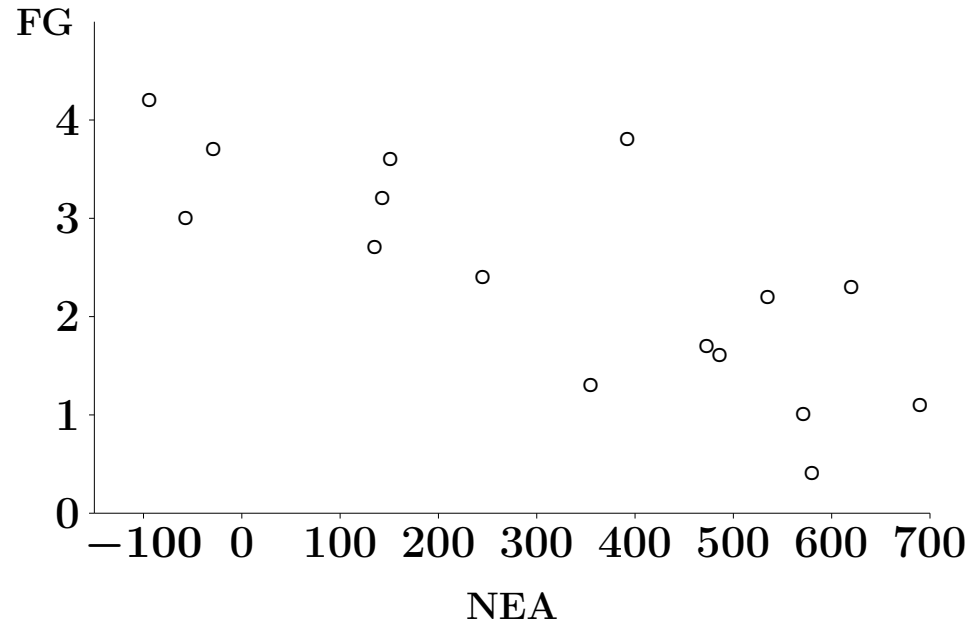
Data:

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

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 is 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/
response variable
(\Rightarrow explanatory).



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

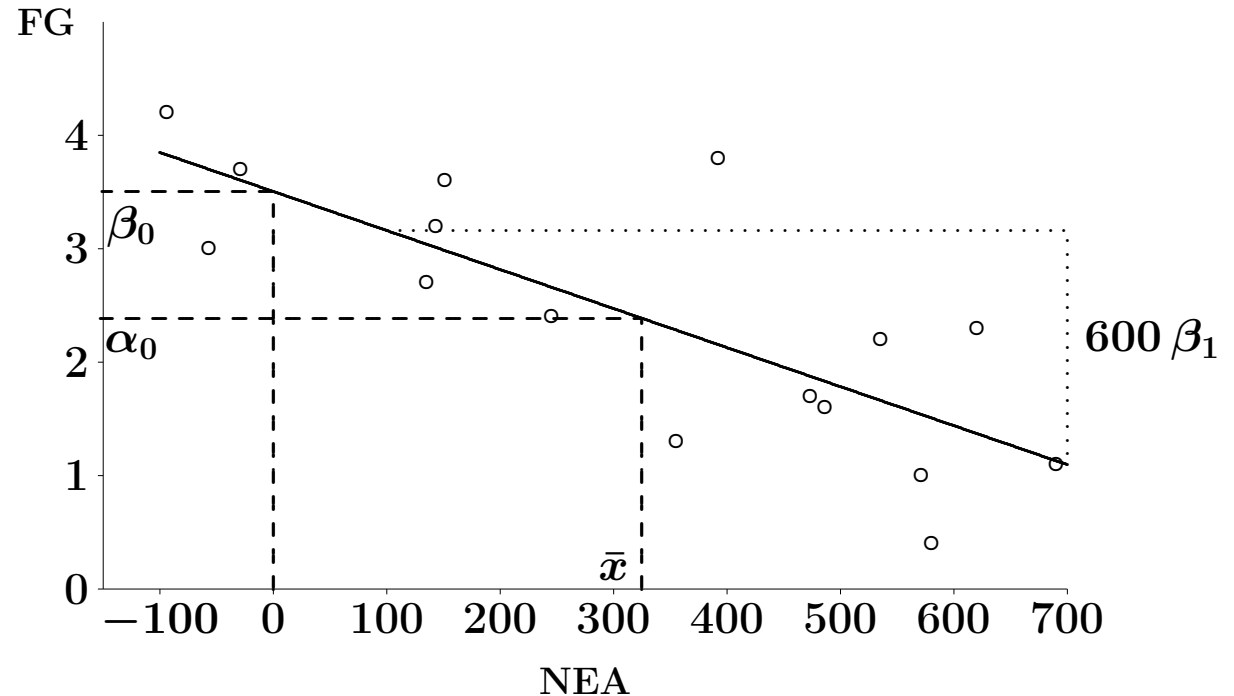
LINEAR RELATION

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 the 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:

$$y = \alpha_0 + \beta_1(x - \bar{x}), \quad \text{where}$$

- $\alpha_0 = y$ -value corresponding to $x = \bar{x}$,
- x values are “**centered**” (by subtracting \bar{x}) to avoid parameter (β_0) out of x 's range,
- **relationships**: $\beta_0 = \alpha_0 - \beta_1 \bar{x}$, or $\alpha_0 = \beta_0 + \beta_1 \bar{x}$.

EXERCISES 2.31 AND 2.32

Exercise 2.31:

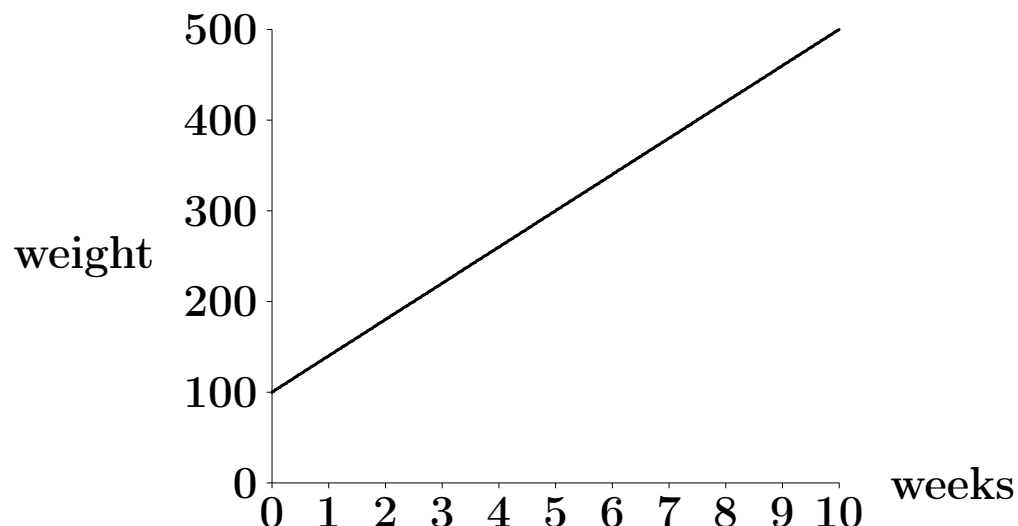
If x = number of seconds since splash and y = distance in meters, the equation is

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

Exercise 2.32:

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

(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 = $100 + 40 \cdot 104 = 4260 \text{ g}$.

LINEAR REGRESSION MODEL

Statistical model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ &= \alpha_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i, \end{aligned}$$

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:

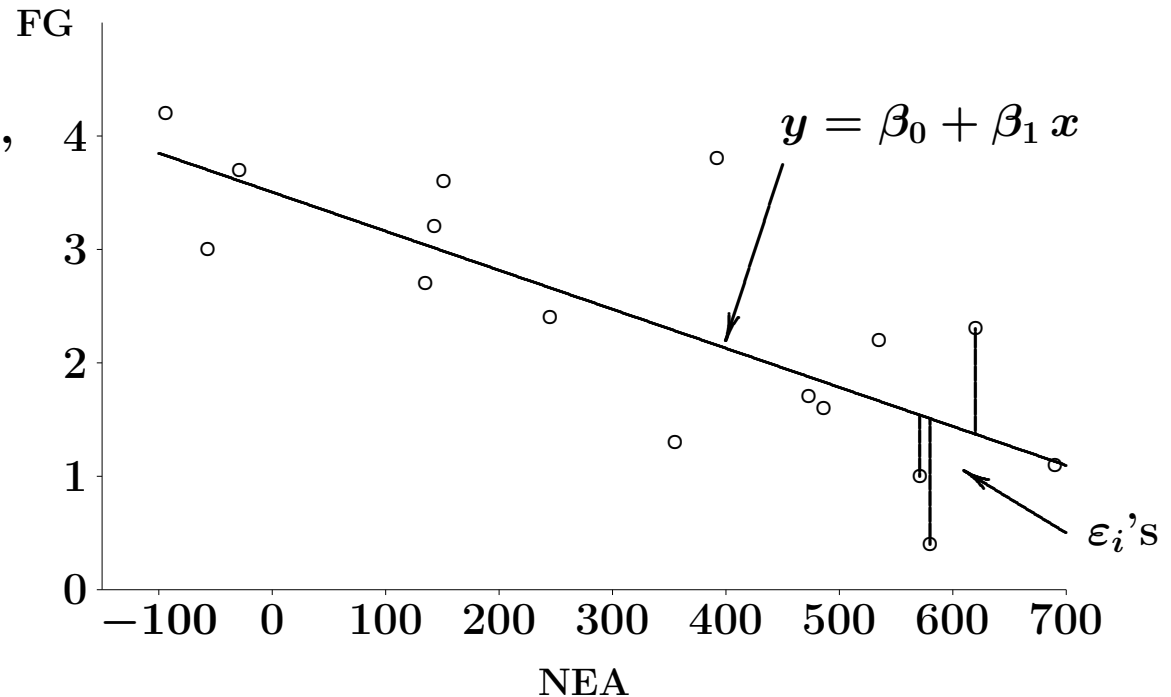
- * the linear relation: $EY_i = \beta_0 + \beta_1 x_i$,

- * normal distribution of errors ε_i ,

- * same standard deviation (or variance) of all observations (homogeneity),

- * independence of errors (and of observations),

— as for ANOVA, all assumptions can be expressed in terms of the errors (ε_i).

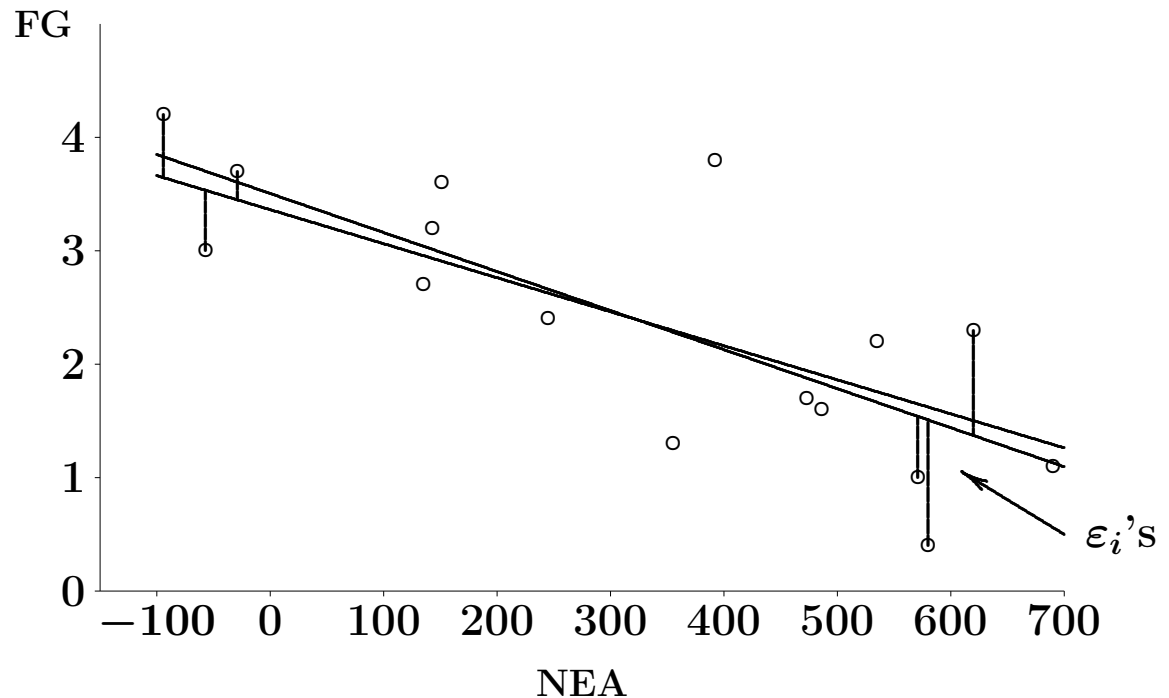


LEAST SQUARES ESTIMATION

How to determine
the regression line
(i.e., 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 this model, plus many others).

PARAMETER ESTIMATES

Parameter estimates:

- slope:⁶

$$\hat{\beta}_1 = \frac{\sum_i (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2},$$

- intercept:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

- estimated line:

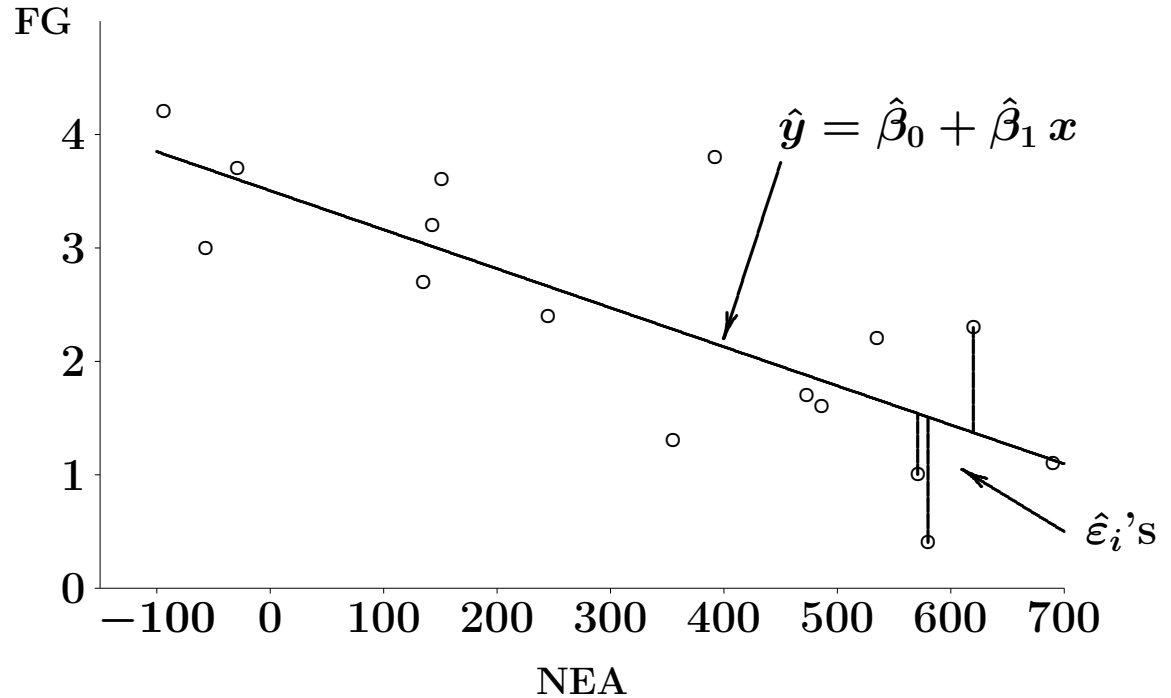
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

- $\hat{\alpha}_0 = \bar{Y}$ (\Rightarrow estimated line passes through (\bar{x}, \bar{Y})),

- residual: $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, (“observed – predicted”)

- error variance:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2} \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n-2} \sum_i \hat{\varepsilon}_i^2.$$



Minitab/Stata/R give estimates and associated standard errors⁷ for mean parameters.

⁶ The formula can also be written: $\hat{\beta}_1 = r s_y / s_x$, where r is the correlation coefficient (slide 10L–13).

⁷ Standard error formulas (not to be used for hand calculation...):

$$\text{slope : } SE(\hat{\beta}_1) = s / \sqrt{\sum_i (x_i - \bar{x})^2}; \quad \text{intercept : } SE(\hat{\beta}_0) = s \sqrt{1/n + \bar{x}^2 / \sum_i (x_i - \bar{x})^2}.$$

CONFIDENCE INTERVALS AND TESTS

Statistical inference about the parameters of the regression line follows the “usual way”, using estimates and their standard errors:

- **degrees of freedom** for s^2 : $df = n - 2$ (also denoted DFE in ANOVA table),
- **confidence intervals** by the familiar formula and using a suitable t^* -value, i.e.
 95% CI: estimate $\pm t^* \cdot \text{SE}(\text{estimate})$, $t^* = t_{.975}(\text{DFE})$,
- example: **test** of slope equal to known and fixed value (b):
 - * $H_0: \beta_1 = b$, against $H_a: \beta_1 \neq b$ (two-sided alternative), or a one-sided H_a ,
 - * **test**: $t = (\hat{\beta}_1 - b) / \text{SE}(\hat{\beta}_1) \sim t(\text{DFE})$ -distribution under H_0 ,
- alternative test of $b = 0$ (horizontal line \sim **no linear relation between x and y**) by ANOVA table:

| 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 | $s^2 = \text{MSE}$, as usual | |

- F -test **equivalent** to t -test (**same P**), because $F = t^2$,
- ANOVA table **not really needed** for simple linear regression (but for models with more x -variables).

PREDICTION

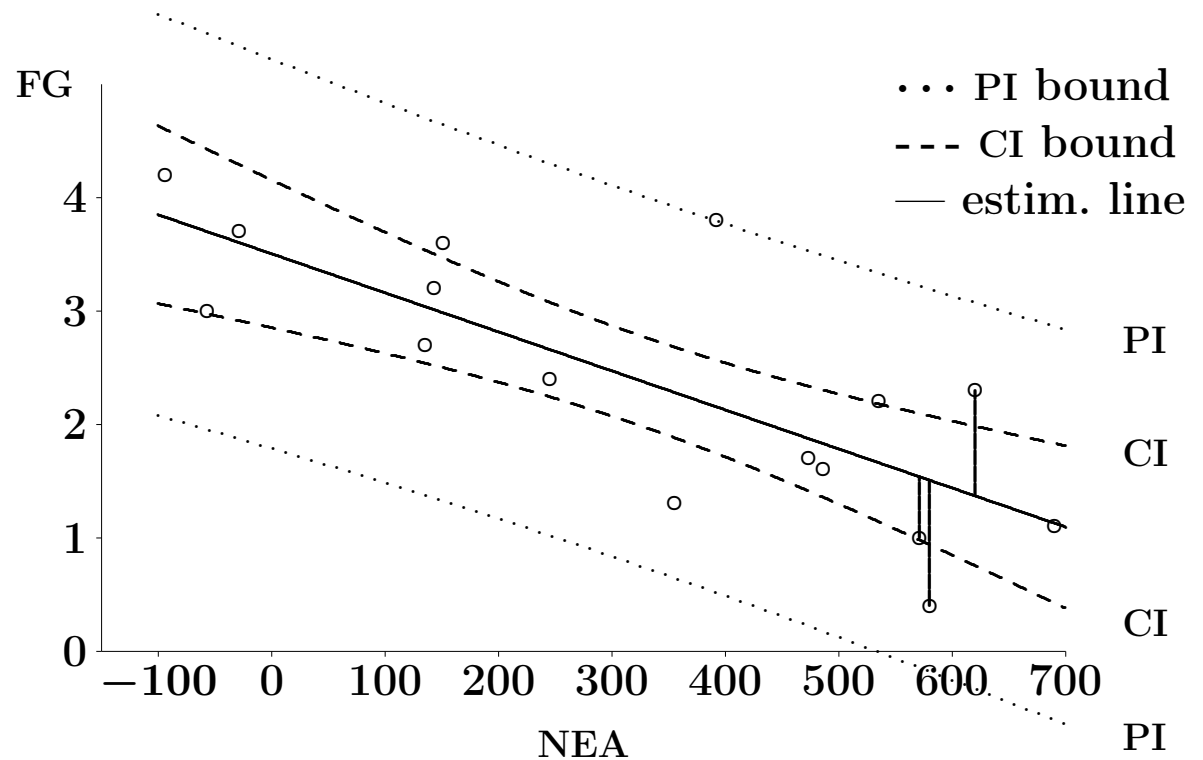
Prediction / Estimation:⁸

- 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 **PI** (**prediction interval**) 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.⁹



⁸ Stata unfortunately uses the terminology: prediction \sim estimation, forecasting \sim prediction.

⁹ Formulas for standard/prediction errors when used for confidence interval (CI) and prediction interval (PI):

$$\text{CI}(x^*) : \text{SE}(\hat{\mu}) = s \sqrt{1/n + (x^* - \bar{x})^2 / \sum_i (x_i - \bar{x})^2} ; \quad \text{PI}(x^*) : \text{SE}(\hat{y}) = s \sqrt{1 + 1/n + (x^* - \bar{x})^2 / \sum_i (x_i - \bar{x})^2}.$$

TOOLS FOR MODEL CHECKING: RESIDUALS

Residuals — our “estimates” of the random variables ε_i in the model,

- calculated as “observed – expected” (or “observed – fitted”), e.g.,
 - * linear regression: $\hat{\varepsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$; 1-way ANOVA: $\hat{\varepsilon}_{ij} = X_{ij} - \bar{X}_i$,
- always: SSE = sum of squared residuals,
- **properties of residuals** if the model is correct:
 - * normally distributed with mean 0 and a computable standard error¹⁰,
 - * residuals are **not independent**.

Other **versions of residuals**, and other variables (“regression diagnostics”):

- **standardized residuals** (i.e., divided by their standard error): $r_i = \hat{\varepsilon}_i / \text{SE}(\hat{\varepsilon}_i)$, approximately distributed as $N(0, 1)$, thus the term “standardized”,
- (**advanced**) **deletion residuals**: predicted value from model *without* current observation, also standardized and can be used for formal outlier tests (VHM 802/812),
- (**advanced**) **influence statistics**: special statistics to assess the impact of a single observation on the fitted regression line¹¹, because it may be “problematic” if estimates or conclusion depends strongly on a single or a few observation(s).¹²

¹⁰ In some (balanced) designs, the standard error is the same for all residuals, but usually it is not.

¹¹ Several different statistics (leverage, Cook’s distance, DF(F)ITS) exist, each with their specific interpretations, but it is beyond this course to go into details with them (→ VHM 802/812).

¹² Also, to assess if a particular observation is influential: **analyze the data with and without**, and compare the results.

MODEL CHECKING IN REGRESSION/ANOVA

Proposed **use of residuals** for model checking (see Appendix for examples of graphs):

- **variance homogeneity**: plot residuals ($\hat{\varepsilon}_i$ or r_i) against model's fitted values (\hat{y}_i):
— should get a noisy pattern with no “fan” shapes,
- **linear relation**: plot residuals ($\hat{\varepsilon}_i$ or r_i) against explanatory variable x_i :¹³
— should get a noisy pattern with no “parabolic” shapes,
- **outliers**: check very large or small values of standardized residuals (r_i):
— extreme r_i -values can be assessed (approximately) in $N(0, 1)$:
 - * values outside $(-2, 2)$ “suspect” in a small dataset,
 - * values outside $(-3.5, 3.5)$ “suspect” in moderate-sized dataset,
- **normal distribution**: normal probability plot of standardized residuals (r_i),¹⁴
- **data errors**: plot residuals ($\hat{\varepsilon}_i$ or r_i) against data order (if applicable).

“Unusual observations” (in Minitab listing):

- standardized residuals beyond $(-2, 2)$ (indicated with R),
- high **leverage** values (indicated with X): extreme among the (x_i) values, so the observation is **potentially influential**.

¹³ In simple linear regression, plots of the residuals against x_i and \hat{y}_i are practically the same (so one of them will do).

¹⁴ Note that P -values for normality tests only apply approximately to residuals, of any type, because of their lack of independence (preceding slide).

CORRELATION

Correlation ρ = a parameter/property of a two-dimensional, continuous distribution (simultaneous distribution of two quantitative variables), expressing the strength and direction of linear association between them in their population.

Sample (Pearson) correlation coefficient: $r = \frac{1}{n-1} \sum_i \left(\frac{X_i - \bar{X}}{s_x} \right) \left(\frac{Y_i - \bar{Y}}{s_y} \right)$
= a descriptive statistic for a sample of pairs of variables (quantitative, response variables), and an estimate of the population correlation ρ : $\hat{\rho} = r$.

Properties of correlations (both types of correlations, but displayed in terms of r):

- $-1 \leq r \leq 1$, with: $\left. \begin{array}{l} r > 0 \\ r = 0 \\ r < 0 \end{array} \right\} \sim \left\{ \begin{array}{l} \text{positive} \\ \text{no} \\ \text{negative} \end{array} \right\}$ linear association,
- $r = -1$ and $r = 1$ correspond to perfect linear association (all points on a straight non-horizontal and non-vertical line),
- correlation between X and Y is same as between Y and X ,
- r defined from standardized variables \Rightarrow unaffected by changes in mean or std. dev.,
- X and Y independent variables $\Rightarrow \rho = 0$ (and $r \approx 0$),
- extended variance addition formulae (IPS Section 4.4):

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2\rho \text{sd}(X)\text{sd}(Y).$$

STATISTICAL INFERENCE FOR CORRELATION

Definition: A pair of variables (X, Y) has a **joint normal distribution** $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$, if

$X \sim N(\mu_x, \sigma_x)$, and $Y \sim N(\mu_y, \sigma_y)$, and the correlation between X and Y is ρ .

Statistical inference for correlation:

- feasible only based on an i.i.d. sample (or SRS) $(X_1, Y_1), \dots, (X_n, Y_n)$ from the joint normal distribution $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$,
- **in this model** (only!): $\rho = 0 \Rightarrow X$ and Y independent, and regressions of Y on X (or reversely X on Y) have slope 0,
- **hypothesis H_0** : $\rho = 0$ can be tested against a one- or two-sided alternative H_a by the t -statistic,

$$t = r \sqrt{\frac{n-2}{1-r^2}}, \quad \text{where } t \sim t(n-2) \text{ under } H_0,$$

which is **exactly the same!!** as the t -test for slope = 0 in any of the two regressions,

- additional **inference for r** , such as CIs and tests for $\rho = \text{known value}$: less easy to calculate, and not accessible in all statistical software,¹⁵
- **nonparametric** correlation: **Spearman's** rank correlation coefficient (i.e., r computed for ranks) \rightarrow lab problem.

¹⁵ Minitab 19+ provides an (approximate) confidence interval, apparently computed by the Fisher ζ (zeta) transformation method.

CORRELATION VS. REGRESSION

Correlation and least-squares regression are very closely related:

- r = slope of least-squares regression line when both variables measured in standardized units ($\hat{\beta}_1 = r s_y / s_x$),
- test for $\rho = 0$ in jointly normal model is same as test for slope = 0 in the two conditional regressions,
- in the ANOVA table for linear regression: $r^2 = \text{SSM} / \text{SST} \Rightarrow r^2$ interpretable as the **the proportion of variation explained by the regression**, out of the total variation:
 - * r^2 large means good **predictive power** of the model,
 - * r^2 large does **not necessarily mean a good model**,¹⁶
 - * r^2 (usually denoted R^2) is widely misused to indicate the model's “quality”.

How to choose between correlation and regression?

- which **model assumption** is more reasonable: normal distribution for pairs (X, Y) (correlation)?, or normal distribution for errors in linear regression?
 - for example, with **only one response variable**: always regression,
- for **two response variables**: is the interest in predicting one from the other (regression)?, or primarily to measure/test their degree of linear association (correlation)?

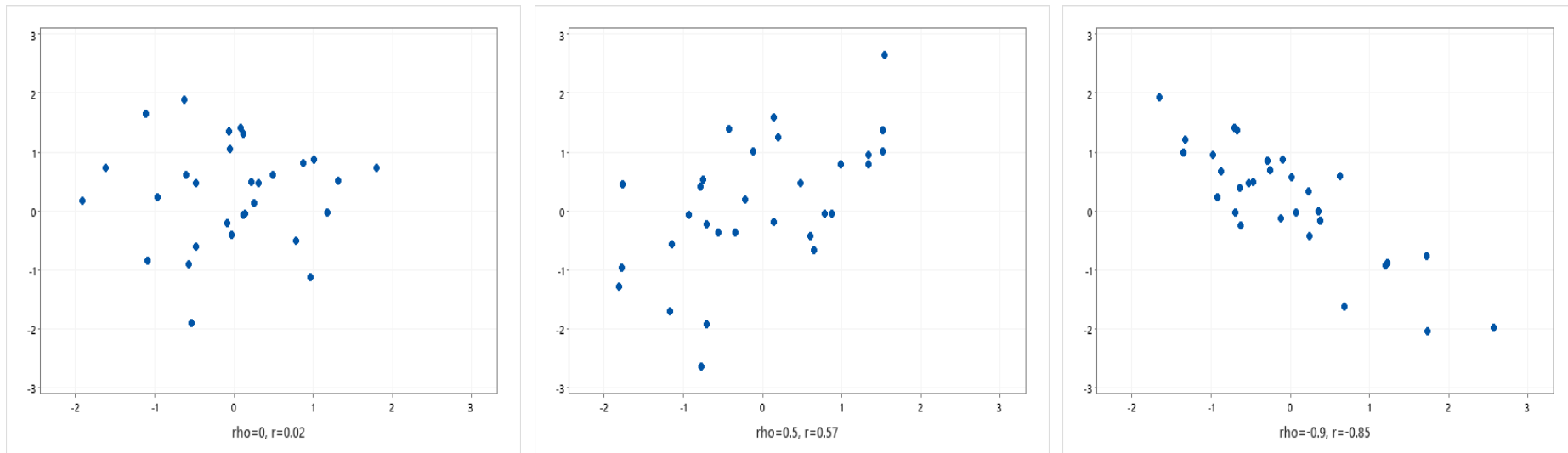
¹⁶ For an illustration, see Extra exercise 21 ([x:21](#)).

CORRELATION II

Some cautions about correlation:

- strictly speaking, only meaningful for **quantitative, response variables**,¹⁷
- only meaningful for roughly **linear associations**,
- the Pearson correlation coefficient r is **not resistant**.

Simulated patterns ($n = 30$):¹⁸



¹⁷ Some practical (but pretty **advanced**) recommendations for use and interpretation:

- can use with fixed (non-response) variable (x): interpret through regression on x ,
- can use with ordinal (\sim quantitative) variable, but strong scale assumption,
- can use with one (!) binary variable: interpret by R^2 in two-sample analysis,
- do not use with nominal variables, or with two binary variables.

¹⁸ See also PLS 4e: Figure 3.5; S: p. 173; IPS 7e: Figure 2.16.

LAST COMMENTS ABOUT CORRELATION AND REGRESSION

Some cautionary points from textbooks¹⁹:

- Linear regression and correlation are based on **linear relationships** — always check if that is reasonable; **note**: some non-linear relations can be transformed to linear ones and analyzed by a linear regression for the transformed variables ²⁰,
- Watch out for **outliers and influential observations**,
- Regression or correlation \nrightarrow **causation**,
- **Lurking variables** can distort any relationship between variables (not new, but particularly important here),
- Beware of **extrapolation** (too far outside the data range).

Variants (extensions) of linear regression:

- **multiple linear regression**: more than one x -variable in model \rightarrow next week!,
- **measurement error models**: if x is a **response var.** and measured with **error/noise**, and interest is in linear regression on **true** value of x (without error), not prediction.

¹⁹ PSLS 4e: Chapter 4; IPS 7e: Section 2.4.

²⁰ Transformation to achieve a linear relation is discussed in supplemental material for IPS 6e (from IPS6e website (discontinued), available at Moodle site); some typical and not too uncommon examples:

$$\begin{aligned}y &= a \times x^b &\longrightarrow & \log(y) = \log(a) + b \times \log(x), \\y &= a \times b^x &\longrightarrow & \log(y) = \log(a) + \log(b) \times x, \\y &= a/(1 + b \times x) &\longrightarrow & 1/y = 1/a + (b/a) \times x.\end{aligned}$$

SUMMARY NOTES

Key words and concepts for 2 quantitative (continuous) variables:

- scatterplot, response and explanatory variable, dependent (y) and independent (x) variable,
- **linear relation**: intercept, slope, prediction, extrapolation, transformation of y and/or x ,
- **linear regression model**: normally distributed, vertical errors about line, least squares estimation, standard deviation about line, t -based inference, ANOVA table, confidence and prediction intervals,
- **model checking**: residuals, standardized residuals, residual plots, outliers, variance homogeneity,
- **correlation**: population parameter/estimate (Pearson's correlation coefficient), strength of linear association, range $(-1,1)$, independence, addition formula for variances,
- **correlation model**: normal distributions, t -test for no association, links with linear regression, squared correlation (r^2 or R^2).

APPENDIX: SCHEMATIC RESIDUAL PLOTS (VERSUS FITTED)

(A) “Perfect”; (B) Fan/Cone shape; (C) Curvilinearity; (D) Missed variable

