Epidemiology/Biostats VHM 812/802 Course
Winter 2020, Atlantic Veterinary College, PEI
Henrik Stryhn

Index of Lecture 1a

INTRODUCTION TO REGRESSION

Linear regression — in a broad sense:

- defining feature: error terms $\sim$ normal distribution,
- one or several predictors (independent or x-variables),
- predictors of all types (continuous, dichotomous, ordinal, nominal)
  $\Rightarrow$ includes (e.g.) two-sample and ANOVA-type models,

— usually termed linear models[1].

Today's lecture:

- review of *simple* linear regression, expanding on model checking tools that *apply generally* to linear models,
- transformation, in particular power transformation and the Box-Cox method for selecting a transformation; also applies generally to linear models,
- computer-assisted using Stata 16,
- notation of lecture(s) mainly follows GO, e.g. variables (e.g., x and y) are not capitalized.

Textbook reading:

- VER: 14.1–3 + 14.8–10 deal with multiple regression,
- GO: model checking procedures in Chapter 6 discussed in terms of a 1-way ANOVA.

---

[1] Also general linear models (*not* generalized!), e.g. Minitab & SAS software.

DATASET DAISY2

- VER2 dataset (from VER website),[2]
- real data ∼ single cohort study involving more than 8000 cows in 42 herds,
- contributed by John Morton, Australia,
- purpose of study: evaluate the effect of various diseases on milk yield and reproductive performance,
- focus here (and in entire VER Ch. 14) on subdataset (`daisy2red`) from 7 herds with high rates of reproductive diseases (1574 lactations from 1446 cows):

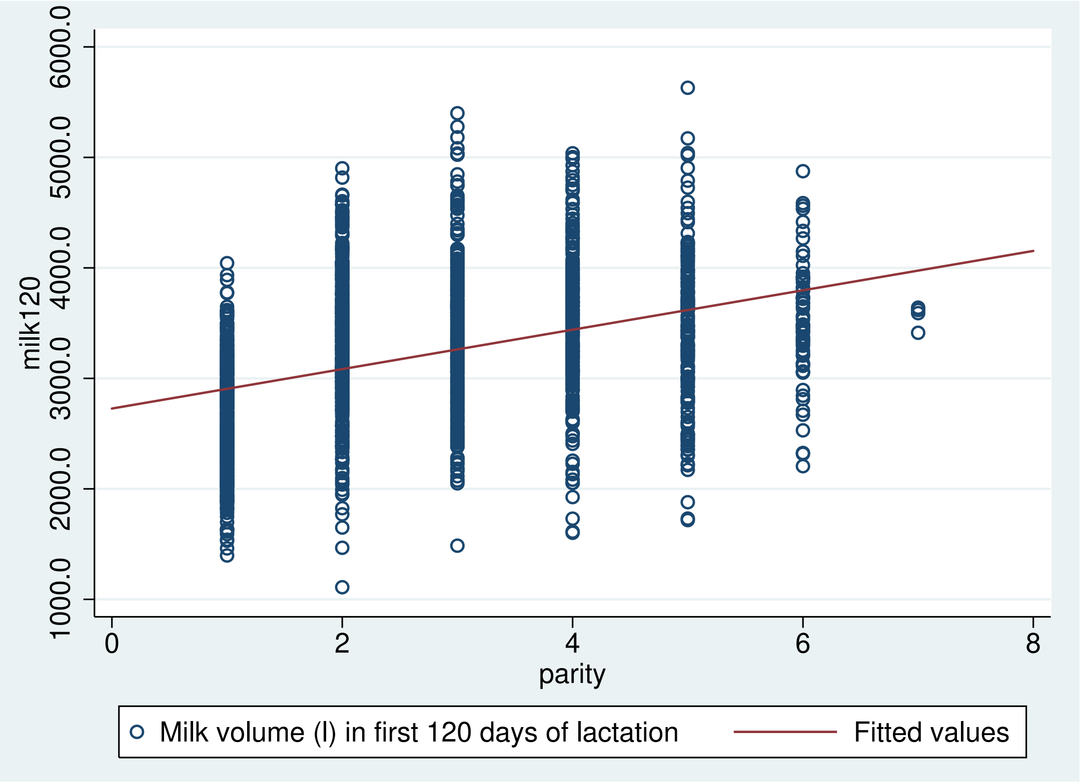
Variable	Description	Values
herd	herd number	(nominal)
cow	cow number	(nominal)
parity	lactation number	1−7
milk120*	milk volume in first 120 days of lact.	1110−5630 l
wpc	wait period to conception interval	1−298 days
twin	twin birth?	0/1
dyst	dystocia at calving?	0/1
vag_disch	vaginal discharge observed?	0/1
rp	retained placenta at calving?	0/1
herd_size	herd size	125−333
calv_dt	calving date	(date)

* 38 missing values

---

[2] Datasets are available at VHM 802 Exercises page (multiple formats).

SIMPLE LINEAR REGRESSION – MODEL



<ins>Statistical model</ins>:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, 1536 \sim \text{ lactations},$$

where the errors $\varepsilon_1, \ldots, \varepsilon_{1536}$ are i.i.d.[3] and $\sim \mathrm{N}(0, \sigma^2)$.

- β_1 = slope (1 unit increase in x corresponds to β_1 units change in y),
- β_0 = intercept (value at $x = 0$),
- σ = stand. deviation ("dispersion") about the line,
- ε_i = (vertical) error for i^{th} observation.

---

[3] i.i.d. = independent and identically distributed.

SIMPLE LINEAR REGRESSION – ANALYSIS

Least squares estimation:

- idea: "best" line minimizes the sum of squared errors
$$\sum_i \varepsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2,$$
- $\hat{\beta}_1$ and $\hat{\beta}_0$ unbiased and "optimal" under certain model assumptions,
- easy calculation formulae (simple regression only!),
$$\begin{aligned} \hat{\beta}_1 &= r s_y / s_x, \text{ where } r = \text{correlation betw. } x \text{ and } y \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{(estimated line).} \end{aligned}$$

Statistical inference about regression parameters:
the 4-step procedure, with t(DFE) as reference distribution.

ANOVA table for simple linear regression:

Source of variation	DF: Degrees of freedom	SS: Sum of squares	MS: Mean square	F
Reg. model	DFM = 1	SSM	MSM = SSM/1	MSM/MSE
Error/Resid.	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 (same P) to t-test for $\beta_1 = 0$: $F = t^2$,
- r^2 = SSM/SST, coefficient of determination, or proportion of variation explained (often denoted R^2).

4-STEP APPROACH TO TESTS AND CIs

- Data: $y_1, \ldots, y_n$,

1) Statistical model containing a (mean) parameter β,

2) Estimate $\hat{\beta}$ for β, based on $y_1, \ldots, y_n$.

3) Standard error $\mathrm{SE}(\hat{\beta})$, either

   ∗ estimated from the data, or

   ∗ known value (rarely realistic in practice),

   Note: in normal models we often have

$$\mathrm{SE}(\hat{\beta}) = A\,\sigma \text{ (or, } \mathrm{Var}(\hat{\mu}) = A^2\sigma^2),$$

   where σ is the standard deviation in the model, and A is a constant determined by the form of $\hat{\beta}$,

4) Reference distribution of $(\hat{\beta} - \beta)/\mathrm{SE}(\hat{\beta})$,
   Note: in normal models with estimated $\mathrm{SE}(\hat{\beta})$ the reference distribution is usually a $t(\mathrm{DFE})$-distribution (otherwise usually the standard normal (z) distribution $\mathrm{N}(0, 1)$),

- Confidence interval $(1-\alpha)$ for β : $\hat{\beta} \pm t^* \,\mathrm{SE}(\hat{\beta})$,[4]
- Test of H_0: $\beta = \beta^*$ against H_a: $\beta \neq \beta^*$, (β^* known value)

$$\text{test statistic} \qquad t = \frac{\hat{\beta} - \beta^*}{\mathrm{SE}(\hat{\beta})},$$

$$P\text{-value} \qquad P = 2 \times \mathrm{P}(t \geq |t_{\mathrm{obs}}|),$$

where $t \sim$ the reference distribution.

---

[4] In VHM 801 notation, $t^* = t_{1-\alpha/2} = t_{1-\alpha/2}(\mathrm{DFE})$, the $(1-\frac{\alpha}{2})$-percentile of $t(\mathrm{DFE})$.

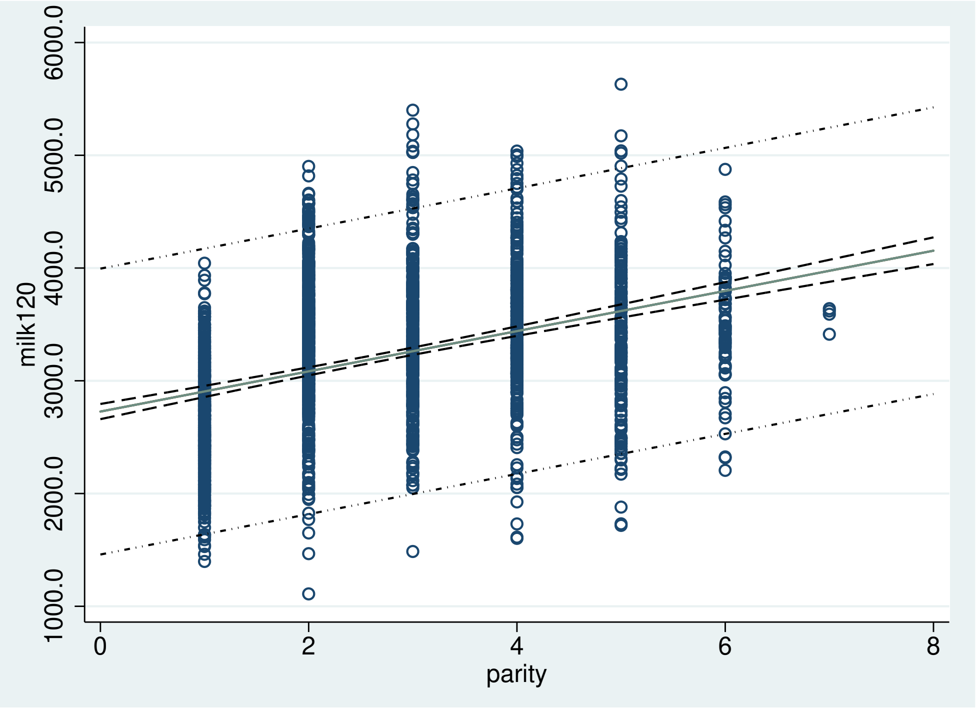
PREDICTION

Objective: give value and interval range (with confidence level $1-\alpha$) for a *new observation* with x-value x^*,

- predicted value (point on line): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$,
- prediction interval (PI) wider than confidence interval[5] (CI) for point on the line, because it involves two types of variability:
  - $*$ SE($\hat{y}$) $\sim$ uncertainty in $\hat{\beta}$'s (which are not exactly equal to true β's),
  - $*$ $\sigma^2 \sim$ variability of the new observation itself,

  in a formula: prediction error $= \sqrt{(\mathrm{SE}(\hat{y}))^2 + \mathrm{MSE}}$,

  – use instead of SE($\hat{y}$) in 4-step approach to CIs.



---

[5] Stata terminology: prediction $\sim$ estimation, forecasting $\sim$ prediction.

MODEL ASSUMPTIONS

Statistical assumptions:

- the linear relation: $\mathrm{E}y_i = \beta_0 + \beta_1 x_i$, or $\mathrm{E}\varepsilon_i = 0$, [6]
- normal distribution of errors [7] [8]: $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$,
- same variance (and standard deviation) of all errors (and observations[8]) – variance homogeneity or homoscedasticity, as opposed to heteroscedasticity,
- independence of errors (and of observations),
- x's considered fixed (measured without error, e.g. because controlled by experimenter);

  If x is an observed (response) variable,

  - the regression model is valid for *prediction* using observed x-values,
  - accounting for variability in x's requires a measurement error model (advanced).[9]

---

[6] Two types of linearity exist: in x and in the parameters (β_0, β_1); the former is relevant for model checking, the latter defines the class of "linear models". For example, the equation, $\mathrm{E}y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$, defines a linear model but is not linear in x.

[7] Strictly speaking, normality for the *residuals* (next slide) is a consequence of the model, not an assumption.

[8] If the errors (ε_i) are normally distributed with the same variance, then the same will be the case for the observations (y_i); however, this is of no use for model checking because their means $(\mathrm{E}y_i)$ differ; *checking normality of* (y_i) *is pointless!!*

[9] (technical) It is true generally that the regression model estimates are biased towards the null for the true regression equation parameters, with a bias proportional to the variability in the x's.

RESIDUALS

Overview: Residuals are estimates of the (unknown) errors and comprise the most useful tool for model checking, both for individual observations and overall; this is because the model assumptions are expressed through the errors (being i.i.d. and $\sim \mathrm{N}(0, \sigma^2)$).

- Raw/Simple residuals defined as:

  $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ ("observed – expected"),

  properties (if model correct):
  normally distributed but *not* independent, with

  - mean 0, that is: $\mathrm{E}\,\hat{\varepsilon}_i = 0$,
  - computable variance, only constant in special cases,

- Standardised residuals:[10]

  $$r_i = \hat{\varepsilon}_i / \mathrm{SE}(\hat{\varepsilon}_i) \;\approx\; \mathrm{N}(0,1) \;\text{ (if model correct)},$$

  more powerful than raw residuals and with direct interpretation,

  - 95% of values expected between -2 and 2,
  - values outside $\pm\, 3.5$ rare in moderately-sized dataset,
  - values outside $\pm\, 5$ (almost) always suspect.

---

[10] The term "studentized residuals" is also used but this often leads to confusion because some sources further distinguish between two types of studentized residuals: internally studentized residuals (= standardised residuals), and externally studentized residuals (= deletion residuals, next slide).

DELETION RESIDUALS

3-step calculation of deletion residual[11] d_i (for obs. i):

- compute fitted value $\tilde{y}_i$ for obs. i based on estimated model for all observations *excluding* observation i (idea: eliminate influence of obs. i on estimates),
- compute residual: $y_i - \tilde{y}_i$,
- standardise residual by dividing by its standard error (also based on model without obs. i).

Interpretation and use:

- for extreme obs., d_i usually somewhat more extreme than r_i (difference can be large, espec. in small datasets),
- can be used for outlier test[12]:
  - ∗ test statistic $= d_i$ (as provided by software),
  - ∗ reference distribution $= t(\text{DFE}-1)$,
    (in which one computes the tail probability),
  - ∗ unless strong "external suspicion" exists that obs. i is outlying[13], one should apply a Bonferroni correction for examining all observations as possible outliers:
    − change signif. level to $0.05/n$, or multiply P by n, where $n =$ number of observations (including i).

---

[11] In Stata, unfortunately termed studentized residuals (see previous footnote).
[12] Null hypothesis H_0: obs. i is in agreement with model from rest of the data.
[13] The suspicion must not be based on the observed value y_i.

ASSESSMENT OF NORMALITY AND LINEARITY

Normality is usually assessed by the standardised residuals:

- graphically: normal (quantile) plot/histogram for r_i's,
- descriptively: compute skewness and kurtosis for r_i's,
- formally using a statistical test for normality: should not be interpreted too rigidly, because
  - the residuals are not independent (one of the assumptions behind all normality tests)
  - the deviations from normality may be statistically significant (in a large data set) but of little importance for the statistical analysis.

Linearity or lack of fit (inadequacy of mean part of model) may also assessed by the standardised residuals:

- graphically: plot r_i's vs. x_i's and look for patterns deviating from horizontal line (maybe using lowess smoother),
- standard residual plot: plot r_i's vs. $\hat{y}_i$'s and look for any patterns beyond noise in a horizontal band which might be associated with missing predictors,[14]
- further graphical exploration: plot r_i's against any other variables of interest that might be related to the outcome (e.g., observation order).

---

[14] In simple linear regression, this plot contains the same information as the plot against the x_i's.

ASSESSMENT OF HOMOSCEDASTICITY

First thing to do:
plot standardised residuals against fitted values (or predictors), and look for cone/fan shape indicating residuals to be more variable at one end of the scale than the other.

Descriptive statistics: compute means and standard deviations of r_i's across groups defined in any "interesting" way.

Test for H_0: homoscedasticity? — no problem, many tests exist. . . ,

- no overall best test (to my knowledge),
- the test may be more sensitive to model deviations than the least squares regression itself (which is "somewhat robust"),
- testing for homoscedasticity seems to be most popular in econometrics (and they spell it with a "k"). . . ,
- some commonly used tests for ungrouped[15] ("regression") models (available in Stata):
  - Breusch-Pagan/Cook-Weisberg test (`hettest`),
    White's test (`imtest`),
- personal view: use these as "descriptive statistics" contributing to your information about the data/model, not as the ultimate truth (so don't use P-values too rigidly),
- truly robust methods exist:
  - ∗ robust standard errors (much later lecture),
  - ∗ robust regression – different statistical approach (not in course).

---

[15] For grouped ("ANOVA") models the most commonly used tests are: Levene's test (`sdtest`), Bartlett's test (`oneway`; very sensitive to model deviations).

TRANSFORMATION IN REGRESSION MODELS

Potential aims of transformation:

1) obtain linear relation,

2) deal with unequal variance (depending on the mean),

3) deal with non-normal errors.

Aims may be conflicting (suggest different transformations)
$\Rightarrow$ transformation is "an art" (and a trial and error process).

Types of transformations:

- consider only transform of y (also possible: x, or both y and x),
- power transformations: $y \mapsto y^\lambda$ for some power λ, [16]
- "standard" variance-stabilising transformations:

Data type (y)	Mean	Variance	Transformation	Power^a
measurement/conc.	$\mathrm{E}y = \mu$	$\mathrm{Var}y \propto \mu^2$	$\log(y)$ or $\ln(y)$	$\lambda = 0$
count	$\mathrm{E}y = \lambda$	$\mathrm{Var}y \propto \lambda$	$\sqrt{y}$	$\lambda = 0.5$
proportion	$\mathrm{E}y = p$	$\mathrm{Var}y \propto p(1-p)$	$\arcsin(\sqrt{y})$	n/a

^{a} within Box-Cox family of power transform.: $y \mapsto \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\ \ln(y) & \text{for } \lambda = 0. \end{cases}$

Statistical inference for transformation:

- estimation[17] of transformation power λ within Box-Cox family,
- associated CI (for λ) gives range of "plausible" values and can be used in significance testing for specific λ-values (e.g., $\mathrm{H}_0 : \lambda = 1$).[18]

---

[16] Stata uses instead the Greek letter θ (theta) to represent the power: $y \mapsto y^\theta$.

[17] (technical) Maximum likelihood estimation, by maximising the so-called (log) profile likelihood function, either by an automatic software routine (Stata, Minitab) or by manual maximisation across a grid of λ-values (S-Plus/R).

[18] (technical) Both likelihood- and Wald-methods usually give sensible results.

BOX-COX TRANSFORMATION

Applied view of Box-Cox transformation[19]:

- Box-Cox analysis (`boxcox` command in Stata) gives optimal transformation (among those considered...):
  - ∗ "optimal" ∼ make residuals fit as well as possible to a normal distribution with homogenenous variance,
  - ∗ transformation to make distribution of outcome close to normal is something else (*not recommended!*)[20],
- once optimal λ-value found, transform using simple formulae:

$$y \mapsto \begin{cases} y^{\lambda} & \text{for } \lambda > 0, \\ \ln(y) & \text{for } \lambda = 0, \\ -1/y^{|\lambda|} & \text{for } \lambda < 0, \end{cases}$$

- common practice to approximate optimal λ-value by a close "nice" value, e.g. 0.5, 0, -0.5 or -1 (to avoid too strange transformations),
- Box-Cox analysis requires all $y_i > 0$:
  - ∗ add a small value to meet requirement if only few $y_i = 0$ or $y_i < 0$,
  - ∗ Box-Cox type analysis possible (in S-Plus/R) also for transformations of the form: $y \mapsto \ln(y + \alpha)$. [20]
- redo model checks for transformed data! ("best" ≠ "good")

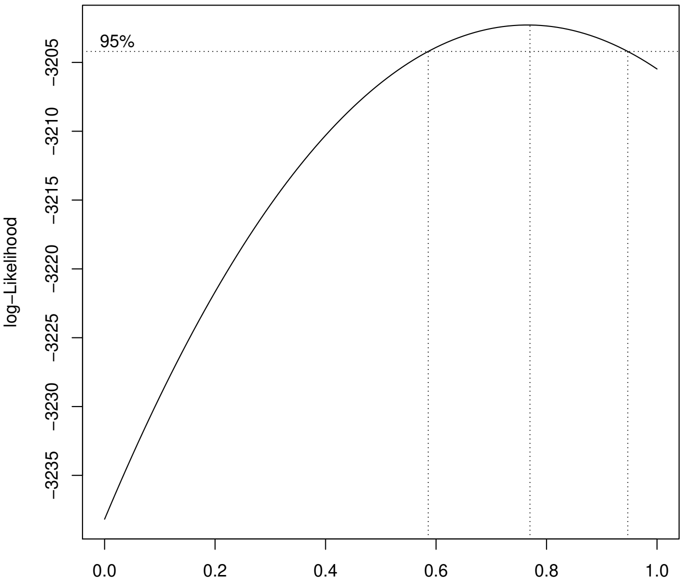
---

[19] Strictly speaking, it is the method of analysis rather than the transformation itself that carries the name of Box and Cox, after their 1964 paper in JRSS B.

[20] The Stata `ladder` and `lnskew0` commands should not be used for inference.

BOX-COX ANALYSIS FOR `DAISY2`

Profile log-likelihood for regression of `milk120` on `parity`:



Conclusion: graph shows optimal λ-value around 0.75 and a 95% CI that excludes both 0.5 and 1.[21]

Comparison of model fits at different scales:

Scale of analysis	Original	Power	Square-root
Residual statistic	$\lambda = 1$	$\lambda = 0.75$	$\lambda = 0.5$
skewness	0.129	-0.023	-0.179
kurtosis (Stata)	3.090	3.071	3.138
normality (P^1)	0.065	0.614	0.012
homoscedast. (P^2)	0.007	0.069	0.362

^{1} Shapiro-Wilk test (`swilk`); ^{2} BPCW test (`hettest`)

Conclusion: cannot achieve both perfect skewness and homoscedasticity: $\lambda = 0.75$ is a fair compromise, but model violations at $\lambda = 1$ and 0.5 hardly very serious for analysis.

---

[21] Estimation in Stata yields $\hat{\lambda} = 0.765$ with a 95% CI of $(0.585, 0.946)$.

BACKTRANSFORMATION IN REGRESSION MODELS

Main message: Results from transformed scale analysis must! (nearly always) be backtransformed to original scale.

General rules (valid for any monotonic transformation):

- backtransformed means $\sim$ medians (*not* means) at original scale,
- CIs can be backtransformed by backtransforming both endpoints,
- difficult to get means and SEs at original scale,[22]
- backtransform regression parameters (β's) only for log-transform (below); *never* backtransform their SEs.

Special procedures for log-transformation[23]; consider the model

$$\begin{aligned} \ln(y_i) &= \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{or} \\ y_i &= \mathrm{e}^{\beta_0} \cdot \mathrm{e}^{\beta_1 x_i} \cdot \mathrm{e}^{\varepsilon_i}. \end{aligned}$$

Disregarding the error terms (involving ε_i), we get the interpretations:

- $\mathrm{e}^{\beta_0} \sim$ median value at original scale for $x = 0$,
- $\mathrm{e}^{\beta_1} \sim$ *multiplicative* effect of a 1-unit increase in x;
  example: if $\beta_1 = 0.4$, then $\mathrm{e}^{\beta_1} \approx 1.49 \sim$ multiplication by 1.49, or a relative increase by 49%,
- if the x's are also on logarithmic scale (say $x = \ln(z)$), then a 1-unit increase in $x \sim$ an increase in z by a factor of $\mathrm{e}^1 = 2.72$; instead, we may consider a change in x by $\ln(2) = 0.693$, corresponding to an increase in z by a factor of 2 (i.e., a doubling);
  example: if $\beta_1 = 0.4$, then $\mathrm{e}^{\beta_1 \cdot \ln(2)} \approx 1.32 \sim$ multiplication by 1.32, or a relative increase by 32%, for a doubling in z.

---

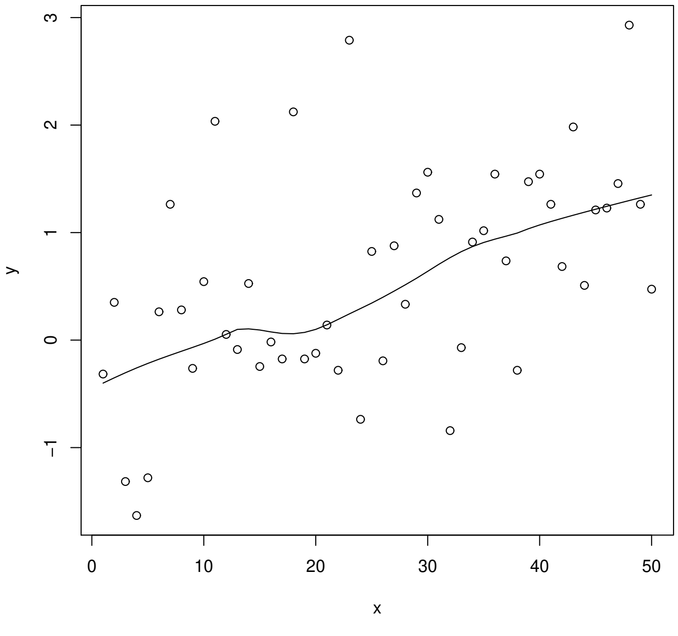
[22] Simulation approaches, beyond the scope of the course, can be used.

[23] Described here for natural log, but works also for other logarithms by replacing the exponential function by the appropriate inverse logarithmic function.

LOWESS CURVE

Applied view of `lowess` method[24]:

- = descriptive graphical tool to explore the relationship between two quantitative variables,
- – typically overlaid a scatterplot, say of y against x,
- – main role/purpose: emphasize local and/or global trends in the relation between the variables that may not be easily visible from a scatter of points (e.g. with substantial noise or a large number of points), such as:
  - ∗ linearity (versus non-linearity),
  - ∗ constant y (versus non-constant y),
- – may be applied to observed variables or to statistics derived from an analysis (e.g. residuals),
- – result depends on the order of x-values $\Rightarrow$ problems with ties among x-values,
- – example (artificial data from original paper) with R function `lowess` and settings: `f=0.5,iter=2,d=1`:



---

[24] The acronym stands for: local weighted scatterplot smoothing, and the method is usually referred to Cleveland (1979), *J. Amer. Statist. Assoc.* **74**, 829-836.

LOWESS SMOOTHING – DETAILS

Several related algorithms (e.g., `lowess`, `loess`) exist, all with different parameters to affect or finetune the result ⇒
no "correct" or "best" method available, and most important consideration for one's choice is usefulness/visual impression of the result.[25]

Key ideas/components:

- regression: fitted value $\hat{y}_i$ of ith point obtained from a (linear or polynomial) weighted least-squares regression of y on x,
- locality: main contributions to fitted value $\hat{y}_i$ from observations y_j with x_j close to x_i, achieved in two ways:
  - ∗ only a fraction f (with $0 < f \leq 1$) of the points contribute at all, e.g. $f = 0.5 \sim 50\%$ closest points (to x_i),
  - ∗ weights for contribution of y_j decrease with distance $|x_i - x_j|$,[26]
- robustness: in iterative refinements of estimates, points are further (down)weighted by the size of their residuals in the regression,
- smoothness: the parameters affecting the smoothness of the resulting curve are: fraction f ($f \uparrow \sim$ more smooth), polynomial order d ($d \uparrow$), residual weighting iterations (`iter` ↑).

Stata implementation `lowess`: some limitations and added flexibility:

- no iterative residual weighting, and only linear regression ($d{=}1$),
- `bwidth` parameter (roughly) equals the fraction f,
- weighting can be turned off and regression can be replaced by a weighted mean (not clear why one would want those. . .),
- adaptation to binary (0/1) outcome (more later in course).

---

[25] Recall that aims are descriptive and *not* involving statistical inference.

[26] The original `lowess` function (and still most common version) uses "tri-cubic" weights, $w_j = (1 - (|x_i - x_j|/\Delta)^3)^3$, where Δ is chosen to yield a range $\sim f$.