

Index of Lecture 1a: Simple linear regression and extensions

Page	Title
1	Introduction to regression
2	Dataset daisy2
3	Simple linear regression – Model
4	Simple linear regression – Analysis
5	4-step approach to tests and CIs
6	Prediction
7	Model assumptions
8	Residuals
9	Deletion residuals
10	Assessment of normality and linearity
11	Assessment of homoscedasticity
12	Transformation in regression models
13	Box-Cox transformation
14	Box-Cox analysis for daisy2red
15	Backtransformation in regression models
16	Lowess curve
17	Lowess smoothing – details

INTRODUCTION TO REGRESSION

Linear regression — in a broad sense, and usually termed **linear models**¹:

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

Today's lecture:

- review of **simple** linear regression (i.e., only one x -variable), 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 17,
- notation of lecture(s) mainly follows GO, e.g. variables (x, 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, beware of this confusion!), e.g. Minitab & SAS software.

DATASET DAISY2

- VER2 dataset (from VER website: www.upei.ca/ver),²
- real data³ ~ single cohort study involving more than 8000 cows in 42 herds,
- **purpose of study**: evaluate the effect of various diseases on milk yield and reproductive performance,
- focus here (and in entire VER Chapter 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 <i>l</i>
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 both at the VHM 812 Moodle site (.dta) and at the VHM 802 Exercises page (also .csv).

³ The data have been contributed by John Morton, Australia.

SIMPLE LINEAR REGRESSION – MODEL

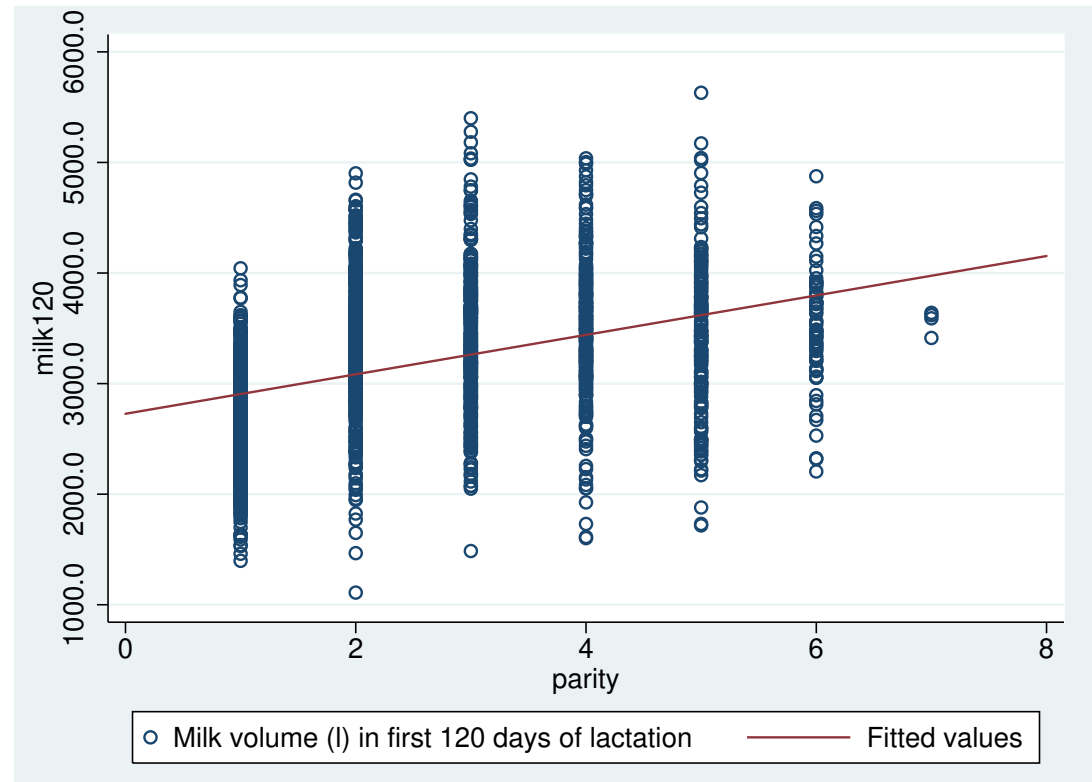
Statistical model:

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

where the errors $\varepsilon_1, \dots, \varepsilon_{1536}$ are i.i.d.⁴ and $\sim N(0, \sigma^2)$.⁵

Interpretations:

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



⁴ The abbreviation i.i.d. stands for “independent and identically distributed”.

⁵ This course uses the most common notation, where the second parameter of a normal distribution is the variance.

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!)

$$\begin{cases} \hat{\beta}_1 = r s_y / s_x, \text{ where } r = \text{correlation}(x, y) \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (\text{estimated line}). \end{cases}$$

Statistical inference about regression parameters:

the 4-step procedure, with $t(\text{DFE})$ as reference distribution.

ANOVA table for simple linear regression:

Source of variation	DF: Degrees of freedom	SS: Sum of squares	MS: Mean square	<i>F</i> : Test statistic
Regression model	DFM = 1	SSM	MSM = SSM/1	MSM/MSE
Error/Residual	DFE = $n - 2$	SSE = $\sum_i \hat{\varepsilon}_i^2$	MSE = SSE/DFE	
Total	DFT = $n - 1$	SST		

- estimated error variance = $s^2 = \text{MSE}$, as usual,
- F*-test equivalent (same *P*) to *t*-test for $\beta_1 = 0$: $F = t^2$,
- $r^2 = \text{SSM}/\text{SST}$, coefficient of determination, or prop. of variation explained (often R^2).

4-STEP APPROACH TO TESTS AND CIs

- **Data** y_1, \dots, y_n ,
- **Statistical model** containing a (mean) parameter β .
- **Estimate** $\hat{\beta}$ for β , based on y_1, \dots, y_n .
- **Standard error** $SE(\hat{\beta})$, either
 - * estimated from the data, or
 - * known value (rarely realistic in practice),

note: in normal models (with error standard deviation σ) we have

$$\text{Var}(\hat{\beta}) = A \sigma^2 \quad \text{and} \quad SE(\hat{\beta}) = \sqrt{A} \sigma,$$

where A is a constant determined by the form of $\hat{\beta}$,

- **Reference distribution** of $(\hat{\beta} - \beta^*) / SE(\hat{\beta})$,
note: in normal models with estimated $SE(\hat{\beta})$, usually the $t(\text{DFE})$ -distribution,
- **Confidence interval** $(1 - \alpha)$ for β : $\hat{\beta} \pm t^* SE(\hat{\beta})$, $(t^* = t_{1-\alpha/2} = t_{1-\alpha/2}(\text{DFE}))$
- **Test** of $H_0: \beta = \beta^*$ against $H_a: \beta \neq \beta^*$, (where $\beta^* = \text{known value}$)

$$\text{test statistic: } t = \frac{\hat{\beta} - \beta^*}{SE(\hat{\beta})}, \quad P\text{-value: } P = 2 \times P(t \geq |t_{\text{obs}}|),$$

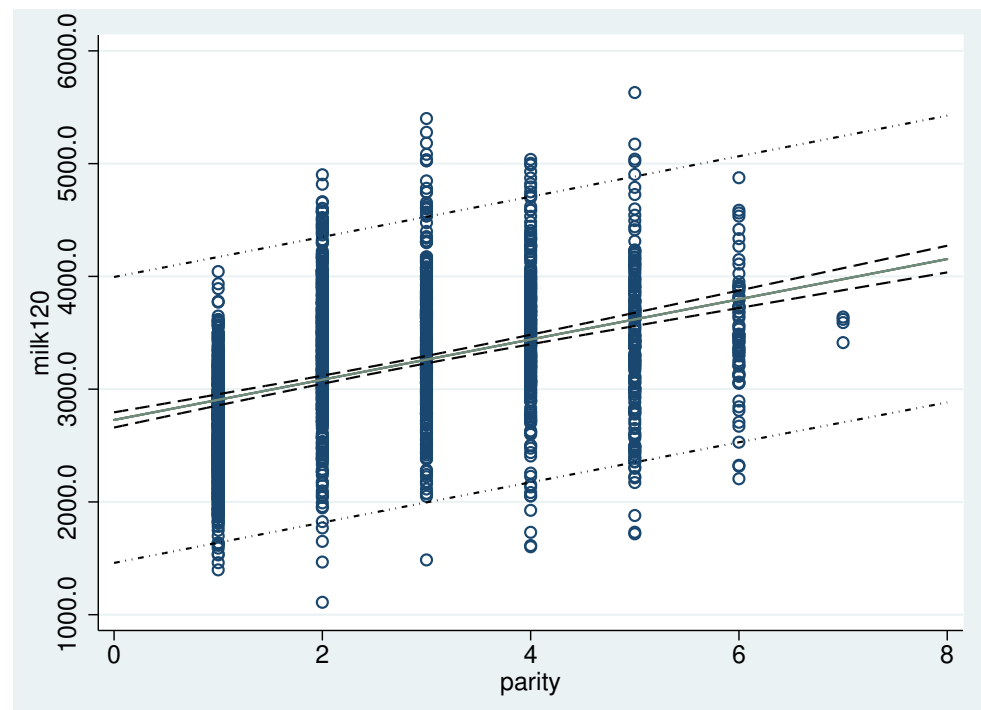
where $t \sim$ the reference distribution.

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): $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$,
- a **prediction interval** (PI) is wider than a confidence interval⁶ (CI) for estimated 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 (“about the line”),

in a **formula**: prediction error
= $\sqrt{(SE(\hat{y}))^2 + MSE}$,
to replace $SE(\hat{y})$ in the 4-step approach to CIs.



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

MODEL ASSUMPTIONS

Statistical assumptions:

- the **linear relation**⁷: $Ey_i = \beta_0 + \beta_1 x_i$, or $E\varepsilon_i = 0$,
- **normal distribution** of errors^{8 9}: $\varepsilon_i \sim N(0, \sigma^2)$,
- **same variance** (and standard deviation) of all errors (and observations⁹) – 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** based on observed x -values,
- * accounting for variability in x 's requires a measurement error model (advanced).¹⁰

⁷ 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, $Ey_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$, defines a linear model but is not linear in x .

⁸ Strictly speaking, normality for the **residuals** (next slide) is a consequence of the model, not an assumption.

⁹ 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 (Ey_i) differ; strictly speaking, **checking normality of (y_i) is pointless!!**

¹⁰ (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 $\sim N(0, \sigma^2)$).

- **Raw/Simple residuals** defined as:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad (\text{“observed} - \text{expected”}),$$

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

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

- **Standardised residuals:**¹¹

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

more powerful than raw residuals and with direct interpretation,

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

¹¹ The term “studentized residuals” is also used, but 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 residual¹² d_i (for observation 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 from model without obs. i).

Interpretation and use:

- for extreme observations, d_i usually somewhat more extreme than r_i (difference can be large, especially in small datasets),
- can be used for an **outlier test**¹³:
 - * 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 outlying¹⁴, one should apply a Bonferroni correction for examining all observations as possible outliers:
 - change significance level to $0.05/n$, or multiply P by n ,where n = number of observations (including i).

¹² In Stata, unfortunately termed studentized residuals (see previous footnote).

¹³ Null hypothesis H_0 : obs. i is in agreement with model from rest of the data.

¹⁴ 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 other summary descriptors for r_i 's,
- **formally** using a statistical test for normality: should not be interpreted too rigidly, because:
 - * the residuals are not independent (all normality tests make this assumption),
 - * 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 be 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, slide 1aL-16),
- **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,¹⁵
- **further graphical exploration**: plot r_i 's against any other variables of interest that might be related to the outcome (e.g., observation order as in Minitab).

¹⁵ 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),
- tests may be more sensitive to model deviations than the least squares regression itself,
- testing for homoscedasticity seems most popular in econometrics (but with a “k”),
- some commonly used tests for ungrouped¹⁶ (“regression”) models (available in Stata):
 - Breusch-Pagan/Cook-Weisberg test (hettest), and 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** (later lecture), and **robust regression** (not in course).

¹⁶ 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 (when variance depends 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 of y :¹⁷

- power transformations: $y \mapsto y^\lambda$ for some power λ ,¹⁸
- “standard” (variance-stabilising) transformations:

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

^a within Box-Cox family of power transformations:

$$y \mapsto \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\ \ln(y) & \text{for } \lambda = 0. \end{cases}$$

Statistical inference for transformation:

- estimation¹⁹ 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., $H_0 : \lambda = 1$).²⁰

¹⁷ One may also consider transformation of x , or of both y and x with the same transformation.

¹⁸ Stata uses instead the Greek letter θ (theta) to represent the power: $y \mapsto y^\theta$.

¹⁹ (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 (R).

²⁰ (technical) Both likelihood and Wald (based on estimate and SE) methods usually give sensible results.

BOX-COX TRANSFORMATION

Applied view of Box-Cox transformation²¹:

- Box-Cox analysis (boxcox command in Stata) gives **optimal transformation**:
 - * “optimal” ~ make **residuals** fit as well as possible to a normal distribution with homogenous variance,
 - * transformation to make distribution of outcome close to normal is something else (**not recommended!**)²²,
- once optimal λ -value found, transform using simpler formulas:
$$y \mapsto \begin{cases} y^\lambda & \text{for } \lambda > 0, \\ \ln(y) & \text{for } \lambda = 0, \\ -1/y^{|\lambda|} & \text{for } \lambda < 0, \end{cases}$$
- 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 the requirement, if only few $y_i = 0$ or $y_i < 0$,
 - * Box-Cox type analysis possible (e.g., in R) also for transformations of the form:
 $y \mapsto \ln(y + \alpha)$.²²
- **always** redo model checks for transformed data! (“best” \neq “good”)

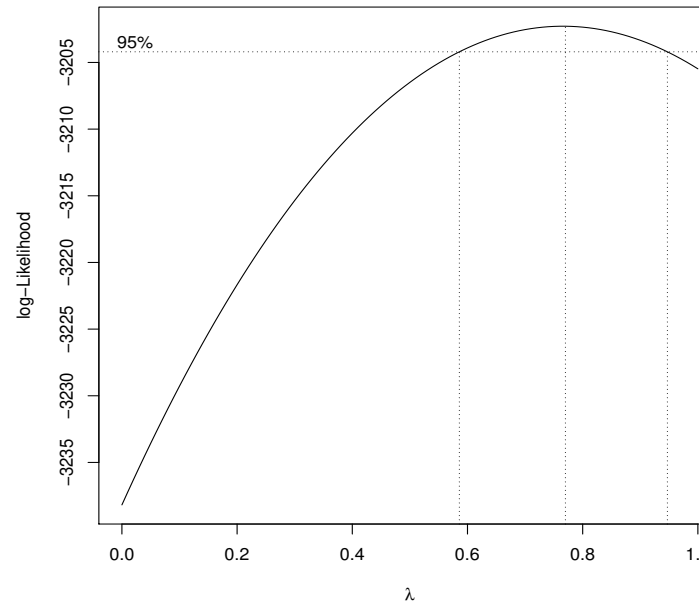
²¹ Strictly speaking, it is the method of analysis rather than the transformation itself that carries the name of Box and Cox, after their 1964 paper in JRSS B.

²² The Stata ladder and lnskev0 commands should not be used for inference.

Box-Cox ANALYSIS FOR DAISY2RED

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. ²³



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
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 (hetttest)

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.

²³ 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,²⁴
- backtransform regression parameters (β 's) only for log-transform (below); **never** backtransform their SEs.

Special procedures for log-transformation²⁵; consider the model

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

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

- * $e^{\beta_0} \sim$ **median** value at original scale for $x=0$,
- * $e^{\beta_1} \sim$ **multiplicative** effect of a 1-unit increase in x ;
example: if $\beta_1=0.4$, then $e^{\beta_1} \approx 1.49 \sim$ 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 $e^1 = 2.72$, and a change in x by $\ln(2) = 0.693 \sim$ increase in z by a factor of 2.²⁶

²⁴ Simulation approaches, beyond the scope of the course, can be used.

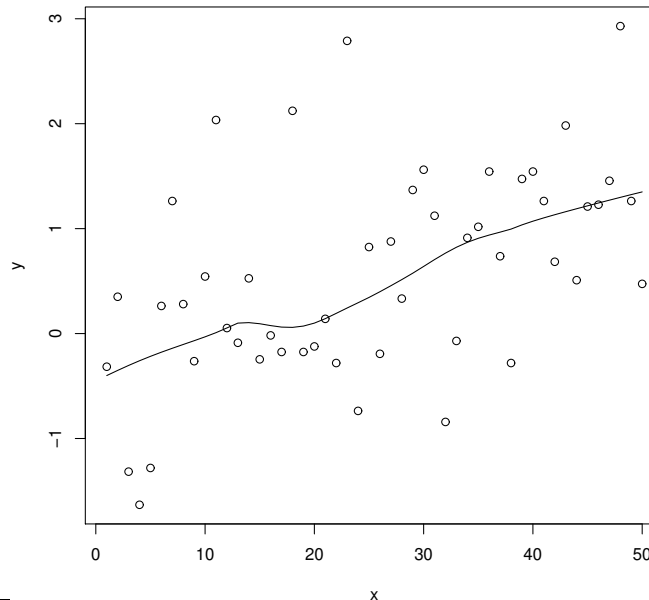
²⁵ Described here for natural log, but works also for other logarithms by replacing the exponential function by the appropriate inverse logarithmic function.

²⁶ **Example:** if $\beta_1=0.4$, then $e^{\beta_1 \cdot \ln(2)} \approx 1.32 \sim$ multiplication by 1.32, or a relative increase by 32%, for a doubling in z .

LOWESS CURVE

Applied view of `lowess` method²⁷:

- = **descriptive graphical tool** to explore the relationship between 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 points²⁸, e.g.
 - * **linearity** (versus non-linearity), or **constant y** (versus non-constant y),
 - may be applied to observed **variables** or to computed **statistics** (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`:



settings:
f=0.5, iter=2, d=1

²⁷ The acronym stands for: **l**ocal **w**eighted scatterplot **s**moothering, and the method is usually referred to Cleveland (1979), *J. Amer. Statist. Assoc.* 74, 829-836.

²⁸ Trends may be masked by substantial noise or a large number of points.

LOWESS SMOOTHING – DETAILS

Several related algorithms (e.g., lowess, loess) exist, all with different parameters to affect or finetune the result \Rightarrow no “correct” or “best” method available, and most important consideration for one’s choice is **usefulness/visual impression** of the result.²⁹

Key ideas/components:

- **regression**: fitted value \hat{y}_i for i^{th} point obtained from a (linear or polynomial) weighted least-squares regression of y on x ,
- **locality**: main contributions to fitted value \hat{y}_i are from observations y_j with x_j close to x_i , because:
 - * only a fraction f 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 the distance $|x_i - x_j|$,³⁰
- **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 \uparrow).

Stata implementation lowess: some limitations and added flexibility³¹:

- no iterative residual weighting, and only linear regression ($d=1$),
- the bwidth parameter (roughly) equals the fraction f ,
- adaptation to binary (0/1) outcome (more later in course).

²⁹ Recall that aims are descriptive and **not** involving statistical inference.

³⁰ 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 the fraction f .

³¹ The weighting can be turned off, and the regression can be replaced by a weighted mean (not clear why of interest).