

Index of 12-L

Page	Title
1	Practical information
2	Tools for model checking: residuals
3	Model checking in regression/ANOVA
4	Correlation
5	Statistical inference for correlation
6	Correlation vs. regression
7	Correlation II
8	Last comments about correlation and regression
9	Introduction to Bayesian approach
10	Bayesian approach
11	How it works: Bayes' formula in action
12	Example: inference about a proportion
13	Choice of prior distribution
14	Proportion example: informative priors
15	Home assignment III
16	Summary notes
17	Review: P -values
18	Appendix: Schematic residual plots (versus fitted)

PRACTICAL INFORMATION

Home assignments:

- no. III returned today: solution posted, review planned,
- no. IV at webpage, and due in one week (Nov 21);
 - * don't wait too long to start looking for articles...
 - * read questions/instructions carefully...

Today's lecture:

- regression (last parts):¹
 - * residual analysis, with Minitab demonstrations,
 - * “warnings” and extensions (suppl. notes: Moodle),
- correlation (full review):¹
 - * correlation coefficient and statistical inference for correlation,
- links between models/procedures for correlation and regression,
- Bayesian statistics:
 - * for your information and not part of the syllabus,
 - * not in textbook: use as reference the lecture notes + supplementary notes (at webpage),
- if time permits, another “misconception clinic”.

¹ See Lecture 11 for textbook references.

TOOLS FOR MODEL CHECKING: RESIDUALS

Residuals:

- = “estimates” of random variables ε_i in model,
- calculated as “observed – expected”, 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 error (may be the same for all residuals, depending on design),
 - * residuals are *not independent*.

Other versions/variables:

- *standardized residuals* (i.e., divided by their standard error):
 $r_i = \hat{\varepsilon}_i / \text{SE}(\hat{\varepsilon}_i)$, approximately distributed as $N(0, 1)$,
- (*advanced*) *deletion residuals*: predicted value from model *without* current observation, also standardized,²
- (*advanced*) *influence statistics*: special statistics to assess the impact of a single observation on the fitted regression line,³
 - * idea: it may be “problematic” if estimates or conclusion depends strongly on a single or a few observation(s).⁴

² Deletion residuals can be used for formal outlier tests (VHM 802/812).

³ Several different statistics (leverage, Cook’s distance, DF(F)ITS) with their specific interpretations, but beyond this course to go into details with them.

⁴ The best way to assess if a particular observation is influential, is to analyze the data with and without this observation, and compare the results.

MODEL CHECKING IN REGRESSION/ANOVA

Proposed use of residuals for model checking:

- 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 :⁵
— should get a noisy pattern with no “parabolic” shapes,
- outliers:
check very large or small values of stand. 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 stand. residuals (r_i),⁶
- data errors:
plot residuals ($\hat{\varepsilon}_i$ or r_i) against data order (if applicable).

“Unusual observations” (in Minitab listing):

- stand. 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*.

⁵ In simple linear regression, plots of the residuals against x_i and \hat{y}_i are practically the same (so one of them will suffice).

⁶ Note that P -values for normality tests only apply approximately to residuals, of any type, because of their dependence.

CORRELATION

Correlation, usually denoted by ρ (“rho”),
 = 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 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 : $\hat{\rho} = r$. (*Pearson* correlation coef.)

Properties of correlations (both types of correlations):

- $-1 \leq r \leq 1$,
- $\left. \begin{array}{l} r > 0 \\ r = 0 \\ r < 0 \end{array} \right\} \sim \left\{ \begin{array}{l} \text{positive} \\ \text{no} \\ \text{negative} \end{array} \right\}$ linear association,
- $r = -1$ and $r = 1$ correspond to perfect linear association (all points on a straight non-horizontal/-vertical line),
- correlation between X and Y is same as betw. Y and X ,
- correlation defined from standardized variables \Rightarrow unaffected by changes in mean or standard deviation,
- 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) follows 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 .⁷

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 $N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$,
- 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,

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

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

- standard errors for r , and tests for $\rho = \text{known value}$: difficult to calculate, and not easily accessible,
- nonparametric correlation: *Spearman's* rank correlation coefficient (i.e., r computed for ranks) \rightarrow lab problem.

⁷ and the conditional distributions of X given Y are normal distributions, and the same for Y given X .

CORRELATION VS. REGRESSION

Correlation and least-squares regression 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 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,⁸
 - * r^2 (usually denoted R^2) is widely misused to indicate the model's "quality".

How to choose between correlation and regression?

look at which model assumptions are most reasonable:

- normal distribution for pairs (X, Y) (correlation), or
 - normal distribution for errors in linear regression,
- for example,
- only one response variable: always regression,
 - two response variables: interest in predicting one from the other (regression)? — or primarily measure/test their degree of linear association (correlation)?

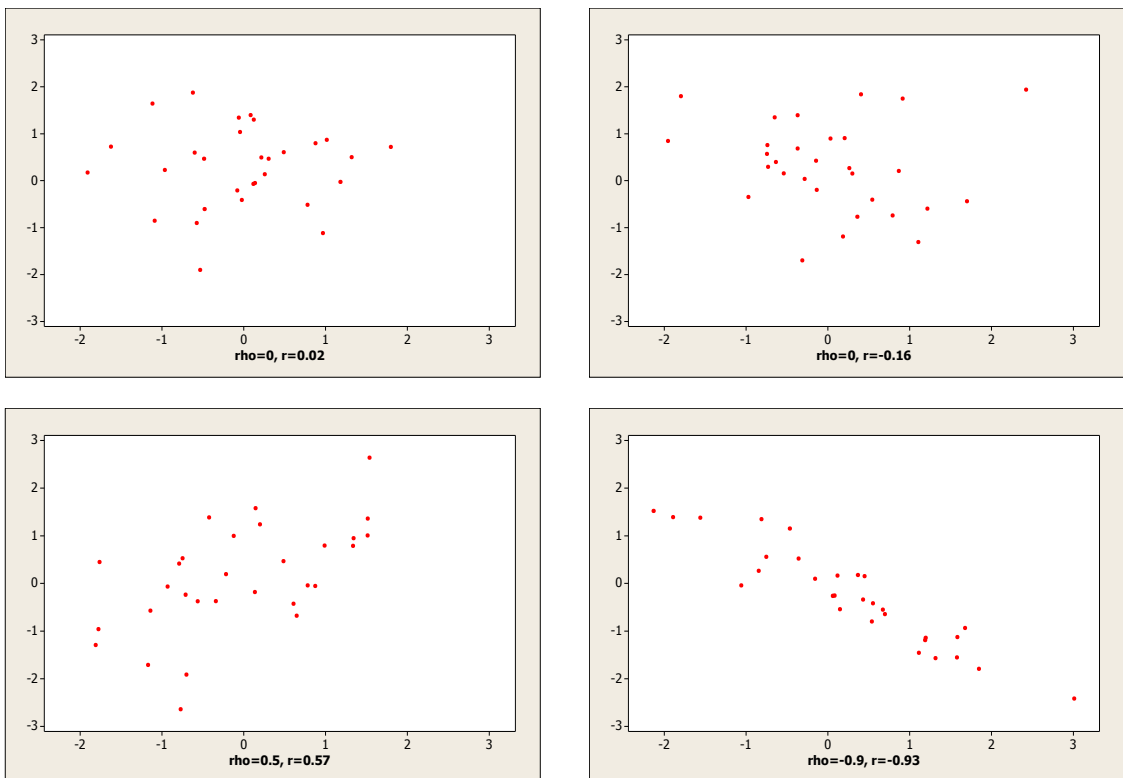
⁸ For an illustration, see Extra exercise 20 (x:20).

CORRELATION II

Some cautions about correlation:

- strictly speaking, only meaningful for *quantitative, response* variables,⁹
- only meaningful for roughly *linear associations*,
- the sample correlation coefficient r is not resistant.

Simulated patterns ($n = 30$):¹⁰



⁹ Some practical recommendations for use and interpretation:

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

¹⁰ See also PSLS 4e: Figure 3.5; S: p. 173; IPS 7e: Figure 2.16.

LAST COMMENTS ABOUT CORRELATION AND REGRESSION

Some cautionary points from textbooks¹¹:

- Linear regression and correlation are based on *linear relationships* — always check if that is reasonable, note: some non-linear relations¹² can be transformed to linear ones and analyzed by a linear regression for the transformed variables, e.g.,

$$y = a \times x^b \longrightarrow \log(y) = \log(a) + b \times \log(x),$$

$$y = a \times b^x \longrightarrow \log(y) = \log(a) + \log(b) \times x,$$

$$y = \frac{a}{1+b \times x} \longrightarrow 1/y = 1/a + (b/a) \times x.$$

- 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),
- *Averaging and/or restricted range* affect estimation of correlation and linear regression.

Variants (extensions) of linear regression:

- multiple linear regression: more than one x -variable in model; textbooks¹³ (and VHM 802/812), generalization of simple linear regression, and not impossible for you to do on your own,
- measurement error models: if x is a *response variable* and measured with (considerable) *error/variation/noise*, and interest is in linear regression on *true* value of x (without error), not prediction.

¹¹ PSLS 4e: Chapter 4; IPS 7e: Section 2.4.

¹² Transformation to achieve a linear relation is discussed in supplemental material for IPS 6e (from IPS6e website (discontinued), available at Moodle site).

¹³ PSLS 3e: Chapter 28 (Suppl.); S: Section 10.4; IPS 7e: Chapter 11.

INTRODUCTION TO BAYESIAN APPROACH

Two rivaling schools of statistics exist:

- Classical (likelihood-based, frequentist)
- Bayesian¹⁴

Statisticians often are strongly in favour of one of them, or have a more pragmatic view (using whatever works for the data at hand).



Fundamentally different concepts:

- probability (frequency interpretation vs. Bayesian subjective probabilities¹⁵),
 - parameters in statistical models (next slide),
- ⇒ different statistical inferences.

Applications of Bayesian methods:

- getting more common in veterinary science/epidemiology,
- vast majority of statistics courses taughts are on classical methods,
- nowadays usually based on MCMC estimation,
 - * simulation-based method using specialised software,
 - * involves new issues related to model checking.

¹⁴ After Reverend Thomas Bayes, 1702–1761, who “invented” Bayes’ theorem.

¹⁵ Unique to each person, also valid for events with no frequency interpretation, e.g. “there is at least one typo in these notes”; see also 3L–3.

BAYESIAN APPROACH

One-line characterisation of Bayesian modelling:

“parameters are random quantities associated with probability distributions”.¹⁶

Rough outline of Bayesian scientific approach:

- 1) Pose a question in terms of a parameter θ ,
- 2) Based on existing knowledge and/or own belief, set up a *prior distribution* for θ ,
- 3) Carry out an experiment to collect (additional) information about θ ,
- 4) Combine the prior distribution and the data into a *posterior distribution* reflecting the knowledge about θ after the experiment.

Classical versus Bayesian approach:

Concept	Classical approach	Bayesian approach
parameter	constant	distribution
prior information on parameters	none	prior distribution
base of inference	likelihood function	posterior distribution
parameter value	(ML) estimate	statistic of posterior, e.g.: median, mode, mean
parameter range	confidence interval	prob. range of posterior
hypothesis statement	test	(Bayesian factors/ <i>P</i> -values)

¹⁶ In classical statistical models, parameters are unknown constants.

HOW IT WORKS: BAYES' FORMULA IN ACTION

Bayes' formula (see 3L–10) for events A and B :

$$P(A|B) = P(B|A) \cdot P(A)/P(B).$$

Also for probability functions and densities f :

$$f(X|Y) = f(Y|X) \cdot f(X)/f(Y).$$

Bayesian framework:

- $f(Y|\theta)$ — the likelihood function (often denoted $L(Y|\theta)$), giving the probability or density of the observed data Y when the parameter takes value θ ,
- $f(\theta)$ — the prior distribution for θ ,
- $f(\theta|Y)$ — the posterior distribution for θ ,

With these definitions, Bayes' formula gives:

$$f(\theta|Y) = f(Y|\theta) \cdot f(\theta)/f(Y) = \text{const}(Y) f(Y|\theta) f(\theta),$$

where $\text{const}(Y)$ means a variable depending on Y but not on θ .

Conclusion: The posterior distribution (for θ) is essentially obtained by multiplying the prior distribution with the likelihood function.

EXAMPLE: INFERENCE ABOUT A PROPORTION

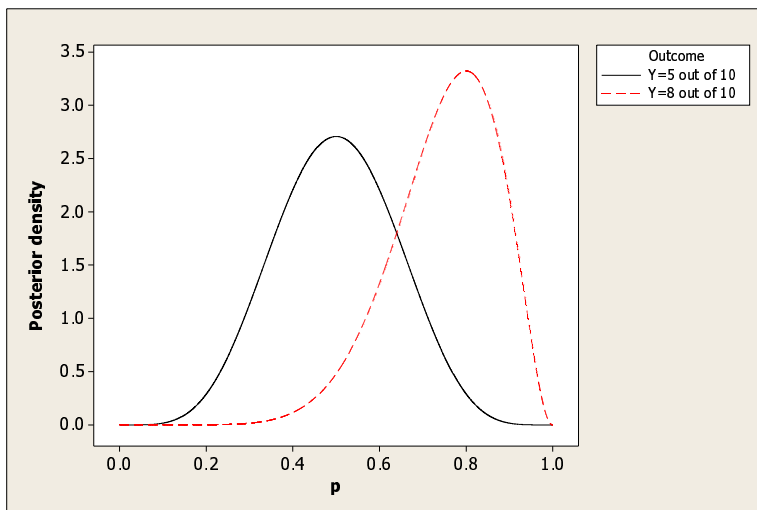
Consider a population of animals and a disease either present or absent in the animals; assume a binomial setting. If $Y=5$ or $Y=8$ animals out of 10 tested positive:

Classical answer:

- 5 out of 10: $\hat{p} = 0.5$, 95% CI (plus 4) for p : (0.238,0.762),
- 8 out of 10: $\hat{p} = 0.8$, 95% CI (plus 4) for p : (0.478,0.951),

Bayesian approach:

- choose a prior: uniform distribution on (0,1), (just one possibility — a “non-informative” prior),
- combine prior and data (Y out of 10 positives) into posterior: beta-distribution¹⁷ ($Y+1, 10-Y+1$), see figure:



- posterior “estimates” and intervals (2.5%–97.5% range):
 - * 5 out of 10: median=0.5, interval: (0.234,0.766),
 - * 8 out of 10: median=0.764, interval: (0.482,0.940).

¹⁷ Beta-distributions (a, b) are distributions on (0,1) with parameters a and b .

CHOICE OF PRIOR DISTRIBUTION

— the cornerstone as well as the strength and weakness of the Bayesian approach; some possibilities:

- subjective prior: the researcher's personal belief,
- non-informative (vague, flat) prior: giving no (or only very weak) preference to a particular θ -value; commonly used despite some philosophical problems, will often lead to results similar to classical statistics,
- convenient prior: to allow simple formulae for posterior distributions in specific models (less used after the invention of MCMC methods; e.g. beta-distribution for p),
- posterior from previous experiment: allows for successive updates of information for every new experiment,
- structural prior: to specify certain regularities in complex data structures (image analysis, genetics),
- expert opinion: quantification of expert opinions obtained in questionnaires or by round-table discussion.

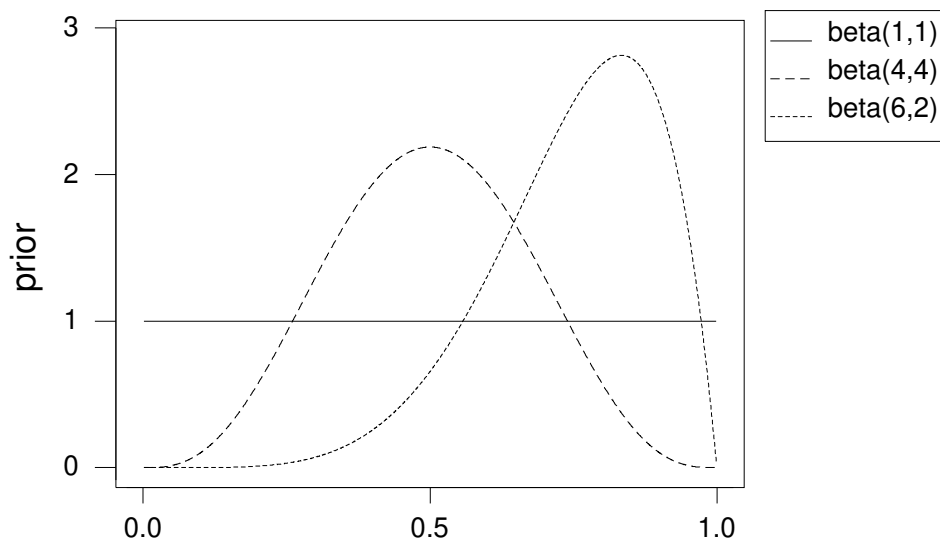
Personal views:

- In some situations (e.g., clinical trials, diagnostic testing), it is very appealing to include prior information; the question is how to quantify it in an “objective” way.
- In many simple situations, the prior distribution is rather unnatural and only complicates matters.

PROPORTION EXAMPLE: INFORMATIVE PRIORS

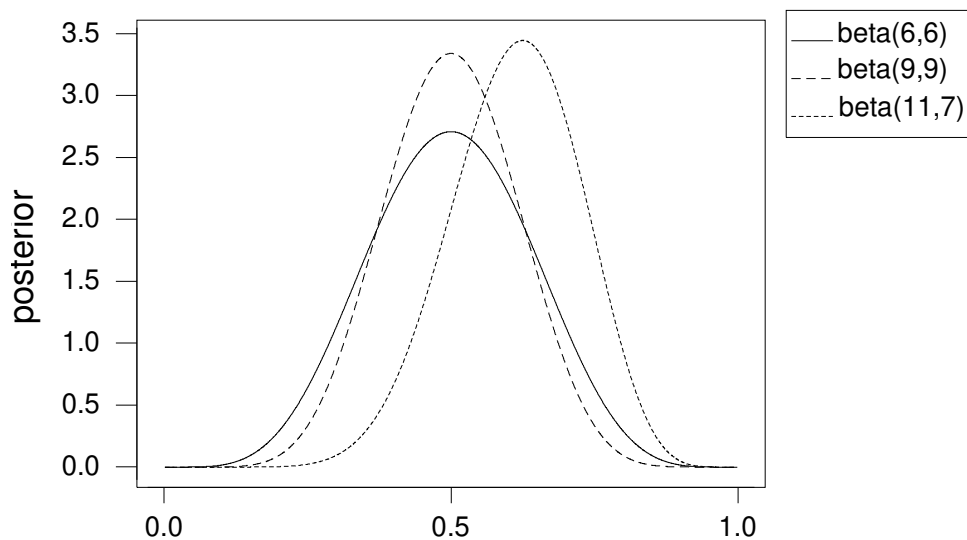
Consider again testing for presence of disease, and observing $Y = 5$ positives out of $n = 10$. Some possible priors for p :

- 1) uniform distribution on $(0,1)$ — same as: beta-distr.(1,1),
- 2) beta-distr.(4,4) — symmetrical around 0.5,
- 3) beta-distr.(6,2) — left skewed with mean 0.75.



Posterior distributions:

- 1) beta-distr.(6,6); 2) beta-distr.(9,9); 3) beta-distr.(11,7);
- * rule: beta-distr.(a, b) & $B(n, Y) \Rightarrow$ beta-distr.($a + Y, b + n - Y$).



HOME ASSIGNMENT III

- generally: some deduction for omitting conclusions or models/assumptions/conditions,
- assumptions behind t/F -procedures:
 - * exact inference: if observations normally distributed,
 - * approximate inference from sample means being approx. normally distributed: generally, with robustness depending on sample size (7L–3).
- sensitivity analysis: explore impact of changes to data or procedures; present example: analyze with/without potential outliers,
- Question 1: can only get inference for the mean by t -distribution methods,
- Question 2: ANOVA is possible and valid (3 groups), but question was worded towards successive pairwise comparisons (so justifiable as pre-planned),
- Questions 2/4: rank-based non-parametric tests can be expressed in terms of medians (and give CIs for medians), but this requires a *strong* assumption about equal shapes (Δ -assumption; slide 8L–7) across groups,
- Questions 3/4: no need for further exploration by X^2 -contributions or pairwise comparisons, respectively, after non-significant overall X^2 - or F -test (unless pre-planned hypotheses exist).

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).

REVIEW: <i>P</i> -VALUES

Significance as determined from *P*-value¹⁸:

Significant			Non-significant
***	**	*	NS
0.1%	1%	5%	
<i>P</i> -value			

Interpretation and suggested wording (personal preferences):

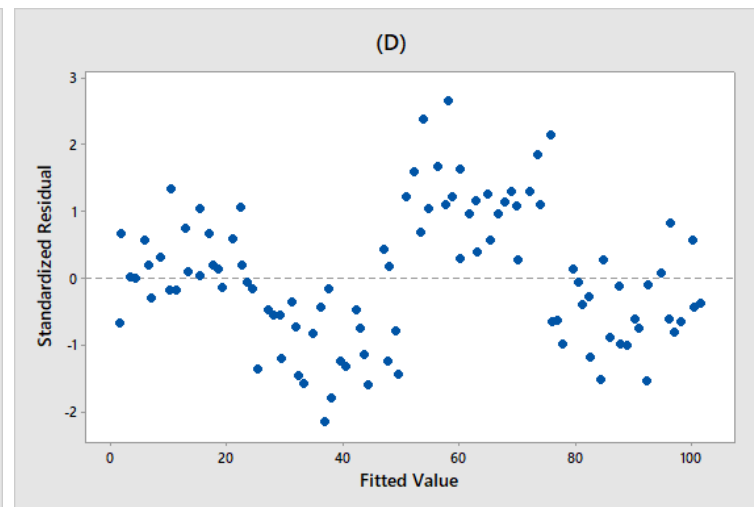
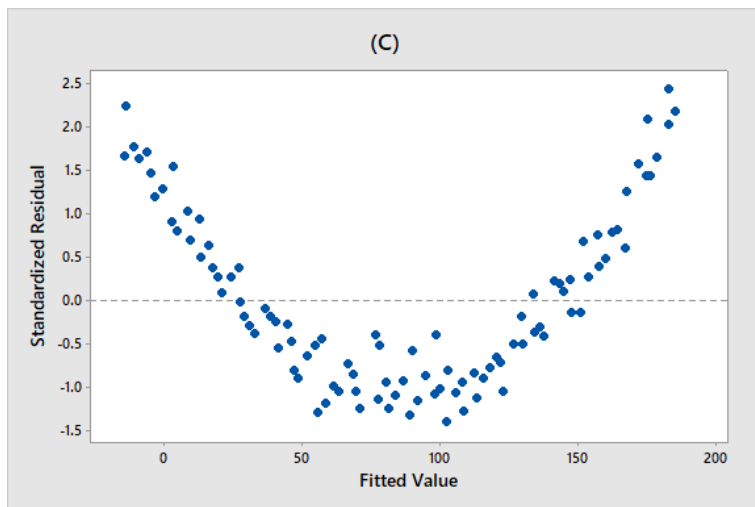
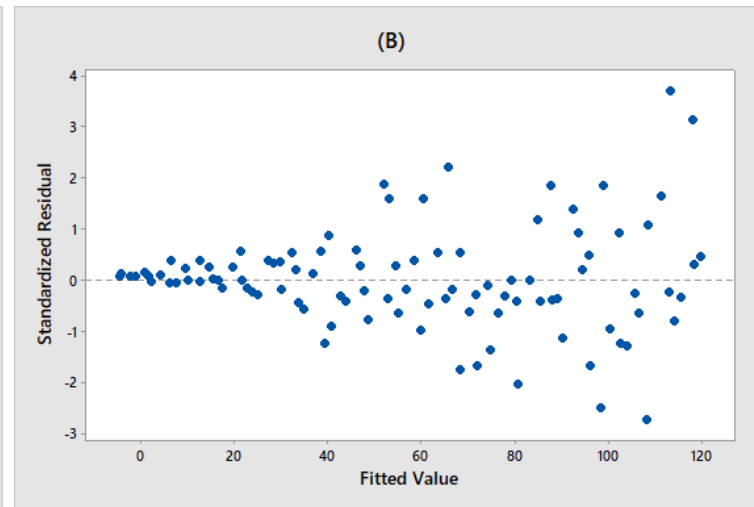
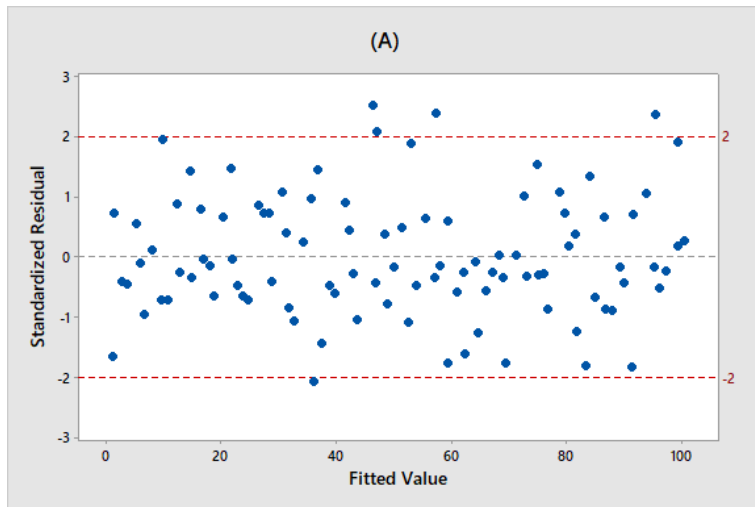
- *P*-values above 0.05 are usually denoted *non-significant* and interpreted as indicating that the result (observed value of the test statistic) could have occurred by coincidence if H_0 is true, so that H_0 cannot be rejected,
- *P*-values just around 0.05 (no matter if just above or below) are *borderline significant*, and give *indication* that H_0 may not be true,
- formally, a *P*-value below 0.05 is *evidence against* H_0 ,
- *P*-values at 0.01 or below are *clearly significant* and give clear indication that H_0 is false, so that H_0 should probably be rejected,
- *P*-values at 0.001 or below are *strongly significant* and show that the data are incompatible with H_0 , so that H_0 without doubt should be rejected.

¹⁸ Warning ! — do not confuse *P*-values and values for *p*'s:

- *P*-values are values (certain probabilities) computed to interpret test results,
- *p*'s are (probability) parameters in particular models, e.g. $B(n, p)$.

APPENDIX: SCHEMATIC RESIDUAL PLOTS (VERSUS FITTED)

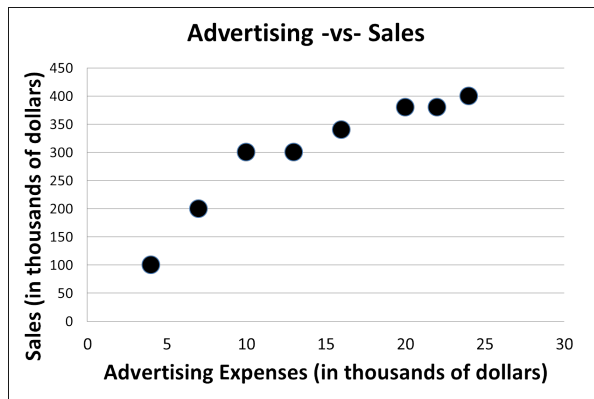
(A) “Perfect”; (B) Fan/Cone shape; (C) Curvilinearity; (D) Missed variable



Chapter 10: Summary Worksheet

Alright marketing majors, your job is on the line. Answer the following questions using the given information for monthly sales and monthly advertising expenditures for 8 different months given below.

Advertising (thousands)	Sales (thousands)
22	380
10	300
4	100
13	300
20	380
16	340
7	200
24	400



correlation coefficient:

$$r = 0.94$$

regression equation:

$$\hat{y} = 13.3x + 107$$

- Is the linear correlation significant? Anything suspicious here?
- What percentage of the variation in sales can be explained by the linear relation to advertising?
- Sketch an approximation to the least squares line on the scatterplot.
- How much in sales would you expect if you spent \$20,000 for advertising?
- How much in sales would you expect if you spent \$30,000 for advertising? Is this a risky prediction? Why or why not?
- What does the slope of the regression equation represent?
- What does the y -intercept represent? Is it meaningful?
- What is the natural choice for the causative variable and the response variable.
- Can we say that an increase in advertising expenditures causes an increase in sales?
- You convince the boss to spend extra money on advertising, including a little extra for yourself. On month 9 you spend \$30,000 on advertising and sales are \$400,010.

- Place the new point on the scatterplot.
- Is this data point an *outlier*?
- What happens to r ?
- What happens to the regression line?
- What happened?
- What argument can you make to save your job?

