WELCOME TO:

STAT.1010:
Riippuvuusanalyysi
Statistical Analysis of Contingency and Regression

Bernd Pape
University of Vaasa
Department of Mathematics and Statistics

TERVETULOA!

www.uwasa.fi/∼bepa/Riippu.html

Literature:

- Amir D. Aczel:
  Complete Business Statistics
- Milton/ Arnold:
  Introduction to Probability and Statistics
- Moore/ McCabe:
  Introduction to the Practice of Statistics
- Conrad Carlberg:
  Statistical Analysis: Microsoft Excel

Old lecture notes in Finnish by Pentti Suomela with SPSS as software may be downloaded from the course homepage.

There you will also find a collection of statistical formulas and tables, which may and should be brought to the exam!

Course Homepage:

www.uwasa.fi/∼bepa/Riippu.html

1. Introduction

1.1. Confidence Intervals and Hypothesis Tests

Confidence Intervals

A point estimate is a single value calculated from the observation values in your sample in order to estimate some parameter of the underlying population. For example the sample mean $\bar{x} = \sum_{i=1}^{n} x_i$, where n is the number of observations x_i in sample, is a point estimate of the underlying population mean μ.

A problem with point estimates is that we are almost sure that they are not the true parameter, because whenever we take a new sample with different observations, we will most probably get a different point estimate leaving us with many point estimates for many different samples, while there is only a single true parameter in the population which cannot simultaneously be identical to all those point estimates from the different samples.

By adding and subtracting margins of error to your point estimate you convert your point estimate into an interval estimate. This increases the chance of the true parameter being covered for the price of a less precise estimate of its value.

If the sampling distribution of your estimator is known, then the margins of error can be determined such, that the resulting interval has a precisely determined probability $1-\alpha$, say, that the interval covers the true parameter value. We have then found a confidence interval at confidence level $1-\alpha$.

The sampling distribution of an estimator is a smoothed histogram of its value in many samples scaled such, that calculating its integral between two numbers will yield the probability that the estimate comes out with a value somewhere between those numbers.

Example

We learned in STAT1030 that the standardized sample mean in a sample of n observations

$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

with sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is Student-t distributed with $n-1$ degrees of freedom.

Let $t_{\alpha/2}(n-1)$ denote the value of T_n for which

$$P(T_n > t_{\alpha/2}(n-1)) = \frac{\alpha}{2}$$

such that by symmetry of the Student t-distribution

$$P(|T_n| > t_{\alpha/2}(n-1)) = P\left(|\bar{X} - \mu| > t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right) = \alpha$$

and the $1-\alpha$ confidence interval for μ becomes

$$CI_{1-\alpha} = \left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right).$$

$t_{\alpha/2}(n-1)$ is determined such that the area (=integral) under the density curve of the Student t-distribution with $n-1$ degrees of freedom between this value and $+\infty$ is exactly $\frac{\alpha}{2}$. These values are tabulated and available from Excel by typing =T.INV.2T(α; $n-1$) in any cell (or TINV(α; $n-1$) before Excel 2007).

Hypothesis Tests

Whenever we calculate something based upon sample observations only, it is called a statistic. An estimator is a statistic used for the special purpose of estimating an underlying population parameter, such as $\bar{x}$ for μ.

Now suppose that rather than using a statistic in order to estimate some unknown parameter, you have already an opinion about what the value of that parameter should be and you want to cross-check whether your opinion can be reasonably maintained in the light of the sample statistics you got.

For example, theory claims that something should be one on average, but in your sample you find hat $\bar{x} = 2$. Does this mean that the theory is wrong or is this just because you didn't see the full population? You can make informed decisions about this if you know the sampling distribution of your statistic under the assumption that the null hypothesis (e.g. $\mu = 1$) is true. This is called hypothesis testing.

The approach is to use the sampling distribution under the null in order to calculate the probability of getting a sample statistic at least as extreme as the value you've got even though the null hypothesis holds true. This is called the p-value of the test.

If the p-value is large, it means that the probability of getting your sample statistic under the presumed parameter value is large, and you accept the null hypothesis that the population parameter is what you claimed it to be.

If the p-value is small, it means that the probability of getting your sample statistic under the presumed parameter value is small, and you reject the null hypothesis against the alternative hypothesis that the population parameter is something else.

How exactly the p-value is determined depends upon whether you use a one-sided or a two-sided test.

In the case of testing whether the arithmetic mean has the specific value μ_0 under the null hypothesis ($H_0 : \mu = \mu_0$) the alternative hypothesis H_1 in a <u>two sided</u> test is

$$H_1 : \mu \neq \mu_0,$$

whereas there are two options for a <u>one-sided</u> test:

$$H_1 : \mu < \mu_0 \qquad \text{OR} \qquad H_1 : \mu > \mu_0.$$

For example, arithmetic sample means larger than the hypothesized population mean, $\bar{x} > \mu_0$, are evidence against H_0 in two sided tests and in one-sided tests of the form $H_1 : \mu > \mu_0$, but not in one-sided tests of the form $H_1 : \mu < \mu_0$.

Similarly, arithmetic sample means smaller than the hypothesized population mean, $\bar{x} < \mu_0$, are evidence against H_0 in two sided tests and in one-sided tests of the form $H_1 : \mu < \mu_0$, but not in one-sided tests of the form $H_1 : \mu > \mu_0$.

Hence we calculate the p-value of the test as

$$p = P(\bar{X} \leq \bar{x} \,|\, \mu = \mu_0) \qquad \text{for } H_1 : \mu < \mu_0,$$
$$p = P(\bar{X} \geq \bar{x} \,|\, \mu = \mu_0) \qquad \text{for } H_1 : \mu > \mu_0,$$

$$p = P(|\bar{X} - \mu| \geq |\bar{x} - \mu| \,|\, \mu = \mu_0) \text{for } H_1 : \mu \neq \mu_0,$$

where $\bar{X}$ denotes the random variable containing the arithmetic mean whose value differs from sample to sample, $\bar{x}$ denotes the particular value of the arithmetic mean we got for our sample at hand, and $|\mu = \mu_0$ stands for the condition that the population mean has indeed the value μ_0, as stated in the null hypothesis.

The null hypothesis is rejected if the p-value falls below some prespecified level α, the so-called significance level of the test, otherwise H_0 is accepted. α denotes the probability of committing a Type I error, which means falsely rejecting a true null hypothesis. We want this probability to remain small, hence we choose often $\alpha = 5\%$, if not even smaller.

Example

Using again that the standardized sample mean

$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

in a sample of n observations is $t(n-1)$-distributed, we get e.g. for the one-sided test against $H_1 : \mu > \mu_0$,

$$p = P(\bar{X} \geq \bar{x} \,|\, \mu = \mu_0) = P(T_n \geq t),$$

where $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ obtained in our sample. p may hence be obtained from integrating the Student t-distribution with n–1 degrees of freedom between t and $+\infty$ and is available from Excel using the command T.DIST.RT($t; n-1$). Similarly,

$$\begin{array}{ll} p = \text{T.DIST.RT}(-t; n-1) & \text{for } H_1 : \mu < \mu_0, \\ p = \text{T.DIST.2T}(|t|\,; n-1) & \text{for } H_1 : \mu \neq \mu_0. \end{array}$$

Note that p-values of two-sided tests are twice the p-values of one-sided tests, because, by symmetry of the t-distribution:

$$P(|T| \geq |t|) = P(T \leq -|t|) + P(T \geq |t|) = 2P(T \geq |t|).$$

Example (continued)

If software is not available, we can still use the critical values t_α from statistical tables in order to decide whether we may reject H_0 at significance level α or not. In order to do that, express the condition $p < \alpha$ required for rejection of H_0 in terms of threshold values t_α, recalling that $P(T > t_\alpha) = \alpha$.

Hence the condition $p < \alpha$ for rejecting $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ reads

$$\underbrace{P(T>t)}_{p} < \underbrace{P(T>t_\alpha)}_{\alpha},$$

which is equivalent to the condition $t > t_\alpha$ for positive values of t (the only area of interest against $H_1 : \mu > \mu_0$), because there the t-distribution is monotonically decreasing.

Similarly the condition for rejecting H_0 against $H_1 : \mu < \mu_0$ is $t < -t_\alpha$. $(P(T<t) < P(T<-t_\alpha))$

The condition for rejecting $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is $|t| > t_{\alpha/2}$, because it means the same as the conditions $t > t_{\alpha/2}$ for $t > 0$, or $t < -t_{\alpha/2}$ for $t < 0$. $(P(|T|>|t|) < P(|T|>t_{\alpha/2}))$

1.2. Scales of Measurement

Recall that the applicability of different statistical methods depends upon the measurement scale of the variable in question.

Variables on nominal scale cannot be used in calulcations and don't reveal any order either. They can only be used for sorting statistical units into groups.
Example: Gender, profession, colours,. . .

Variables on ordinal scale cannot be directly used in calculations either, but they reveil an implicit order. The ordering implies that it is meaningful to define ranks, on the basis of which it is possible to calculate quantiles such as e.g. the median.
Example: agree/partially agree/disagree,
bad/average/good,. . .

Variables on interval scale can directly be used in calculations involving only sums and differences between observation values such as the arithmetic mean $\bar{x}$, however the aribtrariness of the point of origin in these variables precludes meaningful calculation of statistics involving ratios of observation values.
Example: clock time, temperature,...

Variables on ratio scale share the properties of variables on interval scale but allow additionally for meaningful calculation of statistics based upon ratios of observation values, such as the coefficient of variation or the harmonic mean, because the point of origin is uniquely defined as absence of the quantity measured.
Example: Money, Weight, Time intervals,. . .

Note. Variables on interval scale can always be transformed into ratio scale by taking differences of the original variable.
Example: Clock time $\rightarrow$ Time intervals.

1.3. Interdependence of Statistical Variables

1.3.1. Both Variables on Nominal Scale

The analysis starts off from a so called contingency table (ristiintaulukko) which displays the absolute counts (havaittut frekvenssit) f_{ij} of statistical units belonging both to class i of variable X and to class j of variable Y.

Dividing these counts by the total number of observations n, yields the so called relative frequencies (suhteelliset frekvenssit) $p_{ij} = \frac{f_{ij}}{n}$, which applying the relative frequency approach may be interpreted as a proxy for the probability that a randomly chosen statistical unit belongs to class i of X and to class j of Y simultaneously. Therefore we call the observed frequencies (both absolute and relative) also the joint distribution (yhteysjakauma) of X and Y.

The probabilities of a statistical unit to belong to a certain class of X, regardless of its classification with respect to Y are given by $p_{i\bullet} = \sum_j p_{ij}$, which together make up the marginal distribution (reunajakauma) of X. Similiarly, the collection of $p_{\bullet j} = \sum_i p_{ij}$ of probabilities for statistical units to belong to class j of Y regardless of their classification according to X are called the marginal distribution of Y.

Now we know from probability calculus that two events are independent when their joint probability equals the product of their marginal probabilities, that is, $p_{ij} = p_{i\bullet}p_{\bullet j}$, which leads us to assume independence of X and Y when the observed frequencies f_{ij} equal the so called expected frequencies (odotettut frekvenssit) e_{ij}, where

$$e_{ij} = np_{ij} = n \cdot p_{i\bullet} \cdot p_{\bullet j} = n \cdot \frac{f_{i\bullet}}{n} \cdot \frac{f_{\bullet j}}{n} = \frac{f_{i\bullet} f_{\bullet j}}{n}$$

with $f_{i\bullet} = \sum_j f_{ij}$ and $f_{\bullet j} = \sum_i f_{ij}$.

The test statistic used in order to assess whether $f_{ij} \approx e_{ij}$ is Pearson's χ^2 statistics:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - e_{ij})^2}{e_{ij}},$$

where we have assumed that the row variable (rivimuuttuja) X has r classes, and the column variable (sarakemuuttuja) Y has s classes, such that the overall dimension of the contingency table is $(r \times s)$.

The null and alternative hypotheses of the χ^2 independence test (riippumattomuustesti) are:

H_0: X and Y are statistically independent
H_1: X and Y are statistically dependent

If the null hypothesis holds true, then χ^2 is approximately χ^2-distributed with

$df = (r-1)(s-1)$ degrees of freedom.

Large values of χ^2 lead to a rejection of the null hypothesis, which means that we believe there is dependence also out of sample.

When using statistical tables, one first decides for a significance level (merkitsevyystaso) α, which denotes the risk one is ready to take of falsely rejecting a null hypothesis which in fact holds true, and compares the obtained value for the χ^2-statistics with the corresponding critical value (kriittinen arvo) $\chi^2_\alpha(df)$ of the table.

Statistical software programmes report a p-value, as described in the introduction, which must be compared with the significance level. We accept H_0 if $p \geq \alpha$ and reject H_0 if $p < \alpha$. In Excel: p =CHIDIST($\chi^2; df$).

Usually $\alpha = 0.05$, such that:

H_0 is accepted if $\chi^2 \leq \chi^2_\alpha(df)$ or $p \geq 0.05$,
H_0 is rejected if $\chi^2 > \chi^2_\alpha(df)$ or $p < 0.05$.

Note that statistical significance of χ^2 alone, e.g. rejection of H_0, does not yet make any statement about the strength of dependence between the variables.

Table. Tail fractiles χ^2_α of the χ^2-distribution: $P(\chi^2 > \chi^2_\alpha(df)) = \alpha$.

					χ_α^2(df)					
α	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.001
1	0.000039	0.000157	0.000982	0.003932	0.0158	2.706	3.841	5.024	6.635	10.827
2	0.0100	0.0201	0.0506	0.103	0.211	4.605	5.991	7.378	9.210	13.815
3	0.0717	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	16.266
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	18.466
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	20.515
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	22.457
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	24.321
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	26.124
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	27.877
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	29.588
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	31.264
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	32.909
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	34.527
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	36.124
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	37.698
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	39.252
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	40.791
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	42.312
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	43.819
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	45.314
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	46.796
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	48.268
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	49.728
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	51.179
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	52.619
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	54.051
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	55.475
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	56.892
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	58.301
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	59.702
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	73.403
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	86.660
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	99.608
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	112.317
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	124.839
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	137.208
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	149.449

Example: The 5% critical value for a two-way table with 2 levels per variable is 3.84 and the p-value for a χ^2-statistic of 5.024 is 0.025.

Recall that in the special case of two way tables, that is $r = s = 2$, the calculation of the χ^2 statistic simplifies to

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1\bullet}f_{2\bullet}f_{\bullet 1}f_{\bullet 2}} \sim \chi^2(1) \text{ under } H_0.$$

f_{11}	f_{12}	$f_{1\bullet}$
f_{21}	f_{22}	$f_{2\bullet}$
$f_{\bullet 1}$	$f_{\bullet 2}$	n

The χ^2 statistics is only approximately $\chi^2(1)$-distributed (the χ^2-distribution comes about by approximating the binomially distributed marginal sums with the normal distribution). The approximation becomes more precise by applying the following continuity correction (jatkuvuuskorjaus):

$$\chi^2 = \frac{n(|f_{11}f_{22} - f_{12}f_{21}| - n/2)^2}{f_{1\bullet}f_{2\bullet}f_{\bullet 1}f_{\bullet 2}}$$

Note that the shortcuts described on this page are valid only for two-way tables and may not be applied to contingency tables of any other size than 2×2.

A small p-value tells us that the assumption of independence probably does not hold, that is, the row and the column variable are probably dependent. However, the p-value says nothing about how strong this dependence actually is.

The most useful measure of dependence for categorical data is Cramer's V defined as

$$V = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n(k-1)}},$$

where k is the smaller of the number of rows r and columns s. It ranges from 0 (complete independence) to 1 (perfect dependence). As a rule of thumb, there is no substantial dependence if $V < 0.1$.

Example:

Satisfaction with the companies management:

f_{ij}	Vocational training		
	Yes	No	Sum
Satisfied	87	112	199
Don't know	34	30	64
Unsatisfied	22	96	118
Sum	143	238	381

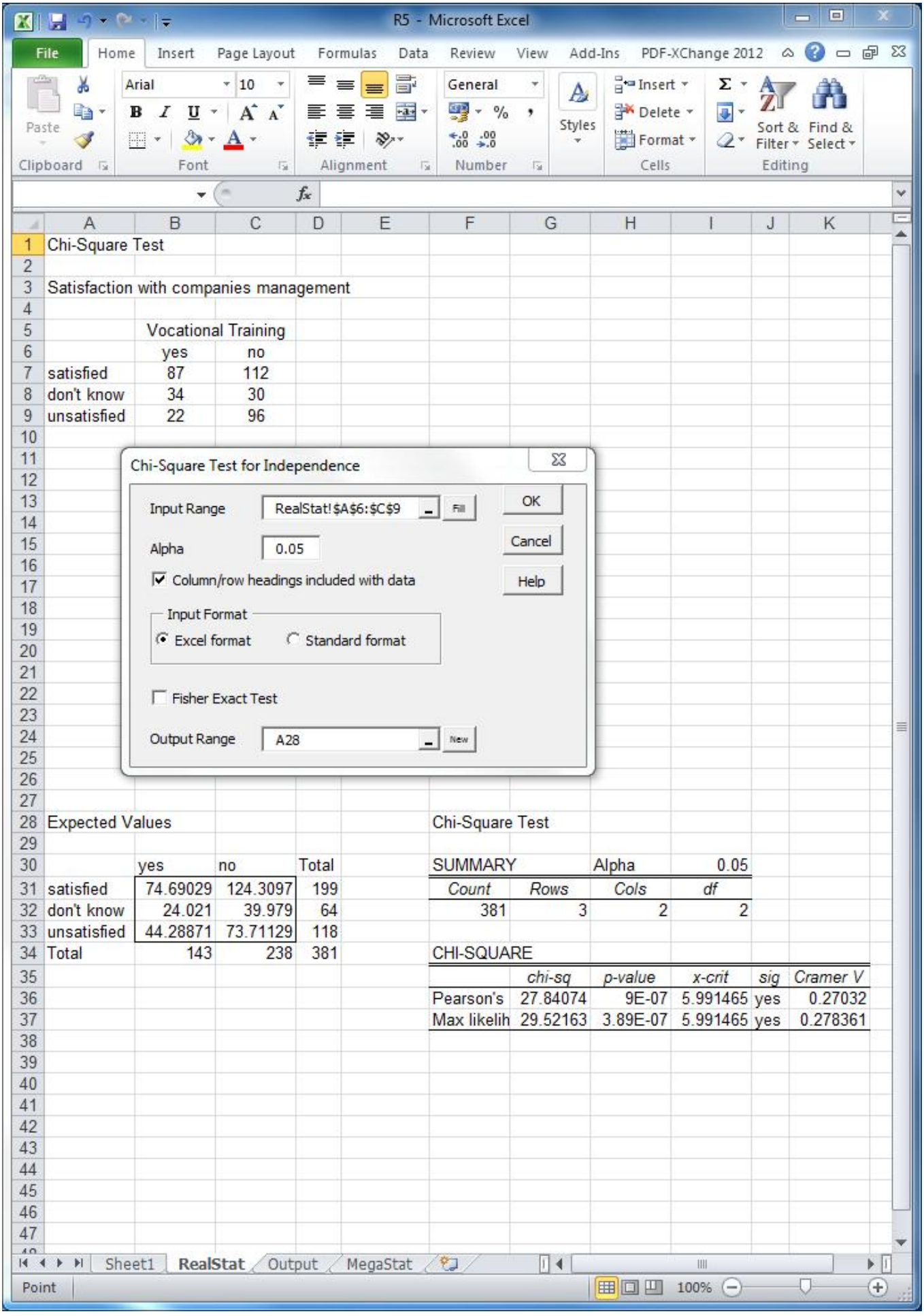
Expected frequencies under independence:

e_{ij}	Vocational training		
	Yes	No	Sum
Satisfied	74.7	124.3	199
Don't know	24.0	40.0	64
Unsatisfied	44.3	73.7	118
Sum	143	238	381

$df = (3-1)(2-1) = 2, \qquad \chi^2_{0.05}(2) = 5.99,$

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = 27.841 > \chi^2_{0.05}(2)$$

$\Rightarrow$There is dependence, $V = \sqrt{\frac{27.841}{381(2-1)}} = 0.27.$


R5 - Microsoft Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-Ins
PDF-XChange 2012
Chi-Square Test
Satisfaction with companies management
Vocational Training
yes
no
satisfied 87 112
don't know 34 30
unsatisfied 22 96
Chi-Square Test for Independence
Input Range
RealStat!$A$6:$C$9
Alpha
0.05
Column/row headings included with data
Input Format
Excel format
Standard format
Fisher Exact Test
Output Range
A28
OK
Cancel
Help
Expected Values
yes no Total
satisfied 74.69029 124.3097 199
don't know 24.021 39.979 64
unsatisfied 44.28871 73.71129 118
Total 143 238 381
Chi-Square Test
SUMMARY Alpha 0.05
Count Rows Cols df
381 3 2 2
CHI-SQUARE
chi-sq p-value x-crit sig Cramer V
Pearson's 27.84074 9E-07 5.991465 yes 0.27032
Max likelih 29.52163 3.89E-07 5.991465 yes 0.278361
Sheet1
RealStat
Output
MegaStat
Point
100%

1.3.2. Both Variables on Ordinal Scale

χ^2 may still be applied, but it doesn't take the order of the ranked classification into account. The concordance (samansuuntaisuus) of the ranked classification is measured by Spearman's rank correlation coefficient r_s, also known as Spearman's ρ, and Kendall's τ.

In order to calculate these measures, one determines first the X and Y's ranks (sijaluvut):

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},\ y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}.$$

If there are no ties, then Spearman's ρ and Kendall's τ are determined as

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \text{ and } \tau = 1 - \frac{4Q}{n(n-1)},$$

where d_i is the difference between ranks and Q is the number of discordant pairs (parittaisten sijanvaihdosten lukumäärä), that is pairs, where an increase in X corresponds to a decrease in Y.

Example: (Snedecor & Cochran)
Ranking of seven rats' conditions by two observers:

Rat Number	Ranking by Obs. 1	Ranking by Obs. 2	Difference d_i	d_i^2
1	4	4	0	0
2	1	2	-1	1
3	6	5	1	1
4	5	6	-1	1
5	3	1	2	4
6	2	3	-1	1
7	7	7	0	0
			$\sum d_i = 0$	$\sum d_i^2 = 8$

$$r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 8}{7(49-1)} = 0.857.$$

In order to compute Kendall's τ, rearrange the two rankings so that one of them is in increasing order:

Rat No.	2	6	5	1	4	3	7
Obs. 1	1	2	3	4	5	6	7
Obs. 2	2	3	1	4	6	5	7

Taking each rank given by observer 2 in turn, count the smaller ranks to the *right* of it and add these counts. For rank 2 the count is 1, since only rat 5 has a smaller rank. The six counts are 1, 1, 0, 0, 1, 0, (no need to count the extreme right rank), such that

$$Q = 3 \quad \text{and} \quad \tau = 1 - \frac{4Q}{n(n-1)} = 1 - \frac{12}{42} = \frac{5}{7} \approx 0.714.$$

Recall that Spearman's rank correlation coefficient is indeed Pearson's linear correlation coefficient applied to ranks, which simplifies to the form given above only in the special case that there are no ties (that is, there are no multiple observations for the same rank).

Both coefficients obey $-1 \leq r_S, \tau \leq 1$, where

$r_s = \tau = 1 \quad \Leftrightarrow$ ranks in same order,
$r_s = \tau = -1 \quad \Leftrightarrow$ ranks in opposite order,
$r_s = \tau = 0 \quad \Leftrightarrow$ independent ranks.

We usually test independence of the ranks:

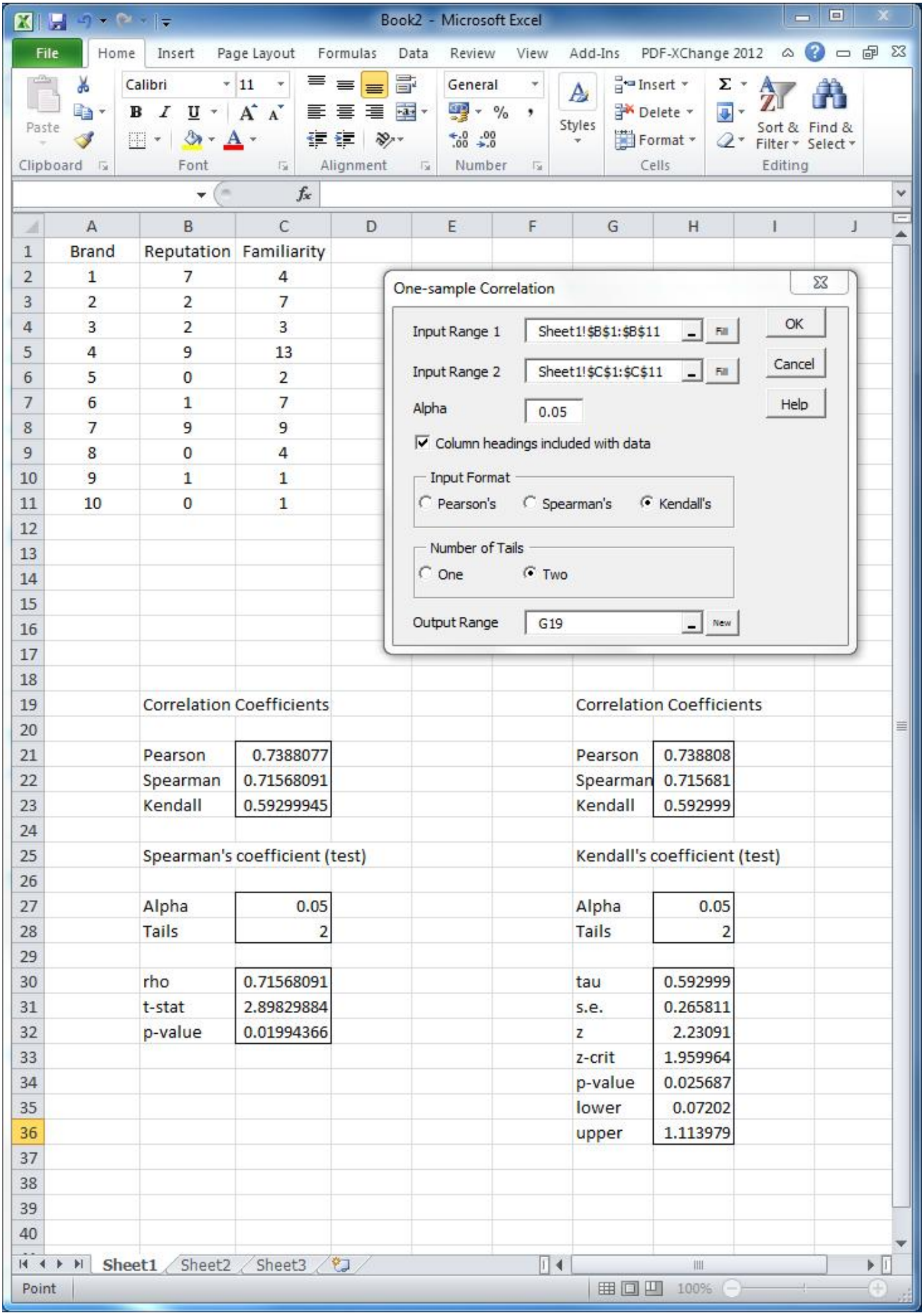
$$H_0: \ \rho_S = 0 \text{ or } \tau = 0.$$

The Real Statistics toolpack offers the exact p-values for this test.

If software is not available and you have a sufficiently large sample size n, you can still use that the test statistic to test $H_0: \rho_S = 0$ is in large samples approximately

$$z = r_S\sqrt{n} \sim N(0,1) \quad \text{under } H_0.$$

E.g. $|z| > 1.96$ is an indication that r_s is significant at $\alpha = 5\%$. The same approximation works for the linear correlation coefficient r, but for τ: $z \approx \frac{3}{2}\tau\sqrt{n}$.



	A	B	C	D	E	F	G	H	I	J
1	Brand	Reputation	Familiarity							
2	1	7	4							
3	2	2	7							
4	3	2	3							
5	4	9	13							
6	5	0	2							
7	6	1	7							
8	7	9	9							
9	8	0	4							
10	9	1	1							
11	10	0	1							
12										
13										
14										
15										
16										
17										
18										
19		Correlation Coefficients					Correlation Coefficients			
20										
21		Pearson	0.7388077				Pearson	0.738808		
22		Spearman	0.71568091				Spearman	0.715681		
23		Kendall	0.59299945				Kendall	0.592999		
24										
25		Spearman's coefficient (test)					Kendall's coefficient (test)			
26										
27		Alpha	0.05				Alpha	0.05		
28		Tails	2				Tails	2		
29										
30		rho	0.71568091				tau	0.592999		
31		t-stat	2.89829884				s.e.	0.265811		
32		p-value	0.01994366				z	2.23091		
33							z-crit	1.959964		
34							p-value	0.025687		
35							lower	0.07202		
36							upper	1.113979		
37										
38										
39										
40										

1.3.3. Both Variables on Interval Scale

Linear association (lineaarinen riippuvuus) between two variables may be assessed using Pearson's linear correlation coefficient $r = r_{xy}$ if both variables are at least on interval scale.

Recall:

• Pearson's linear correlation coefficient is symmetric in the sense that it makes no difference which variable you call x and which you call y in calculating the correlation.

• r_{xy} does not change when we change the units of x, y, or both.

• r_{xy} measures only the strength of linear relationships. It does not describe curved relationships, no matter how strong they are.

• r_{xy} is always a number between -1 and 1 with the sign of r indicating the sign of the linear relationship.

• Pearson's linear correlation coefficient is more sensitive to outliers than Spearman's rank correlation coefficient and Kendall's τ.

Recall that if y is the sum of a linear function of x and some error term e with zero mean, that is,

$$y = \hat{y} + e, \text{ where } \quad \hat{y} = b_0 + b_1 x, \quad \bar{e} = 0,$$

then we may determine the coefficients of the so called regression line (regressiosuora) by means of the method of least squares (OLS) (pns-menetälmä) as

$$b_1 = r_{xy} \cdot \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \bar{x},$$

where s_x and s_y denote the standard deviations of x and y, respectively.

Recall:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right),$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right),$$

$$r_{xy} = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left[\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\displaystyle\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}.$$

Coefficients of Determination

Pearson's linear correlation coefficient r_{xy} is related to the coefficient of determination R^2 (selityskerroin/-aste) of such a regression by

$$R^2 := r_{xy}^2,$$

where R^2 measures the fit (yhteensopivuus) of the regression line as:

$$R^2 = \frac{\text{variance of predicted values } \widehat{y}}{\text{variance of observed values } y} = \frac{s_{\widehat{y}}^2}{s_y^2}.$$

A better measure of fit when comparing regressions with varying numbers of regressors is the so called adjusted R^2 (tarkistettu selitysaste) given by

$$\overline{R^2} = 1 - \frac{n-1}{n-2}(1 - r_{xy}^2)$$

in the case of only one regressor. It approaches the ordinary R^2 for large n. Note that unlike R^2, $\overline{R^2}$ may become negative for correlations close to zero:

$$r_{xy} = 0 \quad \Rightarrow \quad \overline{R^2} = -\frac{1}{n-2}.$$

Linear Regression in Excel

You can perform linear regression either with excel's own data analysis tool or with the Real Statistics data analysis tool by Charles Zaiontz available at `www.real-statistics.com`. Excel's own tool offers additional plots and the real statistics tool offers additional analysis, both of which will be discussed later in this course.

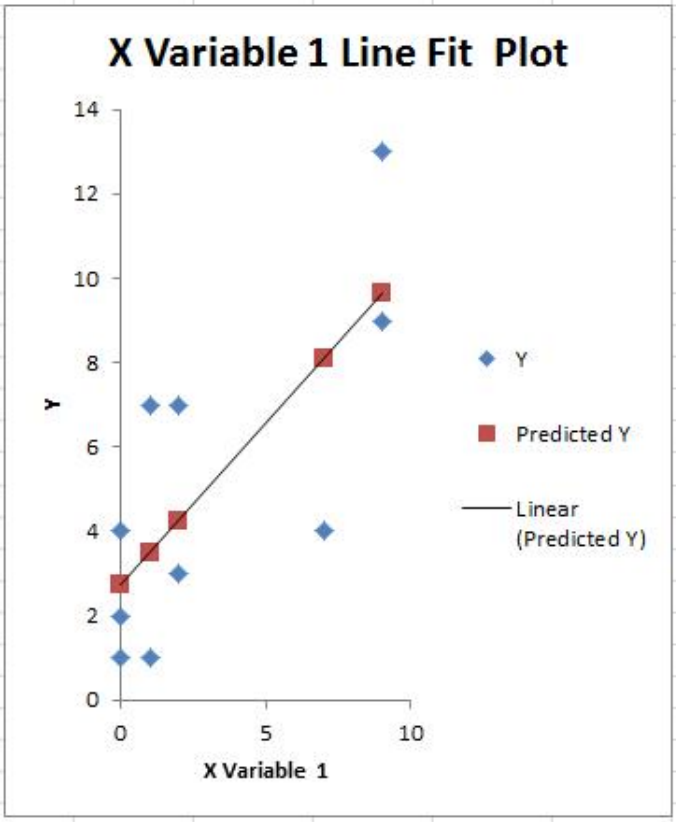
The regression output contains always:
• the coefficients of determination R^2, $\overline{R^2}$;
• the standard error of the estimate s_e, which is an estimator for the unknown standard deviation of the error term out of sample;
• an Analysis of Variance table; which is an F-test of $H_0: \ R^2 = 0$;
• A Parameter Estimates table containing the regression parameters and t-tests of the hypotheses that the respective parameter is 0.

R7 - Microsoft Excel

File | Home | Insert | Page Layout | Formulas | Data | Review | View | Add-Ins | PDF-XChange 2012

A1: SUMMARY OUTPUT

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT							X:	Y:
2								7	4
3	*Regression Statistics*							2	7
4	Multiple R	0.7388077						2	3
5	R Square	0.5458368						9	13
6	Adjusted R Square	0.4890664						0	2
7	Standard Error	2.7673682						1	7
8	Observations	10						9	9
9								0	4
10	ANOVA							1	1
11		*df*	*SS*	*MS*	*F*	*Significance F*		0	1
12	Regression	1	73.63339	73.63339	9.614814	0.014649			
13	Residual	8	61.26661	7.658327					
14	Total	9	134.9						
15									
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	2.7197758	1.164077	2.336422	0.047681	0.035409	5.404143	0.035409	5.404143
18	X Variable 1	0.7678143	0.24762	3.100776	0.014649	0.196802	1.338827	0.196802	1.338827
19									
20									
21									
22	RESIDUAL OUTPUT								
23									
24	*Observation*	*Predicted Y*	*Residuals*						
25	1	8.0944756	-4.09448						
26	2	4.2554043	2.744596						
27	3	4.2554043	-1.2554						
28	4	9.6301041	3.369896						
29	5	2.7197758	-0.71978						
30	6	3.4875901	3.51241						
31	7	9.6301041	-0.6301						
32	8	2.7197758	1.280224						
33	9	3.4875901	-2.48759						
34	10	2.7197758	-1.71978						
35									
36									
37									
38									
39									
40									
41									
42									
43									
44									
45									
46									
47									



Sheet1 | RS Correlations | Output | Regression | RS R

Ready

100%

The ANOVA table for simple linear regression

Analysis of variance (ANOVA) summarizes information about sources of variation in the data based on the framework

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}.$$

The idea is that we may split up the deviation of the observed values y_i from their arithmetic mean $\bar{y}$ into a sum of the deviation of the regression fit $\widehat{y}_i$ from $\bar{y}$ and the deviation of y_i from $\widehat{y}_i$:

$$(y_i - \bar{y}) = (\widehat{y}_i - \bar{y}) + (y_i - \widehat{y}_i).$$

If we square each of the three deviations above and then sum over all n observations, it is an algeabraic fact that the sums of squares add:

$$\sum (y_i - \bar{y})^2 = \sum (\widehat{y}_i - \bar{y})^2 + \sum (y_i - \widehat{y}_i)^2,$$

which we rewrite as

$$\text{SST} = \text{SSR} + \text{SSE},$$

where

$\text{SST} = \sum (y_i - \bar{y})^2$, $\text{SSR} = \sum (\widehat{y}_i - \bar{y})^2$, $\text{SSE} = \sum (y_i - \widehat{y}_i)^2$.

In the abbreviations SST, SSR, and SSE, SS stands for sum of squares, and the T, R, and E stand for total, regression, and error.

Because s_y^2 =SST/(n–1) and $s_{\widehat{y}}^2$ =SSR/(n–1)

$$R^2 = \frac{s_{\widehat{y}}^2}{s_y^2} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Each sum of squares comes with associated degrees of freedom, telling how many quantities used in their calculation can vary freely without changing any estimators of population parameters used in the same calculation.

$$\text{DFT} = n - 1$$

(n y-values minus one for calculating $\bar{y} = \sum y_i$)

$$\text{DFR} = 1$$

(There are n different $\widehat{y}_i$, but they are all produced by varying the single variable x.)

$$\text{DFE} = n - 2$$

(n y-values minus 2 for calculating b_0 and b_1.)

Just like SST is the sum of SSR and SSE, the total degrees of freedom is the sum of the degrees of freedom for the regression model and for the error:

$$\text{DFT} = \text{DFR} + \text{DFE},$$

The ratio of the sum of squares to the degrees of freedom is called the mean square:

$$\text{mean square} = \frac{\text{sum of squares}}{\text{degrees of freedom}}.$$

We know already MST$= \sum (y_i - \bar{y})^2/(n-1)$, which is just the sample variance s_y^2.

$$\text{MSE} = \frac{\sum (y_i - \widehat{y}_i)^2}{n-2}$$

is called the mean square error. Finally,

$$\text{MSR} = \frac{\sum (\widehat{y}_i - \bar{y})^2}{1} = \text{SSR}.$$

These can be used to assess whether $\beta_1 \neq 0$ out of sample, as is shown on the next slide.

ANOVA F-test for simple linear regression

Recall that while the methods of least squares makes no assumptions about the data generating process behind the observations x_i and y_i and may thus always be applied, hypothesis tests about the coefficients of the regression line $\widehat{y} = \beta_0 + \beta_1 x$ require the error terms $e_i = y_i - (b_0 + b_1 x_i)$ to be independent and normally distributed with mean 0 and common standard deviation σ. Under this assumption:

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F(1, n-2) \text{ under } H_0 : \ \beta_1 = 0.$$

When $\beta_1 \neq 0$, MSR tends to be large relative to MSE. So large values of F are evidence against H_0 in favour of the two-sided alternative $\beta_1 \neq 0$. For simple linear regression, this test is equivalent to the two-sided t-test for a significant slope coefficient to be discussed on the next slides.

Student t-tests for Regression Parameters

Recall that the regresion output

$$b_1 = r_{xy} \cdot \frac{s_y}{s_x}, \quad b_0 = \bar{y} - b_1 \bar{x}, \quad s_e^2 = \frac{\sum e_i^2}{n-2}$$

are only estimates of the true regression parameters β_1 and β_0 and σ^2, which vary from sample to sample. That is, we may regard them as sample-specific outcomes of random variables with associated expected values and variances.

Under certain conditions to be discussed soon, the expected values of these estimators are

$$E(b_1) = \beta_1, \quad E(b_0) = \beta_0, \text{ and } E(s_e^2) = \sigma^2,$$

which is why we chosed them in the first place. The standard deviations of the estimators for the regression coefficients turn out to be

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{SSX}} \quad \text{and} \quad \sigma_{b_0} = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SSX}},$$

where $SSX := \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Replacing σ with the standard error of the estimate s_e yields the standard errors for the estimated regression coefficients:

$$SE_{b_1} = \frac{s_e}{\sqrt{SSX}} \quad \text{and} \quad SE_{b_0} = s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SSX}},$$

which belong to the standard regression output. These may in turn be used in order to generate confidence intervals and tests for the regression slope and intercept as follows:

To test $H_0 : \beta_{1/0} = 0$, compute the test statistic

$$T_{1/0} = \frac{b_{1/0}}{SE_{b_{1/0}}}.$$

Reject H_0 against
- $H_1 : \beta_{1/0} \neq 0$ (two-sided) if $|T_{1/0}| \geq t_{\alpha/2}(n–2)$
- $H_1 : \beta_{1/0} \gtrless 0$ (one-sided) if $T_{1/0} \gtrless \pm t_\alpha(n–2)$.

A level $(1-\alpha)$ confidence interval for β_0 is

$$b_0 \pm t_{\alpha/2}(n-2) \cdot SE_{b_0}.$$

A level $(1-\alpha)$ confidence interval for β_1 is

$$b_1 \pm t_{\alpha/2}(n-2) \cdot SE_{b_1}.$$

Fuel Efficiency as a Function of Speed (continued)

Number of Observations Read	60
Number of Observations Used	60

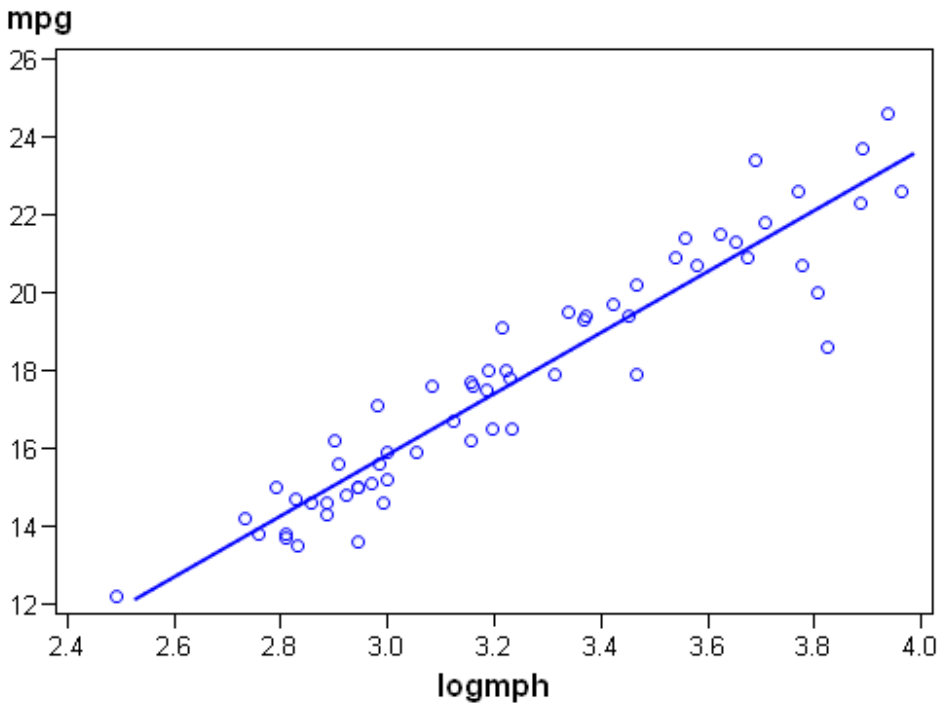
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	493.99177	493.99177	494.50	<.0001
Error	58	57.94073	0.99898		
Corrected Total	59	551.93250			

Root MSE	0.99949	**R-Square**	0.8950
Dependent Mean	17.72500	**Adj R-Sq**	0.8932
Coeff Var	5.63887		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	-7.79632	1.15491	-6.75	<.0001	-10.10813	-5.48451
logmph	logmph	1	7.87424	0.35410	22.24	<.0001	7.16543	8.58305



In the preceding example, the t-statistics came about by dividing the coefficient estimates $b_0 = -7.796$ and $b_1 = 7.874$ by their respective standard errors $SE_{b_0} = 1.155$ and $SE_{b_1} = 0.354$.

The 95% confindence intervals for β_0 and β_1 require the $\frac{\alpha}{2} = 2.5\%$ critical values of the t-distribution with $n - 2 = 58$ degrees of freedom (the same as for the residuals), which may be obtained from a table or by calling T.INV.2T(0.05,58) in Excel as

$$t_{\frac{\alpha}{2}}(n-2) = t_{0.025}(58) \approx 2.002.$$

The 95% confindence intervals for β_0 and β_1 are therefore:

$$\begin{aligned} b_1 \pm t_{0.025}(58) \cdot SE_{b_1} &= 7.874 \pm 2.002 \cdot 0.354 \\ &= (7.165,\ 8.583), \\ b_0 \pm t_{0.025}(58) \cdot SE_{b_0} &= -7.796 \pm 2.002 \cdot 1.155 \\ &= (-10.108,\ -5.485). \end{aligned}$$

The fact that zero is not included in any of these confidence intervals implies that we can reject $H_0 : \beta_1 = 0$ and $H_0 : \beta_0 = 0$ in both two-sided and one-sided tests.

Confidence intervals for the mean response and for individual observations

For any specific value of x, say x^*, the mean of the response y in this subpopulation is

$$\mu_y = \beta_0 + \beta_1 x^*,$$

which we estimate from the sample as

$$\widehat{\mu}_y = b_0 + b_1 x^*.$$

Alternatively we may interpret this expression as a prediction for an individual observation $\widehat{y} = b_0 + b_1 x^*$ for $x = x^*$. The prediction interval for an individual observation, however, is wider than the confidence interval for the mean due to the additional variation of individual responses about the mean response.

A level $(1-\alpha)$ confidence interval for $\widehat{\mu}_y$ is

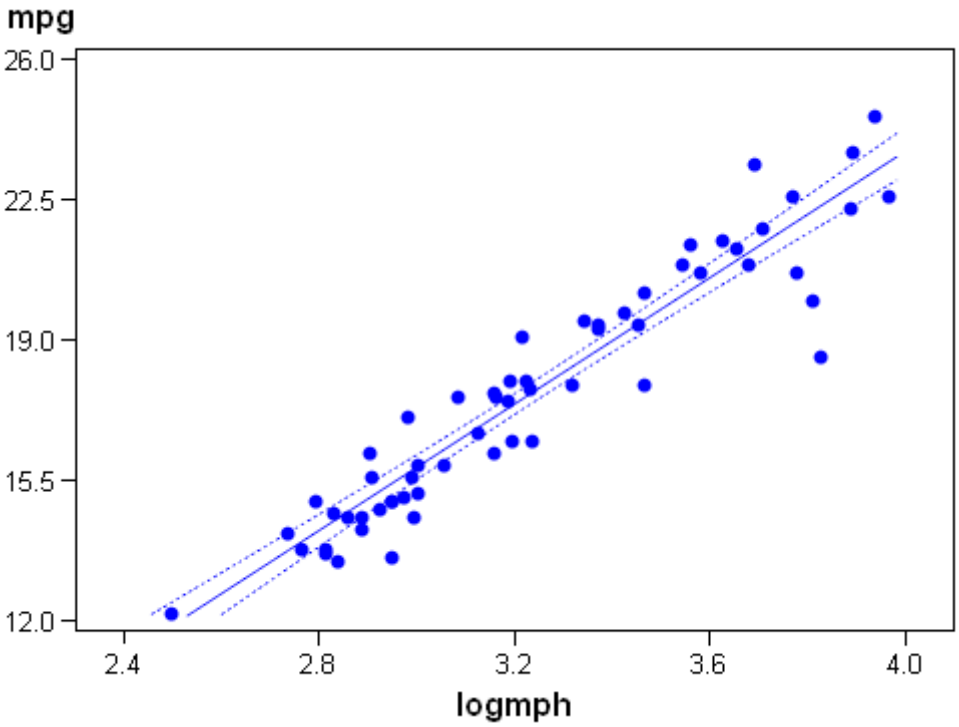
$$\widehat{\mu}_y \pm t_{\alpha/2}(n-2) \cdot SE_{\widehat{\mu}}, \quad SE_{\widehat{\mu}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SSX}}.$$

A level $(1-\alpha)$ prediction interval for $\widehat{y}$ is

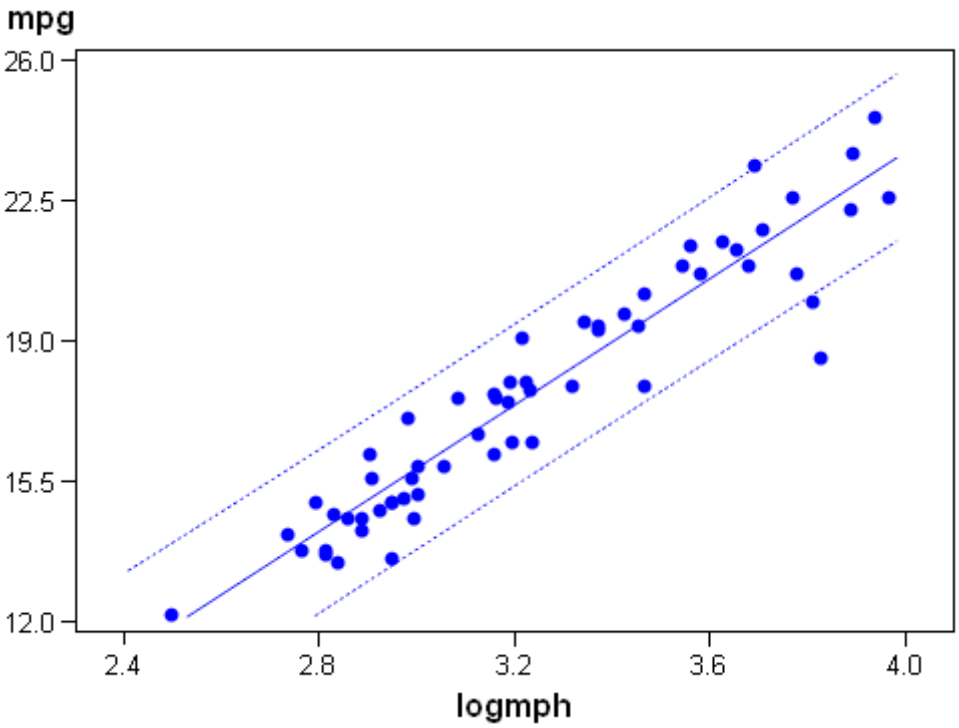
$$\widehat{y} \pm t_{\alpha/2}(n-2) \cdot SE_{\widehat{y}}, \; SE_{\widehat{y}} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SSX}}.$$

Fuel Efficieny as a Function of Speed (continued)

95% confidence limits for the mean response:



95% confidence limits for individual predictions:



1.4. Prerequisites in statistical inference

Statistical tests and confidence intervals are derived on the basis of some central assumptions. We usually assume that our observations are random samples of some prespecified distribution, most commonly the normal distribution or one of its derivatives. This, in turn, requires our data to have certain characteristics before a statistical method can be meaningfully applied.

A general precondition is that the statistical units/ observations are:

- independent of each other,
- are equally reliable,
- and the sample size is sufficiently large.

Beyond these general prerequesites, there are preconditions that apply to the specific statistical method to be used.

1. Contingeny tables

Pearson's χ^2 used in independence and homogeneity tests is approximately χ^2-distributed, if there are sufficiently many observations, that is:

- all expected frequencies are greater than 1,
- no more than 20% of the expected counts are smaller than 5.

If any of those conditions is not met, there are two options:

It's best to use Fishers exact test, which is available as an option from the Chi-Square Test of the Independence tool. It doesn't use the χ^2-approximation at all and works also in small samples, where the assumptions for the χ^2-test are not satisfied. It delivers always the precise p-value (so it's better than the χ^2-test), but the Real Statistics excel add in calculates it only for tables with no more than 9 cells. If you have a 2×2 table and no software is available, you should use the continuity correction discussed earlier.

2. Correlations

The tests for independence of ranks H_0 : $\rho_S = 0$ or $\tau = 0$ are exact and work also in small samples. The same holds for testing the linear correlation coefficient

$$H_0 : \ \rho = 0 \ (x, y \text{ are linearly independent})$$

with the t-test

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

when both x and y are normally distributed. Otherwise the test is only approximate and requires a sufficiently large sample size. For small r and large n we get the approximate z-test:

$$Z = r\sqrt{n} \sim N(0,1) \text{ under } H_0.$$

The Real Statistics data analysis tool allows you also to test $H_0 : \rho = \rho_0 \neq 0$ (known as Fisher's test) in large samples, but again the result is only approximate. The more ρ_0 deviates from 0, the larger the sample size has to be.

Example. Consider again the reputation and familiarity of 10 different brands:

reputation	7	2	2	9	0	1	9	0	1	0
familiarity	4	7	3	13	2	7	9	4	1	1

Pearson's linear correlation coefficient for this sample is $r = 0.7388$. We test $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$ by calculating the test statistics

$$t = \frac{0.7388 \cdot \sqrt{10-2}}{\sqrt{1 - 0.7388^2}} \approx 3.10.$$

From a statistical table or by calling T.INV.2T in excel we obtain the critical values

$$t_{0.02/2}(8) = 2.896 \quad \text{and} \quad t_{0.01/2}(8) = 3.355,$$

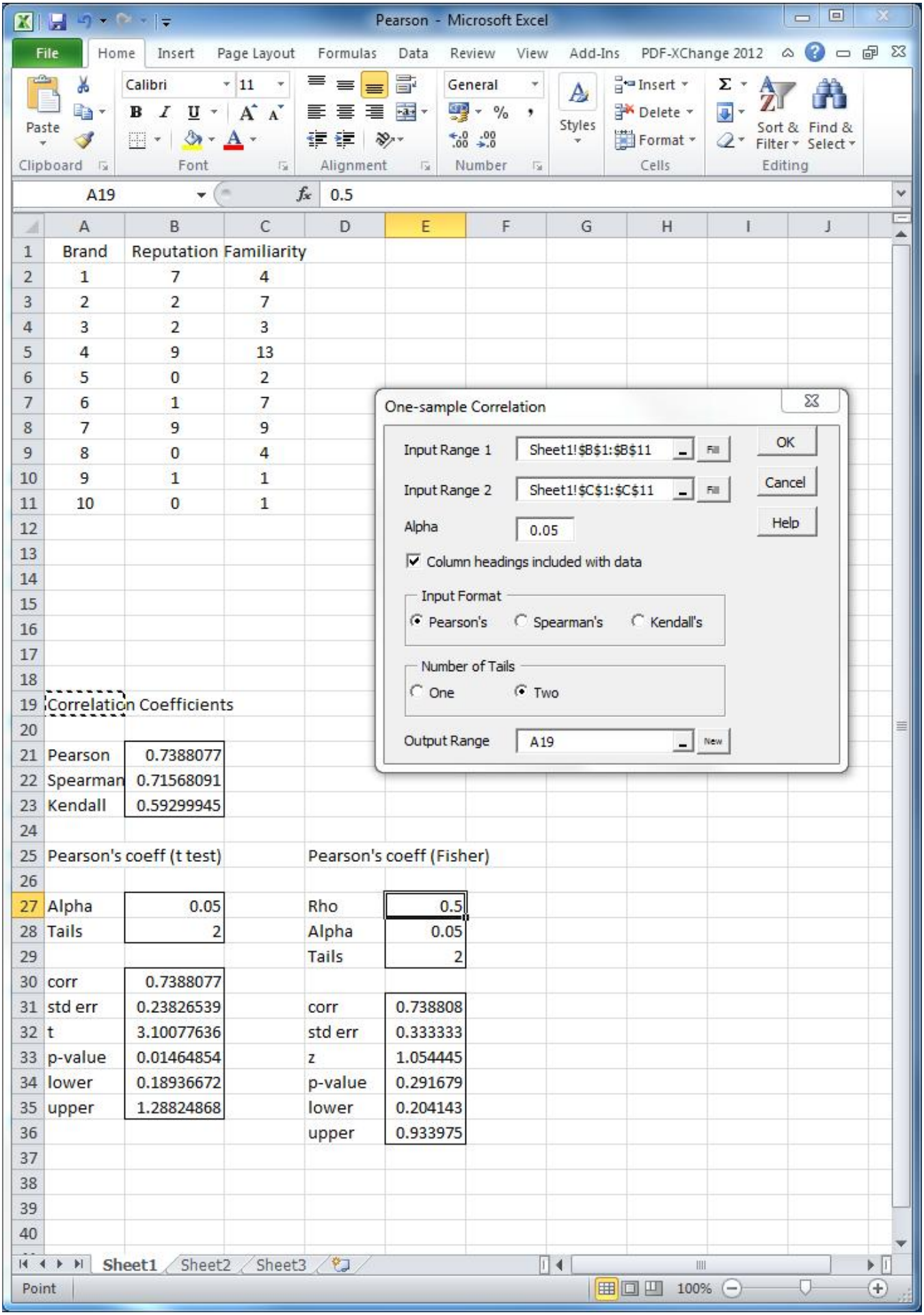
so the two-sided p-value is somewhere between 1% and 2%. The exact p-value is

$$\text{T.DIST.2T}(3.100776; 8) = 0.014649.$$

Applying the large sample approximation yields

$$Z = 0.7388\sqrt{10} = 2.336 \;\Rightarrow\; p \approx 2\%, \quad \text{since}$$

$P(Z \leq 2.336) = \text{NORMSDIST}(2.336) \approx 0.99$, such that $P(Z > |2.336|) \approx 2(1-0.99) = 0.02$.



	A	B	C	D	E	F	G	H	I	J
1	Brand	Reputation	Familiarity							
2	1	7	4							
3	2	2	7							
4	3	2	3							
5	4	9	13							
6	5	0	2							
7	6	1	7							
8	7	9	9							
9	8	0	4							
10	9	1	1							
11	10	0	1							
12										
13										
14										
15										
16										
17										
18										
19	Correlation Coefficients									
20										
21	Pearson	0.7388077								
22	Spearman	0.71568091								
23	Kendall	0.59299945								
24										
25	Pearson's coeff (t test)			Pearson's coeff (Fisher)						
26										
27	Alpha	0.05		Rho	0.5					
28	Tails	2		Alpha	0.05					
29				Tails	2					
30	corr	0.7388077								
31	std err	0.23826539		corr	0.738808					
32	t	3.10077636		std err	0.333333					
33	p-value	0.01464854		z	1.054445					
34	lower	0.18936672		p-value	0.291679					
35	upper	1.28824868		lower	0.204143					
36				upper	0.933975					
37										
38										
39										
40										

3. Regression Analysis

Example: (N. Weiss: Introductory Statistics) Consider the following sample of the prices (y in \$100) of cars as a function of their age (x in years):

x	5	4	6	5	5	5	6	6	2	7	7
y	85	103	70	82	89	98	66	95	169	70	48

A regression of price upon age yields:

$$\widehat{y} = 195.47 - 20.26x.$$

Note that the predictions of a regression line are not completely accurate as for example it predicts the price of 5 years old cars as

$$\$19547 - \$2026 \cdot 5 = \$9417,$$

whereas the true prices vary from car to car between \$8200 and \$9800. The distribution of a response variable Y for a specific value of the predictor variable X is called the conditional distribution (ehdollinen jakauma) of Y given the value $X = x$ with conditional mean $E(Y|X = x)$ and conditional variance $V(Y|X = x)$.

The assumptions for regression inferences are:

1. Normal populations:
For each value of the predictor variable X, the conditional distribution of the response variable Y is a normal distribution.

2. Population regression line:
There are constants β_0 and β_1 such that, for each value x of the predictor variable X, the conditional mean of the response variable is:

$$y = E(Y|X=x) = \beta_0 + \beta_1 x.$$

3. Equal standard deviations (variances):
The conditional standard deviations of the response variable Y are the same for all values of the predictor variable X:

$$V(Y|X=x) = \sigma^2 = \text{const.}$$

The condition of equal standard deviations is called homoscedasticity.

4. Independent observations:
The observations of the response variable are independent of one another.

When assumptions 1–4 for regression inferences hold, then the random errors

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

are independent and normally distributed with mean zero and variance σ^2. The statistical model for simple linear regression may therefore equivalently be stated as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \text{ iid } N(0, \sigma^2),$$

where 'iid' stands for independent and identically distributed. Note that this model has three parameters: β_0, β_1 and σ.

It turns out that the least square estimators

$$b_1 = r_{xy} \cdot \frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

are unbiased estimators of β_1 and β_0, which are themselves normally distributed. An unbiased estimator of the unknown variance σ^2 of the error term is given by the mean square error MSE=$\frac{\sum e_i^2}{n-2}$ where $e_i = y_i - (b_0 + b_1 x_i)$.

We define the standard error of the estimate as $\qquad s_e = \sqrt{\text{MSE}} = \sqrt{\dfrac{\text{SSE}}{n-2}} = \sqrt{\dfrac{\sum e_i^2}{n-2}}.$

Regression Diagnostics: Residual Analysis

1. Normality of Residuals

Normality of residuals may be checked either graphically, by considering the shape parameters of the distribution of residuals, or by performing statistical tests.

Graphs

A first visual check of the normality assumption is taking a look at the histogram of residuals. If the histogram is not bell-shaped, the residuals are not normally distributed.

A bell shaped distribution does, however, not guarantee that the distribution of residuals is normal, for example the t-distribution is also bell-shaped.

Excel's data analysis tool can produce histograms, but it is not very good at finding meaningful bin sizes and also clumsy to use.

Normal probability plots are plots with the ranked observations $x_{(i)}$ on the horizontal axis and the z-values $z_{q_i} = \Phi^{-1}(q_i)$ from the normal distribution corresponding to the respective quantile q_i (observed cumulative probability) of $x_{(i)}$ on the vertical axis. In this form the normal probability plot is also called a quantile-quantile (Q-Q) plot.

Alternatively one may plot the expected normal cumulative probabilities $\Phi(x_{(i)})$ on the vertical axis against the observed cumulative probabilities q_i on the horizontal axis in so called probability-probability (P-P) plots.

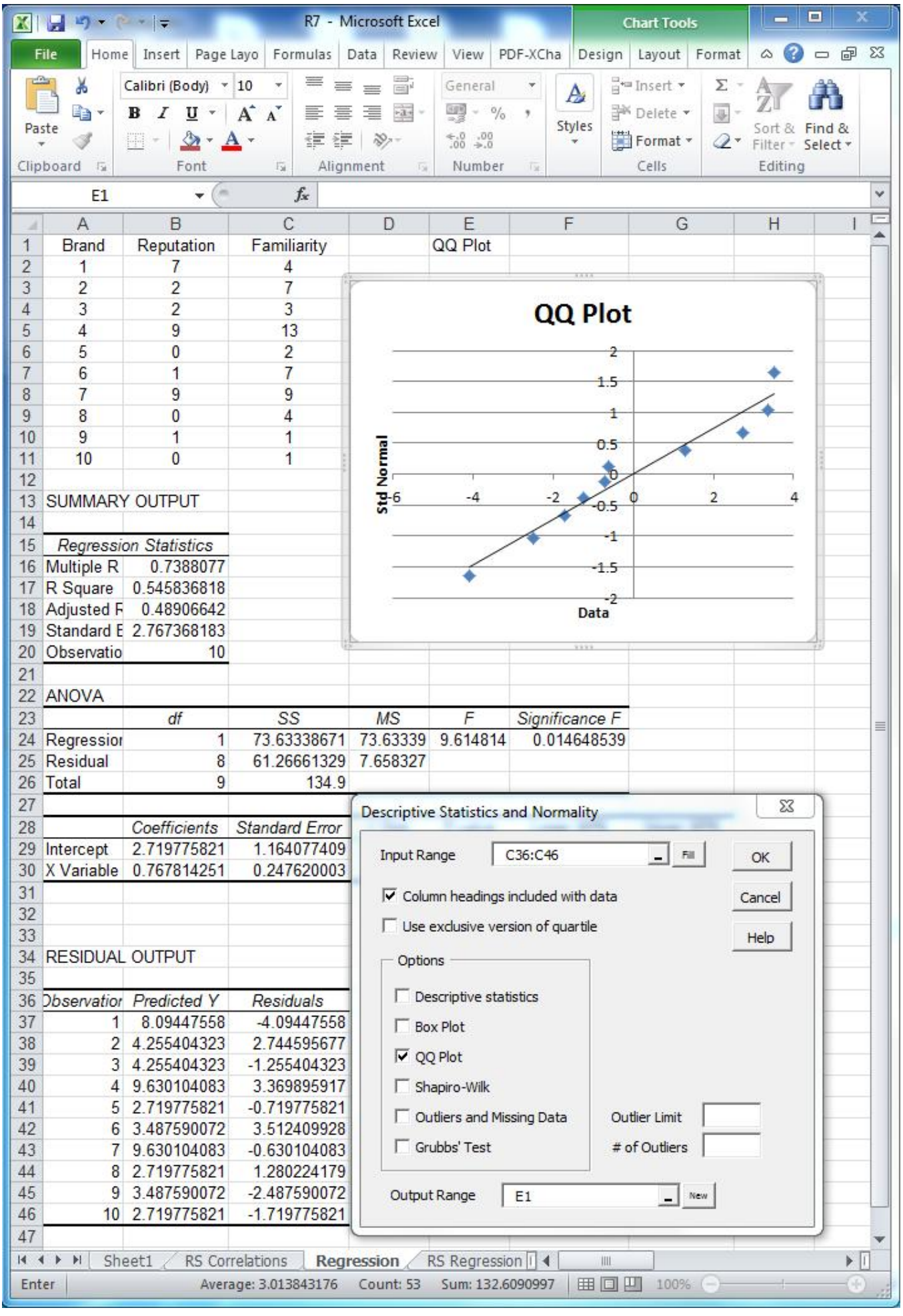
In either case, if the observations are normally distributed, then the normal probability plot should be a straight line. Deviations from this line allow for detection of outliers and qualitative identification of skewness and kurtosis.

Q-Q plots are more generally used than P-P plots, because they stress deviations in the tails, where hypothesis tests are usually done (P-P plots stress deviations in the center).

Note. Excel's regression tool has an option to produce a normal probability plot. This is not the relevant plot of the regression residuals though, but a much less useful normality plot for the unconditional y-values.

To get the relevant normality plot for the residuals, you must first produce those residuals from either Excel's or the Real Statistics data analysis tool and then apply Descriptive Statistics and Normality/ QQ Plot within the Real Statistics data analysis toolbox upon the residuals.

Alternatively you may obtain a P-P plot from Excel by running a second regression with arbitrary x-values and the previously obtained residuals as y-values, asking for a normal probability plot within the regression window and ignoring all other output.



	A	B	C	D	E	F	G	H	I
1	Brand	Reputation	Familiarity		QQ Plot				
2	1	7	4						
3	2	2	7						
4	3	2	3						
5	4	9	13						
6	5	0	2						
7	6	1	7						
8	7	9	9						
9	8	0	4						
10	9	1	1						
11	10	0	1						
12									
13	SUMMARY OUTPUT								
14									
15	*Regression Statistics*								
16	Multiple R	0.7388077							
17	R Square	0.545836818							
18	Adjusted F	0.48906642							
19	Standard E	2.767368183							
20	Observatio	10							
21									
22	ANOVA								
23		*df*	*SS*	*MS*	*F*	*Significance F*			
24	Regressior	1	73.63338671	73.63339	9.614814	0.014648539			
25	Residual	8	61.26661329	7.658327					
26	Total	9	134.9						
27									
28		*Coefficients*	*Standard Error*						
29	Intercept	2.719775821	1.164077409						
30	X Variable	0.767814251	0.247620003						
31									
32									
33									
34	RESIDUAL OUTPUT								
35									
36	*Observatior*	*Predicted Y*	*Residuals*						
37	1	8.09447558	-4.09447558						
38	2	4.255404323	2.744595677						
39	3	4.255404323	-1.255404323						
40	4	9.630104083	3.369895917						
41	5	2.719775821	-0.719775821						
42	6	3.487590072	3.512409928						
43	7	9.630104083	-0.630104083						
44	8	2.719775821	1.280224179						
45	9	3.487590072	-2.487590072						
46	10	2.719775821	-1.719775821						
47									

Shape Paprameters

Skewness

Recall that non-symmetric unimodal distribution are skewed to the right if the observations concentrate upon the lower values or classes (Md$< \bar{x}$), such that it has a long tail to the right, and skewed to the left, if the observations concentrate upon the higher values or classes (Md$> \bar{x}$), such that the distribution has a long tail to the left. This asymmetry is indicated by the (coefficient of) skewness:

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}.$$

In general, the distribution is skewed to the left (right) if g_1 is smaller (larger) than zero. Unimodal distributions with $g_1 = 0$ are symmetric. That is, $g_1 \neq 0$ (in particular when $|g_1| > 2\sqrt{6/n},\ n$ =sample size) is evidence that X is not normally distributed.

Skewness renders PP- and QQ-plots curved rather than linear.

Kurtosis

The (coefficient of) Kurtosis, defined as

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3,$$

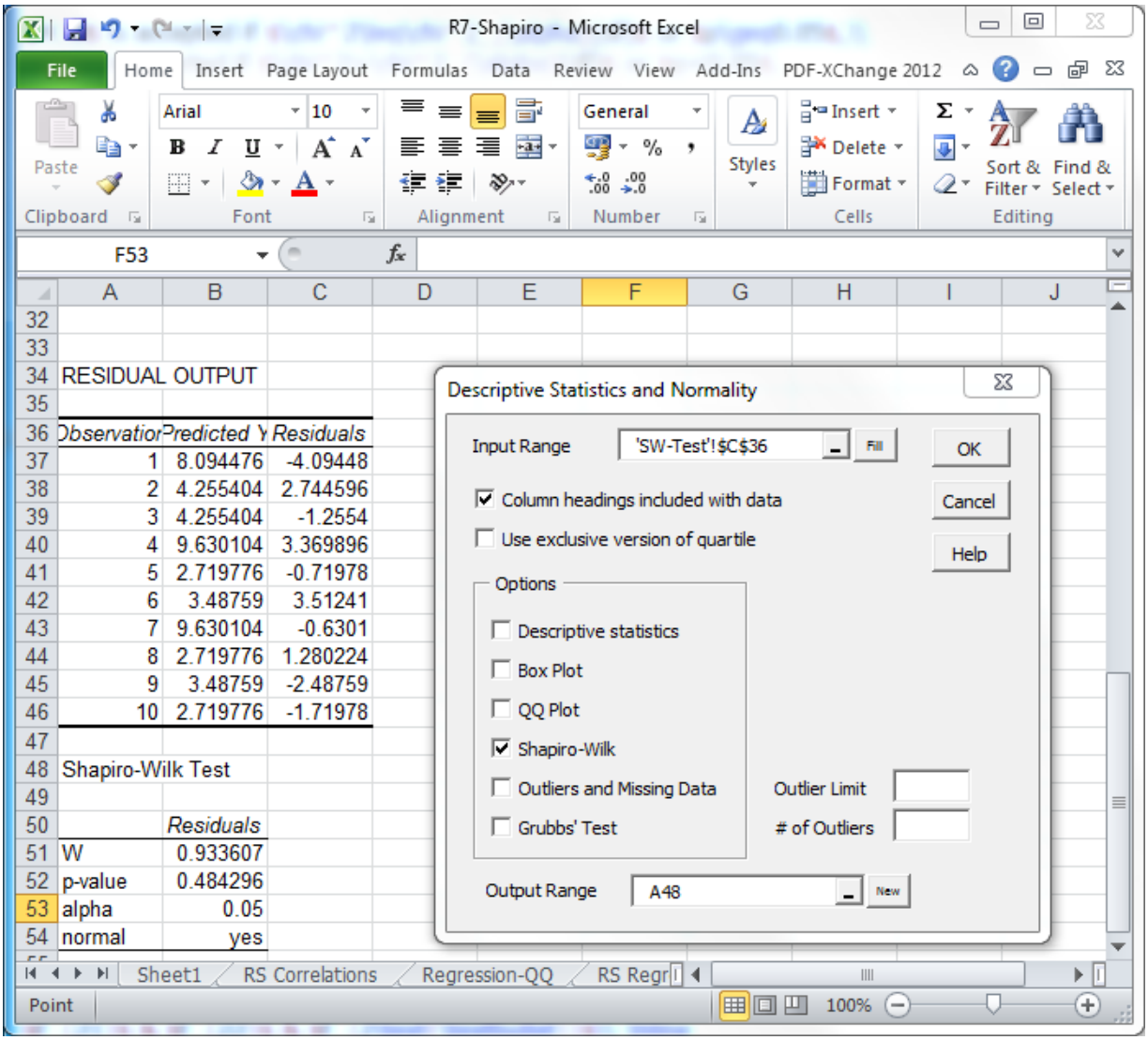
is a measure of peakedness (at least for unimodal distributions). That is, unimodal distributions with low kurtosis ($g_2 < 0$), called platykurtic, are rather evenly spread across all possible values or classes, and unimodal distributions with high kurtosis ($g_2 > 0$), called leptokurtic, have a sharp peak at their mode. Distributions with $g_2 \approx 0$ are called mesokurtic.

The kurtosis of the normal distribution is exactly zero. Therefore, the sign of g_2 tells for unimodal distributions whether they are more ($g_2 > 0$) or less ($g_2 < 0$) sharp peaked than the normal distribution. A clear warning sign against normality is when $|g_2| > 4\sqrt{6/n}$.

Kurtosis renders PP- and QQ-plots S-shaped.

Normality Tests

The most popular test for normality is called Shapiro-Wilk Test available from the Descriptive Statistics and Normality tool. The null hypothesis is that the data is normally distributed and the alternative hypothesis is that it is not. So small p-values (e.g. $p < 0.05$) imply that the data is not normally distributed.



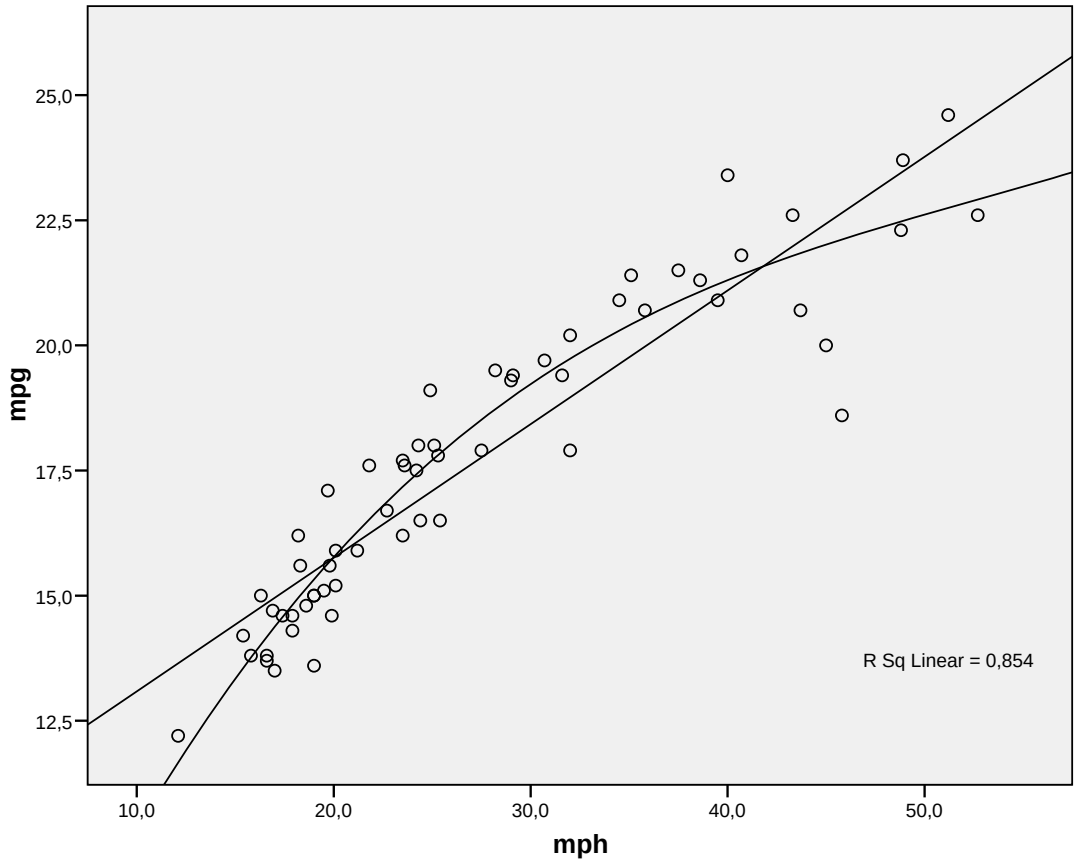
2. Linear Regression Relationship

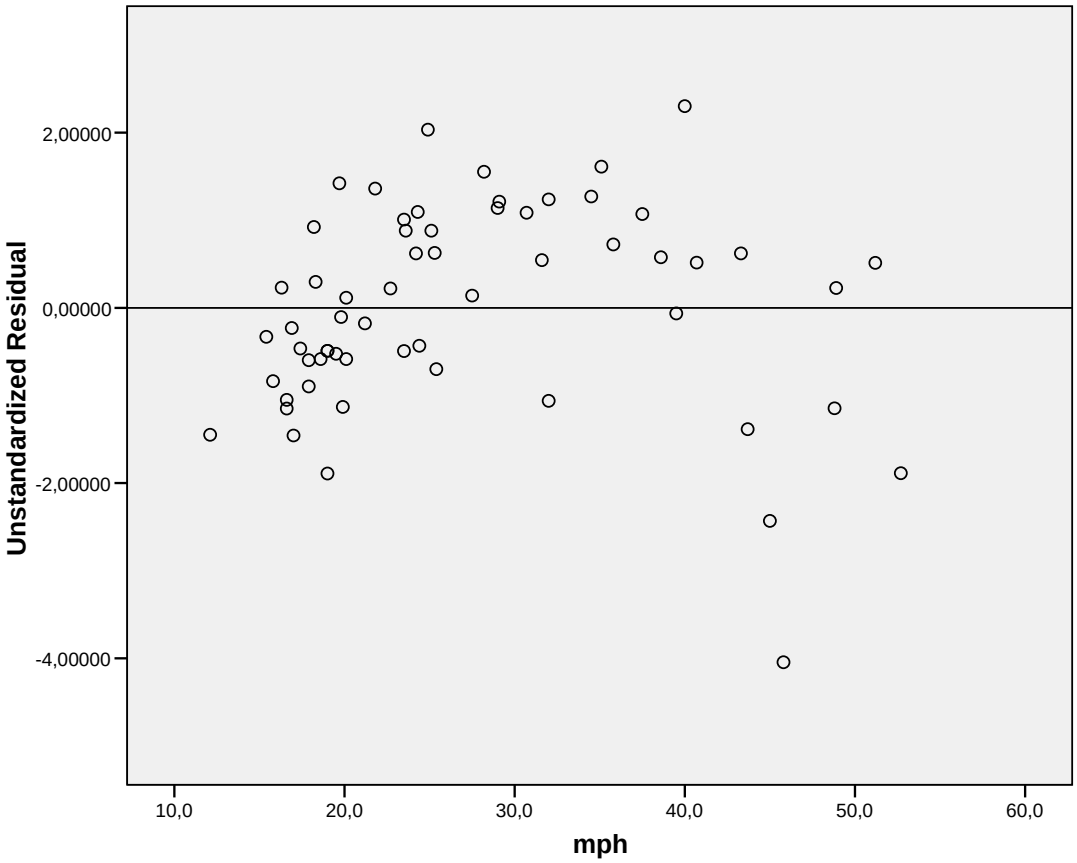
Deviations from straight-line relationships are immediately evident from the scatterplot of the predictor and the response variables. Such deviations are also visible as systematic patterns instead of random scatters in so called residual plots, where the residuals e_i are plotted either against the values of the predictor variable x_i or against the predicted response values $\widehat{y}_i$. These are available from Excel's regression tool, however usually you will have to rescale the axes before getting a readable result.

3. Constant residual variance

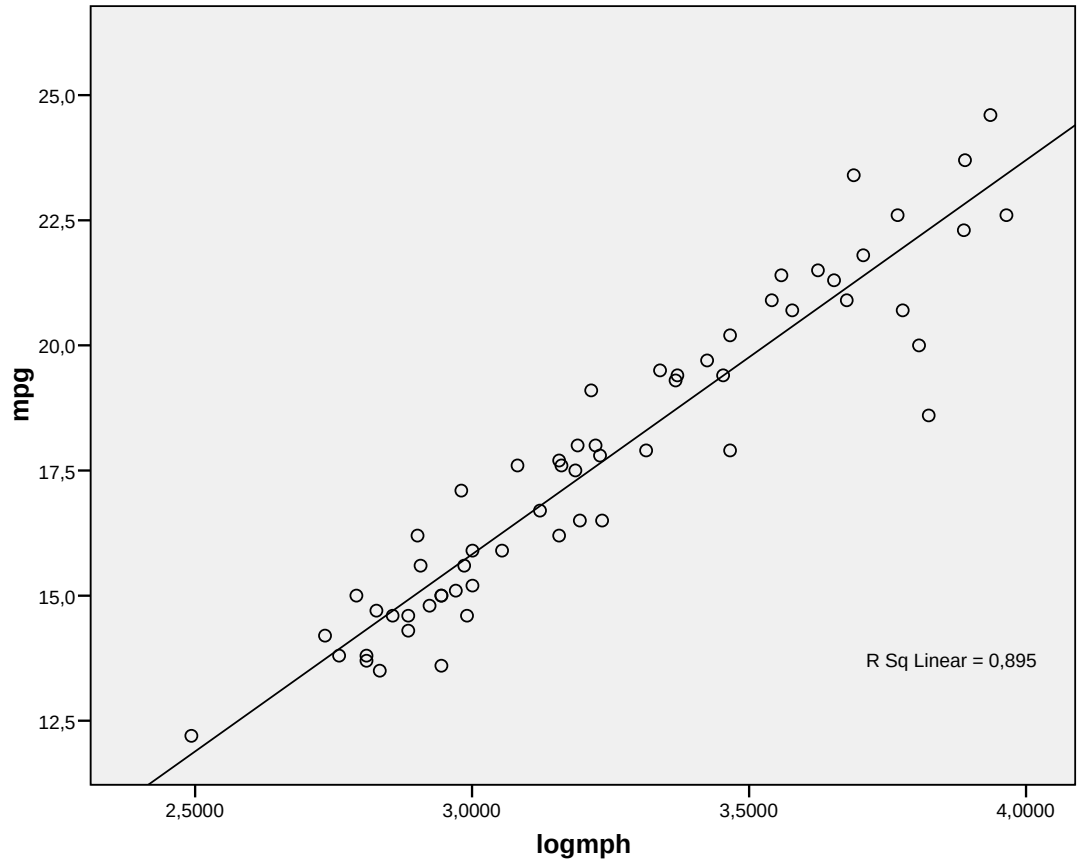
This may also be checked from residual plots: Any systematic pattern in the scatter of the residuals around zero contradicts the assumption of constant residual variance.

Fuel Efficiency as a Function of Speed (Moore/McCabe)

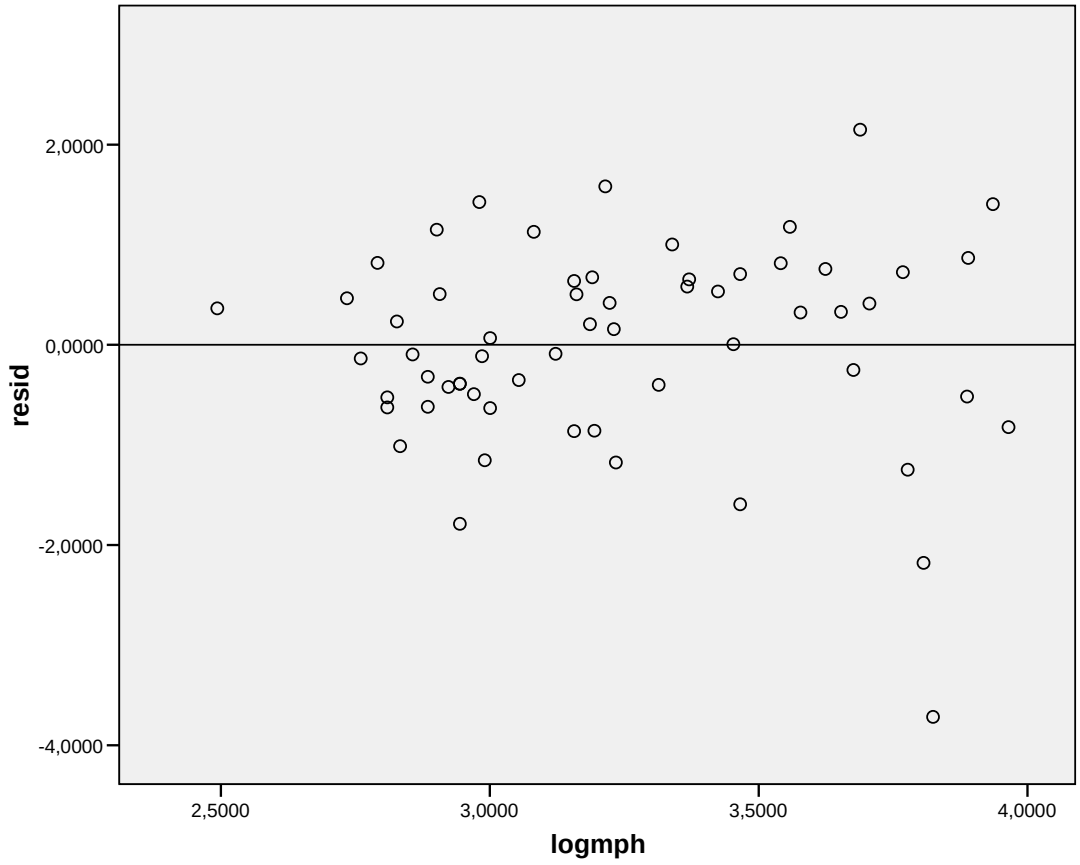




Miles per gallon versus logarithm of miles per hour



Residual Plot



4. Independent Observations: Residual Autocorrelation and the Durbin-Watson Test

We say that the regression residuals are autocorrelated if the correlation of any residual with any of its preceding residuals is nonzero. Residual autocorrelation is the most serious violation of the assumptions of the statistical model for linear regression.

Common reasons for residual autocorrelation:

- Two time-series are regressed upon each other.
- The dependence of Y upon X is nonlinear.
- Additional regressors are missing.
- There are trends or seasonal variation in the data.
- Missing data has been replaced by estimates.

The 1. Order Autocorrelation (1. kertaluvun autokorrelaatio) ρ_1, that is the autocorrelation of any residual ϵ_i with its preceding value ϵ_{i-1}, is assessed by the Durbin-Watson test statistic:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2},$$

which estimates $2(1 - \rho_1)$, that is:

$$\begin{array}{lcl} d \approx 2 & \Rightarrow & \text{Residuals uncorrelated (Ok),} \\ d < 2 & \Rightarrow & \epsilon_i \text{ positively autocorrelated,} \\ d > 2 & \Rightarrow & \epsilon_i \text{ negatively autocorrelated.} \end{array}$$

The critical values d_α depend upon the data and are therefore not known exactly. But there are tabulated upper limits d_U and lower limits d_L, which depend only upon the number of regressors (in our case 1) and the number of data points, such that

$$d_L < d_\alpha < d_U.$$

To perform a Durbin-Watson test:

1. Choose a significance level (e.g. $\alpha = 0.05$).
2. Calculate d
   (available from the Durbin-Watson test option within Real Statistics regression).

3. Look up:
   $d_L(\frac{\alpha}{2})$ and $d_U(\frac{\alpha}{2})$ for a two-sided test,
   $d_L(\alpha)$ and $d_U(\alpha)$ for a one-sided test.

4. (i) Two-sided: $H_0: \rho_1 = 0$ vs. $H_1: \rho_1 \neq 0$

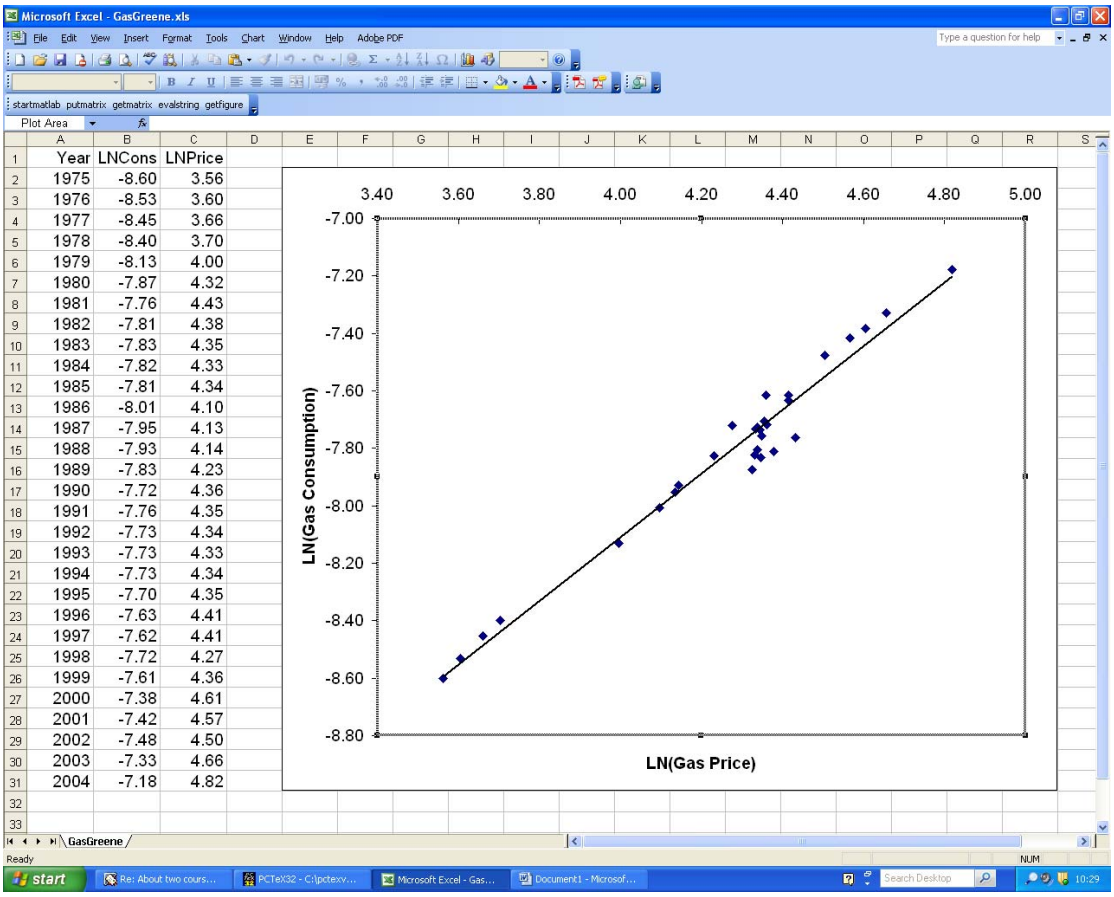
$d \leq d_L$ or $d \geq 4 - d_L$	$\Rightarrow$	reject H_0.
$d_U \leq d \leq 4 - d_U$	$\Rightarrow$	accept H_0.
otherwise	$\Rightarrow$	inconclusive.

   (ii) One-sided: $H_0: \rho_1 = 0$ vs. $H_1: \rho_1 > 0$

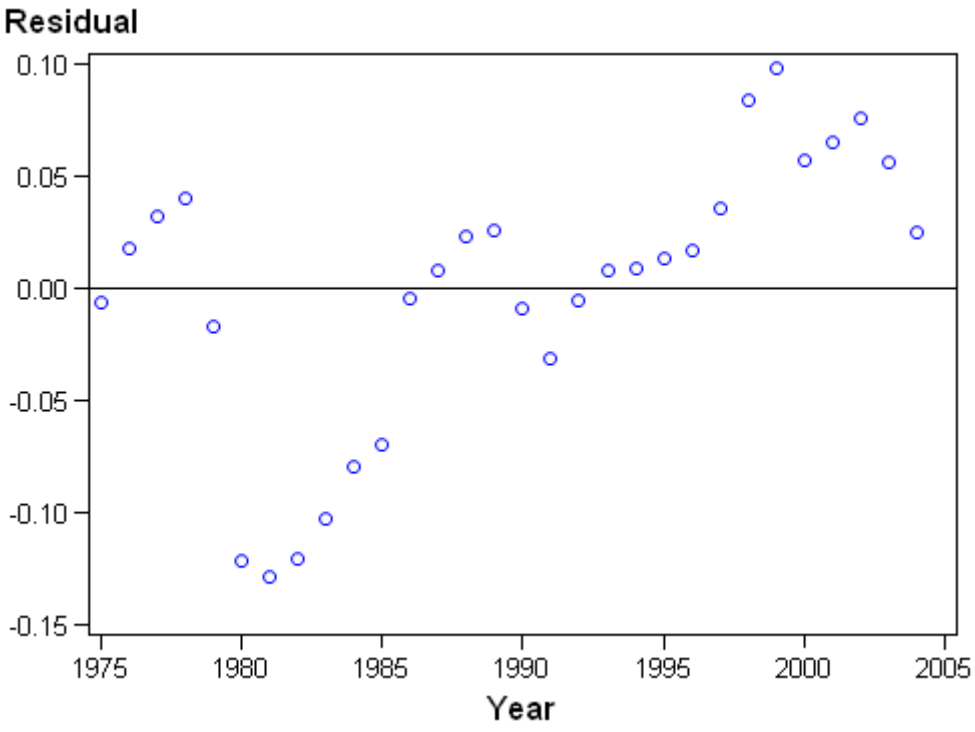
$d \leq d_L$	$\Rightarrow$	reject H_0.
$d \geq d_U$	$\Rightarrow$	accept H_0.
otherwise	$\Rightarrow$	inconclusive.

   (iii) One-sided: $H_0: \rho_1 = 0$ vs. $H_1: \rho_1 < 0$

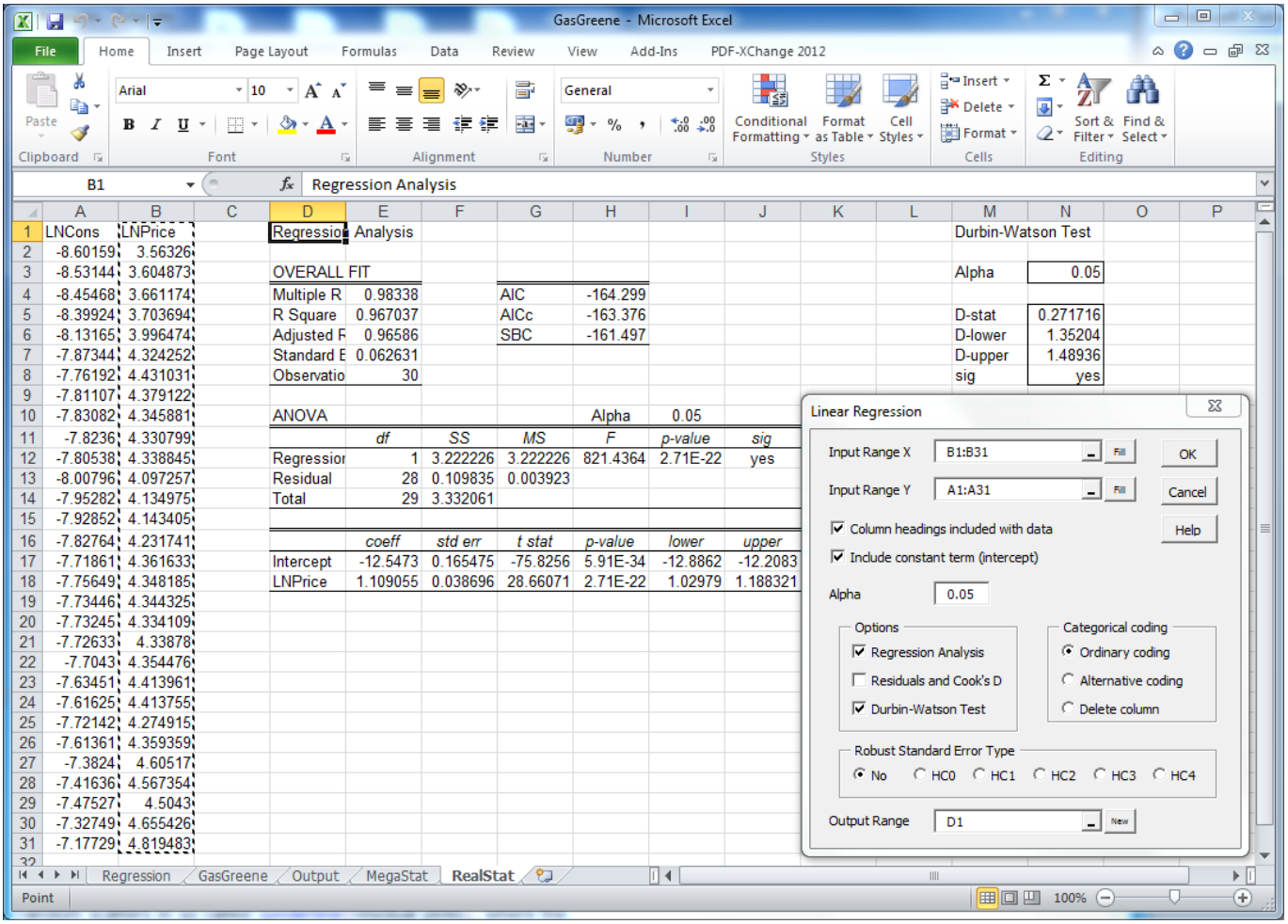
$d \geq 4 - d_L$	$\Rightarrow$	reject H_0.
$d \leq 4 - d_U$	$\Rightarrow$	accept H_0.
otherwise	$\Rightarrow$	inconclusive.



Year	LNCons	LNPrice
1975	-8.60	3.56
1976	-8.53	3.60
1977	-8.45	3.66
1978	-8.40	3.70
1979	-8.13	4.00
1980	-7.87	4.32
1981	-7.76	4.43
1982	-7.81	4.38
1983	-7.83	4.35
1984	-7.82	4.33
1985	-7.81	4.34
1986	-8.01	4.10
1987	-7.95	4.13
1988	-7.93	4.14
1989	-7.83	4.23
1990	-7.72	4.36
1991	-7.76	4.35
1992	-7.73	4.34
1993	-7.73	4.33
1994	-7.73	4.34
1995	-7.70	4.35
1996	-7.63	4.41
1997	-7.62	4.41
1998	-7.72	4.27
1999	-7.61	4.36
2000	-7.38	4.61
2001	-7.42	4.57
2002	-7.48	4.50
2003	-7.33	4.66
2004	-7.18	4.82



The Durbin-Watson test is an option within the Real Statistics Linear Regression tool:



Note that Real Statistics runs this as a one sided test against $H_1 : \rho_1 < 0$. For a two-sided test against $H_1 : \rho_1 \neq 0$, replace your α with $\alpha/2$ in the corresponding field and check also 4 minus the critical values from the output. For a one sided test against $H_1 : \rho_1 > 0$, use your original α and check only 4 minus the critical values from the output.