Nonparametric Tests

- ~~Nonparametric tests are useful when normality or the CLT can not be used.~~
- Nonparametric tests base inference on the **sign** or **rank** of the data as opposed to the actual data values.
- When normality can be assumed, nonparametric tests are less efficient than the corresponding t-tests.
- Sign test (binomial test on +/-)
- Wilcoxon signed rank (paired t-test on ranks)
- Wilcoxon rank sum (unpaired t-test on ranks)

Nonparametric Tests

In the tests we have discussed so far (for continuous data) we have assumed that either the measurements were **normally distributed** or the **sample size was large** so that we could apply the central limit theorem. What can be done when neither of these apply?

- Transform the data so that normality is achieved.
- Use another probability model for the measurements e.g. exponential, Weibull, gamma, etc.
- Use a nonparametric procedure

Nonparametric methods generally make fewer assumptions about the probability model and are, therefore, applicable in a broader range of problems.

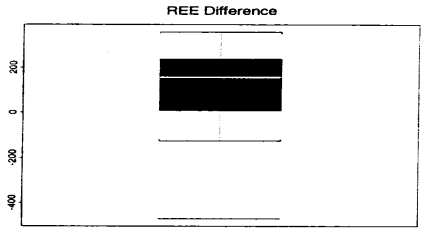
BUT! No such thing as a free lunch...

Nonparametric Tests

These data are REE (resting energy expenditure, kcal/day) for patients with cytic fibrosis and healthy individuals matched on age, sex, height and weight.

Pair	REE - CF	REE - healthy	Difference
1	1153	996	157
2	1132	1080	52
3	1165	1182	-17
4	1460	1452	8
5	1162	1634	-472
6	1493	1619	-126
7	1358	1140	218
8	1453	1123	330
9	1185	1113	72
10	1824	1463	361
11	1793	1632	161
12	1930	1614	316
13	2075	1836	239

Nonparametric Tests



	with #5	w/o #5
mean	99.9	147.6
std.dev	225.7	152.9
n	13	12
t	1.59	3.34

What's your conclusion?

Nonparametric Tests

Let's simplify by just looking at the direction of the difference ...

Pair	REE - CF	REE - healthy	Difference	Sign
1	1153	996	157	+
2	1132	1080	52	+
3	1165	1182	-17	-
4	1460	1452	8	+
5	1162	1634	-472	-
6	1493	1619	-126	-
7	1358	1140	218	+
8	1453	1123	330	+
9	1185	1113	72	+
10	1824	1463	361	+
11	1793	1632	161	+
12	1930	1614	316	+
13	2075	1836	239	+

Nonparametric Tests

We want to test:

$$H_o: \mu_d = 0$$
$$H_a: \mu_d > 0$$

Can we construct a test based only on the **sign** of the difference (no normality assumption)?

If $\mu_d = 0$ then we might expect half the differences to be positive and half the differences to be negative.

- What is a reasonable probability model for the sign of the differences?
- Re-express the Ho given above in terms of that probability model

Sign test

In this example we find 10 positive differences out of 13. What's the probability of that (or more extreme) if Ho is true?

```
. bitesti 13 10 .5

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
       13         10          6.5       0.50000      0.76923

  Pr(k >= 10)           = 0.046143  (one-sided test)
  Pr(k <= 10)           = 0.988770  (one-sided test)
  Pr(k <= 3 or k >= 10) = 0.092285  (two-sided test)
```

- What is the p-value for our sign test?
- What do you conclude ($\alpha = .05$)?

Sign test

- What we really tested was that the **median difference** was zero.
- Note that we didn't make any assumption about the distribution of the underlying data
- The hypothesis that the **Sign Test** addresses is:

  Ho : median difference = 0

  Ha : median difference > (<, ≠) 0

Q: If it is more generally applicable then why not always use it?

A: It is less **efficient** than the t-test when the population is normal. Using a sign test is like using only 2/3 of the data (when the "true" probability distribution is normal)

Sign test

Sign Test Overview:

1. Testing for a single sample (or differences from paired data).
2. Hypothesis is in terms of μ, the **median.**
3. Assign + to all data points where $X_i > \mu_o$ for H_o: $\mu = \mu_o$.
4. Let T= total number of +'s out of n observations.
5. Under H_0, T is binomial with n and p=1/2 (i.e. testing Ho: p = 0.5 on T is the same testing Ho: $\mu = \mu_o$ on X)
6. Get the p-value from binomial distribution or approximating normal, $T/n \sim N(1/2, 1/4n)$
7. This is a valid test of the median without assuming a probability model for the original measurements.

Nonparametric Tests

Q: Can we use some sense of the magnitude of the observations, without using the observations themselves?

A: Yes! We can consider the **rank** of the observations

Pair	REE - CF	REE - healthy	Difference	Sign	rank of $\|d_i\|$
1	1153	996	157	+	6
2	1132	1080	52	+	3
3	1165	1182	-17	-	2
4	1460	1452	8	+	1
5	1162	1634	-472	-	13
6	1493	1619	-126	-	5
7	1358	1140	218	+	8
8	1453	1123	330	+	11
9	1185	1113	72	+	4
10	1824	1463	361	+	12
11	1793	1632	161	+	7
12	1930	1614	316	+	10
13	2075	1836	239	+	9

Nonparametric Tests

A nonparametric test that uses the ranked data is the **Wilcoxon Signed-Rank Test**.

1. Rank the absolute value of the differences (from the null median).
2. Let R_+ equal the sum of ranks of the positive differences.
3. Then

$$E(R_+) = \frac{n(n+1)}{4}$$
$$V(R_+) = n(n+1)(2n+1)/24$$

4. Let

$$Z = \frac{R_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

5. Use normal approximation to the distribution of Z (i.e. compute p-value based on normal dist. i.e. $Z \sim N(0,1)$).

Wilcoxon Signed Rank Test

Note:

- If any $d_i = 0$ we drop them from the analysis (but assuming continuous data, so shouldn't be many).
- For "large" samples (number of non-zero $d_i \geq 15$), can use a normal approximation.
- If there are many "ties" then a correction to V(R+) must be made; computer does this automatically.
- Efficiency relative to t-test is about 95% if the true distribution is normal.

Wilcoxon Signed Rank Test

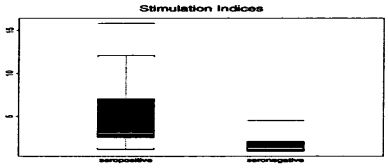
For the REE example we find R+ = 6+3+1+8+11+4+12+7+10+9 = 71

```
. signrank cf = healthy

Wilcoxon signed-rank test

        sign |      obs   sum ranks    expected
-------------+--------------------------------
    positive |       10          71        45.5
    negative |        3          20        45.5
        zero |        0           0           0
-------------+--------------------------------
         all |       13          91          91

unadjusted variance      204.75
adjustment for ties        0.00
adjustment for zeros       0.00
                        ----------
adjusted variance        204.75

Ho: cf = healthy
             z =   1.782
    Prob > |z| =   0.0747
```

Conclusion?

Nonparametric Tests
2 samples

The same issues that motivated nonparametric procedures for the 1-sample case arise in the 2-sample case, namely, non-normality in small samples, and the influence of a few observations. Consider the following data, taken from Miller (1991):



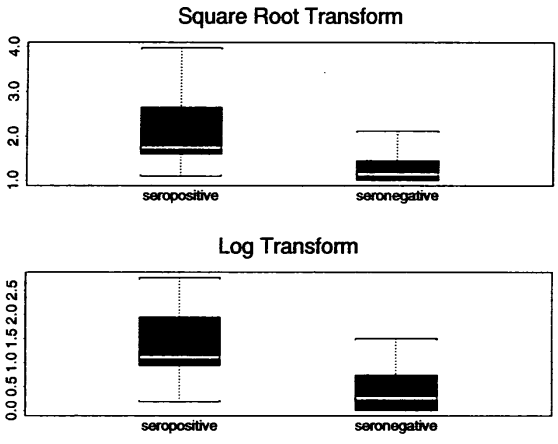
These data are immune function measurements obtained on healthy volunteers. One group consisted of 16 Epstein-Barr virus (EBV) seropositive donors. The other group consisted of 10 EBV seronegative donors. The measurements represent lymphocyte blastogenesis with p3HR-1 virus as the antigen (Nikoskelain et al (1978) *J. Immunology*, **121**:1239-1244).

Nonparametric Tests
2 samples

#	Seropositive	Seronegative
1	2.9	4.5
2	12.1	1.3
3	2.6	1.0
4	2.5	1.0
5	2.8	1.3
6	15.8	1.9
7	3.2	1.3
8	1.8	2.1
9	7.8	2.1
10	2.9	1.0
11	3.2	
12	8.0	
13	1.5	
14	6.3	
15	1.2	
16	3.5	

Nonparametric Tests
2 samples

Can we transform to normality?



Nonparametric Tests
2 samples

Does the 2-sample t statistic depend heavily on the transformation selected?

Does our interpretation depend on the transformation selected?

	RAW	SQRT	LOG
$\bar{Y}_1$	4.88	2.06	1.31
s_1^2	17.11	0.68	0.54
$\bar{Y}_2$	1.75	1.28	0.44
s_2^2	1.13	0.12	0.23
t	2.88	3.34	3.68
df	17	21	23
p-value	0.01	0.003	0.001

Nonparametric Tests
Wilcoxon Rank-Sum Test

Idea: If the distribution for group 1 is the same as the distribution for group 2 then pooling the data should result in the two samples "mixing" evenly. That is, we wouldn't expect one group to have many large values or many small values in the pooled sample.

Procedure:

1. Pool the two samples
2. Order and rank the pooled sample.
3. Sum the **ranks** for each sample.

   R_1 = rank sum for group 1

   R_2 = rank sum for group 2
4. The average rank is $(n_1+n_2+1)/2$.
5. Under H_o: same distribution, $E(R_1) = n_1(n_1+n_2+1)/2$ (why?)

6. The variance of R_1 is

$$V(R_1) = \left(\frac{n_1 n_2}{12}\right)(n_1 + n_2 + 1)$$

(an adjustment is required in the case of ties; this is done automatically by most software packages.)

7. We can base a test on the approximate normality of

$$Z = \frac{R_1 - E(R_1)}{\sqrt{V(R_1)}}$$

This is known as the **Wilcoxon Rank-Sum Test**.

Wilcoxon Rank-Sum Test

Order and rank the **pooled** sample ...

#	Sero +	Rank S+	Sero -	Rank S-
1	2.9	16.5	4.5	21.0
2	12.1	25.0	1.3	6.0
3	2.6	14.0	1.0	2.0
4	2.5	13.0	1.0	2.0
5	2.8	15.0	1.3	6.0
6	15.8	26.0	1.9	10.0
7	3.2	18.5	1.3	6.0
8	1.8	9.0	2.1	11.5
9	7.8	23.0	2.1	11.5
10	2.9	16.5	1.0	2.0
11	3.2	18.5		
12	8.0	24.0		
13	1.5	8.0		
14	6.3	22.0		
15	1.2	4.0		
16	3.5	20.0		
		273		78

Wilcoxon Rank-Sum Test

The sum of the ranks for group 1 is $R_1 = 273$

The null hypothesis is, H_o: same distribution,

```
. ranksum immune, by(ebv)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

         ebv |      obs    rank sum    expected
-------------+---------------------------------
           0 |       10          78         135
           1 |       16         273         216
-------------+---------------------------------
    combined |       26         351         351

unadjusted variance      360.00
adjustment for ties       -1.35
                     ----------
adjusted variance        358.65

Ho: immune(ebv==0) = immune(ebv==1)
             z =  -3.010
    Prob > |z| =   0.0026
```

Conclusion?

Compare to t-tests.

Wilcoxon Rank-Sum Test

Notes:

1. The Wilcoxon test is testing for a difference in <u>location</u> between the two distributions, not for a difference in <u>spread</u>. In fact, the actual hypothesis that is being tested is Ho: P(randomly chosen Y_1 > randomly chosen Y_2) = 0.5 (!).
2. Use of the normal approximation is valid if each group has <u>≥</u> 10 observations. Otherwise, the exact sampling distribution of R_1 can be used. Tables and computer routines are available in this situation.
3. The Wilcoxon rank-sum test is also known as the Mann-Whitney Test. These are equivalent tests.

Summary

- ~~Nonparametric tests are useful when normality or the CLT can not be used.~~
- Nonparametric tests base inference on the **sign** or **rank** of the data as opposed to the actual data values.
- When normality can be assumed, nonparametric tests are less efficient than the corresponding t-tests.
- Without imposing other assumptions on the distributions being compared (e.g., symmetry) there may not be an obvious summary statistic (e.g., mean, median, median pairwise mean) to interpret when the null hypothesis is rejected, or not.

Inference for two-way tables

General R x C tables

- Tests of **homogeneity** of a factor across groups or **independence** of two factors rely on **Pearson's X^2 statistic**.
- $\mathbf{X^2}$ is compared to a $\chi^2((r-1)x(c-1))$ distribution
- Expected cell counts should be larger than 5.

2 x 2 tables

- Cohort (prospective) data (H_0: relative risk for incidence = 1)
- Case-control (retrospective) data (H_0: odds ratio = 1)
- Cross-sectional data (H_0: relative risk for prevalence = 1)
- Paired binary data – McNemar's test (H_0: odds ratio = 1)
- For rare disease OR ≈ RR
- Fisher's exact test

Categorical Data

Types of Categorical Data

- Nominal
- Ordinal

Often we wish to assess whether two **factors** are related. To do so we construct an R x C table that **cross-classifies** the observations according to the two factors. Such a table is called a **contingency table.**

We can test whether the factors are "related" using a χ^2 test.

We will consider the special case of 2 x 2 tables in detail.

Categorical Data

Contingency tables arise from two different, but related, situations:

1) We *sample members of 2 (or more) groups* and classify each member according to some qualitative characteristic.

	Measurement of interest					
	1	2	3	4	5	total
Group 1	p_{11}	p_{12}	…			1.0
Group 2	p_{21}	p_{22}	…			1.0

The hypothesis is

H_0: groups are homogeneous ($p_{1j}=p_{2j}$ for all j)

H_A: groups are not homogeneous

Categorical Data

Example 1: From Doll and Hill (1952) - retrospective assessment of smoking frequency. The table displays the daily average number of cigarettes for lung cancer patients and control patients.

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	7	55	489	475	293	38	1357
	0.5%	4.1%	36.0%	35.0%	21.6%	2.8%	
Control	61	129	570	431	154	12	1357
	4.5%	9.5%	42.0%	31.8%	11.3%	0.9%	
Total	68	184	1059	906	447	50	2714

Categorical Data

Contingency tables arise from two different, but related, situations:

2) We *sample members of a population* and cross-classify each member according to two qualitative characteristics.

		Factor 1				
		1	2	3	4	Total
Factor 2	1	p_{11}	p_{12}	p_{13}	p_{14}	$p_{1.}$
	2	p_{21}	…			
	3	:				
	Total	$p_{.1}$				

The hypothesis is

H_0: factors are independent ($p_{ij}=p_{i.}p_{.j}$)

H_A: factors are not independent

Categorical Data

Example 2. Education versus willingness to participate in a study of a vaccine to prevent HIV infection if the study was to start tomorrow. Counts, row percents and row totals are given.

	definitely not	probably not	probably	definitely	Total
< high school	52 7.4%	79 11.3%	342 48.9%	226 32.3%	699
high school	62 6.9%	153 17.1%	417 46.6%	262 29.3%	894
some college	53 4.2%	213 16.8%	629 49.5%	375 29.5%	1270
college	54 4.9%	231 21.0%	571 51.9%	244 22.2%	1100
some post college	18 6.5%	46 16.6%	139 50.2%	74 26.7%	277
graduate/ prof	25 4.1%	139 22.8%	330 54.1%	116 19.0%	610
Total	264 5.4%	861 17.8%	2428 50.1%	1297 26.7%	4850

Test of Homogeneity

In **example 1** we want to test whether the smoking frequency is the same for each of the populations sampled. We want to test whether the **groups** are **homogeneous** with respect to a characteristic. The concept is similar to a t-test, but the response is categorical.

H_0: smoking frequency same in both groups

H_A: smoking frequency not the same

Q: What does H_0 predict we would observe if all we knew were the marginal totals?

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer							1357
Control							1357
Total	68	184	1059	906	447	50	2714

Test of Homogeneity

A: H_0 predicts the following **expectations**:

	Daily # cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	34	92	529.5	453	223.5	25	1357
Control	34	92	529.5	453	223.5	25	1357
Total	68	184	1059	906	447	50	2714

Each group has the same proportion in each cell as the overall **marginal proportion.** The "equal" expected number for each group is the result of the equal sample size in each group (what would change if there were half as many cases as controls?)

Test of Homogeneity

We have

- Observed counts, O_{ij}
- Expected counts (assuming Ho true), E_{ij}

➢ Heuristically, if the O_{ij} are "near" the E_{ij} that seems consistent with Ho; if the O_{ij} are "far" from E_{ij} we might suspect Ho is not true.

➢ The **Pearson's Chi-square Statistic** (X^2) measures the difference between the observed and expected counts and provides an overall assessment of Ho.

$$X^2 = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2\left((r-1)\times(c-1)\right)$$

Chi-square distribution with (r-1)*(c-1) degrees of freedom (BM table D)

Table entry for p is the critical value χ^* with probability p lying to its right.

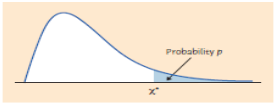


TABLE D Chi-square distribution critical values

	p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
00	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

Test of Homogeneity

Example 1. Smoking history vs lung cancer

```
. tabi 7 55 489 475 293 38 \ 61 129 570 431 154 12

           |                                col
       row |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         1 |         7         55        489        475        293 |     1,357
         2 |        61        129        570        431        154 |     1,357
-----------+-------------------------------------------------------+----------
     Total |        68        184      1,059        906        447 |     2,714


           |    col
       row |         6 |     Total
-----------+-----------+----------
         1 |        38 |     1,357
         2 |        12 |     1,357
-----------+-----------+----------
     Total |        50 |     2,714

          Pearson chi2(5) = 137.7193   Pr = 0.000
```

Conclusion?

Test of Independence

The **Chi-squared Test of Independence** is mechanically the same as the test for homogeneity. The difference is conceptual - the R x C table is formed by sampling from a population (not subgroups) and cross-classifying the factors of interest. Therefore, the null and alternative hypotheses are written as:

H_0: The two factors are independent

H_A: The two factors are not independent

Independence implies that each row has the same relative frequencies (or each column has the same relative frequency).

Example 2 is a situation where individuals are classified according to two factors. In this example, the assumption of independence implies that willingness to participate doesn't depend on the level of education (and visa-versa).

Test of Independence

	definitely not	probably not	probably	definitely	Total
< high school	52 7.4%	79 11.3%	342 48.9%	226 32.3%	699
high school	62 6.9%	153 17.1%	417 46.6%	262 29.3%	894
some college	53 4.2%	213 16.8%	629 49.5%	375 29.5%	1270
college	54 4.9%	231 21.0%	571 51.9%	244 22.2%	1100
some post college	18 6.5%	46 16.6%	139 50.2%	74 26.7%	277
graduate/ prof	25 4.1%	139 22.8%	330 54.1%	116 19.0%	610
Total	264 5.4%	861 17.8%	2428 50.1%	1297 26.7%	4850

Q: Based on the observed row proportions, how does the independence hypothesis look?

Q: How would the expected cell frequencies be calculated?

Q: How many degrees of freedom would the chi-square have?

Test of Independence

```
. tabi 52 79 342 226 \ 62 153 417 262 \ 53 213 629 375 \ 54 231 571
244 \ 18 46 139 74 \ 25 139 330 116

           |                    col
       row |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |        52         79        342        226 |       699
         2 |        62        153        417        262 |       894
         3 |        53        213        629        375 |     1,270
         4 |        54        231        571        244 |     1,100
         5 |        18         46        139         74 |       277
         6 |        25        139        330        116 |       610
-----------+--------------------------------------------+----------
     Total |       264        861      2,428      1,297 |     4,850

         Pearson chi2(15) =  89.7235   Pr = 0.000
```

Conclusion?

Summary
χ^2 Tests for R x C Tables

1. Tests of **homogeneity** of a factor across groups or **independence** of two factors rely on **Pearson's X^2 statistic**.
2. X^2 is compared to a $\chi^2((r-1)x(c-1))$ distribution (BM, table D or `display chiprob(df,X2)`).
3. Expected cell counts should be larger than 5.
4. We have considered a global test without using possible factor ordering. Ordered factors permit a test for trend (see Agresti, 1990).

2 x 2 Tables

Example 1: Pauling (1971)

Patients are randomized to either receive Vitamin C or placebo. Patients are followed-up to ascertain the development of a cold.

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

Q: Is treatment with Vitamin C associated with a reduced probability of getting a cold?

Q: If Vitamin C is associated with reducing colds, then what is the magnitude of the effect?

2 x 2 Tables

Example 2: Keller (AJPH, 1965)

Patients with (cases) and without (controls) oral cancer were surveyed regarding their smoking frequency (note: this table collapses over the smoking frequency categories shown in Keller).

	Case	Control	Total
Smoker	484	385	869
Non-Smoker	27	90	117
Total	511	475	986

Q: Is oral cancer associated with smoking?

Q: If smoking is associated with oral cancer, then what is the magnitude of the risk?

2 x 2 Tables

Example 3: Norusis (1988)

In 1984, a random sample of US adults were cross-classified based on their income and reported job satisfaction:

	Dissatisfied	Satisfied	Total
< $15,000	104	391	495
≥ $15,000	66	340	406
Total	170	731	901

Q: Is salary associated with job satisfaction?

Q: If salary is associated with satisfaction, then what is the magnitude of the effect?

2 x 2 Tables

Example 4: Sartwell et al (1969)

Is oral contraceptive use associated with thromboembolism? 175 cases with blood clots of unknown origin were matched to controls based on age, race, time and place of hospitalization, parity, marital status and SES.

		Control OC Use	
		Yes	No
Case OC Use	Yes	10	57
	No	13	95

Q: Is OC use associated with thromboembolism?

Q: If OC use is associated with thromboembolism then what is the magnitude of the effect?

2 x 2 Tables

Each of these tables can be represented as follows:

	D	not D	Total
E	a	b	$(a + b) = n_1$
not E	c	d	$(c + d) = n_2$
Total	$(a + c) = m_1$	$(b + d) = m_2$	N

The question of association can be addressed with **Pearson's** X^2 (except for example 4) We compute the **expected** cell counts as follows:

Expected:

	D	not D	Total
E	n_1m_1/N	n_1m_2/N	$(a + b) = n_1$
not E	n_2m_1/N	n_2m_2/N	$(c + d) = n_2$
Total	$(a + c) = m_1$	$(b + d) = m_2$	N

2 x 2 Tables

Recall, Pearson's chi-square is given by:

$$X^2 = \sum_{i=1}^{4} (O_i - E_i)^2 / E_i$$

Q: How does this X^2 test in Example 1 compare to simply using the 2 sample binomial test of

$$H_0 : P(D \mid E) = P(D \mid \overline{E})?$$

Q: How does the X^2 test in Example 2 compare to simply using the 2 sample binomial test of

$$H_0 : P(E \mid D) = P(E \mid \overline{D})?$$

2 x 2 Tables – Prospective study

Example 1: Pauling (1971)

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

H_0 : probability of disease does not depend on treatment
H_A : probability of disease does depend on treatment

2 x 2 Tables – Prospective study

```
. csi 17 31 122 109

                |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
          Cases |        17          31  |         48
       Noncases |       122         109  |        231
----------------+------------------------+------------
          Total |       139         140  |        279
                           :
                           :
                           :
                            chi2(1) =     4.81  Pr>chi2 = 0.0283
```

The X^2 value is 4.81 and the p-value is $P(\chi^2(1) > 4.81) = 0.028$. Therefore, using $\alpha = .05$, we reject the hypothesis that the risk of disease is equal in both treatment groups and conclude that vitamin C is protective.

How does this compare to the two sample test of binomial proportions?

```
. prtesti 139 .1223 140 .2214

Two-sample test of proportion                     x: Number of obs =      139
                                                  y: Number of obs =      140
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |      .1223   .0277894                      .0678338    .1767662
           y |      .2214   .0350899                      .1526251    .2901749
-------------+----------------------------------------------------------------
        diff |     -.0991    .044761                       -.18683     -.01137
             |  under Ho:   .0451895    -2.19   0.028
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =  -2.1930
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0142         Pr(|Z| < |z|) = 0.0283          Pr(Z > z) = 0.9858
```

Therefore, we reject H_0 **with the exact same result as the χ^2 test**. (Note: $2.19^2 = 4.81$)

2 x 2 Tables – Prospective Study

Example 1 fixed the number of E and not E, then evaluated the disease status after a fixed period of time. This is a **prospective study**. Given this design we can estimate the **relative risk**:

$$RR = \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

The range of RR is [0, ∞). By taking the logarithm, we have (- ∞, +∞) as the range for ln(RR) and a better approximation to normality for the estimated $\ln(\hat{RR})$:

$$\ln(\hat{RR}) = \ln\left(\frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \overline{E})}\right)$$

$$= \ln\left(\frac{a/n_1}{c/n_2}\right)$$

$$\ln(\hat{RR}) \sim N\left(\ln(p_1/p_2), \frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}\right)$$

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

The estimated relative risk is:

$$\hat{RR} = \frac{\hat{P}(D|E)}{\hat{P}(D|\overline{E})} = \frac{17/139}{31/140}$$
$$= 0.55$$

We can obtain a confidence interval for the relative risk by first obtaining a confidence interval for the log RR. For Example 1, a 95% confidence interval for the log relative risk is given by:

$$\ln(\hat{RR}) \pm 1.96 \times \sqrt{\frac{1-\hat{p}_1}{\hat{p}_1 n_1} + \frac{1-\hat{p}_2}{\hat{p}_2 n_2}}$$

$$\ln(0.55) \pm 1.96 \times \sqrt{\frac{122}{(17)(139)} + \frac{109}{(31)(140)}}$$

The resulting 95% CI for the log RR is

-0.593 ± 1.96 × 0.277
-0.593 ± 0.543
(-1.116, -0.050)

To obtain a 95% confidence interval for the **relative risk** we exponentiate the end-points of the interval for the **log - relative risk**. Therefore,

(exp(-1.116), exp(-0.050))

(.33 , .95)

is a 95% confidence interval for the relative risk.

2 x 2 Tables – Prospective Study

```
. csi 17 31 122 109

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        17          31  |         48
        Noncases |       122         109  |        231
-----------------+------------------------+------------
           Total |       139         140  |        279
                 |                        |
            Risk |  .1223022    .2214286  |    .172043
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0991264       |   -.1868592   -.0113937
      Risk ratio |         .5523323       |    .3209178    .9506203
 Prev. frac. ex. |         .4476677       |    .0493797    .6790822
 Prev. frac. pop |         .2230316       |
                 +-------------------------------------------------
                              chi2(1) =     4.81  Pr>chi2 = 0.0283
```


2 x 2 Tables – Case-Control Study

In Example 2 we fixed the number of **cases** and **controls** then ascertained exposure status (i.e. we measured P(E|D)). Such a design is known as **case-control study**. Based on this we are able to estimate P(E|D) but not P(D|E). That means we can't (directly) estimate the relative risk ☹.

However, we can estimate the **exposure odds ratio** 😐 **…** What's an odds ratio?

$$OR = \frac{P(E \mid D)/(1 - P(E \mid D))}{P(E \mid \overline{D})/(1 - P(E \mid \overline{D}))}$$

… and Cornfield (1951) showed the exposure odds ratio is equivalent to the **disease odds ratio** 😐 **…** That's odd!

$$\frac{P(E \mid D)/(1 - P(E \mid D))}{P(E \mid \overline{D})/(1 - P(E \mid \overline{D}))} = \frac{P(D \mid E)/(1 - P(D \mid E))}{P(D \mid \overline{E})/(1 - P(D \mid \overline{E}))}$$

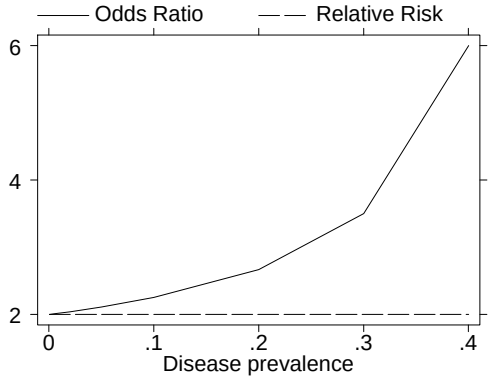
Odds Ratio

… and, for rare diseases, P(D | E) ≈ 0 so that the **disease odds ratio** approximates the relative risk! ☺

$$\frac{P(D \mid E)/(1-P(D \mid E))}{P(D \mid \overline{E})/(1-P(D \mid \overline{E}))} \approx \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

- Case-Control data ⇒ able to estimate the exposure odds ratio ⇒ exposure odds ratio equal to the disease odds ratio ⇒ for rare diseases, odds ratio approximates the relative risk.

For rare diseases, the sample odds ratio approximates the population relative risk.



2 x 2 Tables – Case-Control Study

Like the relative risk, the odds ratio has [0, ∞) as its range. The **log odds ratio** has (- ∞, +∞) as its range and the normal distribution is a good approximation to the sampling distribution of the estimated log odds ratio.

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

$$\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{ad}{bc}$$

Confidence intervals are based upon:

$$\ln\left(\hat{OR}\right) \sim N\left(\ln(\text{OR}), \frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)} + \frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}\right)$$

Therefore, a (1 - α) confidence interval for the log odds ratio is given by:

$$\ln\left(\frac{ad}{bc}\right) \pm z_{1-\alpha/2} \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

2 x 2 Tables – Case-Control Study

```
. cci 484 27 385 90

                                                        Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |       484          27  |        511       0.9472
        Controls |       385          90  |        475       0.8105
-----------------+------------------------+------------------------
           Total |       869         117  |        986       0.8813
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         4.190476       |    2.633584    6.836229 (exact)
 Attr. frac. ex. |         .7613636       |    .6202893    .8537205 (exact)
 Attr. frac. pop |          .721135       |
                 +-----------------------------------------------------
                              chi2(1) =    43.95  Pr>chi2 = 0.0000
```

Interpreting Odds ratios

1. What is the outcome of interest? (i.e. disease)
2. What are the two groups being contrasted? (i.e. exposed and unexposed)

$$\text{OR} = \frac{\text{odds of OUTCOME in EXPOSED}}{\text{odds of OUTCOME in UNEXPOSED}}$$

- Similar to RR for rare diseases
- Meaningful for both cohort and case-control studies
- OR > 1 $\Rightarrow$ increased odds of OUTCOME with EXPOSURE
- OR < 1 $\Rightarrow$ decreased odds of OUTCOME with EXPOSURE

Interpreting Odds ratios

Be aware of how the table is laid out …

	Case	Control	Total
Non-Smoker	27	90	117
Smoker	484	385	869
Total	511	475	986

Odds ratio = .239 $\Rightarrow$ Interpret.

2 x 2 Tables – Cross-sectional Study

Example 3 is an example of a **cross-sectional** study since only the total for the table is fixed in advance. The row totals or column totals are not fixed in advance.

Either the relative risk or odds ratio may be used to summarize the association when using a cross-sectional design.

The major distinction from a **prospective** study is that a **cross-sectional** study will reveal the number of cases currently in the sample. These are known as **prevalent** cases. In a prospective study we count the number of new cases, or **incident** cases.

Study	Probability	Description
Cohort	incidence	probability of obtaining the disease
Cross-sectional	prevalence	probability of having the disease

2 x 2 Tables – Cross-sectional Study

```
. csi 104 391 66 340, or

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |       104         391  |        495
        Noncases |        66         340  |        406
-----------------+------------------------+------------
           Total |       170         731  |        901
                 |                        |
            Risk |  .6117647    .5348837  |   .5493896
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |          .076881       |   -.0048155    .1585775
      Risk ratio |         1.143734       |    .9967902     1.31234
 Attr. frac. ex. |         .1256708       |   -.0032201    .2380023
 Attr. frac. pop |         .0264036       |
      Odds ratio |         1.370224       |    .9752222    1.925102 (Cornfield)
                 +-------------------------------------------------
                              chi2(1) =     3.29  Pr>chi2 = 0.0696
```

Fisher's Exact Test

Motivation: When a 2×2 table contains cells that have fewer than 5 expected observations, the normal approximation to the distribution of the log odds ratio (or other summary statistics) is known to be poor. This can lead to incorrect inference since the p-values based on this approximation are not valid.

Solution: Use Fisher's Exact Test

	D+	D-	Total
E+			n_1
E-			n_2
Total	m_1	m_2	N

Fisher's Exact Test

Example: Cardiovascular disease. A retrospective study is done among men aged 50-54 who died over a 1-month period. The investigators tried to include equal numbers of men who died from CVD and those that did not. Then, asking a close relative, the dietary habits were ascertained.

	High Salt	Low Salt	Total
non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

A calculation of the odds ratio yields:

Interpret. $$\text{OR} = \frac{2 \times 30}{5 \times 23} = 0.522$$

Fisher's Exact Test

Example: Cardiovascular disease.

If we consider the margins fixed, there are only a limited number of possible tables. Using the hypergeometric distribution, "we" can compute the probability of each table under Ho.

Possible Tables (with probability under Ho):

0		25
		35
7	53	60

.017

1		25
		35
7	53	60

.105

2		25
		35
7	53	60

.252

3		25
		35
7	53	60

.312

4		25
		35
7	53	60

.214

5		25
		35
7	53	60

.082

6		25
		35
7	53	60

.016

7		25
		35
7	53	60

.001

Fisher's Exact Test

To compute a p-value we then use the usual approach of summing the probability of all events (tables) as extreme or more extreme than the observed data.

- For a one tailed test we sum the probabilities of all tables with *a* less than or equal to (greater than or equal to) the observed *a*.
- For a two-tailed test of $p_1 = p_2$ we sum all tables that are less likely than the observed.

You will *never* do this by hand ….

Fisher Exact test using Stata

Fisher's exact test.

```
. cci 5 30 2 23,exact
                                                        Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |         5          30  |         35       0.1429
       Controls |         2          23  |         25       0.0800
----------------+------------------------+------------------------
          Total |         7          53  |         60       0.1167
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |         1.916667       |    .2789585    21.62382 (exact)
 Attr. frac. ex. |         .4782609       |   -2.584763    .9537547 (exact)
 Attr. frac. pop |          .068323       |
                +-------------------------------------------------
                              1-sided Fisher's exact P = 0.3747
                              2-sided Fisher's exact P = 0.6882
```


Fisher Exact test using Stata

The usual chi-squared test,
for comparison.

```
. cci 5 30 2 23
                                                        Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+------------------------
          Cases |         5          30  |         35       0.1429
       Controls |         2          23  |         25       0.0800
----------------+------------------------+------------------------
          Total |         7          53  |         60       0.1167
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
     Odds ratio |         1.916667       |    .2789585    21.62382 (exact)
 Attr. frac. ex. |         .4782609       |   -2.584763    .9537547 (exact)
 Attr. frac. pop |          .068323       |
                +-------------------------------------------------
                          chi2(1) =     0.56  Pr>chi2 = 0.4546
```

Paired Binary Data

Example 4 measured a binary response on matched pairs. This is an example of **paired binary data**. One way to display these data is the following:

	OC	No OC	Total
Case	67	108	175
Control	23	152	175
Total	90	260	350

Q: Can't we simply use X^2 Test of Homogeneity to assess whether this is evidence for an increase in knowledge?

A: NO!!! The X^2 tests assume that the rows are **independent** samples. In this design, the controls are constrained to be similar to the controls in many respects.

Paired Binary Data

For paired binary data we display the results as follows:

		Control OC	
		Yes	No
Case OC	Yes	n_{11}	n_{10}
	No	n_{01}	n_{00}

This analysis explicitly recognizes the heterogeneity of subjects. Thus, those that score (0,0) and (1,1) provide no information about the effect of OC use since they may be "weak" or "strong" individuals. These are known as the **concordant pairs**. The information regarding OC use is in the **discordant pairs**, (0,1) and (1,0).

p_1 = "success" probability for cases
p_2 = "success" probability for controls

$$H_0 : p_1 = p_2$$
$$H_A : p_1 \neq p_2$$

Paired Binary Data - McNemar's Test

Under the null, H_0 : $p_1 = p_2$, we expect equal numbers of "01" and "10" discordant pairs (i.e., $E[n_{01}] = E[n_{10}]$). Specifically, under the null:

$$M = n_{01} + n_{10}$$

$$n_{01} \mid M \sim Bin\left(M, \frac{1}{2}\right)$$

$$Z = \frac{n_{01} - M\frac{1}{2}}{\sqrt{M\frac{1}{2}\left(1-\frac{1}{2}\right)}}$$

Under H_0, $Z^2 \sim \chi^2(1)$, and forms the basis for **McNemar's Test for Paired Binary Responses**.

The odds ratio comparing the odds of OC use for cases to OC use for controls is estimated by:

$$\hat{OR} = \frac{n_{10}}{n_{01}}$$

Confidence intervals: see Breslow and Day (1981), sec. 5.2, or Armitage and Berry (1987), chap. 16.

Example 4:

		Control OC	
		Yes	No
Case OC	Yes	10	57
	No	13	95

We can test H_0: $p_1 = p_2$ using **McNemar's Test:**

$$Z = \frac{n_{01} - M\frac{1}{2}}{\sqrt{M\frac{1}{2}\left(\frac{1}{2}\right)}}$$

$$= \frac{13 - (13+57)/2}{\sqrt{(13+57)/4}}$$

$$= 5.26$$

Comparing 5.26^2 to a χ^2 (1) we find that $p < 0.001$. Therefore we reject the null hypothesis of equal OC use probabilities for cases and controls.

We estimate the odds ratio as $\hat{OR} = 57/13 = 4.38$.

Matched case-control data in Stata

```
. mcci 10 57 13 95

                | Controls               |
Cases           |   Exposed   Unexposed  |      Total
----------------+------------------------+------------
         Exposed |        10          57  |         67
       Unexposed |        13          95  |        108
----------------+------------------------+------------
           Total |        23         152  |        175

McNemar's chi2(1) =     27.66    Prob > chi2 = 0.0000
Exact McNemar significance probability       = 0.0000

Proportion with factor
        Cases       .3828571
        Controls    .1314286     [95% Conf. Interval]
                   ---------     --------------------
        difference  .2514286      .1597329   .3431243
        ratio       2.913043      1.918355   4.423488
        rel. diff.  .2894737      .1985361   .3804113

        odds ratio  4.384615      2.371377   8.731311   (exact)
```


Paired Binary Data

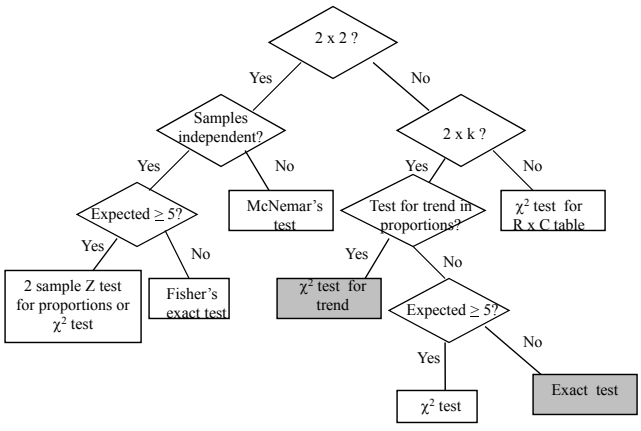
Paired data analyses arise in a number of situations …

- Matched case-control studies (as above)
- Repeated tests on an individual over time (e.g. before-after)
- Paired observations on an individual (e.g. two eyes)
- Twin studies
- Other …

Summary for 2 x 2 Tables

- Cohort Analysis (Prospective)
  1. H_0: $P(D \mid E) = P(D \mid \overline{E})$
  2. RR for incident disease
  3. χ^2 test (or Fisher's Exact)
- Case Control Analysis (Retrospective)
  1. H_0: $P(E \mid D) = P(E \mid \overline{D})$
  2. OR (≈ RR for rare disease)
  3. χ^2 test (or Fisher's Exact)
- Cross-sectional Analysis
  1. H_0: $P(D \mid E) = P(D \mid \overline{E})$
  2. RR for prevalent disease
  3. χ^2 test (or Fisher's Exact)
- Paired Binary Data
  1. H_0: $P(D \mid E) = P(D \mid \overline{E})$
  2. OR
  3. McNemar's test (or exact Binomial)

Categorical data -summary

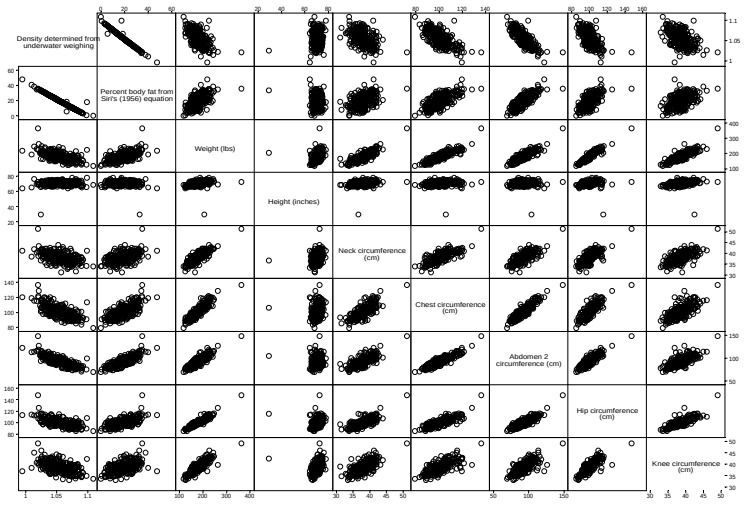


Inference in Correlation and Linear Regression

- Correlation
  - Pearson's, Spearman's
  - Hypothesis test for ρ
- Linear Regression
  - Summarize linear association
  - Prediction
- Hypothesis testing for regression parameters
- Confidence intervals
  - parameters
  - fitted values
  - new observation (prediction interval)
- Sums of Squares
  - Regression SS, Residual SS, Total SS, R^2
- Assumptions in linear regression
  - **L**inearity
  - **I**ndependence
  - **N**ormality
  - **E**qual variances
- Model Checking
  - Checking systematic component (linearity)
  - Checking the random component (normality, equal variance)

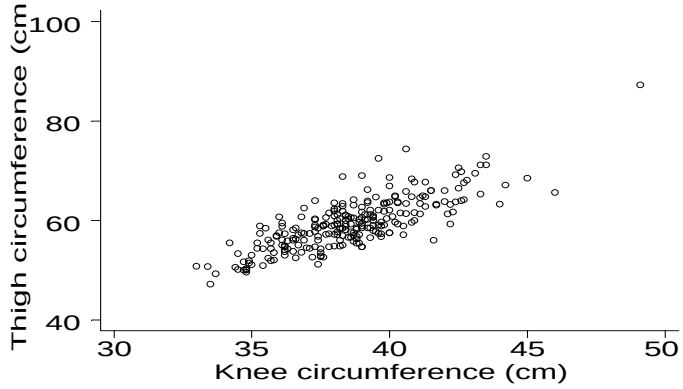
Body Fat Dataset

```
Contains data from bodyfat.dta        obs:           252

   1. density   float  %9.0g                 Density determined from
                                              underwater weighing
   2. pctfat    float  %9.0g                 Percent body fat from Siri's
                                              (1956) equation
   3. age       float  %9.0g                 Age (years)
   4. weight    float  %9.0g                 Weight (lbs)
   5. height    float  %9.0g                 Height (inches)
   6. neck      float  %9.0g                 Neck circumference (cm)
   7. chest     float  %9.0g                 Chest circumference (cm)
   8. abdomen   float  %9.0g                 Abdomen 2 circumference (cm)
   9. hip       float  %9.0g                 Hip circumference (cm)
  10. thigh     float  %9.0g                 Thigh circumference (cm)
  11. knee      float  %9.0g                 Knee circumference (cm)
  12. ankle     float  %9.0g                 Ankle circumference (cm)
  13. biceps    float  %9.0g                 Biceps (extended)
                                              circumference (cm)
  14. forarm    float  %9.0g                 Forearm circumference (cm)
  15. wrist     float  %9.0g                 Wrist circumference (cm)
```



Correlation

We want to measure the "strength of association" between two (quantitative) variables. For this purpose, we will use the correlation coefficient.



Pearson's Correlation Coefficient

The **correlation** between two variables X and Y is defined as:

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{V(X)V(Y)}}$$

Properties:

- Symmetric – no distinction between X and Y
- The correlation is constrained: $-1 \le \rho \le +1$
- $|\rho| = 1$ means "perfect linear relationship":

  $Y = a + bX$
- The correlation is a scale free measure.
- We estimate the correlation as:

$$R = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{s_X s_Y} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{s_X s_Y}$$

Inference for Pearson's Correlation Coefficient

To test the hypothesis:

$$H_0 : \rho = 0$$
$$H_A : \rho \neq 0$$

We use the statistic:

$$T = \sqrt{n-2}\,\frac{R}{\sqrt{1-R^2}}$$

Under the null hypothesis:

$$T \sim t(n-2)$$

which forms the basis for testing.

NOTE: For the validity of the test we assume that both X and Y are normally distributed (bivariate normality).

Inference for Pearson's Correlation Coefficient

E.g. Knee circumference and thigh circumference

n = 252

R = 0.799

$H_0 : \rho = 0$
$H_A : \rho \neq 0$

$$T = \sqrt{n-2}\frac{R}{\sqrt{1-R^2}} = \sqrt{252-2}\frac{.799}{\sqrt{1-.799^2}} = 21$$

Conclusion: reject H_0 with $p < .0001$

Inference for Pearson's Correlation Coefficient

```
. pwcorr knee thigh, sig

             |     knee    thigh
-------------+------------------
        knee |   1.0000
             |
             |
       thigh |   0.7992   1.0000
             |   0.0000
```

Spearman Rank Correlation

- A nonparametric analogue to Pearson's correlation coefficient is Spearman's rank correlation coefficient. Use Spearman's correlation when the assumption of (bivariate) normality is not met.
- A measure of monotonic association (not necessarily linear)
- Based on the ranked data
  - Rank each sample separately
  - Compute Pearson's correlation on the ranks
  - $-1 \leq R_s \leq 1$
  - $T = \sqrt{n-2}\,\frac{R_s}{\sqrt{1-R_s^2}} \sim t(n-2)$

Spearman Rank Correlation

```
. spearman knee thigh

 Number of obs =     252
Spearman's rho =       0.7699

Test of Ho: knee and thigh are
independent
    Prob > |t| =       0.0000
```

Correlation – Restricted Range

What happens if we restrict the range of the data for one or the other variables when computing correlation?

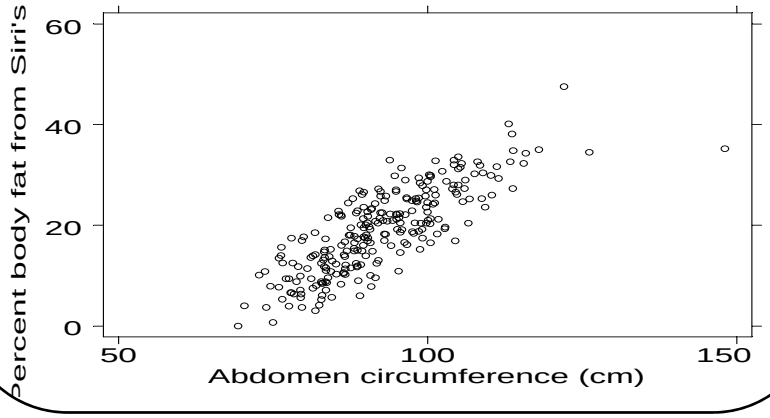
E.g. knee circumference vs thigh circumference

range	R	p
All	.80	<.001
knee < 45	.78	<.001
knee < 40	.68	<.001
knee < 35	.19	.48

Linear Regression

The correlation coefficient was used to summarize the strength of the relationship between interchangeable X and Y.

Sometimes, however, X and Y are not interchangeable. We may want to predict Y from X.



Linear Regression

- If a scatterplot suggests a linear relationship between X and Y we can draw a **linear regression line** to describe how the mean of Y ~~changes~~ **differs** when X ~~changes~~ **differs** or to predict the mean of Y for any given value of X.
- In **linear regression** one variable (X) is used to predict or explain another (Y) (the situation is asymmetric).

  X independent, predictor $\Rightarrow$ Y dependent, response

- We assume that we collect a sample of **pairs** of observations,

  (X_i, Y_i) for $i = 1, 2, \ldots, n$

  Note: here, X and Y are both quantitative; more generally, X need not be.

- Modeling the relationship between X and Y requires the specification of two components:
  - **Systematic Component**
  - **Random Component**

Assumptions for Linear Regression

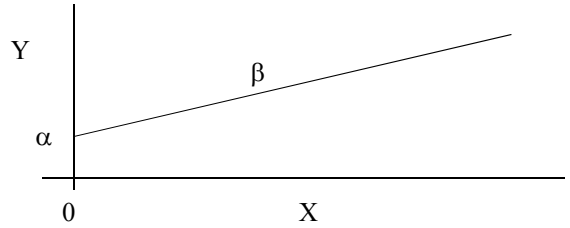
Systematic component:

$$E(Y_i \mid X_i) = \alpha + \beta X_i$$

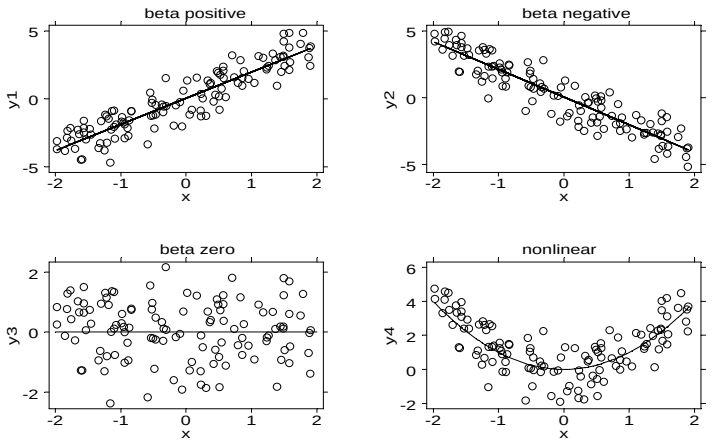
"expected (mean) population value of Y at X_i"

α = intercept = value of mean of Y when X = 0

β = slope = expected ~~change~~ **difference** in mean of Y for each 1 unit ~~change~~ **difference** in X



Examples of Systematic component



Assumptions for Linear Regression

Random part: $Y_i = E(Y_i|X_i) + \varepsilon_i$

$= \alpha + \beta X_i + \varepsilon_i$

1. Equal variance (i.e. variance doesn't depend on X)

$$V(Y_i \mid X_i) = V(\varepsilon_i) = \sigma^2$$

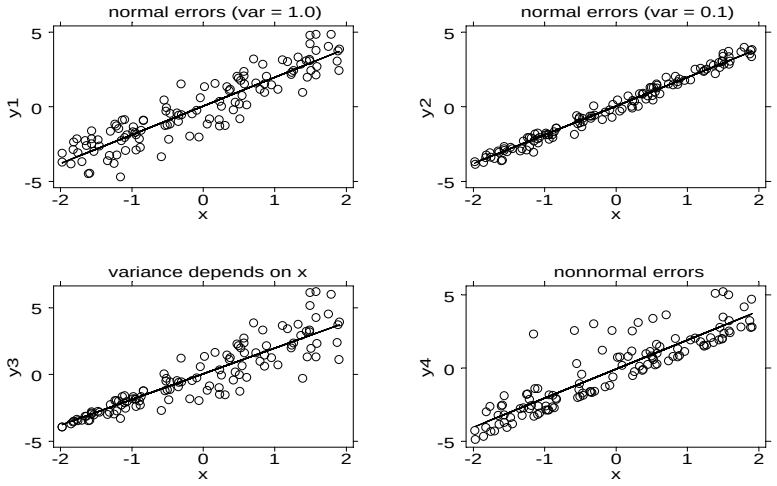
2. Responses are independent.

Y_i , Y_j (actually, ε_i , ε_j) are independent for all i, j.

3. "Errors" are normally distributed.

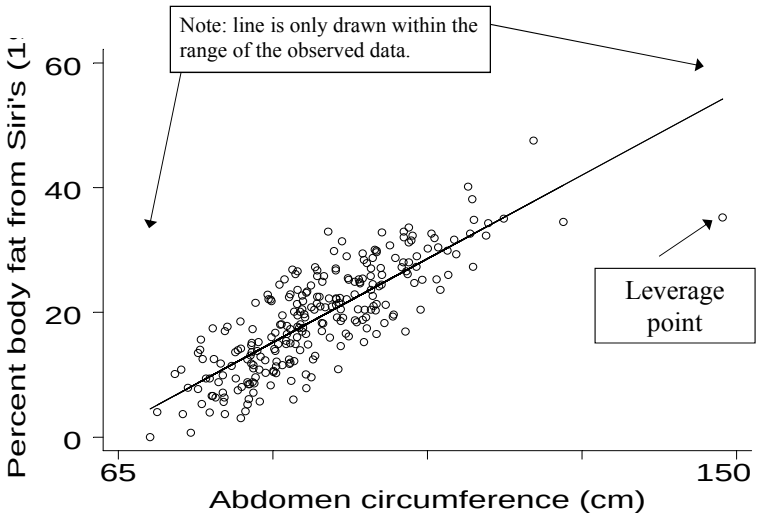
$$\varepsilon_i \sim N(0, \sigma^2)$$

Assumptions for Linear Regression




```
. summarize abdomen

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
 abdomen |     252    92.55595   10.78308       69.4      148.1

. regress pctfat abdomen

  Source |       SS       df       MS                  Number of obs =     252
---------+------------------------------                F(  1,   250) =  488.93
   Model |  11631.5264     1  11631.5264                Prob > F      =  0.0000
Residual |  5947.46321   250  23.7898528                R-squared     =  0.6617
---------+------------------------------                Adj R-squared =  0.6603
   Total |  17578.9896   251   70.035815                Root MSE      =  4.8775

------------------------------------------------------------------------------
  pctfat |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 abdomen |   .6313044   .0285507     22.112   0.000       .5750739    .6875349
   _cons |  -39.28018   2.660337    -14.765   0.000      -44.51971   -34.04065
------------------------------------------------------------------------------
```

Y = -39.28 + .6312 X



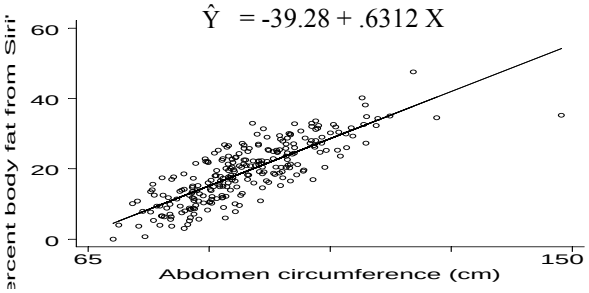
Regression - Predicted Values

Given the estimates (a, b) we can find the **predicted value**, $\hat{Y}_i$, for any value of X_i.

$$\hat{Y}_i = a + bX_i$$

The interpretation of $\hat{Y}_i$ is as the **estimated mean value of** Y_i **for a large sample of values taken at** $X = X_i$.

$\hat{Y}$ = -39.28 + .6312 X



Predicted body fat when abdominal circumference is 90 cm
= -39.28 + .6312*90 = 17.53 percent

Regression - Residuals

We also wish to estimate σ^2. Recall that $\sigma^2 = \text{Var}(\varepsilon_i)$. We call the ε_i the "residuals".

We don't know the ε_i exactly since these are based on α and β. BUT, we do have a reasonable estimate based on a and b:

$$r_i = Y_i - a - bX_i$$
$$= Y_i - \hat{Y}_i$$

Since the average of the r_i is 0 (guaranteed by least squares), a reasonable estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_i r_i^2}{n-2} = \frac{\sum_i (Y_i - a - bX_i)^2}{n-2}$$

- We will also use the estimated residuals to assess the adequacy of our model.

Inferences about Regression Parameters

For the simple linear model we can test hypotheses regarding β:

$H_0 : \beta = 0$
$H_A : \beta \neq 0$

using a standardized test statistic: $T = \frac{b-0}{\sqrt{V(b)}}$

Similarly, hypotheses about α (less common):

$H_0 : \alpha = 0$
$H_A : \alpha \neq 0$

are based on the test statistic: $T = \frac{a-0}{\sqrt{V(a)}}$

We just need estimates of V(a) and V(b)…

Inferences about Regression Parameters

The variance of the estimated regression coefficients ($a = \hat{\alpha}$, $b = \hat{\beta}$) is given by:

$$V(a) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{L_{xx}}\right)$$

$$V(b) = \sigma^2\left(\frac{1}{L_{xx}}\right)$$

computer does these calculations

where $L_{xx} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = (n-1)s_x^2$ and we replace σ by its estimate.

Inferences about Regression Parameters

Bodyfat example: Regress abdominal fat (Y) on abdomen circum (X).

H_o: $\beta = 0$

H_a: $\beta \neq 0$

a = -39.28
b = 0.6312
$\hat{\sigma} = 4.877$
$L_{xx} = 251 * 10.78^2 = 29184.5$

(see Stata output on page 420)

$$T = \frac{.6312 - 0}{4.877\sqrt{\frac{1}{29184.5}}} = 22.1$$

Conclusion?

NOTE: The tests for Ho:β=0 and Ho:ρ=0 are mathematically equivalent.

Confidence Intervals for Regression Parameters

Given that the errors ε_i are independent, have equal variances, and are normally distributed, then:

$$a \sim N\left(\alpha,\ \sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{L_{xx}}\right)\right)$$

$$b \sim N\left(\beta,\ \sigma^2\left(\frac{1}{L_{xx}}\right)\right)$$

Since σ is unknown, confidence intervals for the regression parameters use the t(n - 2) distribution:

$$\text{CI for } \alpha: \quad a \ \pm\ t_{1-\alpha/2}(n-2)\times\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{L_{xx}}}$$

$$\text{CI for } \beta: \quad b \ \pm\ t_{1-\alpha/2}(n-2)\times\hat{\sigma}\sqrt{\frac{1}{L_{xx}}}$$

Confidence Intervals for Regression Parameters

Bodyfat example: (n = 252)

$$a = -39.28$$

$$b = 0.6312$$

$$\hat{\sigma} = 4.877$$

$$L_{xx} = 29184.5$$

A 95% confidence interval for β is

0.6312 ± 1.97*4.877*sqrt(1/29184.5)

(.575,.687)

Confidence Intervals for Predicted Means

The predicted value, $\hat{Y}_i$, is the estimated mean response at X_i and is estimated as:

$$\hat{Y}_i = a + bX_i$$

Further $\hat{V}(\hat{Y}_i \mid X_i) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{L_{xx}}\right)$

so, a confidence interval for $E(Y_i \mid X_i) = \alpha + \beta X_i$ is given by:

$$\hat{Y}_i \pm t_{1-\alpha/2}(n-2) \times \sqrt{\hat{V}(\hat{Y}_i | X_i)}$$

Confidence Intervals for Predicted Means

Bodyfat example: (n = 252)

$$a = -39.28$$
$$b = 0.6312$$
$$\hat{\sigma} = 4.877$$
$$\overline{X} = 92.56$$
$$L_{xx} = 29184.5$$

Consider the mean bodyfat for an abdomen circumference of 100 cm:

$$\hat{Y}_i = a + b \times X_i$$
$$= -39.28 + 0.6312 \times 100 = 23.82$$

$$\hat{V}\left(\hat{Y}_i | X_i\right) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{L_{xx}}\right)$$
$$= (4.877)^2\left(\frac{1}{252} + \frac{(100-92.56)^2}{29184.5}\right) = 0.139$$

$t_{1-\alpha/2}(n-2) = 1.97$

Thus a 95% confidence interval for $E(Y_i \mid X = 100)$ is:

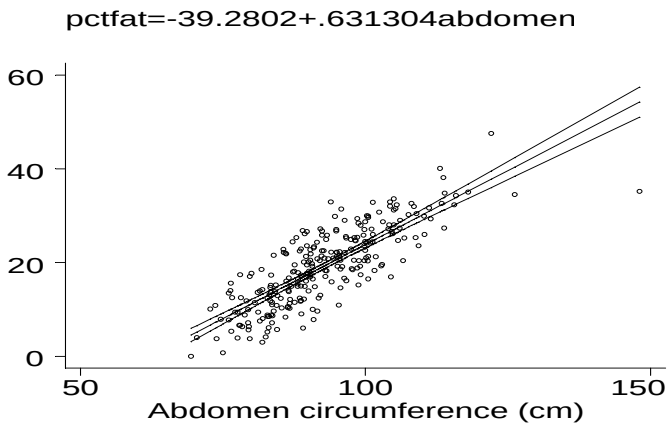
$$\hat{Y}_i \pm t_{1-\alpha/2}(n-2) \times \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{L_{xx}}}$$

$$= 23.82 \pm 1.97 \times \sqrt{0.139}$$

$$= 23.82 \pm 0.74$$

$$= (23.08,\ 24.56)$$

Confidence Intervals for Predicted Means



Prediction Intervals

The confidence interval for E(Y|X) that we have developed gives us an interval that we expect the (population) **mean** of Y at X to fall in.

Suppose that we wanted an interval (range of values) that we would expect a **single** "new" observation to fall in…

- How should the **prediction** of an single new observation at X = 100 (say) compare to the prediction of the mean of all observations at X = 100 (same, higher, lower)?
- How should the **uncertainty** about the prediction of an single new observation at X = 100 (say) compare to the uncertainty about the prediction of the mean of all observations at X = 100 (same, higher, lower)?

Prediction Intervals

In predicting a single new observation we have the uncertainty about the population mean PLUS the intrinsic variability of individual observations (σ^2). The variability in predicting a single new observation is the sum of these:

$$\begin{aligned} \text{Var}(\hat{Y}_{\text{single}}) &= \sigma^2 + \text{Var}(\hat{Y}_{\text{mean}}) \\ &= \sigma^2\left(1+\frac{1}{n}+\frac{(X-\bar{X})^2}{L_{xx}}\right) \end{aligned}$$

Thus, for an **individual** observation the interval:

$$(a+bX_i) \pm t_{1-\alpha/2}(n-2)\times\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{L_{xx}}}$$

$$\hat{Y}_i \pm t_{1-\alpha/2}(n-2)\times\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{L_{xx}}}$$

is a $(1 - \alpha)$ **prediction interval** for a new observation taken at X_i.

Prediction Intervals

Bodyfat example: (n = 252)

$$a = -39.28$$
$$b = 0.6313$$
$$\hat{\sigma} = 4.877$$
$$\overline{X} = 92.56$$
$$L_{xx} = 29{,}184.5$$

Consider an **individual** bodyfat measurement for a new individual with an abdomen circumference of 100cm:

$$\hat{Y}_i = a + b \times 100 = 23.82$$

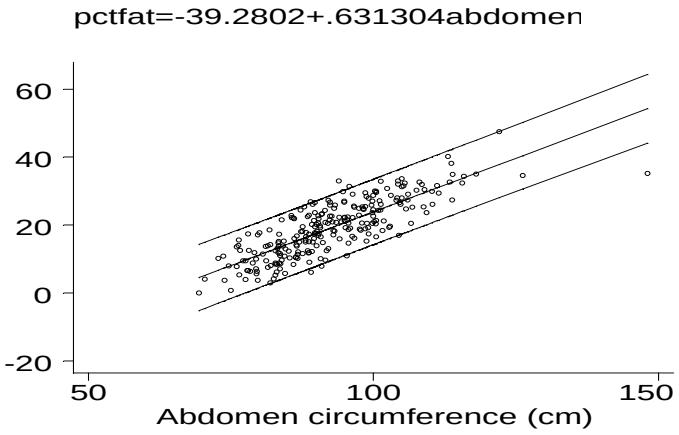
A **95% prediction interval** is given by $\hat{Y}_i \pm t_{1-\alpha/2}(n-2) \times \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{L_{xx}}}$

$$23.82 \pm 1.97 \times 4.877 \sqrt{1 + \frac{1}{252} + \frac{(100 - 92.56)^2}{29{,}184.5}}$$

$$23.82 \pm 9.64$$

$$(14.18 \ , \ 33.46)$$

Prediction Intervals



To get confidence intervals on predicted values and prediction intervals, first edit the dataset to add the X values you want (leave Y missing), then fit the regression, and use predict.

```
. use “bodyfat.dta", clear
. edit                          // add “fake” observations
. reg pctfat abdomen
. predict fathat                // gives E(Y|X)
. predict sefathat, stdp        // gives (se for) CI for E(Y|X)
. predict senew, stdf           // gives (se for) PI

. list pctfat abdomen fathat sefathat senew if abdomen==100

        pctfat   abdomen      fathat    sefathat       senew
 253.        .       100    23.85025    .3735964    4.891771
```


Sum of Squares (SS)

It is clear that

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

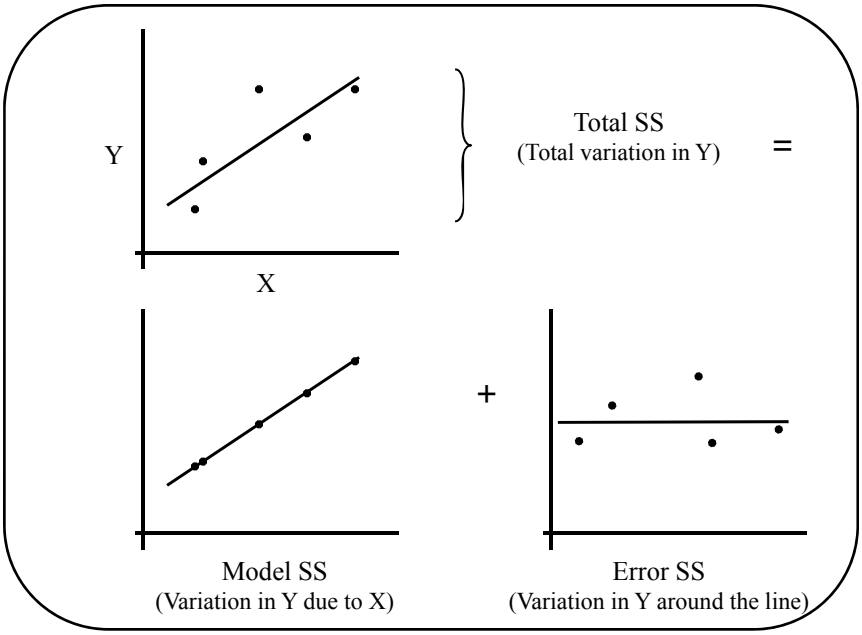
It can also be shown that

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ = **Total SS** - describes the total variation of the Y_i

$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ = **Error SS** - describes the variation of the Y_i around the regression line.

$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ = **Model SS** - describes the structural variation; how much of the variation is due to the regression relationship.



R^2

Total SS = Model SS + Error SS

This decomposition allows a characterization of the usefulness of the covariate X in predicting the response variable Y_i.

Q: If you didn't know X, what would you predict for mean of Y?
A: $\overline{Y}$

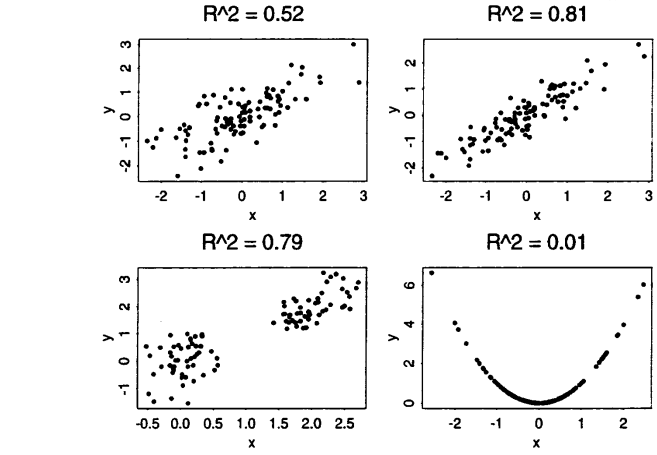
Q: How much unexplained variation is left after you make that prediction?
A: Total SS

Q: What did we gain by using X?
A: The proportion of the **Total** variation that can be explained by the regression of Y on X is R^2 = Model SS/Total SS

Alternatively, we can say that the unexplained (residual) variation decreased by a proportion R^2 (i.e. R^2 = 1- Error SS/Total SS)

This R^2 is, in fact, the correlation coefficient squared.

Examples of R^2



Regression - Model Checking

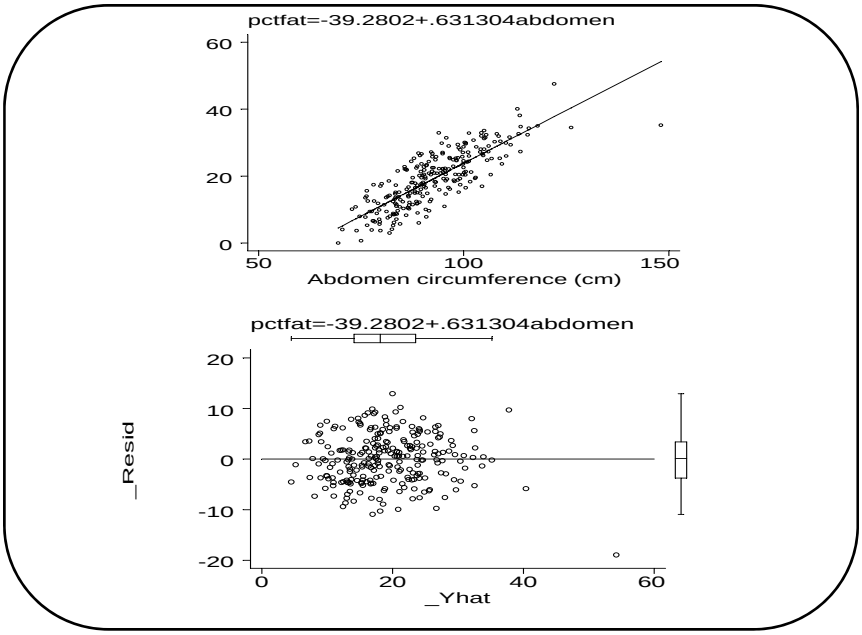
Given the data Y_i and the fitted values, $\hat{Y}_i$,we define the residual as:

$$r_i = Y_i - \hat{Y}_i$$

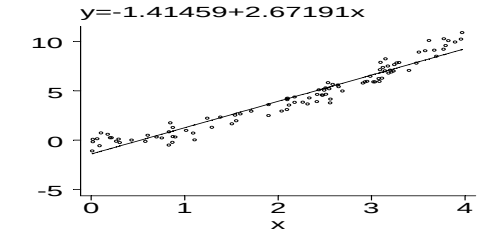
This captures the component of the measurement Y_i that cannot be "explained" by X_i. We will use the residuals to assess our model in terms of the adequacy of both the systematic and random components.

Assumptions and Diagnostics

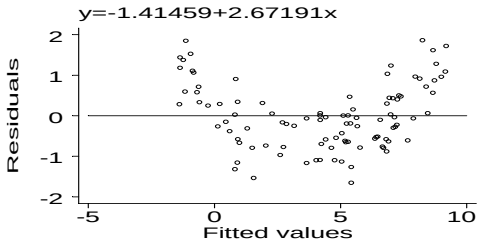
Assumption	Model Checking
Linearity	• residual vs X or $\hat{Y}$ **Q:** Is there any trend?
Independence	**Q:** Any scientific concerns?
Normality	• residual histogram / qq-plot **Q:** Symmetric? Normal?
Equal Variance	• residual vs X or $\hat{Y}$ **Q:** Is there any pattern?



True model: y = x^1.7

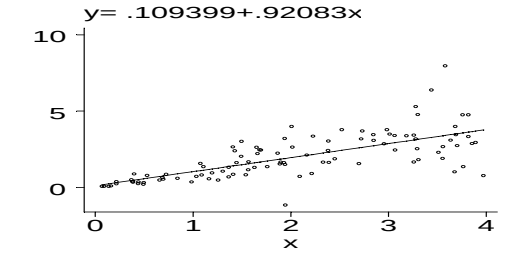


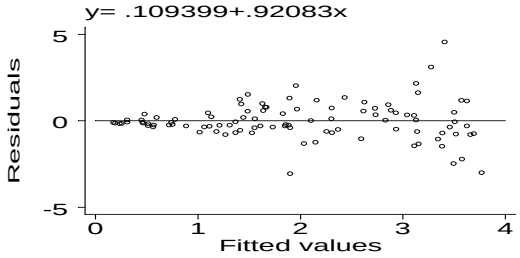
fitted model



plot residuals vs prediction

True model: y = x + errors increasing with x

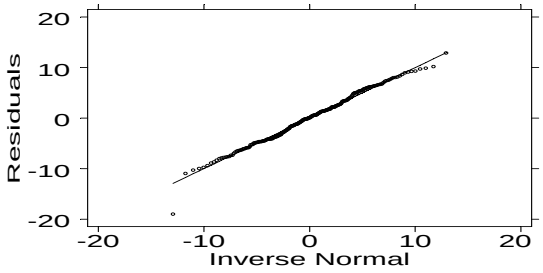




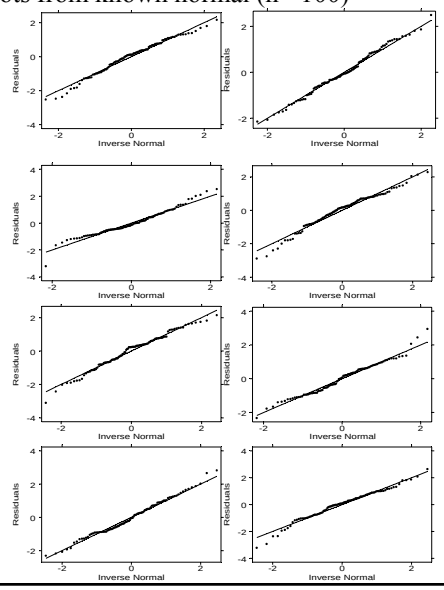
Quantile-Quantile plot (QQplot)

- Let r_i be the i'th ordered residual (smallest to largest)
- Let p_i be the percentile of the i'th ordered residual. $p_i = i/(n+1)$
- Plot r_i versus $E(r_i) = s_r \times Z_{pi}$
- If residuals are normal, plot should be a straight line

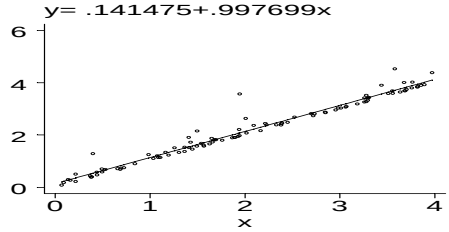
E.g. Bodyfat vs abdominal circumference



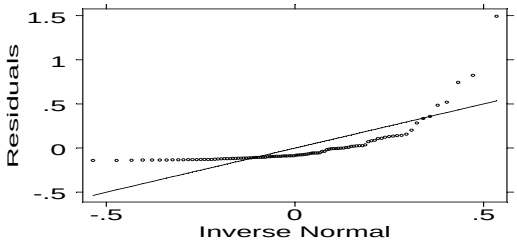
QQplots from known normal (n= 100)



True model: y = x + chi-squared errors



QQ plot



Impact of Violations

Nonlinearity:

1. Estimates - rubbish. Biased estimation.
2. Tests/CIs - also rubbish. Systematic deviations spill over into estimates of variability.
3. Correction - transform or choose a nonlinear model.

Nonnormality:

1. Estimates - effect is minimal for most departures. Outliers can be a disaster. If points exist far from the main body of X values, they can exert undue influence on estimates (particularly $\hat{\beta}$).
2. Tests/CIs - again minimal for most departures
3. Correction - delete outliers (if warranted) or nonparametric regression.

Impact of Violations

Unequal Variances:

1. Estimates - minimal impact. (still unbiased, consistent)
2. Tests/CIs - variance estimates are wrong, but the effect is usually not dramatic.
3. Correction - transform or weighted least squares.

Dependence:

1. Estimates - range of possibilities, but often the estimates are unbiased.
2. Tests/CIs - variance estimates are wrong. Often they will overestimate the precision and inflate test statistics (p-values too small).
3. Correction - regression for dependent data.

Summary

Correlation
- Pearson's
- Spearman's
- Hypothesis test for ρ

Purposes of Linear Regression
- Summarize linear association
- Prediction

Assumptions in linear regression
- **L**inearity
- **I**ndependence
- **N**ormality
- **E**qual variances

Fitting a linear regression
- Least squares

Hypothesis testing for regression parameters
- t test - single parameter

Confidence intervals
- parameters
- fitted values
- new observation (prediction interval)

Sums of Squares
- Regression SS
- Residual SS
- Total SS
- R^2

Model Checking…
- Checking systematic component
- Checking the random component

Model Checking…

Anscombe's Quartet (1973)

- Statistician Francis Anscombe created four datasets with nearly identical simple statistical properties. He used the illustration to demonstrate the effects of outliers and non-linear patterns.
- And to warn us of the importance of graphing our data!

Model Checking…

Anscombe's Quartet (1973)

Each of the four dataset has the following summaries:

- $E[Y] = 3 + 5\,X$ (2-3 decimal places)
- $\overline{X} = 9$ (exact)
- $\overline{Y} = 7.50$ (2 decimal places)
- $S_x = 11$ (exact)
- $S_y = 4.12$ (2 decimal places)
- $R = 0.816$ (2 decimal places)

Model Checking…

Anscombe's Quartet (1973)