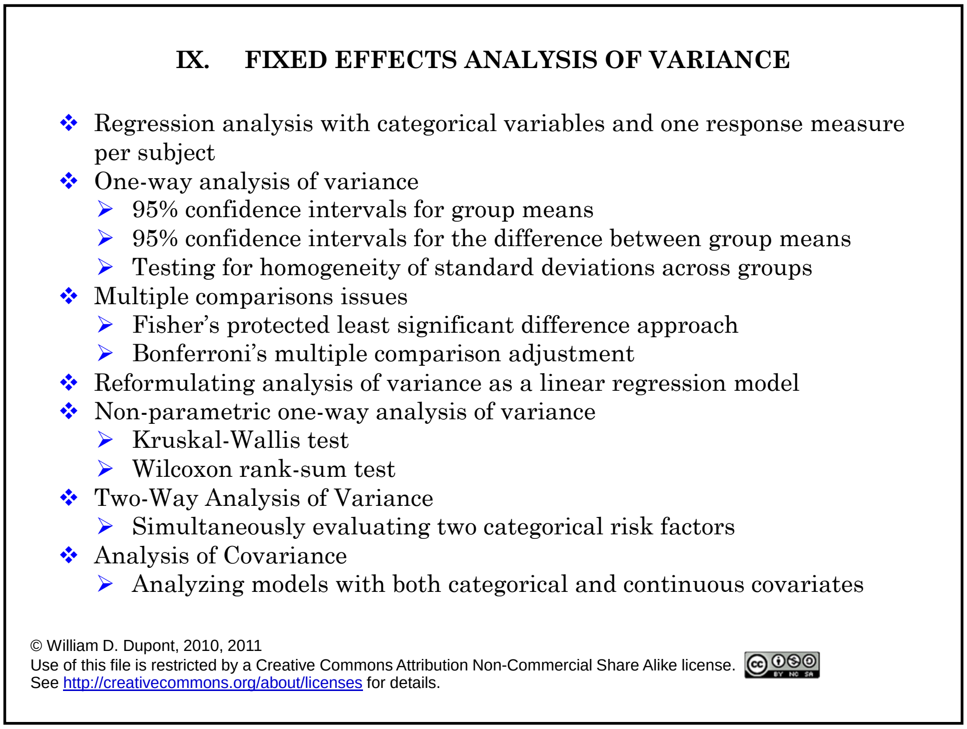
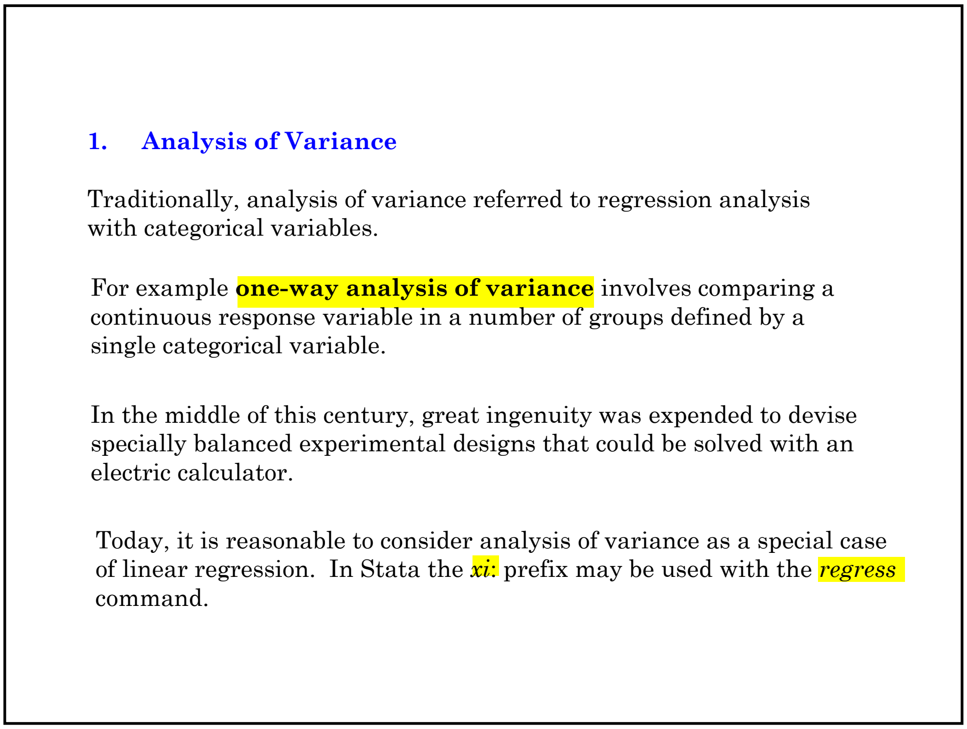


IX. FIXED EFFECTS ANALYSIS OF VARIANCE

- Regression analysis with categorical variables and one response measure per subject
- One-way analysis of variance
  - 95% confidence intervals for group means
  - 95% confidence intervals for the difference between group means
  - Testing for homogeneity of standard deviations across groups
- Multiple comparisons issues
  - Fisher's protected least significant difference approach
  - Bonferroni's multiple comparison adjustment
- Reformulating analysis of variance as a linear regression model
- Non-parametric one-way analysis of variance
  - Kruskal-Wallis test
  - Wilcoxon rank-sum test
- Two-Way Analysis of Variance
  - Simultaneously evaluating two categorical risk factors
- Analysis of Covariance
  - Analyzing models with both categorical and continuous covariates



1. Analysis of Variance

Traditionally, analysis of variance referred to regression analysis with categorical variables.

For example **one-way analysis of variance** involves comparing a continuous response variable in a number of groups defined by a single categorical variable.

In the middle of this century, great ingenuity was expended to devise specially balanced experimental designs that could be solved with an electric calculator.

Today, it is reasonable to consider analysis of variance as a special case of linear regression. In Stata the *xi:* prefix may be used with the *regress* command.

A critical assumption of these analyses is that the error terms for each observation are independent and have the same normal distribution. This assumption is often reasonable as long as we only have one response observation per patient.

These analyses assume that all parameters are attributes of the underlying population, and that we have obtained a representative sample of this population. These parameters measure attributes that are called **fixed-effects**.

In contrast, we often have multiple observations per patient. In this case some of the parameters measure attributes of the individual patients in the study. Such attributes are called **random effects**. A model that has both random and fixed effects is called a **mixed effects** model or a **repeated measures** model.

2. One-Way Analysis of Variance

Let n_i be the number of subjects in the i^{th} group

$n = \sum n_i$ be the total number of study subjects

y_{ij} be a continuous response variable on the j^{th} patient from the i^{th} group.

We assume for $i = 1,2,\ldots k$; $j = 1,2,\ldots,n_i$ that

$$y_{ij} = \beta_i + \varepsilon_{ij} \qquad \{9.1\}$$

where

$\beta_1,\beta_2,\ldots\beta_k$ are unknown parameters, and

ε_{ij} are mutually independent, normally distributed error terms with mean 0 and standard deviation σ.

Under this model, the expected value of y_{ij} is $\mathrm{E}\left[y_{ij} \mid i\right] = \beta_i$

Models like {9.1} are called **fixed-effects** models because the parameters $\beta_1,\beta_2,\ldots\beta_k$ are fixed constants that are attributes of the underlying population.

The response y_{ij} differs from β_i only because of the error term ε_{ij}. Let

$b_1,b_2,\ldots b_k$ be the least squares estimates of $\beta_1,\beta_2,\ldots\beta_k$, respectively,

$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ be the sample mean for the i^{th} group,

and

$s^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2/(n-k)$ be the mean squared error (MSE) estimate of σ^2 {9.2}

We estimate σ by s, which is called the root MSE. It can be shown that $b_i = \bar{y}_i$, $E\left[b_i\right]=\beta_i$, and $\mathrm{E}\left[s^2\right]=\sigma^2$. A 95% confidence interval for β_i is

given by $\bar{y}_i \pm t_{n-k,0.025}\left(s/\sqrt{n_i}\right)$ {9.3}

Note that model {9.1} assumes that the standard deviation of ε_{ij} is the same for all groups. If it appears that there is appreciable variation in this standard deviation among groups then the 95% confidence interval for β_i should be estimated by

$\bar{y}_i \pm t_{n_i-1,0.025}\left(s_i\big/\sqrt{n_i}\right)$ {9.4}

where s_i is the sample standard deviation of y_{ij} within the i_{th} group.

We wish to test the null hypothesis that the expected response is the same in all groups. That is, we wish to test whether

$\beta_1 = \beta_2 = \ldots = \beta_k$ {9.5}

We can calculate a statistic that has a ***F* distribution** with k-1 and n-k degrees of freedom when this null hypothesis is true.

We reject the null hypothesis in favor of a multi-sided alternative hypothesis when the F statistic is sufficiently large.

The *P* value associated with this test is the probability that this statistic exceeds the observed value when this null hypothesis is true.

When there are just two groups, the *F* statistic will have 1 and $n - 2$ degrees of freedom. In this case, the one-way analysis of variance is equivalent to an independent *t* test.

The square root of this *F* statistic equals the absolute value of the *t* statistic with $n - 2$ degrees of freedom.

A test due to Levene (1960) can be performed to test the assumption that the standard deviation of ε_{ij} is constant within each group. If this test is significant, or if there is considerable variation in the values of s_i, then you should use equation {9.4} rather than equation {9.3} to calculate confidence intervals for the group means.

$$\bar{y}_i \pm t_{n-k,0.025}\left(s/\sqrt{n_i}\right) \qquad \{9.3\}$$

$$\bar{y}_i \pm t_{n_i-1,0.025}\left(s_i/\sqrt{n_i}\right) \qquad \{9.4\}$$

3. Multiple Comparisons

If, the analysis of variance *F* statistic is significant and the number of groups is not too large, we can make pair-wise comparisons of the different groups.

If the standard deviations within the *k* groups appears similar we can increase the power of the test that $\beta_i = \beta_j$ by using the formula

$$t_{n-k} = \left(\bar{y}_i - \bar{y}_j\right) / \left(s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \right) \qquad \{9.6\}$$

where *s* is the root MSE estimate of σ obtained from the analysis of variance.

Under the null hypothesis that $\beta_i = \beta_j$ equation {9.6} will have a *t* distribution with *n-k* degrees of freedom.

This test is more powerful then the independent *t* test but is less robust.

A 95% confidence interval for the difference in population means between groups *i* and *j* is

$$\bar{y}_i - \bar{y}_j \pm t_{n-k,0.025}\left(s\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right) \qquad \{9.7\}$$

Alternately, a confidence interval based on the independent *t* test may be used if it appears unreasonable to assume a uniform standard deviation in all groups

$$\bar{y}_i - \bar{y}_j \pm t_{n_i+n_j-2,0.025}\left(s_p\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right) \qquad \{9.8\}$$

If the F test is not significant you should not report pair-wise significant differences unless they remain significant after a **Bonferroni multiple comparisons adjustment** (multiplying the P value by the number of pair wise tests.

If the number of groups is large and there is no natural ordering of the groups then a multiple comparisons adjustment may be advisable even if the F test is significant.

4. Fisher's Protected Least Significant Difference (LSD) Approach to Multiple Comparisons

The idea of only analyzing subgroup effects (e.g. differences in group means) when the main effects (e.g. F test) are significant is known as known as **Fisher's Protected Least Significant Difference (LSD) Approach to Multiple Comparisons.**

The F statistic tests the hypothesis that all of the group response means are simultaneously equal.

If we can reject this hypothesis it follows that some of the means must be different.

Fisher argued that in this situation you should be able to investigate which ones are different without having to pay a multiple comparisons penalty.

This approach is not guaranteed to preserve the experiment-wide Type I error probability, but makes sense in well structured experiments where the number of groups being examined is not too large.

5. Reformulating Analysis of Variance as a Linear Regression Model

A one-way analysis of variance is, in fact, a special case of the multiple regression model. Let

y_h denote the response from the h^{th} study subject, $h = 1,2,\ldots n$, and let

$$x_{hi} = \begin{cases} 1: & \text{if the } h^{th} \text{ patient is in the } i^{th} \text{ group} \\ 0: & \text{otherwise} \end{cases}$$

Then model (9.1) can be rewritten

$$y_h = \alpha + \beta_2 x_{h2} + \beta_3 x_{h3} + \ldots + \beta_k x_{hk} + \varepsilon_h \qquad \{9.9\}$$

where ε_h are mutually independent, normally distributed error terms with mean 0 and standard deviation σ. Note that model {9.9} is a special case of model (3.1). Thus, this analysis of variance is also a regression analysis in which all of the covariates are zero-one indicator variables.

Also,

$$\mathrm{E}\left[y_h \mid x_{h2}, x_{h3}, \cdots, x_{hk}\right] = \begin{cases} \alpha & \text{if the } h^{th} \text{ patient is from group 1} \\ \alpha + \beta_i & \text{if the } h^{th} \text{ patient is from group } i > 1 \end{cases}$$

Thus, α is the expected response of patients in the first group and β_i is the expected difference in the response of patients in the i_{th} and first groups.

The least squares estimates of α and β_i are $\bar{y}_1$ and $\bar{y}_i - \bar{y}_1$, respectively.

We can use any multiple linear regression program to perform a one-way analysis of variance, although most software packages have a separate procedure for this task.

6. Non-parametric Methods

a) Kruskal-Wallis Test

The **Kruskal-Wallis** test is the non-parametric analog of the one-way analysis of variance (Kruskal and Wallis 1952).

Model {9.1} assumes that the ε_{ij} terms are normally distributed and have the same standard deviation. If either of these assumptions is badly violated then the Kruskal-Wallis test should be used.

Suppose that patients are divided into k groups as in model {9.1} and that y_{ij} is a continuous response variable on the j^{th} patient from the i^{th} group.

The null hypothesis of this test is that the distributions of the response variables are the same in each group.

Let

n_i be the number of subjects in the i^{th} group,

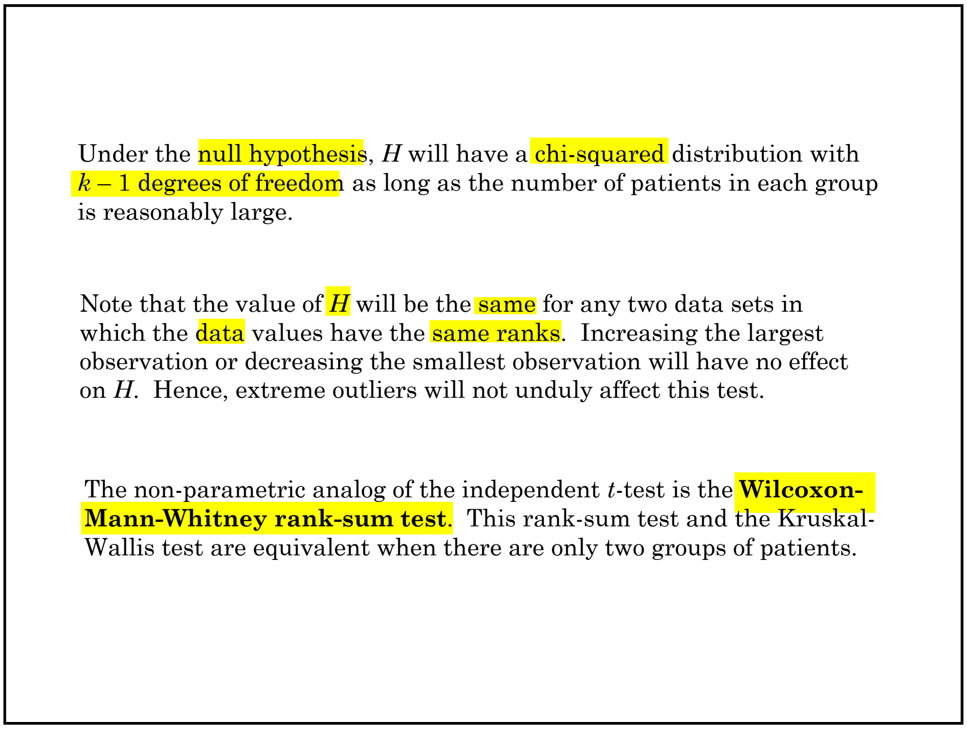
$n = \sum n_i$ be the total number of study subjects.

We rank the values of y_{ij} from lowest to highest and let R_i be the sum of the ranks for the patients from the i^{th} group.

If all of the values of y_{ij} are distinct (no ties) then the Kruskal-Wallis test statistic is

$$H = \frac{12}{n(n+1)}\left(\sum \frac{R_i^2}{n_i}\right) - 3(n+1) \qquad \{9.10\}$$

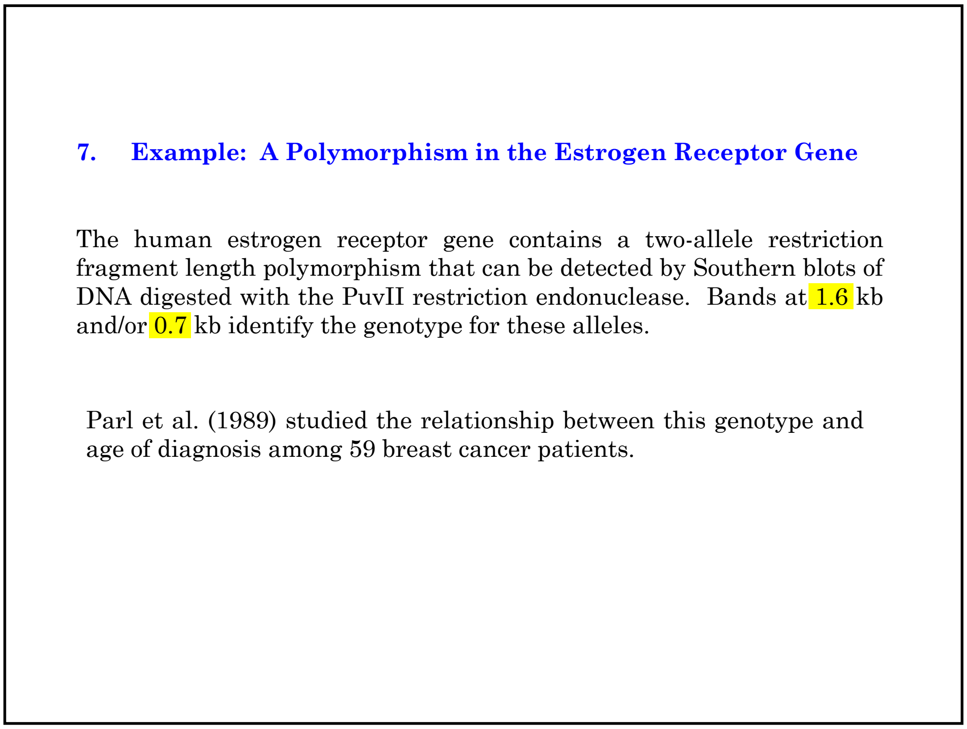
When there are ties a slightly more complicated formula is used (see Steel and Torrie 1980).



Under the null hypothesis, H will have a chi-squared distribution with $k - 1$ degrees of freedom as long as the number of patients in each group is reasonably large.

Note that the value of H will be the same for any two data sets in which the data values have the same ranks. Increasing the largest observation or decreasing the smallest observation will have no effect on H. Hence, extreme outliers will not unduly affect this test.

The non-parametric analog of the independent t-test is the **Wilcoxon-Mann-Whitney rank-sum test**. This rank-sum test and the Kruskal-Wallis test are equivalent when there are only two groups of patients.



7. Example: A Polymorphism in the Estrogen Receptor Gene

The human estrogen receptor gene contains a two-allele restriction fragment length polymorphism that can be detected by Southern blots of DNA digested with the PuvII restriction endonuclease. Bands at 1.6 kb and/or 0.7 kb identify the genotype for these alleles.

Parl et al. (1989) studied the relationship between this genotype and age of diagnosis among 59 breast cancer patients.

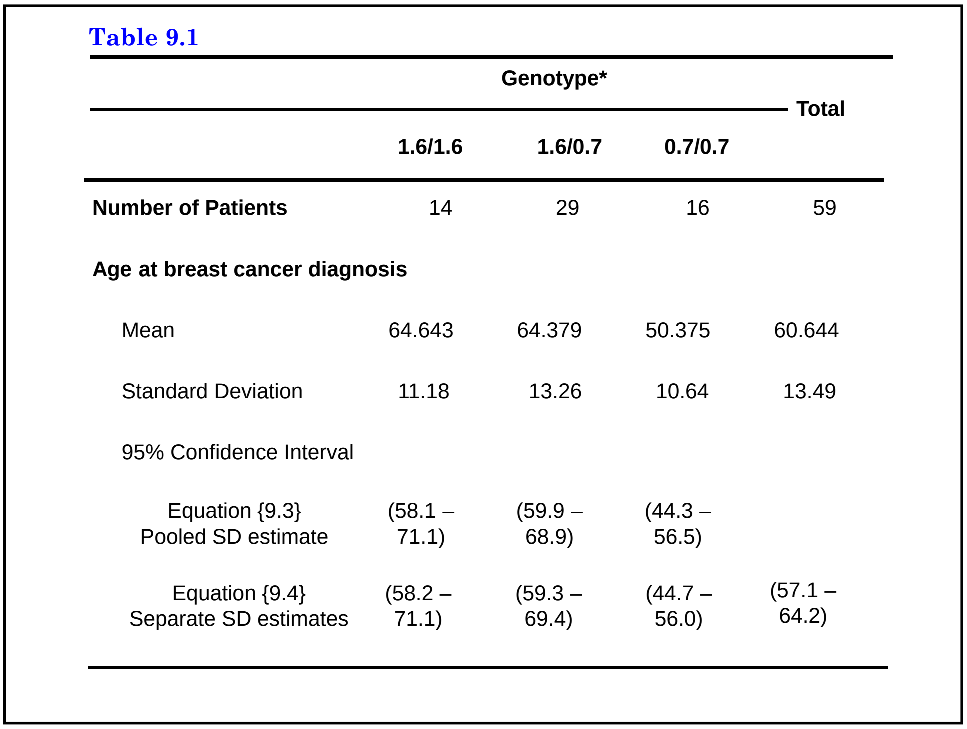
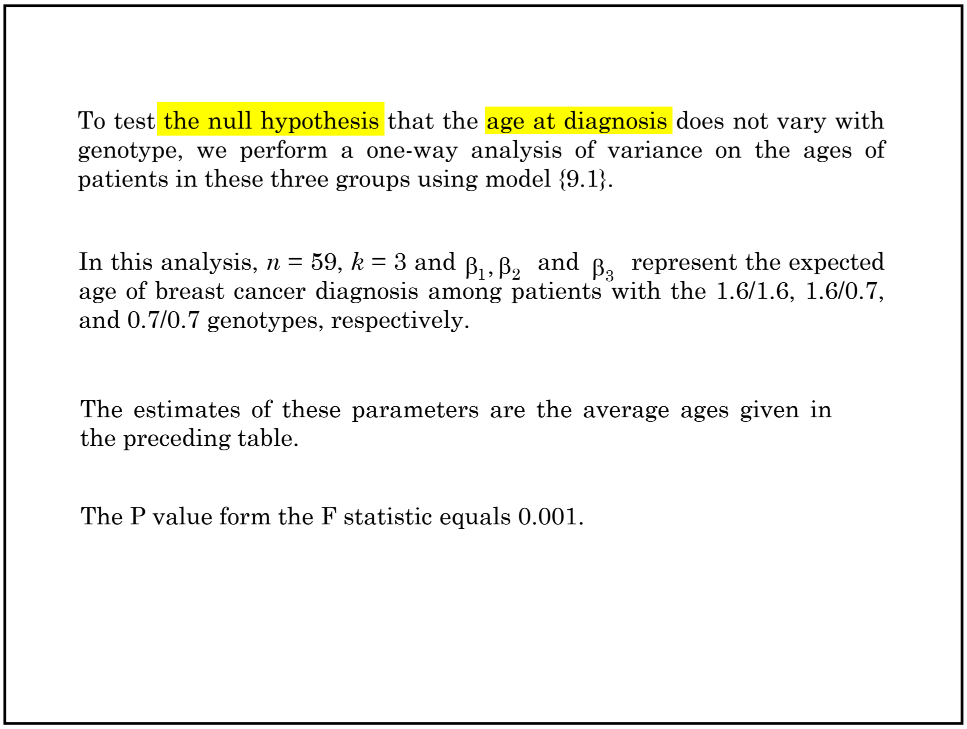


Table 9.1

	Genotype*			Total
	1.6/1.6	1.6/0.7	0.7/0.7	
Number of Patients	14	29	16	59
Age at breast cancer diagnosis				
Mean	64.643	64.379	50.375	60.644
Standard Deviation	11.18	13.26	10.64	13.49
95% Confidence Interval				
Equation {9.3} Pooled SD estimate	(58.1 – 71.1)	(59.9 – 68.9)	(44.3 – 56.5)	
Equation {9.4} Separate SD estimates	(58.2 – 71.1)	(59.3 – 69.4)	(44.7 – 56.0)	(57.1 – 64.2)



To test the null hypothesis that the age at diagnosis does not vary with genotype, we perform a one-way analysis of variance on the ages of patients in these three groups using model {9.1}.

In this analysis, $n = 59$, $k = 3$ and β_1, β_2 and β_3 represent the expected age of breast cancer diagnosis among patients with the 1.6/1.6, 1.6/0.7, and 0.7/0.7 genotypes, respectively.

The estimates of these parameters are the average ages given in the preceding table.

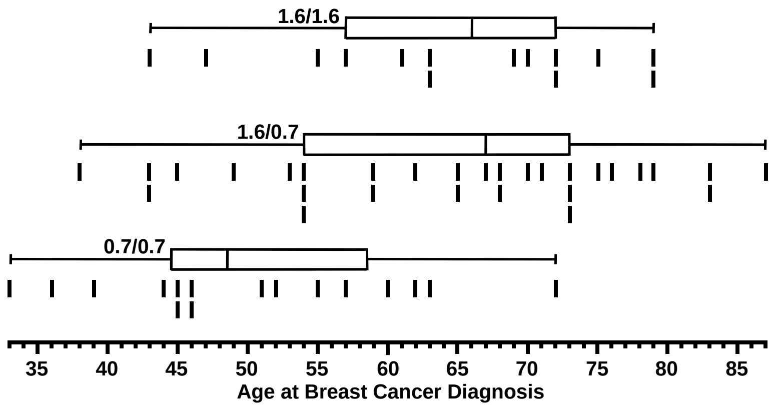
The P value form the F statistic equals 0.001.

Table 9.2

Comparison	Difference in Mean Age of Diagnosis	95% Confidence Interval	*P* Value	
			Eq. {0.7}*	Rank-sum**
1.6/0.7 vs. 1.6/1.6	-0.264	(-8.17 to 7.65)	0.95	0.96
0.7/0. 7 vs. 1.6/1.6	-14.268	(-23.2 to -5.37)	0.002	0.003
0.7/0. 7 vs. 1.6/0.7	-14.004	(-21.6 to -6.43)	< 0.0005	0.002

* Equation 7 uses the pooled estimate of *s*

** Wilcoxon-Mann-Whitney rank-sum test



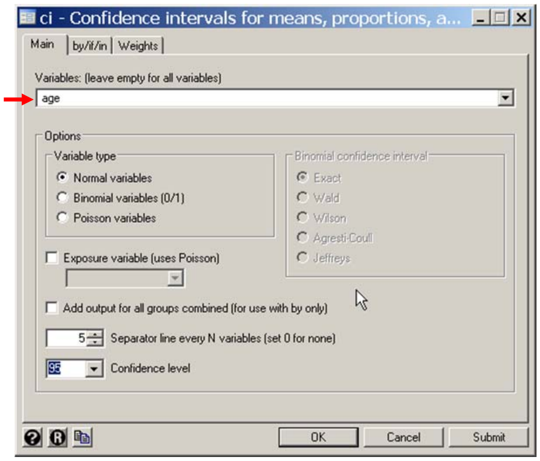
8. One-Way Analyses of Variance using Stata

The following Stata log file and comments illustrate how to perform the one-way analysis of variance discussed in the preceding section.

```
  * 10.8.ERpolymorphism.log
. *
. *  Do a one-way analysis of variance to determine whether age
. *  at breast cancer diagnosis varies with estrogen receptor (ER)
. *  genotype using the data of Parl et al. (1989).
. *
. use C:\WDDtext\10.8.ERpolymorphism.dta                                {1}
. * Statistics > Summaries, tables, ... > Summary ... > Confidence intervals
. ci age                                                                {2}

    Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
 -------------+---------------------------------------------------------------
         age |      59     60.64407    1.756804        57.12744    64.16069
```

{1} This data set contains the **age of diagnosis** and **estrogen receptor genotype** of the 59 breast cancer patients studied by Parl et al. (1989). The **genotypes 1.6/1.6, 1.6/0.7 and 0.7/0.7** are coded 1, 2 and 3 in the variable *genotype,* respectively.

{2} This ***ci*** command calculates the mean age of diagnosis (*age*) together with the associated **95% confidence interval**. This confidence interval is calculated using equation {9.4}. The estimated **standard error of the mean** and the **number of patients** with non-missing ages is also given.



Variables: (leave empty for all variables) age

Options

Variable type
● Normal variables
○ Binomial variables (0/1)
○ Poisson variables

Binomial confidence interval
● Exact
○ Wald
○ Wilson
○ Agresti-Coull
○ Jeffreys

□ Exposure variable (uses Poisson)

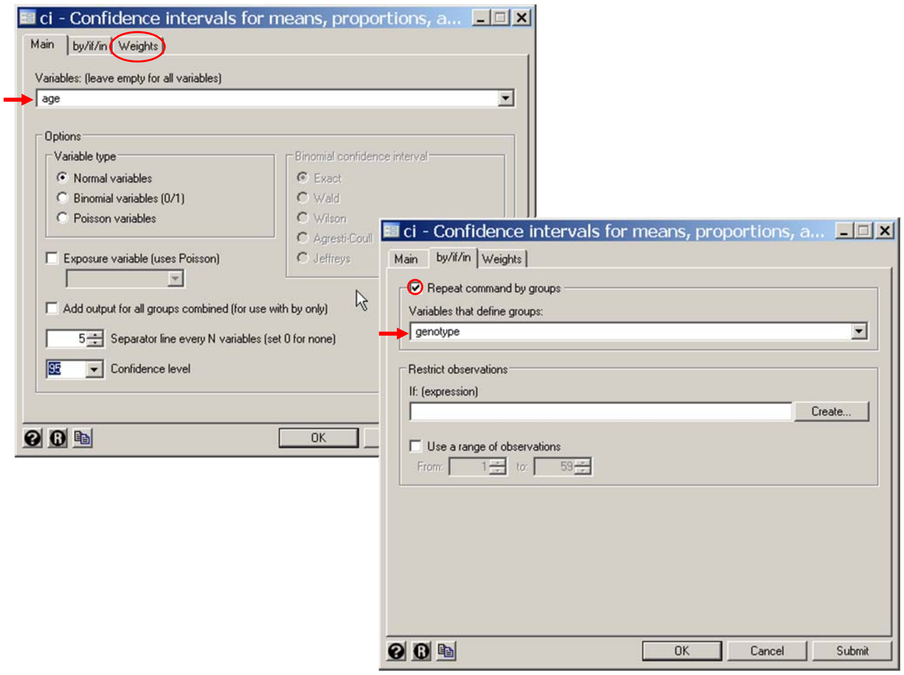
□ Add output for all groups combined (for use with by only)

5 Separator line every N variables (set 0 for none)

95 Confidence level

```
. * Statistics > Summaries, tables, ... > Summary ... > Confidence intervals
. by genotype: ci age                                                     {3}

_______________________________________________________________________________
-> genotype = 1.6/1.6

    Variable |      Obs          Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         age |       14      64.64286    2.988269         58.1871    71.09862

_______________________________________________________________________________
-> genotype = 1.6/0.7

    Variable |      Obs          Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         age |       29      64.37931    2.462234        59.33565    69.42297

_______________________________________________________________________________
-> genotype = 0.7/0.7

    Variable |      Obs          Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         age |       16        50.375    2.659691          44.706      56.044
```

{3} The command prefix ***by genotype:*** specifies that means and 95% **confidence intervals** are to be calculated for **each** of the three **genotypes**.



ci - Confidence intervals for means, proportions, a...

Variables: (leave empty for all variables) age

Options

Variable type: ● Normal variables ○ Binomial variables (0/1) ○ Poisson variables

□ Exposure variable (uses Poisson)

□ Add output for all groups combined (for use with by only)

5 Separator line every N variables (set 0 for none)

95 Confidence level

ci - Confidence intervals for means, proportions, a...

☑ Repeat command by groups

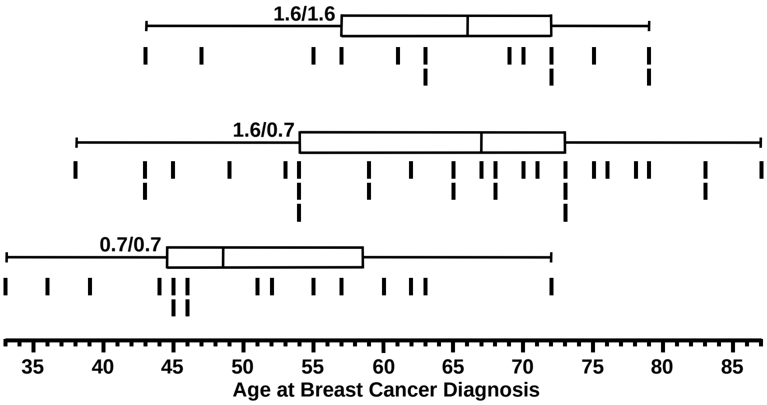
Variables that define groups: genotype

Restrict observations

If: (expression)

□ Use a range of observations From: 1 to: 59

```
. *
. *  The following graph type is not available in Stata version 8.0
. *
. graph7  age, by(genotype) box oneway                                   {4}
```



{4} The *graph7* command implements Stata Version 7 commands using version 7 syntax. The following graph is one that is not available in Version 8. The ***box*** and ***oneway*** options of this *graph* command create a graph that is similar to the Figure. See also Sections 10.7 and 10.8 of text for a prettier way of drawing this graph.

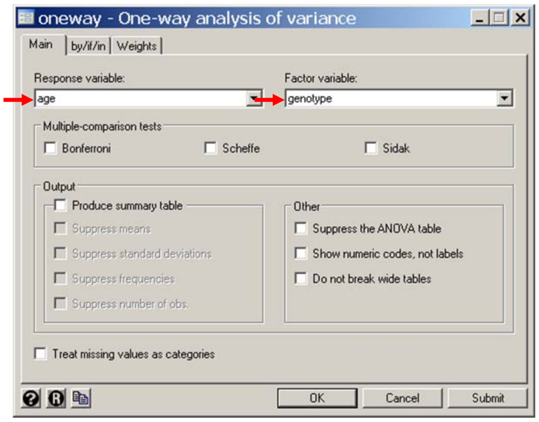
```
. * Statistics > Linear models and related > ANOVA/MANOVA > One-way ANOVA
. oneway age genotype                                                      {5}

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2315.73355      2   1157.86678      7.86     0.0010    {6}
 Within groups      8245.79187     56   147.246283                         {7}
------------------------------------------------------------------------
    Total           10561.5254     58   182.095266

Bartlett's test for equal variances:  chi2(2) =   1.0798  Prob>chi2 = 0.583 {8}
```

{5} This ***oneway*** command performs a **one-way analysis of variance** of *age* with respect to the three distinct values of ***genotype***.

{6} The ***F*** statistic from this analysis equals **7.86**. If the mean age of diagnosis in the target population is the same for all three genotypes, this statistic will have an ***F*** distribution with $k - 1 =$ 3 – 1= 2 and $n - k$ = **56** degrees of freedom. The probability that this statistic exceeds 7.86 is **0.001.**

{7} The **MSE** estimate of is = **147.246**.

{8} **Bartlett's test** for **equal variances** (i.e. equal standard deviations) gives a *P* value of **0.58**.



oneway - One-way analysis of variance

Response variable: age

Factor variable: genotype

Multiple-comparison tests

□ Bonferroni □ Scheffe □ Sidak

Output

□ Produce summary table

□ Suppress means

□ Suppress standard deviations

□ Suppress frequencies

□ Suppress number of obs.

Other

□ Suppress the ANOVA table

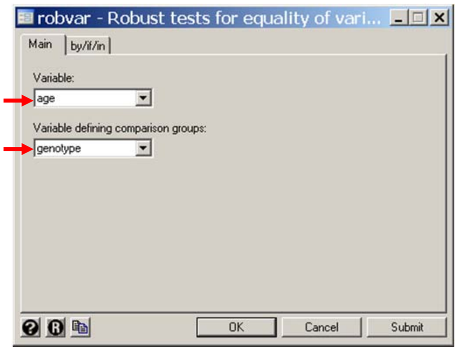
□ Show numeric codes, not labels

□ Do not break wide tables

□ Treat missing values as categories

```
. *
. *  Test whether the standard deviations of age are equal in
. *  patients with different genotypes.
. *
. * Statistics > Summaries, ... > Classical ... > Robust equal variance test
. robvar age, by(genotype)


            |     Summary of Age at Diagnosis
   Genotype |        Mean   Std. Dev.       Freq.
------------+------------------------------------
    1.6/1.6 |   64.642857   11.181077          14
    1.6/0.7 |    64.37931   13.259535          29
    0.7/0.7 |      50.375   10.638766          16
------------+------------------------------------
      Total |   60.644068   13.494268          59

W0  =  0.83032671   df(2, 56)     Pr > F = 0.44120161

W50 =  0.60460508   df(2, 56)     Pr > F = 0.54981692

W10 =  0.79381598   df(2, 56)     Pr > F = 0.45713722
```

This ***robvar*** command performs a test of the equality of variance among groups defined by ***genotype*** using methods of Levene (1960) and Brown and Forsythe (1974). These tests are less sensitive to departures from normality than Bartlett's test. There is no evidence of heterogeneity of variance for age in these three groups.



robvar - Robust tests for equality of vari...

Variable: age

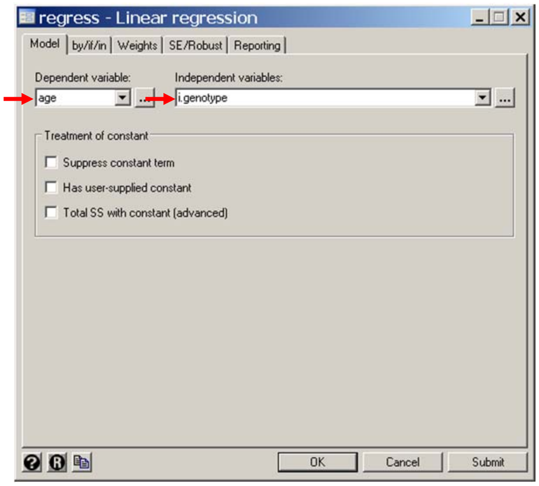
Variable defining comparison groups: genotype

```
. *
. *  Repeat analysis using linear regression
. *
. *  Statistics > Linear models and related > Linear regression
. regress age i.genotype                                                  {9}

      Source |       SS       df       MS              Number of obs =      59
-------------+------------------------------           F(  2,    56) =    7.86
       Model |  2315.73355     2  1157.86678           Prob > F      =  0.0010
    Residual |  8245.79187    56  147.246283           R-squared     =  0.2193
-------------+------------------------------           Adj R-squared =  0.1914
       Total |  10561.5254    58  182.095266           Root MSE      =  12.135

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    genotype |
          2  |  -.2635468   3.949057    -0.07   0.947    -8.174458    7.647365 {10}
          3  |  -14.26786   4.440775    -3.21   0.002     -23.1638   -5.371915
             |
       _cons |   64.64286   3.243084    19.93   0.000     58.14618    71.13953 {11}
------------------------------------------------------------------------------
. oneway age genotype

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2315.73355      2   1157.86678      7.86     0.0010
Within groups       8245.79187     56   147.246283
------------------------------------------------------------------------
    Total           10561.5254     58   182.095266
```

{9} This ***regress*** command preforms exactly the same one-way analysis of variance as the ***oneway*** command given above. Note that the *F* statistic, the *P* value for this statistic and the MSE estimate of are identical to that given by the ***oneway*** command. The **syntax** of the ***xi:*** prefix is explained in Section **5.10**. The model used by this command is equation {9.9} with $k = 3$.

{10} The estimates of β_2 and β_3 in this example are $\bar{y}_2 - \bar{y}_1 = 64.379 - 64.643 = -0.264$ and $\bar{y}_3 - \bar{y}_1 = 50.375 - 64.643 = -14.268$, respectively. They are highlighted in the column labeled ***Coef.*** The 95% confidence intervals for β_2 and β_3 are calculated using equation {9.7}. The ***t*** statistics for testing the null hypotheses that $\beta_2 = 0$ and $\beta_3 = 0$ are -0.07 and -3.21, respectively. They are calculated using equation {9.6}. The highlighted values in this output are also given in **Table 9.2**.

{11} The estimate of α is $\bar{y}_1 = 64.643$. The 95% confidence interval for α is calculated using equation {9.3}. These statistics are also given in Table 10.1.



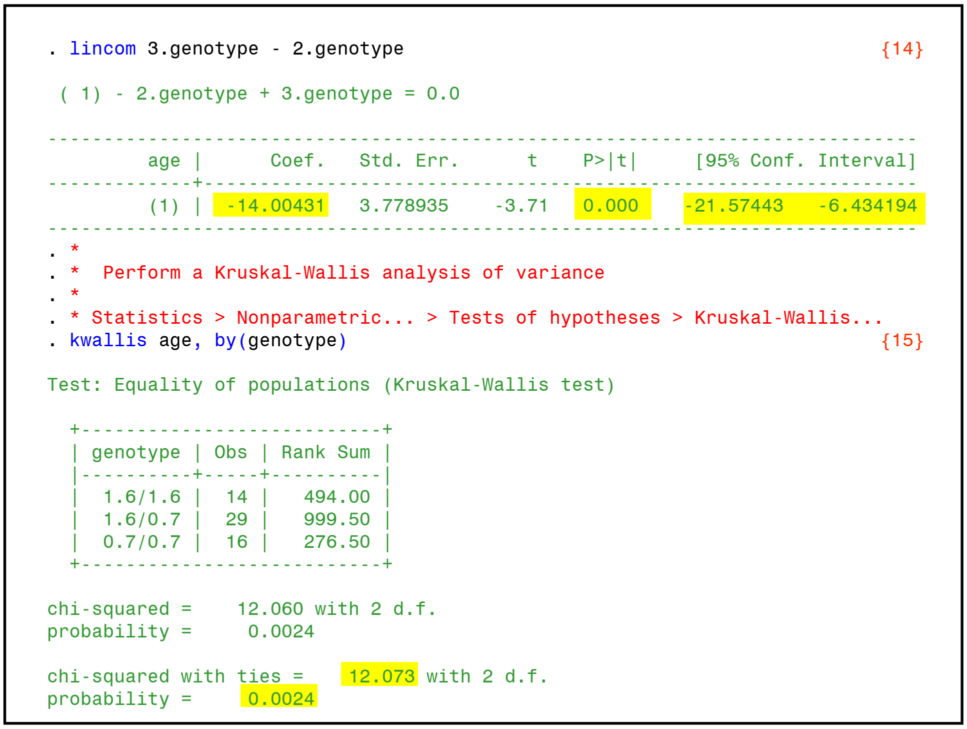
Dependent variable: age

Independent variables: i.genotype

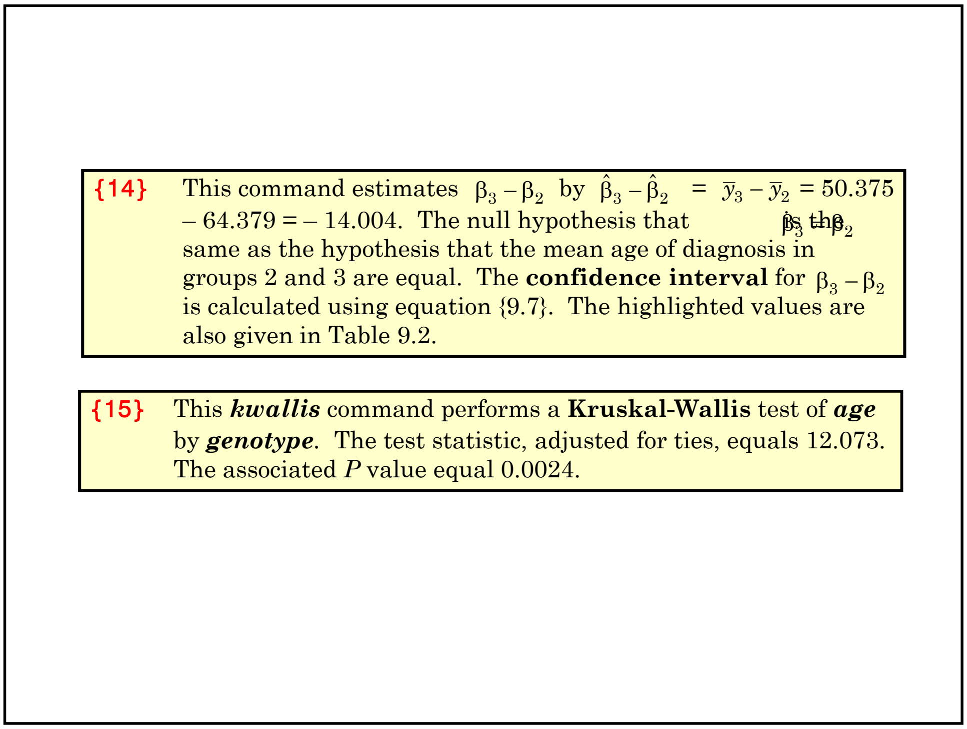
```
. lincom _cons + _Igenotype_2                                                {12}

 ( 1)  _Igenotype_2 + _cons = 0.0

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.       t     P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   64.37931   2.253322     28.57    0.000      59.86536    68.89326  {13}
------------------------------------------------------------------------------

. lincom _cons + _Igenotype_3

 ( 1)  _Igenotype_3 + _cons = 0.0

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.       t     P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |     50.375   3.033627     16.61    0.000      44.29791    56.45209
------------------------------------------------------------------------------
```

{12} This ***lincom*** command estimates $\alpha + \beta_2$ by $\hat{\alpha} + \hat{\beta}_2 = \bar{y}_2$. A 95 % confidence interval for this estimate is also given. Note that $\alpha + \beta_2$ equals the population mean age of diagnosis among women with the 1.6/0.7 genotype. Output from this and the next ***lincom*** command are also given in Table 9.1.

{13} This confidence interval is calculated using equation {9.3}.

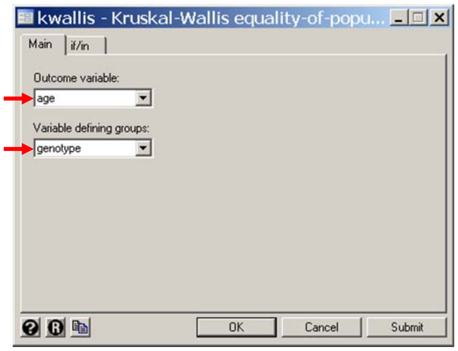


```
. lincom 3.genotype - 2.genotype                                           {14}

 ( 1) - 2.genotype + 3.genotype = 0.0

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -14.00431   3.778935    -3.71   0.000    -21.57443   -6.434194
------------------------------------------------------------------------------
. *
. *  Perform a Kruskal-Wallis analysis of variance
. *
. * Statistics > Nonparametric... > Tests of hypotheses > Kruskal-Wallis...
. kwallis age, by(genotype)                                                {15}

Test: Equality of populations (Kruskal-Wallis test)

  +---------------------------+
  | genotype | Obs | Rank Sum |
  |----------+-----+----------|
  |  1.6/1.6 |  14 |   494.00 |
  |  1.6/0.7 |  29 |   999.50 |
  |  0.7/0.7 |  16 |   276.50 |
  +---------------------------+

chi-squared =    12.060 with 2 d.f.
probability =     0.0024

chi-squared with ties =    12.073 with 2 d.f.
probability =     0.0024
```



{14} This command estimates $\beta_3 - \beta_2$ by $\hat{\beta}_3 - \hat{\beta}_2 = \bar{y}_3 - \bar{y}_2 = 50.375 - 64.379 = -14.004$. The null hypothesis that $\beta_3 = \beta_2$ is the same as the hypothesis that the mean age of diagnosis in groups 2 and 3 are equal. The **confidence interval** for $\beta_3 - \beta_2$ is calculated using equation {9.7}. The highlighted values are also given in Table 9.2.

{15} This ***kwallis*** command performs a **Kruskal-Wallis** test of ***age*** by ***genotype***. The test statistic, adjusted for ties, equals 12.073. The associated *P* value equal 0.0024.

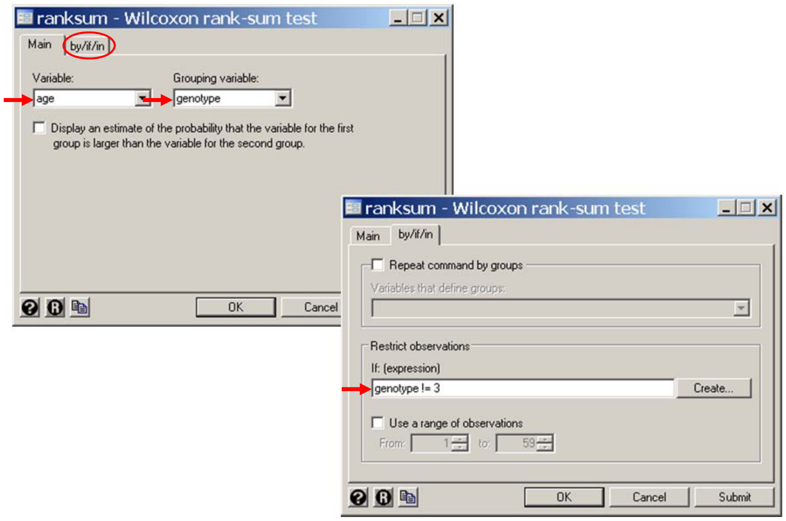


Outcome variable: age

Variable defining groups: genotype

```
. * Statistics > Nonparametric... > Tests... > Wilcoxon rank-sum test
. ranksum age if genotype !=3, by(genotype)                           {16}

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

    genotype |      obs    rank sum    expected
-------------+---------------------------------
     1.6/1.6 |       14         310         308
     1.6/0.7 |       29         636         638
-------------+---------------------------------
    combined |       43         946         946

unadjusted variance     1488.67
adjustment for ties       -2.70
                     ----------
adjusted variance       1485.97

Ho: age(genotype==1.6/1.6) = age(genotype==1.6/0.7)
             z =   0.052
    Prob > |z| =   0.9586
```

{16} This command performs a **Wilcoxon-Mann-Whitney rank-sum** test on the age of diagnosis of women with the 1.6/1.6 genotype versus the 1.6/0.7 genotype. The *P* value for this test is **0.96**. The next two commands perform the other two pair-wise comparisons of age by genotype using this rank-sum test. The highlighted *P* values are included in **Table 10.2**.



ranksum - Wilcoxon rank-sum test

Main | by/if/in

Variable: age

Grouping variable: genotype

□ Display an estimate of the probability that the variable for the first group is larger than the variable for the second group.

ranksum - Wilcoxon rank-sum test

Main | by/if/in

□ Repeat command by groups

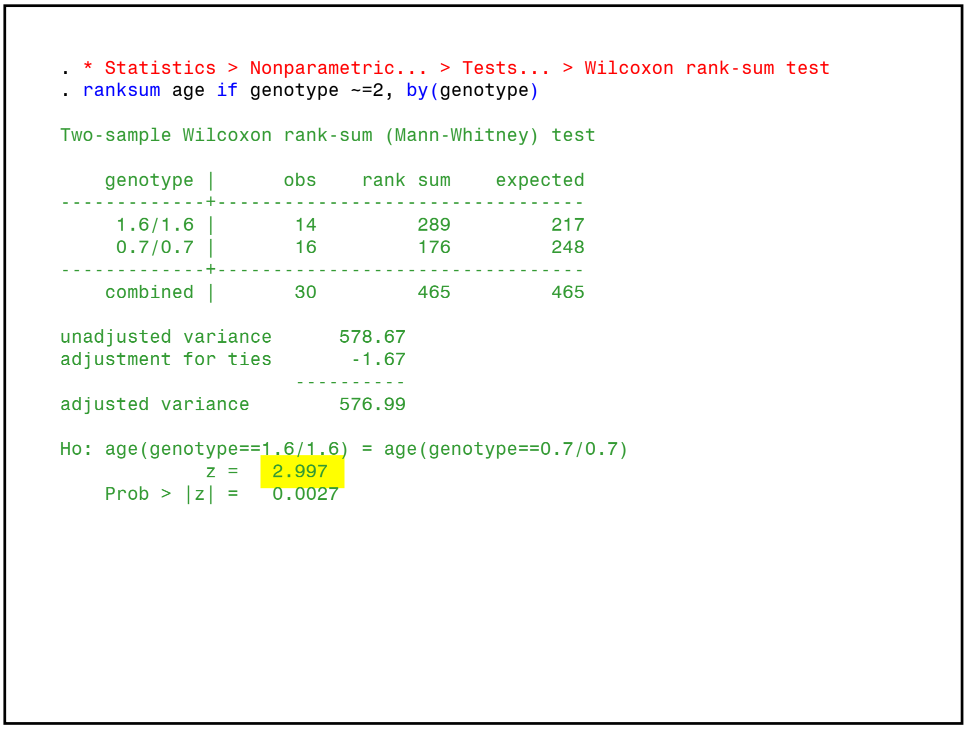
Variables that define groups:

Restrict observations

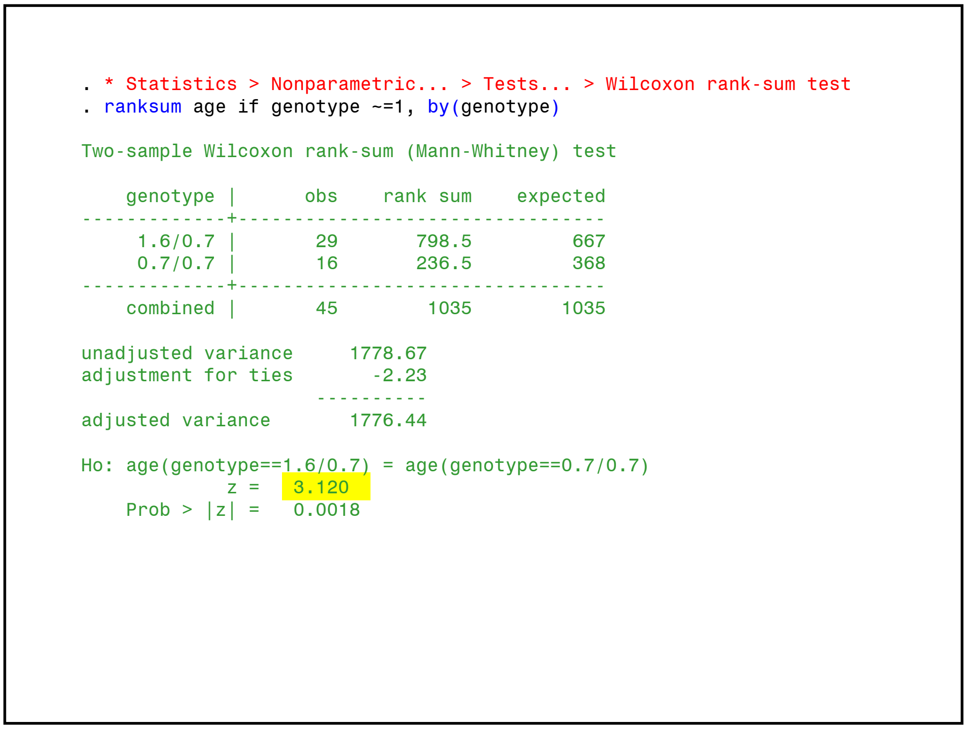
If: (expression) genotype != 3

□ Use a range of observations

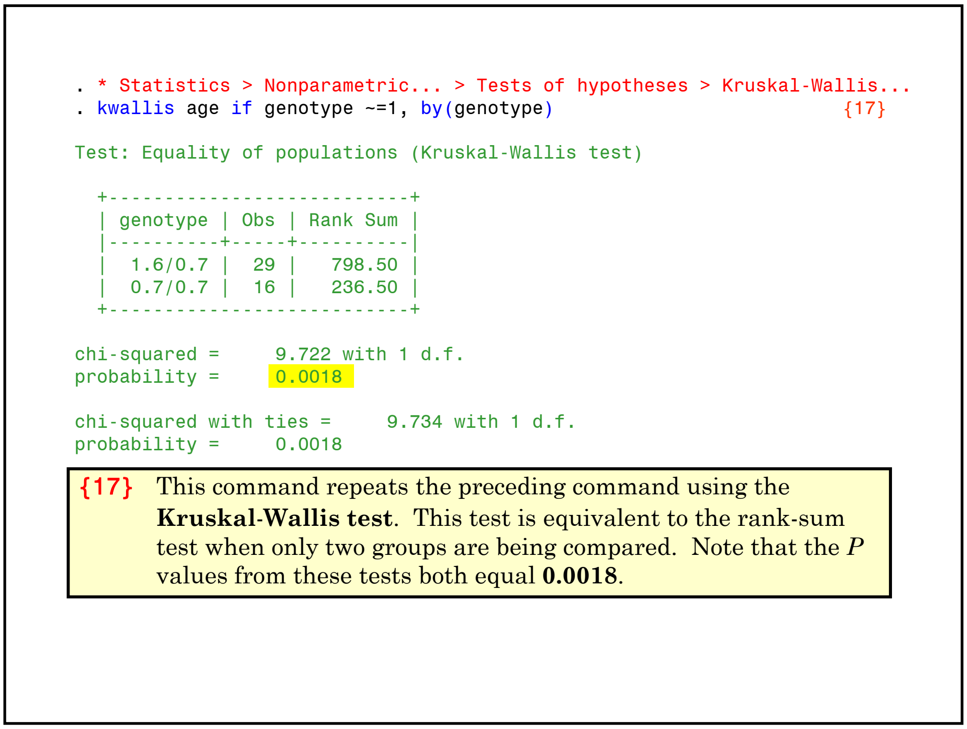
From: 1 to: 59



```
. * Statistics > Nonparametric... > Tests... > Wilcoxon rank-sum test
. ranksum age if genotype ~=2, by(genotype)

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

    genotype |      obs    rank sum    expected
-------------+---------------------------------
     1.6/1.6 |       14         289         217
     0.7/0.7 |       16         176         248
-------------+---------------------------------
    combined |       30         465         465

unadjusted variance      578.67
adjustment for ties       -1.67
                     ----------
adjusted variance        576.99

Ho: age(genotype==1.6/1.6) = age(genotype==0.7/0.7)
             z =   2.997
    Prob > |z| =   0.0027
```



```
. * Statistics > Nonparametric... > Tests... > Wilcoxon rank-sum test
. ranksum age if genotype ~=1, by(genotype)

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

    genotype |      obs    rank sum    expected
-------------+---------------------------------
     1.6/0.7 |       29       798.5         667
     0.7/0.7 |       16       236.5         368
-------------+---------------------------------
    combined |       45        1035        1035

unadjusted variance     1778.67
adjustment for ties       -2.23
                      ----------
adjusted variance       1776.44

Ho: age(genotype==1.6/0.7) = age(genotype==0.7/0.7)
             z =   3.120
    Prob > |z| =   0.0018
```



```
. * Statistics > Nonparametric... > Tests of hypotheses > Kruskal-Wallis...
. kwallis age if genotype ~=1, by(genotype)                          {17}

Test: Equality of populations (Kruskal-Wallis test)

  +--------------------------+
  | genotype | Obs | Rank Sum |
  |----------+-----+----------|
  |  1.6/0.7 |  29 |   798.50 |
  |  0.7/0.7 |  16 |   236.50 |
  +--------------------------+

chi-squared =     9.722 with 1 d.f.
probability =     0.0018

chi-squared with ties =     9.734 with 1 d.f.
probability =     0.0018
```

{17} This command repeats the preceding command using the **Kruskal-Wallis test**. This test is equivalent to the rank-sum test when only two groups are being compared. Note that the *P* values from these tests both equal **0.0018**.

9. Two-Way Analysis of Variance, Analysis of Covariance, and Other Models

Fixed-effects analyses of variance generalize to a wide variety of complex models. For example, suppose that hypertensive patients were treated with either a placebo, a diuretic alone, a beta-blocker alone, or with both a diuretic and a beta-blocker. Then a model of the effect of treatment on diastolic blood pressure (DBP) might be

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad \{9.11\}$$

where

α ,β_1 and β_2 are unknown parameters,

$$x_{i1} = \begin{cases} 1: & i^{th} \text{ patient is on a diuretic} \\ 0: & \text{otherwise} \end{cases}$$

$$x_{i2} = \begin{cases} 1: & i^{th} \text{ patient is on a beta-blocker} \\ 0: & \text{otherwise} \end{cases}$$

y_i is the DBP of the *i*th patient after some standard interval therapy , and

ε_i are error terms that are independently and normally distributed with mean zero and standard deviation σ

Model {9.11} is an example of a fixed-effects, **two-way analysis of variance.**

It is called two-way because each patient is simultaneously influenced by two covariates — in this case whether she did, or did not, receive a diuretic or a beta-blocker.

A critical feature of this model is that each patient's blood pressure is only observed once.

It is this feature that makes the independence assumption for the error term reasonable and makes this a fixed-effects model. In this model,

α is the mean DBP of patients on placebo,

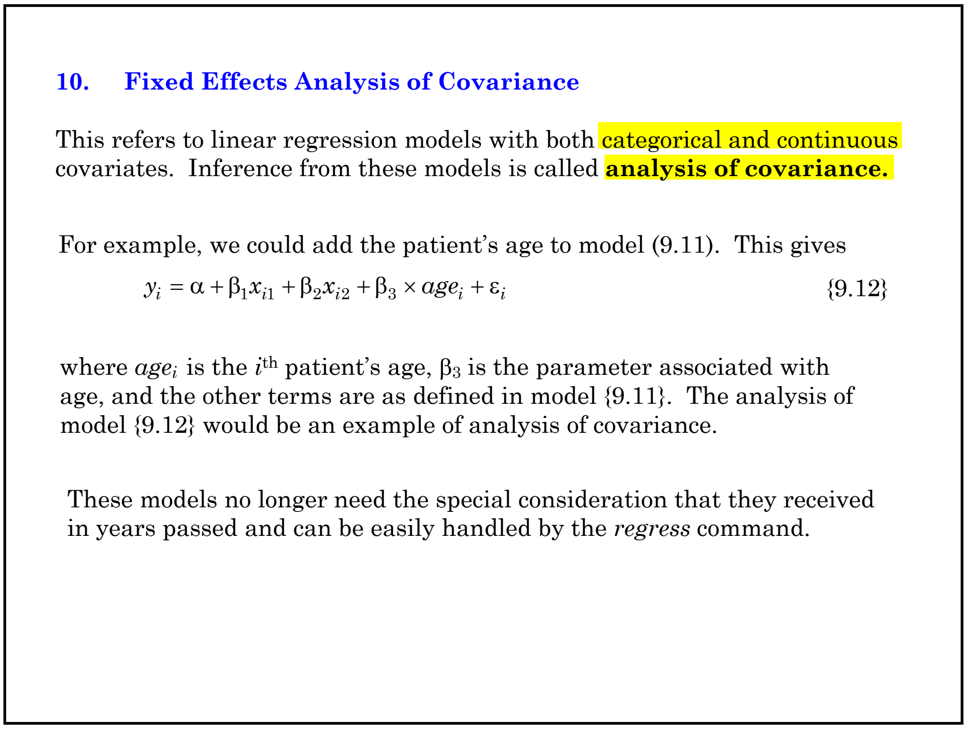
$\alpha + \beta_1$ is the mean DBP of patients on the diuretic alone,

$\alpha + \beta_2$ is the mean DBP of patients on the beta-blocker alone, and

$\alpha + \beta_1 + \beta_2$ is the mean DBP of patients on both treatments.

The model is additive since it assumes that the mean DBP of patients on both drugs is $\alpha + \beta_1 + \beta_2$.

If this assumption is unreasonable, we can add an interaction term as in Section 3.12.



10. Fixed Effects Analysis of Covariance

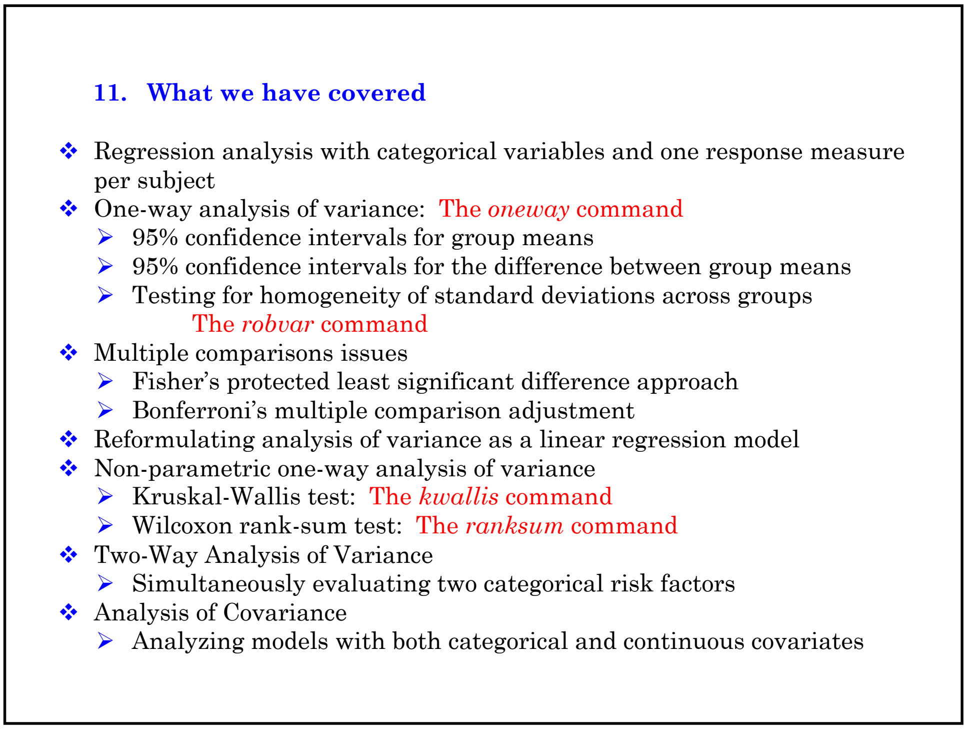
This refers to linear regression models with both categorical and continuous covariates. Inference from these models is called **analysis of covariance.**

For example, we could add the patient's age to model (9.11). This gives

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 \times age_i + \varepsilon_i \qquad \{9.12\}$$

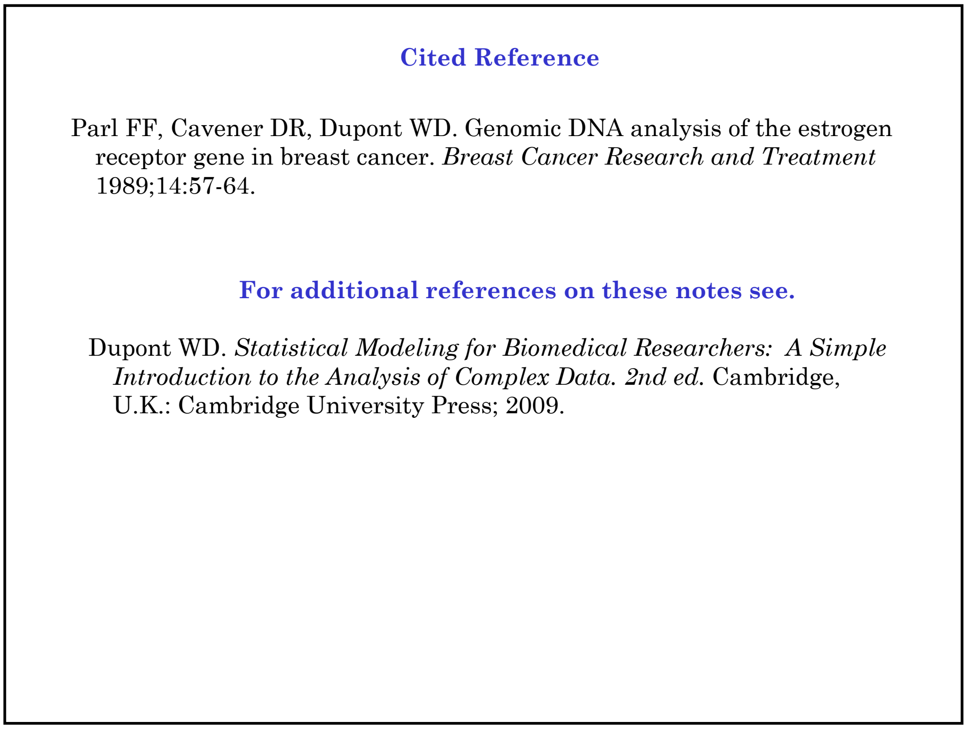
where age_i is the i^{th} patient's age, β_3 is the parameter associated with age, and the other terms are as defined in model {9.11}. The analysis of model {9.12} would be an example of analysis of covariance.

These models no longer need the special consideration that they received in years passed and can be easily handled by the *regress* command.



11. What we have covered

- Regression analysis with categorical variables and one response measure per subject
- One-way analysis of variance: The *oneway* command
  - 95% confidence intervals for group means
  - 95% confidence intervals for the difference between group means
  - Testing for homogeneity of standard deviations across groups
    The *robvar* command
- Multiple comparisons issues
  - Fisher's protected least significant difference approach
  - Bonferroni's multiple comparison adjustment
- Reformulating analysis of variance as a linear regression model
- Non-parametric one-way analysis of variance
  - Kruskal-Wallis test: The *kwallis* command
  - Wilcoxon rank-sum test: The *ranksum* command
- Two-Way Analysis of Variance
  - Simultaneously evaluating two categorical risk factors
- Analysis of Covariance
  - Analyzing models with both categorical and continuous covariates



Cited Reference

Parl FF, Cavener DR, Dupont WD. Genomic DNA analysis of the estrogen receptor gene in breast cancer. *Breast Cancer Research and Treatment* 1989;14:57-64.

For additional references on these notes see.

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd ed.* Cambridge, U.K.: Cambridge University Press; 2009.