Section 12.1 - Inference for One-Way Analysis of Variance

Statistics 104

Autumn 2004

Analysis of Variance

An approach for analyzing data sets with a continuous response variable and categorical predictor variables.

One-Way Analysis of Variance (ANOVA)

ANOVA with a single categorical predictor variable.

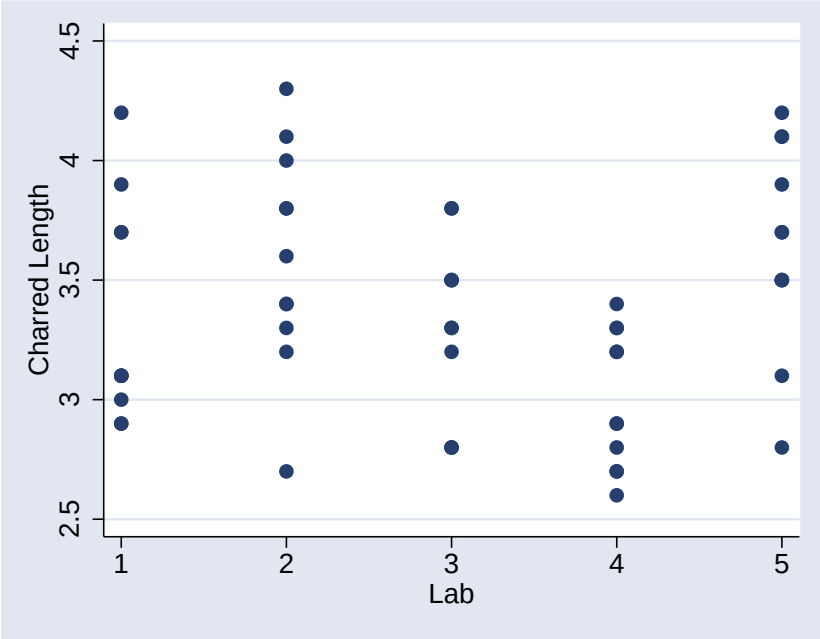
Example: Fabric flammability testing

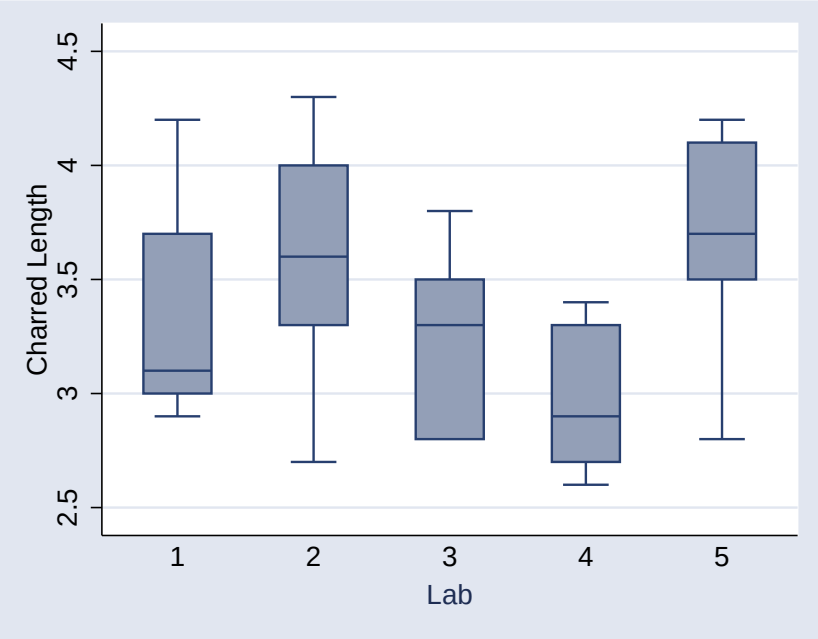
5 labs were involved in testing the flammability of fabric. One question of interest was whether the 5 labs give similar responses in their testing.

Each lab tested 11 pieces of cloth

$y =$ length of charred part of material

$x =$ laboratory





Model:

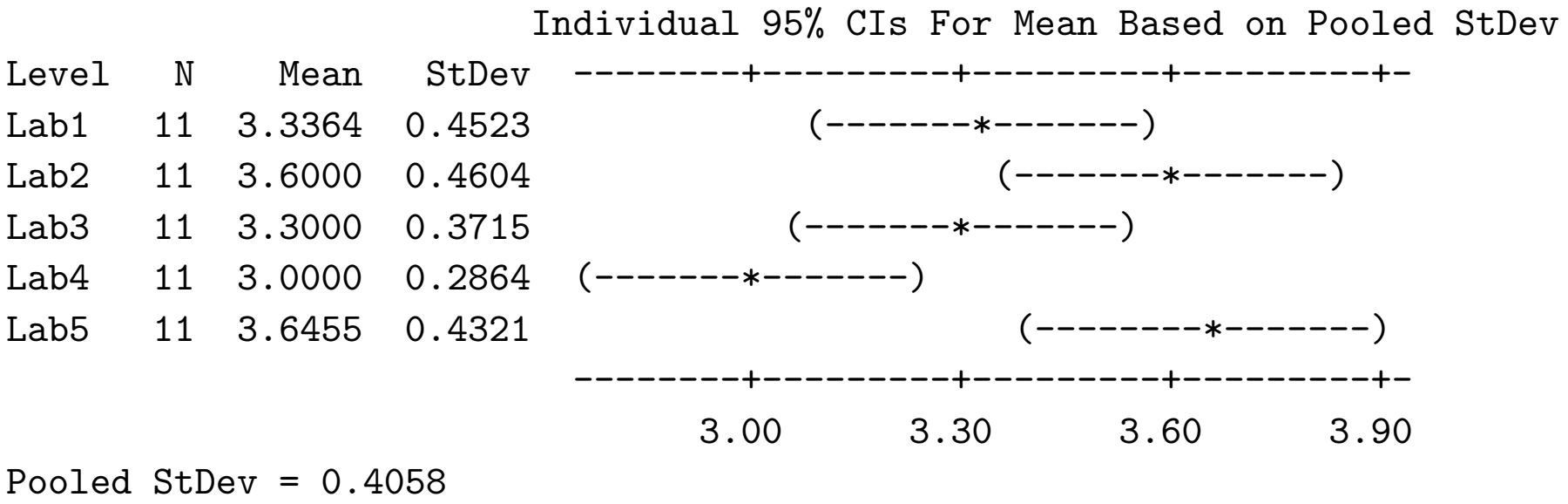
Each population (level of the predictor variable) has its own mean. The observations from each population are normally distributed with the same standard deviation

$$\text{Pop } i : y_{ij} \sim N(\mu_i, \sigma); j = 1, \ldots, n_i, i = 1, \ldots, I$$

This is a similar model as for the pooled two sample t procedure, except that there are more than two populations.

While we will be assuming that the standard deviations for each population are the same, this assumption can be relaxed.

Summary statistics for each lab



It appears that there might be some differences between different labs.

Fitting the model:

Need to estimate the means for each population $(\mu_1, \mu_2, \ldots, \mu_I)$ and the common standard deviation σ.

Estimate μ_i with $\bar{y}_i$, the average of all observations from population i.

Let s_i be the sample standard deviation from population i.

Then the pooled estimate of the standard deviation is given by

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + \ldots + (n_I - 1)s_I^2}{N - I} \\ s_p &= \sqrt{s_p^2} \end{aligned}$$

where $N = n_1 + \ldots + n_I$ is the total number of observations.

This is just like s_p for the two-sample pooled t procedures, except there are I groups, instead of 2.

The confidence intervals for μ_i shown in the earlier Minitab output were calculated with the following formula

$$\text{Lab } i: \quad \bar{y}_i \pm t^* \frac{s_p}{\sqrt{n_i}}$$

where t^* is based on $N - I$ degrees of freedom.

With these confidence intervals, some overlap greatly, suggesting that those labs might be similar, and some don't (e.g. Labs 4 & 5), suggesting that those labs might be different.

Suggests looking at the hypotheses

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I \quad \text{vs} \quad H_A : \text{at least one } \mu_i \text{ is different}$$

(don't want to specify which one is different)

As in the regression setting, we can describe these two hypotheses through a set of possible models.

- Full Model (H_A):

$$y_{ij} \sim N(\mu_i, \sigma)$$

- Reduced Model (H_0):

$$y_{ij} \sim N(\mu, \sigma)$$

These models can also be written as

- Full Model (H_A):

$$y_{ij} = \mu_i + \epsilon_{ij}$$

- Reduced Model (H_0):

$$y_{ij} = \mu + \epsilon_{ij}$$

Note that the framework for this model is the similar to the use of indicator variables in regression. These models can be written in that framework and describe exactly the same model.

As in the regression setting, these models can be compared through a Sums of Squares decomposition.

$$\begin{aligned} SST &= \sum (y_{ij} - \bar{y})^2 && \text{(Total sums of squares)} \\ SSM &= \sum (\bar{y}_i - \bar{y})^2 && \text{(Model SS)} \\ SSE &= \sum (y_{ij} - \bar{y}_i)^2 && \text{(Error or Residual)} \end{aligned}$$

$$SST = SSM + SSE$$

Similarly, there is also a decomposition of degrees of freedom

$$\begin{aligned} DFT &= N-1 \\ DFM &= I-1 \\ DFE &= N-I \end{aligned}$$

$$DFT = DFM + DFE$$

As in the regression setting, SSM is a measure of how much better the Full model describes the data over the Reduced model.

ANOVA Table

Source	DF	SS	MS	F
Model	$I-1$	$SSM = \sum(\bar{y}_i - \bar{y})^2$	$MSM = \frac{SSM}{DFM}$	$F = \frac{MSM}{MSE}$
Error	$N-I$	$SSE = \sum(y_{ij} - \bar{y}_i)^2$	$MSE = \frac{SSE}{DFE}$	
Total	$N-1$	$SST = \sum(y_{ij} - \bar{y})^2$		

The null hypothesis can be examined with the F test

$$F = \frac{MSM}{MSE}$$

and compared to an $F(I-1, N-I)$ distribution.

There are a couple of ways of getting this analysis out in Stata.

```
. oneway charred lab, tabulate

           |         Summary of Charred
       Lab |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |   3.3363636     .452267           11
         2 |         3.6   .46043457           11
         3 |         3.3   .37148352           11
         4 |           3   .28635643           11
         5 |   3.6454545    .4321195           11
-----------+------------------------------------
     Total |   3.3763636   .45581169           55

                       Analysis of Variance
    Source              SS         df      MS             F     Prob > F
------------------------------------------------------------------------
Between groups      2.98654522      4   .746636305       4.53     0.0033
 Within groups      8.23272705     50   .164654541
------------------------------------------------------------------------
    Total           11.2192723     54   .207764301
Bartlett's test for equal variances:  chi2(4) =   2.6755  Prob>chi2 = 0.614
```

```
. anova charred lab

                   Number of obs =      55     R-squared     =  0.2662
                   Root MSE      = .405776     Adj R-squared =  0.2075

          Source |  Partial SS    df       MS           F     Prob > F
      -----------+----------------------------------------------------
           Model |  2.98654522     4  .746636305       4.53     0.0033
                 |
             lab |  2.98654522     4  .746636305       4.53     0.0033
                 |
        Residual |  8.23272705    50  .164654541
      -----------+----------------------------------------------------
           Total |  11.2192723    54  .207764301
```

The `oneway` procedure allows for the summaries for each group to be given with the `tabulate` option. The `anova` procedure can be used when there are more than one predictor variable. It can also be used to examine the analysis you would get if you used indicator variables in a regression to analyze the data (`regress` option).

```
. anova charred lab, regress

      Source |       SS       df       MS          Number of obs =      55
-------------+------------------------------       F(  4,    50) =    4.53
       Model |  2.98654522     4  .746636305        Prob > F      =  0.0033
    Residual |  8.23272705    50  .164654541        R-squared     =  0.2662
-------------+------------------------------       Adj R-squared =  0.2075
       Total |  11.2192723    54  .207764301        Root MSE      =  .40578

------------------------------------------------------------------------------
 charred        Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------------------------------------------------------------------
_cons         3.645455   .1223462    29.80   0.000     3.399715    3.891194
lab
      1    -.3090909   .1730237    -1.79   0.080    -.6566192    .0384374
      2    -.0454545   .1730237    -0.26   0.794    -.3929828    .3020738
      3    -.3454545   .1730237    -2.00   0.051    -.6929828    .0020737
      4    -.6454545   .1730237    -3.73   0.000    -.9929828   -.2979262
      5    (dropped)
------------------------------------------------------------------------------
```

Since we observe a large F statistic (and a small p-value), it appears that there are a difference between some of the labs. However this analysis doesn't tell us which ones.

There is a slightly different way at looking at the sums of squares.

$$\begin{aligned} SST &= \sum_{obs}(y_{ij} - \bar{y})^2 \\ SSM &= \sum_{obs}(\bar{y}_i - \bar{y})^2 = \sum_{groups} n_i(\bar{y}_i - \bar{y})^2 \\ SSE &= \sum_{obs}(y_{ij} - \bar{y}_i)^2 = \sum_{groups} (n_i - 1)s_i^2 \end{aligned}$$

A couple of things drop out of this view. First

$$MSE = s_p^2 \quad \text{and} \quad RootMSE = s_p$$

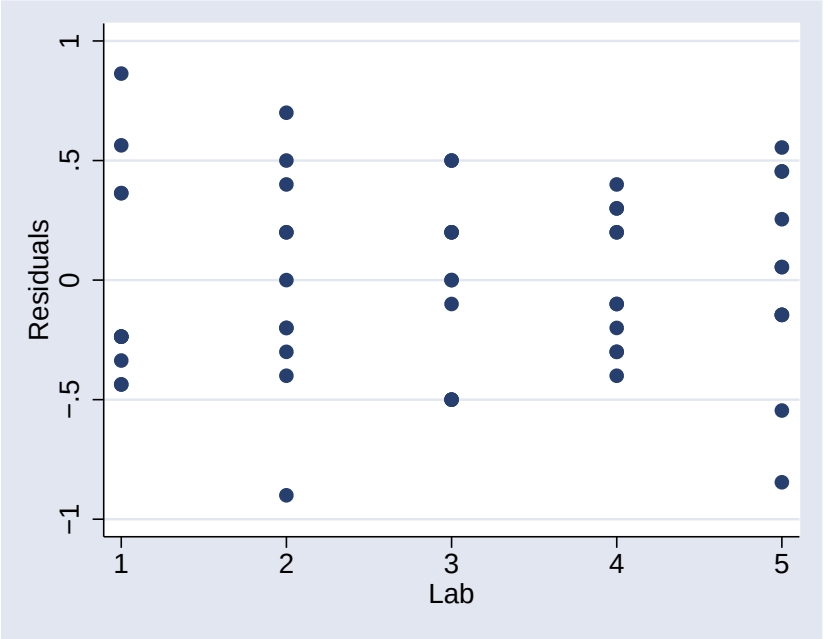
Also we can think of SSM as a measure of how far each group average is from the overall average.

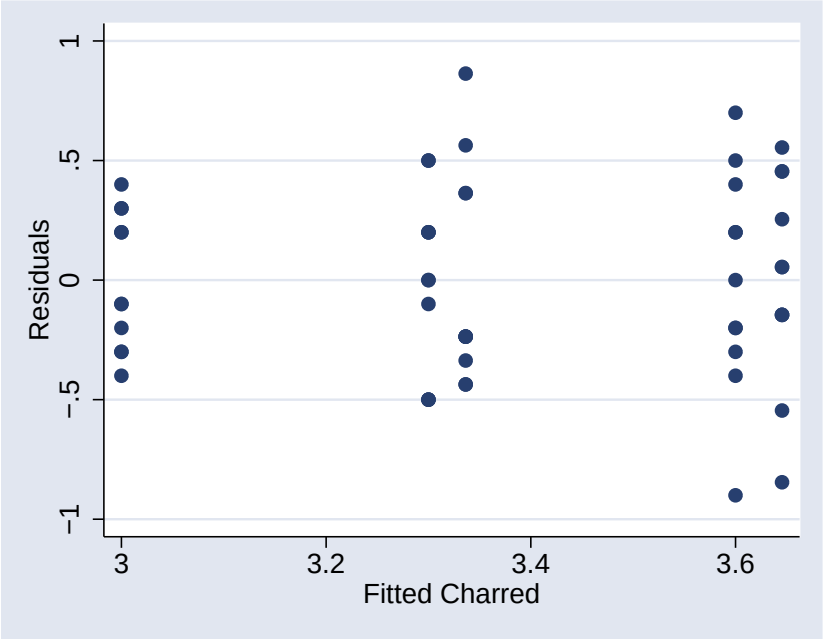
As with regression, the coefficient of determination is also defined for ANOVA. As before

$$R^2 = \frac{SSM}{SSE}$$

For the example, $R^2 = 0.27$, which is not particularly large. This is not particularly surprising since the data for each lab is fairly variable.

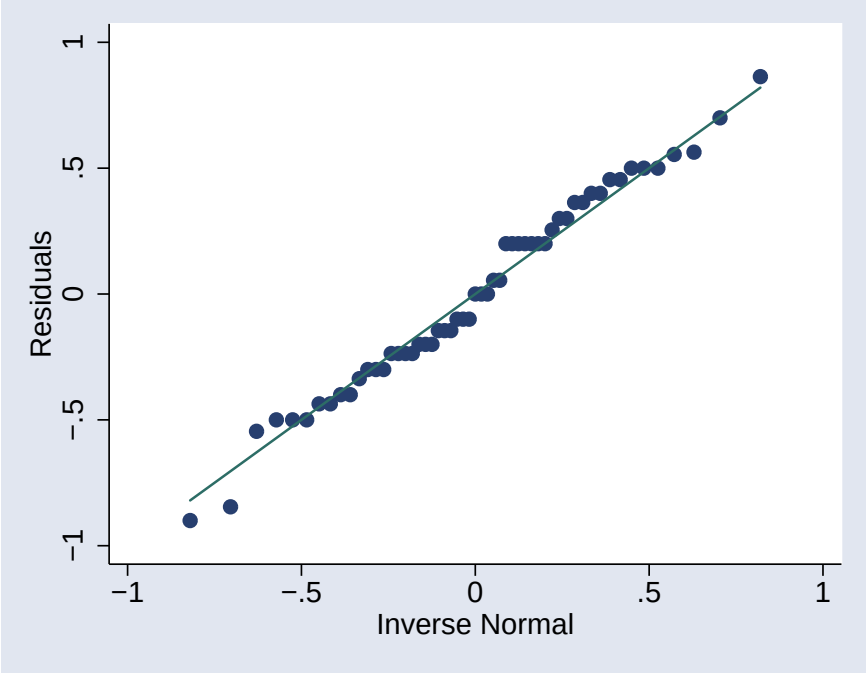
To get R^2 directly from Stata, you must use the `anova` command. Also this command is needed to get the fits, residuals, etc from the `predict` command.






```
                                Individual 95% CIs For Mean Based on Pooled StDev
Level    N    Mean   StDev  --------+---------+---------+---------+-
Lab4    11  3.0000  0.2864  (-------*-------)
Lab3    11  3.3000  0.3715          (-------*-------)
Lab1    11  3.3364  0.4523           (-------*-------)
Lab2    11  3.6000  0.4604                    (-------*-------)
Lab5    11  3.6455  0.4321                     (--------*-------)
                                --------+---------+---------+---------+-
                                      3.00      3.30      3.60      3.90
```

Now there is a slight suggestion that there might be a increasing variance of the deviations as the mean increases, but it is not particularly worrisome here.



The normality assumption looks reasonable here.

What if $I = 2$?

As mentioned earlier, the model for One-Way ANOVA is similar to the model for the pooled two-sample t procedures.

It is possible to show that an F test from a One-Way ANOVA will give the same answer as the pooled two-sample t test on the hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_A : \mu_1 \neq \mu_2$$

It is possible to show that $t^2 = F$ and the p-values will be the same.

```
. ttest btuin, by(damper)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |  Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     EVD |   40     9.90775    .4774831    3.019868    8.941949    10.87355
     TVD |   50      10.143    .3913156    2.767019    9.356622    10.92938
---------+--------------------------------------------------------------------
combined |   90    10.03844    .3023127     2.86799    9.437755    10.63913
---------+--------------------------------------------------------------------
    diff |        -.2352499    .6113255               -1.450131     .979631
------------------------------------------------------------------------------
Degrees of freedom: 88

                    Ho: mean(EVD) - mean(TVD) = diff = 0

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
       t =  -0.3848                 t =  -0.3848                 t =  -0.3848
   P < t =   0.3506           P > |t| =   0.7013           P > t =   0.6494
```

```
. oneway btuin damper, tabulate

           |          Summary of BTUIn
    Damper |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
       EVD |     9.90775    3.019868          40
       TVD |      10.143   2.7670195          50
-----------+------------------------------------
     Total |   10.038444   2.8679903          90

                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1.22983418      1   1.22983418      0.15     0.7013
 Within groups      730.827941     88   8.30486297
------------------------------------------------------------------------
    Total           732.057775     89   8.22536826
```