Displays for Statistics 5303

Lecture 7

September 18, 2002

Christopher Bingham, Instructor

612-625-7023 (St. Paul)
612-625-1024 (Minneapolis)

Class Web Page
http://www.stat.umn.edu/~kb/classes/5303

Coping with serial correlation

This example is based on the data in Example 6.3. They are differences in temperature readings for 64 consecutive simultaneous measurements with two thermocouples. The data file has a single column of differences.

```
Cmd> diffs <- read("","exmpl6.3") # or matread()
exmpl6.3     64
) A data set from Oehlert (2000) \emph{A First Course in Design
) and Analysis of Experiments}, New York: W. H. Freeman.
)
) Data originally from Christensen, R. and L.~G. Blackwood .
) (1993)``Tests for precision and accuracy of multiple measuring
) devices.''{\em Technometrics\/}~{\em 35}, 411--420.
) Table 6.2, p. 121
) Temperature differences in degrees Celsius between
) two thermocouples for 64 consecutive readings.
Read from file "TP1:Stat5303:Data:OeCh06.dat"
```

Here we just want to estimate the mean μ. This is the same as fitting the model

$$d_i = \mu + \varepsilon_i,\ i = 1,\ldots,n = 64.$$

I want to do this using **anova()** so I can use commands like **resvsrankits()** that work only after **anova()**, **regress()** and other similar commands. To fit only a mean, you use the model **"diffs = 1"**.

```
Cmd> anova("diffs = 1",fstat:T)
Model used is diffs = 1
             DF          SS          MS           F     P-value
CONSTANT      1      631.52      631.52  1.0022e+06           0
ERROR1       63      0.0397  0.00063016
```

This is one of the few times when the **CONSTANT** line might be useful.

- $SS = N\bar{y}_{\cdot\cdot}^2$
- The F-statistic tests H_0: $\mu = 0$. $F = t^2$, where t = one sample t-statistic:

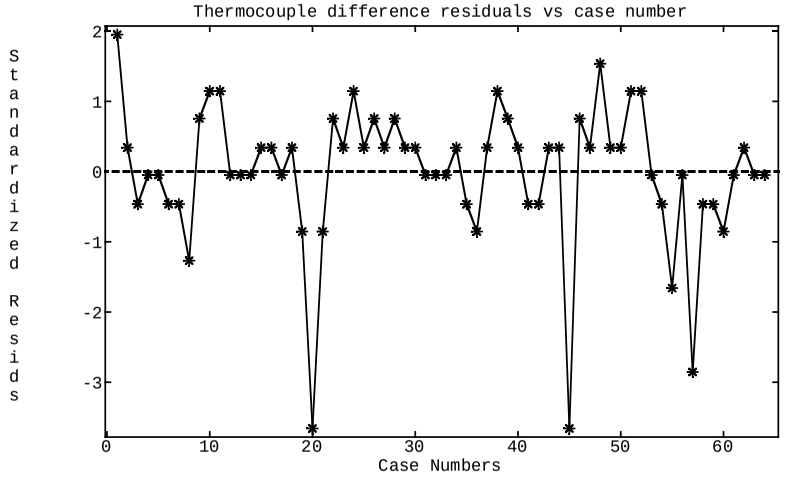
```
Cmd> tval(diffs)^2 # numerical check; same as F
(1)  1.0022e+06
```

But that's not what we are interested in here. We are checking assumptions.

The observations were entered in time order. This allows some checking of the **independence assumption.**

You can at least check that successive observations are uncorrelated. If they are correlated, we say there is data is **autocorrelated** or has **serial correlation**. I used **resvsindex()** to plot residuals against case number:

```
Cmd> resvsindex(lines:T,\
     title:"Thermocouple difference residuals vs case number")
```



lines:T directs drawing lines between successive points.

There a tendency for positive residuals to follow positive residuals and negative residuals to follow negative residuals (or high response levels to follow high and low to follow low).

This is a sign of *positive serial correlation*, the most common kind.

In a significance test of the hypothesis of no serial correlation, it is common to do a one-tail test, protecting against positive serial correlation, but not against the much rarer negative serial correlation.

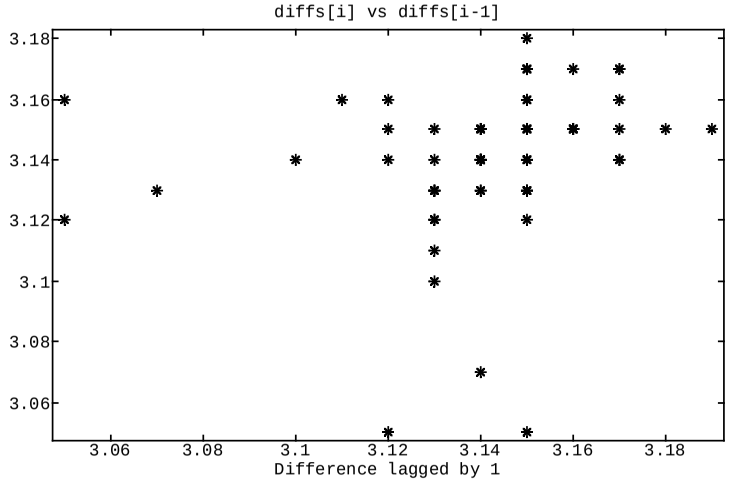
Here are two or three ways to check the null hypothesis of that successive values are independent against the alternative that there is non-zero first order serial correlation.

A simple graphical check is to plot y_i vs y_{i-1}, the value for the preceding time.

If you see an apparent linear pattern, or even any pattern, there is serial correlation.

When there is no serial correlation you should get a featureless blob of points.

```
Cmd> plot(diffs[-64],diffs[-1],xlab:"Difference lagged by 1",\
  title:"diffs[i] vs diffs[i-1]")
```



diffs[-64] (on the x-axis) is **diffs** without case 64, running from case 1 to case 63. **diffs[-1]** (on the y-axis) is **diffs** running from case 2 to case 64, one step ahead of **diffs[-64]**.

There is some tendency for high differences to be followed by high (plotted in the upper right hand corner) and low followed by low (lower left hand corner), just as we saw in the time plot.

You can use **cor()** to estimate the first order serial correlation $\rho_1 \equiv \text{Corr}[y_i, y_{i-1}]$ or **regress()** to fit a regression of d_i on d_{i-1}.

When the residuals are independent, the estimated first order serial correlation has an approximate standard error $1/\sqrt{n}$.

```
Cmd> cor(diffs[-1],diffs[-64])[1,2] #serial correlation
(1,1)     0.21994        Estimated 1st order ser. corr

Cmd> zstat <- 0.21994/(1/sqrt(64)) # z-statistic=estimate/SE

Cmd> 1 - cumnor(zstat) # one sided P-value
(1)    0.039245      Significant at 5% level
```


Now regress responses for cases 2 - 64 on responses for cases 1 - 63, lagged one step behind.

```
Cmd> regress("{diffs[-1]} = {diffs[-64]}",pval:T)
Model used is {diffs[-1]} = {diffs[-64]}
                    Coef      StdErr           t     P-Value
CONSTANT          2.4709     0.38025      6.4981  1.6761e-08
{diffs[-64]}     0.21315     0.12105      1.7609    0.083262

N: 63, MSE: 0.00058167, DF: 61, R^2: 0.04837
Regression F(1,61): 3.1009, P-value: 0.083262, Durbin-Watson:
1.9948
To see the ANOVA table type 'anova()'
```

The printed P-value .08326 is a *two* tail P-value.

The one tail P-value is .08326/2 = .04163, not much different from the P-value for the z-statistic.

The slope 0.21315 is an estimate of the first order serial correlation that is different from, but close to .21994.

Another test statistic is the **Durbin-Watson** statistic DW, computed from the residuals from the previous ANOVA. Since these were destroyed by **regress()** I needed to run **anova()** again to restore them.

```
Cmd> anova("diffs=1",silent:T) # redo anova() silently
```

$$DW = \frac{\sum_{i=2}^{n}(r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2},$$

where r_i is the residual for case i.

- When there is no serial correlation, $\mu_{DW} = 2$
- When $\rho_1 > 0$, $\mu_{DW} < 2$.

So a test based on DW rejects independence in favor of positive serial correlation when DW is significantly < 2.

```
Cmd> dw <- sum((RESIDUALS[-1]- RESIDUALS[-64])^2)/\
                sum(RESIDUALS^2)

Cmd> dw # observed DW statistic
(1)      1.5139
```


It can be shown that **(2 - DW)/2** is another estimate of the first order serial correlation

```
Cmd> (2 - dw)/2
(1)     0.24307      Not far from .0213 and .0220
```

It would be nice to be able to check whether DW is significantly below 2.

MacAnova doesn't have a function to compute critical values for DW, but you can estimate the P-value by **simulating** samples of independent normal data (for which H_0 is true), computing values of DW for each sample and seeing the proportion of values that are less than the observed.

```
Cmd> M <- 5000; DW <- rep(0,M) # do 5000 repetitions

Cmd> for(i,1,M){
        tmpdiffs <- rnorm(64)
        anova("tmpdiffs=1",silent:T)
        DW[i] <- sum((RESIDUALS[-1]- RESIDUALS[-64])^2)/\
           sum(RESIDUALS^2)
    ;;}

Cmd> sum(DW <= dw) # number <= observed
(1)         120

Cmd> 120/M # proportion <= observed = one tail P-value
(1)       0.024
```

P-value < .05, confirming the other tests.

Caution: These tests for serial correlation, including the Durbin-Watson test are appropriate *only when the data have a relevant time order*. The results make sense only if the statistic is based on an actual time order.

The Durbin-Watson statistic DW is always part of **regress()** output, whether or not the order of cases is meaningful. At present the only easy way in MacAnova to test its significance is by simulation.

DW is also used with independent data as a test of constant mean.

If there is a trend in the mean, whether linear or curvilinear, there will be a tendency for bunches of successive residuals $y_i - \overline{y}_.$ to have the same sign.

For instance, when there is a strong increasing linear trend with time, the first half of the $y_i - \overline{y}_.$ will probably be negative and the last half positive.

This is just what makes the DW statistic small.

In this situation where you are not testing for serial correlation but for non-constant mean, the test is often called the **von Neumann test**.

More on contrasts

A *contrast* is a linear combination of μ's or α's

$$w(\{\mu_i\}) \equiv \sum_i w_i \mu_i, \text{ with } \sum_i w_i = 0$$

Because $\sum_i w_i = 0$, $w(\{\mu_i\})$ doesn't depend on μ^* and $w(\{\mu_i\}) = \sum_i w_i \alpha_i = w(\{\alpha_i\})$

Under the ANOVA assumptions (independent samples with equal σ), the variance of an observed contrast

$$w(\{\overline{y_{i\bullet}}\}) = \sum_i w_i \overline{y_i}$$

is

$$\text{Var}[w(\{\overline{y_{i\bullet}}\})] = \sigma^2 \sum_{1 \le i \le g} w_i^2 / n_i$$

When $n_1 = n_2 = \ldots n_g = n$, this simplifies to

$$\text{Var}[w(\{\overline{y_{i\bullet}}\})] = = \sigma^2 (\sum_{1 \le i \le g} w_i^2)/n$$

The *estimated* standard error is

$$\hat{S}E[w(\{\overline{y_{i\bullet}}\})] = s_p \times \sqrt{\{\sum_{1 \le i \le g} w_i^2 / n_i\}}$$

where $s_p^2 = MS_E$ from ANOVA.

For a comparison $\alpha_i - \alpha_j$ of two treatment effects, the estimated contrast is $\hat{\alpha}_i - \hat{\alpha}_j = \hat{\mu}_i - \hat{\mu}_j = \overline{y_{i\bullet}} - \overline{y_{j\bullet}}$. This has standard error $\sigma\sqrt{\{1/n_i + 1/n_j\}}$.

When $n_i = n_j = n$, $SE[\hat{\alpha}_i - \hat{\alpha}_j] = \sigma\sqrt{(2/n)}$.

Continuing with the example of log resin times to failure:

```
Cmd> anova("logy=treat",fstat:T)
Model used is logy=treat
WARNING: summaries are sequential
             DF          SS          MS           F     P-value
CONSTANT      1      79.425      79.425  8653.95365  1.6145e-40
treat         4      3.5376     0.88441    96.36296  2.2419e-17
ERROR1       32     0.29369   0.0091779

Cmd> dfe <- DF[3]; dfe #error degrees of freedom
      ERROR1
          32

Cmd> mse <- SS[3]/dfe; mse # Mean square error = s_p^2
      ERROR1
   0.0091779

Cmd> n <- tabs(,treat); n # sample sizes
(1)           8           8           8           7           6

Cmd> w <- vector(vector(1,1)/2,-vector(1,1,1)/3);w #contrast
(1)         0.5         0.5    -0.33333    -0.33333    -0.33333

Cmd> sqrt(mse*sum(w^2/n)) # estimated standard error
(1)    0.031886

Cmd> contrast(treat,w)$se # black box approach
(1)    0.031886      Same as direct computation
```

Under the normality assumption, you can use Student's t with $df_{error} = N - g$ degrees of freedom in tests or confidence intervals.

Specifically, when $W(\{\mu_i\}) = \sum_i w_i\mu_i = 0$, $t = \sum_i w_i\overline{y}_{i\bullet}/(s_p \times \sqrt{\{\sum_{1\le i\le g} w_i^2/n_i\}})$, $s_p = \sqrt{MSE}$ is distributed as t_{N-g}.

```
Cmd> tstat <- contrast(treat,w)$estimate/contrast(treat,w)$se

Cmd> twotailt(tstat, dfe) # find two-tail P-value
(1)  3.0663e-18          Essentially 0
```

A $1 - \alpha$ confidence interval for $\sum_i w_i\mu_i$ of has the usual form

$$\underline{\text{estimate}} \pm \underline{\text{critical value}} \times \underline{\text{std error}}$$

$$= \sum_i w_i\overline{y}_{i\bullet} \pm t_{1-\alpha/2} s_p \times \sqrt{\{\sum_{1\le i\le g} w_i^2/n_i\}}$$

```
Cmd> contrast(treat,w)$estimate + \
 vector(-1,1)*invstu(1 - .025,dfe)*contrast(treat,w)$se
(1)     0.50619     0.63609
```

`vector(-1,1)` is a MacAnova way to express ± 1.

The sum of squares SS for a contrast is

$$SS_w = w(\{\overline{y}_{i\bullet}\})^2/\{\sum_{1\le i\le g} w_i^2/n_i\}$$

```
Cmd> vector(sum(w*muhats)^2/sum(w^2/n), contrast(treat,w)$ss)
(1)      2.9446      2.944
```

When the sample sizes are equal, this simplifies to

$$SS_w = n\,(\sum_{1\le i\le g} w_i\overline{y}_i)^2/\sum_{1\le i\le g} w_i^2$$

Sometimes related contrasts are grouped together in *families*. For example, the set of all pairwise contrasts $\alpha_i - \alpha_j$ is a family.

Another family of contrasts, which I illustrated a little of last lecture, are **polynomial contrasts**. These may be useful when the treatments are determined by levels of a quantitative variable x.

We looked at a **linear** contrast, whose coefficients are proportional to

$$w_i = x_i - \overline{x},\ \overline{x} = \sum_{1 \le i \le g} n_i x_i / N$$

Similarly there are contrasts that focus on the quadratic, cubic and higher order terms.

When doses are equally spaced, there are tables of the contrast coefficients (see, for example, Table D6, p. 630). But these don't help when the values predictor variable are not equally spaced.

However, in either case, you can get the polynomial contrast SS's (but not their values and standard errors) by fitting a polynomial trend in **anova()**.

First you need a vector of temperatures for each case. Start by entering the temperatures for each treatment group and then use **treat** as a subscript to get the full length N = 37 vector **temper** .

```
Cmd> temperature <- vector(175,194,213,231,250)

Cmd> temper <- temperature[treat]

Cmd> list(treat,temperature,temper)
temper          REAL   37
temperature     REAL   5
treat           REAL   37    1     FACTOR with 5 levels

Cmd> anova("logy=P4(temp)",fstat:T)
Model used is logy=P4(temp)
WARNING: summaries are sequential
              DF          SS          MS           F     P-value
CONSTANT       1      79.425      79.425  8653.95365           0
{temp}         1      3.4593      3.4593   376.91283  2.8767e-19
{(temp)^2}     1    0.078343    0.078343     8.53610   0.0063378
{(temp)^3}     1  1.8572e-05  1.8572e-05     0.00202      0.9644
{(temp)^4}     1  8.2568e-06  8.2568e-06     0.00090     0.97626
ERROR1        32     0.29369   0.0091779
```

The **SS** column has the contrast sums of squares and the **F** column has the corresponding t^2. There is no easy way to get the values or SE's of the contrasts.

In testing these, you always work backward starting from the highest order, and stop when you find the first significant polynomial contrast (**{(temp)^2}** here).

An important property of some sets or families of contrasts is **orthogonality**.

Definition

Two contrasts $\{w_i^{(1)}\}$ and $\{w_i^{(2)}\}$ are *orthogonal* when

$$\boxed{\sum_{1\le i\le g} w_i^{(1)} w_i^{(2)}/n_i = 0}$$

When the sample sizes are equal this simplifies to

$$\boxed{\sum_{1\le i\le g} w_i^{(1)} w_i^{(2)} = 0}$$

When the ANOVA assumptions, including normal errors, are satisfied, two orthogonal contrasts

$$w^{(1)}(\{\overline{y_{i\bullet}}\}) = \sum_{1\le i\le g} w_i^{(1)} \overline{y_{i\bullet}}$$
$$w^{(2)}(\{\overline{y_{i\bullet}}\}) = \sum_{1\le i\le g} w_i^{(2)} \overline{y_{i\bullet}}$$

are **independent**. With independent errors, but not normality, orthogonal contrasts are **uncorrelated**.

The polynomial contrasts are all mutually orthogonal.

When there are $g = 4$ treatments defined by all combinations of two factors A at two levels and B at two levels, the 4 means μ_{11}, μ_{12}, μ_{21} and μ_{22} can be arranged in a 2 by 2 table:

	B_1	B_2
A_1	μ_{11}	μ_{12}
A_2	μ_{21}	μ_{22}

An important family of contrasts for this case are defined by the weights

- $\{1/2, 1/2, -1/2, -1/2\}$ or $\{1, 1, -1, -1\}$
  Compares A_1 and A_2 ignoring B
- $\{1/2, -1/2, 1/2, -1/2\}$ or $\{1, -1, 1, -1\}$
  Compares B_1 and B_2 ignoring A
- $\{1, -1, -1, 1\}$
  Compares B_1-B_2 at A_1 with B_1-B_2 at A_2
  or A_1-A_2 at B_1 with A_1-A_2 at B_2

The first is an **A main effect contrast** because it compares the effects of A_1 and A_2 ignoring B.

The second is a **B main effect contrast** because it compares the effects of B_1 and B_2 ignoring A.

The third is an **A×B interaction contrast** which is used to see if the effect of A depends on the level of B (or the effect of B depends on the level of A).

When the 4 sample sizes are equal, these contrasts are orthogonal.

```
Cmd> w_a <- vector(1,1,-1,-1)/2

Cmd> w_b <- vector(1,-1,1,-1)/2

Cmd> w_ab <- vector(1,-1,-1,1)

Cmd> vector(sum(w_a*w_b),sum(w_a*w_ab),sum(w_b*w_ab))
(1)          0          0          0
```

All sums of products are 0.