Exploratory Data Analysis

Tabular Data Analysis and Smoothing

David B King, Ph.D.

October 11, 2015

What's Next?

1. Review: Two-way analysis
2. Example of Median Polish: Smoke Data
3. Non-additive Fits for two-way tables (EDTTS Chapter 3)
4. Three-way analysis (EDTTS Chapter 4)
5. Example of Multiple Carriers

Median Polish For Additive Model

$$y_{ij} = \widehat{y_{ij}} + e_{ij} = g(i,j) + e_{ij}$$

- Simple Additive Model: *main-effects model* (simple interpretation)

$$\widehat{y_{ij}} = m + a_i + b_j$$

- Mean: minimize $\sum_{ij} e_{ij}^2$; no iteration; affected by outliers
- Median Polish: $\sum_{ij} |e_{ij}|$; iterative fit; resistant to outliers
  – Fit is not exactly, but is often very close, to the one that minimizes the sum of the absolute residuals.
- Residuals: tell us how much more variance left after the fit $m + a_i + b_j$
  – Characteristic pattern: Negative/positive in opposite corners, nearly 0 in center row/column
  – Diagnostic plot (transformation, interaction term)
- Forget-it Plot

Forget-It Plot

Why do we rotate 45^o in the forget it plot?

- Answer: Under the simple additive model:

$$\widehat{y_{ij}} = m + a_i + b_j$$

- Equation is like the equation for the plane:

$$z(x, y) = m + x + y$$

- The lines of constant altitude $Z(x, y) = z_0$ are perpendicular to the gradient

$$\overrightarrow{\nabla z} = \frac{\partial z}{\partial x}\hat{i} + \frac{\partial z}{\partial y}\hat{j} = 1\hat{i} + 1\hat{j}$$

Quantifying the Goodness of Median Polish

We know that non-parametric techniques that are based upon ordered statistics or letter values can often times not be competitive compared with approaches such as ANOVA in terms of power. But what they lose in power, they gain in robustness and resistance.

How do we measure resistance?

- Answer: The **Worst Case Breakdown Bound (WCBB)**, which is the fraction of total data points we can arbitrarily change (send to infinity) without affecting the statistic we are interested in.
- Example: What is the the WCBB for the median $\tilde{x}$ when considering the ordered observations?

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

Worst Case Breakdown Bound (WCBB)

For the median $\tilde{x}$ the **worst case scenario** would be sending the right half of the ordered observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ to ∞ or the left half of the observations to $-\infty$, and this change alone would change the value of the median $\tilde{x}$.
Hence the WCBB for the median $\tilde{x}$ is

$$WCBB(\tilde{x}) = \begin{cases} \frac{k}{2k+1} & \text{if } n = 2k+1 \text{ (odd) or,} \\ \frac{k}{2k+2} & \text{if } n = 2k+2 \text{ (even)} \end{cases}$$

If we define the oddness function

$$d(n) = \begin{cases} 0 & \text{if n odd and,} \\ 1 & \text{if if n even} \end{cases}$$

then

$$WCBB(\tilde{x}) = \frac{1}{2} - \frac{2 - d(n)}{2n}.$$

Well Placed Breakdown Bound (WPBB)

The formula for the WCBB comes from thinking about all the ways that we can mess with the data in the worst possible manner in order to affect our statistic in the worst possible way. WCBB takes a very pessimistic view!
The **Well Placed Breakdown Bound (WPBB)** takes an optimistic view. For example, the WPBB measure for the median $\tilde{x}$ comes from thinking optimistically about all the ways we could alter our ordered data

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

without having it affect out median? I.e. what is the maximum fraction of points that we could send off to ∞ on the right hand side of the median and $-\infty$ on the left hand side of the median without changing the value of the median?

Well Placed Breakdown Bound (WPBB)

The formula for the WCBB comes from thinking about all the ways that we can mess with the data in the worst possible manner in order to affect our statistic in the worst possible way. WCBB takes a very pessimistic view!
The **Well Placed Breakdown Bound (WPBB)** takes an optimistic view. For example, the WPBB measure for the median $\tilde{x}$ comes from thinking optimistically about all the ways we could alter our ordered data

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

without having it affect out median? I.e. what is the maximum fraction of points that we could send off to ∞ on the right hand side of the median and $-\infty$ on the left hand side of the median without changing the value of the median?

$$WPBB(\tilde{x}) = \begin{cases} \frac{2k}{2k+1} & \text{if } n = 2k+1 \text{ (odd) or,} \\ \frac{2k}{2k+2} & \text{if } n = 2k+2 \text{ (even)} \end{cases} = 1 - \frac{2 - d(n)}{n}.$$

Thus $WPBB(\tilde{x}) = 2WCBB(\tilde{x})$.

WCBB for Median Polish

The formula for the WCBB when we fit tabular data to the additive model

$$y_{ij} = m + a_i + b_j + \epsilon_{ij}$$

using median polish?

- Want to think of ways we can alter data in the table in the worst possible fashion without affecting the overall median m, the row effects a_i and the column effects b_j.

WCBB for Median Polish

The formula for the WCBB when we fit tabular data to the additive model

$$y_{ij} = m + a_i + b_j + \epsilon_{ij}$$

using median polish?

- Want to think of ways we can alter data in the table in the worst possible fashion without affecting the overall median m, the row effects a_i and the column effects b_j.
- Worse possible scenario would come if we put all bad values in the same row or column.

WCBB for Median Polish

The formula for the WCBB when we fit tabular data to the additive model

$$y_{ij} = m + a_i + b_j + \epsilon_{ij}$$

using median polish?

- Want to think of ways we can alter data in the table in the worst possible fashion without affecting the overall median m, the row effects a_i and the column effects b_j.
- Worse possible scenario would come if we put all bad values in the same row or column.
- The minimum number of bad values we could put in the table should be a function of $\min(R, C)$ where $R = \#$ of rows and $C = \#$ of columns.

WCBB for Median Polish

The formula for the WCBB when we fit tabular data to the additive model

$$y_{ij} = m + a_i + b_j + \epsilon_{ij}$$

using median polish?

- Want to think of ways we can alter data in the table in the worst possible fashion without affecting the overall median m, the row effects a_i and the column effects b_j.
- Worse possible scenario would come if we put all bad values in the same row or column.
- The minimum number of bad values we could put in the table should be a function of $\min(R, C)$ where $R = \#$ of rows and $C = \#$ of columns.
- We could replace up to roughly $\min(R, C)/2$ of the data in either the rows or columns with ∞ without affecting m, a_i or b_j hence

$$WCBB(\hat{m}, \hat{a}_i, \hat{b}_j) = \frac{\min(R, C) - 2 + d(\min(R, C))}{2RC}.$$

WPBB for Median Polish

- Want to think of the largest number of bad data values B we could put in our table in the best possible fashion without affecting the overall median m, the row effects a_i and the column effects b_j.

WPBB for Median Polish

- Want to think of the largest number of bad data values B we could put in our table in the best possible fashion without affecting the overall median m, the row effects a_i and the column effects b_j.
- Best possible scenario would come if we strategically distribute at most

$$B = \frac{1}{2}\min\{R(C-2+d(C)), C(R-2+d(R))\}$$

evenly throughout the table so that each row and column receives an "equal amount of badness."

WPBB for Median Polish

- Want to think of the largest number of bad data values B we could put in our table in the best possible fashion without affecting the overall median m, the row effects a_i and the column effects b_j.
- Best possible scenario would come if we strategically distribute at most

$$B = \frac{1}{2}\min\{R(C-2+d(C)), C(R-2+d(R))\}$$

evenly throughout the table so that each row and column receives an "equal amount of badness."

- The Best Case or Well Placed Breakdown Bound is therefore

$$WPBB = \frac{B}{RC} = \frac{\min\{R(C-2+d(C)), C(R-2+d(R))\}}{2RC}.$$

Diagnostic Plot to test for Nonadditivity

After fitting the table to the additive model

$$y_{ij} = m + a_i + b_j + \epsilon_{ij}$$

we want to produce a plot which shows us if we have an "interaction effect" between the rows and columns.
The diagnostic plot is a plot of the ordered pairs $(\hat{\text{cv}}_{ij}, \hat{\epsilon}_{ij})$ where

$$\hat{\text{cv}}_{ij} = \text{comparison value} = \frac{\hat{a}_i \hat{b}_j}{\hat{m}}$$

Diagnostic Plot R Code

```
diagplot = function(tab){
# tab is the data in the form of a $R$ by $C$ table
require(car)
fit=medpolish(tab)
a = fit$row
b = fit$column
m = fit$overall
res = c(fit$residuals)
CV = c((a %o% b)/m)
plot(CV,res,xlab="Comparison value",ylab="Residual",main="Tukey Additivity Plot
abline(v=0,h=0,lty=2)
abline(0,-1,col="red")
F=rlm(res~ CV)
abline(F,col="blue")
plot(temp.comp, temp.MP$res,xlab="Comparison value",ylab="Residual",cex=0.5)
return(F)
}
```

Non-Additive Fits for Two Way Tables

Chapter 3 of EDTTS

Non-Additive Fits for Two Way Tables

One Step beyond an Additive Fit: "extended fit" or "fit with ODOFNA" (Tukey1949):

$$\text{Model: } y_{ij} = \mu + \alpha_i + \beta_j + \kappa\alpha_i\beta_j + \epsilon_{ij}$$

$$\text{Fit: } y_{ij} = m + a_i + b_j + ka_ib_j + e_{ij}$$

One extra degree of freedom for estimating κ, facilitated through the diagnostic plot:

plot residuals e_{ij} versus comparison values $\mathrm{cv}_{ij} = a_ib_j/m$.

If the slope $s = (1 - p)$, then either:

(a) Transform y to z via $z = y^p$ ($p = 0$ log transformation) and fit simple additive model to z

or

(b) Estimate k by s/m and use $y_{ij} = m + a_i + b_j + ka_ib_j$.

Non-Additive Fits for Two Way Tables

Procedure:

- Fit the data to the additive model

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

- Construct diagnostic plots for residuals
- If you see patterns in your residuals after fitting the data to simple additive model you really have three options:

1. Transform y for additivity via $z = y^p$ and fit simple additive model to z
2. Estimate $k = s/m$ and use $y_{ij} = m + a_i + b_j + ka_ib_j$.
3. If you see a cris-cross pattern in residuals you can take log of data and fit the log of data to additive model.

Why does log reveal interaction effect?

Suppose that data fit this model perfectly:

$$y_{ij} = m + a_i + b_j + (a_i b_j)/m$$

Hence

$$y_{ij} = m(1 + a_i/m)(1 + b_j/m)$$

and so

$$\log(y_{ij}) = \log(m) + A_i + B_i$$

with $A_i = \log(1 + a_i/m)$ and $B_i = \log(1 + b_j/m)$.
We're back to additive model!!

Diagnosing Non-additivity in residuals

But the log transform only works well if $\epsilon_{ij} = \kappa(a_i b_j)$.
Why does the

$$k = \frac{s}{m}$$

rule work if your diagnostic plots show a slope s?

Diagnosing Non-additivity in residuals

Suppose you construct the diagnostic plot of

$$\epsilon_{ij} \text{ vs } \mathrm{cv}_{ij}$$

and the resistant line drawn through residuals has slope s.
Since the resistant like has slope s this implies that residuals have internal structure of the form

$$\epsilon_{ij} = \kappa a_i b_j + r_{ij} \Longrightarrow \hat{\epsilon}_{ij} \approx \kappa m \left(\frac{a_i b_j}{m} \right) + \hat{r}_{ij}$$

and so the fitted line would have slope

$$s = \kappa m \Longrightarrow \hat{\kappa} = \frac{s}{m}$$

Example: Smoking data

Smoking prevalence data $\times 10$:

- Column (Year): 1974, 1979, 1983, 1985, 1987, 1988, 1990-1992
- Row: corresponds to level of education

```
White Males
< 12   516 480 479 452 453 448 417 417 414
 12    422 386 371 348 346 342 330 324 329
13-15  414 364 326 323 280 282 254 260 259
>=16   281 228 211 192 174 171 145 147 150

Afri-Amer Males
< 12   583 501 460 511 494 453 414 478 445
 12    512 484 472 419 436 483 374 396 387
13-15  457 393 447 423 324 348 283 327 270
>=16   418 379 313 320 209 215 206 183 269
```

Example: Smoking data

White Females

< 12	370	361	355	371	370	352	336	337	331
12	321	299	309	294	294	293	268	275	295
13-15	305	306	280	271	262	238	214	223	236
>=16	258	219	189	168	164	151	137	133	142

African-American Females

< 12	364	319	369	392	350	339	268	333	332
12	419	330	352	323	281	301	240	260	259
13-15	332	288	265	237	272	268	231	248	270
>=16	352	434	287	275	195	222	169	144	258

Example: Smoking data

a. Computed 16 centercepts and 16 slopes for the 16 smoking trends (4 gender-race groups x 4 education levels).
b. Place the 16 centercepts into a 4x4 table and median polish them.
c. Place the 16 slopes into a 4x4 table and median polish them.
d. Create the diagnostic plot for each table.
e. Stem-and-leaf the residuals for each table.
f. Create the "forget-it" plots.
g. What are your interpretations of the median polish results?

One Step beyond an Additive Fit

Example (EDTTS page 73): specific volume of peroxide-cured rubber at four temperatures and six pressures.

TABLE 3-1. Specific volume (in cubic centimeters per gram) of peroxide-cured rubber at four temperatures (in degrees Celsius) and six pressures (in kilograms per square centimeter above atmospheric pressure).

	Pressure					
Temperature	500	400	300	200	100	0
0	1.0637	1.0678	1.0719	1.0763	1.0807	1.0857
10	1.0697	1.0739	1.0782	1.0828	1.0876	1.0927
20	1.0756	1.0801	1.0846	1.0894	1.0944	1.0998
25	1.0786	1.0830	1.0877	1.0926	1.0977	1.1032

One Step beyond an Additive Fit

Example (EDTTS page 73): specific volume of peroxide-cured rubber at four temperatures and six pressures

- Note that "+ - / - +" pattern in the table of residuals

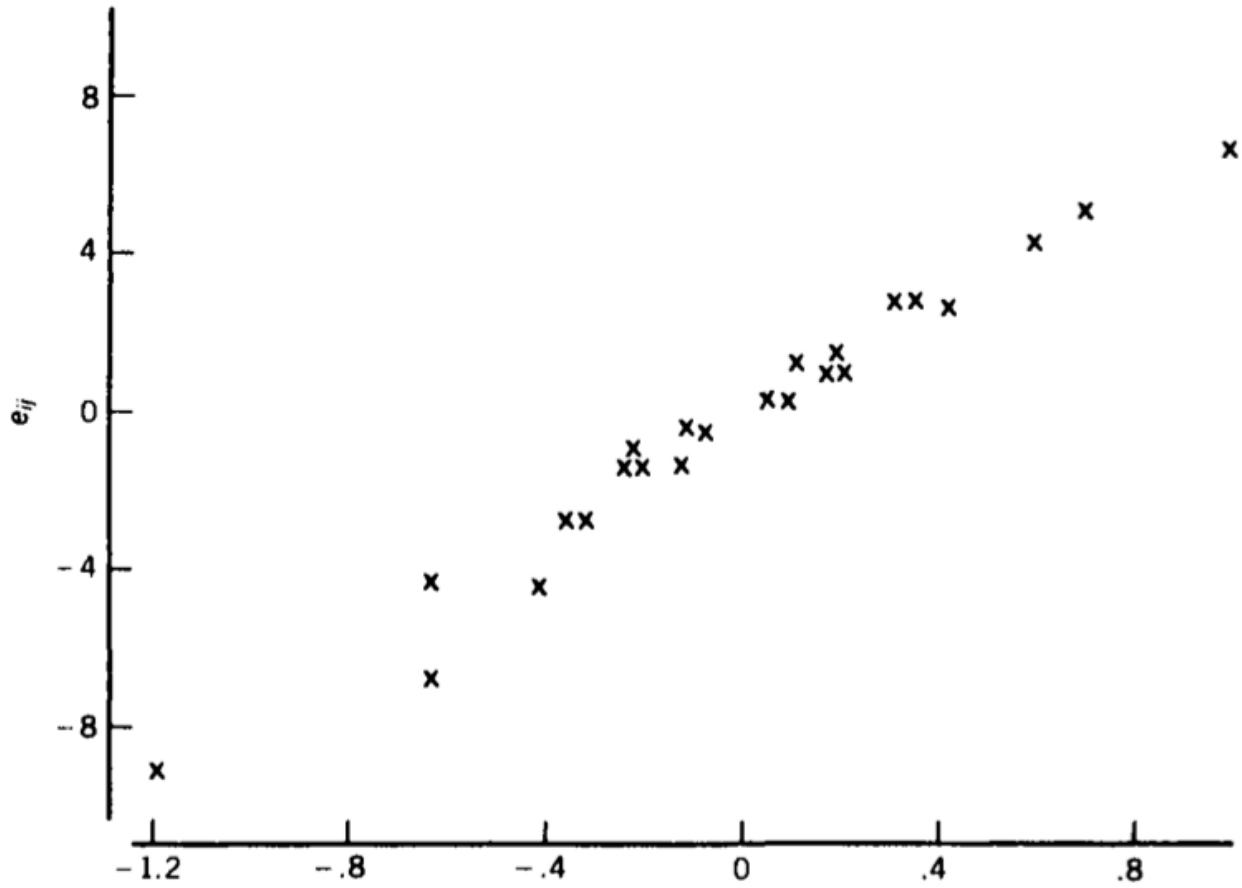
TABLE 3-2. Simple additive fit by median polish for the rubber data (unit: 10^{-4} cm^3 / g).

	Pressure						
Temperature	500	400	300	200	100	0	a_i
0	7.0	4.5	1.5	−1.5	−6.5	−9.0	−96.5
10	3.0	1.5	0.5	−0.5	−1.5	−3.0	−32.5
20	−3.0	−1.5	−0.5	0.5	1.5	3.0	32.5
25	−4.5	−4.0	−1.0	1.0	3.0	5.5	64.0
b_j	−111.0	−67.5	−23.5	23.5	72.5	125.0	10837.5 = m

One Step beyond an Additive Fit

Example (EDTTS page 73): specific volume of peroxide-cured rubber at four temperatures and six pressures

- Estimated slope in diagnostic plot = 7.81



One Step beyond an Additive Fit

Example (EDTTS page 73): specific volume of peroxide-cured rubber at four temperatures and six pressures

- Method (a), $z = y^{-6.81}$, doesn't make a lot of sense
- Method (b): $\widehat{y_{ij}} = 1.08375 + a_i + b_j + 7.21 a_i b_j$

TABLE 3-3. Extended fit for rubber data (unit: 10^{-4} cm^3 / g).

	Values for ka_ib_j						
Temperature	500	400	300	200	100	0	a_i
0	7.7	4.7	1.6	−1.6	−5.0	−8.7	−96.5
10	2.6	1.6	0.6	−0.6	−1.7	−2.9	−32.5
20	−2.6	−1.6	−0.6	0.6	1.7	2.9	32.5
25	−5.1	−3.1	−1.1	1.1	3.3	5.8	64.0
b_j	−111.0	−67.5	−23.5	23.5	72.5	125.0	10837.5 = m
	Residuals (r_{ij})						
0	−0.73	−0.20	−0.14	0.14	−1.45	−0.30	
10	0.40	−0.08	−0.05	0.05	0.20	−0.07	
20	−0.40	0.08	0.05	−0.05	−0.20	0.07	
25	0.63	−0.88	0.08	−0.08	−0.35	−0.27	

Internal vs External Structure

- Internal: what patterns/trends lie in the residuals
- External: how are main effects ($\{a_i\}$, $\{b_j\}$) related to levels of the factors (e.g., temperature, pressure)?
- Specific volume of rubber example:

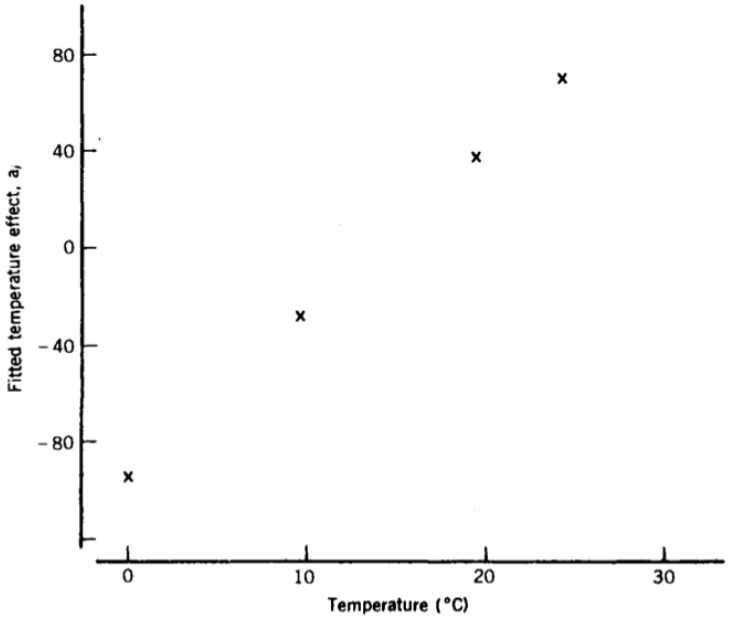


Figure 3-2. Row effects versus temperature in the rubber data. Least-squares regression line: $a(t) = .000643t - .00966$. [Units of a_i: 10^{-4} cm^3/g. Note that $a(t)$ is in cm^3/g, the units of the original data.]

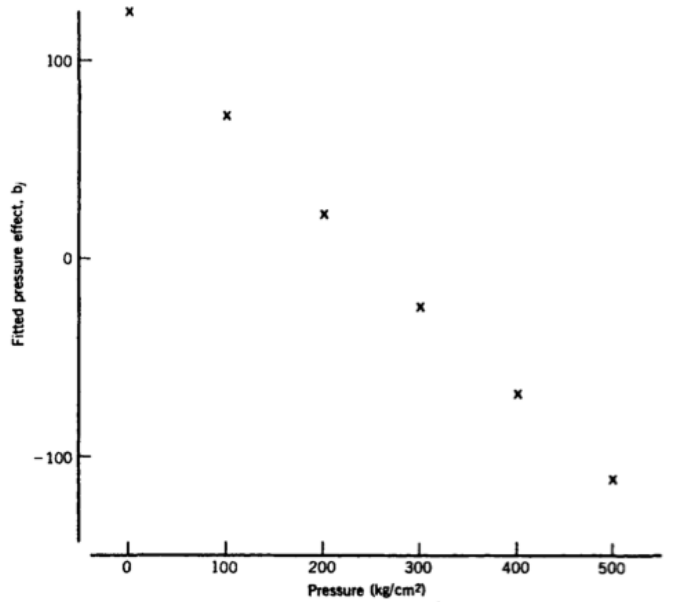


Figure 3-3. Column effects versus pressure in the rubber data. Least-squares regression line: $b(p) = .01208 - .000047p$. [Units of b_j: 10^{-4} cm^3/g. Note that $b(p)$ is in cm^3/g, the units of the original data.]

Internal vs External Structure

- Specific volume of rubber example:

b_j versus $p_j \Rightarrow b_j \downarrow$ as
$p_j \uparrow, b(p) \approx (12.08 - 0.047p) \times 10^{-3} cm^3/g.$
a_i versus $t_i \Rightarrow a_i \uparrow$ as
$t_i \uparrow, a(t) \approx (-9.66 + 0.643t) \times 10^{-3} cm^3/g.$

$$\hat{y}(t,p) = 1.08375 + a(t) + b(p) + 7.21a(t)b(p)$$
$$= 10^{-4}(10853 + 6.9928t - 0.43778p - 0.0021834tp)$$

TABLE 3-4. Fitted values from equation (9) and residuals for rubber data, based on temperature and pressure (unit: 10^{-4} cm^3/g).

	Pressure					
Temperature	500	400	300	200	100	0
	Predicted values					
0	10634	10678	10722	10766	10810	10853
10	10693	10739	10785	10831	10877	10923
20	10752	10801	10849	10897	10945	10993
25	10782	10831	10880	10930	10979	11028

Internal vs External Structure

- Residuals not bad except they tend to be positive in first and last columns and negative in the middle columns, suggesting that the relationship for columns, $b(p)$, should be quadratic, not just linear, in p.
- Fit is better for predicting y at other combinations of $0 \leq t \leq 25$ and $0 \leq p \leq 500$ (beware of going outside the experimental region).
- Also it requires fitting only 4 constants (6, using quadratic function of p), whereas the extended fit requires $4 + 6 = 10$ degrees of freedom.
- Might fit better constants using linear regression, now that the exploratory analysis has uncovered this relationship between y and t, p.

Assessing and comparing fits

- Graphical displays of residuals: Stem-and-leaf; boxplot
- Reduction in Total Absolute Variation:
  - Compare: Classical R^2:

$$R^2 = 1 - \frac{\sum_{ij} e_{ij}^2}{\sum_{ij}(y_{ij} - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

    Multiplied by 100, this gives the percent variance explained by the fit, in squared units of y_{ij}.
  - Better for exploratory purposes: percent reduction in total absolute variation:

$$P = 1 - \frac{\sum_{ij} |e_{ij}|}{\sum_{ij} |y_{ij} - y_M|}, \quad y_M = \text{median}_{ij}\{y_{ij}\}$$

    Advantages: more resistant to outliers; units are the same as those for the data y_{ij}.
- Example: specific volume of rubber:
  – $P = 96.5\%$ for median polish fit;
  – $P = 99.7\%$ for extended fit (i.e., extended fit picks up 3.2% of the remaining unexplained 3.5% variation)

An aid to choosing fits

More parameters $\Rightarrow$ better fit!

- Classical comparison: Consider [SS(reduced) - SS(full)]/SS(full) where reduced refers to the model with fewer parameters and full refers to the full model with all the parameters; compare this statistic to an F-distribution with $r - f$ and f degrees of freedom, where r and f are the corresponding degrees of freedom in the reduced and full models, respectively.
- Alternatively, consider MSE/df = mean squared error per degree of freedom; the smaller, the better.
- By analogy, use $(1 - P)/df$.
- Ex:
  – for simple median polish, (100-96.5)/15 = 0.230% per df;
  – for extended fit, (100-99.7)/14 = 0.021% per df (major improvement).

Other Non-additive Fits

- Multiplicative Fits:

$$\widehat{y_{ij}} = q + hc_i d_j$$

- Additive+multiplicative fits

General model: $y_{ij} = \mu + \alpha_i \beta_j + \kappa \gamma_i \delta_j + \epsilon_{ij}$

Note that if $\kappa = 0$, then we have the simple additive model; if $\gamma_i = \alpha_i$ and $\delta_j = \beta_j$, then we have the extended fit.

Three Way Analyses

Chapter 4 of EDTTS

Structure of three-way table

Suppose factor A has I levels, factor B has J levels, and factor C has K levels. The possible combinations of one level from each factor form $I \times J \times K$ cells that contain the observations, often one value per cell.

TABLE 4-1. Two of the six possible two-way arrangements of a three-way table.

a. Factor A alone and factors B and C combined, with one block for each level of B

	B:	1			2			···	J		
A	C:	1	···	K	1	···	K	···	1	···	K
1		y_{111}	···	y_{11K}	y_{121}	···	y_{12K}	···	y_{1J1}	···	y_{1JK}
⋮		⋮		⋮	⋮		⋮		⋮		⋮
I		y_{I11}	···	y_{I1K}	y_{I21}	···	y_{I2K}	···	y_{IJ1}	···	y_{IJK}

b. Factor C alone and factors A and B combined, with one block for each level of A

A	B	C: 1	2	···	K
1	1	y_{111}	y_{112}	···	y_{11K}
	⋮	⋮	⋮		⋮
	J	y_{1J1}	y_{1J2}	···	y_{1JK}
2	1	y_{211}	y_{212}	···	y_{21K}
	⋮	⋮	⋮		⋮
	J	y_{2J1}	y_{2J2}	···	y_{2JK}
⋮	⋮	⋮	⋮		⋮
I	1	y_{I11}	y_{I12}	···	y_{I1K}
	⋮	⋮	⋮		⋮
	J	y_{IJ1}	y_{IJ2}	···	y_{IJK}

Decompositions and Models for three-way analysis

General form:

$$y_{ijk} = f(i, j, k) + \epsilon_{ijk}$$

If the three factors exert their influences separately and additively, the model takes the simple form: *main-effects-only* or *simple additive* model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$$

The most general additive decomposition of this form for a three-way table is the *full-effects* model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijk}$$

Median Polish Analysis for the Main-effects-only Model

Two ways to do main-effects-only decomposition

- Direct extension of two-way median-polish to three-way table
- Do two-way median polish twice:

  First on a $ij \times k$ table and then on a $i \times j$ table

If we do two way median polish twice after first fit we have

$$y_{ijk} = m + a_{ij} + b_k + r_{ijk}$$

After second median polish we take $r'_{ij} = \mathrm{med}_k r_{ijk}$ and fit the additive model

$$\tilde{y}'_{ij} = a_{ij} + r'_{ij} = b_i + c_j + \epsilon_{ij}.$$

Tukey's Three Way Median Polish Procedure

At iteration 0 assign the values

$$a_i^{(0)} = b_j^{(0)} = a_k^{(0)} = 0 \text{ and,}$$
$$r_{ijk}^{(0)} = y_{ijk}$$

For iteration $p = 1, \ldots$ Tukey's procedure proceeds as follows:

$$a_i^{(p)} = \text{med}_{jk}\{r_{ijk}^{(p-1)}\}$$
$$b_j^{(p)} = \text{med}_{ik}\{r_{ijk}^{(p-1)} - a_i^{(p)}\}$$
$$c_k^{(p)} = \text{med}_{ij}\{r_{ijk}^{(p-1)} - a_i^{(p)} - b_j^{(p)}\}$$
$$r_{ijk}^{(p)} = r_{ijk}^{(p-1)} - a_i^{(p)} - b_j^{(p)} - c_k^{(p)}$$

Tukey's Three Way Median Polish Procedure

For iteration $p = 1, \ldots$ we then compute:

$$m_a^{(p)} = \text{med}_i\{a_i^{(p)}\}$$
$$m_b^{(p)} = \text{med}_j\{b_j^{(p)}\}$$
$$m_c^{(p)} = \text{med}_k\{c_k^{(p)}\}$$

and set fixed effect estimates equal to

$$\hat{\alpha}_i^{(p)} = a_i^{(p)} - m_a^{(p)}$$
$$\hat{\beta}_j^{(p)} = b_j^{(p)} - m_b^{(p)}$$
$$\hat{\gamma}_k^{(p)} = c_k^{(p)} - m_c^{(p)}$$
$$\hat{\mu}^{(p)} = m_a^{(p)} + m_b^{(p)} + m_c^{(p)}$$

After the p^{th} iteration the final model is

$$y_{ijk} = \hat{\mu}^{(p)} + \hat{\alpha}_i^{(p)} + \hat{\beta}_j^{(p)} + \hat{\gamma}_k^{(p)} + r_{ijk}^{(p)}.$$

Plot effects and residuals

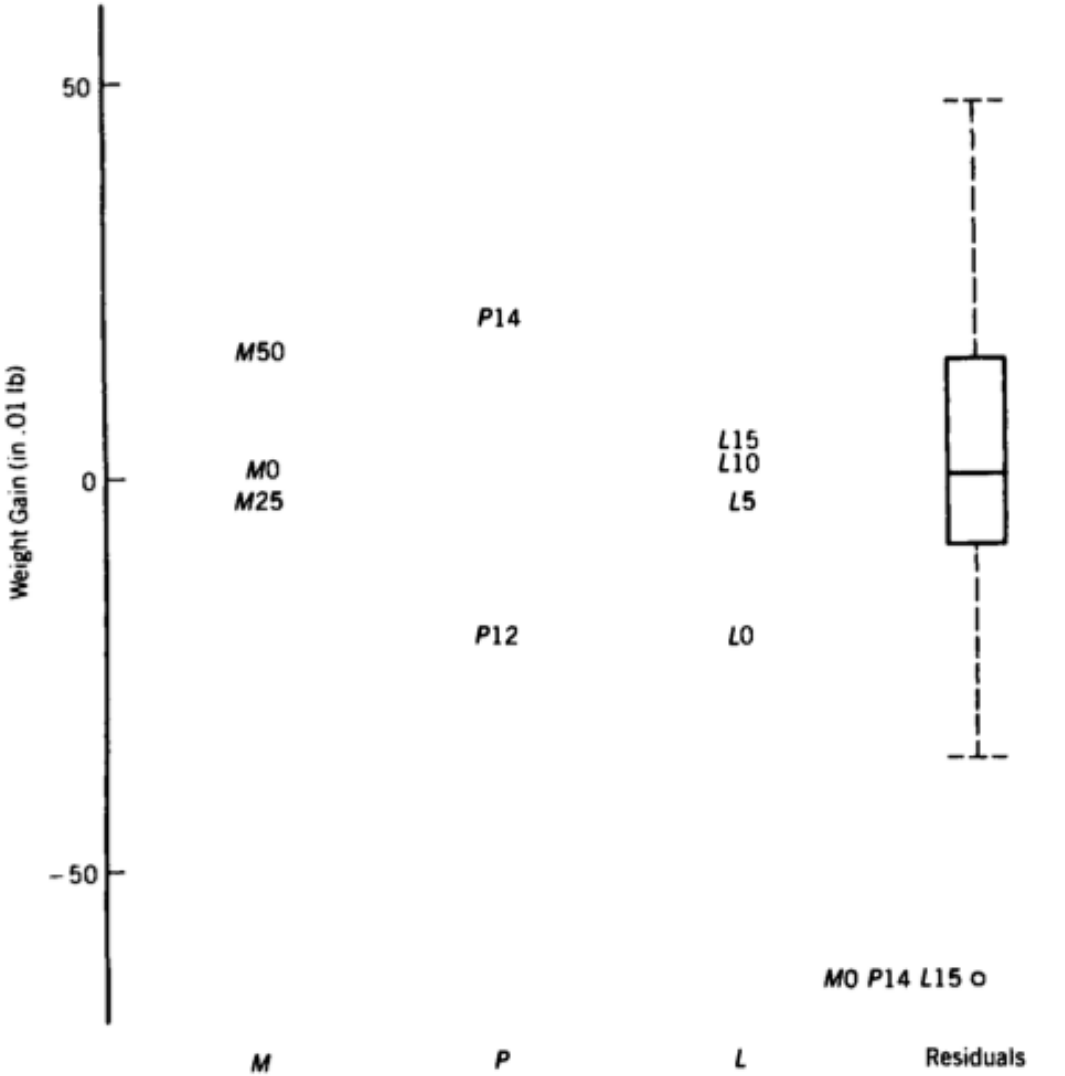


Figure 4-1. Dot plots of the effects and boxplot of the residuals from a main-effects-only analysis of the pig feeding data by median polish.

Nonadditivity and a diagnostic plot

A plot of the residuals

$$r_{ijk} = y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_k$$

against the comparison values

$$CV_{ijk} = \frac{\hat{\alpha}_i\hat{\beta}_j + \hat{\alpha}_i\hat{\gamma}_k + \hat{\beta}_j\hat{\gamma}_k}{\hat{\mu}}$$

might well approximately yield a line with slope $1 - p$, indicating that a transformation to the pth power could be useful in removing nonadditivity.

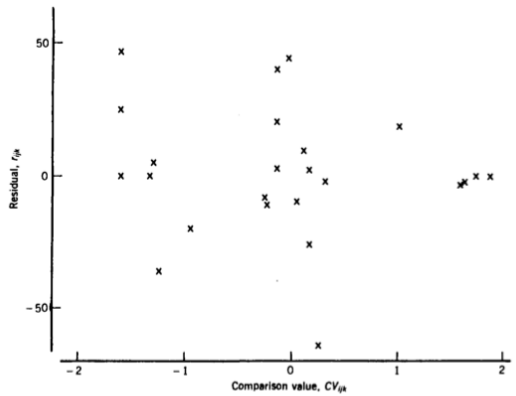


Figure 4-2. Diagnostic plot for main-effects-only analysis of the pig feeding data by median polish.

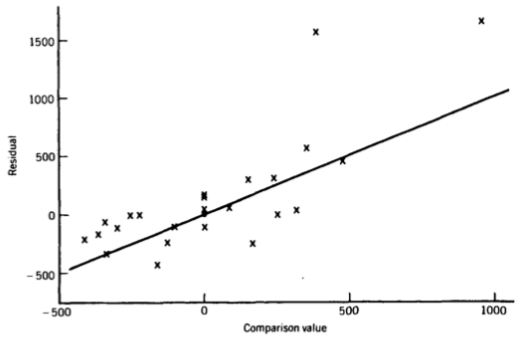


Figure 4-4. Diagnostic plot for main-effects-only analysis of the yarn breakage data. The dot at the origin represents three points. The line (through the origin) has slope 1.00 (calculated as the median of the slopes of the lines joining each point and the origin).

Constructed Example

In this constructed table each of the three factors has three levels, and the common value and main effects are as follows:

$$\mu = 20$$
$$(\alpha_1, \alpha_2, \alpha_3) = (-1, 0, 3)$$
$$(\beta_1, \beta_2, \beta_3) = (-1, 0, 1)$$
$$(\gamma_1, \gamma_2, \gamma_3) = (-1, 0, 2)$$

So

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k$$
$$y_{ijk} = x_{ijk}^2$$

Constructed Example

TABLE 4-25. Constructed data, perfectly additive in the square root scale.

		C		
A	B	1	2	3
1	1	289	324	400
	2	324	361	441
	3	361	400	484
2	1	324	361	441
	2	361	400	484
	3	400	441	529
3	1	441	484	576
	2	484	529	625
	3	529	576	676

TABLE 4-26. Main-effects analysis of the constructed data by median polish.

		C				
A	B	1	2	3	A-Effects	B-Effects
1	1	8	4	−4		−39
	2	4	2	−2	−41	0
	3	0	0	0		41
2	1	2	0	−4		−39
	2	0	0	0	0	0
	3	−2	0	4		41
3	1	−10	−6	2		−39
	2	−6	0	12	129	0
	3	−2	6	22		41
C-effects		−39	0	84	Overall	400

Constructed Example

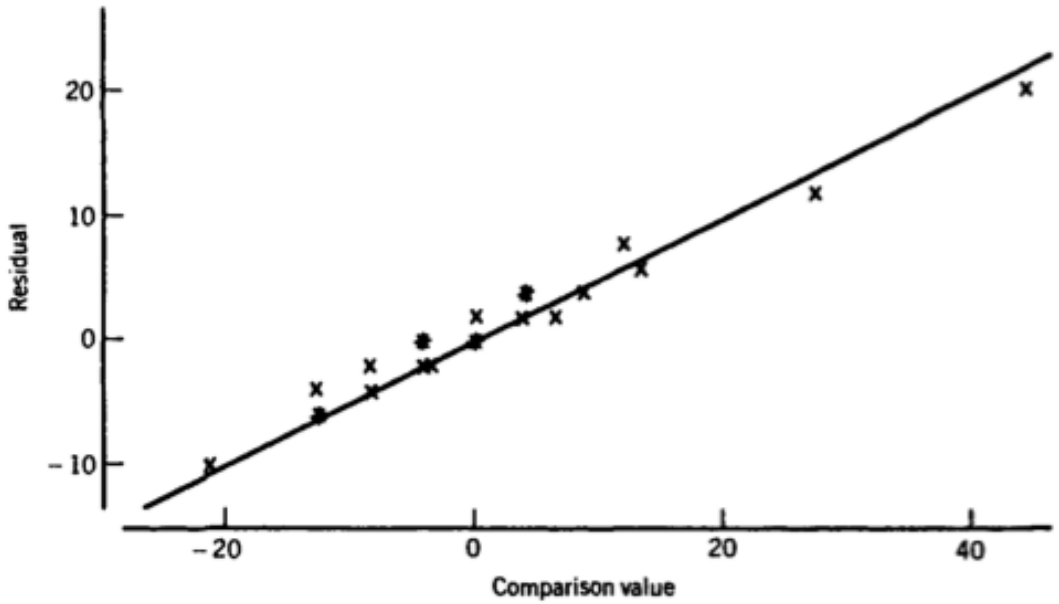


Figure 4-7. Diagnostic plot for the constructed data, perfectly additive in the square root scale. The reference line has slope 0.5.

More Topics in EDTTS Ch 4

- Median-polish analysis for the full-effects model (really complicated)
- If you want to go through lots of sticky accounting check out EDTTS section 4F WOW!!
- Analysis using means: Example with outliers
- Review section 4E for the analysis using means case.

Example: Hearing data

Hearing data. (Source: Probability sample in the U.S., 1965.) Cuthbert Daniel wrote an article in Technometrics (1978: 385-395) in which he analyzed the impact of a single outlier, or a set of outliers, on a classical two-way (rows and columns) fit (i.e., fitting by means). He illustrated his findings on the hearing prevalence rates: Percent of males aged 55-64 with hearing levels at least 16dB above audiometric 0, at 500, 1000, 2000, 3000, 4000, 6000 Hz (cycles per second) and normal speech, for 7 occupational groups: (1) professional; (2) farmers; (3) clerical-sales; (4) craftsmen; (5) operators; (6) service; (7) laborers.

hz	profl	farm	sales	crafts	oper	serv	labor
500	2.1	6.8	8.4	1.4	14.6	7.9	4.8
1000	1.7	8.1	8.4	1.4	12.0	3.7	4.5
2000	14.4	14.8	27.0	30.9	36.5	36.4	31.4
3000	57.4	62.4	37.4	63.3	65.5	65.6	59.8
4000	66.2	81.7	53.3	80.7	79.7	80.8	82.4
6000	75.2	94.0	74.3	87.9	93.3	87.8	80.5
norm	4.1	10.2	10.7	5.5	18.1	11.4	6.1

Example: Hearing data

Daniel studied the pattern of least squares residuals (i.e., after fitting the table by means) and identified observations in these cells that he suspected were "outliers": [3,2], [4,3], [5,3], [6,3], [3,1]. He also suggested that residuals that exceed 3 x RMS = 10.2 should be viewed with suspicion.

a. Conduct the "means" analysis.

b. Conduct a median polish on these data. Construct a "back-to-back" stem-and-leaf of the residuals from (a) above and those here, to compare their distributions. What do you observe? (You can also plot the 49 points, with x-axis as LS residual and y-axis as median polish residual.) Do Daniel's suspected outliers correspond to large residuals?

c. Construct the diagnostic plot. Does a transformation appear to be indicated? If so, transform.

d. Plot the fit (forget-it plot). Which factor, frequency or occupation, has the largest effect on hearing prevalence?

Smoothing

- Situation:
  (x, y) data pairs: $(x_1, y_1), \ldots, (x_n, y_n)$
- How does y vary with x in presence of
  - measurement error in y
  - outliers or "exotic" values
- Model: $y_i = f(x_i) + e_i$
  - $f(\cdot)$ reasonably smooth dependence on x
  - What is f?

Smoothing

- Enhances relationship between y and x
- Seeks to emphasize broad, long-term trends
- Reduces distractions due to measurement error, outliers
- Allows closer examination of underlying trends

Smoothing achieves these goals by emphasizing low frequencies (slowly varying trends) and reducing high frequencies (noise).

Smoother

- The ultimate/fundamental smoother: straight line fit (plotting residuals from line is like putting data under a microscope)
- Straight line is an example of *parametric* fit: estimate parameters (intercept and slope)
- Simple *nonparametric* fit:
  - Running means of length $(2k+1)$: ith smoothed value is

$$\tilde{y}_i = \frac{y_{i-k} + \cdots + y_{i-1} + y_i + y_{i+1} + \cdots + y_{i+k}}{2k+1}$$

  - Running medians of length $(2k+1)$: ith smoothed value is

$$\tilde{y}_i = \text{median}\{y_{i-k}, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_{i+k}\}$$

Running means/medians

Example:

x	y	3-means	5-means	$3R$
1	-2			
2	-1	-2		-2
3	-3	2	3.0	-1
4	10	6	5.4	10
5	11	11	8.0	11
6	12	11		11
7	10			

Note:

- Smoothing using $(2k+1)$-means/medians: lose k points at each end; need to define an "end value rule".
- Running means: Bigger k (wider span) $\rightarrow$ more smoothing $\rightarrow$ less noisy results $\rightarrow$ more bias (could miss important features)

Running means/medians

Example:

x	y	3-means	5-means	$3R$
1	-2			
2	-1	-2		-2
3	-3	2	3.0	-1
4	10	6	5.4	10
5	11	11	8.0	11
6	12	11		11
7	10			

Note:

- Running medians: more robust
  $3R$ = running medians of length 3: more faithful to sudden shifts and edges in data
- We expect to iterate as before: sometimes resmoothing the smooth, sometimes calculating residuals which are smoothed

Running means

Why wider span gives smoother, but more biased result?

Suppose:

1. $y_i = f(x_i) + e_i$
2. e_i are independent, mean 0 and variance σ^2
3. $f(x_i) \approx f(x_{i-1}) \approx f(x_{i+1})$

Then:

$$\tilde{y}_i = \frac{y_{i-1} + y_i + y_{i+1}}{3} \approx f(x_i) + \frac{e_{i-1} + e_i + e_{i+1}}{3}$$

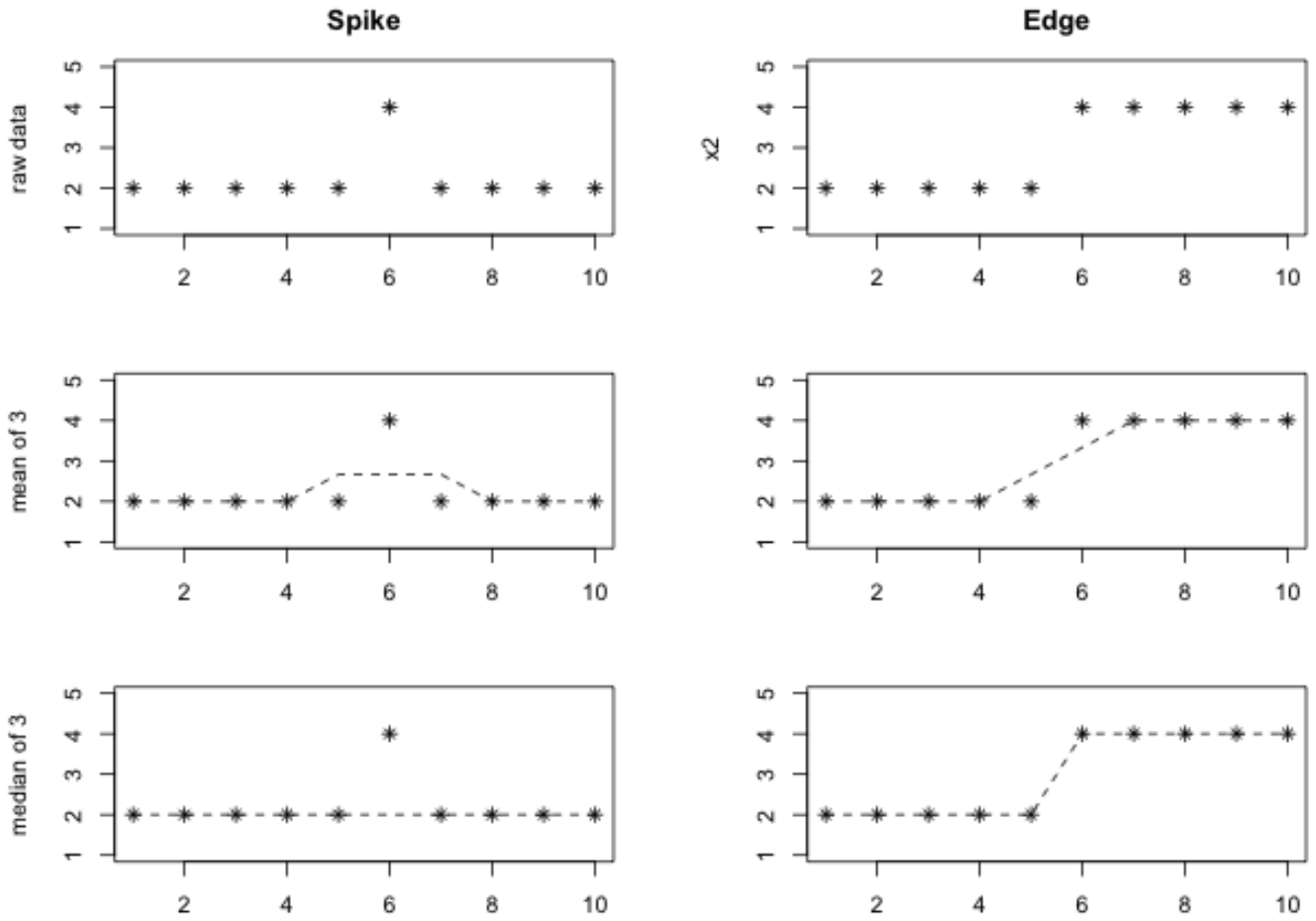
$$Var(\tilde{y}_i) \approx \frac{1}{3} Var(e_i) = \frac{\sigma^2}{3}$$

So

- smoothed value has 1/3 of the variance of original (with 5-means, 1/5 the variance).
- more bias: because (3) is not really quite true, and even less true with wider spans.

Running Means and Running Medians of spikes and edges

In particular, running means will squeeze peaks, raise valleys, smooth over abrupt features.



Other types of linear smoothers (can be written $\sum w_i y_i$)

Kernel Smoothers:

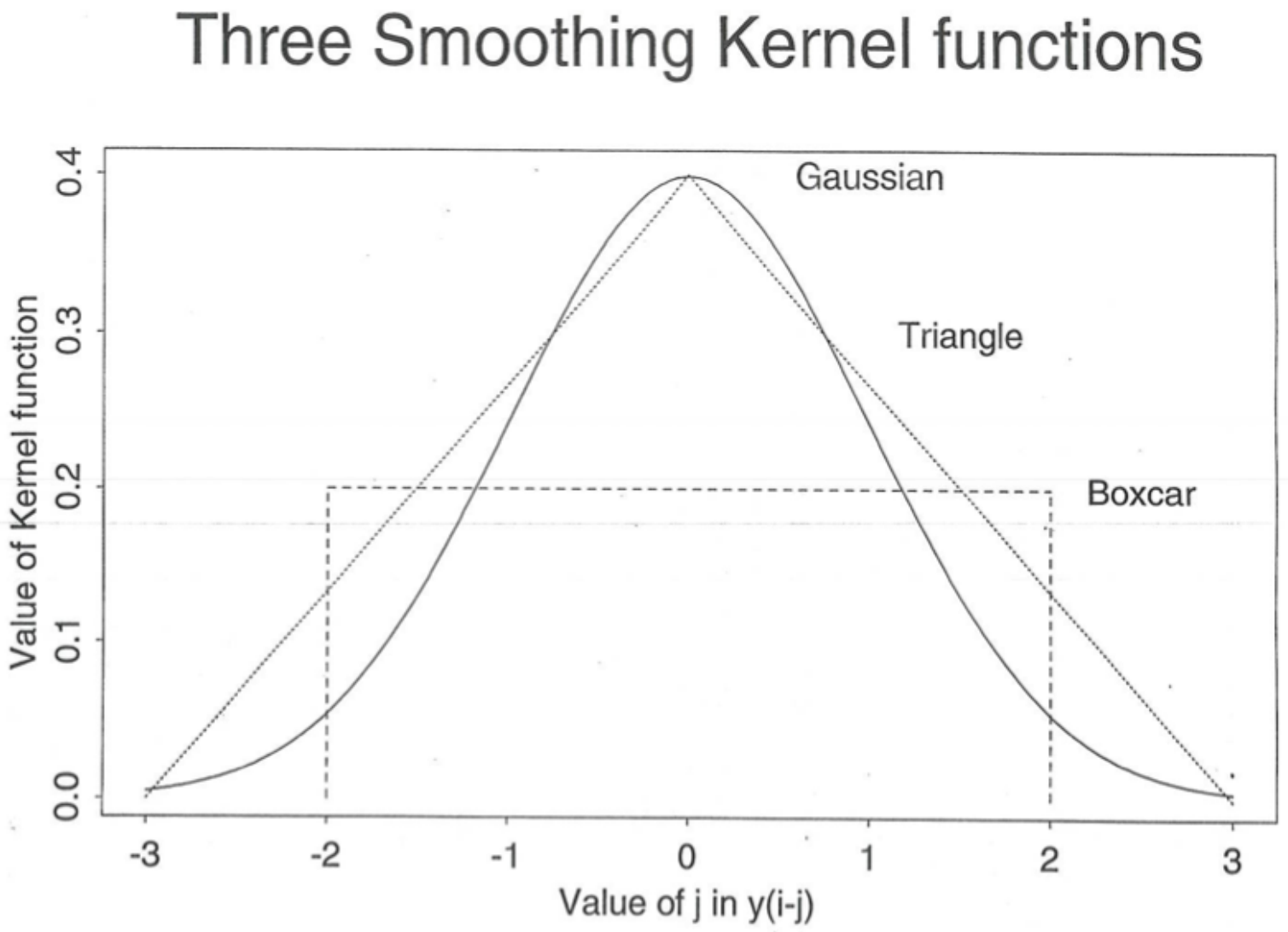
$$\tilde{y}_i = \sum_{j=-k}^{k} K\left(\frac{x_i - x_j}{h}\right) \cdot y_{i-j}$$

$h =$ bandwidth, $K =$ kernel function

- large $h \rightarrow$ blur, lose detailed structure
  small $h \rightarrow$ no smoothing!
- Some K functions: Gaussian density (bell curve); "boxcar" (running means); triangle
- In general, choice of h is more critical than choice of K (see Silverman 1986, Density estimation, §3.4)
- R function: ksmooth (Kernel regression smoother)

```
ksmooth(x, y, kernel = c("box","normal"), bandwidth=0.5,
range.x = range(x),
n.points = max(100, length(x)), x.points)
```

Other types of linear smoothers: Kernel Smoothers



Other types of linear smoothers (can be written $\sum w_i y_i$)

LOESS or LOWESS (locally weighted scatterplot smoothing): To calculate $\tilde{y}_i$, smoothed ith value

1. Choose f = fraction of n data points to be used for smoothing (wider $f \rightarrow$ bigger span)
2. Let X = set of all $\lfloor f \cdot n \rfloor$ x's closes to x_i (one of which is x_i itself). Note: if $x_i = x_1$ and x's are sorted, then $X = \{x_1, x_2, \ldots, x_{fn}\}$.
3. Fit a weighted regression line using the x's in X and their corresponding y's with weights

$$w_j = \max\left\{0, (1 - u_j^3)^3\right\}$$

$$u_j = \frac{\text{dist}(x_i, x_j)}{\max_{\{x_k \in X\}} \text{dist}(x_i, x_k)}$$

   u_j is the distance of x_j to the target x_i, normalized by the distance of x_i to the furthest x in the set X.

4. $\tilde{y}_i$ is the predicted value of y from the weighted regression at $x = x_i$.

LOWESS Smoother

Scatter plot smoothing

```
lowess(x, y = NULL, f = 2/3, iter = 3,
               delta = 0.01 * diff(range(x)))
```

This function performs the computations for the LOWESS smoother which uses locally-weighted polynomial regression.

Other types of linear smoothers (can be written $\sum w_i y_i$)

Friedman's SuperSmoother: Smooth the (x, y) values by Friedman's "super smoother".

```
supsmu(x, y, wt, span = "cv", periodic = FALSE, bass = 0)
```

supsmu is a running lines smoother which chooses between three spans for the lines. The running lines smoothers are symmetric, with $k/2$ data points each side of the predicted point, and values of k as $0.5n$, $0.2n$ and $0.05n$.

- wt: case weights, by default all equal
- span: the fraction of the observations in the span of the running lines smoother.
- periodic: if TRUE, the x values are assumed to be in [0, 1] and of period 1.
- bass: controls the smoothness of the fitted curve. Values of up to 10 indicate increasing smoothness.

Iterating a linear smoother

Iterating a linear smoother: resmoothing the smooth with another linear smoother

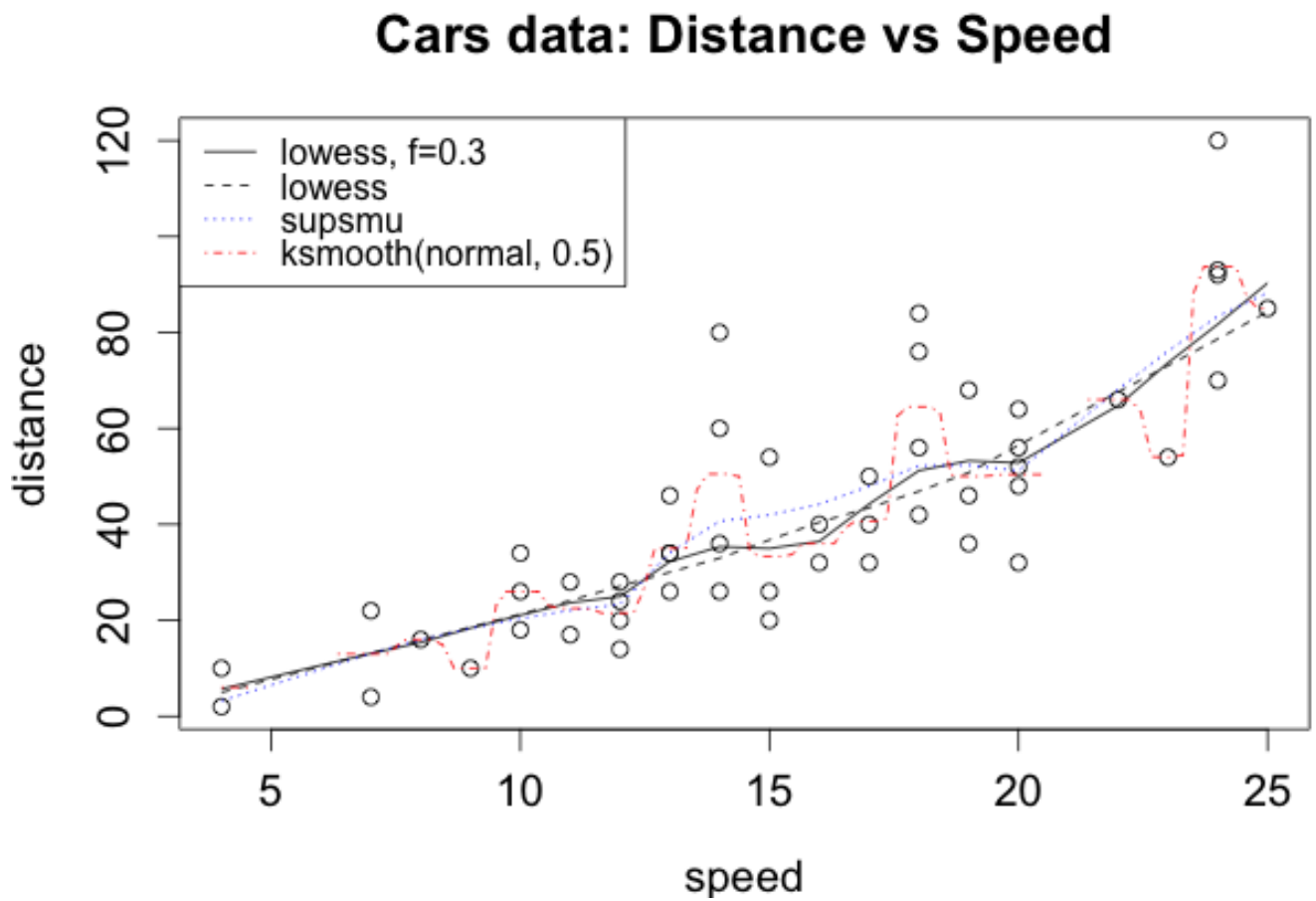
- Example:

$$\begin{aligned}\tilde{y}_4 &= \frac{y_3 + y_4 + y_5}{3} \\ \tilde{\tilde{y}}_4 &= \frac{\tilde{y}_3 + \tilde{y}_4 + \tilde{y}_5}{3} \\ &= \frac{y_2 + 2y_3 + 3y_4 + 2y_5 + y_6}{9}\end{aligned}$$

  is equivalent to a triangle kernel smoother.

- In general, repeating a linear smoother widens the span and sharpens the central peak.

Example


Cars data: Distance vs Speed
lowess, f=0.3
lowess
supsmu
ksmooth(normal, 0.5)
distance
0
20
40
60
80
120
5
10
15
20
25
speed

Linear smoother vs Nonlinear smoothers

Problems with linear smoothers:

- Smooth over sharp features
- Strongly affected by outliers

Nonlinear smoothers:

- cannot be expressed as $\sum w_i y_i$
- flexible (no linear constraints)
- usually involve medians instead of means
- catch depths of troughs, heights of peaks
- reduce influence of outliers
- easy to do by hand

Tukey's Running median Smoothing

smooth(x, kind = c("3RS3R", "3RSS", "3RSR", "3R", "3", "S"), twiceit = FALSE, endrule = "Tukey", do.ends = FALSE)

```
xx <- rnorm(20); xx
```

```
0.98 0.54 -0.75 0.34 0.88 0.48 0.08 -0.43 -0.57 0.68
1.56 -0.58 0.22 0.71 0.71 -0.15 1.29 0.64 1.17 1.85
```

```
#Smooth by 3:
xx3 <- smooth(xx, kind="3")
3 Tukey smoother resulting from smooth(x=xx, kind="3")
used 1 iterations
```

```
0.94 0.54 0.34 0.34 0.48 0.48 0.08 -0.43 -0.43 0.68
0.68 0.22 0.22 0.71 0.71 0.71 0.64 1.17 1.17 1.17
```

Tukey's Running median Smoothing

Some specific problems with $3R$:

- Plateaus (hence "splitting")

  An artifact of $3R$ is the presence of two adjacent smoothed y's with the same value. "Split" between them and apply end-value rule to each one so the values will differ.

Tukey's Running median Smoothing

Some specific problems with $3R$:

- "End value rules": Construct y_0, y_{n+1}, Tukey EDA, p221
  - "the change from the end smoothed value to the next-to-end smoothed value is between 0 and +2 times the change from the next-to-end smoothed-value to the next-to-end-but-one smoothed value."
  - "subject to this being true, the end smoothed-value is as close to the end input-value as possible."

"This means that we can look at two differences:

end input-value MINUS next-to-end smoothed value

and

next-to-end smoothed value MINUS next-but-one-to-end smoothed value

Tukey's Running median Smoothing

Some specific problems with $3R$:

- "End value rules": Construct y_0, y_{n+1}, Tukey EDA, p221
  -

  and if the first is between 0 and +2 times the second, we can copy on. Otherwise, we can make

  end smoothed value MINUS next-to-end smoothed value

  either zero or two times

  next-to-end smoothed value MINUS next-but-one-to-end smoothed value."

$$\tilde{y_1} = \text{median}\{y_1, \tilde{y_2}, y_0\}$$

where $y_0 = 3\tilde{y_2} - 2\tilde{y_3}$ is a linear extrapolation of y_3 and y_2 to x_0 as if x_0, x_1, x_2, x_3 are equally spaced.

Tukey's Running median Smoothing

Some specific problems with $3R$:

- "Twicing": smooth the residuals and add back to the original smooth

Tukey's Running median Smoothing

3RS3R, 3RSS, 3RSR (3RSSHT)

- 3 = smooth by medians of length 3
- R = repeat the previous smooth until no change
- S = split 2-plateaus
- H = hanning (1/4, 1/2, 1/4)
- T = twice (repeat on residuals)

smoothEnds(y, k=3): apply end-value rule to ends only

runmed: apply end-value rule also:

runmed(x, k, endrule = c("median", "keep", "constant"), algorithm = NULL, print.level = 0)

Extensions:

Extending to several variables: *Backfitting*

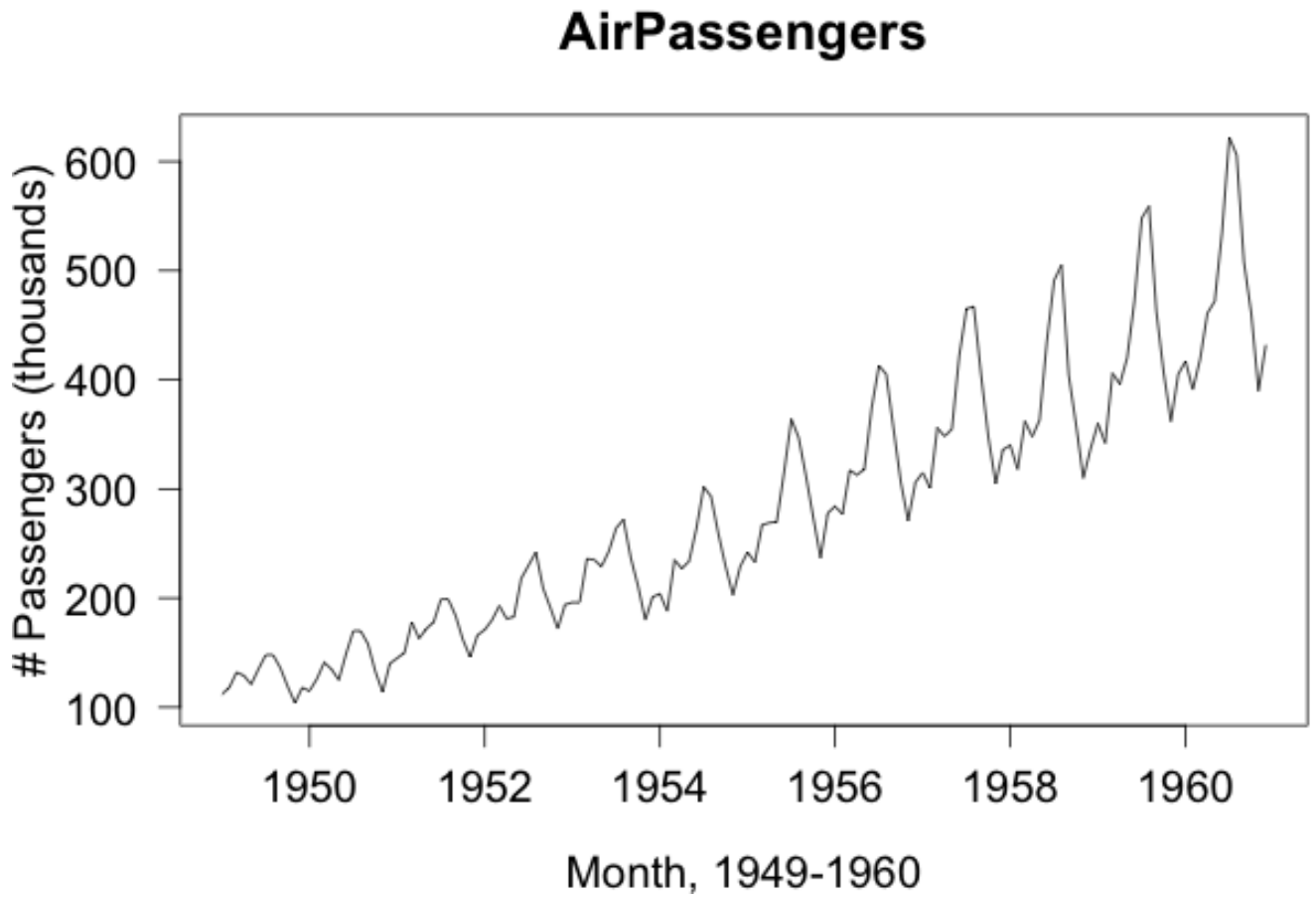
$$\tilde{y} = f(x_1) + f(x_2)$$

1. Initial smooth: fit y as a linear function of x_1, x_2
2. Iterate: smooth residuals as a function of x_i
3. Repeat step (2) for each x_i, in turn, until residuals are "flat"

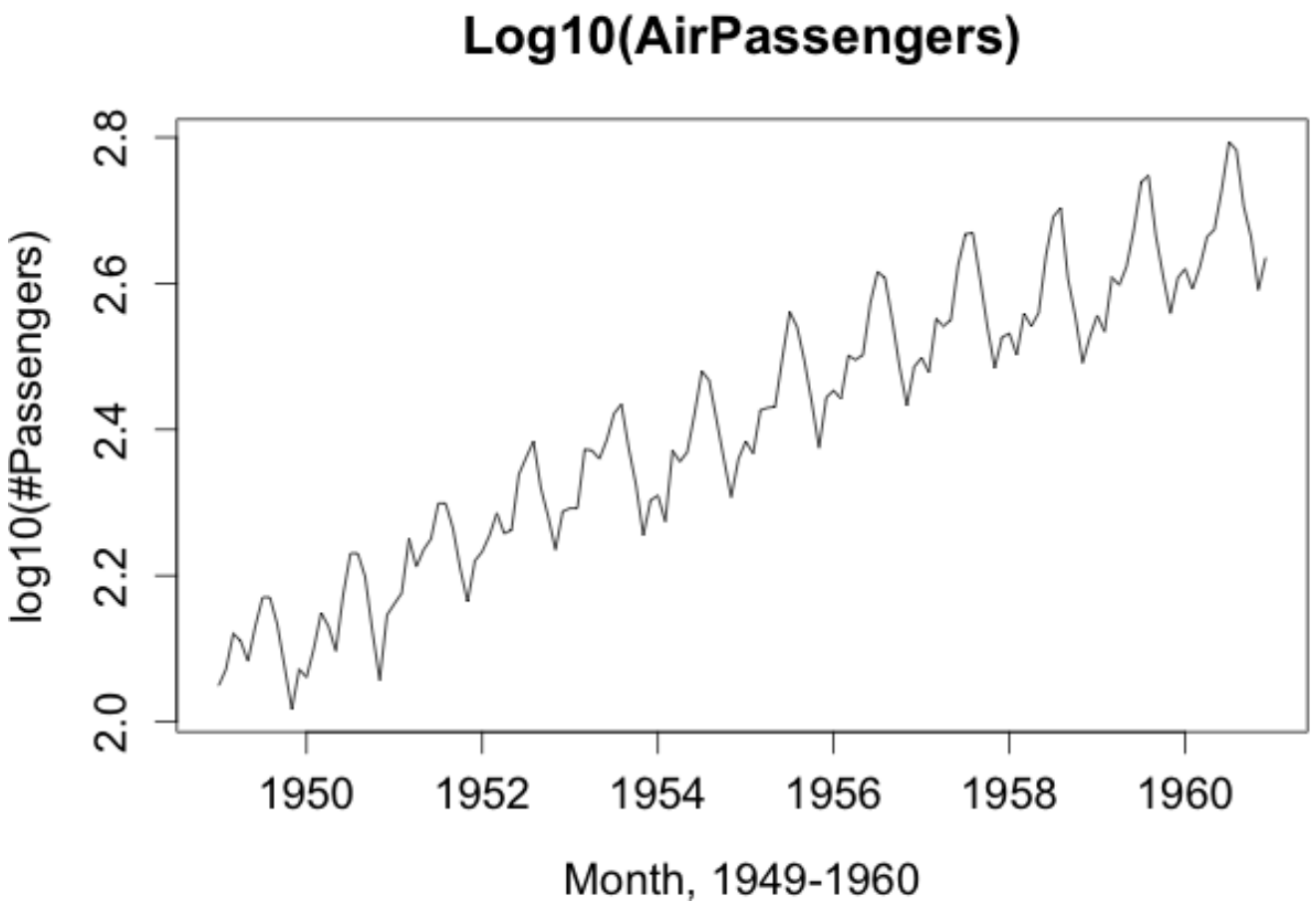
See Hastie and Tibshirani (1993), *Generalized Additive Models*.

Example

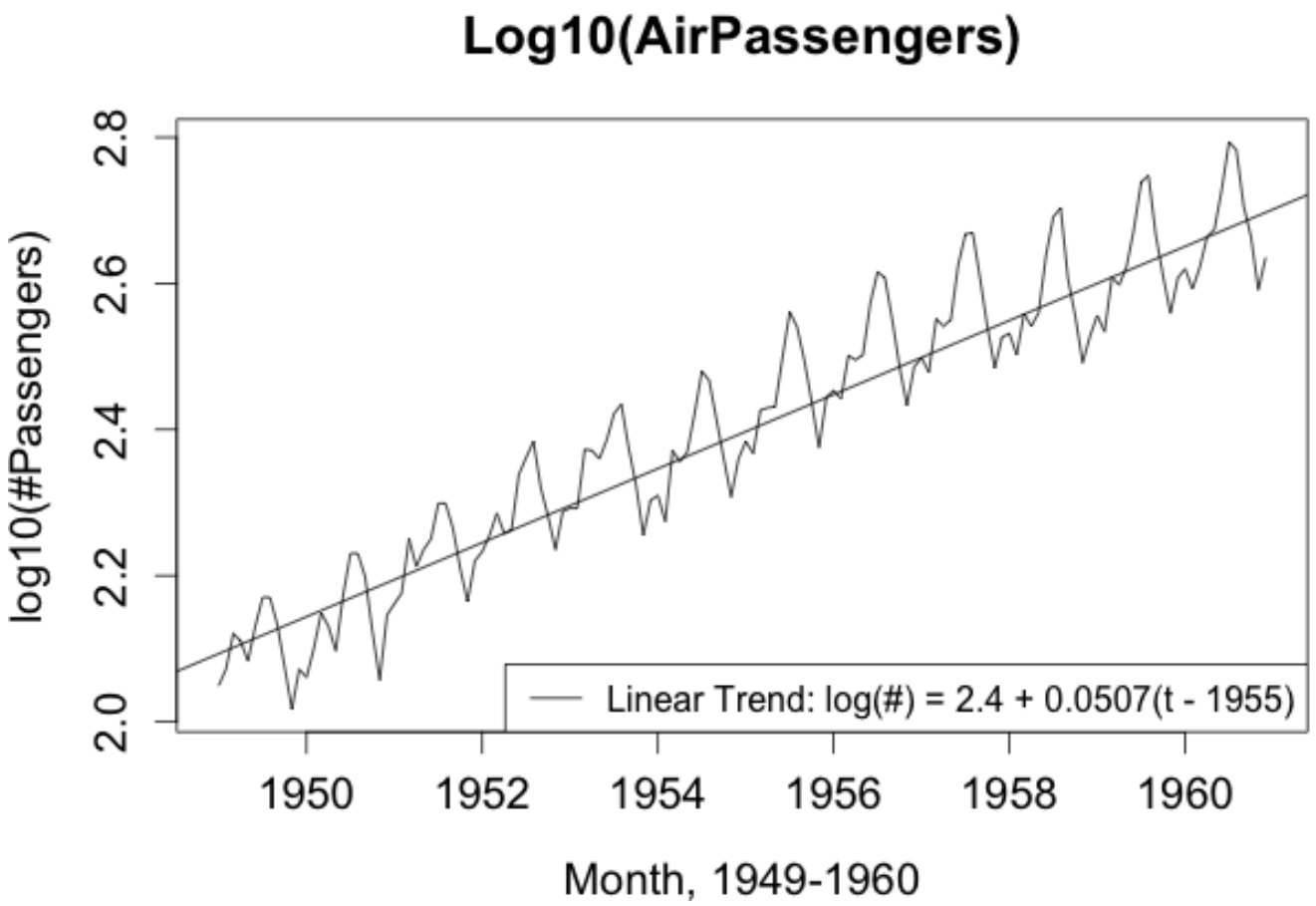
The classic Box & Jenkins airline data. Monthly totals of international airline passengers, 1949 to 1960.



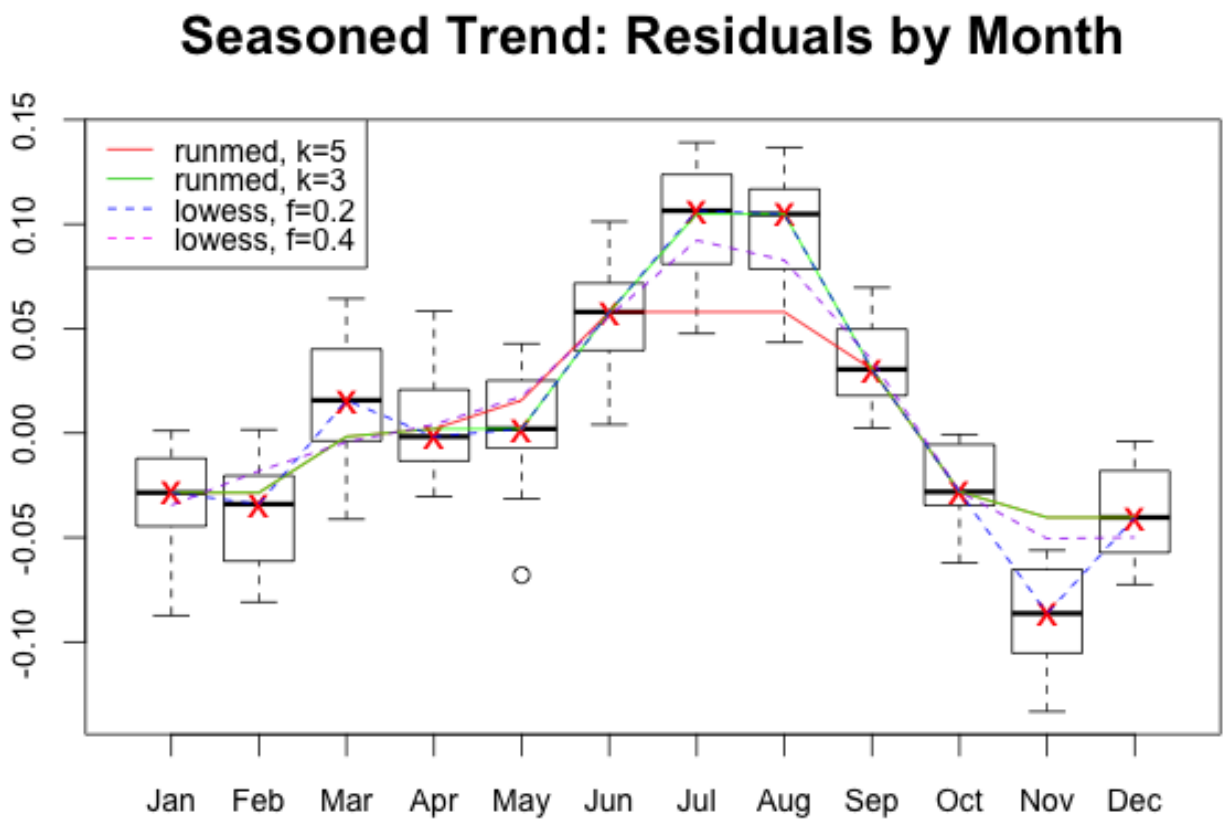
Example


Log10(AirPassengers)
log10(#Passengers)
2.0
2.2
2.4
2.6
2.8
1950
1952
1954
1956
1958
1960
Month, 1949-1960

Example


Log10(AirPassengers)
log10(#Passengers)
2.0
2.2
2.4
2.6
2.8
1950
1952
1954
1956
1958
1960
Month, 1949-1960
Linear Trend: log(#) = 2.4 + 0.0507(t - 1955)

Example


Seasoned Trend: Residuals by Month
runmed, k=5
runmed, k=3
lowess, f=0.2
lowess, f=0.4
0.15
0.10
0.05
0.00
-0.05
-0.10
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Summary

R functions for smoothing: help.search("smooth")

1. Basic library:
   - ksmooth: Kernel regression smoother
   - lowess: Scatter plot smoothing
   - smooth.spline: Fits a cubic smoothing spline to the supplied data.
   - predict.smooth.spline: predict from spline smooth
   - runmed: running medians
   - scatter.smooth: Plot and add a smooth curve computed by loess to a scatter plot.
   - smooth: Tukey's running median smoothing
   - smoothEnds: end-value smoothing for running medians
   - supsmu: Friedman's SuperSmoother
   - pspline: Smoothing splines using a pspline basis

Summary

R functions for smoothing: help.search("smooth")

2. library("graphics"):
   - panel.smooth: Simple Panel Plot
   - smoothScatter: Scatterplots with Smoothed Densities Color Representation
   - smooth: Tukey's running median smoothing
   - smoothEnds: end-value smoothing for running medians
   - runmed: running medians