QÜESTIIÓ, Vol 16, 1,2,3 pp. 135-145, 1992

QUANTILE PLOTS IN THE ANALYSIS OF HETEROSCEDASTIC MODELS

M. PEPIÓ and C. POLO
Laboratorio de Estadística
E.T.S.E.I.T. - U.P.C.

Recent developments in quality engineering methods have led to considerable interest in the analysis of variance, building a dispersion model, identifying important effects from replicated experiments and checking for significance by means of a half-normal plot. A methodology based on a chi-squared quantile plot is presented here for checking first the presence of heteroscedasticity, outliers and other data peculiarities, and after the estimation stage a new stepwise procedure tests for significant effects.

Quantile Plots in the Analysis of Heteroscedastic Models.

Keywords: Chi-squared Plot; Dispersion Model; Stepwise Test.

1. INTRODUCTION

In the process of ascertaining the functional relationship amongst the observed response and the experimental factors, two points are to be considered.

-Article rebut el juliol de 1991.
-Acceptat el novembre de 1992.

On the one hand, to include in the model all knowledge anticipated by the physical nature of the phenomena under analysis and, on the other hand, to look graphically into the data to reveal specific shapes to modelize the response. From a data analytic viewpoint, one would like to have procedurres which use some sort of statistical model for aiding the process of making inferences, while at the same time not requiring a commitment on any narrow specification of objectives, including the unquestioned acceptance of all the assumptions made in the model. Thus the techniques should have value not only for identifying possibly real effects but also for indicating the presence of outliers, heteroscedasticity and other peculiarities which are often assumed to be non-existent by the formal model.

2. QUANTILE PLOTS

Quantile plots [2, 3, 4] are general purpose displays that portray many distributional features of a set of data. Quantile plots not only are useful for graphically describing the distribution of a set of data values, but they can also be used to assess the fidelity of a set of data to a hypothesized probability distribution.

For many analyses the calculation and plotting of quantiles will suffice to enable an experimenter to discover most of the salient distributional features of a data set. This technique allows an analyst to see more complex variations in the data than those that are provided by simple summary statistics. The entire range of the distribution can be examined, and subtle shifts in shape, location and spread are easily detectable by departures from linearity, zero intercept and unit slope.

The most widely used quantile plot is likely the Normal Probability Plot, comparing the empirical distribution function of a set of data with the Normal distribution function. Be $y_{(1)} \leq \cdots \leq y_{(n)}$ the ordered observations, p_i a cumulative proportion associated with the $i-$th ordered observation —$(i - 0.5)/n$ and $i/(n + 1)$ are very common— and $q(p_i)$ the standard normal quantile, the Normal Probability Plot is a plot of the points $\{y_{(i)}, q(p_i)\}$, $i = 1, \ldots, n$.

3. QUANTILE PLOT FOR VARIABILITY

Common variance is a standard assumption in many statistical analysis (Linear Model, Anova, etc.) checked afterwards by means of a plot of residuals versus the predicted values. Nevertheless, if there are r replications within each cell of a multiway cross-classification or within each treatment of an $n = 2^k$, possibly fractional, factorial experiment, then the analysis leads to n sums of squared derivations from the within-replication mean, $s_1^2, \ldots, s_n^2$, each with $\nu = r - 1$ degrees of freedom.

Given the assumption of normality and no correlation, to test the homoscedasticity and assess the relative magnitudes of $s_1^2, \ldots, s_n^2$, these once divided by the unknown variance can be considered as a random sample from a central chi-squared distribution, and the appropiate quantile plot is thus a plot of the ordered values $s_{(1)}^2 \leq \cdots \leq s_{(n)}^2$, against the quantiles of a χ^2 distribution with $r - 1$ degrees of freedom. If all observations have the same variance the plot configuration would be linear with zero intercept and the slope would be an estimate of the common variance. But if the configuration is suggestive of two or more straight lines a reasonable interpretation would be that the treatments or cells belonging to the same linear piece have an underlying common variance but those that belong to two different pieces do not share a common variance.

4. DISPERSION MODEL

If lack of homoscedasticity is detected a commonly used method for identifying important dispersion effects from replicated experiments is based on a least squares analysis of the logarithm of the within-replication variance [1].

Let $Y_{ij} (i = 1, \ldots, n; j = 1, \ldots, r)$ be the responses of r replications of an $n = 2^k$ factorial experiment. The data are supposed to follow a location-dispersion model

$$Y_{ij} = m_i + \sigma_i \varepsilon_{ij} \tag{1}$$

where the ε_{ij} random values are independent, normally distributed, with zero mean and unit variance. A model [5, 7] relates the variances σ_i^2 to the dispersion effects θ_k by means of

$$h(\sigma^2) = \boldsymbol{A}\theta \tag{2}$$

where σ^2 is the vector of variances, θ the vector of unknown parameters, $\boldsymbol{A}$ denotes the regression matrix associated with the design and $\boldsymbol{a}_i$ will be used to denote the ith row of $\boldsymbol{A}$. For inference about the dispersion effects we use the sums of squared deviations from the within-replications mean $\overline{Y}_{i\cdot}$,

$$S_i = \sum_{j=1}^{r} (Y_{ij} - \overline{Y}_{i\cdot})^2 \qquad i = 1, \ldots, n \tag{3}$$

and then $S_i/\sigma_i^2 = u_i$ is distributed χ^2 with $\nu = r - 1$ degrees of freedom. Thus, in order to have an additive model, we take logarithms ($h(\bullet) \equiv \ln$), getting

$$\ln S_i = \ln \sigma_i^2 + \ln u_i = \boldsymbol{a}_i \theta + \ln u_i$$

where the random component $\ln u_i$ is $\ln \chi^2$.

With the least squares estimator $\hat{\theta}$,

$$\hat{\sigma}_i^2 = \frac{1}{r-1} e^{\boldsymbol{a}_i \hat{\theta}} \tag{4}$$

is an unbiased estimator for the variance of each treatment.

If $\hat{\sigma}_i^2$ is a good estimate of σ_i^2 then a plot of $S_i/\hat{\sigma}_i^2$ against the quantiles of the χ^2 distribution with $\nu = r - 1$ must show the whole set of points near to a straight line through the origin. Then a stepwise method to find out the significant dispersion effects θ_i is to introduce in the regression function the estimated effects one by one as per biggest absolute value to obtain the estimates $\hat{\sigma}_i^2$ and the quantile plot. If the plot shows any kind of departure from the straight line the regression function is incomplete requiring at least a new term. The procedure ends when the plot may be considered as sufficiently null and no other departures are uncovered.

5. AN EXAMPLE

Pignatiello and Ramberg [6] presented an experiment concerning the development of a heat-treatment process of leaf springs in trucks so that the free height, Y, of a spring in an unloaded condition be as close as possible to the target value of eight inches with minimum variability. The design factors were B, furnace temperature; C, heating time; D, transfer time; E, hold down time, and O, quench-oil temperature. Among the five factors, quench-oil temperature is not easily controllable, and Pignatello and Ramberg treated it as both a control factor and a noise factor. In our analysis we treat quench-oil temperature as

a control factor to estimate the regression function for the dispersion model and check for its significance. This results in three replications of a 2^{5-1} factorial. The data are presented in Table 1, jointly with the treatment means, $\overline{Y}_{i.}$, and the sums of squared deviations, S_i.

Table 1

Data From a Replicated 2^{5-1} factorial experiment.

Nº	B	C	BC	D	BD	CD	E	O	Y_{ij}			$\overline{Y}_{i.}$	S_i
1	+1	−1	−1	+1	−1	+1	+1	−1	−1	7,78	7,78	7,81	7,79	0,0006
2	+1	+1	−1	−1	−1	−1	+1	+1	−1	8,15	8,18	7,88	8,07	0,0546
3	+1	−1	+1	−1	−1	+1	−1	+1	−1	7,50	7,56	7,50	7,52	0,0024
4	+1	+1	+1	+1	−1	−1	−1	−1	−1	7,59	7,56	7,75	7,63	0,0208
5	+1	−1	−1	+1	+1	−1	−1	+1	−1	7,94	8,00	7,88	7,94	0,0072
6	+1	+1	−1	−1	+1	+1	−1	−1	−1	7,69	8,09	8,06	7,95	0,0993
7	+1	−1	+1	−1	+1	−1	+1	−1	−1	7,56	7,62	7,44	7,54	0,0168
8	+1	+1	+1	+1	+1	+1	+1	+1	−1	7,56	7,81	7,69	7,69	0,0312
9	+1	−1	−1	+1	−1	+1	+1	−1	+1	7,50	7,25	7,12	7,29	0,0745
10	+1	+1	−1	−1	−1	−1	+1	+1	+1	7,88	7,88	7,44	7,73	0,1290
11	+1	−1	+1	−1	−1	+1	−1	+1	+1	7,50	7,56	7,50	7,52	0,0024
12	+1	+1	+1	+1	−1	−1	−1	−1	+1	7,63	7,75	7,56	7,65	0,0184
13	+1	−1	−1	+1	+1	−1	−1	+1	+1	7,32	7,44	7,44	7,40	0,0096
14	+1	+1	−1	−1	+1	+1	−1	−1	+1	7,56	7,69	7,62	7,62	0,0084
15	+1	−1	+1	−1	+1	−1	+1	−1	+1	7,18	7,18	7,25	7,20	0,0033
16	+1	+1	+1	+1	+1	+1	+1	+1	+1	7,81	7,50	7,59	7,63	0,0510

Fig. 1 is quantile plot of the ordered S_i, displaying, as we may see, two distinct set of points. Accordingly, we apply least squares to estimate the dispersion effects, $\hat{\theta}$, as shown in Table 2.

Table 2

Estimates of the dispersion coefficients.

Alias	$\hat{\theta}$	$\hat{\theta}^*$	$\hat{\theta}'$
I	-4.2382	−4.3739	−4.2924
$B+CDE$	0.9443	1.0800	0.9985
$C+BDE$	−0.2832	−0.1475	−0.2290
$BC+DE$	0.0005	−0.1352	−0.0537
$D+BCE$	0.1241	0.2598	0.1784
$BD+CE$	−0.2133	−0.3490	−0.2676
$CD+BE$	0.3364	0.2007	0.2822
$E+BCD$	0.1078	0.2435	0.1620
O	0.1401	0.0044	0.0859
BO	−0.2952	−0.1595	−0.2410
CO	−0.2974	−0.1617	−0.2431
DO	−0.5549	−0.4192	−0.5006
EO	0.0648	0.2005	0.1190
BCO	0.5448	0.4091	0.4906
BDO	0.2154	0.0797	0.1611
CDO	0.4282	0.2924	0.3739

The biggest absolute effect is associated to factor B, furnace temperature, giving the estimates

$$\hat{\sigma}_1^2 = \frac{1}{2}\,\mathrm{EXP}\,[-4.2382 + 0.9443^* B]$$

and Fig. 2(a) plots the ordered $S_i/\hat{\sigma}_{1i}^2$ against the quantiles of the χ^2 distribution of $\nu = 2$. This figure shows all points except one near a straight line through the origin. The anomalous point corresponds to S_9. We proceed to introduce the next biggest absolute effect into the regression function to obtain the estimates

$$\hat{\sigma}_2^2 = \frac{1}{2}\,\mathrm{EXP}\,[-4 - 2382 + 0.9443^* B - 0.5549^* D^* O]$$

and the quantile plot of Fig. 2(b). This figure displays a pattern similar to that of panel 2(a) and the deviant point corresponds again to the same treatment. Then S_9 is considered an outlier and in order to correct it the bigger or the smaller replicate must be replaced by the average of the other two. If the bigger, 7.5, is the replaced one, then $S_9^* = 0.0085$ and the estimated effects are denoted $\hat{\theta}^*$ and exhibited in Table 2. The quantile plots of $S_i^*/\hat{\sigma}_{1i}^{*2}$ and $S_i^*/\hat{\sigma}_{2i}^{*2}$ are shown in panels (a) and (b) of Fig. 3, and none is a null one. But when we replace the smallest data value, 7.12, we obtain $S_9' = 0.0313$, the effects $\hat{\theta}'$ of Table 2 and the quantile plots of Fig. 4. The first three panels, (a), (b) and (c), are associated to the regression function with only one, two and three effects introduced according their biggest absolute value. Looking at these plots we see that the anomaly has been removed and can infer that only factor B effects dispersion, thus

$$\hat{\sigma}^2 = \frac{1}{2}\,\mathrm{EXP}\,[-4.2924 + 0.9985^* B] \tag{5}$$

is the best estimate of the variability associated to each treatment. Fig. 4(d) is a quantile plot of the sums of squared deviations, S_i', once amended the anomalous value of number 9 treatment. This exhibit also shows two distinct straight lines proving the heteroscedasticity of the process.

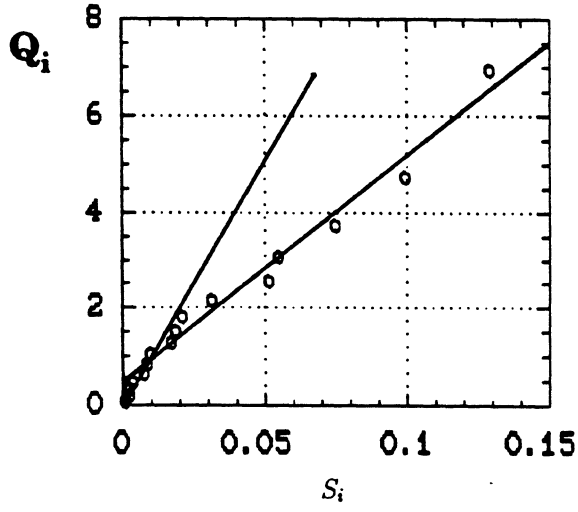
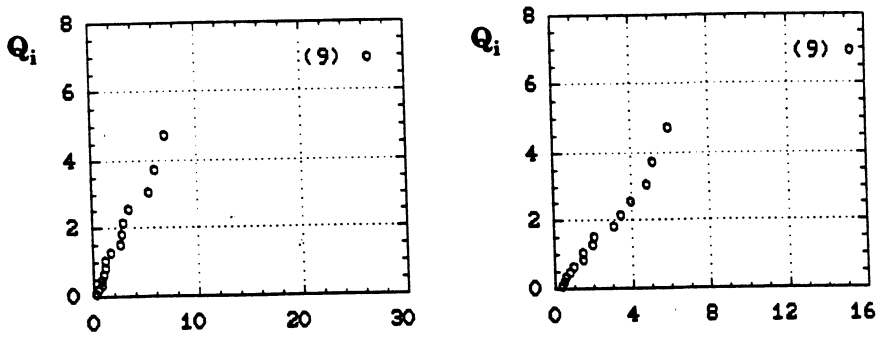
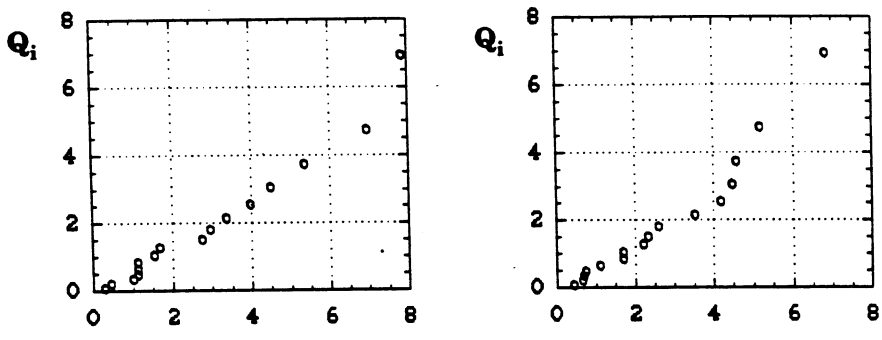


Figure 1.
Chi-squared quantile plot corresponding to Table 1.



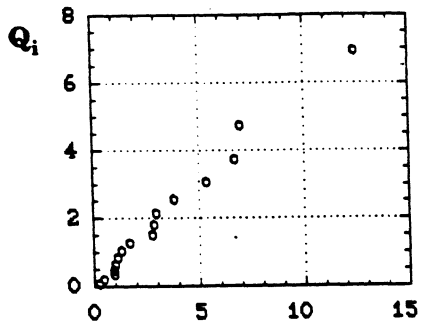
(a) $S_i/\hat{\sigma}^2_{1i}$ (b) $S_i/\hat{\sigma}^2_{2i}$

Figure 2.
Q-plots for the stepwise regresion.

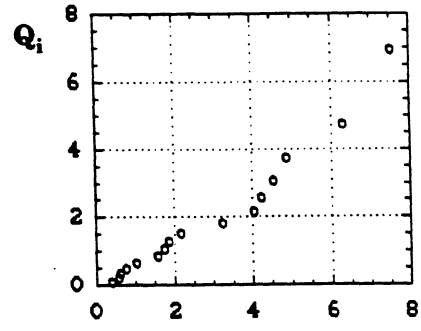


(a) $S^*_i/\hat{\sigma}^2_{1i}$ (b) $S^*_i/\hat{\sigma}^2_{2i}$

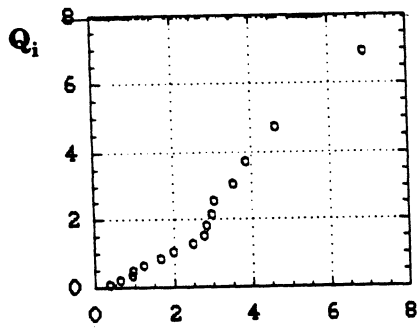
Figure 3.
Chi-squared probability plots disentangling the outlier.



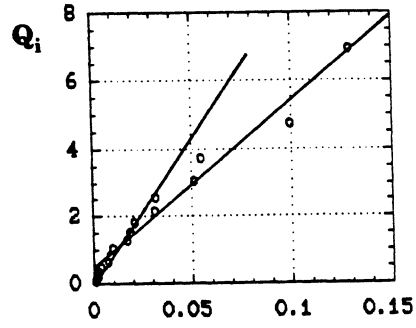
(a) $S_i'/\hat{\sigma}_{1i}^2$



(b) $S_i'/\hat{\sigma}_{2i}^2$

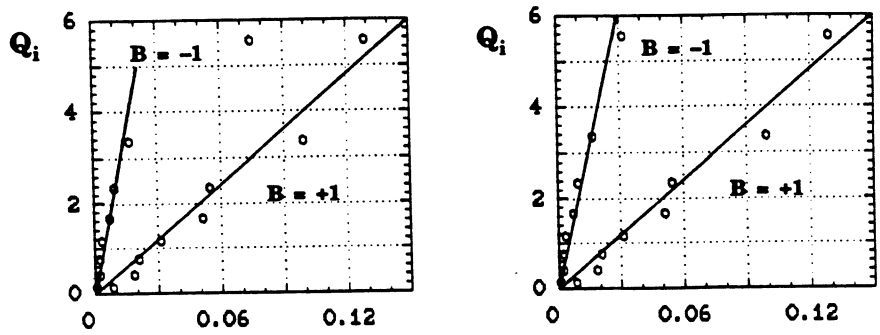


(c) $S_i'/\hat{\sigma}_{3i}^2$



(d) S_i'

Figure 4.
Plots for removing the outlier and looking for significance.



(a) S_i^* (b) S_i'

Figure 5.
Q-plots of two subsets of sums of squared deviations.

Since the process variability only depends on factor B, the sums of squared deviations can be split in two sets of equal variance according the B levels, and a quantile plot of each set has to show a linear configuration. Fig. 5 exhibits these plots: panel (a) displays data as measured and shows clearly two straight lines and one deviant point, S_9. Panel (b) of figure 5 corresponds to the sums of squared deviations after amending that anomaly, its configuration corroborates the conclusions of our analysis.

6. REFERENCES

[1] **Bartlett, M.S. and Kendall, D.G.** (1946). "The Statistical Analysis of Variance — Heterogeneity and the Logaritmic Transformation". *J.R.S.S., Ser. B,* **8**, 128–138.

[2] **Daniel, C.** (1959). "Use of Half-Normal Plots in Interpreting Factorial Two-level Experiments". *Technometrics,* **1**, 311–341.

[3] ——— (1976). "Applications of Statistics to Industrial Experimentation". J. Wiley.

[4] **Gnanadesikan, R.** (1988). "Graphical Methods for Internal Comparison in ANOVA and MANOVA". *Handbook of Statistics,* **1**, North-Holland.

[5] **Pepió, M. and Polo, C.** (1989). "Estimación eficiente para la optimización de un proceso". *Qüestiió,* Vol **13, 1-2-3**, 193–211.

[6] **Pignatiello, J.J. and Ramberg, J.S.** (1985). "Discussion on Kackar's paper". *Journal of Quality Tecnology,* **17** (4), 198–209.

[7] **Polo, C. and Pepió, M.** (1991). "Modelización y Optimización de Procesos". CPDA.