A TEST OF THE HYPOTHESIS THAT THE REGRESSION OF Y ON X IS LINEAR, AGAINST THE ALTERNATIVE HYPOTHESIS THAT THE REGRESSION OF $Y^{\frac{1}{p}}$ ON X IS LINEAR

BU-487-M

by

November 1973

T. M. Hsuan and D. S. Robson

Abstract

Tukey's single degree of freedom non-additivity test can be extended to the general linear model for testing whether the fit to a linear model is improved by a specified (non-linear) transformation $T^{\lambda}(y)$ of the dependent variable. A single degree of freedom sum of squares is formulated to test whether the regression of Y on X is linear against the alternative that the regression of $Y^{\frac{1}{p}}$ on X is linear, and the sum of squares completely accounts for residual if and only if $Y = (\alpha+\beta X)^{p}$.

A TEST OF THE HYPOTHESIS THAT THE REGRESSION OF Y ON X IS LINEAR, AGAINST THE ALTERNATIVE HYPOTHESIS THAT THE REGRESSION OF Y^{p} [1] ON X IS LINEAR

by

T. M. Hsuan and D. S. Robson

BU-487-M November 1973

Introduction

The idea presented in BU-334-M of testing for additivity on a transformed scale in the balanced two-way classification can be extended to the general linear model $Y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$ by constructing a single degree of freedom sum of squares for testing whether the fit to a linear model is improved by a specified (non-linear) transformation $T^{1}(y)$ of the dependent variable. At present the procedure for testing goodness-of-fit to $X\beta$ appears to consist of altering the design matrix X by appending or deleting columns and then testing whether R^2 is significantly altered. The proposed approach may be viewed as leaving X unchanged but altering the scale of Y .

A natural procedure to follow in the latter case is to fit the linear model on both scales, y and $T^{1}(y)$, and compare the resulting values of R^2, but only approximate tests can be derived for comparing $R^2(y)$ with $R^2(T^{1}(y))$. To circumvent this difficulty we propose fitting first on the y-scale, for which the null hypothesis specifies $NIID(0,\sigma^2 I)$ errors, and then transforming these best fitting coefficients $\hat{\beta} = (X'X)^{-1}X'Y$ of the linear model $X\beta$ into estimates $\tilde{\beta}_T$ for the non-linear model $T(X\beta)$ by $\hat{\beta} = (X'X)^{-1}X'T(X\tilde{\beta}_T)$. Estimates $\tilde{\beta}_T$ obtained in this manner depend on the observations Y only through $\hat{\beta}$; i.e., $\tilde{\beta}_T = \tilde{\beta}_T(X,\hat{\beta})$, and hence the predicted values $\tilde{Y}_T = T(X\tilde{\beta}_T)$ are statistically independent of the least squares residuals $e = Y - X\hat{\beta}$ under H_0 . The single degree of freedom

sum of squares

$$\tilde{S}_T^2 = (e'\tilde{Y}_T)^2 \Big/ \tilde{Y}_T'(I-X(X'X)^{-1}X')\tilde{Y}_T$$

is therefore distributed as $\sigma^2\chi^2_{1\ df}$ independently of the remainder $e'e - \tilde{S}_T^2$ under H_0 . This does provide a test of the $T(X\beta)$ model in the sense that if $Y = T(X\beta)$ then $e'e = \tilde{S}_T^2$.

There are some technical matters to be cleared up for this procedure: does a solution $\tilde{\beta}_T$ exist for all T, and if more than one solution exists is $T(X\tilde{\beta}_T)$ unique for all T? This point is illustrated here by constructing a test of Y linear on X, against the alternative that $E(Y|X) = (\alpha+\beta X)^p$.

The Power Transformation in Simple Linear Regression

Our objective here is to construct a single degree of freedom sum of squares for testing H_0 that the regression of Y on X is linear against the alternative hypothesis that $E(Y|X) = (\alpha_p+\beta_p X)^p$ and to investigate whether $\tilde{\alpha}_p$ and $\tilde{\beta}_p$ exist for all real p . The estimation equations $X'Y = X'T(X\tilde{\beta}_T)$ are

$$\sum_{i=1}^{n} (\tilde{\alpha}_p+\tilde{\beta}_p x_i)^p = \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} x_i (\tilde{\alpha}_p+\tilde{\beta}_p x_i)^p = \sum_{i=1}^{n} x_i y_i \quad .$$

If the solutions $\tilde{\alpha}_p$ and $\tilde{\beta}_p$ satisfying the above equations exist, then we can calculate $\tilde{S}_T^2$. Writing

$$(\tilde{\alpha}_p+\tilde{\beta}_p x_i)^p = \tilde{\alpha}_p^p\left(1+\frac{\tilde{\beta}_p}{\tilde{\alpha}_p} x_i\right)$$

let $\lambda = \dfrac{\tilde{\beta}_p}{\tilde{\alpha}_p}$ and assume $x_i \geq 0$, $y_i \geq 0$ and not all x_i's are equal. The estimation equations become

$$\tilde{\alpha}_p^p \sum_{i=1}^{n} (1+\lambda x_i)^p = \sum_{i=1}^{n} y_i$$

$$\tilde{\alpha}_p^p \sum_{i=1}^{n} x_i (1+\lambda x_i)^p = \sum_{i=1}^{n} x_i y_i$$

then $$\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i} = \frac{\sum_{i=1}^{n} x_i (1+\lambda x_i)^p}{\sum_{i=1}^{n} (1+\lambda x_i)^p} \text{ def.} = f_p(\lambda) \ .$$

If there exist λ_o satisfying $f_p(\lambda_o) = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i}$ then

$$\tilde{\alpha}_p^p = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} (1+\lambda_o x_i)^p} \quad \text{and} \quad \tilde{\beta}_p = \tilde{\alpha}_p \lambda_o \ ,$$

hence, the problem reduces to checking whether there exist solutions for

$$f_p(\lambda) = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i} \ .$$

$$\frac{\partial}{\partial \lambda} f_p(\lambda)$$

$$= \frac{1}{\left[\Sigma(1+\lambda x_i)^p\right]^2} \left[\left(\Sigma(1+\lambda x_i)^p\right)\left(\Sigma p x_i^2 (1+\lambda x_i)^{p-1}\right) - \left(\Sigma x_i (1+\lambda x_i)^p\right)\left(\Sigma p x_i (1+\lambda x_i)^{p-1}\right)\right]$$

$$= \frac{p}{\left[\Sigma(1+\lambda x_i)^p\right]^2}\left[\sum_{i\neq j} x_j^2(1+\lambda x_i)^p(1+\lambda x_j)^p - \sum_{i\neq j} x_i x_j (1+\lambda x_i)^p(1+\lambda x_j)^p\right]$$

$$= \frac{p}{\left[\Sigma(1+\lambda x_i)^p\right]^2}\left[\sum_{1<j} (1+\lambda x_i)^{p-1}(1+\lambda x_j)^{p-1}(x_i - x_j)^2\right]$$

Denote $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ as the ordered x_i's .

Since some of the $(1+\lambda x_i)^p$, $i=1,\cdots,n$ may not be real when $\lambda < -\frac{1}{x_{(n)}}$, then we can assume that $\lambda \geq -\frac{1}{x_{(n)}}$. We have

$$\frac{\partial}{\partial\lambda} f_p(\lambda) > 0 \quad \text{if} \quad p > 0$$

$$\frac{\partial}{\partial\lambda} f_p(\lambda) < 0 \quad \text{if} \quad p < 0$$

$$\frac{\partial}{\partial\lambda} f_p(\lambda) = 0 \quad \text{if} \quad p = 0 ;$$

hence, $f_p(\lambda)$ is an increasing function of λ for $p > 0$, $f_p(\lambda)$ is a decreasing function of λ for $p < 0$ and $f_0(\lambda) = \bar{x}$ for all λ .

$$f_p\left(-\frac{1}{x_{(n)}}\right) = \frac{\sum_{i=1}^{n} x_i\left(1 - \frac{x_i}{x_{(n)}}\right)^p}{\sum_{i=1}^{n}\left(1 - \frac{x_i}{x_{(n)}}\right)^p}$$

$$\lim_{p\to\infty} f_p\left(-\frac{1}{x_{(n)}}\right) = x_{(1)}$$

$$\lim_{p\to-\infty} f_p\left(-\frac{1}{x_{(n)}}\right) = x_{(n)}$$

$$\lim_{\lambda\to\infty} f_p(\lambda) = \lim_{\lambda\to\infty} \frac{\sum_{i=1}^{n} x_i(\lambda x_i)^p}{\sum_{i=1}^{n}(\lambda x_i)^p} = \frac{\sum_{i=1}^{n} x_i^{p+1}}{\sum_{i=1}^{n} x_i^p}$$

$$\lim_{p\to\infty}\lim_{\lambda\to\infty} f_p(\lambda) = x_{(n)}$$

$$\lim_{p\to-\infty}\lim_{\lambda\to\infty} f_p(\lambda) = x_{(1)} \ .$$

Also, for any fixed λ

$$\frac{\partial}{\partial p} f_p(\lambda)$$

$$= \frac{1}{\left[\Sigma(1+\lambda x_i)^p\right]^2}\left[\left(\Sigma(1+\lambda x_i)^p\right)\left(\Sigma x_i (1+\lambda x_i)^p \log(1+\lambda x_i)\right) - \left(\Sigma x_i (1+\lambda x_i)^p\right)\left(\Sigma(1+\lambda x_i)^p \log(1+\lambda x_i)\right)\right]$$

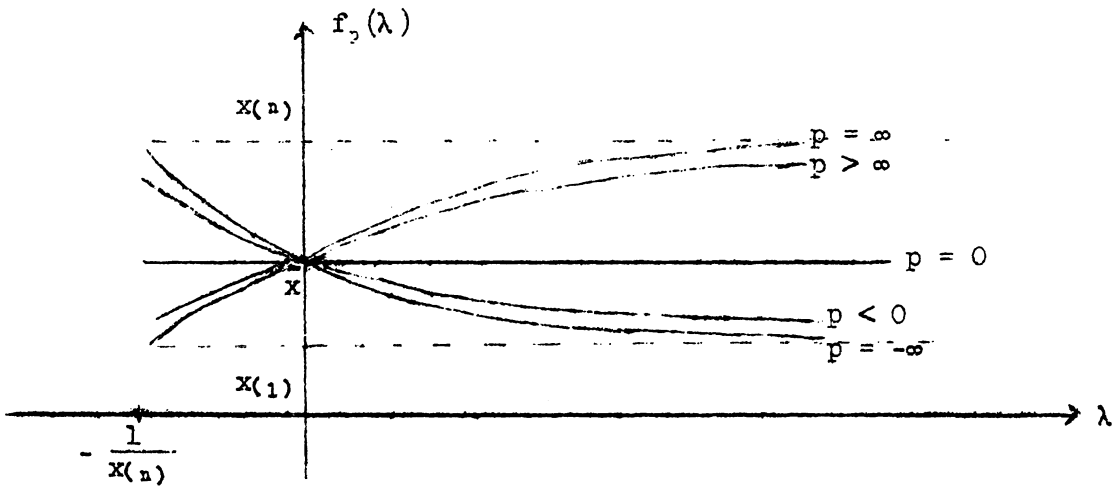
$$= \frac{1}{\left[\Sigma(1+\lambda x_i)^p\right]^2}\left[\sum_{i<j} (1+\lambda x_{(i)})^p (1+\lambda x_{(j)})^p (x_{(i)} - x_{(j)})\left(\log(1+\lambda x_{(i)}) - \log(1+\lambda x_{(j)})\right)\right]$$

$$\frac{\partial}{\partial p} f_p(\lambda) > 0 \quad \text{if} \quad \lambda > 0$$

$$\frac{\partial}{\partial p} f_p(\lambda) < 0 \quad \text{if} \quad -\frac{1}{x_{(n)}} < \lambda < 0$$

$$f_p(0) = \bar{x} \ .$$

Hence, $f_p(\lambda)$ is an increasing function of p for $\lambda > 0$, $f_p(\lambda)$ is a decreasing function of p for $\lambda < 0$, and $f_p(0) = \bar{x}$. These results can be illustrated by the following schematic graph.



Because of $x_{(1)} < \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i} < x_{(n)}$, we can conclude that $f_p(\ ) = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i}$

has a unique solution if $\dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i}$ is in between $\dfrac{\sum_{i=1}^{n} x_i^{p+1}}{\sum_{i=1}^{n} x_i^{p}}$ and $\dfrac{\sum_{i=1}^{n} x_i \left(1 - \dfrac{x_i}{x_{(n)}}\right)^p}{\sum_{i=1}^{n} \left(1 - \dfrac{x_i}{x_{(n)}}\right)^p}$;

particularly, a unique solution exists when $p = \infty$ and $p = -\infty$; the solution exists only when $\dfrac{\Sigma x_i y_i}{\Sigma y_i} = \bar{x}$ for $p = 0$ and then there are infinitely many solutions.

Letting $\tilde{y}_i = (\tilde{\alpha}_p + \tilde{\beta}_p x_i)^p$ and assuming $\tilde{\alpha}_p$ and $\tilde{\beta}_p$ exist then the single degree of freedom sum of squares

$$\tilde{S}_T^2 = \frac{\left(\sum_1^n e_i \tilde{y}_i\right)^2}{\sum_1^n \tilde{y}_i^2 - n\bar{y}^2 - \dfrac{\left[\sum_1^n (x_i - \bar{x}) y_i\right]^2}{\sum_1^n (x_i - \bar{x})^2}}$$

is distributed as $\sigma^2 \chi^2_{1\ df}$ independently of $\Sigma e_i^2 - \tilde{S}_T^2$ under H_0.

References

Robson, D. S. Tests for non-additivity viewed as tests of the hypothesis of no interaction. A preliminary report. BU-334-M, October 1970.

Robson, D. S. One degree of freedom for testing a test of the hypothesis that the regression of Y on X is linear, against the alternative hypothesis that the regression of log Y on X is linear. BU-438-M, December 1972.