Uso de Transformações na Análise de Variância e Análise de Regressão

Estevão Kusaba Rodrigues
Mateus Godoi
Melissa Tiemi Tanaka
Sávio Campos de Souza
Tiago Sotelo

Introdução

Suposições não verdadeiras

Teste F para análise de variância não é sensível a alguns desvios das suposições

Desvio acentuado causa resultados alterados

Transformação

Existem várias transformações, neste trabalho estudaremos: $logY, \sqrt{Y}, \frac{1}{Y}, arcsen\sqrt{Y}$

O uso de transformações na década de 30

Objetivos

Alertar sobre o uso indevido das técnicas de análise de variância e regressão

Apresentar alguns tipos de transformações

Orientar como detectar a necessidade do uso de transformações

Mostrar os cuidados que devem ser tomados ao utilizar transformações

Causas dos Desvios das Suposições

Não-Correlação dos erros

Homogeneidade de variância

Normalidade

Aditividade

Causas dos Desvios das Suposições

Não correlação dos erros

Pode ser assegurada normalmente com esquema de aleatorização adequado

O desvio dessa suposição ocorre mais comumente quando amostras são tomadas sequencialmente no tempo

Causas dos Desvios das Suposições

Homogeneidade de variância

Danos em parte do experimento

Diferença na natureza dos tratamentos

Situações em que a variância é função da média, associado a não-normalidade (distribuição de Poisson)

Outliers genuínos e não-genuínos

Causas dos Desvios das Suposições

Normalidade

Normalidade perfeita raramente acontece

Variáveis discretas não tem distribuição normal

Para amostras com números inteiros e pequenos, a distribuição tende mais a Poisson

Causas dos Desvios das Suposições

Aditividade

Não-aditividade muitas vezes é inerente ao modelo

Causada pela presença de outliers

Como Detectar Desvios das Suposições

Análise de Resíduos

Análise descritiva

Testes de significância

Como Detectar Desvios das Suposições

Análise de Resíduos

- Métodos de análise gráfica e métodos numéricos
- Usada como procedimento de rotina
- Detecção de não-aditividade por outliers

Como Detectar Desvios das Suposições

Análise descritiva

- Construção de Histograma (não normalidade)
- Papel de probabilidade Normal (não normalidade)

Como Detectar Desvios das Suposições

Análise descritiva

- Inspeção das variâncias dos grupos analisados (Heterocedasticidade)
- Amplitude de variação (Heterocedasticidade)
- Coeficiente de variação (Heterocedasticidade)

Como Detectar Desvios das Suposições

Análise descritiva

- Gráficos de Perfil (não aditividade)

Como Detectar Desvios das Suposições

Testes de significância

falha de uma suposição altera o nível de significância

Perda de sensibilidade, existe um teste mais poderoso, e de precisão dos estimadores

Solução Teórica

Suponha que existe uma relação entre a média $E[Y] = \mu$ e a variância $Var[Y] = \sigma^2$

$$\sigma^2 = f(\mu)$$

Procuramos uma transformação de Y, $Z = g(Y)$ tal que $Var[Z]$ seja constante.

Desenvolvendo em série de Taylor em torno de μ,obtemos:

$$Z = g(Y) = g(\mu) + (Y - \mu)g'(\mu)$$

Solução Teórica

Para esse grau de aproximação, temos que:

$$E[Z] = E[g(\mu) + (Y - \mu)g'(\mu)] = g(\mu)$$

e

$$Var[Z] = E[(Z - E[Z])^2] = E[((Y - \mu)g'(\mu))^2] = [g'(\mu)]^2 Var[Y]$$

Lembrando que $Var[Y] = \sigma^2 = f(\mu)$, obtemos:

$$Var[Z] = [g'(\mu)]^2 f(\mu) = K$$

Onde, por hipótese, K é uma constante positiva.

Solução Teórica

Portanto,

$$g'(\mu) = \sqrt{\frac{K}{f(\mu)}} \quad \Longrightarrow \quad g(\mu) = \int \sqrt{\frac{K}{f(\mu)}} d\mu$$

De forma mais geral, $g(y) = \int \sqrt{\frac{K}{f(y)}} dy$

Veremos a seguir alguns exemplos.

Exemplo 1

Considere que $Y \sim Poisson(\mu)$

$$Var[Y] = f(\mu) = \mu$$

Então:

$$g(\mu) = \int \frac{\sqrt{K_1}}{\sqrt{\mu}} d\mu = K_1\sqrt{\mu} + K_2$$

$Z = \sqrt{Y}$ estabiliza a variância.

$$Var[Z] = [g'(\mu)]^2 Var[Y] = (\frac{1}{2\sqrt{\mu}})^2 \mu = \frac{1}{4}$$

Exemplo 2

Seja Y uma variável aleatória tal que $E[Y] = \mu$ e $Var[Y] = \mu^2$

$$g(\mu) = \int \frac{\sqrt{K_1}}{\sqrt{\mu^2}} d\mu = K_1 log(\mu) + K_2$$

Então, a transformação $Z = log(Y)$ estabiliza a variância.

$$Var[Z] = [g'(\mu)]^2 Var[Y] = \frac{1}{\mu^2}\mu^2 = 1$$

Exemplo 3

Seja $Y^* \sim Binomial(n, \mu)$. Consideremos a proporção de sucessos $Y = \frac{Y^*}{n}$

$$Var[Y] = f(\mu) = \frac{\mu(1-\mu)}{n}$$

$$g(\mu) = \int \frac{\sqrt{nK}}{\sqrt{\mu(1-\mu)}} d\mu = K_1 arcsen(\sqrt{\mu}) + K_2$$

Então, $Z = arcsen(\sqrt{\mu})$ estabiliza a variância.

$$Var[Z] = [g'(\mu)]^2 Var[Y] = \frac{1}{[2\sqrt{\mu(1-\mu)}]^2} . \frac{\mu(1-\mu)}{n} = \frac{1}{4n}$$

Transformação Logarítmica

justificativa: valores distintos de logaritmos em bases distintas diferem apenas por um fator constante

$(\log_a b = k \times log_c b)$

Os resultados da estatística F da análise de variância não se alteram com transformações lineares

Quando aparecer valor zero ou negativo é preciso usar log(Y+1) ou log(Y+K)

Transformação Logarítmica

Tende a estabilizar a variância de grupos que na variável original tenham variâncias muito distintas.

É apropriada quando na escala original, o desvio padrão é proporcional à média.

Torna distribuições não-normais mais próximas da normal, em muitos casos.

Transformação raiz quadrada

Se a variável resposta (Y) apresentar valores negativos é preciso

aplicar transformação $\sqrt{Y + K}$

Sendo K uma constante conveniente

Se na variável original a variância é proporcional à média, a transformação raiz quadrada estabiliza a variância

A distribuição de Poisson é um caso típico

Exemplos: transformação raiz quadrada

Se alguns valores da variável resposta (Y) são pequenos, tais que as médias dos grupos comparados esteja entre 2 e 10, e especialmente se aparecer zeros, a transformação $\sqrt{Y + \frac{1}{2}}$ é recomendada.

Para contagens pequenas (<10) sugere-se usar a transformações $\sqrt{Y+1}$ ou $\sqrt{Y} + \sqrt{Y+1}$ ou $\sqrt{Y + \frac{3}{2}}$

são mais efetivas para estabilizar a variância.

Quando a média e a variância são inversamente proporcionais a transformação $\sqrt{Y_{max}} - \sqrt{Y_{max} - Y}$

tem se mostrado eficiente para a estabilização da variância

Transformação recíproca

$$Z = \frac{1}{Y}$$

ou se há observação com valor zero: $Z = \frac{1}{Y+1}$ ou $Z = \frac{1}{Y+K}$

Estabiliza a variância se a variância de Y for proporcional à potência quarta da média

$Z = \frac{1}{Y+K}$ para K conveniente, pode aproximar distribuição não-normal a normal

Áreas de uso: análise de tempo de sobrevivência, estudos de tempo de cura, farmacológicos ou de plantas por unidade de área, quando a densidade dos grupos comparados é muito variável.

Transformação Angular $arcsen\sqrt{Y}$

Utilizada para estabilizar a variância da variável “proporção de sucessos”, quando a variável número de sucesso segue distribuição Binomial

Recomendada quando as porcentagens dos grupos a serem comparados cobrem uma grande amplitude de variação. Se todas porcentagens variam entre 30% a 70% a transformação não é necessária. Porque há pouca variação do $p(1-p)$ e a variância se mantém razoavelmente constante.

Transformações “Probit” e “Logit”

“Probit”: a função distribuição da N(0,1) é usada para modelar frequências acumuladas, principalmente para relacionar estas frequências com variáveis explicativas (FINNEY, 1971)

“Logit”: similar à transformação “Probit” só que baseada na distribuição logística ao invés da N(0, 1). (BERKSON 1944)

Método Box-Cox

• Transformação que torna as condições de normalidade e homocedasticidade satisfeitas simultaneamente.

• Trabalhar com uma família de transformações (Moore & Tukey, Anscombe & Tukey, 1954)

• Tukey (1957) sugeriu a classe das funções do tipo potência:

$$y(\lambda) = \begin{cases} y^{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}, \lambda \in \mathbb{R}$$

Problemas de descontinuidade

Tukey (1957):

$$y(\lambda) = \begin{cases} y^{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

Box & Cox (1964):

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

Procedimento apropriado

Box & Cox (1954):

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

$$\dot{y} = \ln^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} \ln y_i\right]$$

Procedimento computacional

O estimador de máxima verossimilhança de λ é tal que SSE(λ) é mínima.

- Ajuste modelos para y(λ) para diferentes valores de λ, e plote SSE(λ) vs λ;
- Escolha o λ correspondente ao menor SSE(λ) (ou outro conveniente);
- Ajuste o modelo final $y(\lambda) = X\beta + \varepsilon$ com o λ escolhido.

Intervalo de confiança

O intervalo de confiança para λ consiste dos λ que satisfazem a desigualdade:

$$L(\hat{\lambda}) - L(\lambda) \leq \frac{1}{2}\chi^2_{\alpha,1}/n$$

Isso equivale a traçar uma reta no gráfico da função $L(\lambda)$ vs λ na altura:

$$L(\hat{\lambda}) - \frac{1}{2}\chi^2_{\alpha,1}$$

No gráfico dos SSE(λ) vs λ, a reta seria traçada na altura:

$$SS^* = SSE(\hat{\lambda})e^{\chi^2_{\alpha,1}/n}$$

Exemplo (Montgomery, Peck, Vining)

TABLE 5.2 Demand (y) and Energy Usage (x) Data for 53 Residential Cnstomers, August

Customer	x (kWh)	y (kW)	Customer	x (kWh)	y(kW)
1	679	0.79	27	837	4.20
2	292	0.44	28	1748	4.88
3	1012	0.56	29	1381	3.48
4	493	0.79	30	1428	7.58
5	582	2.70	31	1255	2.63
6	1156	3.64	32	1777	4.99
7	997	4.73	33	370	0.59
8	2189	9.50	34	2316	8.19
9	1097	5.34	35	1130	4.79
10	2078	6.85	36	463	0.51
11	1818	5.84	37	770	1.74
12	1700	5.21	38	724	4.10
13	747	3.25	39	808	3.94
14	2030	4.43	40	790	0.96
15	1643	3.16	41	783	3.29
16	414	0.50	42	406	0.44
17	354	0.17	43	1242	3.24
18	1276	1.88	44	658	2.14
19	745	0.77	45	1746	5.71
20	435	1.39	46	468	0.64
21	540	0.56	47	1114	1.90
22	874	1.56	48	413	0.51
23	1543	5.28	49	1787	8.33
24	1029	0.64	50	3560	14.94
25	710	4.00	51	1495	5.11
26	1434	0.31	52	2221	3.85
			53	1526	3.93

Exemplo (Montgomery, Peck, Vining)

TABLE 5.7 Values of the Residual Sum of Squares for Various Values of λ, Example 5.3

λ	$SS_{Res}(\lambda)$
–2	34,101.0381
–1	986.0423
–0.5	291.5834
0	134.0940
0.125	118.1982
0.25	107.2057
0.375	100.2561
0.5	96.9495
0.625	97.2889
0.75	101.6869
1	126.8660
2	1,275.5555

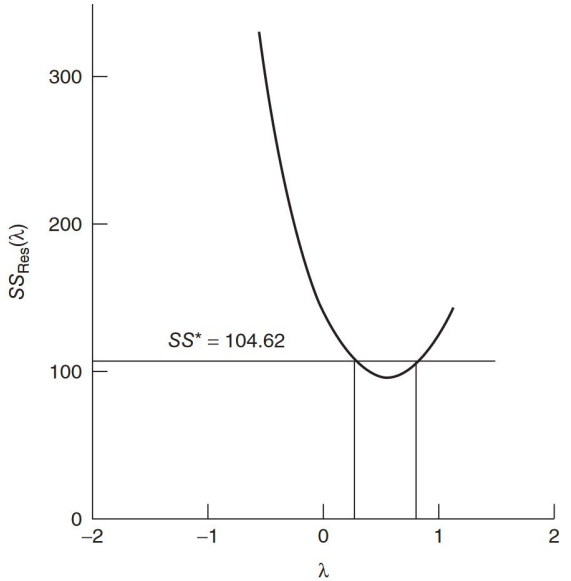

Figure 5.9 Plot of residual sum of squares $SS_{Res}(\lambda)$ versus λ.

Exemplo (Montgomery, Peck, Vining)

$$SS^* = SS_{Res}(\hat{\lambda})e^{\chi^2_{\alpha,1}/n}$$

$$\text{SS}^* = 96.9495e^{3.84/53}$$

$$\text{SS}^* = 96.9495(1.0751)$$

$$\text{SS}^* = 104.23$$

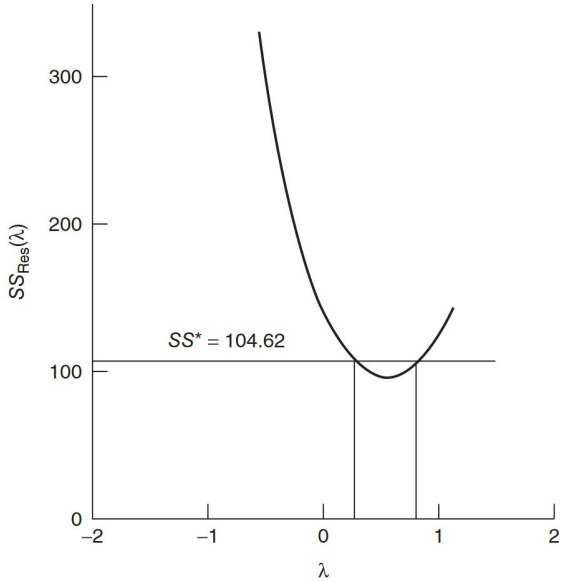

Figure 5.9 Plot of residual sum of squares $SS_{\text{Res}}(\lambda)$ versus λ.

Referências

A. L. Siqueira (1983) Uso de transformação em análise de variância e análise de regressão.

D. C. Montgomery, E. A. Peck, G. G. Vining (2012) Introduction to Linear Regression Analysis,

5th Edition.