Correlação & Regressão 2 Múltiplas

Prof. Lori Viali, Dr.
viali@mat.ufrgs.br
http://www.mat.ufrgs.br/~viali/

Testando a Significância Global da Regressão

Nem sempre se quer testar os coeficientes individuais da regressão. Pode ser necessário e é conveniente testar o modelo como um todo, isto é testar se:

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

Este caso pode ser tratado através da análise de variância (ANOVA).

O modelo de Regressão Múltipla Geral é dado por:

$$Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \ldots + \beta_k X_{ki} + U_i$$

Para testar a hipótese nula de que:

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

Isto é, todos os coeficientes são nulos, contra a alternativa de que nem todos são simultaneamente nulos, determina-se:

$$F = \frac{SQE/(k-1)}{SQR/(n-k)}$$

A expressão tem uma distribuição F com k - 1 e n - k graus de liberdade.

$$F = \frac{SQE/(k-1)}{SQR/(n-k)} = \frac{(n-k)SQE}{(k-1)SQR} =$$

$$= \frac{(n-k)SQE}{(k-1)(SQT - SQE)} =$$

$$= \frac{(n-k)(SQE/SQT)}{(k-1)[1-(SQE/SQT)]} =$$

$$= \frac{(n-k)R^2}{(k-1)(1-R^2)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Onde:

SSR (**S**oma dos **Q**uadrados dos **R**esíduos)

(RSS = Residual Sum of Squares)

$$VR = SQR = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

SQE (**S**oma dos **Q**uadrados **E**xplicados)

(ESS = Explained Sum of Squares)

$$VE = SQE = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

e:

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

G.L.	SQT	=	SQR	+	SQE
	n -1	=	(n - k - 1)	+	k

O resultado anterior mostra que F e R^2 variam diretamente. Assim se $R^2 = 0$, então F é zero. Quanto maior o valor de R^2 maior será o valor de F. Desta forma o teste F que é de ajuste do modelo também testa a significância do coeficiente de determinação.

Decidindo entre modelos competitivos

A decisão entre um modelo linear ou um modelo log-linear (o lagaritmo do regressor é uma função dos logaritmos dos regressores) é uma questão básica na análise empírica. Para testar:

H_0: Modelo Linear;

H_1: Modelo Log-Linear.

Pode-se utilizar o teste MWD.

O teste MWD foi proposto por MacKinnon, White e Davidson e envolve as seguintes etapas:

Estimar o modelo linear e determinar os valores $\hat{Y}$;

Estimar o modelo log-linear e obter os valores $\widehat{\ln Y}$;

Obtenha $Z_1 = \ln \hat{Y} - \ln Y$;

Fazer uma regressão de Y sobre os valores de X e Z otidos como acima. Rejeitar H_0 se o coeficiente de Z_1 for estatisticamente significativo através do teste t tradicional;

Obter $Z_2 = (\text{antilog } \ln Y - \hat{Y})$

Regredir o ln de Y sobre os logaritmos de Xs e Z_2. Rejeitar H_1 se o coeficiente de Z_2 for significativo pelo teste t.

Relaxando as Hipóteses do Modelo Clássico

O modelo clássico de Regressão Linear é baseado em um conjunto de hipóteses simplificadoras:

- É linear nos parâmetros;
- Os regressores X_i são fixos em amostragens repetidas;
- A expectância dos U_i é zero;
- A variância de U_i é constante e homocedástica.

- Se U_i não são autocorrelacionados;
- Se os X_i são aleatórios eles são independentes ou não-correlacionados com U_i;
- O número de observações (n) deve ser maior que o número de regressoes (k);
- Não há relação linear entre os regressoes, isto é, multicolinearidade;
- Os termos U_i são normais.

Três questões devem ser respondidas:

Qual o desvio mínimo em relação a uma hipótese, para que isto faça diferença?

Como verificar se uma hipótese foi, de fato, violada, numa situação específica?

Que correção adotar quando uma ou mais hipóteses não forem verdadeiras?

Multicolinearidade

O termo multicolinearidade foi cunhado por Ragnar Frisch na obra "Statistical Confluence Analysis by Means of Complete Regression Systems" do Instituto de Economia da Universidade de Oslo que foi publicada em 1934.

O termo significa a existência de uma relação "perfeita" linear entre algumas ou todas as variáveis explicativas do modelo.

Assim para uma regressão que envolva "k" variáveis explicativas: X_1, X_2, …, X_k, diremos que existe uma relação linear exata se:

$$\lambda_1 X_1 + \lambda_2 X_2 + \lambda_k X_k = 0$$

Onde λ_1, λ_2, …, λ_k são constantes não simultaneamente nulos.

A idéia de multicolinearidade inclui ainda:

$$\lambda_1 X_1 + \lambda_2 X_2 + \lambda_k X_k + V_i = 0$$

Onde o termo V_i é estocástico.

O termo multicolinar como definido inclui apenas relacionamento linear mas isto não exclui outras relações como por exemplo:

$$X_2 = X_1 . X_1$$

A existência da multicolinearidade pefeita torna os coeficientes da regressão indeterminados e seus erros padrão infinitamente grandes. Se a multicolinearidade não for alta (não perfeita) os coeficientes de regressão poderão ser determinados mas os erros padrão serão grandes.

Conseqüências da multicolinearidade

Se as hipóteses do modelo são satisfeitas os estimadores de MQO dos coeficientes da regressão são MELNV. Pode-se mostrar que mesmo que as variáveis sejam altamente colineares os MQO ainda mantém a propriedade MELNV. Assim as conseqüências práticas podem ser:

As estimativas apresentarem grandes variâncias e como resultante ter-se-á:

Intevalos de confiança maiores;

Alguns coeficientes podem ser não significativos;

O R^2 ainda ser alto, mesmo com coeficientes não significativos.

Percepção da multicolinearidade

Este é um fenômeno essencialmente amostral, conseqüência decorrente em boa parte de dados não-experimentais coletados na maioria das Ciências Sociais. A seguir algumas regras práticas para detectar sua presença:

- Um R^2 alto com poucos regressores significativos;
- Altas correlações dois a dois entre os regresssores;
- Índice de Condição (IC)

Índice de Condição

O número de condição "k" é definido como:

$$k = \frac{\text{Autovalor Máximo}}{\text{Autovalor Mínimo}}$$

O Índice de Condição (IC) é definido, então, como:

$$IC = \sqrt{\frac{\text{Autovalor Máximo}}{\text{Autovalor Mínimo}}} = \sqrt{k}$$

Pode-se adotar, então, a seguinte regra empírica. Se k estiver entre 100 e 1000 existe multicolinearidade de moderada a forte. Se estiver acima de 1000 a multicolinearidade é grave. Da mesma pode-se utilizar o IC. Se ele estiver entre 10 e 30 colinearidade moderada a forte e acima de 30 grave.

Heteroscedasticidade

Uma hipótese importante do modelo clássico de regressão linear é a de que a variância de cada termo residual (U_i) é constante e igual a σ^2.

Homo (igual) scedasticidade (dispersão) , ou

$$E(U_i^2) = \sigma^2 \quad i = 1, 2, \ldots, n$$

Alternativamente a homoscedasticidade pode ser expressa por:

$$\mathcal{V}(Y_i / X) = \sigma^2$$

A heteroscedasticidade é, então dada por:

$$\mathcal{V}(Y_i / X) = \sigma_i^2$$

Alguns causas da heteroscedasticidade podem ser:

- Situações de aprendizagem e erro;
- Aumento de renda com aumento da liberdade de escolha de como dispor a renda;
- Melhora nas técnicas de coleta de dados, menos erros, menor variabilidade;

A heteroscedasticidade é mais comum quando os dados são provenientes de cortes de séries temporais.

O que acontece com os estimadores dos MQO e com suas variâncias na presença de heteroscedasticidade?

Vamos supor o modelo de Regressão Linear Simples: $Y_i = \alpha + \beta X_i + U_i$ e que:

$$E(U_i^2) = \sigma_i^2$$

A inclinação da linha de regressão é dada por:

$$b = \frac{S_{XY}}{S_{XX}} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

Neste caso, a variância do estimador será dada por:

$$\mathcal{V}(b) = \frac{\sum (X_i - \bar{X})^2 \sigma_i^2}{[\sum (X_i - \bar{X})^2]^2}$$

Se $\sigma_i^2 = \sigma^2$, então a expressão acima ficará reduzida ao caso usual.

Neste caso o estimador MQO continua linear e não tendencioso, mas não será mais de variância mínima.

Ele não é eficiente, pois não leva em consideração a informação de que para cada x a variância de Y é diferente. Para obter um estimador eficiente é preciso fazer uso do método dos MQG.

MQG (Mínimos Quadrados Generalizados)

O MQO não leva em conta as diferentes variabilidades dos resíduos, conferindo a mesma importância para cada observação. O MQG leva em conta explicitamente tal informação e por isto é capaz de produzir estimadores eficientes na presença de heteroscedasticidade.

Detectando a Heterocedasticidade

Como saber se existe heteroscedasticidade nos dados? Não existe um método seguro com valores amostrais. Como, em geral, só existe um Y para cada X, dectetar a presença de heroscedasticidade não é simples.

A maioria dos métodos se baseia no exame dos resíduos.

Testes formais

- Teste de Park;
- Teste de Glejser;
- Teste de Spearman de correlação da ordem;
- Teste de Goldfeld-Quandt;
- Teste de Breusch-Pagan-Godfrey;
- Teste Geral de Heteroscedasticidade de White;

Medidas Corretivas

As medidas corretivas devem levar em conta as duas seguintes situações:

Quando as variabilidades resíduais forem conhecidas e

Quando elas não forem conhecidas.

Se as variabilidades residuais forem conhecidas então deve-se utilizar o Método dos Mínimos Quadrados Generalizados ou Ponderados, onde a ponderação é dada por:

$$w_i = \frac{1}{\sigma_i^2}$$

Se as variabilidades residuais não forem conhecidas pode-se adotar os seguintes procedimentos:

Variâncias e erros-padrão consistentes em heteroscedasticidade segundo White;

Hipóteses plausíveis a respeito do padrão de heteroscedasticidade;

Autocorrelação

Uma hipótese importante do modelo clássico de regressão linear é a de que não existe autocorrelação ou correlação serial entre os resíduos U_i.

No entanto, a correlação pode ocorrer, então deve-se responder:

- Qual a sua natureza?
- Quais as conseqüências teóricas e práticas?
- Como corrigir o problema quando ele ocorre?

A Natureza

O termo autocorrelação pode ser entendido como a "correlação entre os termos de observações no tempo" [séries temporais] ou "espaciais" [dados de corte].

No modelo clássico a suposição é de que:

$E(U_i U_j) = 0 \quad se \; i \neq j$

Isto é, um dado resíduo "i" não é influenciado por um outro dado resíduo "j".

Causas da Autocorrelação

- Inércia ou rigidez. Séries como PNB, Índices de Preços, Produção, Emprego e Desemprego são cíclicas;
- Viés de especificação: variáveis excluídas.
- Viés de especificação: forma funcional incorreta;
- Fenômeno da Teia de Aranha.

A oferta de produtos agrícolas reflete um fenômeno denominado de Teia de Aranha, em que a oferta reage ao preço como uma defasagem de um período de tempo, pois as decisões relativas à oferta levam um certo tempo para serem implementadas.

- *Defasagens. Em uma regressão de série temporal do consumo sobre a renda, não é raro verificar que o consumo no período corrente depende, entre outras coisas, do consumo no período anterior;*
- *Manipulações de dados. Dados trimestrais agregados de médias de dados mensais;*

Estimativas por MQO com Autocorrelação

O que ocorre com os estimadores de MQO se $E(U_iU_j) \neq 0$ *(para* $i \neq j$*) e as demais hipóteses forem mantidas?*

Neste caso os estimadores, a exemplo, do caso heteroscedástico, são ainda lineares e não tendeciosos.

No entanto sua variância será afetada. Neste caso eles não mais terão variância mínima, isto é, eles não serão eficientes.

Aqui, também, a exemplo da heteroscedasticidade pode-se encontrar um estimador que seja eficiente.

Para isto será necessário utilizar MQG – Mínimos Quadrados Generalizados, que incorpora qualquer informação adicional que tivermos através da transformação das variáveis.

Detectando a Autocorrelação

A autocorrelação é um problema potencialmente sério e medidas corretivas devem ser tomadas. Entretanto, inicialmente, é necessário, verificar se ela existe. Alguns testes para detectar a autocorrelação.

- *Método Gráfico. Representar graficamente os resíduos* (U_t) *e os resíduos padronizados* (U_t/s)*;*
- *Teste das carreiras ou de Geary.*
- *O teste d de Durbin-Watson*

Medidas Corretivas

Quando a estrutura da autocorrelação é conhecida utilizar a transformação de Prais-Winsten e a Equação de Diferença Generalizada ou de Quase-Diferença.

Quando o autocorrelação não é conhecida.

Embora simples de aplicar a regressão de diferença generalizada é geralmente difícil de rodar, pois, na prática, poucas vezes se conhece o valor de ρ. Por isto foram criados métodos alternativos.

Método da primeira diferença. Para aplicá-lo é necessário fazer o teste de Berenblutt-Webb de que $\rho = 1$.

O processo iterativo de Cochrane-Orcutt para estimar ρ.

O método de Cochrane-Orcutt em duas etapas. É uma versão abreviada do processo iterativo.

Método de Durbin em duas etapas para estimar ρ.

PARK, R. E. Estimation with Heteroscedastic Error Terms. Econometrica. v. 34, n. 34, Out de 1966. p. 888.

GLEJSER, H. A New Test for Heteroscedasticity. Journal of the American Statistical Association. v. 64, 1969. p. 316-23.

GOLDFELD, S. M., QUANDT, R. E. Nonlinear Methods of Econometrics. Amesterdã: North-Holland, 1972.

BREUSCH, T., PAGAN, A. A Simple Test for Heteroscedasticity and Random Coefficient Variation. Econometrica. v. 47, 1979. p. 1287-94.

GODFREY, L. Testing for Multiplicative Heteroscedasticity. Jornal of Econometrics. v. 8, 1978. p. 227-36.

WHITE, H. A Heteroscedasticity Consistent Covariance Matrix Estimator and a Direct Test of Heteroscedasticity. Econometrica. v. 48, 1980. p. 817-18.

GEARY, R. C. Relative Efficiency of Count of Sign Changes for Assessing Residual Autoregression in Least Squares Regression. Biometrika, v. 57, 1970. P. 123-27.

DURBIN, J., WATSON, G. S. Testing for Serial Correlation in Least-Squares Regression. Biometrika. v. 38, 1951. p. 159-71.

BERENBLUTT, I. I., WEBB, G. I. A New Test for Autocorrelated Errors in the Linear Regression Model. Journal of the Royal Statistical Society. Série B, v. 35, n. 1, 1973. P. 33-50.

COCHRANE, D. ORCUTT, G. H. Application of Least Squares Regressions to Relationships Containing Autocorrelated Error Terms. Journal of the Royal Statistical Society. v. 44, 1949. P. 32-61.

DURBIN, J. Estimation of Parameters in Time-Series Regression Models. Journal of the Royal Statistical Society. Série B. v. 22, 1960. p. 139-153.