Regressão Múltipla – Tópicos Adicionais

HO 231 – Econometria

Profa. Rosangela Ballini

Instituto de Economia - UNICAMP

Ementa

- Previsão de *Y* com Regressor em Logaritmo
- Comparação de Modelos (Teste de Wald e R^2 ajustado)
- Critério de Informação AIC
- Multicolinearidade

Bibliografia

Wooldridge, J. M. 2003. *Introductory Econometrics*. Caps. 3-6.

Modelos Logarítmicos - Exemplo

- A partir do arquivo Dados_CO2.xlsx estime:

1. $CO2 = \beta_0 + \beta_1 PIB + \beta_2 Setor2 + u$
2. $\ln(CO2) = \beta_0 + \beta_1 PIB + \beta_2 Setor2 + u$
3. $\ln(CO2) = \beta_0 + \beta_1 \ln(PIB) + \beta_2 Setor2 + u$

Modelos Logarítmicos - Exemplo

- Modelos com mesmo regressor podem ser comparados com o R^2;

Variable	linear	loglin	loglog
pib	.00032095	.00009541	
	12.27	10.30	
	0.0000	0.0000	
setor2	.10508293	.04189997	.02460615
	5.29	5.96	5.26
	0.0000	0.0000	0.0000
lnpib			.82701542
			21.54
			0.0000
_cons	-1.1262519	-1.3743163	-6.5469295
	-1.68	-5.79	-21.14
	0.0948	0.0000	0.0000
F	87.854872	69.502992	270.07999
r2	.51420759	.45574838	.76492579

legend: b/t/p

Variações absolutas constantes: para cada variação de 1 US$ no PIB per capita, espera-se um acréscimo de 0,3 kg nas emissões de CO^2;

Variações absolutas em X e relativas em Y: Para cada variação de 1 US$ no PIB per capita, espera-se um acréscimo de 0,009% nas emissões de CO^2;

Variações relativas em *X* e *Y*: Para cada variação de 1% no PIB, espera-se uma variação de 0,83% nas emissões de CO^2.

O R2 do modelo linear com os dos demais modelos não são diretamente comparáveis. Entretanto, a expressiva diferença entre o R2 do modelo log-log em relação aos demais, sugere a maior qualidade deste ajuste. Outras estatísticas podem, entretanto, auxiliar na decisão...

Normalidade dos Erros - Histograma

- O objeto lm contém uma lista de componentes, entre eles **residuals**;

```
> res=loglog$residuals
```

res é uma variável que recebe os valores dos resíduos do modelo loglog ajustado.

- O histograma permite visualizar a forma de distribuição observada dos dados e compará-la com uma distribuição teórica;
- O comando **hist.()** elabora um histograma da variável desejada.
- A opção **density** adiciona a função densidade à frequência observada;

```
> hist(res, freq=F)
> par(new=TRUE)
> plot(density(res),main='',xlab='',ylab='')
```

Teste de Normalidade dos Erros

$$\begin{cases} H_0: \text{ Erros normais} \\ H_1: \text{ Erros não normais} \end{cases}$$

Rejeitar H_0 significa encontrar evidências que os erros não estão normalmente distribuídos;

- Entre os diversos testes de normalidade, o teste de Shapiro-Wilk apresenta um bom poder (capacidade de rejeitar H_0 em populações não normais);

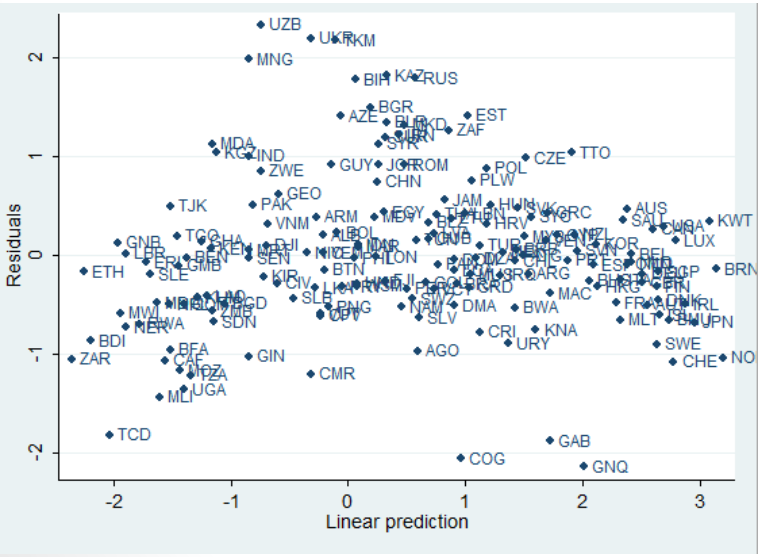
```
> shapiro.test(res)

Shapiro-Wilk normality test data: res
        W = 0.97117, p-value = 0.001372
```

Há evidências para rejeitar H_0, ou seja, para afirmar que os erros do modelo não estejam normalmente distribuídos.
Assim, a validade das estatísticas *t* e *F* estará condicionada às propriedades assintóticas dos estimadores (para grandes amostras). Em outras palavras, as estatísticas *t* e *F* não seriam apropriadas para amostras pequenas.

Análise Exploratória – *scatter*

- A dispersão dos resíduos em torno dos valores previstos permite, por exemplo, identificar eventuais valores extremos ou falhas de especificação;

O comando **fitted.values**, fornece os valores previstos na variável *ln(co2)*.



A dispersão dos resíduos em torno dos valores previstos sugere um possível comportamente quadrático do regressando em função dos regressores (provavelmente PIB per capita).

Há ainda valores extremos, como Congo (COG) que possui um valor de lnco2 muito inferior ao esperado, e Uzbesquitão (UZB), com um valor muito superior ao esperado.

Previsão de *Y* em Equações (Semi-)Logarítmicas

- Seja a relação semi-logarítmica: $\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$
  - Os valores previstos de ln(*Y*) serão:

$$\widehat{\ln(Y)} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

- A simples previsão de *Y* pelo antilogaritmo de ln(*Y*) será viesada:

$$\hat{Y} \neq e^{\widehat{\ln(Y)}}$$

- Se o modelo satisfaz as hipóteses RLM1 a RLM6, tem-se:

$$E(Y|X_1, X_2) = \exp\left(\frac{\sigma^2}{2}\right).\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

- Sendo a previsão dada por:

$$\hat{Y} = \exp\left(\frac{\hat{\sigma}^2}{2}\right)\exp(\widehat{\ln(Y)})$$

A qual é uma estimativa consistente. Porém, depende da normalidade do termo erro.

- Se o termo erro é independente das variáveis explicativas, tem-se:

$$E(Y|X_1, X_2) = \alpha_0 . \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

em que α_0 é o valor esperado de *exp(u)*.

- Dessa forma, o valor previsto de y é dada por:

$$\hat{Y} = \hat{\alpha}_0 \exp(\widehat{\ln(Y)})$$

- Duas formas para estimar α_0:

1. Estimador do método dos momentos:

$$\hat{\alpha}_0 = \frac{\sum \exp(\hat{u})}{n}$$

2. Ou usando as estimativas de MQO:

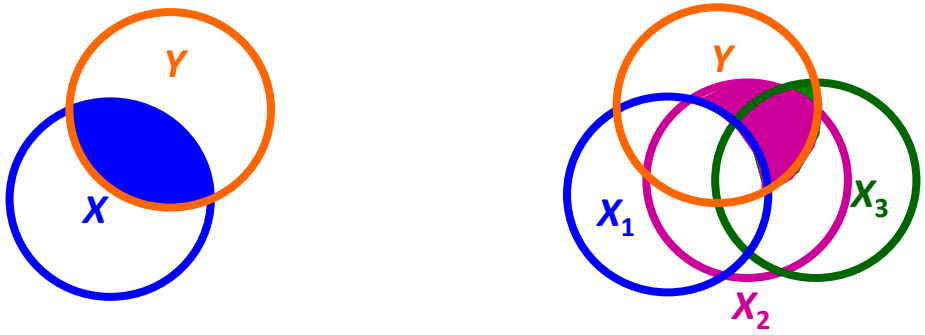
$$\check{\alpha}_0 = \left(\sum \hat{m}_i^2 \right)^{-1} \left(\sum \hat{m}_i y_i \right)$$

em que $\hat{m}_i = \exp(\widehat{\ln(y)})$

Comparação de Modelos *Nested*

- Sejam dois modelos, um deles contendo um sub-conjunto de regressores do outro (*nested models*):

$$Y = \alpha + \beta X_1 + e \qquad Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$



Pergunta: Podemos de fato afirmar que as variáveis adicionais X_2 e X_3 acrescentem informação relevante para explicar *Y*?

Teste de Wald:

- Teste usado para verificar se um conjunto de q regressores contribui para explicar a variável dependente:

$$\begin{cases} H_0 : \beta_{j+1} = .. = \beta_{j+q} = 0 \\ H_1 : \beta_{j+s} \neq 0 \end{cases}$$

Teste F: incorpora a restrição no procedimento de estimativa dos parâmetros.

Esta questão é verificada por meio do teste F :

$$F = \frac{\frac{(SQRes_R - SQRes_U)}{(g_R - g_U)}}{SQRes_U / g_U}$$

Sendo, g_R graus de liberdade do modelo restrito e g_U graus de liberdade irrestrito.

- Testes para grandes amostras: estatística multipliador de Lagrange

- LM: exige apenas a estimação do modelo restrito

- Etapas para obtermos LM:

1. Regrida *Y* sobre o conjunto restrito de *q* variáveis e salve os resíduos $\tilde{u}$
2. Regrida $\tilde{u}$ sobre todas as variáveis independents e obtenha o R_u^2
3. Calcule $LM = n * R_u^2$
4. Para *q* graus de liberdade, a estatística LM tem distribuição χ_q^2, obtenha o p-valor. Se p-valor menor que um determinado nível de significância rejeita-se H_0

Modelos *Nested* - Exemplo

- Sejam dois nested-models para o logaritmo das emissões:

$$\ln(co2) = \alpha + \beta setor2 + e \qquad \ln(co2) = \alpha + \beta_1 setor2 + \beta_2 pib + \beta_3 pib^2 + e$$

- Para sabermos se as variáveis *pib* e *pib*2 acrescentam informações relevantes à previsão das emissões, devemos testar:

$$\begin{cases} H_0 : \beta_2 = \beta_3 = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

- Teste de Wald no R: função **linearHypothesis**:

```
> linearHypothesis(unrestr,c("pib=0","pib2=0"),test="F")
```

```
> linearHypothesis(unrestr,c("pib=0","pib2=0"),test="Chisq")
```

Testa a hipótese que os coeficientes associados às variáveis *pib* e *pib*2 são, simultaneamente, iguais a zero.

A probabilidade de erro ao afirmarmos que as variáveis acrescentam informação relevante é praticamente nula.

Comparação de Modelos *Non-Nested*

- Sejam dois modelos, contendo conjuntos distintos de regressores (modelos *non-nested*):

$$Y = \alpha + \beta X_1 + e \qquad Y = \alpha + \beta_2 X_2 + \beta_3 X_3 + e$$

Pergunta: Qual é o melhor modelo para explicar o comportamento de *Y*?

Coeficiente de Determinação Ajustado ($\bar{R}^2$):

$$\bar{R}^2 = 1 - \frac{SQRes/[n-(k+1)]}{STQ/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-(k+1)}$$

O R^2 ajustado pondera o coeficiente de determinação (R^2) pelo número de variáveis explicativas e pelo número de observações da amostra.
Particularmente útil quando compara-se modelos de regressão múltipla que prevêem a mesma variável dependente, pois penaliza aquele modelo com maior número de variáveis independentes.

Modelos *Non-Nested* - Exemplo

- Sejam dois non-nested models para o logaritmo das emissões:

$$\ln(co2) = \alpha + \beta \ln(pib) + e \qquad \ln(co2) = \alpha + \beta_2 pib + \beta_3 pib^2 + e$$

- O R^2 ajustado é apresentado juntamente às estimativas da tabela ANOVA:

Observations	169	169
R2	0.72571	0.47437
Adjusted R2	0.72407	0.46804
F Statistic	441.85050***	74.90614***
	(df = 1; 167)	(df = 2; 166)

O primeiro modelo tem um maior poder de explicação da variabilidade do *lnco2*. Após ponderarmos pelo número de variáveis de cada modelo, o poder de explicação do primeiro modelo é de 72,4% e o do segundo modelo é de 46,8%.

Critério de Informação Akaike

- Critério de Informacão de Akaike (AIC) é conhecido como uma estatística de penalização de log-verossimilhança
- AIC é obtido por:

$$AIC = -2loglikelihood + 2 * (p + 1)$$

Em que p é o número de parâmetros e $loglikelihood$ é dado por:

$$loglikelihood = -\frac{n}{2}\log(2\pi) - nlog(\sigma) - \sum \frac{(y-\mu)^2}{2\sigma^2}$$

Supondo uma distribuição normal para os resíduos.

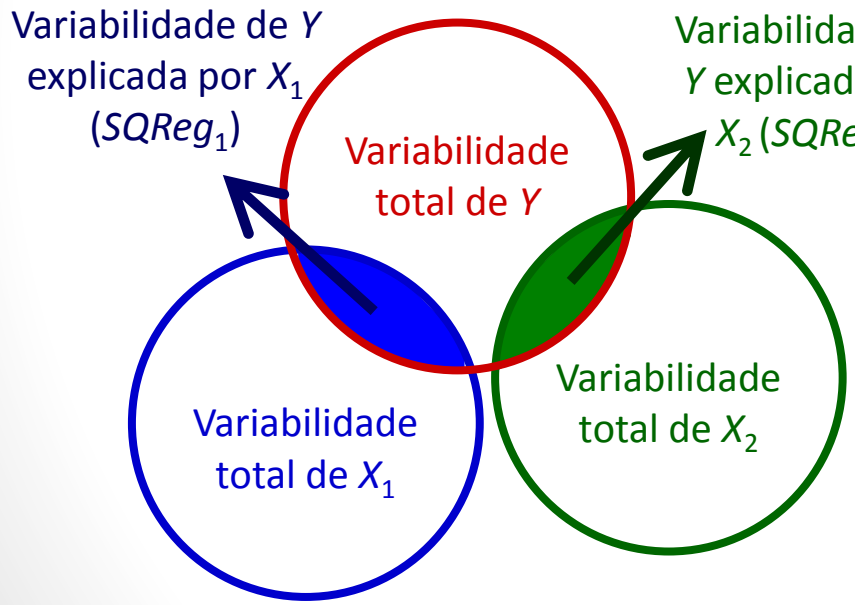
- Escolhe-se o modelo com menor AIC

Multicolinearidade - Conceito

Seja o modelo definido por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i$$

Se X_1 e X_2 são independentes, a variabilidade de Y explicada pelo modelo de RLM divide-se em duas partes disjuntas:
o efeito isolado de X_1 ($SQReg_1$) e o efeito isolado de X_2 ($SQReg_2$).



$$Y = \alpha_0 + \beta_1 X_1 + u_{1i} \Rightarrow SQ\,\mathrm{Re}\,g_1$$

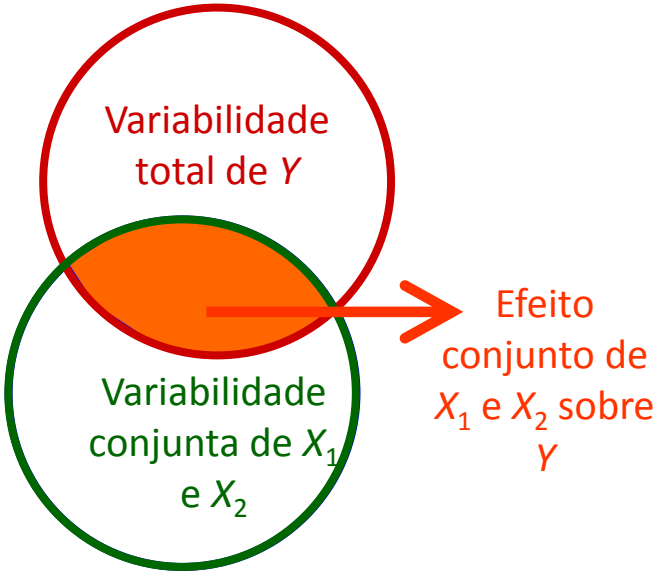
$$Y = \alpha_1 + \beta_2 X_2 + u_{2i} \Rightarrow SQ\,\mathrm{Re}\,g_2$$

Tem-se os mesmos coeficientes angulares das regressões simples

Multicolinearidade - Conceito

No outro extremo, caso tenha uma relação linear exata entre X_1 e X_2 (perfeita colinearidade), ou seja:

$$X_{1_i} = \lambda_2 X_{2_i}$$



Nesta situação, seria impossível estimar os efeitos isolados de X_1 e X_2 sobre Y.

Multicolinearidade - Conceito

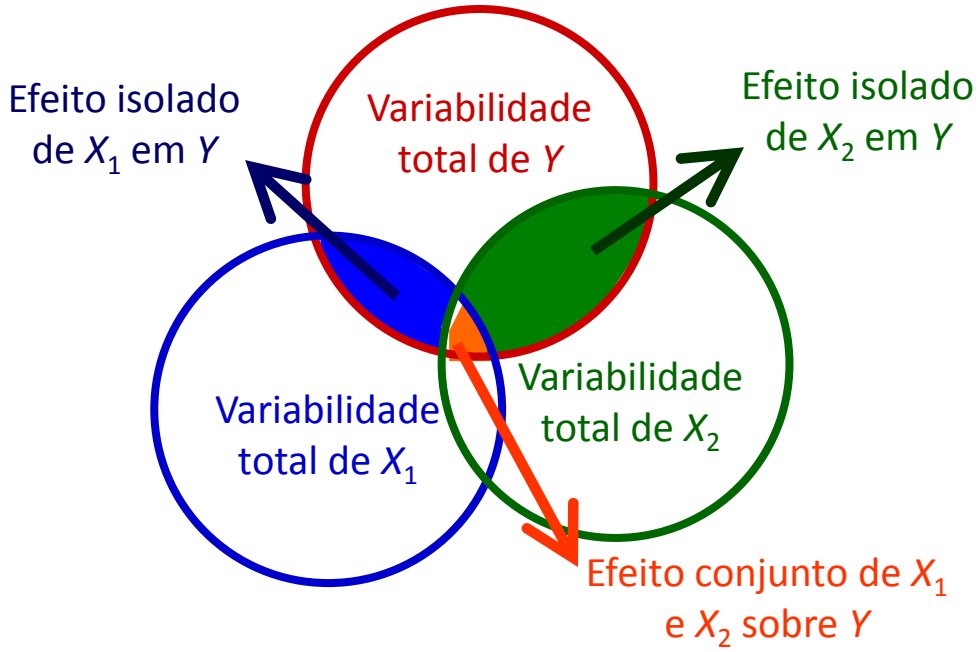
Frequentemente observa-se relação entre as variáveis independentes X_1 e X_2

Nesses casos, o efeito sobre Y poderá ser dividido em 3 partes:

i) efeito isolado de X_1;

ii) isolado de X_2; e

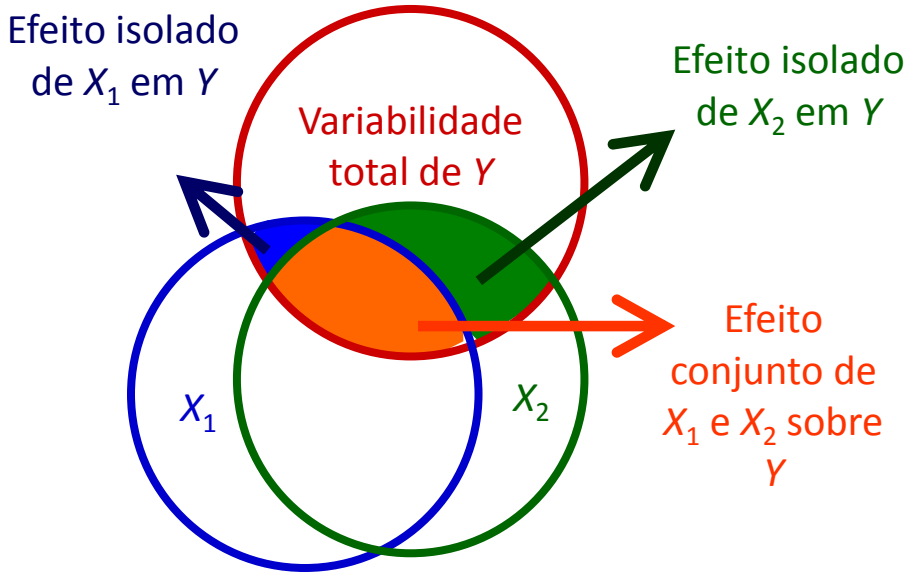
iii) efeito conjunto de X_1 e X_2.



Multicolinearidade - Conceito

Questão: caso haja uma forte relação linear entre X_1 e X_2 (*multicolinearidade*) pode comprometer a identificação dos efeitos isolados de X_1 e X_2 sobre Y.

Ou seja, a maior parcela da variabilidade de Y é explicada pelo efeito conjunto de X_1 e X_2.



Algebricamente, essa relação de multicolinearidade seria dada por:

$$X_{1_i} = \lambda_2 X_{2_i} + v_i$$

Multicolinearidade - Definição

Colinearidade Perfeita

Duas variáveis são ditas perfeitamente colineares quando há uma relação linear exata entre essas. De maneira genérica, representa-se a linearidade perfeita por:

$$X_{j_i} = \lambda_1 X_{1_i} + \lambda_2 X_{2_i} + \ldots + \lambda_k X_{k_i}$$

Multicolinearidade

Há multicolinearidade em um modelo de regressão múltipla quando duas ou mais variáveis independentes são fortemente relacionadas linearmente entre si. Nesse caso, tem-se:

$$X_{j_i} = \lambda_1 X_{1_i} + \lambda_2 X_{2_i} + \ldots + \lambda_k X_{k_i} + v_i$$

Multicolinearidade - Definição

Conseqüências da Multicolinearidade

- Existência de uma colinearidade exata entre duas ou mais variáveis independentes torna impossível a obtenção dos coeficientes dos parâmetros por MQO.

- Presença de multicolinearidade os estimadores de MQO continuam sendo os MELNV.

- Multicolinearidade torna muitas vezes as estimativas dos coeficientes dos parâmetros (β's) insignificantes, já que cada um pressupõe, por definição, a variação em Y dada uma variação unitária em X, mantendo-se constantes as demais informações. Ou seja, se duas variáveis independentes são fortemente correlacionadas, tornár-se-á muito difícil haver variação em uma sem que haja em outra.

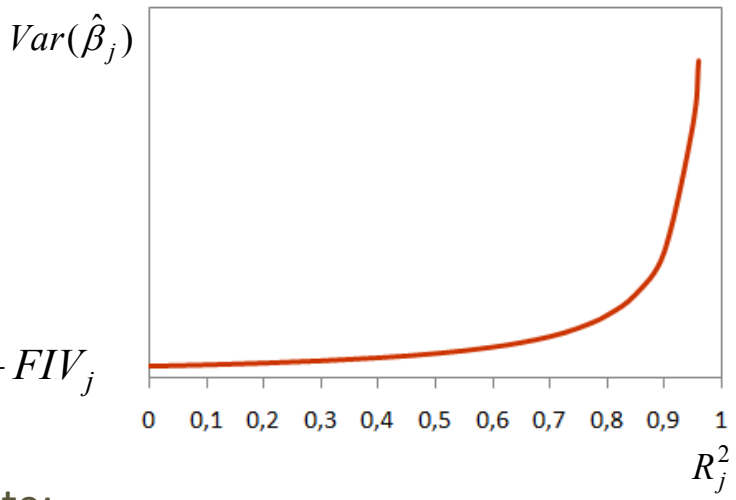
Fator Inflacionário da Variância

Variância dos estimadores será:

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

Alternativamente, as estimativas das variâncias individuais podem ser obtidas por:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{j_i}^2 (1-R_j^2)} = \frac{\sigma^2}{\sum_{i=1}^{n} x_{j_i}^2} \frac{1}{(1-R_j^2)} = \frac{\sigma^2}{\sum_{i=1}^{n} x_{j_i}^2} FIV_j$$



Em que R^2_j é o coeficiente de determinação do ajuste:

$$X_{j_i} = \lambda_0 + \lambda_1 X_{1_i} + \ldots + \lambda_{k-1} X_{k_i} + v_i$$

Fator Inflacionário da Variância - FIV

Representa o quanto a variância de $\hat{\beta}_j$ está sendo inflacionada pela relação de multcolinearidade entre X_j e as demais variáveis independentes. Quando não houver relação entre as variáveis independentes (R^2_j =0) o FIV_j será igual a 1 e, à medida que aproxima-se de uma relação exata (R^2_j =1), o FIV_j tenderá a infinito.

Multicolinearidade - Identificação

Conflito entre estatísticas R^2 e F do modelo e testes t para os parâmetros β: as estatísticas R^2 e F podem indicar um modelo significativo, enquanto os testes t dos parâmetros β's seriam insignificantes.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_i \Rightarrow H_0 : \beta_1 = \beta_2 = 0 \quad \mathbf{X} \quad H_0 : \beta_1 = 0 \text{ e } H_0 : \beta_2 = 0$$

Ajuste linear significativo entre as variáveis independentes: uma das variáveis independentes apresentaria forte relação de linearidade com as demais.

$$X_{j_i} = \lambda_0 + \lambda_1 X_{1_i} + \ldots + \lambda_{k-1} X_{k_i} + v_i \Rightarrow R_j^2$$

Fator Inflacionário da Variância: uma consequência de uma relação linear forte entre um dos regressores e os demais será um elevado valor para o respectivo FIV.

$$R_j^2 \Rightarrow FIV_j$$

Multicolinearidade - Exemplo

Seja o modelo para a relação entre emissões de CO^2, PIB e população:

$$\ln(co2) = \beta_0 + \beta_1 + \ln(pib) + \beta_2 \ln(pop) + u$$

A partir dos valores de 2010 da base Dados_Co2PibPop.xlsx, ajuste o modelo e analise os resultados.

Comando para o cálculo do fator inflacionário:

```
> vif(modelo)
```

Multicolinearidade - Correção

Medidas paliativas na presenção de multicolinearidade:

Aumentar o tamanho da amostra: aumentando o tamanho da amostra aumenta-se a variabilidade de X_j e, conseqüentemente, reduz a variância do estimador de β_j.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} x_{j_i}^2} FIV_j$$

Tranformar as variáveis: a multicolinearidade pode ser eliminada a partir de funções das variáveis independentes.

$$Y = \delta_0 + \delta_0 Z_i + u_i \qquad \text{onde} \qquad Z_i = f(X_{1_i}, X_{2_i})$$

Omitir regressor que apresentar alta colinearidade com os demais: uma solução simples, mas perigora, é excluir uma ou mais variáveis que apresentam multicolinearidade. A exclusão de variáveis essenciais para compreensão do problema pode, entretanto, gerar **viés de especificação**.

Correção – Exemplo de Transformação

Nova espeficação para contornar eventuais impactos da multicolinearidade:

$$\ln(co2) = \beta_0 + \beta_1 + \ln(pib) + \beta_2 \ln(pop) + u$$

$$\ln\left(\frac{co2}{pop}\right) = \delta_0 + \delta_1 \ln\left(\frac{pib}{pop}\right) + v$$

Correção – Exemplo de Eliminação

Podemos comparar os ajustes com e sem a inclusão de regressores:

$$\ln(co2) = \beta_0 + \beta_1 + \ln(pib) + \beta_2 \ln(pop) + u$$

$$\ln(co2) = \tilde{\beta}_0 + \tilde{\beta}_1 + \ln(pib) + u$$

A exclusão do regressor *lnpop* no segundo modelo reduz o erro padrão do estimador do coeficiente associado a *lnpib?*. Ou seja, a estimativa de β_1 é mais eficiente?

Sua exclusão tende a causar viés na estimativa de $\beta_{1,}$ uma vez que *lnpib* e *lnpop* estão correlacionados?

Exercícios

1) A partir do arquivo *wage2.dta*:
   a) Compare a qualidade do ajuste de um modelo linear e outro log-lin para a renda (*wage*) como função linear dos anos de educação (*educ*), quoeficiente de inteligência (*IQ*) e experiência profissional (*exper*);
   b) Compare a normalidade dos resíduos nos dois ajustes;
   c) Realize previsões para a renda (*wage*) a partir do modelo log-lin;
   d) Analise a contribuição de um termo quadrático para a experiência profissional;
   e) Analise a multicolinearidade entre os regressores e seus potenciais impactos sobre as estimativas;
   f) Compare os impactos da exclusão dos regressores *tenure* e *age* na variância e viés do coeficiente associado a *exper*.

Exercícios

2) A partir do arquivo Dados_PIB.XLS, que contém informações do WDB sobre sobre gastos com P&D, PIB per capita, razão de dependência para jovens e gastos com educação dos países em 2010, pede-se:
   a) Compare a qualidade dos ajustes linear, log-lin e log-log;
   b) Compare a normalidade dos resíduos de cada ajuste;
   c) Realize previsões para o PIB per capita nas funções logarítmicas;
   d) Avalie a necessidade de um termo quadrático para a razão de dependência;
   e) Analise a multicolinearidade entre razão de dependência e gastos com educação. Quais seriam seus eventuais impactos sobre as estimativas?
   f) Compare a variância e estimativas dos coeficientes para ajustes com e sem o regressor razão de dependência.