# CAPÍTULO 9 – MODELOS DE REGRESSÃO COM VARIÁVEIS BINÁRIAS

## 1-OBJETIVOS

Considerar modelos em que uma ou mais variáveis explicativas são variáveis nominais (também chamadas de indicadores, variáveis qualitativas, variáveis binárias ou variáveis "dummy"). O caso mais simples é quando fazemos a variável igual a 1 para uma categoria e 0 para a categoria mutuamente exclusiva à primeira. Por exemplo, podemos definir SEXO = 1 se feminino, e 0 se masculino.

Os modelos de regressão que contêm apenas variáveis binárias ou qualitativas são chamados de modelos de Análise de Variância (modelos ANOVA).

## 2- CUIDADOS NO USO DE MODELOS COM VARIÁVEIS QUALITATIVAS

Suponha que desejamos inserir no modelo uma variável qualitativa com m categorias. É importante notar que o modelo deverá ser especificado da seguinte forma:

1) Modelo com termo constante e (m-1) variáveis dummy
2) Modelo SEM termo constante e m variáveis dummy

Por que? Poderia parecer natural, à primeira vista, escrever o modelo como:

$$Y_i = \alpha + \beta_1.D_{1i} + \beta_2.D_{2i} + ... + \beta_m.D_{mi} + \varepsilon_i \qquad (1)$$

onde $D_j$ = 1 ou 0 se a observação pertence à j-ésima categoria da variável X, para j = 1, 2,..., n.

Mas, qual a matriz de design X para o modelo da equação (1) acima? VERIFIQUE que a primeira coluna de X é composta apenas de 1's. A segunda coluna contém 1's e 0's, assim como todas as demais colunas. Mas, a soma das colunas 2 até m+1 é igual à 1ª. coluna, pois somando todas as variáveis dummy em todos os seus níveis encontramos uma coluna de 1's.

Logo, o modelo representado por (1) NÃO PODE ser ajustado, pois sua matriz de design exibe colinearidade perfeita. As alternativas são as indicadas em 1) e 2) acima.

Como funciona o modelo 1)?

Escolhe-se uma categoria como categoria "base". Apenas a título de exemplo, suponha que ela é a m-ésima categoria da nossa variável qualitativa. Então o modelo a ser ajustado terá constante e (m-1) variáveis dummy, cada uma correspondendo às categorias 1, 2, ..., m-1 respectivamente.

Ou seja, a equação do modelo é:

$$Y_i = \alpha + \beta_1.D_{1i} + \beta_2.D_{2i} + ... + \beta_{m-1}.D_{(m-1)i} + \varepsilon_i \quad (2)$$

Suponha que, para uma dada observação estamos na 1ª. categoria da variável qualitativa e assim $D_{1i} = 0$ e todas as outras variáveis dummy são zero. Então o valor ajustado nesta observação será:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1.(1) = \hat{\alpha} + \hat{\beta}_1 \quad (3)$$

Suponha agora que uma observação corresponde à m-ésima categoria (que é a **categoria base**). Então, os valores de TODAS as variáveis dummy nesta observação são zero, e o valor ajustado é:

$$\hat{Y}_i = \hat{\alpha} \quad (4)$$

A partir de (3) e (4) fica fácil interpretar o significado dos coeficientes na regressão (2). $\alpha$ indica o valor médio na categoria omitida (neste caso a m-ésima). Cada $\beta_i$ indica a diferença entre o valor da i-ésima categoria e a média da categoria omitida.

Modelo 2)

A especificação do modelo 2) é:

$$Y_i = \beta_1.D_{1i} + \beta_2.D_{2i} + ... + \beta_m.D_{(m)i} + \varepsilon_i \quad (5)$$

Note que:

- Na equação (5) não existe termo constante;
- Agora existem variáveis dummy para TODAS as m categorias.

Qual a interpretação dos coeficientes na equação (5)? Cada $\beta_i$ indica o valor médio da respectiva categoria.

**Exemplo 9.1.**

Neste exemplo a variável dependente é o percentual de votos nulos e brancos (soma dos dois percentuais) no 1º. Turno das eleições municipais para prefeito no município do Rio de Janeiro em 2008. A variável explicativa é a região da cidade em que está situada a seção eleitoral, dividida em 5 categorias: Centro, Sul, Norte, Oeste, Subúrbio.

Existem 10702 observações na amostra, cada uma corresponde a uma seção eleitoral, ou seja, uma urna de votação.

Inicialmente ajustamos o modelo:

$$Y_i = \alpha + \beta_1.CENTRO_i + \beta_2.NORTE_i + \beta_3.SUL_i + \beta_4.OESTE_i + \varepsilon_i$$

Ou seja, a categoria base é "SUBURBIO". Os resultados desta regressão estão abaixo:

**Coefficients(a)**

| Model | | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|---|
| | | B | Std. Error | B | Std. Error |
| | (Constant) | 13,405 | ,037 | 360,889 | ,000 |
| | indicador centro da cidade | ,696 | ,125 | 5,565 | ,000 |
| | indicador zona norte | -2,423 | ,096 | -25,262 | ,000 |
| | indicador zona sul | -3,529 | ,073 | -48,395 | ,000 |
| | indicador zona oeste | -,190 | ,055 | -3,479 | ,001 |

a Dependent Variable: soma dos percentuais de nulos e brancos

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,472(a) | ,223 | ,222 | 2,43974 |

a Predictors: (Constant), indicador zona oeste, indicador centro da cidade, indicador zona norte, indicador zona sul

**ANOVA(b)**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 18244,922 | 4 | 4561,230 | 766,295 | ,000(a) |
| | Residual | 63671,940 | 10697 | 5,952 | | |
| | Total | 81916,862 | 10701 | | | |

a Predictors: (Constant), indicador zona oeste, indicador centro da cidade, indicador zona norte, indicador zona sul
b Dependent Variable: soma dos percentuais de nulos e brancos

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| perc_nulos_brancos | 12,6959 | 2,76678 | 10702 |

Que estória você pode contar a partir destes dados?

- O modelo é, em geral, altamente significante (veja a estatística F), ou seja, os percentuais de nulos + brancos variam de acordo com a região da cidade. Apesar disso, o $R^2$ do modelo é baixo. Que tal investigar a relação entre a estatística F e o $R^2$ e descobrir por que?

- No SUBURBIO (categoria base, variável dummy omitida), o percentual médio de votos brancos + nulos é 13,405%, que é superior à média geral do município (12,696%).
- No CENTRO, o percentual médio de brancos + nulos é 13,405 + 0,696 = 14, 101%.
- Na zona NORTE, o percentual médio de brancos + nulos é 13,405 - 2,423= 10,982%, o SEGUNDO MENOR de todas as regiões da cidade.
- Na zona SUL, o percentual médio de brancos + nulos é 13,405 - 3,529= 9,876%, o MENOR de todas as regiões da cidade.
- Na zona OESTE, o percentual médio de brancos + nulos é 13,405 - 0,190= 13,215%, superior à média geral do município.

Para comparação ajustamos o modelo que emprega todas as categorias da variável "região da cidade" e não contém termo constante.

$$Y_i = \beta_1.CENTRO_i + \beta_2.NORTE_i + \beta_3.SUL_i + \beta_4.OESTE_i + \beta_5.SUBURBIO_i + \varepsilon_i$$

Os resultados seguem abaixo.

**ANOVA(c,d)**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 1743244,336 | 5 | 348648,867 | 58573,634 | ,000(a) |
| | Residual | 63671,940 | 10697 | 5,952 | | |
| | Total | 1806916,276(b) | 10702 | | | |

a Predictors: indicador suburbio, indicador zona oeste, indicador zona sul, indicador zona norte, indicador centro da cidade
b This total sum of squares is not corrected for the constant because the constant is zero for regression through the origin.
c Dependent Variable: soma dos percentuais de nulos e brancos
d Linear Regression through the Origin

**Coefficients(a,b)**

| Model | | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|---|
| | | **B** | Std. Error | B | Std. Error |
| | indicador centro da cidade | **14,102** | ,119 | 118,030 | ,000 |
| | indicador zona norte | **10,982** | ,088 | 124,174 | ,000 |
| | indicador zona sul | **9,877** | ,063 | 157,413 | ,000 |
| | indicador zona oeste | **13,215** | ,040 | 329,390 | ,000 |
| | indicador suburbio | **13,405** | ,037 | 360,889 | ,000 |

**a Dependent Variable: soma dos percentuais de nulos e brancos**
**b Linear Regression through the Origin**

Note que os coeficientes da regressão estimada por este modelo são exatamente os mesmos obtidos no modelo anterior.

**Que modelo usar?**

No fundo é uma questão de gosto...

A maioria dos pesquisadores tende a usar o modelo representado pela equação (2), que inclui o termo constante. Isso acontece pois nesta forma é possível verificar facilmente se a categorização faz diferença (em relação à categoria base).

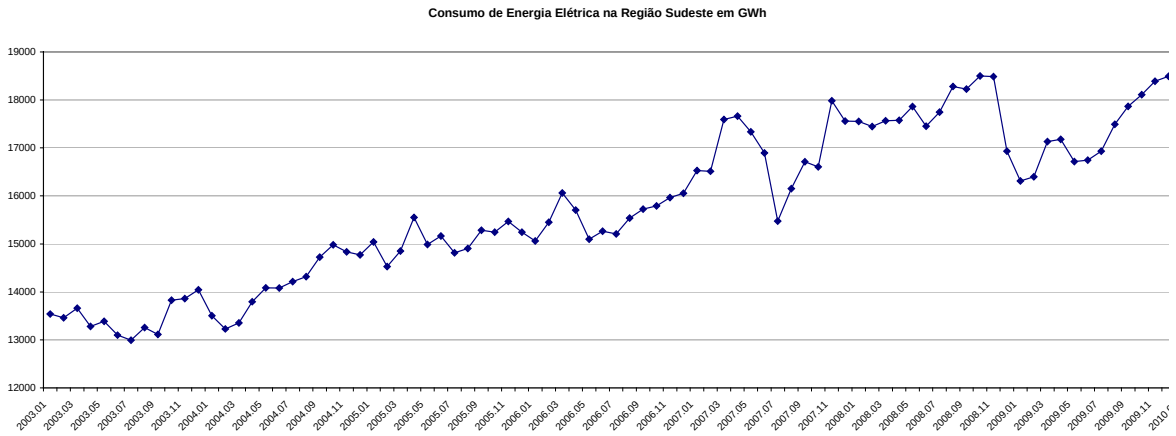
**Nota**

É possível combinar no mesmo nodelo mais de uma variável qualitativa. O único ponto a considerar é que cada variável qualitativa deve ser expressa com o seu número de categorias menos UM. Também, neste caso, o termo constante indicará a média quando a observação estiver nos níveis das categorias base para as duas variáveis qualitativas.

**Exemplo 9.2.**

Uso de variáveis "dummy" para representar a sazonalidade de uma série temporal.

Considere a série de consumo mensal de energia elétrica na região Sudeste a partir de janeiro de 2003 mostrada no gráfico a seguir.




Vamos calcular os fatores sazonais mensais para esta série usando o primeiro método descrito neste capítulo. As variáveis INDIC_01 a INDIC_12 são, respectivamente, as "dummies" para os meses de Janeiro a Dezembro.

Método 1 – modelo com constante – categoria omitida = janeiro

O modelo ajustado é:

**Coefficients(a)**

| Model | | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|---|
| | | B | Std. Error | B | Std. Error |
| | (Constant) | 15688,000 | 604,355 | 25,958 | ,000 |
| | indic_02 | -398,571 | 884,684 | -,451 | ,654 |
| | indic_03 | 57,571 | 884,684 | ,065 | ,948 |
| | indic_04 | 133,286 | 884,684 | ,151 | ,881 |
| | indic_05 | -49,000 | 884,684 | -,055 | ,956 |
| | indic_06 | -159,286 | 884,684 | -,180 | ,858 |
| | indic_07 | -347,429 | 884,684 | -,393 | ,696 |
| | indic_08 | 18,714 | 884,684 | ,021 | ,983 |
| | indic_09 | 262,571 | 884,684 | ,297 | ,767 |
| | indic_10 | 463,714 | 884,684 | ,524 | ,602 |
| | indic_11 | 739,143 | 884,684 | ,835 | ,406 |
| | indic_12 | 468,143 | 884,684 | ,529 | ,598 |

a Dependent Variable: consumo_ee_sudeste

Note que as estatísticas t são pequenas, indicando que os fatores sazonais não são significantes neste caso. O que o modelo nos diz é que esta série, no período indicado, não é sazonal, ou pelo menos, que não conseguimos identificar as componentes sazonais mensais.

As estatísticas do modelo também indicam que apenas os fatores sazonais não conseguem "explicar" o consumo de energia – note que a estatística F na próxima tabela é muito pequena, e o modelo como um todo não é significativo.

**ANOVA(b)**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8832858,961 | 11 | 802987,178 | ,275 | ,989(a) |
| | Residual | 213302712,286 | 73 | 2921954,963 | | |
| | Total | 222135571,247 | 84 | | | |

a Predictors: (Constant), indic_12, indic_11, indic_10, indic_09, indic_08, indic_07, indic_06, indic_05, indic_04, indic_03, indic_02

b Dependent Variable: consumo_ee_sudeste

Na verdade, é provável que o "culpado" por isso seja o nível da série, que está aumentando, e não está sendo considerado na modelagem. Vamos adicionar uma tendência linear ao modelo, através de uma variável que assume os valores 1, 2, 3, ... 85. Esta variável será chamada de "tempo" no modelo a seguir.

O novo modelo (que inclui a tendência linear) tem os seguintes diagnósticos:

**Coefficients(a)**

| Model | | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|---|
| | | B | Std. Error | B | Std. Error |
| | (Constant) | 13054,619 | 233,489 | 55,911 | ,000 |
| | indic_02 | -92,364 | 300,804 | -,307 | ,760 |
| | indic_03 | 302,537 | 300,704 | 1,006 | ,318 |
| | indic_04 | 317,010 | 300,626 | 1,054 | ,295 |
| | indic_05 | 73,483 | 300,570 | ,244 | ,808 |
| | indic_06 | -98,044 | 300,537 | -,326 | ,745 |
| | indic_07 | -347,429 | 300,526 | -1,156 | ,251 |
| | indic_08 | -42,527 | 300,537 | -,142 | ,888 |
| | indic_09 | 140,089 | 300,570 | ,466 | ,643 |
| | indic_10 | 279,990 | 300,626 | ,931 | ,355 |
| | indic_11 | 494,177 | 300,704 | 1,643 | ,105 |
| | indic_12 | 161,936 | 300,804 | ,538 | ,592 |
| | tempo | 61,241 | 2,587 | 23,677 | ,000 |

a Dependent Variable: consumo_ee_sudeste

**ANOVA(b)**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| | Regression | 197858692,664 | 12 | 16488224,389 | 48,901 | ,000(a) |
| | Residual | 24276878,583 | 72 | 337178,869 | | |
| | Total | 222135571,247 | 84 | | | |

a Predictors: (Constant), tempo, indic_07, indic_08, indic_06, indic_05, indic_09, indic_10, indic_04, indic_11, indic_03, indic_12, indic_02
b Dependent Variable: consumo_ee_sudeste

**Model Summary(b)**

| Model | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|
| | ,891 | ,872 | 580,671 |

a Predictors: (Constant), tempo, indic_07, indic_08, indic_06, indic_05, indic_09, indic_10, indic_04, indic_11, indic_03, indic_12, indic_02
b Dependent Variable: consumo_ee_sudeste

O que podemos concluir?

- Os fatores sazonais são ainda não significantes, mas a qualidade do ajuste melhorou sensivelmente. O modelo agora é, como um todo, significativo (veja a estatística F). As

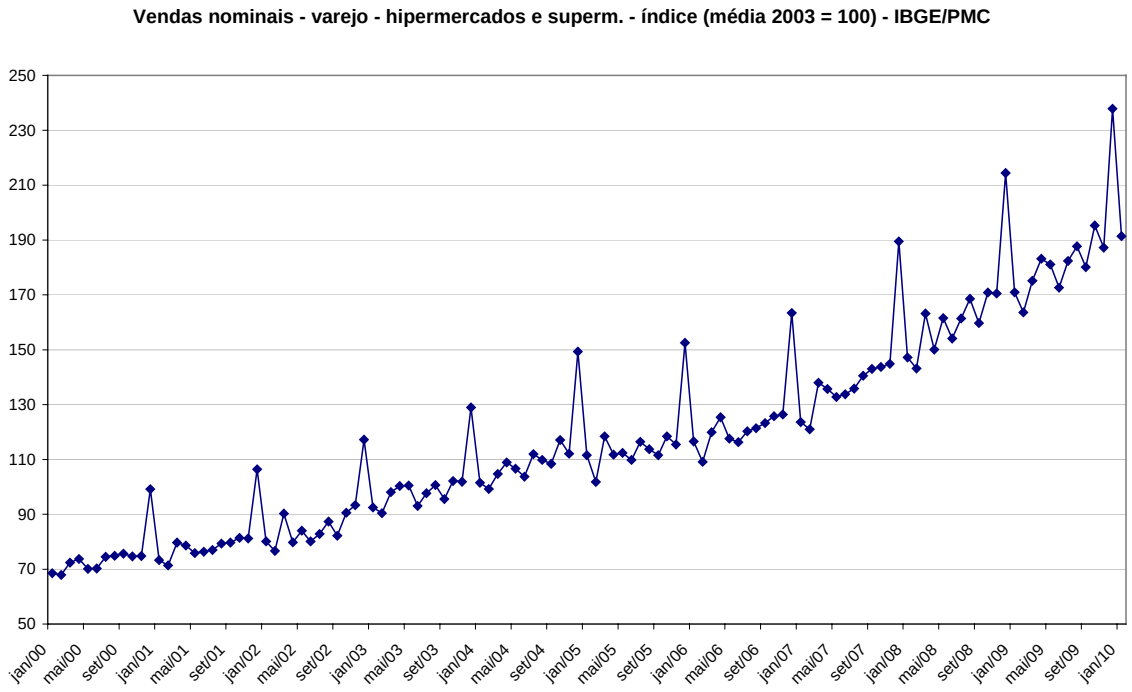
estatísticas t dos fatores sazonais são não significantes, mas são maiores (em módulo) que na situação anterior. Talvez o procedimento mais adequado (se o objetivo é encontrar um modelo parcimonioso) seja empregar apenas alguns dos fatores sazonais, e não todos, escolhendo-os, por exemplo, através de um procedimento "stepwise".

A estrutura do modelo ajustado é:

$$(Consumo\_estimado_t) = 13054{,}62 + 61{,}24 * t - 92{,}36 * Indic\_02_t + 302{,}54 * Indic\_03_t + 317{,}01 * Indic\_04_t +$$
$$+ 73{,}48 * Indic\_05_t - 98{,}04 * Indic\_06_t - 347{,}43 * Indic\_07_t - 42{,}53 * Indic\_08_t +$$
$$+ 140{,}09 * Indic\_09_t + 279{,}99 * Indic\_10_t + 494{,}18 * Indic\_11_t + 161{,}94 * Indic\_12_t$$

**Exemplo 9.3.**

Considere a série de Vendas nominais no varejo (hipermercados e supermercados) – número índice (média 2003 = 100), fornecida pela Pesquisa Mensal do Comércio (PMC) do IBGE e mostrada no próximo gráfico.




A sazonalidade na série fica particularmente óbvia por conta do mês de Dezembro, onde ocorre um "pico" nas vendas do ano. No entanto, deve-se notar também que a série tem uma tendência muito expressiva, e modela-la sem levar em conta esta tendência pode nos levar a encontrar fatores sazonais que não fazem sentido, pois estão capturando também a componente da tendência, além da sazonalidade.

Vamos experimentar diversos modelos e comentar os resultados.

Modelo 1 – SEM TENDÊNCIA, apenas componentes sazonais

Estrutura: $Y_i = \alpha + \beta_1.D_{1i} + \beta_2.D_{2i} + ... + \beta_{m-1}.D_{(m-1)i} + \varepsilon_i$ onde ajustamos m-1 = 11 variáveis dummy. Escolhemos neste caso o mês de janeiro como variável omitida, e assim serão ajustadas as dummies para fevereiro a dezembro. Os resultados do ajuste são:

**Coefficients(a)**

| Model | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|
| | B | Std. Error | B | Std. Error |
| (Constant) | 116,116 | 11,144 | 10,419 | ,000 |
| indic_02 | -11,697 | 16,150 | -,724 | ,470 |
| indic_03 | -,141 | 16,150 | -,009 | ,993 |
| indic_04 | -1,389 | 16,150 | -,086 | ,932 |
| indic_05 | -1,876 | 16,150 | -,116 | ,908 |
| indic_06 | -5,106 | 16,150 | -,316 | ,752 |
| indic_07 | -,095 | 16,150 | -,006 | ,995 |
| indic_08 | 2,284 | 16,150 | ,141 | ,888 |
| indic_09 | -,199 | 16,150 | -,012 | ,990 |
| indic_10 | 5,858 | 16,150 | ,363 | ,718 |
| indic_11 | 4,647 | 16,150 | ,288 | ,774 |
| indic_12 | 39,735 | 16,150 | 2,460 | ,015 |

a Dependent Variable: Vendas nominais - varejo - hipermercados e superm. - índice (média 2003 = 100) - IBGE/PMC - PMC12_VNSUPT12

**ANOVA(b)**

| Model | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 17236,646 | 11 | 1566,968 | 1,147 | ,333(a) |
| Residual | 148909,937 | 109 | 1366,146 | | |
| Total | 166146,582 | 120 | | | |

a Predictors: (Constant), indic_12, indic_11, indic_10, indic_09, indic_08, indic_07, indic_05, indic_04, indic_03, indic_02, indic_06

b Dependent Variable: Vendas nominais - varejo - hipermercados e superm. - índice (média 2003 = 100) - IBGE/PMC - PMC12_VNSUPT12

Note como o ajuste do modelo é ruim. A estatística F é muito pequena, indicando que os regressores são, em conjunto, não significantes. Dentre os fatores sazonais, apenas o relativo ao mês de dezembro é significativo (veja as estatísticas t). O $R^2$ desde modelo é terrível, apenas 10,4%, e o $R^2$ ajustado é 1,3%.

Vamos tentar melhorar este resultado incorporando uma tendência linear ao modelo. Criamos uma variável "tempo" definida como 1, 2, ..., 121, que é apenas um indicador do instante de tempo. Esta variável poderia ter sido criada de outra forma, qualquer transformação linear da variável "tempo" definida acima serviria. A estrutura do modelo agora é:

Modelo 2 – TENDÊNCIA LINEAR e componentes sazonais

Estrutura: $Y_i = \alpha + \lambda.t + \beta_1.D_{1i} + \beta_2.D_{2i} + ... + \beta_{m-1}.D_{(m-1)i} + \varepsilon_i$ onde ajustamos m-1 = 11 variáveis dummy. Escolhemos neste caso o mês de janeiro como variável omitida, e assim serão ajustadas as dummies para fevereiro a dezembro. Os resultados do ajuste são:

**Coefficients(a)**

| Model | | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|---|
| | | B | Std. Error | B | Std. Error |
| 1 | (Constant) | 56,560 | 3,155 | 17,927 | ,000 |
| | indic_02 | -6,816 | 4,044 | -1,685 | ,095 |
| | indic_03 | 3,764 | 4,043 | ,931 | ,354 |
| | indic_04 | 1,540 | 4,043 | ,381 | ,704 |
| | indic_05 | ,076 | 4,042 | ,019 | ,985 |
| | indic_06 | -4,130 | 4,042 | -1,022 | ,309 |
| | indic_07 | -,095 | 4,042 | -,024 | ,981 |
| | indic_08 | 1,307 | 4,042 | ,323 | ,747 |
| | indic_09 | -2,152 | 4,042 | -,532 | ,596 |
| | indic_10 | 2,929 | 4,043 | ,724 | ,470 |
| | indic_11 | ,741 | 4,043 | ,183 | ,855 |
| | indic_12 | 34,853 | 4,044 | 8,619 | ,000 |
| | tempo | ,976 | ,024 | 40,397 | ,000 |

a Dependent Variable: Vendas nominais - varejo - hipermercados e superm. - índice (média 2003 = 100) - IBGE/PMC - PMC12_VNSUPT12

**ANOVA(b)**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 156903,550 | 12 | 13075,296 | 152,778 | ,000(a) |
| | Residual | 9243,032 | 108 | 85,584 | | |
| | Total | 166146,582 | 120 | | | |

a Predictors: (Constant), tempo, indic_07, indic_08, indic_06, indic_09, indic_05, indic_10, indic_04, indic_11, indic_03, indic_12, indic_02

b Dependent Variable: Vendas nominais - varejo - hipermercados e superm. - índice (média 2003 = 100) - IBGE/PMC - PMC12_VNSUPT12

O resultado do modelo é bem superior ao anterior. Agora o modelo é significante (veja a estatística F) e ambos o $R^2$ e o $R^2$ ajustado são altos (94,4 e 93,8% respectivamente). No entanto, apenas os fatores sazonais para fevereiro e dezembro são significantes ao nível 10%.

O que fazer? Podemos tentar mudar a "cara" da tendência e tentar ajustar, por exemplo, uma tendência quadrática. Isso nos leva ao próximo modelo.

Modelo 3 – TENDÊNCIA QUADRÁTICA e componentes sazonais

Estrutura: $Y_i = \alpha + \lambda_1.t + \lambda_2.t^2 + \beta_1.D_{1i} + \beta_2.D_{2i} + ... + \beta_{m-1}.D_{(m-1)i} + \varepsilon_i$ Ajustamos m-1 = 11 variáveis dummy. Escolhemos neste caso o mês de janeiro como variável omitida, e assim serão ajustadas as dummies para fevereiro a dezembro. Os resultados do ajuste são:

**Coefficients(a)**

| Model | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|
| | B | Std. Error | B | Std. Error |
| (Constant) | 71,441 | 2,115 | 33,775 | ,000 |
| indic_02 | -5,335 | 2,369 | -2,252 | ,026 |
| indic_03 | 5,304 | 2,369 | 2,239 | ,027 |
| indic_04 | 3,125 | 2,369 | 1,319 | ,190 |
| indic_05 | 1,694 | 2,369 | ,715 | ,476 |
| indic_06 | -2,493 | 2,369 | -1,052 | ,295 |
| indic_07 | 1,549 | 2,369 | ,654 | ,515 |
| indic_08 | 2,945 | 2,369 | 1,243 | ,217 |
| indic_09 | -,534 | 2,369 | -,225 | ,822 |
| indic_10 | 4,514 | 2,369 | 1,905 | ,059 |
| indic_11 | 2,281 | 2,369 | ,963 | ,338 |
| indic_12 | 36,334 | 2,369 | 15,335 | ,000 |
| tempo | ,180 | ,057 | 3,169 | ,002 |
| tempo_quadrado | ,007 | ,000 | 14,429 | ,000 |

a Dependent Variable: Vendas nominais - varejo - hipermercados e superm. - índice (média 2003 = 100) - IBGE/PMC - PMC12_VNSUPT12

Os fatores sazonais para fevereiro, março, outubro e dezembro são significantes agora. Também existem outros fatores "quase" significantes (com 79% ou 80% de significância, que são os relativos a Abril e Agosto).

Note também que os parâmetros que caracterizam a tendência são altamente significantes.

Para tentar ainda mais amortecer esta tendência, poderíamos ajustar o modelo à série logaritmada de vendas.

Modelo 4 – TENDÊNCIA QUADRÁTICA e componentes sazonais aplicada à série do logaritmo de Vendas

**Coefficients(a)**

| Model | | Unstandardized Coefficients | | t | Sig. |
|---|---|---|---|---|---|
| | | B | Std. Error | B | Std. Error |
| 1 | (Constant) | 4,239 | ,014 | 296,004 | ,000 |
| | indic_02 | -,048 | ,016 | -3,001 | ,003 |
| | indic_03 | ,049 | ,016 | 3,039 | ,003 |
| | indic_04 | ,028 | ,016 | 1,741 | ,085 |
| | indic_05 | ,013 | ,016 | ,790 | ,431 |
| | indic_06 | -,022 | ,016 | -1,369 | ,174 |
| | indic_07 | ,013 | ,016 | ,822 | ,413 |
| | indic_08 | ,024 | ,016 | 1,503 | ,136 |
| | indic_09 | -,003 | ,016 | -,213 | ,832 |
| | indic_10 | ,036 | ,016 | 2,262 | ,026 |
| | indic_11 | ,021 | ,016 | 1,287 | ,201 |
| | indic_12 | ,270 | ,016 | 16,815 | ,000 |
| | tempo | ,006 | ,000 | 16,171 | ,000 |
| | tempo_quadrado | 1,63E-005 | ,000 | 5,322 | ,000 |

a Dependent Variable: log_vendas_varejo_hipermercados

Em vermelho estão indicados os coeficientes das dummies significantes. Note que as dummies para Junho e Agosto são significantes a nível 17%, e Novembro é significante a 20%. O modelo é altamente significante, como indicado na tabela ANOVA a seguir. O $R^2$ e o $R^2$ ajustado são, respectivamente, 99,4 e 98,5%.

**ANOVA(b)**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 10,957 | 13 | ,843 | 627,171 | ,000(a) |
| | Residual | ,144 | 107 | ,001 | | |
| | Total | 11,100 | 120 | | | |

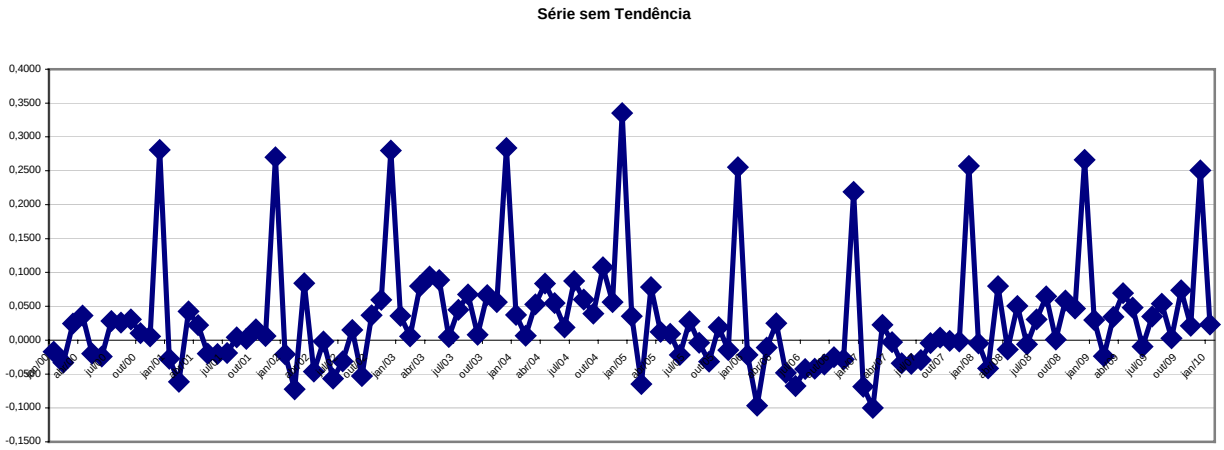
a Predictors: (Constant), tempo_quadrado, indic_07, indic_08, indic_06, indic_09, indic_05, indic_10, indic_04, indic_11, indic_03, indic_12, indic_02, tempo
b Dependent Variable: log_vendas_varejo_hipermercados

E o que estes fatores sazonais representam neste caso? A série (na escala log) menos a tendência deve ser estacionária, ou quase isso, ou seja, não deve "subir" nem "descer". Então, o que deve ficar aparente nesta série em que a tendência foi eliminada são os fatores sazonais.

Vamos calcular:

$$Z_t = \hat{Y}_t - 4{,}2385 - 0{,}0062336 * t - (1{,}62883e-005) * t^2$$

Onde $\hat{Y}_t$ é o logaritmo da série de vendas de varejo em hipermercados. $Z_t$ é a série sem tendência. O gráfico de $Z_t$ está a seguir. Note que os fatores sazonais tornam-se bastante claros e a série é estacionária na média, ou seja, não tem tendência.




Qual a interpretação dos fatores sazonais?

Se olharmos para o modelo da variável logaritmada, notamos que o nível em Janeiro (categoria omitida) é dado por:

$$\hat{Y}_t = 4{,}2385 + 0{,}0062336 * t + (1{,}62883e-005) * t^2$$

Isto ocorre pois em Janeiro os valores de todas as variáveis dummy são zero. A equação acima nos permite encontrar o valor ajustado para t = 1 (Jan/2000), t = 13 (Jan/2001), t = 25 (Jan/2002) e etc, bastando substituir o valor de t apropriado.

E as previsões dos valores de Dezembro? São dadas por:

$$\hat{Y}_t = 4{,}2385 + 0{,}0062336 * t + (1{,}62883e-005) * t^2 + 0{,}2697 * \text{INDIC\_DEZEMBRO}_t$$

Onde $\text{INDIC\_DEZEMBRO}_t$ = 1 se t é um mês de Dezembro e 0 se t não é um mês de Dezembro.

Por exemplo, a previsão para Dezembro de 2000 (t = 12) será:

$$\hat{Y}_t = 4{,}2385 + 0{,}0062336 * (12) + (1{,}62883e-005) * (12)^2 + 0{,}2697$$

Em Dezembro de 2008 (t= 108), a previsão é:

$$\hat{Y}_t = 4{,}2385 + 0{,}0062336 * (108) + (1{,}62883e-005) * (108)^2 + 0{,}2697$$

Note que o último termo desta equação (0,2697) só afeta os meses de Dezembro, ou seja, só se aplica à equação nos instantes t que correspondem a um mês de Dezembro.

**Para casa:**

Use o Excel (ou algo parecido) para calcular os valores ajustados pelo modelo para cada mês.

**Para casa II:**

**Considere as planilhas:**

Dados_mensais_exemplos_capitulo_9.xls

Exemplo_taxa_desemprego_RMSP.xls

PIB_trimestral.xls

Exercite o que você aprendeu neste capítulo sobre fatores sazonais, tendências lineares, quadráticas, etc... e ajuste modelos para as séries nestas planilhas que não foram usadas nos exemplos deste texto.

O que muda quando precisamos estimar fatores sazonais trimestrais?