CAPÍTULO 7 – ANÁLISE DE REGRESSÃO MÚLTIPLA – O PROBLEMA DA ESTIMAÇÃO

1-O MODELO DE 3 VARIÁVEIS – NOTAÇÃO E PREMISSAS

Considere o modelo de regressão múltipla com 2 variáveis explicativas dado por:

$$Y_i = \beta_1 + \beta_2.X_{2i} + \beta_3.X_{3i} + \varepsilon_i \quad \text{para i = 1, 2, ..., n} \qquad (1)$$

Na equação (1), Y é a variável dependente, X_2 e X_3 são as variáveis explicativas (ou regressores) e ε é o termo de erro aleatório. i denota a i-ésima observação. Os coeficientes β_2 e β_3 são chamados de *coeficientes parciais de regressão* e β_1 é o intercepto ou coeficiente linear.

Novamente relembramos as hipóteses subjacentes ao modelo da equação (1):

1. A média dos erros é zero (condicional aos valores dos X's): $E(\varepsilon_i | X_{2i}, X_{3i}) = 0$ para todo i
2. Os erros são descorrelatados: $COV(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$
3. Homocedasticidade (variância constante): $VAR(\varepsilon_i) = \sigma^2$ para todo i
4. Covariância nula entre os erros e cada variável: $COV(\varepsilon_i, X_{2i}) = COV(\varepsilon_i, X_{3i}) = 0$ para todo i
5. Suposição de que o modelo está corretamente especificado.
6. Inexistência de colinearidade entre os regressores, ou seja, não há relação linear exata entre X_2 e X_3. Em particular, esta hipótese implica nas colunas da matriz do modelo X (vide apêndice C) serem linearmente independentes.

Por que esta hipótese 6 (ausência de colinearidade perfeita) é importante? Pois nos permite simplificar a estrutura do modelo. Se X_3 é uma função linear perfeita de X_2, na prática não existem duas variáveis explicativas, existe só uma. Suponha que $X_3 = 2.X_2$ exatamente. Então o modelo de regressão (1) torna-se:

$$Y_i = \beta_1 + \beta_2.X_{2i} + \beta_3.X_{3i} + \varepsilon_i = \beta_1 + \beta_2.X_{2i} + \beta_3.(2.X_{2i}) + \varepsilon_i = \beta_1 + (\beta_2 + 2.\beta_3)X_{2i} + \varepsilon_i = \\ = \beta_1 + \alpha.X_{2i} + \varepsilon_i \qquad (2)$$

Ou seja, na prática (1) reduz-se a um modelo com apenas uma variável explicativa, e não conseguimos separar a influência de X_2 e X_3, que está "misturada" dentro do parâmetro α.

Mais sobre colinearidade...

- Na prática, é comum existir correlação entre os regressores – o que não pode existir é correlação perfeita (+ 1 ou -1) entre eles, pois isso impediria a inversão da matriz X^tX que é necessária para calcular os estimadores MQO.
- Multicolinearidade se refere a relações LINEARES entre os regressores. Não diz nada, a princípio, sobre relações como $X_3 = X_2^2$.

2- O Significado dos Coeficientes de Regressão Parciais

Da hipótese sobre a média dos erros segue que:

$$E(Y_i|X_{2i},X_{3i}) = \beta_1 + \beta_2.X_{2i} + \beta_3.X_{3i} \qquad (3)$$

Os coeficientes β_2 e β_3 são chamados de *coeficientes parciais de regressão* ou coeficientes angulares parciais. O que eles significam?

- β_2 mede a variação no valor médio de Y, E(Y) por unidade de variação de X_2 mantendo-se X_3 constante. Ou seja, é o efeito líquido de uma unidade de variação em X_2 sobre a média de Y excluindo-se os efeitos de X_3.
- Similarmente, β_3 mede a variação no valor médio de Y, E(Y) por unidade de variação de X_3 mantendo-se X_2 constante.
- Você pode olhar para β_2 e β_3 e reconhecer as derivadas parciais de $E(Y|X_2, X_3)$ em relação a X_2 e X_3 respectivamente.

A questão principal é: como manter fixa a influência de um dos regressores, olhando só para o efeito do outro regressor? Um algoritmo possível é mostrado a seguir, com um exemplo. Veremos que ele não será necessário para calcular os β's, sua inclusão aqui é para propósitos ilustrativos.

Exemplo 7.1.

A planilha cars_spss.xls foi retirada de um arquivo de exemplos que acompanha o software estatístico spss. As variáveis presentes no arquivo são:

milhas_por_galao = indicador do consumo de gasolina do carro

motor = indicador do tamanho (em polegadas cúbicas) do motor

hp = potência do motor em hp

peso = peso do carro em libras

tempo_aceleracao = tempo para acelerar de 0 a 60 milhas por hora em segundos

ano = ano do modelo

pais_origem = país de origem, variável categórica

numero_cilindros = número de cilindros do motor

filtro_8_cilindros = filtro = 0 se o motor tem 8 cilindros, 1 do contrário

Neste exemplo analisamos **APENAS** os carros com motor abaixo de 8 cilindros.

Os gráficos a seguir mostram a relação entre milhas_por_galao e peso e milhas_por_galao e hp.

Gráfico 1

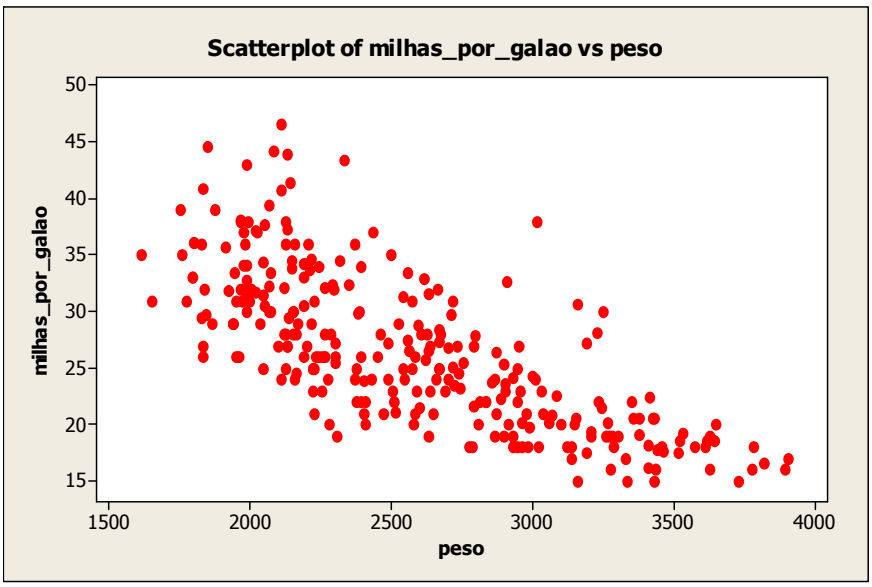
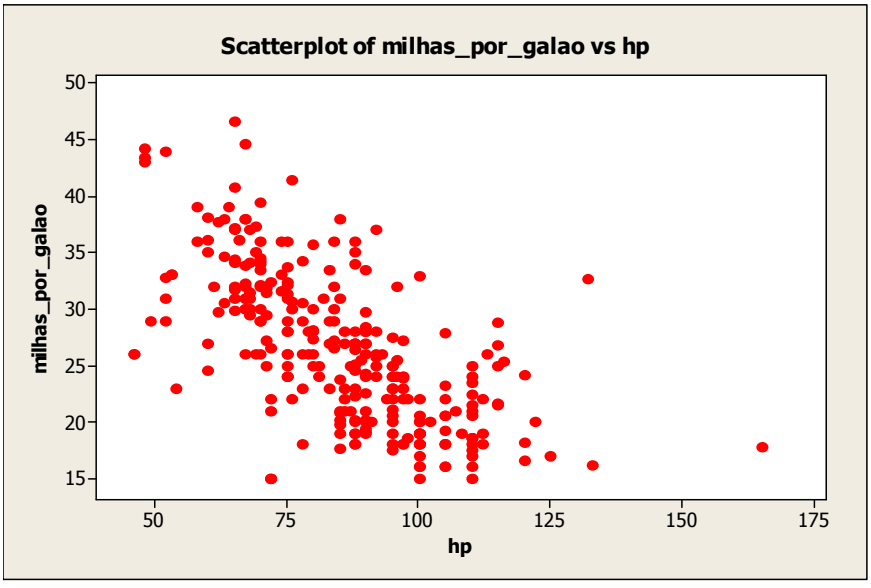


Gráfico 2



Um modelo de regressão de “milhas_por_galao” em “hp” e “peso” fornece a seguinte equação e diagnósticos básicos:

```
The regression equation is
milhas_por_galao = 52,9 - 0,0990 hp - 0,00698 peso


282 cases used, 9 cases contain missing values


Predictor         Coef     SE Coef        T       P
Constant        52,904       1,365    38,76   0,000
hp            -0,09897     0,02062    -4,80   0,000
peso       -0,0069813   0,0006945   -10,05   0,000


S = 4,33857   R-Sq = 58,5%   R-Sq(adj) = 58,2%


Analysis of Variance

Source           DF       SS      MS       F      P
Regression        2   7394,3  3697,2  196,41  0,000
Residual Error  279   5251,7    18,8
Total           281  12646,0


Source  DF  Seq SS
hp       1  5492,1
peso     1  1902,2
```

Ou seja, os coeficientes de “hp” e “peso” são, respectivamente, -0,09897 e -0,0069813 .

Suponha que desejamos identificar a influência de “hp” sobre “milhas_por_galao” mantendo constante o efeito (linear) de “peso”. Como fazer isso?

1) Faça a regressão de “milhas_por_galao” em “peso” e calcule os resíduos. O resultado é:

```
The regression equation is
milhas_por_galao = 50,7 - 0,00941 peso


288 cases used, 3 cases contain missing values


Predictor         Coef     SE Coef        T       P
Constant        50,730       1,316    38,54   0,000
peso        -0,0094074   0,0005021   -18,74   0,000


S = 4,51187   R-Sq = 55,1%   R-Sq(adj) = 55,0%
```

O que este modelo nos diz? Os resíduos são: $R1_i = mph_i - 50,7 + 0,00941*peso_i$ onde mph_i indica o consumo (em milhas por galão) do i-ésimo carro.

2) Faça a regressão de “hp” em “peso” e calcule os resíduos.

```
The regression equation is
hp = 22,6 + 0,0244 peso
```

```
285 cases used, 6 cases contain missing values


Predictor       Coef    SE Coef       T       P
Constant      22,572      3,715    6,08   0,000
peso        0,024364   0,001416   17,21   0,000


S = 12,6751    R-Sq = 51,1%    R-Sq(adj) = 51,0%
```

Os resíduos deste modelo são $R2_i = hp_i – 22,6 – 0,0244*peso_i$ e indicam a parte de "hp" que sobra após removermos a influência linear de "peso".

3) Faça a regressão (sem constante) dos resíduos em 1) em relação aos resíduos em 2). O coeficiente angular desta regressão nos dá o efeito líquido de "hp" sobre "milhas_por_galao", ou seja, vai fornecer o valor -0,09897 que encontramos acima na regressão múltipla. A "mágica" dos MQO é que eles nos fornecem estes coeficientes sem que a gente tenha que calcular estas regressões intermediárias.

```
The regression equation is
RESI1 = - 0,0990 RESI2


282 cases used, 9 cases contain missing values


Predictor       Coef   SE Coef      T      P
Noconstant
RESI2       -0,09897   0,02054  -4,82  0,000
```

Como vimos acima, o coeficiente desta regressão é o mesmo que o coeficiente de "hp" na regressão original. Na verdade você pode escrever as equações e obter o coeficiente -0.09897 explicitamente e ver que isso sempre vai dar certo. Note que, da 1ª regressão:

mph = 50,730 – 0,0094074*(peso) + RESI1 e então, RESI1 = mph - 50,730 + 0,0094074*(peso)

Da 2ª. regressão: hp = 22,572 + 0,024364*(peso) + RESI2 e então RESI2 = hp – 22,572 - 0,024364*(peso)

Fazendo a regressão de RESI1 em RESI2 leva a:

RESI1 = -0,09897*(RESI2)

Subsitituindo os valores de RESI1 e RESI2 encontrados nas duas primeiras equações de regressão leva a:

mph - 50,730 + 0,0094074*(peso) = -0,09897*(hp – 22,572 -0,024364*(peso))

mph = 50,730 + 0,09897*(22,572) + peso*(- 0,0094074 + 0,09897*0,024364) + hp*(-0,09897)

mph = 52,964 -0,0070*peso - 0,09897*hp

O que concorda, a menos de erros de arredondamento, com os coeficientes encontrados na regressão múltipla.

Se você quiser obter o coeficiente de "milhas_por_galao" em "peso" (sem recorrer à regressão múltipla) pode usar um procedimento análogo.

3 – OS ESTIMADORES DE MQO

O livro do Gujarati explicita os estimadores de MQO neste caso. Como já derivamos estes estimadores em forma matricial, acho mais conveniente relembrar a solução matricial e escrever a matriz do modelo desta forma mais geral.

Lembre-se que a solução matricial para os estimadores MQO é (vide apêndice C):

$$\left(X^t X\right)^{-1}.\left(X^t X\right).\hat{\beta} = \left(X^t X\right)^{-1}.X^t.y \Rightarrow \hat{\beta} = \left(X^t X\right)^{-1}.X^t.y \qquad (4)$$

Neste caso:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ .... \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{1,2} & x_{1,3} \\ 1 & x_{2,2} & x_{2,3} \\ ... & ... & ... \\ 1 & x_{n,2} & x_{n,3} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \varepsilon = \begin{pmatrix} e_1 \\ e_2 \\ ... \\ e_n \end{pmatrix}$$

A matriz de variância-covariância dos β's é uma matriz 3 x 3 simétrica, que contém as variâncias na diagonal principal e as covariâncias dos β's fora da diagonal principal. Ela é dada por (vide teorema 4.4. nas notas de aula do apêndice C):

$$V = \sigma^2\left(X^t X\right)^{-1} \quad \text{onde } \sigma^2 = \text{VAR}(\varepsilon_i) \text{ para i} = 1,2,...,\text{n} \qquad (5)$$

Note que, na equação (5), as variâncias e covariâncias dos β's dependem de um parâmetro desconhecido, σ^2, que é a variância dos erros. Isso indica que precisamos estimar σ^2 para que a equação (5) tenha alguma utilidade prática. Como fazê-lo? A resposta está a seguir.

Resultado 4.2.

Num modelo de regressão múltipla com intercepto e (k-1) variáveis explicativas, um estimador **não tendencioso** de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k} = \frac{\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2}{n-k} \qquad (6)$$

Neste caso particular (2 regressores e uma constante), k = 3.

O estimador dado pela equação (6) NÃO É o estimador de máxima verossimilhança de σ^2 sob a hipótese de normalidade, que tem denominador n sempre. É claro que, à medida que o número de observações (n) cresce, o estimador de máxima verossimilhança e os estimador não tendencioso tendem a ser bem "parecidos".

Propriedades dos estimadores MQO

- A superfície de regressão passa pelos pontos médios de Y e de todas as variáveis explicativas, neste caso: $(\overline{Y}, \overline{X}_2, \overline{X}_3)$.

- Assim, no caso geral temos:

$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2.\overline{X}_2 + \hat{\beta}_3.\overline{X}_3 + \ldots + \hat{\beta}_{k-1}.\overline{X}_{k-1} \qquad (7)$$

Reescrevendo (7):

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2.\overline{X}_2 - \hat{\beta}_3.\overline{X}_3 - \ldots - \hat{\beta}_{k-1}.\overline{X}_{k-1} \qquad (8)$$

Para uma observação qualquer, o valor ajustado por MQO é:

$$\begin{aligned}
&\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2.X_{2i} + \hat{\beta}_3.X_{3i} + \ldots + \hat{\beta}_{k-1}.X_{k-1,i} = \\
&= \overline{Y} - \hat{\beta}_2.\overline{X}_2 - \hat{\beta}_3.\overline{X}_3 - \ldots - \hat{\beta}_{k-1}.\overline{X}_{k-1} + \hat{\beta}_2.X_{2i} + \hat{\beta}_3.(X_{3i}) + \ldots + \hat{\beta}_{k-1}.X_{k-1,i} = \\
&= \overline{Y} + \hat{\beta}_2.(X_{2i} - \overline{X}_2) + \hat{\beta}_3.(X_{3i} - \overline{X}_3) + \ldots + \hat{\beta}_{k-1}.(X_{k-1,i} - \overline{X}_{k-1}) \qquad (9) \\
&\text{Isto é :} \\
&\hat{Y}_i - \overline{Y} = \hat{\beta}_2.(X_{2i} - \overline{X}_2) + \hat{\beta}_3.(X_{3i} - \overline{X}_3) + \ldots + \hat{\beta}_{k-1}.(X_{k-1,i} - \overline{X}_{k-1})
\end{aligned}$$

A equação (9) nos diz que o modelo pode ser escrito como uma regressão múltipla SEM constante usando como variáveis explicativas os desvios das variáveis em relação às suas médias.

4- O COEFICIENTE DE DETERMINAÇÃO (R^2) E O COEFICIENTE DE CORRELAÇÃO MÚLTIPLA

Considere um caso muito simples – o modelo constante. Queremos estimar um modelo constante, ou seja, yi = c + erro. Sem o conhecimento de nenhuma variável explicativa, ou seja, sem o modelo de regressão, a nossa melhor estimativa seria o valor médio das observações de y. Na verdade você pode aplicar o critério de mínimos quadrados ordinários para verificar que o estimador de c é a média de Y. Aqui:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{c})^2 = \sum_{i=1}^{n}\left(y_i^{\,2} - 2.\hat{c}.y_i + \hat{c}^2\right) = \sum_{i=1}^{n} y_i^{\,2} - 2\hat{c}.n.\overline{y} + n.\hat{c}^2$$

Derivando em relação a c^ e igualando a zero fornece:

$$\frac{dRSS}{d\hat{c}} = 0 \Rightarrow -2.n.\bar{y} + 2.n\hat{c} = 0 \Rightarrow \hat{c} = \bar{y}$$

A soma de quadrados (mínima) no modelo constante será conhecida como Soma de Quadrados Total, e denotada SST.

Note que:

$$SST = SYY = \sum (y_i - \bar{y})^2 \qquad (10)$$

A soma dos quadrados devidos à regressão (ou soma dos quadrados explicados pela regressão), SSReg (do inglês "sum of squares due to regression") é dada por:

$$SS\,\mathrm{Re}\,g = \sum \left(\hat{y}_i - \bar{\hat{y}}\right)^2 \quad \text{onde } \bar{\hat{y}}_i \text{ é a média dos valores ajustados, } \bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\beta}_1 + \hat{\beta}_2.x_{2i} + \hat{\beta}_3.x_{3i}\right)$$

$$\bar{\hat{y}} = \hat{\beta}_1 + \hat{\beta}_2.\bar{x}_2 + \hat{\beta}_3.\bar{x}_3 = \bar{y} \quad \text{pelas propriedades dos MQO, como já visto.}$$

Então, a soma dos quadrados explicados pela regressão é dada por:

$$SS\,\mathrm{Re}\,g = \sum (\hat{y}_i - \bar{y})^2 \qquad (11)$$

Após uma certa álgebra pode-se provar que (faça-o!):

$$SST = SYY = \sum (y_i - \bar{y})^2 = \text{SSReg} + \text{RSS} \qquad (12)$$

Ou seja, a soma dos quadrados total é composta de duas partes:

- A soma dos quadrados devido à regressão e,
- A soma do quadrado dos resíduos.

Intuitivamente, o "peso" destas duas partes na SST deve ser **um** (mas não o único) indicador do quanto o ajuste da regressão é "bom". Se a soma do quadrado dos residuos é "grande" (em relação a SSReg), o ajuste deve ser "ruim". Do contrário, se RSS é "pequena" (e SSReg é "grande"), o ajuste do modelo deve ser "bom".

Definição 4.3. (Coeficiente de Determinação R^2)

O coeficiente de determinação (R^2) de uma regressão é um número entre 0 e 1 definido como:

$$R^2 = \frac{SS\,\mathrm{Re}\,g}{SST} = \frac{SS\,\mathrm{Re}\,g}{SYY} = \frac{SYY - RSS}{SYY} = 1 - \frac{RSS}{SYY} \qquad (13)$$

Quanto mais próximo de 1. "melhor" o ajuste do modelo de regressão.

Num modelo de regressão simples, R^2 é o quadrado do coeficiente de correlação entre X e Y. Num modelo de regressão múltipla, existe uma quantidade análoga a r, que é chamada de coeficiente de correlação múltipla, que mede o grau de associação entre Y e TODAS as variáveis explicativas em conjunto.

Exemplo 4.1. – continuação

Vamos olhar com mais atenção os resultados do Exemplo 4.1. Em particular estamos interessados no cálculo do R^2 e na tabela ANOVA (que contém as somas de quadrados e estas somas divididas pelos seus graus de liberdade).

Abaixo vemos que:

```
S = 4,33857   R-Sq = 58,5%   R-Sq(adj) = 58,2%
```

Então cerca de 59% da variação nos dados é explicada pelo modelo com as 2 variáveis. Não é bom, mas também não é trágico. Note que o desvio padrão estimado é 4,339, assim a variância estimada é $(4{,}339)^2$ = 18,823. Mas, sabemos que este estimador é a RSS dividida por (n-3). A regressão usou n = 282 observações (vide exemplo 4.1) e então n-3 = 279. Veja agora a tabela ANOVA a seguir:

```
Analysis of Variance

Source           DF       SS      MS       F      P
Regression        2   7394,3  3697,2  196,41  0,000
Residual Error  279   5251,7    18,8
Total           281  12646,0
```

A soma do quadrado dos resíduos é 5251,7, seus graus de liberdade são (n-3) = 279. Dividindo RSS por 279 encontramos 18,8 (na verdade 18,823), que é o estimador da variância, que aparece na tabela ANOVA como o MS (mean squared) associado aos resíduos.

E o R^2? Pela definição e usando os valores da tabela ANOVA:

$$R^2 = \frac{SS\,\mathrm{Re}\,g}{SST} = \frac{SS\,\mathrm{Re}\,g}{SYY} = \frac{7394{,}3}{12646{,}0} = 0.5847$$

$$\text{Também o } R^2 \text{ pode ser calculado como:}$$

$$R^2 = 1 - \frac{RSS}{SYY} = 1 - \frac{5251{,}7}{12646{,}0} = 1 - 0{,}4153 = 0{,}5847$$

5 – O R^2 AJUSTADO

Um problema no uso do R^2 como medida da qualidade de um modelo de regressão é que ele é não decrescente no número de variáveis explicativas. Ou seja, à medida que colocamos mais variáveis explicativas, o R^2 aumenta (ou pelo menos, não decresce). Por que?

$$R^2 = \frac{SS\,\mathrm{Re}\,g}{SST} = \frac{SS\,\mathrm{Re}\,g}{SYY} = 1 - \frac{RSS}{SYY} = 1 - \frac{\sum \hat{\varepsilon}_i^{\,2}}{\sum (y_i - \bar{y})^2}$$

O denominador nesta última expressão (SYY) não depende do número de variáveis explicativas, mas o numerador depende. À medida que aumentamos o número de regressores, RSS tende a cair (ou pelo menos ficar igual), e assim, R^2 tende a crescer quando adicionamos mais regressores ao modelo.

Então, para comparar dois modelos para a MESMA variável dependente que tenham número de regressores diferentes, faz sentido "penalizar" o R^2 à medida que aumentamos o número de variáveis explicativas. Isso nos leva à definição do R^2 ajustado.

Definição 5.1. (R^2 ajustado)

$$R^2 adj = 1 - \frac{RSS/(n-k)}{SYY/(n-1)} \qquad (14)$$

Onde k é o número de parâmetros num modelo com (k-1) variáveis explicativas (e um termo constante) e n é o número de observações.

Por que falamos em R^2 "ajustado"? Porque as somas dos quadrados que originalmente aparecem na definição do R^2 são ajustadas pelos seus graus de liberdade. A RSS está associada a (n-k) graus de liberdade, pois o modelo tem k parâmetros, significando que "perdemos" k graus de liberdade em relação ao número de observações original. SYY está associada a (n-1) graus de liberdade, pois podemos interpretá-la como a soma de quadrados dos resíduos de um modelo constante (só um parâmetro), e então perde-se 1 grau de liberdade apenas.

Podemos reescrever o R^2 ajustado em termos de estimadores da variância. Note que:

$\hat{\sigma}^2 = \frac{RSS}{n-k}$ é o estimador da variância do erro e $S_Y^2 = \frac{SYY}{n-1}$ é a variância amostral dos Y's.

Relação entre R^2 e o R^2 ajustado

Com um pouco de álgebra é fácil mostrar que:

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

Para k > 1, o R^2 ajustado é menor que o R^2. Também, o R^2 ajustado pode ser negativo, o que não ocorre com o R^2 "usual". Note que se $R^2 = 1$, este também será o valor do R^2 ajustado. Se $R^2 = 0$, o R^2 ajustado será negativo se k >1.

No exemplo 4.1. note que o R^2 ajustado é R-Sq(adj) = 58,2%, menor que o R^2 (58,5%). A diferença é pequena pois neste caso n-1 = 281 e n-3 = 279.

Cuidados ao usar o R^2 e o R^2 ajustado

Ao comparar dois modelos através do R^2 e do R^2 ajustado, é preciso ter em mente que:

- A variável explicativa deve ser a mesma nos 2 modelos. Não se pode comparar desta forma um modelo para Y e outro para ln(Y), por exemplo;
- O número de observações (n) deve ser o mesmo nos 2 modelos.