Capítulo 4

Qualidade do ajuste

Uma vez escolhido um modelo, isto é, uma vez decididas as variáveis explicativas significativas queremos saber o grau de certeza ou o grau de acerto do modelo. Queremos construir uma medida quantitativa da qualidade do modelo. Qualificar um modelo pode ser realizado de diversas maneiras, por exemplo, observando o desvio padrão das estimativas dos parâmetros de regressão; se estes forem pequenos em relação à estimativa significa uma boa qualidade na estimação.

Para fugir de termos ambíguos foram pensados e amplamente estudados índices que quantificam a qualidade de um modelo de regressão. Um deles, amplamente utilizado é conhecido como coeficiente de determinação e denotado por R^2. Este coeficiente é também conhecido como coeficiente de determinação de Pearson ou ainda como coeficiente de determinação múltipla.

Figura 4.1: Karl Pearson

Primeiro vamos definir e estudar o coeficiente de correlação baseados nos momentos ou coeficiente de correlação de Pearson, depois definiremos o coeficiente de determinação, estudaremos suas propriedades para explicar suas vantagens e limitações mas também consideraremos outras utilidades deste coeficiente. Veremos como o R^2 permite-nos escolher as variáveis explicativas que mais influenciam a resposta e por último nos dedicaremos a estudar algumas transformações tanto nas variáveis explicativas quanto na variável resposta que, eventualmente, possam melhorar a qualidade do ajuste do modelo de regressão.

4.1 Coeficiente de correlação ρ

Em 1885, Sir Francis Galton[1] definiu o primeiro modelo de regressão e completou a teoria de correlação bivariada. Uma década mais tarde, Karl Pearson[2] desenvolveu o índice que ainda usamos para medir a correlação ρ, também chamado de coeficiente de correlação de Pearson.

[1]Francis Galton (1822 - 1911) foi um antropólogo, meteorologista, matemático e estatístico inglês.

[2]Karl Pearson (1857 - 1936) foi um grande contribuidor para o desenvolvimento da Estatística como uma disciplina científica séria e independente. Ele foi o fundador do Departamento de Estatística Aplicada (Department of Applied Statistics) na University College London em 1911; foi o primeiro departamento universitário dedicado à estatística em todo o mundo.

Primeiramente, precisamos de métodos para medir o grau de dependência entre duas variáveis na população e na amostra. A noção de coeficiente de correlação é estendida a uma medida de dependência entre uma variável e um conjunto de variáveis por meio do coeficiente de correlação múltipla. O coeficiente de correlação parcial é uma medida de dependência quando os efeitos de outras variáveis de correlação foram removidos. Esses vários coeficientes de correlação calculados a partir de amostras são usados para estimar parâmetros correspondentes de distribuições e para testar hipóteses, como hipóteses de independência.

Lembremos que se X e Y forem duas variáveis aleatórias tais que $\mathrm{E}(X^2)$ e $\mathrm{E}(Y^2)$ sejam ambas finitas a covariância entre X e Y é definida com

$$\mathrm{Cov}(X,Y) \;=\; \mathrm{E}\big((X-\mu_X)(Y-\mu_Y)\big) \;=\; \mathrm{E}(XY) \;-\; \mu_X\mu_Y,$$

sendo que $\mu_X = \mathrm{E}(X)$ e $\mu_Y = \mathrm{E}(Y)$. A covariância é uma medida de dependência linear entre duas variáveis aleatórias. Usando propriedades de valores esperados é muito fácil deduzir as seguintes propriedades.

Teorema 4.1. *Sejam X e Y duas variáveis aleatórias tais que $\mathrm{Cov}(X,Y)$ exista. Então:*

1- Para quaisquer constantes a, b, c e d, temos que

$$Cov(aX+b, cY+d) \;=\; ac\,\mathrm{Cov}(X,Y)\cdot \tag{4.1}$$

2- Se X e Y forem independentes, temos que $\mathrm{Cov}(X,Y)=0$.

Demonstração. Exercício. □

O inverso para o item 2 do Teorema 4.1 não é verdade. De fato, é simples encontrar um exemplo onde $Y = g(X)$, ou seja, onde Y é uma função de X mas $\mathrm{Cov}(X,Y)=0$; veja uma situação assim no Exercício 2. Um outro exemplo seria o caso de X ser uma variável aleatória qualquer e I uma oura variável aleatória satisfazendo que $P(I=1)=P(I=-1)=1/2$, com I independente de X. Seja $Y = IX$. Assim, $Y = \pm X$, cada um com probabilidade $1/2$, independente do valor de X. Então X e Y são não correlacionadas, mas não são independentes. Poderíamos substituir I por qualquer variável aleatória de média zero independente de X.

Dadas as variáveis aleatórias $X_1, \cdots, X_n$, muitas vezes é conveniente representar as variâncias e covariâncias destas variáveis através de uma matriz quadrada $n \times n$. Com esse objetivo construímos o vetor coluna $X = (X_1, \cdots, X_n)^\top$ e então nós definimos a matriz de variância-covariância ou também chamada de matriz de covariância de X como sendo a matriz $n \times n$ $C = \mathrm{Cov}(X)$ cujos elementos diagonais são $C_{ii} = \mathrm{Var}(X_i)$, $i = 1, \cdots, n$ e cujos elementos fora da diagonal são $C_iij = \mathrm{Cov}(X_i, X_j)$, $i \neq j$. Matrizes de variância-covariância podem ser manipuladas utilmente para transformações lineares de X: Se $Y = XB + \epsilon$ para alguma matriz B de dimensão $m \times n$ e vetor ϵ de comprimento m temos então

$$\mathrm{Cov}(Y) \;=\; B\,\mathrm{Cov}(X)B^\top\cdot$$

Da mesma forma, se definirmos o vetor de média de X como sendo

$$\mathrm{E}(X) \;=\; \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_n) \end{pmatrix},$$

temos $\mathrm{E}(Y) = B\,\mathrm{E}(X) + \epsilon$.

Enquanto a covariância dá alguma indicação da associação linear entre duas variáveis aleatórias, seu valor depende da escala das duas variáveis aleatórias. Uma medida baseada na covariância que não depende da escala das variáveis é o coeficiente de correlação, como definido em (1.5). A vantagem da correlação é o fato de que ela é essencialmente invariante às transformações lineares, ao contrário da covariância. Isto é, se $U = aX + b$ e $V = cY + d$ então

$$\rho \;=\; \mathrm{Corr}(U, V) = \mathrm{Corr}(X, Y),$$

se a e c tiverem o mesmo sinal; se a e c tiverem sinais diferentes, então $\mathrm{Corr}(U, V) = -\mathrm{Corr}(X, Y)$.

Mencionamos que se X e Y são variáveis aleatórias independentes com $\mathrm{E}(X^2)$ e $\mathrm{E}(Y^2)$ finitas então $\mathrm{Corr}(X, Y) = 0$ desde $\mathrm{Cov}(X, Y) = 0$. Entretanto, como com a covariância, uma correlação de 0 não implica independência. Correlação é meramente uma medida de dependência linear entre variáveis aleatórias, mede essencialmente o grau em que podemos aproximar uma variável aleatória por uma função linear de outra.

Teorema 4.2. *Sejam X e Y ambas variáveis aleatórias tais que $\mathrm{E}(X^2)$ e $\mathrm{E}(Y^2)$ sejam finitas e definimos*

$$g(a, b) \;=\; \mathrm{E}\Big((Y - a - bX)^2\Big).$$

Então, $g(a, b)$ é minimizada quando

$$b_0 \;=\; \frac{\mathrm{Corr}(X, Y)}{\mathrm{Var}(X)} \;=\; \mathrm{Corr}(X, Y)\left(\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}\right)^{1/2}$$

e

$$a_0 \;=\; \mathrm{E}(Y) \;-\; b_0\,\mathrm{Var}(X),$$

com $g(a_0, b_0) = \mathrm{Var}(Y)\big(1 - \mathrm{Corr}^2(X, Y)\big)$.

Demonstração. Exercício. □

Este teorema pode ser interpretado considerando-se a previsão de Y como uma função linear h de X e considerando o erro quadrático médio de predição $\mathrm{E}\big((Y - h(X))^2\big)$. Se tomarmos $h(x) = a$, ou seja, como sendo uma constante a, então como uma função de a, $\mathrm{E}\big((Y - a)^2\big)$ é minimizado em $a = \mathrm{E}(Y)$ com $\mathrm{E}\big((Y - \mathrm{E}(Y))^2\big) = \mathrm{Var}(Y)$. Tomando $h(x)$ como sendo uma função linear, o erro quadrático médio de previsão, de acordo com o Teorema 4.2, é $\mathrm{Var}(Y)\big(1 - \mathrm{Corr}^2(X, Y)\big)$. Assim, a redução do erro quadrático médio de predição ao predizer Y por uma função linear de X depende explicitamente da correlação.

Com alguma imaginação, é possível derivar uma medida mais útil de dependência entre duas variáveis aleatórias. Sejam X e Y variáveis aleatórias e considere

$$\mathrm{Corr}\big(\phi(X), \psi(Y)\big).$$

Se X e Y forem independentes esta correlação, quando bem definida, será sempre 0, uma vez que $\phi(X)$ e $\psi(Y)$ serão sempre independentes. Por outro lado, se $Y = \phi(X)$, então $\mathrm{Corr}\big(\phi(X), Y\big) = 1$, mesmo se $\mathrm{Corr}(X, Y) = 0$. Isso sugere que podemos definir a correlação máxima entre X e Y a ser

$$\sup_{\phi,\psi} \mathrm{Corr}\big(\phi(X), \psi(Y)\big),$$

onde o supremo é assumido sobre todas as funções ϕ e ψ com $\mathrm{Var}(\phi(X)) = \mathrm{Var}\big(\psi(Y)\big) = 1$. A condição que $\mathrm{Var}\big(\phi(X)\big) = \mathrm{Var}\big(\psi(Y)\big) = 1$ é necessário para excluir transformações constantes ϕ e ψ. Claramente, $\sup_{\phi,\psi} \mathrm{Corr}\big(\phi(X), \psi(Y)\big) \geq 0$ com $\sup_{\phi,\psi} \mathrm{Corr}\big(\phi(X), \psi(Y)\big) = 0$ se, e somente se, X e Y forem independentes. É claro que as funções ϕ e ψ maximizando a correlação não são únicas, pois $\mathrm{Corr}\big(a\phi(X) + b, c\psi(Y) + d\big)$ é o mesmo que $\mathrm{Corr}\big(\phi(X), \psi(Y)\big)$ se a e c tem o mesmo sinal. Infelizmente, a correlação máxima normalmente não é fácil de calcular.

4.1.1 Coeficiente de correlação amostral

Sejam $(X_1, Y_1), \cdots, (X_n, Y_n)$ vetores aleatórios distribuídos independentemente, cada um normal bivariada com médias μ_X, μ_Y, variâncias σ_X^2, σ_Y^2 e correlação ρ. O problema é estimar ρ de maneira não viciada quando todos os parâmetros são desconhecidos. Dedicamos estas seção à encontrar o chamado coeficiente de correlação amostral e sua função de densidade.

Teorema 4.3. *Seja $(X_1, Y_1), \cdots, (X_n, Y_n)$ uma amostra aleatória da distribuição normal bivariada, segundo definida em (1.8). Então, o estimador do coeficiente de correlação ou coeficiente de correlação amostral $\widehat{\rho}_n$ é dado por*

$$\widehat{\rho}_n = \frac{\sum_{i=1}^{n}(X_1 - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}. \tag{4.2}$$

Demonstração. Execício. □

O coeficiente de correlação ρ é chamado de coeficiente de correlação, de coeficiente de correlação de Pearson ou ainda coeficiente de correlação de produtos de momentos de Pearson. Acima definimos o coeficiente de correlação amostral de Pearson $\widehat{\rho}_n$ do qual vamos encontrar a função de densidade assim como elucidaremos propriedades. Primeiro, como chegamos a expressão em (4.2)? Embora deixamos ao leitor a prova da expressão do $\widehat{\rho}_n$ podemos indicar que para sua obtenção podemos utilizar os procedimentos de estimação dos momentos e de máxima verossimilhança.

No começo deste capítulo mencionados que o coeficiente de correlação foi apresentado por Karl Pearson mais foi R. A. Fisher quem investigou e encontrou de diversas formas a função de densidade do coeficiente de correlação amostral, estes resultados foram apresentados no artigos Fisher (1915) e Fisher (1928). No seguinte teorema apresentamos este resultado de maneira moderna, deve-se este resultado a Anderson (1984).

Teorema 4.4. *Seja $(X_1, Y_1), \cdots, (X_n, Y_n)$ uma amostra aleatória da distribuição normal bivariada com ρ sendo o coeficiente de correlação entre X e Y. Então, a função de densidade do coeficiente de correlação amostral é dada por*

$$f(\widehat{\rho}_n) = \frac{2^{n-3}}{\pi\Gamma(n-2)}(1-\rho^2)^{\frac{n-1}{2}}(1-\widehat{\rho}_n^2)^{\frac{n-4}{2}} \sum_{k=0}^{\infty} \Gamma^2\left(\frac{n-1+k}{2}\right) \frac{\left(2\rho\widehat{\rho}_n\right)^k}{k!}. \tag{4.3}$$

Demonstração. content... □

Na Figura 4.2 mostramos duas figuras. A esquerda apresentamos a forma da função de densidade do coeficiente de correlação amostral quando o tamanho da amostra é $n = 50$ em três situações diferentes. As situações referem-se à diferentes valores do coeficiente de correlação populacional ou teórico ρ, escolhemos os valores $\rho = -0.5$, $\rho = 0$ e $\rho = 0.5$. A forma desta função de densidade aparece na cor azul para o caso $\rho = -0.5$, a linha é preta quando $\rho = 0$ e vermelha quando $\rho = 0.5$, sendo escolhidos assim para melhorar a visualização. Podemos apreciar que esta função é simétrica com eixo de simetria o valor de ρ e atingindo o máximo nesse ponto. Nesta figura, a direita mostramos gráficos correspondentes à funções de $\widehat{\rho}_n$, as quais serão explanados na Seção 4.1.2.

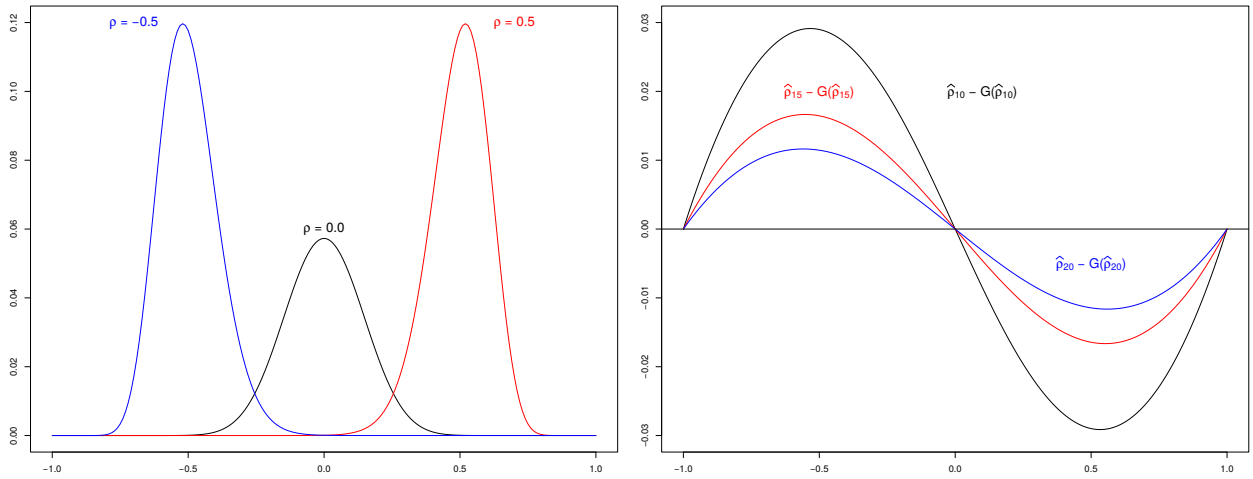

Figura 4.2: Função de densidade do estimador do coeficiente de correlação de Pearson.

4.1.2 Momentos do coeficiente de correlação amostral

Há muito se sabe que, mesmo para dados normais, a estimativa do coeficiente de correlação tem um pequeno viés (Fisher, 1915). Esse viés nos dados normais é conservador, levando à subestimação do verdadeiro coeficiente de correlação absoluta. É importante ressaltar que esse viés de dados normal é uma preocupação apenas para amostras pequenas; o viés absoluto se torna desprezível (menor que 0,01) para um tamanho de amostra maior que 20. Para corrigir esse viés, Fisher (1915) e outros (por exemplo, Olkin & Pratt, 1958) recomendaram vários ajustes aproximados. O problema, porém, é que cada um desses ajustes assume uma normalidade bivariada e poderia causar mais danos do que benefícios quando a anormalidade está presente.

Lamentavelmente, o coeficiente de correlação amostral é viciado. Com o objetivo de encontrar quais transformações faram dele um estimador com esperança ρ, o coeficiente de correlação, Olkin & Pratt (1958) procuraram obter a forma da função $G(\widehat{\rho}_n)$ de maneira que $\mathrm{E}\big(G(\widehat{\rho}_n)\big) = \rho$. Mostramos aqui resultados que nos permitem encontrar a forma da função G em amostras fixas e depois inferir resultados assintóticos.

Teorema 4.5. *Seja* $(X_1, Y_1), \cdots, (X_n, Y_n)$ *uma amostra aleatória da distribuição normal bivariada com coeficiente de correlação* ρ*. Então, se*

$$G(\widehat{\rho}_n) = \widehat{\rho}_n F(1/2, 1/2; (n-2)/2; 1-\widehat{\rho}_n), \tag{4.4}$$

onde F *é a função hipergeométrica, obtemos que* $\mathrm{E}\big(G(\widehat{\rho}_n)\big) = \rho$.

Demonstração. content... □

Podemos escrever, de maneira alternativa, a função G da seguinte forma

$$G(\widehat{\rho}_n) = \widehat{\rho}_n \frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-3}{2}\right)} \int_0^1 \frac{t^{-\frac{1}{2}}(1-t)^{\frac{n-3}{2}-1}}{\left(1-t(1-\widehat{\rho}_n^2)\right)^{\frac{1}{2}}}\,\mathrm{d}t$$

ou ainda

$$G(\widehat{\rho}_n) = \widehat{\rho}_n \frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-3}{2}\right)} \int_0^\infty \frac{t^{-\frac{1}{2}}(1+t)^{-\frac{n-3}{2}-1}}{\left(1+t\widehat{\rho}_n^2\right)^{\frac{1}{2}}}\,\mathrm{d}t. \tag{4.5}$$