UNIVERSIDADE FEDERAL DO CARIRI

CENTRO DE CIÊNCIAS E TECNOLOGIA

UNIDADE CURRICULAR: ESTATÍSTICA

Disciplina: Probabilidade e Estatística

Docente: Rosilda Benício (r*osilda.benicio@ufca.edu.br*)

UNIDADE IV: CORRELAÇÃO E REGRESSÃO

Sumário

UNIDADE IV: CORRELAÇÃO E REGRESSÃO

Correlação

Dizemos que duas variáveis X e Y estão correlacionadas se a mudança de uma provoca mudança na outra. *Positivamente correlacionadas* quando elas caminham num mesmo sentido, ou seja, elementos com valores pequenos de X tendem a ter valores pequenos de Y e elementos com valores grandes de X, tendem a ter valores grandes de Y. E *negativamente correlacionadas* quando elas caminham em sentidos opostos, ou seja, elementos com valores pequenos de X tendem a ter valores grandes de Y e elementos com valores grandes de X tendem a ter valores pequenos de Y. Por exemplo, o número de falhas numa obra e a satisfação do construtor; número de dias de atraso na entrega de uma obra e número de dias chuvosos.

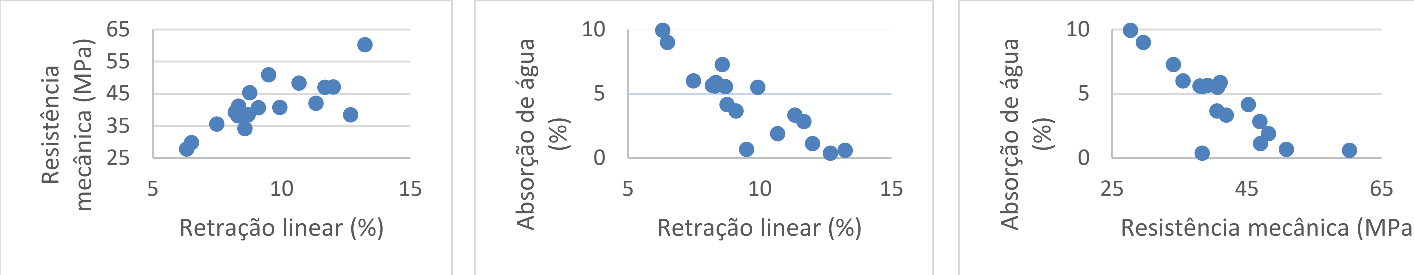
O conceito de correlação refere-se a uma associação numérica entre duas variáveis, não implicando, necessariamente, relação de causa-e-efeito. A análise de dados para verificar correlações é usualmente feita em termos exploratórios; verificação de uma correlação serve como elemento auxiliar na análise do problema em estudo.

Diagramas de dispersão

Uma forma de visualizarmos se duas variáveis apresentam-se correlacionadas é através do diagrama de dispersão, onde os valores das variáveis são representados por pontos, num sistema cartesiano.

Exemplo: No processo de queima de massa cerâmica para pavimento, corpos de prova foram avaliados por três variáveis: X_1 = retração linear (%), X_2= resistência mecânica (MPa) e X_3 = absorção de água (%). Os resultados de 18 ensaios são apresentados a seguir:

Ensaio	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
1	8,70	38,42	5,54
2	11,68	46,93	2,83
3	8,30	38,05	5,58
4	12,00	47,04	1,10
5	9,50	50,90	0,64
6	8,58	34,10	7,25
7	10,68	48,23	1,88
8	6,32	27,74	9,92
9	8,20	39,20	5,63
10	13,24	60,24	0,58
11	9,10	40,58	3,64
12	8,33	41,07	5,87
13	11,34	41,94	3,32
14	7,48	35,53	6,00
15	12,68	38,42	0,36
16	8,76	45,26	4,14
17	9,93	40,70	5,48
18	6,50	29,66	8,98



A figura acima, sugere que existe correlação positiva entre resistência mecânica e retração linear. E correlação negativa entre absorção de água e retração linear; e entre resistência mecânica e absorção de água.

Coeficiente de correlação linear de Pearson

Para qualquer conjunto de dados, podemos demonstrar que o valor do coeficiente de correlação de Pearson, r, estará no intervalo de -1 a 1. Será positivo quando os dados apresentarem correlação linear positiva; e será negativo quando os dados apresentarem correlação linear negativa.

O valor de r será tão mais próximo de 1 (ou –1) quanto mais forte for a relação linear nos dados observados. Teremos: $r = +1$ (correlação positiva perfeita) e $r = -1$ (correlação negativa perfeita). Quando não houver correlação nos dados, r acusará um valor próximo de zero.

Alguns autores admitem a seguinte interpretação:

Valor de ρ (+ ou -)	Interpretação da correlação
0,00 – 0,39	fraca
0,40 – 0,69	moderada
0,70 – 0,89	forte
0,90 – 1,00	muito forte

Observação: É importante fazer o teste de hipótese para a correlação, uma vez que a interpretação pode ser incoerente com a realidade, a depender do tamanho da amostra observada.

Para efetuar o cálculo do coeficiente de correlação r, em geral, é conveniente usar a expressão a seguir:

$$r = \frac{n\sum(x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Ilustraremos o uso dessa expressão com as 3 observações: (3,6), (4,4) e (5,2):

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	3	6	9	36	18
2	4	4	16	16	16
3	5	2	25	4	10
Soma	12	12	50	56	44

$$r = \frac{3(44) - (12)(12)}{\sqrt{3(50) - (12)^2}\,\sqrt{3(56) - (12)^2}} = -\frac{12}{\sqrt{6}\,\sqrt{24}} = -1$$

Exemplo: Com as 18 observações das variáveis retração linear (%), resistência mecânica (MPa) e absorção de água (%), calculamos o coeficiente de correlação de Pearson para cada par de variáveis, como é mostrado a seguir: Retração linear (RL), Resistência mecânica (RM) e Absorção de água (AA).

	RL	RM	AA
RL	1,00	0,75	-0,88
RM	0,75	1,00	-0,84
AA	-0,88	-0,84	1,00

Observamos que, entre resistência mecânica e retração linear, temos correlação positiva de moderada a forte. Entre retração linear e absorção de água, e entre resistência mecânica e absorção de água, temos correlações negativas fortes.

Observação: Para evitar o efeito da unidade de medida, consideramos os dados em termos da quantidade de desvio padrão que se afastam da média. Assim, a padronização de (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) é feita da seguinte forma:

$$x_i' = \frac{x_i - \bar{x}}{s_x};\ y_i' = \frac{y_i - \bar{y}}{s_y}\ , \qquad i = 1, 2, \dots, n$$

onde:
$\bar{x}$: média dos $x_i's$; s_x: desvio padrão dos $x_i's$ e $\bar{y}$: média dos $y_i's$; s_y: desvio padrão dos $y_i's$

$$r = \frac{\sum_{i=1}^{n}(x_i' y_i')}{n-1}$$

Efetuar o cálculo de r através dos valores padronizados, além de ser bastante trabalhoso tem o inconveniente de incorporar erros de arredondamento no cálculo dos valores padronizados, podendo comprometer o resultado final. A outra forma de calcular é mais usual.

Covariância e Coeficiente de correlação populacional

A covariância é a medida de variabilidade conjunta entre duas variáveis, X e Y. Se as variáveis têm covariância positiva tendem a mostrar um comportamento semelhante, ou seja, os menores (maiores) valores da variável X correspondem aos menores (maiores) valores da variável Y. Se a covariância é negativa, então as variáveis tendem a mostrar um comportamento oposto.

Definição: Sejam X e Y variáveis aleatórias integráveis. Então a covariância entre X e Y é definida por
$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
Se $Cov(X,Y) = 0$, dizemos que X e Y são não-correlacionadas.
Se X e Y são independentes e integráveis, ou seja, não correlacionadas, $E(XY) = \mu_X\mu_Y$.
Observação: $E(XY) = \mu_X\mu_Y$ não implica em independência.

A covariância padronizada, chama-se coeficiente de correlação. De forma análoga, como definimos a medida descritiva de correlação entre as observações, podemos definir, em termos probabilísticos, o parâmetro correlação entre duas variáveis aleatórias, X e Y, ou seja:

$$\rho = corr(X,Y) = E\left\{\left(\frac{X-\mu_X}{\sigma_X}.\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}$$

onde $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X = \sqrt{V(X)}$ e $\sigma_Y = \sqrt{V(Y)}$.

Inferência sobre ρ

Dada uma amostra aleatória simples (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) de n observações do par de variáveis aleatórias (X, Y), o coeficiente r calculado, pode ser considerado uma estimativa do verdadeiro e desconhecido coeficiente ρ. É comum o interesse em verificar as seguintes hipóteses:

$$\begin{cases} H_0: \rho = 0\ (as\ variáveis\ X\ e\ Y\ são\ não\ correlacionadas) \\ H_1: \rho \neq 0\ (as\ variáveis\ X\ e\ Y\ são\ correlacionadas) \end{cases}$$

podendo, ainda, a hipótese alternativa indicar o sentido da correlação (teste unilateral), tal como $H_1': \rho > 0$ (X e Y são correlacionadas positivamente) ou $H_1'': \rho < 0$ (X e Y são correlacionadas negativamente). O teste é aplicado nos casos em que esperamos o coeficiente de correlação com determinado sinal (+ ou -).

Restringindo-se à verificação de correlação linear e supondo X e Y com distribuições normais, podemos realizar o teste calculando $t = \frac{r.\sqrt{n-2}}{\sqrt{1-r^2}}$ e usando como distribuição de referência a t de Student, com $(n-2)$ graus de liberdade.

	t	p-valor
Retração linear x Resistência mecânica	4,550	0,000328
Resistência mecânica x Absorção de água	- 6,141	1,792e-06
Retração linear x Absorção de água	-7,297	1,421e-05

Considerando o teste bilateral ao nível de significância de 5%, com 16 graus de liberdade, $t_c = 2{,}120$. Em valor absoluto, $t_c < t$, rejeita H_0.

Observação: Ao invés de realizar os cálculos acima, podemos utilizar a tabela do valor absoluto mínimo para o coeficiente de correlação **r** de Pearson ser significativo.

Exemplo: Calculamos, anteriormente, as seguintes correlações baseadas em $n = 18$ observações:

	Retração linear	Resistência mecânica	Absorção de água
Retração linear	1,00	0,75	-0,88
Resistência mecânica	0,75	1,00	-0,84
Absorção de água	-0,88	-0,84	1,00

Considerando testes bilaterais ao nível de significância de 5%, verificamos na tabela que, para $n = 18$, o valor absoluto mínimo para a correlação ser significativa (rejeitar H_0) é 0,468. Como os três coeficientes calculados são, em valor absoluto, superiores a 0,468, concluímos que as três medidas usadas para avaliar a qualidade da cerâmica são realmente correlacionadas.

Exercícios

1. Calcule o coeficiente de correlação de Pearson entre retração linear (%) e resistência mecânica (MPa) para as 5 primeiras observações apresentadas no primeiro exemplo. Apenas com estas observações, o teste estatístico detecta correlação entre as duas variáveis?

2. O ensaio de esclerometria, um método de dureza superficial não destrutivo é empregado na determinação da resistência à compressão do concreto. Para alguns ensaios é possível realizar no corpo-de-prova primeiramente o ensaio de esclerometria e, em seguida, o ensaio para obter a resistência à compressão. Ensaios de esclerometria e resistência a compressão na idade de 7 dias foram realizados, para 18 corpos de prova. Os resultados obtidos estão expostos no quadro abaixo.

IE	15	16	14	16	15	14	16	16	16
Resistência (MPa)	25,6	28,7	27,4	28,3	29,2	28,9	31	30,4	29,9
IE	20	19	19	19	19	17	17	18	17
Resistência (MPa)	36,6	40,6	38,9	34,3	35	34,9	35,5	37	34,6

a) Calcule o coeficiente de correlação para as variáveis de interesse, e interprete os resultados obtidos.
b) A um nível de significância de 2%, você diria que as variáveis estão significativamente correlacionadas?

Regressão Linear Simples

Iniciaremos o estudo da regressão com a formulação mais simples, relacionando uma variável Y, chamada de variável resposta ou dependente, com uma variável X, denominada de variável explicativa ou independente. Como, por exemplo:

Variável independente, X	**Variável dependente, Y**
Temperatura do forno (°C)	Resistência mecânica da cerâmica (MPa)
Área construída do imóvel (m²)	Preço do imóvel (R$)

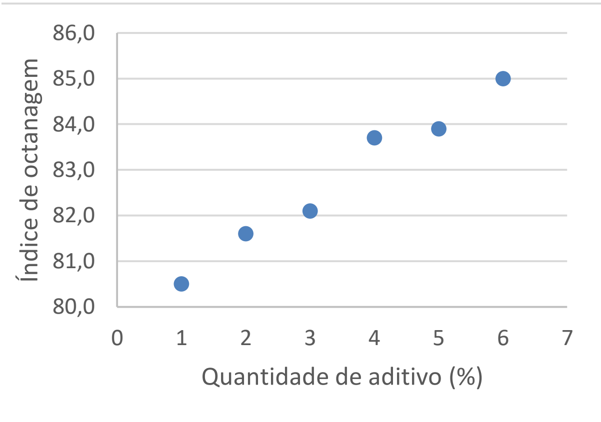
Observe que esses exemplos se distinguem dos exemplos sobre correlação por suporem uma relação de casualidade entre X e Y. É esta a diferença básica de um estudo de correlação e uma análise de regressão. A aplicação da análise de regressão é geralmente feita sob um referencial teórico, que justifique uma relação matemática de casualidade. Além disso, a variável X normalmente é controlada (não aleatória) e Y é uma variável aleatória.

Assim como num estudo de correlação, a análise de regressão também parte de um conjunto de observações pareadas $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, relativas as variáveis X e Y.

Exemplo: Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um novo aditivo (X). Para isso, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo. Os resultados são mostrados na figura seguinte.

(*Octanagem - propriedade de a gasolina resistir à compressão sem entrar em autoignição.)*

X	Y
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85,0



Observe que é razoável supor uma relação aproximadamente linear entre X e Y para os níveis de aditivo ensaiados (de 1 a 6%). Contudo, os pontos não estão exatamente sobre uma reta, provavelmente por causa da existência de fatores não controláveis no processo. Vamos supor, então, que o valor esperado de Y varie com X, de acordo com uma equação de primeiro grau, ou seja:

$$E\{Y\} = \alpha + \beta X$$

em que α e β são os parâmetros do modelo.

Seja um conjunto de observações $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. O chamado modelo de regressão linear simples para as observações é dado por

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

Onde: Y_i é a variável aleatória associada à i-esima observação do Y; e ϵ_i é o erro aleatório da i-esima observação, isto é, o efeito de infinidade de fatores que estão afetando a observação de Y de forma aleatória.

Note que é razoável supor $E\{\epsilon_i\} = 0 \quad (i = 1,2,3, \dots, n)$. Assim, considerando X uma variável controlável (não aleatória), temos, pelas propriedades do valor esperado

$E\{Y\} = \alpha + \beta x_i$, é compatível com
$E\{Y\} = \alpha + \beta X$ para a i-esima observação.

Método dos mínimos quadrados

Para a construção do modelo $E\{Y\} = \alpha + \beta X$, precisamos obter estimativas para α e β, a partir de um conjunto de observações $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Ou seja, queremos encontrar a reta que passe o mais próximo possível dos pontos observados.

Há vários métodos para estimar os parâmetros do modelo. O mais usual é o método de mínimos quadrados, que consiste em fazer com que a soma dos erros quadráticos seja a menor possível. Considerando o modelo $E\{Y\} = \alpha + \beta x_i$, temos que o erro aleatório da i-esima observação ($i = 1, 2, 3, \ldots, n$) é dado por

$$\epsilon_i = Y_i - (\alpha + \beta x_i).$$

O método consiste em obter os valores de α e β que minimizam a expressão:

$$S = \sum \epsilon_i^2 = \sum \{Y_i - (\alpha + \beta x_i)\}^2$$

que pode ser feito igualando derivadas parciais a zero, ou seja:

$$\frac{\partial S}{\partial \alpha} = 0 \text{ e } \frac{\partial S}{\partial \beta} = 0$$

resultando nas seguintes estimativas para α e β, as quais chamaremos de a e b, respectivamente:

$$b = \frac{n \sum (x_i y_i) - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

$$a = \frac{\sum y_i - b \sum x_i}{n}$$

em que $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ é a amostra efetivamente observada.

A chamada equação (reta) de regressão é dada por: $\hat{y} = a + bx$

Para cada valor x_i ($i = 1, 2, 3, \ldots, n$), temos pela equação de regressão, o valor predito: $\hat{y}_i = a + bx_i$

A diferença entre os valores observados e os preditos é chamada de resíduo: $e_i = y_i - \hat{y}_i$

O resíduo relativo à i-esima observação (e_i) pode ser considerado uma estimativa do erro aleatório (ϵ_i) desta observação.

Desvio padrão dos resíduos: $S_e = \sqrt{\frac{\sum (e_i - \bar{e})^2}{n-2}}$. Ao se ajustar a reta espera-se que ela explique o conjunto de dados coletados. Quanto mais próximo a zero estiver S_e, o ajuste da reta se aproxima da completude.

Exemplo (Continuação): Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um novo aditivo (X). Para isso, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo.

Dados			Cálculos intermediários	
Ensaio (i)	x_i	y_i	x_i^2	$x_i y_i$
1	1	80,5	1	80,5
2	2	81,6	4	163,2
3	3	82,1	9	246,3
4	4	83,7	16	334,8
5	5	83,9	25	419,5
6	6	85,0	36	510,0
Soma	21	496,8	91	1754,3

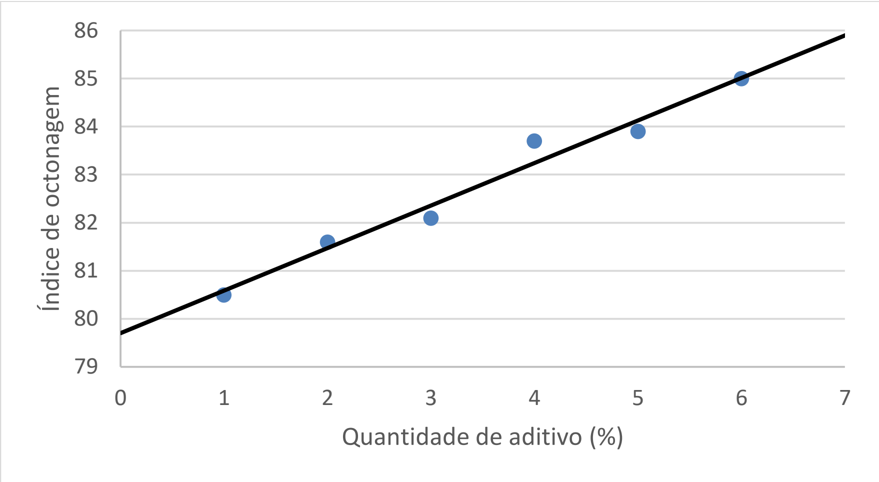
$$b = \frac{6(1754{,}3) - (21)(496{,}8)}{6\,(91) - (21)^2} = \frac{93}{105} = 0{,}886$$

$$a = \frac{496{,}8 - (0{,}886)(21)}{6} = 79{,}7$$

Assim, temos a seguinte equação de regressão:

$\hat{y} = 79{,}7 + 0{,}886x.$

Para traçar a reta no plano, basta atribuir dois valores para x e calcular os correspondentes valores de $\hat{y}$. Veja na figura abaixo.



A partir dos seis ensaios experimentais, construímos um modelo, o qual nos permite predizer o índice de octanagem da gasolina ($\hat{y}$) a partir de uma quantidade do novo aditivo (x). Por exemplo, se for adicionado $x = 5{,}5\%$ de aditivo, esperamos um índice de octanagem de $\hat{y} = 79{,}7 + (0{,}886).(5{,}5) = 84{,}573$. A tabela seguinte mostra que os valores preditos pelo modelo estão bastante próximos dos valores observados no experimento.

x_i	y_i	$\hat{y}_i$	e_i
1	80,5	80,586	-0,086
2	81,6	81,472	0,128
3	82,1	82,358	-0,258
4	83,7	83,244	0,456
5	83,9	84,130	-0,230
6	85,0	85,016	-0,016

O coeficiente b fornece uma estimativa da variação esperada de Y, a partir da variação de uma unidade em X. O sinal deste coeficiente indica o sentido da variação. No exemplo, podemos dizer: a cada 1% a mais do novo aditivo, esperamos um aumento de 0,886 no índice de octanagem.

Análise de variância do modelo

Se X não influencia Y, então o valor esperado de Y pode ser estimado simplesmente pela média aritmética ($\bar{y}$) das observações de Y. Mas se existe influência de X sobre Y, então deve haver algum ganho em considerar a equação de regressão ($\hat{y} = a + bx$). Este ganho pode ser avaliado ao comparar os resíduos nas duas situações.

Para $i = 1, 2, 3, \dots, n$, sejam:

a) $y_i - \bar{y}$ (desvios em relação à média aritmética – não levam em consideração a relação entre X e Y);

b) $y_i - \hat{y}_i$ (desvios em relação aos valores preditos pela equação de regressão – consideram uma relação linear entre X e Y);

c) $\hat{y}_i - \bar{y}$ (desvios dos valores preditos em relação à média aritmética.

As somas dos quadrados dos desvios satisfazem à seguinte equação:

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

$$variação\ total = variação\ explicada\ pela\ eq.de\ regressão + variação\ não\ explicada$$

Chamaremos de coeficiente de determinação a seguinte razão:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{variação\ explicada}{variação\ total};\ 0 \leq R^2 \leq 1$$

O coeficiente de determinação é uma medida descritiva da proporção da variação de Y que pode ser explicada por variações de X, segundo o modelo especificado. Verificando se o modelo proposto é ou não adequado para descrever o fenômeno. Quanto mais próximo a 1 estiver o R^2, mais adequado é o modelo.

Exemplo: (Continuação):

x_i	y_i	$\bar{y}$	$\hat{y}_i$	$(y_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \bar{y})^2$
1	80,5	82,8	80,59	5,29	0,01	4,90
2	81,6	82,8	81,47	1,44	0,02	1,77
3	82,1	82,8	82,36	0,49	0,07	0,20
4	83,7	82,8	83,24	0,81	0,21	0,20
5	83,9	82,8	84,13	1,21	0,05	1,77
6	85,0	82,8	85,01	4,84	0,00	4,90
Soma de quadrados				14,08	0,35	13,73

Coeficiente de determinação:

$$R^2 = \frac{13,73}{14,08} = 0,975 = 97,5\%$$

Considerando os seis ensaios realizados, a variância da octanagem da gasolina é explicada, em parte, pela variação da quantidade de aditivo adicionado ($R^2 = 97,5\%$ de explicação) e em parte ($1 - R^2 = 2,5\%$) devido a outros fatores intervenientes no processo.

No caso do modelo de regressão simples, R^2 coincide, numericamente, com o quadrado do coeficiente de correlação de Pearson (r).

A equação de regressão obtida, apenas estabelece uma relação funcional entre as variáveis dependente e independente, para representar o fenômeno em estudo. Portanto, a simples obtenção da equação não responde ao pesquisador se a variação da variável independente influencia significativamente na variação da variável dependente. Para verificar a adequabilidade, deve-se testar, pela ANOVA:

$$\begin{cases} H_0: \beta = 0\ (a\ variável\ independente\ não\ influencia\ ) \\ H_1: \beta \neq 0\ (a\ variável\ independente\ influencia\ ) \end{cases}$$

Processo simplificado de cálculo:

- Soma de quadrados totais (corrigida pela média aritmética):

$$SQT = \sum(y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

- Soma de quadrados do erro ou soma de quadrado dos resíduos:

$$SQE = \sum(y_i - \hat{y}_i)^2 = \sum y_i^2 - a\sum y_i - b\sum x_i y_i$$

- Soma de quadrados da regressão:

$$SQR = SQT - SQE$$

- Coeficiente de determinação:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

- Análise de variância (Anova) da regressão linear simples:

Fonte de variação	$\boldsymbol{SQ}$	$\boldsymbol{gl}$	$\boldsymbol{QM}$	$\boldsymbol{F}$
Regressão	$SQR = \sum(\hat{y}_i - \bar{y})^2$	1	$QMR = SQR/1$	$f_0 = \dfrac{QMR}{QME}$
Erro	$SQE = \sum(y_i - \hat{y}_i)^2$	$n-2$	$QME = SQE/(n-2)$	
Total	$SQT = \sum(y_i - \bar{y})^2$	$n-1$	$QMT = SQT/(n-1)$	

Exemplo (Continuação): Tabela da Anova

Fonte de variação	$\boldsymbol{SQ}$	$\boldsymbol{gl}$	$\boldsymbol{QM}$	$\boldsymbol{F}$
Regressão	13,73	1	13,729	$f_0 = 156{,}26$
Erro	0,35	4	0,088	
Total	14,08	5		

Admitindo $\alpha = 1\%$, temos, de acordo com a tabela, $f = 21{,}20$. Logo, H_0 é rejeitada. Ou seja, a variação da quantidade de aditivo adicionado não influencia significativamente na variação da octanagem da gasolina.

Análise dos resíduos

Na seção anterior, estabelecemos um modelo para um conjunto de observações $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, relativo às variáveis X e Y, da forma

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

onde:
Y_i é a variável aleatória associada à i-esima observação do Y;
ϵ_i é o erro aleatório da i-esima observação; e
α e β são parâmetros a serem estimados com os dados.

Assim, estamos assumindo que X causa Y através de uma relação linear, e toda a variação em torno dessa relação deve-se ao efeito (erro) aleatório. É necessário supor que as observações de Y sejam independentes, e o termo de erro tenha distribuição normal, com média nula e variância constante. Um processo gráfico é usado para verificar se estas suposições podem ser válidas, e caso contrário, o que pode ser feito para corrigir as distorções.

Após a estimação dos parâmetros α e β, podemos calcular os resíduos, ou seja:

$$e_i = y_i - \hat{y}_i \qquad (i + 1, 2, 3, \ldots, n)$$

A partir disso, constrói-se gráficos para verificar:

- Adequação do modelo;
- Presença de pontos discrepantes;
- Relação não linear (logarítmica; exponencial...).

Exercícios

1. No processo de queima de massa cerâmica, avaliou-se o efeito da temperatura do forno (X) sobre a resistência mecânica da massa queimada (Y). Foram realizados 6 ensaios com níveis de temperatura equidistantes, os quais designaremos por 1, 2, 3, 4, 5 e 6. Os valores obtidos de resistência mecânica (MPa) foram: 41, 42, 50, 53, 54 e 60, respectivamente. Pede-se:

a) Calcule o coeficiente de correlação de Pearson.
b) As estimativas de α e β da equação de regressão $E(Y) = \alpha + \beta\, x$.
c) O coeficiente R^2.
d) O desvio padrão dos resíduos.

2. A tabela a seguir relaciona os pesos (em centenas de kg) e as taxas de rendimento de combustível em rodovia (km/litro), numa amostra de dez carros de passeio novos.

Peso	12	13	14	14	16	18	19	22	24	26
Rendimento	16	14	14	13	11	12	09	09	08	06

a) Calcule o coeficiente de correlação de Pearson.

b) Considerando o resultado do item (a), como você avalia o relacionamento entre peso e rendimento, na amostra observada?

c) Para estabelecer a equação de regressão, qual deve ser a variável dependente e a variável independente? Justifique sua resposta.

d) Estabeleça a equação de regressão.

e) Apresente o diagrama de dispersão e a reta de regressão obtida no item anterior.

f) Você considera adequado o ajuste do modelo de regressão do item (d)? Dê uma medida dessa adequação.

g) Qual é o rendimento esperado para um carro de 2000 kg? Justifique sua resposta. *Lembrete*: *os dados de peso na tabela estão em centenas de kg.*

h) Você considera seu estudo capaz de predizer o rendimento esperado de um veículo com peso de 7000 kg? Justifique sua resposta.

3. Para verificar a viabilidade de incluir os resíduos da queima de carvão mineral na composição do cimento, foram feitos ensaios com cimento contendo de 0 a 9% de cinza de carvão; e medida a resistência à compressão (em MPa), após 28 dias. Os resultados foram os seguintes:

Carvão (%)	0	1	2	3	4	5	6	7	8	9
Resistência (MPa)	38,5	40,2	42,1	37,5	41,1	36,9	38,2	36,7	39,5	35,9

a) Calcule o coeficiente de correlação de Pearson.

b) Estabeleça a equação de regressão.

c) Calcule R^2.

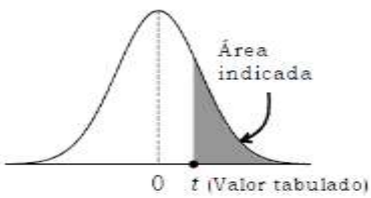
d) Os resultados mostram evidência de que o uso de cinza de carvão mineral na composição do cimento diminui a sua resistência aos 28 dias?

Tabela: Valor absoluto mínimo para o coeficiente de correlação **r** de Pearson ser significativo.

Nível de significância, **α**, num teste unilateral						
	0,100	**0,050**	**0,025**	**0,010**	**0,005**	**0,001**
Nível de significância, **α**, num teste bilateral						
n	**0,200**	**0,100**	**0,050**	**0,020**	**0,010**	**0,002**
5	0,687	0,805	0,878	0,934	0,959	0,986
6	0,608	0,729	0,811	0,882	0,917	0,963
7	0,551	0,669	0,754	0,833	0,875	0,935
8	0,507	0,621	0,707	0,789	0,834	0,905
9	0,472	0,582	0,666	0,750	0,798	0,875
10	0,443	0,549	0,632	0,715	0,765	0,847
11	0,419	0,521	0,602	0,685	0,735	0,820
12	0,398	0,497	0,576	0,658	0,708	0,795
13	0,380	0,476	0,553	0,634	0,684	0,772
14	0,365	0,458	0,532	0,612	0,661	0,750
15	0,351	0,441	0,514	0,592	0,641	0,730
16	0,338	0,426	0,497	0,574	0,623	0,711
17	0,327	0,412	0,482	0,558	0,606	0,694
18	0,317	0,400	0,468	0,543	0,590	0,678
19	0,308	0,389	0,456	0,529	0,575	0,662
20	0,299	0,378	0,444	0,516	0,561	0,648
21	0,291	0,369	0,433	0,503	0,549	0,635
22	0,284	0,360	0,423	0,492	0,537	0,622
23	0,277	0,352	0,413	0,482	0,526	0,610
24	0,271	0,344	0,404	0,472	0,515	0,599
25	0,265	0,337	0,396	0,462	0,505	0,588
26	0,260	0,330	0,388	0,453	0,496	0,578
27	0,255	0,323	0,381	0,445	0,487	0,568
28	0,250	0,317	0,374	0,437	0,479	0,559
29	0,245	0,311	0,367	0,430	0,471	0,550
30	0,241	0,306	0,361	0,423	0,463	0,541
35	0,222	0,283	0,334	0,392	0,430	0,504
40	0,207	0,264	0,312	0,367	0,403	0,474
45	0,195	0,248	0,294	0,346	0,380	0,449
50	0,184	0,235	0,279	0,328	0,361	0,427
60	0,168	0,214	0,254	0,300	0,330	0,391
70	0,155	0,198	0,235	0,278	0,306	0,363
80	0,145	0,185	0,220	0,260	0,286	0,340
90	0,136	0,174	0,207	0,245	0,270	0,322
100	0,129	0,165	0,197	0,232	0,256	0,305

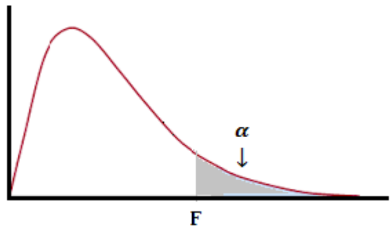
Nota: Tabela construída com base na estatística $t = \frac{r.\sqrt{n-2}}{\sqrt{1-r^2}}$ que tem distribuição t de *Student* com $gl = n - 2$, sob as suposições de os dados terem distribuição normal e a correlação ser linear.

Distribuição **t de *Student***



	Área da cauda superior								
gl	**0,2500**	**0,1000**	**0,0500**	**0,0250**	**0,0100**	**0,0050**	**0,0025**	**0,0010**	**0,0005**
1	1,000	3,078	6,314	12,710	31,820	63,660	127,300	318,300	636,600
2	0,816	1,886	2,920	4,303	6,965	9,925	14,090	22,330	31,600
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,210	12,920
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,660
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
35	0,682	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
45	0,680	1,301	1,679	2,014	2,412	2,690	2,952	3,281	3,520
50	0,679	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
z	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Distribuição **F de Snedecor**
($\boldsymbol{\alpha} = \mathbf{1}\%$)



gl	graus de liberdade no numerador									
denom.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	4052,20	4999,30	5403,50	5624,30	5764,00	5859,00	5928,30	5981,00	6022,40	6055,90
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,93	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,52	5,56	5,04	4,70	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,90	3,81
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,02	3,84	3,71	3,60	3,51
19	8,19	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,77	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,79	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,05	3,73	3,50	3,33	3,20	3,09	3,01
30	7,56	5,39	4,51	4,02	3,70	3,47	3,31	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,79	2,70
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
80	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,64	2,55
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50

Distribuição **F de Snedecor** ($\alpha = 2,5\%$)										
gl	graus de liberdade no numerador									
denom.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	647,80	799,50	864,20	899,60	921,80	937,10	948,20	956,60	963,30	968,60
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40
3	17,44	16,04	15,44	15,10	14,89	14,74	14,62	14,54	14,47	14,42
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,91	8,84
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
7	8,07	6,54	5,89	5,52	5,29	5,12	5,00	4,90	4,82	4,76
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
9	7,21	5,72	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,86	3,78	3,72
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15
15	6,20	4,77	4,15	3,80	3,58	3,42	3,29	3,20	3,12	3,06
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,13	3,05	2,99
17	6,04	4,62	4,01	3,67	3,44	3,28	3,16	3,06	2,99	2,92
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82
20	5,87	4,46	3,86	3,52	3,29	3,13	3,01	2,91	2,84	2,77
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,74
22	5,79	4,38	3,78	3,44	3,22	3,06	2,93	2,84	2,76	2,70
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67
24	5,72	4,32	3,72	3,38	3,16	3,00	2,87	2,78	2,70	2,64
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61
26	5,66	4,27	3,67	3,33	3,11	2,95	2,82	2,73	2,65	2,59
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,58	2,51
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
50	5,34	3,98	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27
80	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,36	2,28	2,21
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18

Distribuição **F de Snedecor** ($\alpha = 5\%$)										
gl	graus de liberdade no numerador									
denom.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	161,40	199,50	215,70	224,60	230,20	234,00	236,80	238,90	240,50	241,90
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,39	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,97	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,10	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,97	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,56	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,33	3,47	3,07	2,84	2,69	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,38	2,32	2,28
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,26
25	4,24	3,39	2,99	2,76	2,60	2,49	2,41	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,17
40	4,09	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,98	1,93

Distribuição **F de Snedecor** ($\alpha = 10\%$)										
gl	graus de liberdade no numerador									
denom.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,20
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39
3	5,54	5,46	5,39	5,34	5,31	5,29	5,27	5,25	5,24	5,23
4	4,55	4,33	4,19	4,11	4,05	4,01	3,98	3,96	3,94	3,92
5	4,06	3,78	3,62	3,52	3,45	3,41	3,37	3,34	3,32	3,30
6	3,78	3,46	3,29	3,18	3,11	3,06	3,01	2,98	2,96	2,94
7	3,59	3,26	3,07	2,96	2,88	2,83	2,79	2,75	2,73	2,70
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,25	2,21	2,19
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14
14	3,10	2,73	2,52	2,40	2,31	2,24	2,19	2,15	2,12	2,10
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03
17	3,03	2,65	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,01	1,98
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96
20	2,98	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,97	1,94
21	2,96	2,58	2,37	2,23	2,14	2,08	2,02	1,98	1,95	1,92
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90
23	2,94	2,55	2,34	2,21	2,12	2,05	2,00	1,95	1,92	1,89
24	2,93	2,54	2,33	2,20	2,10	2,04	1,98	1,94	1,91	1,88
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,90	1,87
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86
27	2,90	2,51	2,30	2,17	2,07	2,01	1,95	1,91	1,87	1,85
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84
29	2,89	2,50	2,28	2,15	2,06	1,99	1,94	1,89	1,86	1,83
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76
50	2,81	2,41	2,20	2,06	1,97	1,90	1,84	1,80	1,76	1,73
60	2,79	2,39	2,18	2,04	1,95	1,88	1,82	1,78	1,74	1,71
80	2,77	2,37	2,15	2,02	1,92	1,85	1,79	1,75	1,71	1,68
100	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,70	1,66