Aula de Exercícios - Testes & Regressão Linear

Organização: Airton Kist, Guilherme Ludwig
Digitação: Guilherme Ludwig

Estimadores de Máxima Verossimilhança

Exemplo

Considere Y uma v.a. com distribuição de Poisson, com parâmetro $\lambda > 0$. Obtenha o EMV de λ, baseado numa amostra de tamanho n.

Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 303.

Estimadores de Máxima Verossimilhança

Lembre-se que se temos uma amostra aleatória (isto é, independente e identicamente distribuida) $X_1, X_2, \ldots, X_n$ da variável aleatória X com distribuição de Poisson (λ desconhecido), então a função de verossimilhança é, por definição:

$$L(\lambda|x_1, \ldots, x_n) = P(X_1 = x_1|\lambda) \ldots P(X_n = x_n|\lambda)$$

Note que $L(\lambda|x_1, \ldots, x_n)$ é função de λ, com $x_1, \ldots, x_n$ observados. O nosso estimador $\hat{\lambda}_{MV}$ é o valor de λ que maximiza essa função.

Estimadores de Máxima Verossimilhança

Um ponto de máximo de uma função contínua $f(x)$ é um ponto x_0 que satisfaz

$$f'(x_0) = 0$$

$$f''(x_0) < 0$$

Note ainda que se x_0 é ponto de máximo de $f(x)$, e $f(x) > 0$ $\forall x \in \mathbb{R}$, então x_0 também é ponto de máximo de $\log(f(x))$, pois

$$\frac{d}{dx}\log(f(x))\bigg|_{x=x_0} = \frac{f'(x)}{f(x)}\bigg|_{x=x_0} = \frac{f'(x_0)}{f(x_0)} = 0$$

$$\frac{d^2}{dx^2}\log(f(x))\bigg|_{x=x_0} = \frac{f''(x_0)\cdot f(x_0) - (f'(x_0))^2}{(f(x_0))^2} < 0$$

Estimadores de Máxima Verossimilhança

Seja $\ell(\lambda|x_1, \ldots, x_n) = \log\left(L(\lambda|x_1, \ldots, x_n)\right)$ a função de log-verossimilhança. Encontraremos o EMV usando ou ℓ ou L, a que for mais conveniente.

No caso da Poisson, temos que

$$L(\lambda|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \left(e^{-n\lambda}\right)\left(\lambda^{\sum_{i=1}^{n} x_i}\right)\left(\prod_{i=1}^{n} \frac{1}{x_i!}\right)$$

que é difícil de derivar, mas sua log-verossimilhança é

$$\ell(\lambda|x_1, \ldots, x_n) = -n\lambda + \log(\lambda)\sum_{i=1}^{n} x_i + \log\left(\prod_{i=1}^{n} \frac{1}{x_i!}\right)$$

Estimadores de Máxima Verossimilhança

Agora, derivamos ℓ com relação a λ e igualamos a zero, para encontrar o ponto crítico:

$$\frac{d}{d\lambda}\ell(\lambda|x_1,\ldots,x_n) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \Leftrightarrow \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Note também que

$$\frac{d^2}{d\lambda^2}\ell(\lambda|x_1,\ldots,x_n) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0$$

pois $x_i > 0$, $\forall i$. Então o ponto encontrado, dado por $\bar{x}$, é de máximo, e o estimador de máxima verossimilhança é

$$\hat{\lambda}_{MV} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Teste T

Exemplo

Num estudo comparativo do tempo médio de adaptação, uma amostra aleatória, de 50 homens e 50 mulheres de um grande complexo industrial, produziu os seguintes resultados:

	Homens	Mulheres
Média	3,2 anos	3,7 anos
Desvio padrão	0,8 anos	0,9 anos

Que conclusões você poderia tirar para a população dessa indústria? Quais suposições você deve fazer?
Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 365.

Teste T

Queremos determinar se há diferença entre o tempo de adaptação de homens e mulheres. Devemos supor que

- O tempo de adaptação tem distribuição Normal.
- A amostra foi colhida de maneira independente.

Queremos testar a hipótese que as médias são iguais, isto é, $H_0 : \mu_H = \mu_M$, ou equivalentemente, $H_0 : \mu_H - \mu_M = 0$. Mas antes, devemos testar a hipótese $H_0' : \sigma_H^2 = \sigma_M^2$.

Teste T

O teste F para igualdade de variâncias é baseado na estatística $W = S_H^2/S_M^2 \sim F(n_H - 1, n_M - 1)$. Temos que $W \approx 0.79$. A região crítica do teste, a 5% de significância, é $RC(0.05) = \{[W < 0.567] \cup [W > 1.762]\}$. Repare que os pontos (0.567,1.762) são tais que $F_W(0.567; 49, 49) = 0.025$ e $F_W(1.762; 49, 49) = 0.975$, F_W é a função de distribuição acumulada da F, com 49 e 49 graus de liberdade.

Como $W = 0.79 \notin \{[W < 0.567] \cup [W > 1.762]\}$, não rejeitamos a hipótese nula $H_0' : \sigma_H^2 = \sigma_M^2$. Devemos aplicar o teste T para variâncias iguais, desconhecidas.

Teste T

O teste T com variâncias iguais mas desconhecidas é baseado na seguinte estatística:

$$T = \frac{\bar{X}_H - \bar{X}_M}{S_p \sqrt{\frac{1}{n_H} + \frac{1}{n_M}}} \sim t_{(n_H + n_M - 2)}$$

onde S_p, o desvio padrão comum (*pooled standard deviation*) é dado por

$$S_p^2 = \frac{(n_H - 1)S_H^2 + (n_M - 1)S_M^2}{n_H + n_M - 2}$$

No problema apresentado, $s_p = 0.8514$.

Teste T

A estatística observada foi

$$t_0 = \frac{3.2 - 3.7}{0.8514\sqrt{\frac{1}{50} + \frac{1}{50}}} = -2.9363$$

Note que a região crítica agora é dada por

$$RC(0.05) = \{[T < -1.984] \cup [T > 1.984]\}$$

onde $q = -1.984$ é o ponto tal que $P([T < q]) = 0.025$, etc.
E como $-2.9363 \in RC$, rejeitamos a hipótese nula. Ou seja, há evidência em favor da diferença entre o tempo médio de adaptação dos homens e das mulheres.

Teste T

Exemplo

Para investigar a influência da opção profissional sobre o salário inicial de recém-formados, investigaram-se dois grupos de profissionais: um de liberais em geral e outro de formados em Administração de Empresas. Com os resultados abaixo, expressos em salários mínimos, quais seriam suas conclusãoes?

Liberais	6,6	10,3	10,8	12,9	9,2	12,3	7,0	
Admin.	8,1	9,8	8,7	10,0	10,2	8,2	8,7	10,1

Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 366.

Teste T

Temos duas amostras de populações independentes. Vamos assumir que os salários tem distribuição normal, com média μ_L e variância σ_L^2 para os profissionais liberais, e média μ_A e variância σ_A^2 para os administradores.

Queremos testar a hipótese $H_0 : \mu_L = \mu_A$. Mas antes de tudo, queremos determinar se não rejeitamos a hipótese secundária $H_0' : \sigma_L^2 = \sigma_A^2$, para decidirmos qual tipo de teste T utilizaremos, pois a variância é desconhecida.

Teste T

Observe que a tabela nos dá os seguintes valores: $\bar{x}_L = 9.87$, $\bar{x}_A = 9.22$, $s_L = 2.43$ e $s_a = 0.88$, com $n_L = 7$ e $n_A = 8$.

O teste F para igualdade de variâncias é baseado na estatística $W = S_L^2/S_A^2 \sim F(n_L - 1, n_A - 1)$. Temos que $W = 7.513$. A região crítica do teste, a 5% de significância, é $RC(0.05) = \{[W < 0.175] \cup [W > 5.119]\}$. Novamente, os pontos (0.175,5.119) são tais que $F_W(0.175; 6, 7) = 0.025$ e $F_W(5.119; 6, 7) = 0.975$, F_W é a função de distribuição acumulada da F, com 6 e 7 graus de liberdade.

Como $W = 7.513 \in \{[W < 0.175] \cup [W > 5.119]\}$, rejeitamos a hipótese nula $H_0' : \sigma_L^2 = \sigma_A^2$. Devemos aplicar o teste T para variâncias diferentes, desconhecidas.

Teste T

A estatística do teste T com variâncias desconhecidas e desiguais é dada por

$$T = \frac{\bar{X}_L - \bar{X}_A}{\sqrt{S_L^2/n_L + S_A^2/n_A}}$$

com ν graus de liberdade, dados por

$$\nu = \frac{(C + D)^2}{C^2/(n_L - 1) + D^2/(n_A - 1)}$$

onde $C = s_L^2/n_L$ e $D = s_A^2/n_A$. Temos que $C = 0.84$ e $D = 0.10$, logo $\nu = 7.39 \approx 7$.

Teste T

A estatística T observada é dada por

$$t_0 = \frac{9.87 - 9.22}{\sqrt{(2.43)^2/7 + (0.88)^2/8}} = 0.67$$

A região crítica, a 5% de significância, é dada por uma t_ν com $\nu = 7$ graus de liberdade. Temos que

$$RC(0.05) = \{[T < -2.364] \cup [T > 2.364]\}$$

e como $t_0 \notin RC$, não rejeitamos a hipótese de igualdade de médias.

Teste de Aderência

Exemplo

Um modelo genético especifica que animais de certa população devam estar classificados em quatro categorias, com probabilidades $p_1 = 0.656$, $p_2 = 0.093$, $p_3 = 0.093$ e $p_4 = 0.158$. Dentre 197 animais, obtivemos as seguintes frequências observadas: $O_1 = 125$, $O_2 = 18$, $O_3 = 20$ e $O_4 = 34$. Teste se esses dados estão de acordo com o modelo genético postulado.
Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 395.

Teste de Aderência

Temos as probabilidades de cada categoria, e o número observado de indivíduos. Então o número esperado de indivíduos numa categoria é dado por $n \cdot p_c$, ou seja,
$E_1 = n \cdot p_1 = 197 \cdot 0.656 = 129.23$, $E_2 = E_3 = 197 \cdot 0.093 = 18.32$ e $E_4 = 197 \cdot 0.158 = 31.13$.

A estatística do teste de aderência é

$$Q = \sum_{i=1}^{s} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(s-1)}$$

Teste de Aderência

Nossa estatística observada Q_0 é dada por:

$$Q_0 = \frac{(125 - 129.23)^2}{129.23} + \frac{(18 - 18.32)^2}{18.32} + \frac{(20 - 18.32)^2}{18.32} + \frac{(34 - 31.13)^2}{31.13}$$

$$Q_0 = 0.5627$$

Note que a região crítica para os testes baseados na χ^2 é dada por $P(Q > q) = \alpha$. Para $\alpha = 0.05$ e 3 graus de liberdade, o valor de $q = 7.8147$. Portanto, não rejeitamos a hipótese de que os dados seguem a distribuição proposta.

Teste de Aderência

Exemplo

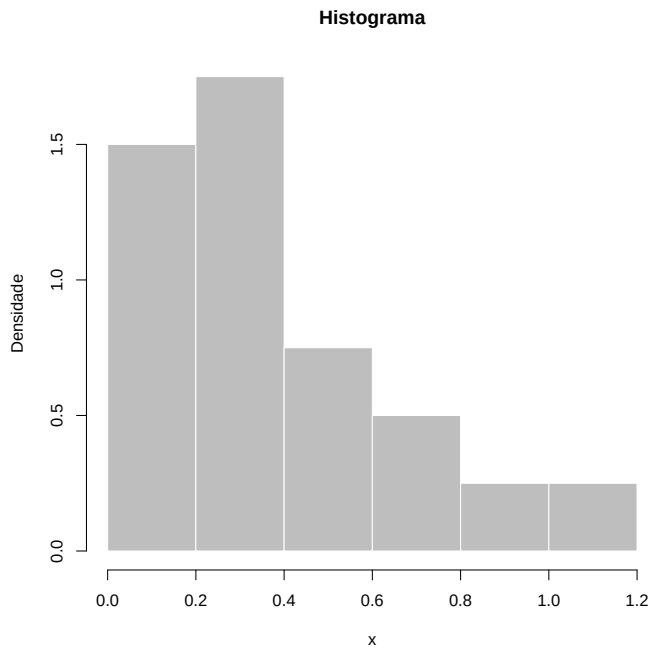
Teste, para o nível $\alpha = 0.01$, se os dados abaixo vem de uma distribuição exponencial com média 0.5.

0.378	0.391	0.458	0.063	0.009
1.007	0.470	0.368	0.831	0.387
0.228	0.389	0.627	0.480	0.093
0.123	0.089	0.646	0.093	0.400

Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 409.

Análise de Regressão

Antes de aplicar o teste, observe o histograma dos dados:



É razoável dizer que eles têm distribuição exponencial?

Teste de Aderência

Para distribuições contínuas, devemos construir a tabela de valores esperados/observados a partir dos quartis esperados/observados. Note que queremos testar se os dados têm distribuição exponencial, com $\lambda = 2$. Temos que os quantis teóricos são dados por:

$$q_1(X) \text{ é tal que } \int_0^{q_1} \lambda e^{-\lambda x} = 0.25$$

$$q_2(X) = \text{Med}(X) \text{ é tal que } \int_0^{q_2} \lambda e^{-\lambda x} = 0.5$$

$$q_3(X) \text{ é tal que } \int_0^{q_3} \lambda e^{-\lambda x} = 0.75$$

Teste de Aderência

Para $\lambda = 2$, temos que $q_1 = 0.1438$, $q_2 = 0.3466$ e $q_3 = 0.6931$.

Defina as categorias A_1, A_2, A_3 e A_4, onde

- um elemento $x \in A_1$ se $x < q_1(X)$,
- $x \in A_2$ se $q_1(X) < x < q_2(X)$,
- $x \in A_3$ se $q_2(X) < x < q_3(X)$ e
- $x \in A_4$ se $x > q_3(X)$.

Se a hipótese nula é verdadeira (isto é, os dados têm distribuição exponencial(2)), então a proporção esperada de cada categoria é $1/4$.

Teste de Aderência

Construimos então a tabela com as frequências observadas e esperadas:

	A_1	A_2	A_3	A_4	Total
O_i	6	1	11	2	20
E_i	5	5	5	5	20

A estatística observada Q_0 é dada por:

$$Q_0 = \frac{(6-5)^2}{5} + \frac{(1-5)^2}{5} + \frac{(11-5)^2}{5} + \frac{(2-5)^2}{5} = 12.4$$

que tem 3 graus de liberdade. Como $12.4 > 7.814 = \chi^2_3(0.95)$, rejeito $H_0 \Rightarrow$ os dados não tem distribuição exponencial(2).

Teste de Homogeneidade

Exemplo

Uma prova de Estatística foi aplicada a 100 alunos de Ciências Humanas e a 100 alunos de Ciências Biológicas. As notas são classificadas segundo os graus A, B, C, D e E (onde D significa que o aluno não recebe créditos e E indica que o aluno foi reprovado). Os resultados estão na seguinte tabela:

	A	B	C	D	E	Total
Humanas	15	20	30	20	15	100
Biológicas	8	23	18	34	17	100

Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 390.

Teste de Homogeneidade

Queremos aplicar o teste de homogeneidade, isto é, se P_H, a população de estudantes de ciências humanas, é igual a P_B, a população de estudantes de ciências biológicas. Ou seja, $H_0 : P_H = P_B$.

Note primeiro que, se H_0 é verdadeira, então deveriamos esperar as proporções de respostas de cada categoria iguais às da última linha da tabela, chamada "Total", que contém a soma das frequências observadas.

	A	B	C	D	E	Total
Humanas	15	20	30	20	15	100
Biológicas	8	23	18	34	17	100
Total	23	43	48	54	32	200

Teste de Homogeneidade

A estatística do teste de homogeneidade é dada por

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}$$

onde que n_{ij}^* é a frequência esperada. Podemos obtê-la considerando o valor na linha total, proporcional ao tamanho da amostra na linha. Por exemplo, o valor esperado de A nas ciências humanas é igual ao das biológicas, que é a proporção de A na amostra toda (23/200) vezes o número de indivíduos no estrato considerado (100 em cada), o que nos dá uma frequência esperada de 11.5.

Teste de Homogeneidade

A tabela esperada (isto é, sob a hipótese nula) é a seguinte:

	A	B	C	D	E	Total
Humanas	11.5	21.5	24	27	16	100
Biológicas	11.5	21.5	24	27	16	100
Total	23	43	48	54	32	200

De modo que a estatística observada do teste é dada por

$$Q_0 = \frac{(15 - 11.5)^2}{11.5} + \ldots + \frac{(17 - 16)^2}{16} = 9.09$$

Teste de Homogeneidade

Sabemos que sob a hipótese nula, a estatística do teste tem distribuição $Q \sim \chi^2$, com $n = (r-1)(s-1)$ graus de liberdade, onde r é o número de linhas e s é o número de colunas.

Observando a tabela da distribuição, com quatro graus de liberdade, vemos que com $\alpha = 0.05$, o valor encontrado é 9.488. Então a região crítica é $[Q > 9.488]$, e como $Q_0 \notin RC(0.05)$, não rejeitamos a hipótese nula $\Rightarrow$ não existe diferença significativa entre as populações.

Análise de Regressão

Exemplo

Considere as seguintes variáveis, X e Z, que representam a idade e a acuidade visual, respectivamente.

x	z	x	z	x	z	x	z	x	z
20	90	25	100	30	70	35	90	40	90
20	100	25	90	30	90	35	80	40	90
20	80	25	80	30	90	35	70	40	60
20	90	25	90	30	80	35	90	40	80

Fonte: Morettin & Bussab, Estatística Básica 5ª edição, pág 441.

Análise de Regressão

Exemplo

(a) Encontre a reta de quadrados mínimos $\hat{z}_i = \alpha + \beta x_i$, onde z mede a acuidade visual e x a idade do i-ésimo indivíduo.

(b) Interprete o significado de α e β nesse problema.

Análise de Regressão

(a) Temos que os estimadores para α e β são dados por:

$$\hat{\alpha} = \bar{z} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i z_i - n\bar{x}\bar{z}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Note agora que $n = 20$, $\bar{x} = 30$, $\bar{z} = 85$, $\sum_{i=1}^{n} x_i z_i = 50450$ e $\sum_{i=1}^{n} x_i^2 = 19000$. Com isso, temos que:

$$\hat{\beta} = \frac{50450 - 20 \cdot 30 \cdot 85}{19000 - 20 \cdot 30^2} = -\frac{11}{20} \Rightarrow \hat{\alpha} = 85 + \frac{11}{20}30 = 101.5$$

Análise de Regressão

(a) Temos que os estimadores para α e β são dados por:

$$\hat{\alpha} = \bar{z} - \hat{\beta}\bar{x}$$

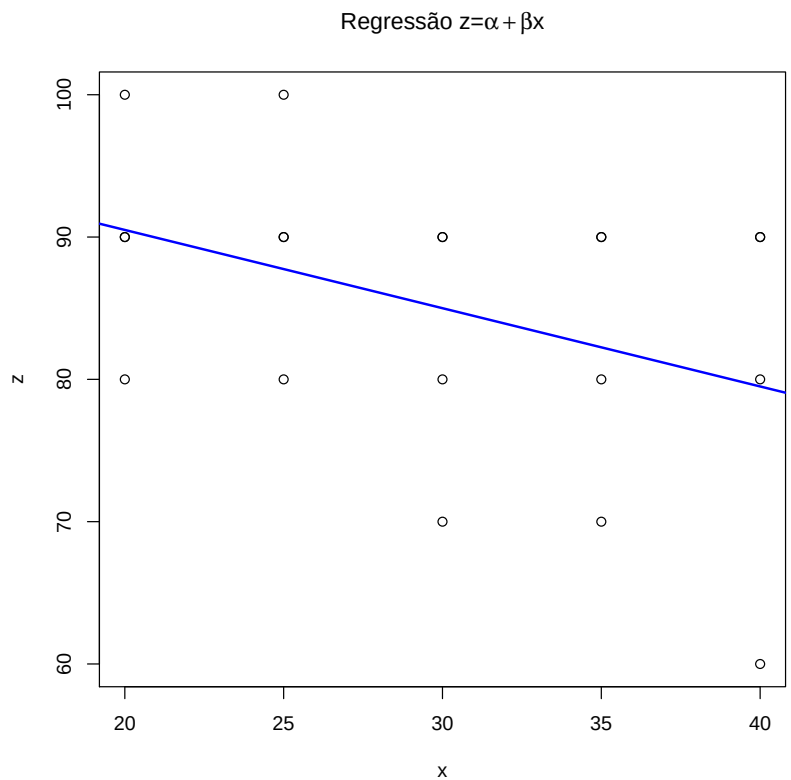
$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i z_i - n\bar{x}\bar{z}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Note agora que $n = 20$, $\bar{x} = 30$, $\bar{z} = 85$, $\sum_{i=1}^{n} x_i z_i = 50450$ e $\sum_{i=1}^{n} x_i^2 = 19000$. Com isso, temos que:

$$\hat{\beta} = \frac{50450 - 20 \cdot 30 \cdot 85}{19000 - 20 \cdot 30^2} = -\frac{11}{20} \Rightarrow \hat{\alpha} = 85 + \frac{11}{20}30 = 101.5$$

Análise de Regressão

(a) (cont.) O gráfico de dispersão dos dados, com a reta ajustada de regressão em azul, é dado por:



Análise de Regressão

(b) Neste problema, a interpretação dos parâmetros é a seguinte:

- α é o intercepto. Ele representa a acuidade visual na idade $z = 0$, se fosse possível medí-la. Mas $z = 0$ não faz parte do intervalo $[20, 40]$ de idades observadas, e não faz sentido falar em acuidade visual de recém nascidos, então sua interpretação deve ser feita com cuidado.
- β é mais interessante. O fato dele ser negativo significa que, a medida que os indivíduos envelhecem, sua acuidade visual diminui. Ela diminui na razão de -0.55 por ano, pelo modelo ajustado.