

# Interpretando a regressão linear por meio de modelos de estatísticas. Resumo ()



Tim McAleer

Seguir



5 de dezembro de 2020 · 7 min de leitura

Imagens retiradas de <https://www.statsmodels.org/>

Toda a codificação feita usando Python e a biblioteca de modelos de estatísticas do Python.

```
In [8]: mod = smf.ols(formula='Lottery ~ Literacy + Wealth + Region', data=df)
```

```
In [9]: res = mod.fit()
```

```
In [10]: print(res.summary())
```

```

                                OLS Regression Results
=====
Dep. Variable:                  Lottery    R-squared:                0.338
Model:                            OLS      Adj. R-squared:           0.287
Method:                 Least Squares    F-statistic:                6.636
Date:                  Sat, 28 Nov 2020    Prob (F-statistic):        1.07e-05
Time:                  14:39:43    Log-Likelihood:            -375.30
No. Observations:                85      AIC:                       764.6
Df Residuals:                    78      BIC:                       781.7
Df Model:                        6
Covariance Type:                nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	38.6517	9.456	4.087	0.000	19.826	57.478
Region[T.E]	-15.4278	9.727	-1.586	0.117	-34.793	3.938
Region[T.N]	-10.0170	9.260	-1.082	0.283	-28.453	8.419
Region[T.S]	-4.5483	7.279	-0.625	0.534	-19.039	9.943
Region[T.W]	-10.0913	7.196	-1.402	0.165	-24.418	4.235
Literacy	-0.1858	0.210	-0.886	0.378	-0.603	0.232
Wealth	0.4515	0.103	4.390	0.000	0.247	0.656

```

=====
Omnibus:                        3.049    Durbin-Watson:           1.785
Prob(Omnibus):                  0.218    Jarque-Bera (JB):        2.694
Skew:                          -0.340    Prob(JB):                0.260
Kurtosis:                       2.454    Cond. No.                 371.
=====

```

Vamos descobrir o que isso significa!

Não se deixe intimidar pelas palavras grandes e pelos números! Este blog está aqui para traduzir todas essas informações para o inglês simples. Nosso objetivo é fornecer uma visão geral de todas as estatísticas. Pesquisas adicionais são altamente recomendadas para uma análise aprofundada de cada componente.

Vamos começar no início.

```
In [8]: mod = smf.ols(formula='Lottery ~ Literacy + Wealth + Region', data=df)

In [9]: res = mod.fit()

In [10]: print(res.summary())
```

Codificando nosso resumo.

A linha de código anterior que está faltando aqui é Então, o que estamos fazendo aqui é usar a função `ols ()` ou Ordinary Least Squares fornecida da biblioteca `statsmodels`. OLS é uma técnica comum usada na análise de regressão linear. Em resumo, ele compara a diferença entre pontos individuais em seu conjunto de dados e a linha de melhor ajuste prevista para medir a quantidade de erro produzida. A função `smf.ols ()` requer duas entradas, a fórmula para produzir a linha de melhor ajuste e o conjunto de dados. `import statsmodels.formula.api as smf`

A fórmula é fornecida como uma string, na seguinte forma: 'variável dependente ~ lista de variáveis independentes separadas pelo símbolo +' Em termos simples, a variável dependente é o fator que você está tentando prever, e do outro lado do fórmula são as variáveis que você está usando para prever. O conjunto de dados neste caso é denominado 'df' e está sendo usado para determinar a aposta per capita na Real Loteria da França de 1830 usando algumas características. Para o propósito desta lição, os dados são irrelevantes, mas estão disponíveis <https://cran.r-project.org/web/packages/HistData/HistData.pdf> para seu interesse.

Nossa primeira linha de código cria um modelo, então o chamamos de 'mod' e a segunda usa o modelo para criar uma linha de melhor ajuste, daí a regressão linear. Chamamos-lhe 'res' porque analisa os resíduos do nosso modelo. Em seguida, imprimimos nosso resumo.

```
OLS Regression Results
=====
Dep. Variable:          Lottery    R-squared:                0.338
Model:                  OLS        Adj. R-squared:           0.287
Method:                 Least Squares    F-statistic:              6.636
Date:                  Sat, 28 Nov 2020    Prob (F-statistic):       1.07e-05
Time:                  14:39:43    Log-Likelihood:          -375.30
No. Observations:      85        AIC:                     764.6
Df Residuals:          78        BIC:                     781.7
Df Model:              6
Covariance Type:       nonrobust
=====
```

## Detalhes e estatísticas

O topo do nosso resumo começa nos dando alguns detalhes que já conhecemos. Nossa **variável dependente** é 'Loteria', estamos usando OLS conhecido como Ordinary Least Squares e a **data e hora em** que criamos o **modelo** . A seguir, ele detalha nosso **número de observações** no conjunto de dados. **Df Residuals** é outro nome para nossos graus de liberdade em nosso modo. Isso é calculado na forma de 'nk-1' ou 'número de observações-número de variáveis de previsão-1'. **O modelo Df** numera nossas variáveis de previsão. Se você está se perguntando por que inserimos apenas 3 variáveis de previsão na fórmula, mas os Resíduos Df e o Modelo estão dizendo que são 6, entraremos nisso mais tarde. Nosso **O tipo de covariância** é listado como não robusto. A covariância é uma medida de como duas variáveis estão vinculadas de maneira positiva ou negativa, e uma covariância robusta é aquela que é calculada de forma a minimizar ou eliminar variáveis, o que não é o caso aqui.

**R-quadrado** é possivelmente a medição mais importante produzida por este resumo. R-quadrado é a medida de quanto da variável independente é explicado por mudanças em nossas variáveis dependentes. Em termos percentuais, 0,338 significaria que nosso modelo explica 33,8% da mudança em nossa variável 'Loteria'. **R-quadrado ajustado** é importante

para analisar a eficácia de múltiplas variáveis dependentes no modelo. A regressão linear tem a qualidade de que o valor de R ao quadrado do seu modelo nunca diminuirá com variáveis adicionais, apenas iguais ou superiores. Portanto, seu modelo pode parecer mais preciso com várias variáveis, mesmo que elas estejam contribuindo mal. O R-quadrado ajustado penaliza a fórmula do R-quadrado com base no número de variáveis, portanto, uma pontuação ajustada mais baixa pode indicar que algumas variáveis não estão contribuindo para o R-quadrado do seu modelo de maneira adequada.

A **estatística F** na regressão linear está comparando seu modelo linear produzido para suas variáveis com um modelo que substitui o efeito de suas variáveis por 0, para descobrir se seu grupo de variáveis é *estatisticamente significativo*. Para interpretar esse número corretamente, é necessário usar um valor alfa escolhido e uma tabela F. **Prob (Estatística F)** usa esse número para informar a precisão da hipótese nula ou se é preciso que o efeito de suas variáveis seja 0. Nesse caso, ele está nos informando de 0,00107% de chance disso. **A probabilidade de log** é um significante numérico da probabilidade de que seu modelo produzido produziu os dados fornecidos. É usado para comparar os valores dos coeficientes de cada variável no processo de criação do modelo. **AIC** e o **BIC** são usados para comparar a eficácia dos modelos no processo de regressão linear,

usando um sistema de penalidade para medir múltiplas variáveis. Esses números são usados para seleção de recursos de variáveis.

	coef	std err	t	P> t	[0.025	0.975]
-----						
Intercept	38.6517	9.456	4.087	0.000	19.826	57.478
Region[T.E]	-15.4278	9.727	-1.586	0.117	-34.793	3.938
Region[T.N]	-10.0170	9.260	-1.082	0.283	-28.453	8.419
Region[T.S]	-4.5483	7.279	-0.625	0.534	-19.039	9.943
Region[T.W]	-10.0913	7.196	-1.402	0.165	-24.418	4.235
Literacy	-0.1858	0.210	-0.886	0.378	-0.603	0.232
Wealth	0.4515	0.103	4.390	0.000	0.247	0.656
=====						
<b>Omnibus:</b>		3.049	Durbin-Watson:			1.785
Prob(Omnibus):		0.218	Jarque-Bera (JB):			2.694
<b>Skew:</b>		-0.340	Prob(JB):			0.260
<b>Kurtosis:</b>		2.454	Cond. No.			371.
=====						

Em nossos coeficientes!

Agora vemos o trabalho da nossa modelo! Vamos decompô-lo.

A interceptação é o resultado de nosso modelo se todas as variáveis foram ajustadas para 0. Na fórmula linear clássica ' $y = mx + b$ ', é nosso  $b$ , uma constante adicionada para explicar um valor inicial para nossa linha.

Abaixo da interceptação estão nossas variáveis. Lembre-se de nossa fórmula? 'Loteria ~ Região + Alfabetização + Riqueza' Aqui vemos nossas variáveis dependentes representadas. Mas por que existem quatro versões diferentes de Region quando inserimos apenas uma? Simplificando, a fórmula espera valores contínuos na forma de números. Ao inserir a região com pontos de dados como strings, a fórmula separa cada string em categorias e analisa a categoria separadamente. Formatar seus dados com antecedência pode ajudá-lo a organizar e analisar isso de maneira adequada.

Nossa primeira coluna informativa é o coeficiente. Para nossa interceptação, é o valor da interceptação. Para cada variável, é a medida de como a mudança nessa variável afeta a variável independente. É o 'm' em ' $y = mx + b$ '. Uma unidade de mudança na variável dependente afetará o valor do coeficiente da variável de mudança na variável independente. Se o coeficiente for negativo, eles têm uma relação *inversa*. À medida que um sobe, o outro cai.

Nosso erro padrão é uma estimativa do desvio padrão do coeficiente, uma medida da quantidade de variação no coeficiente ao longo de seus pontos de dados. O t está relacionado e é uma medida da precisão com a qual o coeficiente foi medido. Um erro padrão baixo em comparação com um



coeficiente alto produz uma estatística  $t$  alta, o que significa uma significância alta para seu coeficiente.

$P > |t|$  é uma das estatísticas mais importantes do resumo. Ele usa a estatística  $t$  para produzir o *valor p*, uma medida da probabilidade de seu coeficiente ser medido por meio de nosso modelo por acaso. O valor  $p$  de 0,378 para Riqueza indica que há 37,8% de chance de a variável Riqueza não afetar a variável dependente, Loteria, e nossos resultados são produzidos por acaso. A análise adequada do modelo irá comparar o valor  $p$  a um *valor alfa* previamente estabelecido, ou um limite com o qual podemos aplicar significância ao nosso coeficiente. Um alfa comum é 0,05, que poucas de nossas variáveis passam nesta instância.

[0,025 e 0,975] são ambas medidas de valores de nossos coeficientes dentro de 95% de nossos dados, ou dentro de dois desvios padrão. Fora desses valores geralmente podem ser considerados outliers.

**Omnibus** descreve a normalidade da distribuição de nossos resíduos usando inclinação e curtose como medidas. Um 0 indicaria normalidade perfeita. **Prob (Omnibus)** é um teste estatístico que mede a probabilidade de os resíduos serem normalmente distribuídos. Um 1 indicaria uma distribuição perfeitamente normal. **Skew** é uma medida de simetria em

nossos dados, com 0 sendo simetria perfeita. **A curtose** mede o nível de pico de nossos dados, ou sua concentração em torno de 0 em uma curva normal. Uma curtose mais alta implica menos valores discrepantes.

**Durbin-Watson** é uma medida de homocedasticidade, ou uma distribuição uniforme de erros em nossos dados. A heterocedasticidade implicaria em uma distribuição desigual, por exemplo, à medida que o ponto de dados aumenta, o erro relativo aumenta. A homocedasticidade ideal ficará entre 1 e 2. **Jarque-Bera (JB)** e **Prob (JB)** são métodos alternativos de medição do mesmo valor que Omnibus e Prob (Omnibus) usando assimetria e curtose. Usamos esses valores para confirmar uns aos outros. **Número de condição** é uma medida da sensibilidade de nosso modelo em comparação com o tamanho das mudanças nos dados que está analisando. A multicolinearidade está fortemente implícita em um alto número de condição. Multicolinearidade um termo para descrever duas ou mais variáveis independentes que estão fortemente relacionadas entre si e estão afetando falsamente nossa variável prevista por redundância.

Nossas definições mal arranham a superfície de qualquer um desses tópicos. A pesquisa independente é fortemente encorajada para uma compreensão desses termos e como eles se relacionam uns com os outros. Esperamos que este blog tenha lhe dado uma compreensão suficiente para

começar a interpretar seu modelo e as maneiras pelas quais ele pode ser melhorado!

---

## Inscriva-se para as 10 melhores histórias

Por The Startup

Torne-se mais inteligente na construção do seu negócio. Inscreva-se para receber as 10 histórias mais lidas do Startup - entregues diretamente em sua caixa de entrada, duas vezes por mês. [Dê uma olhada.](#)

Receba este boletim informativo

Os emails serão enviados para gomcalsam@gmail.com .  
[Você não?](#)

Ciência de Dados

Regressão linear

Statsmodels

Estatísticas

### Saber mais.

Medium é uma plataforma aberta onde 170 milhões de leitores encontram um pensamento perspicaz e dinâmico. Aqui, vozes de especialistas e desconhecidas mergulham no

### Torne o Medium seu.

Siga os escritores, publicações e tópicos que são importantes para você e você os verá na sua página inicial e na sua caixa de entrada. Explorar

### Escreva uma história no Medium .

Se você tem uma história para contar, conhecimento para compartilhar ou uma perspectiva para oferecer - seja bem-vindo. É fácil e grátis postar seu pensamento sobre qualquer tópico. Comece um blog

âmago de qualquer tópico e trazem novas ideias  
à tona. Saber mais

CercaEscreverAjudaJurídico  
de