IAA018 - ESTATÍSTICA APLICADA II
Parte 1

Prof. Arno P. Schmitz

UFPR – Universidade Federal do Paraná

Tipos de Erros

Todo teste de hipótese possui erros associados ao próprio teste

Erro Tipo I: rejeição da hipótese nula quando esta é verdadeira. A probabilidade do "erro tipo I" é a probabilidade de concluir que existe relação entre duas variáveis quando na verdade não existe (que é devida ao acaso, aleatória). A probabilidade do erro do tipo I é expresso pelo nível de significância (α).

Portanto, a melhor técnica é minimizar o "erro tipo I" (α). Isto é conseguido apenas aplicando aos testes níveis de significância menores, tais como 0.05 ou 0.01 (5% ou 1%).

Tipos de Erros

Erro Tipo II: aceitação da hipótese nula quando esta é falsa.

Em um teste de médias de uma amostra aleatória, o “erro tipo II” significa a probabilidade de que a média de uma amostra aleatória seja igual ou superior que o valor calculado para a média amostral.

Exemplo: Testar a hipótese nula de que a média de todas as contas a receber é no mínimo \$260,00 com α = 0.05. O valor histórico da média da população é \$240,00. O desvio padrão de uma amostra é \$43,00 e o tamanho da amostra é 36.

$$\sigma_{\bar{x}} = \frac{43}{\sqrt{36}} = 7{,}17$$

$$\text{IC} = 260 - 1{,}65 \,.\, 7{,}17 = \$248{,}17$$

$$z = \frac{248{,}17 - 240{,}00}{7{,}17} = 1{,}14$$

$$P(erro\ tipo\ II) = P(z \geq 1{,}14)$$
$$= 0{,}50 - 0{,}3729 \cong 0{,}13\ (13\%)$$

Obs: 0,50 porque o teste é “≥” na cauda a direita; o valor de 0,3729 é o valor de “Z” para 1,14.

Tipos de Erros

Erro Tipo II: aceitação da hipótese nula quando esta é falsa.

Para o “erro tipo II” (β) relacionado a análise de regressão, esse erro depende do poder (ou potência) do teste. Pode-se diminuir o risco de cometer um erro do tipo II assegurando que o seu teste tenha potência suficiente.

A probabilidade de rejeitar a hipótese nula quando ela é falsa é igual a 1–β. Esse valor é a potência do teste.

A potência do teste tem o objetivo identificar o quanto o teste de hipóteses controla um erro do tipo II, ou seja, qual a probabilidade de rejeitar a hipótese nula se realmente for falsa. Em termos práticos, são importantes os testes com nível de significância próximos do nível de significância nominal e que o poder seja alto, mesmo em situações de amostras pequenas.

Tipos de Erros

Erro Tipo II: aceitação da hipótese nula quando esta é falsa.

O poder de um teste de hipóteses depende de três fatores:

- **Tamanho da amostra:** Mantendo todos os outros parâmetros iguais, quanto maior o tamanho da amostra, maior o poder do teste.
- **Nível de Significância:** Quanto mais elevado for o nível de significância, maior o poder do teste. Se aumentarmos o nível de significância, reduz-se a região de aceitação. Como resultado, você tem maior chance de rejeitar a hipótese nula. Isto significa que você tem menos chance de aceitar a hipótese nula quando ela é falsa, isto é, menor chance de cometer um erro do tipo II. Então, o poder do teste aumenta. Mas por outro lado não se pode aumentar indefinidamente o nível de significância (por exemplo 5%).
- **O verdadeiro valor do parâmetro a ser testado:** Quanto maior for a diferença entre o "verdadeiro" valor do parâmetro e o valor especificado pela hipótese nula, maior o poder do teste.

Tipos de Erros

Erro Tipo II: aceitação da hipótese nula quando esta é falsa.

Para um teste de hipóteses bi-caudal, que é o mais popular um análise de regressão, a potência do teste é dada por:

$$Poder = 1 - \Phi\left(Z_{\frac{\alpha}{2}} - \frac{\delta}{\sigma}\sqrt{n}\right) + \Phi\left(-Z_{\alpha/2} - \frac{\delta}{\sigma}\sqrt{n}\right)$$

Em que:

Φ = função distribuição acumulada de uma variável aleatória com distribuição normal padrão;

$Z_{\frac{\alpha}{2}}$ = valor tabelado de Z bilateral ao nível de significância escolhido;

δ = diferença entre as hipóteses nula e alternativa que se deseja que o poder do teste calcule;

σ = desvio padrão;

n = tamanho da amostra.

Níveis de Confiança

➔Em estudos socioeconômicos utiliza-se 0.05 (5%) de significância – que é a probabilidade de “erro tipo I”. Portanto, se a probabilidade de erro é 5%, o nível de confiança do teste é de 95% de confiança.

➔Em estudos da área de saúde pode-se usar, em alguns casos, 0.01 (1%) de significância – que é a probabilidade de “erro tipo I”. Logo, neste caso essa probabilidade de “erro tipo I” é menor que em dados socioeconômicos (0.05 ou 5%). Neste caso, para dados em saúde, então, pode ser utilizado o nível de confiança de 99%.

Obs: Outras áreas de pesquisa podem utilizar outros níveis de confiança, mas de maneira geral a maioria das áreas utiliza-se de 95% de confiança ou 5% (0.05) de significância.

Análise de Viés e Variância

Objetivos dos Modelos Estatísticos

Objetivo inferencial: Quais preditores (X) são importantes? Qual a relação entre cada preditor (X) e a variável resposta (Y)? Qual o efeito da mudança de valor de um dos preditores (X) na variável resposta (Y)?

Objetivo preditivo: a partir de uma função de regressão geral

$$r(x) \coloneqq E[Y|X = x]$$

Em que:

Y = variável dependente, variável de resposta;

x = vetor de variável explicativa.

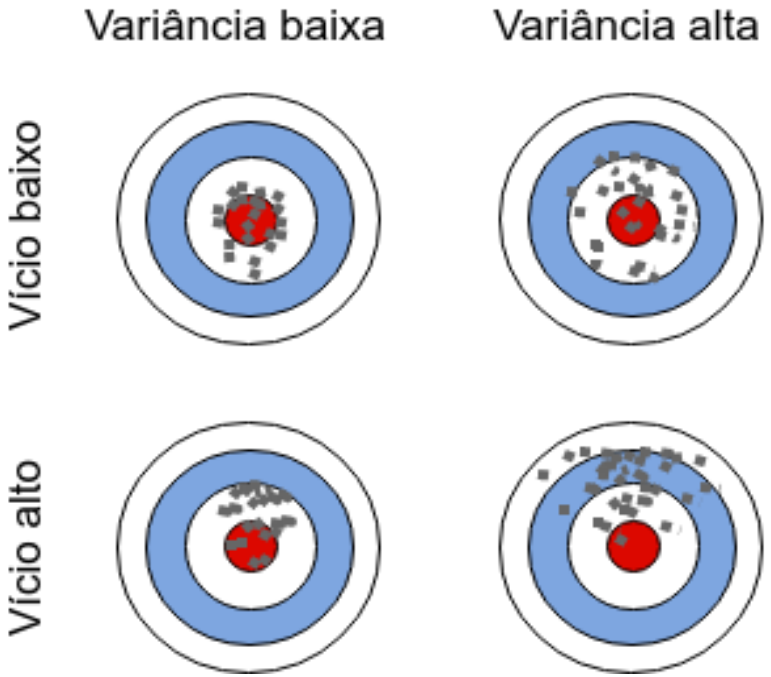
Para $Y \in \mathbb{R}$ como uma variável aleatória e o vetor $x = (x_1, \ldots\ldots, x_d) \in \mathbb{R}^d$

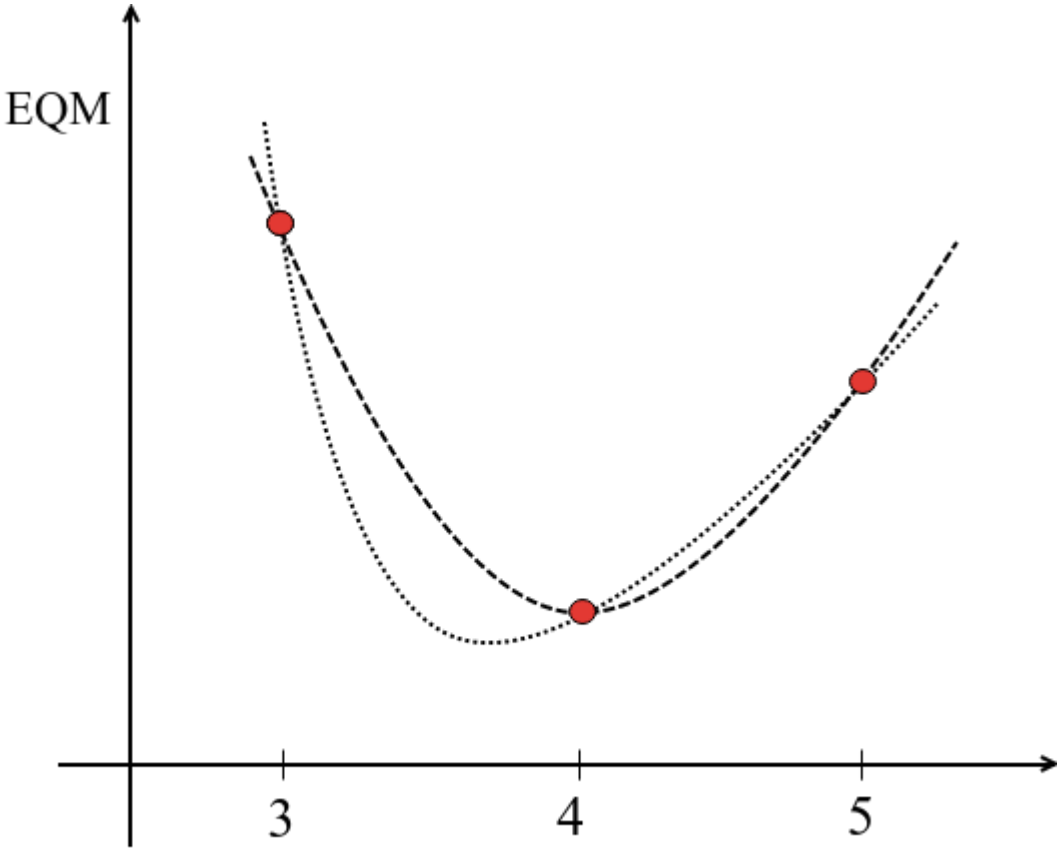
Análise de Viés e Variância

- O que se busca em modelos lineares (por MQO) é minimizar os resíduos, buscar a variância mínima. Mas nem sempre se consegue quando se tem **modelos heterocedásticos (variância não constante na amostra)**. Além disso, tem-se a "**inflação da variância**" dada pela **multicolinearidade**.
- Então, em geral os **modelos de MQO são mais precisos para diagnóstico do que para predição**. No caso de predição, torna-se mais importante um equilíbrio entre viés e variância.
- Nos modelos por **MQO, para reduzir-se o viés (erro) busca-se introduzir mais variáveis, nem sempre significativas no sentido de aumentar o poder explicativo do modelo.**
- A escolha de variáveis não é tarefa fácil, os **modelos stepwise são uma alternativa** dentro dos modelos lineares por MQO, mas estes tendem a reduzir o poder explicativo em favor de menor variância.

Análise de Viés e Variância

- No caso de modelos de **machine learning busca-se a predição de valores "novos"**, fora da amostra utilizada no modelo. Então, **pode-se permitir algum aumento do viés na estimativa dos parâmetros e obter decréscimos na função custo (EQM – Erro Quadrado Médio)**. Esse é o **trade-off entre viés e variância**.





Fonte: http://cursos.leg.ufpr.br/ML4all/apoio/Regularizacao.html

Análise de Viés e Variância

Objetivos dos Modelos Estatísticos

No objetivo preditivo deseja-se criar uma função

$$g: \mathbb{R}^d \longrightarrow \mathbb{R}$$

Que tenha um bom poder preditivo, ou seja, como criar g tal que dadas novas observações *i.i.d.* $(X_{n+1}, Y_{n+1}), \ldots\ldots\ldots, (X_{n+m}, Y_{n+m})$ tenha-se

$$g(x_{n+1}) \approx\ y_{n+1}, \cdots, g(x_{n+m}) \approx y_{n+m}$$

i.i.d. $=$ independente e identicamente distribuídas

Análise de Viés e Variância

O primeiro passo para construir boas funções de predição é criar um critério para medir a performance da função de predição $g: \mathbb{R}^d \longrightarrow \mathbb{R}$. Isso pode ser feito através do risco quadrático:

$$R_{pred}(g) = E[(Y - g(X))^2]$$

Em que:

X e Y representam uma nova observação que não foi utilizada para estimar g.

➔ Quanto menor o risco, melhor é a função de predição g.

➔O risco é definido como a esperança matemática de uma função de perda. Por exemplo, a função de perda quadrática:

$$L\big(g;\ (X,Y)\big) = (Y\ - g(X))^2$$

Análise de Viés e Variância

➔Então a performance do estimador é baseada no risco quadrático e criar uma boa função de predição ”g” equivale a encontrar um bom estimador para a função de regressão “$r(x)$”.

➔Estimar uma função de regressão é o melhor caminho para se criar uma função de predição de novas observações Y com base em covariáveis (x) – variáveis explicativas.

➔ Quando introduzimos uma nova observação (Y e x) na amostra, que não foi utilizada para estimar ”g” , então:

$$R_{pred}(g) = R_{pred}(g) + E[V[Y|X]]$$

Análise de Viés e Variância

➔Então a performance do estimador é baseada no risco quadrático e criar uma boa função de predição "g" equivale a encontrar um bom estimador para a função de regressão "$r(x)$".

➔Estimar uma função de regressão é o melhor caminho para se criar uma função de predição de novas observações Y com base em covariáveis (x) – variáveis explicativas.

➔ Quando introduzimos uma nova observação (Y e x) na amostra, que não foi utilizada para estimar "g" , então:

$$R_{pred}(g) = R_{pred}(g) + E[V[Y|X]]$$

Se a função "g" é entendida como fixa, então $R(g)$ é chamado de risco condicional. Mas se "g" é aleatória, então o risco é chamado de risco esperado.

Análise de Viés e Variância

➔ Denotamos o risco preditivo por R ao invés de R_{pred}.

➔Para uma função ”g” fixa, o risco condicional é baixo pois, dado um novo conjunto de dados (Y e x) $\mathrm{i.i.d}$ em uma amostra “grande” a lei dos grandes números garante que o risco seja baixo.

Sub-ajuste ou underfitting = modelos muito simples que não são suficientes para explicar o comportamento dos dados.

Super-ajuste ou overfitting = modelos que se ajustam demais a uma dada amostra e que possuem poder de generalização baixo.

Exemplo: modelos com p=1 (linear); p=4; e p=50.

Análise de Viés e Variância

➔ Risco observado = erro quadrático médio no conjunto de treinamento; é um estimador muito otimista do risco real. Este tende a levar o estimador ao super-ajuste, isto porque "g" foi escolhida para ajustar bem Y e X.

➔Uma solução é dividir a amostra em, por exemplo 70% (treinamento) e 30% (validação) – Data Splitting.

➔Utiliza-se a amostra de treinamento para estimar "g" (estimar os coeficientes da regressão) e usa-se o conjunto de validação para estimar $R(g)$ via validação cruzada.

➔A boa prática para selecionar as observações da amostra para compor os grupos de treinamento e validação é por **aleatoriedade**.

➔Como o grupo de validação não foi utilizado para estimar os parâmetros de "g" , o estimador de erro quadrático médio é consistente pela lei dos grandes números.

Análise de Viés e Variância

➔ Neste caso $R(g)$ por validação cruzada é dado por:

$$R(g) \approx \frac{1}{n-s}\sum_{i=s+1}^{n} L\big(g;(X_i,Y_i)\big) \coloneqq \hat{R}(g)$$

*Pode ser utilizado também o K-fold cross validation

➔Segundo a lei dos grandes números, a estimação do risco baseada na divisão treinamento versus validação fornece um estimador consistente para o erro condicional.

➔A especificação do modelo não garante melhor poder explicativo e preditivo. Por exemplo: um modelo pode estar corretamente especificado segundo testes de especificação (p.ex. RESET) e apresentar baixo poder explicativo e pouco poder preditivo. Ater-se a teoria e aos estudos já feitos são um bom caminho a seguir.

Análise de Viés e Variância

➔ Quando a função de predição é muito otimista no conjunto de validação e treinamento, uma saída pode ser a divisão da amostra em 3 partes: validação, treinamento e teste.

➔A amostra de teste é usada para estimar o erro do melhor estimador da regressão encontrado segundo o conjunto de validação.

➔Utilizando o conjunto de teste, podemos também fazer um intervalo de confiança para o risco.

➔Uma forma alternativa de se estimar o risco de um certo modelo g é utilizando uma medida de penalização ou complexidade. Quanto mais parâmetros no modelo, mais o erro quadrático médio observado, EQM(g) subestima R(g), isto é, maior a diferença entre EQM(g) e R(g).

Análise de Viés e Variância

➔ A ideia por trás de métodos de penalização é criar uma medida de complexidade para g, P(g), que é utilizada para corrigir essa diferença. Por exemplo, P(g) pode ser o número de parâmetros do modelo. Pode-se então compensar o quão subestimado R(g) é adicionando estas duas quantidades.

➔Funções de penalização: AIC e BIC, por exemplo.

➔O risco pode ser decomposto em 3 partes:

$$E\left[\left(Y-\hat{g}(X)\right)^2\middle|\, X=x\right] = \underbrace{V[Y|X=x]}_{a} + \underbrace{(r(x)-E[\hat{g}(x)])^2}_{b} + \underbrace{V[\hat{g}(x)]}_{c}$$

Análise de Viés e Variância

a) A variância intrínseca da variável resposta (Y), que não depende da função $\hat{g}$ escolhida e, assim, não pode ser reduzida;

b) o quadrado do viés do estimador $\hat{g}$;

c) Variância do estimador;

- "b" e "c" podem ser reduzidos se escolhida a função $\hat{g}$ adequada.

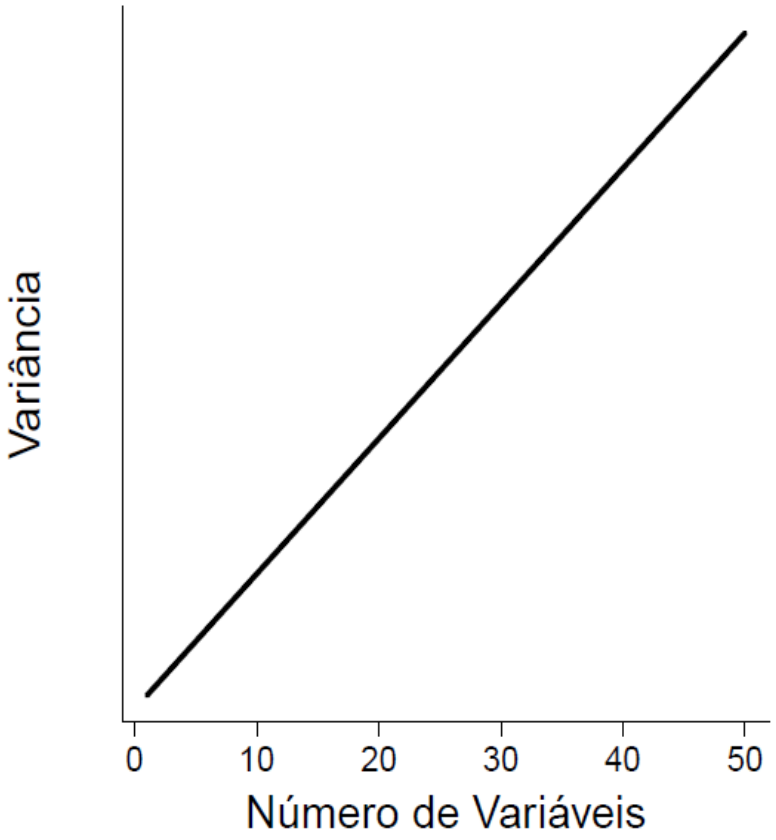
➔Modelos com muitos parâmetros possuem viés relativamente baixo, mas variância alta, já que é necessário estimar todos eles.

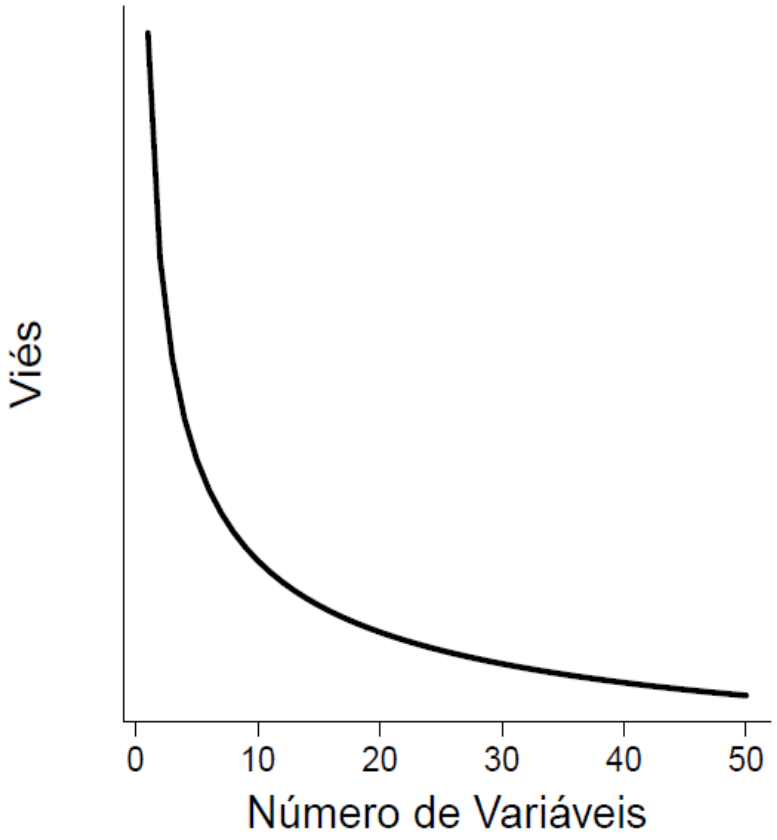
➔Modelos com poucos parâmetros possuem variância baixa, mas viés muito alto, já que são demasiado simplistas para descrever o modelo gerador dos dados.

➔Com a finalidade de obter um bom poder preditivo deve-se escolher um número de parâmetros nem tão alto, nem tão baixo.

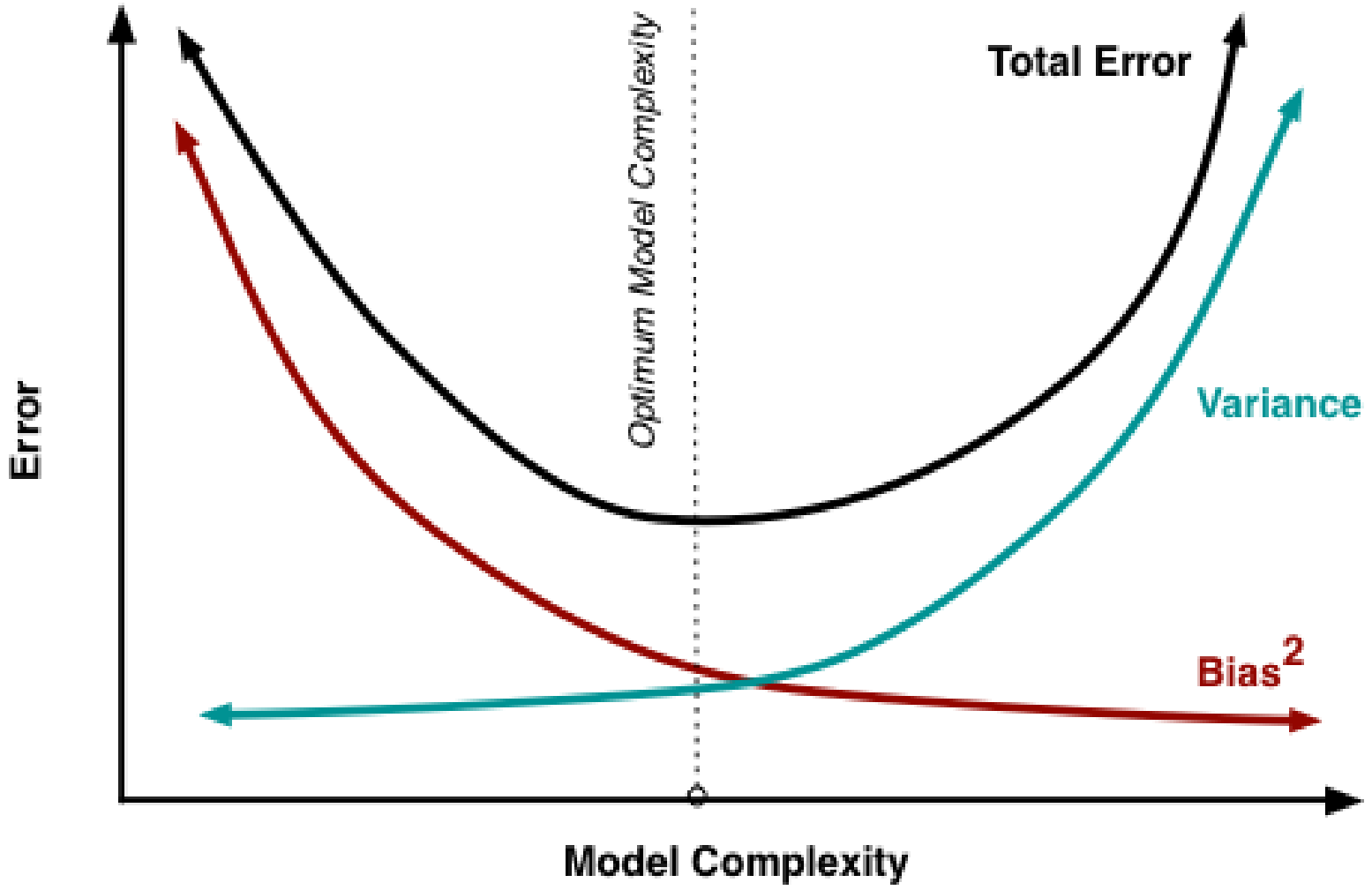
Análise de Viés e Variância

Tradeoff entre Viés e Variância





Análise de Viés e Variância



Fonte:https://stats.stackexchange.com/questions/336433/bias-variance-tradeoff-math

Análise de Viés e Variância

Tradeoff entre Viés e Variância

➔O número de parâmetros (variáveis explicativas do modelo) controlam o balanço entre viés e variância.

➔O valor ótimo de parâmetros depende do tamanho da amostra e de $r(x)$, ou seja, do modelo de regressão utilizado.

➔O número ideal de parâmetros (variáveis explicativas) pode ser escolhido por validação cruzada.

Análise de Viés e Variância

.

Tabela 3.2: Resultados dos métodos de seleção de variáveis no exemplo da Seção 3.8. *: Busca pelo melhor subconjunto.

Método	**Variáveis Selecionadas**	**Tempo de ajuste**	**Risco Estimado**
Mínimos Quadrados	Todas	0.002 segundos	14.63 (0.02)
Melhor AIC*	$x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$	1 hora e 20 minutos	0.30 (0.02)
Forward Stepwise	$x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$	0.46 segundos	0.30 (0.02)
Ridge	Todas	0.19 segundos	0.33 (0.03)
Lasso	x_1, x_2, x_3, x_4, x_5	0.08 segundos	0.25 (0.02)

Modelos de Regularização

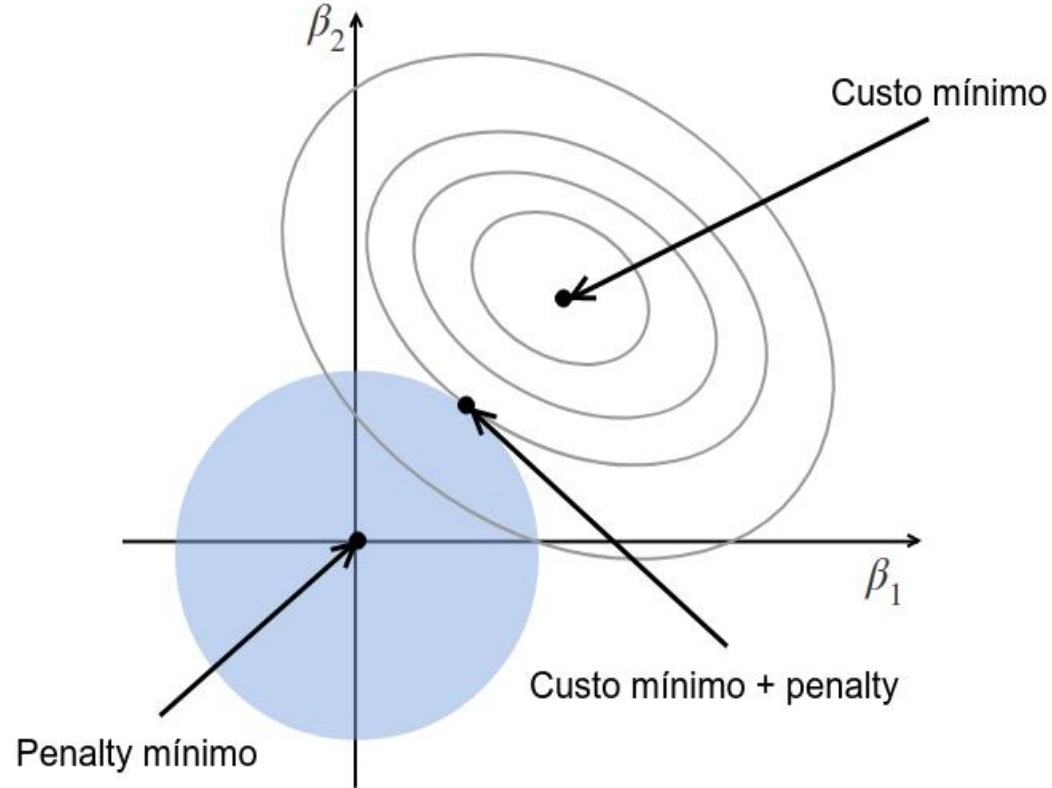
- Neste contexto surgem os **modelos com regularização ou penalização** em que deixa-se de lado a tarefa de excluir variáveis (seja por ausência de significância estatística ou por multicolinearidade).
- Os métodos de regularização são aconselhados, pois permitem cenários contínuos do domínio da função custo (**introdução de quantas variáveis estiverem disponíveis**) e **lidam bem** com casos em que tem-se **muitas variáveis e amostras relativamente pequenas.**
- Esses modelos incorporam uma **restrição (penalty)** às estimativas dos parâmetros de MQO:

$$\hat{\beta}^{Restrito} = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 , s.a. \quad g(\beta) < t$$

- Em que $g(\beta)$ é a função "penalty" (shrinkage penalty); t é um escalar entre zero e infinito.

Modelos de Regularização

- O papel da função penaty é manter as estimativas de β_j próximas de zero (regulando-as). Quanto menor o valor de t, maior o penalty.
- A figura abaixo representa essa situação (nesse caso, $g(\beta) = \beta_1^2 + \beta_2^2$). Nosso objetivo é encontrar os valores de β que representam um custo mínimo, restrito à região em azul.



Modelos de Regularização

- Implementa-se o processo de penalização por meio dos **Multiplicadores de Lagrange**, aumentando a função objetivo, da seguinte forma:

$$\hat{\beta}^{Restrito} = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda g(\beta)$$

- em que λ é um escalar entre zero e infinito. Trata-se de um tuning parameter, determinado isoladamente. A medida que λ cresce, a flexibilidade do modelo diminui (reduzindo a variância e aumentando o viés).

Modelos de Regularização

- Quando a regularização pertence à família das potências, temos a seguinte especificação:

$$\hat{\beta}^{Restrito} = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

dependendo da escolha de q>0, obtêm-se diferentes penalizações:

i. q=2 → penalização Ridge (**Modelo de regressão Ridge**);
ii. q=1 → penalização Lasso (**Modelo de regressão Lasso**);

Modelos de Regularização

- Quando o último termo é alterado como abaixo tem-se a penalização ElasticNet Ou seja, um **modelo de regressão ElasticNet**:

$$\hat{\beta}^{Restrito} = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left(\alpha \left| \beta_j \right| + (1-\alpha) \beta_j^2 \right)$$

Para $0 \leq \alpha \leq 1$

Então, quando α = 1 ElasticNet e Lasso são iguais; quando α = 0 ElasticNet e Ridge são iguais.

FIM DA PARTE 1 !!!