Universidade Federal do Paraná

Laboratório de Estatística e Geoinformação - LEG

Regularização

Prof.: Eduardo Vargas Ferreira

Introdução

- Estudaremos nesta seção formas de aprimorar os modelos lineares através da substituição do ajuste por mínimos quadrados ordinários por procedimentos alternativos;

- Mas, por que estudar tais formas se o modelo linear é tão simples e fácil de estimar?

  - **Precisão da predição:** especialmente quando $p > n$ ("*small n, large p*"), para controlar a variância;

  - **Interpretação do modelo:** removendo características irrelevantes, obtemos modelos mais simples e fáceis de interpretar;

- Nesta aula, abordaremos os mecanismos automáticos de seleção das variáveis (regulando a entrada das mesmas).

Introdução

- Estudaremos nesta seção formas de aprimorar os modelos lineares através da substituição do ajuste por mínimos quadrados ordinários por procedimentos alternativos;

- Mas, por que estudar tais formas se o modelo linear é tão simples e fácil de estimar?

  - **Precisão da predição:** especialmente quando $p > n$ ("*small n, large p*"), para controlar a variância;

  - **Interpretação do modelo:** removendo características irrelevantes, obtemos modelos mais simples e fáceis de interpretar;

- Nesta aula, abordaremos os mecanismos automáticos de seleção das variáveis (regulando a entrada das mesmas).

Introdução

- Estudaremos nesta seção formas de aprimorar os modelos lineares através da substituição do ajuste por mínimos quadrados ordinários por procedimentos alternativos;

- Mas, por que estudar tais formas se o modelo linear é tão simples e fácil de estimar?

  - ⋆ **Precisão da predição:** especialmente quando $p > n$ ("*small n, large p*"), para controlar a variância;

  - ⋆ **Interpretação do modelo:** removendo características irrelevantes, obtemos modelos mais simples e fáceis de interpretar;

- Nesta aula, abordaremos os mecanismos automáticos de seleção das variáveis (regulando a entrada das mesmas).

Introdução

- Estudaremos nesta seção formas de aprimorar os modelos lineares através da substituição do ajuste por mínimos quadrados ordinários por procedimentos alternativos;

- Mas, por que estudar tais formas se o modelo linear é tão simples e fácil de estimar?

  - ⋆ **Precisão da predição:** especialmente quando $p > n$ ("*small n, large p*"), para controlar a variância;

  - ⋆ **Interpretação do modelo:** removendo características irrelevantes, obtemos modelos mais simples e fáceis de interpretar;

- Nesta aula, abordaremos os mecanismos automáticos de seleção das variáveis (regulando a entrada das mesmas).

Introdução

- Estudaremos nesta seção formas de aprimorar os modelos lineares através da substituição do ajuste por mínimos quadrados ordinários por procedimentos alternativos;

- Mas, por que estudar tais formas se o modelo linear é tão simples e fácil de estimar?

  - ⋆ **Precisão da predição:** especialmente quando $p > n$ ("*small n, large p*"), para controlar a variância;

  - ⋆ **Interpretação do modelo:** removendo características irrelevantes, obtemos modelos mais simples e fáceis de interpretar;

- Nesta aula, abordaremos os mecanismos automáticos de seleção das variáveis (regulando a entrada das mesmas).

Métodos de seleção de variáveis

- Será discutido o seguinte problema: temos p variáveis explanatórias e queremos escolher o(s) modelo(s) de regressão linear múltipla com base em todas ou subconjuntos dessas variáveis.
- Supondo que o intercepto sempre faça parte da regressão, então temos 2^p possíveis modelos.
- Isto significa que com $p = 4$ temos que analisar 16 regressões.
- E com $p = 10$ teríamos que analisar 1024 ajustes.
- Nesta última situação é útil dispor de mecanismos para escolher modelos evitando ajustar todas as possibilidades.

Métodos de seleção de variáveis

- Será discutido o seguinte problema: temos p variáveis explanatórias e queremos escolher o(s) modelo(s) de regressão linear múltipla com base em todas ou subconjuntos dessas variáveis.

- Supondo que o intercepto sempre faça parte da regressão, então temos 2^p possíveis modelos.

- Isto significa que com $p = 4$ temos que analisar 16 regressões.

- E com $p = 10$ teríamos que analisar 1024 ajustes.

- Nesta última situação é útil dispor de mecanismos para escolher modelos evitando ajustar todas as possibilidades.

Métodos de seleção de variáveis

- Será discutido o seguinte problema: temos p variáveis explanatórias e queremos escolher o(s) modelo(s) de regressão linear múltipla com base em todas ou subconjuntos dessas variáveis.
- Supondo que o intercepto sempre faça parte da regressão, então temos 2^p possíveis modelos.
- Isto significa que com $p = 4$ temos que analisar 16 regressões.
- E com $p = 10$ teríamos que analisar 1024 ajustes.
- Nesta última situação é útil dispor de mecanismos para escolher modelos evitando ajustar todas as possibilidades.

Métodos de seleção de variáveis

- Será discutido o seguinte problema: temos p variáveis explanatórias e queremos escolher o(s) modelo(s) de regressão linear múltipla com base em todas ou subconjuntos dessas variáveis.
- Supondo que o intercepto sempre faça parte da regressão, então temos 2^p possíveis modelos.
- Isto significa que com $p = 4$ temos que analisar 16 regressões.
- E com $p = 10$ teríamos que analisar 1024 ajustes.
- Nesta última situação é útil dispor de mecanismos para escolher modelos evitando ajustar todas as possibilidades.

Métodos de seleção de variáveis

- Será discutido o seguinte problema: temos p variáveis explanatórias e queremos escolher o(s) modelo(s) de regressão linear múltipla com base em todas ou subconjuntos dessas variáveis.

- Supondo que o intercepto sempre faça parte da regressão, então temos 2^p possíveis modelos.

- Isto significa que com $p = 4$ temos que analisar 16 regressões.

- E com $p = 10$ teríamos que analisar 1024 ajustes.

- Nesta última situação é útil dispor de mecanismos para escolher modelos evitando ajustar todas as possibilidades.

Métodos sequenciais automáticos

Métodos sequenciais automáticos

- Mesmo com um número moderado de variáveis explanatórias, digamos 10, é conveniente dispor de algum roteiro para examinar um subconjunto de todas as possíveis regressões.

- Neste sentido, na literatura tem sido proposto métodos automáticos, os quais comparam basicamente a SQReg extra.

- Estudaremos três métodos:

  - **Best subset selection**: Para $k = 1, \ldots, p$, ajustamos todas os $\binom{p}{k}$ modelos que contém exatamente k preditores;

  - **Seleção Forward**: começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

  - **Eliminação Backward**: começamos com o modelo completo e vamos eliminando variáveis uma a uma.

Métodos sequenciais automáticos

- Mesmo com um número moderado de variáveis explanatórias, digamos 10, é conveniente dispor de algum roteiro para examinar um subconjunto de todas as possíveis regressões.

- Neste sentido, na literatura tem sido proposto métodos automáticos, os quais comparam basicamente a SQReg extra.

- Estudaremos três métodos:

  - **Best subset selection**: Para $k = 1, \ldots, p$, ajustamos todas os $\binom{p}{k}$ modelos que contém exatamente k preditores;

  - **Seleção Forward**: começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

  - **Eliminação Backward**: começamos com o modelo completo e vamos eliminando variáveis uma a uma.

Métodos sequenciais automáticos

- Mesmo com um número moderado de variáveis explanatórias, digamos 10, é conveniente dispor de algum roteiro para examinar um subconjunto de todas as possíveis regressões.

- Neste sentido, na literatura tem sido proposto métodos automáticos, os quais comparam basicamente a SQReg extra.

- Estudaremos três métodos:

  - **Best subset selection**: Para $k = 1, \ldots, p$, ajustamos todas os $\binom{p}{k}$ modelos que contém exatamente k preditores;

  - **Seleção Forward**: começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

  - **Eliminação Backward**: começamos com o modelo completo e vamos eliminando variáveis uma a uma.

Métodos sequenciais automáticos

- Mesmo com um número moderado de variáveis explanatórias, digamos 10, é conveniente dispor de algum roteiro para examinar um subconjunto de todas as possíveis regressões.

- Neste sentido, na literatura tem sido proposto métodos automáticos, os quais comparam basicamente a SQReg extra.

- Estudaremos três métodos:

  - ⋆ **Best subset selection**: Para $k = 1, \ldots, p$, ajustamos todas os $\binom{p}{k}$ modelos que contém exatamente k preditores;

  - ⋆ **Seleção Forward**: começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

  - ⋆ **Eliminação Backward**: começamos com o modelo completo e vamos eliminando variáveis uma a uma.

Métodos sequenciais automáticos

- Mesmo com um número moderado de variáveis explanatórias, digamos 10, é conveniente dispor de algum roteiro para examinar um subconjunto de todas as possíveis regressões.

- Neste sentido, na literatura tem sido proposto métodos automáticos, os quais comparam basicamente a SQReg extra.

- Estudaremos três métodos:

  - ⋆ **Best subset selection**: Para $k = 1, \ldots, p$, ajustamos todas os $\binom{p}{k}$ modelos que contém exatamente k preditores;

  - ⋆ **Seleção Forward**: começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

  - ⋆ **Eliminação Backward**: começamos com o modelo completo e vamos eliminando variáveis uma a uma.

Métodos sequenciais automáticos

- Mesmo com um número moderado de variáveis explanatórias, digamos 10, é conveniente dispor de algum roteiro para examinar um subconjunto de todas as possíveis regressões.

- Neste sentido, na literatura tem sido proposto métodos automáticos, os quais comparam basicamente a SQReg extra.

- Estudaremos três métodos:

  - ⋆ **Best subset selection**: Para $k = 1, \ldots, p$, ajustamos todas os $\binom{p}{k}$ modelos que contém exatamente k preditores;

  - ⋆ **Seleção Forward**: começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

  - ⋆ **Eliminação Backward**: começamos com o modelo completo e vamos eliminando variáveis uma a uma.

Best subset selection

1. Seja $\mathcal{M}_0$ o modelo nulo (que não contém preditores).
2. Para $k = 1, \ldots, p$:
   (a) Ajuste os $\binom{p}{k}$ modelos que contém exatamente k preditores;
   (b) Escolha o melhor dentre esses $\binom{p}{k}$ modelos, denotando-o por $\mathcal{M}_k$ (o melhor será definido segundo algum critério - veremos adiante).
3. Selecione o melhor dos $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ modelos.

Best subset selection

1. Seja $\mathcal{M}_0$ o modelo nulo (que não contém preditores).

2. Para $k = 1, \ldots, p$:

   (a) Ajuste os $\binom{p}{k}$ modelos que contém exatamente k preditores;

   (b) Escolha o melhor dentre esses $\binom{p}{k}$ modelos, denotando-o por $\mathcal{M}_k$ (o melhor será definido segundo algum critério - veremos adiante).

3. Selecione o melhor dos $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ modelos.

Best subset selection

1. Seja $\mathcal{M}_0$ o modelo nulo (que não contém preditores).

2. Para $k = 1, \ldots, p$:

   (a) Ajuste os $\binom{p}{k}$ modelos que contém exatamente k preditores;

   (b) Escolha o melhor dentre esses $\binom{p}{k}$ modelos, denotando-o por $\mathcal{M}_k$ (o melhor será definido segundo algum critério - veremos adiante).

3. Selecione o melhor dos $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ modelos.

Best subset selection

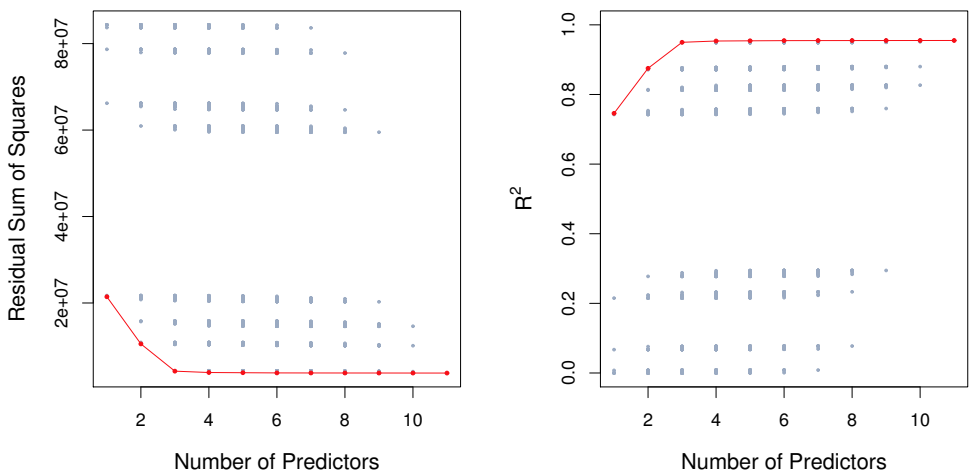
1. Seja $\mathcal{M}_0$ o modelo nulo (que não contém preditores).

2. Para $k = 1, \ldots, p$:

   (a) Ajuste os $\binom{p}{k}$ modelos que contém exatamente k preditores;

   (b) Escolha o melhor dentre esses $\binom{p}{k}$ modelos, denotando-o por $\mathcal{M}_k$ (o melhor será definido segundo algum critério - veremos adiante).

3. Selecione o melhor dos $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ modelos.



Stepwise selection

- Por razões computacionais, o best subset selection não pode ser aplicado quando temos um número grande de variáveis;
- Ele também sofre com problemas estatísticos quando p é grande: chance alta de encontrar um modelo que parece bom aos dados de treino, mas sem poder de predição de dados futuros;
- Ou seja, eu escolherei um modelo que se ajusta MUITO bem aos dados de treino ocasionando o **overftting**;
- Por estes motivos, os métodos por **stepwise** são mais atrativos. Estes não exploram todos os possíveis modelos, mas sim um conjunto mais restrito.

Stepwise selection

- Por razões computacionais, o best subset selection não pode ser aplicado quando temos um número grande de variáveis;
- Ele também sofre com problemas estatísticos quando p é grande: chance alta de encontrar um modelo que parece bom aos dados de treino, mas sem poder de predição de dados futuros;
- Ou seja, eu escolherei um modelo que se ajusta MUITO bem aos dados de treino ocasionando o **overftting**;
- Por estes motivos, os métodos por **stepwise** são mais atrativos. Estes não exploram todos os possíveis modelos, mas sim um conjunto mais restrito.

Stepwise selection

- Por razões computacionais, o best subset selection não pode ser aplicado quando temos um número grande de variáveis;
- Ele também sofre com problemas estatísticos quando p é grande: chance alta de encontrar um modelo que parece bom aos dados de treino, mas sem poder de predição de dados futuros;
- Ou seja, eu escolherei um modelo que se ajusta MUITO bem aos dados de treino ocasionando o **overftting**;
- Por estes motivos, os métodos por **stepwise** são mais atrativos. Estes não exploram todos os possíveis modelos, mas sim um conjunto mais restrito.

Stepwise selection

- Por razões computacionais, o best subset selection não pode ser aplicado quando temos um número grande de variáveis;
- Ele também sofre com problemas estatísticos quando p é grande: chance alta de encontrar um modelo que parece bom aos dados de treino, mas sem poder de predição de dados futuros;
- Ou seja, eu escolherei um modelo que se ajusta MUITO bem aos dados de treino ocasionando o **overftting**;
- Por estes motivos, os métodos por **stepwise** são mais atrativos. Estes não exploram todos os possíveis modelos, mas sim um conjunto mais restrito.

Seleção Forward

1. Começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

2. Em cada passo:

   (a) O modelo base tem q termos. Ajustamos as $(p-q)$ regressões (correspondentes as potenciais $(p-q)$ variáveis explanatórias) e registramos as somas de quadrados da regressão de cada modelo.

   (b) Identificar a variável X_i, cuja incorporação ao modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q+1} = SQReg_q + SQReg_{extra}$.

   (c) Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_{q+1}/(q+1)}.$$

Seleção Forward

1. Começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

2. Em cada passo:

   (a) O modelo base tem q termos. Ajustamos as $(p-q)$ regressões (correspondentes as potenciais $(p-q)$ variáveis explanatórias) e registramos as somas de quadrados da regressão de cada modelo.

   (b) Identificar a variável X_i, cuja incorporação ao modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q+1} = SQReg_q + SQReg_{extra}$.

   (c) Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_{q+1}/(q+1)}.$$

Seleção Forward

1. Começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

2. Em cada passo:

   (a) O modelo base tem q termos. Ajustamos as $(p-q)$ regressões (correspondentes as potenciais $(p-q)$ variáveis explanatórias) e registramos as somas de quadrados da regressão de cada modelo.

   (b) Identificar a variável X_i, cuja incorporação ao modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q+1} = SQReg_q + SQReg_{extra}$.

   (c) Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_{q+1}/(q+1)}.$$

Seleção Forward

1. Começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;

2. Em cada passo:

   (a) O modelo base tem q termos. Ajustamos as $(p-q)$ regressões (correspondentes as potenciais $(p-q)$ variáveis explanatórias) e registramos as somas de quadrados da regressão de cada modelo.

   (b) Identificar a variável X_i, cuja incorporação ao modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q+1} = SQReg_q + SQReg_{extra}$.

   (c) Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_{q+1}/(q+1)}.$$

Exemplo

Variáveis	SQR	QMR	R^2	C_p	AIC	BIC
Constante	2715.76	226.31	0	442.99	263.95	275.67
X_1	1265.69	115.06	0.53	202.55	156.51	170.73
X_2	906.34	82.39	0.67	142.49	112.07	122.25
X_3	1939.40	176.31	0.29	315.15	239.83	261.61
X_4	883.87	80.35	0.67	138.73	109.30	119.22
X_1, X_2	57.90	5.79	0.98	2.68	9.50	10.83
X_1, X_3	1227.07	122.71	0.55	198.09	194.68	221.79
X_1, X_4	74.76	7.48	0.97	5.50	11.87	13.52
X_2, X_3	415.44	41.54	0.85	62.44	65.90	75.08
X_2, X_4	868.88	86.89	0.68	138.23	137.85	157.05
X_3, X_4	175.74	17.57	0.94	22.37	27.88	31.76
X_1, X_2, X_3	48.11	5.35	0.98	3.04	9.90	11.78
X_1, X_3, X_4	50.84	5.64	0.98	3.50	10.44	12.42
X_1, X_2, X_4	47.97	5.33	0.98	3.02	9.86	11.73
X_2, X_3, X_4	73.81	8.20	0.97	7.34	15.17	18.05
X_1, X_2, X_3, X_4	47.86	5.98	0.98	5.00	12.91	16.04

Eliminação Backward

1. Começamos com o modelo completo e eliminamos variáveis uma a uma;

2. Em cada passo:

   (a) Seja o modelo base formado por q termos. Ajustamos as $(q-1)$ regressões e registramos as correspondentes SQReg.

   (b) Identificar a variável X_i, cuja eliminação do modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q-1} = SQReg_q - SQReg_{extra}$.

   (c) Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_q/q}.$$

Eliminação Backward

1. Começamos com o modelo completo e eliminamos variáveis uma a uma;

2. Em cada passo:

   (a) Seja o modelo base formado por q termos. Ajustamos as $(q-1)$ regressões e registramos as correspondentes SQReg.

   (b) Identificar a variável X_i, cuja eliminação do modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q-1} = SQReg_q - SQReg_{extra}$.

   (c) Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_q/q}.$$

Eliminação Backward

1. Começamos com o modelo completo e eliminamos variáveis uma a uma;

2. Em cada passo:

   (a) Seja o modelo base formado por q termos. Ajustamos as $(q-1)$ regressões e registramos as correspondentes SQReg.

   (b) Identificar a variável X_i, cuja eliminação do modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q-1} = SQReg_q - SQReg_{extra}$.

   (c) Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_q/q}.$$

Eliminação Backward

1. Começamos com o modelo completo e eliminamos variáveis uma a uma;

2. Em cada passo:

   (a) Seja o modelo base formado por q termos. Ajustamos as $(q-1)$ regressões e registramos as correspondentes SQReg.

   (b) Identificar a variável X_i, cuja eliminação do modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q-1} = SQReg_q - SQReg_{extra}$.

   (c) Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_q/q}.$$

Conclusão

- Diferentes métodos sequenciais não necessariamente conduzem ao mesmo modelo;
- Mais ainda, em geral nada garante que o *melhor* subconjunto de regressoras seja escolhido;
- Na prática, é melhor identificar vários *bons modelos* candidatos para serem examinados, do que aceitar sem criticas o modelo escolhido pelo computador;
- É importante considerar os modelos no contexto da aplicação prática, e a natureza das variáveis;
- Adicionalmente, os modelos escolhidos devem ser validados, utilizando técnicas de diagnóstico, e submetidos à avaliação junto ao pesquisador antes de se tirar conclusões.

Conclusão

- Diferentes métodos sequenciais não necessariamente conduzem ao mesmo modelo;
- Mais ainda, em geral nada garante que o *melhor* subconjunto de regressoras seja escolhido;
- Na prática, é melhor identificar vários *bons modelos* candidatos para serem examinados, do que aceitar sem criticas o modelo escolhido pelo computador;
- É importante considerar os modelos no contexto da aplicação prática, e a natureza das variáveis;
- Adicionalmente, os modelos escolhidos devem ser validados, utilizando técnicas de diagnóstico, e submetidos à avaliação junto ao pesquisador antes de se tirar conclusões.

Conclusão

- Diferentes métodos sequenciais não necessariamente conduzem ao mesmo modelo;
- Mais ainda, em geral nada garante que o *melhor* subconjunto de regressoras seja escolhido;
- Na prática, é melhor identificar vários *bons modelos* candidatos para serem examinados, do que aceitar sem criticas o modelo escolhido pelo computador;
- É importante considerar os modelos no contexto da aplicação prática, e a natureza das variáveis;
- Adicionalmente, os modelos escolhidos devem ser validados, utilizando técnicas de diagnóstico, e submetidos à avaliação junto ao pesquisador antes de se tirar conclusões.

Conclusão

- Diferentes métodos sequenciais não necessariamente conduzem ao mesmo modelo;
- Mais ainda, em geral nada garante que o *melhor* subconjunto de regressoras seja escolhido;
- Na prática, é melhor identificar vários *bons modelos* candidatos para serem examinados, do que aceitar sem criticas o modelo escolhido pelo computador;
- É importante considerar os modelos no contexto da aplicação prática, e a natureza das variáveis;
- Adicionalmente, os modelos escolhidos devem ser validados, utilizando técnicas de diagnóstico, e submetidos à avaliação junto ao pesquisador antes de se tirar conclusões.

Conclusão

- Diferentes métodos sequenciais não necessariamente conduzem ao mesmo modelo;
- Mais ainda, em geral nada garante que o *melhor* subconjunto de regressoras seja escolhido;
- Na prática, é melhor identificar vários *bons modelos* candidatos para serem examinados, do que aceitar sem criticas o modelo escolhido pelo computador;
- É importante considerar os modelos no contexto da aplicação prática, e a natureza das variáveis;
- Adicionalmente, os modelos escolhidos devem ser validados, utilizando técnicas de diagnóstico, e submetidos à avaliação junto ao pesquisador antes de se tirar conclusões.

Escolhendo o modelo ótimo

- Estamos interessados no modelo com menor erro no teste (o erro do treinamento é uma estimativa pobre do erro do teste);
- Sabemos ainda que o modelo que contém todas as regressoras sempre produzirá menores resíduos e o maior R^2 (quantidades relacionadas ao treinamento);
- Assim, a soma de quadrados dos resíduos e R^2 não são critérios adequados para escolher o melhor modelo. Veremos agora outras abordagens:
  - ⋆ Estatística C_p de Mallows;
  - ⋆ Critério de Informação de Akaike (AIC);
  - ⋆ Critério de Informação Bayesiano (BIC);
  - ⋆ Coeficiente de Determinação Ajustado ($\bar{R}^2$).

Escolhendo o modelo ótimo

- Estamos interessados no modelo com menor erro no teste (o erro do treinamento é uma estimativa pobre do erro do teste);
- Sabemos ainda que o modelo que contém todas as regressoras sempre produzirá menores resíduos e o maior R^2 (quantidades relacionadas ao treinamento);
- Assim, a soma de quadrados dos resíduos e R^2 não são critérios adequados para escolher o melhor modelo. Veremos agora outras abordagens:
  - Estatística C_p de Mallows;
  - Critério de Informação de Akaike (AIC);
  - Critério de Informação Bayesiano (BIC);
  - Coeficiente de Determinação Ajustado ($\bar{R}^2$).

Escolhendo o modelo ótimo

- Estamos interessados no modelo com menor erro no teste (o erro do treinamento é uma estimativa pobre do erro do teste);
- Sabemos ainda que o modelo que contém todas as regressoras sempre produzirá menores resíduos e o maior R^2 (quantidades relacionadas ao treinamento);
- Assim, a soma de quadrados dos resíduos e R^2 não são critérios adequados para escolher o melhor modelo. Veremos agora outras abordagens:
  - ⋆ Estatística C_p de Mallows;
  - ⋆ Critério de Informação de Akaike (AIC);
  - ⋆ Critério de Informação Bayesiano (BIC);
  - ⋆ Coeficiente de Determinação Ajustado ($\bar{R}^2$).

Escolhendo o modelo ótimo

- Estamos interessados no modelo com menor erro no teste (o erro do treinamento é uma estimativa pobre do erro do teste);
- Sabemos ainda que o modelo que contém todas as regressoras sempre produzirá menores resíduos e o maior R^2 (quantidades relacionadas ao treinamento);
- Assim, a soma de quadrados dos resíduos e R^2 não são critérios adequados para escolher o melhor modelo. Veremos agora outras abordagens:
  - ⋆ Estatística C_p de Mallows;
  - ⋆ Critério de Informação de Akaike (AIC);
  - ⋆ Critério de Informação Bayesiano (BIC);
  - ⋆ Coeficiente de Determinação Ajustado ($\bar{R}^2$).

Escolhendo o modelo ótimo

- Estamos interessados no modelo com menor erro no teste (o erro do treinamento é uma estimativa pobre do erro do teste);
- Sabemos ainda que o modelo que contém todas as regressoras sempre produzirá menores resíduos e o maior R^2 (quantidades relacionadas ao treinamento);
- Assim, a soma de quadrados dos resíduos e R^2 não são critérios adequados para escolher o melhor modelo. Veremos agora outras abordagens:
  - ⋆ Estatística C_p de Mallows;
  - ⋆ Critério de Informação de Akaike (AIC);
  - ⋆ Critério de Informação Bayesiano (BIC);
  - ⋆ Coeficiente de Determinação Ajustado ($\bar{R}^2$).

Escolhendo o modelo ótimo

- Estamos interessados no modelo com menor erro no teste (o erro do treinamento é uma estimativa pobre do erro do teste);
- Sabemos ainda que o modelo que contém todas as regressoras sempre produzirá menores resíduos e o maior R^2 (quantidades relacionadas ao treinamento);
- Assim, a soma de quadrados dos resíduos e R^2 não são critérios adequados para escolher o melhor modelo. Veremos agora outras abordagens:
  - ⋆ Estatística C_p de Mallows;
  - ⋆ Critério de Informação de Akaike (AIC);
  - ⋆ Critério de Informação Bayesiano (BIC);
  - ⋆ Coeficiente de Determinação Ajustado ($\bar{R}^2$).

Métodos de seleção de variáveis

(a) **Estatística C_p de Mallows:** Seja d o número total de parâmetros,

$$C_p = \frac{1}{n}\left(SQRes_d + 2d\hat{\sigma}^2\right).$$

preferimos os modelos tais que $C_p < d$, e entre estes, com menores C_p.

(b) O **Critério de Informação de Akaike (AIC)** é dado por

$$AIC = -2\log L + 2d.$$

preferimos o modelo com menor AIC.

(c) **Critério de Informação Bayesiano, BIC:** Já que com bastante frequência o AIC conduz a escolher o modelo com maior número de termos, Schwarz (1978) propôs o BIC como

$$BIC = \frac{1}{n}\left(SQRes_k + \log(n)d\hat{\sigma}^2\right).$$

preferimos o modelo com menor BIC.

Métodos de seleção de variáveis

(a) **Estatística C_p de Mallows:** Seja d o número total de parâmetros,

$$C_p = \frac{1}{n}\left(SQRes_d + 2d\hat{\sigma}^2\right).$$

preferimos os modelos tais que $C_p < d$, e entre estes, com menores C_p.

(b) O **Critério de Informação de Akaike (AIC)** é dado por

$$AIC = -2\log L + 2d.$$

preferimos o modelo com menor AIC.

(c) **Critério de Informação Bayesiano, BIC:** Já que com bastante frequência o AIC conduz a escolher o modelo com maior número de termos, Schwarz (1978) propôs o BIC como

$$BIC = \frac{1}{n}\left(SQRes_k + \log(n)d\hat{\sigma}^2\right).$$

preferimos o modelo com menor BIC.

Métodos de seleção de variáveis

(a) **Estatística C_p de Mallows:** Seja d o número total de parâmetros,

$$C_p = \frac{1}{n}\left(SQRes_d + 2d\hat{\sigma}^2\right).$$

preferimos os modelos tais que $C_p < d$, e entre estes, com menores C_p.

(b) O **Critério de Informação de Akaike (AIC)** é dado por

$$AIC = -2\log L + 2d.$$

preferimos o modelo com menor AIC.

(c) **Critério de Informação Bayesiano, BIC:** Já que com bastante frequência o AIC conduz a escolher o modelo com maior número de termos, Schwarz (1978) propôs o BIC como

$$BIC = \frac{1}{n}\left(SQRes_k + \log(n)d\hat{\sigma}^2\right).$$

preferimos o modelo com menor BIC.

Métodos de seleção de variáveis

(d) **Coeficiente de Determinação Ajustado, $\bar{R}^2$:** É preferível ao R^2 já que não necessariamente aumenta quando incluímos variáveis no modelo.

$$\bar{R}^2 = 1 - \frac{SQRes_d/(n-d-1)}{SQT/(n-1)}.$$

- Note que o BIC substituiu $2d\hat{\sigma}^2$ (utilizado no C_p) por $\log(n)d\hat{\sigma}^2$;
- E como $\log(n) > 2$ para qualquer $n > 7$, geralmente o BIC penaliza mais os modelos com muitas variáveis do que o C_p;

Métodos de seleção de variáveis

(d) **Coeficiente de Determinação Ajustado, $\bar{R}^2$:** É preferível ao R^2 já que não necessariamente aumenta quando incluímos variáveis no modelo.

$$\bar{R}^2 = 1 - \frac{SQRes_d/(n-d-1)}{SQT/(n-1)}.$$

- Note que o BIC substituiu $2d\hat{\sigma}^2$ (utilizado no C_p) por $\log(n)d\hat{\sigma}^2$;
- E como $\log(n) > 2$ para qualquer $n > 7$, geralmente o BIC penaliza mais os modelos com muitas variáveis do que o C_p;

Métodos de seleção de variáveis

(d) **Coeficiente de Determinação Ajustado, $\bar{R}^2$:** É preferível ao R^2 já que não necessariamente aumenta quando incluímos variáveis no modelo.

$$\bar{R}^2 = 1 - \frac{SQRes_d/(n-d-1)}{SQT/(n-1)}.$$

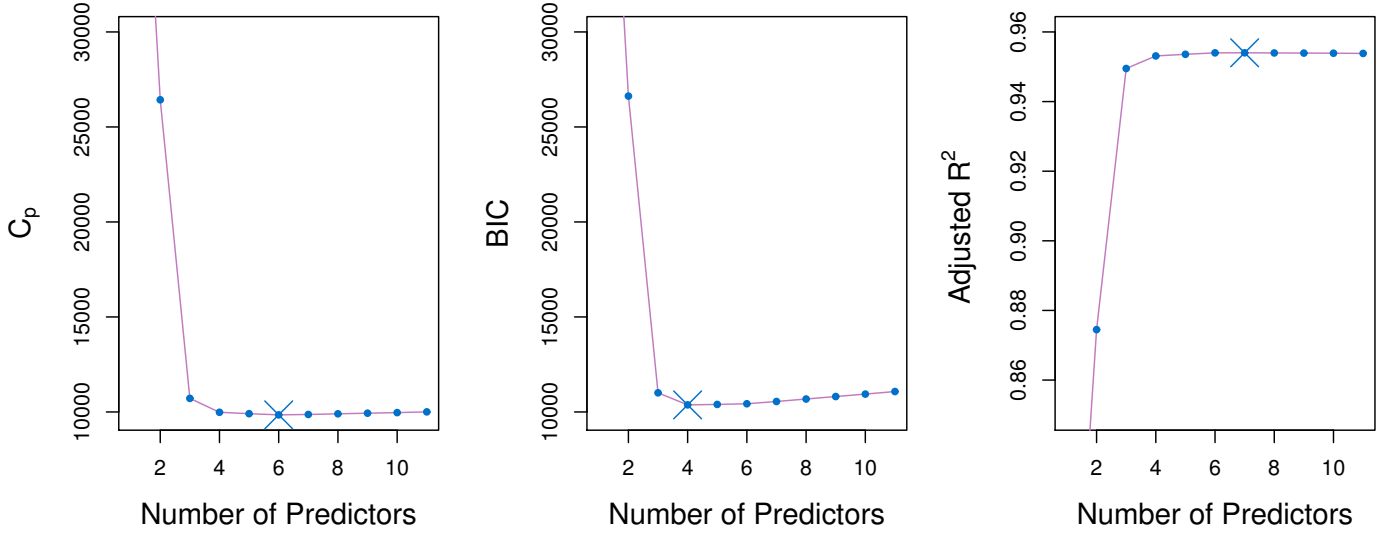
- Note que o BIC substituiu $2d\hat{\sigma}^2$ (utilizado no C_p) por $\log(n)d\hat{\sigma}^2$;
- E como $\log(n) > 2$ para qualquer $n > 7$, geralmente o BIC penaliza mais os modelos com muitas variáveis do que o C_p;

Métodos de seleção de variáveis

(d) **Coeficiente de Determinação Ajustado, $\bar{R}^2$:** É preferível ao R^2 já que não necessariamente aumenta quando incluímos variáveis no modelo.

$$\bar{R}^2 = 1 - \frac{SQRes_d/(n-d-1)}{SQT/(n-1)}.$$

- Note que o BIC substituiu $2d\hat{\sigma}^2$ (utilizado no C_p) por $\log(n)d\hat{\sigma}^2$;
- E como $\log(n) > 2$ para qualquer $n > 7$, geralmente o BIC penaliza mais os modelos com muitas variáveis do que o C_p;



Multiplicadores de Lagrange

Introdução

- O problema de otimização restrita é expresso de maneira geral como

$$\min_{x_1,x_2,\ldots,x_n} (\text{ou } max)\ J = f(x_1, x_2, \ldots, x_n), \quad \text{sujeito a} \quad g(x_1, x_2, \ldots, x_n) = 0$$

- Para resolvê-lo, em algumas situações, podemos incorporar a função de restrição à função a ser otimizada;

- Por exemplo,

$$\max_{x,y} x^2 y, \quad \text{sujeito a} \quad x + y = 5.$$

$$\max_{x} x^2(5 - x), \quad \text{pois} \quad y = 5 - x.$$

Introdução

- O problema de otimização restrita é expresso de maneira geral como

$$\min_{x_1,x_2,\ldots,x_n} (\text{ou } max)\ J = f(x_1, x_2, \ldots, x_n), \quad \text{sujeito a} \quad g(x_1, x_2, \ldots, x_n) = 0$$

- Para resolvê-lo, em algumas situações, podemos incorporar a função de restrição à função a ser otimizada;

- Por exemplo,

$$\max_{x,y} x^2 y, \quad \text{sujeito a} \quad x + y = 5.$$

$$\max_{x} x^2(5 - x), \quad \text{pois} \quad y = 5 - x.$$

Introdução

- O problema de otimização restrita é expresso de maneira geral como

$$\min_{x_1,x_2,\ldots,x_n} (\text{ou } max)\ J = f(x_1, x_2, \ldots, x_n), \quad \text{sujeito a} \quad g(x_1, x_2, \ldots, x_n) = 0$$

- Para resolvê-lo, em algumas situações, podemos incorporar a função de restrição à função a ser otimizada;

- Por exemplo,

$$\max_{x,y} x^2 y, \quad \text{sujeito a} \quad x + y = 5.$$

$$\max_{x} x^2(5 - x), \quad \text{pois} \quad y = 5 - x.$$

Multiplicadores de Lagrange

- Entretanto, a maioria dos problemas de otimização não são tão fáceis e requerem outros métodos de otimização;
- E o método de **Multiplicadores de Lagrange** é um deles;
- A ideia é modificar (aumentar) a função objetivo, $J = f(\mathbf{x})$, através da adição de termos que descrevam as restrições;

$$J_A(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2, \ldots, \lambda_m) = f(x_1, x_2, \ldots, x_n) + \sum_{j=1}^{m} \lambda_j g_j(x_1, \ldots, x_n),$$

em que

  - J_A é a função objetivo modificada;
  - λ_j são os Multiplicadores de Lagrange;

Multiplicadores de Lagrange

- Entretanto, a maioria dos problemas de otimização não são tão fáceis e requerem outros métodos de otimização;
- E o método de **Multiplicadores de Lagrange** é um deles;
- A ideia é modificar (aumentar) a função objetivo, $J = f(\mathbf{x})$, através da adição de termos que descrevam as restrições;

$$J_A(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2, \ldots, \lambda_m) = f(x_1, x_2, \ldots, x_n) + \sum_{j=1}^{m} \lambda_j g_j(x_1, \ldots, x_n),$$

em que

  - J_A é a função objetivo modificada;
  - λ_j são os Multiplicadores de Lagrange;

Multiplicadores de Lagrange

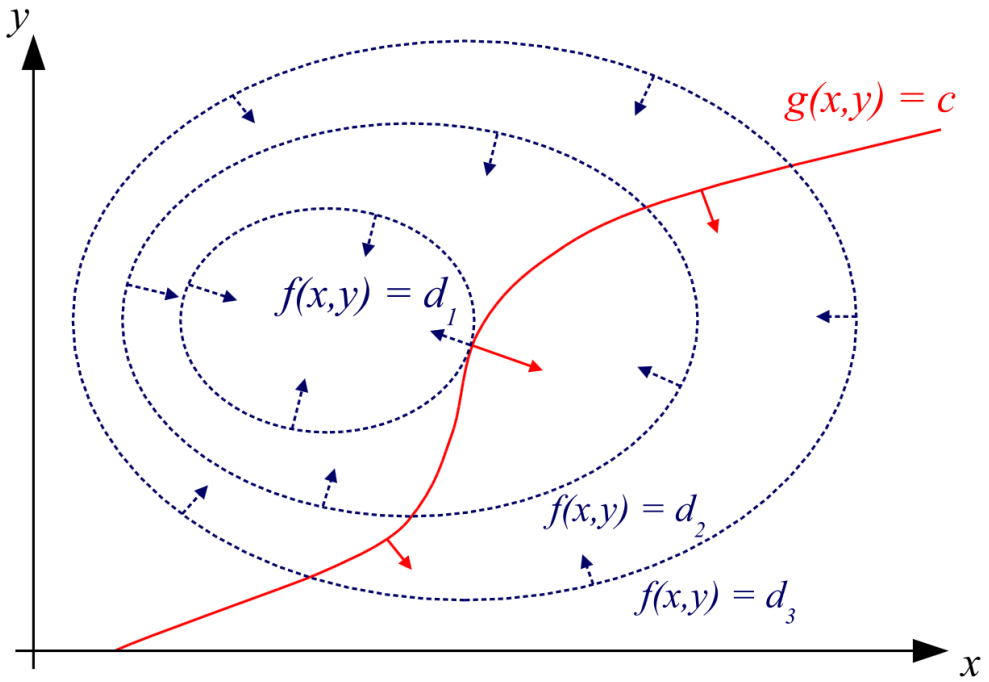
- Entretanto, a maioria dos problemas de otimização não são tão fáceis e requerem outros métodos de otimização;
- E o método de **Multiplicadores de Lagrange** é um deles;
- A ideia é modificar (aumentar) a função objetivo, $J = f(\mathbf{x})$, através da adição de termos que descrevam as restrições;

$$J_A(x_1, x_2, \ldots, x_n, \lambda_1, \lambda_2, \ldots, \lambda_m) = f(x_1, x_2, \ldots, x_n) + \sum_{j=1}^{m} \lambda_j g_j(x_1, \ldots, x_n),$$

em que

- J_A é a função objetivo modificada;
- λ_j são os Multiplicadores de Lagrange;

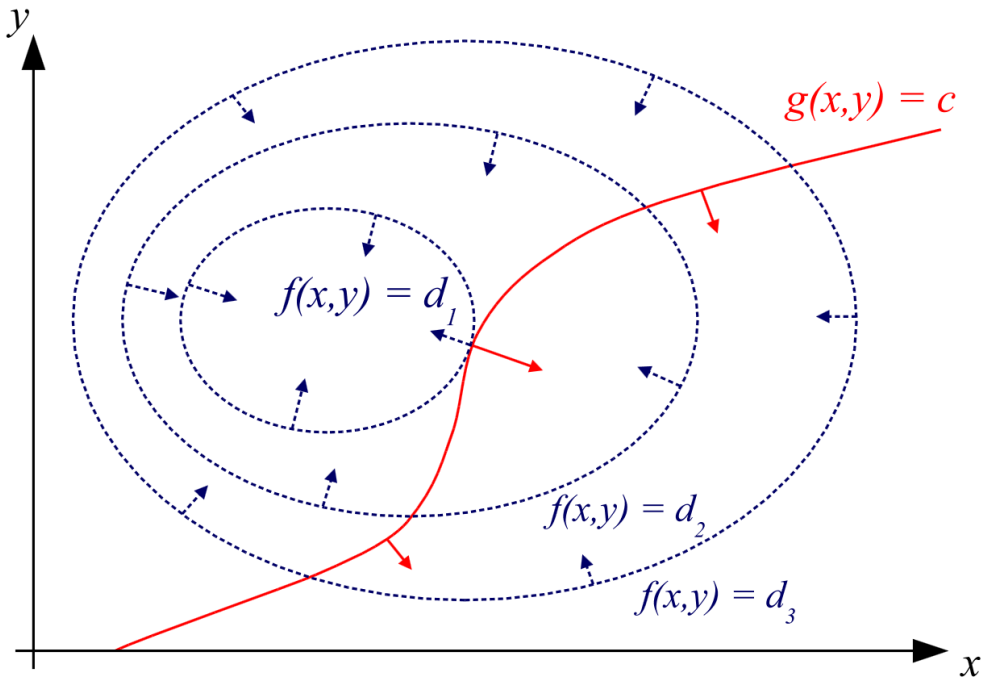
Ideia dos multiplicadores



- Queremos minimizar $f(\mathbf{x})$ sujeito à $g_i(\mathbf{x}) = 0$, $i = 1 : q$;
- E o mínimo restrito é encontrado a partir da equação

$$\mathbf{0} = \nabla f(\mathbf{x}) + \sum_{i=1}^{q} \lambda_i \nabla g_i(\mathbf{x})$$

Ideia dos multiplicadores



- Queremos minimizar $f(\mathbf{x})$ sujeito à $g_i(\mathbf{x}) = 0$, $i = 1 : q$;
- E o mínimo restrito é encontrado a partir da equação

$$\mathbf{0} = \nabla f(\mathbf{x}) + \sum_{i=1}^{q} \lambda_i \nabla g_i(\mathbf{x})$$

Exemplo

- Considere o problema de minimização:

$$\min_x \frac{1}{2}kx^2, \quad \text{sujeito a} \quad x \geq b.$$

- Este problema apresenta única variável;
- Reescrevendo a restrição na forma $b - x \leq 0$, a função J_A fica:

$$J_A(x, \lambda) = \frac{1}{2}kx^2 - \lambda x + \lambda b.$$

- Vamos estudar dois casos: $b = 1$ e $b = -1$.

Exemplo

- Considere o problema de minimização:

$$\min_x \frac{1}{2}kx^2, \quad \text{sujeito a} \quad x \geq b.$$

- Este problema apresenta única variável;
- Reescrevendo a restrição na forma $b - x \leq 0$, a função J_A fica:

$$J_A(x, \lambda) = \frac{1}{2}kx^2 - \lambda x + \lambda b.$$

- Vamos estudar dois casos: $b = 1$ e $b = -1$.

Exemplo

- Considere o problema de minimização:

$$\min_x \frac{1}{2}kx^2, \quad \text{sujeito a} \quad x \geq b.$$

- Este problema apresenta única variável;
- Reescrevendo a restrição na forma $b - x \leq 0$, a função J_A fica:

$$J_A(x, \lambda) = \frac{1}{2}kx^2 - \lambda x + \lambda b.$$

- Vamos estudar dois casos: $b = 1$ e $b = -1$.

Exemplo

- Considere o problema de minimização:

$$\min_x \frac{1}{2}kx^2, \quad \text{sujeito a} \quad x \geq b.$$

- Este problema apresenta única variável;
- Reescrevendo a restrição na forma $b - x \leq 0$, a função J_A fica:

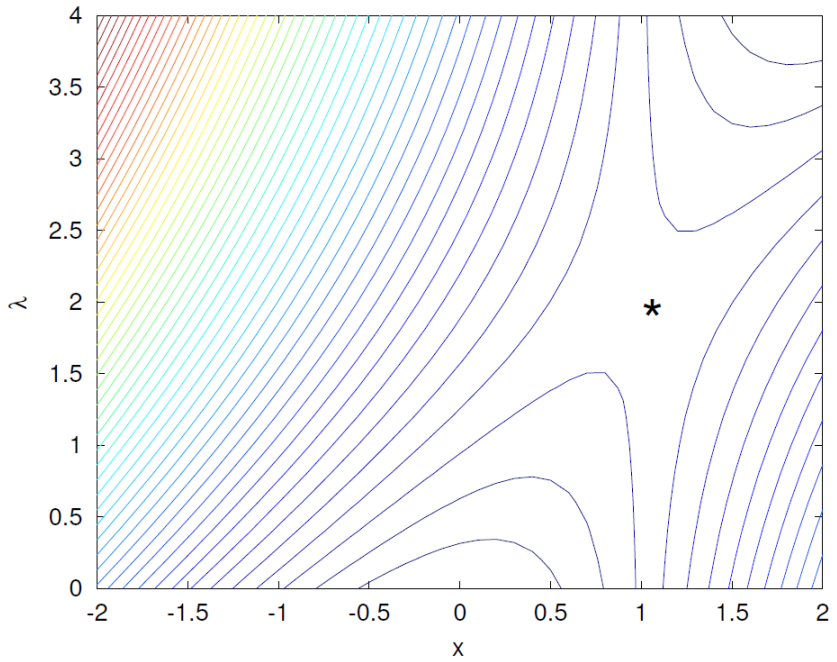
$$J_A(x, \lambda) = \frac{1}{2}kx^2 - \lambda x + \lambda b.$$

- Vamos estudar dois casos: $b = 1$ e $b = -1$.

Exemplo

Caso 1: $b = 1$

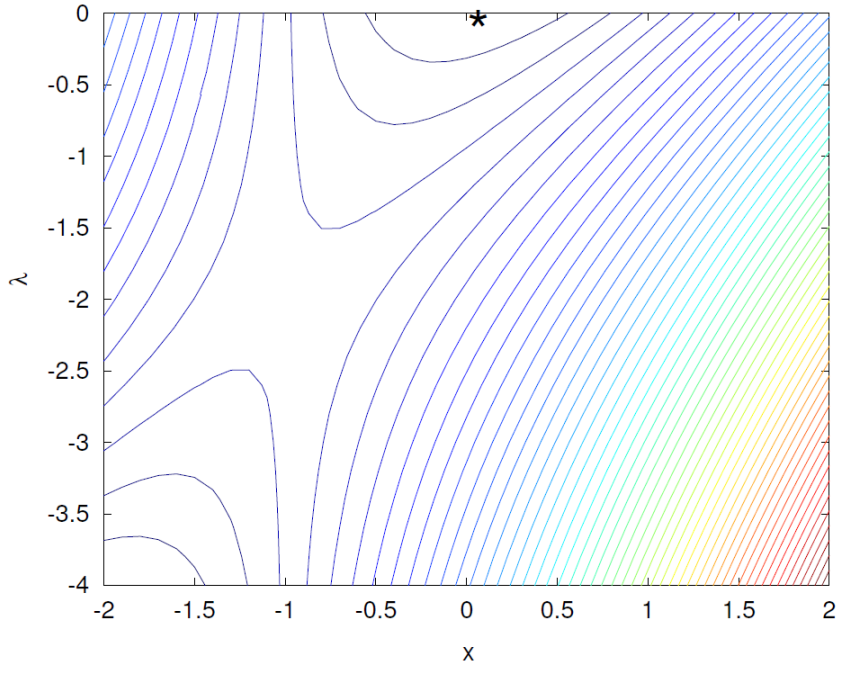
- O mínimo é restrito pela inequação $x \geq 1$;
- E o valor de λ minimiza $J_A(x, \lambda)$ em $x^* = 1$.



Exemplo

Caso 2: $b = -1$

- O mínimo **não** é restrito pela inequação $x \geq -1$;
- Tomando $\lambda = 0$, $J_A(x, \lambda)$ é minimizada em $x = 0$.



O problema "small n, large p"

- Suponha que seu objetivo é verificar se uma substância está ou não relacionada com a incidência de uma doença;
- Entretanto, essa substância é difícil de se conseguir, e se tem o bastante para testar em apenas 250 animais;
- Por outro lado, em cada experimento se gera cerca de 5000 variáveis de interesse, com base na expressão genética;
- Você tem em mãos um problema em que a quantidade de variáveis é muito maior que o número de observações . . . e agora?

O problema "small n, large p"

- Suponha que seu objetivo é verificar se uma substância está ou não relacionada com a incidência de uma doença;
- Entretanto, essa substância é difícil de se conseguir, e se tem o bastante para testar em apenas 250 animais;
- Por outro lado, em cada experimento se gera cerca de 5000 variáveis de interesse, com base na expressão genética;
- Você tem em mãos um problema em que a quantidade de variáveis é muito maior que o número de observações . . . e agora?

O problema "small n, large p"

- Suponha que seu objetivo é verificar se uma substância está ou não relacionada com a incidência de uma doença;
- Entretanto, essa substância é difícil de se conseguir, e se tem o bastante para testar em apenas 250 animais;
- Por outro lado, em cada experimento se gera cerca de 5000 variáveis de interesse, com base na expressão genética;
- Você tem em mãos um problema em que a quantidade de variáveis é muito maior que o número de observações . . . e agora?

O problema "small n, large p"

- Suponha que seu objetivo é verificar se uma substância está ou não relacionada com a incidência de uma doença;
- Entretanto, essa substância é difícil de se conseguir, e se tem o bastante para testar em apenas 250 animais;
- Por outro lado, em cada experimento se gera cerca de 5000 variáveis de interesse, com base na expressão genética;
- Você tem em mãos um problema em que a quantidade de variáveis é muito maior que o número de observações . . . e agora?

O problema "small n, large p"

- A dificuldade é que a maioria dos métodos modernos de análise de dados falha por diferentes razões, p. ex.:
  - ⋆ **Modelos Lineares Generalizados** falham, pois a matriz do modelo não tem Posto completo;
  - ⋆ **Random Forests** falha pois a probabilidade de selecionar variáveis importantes diminui muito.
  - ⋆ **Análise de Clusters** e métodos baseados em distâncias no plano cartesiano, em geral, falham devido à "maldição da dimensionalidade".

O problema "small n, large p"

- A dificuldade é que a maioria dos métodos modernos de análise de dados falha por diferentes razões, p. ex.:
  - ⋆ **Modelos Lineares Generalizados** falham, pois a matriz do modelo não tem Posto completo;
  - ⋆ **Random Forests** falha pois a probabilidade de selecionar variáveis importantes diminui muito.
  - ⋆ **Análise de Clusters** e métodos baseados em distâncias no plano cartesiano, em geral, falham devido à "maldição da dimensionalidade".

O problema "small n, large p"

- A dificuldade é que a maioria dos métodos modernos de análise de dados falha por diferentes razões, p. ex.:

  - ⋆ **Modelos Lineares Generalizados** falham, pois a matriz do modelo não tem Posto completo;

  - ⋆ **Random Forests** falha pois a probabilidade de selecionar variáveis importantes diminui muito.

  - ⋆ **Análise de Clusters** e métodos baseados em distâncias no plano cartesiano, em geral, falham devido à "maldição da dimensionalidade".

O problema "small n, large p"

- A dificuldade é que a maioria dos métodos modernos de análise de dados falha por diferentes razões, p. ex.:

  - ⋆ **Modelos Lineares Generalizados** falham, pois a matriz do modelo não tem Posto completo;

  - ⋆ **Random Forests** falha pois a probabilidade de selecionar variáveis importantes diminui muito.

  - ⋆ **Análise de Clusters** e métodos baseados em distâncias no plano cartesiano, em geral, falham devido à "maldição da dimensionalidade".

O problema "small n, large p"

- Uma suposição razoável é que nem todas variáveis serão bons preditores para a resposta:
  - ⋆ Alguns são muito bons preditores;
  - ⋆ Outros maus preditores;
  - ⋆ Mais tantos que não servem para nada (no sentido de explicar a resposta!);
  - ⋆ Entretanto, ainda dentro dos bons preditores, algumas são correlacionadas e não são "alavancados" por conta disso.
- Podemos resolver essa questão analisando os vetores um a um (geralmente intratável);
- O ideal é que o próprio algoritmo do modelo realize esta seleção.

O problema "small n, large p"

- Uma suposição razoável é que nem todas variáveis serão bons preditores para a resposta:
  - ⋆ Alguns são muito bons preditores;
  - ⋆ Outros maus preditores;
  - ⋆ Mais tantos que não servem para nada (no sentido de explicar a resposta!);
  - ⋆ Entretanto, ainda dentro dos bons preditores, algumas são correlacionadas e não são "alavancados" por conta disso.
- Podemos resolver essa questão analisando os vetores um a um (geralmente intratável);
- O ideal é que o próprio algoritmo do modelo realize esta seleção.

O problema "small n, large p"

- Uma suposição razoável é que nem todas variáveis serão bons preditores para a resposta:
  - ⋆ Alguns são muito bons preditores;
  - ⋆ Outros maus preditores;
  - ⋆ Mais tantos que não servem para nada (no sentido de explicar a resposta!);
  - ⋆ Entretanto, ainda dentro dos bons preditores, algumas são correlacionadas e não são "alavancados" por conta disso.
- Podemos resolver essa questão analisando os vetores um a um (geralmente intratável);
- O ideal é que o próprio algoritmo do modelo realize esta seleção.

O problema "small n, large p"

- Uma suposição razoável é que nem todas variáveis serão bons preditores para a resposta:
  - ⋆ Alguns são muito bons preditores;
  - ⋆ Outros maus preditores;
  - ⋆ Mais tantos que não servem para nada (no sentido de explicar a resposta!);
  - ⋆ Entretanto, ainda dentro dos bons preditores, algumas são correlacionadas e não são "alavancados" por conta disso.
- Podemos resolver essa questão analisando os vetores um a um (geralmente intratável);
- O ideal é que o próprio algoritmo do modelo realize esta seleção.

O problema "small n, large p"

- Uma suposição razoável é que nem todas variáveis serão bons preditores para a resposta:
  - ⋆ Alguns são muito bons preditores;
  - ⋆ Outros maus preditores;
  - ⋆ Mais tantos que não servem para nada (no sentido de explicar a resposta!);
  - ⋆ Entretanto, ainda dentro dos bons preditores, algumas são correlacionadas e não são "alavancados" por conta disso.
- Podemos resolver essa questão analisando os vetores um a um (geralmente intratável);
- O ideal é que o próprio algoritmo do modelo realize esta seleção.

O problema "small n, large p"

- Uma suposição razoável é que nem todas variáveis serão bons preditores para a resposta:
  - ⋆ Alguns são muito bons preditores;
  - ⋆ Outros maus preditores;
  - ⋆ Mais tantos que não servem para nada (no sentido de explicar a resposta!);
  - ⋆ Entretanto, ainda dentro dos bons preditores, algumas são correlacionadas e não são "alavancados" por conta disso.
- Podemos resolver essa questão analisando os vetores um a um (geralmente intratável);
- O ideal é que o próprio algoritmo do modelo realize esta seleção.

O problema "small n, large p"

- Uma suposição razoável é que nem todas variáveis serão bons preditores para a resposta:
  - ⋆ Alguns são muito bons preditores;
  - ⋆ Outros maus preditores;
  - ⋆ Mais tantos que não servem para nada (no sentido de explicar a resposta!);
  - ⋆ Entretanto, ainda dentro dos bons preditores, algumas são correlacionadas e não são "alavancados" por conta disso.
- Podemos resolver essa questão analisando os vetores um a um (geralmente intratável);
- O ideal é que o próprio algoritmo do modelo realize esta seleção.

Uma solução é a Regularização

- Uma forma de restringir o número de variáveis é impor um custo, ou *penalty*, ao algoritmo

$$\min_{\beta} E\left[J(h(\mathbf{x}, \beta) - \mathbf{y})\right] \text{ sujeito a } P(\beta) < t,$$

em que t é um número real entre zero e infinito.

- E qual relação tem a definição acima com Multiplicadores de Lagrange?

$$\min_{\beta} E\left[J(h(\mathbf{x}, \beta) - \mathbf{y})\right] + \lambda P(\beta),$$

em que λ é um número real entre zero e infinito.

- t **e** λ **são inversamente proporcionais**.

Uma solução é a Regularização

- Uma forma de restringir o número de variáveis é impor um custo, ou *penalty*, ao algoritmo

$$\min_{\beta} E\left[J(h(\mathbf{x}, \beta) - \mathbf{y})\right] \text{ sujeito a } P(\beta) < t,$$

em que t é um número real entre zero e infinito.

- E qual relação tem a definição acima com Multiplicadores de Lagrange?

$$\min_{\beta} E\left[J(h(\mathbf{x}, \beta) - \mathbf{y})\right] + \lambda P(\beta),$$

em que λ é um número real entre zero e infinito.

- t **e** λ **são inversamente proporcionais**.

Uma solução é a Regularização

- Uma forma de restringir o número de variáveis é impor um custo, ou *penalty*, ao algoritmo

$$\min_{\beta} E\left[J(h(\mathbf{x}, \beta) - \mathbf{y})\right] \text{ sujeito a } P(\beta) < t,$$

em que t é um número real entre zero e infinito.

- E qual relação tem a definição acima com Multiplicadores de Lagrange?

$$\min_{\beta} E\left[J(h(\mathbf{x}, \beta) - \mathbf{y})\right] + \lambda P(\beta),$$

em que λ é um número real entre zero e infinito.

- t e λ são inversamente proporcionais.

Uma solução é a Regularização

- Uma forma de restringir o número de variáveis é impor um custo, ou *penalty*, ao algoritmo

$$\min_{\beta} E\left[J(h(\mathbf{x}, \beta) - \mathbf{y})\right] \text{ sujeito a } P(\beta) < t,$$

em que t é um número real entre zero e infinito.

- E qual relação tem a definição acima com Multiplicadores de Lagrange?

$$\min_{\beta} E\left[J(h(\mathbf{x}, \beta) - \mathbf{y})\right] + \lambda P(\beta),$$

em que λ é um número real entre zero e infinito.

- t **e** λ **são inversamente proporcionais**.

Como escolher as funções J e P?

- Substituindo o valor esperado pelo estimador não-viesado, o problema de otimização fica então:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \beta) - y_i)] + \lambda P(\beta)$$

- P. ex. se J fosse a perda quadrática, o primeiro termo seria a SSE;
- As funções **penalty** (*shrinkage penalty*) mantém as estimativas β_j próximas de zero.
- λ é o **tuning parameter**, e é determinado separadamente. Ele controla o impacto de J e P nas estimativas dos parâmetros.

Como escolher as funções J e P?

- Substituindo o valor esperado pelo estimador não-viesado, o problema de otimização fica então:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \beta) - y_i)] + \lambda P(\beta)$$

- P. ex. se J fosse a perda quadrática, o primeiro termo seria a SSE;

- As funções **penalty** (*shrinkage penalty*) mantém as estimativas β_j próximas de zero.

- λ é o **tuning parameter**, e é determinado separadamente. Ele controla o impacto de J e P nas estimativas dos parâmetros.

Como escolher as funções J e P?

- Substituindo o valor esperado pelo estimador não-viesado, o problema de otimização fica então:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \boldsymbol{\beta}) - y_i)] + \lambda P(\boldsymbol{\beta})$$

- P. ex. se J fosse a perda quadrática, o primeiro termo seria a SSE;

- As funções **penalty** (*shrinkage penalty*) mantém as estimativas β_j próximas de zero.

- λ é o **tuning parameter**, e é determinado separadamente. Ele controla o impacto de J e P nas estimativas dos parâmetros.

Como escolher as funções J e P?

- Substituindo o valor esperado pelo estimador não-viesado, o problema de otimização fica então:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \beta) - y_i)] + \lambda P(\beta)$$

- P. ex. se J fosse a perda quadrática, o primeiro termo seria a SSE;

- As funções **penalty** (*shrinkage penalty*) mantém as estimativas β_j próximas de zero.

- λ é o **tuning parameter**, e é determinado separadamente. Ele controla o impacto de J e P nas estimativas dos parâmetros.

Família das potências

- A ideia é quanto maior for o valor absoluto de um coeficiente, maior é a penalidade atribuída a ele;
- E menor será a chance daquela estimativa (ou da variável) entrar no modelo final;
- Em geral, o problema de otimização com uma penalidade da família das potências pode ser escrito como:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \beta) - y_i)] + \lambda \sum_{j=1}^{k} |\beta_j|^q$$

- Em Álgebra Linear chamamos $\left(\sum_{j=1}^{k} |\beta_j|^q \right)^{1/q}$ de norma ℓ^q.

Família das potências

- A ideia é quanto maior for o valor absoluto de um coeficiente, maior é a penalidade atribuída a ele;
- E menor será a chance daquela estimativa (ou da variável) entrar no modelo final;
- Em geral, o problema de otimização com uma penalidade da família das potências pode ser escrito como:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \beta) - y_i)] + \lambda \sum_{j=1}^{k} |\beta_j|^q$$

- Em Álgebra Linear chamamos $\left(\sum_{j=1}^{k} |\beta_j|^q \right)^{1/q}$ de norma ℓ^q.

Família das potências

- A ideia é quanto maior for o valor absoluto de um coeficiente, maior é a penalidade atribuída a ele;
- E menor será a chance daquela estimativa (ou da variável) entrar no modelo final;
- Em geral, o problema de otimização com uma penalidade da família das potências pode ser escrito como:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \boldsymbol{\beta}) - y_i)] + \lambda \sum_{j=1}^{k} |\beta_j|^q$$

- Em Álgebra Linear chamamos $\left(\sum_{j=1}^{k} |\beta_j|^q \right)^{1/q}$ de norma ℓ^q.

Família das potências

- A ideia é quanto maior for o valor absoluto de um coeficiente, maior é a penalidade atribuída a ele;
- E menor será a chance daquela estimativa (ou da variável) entrar no modelo final;
- Em geral, o problema de otimização com uma penalidade da família das potências pode ser escrito como:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \boldsymbol{\beta}) - y_i)] + \lambda \sum_{j=1}^{k} |\beta_j|^q$$

- Em Álgebra Linear chamamos $\left(\sum_{j=1}^{k} |\beta_j|^q \right)^{1/q}$ de norma ℓ^q.

$q = 2$ - **Penalização Ridge**

- O problema de otimização é dado por:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \boldsymbol{\beta}) - y_i)] + \lambda \sum_{j=1}^{k} \beta_j^2$$

- A ideia surgiu para solucionar a singularidade da matriz quando $p > n$. Para isso soma-se uma constante λ à sua diagonal;

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- Considerando o caso dos vetores de $\mathbf{X}$ ortonormais temos

$$\hat{\beta}_\lambda^R = \frac{\hat{\beta}^{OLS}}{1 + \lambda}$$

- Este fato ilustra a característica essencial da regressão *Ridge*: **shrinkage**.

$q = 2$ - **Penalização Ridge**

- O problema de otimização é dado por:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \boldsymbol{\beta}) - y_i)] + \lambda \sum_{j=1}^{k} \beta_j^2$$

- A ideia surgiu para solucionar a singularidade da matriz quando $p > n$. Para isso soma-se uma constante λ à sua diagonal;

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- Considerando o caso dos vetores de $\mathbf{X}$ ortonormais temos

$$\hat{\beta}_\lambda^R = \frac{\hat{\beta}^{OLS}}{1+\lambda}$$

- Este fato ilustra a característica essencial da regressão *Ridge*: **shrinkage**.

$q = 2$ - **Penalização Ridge**

- O problema de otimização é dado por:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \boldsymbol{\beta}) - y_i)] + \lambda \sum_{j=1}^{k} \beta_j^2$$

- A ideia surgiu para solucionar a singularidade da matriz quando $p > n$. Para isso soma-se uma constante λ à sua diagonal;

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- Considerando o caso dos vetores de $\mathbf{X}$ ortonormais temos

$$\hat{\beta}_\lambda^R = \frac{\hat{\beta}^{OLS}}{1 + \lambda}$$

- Este fato ilustra a característica essencial da regressão *Ridge*: **shrinkage**.

$q = 2$ - **Penalização Ridge**

- O problema de otimização é dado por:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \boldsymbol{\beta}) - y_i)] + \lambda \sum_{j=1}^{k} \beta_j^2$$

- A ideia surgiu para solucionar a singularidade da matriz quando $p > n$. Para isso soma-se uma constante λ à sua diagonal;

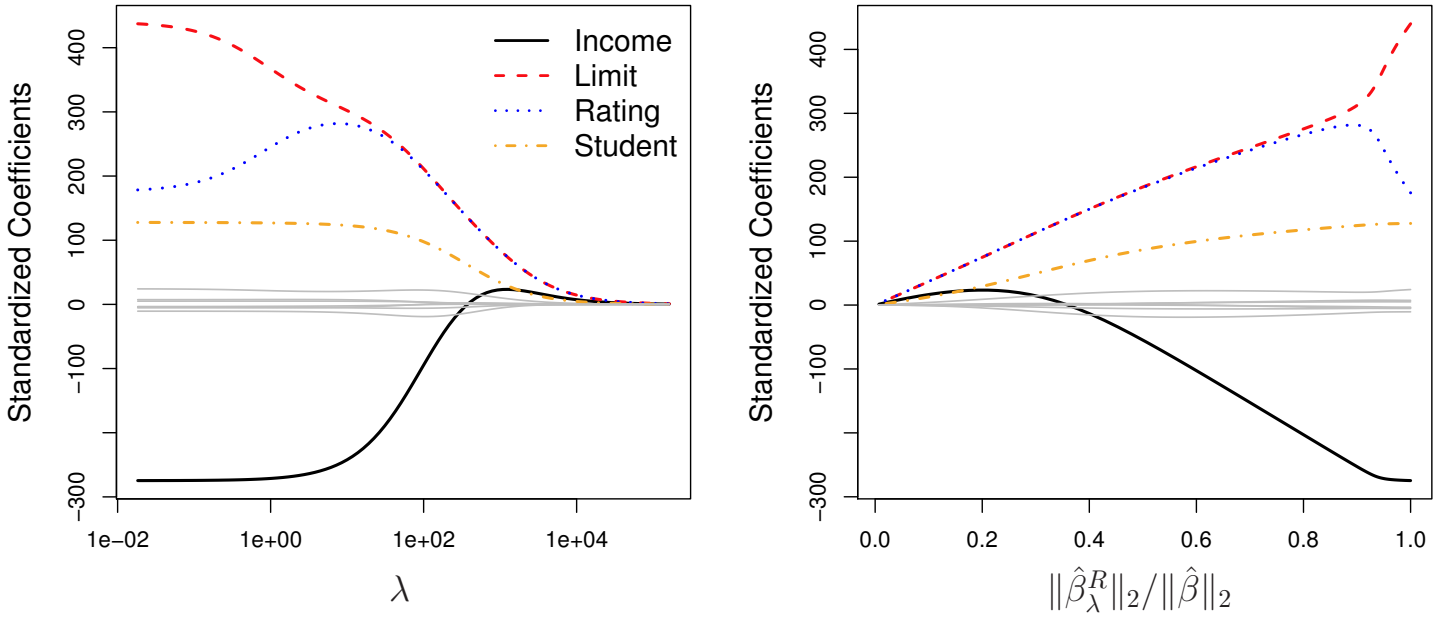
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- Considerando o caso dos vetores de $\mathbf{X}$ ortonormais temos

$$\hat{\beta}_\lambda^R = \frac{\hat{\beta}^{OLS}}{1 + \lambda}$$

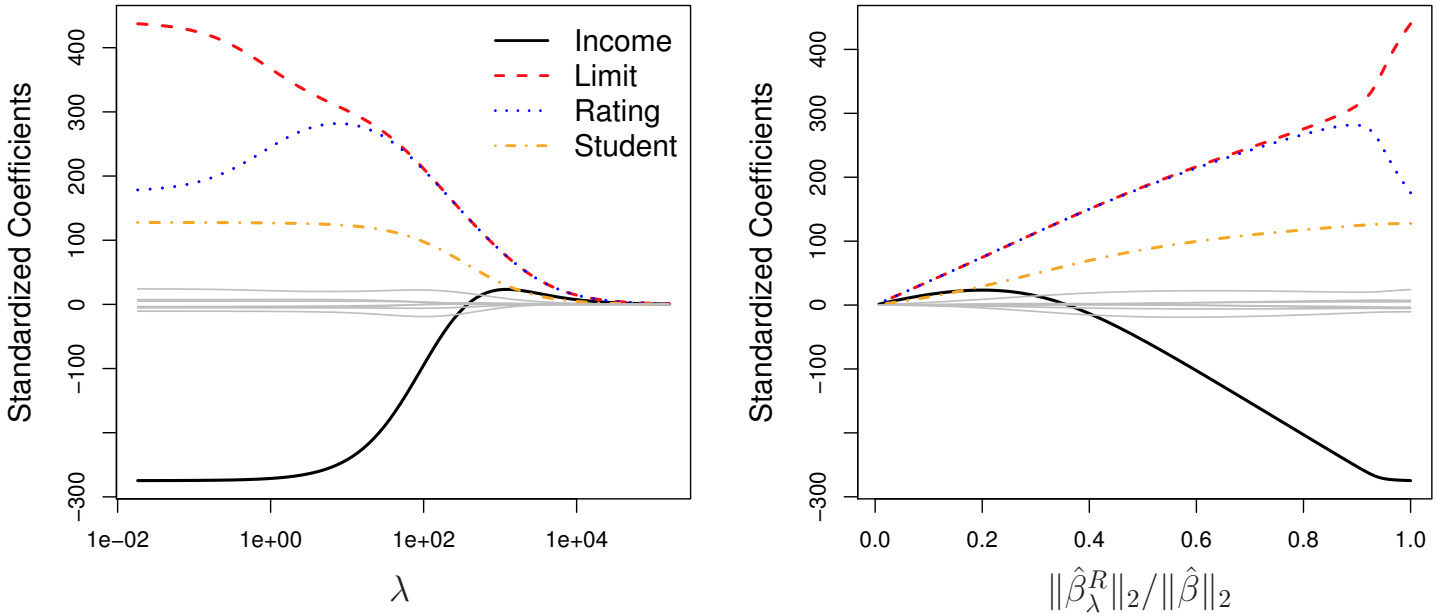
- Este fato ilustra a característica essencial da regressão *Ridge*: **shrinkage**.

Exemplo: `Credit` data set



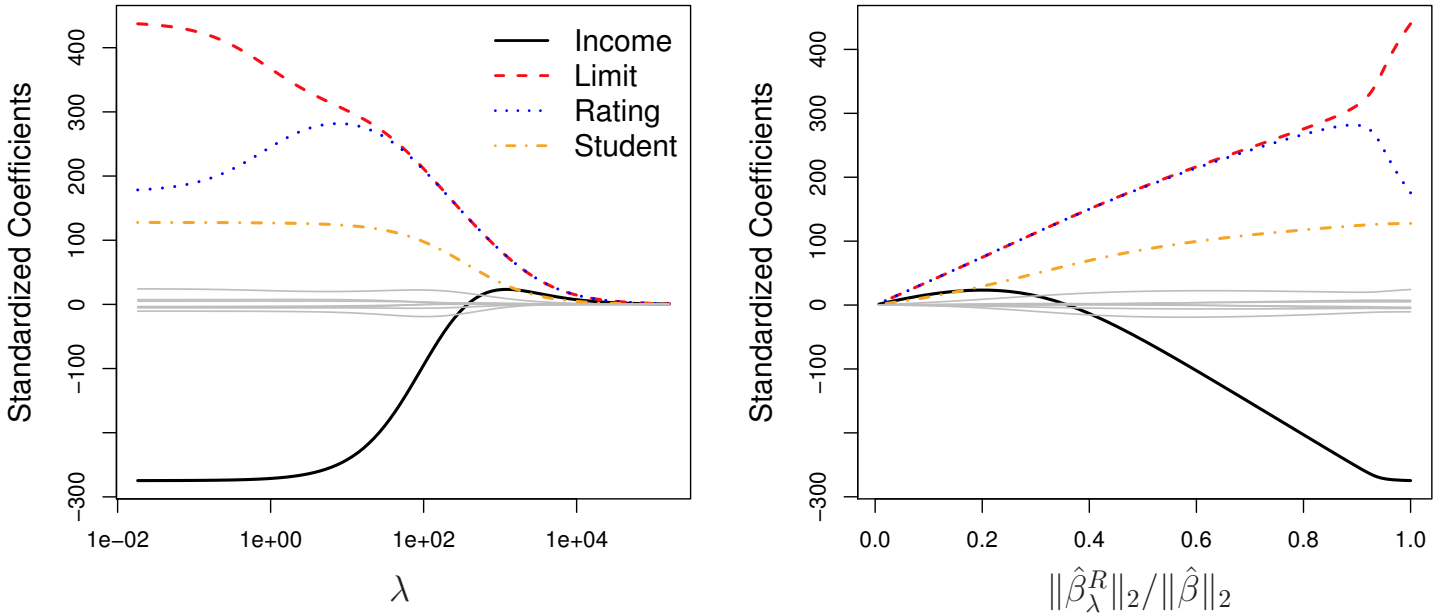
- O gráfico da esquerda, cada curva corresponde à estimativa de cada coeficiente através da regressão *Ridge* plotada como função de λ;
- O lado direito refere-se às mesmas estimativas dos coeficientes da regressão, mas como função de $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$;
- $\hat{\beta}$ denota o vetor de estimativas por mínimos quadrados.

Exemplo: `Credit` data set



- O gráfico da esquerda, cada curva corresponde à estimativa de cada coeficiente através da regressão *Ridge* plotada como função de λ;
- O lado direito refere-se às mesmas estimativas dos coeficientes da regressão, mas como função de $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$;
- $\hat{\beta}$ denota o vetor de estimativas por mínimos quadrados.

Exemplo: `Credit` data set



- O gráfico da esquerda, cada curva corresponde à estimativa de cada coeficiente através da regressão *Ridge* plotada como função de λ;
- O lado direito refere-se às mesmas estimativas dos coeficientes da regressão, mas como função de $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$;
- $\hat{\beta}$ denota o vetor de estimativas por mínimos quadrados.

$q = 1$ - **Penalização Lasso**

- A regressão *Ridge* falha na parcimônia do modelo. Ela inclui todos os p preditores (ainda que com pouco peso);
- **Lasso** é uma alternativa que contorna esta desvantagem. Os coeficientes *Lasso*, $\hat{\beta}^L_\lambda$, minimizam a quantidade

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \beta) - y_i)] + \lambda \sum_{j=1}^{k} |\beta_j|$$

- O *penalty* ℓ^1 funciona também como um **selecionador de variável**.

$q = 1$ - **Penalização Lasso**

- A regressão *Ridge* falha na parcimônia do modelo. Ela inclui todos os p preditores (ainda que com pouco peso);
- **Lasso** é uma alternativa que contorna esta desvantagem. Os coeficientes *Lasso*, $\hat{\beta}_\lambda^L$, minimizam a quantidade

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} [J(h(x_i, \boldsymbol{\beta}) - y_i)] + \lambda \sum_{j=1}^{k} |\beta_j|$$

- O *penalty* ℓ^1 funciona também como um **selecionador de variável**.

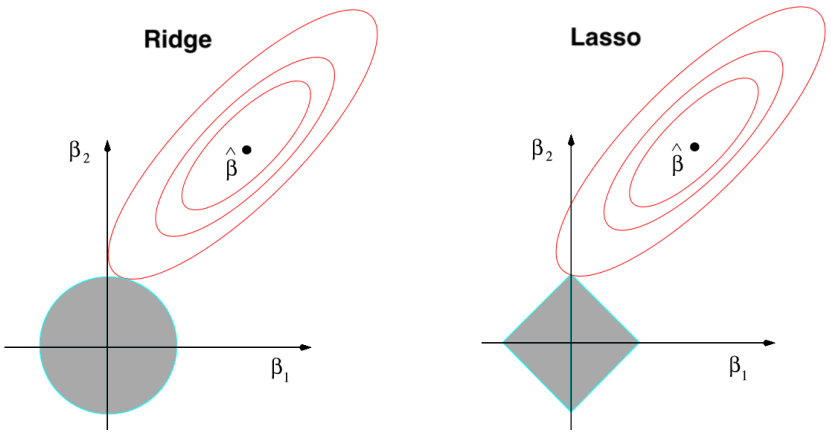
$q = 1$ - **Penalização Lasso**

- A regressão *Ridge* falha na parcimônia do modelo. Ela inclui todos os p preditores (ainda que com pouco peso);

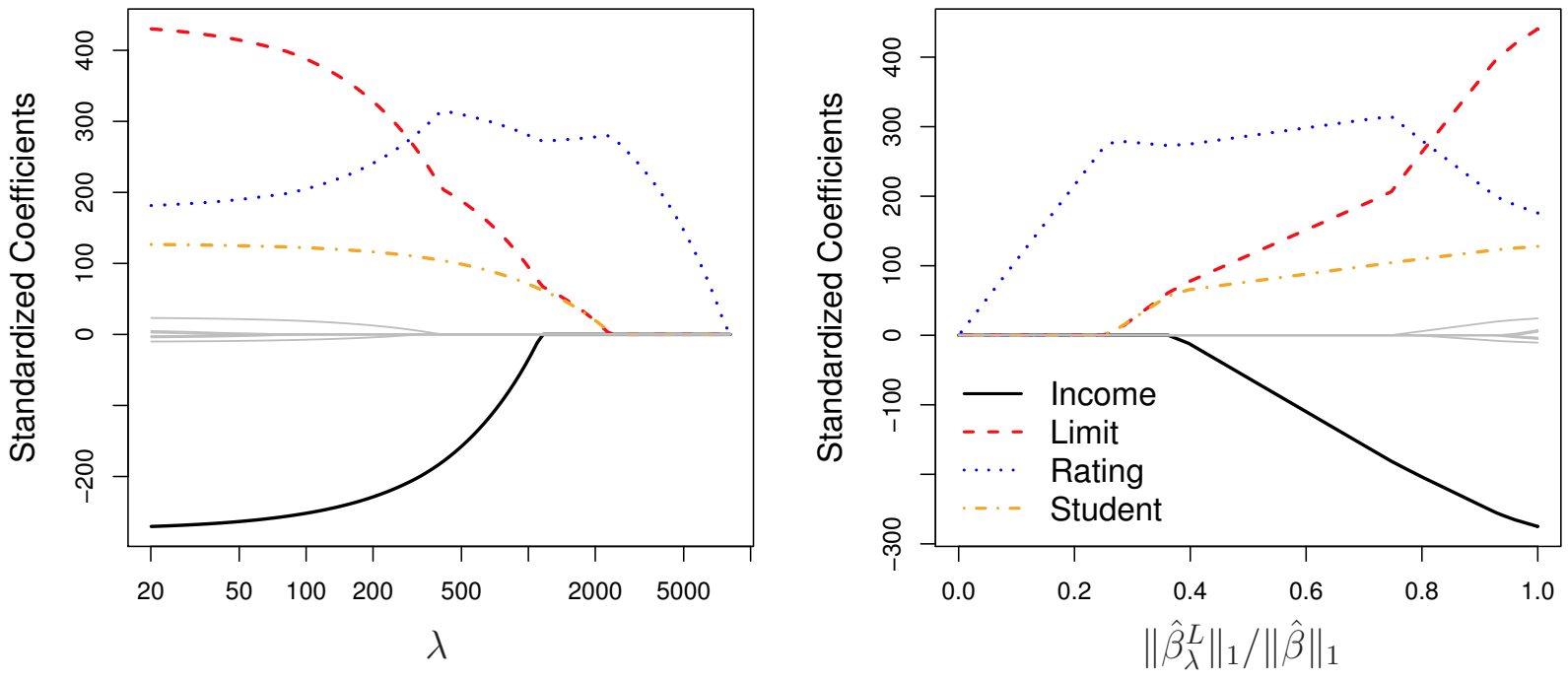
- **Lasso** é uma alternativa que contorna esta desvantagem. Os coeficientes *Lasso*, $\hat{\beta}^L_\lambda$, minimizam a quantidade

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}\left[J(h(x_i,\boldsymbol{\beta})-y_i)\right]+\lambda\sum_{j=1}^{k}|\beta_j|$$

- O *penalty* ℓ^1 funciona também como um **selecionador de variável**.



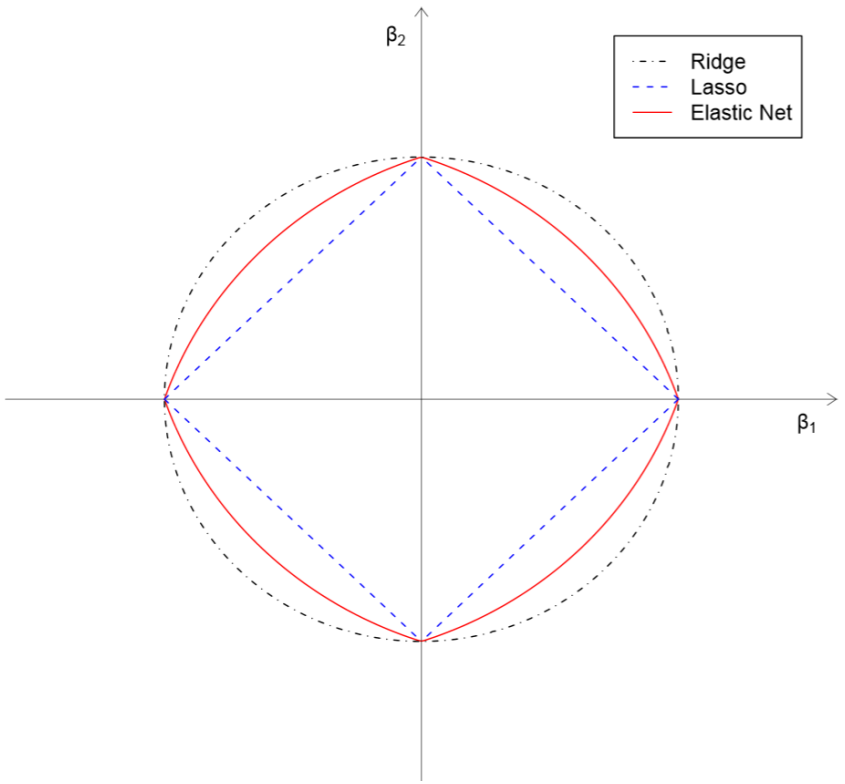
Exemplo: `Credit` data set



Elastic net

- **Elastic net** é um compromisso entre a regressão *Ridge* e *Lasso*. Os coeficientes elastic net, $\hat{\beta}^E_\lambda$, minimizam a quantidade

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}\left[J(h(x_i,\boldsymbol{\beta})-y_i)\right]+\lambda\sum_{j=1}^{k}\left(\alpha|\beta_j|+(1-\alpha)\beta^2\right)$$



$q < 1$ - Penalização horseshoe

- E o que acontece se reduzirmos q ainda mais? Esse estudo deu origem aos estimadores baseados em penalização *horseshoe*;
- Ela favorece ainda mais a presença de 0's (maior esparcidade);
- Ou seja, tende a encontrar as elipses geradas pelos mínimos quadrados em cima dos eixos com mais frequência que *Ridge* e *Lasso*;

$q < 1$ - **Penalização horseshoe**

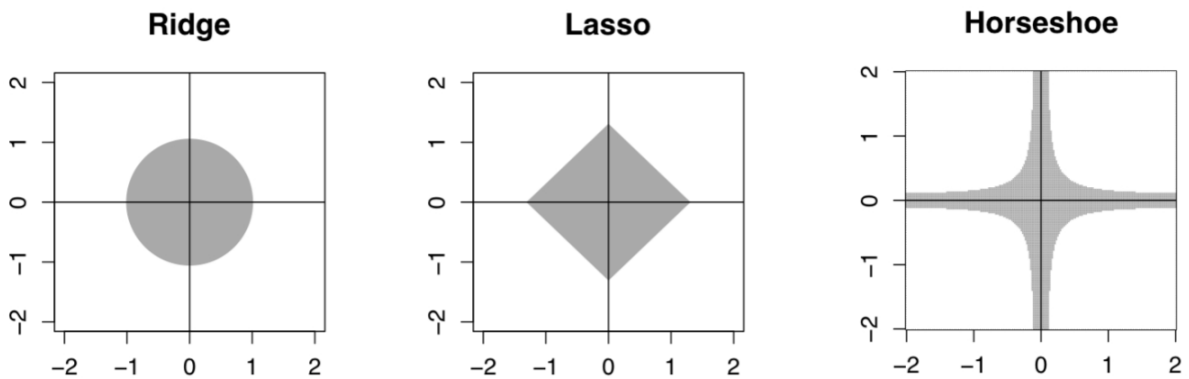
- E o que acontece se reduzirmos q ainda mais? Esse estudo deu origem aos estimadores baseados em penalização *horseshoe*;
- Ela favorece ainda mais a presença de 0's (maior esparcidade);
- Ou seja, tende a encontrar as elipses geradas pelos mínimos quadrados em cima dos eixos com mais frequência que *Ridge* e *Lasso*;

$q < 1$ - **Penalização horseshoe**

- E o que acontece se reduzirmos q ainda mais? Esse estudo deu origem aos estimadores baseados em penalização *horseshoe*;
- Ela favorece ainda mais a presença de 0's (maior esparcidade);
- Ou seja, tende a encontrar as elipses geradas pelos mínimos quadrados em cima dos eixos com mais frequência que *Ridge* e *Lasso*;

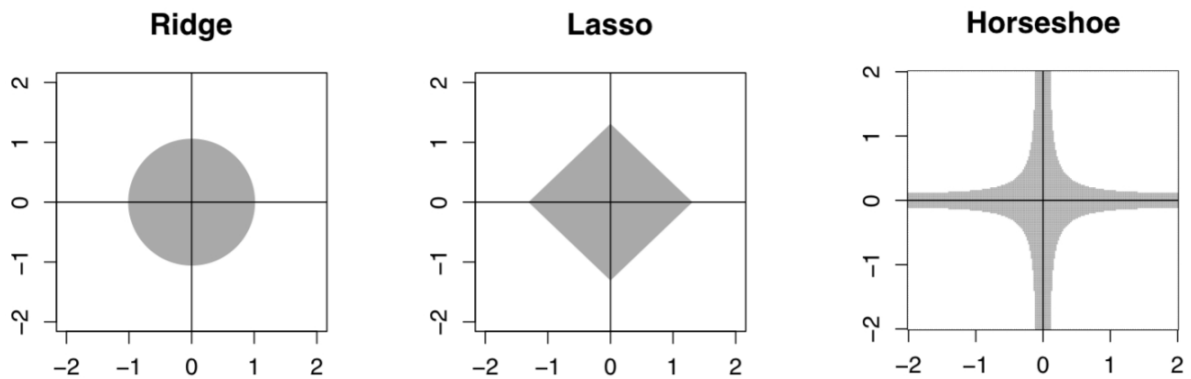
$q < 1$ - Penalização horseshoe

- E o que acontece se reduzirmos q ainda mais? Esse estudo deu origem aos estimadores baseados em penalização *horseshoe*;
- Ela favorece ainda mais a presença de 0's (maior esparcidade);
- Ou seja, tende a encontrar as elipses geradas pelos mínimos quadrados em cima dos eixos com mais frequência que *Ridge* e *Lasso*;



$q < 1$ - Penalização horseshoe

- E o que acontece se reduzirmos q ainda mais? Esse estudo deu origem aos estimadores baseados em penalização *horseshoe*;
- Ela favorece ainda mais a presença de 0's (maior esparcidade);
- Ou seja, tende a encontrar as elipses geradas pelos mínimos quadrados em cima dos eixos com mais frequência que *Ridge* e *Lasso*;



- E quando $q = 0$ voltamos ao **Best subset selection**.

Selecionando o tuning parameter, λ

- Assim como no *subset selection*, para regressão *Ridge* e *Lasso* necessitamos de um método para determinar qual o melhor modelo;
- Neste caso, precisamos encontrar o valor de λ que fornece esta informação;
- **Validação cruzada** (que veremos na próxima aula) fornece uma maneira simples de resolver este problema:
  (a) A partir de uma grade valores de λ, calculamos a taxa de erro de validação (para cada λ);
  (b) Escolhemos o valor de λ que fornece a menor taxa de erro;
  (c) Finalmente, ajustamos novamente o modelo, utilizando todas as observações disponíveis (não somente a de validação), com o valor de λ encontrado anteriormente.

Selecionando o tuning parameter, λ

- Assim como no *subset selection*, para regressão *Ridge* e *Lasso* necessitamos de um método para determinar qual o melhor modelo;
- Neste caso, precisamos encontrar o valor de λ que fornece esta informação;
- **Validação cruzada** (que veremos na próxima aula) fornece uma maneira simples de resolver este problema:
  (a) A partir de uma grade valores de λ, calculamos a taxa de erro de validação (para cada λ);
  (b) Escolhemos o valor de λ que fornece a menor taxa de erro;
  (c) Finalmente, ajustamos novamente o modelo, utilizando todas as observações disponíveis (não somente a de validação), com o valor de λ encontrado anteriormente.

Selecionando o tuning parameter, λ

- Assim como no *subset selection*, para regressão *Ridge* e *Lasso* necessitamos de um método para determinar qual o melhor modelo;
- Neste caso, precisamos encontrar o valor de λ que fornece esta informação;
- **Validação cruzada** (que veremos na próxima aula) fornece uma maneira simples de resolver este problema:
  (a) A partir de uma grade valores de λ, calculamos a taxa de erro de validação (para cada λ);
  (b) Escolhemos o valor de λ que fornece a menor taxa de erro;
  (c) Finalmente, ajustamos novamente o modelo, utilizando todas as observações disponíveis (não somente a de validação), com o valor de λ encontrado anteriormente.

Selecionando o tuning parameter, λ

- Assim como no *subset selection*, para regressão *Ridge* e *Lasso* necessitamos de um método para determinar qual o melhor modelo;
- Neste caso, precisamos encontrar o valor de λ que fornece esta informação;
- **Validação cruzada** (que veremos na próxima aula) fornece uma maneira simples de resolver este problema:
  (a) A partir de uma grade valores de λ, calculamos a taxa de erro de validação (para cada λ);
  (b) Escolhemos o valor de λ que fornece a menor taxa de erro;
  (c) Finalmente, ajustamos novamente o modelo, utilizando todas as observações disponíveis (não somente a de validação), com o valor de λ encontrado anteriormente.

Selecionando o tuning parameter, λ

- Assim como no *subset selection*, para regressão *Ridge* e *Lasso* necessitamos de um método para determinar qual o melhor modelo;
- Neste caso, precisamos encontrar o valor de λ que fornece esta informação;
- **Validação cruzada** (que veremos na próxima aula) fornece uma maneira simples de resolver este problema:
  (a) A partir de uma grade valores de λ, calculamos a taxa de erro de validação (para cada λ);
  (b) Escolhemos o valor de λ que fornece a menor taxa de erro;
  (c) Finalmente, ajustamos novamente o modelo, utilizando todas as observações disponíveis (não somente a de validação), com o valor de λ encontrado anteriormente.

Selecionando o tuning parameter, λ

- Assim como no *subset selection*, para regressão *Ridge* e *Lasso* necessitamos de um método para determinar qual o melhor modelo;
- Neste caso, precisamos encontrar o valor de λ que fornece esta informação;
- **Validação cruzada** (que veremos na próxima aula) fornece uma maneira simples de resolver este problema:

  (a) A partir de uma grade valores de λ, calculamos a taxa de erro de validação (para cada λ);

  (b) Escolhemos o valor de λ que fornece a menor taxa de erro;

  (c) Finalmente, ajustamos novamente o modelo, utilizando todas as observações disponíveis (não somente a de validação), com o valor de λ encontrado anteriormente.

Selecionando o tuning parameter, λ