Modelos de Regressão Múltipla - Parte VII

Erica Castilho Rodrigues

26 de Janeiro de 2016

Regressão Polinomial

Multicolinearidade

Regressão Polinomial

- Vimos como ajustar um modelo não linear fazendo transformações das variáveis, como, por exemplo

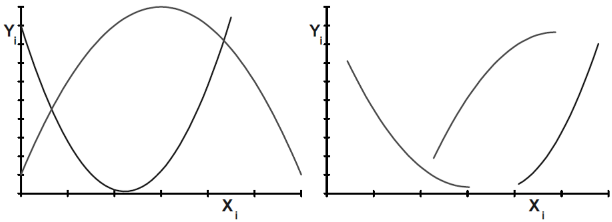
$$Y = \exp\{\beta_0 + \beta_1 x + \epsilon\} \, .$$

- Existem casos em que a relação entre as variáveis é polinomial.
- Por exemplo, parábolas aparecem frequentemente em problemas de engenharia.
- Polinômios podem ajustar uma grande variedade de curvas.

- A relação entre a variável resposta e a explicativa é modelada por um polinômio de ordem p.
- Ajustamos um polinômio de ordem 2, 3 ou 4 e então verificamos se podemos eliminar alguns termos do modelo.
- Com polinômios podemos:
  - determinar se existe uma relação curvilínea entre Y e X;
  - determinar se essa curva é quadrática, cúbica, etc;
  - obter a equação polinomial de Y em função de X (podemos ter várias preditoras).

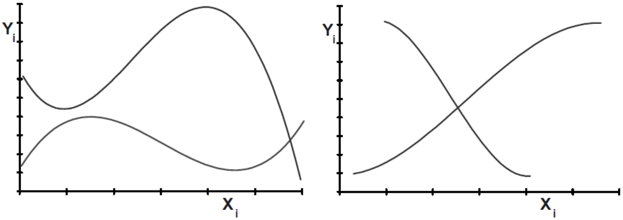
Polinômio mais simples - de segunda ordem e com uma variável explicativa

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$



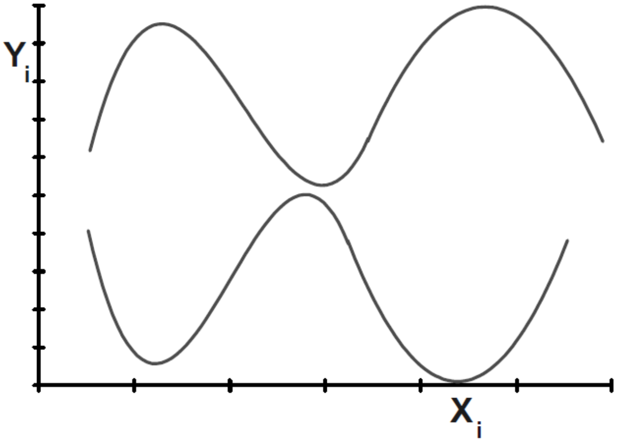
Um polinômio um pouco mais complexo - de terceira ordem e com uma variável explicativa

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i$$



Um polinômio um pouco mais complexo - de quarta ordem e com uma variável explicativa

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \epsilon_i$$



- Começamos primeiro com uma reta, ou polinômio de grau um.
- Vamos incluindo graus maiores e testando a significância.
- Para fazer essa inclusão, basta acrescentar na matriz **X**, uma coluna com os valores de X^2 ou X^3.
- Quando chegamos em um grau que não é mais significativo interrrompemos o processo.
- Todos os termos de grau menor são incluídos no modelo.
- Por exemplo, se X^3 é significativo:
  - vamos incluir X e X^2 no modelo.

- Não devemos colocar muito termos no modelo.
- Se tivermos n observações:
  - um polinômio de grau $n-1$ se ajusta perfeitamente aos dados.
- Esse modelo é bom?

- Não devemos colocar muito termos no modelo.
- Se tivermos n observações:
  - um polinômio de grau $n - 1$ se ajusta perfeitamente aos dados.
- Esse modelo é bom?
- Esse modelo não é bom, pois está ajustando ruído, variação aleatória.
- O grau do polinômio não deve ser maior do que 1/3 do número de observações.
- Se temos uma amostra de tamanho 12:
  - não devemos ajustar um polinômio de grau maior que 4.

- Todas suposições do modelo de regressão simples devem valer para o modelo polinomial:

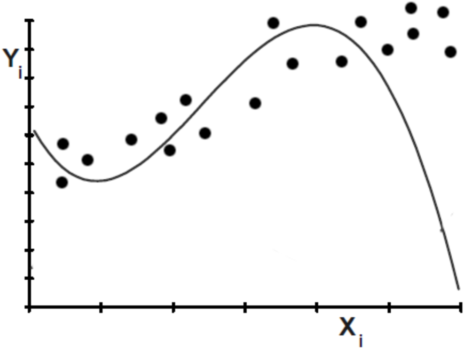
- Todas suposições do modelo de regressão simples devem valer para o modelo polinomial:

- Todas suposições do modelo de regressão simples devem valer para o modelo polinomial:
  - normalidade,

- Todas suposições do modelo de regressão simples devem valer para o modelo polinomial:
    - normalidade, independência e

- Todas suposições do modelo de regressão simples devem valer para o modelo polinomial:
  - normalidade, independência e variância constante dos erros.
- Os polinômios não podem ser usados para fazer predições fora do intervalo dos dados.
- Pontos de inflexão:
  - permitem que a relação mude o sentido, de crescente, vire decrescente.
- A cada grau que aumentamos do polinômio, permitimos que essa mudança ocorra mais uma vez.

- Extrapolação são inúteis a não ser que estejamos ajustando uma parábola:
  - a curva não tem uma tendência bem comportada fora do intervalos dos dados.



- Se aumentarmos o grau do polinômio ele pode representar qualquer curva.
- Porém, devemos usar apenas se não pudermos encontrar nenhuma função mais simples.
- Se uma função exponencial se ajusta bem aos dados, devemos usá-la ao invés do polinômio.
- Os parâmetros da exponencial tem interpretação direta.
- Qual interpretação do termo β_2 em

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \, ?$$

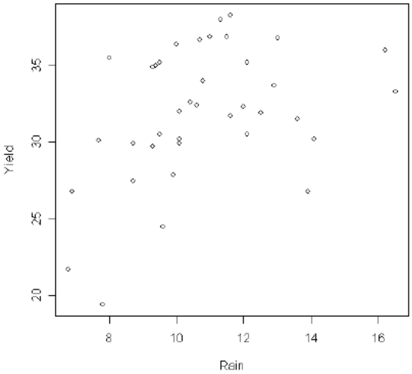
- Se aumentarmos o grau do polinômio ele pode representar qualquer curva.
- Porém, devemos usar apenas se não pudermos encontrar nenhuma função mais simples.
- Se uma função exponencial se ajusta bem aos dados, devemos usá-la ao invés do polinômio.
- Os parâmetros da exponencial tem interpretação direta.
- Qual interpretação do termo β_2 em

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \,?$$

- Não tem interpretação.
- Portanto, só devemos usar polinômios se não encontramos nenhuma função mais simples.

Exemplo:

- Estamos analisando a produção de milho em uma fazenda.
- Queremos verificar como ela varia com a quantidade de chuva.



Exemplo: (continuação)

- A relação entre as variáveis parece linear?

Exemplo: (continuação)

- A relação entre as variáveis parece linear? Não.
- Qual tipo de curva parece se ajustar bem nesse caso?

Exemplo: (continuação)

- A relação entre as variáveis parece linear? Não.
- Qual tipo de curva parece se ajustar bem nesse caso? Uma parábola.
- Vamos então ajustar um polinômio de grau 2.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \quad \epsilon_i \sim^{iid} N(0, \sigma^2)\ .$$

- Comandos em R para ajustar o modelo

```
results = lm(Yield ~ Rain + I(Rain^2))
summary(results)
```

Exemplo: (continuação)

- Os resultados obtidos através do ajuste do modelo são mostrados a seguir.

```
Call:
lm(formula = Yield ~ Rain + I(Rain^2))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.01467 11.44158 -0.438 0.66387
Rain 6.00428 2.03895 2.945 0.00571 **
I(Rain^2) -0.22936 0.08864 -2.588 0.01397 *

Residual standard error: 3.763 on 35 degrees
of freedom

Multiple R-squared: 0.2967,
Adjusted R-squared: 0.2565
F-statistic: 7.382 on 2 and 35 DF,
p-value: 0.002115
```

Exemplo: (continuação)

- Quais as conclusões podemos tirar?

Exemplo: (continuação)

- Quais as conclusões podemos tirar?
- A quantidade de chuva é significativa para explicar a produção de milho.

Exemplo: **(continuação)**

- Quais as conclusões podemos tirar?
- A quantidade de chuva é significativa para explicar a produção de milho.
- O termo quadrático é singificativo.
- O modelo ajustado é dado por

$$\hat{Y} =$$

Exemplo: (continuação)

- Quais as conclusões podemos tirar?
- A quantidade de chuva é significativa para explicar a produção de milho.
- O termo quadrático é singificativo.
- O modelo ajustado é dado por

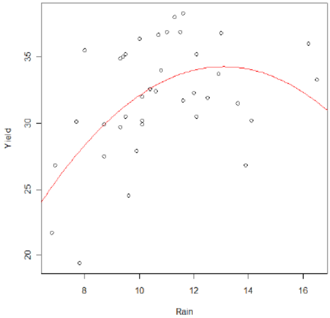
$$\hat{Y} = -5.01 + 6X - 0.23X^2 \, .$$

- Podemos usar os seguintes comandos para plotar a curva sobre o grafo de dispersão

```
plot(Rain,Yield)
x = seq(6, 17, by=0.1)
lines(x, beta[1]+beta[2]*x +
 beta[3]*x^2, col='red')
```

Exemplo: (continuação)

- A figura ao lodo mostra o gráfico com a curva.
- O modelo quadrático parece representar bem a relação entre as variáveis.
- Não parece neccessário testar polinômios de ordem mais alta.



- Podemos usar regressão polinomial quando temos mais de uma variável explicativa.
- Por exemplo, considere o caso em que temos duas variáveis.
- Além dos termos, quadráticos, cúbicos, etc, podemos incluir interação.
- Um modelo possível seria

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{2i} + \beta_4 X_{2i}^2 + \beta_5 X_{1i} X_{2i} + \epsilon_i$$

- Qual figura geométrica essa equação representa?

- Podemos usar regressão polinomial quando temos mais de uma variável explicativa.
- Por exemplo, considere o caso em que temos duas variáveis.
- Além dos termos, quadráticos, cúbicos, etc, podemos incluir interação.
- Um modelo possível seria

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \beta_3 X_{2i} + \beta_4 X_{2i}^2 + \beta_5 X_{1i} X_{2i} + \epsilon_i$$

- Qual figura geométrica essa equação representa? Uma superfície.
- De acordo com os valores dos parâmetros a superfície assume formatos distintos.

- Considere o caso em que temos

$$\beta_2 > 0 \quad \beta_4 > 0$$

- Duas dimensões convexas.
- Considere ainda que

$$\beta_5 = 0$$

- Isso significa que

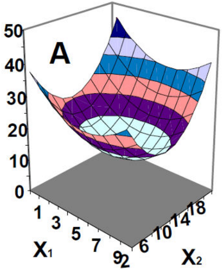
- Considere o caso em que temos

$$\beta_2 > 0 \quad \beta_4 > 0$$

- Duas dimensões convexas.
- Considere ainda que

$$\beta_5 = 0$$

- Isso significa que não há interação.
- A relação entre X_1 e Y não muda com os valores de X_2.



- Considere o caso em que temos

$$\beta_2 > 0 \quad \beta_4 > 0$$

- Duas dimensões convexas.
- Considere ainda que

$$\beta_5 \neq 0$$

- Isso significa que

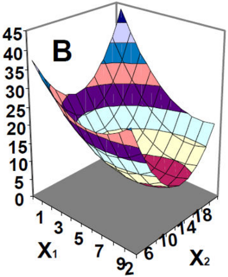
- Considere o caso em que temos

$$\beta_2 > 0 \quad \beta_4 > 0$$

- Duas dimensões convexas.
- Considere ainda que

$$\beta_5 \neq 0$$

- Isso significa que há interação.
- A relação entre X_1 e Y muda com os valores de X_2.



- Considere o caso em que temos

$$\beta_2 > 0 \quad \beta_4 < 0$$

- Uma dimensão é côncava e outra convexa.
- Considere ainda que

$$\beta_5 \neq 0$$

- Isso significa que

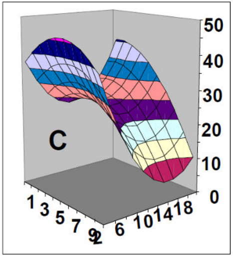
- Considere o caso em que temos

$$\beta_2 > 0 \quad \beta_4 < 0$$

- Uma dimensão é côncava e outra convexa.
- Considere ainda que

$$\beta_5 \neq 0$$

- Isso significa que há interação.
- A relação entre X_1 e Y muda com os valores de X_2.



Multicolinearidade

- Sabemos que, para ajustar o modelo

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$$

a solução de mínimos quadrados é dada por

$$\hat{\boldsymbol{\beta}} =$$

- Sabemos que, para ajustar o modelo

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon$$

a solução de mínimos quadrados é dada por

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\ .$$

- Porém, se $(\mathbf{X}'\mathbf{X})$ é singular:

- Sabemos que, para ajustar o modelo

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon$$

a solução de mínimos quadrados é dada por

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \, .$$

- Porém, se $(\mathbf{X}'\mathbf{X})$ é singular:
  - o estimador não pode ser obtido dessa forma;
  - as equações normais não têm solução única.
- Isto acontece porque?

- Sabemos que, para ajustar o modelo

$$\mathbf{Y} = \beta\mathbf{X} + \epsilon$$

a solução de mínimos quadrados é dada por

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

- Porém, se $(\mathbf{X}'\mathbf{X})$ é singular:
  - o estimador não pode ser obtido dessa forma;
  - as equações normais não têm solução única.
- Isto acontece porque?
- A colunas da matriz $\mathbf{X}$ não são linearmente independentes.
- Uma deas é combinação linear das demais.

- Observe que as colunas de $\mathbf{X}'\mathbf{X}$ são combinações lineares das colunas de $\mathbf{X}$.
- Se as colunas de $\mathbf{X}$ são linearmente de dependentes,
  - combinações lineares dessas colunas também serão.
- Portanto as colunas de $\mathbf{X}'\mathbf{X}$ não serão linearmente independentes e a matriz não tem inversa.

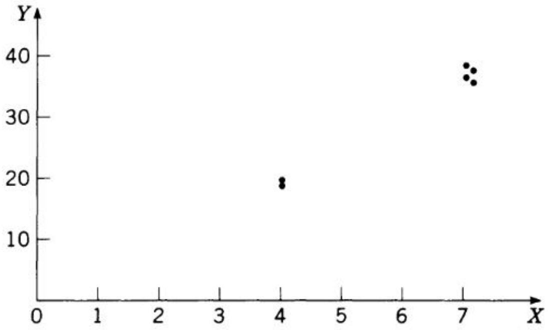
- Dizemos que existe multicolinearidade entre as colunas de X:
  - as variáveis explicativas são fortemente correlacionadas entre si.
- Se as variáveis são altamente correlacionadas, mas a correlação não é um:
  - uma não será exatamente combinação linear das demais.
- A matriz $(\mathbf{X}'\mathbf{X})$ tem inversa, ou seja $\det((\mathbf{X}'\mathbf{X})) \neq 0$.
- Porém $\det((\mathbf{X}'\mathbf{X})) \approx 0$.
- Dizemos que a matriz é mal condicionada.
- A inversa
$$(\mathbf{X}'\mathbf{X})^{-1}$$
é muito instável.
- Alterações pequenas na matriz modificam muito sua inversa.

- Esse tipo de comportamento não é desejável em um modelo de regressão.
- Veremos como identificar o problema e algumas possíveis soluções.
- Consequências para o ajuste do modelo:
  - estimadores dos coeficientes não são confiáveis;
  - estimadores com alta variância e covariância.

Exemplo

- Estamos interessados em ajustar um modelo para o seguinte conjunto de dados

X	4	4	7	7	7.1	7.1
Y	19	20	37	39	36	38



Exemplo (continuação)

- Podemos tentar um modelo quadrático

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \, .$$

- Obtemos as equações normais a seguir

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Y}$$

$$\begin{bmatrix} 6 & 36.2 & 230.82 \\ 36.2 & 230.82 & 1529.82 \\ 230.82 & 1529.82 & 10396.3362 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 189 \\ 1213.4 \\ 8078.34 \end{bmatrix}$$

- Essa matriz é aproximadamente singular.

Exemplo (continuação)

- A reta estimada e os erros padrões dos estimadores são mostrados a seguir

$$\hat{Y} = -151 + 63.5X - 5.2X^2$$

$$EP(\hat{\beta}_0) = 113 \quad EP(\hat{\beta}_1) = 44 \quad EP(\hat{\beta}_2) = 4.0$$

- Os erros padrões são muito altos.
- Os valores estimados dos coeficientes não têm precisão.
- O modelo não é últil.
- Isso ocorre pois X e X^2 são altamente correlacionadas.
- Se ajustarmos um modelo de X^2 em função de X obtemos

$$X^2 = -28.217 + 11.053X \quad R^2 = 99.99\% \, .$$

Indicações da presença de multicolinearidade:

- Coeficientes de correlação linear entre pares de variáveis explicativas ficam muito próximos de -1 ou 1.
- Gráficos de dispersão entre pares de varráveis explicativas apresentam configurações especiais:
  - indicando algum tipo de relação entre elas.
- Coeficientes de regressão apresentam sinais algébricos opostos ao esperado a partir de conhecimento teórico.
- Coeficientes de regressão sofrem grandes alterações quando:
  - uma coluna ou linha da matrix $\mathbf{X}$ é extraída;
  - ou seja, quando uma variável explicativa é retirada do modelo ou uma observação é retirada da amostra.

Indicações da presença de multicolinearidade:

- O teste F rejeita

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

  mas nenhuma das

$$H_0 : \beta_j = 0$$

  é rejeitada pelos testes t individuais.
- Variáveis explicativas que teoricamente são consideradas importantes:
  - apresentam coeficientes de regressão com estatística t muito baixa.
- Os erros padrão dos coeficientes de regressão são muito altos.

- Estes métodos de diagnósticos são informais.
- Possuem limitações importantes:
  - não fornecem uma medida do impacto da multicolinearidade;
  - não identificam a natureza da multicolinearidade.
- Diagramas de dispersão e coeficientes de correlação revelam relações entre pares de variáveis.
- Não mostram a relação entre grupos de variáveis.
- Como por exemplo, a relação entre X_1 e uma combinação linear de X_2, X_3 e X_4.

- Método formal de detectar e medir multicolinearidade:
  - análise dos fatores de inflação da variância (VIF).
- Para o modelo

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

ajustado por mínimos quadrados. A matriz de Covariâncias dos estimadores dos coeficientes é dada por

$$Var(\hat{\beta}) =$$

- Método formal de detectar e medir multicolinearidade:
  - análise dos fatores de inflação da variância (VIF).
- Para o modelo

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

ajustado por mínimos quadrados. A matriz de Covariâncias dos estimadores dos coeficientes é dada por

$$Var(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

os elementos da diagonal são iguais a

$$Var(\hat{\beta}_k) = VIF_k \left(\frac{\sigma^2}{S_k} \right) \quad \text{com} \quad S_k = \sum_{i=1}^{n} (X_{ki} - \overline{X}_k)^2$$

- Se X_k não está relacionada às outras variáveis X, tem-se $VIF_k = 1$.

- Portanto
$$VIF_k = Var(\hat{\beta}_k)\left(\frac{S_k}{\sigma^2}\right)$$
- VIF_k é chamado fator de inflação de variância de β_k.
- Mede quanto a $\frac{\sigma^2}{S_k}$ é inflacionada pela:
    - relação da variável X_k com as demais.
- Se o VIF_k aumenta, $Var(\hat{\beta}_k)$ aumenta.
- O *VIF* também pode ser calculado como o elemento da diagonal da matriz $\mathbf{C}^{-1}$;
    - onde **C** é a matriz de correlações das variáveis explicativas.
- Pode-se mostrar ainda que
$$VIF_k = \frac{1}{(1 - R_k^2)}$$
onde R_k^2 é coeficiente de determinação da regressão de X_k (variável resposta) em função das outras variáveis explicativas.

Exemplo:

- Foi feito um estudo para avaliar a quantidade de gordura corporal (Y).
- Foram selecionadas as seguintes variáveis explicativas:
  - X_1 - medida de expessura da dobra do tríceps;
  - X_2 - circunferência da coxa;
  - X_3 - circunferência do antebraço.
- Os dados de 20 mulheres de 25 a 24 anos foram coletados.

Exemplo:

- Observou-se que:
  - X_1 e X_2 são altamente correlacionadas.
  - As estimativas e os erros padrão dos β's mudam muito de um modelo para o outro.
  - Estimativas negativas dos β's, contrariando o esperado.
  - No modelo com as três preditoras, o teste F é significante (a 5%), mas todos os testes t individuais são não-significantes.
- Todos esses são indícios de multicolinearidade.
- Os valores dos VIF's são apresentados a seguir

$$VIF_1 = 708.84 \quad VIF_2 = 564.34 \quad VIF_3 = 104.61$$

- Valores altos indicam presença de multicolinearidade.

- Podemos tomar algumas medidas para corrigir o problema de multicolineridade.
- Coletar mais dados planejados para “quebrar” a relação entre as variáveis nos dados existentes.
- Considere que X_1 e X_2 são positivamente correlacionadas.
- Devemos coletar novos pares de observações (X_1, X_2) tais que:
  - sejam coletados valores baixos de X_1 e altos de X_2, e vice-versa.
- Isso pode ser uma característica intrínseca das variáveis.
- Por exemplo, as variáveis sócio-econômicas renda da família e tamanho da casa.

- Podemos ainda redefinir as variáveis preditoras.
- Por exemplo, se X_1, X_2 e X_3 são aproximadamente dependentes.
- É possível encontrar alguma função delas,

$$X = \frac{(X_1 + X_2)}{X_3} \quad X = X_1 X_2 X_3$$

  que preserve a relação das preditoras, mas elimine a multicolineridade.
- Uma outra solução é eliminar variáveis explicativas.
- Essa não é uma solução satisfatória,
  - pois as preditoras podem ter grande poder de explicação da resposta;
  - podemos estar descartando informação importante.

- Uma outra possbilidade é usar métodos alternativos de estimação.
- Um dos procedimentos possíveis é a *Ridge Regression*.
- Consiste em somar uma matriz diagonal à matriz $\mathbf{X}'\mathbf{X}$.
- A estimativa dos coeficientes fica

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X} + \kappa\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

onde $\kappa \geq 0$ é uma constante especificada pelo analista.