Análise de Regressão

Adaptado de Mine Çetinkaya-Rundel, OpenIntro

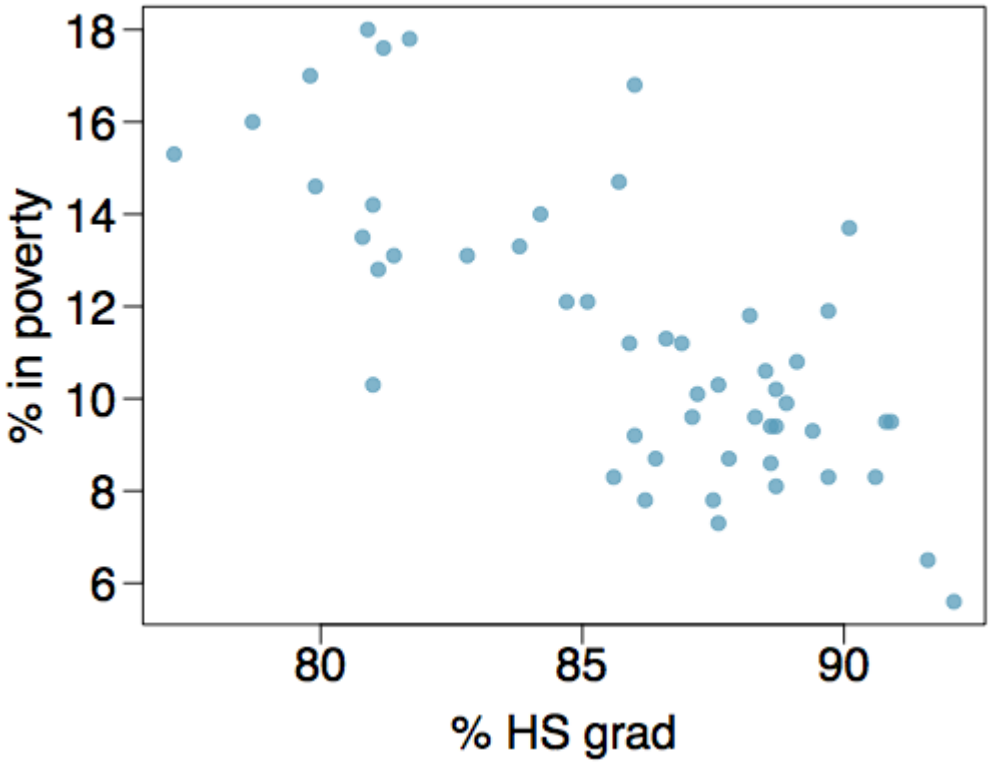
Muitos outros slides interessantes em:
https://www.openintro.org/book/os/

Alda Botelho Azevedo
Fevereiro 2025

Ajuste de Recta, Resíduos e Correlação

Pobreza vs. Taxa de Graduação do Ensino Secundário

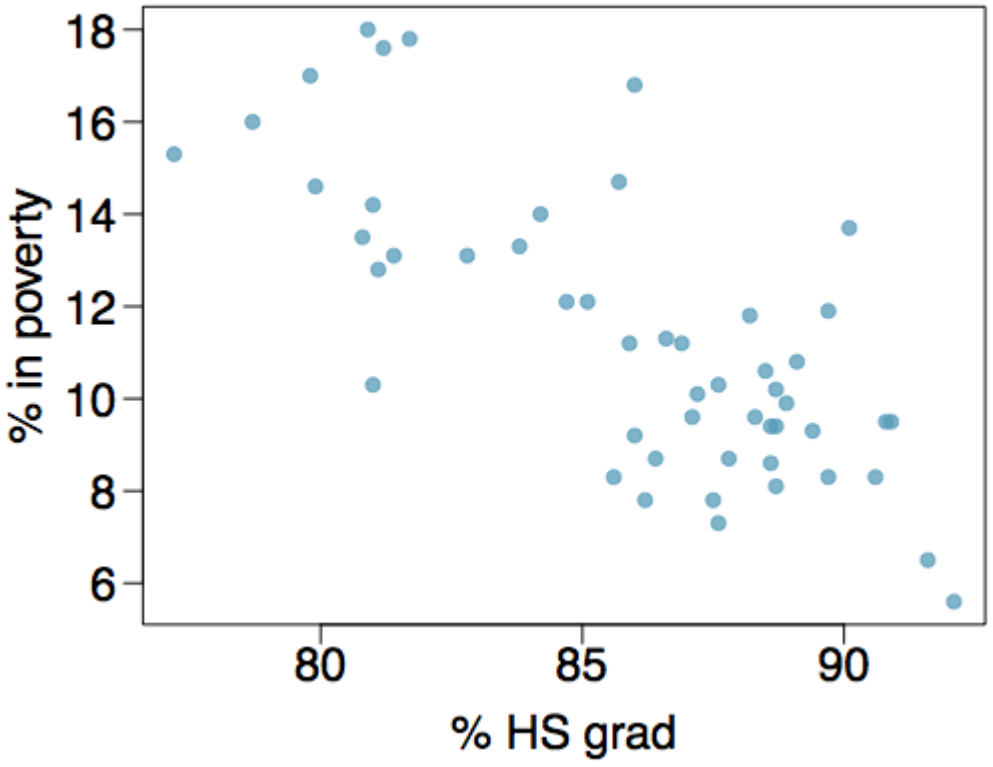
O gráfico de dispersão abaixo mostra a relação entre a taxa de graduação do ensino secundário em todos os 50 estados dos EUA e Washington, D.C. e a percentagem de residentes que vivem abaixo da linha da pobreza (rendimento inferior a $23.050 para uma família de 4 pessoas em 2012).



Var. Dependente?

Pobreza vs. Taxa de Graduação do Ensino Secundário

O gráfico de dispersão abaixo mostra a relação entre a taxa de graduação do ensino secundário em todos os 50 estados dos EUA e Washington, D.C. e a percentagem de residentes que vivem abaixo da linha da pobreza (rendimento inferior a $23.050 para uma família de 4 pessoas em 2012).

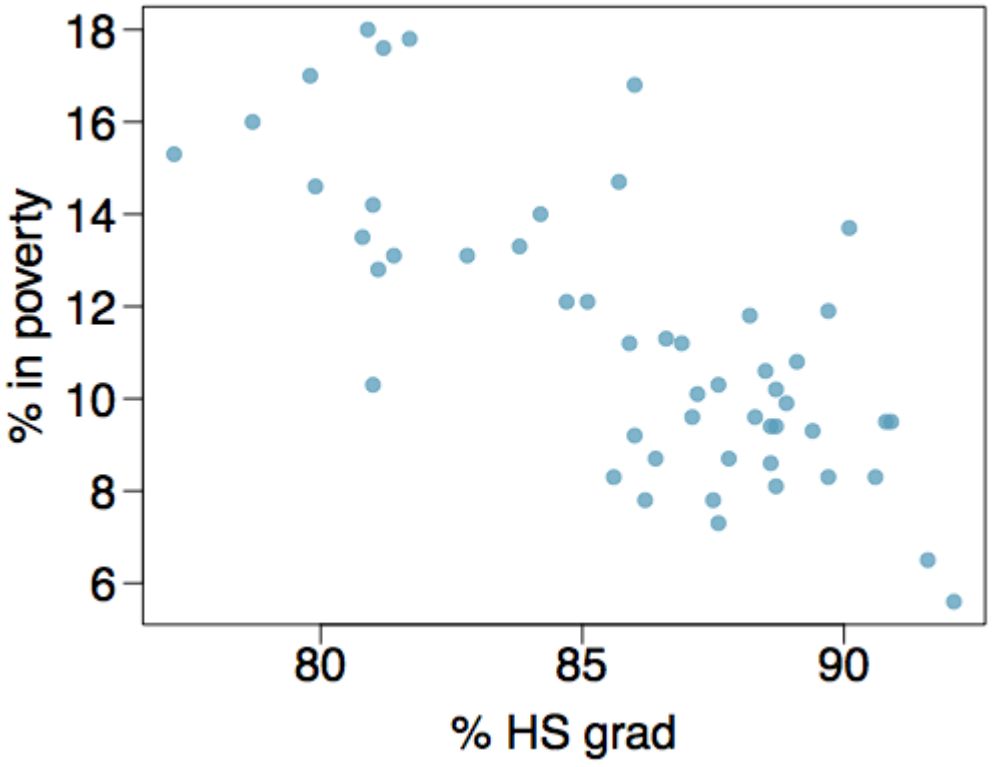


Var. Dependente?

% in poverty

Pobreza vs. Taxa de Graduação do Ensino Secundário

O gráfico de dispersão abaixo mostra a relação entre a taxa de graduação do ensino secundário em todos os 50 estados dos EUA e Washington, D.C. e a percentagem de residentes que vivem abaixo da linha da pobreza (rendimento inferior a $23.050 para uma família de 4 pessoas em 2012).



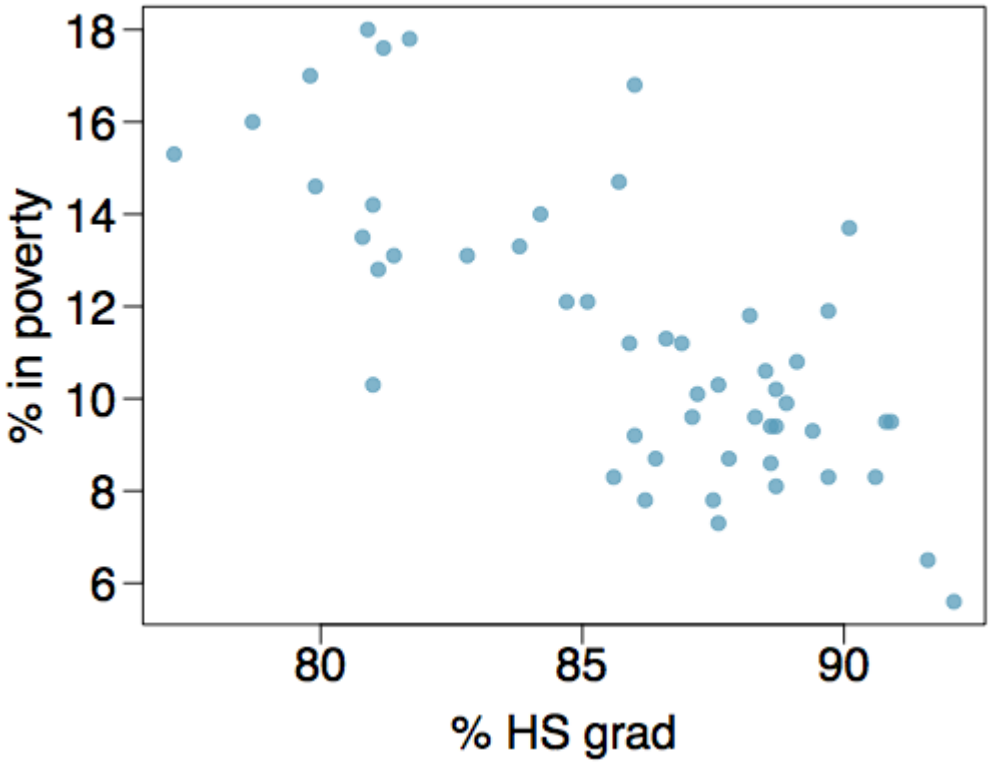
Var. Dependente?

% in poverty

Var. Independente?

Pobreza vs. Taxa de Graduação do Ensino Secundário

O gráfico de dispersão abaixo mostra a relação entre a taxa de graduação do ensino secundário em todos os 50 estados dos EUA e Washington, D.C. e a percentagem de residentes que vivem abaixo da linha da pobreza (rendimento inferior a $23.050 para uma família de 4 pessoas em 2012).



Var. Dependente?

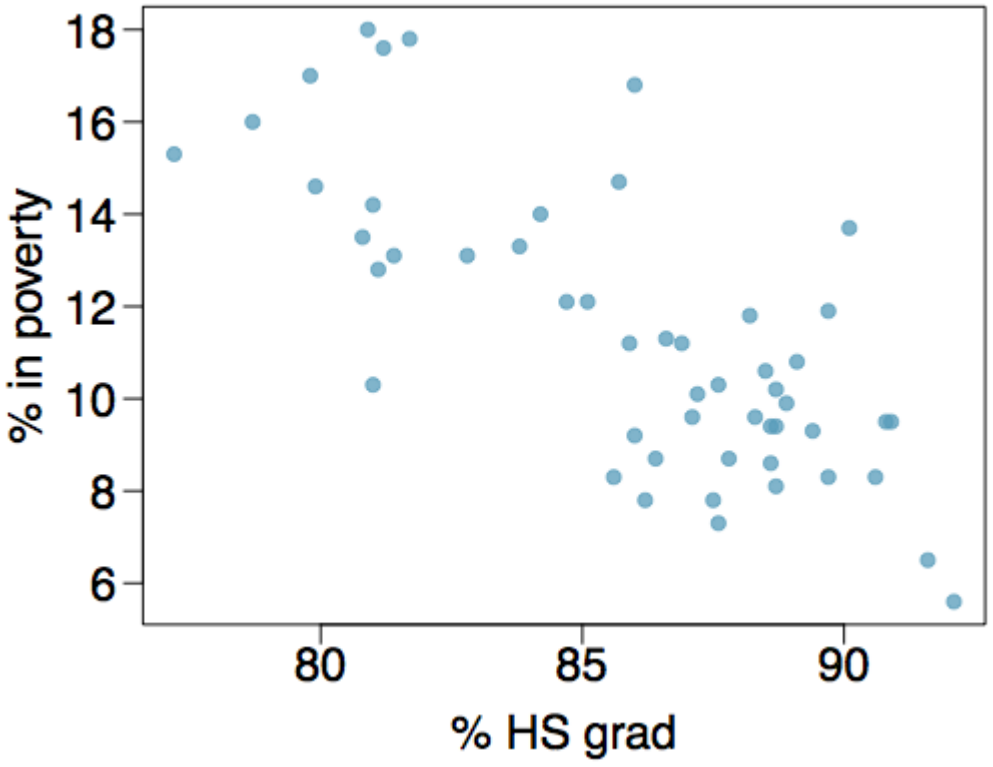
% in poverty

Var. Independente?

% HS grad

Pobreza vs. Taxa de Graduação do Ensino Secundário

O gráfico de dispersão abaixo mostra a relação entre a taxa de graduação do ensino secundário em todos os 50 estados dos EUA e Washington, D.C. e a percentagem de residentes que vivem abaixo da linha da pobreza (rendimento inferior a $23.050 para uma família de 4 pessoas em 2012).



Var. Dependente?

% in poverty

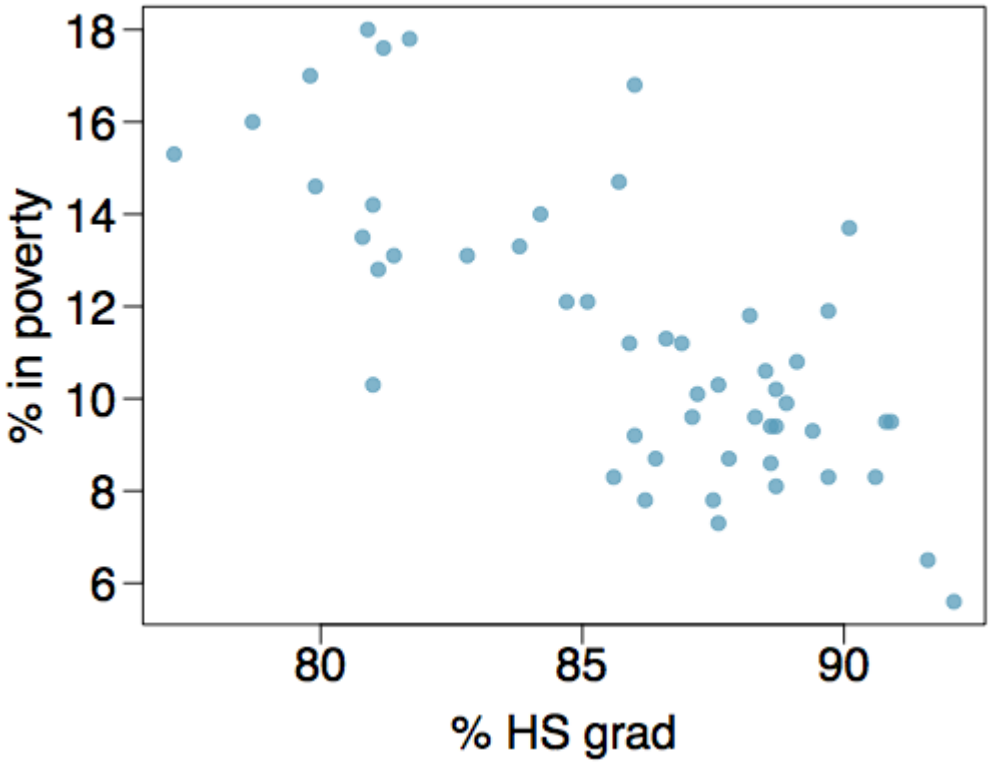
Var. Independente?

% HS grad

Relação entre variáveis?

Pobreza vs. Taxa de Graduação do Ensino Secundário

O gráfico de dispersão abaixo mostra a relação entre a taxa de graduação do ensino secundário em todos os 50 estados dos EUA e Washington, D.C. e a percentagem de residentes que vivem abaixo da linha da pobreza (rendimento inferior a $23.050 para uma família de 4 pessoas em 2012).



Var. Dependente?

% in poverty

Var. Independente?

% HS grad

Relação entre variáveis?

Linear, negativa, moderadamente forte

Poverty vs. HS graduate rate

O modelo linear para prever a pobreza a partir da taxa de graduação do ensino secundário nos **EUA** é:

$$\hat{poverty} = 64.78 - 0.62 * HS_{grad}$$

O "chapéu" (ˆ) é utilizado para indicar que se trata de uma estimativa.

Poverty vs. HS graduate rate

O modelo linear para prever a pobreza a partir da taxa de graduação do ensino secundário nos **EUA** é:

$$\hat{poverty} = 64.78 - 0.62 * HS_{grad}$$

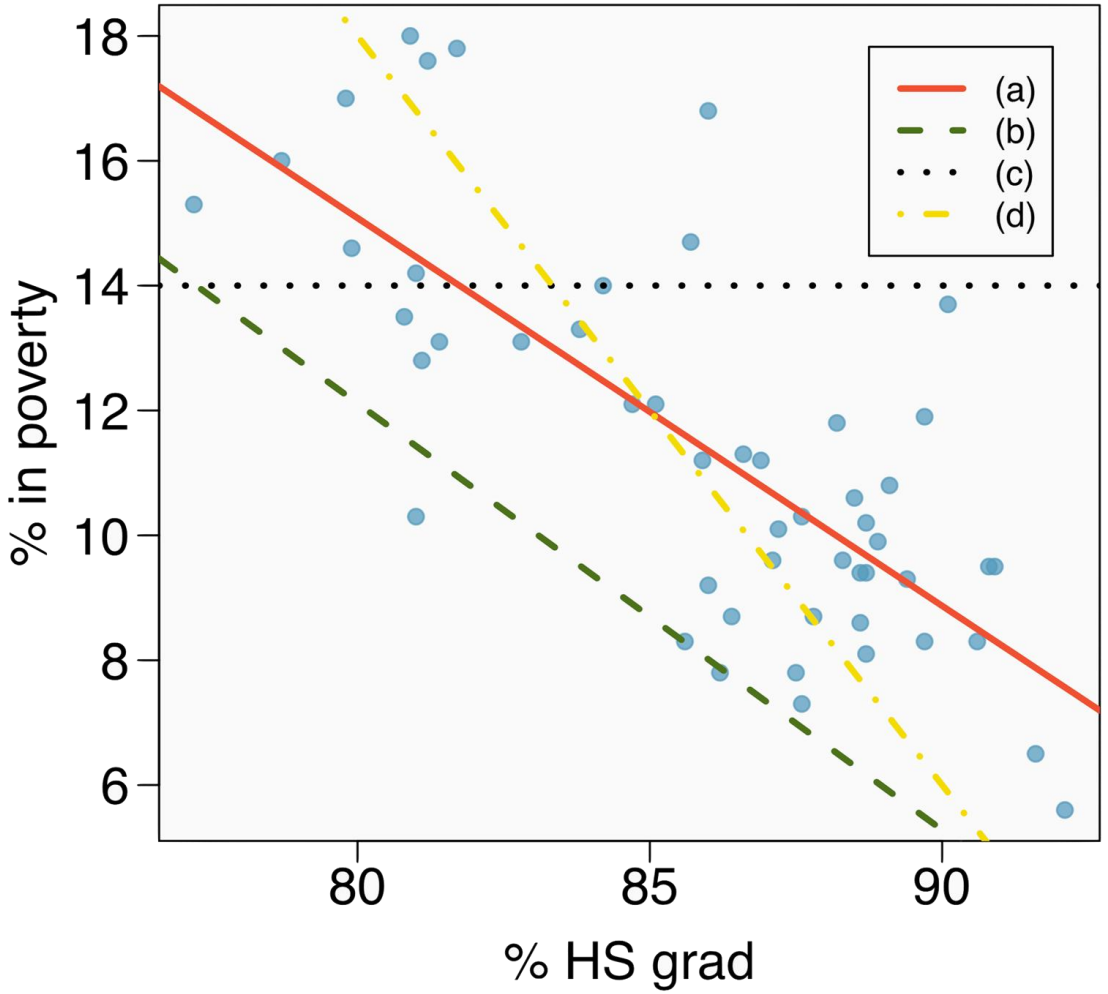
O "chapéu" (ˆ) é utilizado para indicar que se trata de uma estimativa.

A taxa de graduação do Ensino Secundário na Geórgia é de 85.1%. Qual é o nível de pobreza que o modelo prevê para este estado?

64.78 − 0.62 x 85.1 = 12.018

Visualizando a recta

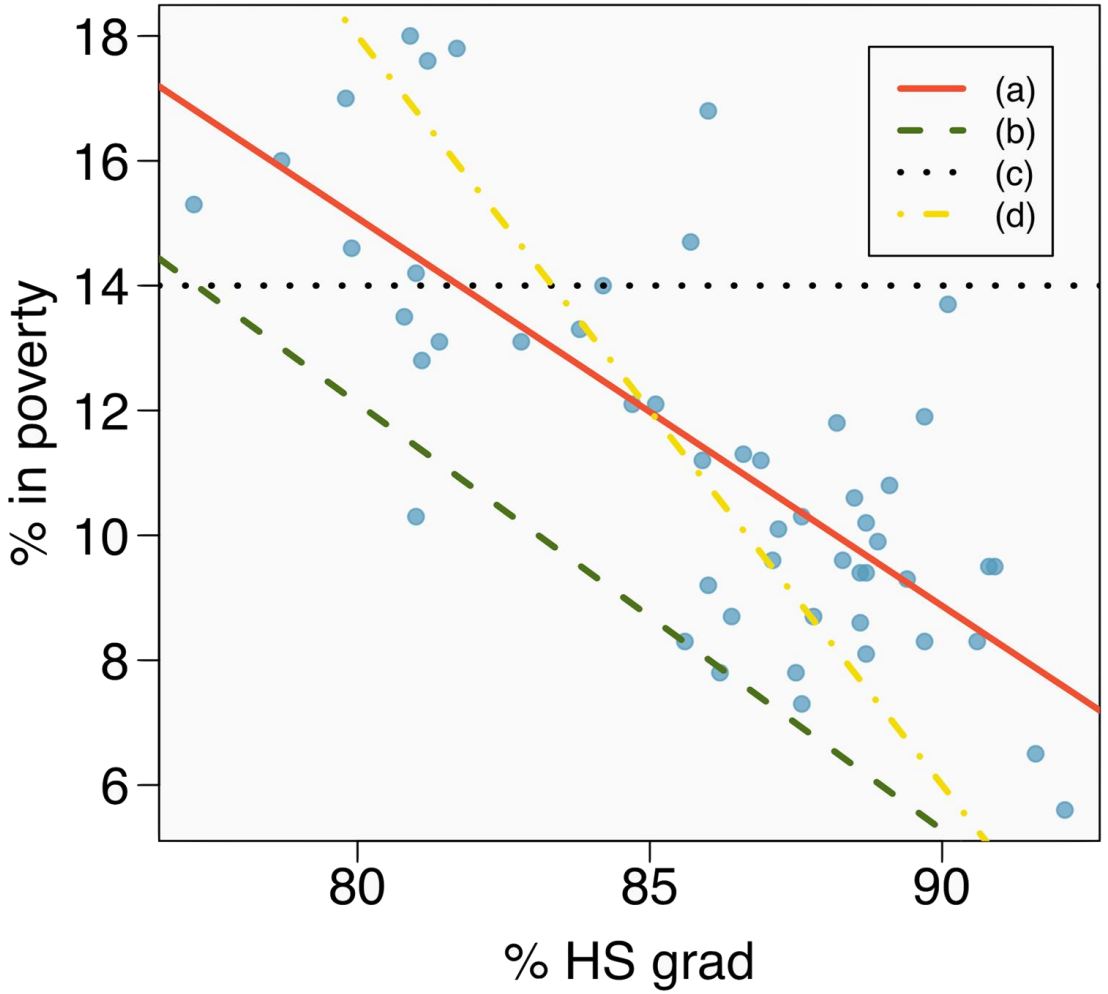
Qual das seguintes opções parece ser a linha que melhor se ajusta à relação linear entre a % de pobreza e a % de graduação no ensino secundário?



Visualizando a recta

Qual das seguintes opções parece ser a linha que melhor se ajusta à relação linear entre a % de pobreza e a % de graduação no ensino secundário?

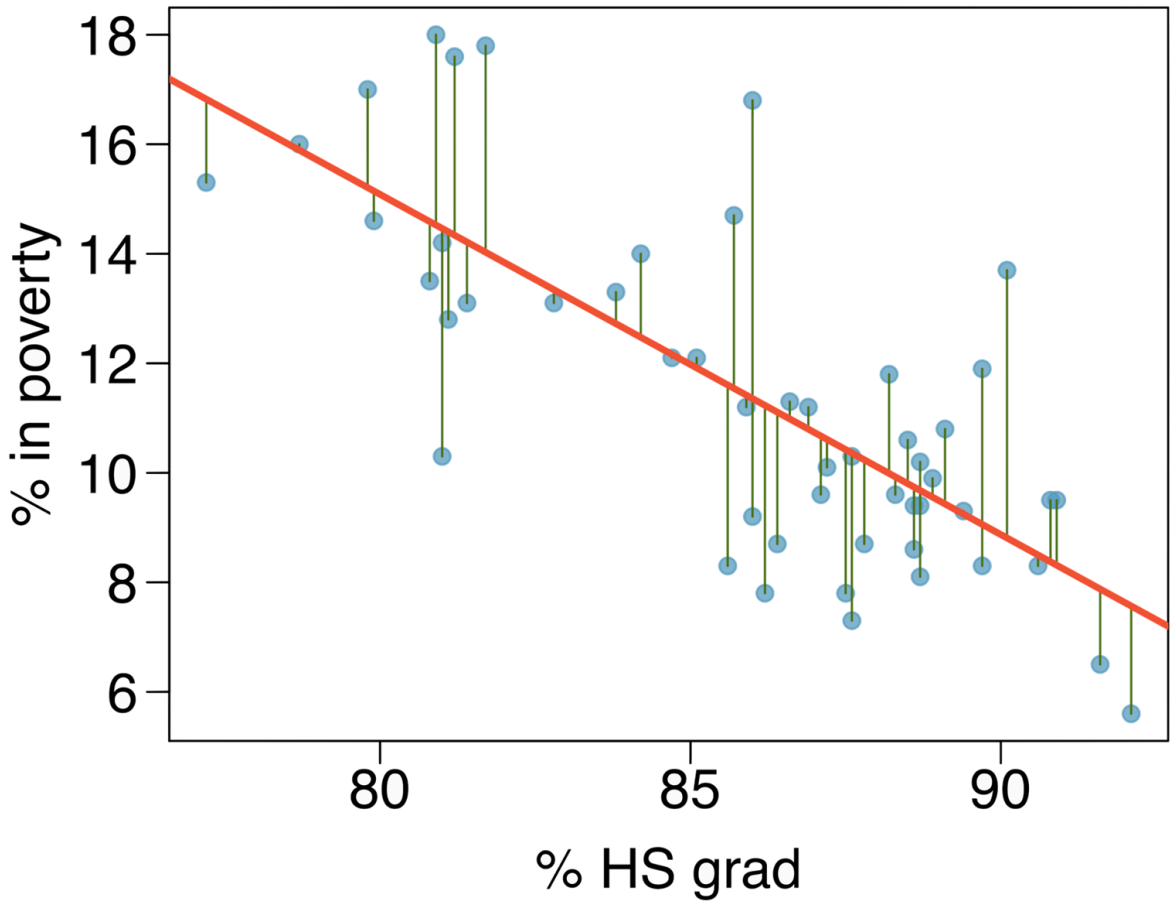
(a)



Resíduos

Os resíduos são os valores restantes após o ajuste do modelo.

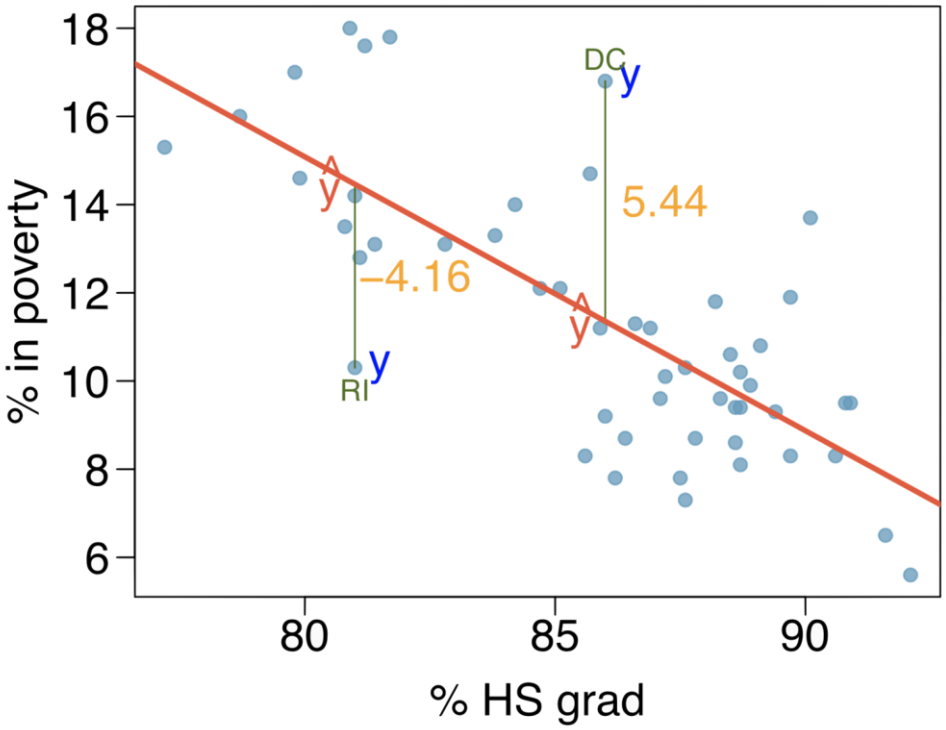
Dados=Ajuste+Resíduos



Resíduos (cont.)

Resíduo é a diferença entre o observado (y_i) e o preditto $\hat{y}_i$.

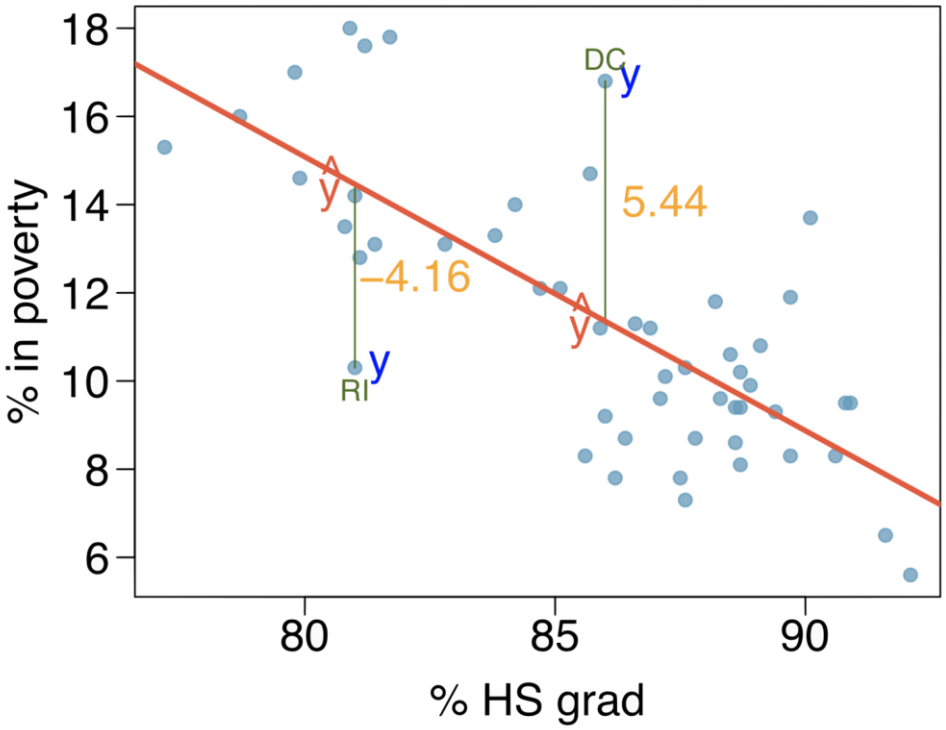
$$e_i = y_i - \hat{y}_i$$



Resíduos (cont.)

Resíduo é a diferença entre o observado (y_i) e o preditto $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i$$



Quantificação da relação

A correlação descreve a força da associação linear entre duas variáveis.

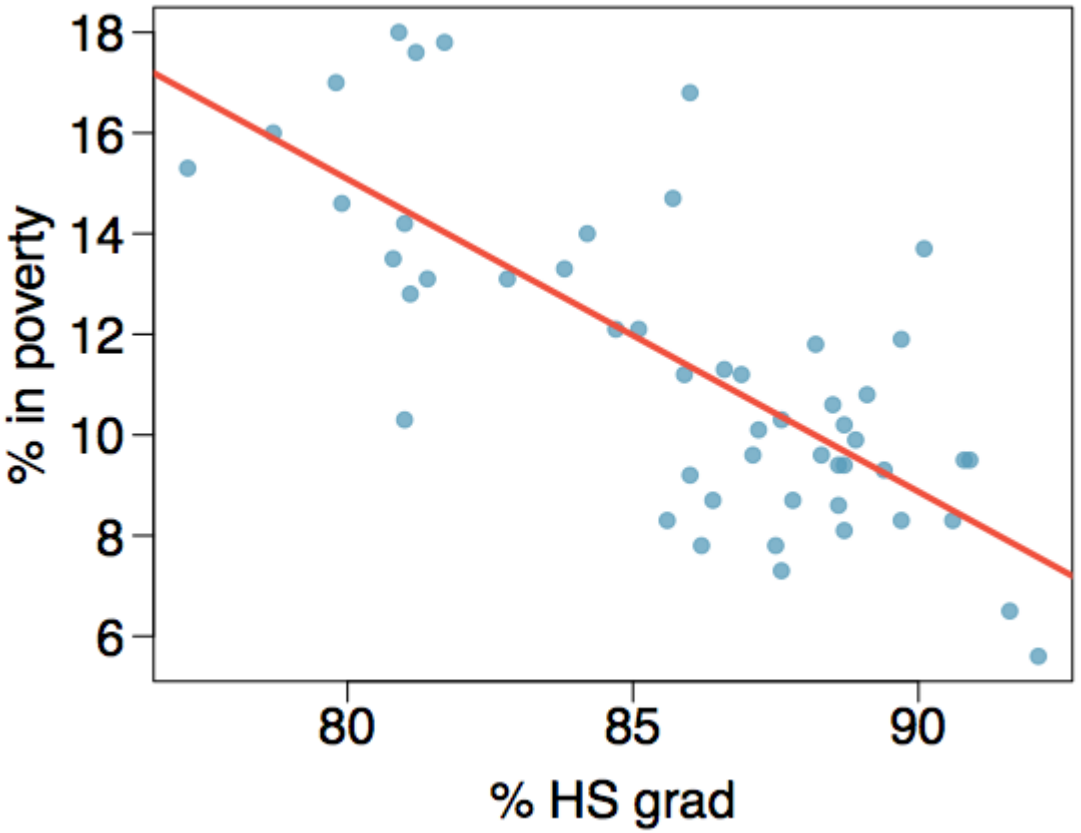
O seu valor varia entre -1 (correlação negativa perfeita) e +1 (correlação positiva perfeita).

Um valor de 0 indica que não existe associação linear entre as variáveis.

Estimando a correlação

Qual das seguintes opções é a melhor estimativa para a correlação entre a percentagem de pessoas em situação de pobreza e a percentagem de graduados do ensino secundário?

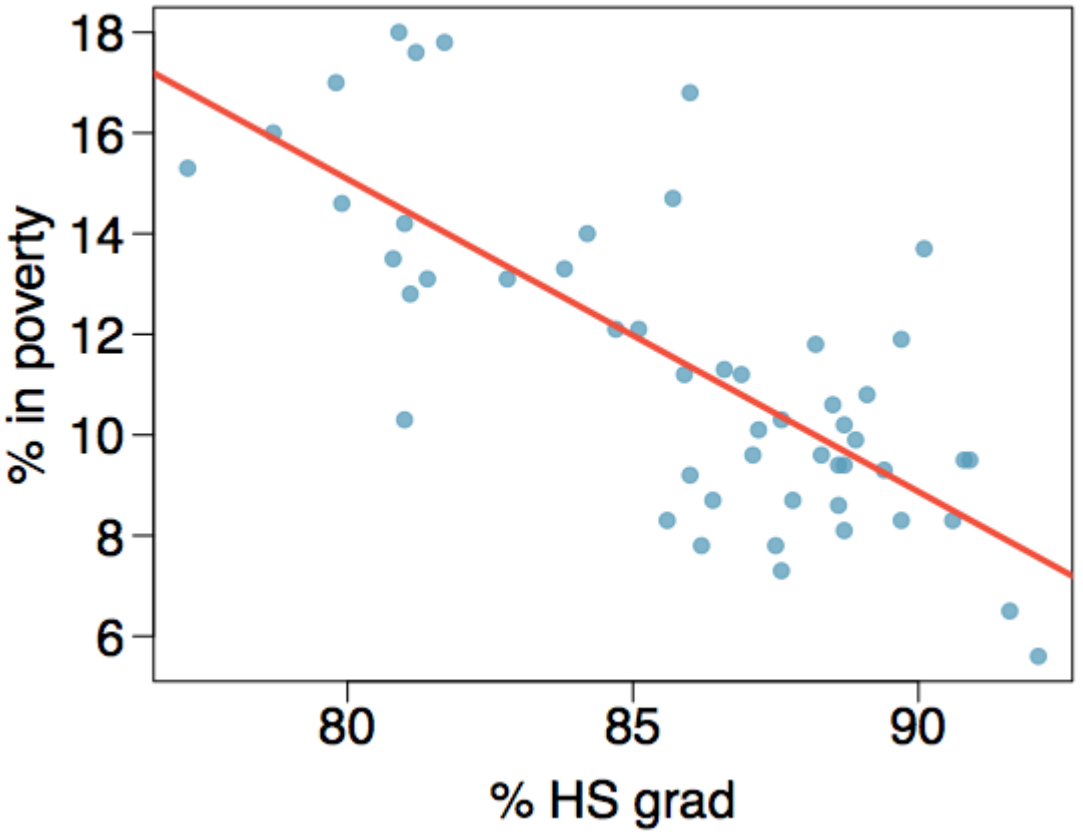
(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



Estimando a correlação

Qual das seguintes opções é a melhor estimativa para a correlação entre a percentagem de pessoas em situação de pobreza e a percentagem de graduados do ensino secundário?

(a) 0.6
(b) -0.75
(c) -0.1
(d) 0.02
(e) -1.5



Estimando a correlação

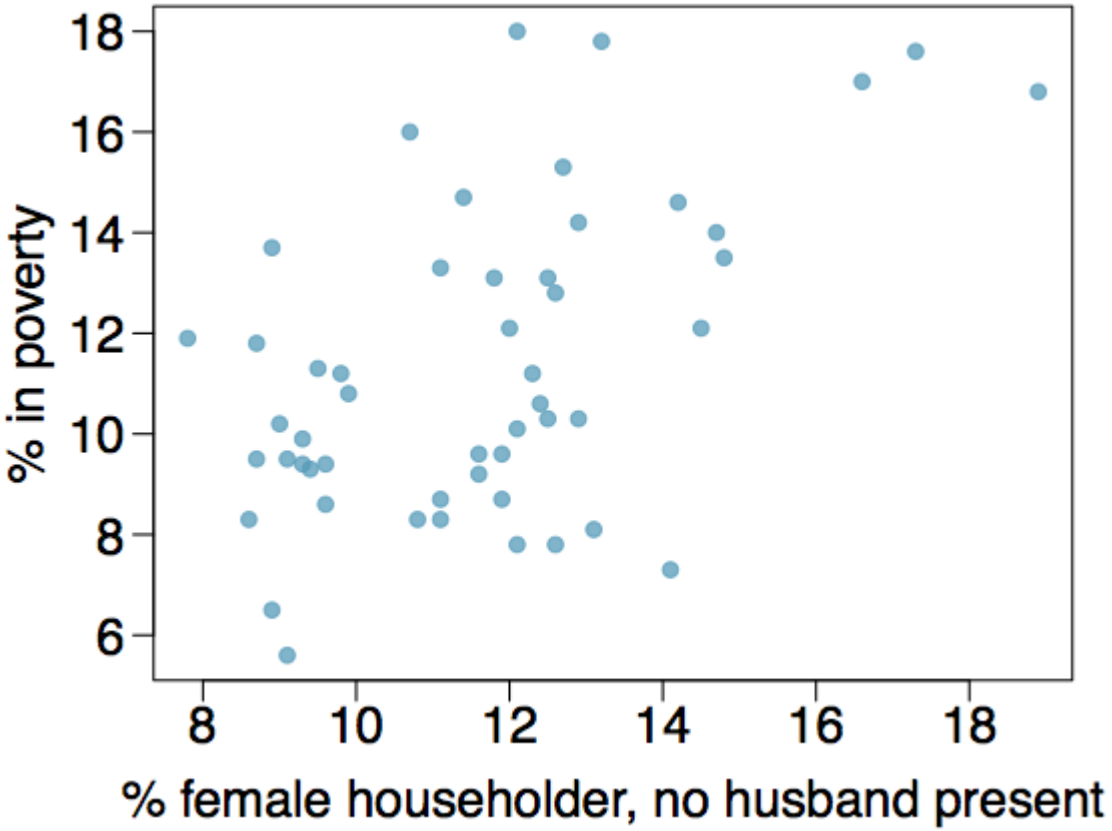
Qual das seguintes opções é a melhor estimativa para a correlação entre a percentagem de pessoas em situação de pobreza e a percentagem de agregados familiares liderados por mulheres?

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5

Estimando a correlação

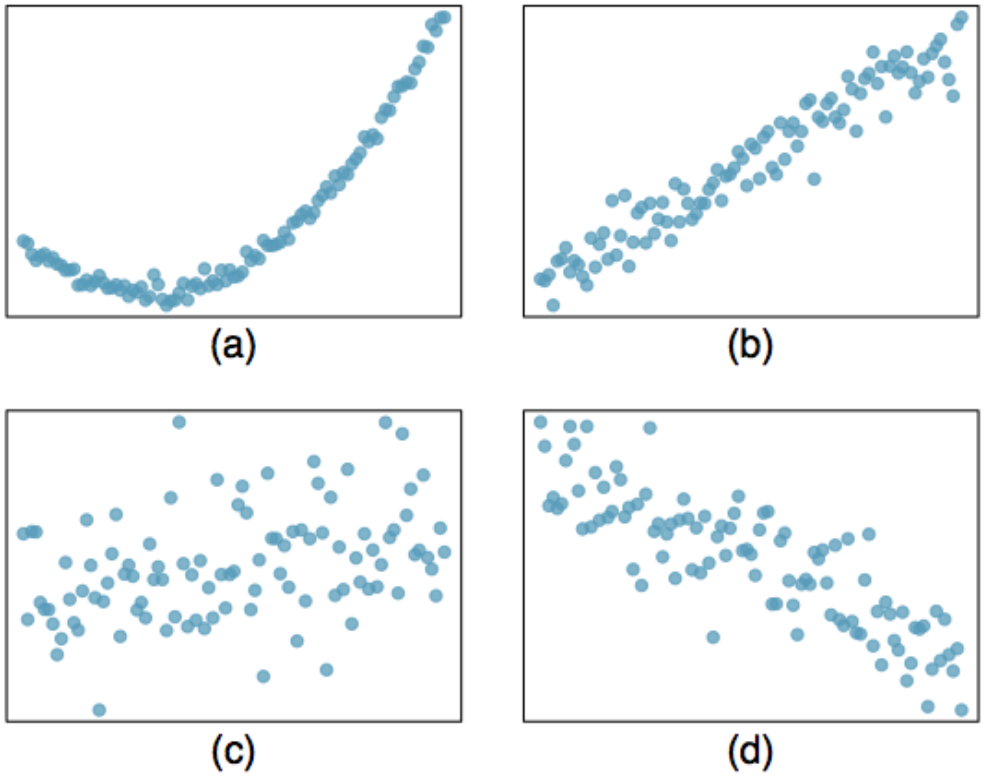
Qual das seguintes opções é a melhor estimativa para a correlação entre a percentagem de pessoas em situação de pobreza e a percentagem de agregados familiares liderados por mulheres?

(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5



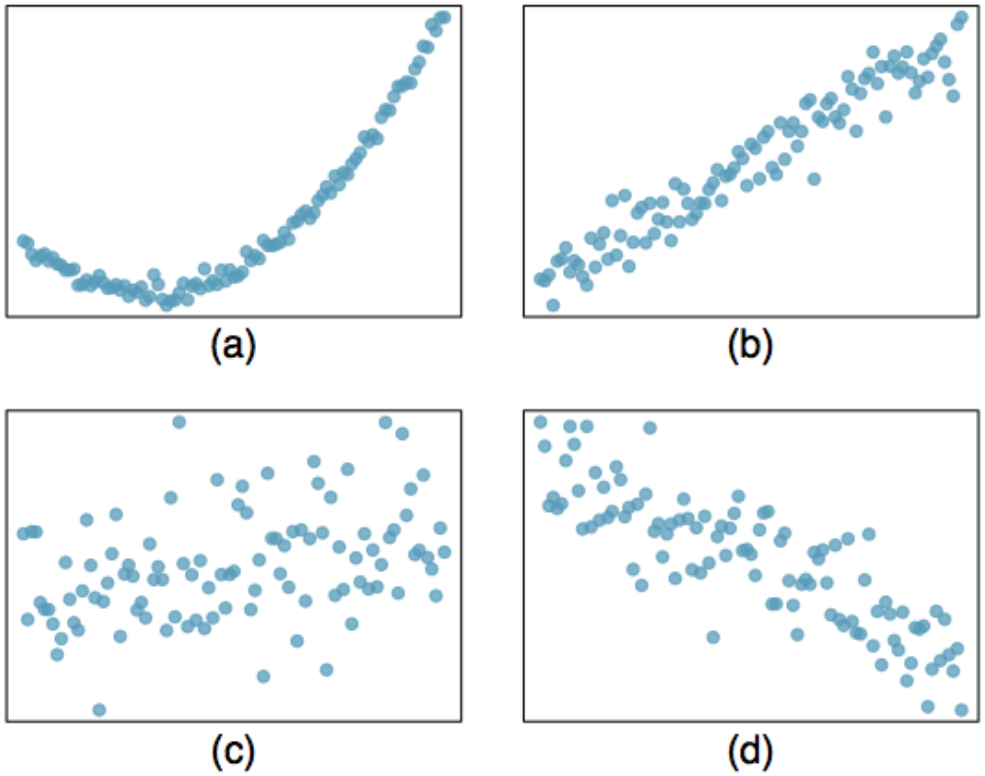
Estimando a correlação

Qual das seguintes opções tem a correlação mais forte, ou seja, o coeficiente de correlação mais próximo de +1 ou -1?



Estimando a correlação

Qual das seguintes opções tem a correlação mais forte, ou seja, o coeficiente de correlação mais próximo de +1 ou -1?



(b)

Ajustando uma recta pelo método dos mínimos quadrados

Uma medida para a melhor recta

- Queremos uma linha que tenha resíduos pequenos:

Opção 1: Minimizar a soma das magnitudes (valores absolutos) dos resíduos

$$|e_1| + |e_2| + \ldots + |e_n|$$

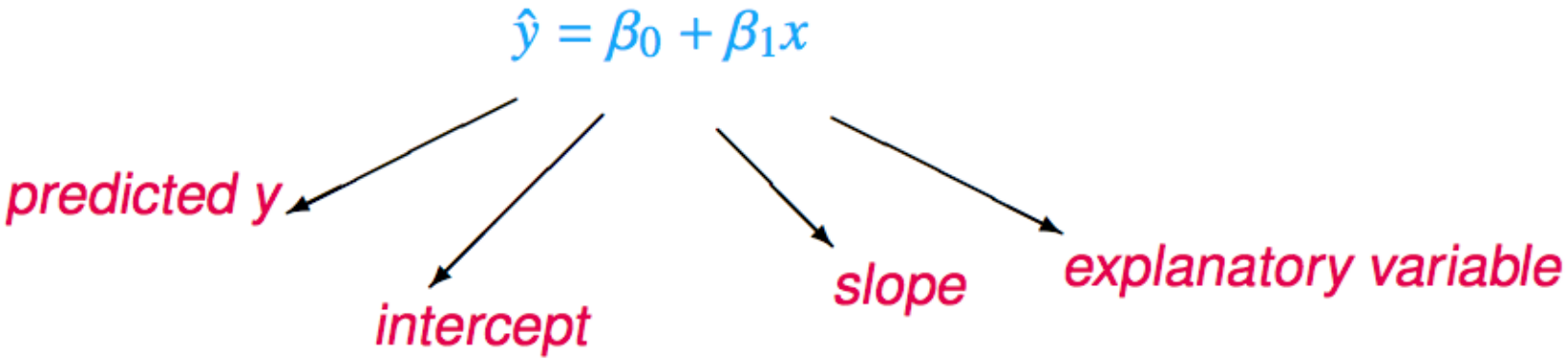
Opção 2: Minimizar a soma dos quadrados dos resíduos -- mínimos quadrados

$$e_1^2 + e_2^2 + \ldots + e_n^2$$

- Por que mínimos quadrados?

1. É o método mais utilizado.
2. É mais fácil de calcular manualmente e usando software.
3. Em muitas aplicações, um resíduo duas vezes maior que outro é geralmente mais do que o dobro pior.

The least squares line



Notação:

- Intercepto (ou ordenada na origem):
  - Parâmetro: β_0
  - Estimativa pontual: b_0

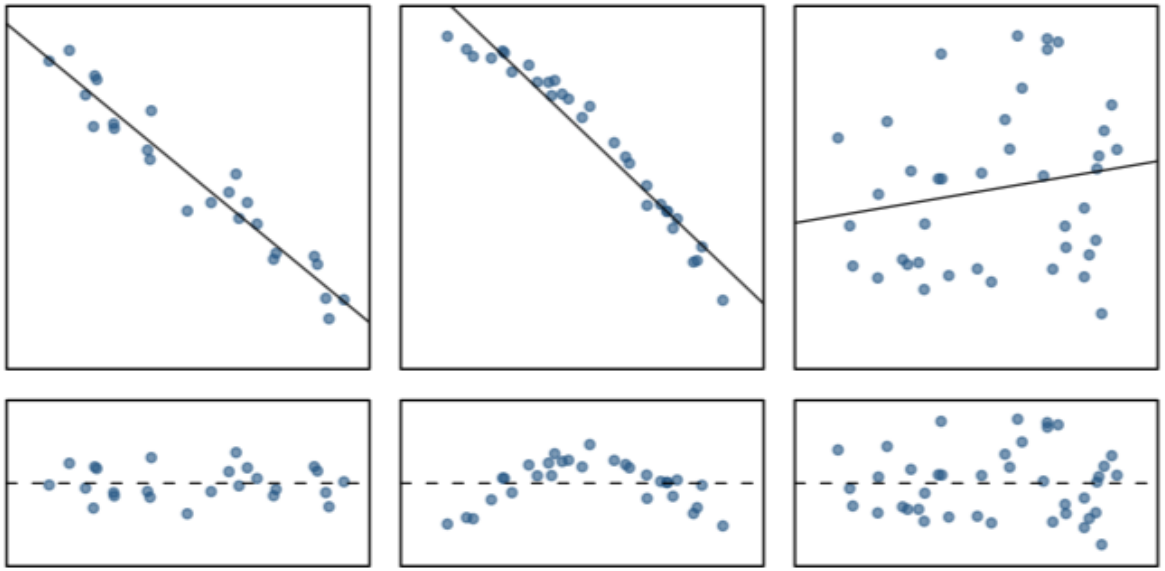
- Declive:
  - Parâmetro : β_1
  - Estimativa pontual: b_1

Condições para a linha dos mínimos quadrados

1. Linearidade
2. Resíduos aproximadamente normais
3. Variabilidade constante
4. Não haver outliers extremos

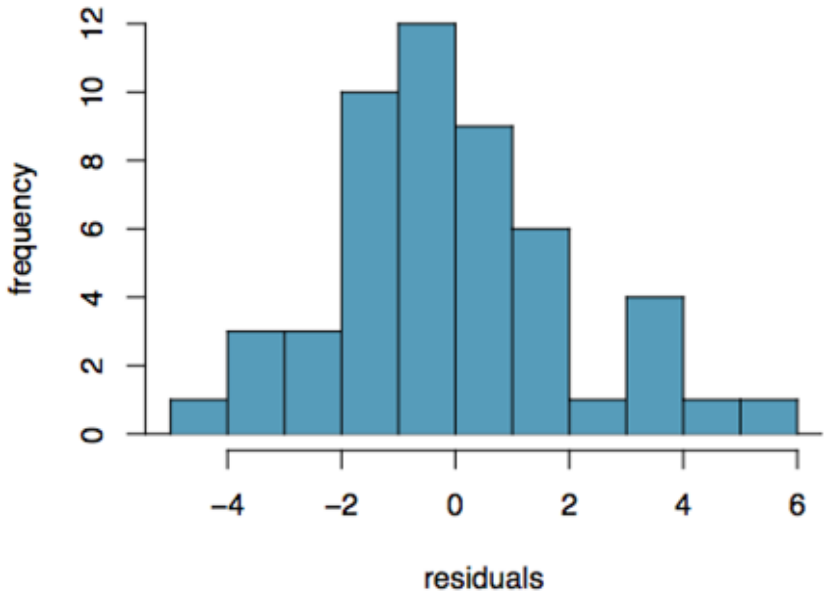
Condições: (1) Linearidade

- A relação entre a variável explanatória e a variável de resposta deve ser linear.

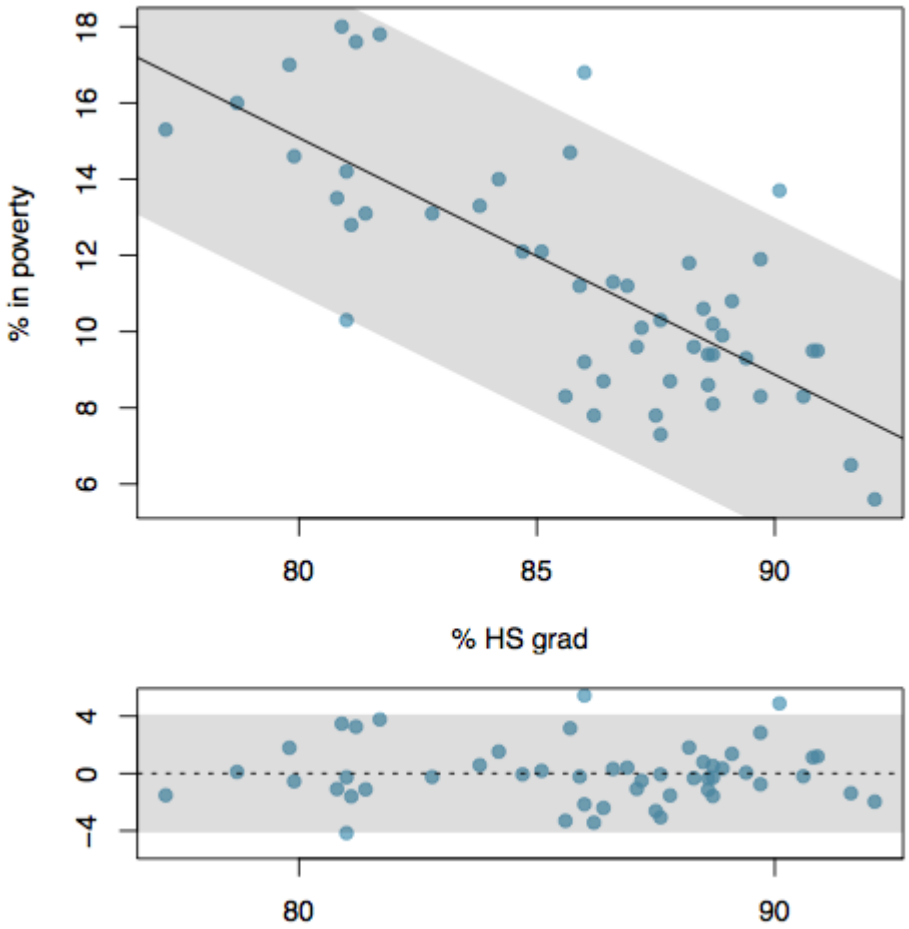


Condições: (2) Resíduos aproximadamente normais.

- Resíduos aproximadamente normais.
- Essa condição pode não ser satisfeita quando há observações incomuns que não seguem a tendência do restante dos dados.
- Verificar usando um histograma ou um gráfico de probabilidade normal dos resíduos.



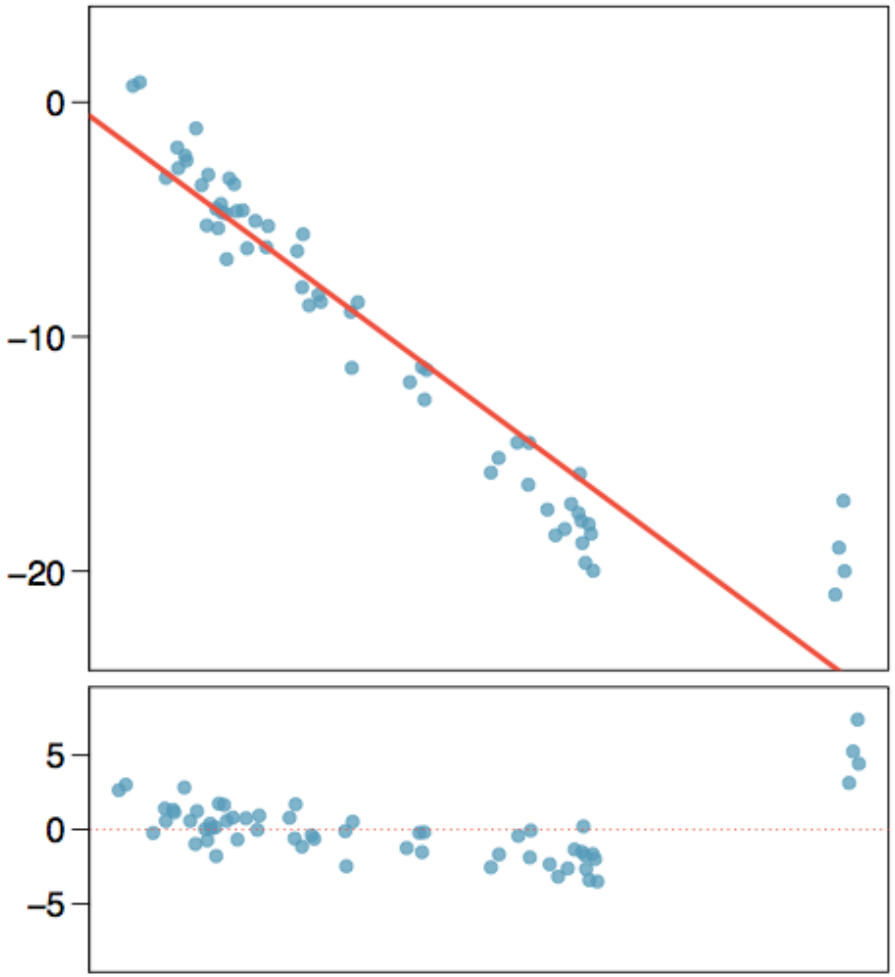
Condições: (3) Variabilidade constante



- A variabilidade dos pontos em torno da recta de mínimos quadrados deve ser aproximadamente constante.
- Isso implica que a variabilidade dos resíduos em torno da linha 0 também deve ser aproximadamente constante. Também chamada de **homocedasticidade**.
- Verifique usando um histograma ou um gráfico de probabilidade normal dos resíduos.

Condições: (4) Não haver outliers extremos

Como é que os outliers influenciam a recta dos mínimos quadrados neste gráfico?

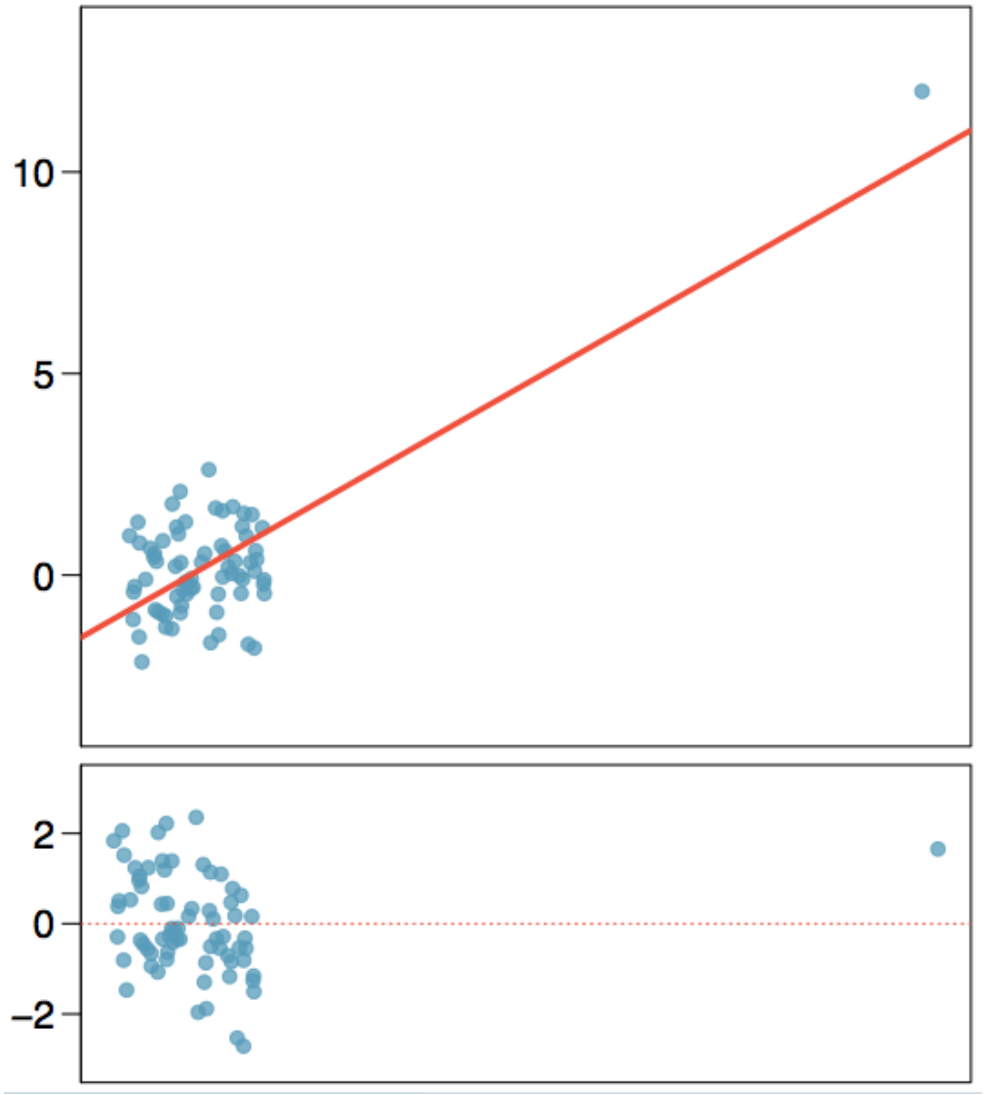


Para responder a esta pergunta, pense em onde estaria a linha de regressão com e sem os outliers.
Sem os outliers, a recta de regressão seria mais inclinada e estaria mais próxima do grupo maior de observações.
Com os outliers, a linha é puxada para cima e afastada de algumas das observações no grupo maior.

Tipos de outliers

Como é que os outliers influenciam a recta dos mínimos quadrados neste gráfico?

Sem o outlier não haveria relação evidente entre x e y.



Terminologia

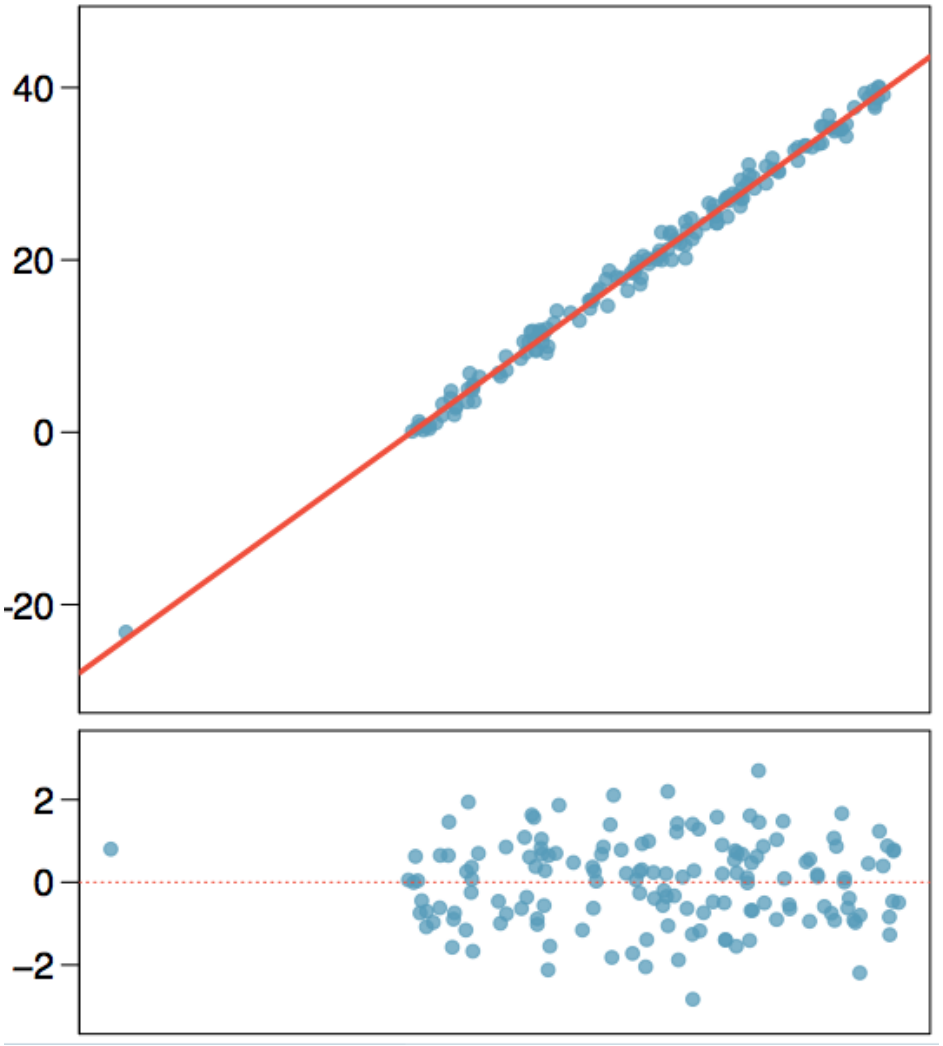
- Os outliers são pontos que se encontram afastados do conjunto principal de pontos.
- Outliers que estão horizontalmente afastados do centro do conjunto de pontos são chamados pontos de alavancagem elevada. Pontos de alavancagem elevada que realmente influenciam a inclinação da linha de regressão são chamados pontos influentes. Para determinar se um ponto é influente, visualize a recta de regressão com e sem esse ponto. A inclinação da linha muda consideravelmente?
  - Se sim, então o ponto é influente.
  - Se não, então poderá não ser um ponto influente.

Nota: existem métricas para detectar pontos influentes (distância de Cook, resíduos podronizados e estudantizados, DFBETA, …). Para uma análise robusta, usam-se várias métricas combinadas, em vez de apenas uma.

Tipos de outliers

Qual das opções abaixo melhor descreve o tipo de outlier?

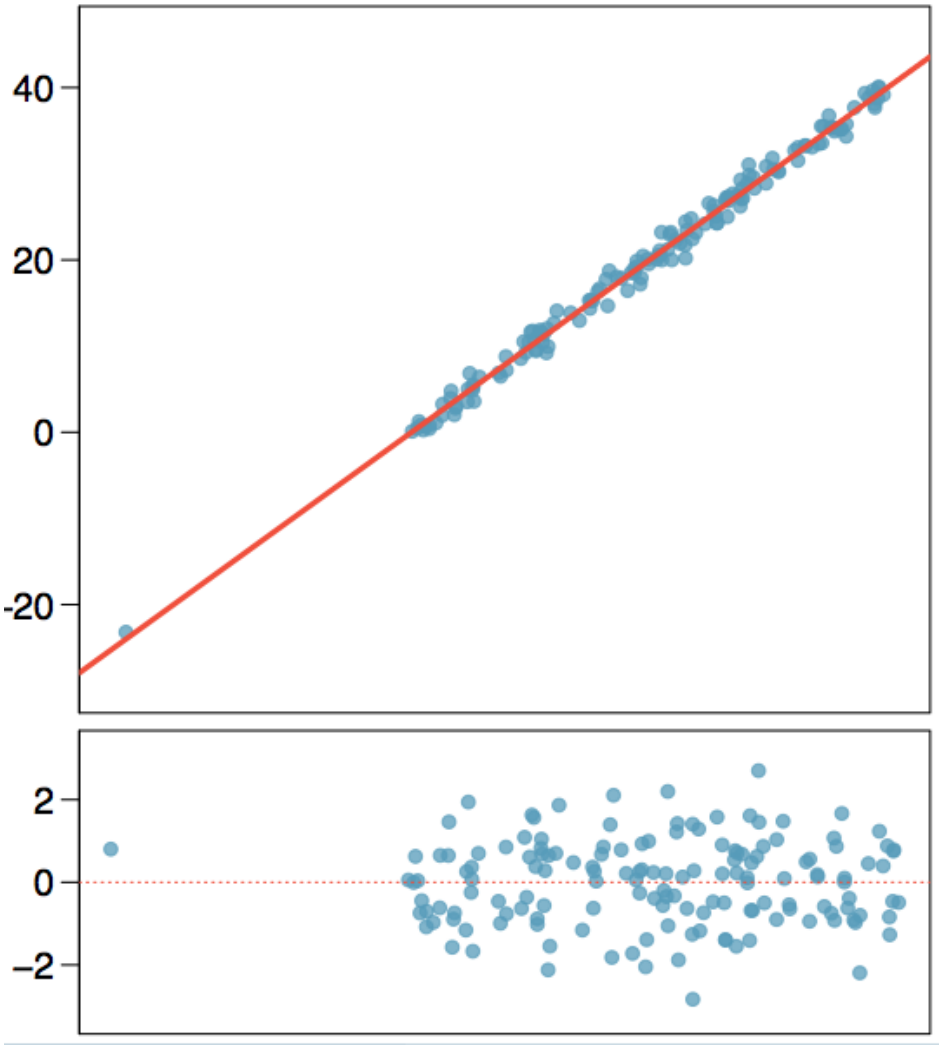
(a) ponto influente
(b) alavancagem
(c) nenhum dos anteriores
(d) não existe nenhum outlier



Tipos de outliers

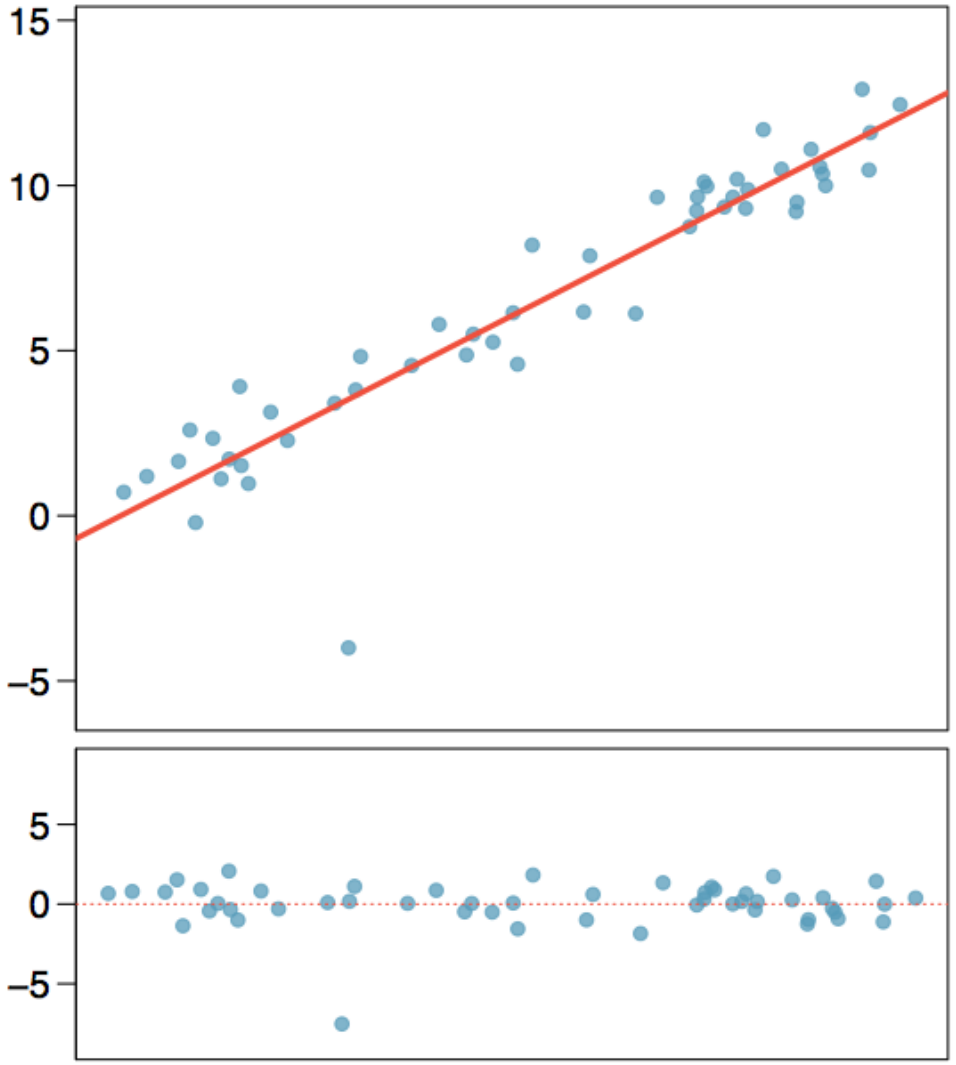
Qual das opções abaixo melhor descreve o tipo de outlier?

(a) ponto influente
(b) alavancagem
(c) nenhum dos anteriores
(d) não existe nenhum outlier



Tipos de outliers

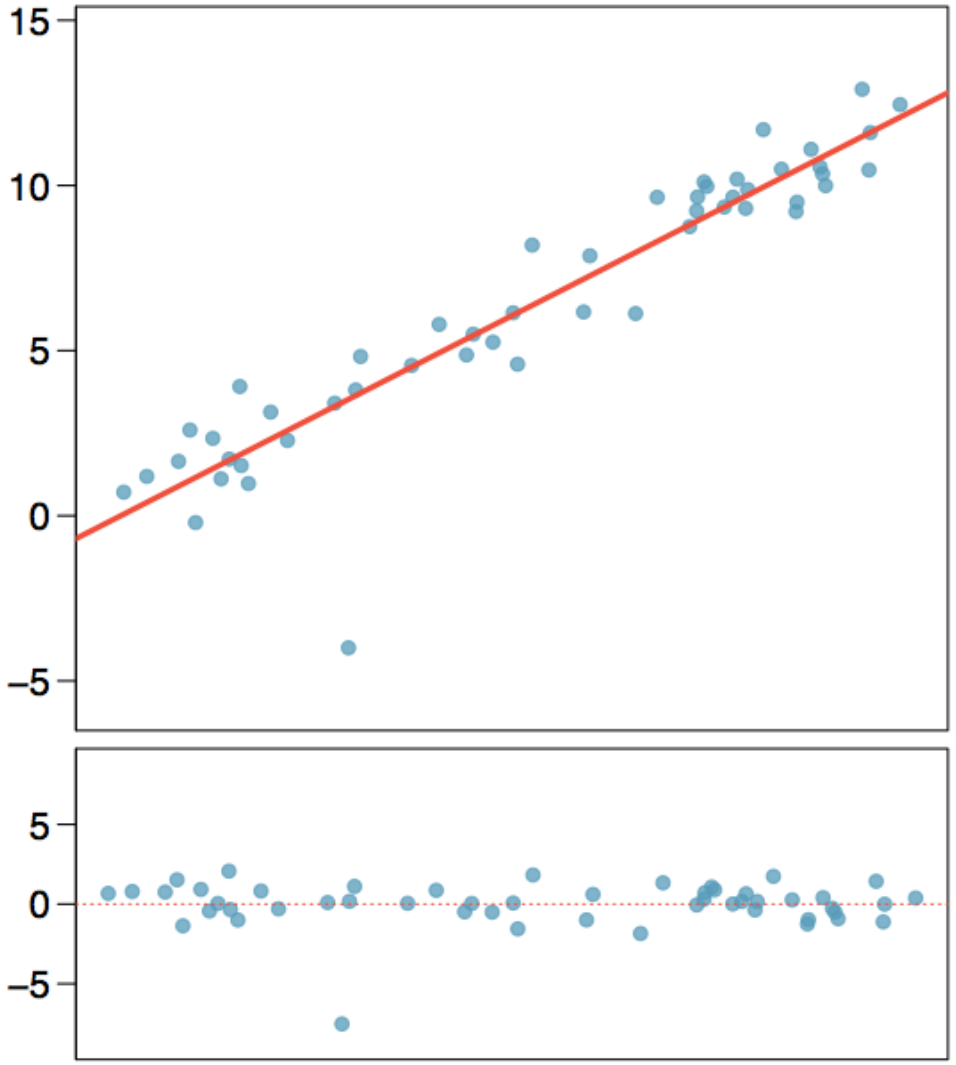
Será que este outlier influencia o declive da recta da regressão?



Tipos de outliers

Será que este outlier influencia o declive da recta da regressão?

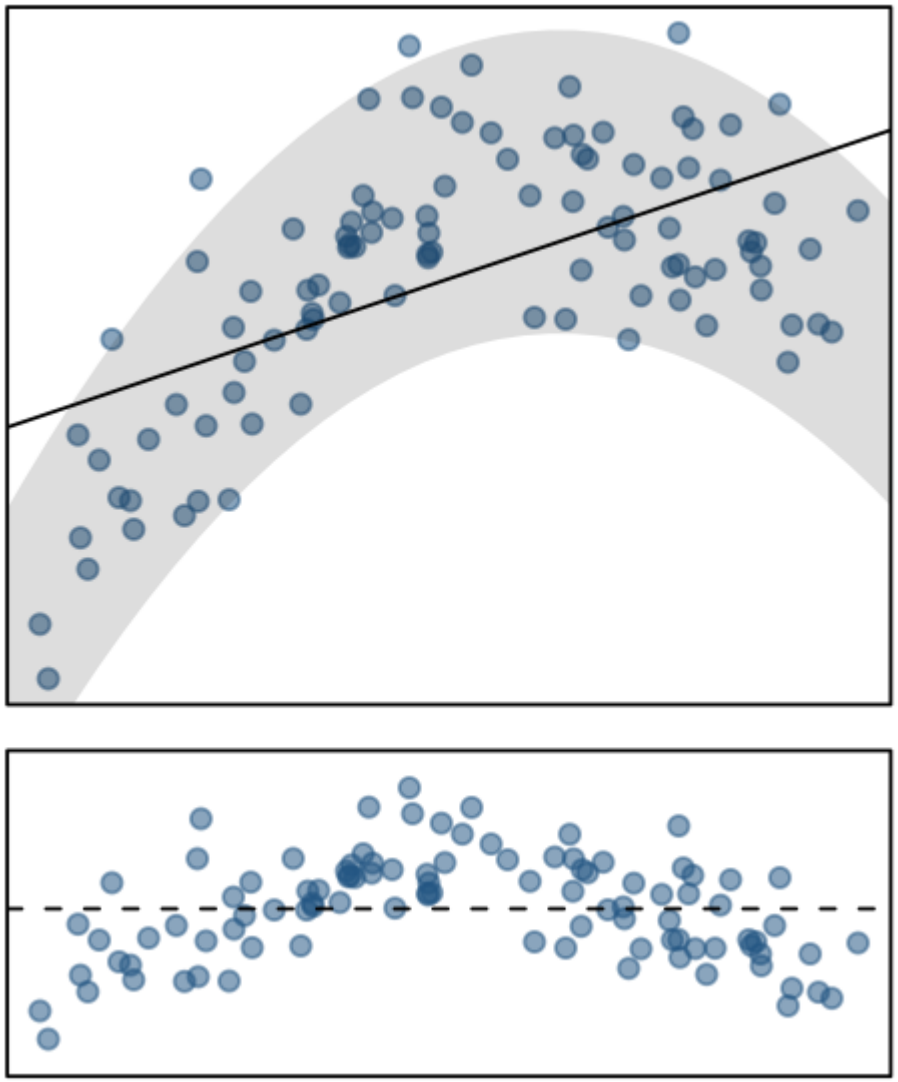
Nem por isso…



Verificação das condições

Que condição está a ser violada?

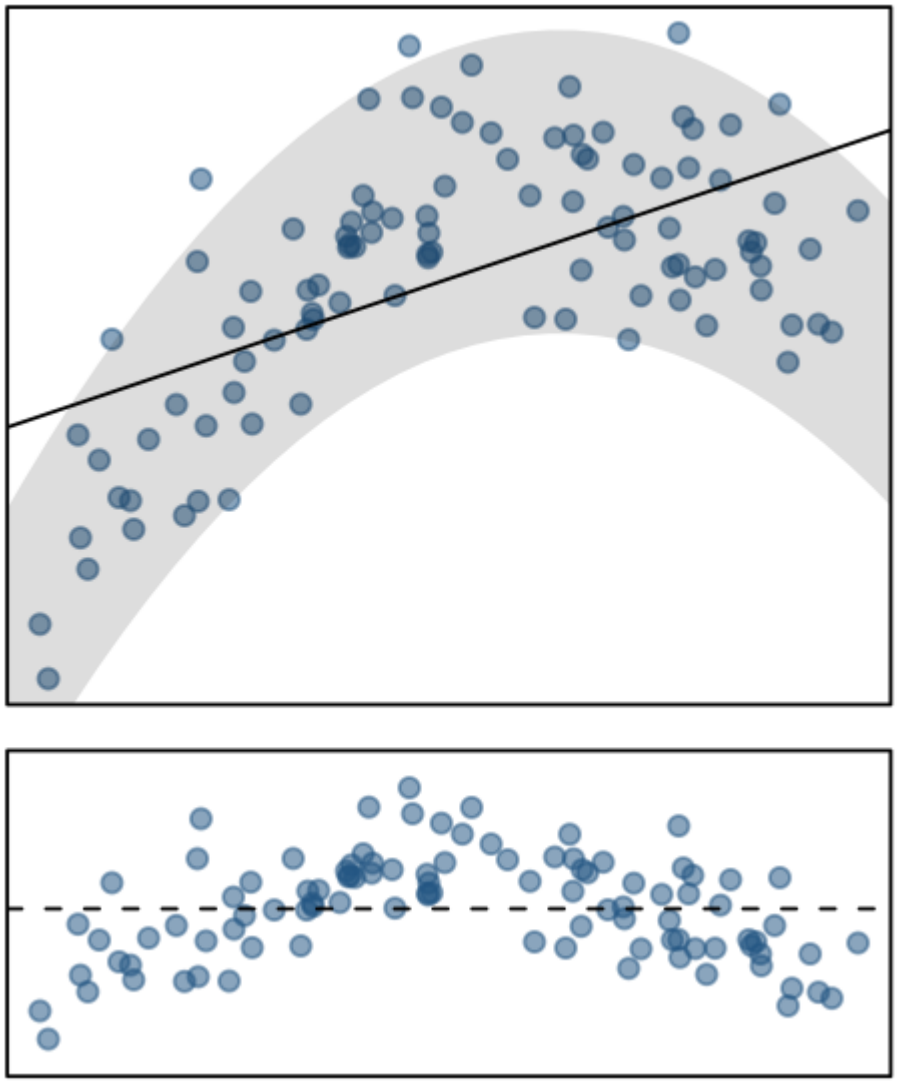
(a) Variabilidade constante
(b) Relação linear
(c) Normalidade dos resíduos
(d) Não haver outliers extremos



Verificação das condições

Que condição está a ser violada?

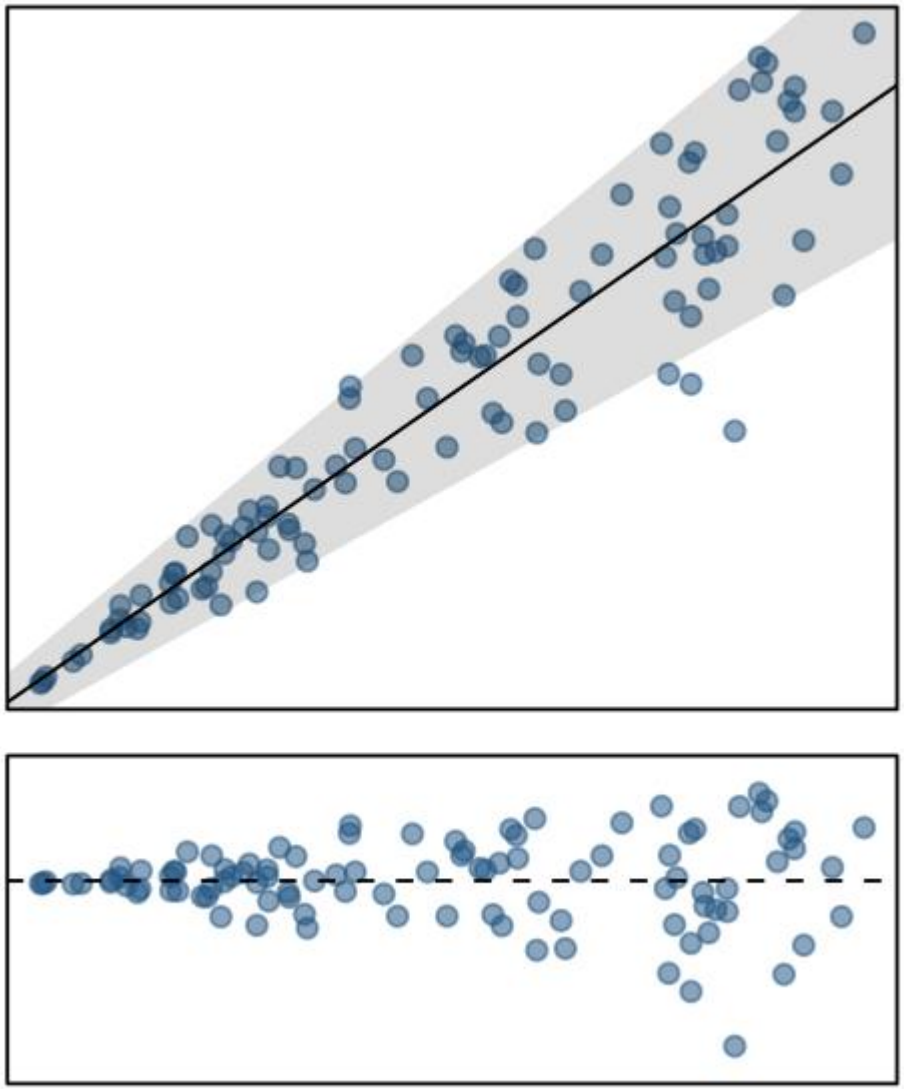
(a) Variabilidade constante
(b) Relação linear
(c) Normalidade dos resíduos
(d) Não haver outliers extremos



Verificação das condições

Que condição está a ser violada?

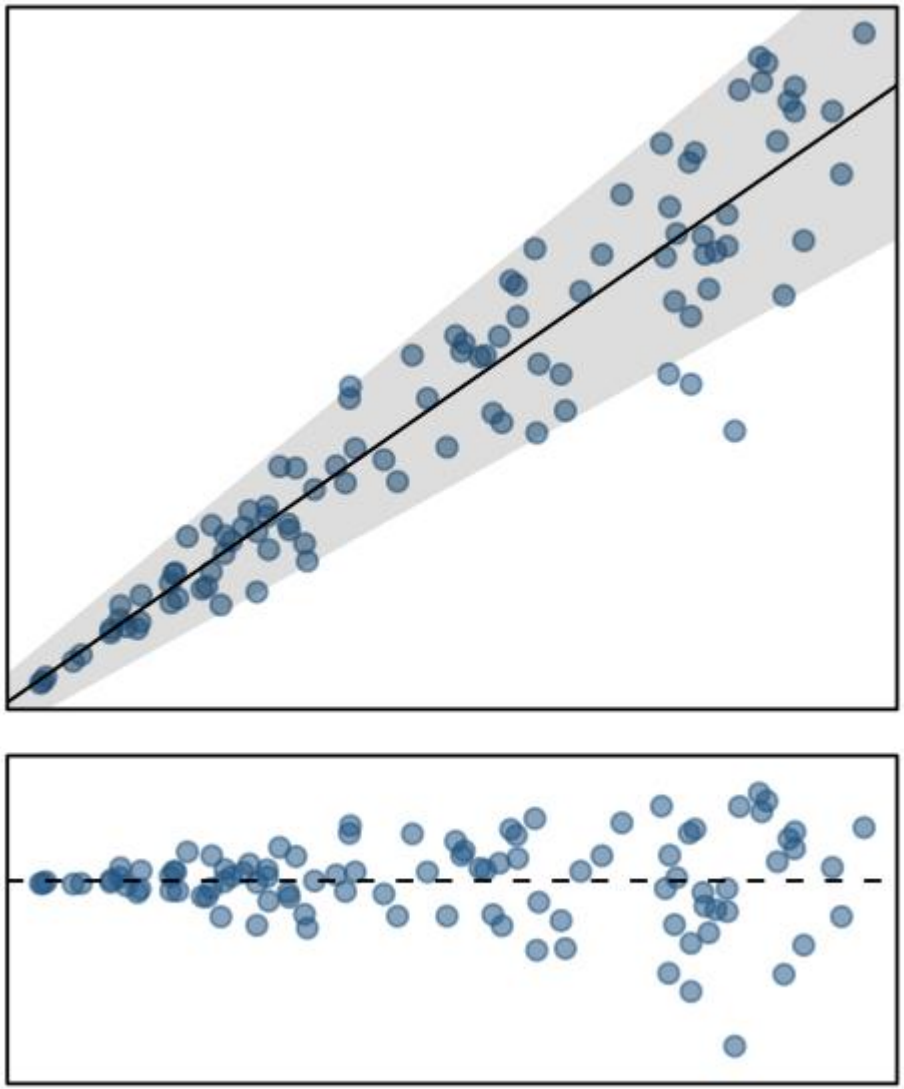
(a) Variabilidade constante
(b) Relação linear
(c) Normalidade dos resíduos
(d) Não haver outliers extremos



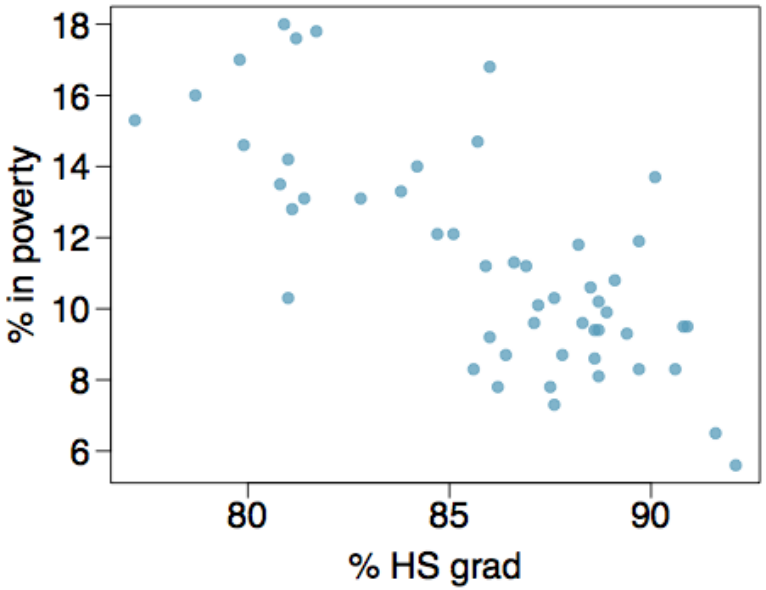
Verificação das condições

Que condição está a ser violada?

(a) Variabilidade constante
(b) Relação linear
(c) Normalidade dos resíduos
(d) Não haver outliers extremos



Dado que...



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

Declive

O declive da regressão pode ser calculado como:

$$b_1 = \frac{s_y}{s_x} R$$

Declive

O declive da regressão pode ser calculado como:

$$b_1 = \frac{s_y}{s_x} R$$

Em contexto…

	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

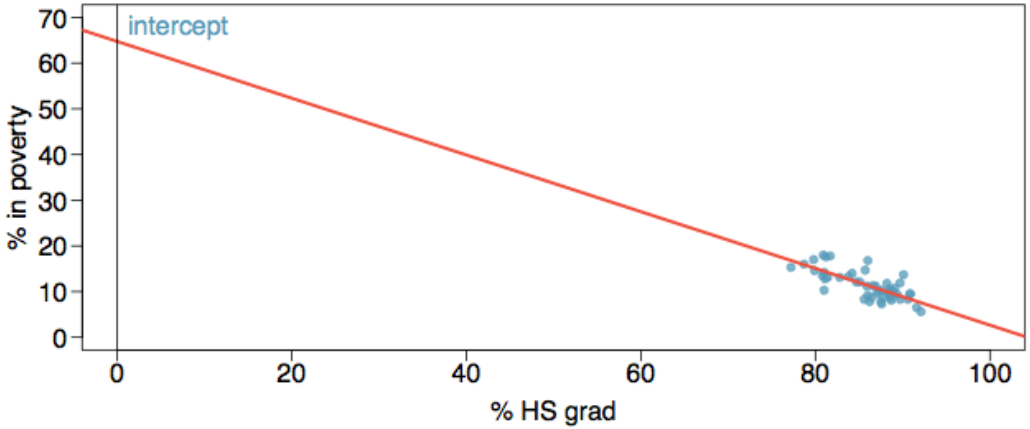
Interpretação

Para cada ponto percentual adicional na taxa de graduados do ensino médio, espera-se que a percentagem de pessoas vivendo na pobreza seja, em média, 0,62 pontos percentuais menor.

Intercepto

O intercepto é o ponto onde a recta de regressão intersecta o eixo y. O cálculo do intercepto usa o fato de que uma recta de regressão sempre passa por ($\bar{x}$, $\bar{y}$).

$$b_0 = \bar{y} - b_1 \bar{x}$$



b_0 = 11.35 - (-0.62) x 86.01
= 64.68

Qual das seguintes opções é a interpretação correta do intercepto?

(a) Para cada aumento de um ponto percentual na taxa de diplomados do ensino secundário, a percentagem de pessoas a viver na pobreza deverá aumentar, em média, 64,68%.

(b) Para cada diminuição de um ponto percentual na taxa de diplomados do ensino secundário, a percentagem de pessoas a viver na pobreza deverá aumentar, em média, 64,68%.

(c) Não haver diplomados do ensino secundário resulta em 64,68% dos residentes a viver abaixo do limiar da pobreza.

(d) Os estados sem diplomados do ensino secundário deverão ter, em média, 64,68% dos residentes a viver abaixo do limiar da pobreza.

(e) Nos estados sem diplomados do ensino secundário, a percentagem de pessoas a viver na pobreza deverá aumentar, em média, 64,68%.

Qual das seguintes opções é a interpretação correta do intercepto?

(a) Para cada aumento de um ponto percentual na taxa de diplomados do ensino secundário, a percentagem de pessoas a viver na pobreza deverá aumentar, em média, 64,68%.

(b) Para cada diminuição de um ponto percentual na taxa de diplomados do ensino secundário, a percentagem de pessoas a viver na pobreza deverá aumentar, em média, 64,68%.

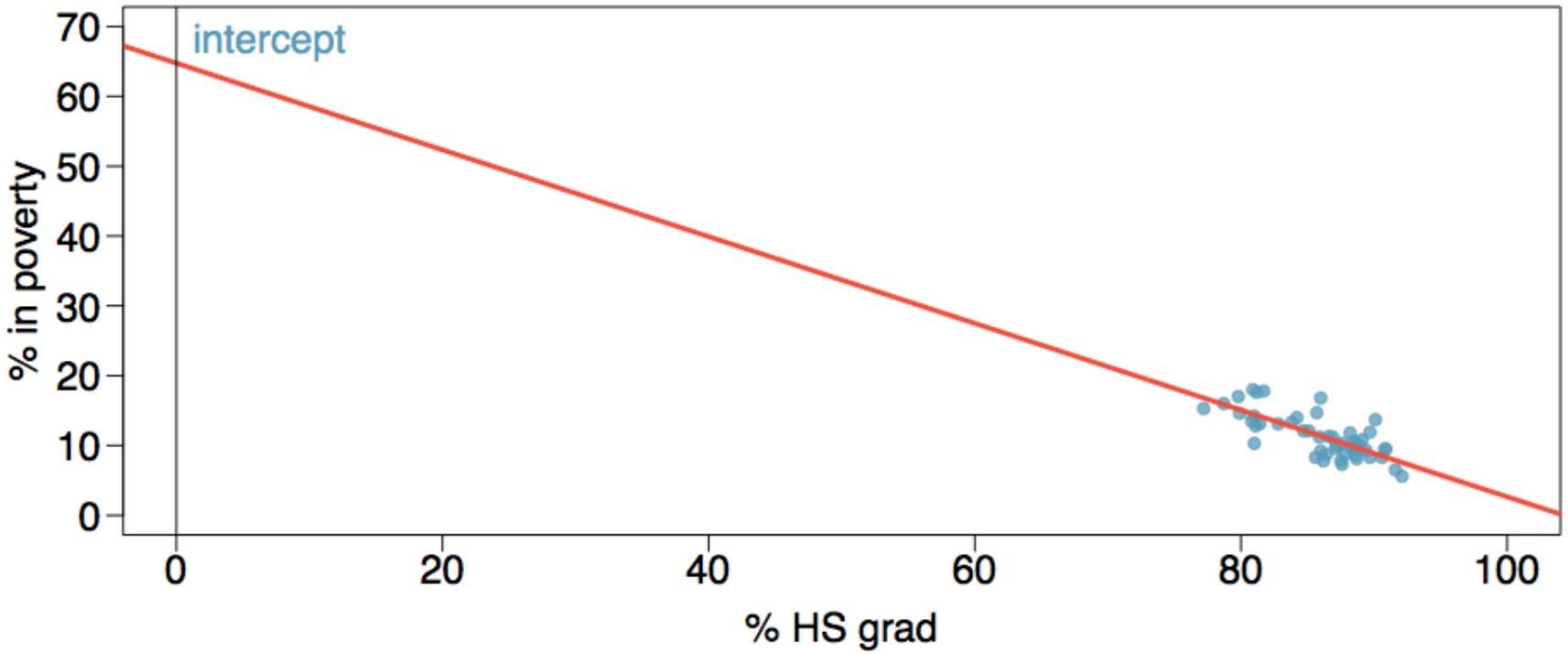
(c) Não haver diplomados do ensino secundário resulta em 64,68% dos residentes a viver abaixo do limiar da pobreza.

(d) Os estados sem diplomados do ensino secundário deverão ter, em média, 64,68% dos residentes a viver abaixo do limiar da pobreza.

(e) Nos estados sem diplomados do ensino secundário, a percentagem de pessoas a viver na pobreza deverá aumentar, em média, 64,68%.

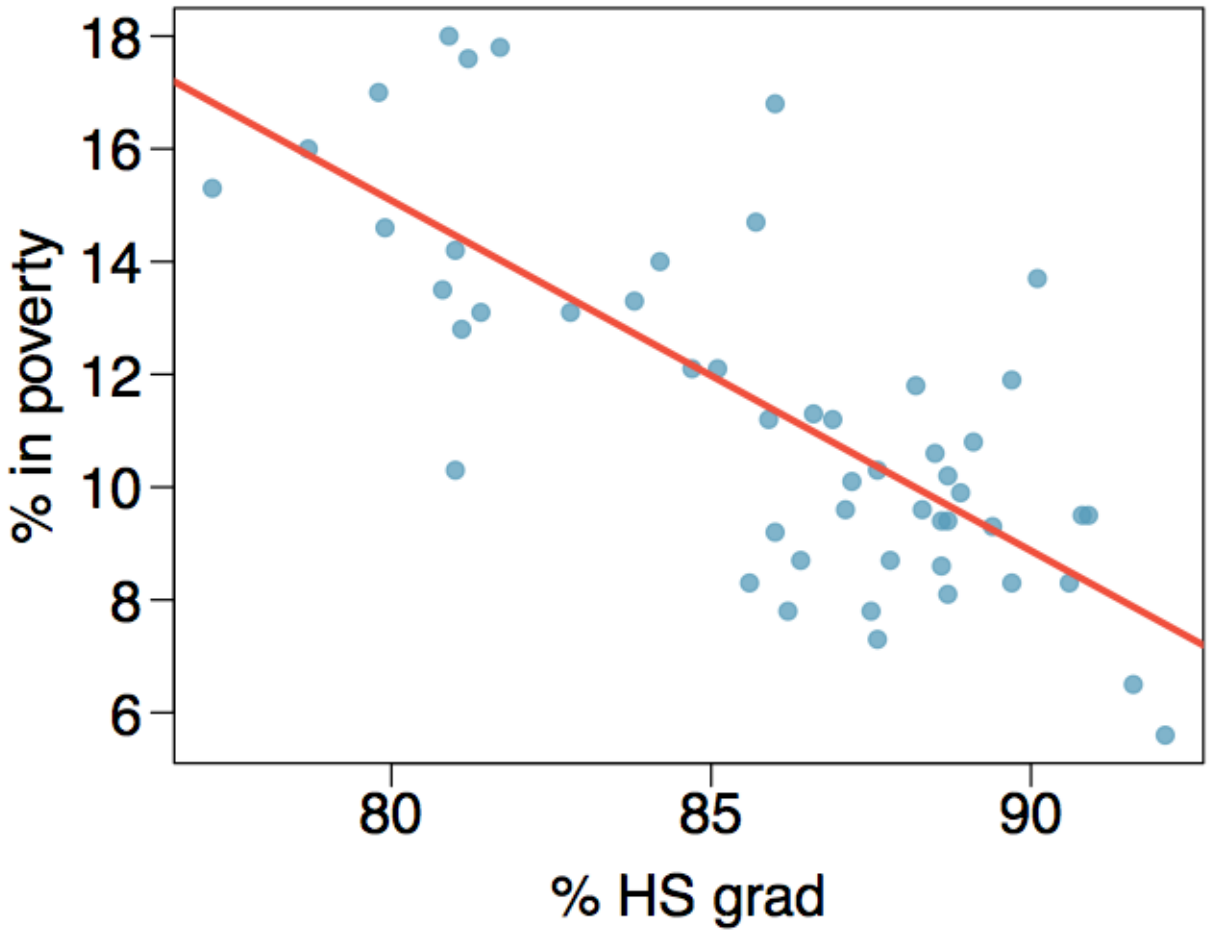
Mais sobre o intercepto

Como não há estados no conjunto de dados sem diplomados do ensino secundário, o intercepto não é relevante, não é muito útil e também não é fiável, uma vez que o valor previsto do intercepto está muito distante da maioria dos dados.



Linha de regressão

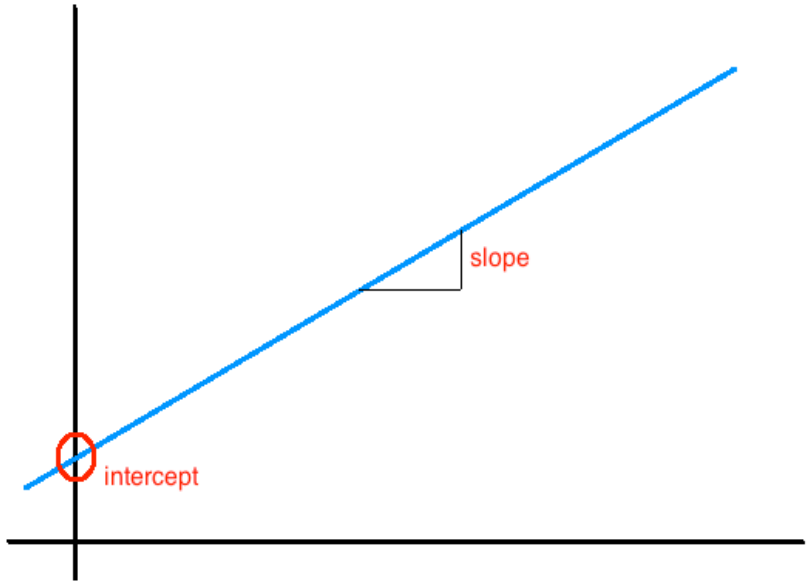
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$



Interpretação da inclinação e do intercepto

- *Intercepto*: Quando *x* = 0, espera-se que *y* seja igual ao intercepto.

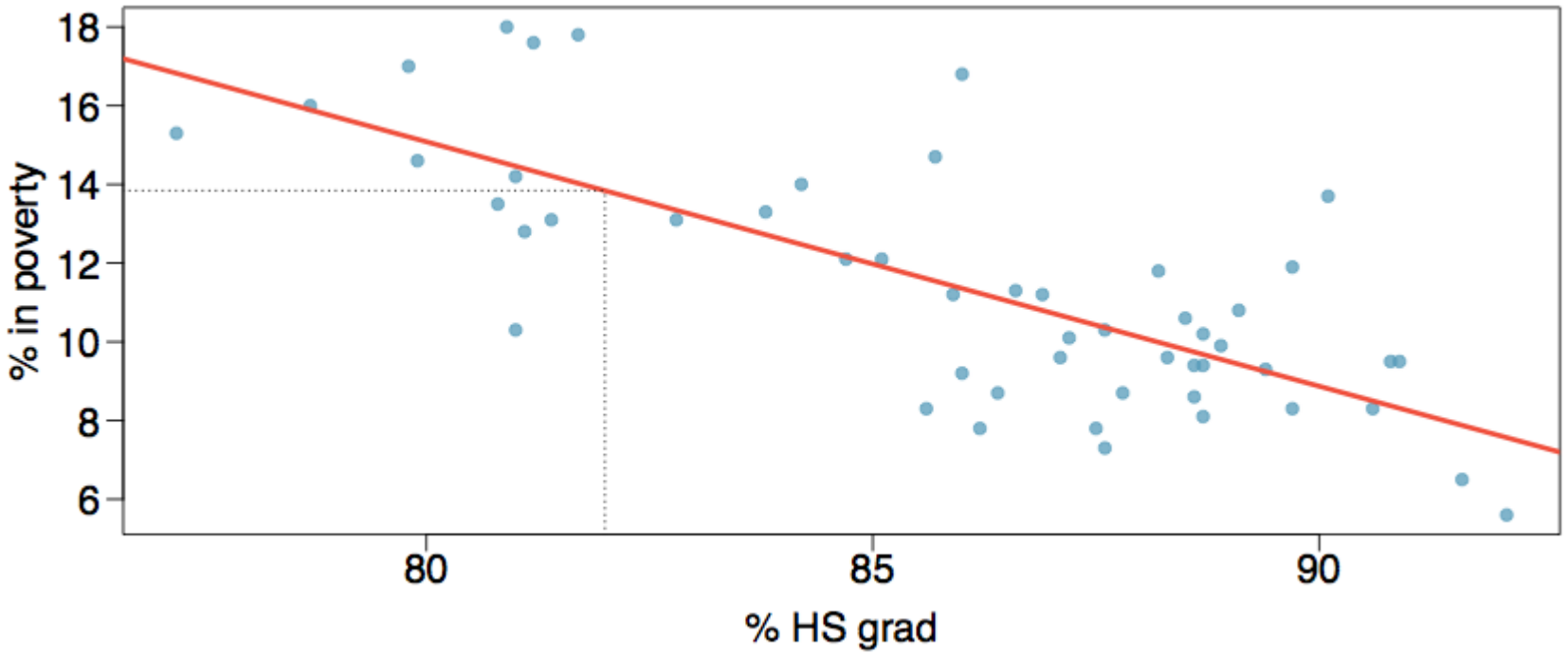
- *Declive*: Para cada unidade em x, espera-se que y aumente ou diminua, em média, pelo valor do declive.



Nota: Essas afirmações não são causais, a menos que o estudo seja um experimento controlado randomizado.

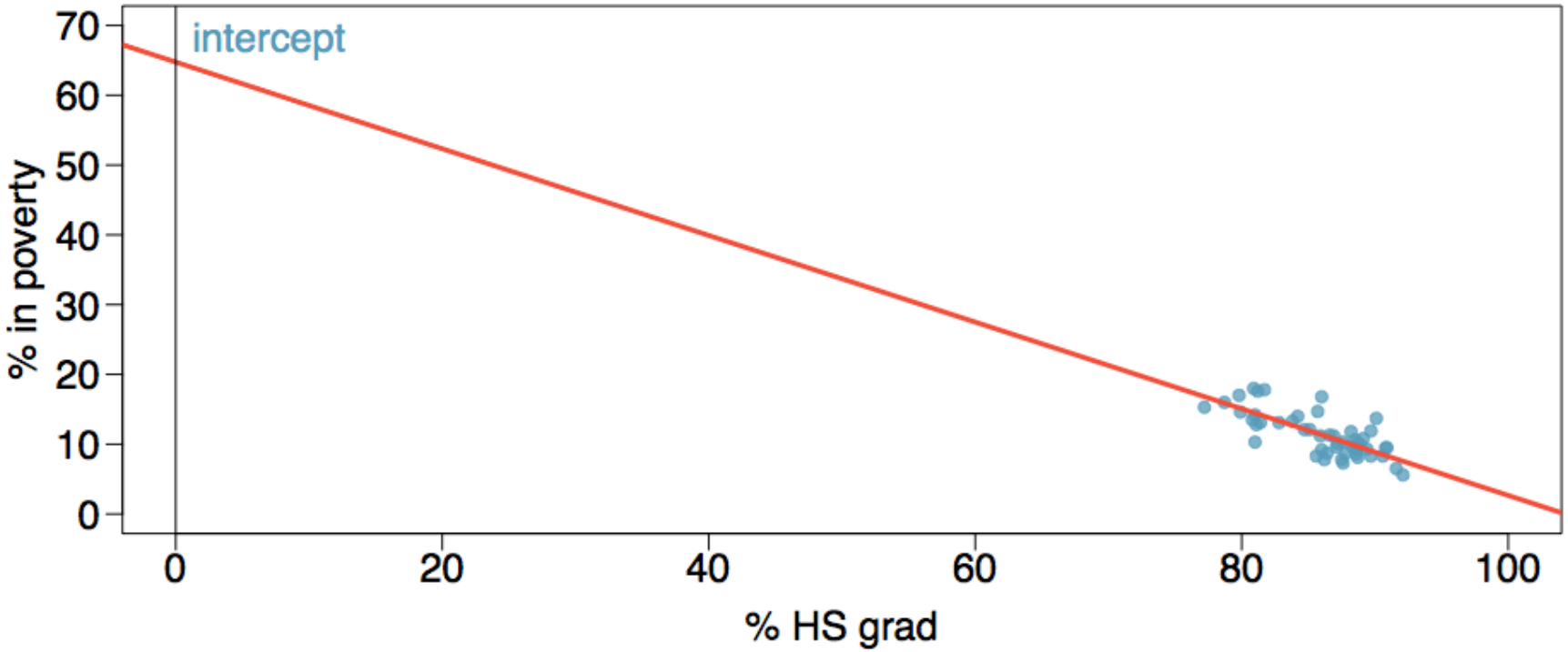
Previsão

- Usar o modelo linear para prever o valor da variável de resposta para um determinado valor da variável explicativa é chamado de previsão, simplesmente substituindo o valor de x na equação do modelo linear.
- Haverá alguma incerteza associada ao valor previsto.

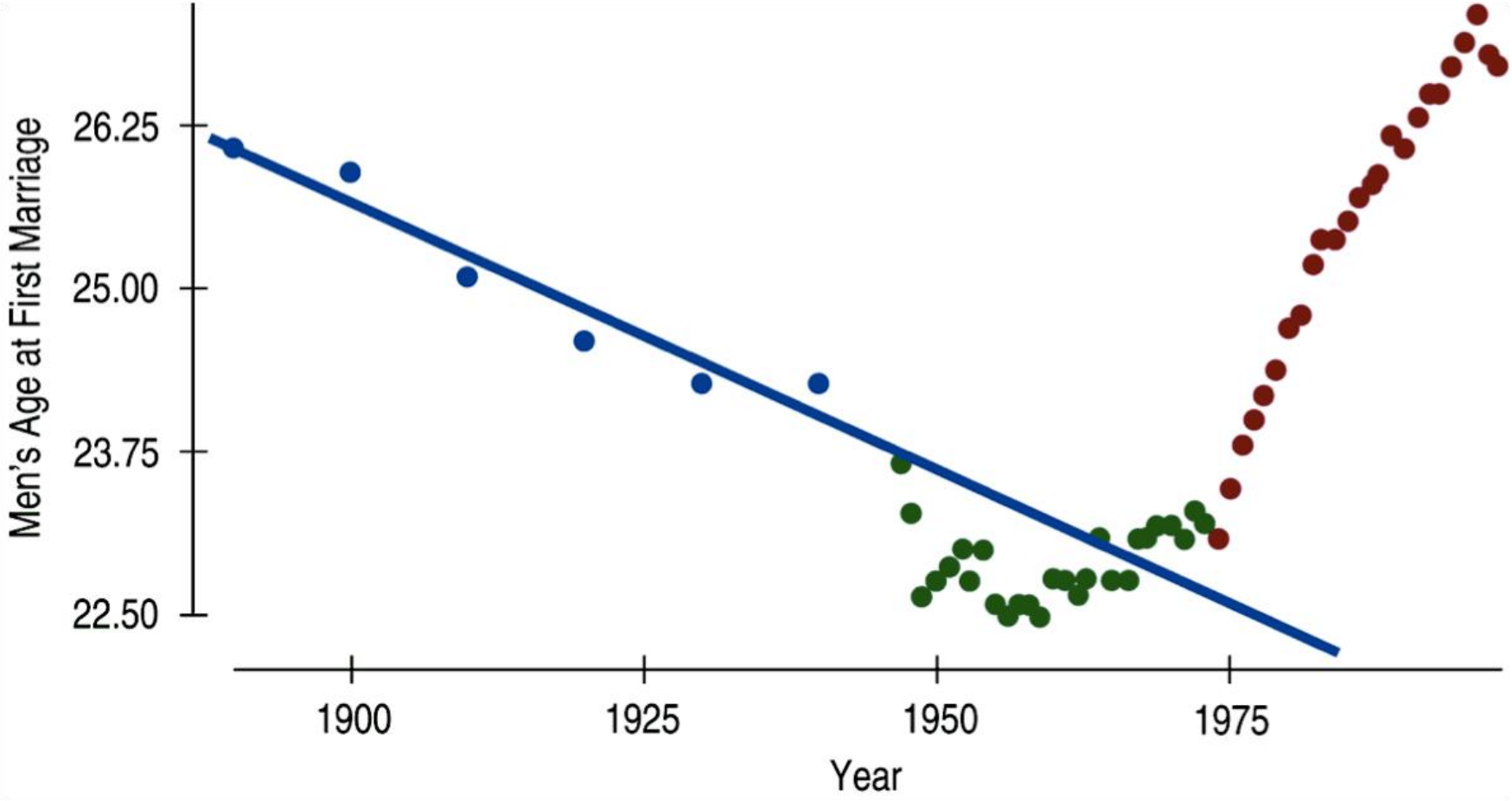


Extrapolação

- Aplicar uma estimativa do modelo a valores fora do domínio dos dados originais é chamado de extrapolação.
- Às vezes, o intercepto pode ser uma extrapolação.



Exemplos de extrapolação



Exemplos de extrapolação

BBC NEWS

Watch One-Minute World News

News Front Page

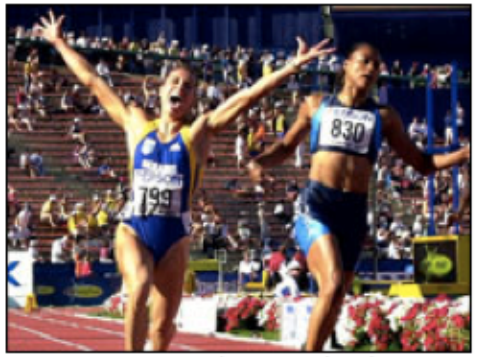
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend Printable version

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.



Women are set to become the dominant sprinters

An Oxford University study found that women are running faster than they have ever done over 100m.

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times are also going to improve."

Exemplos de extrapolação

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

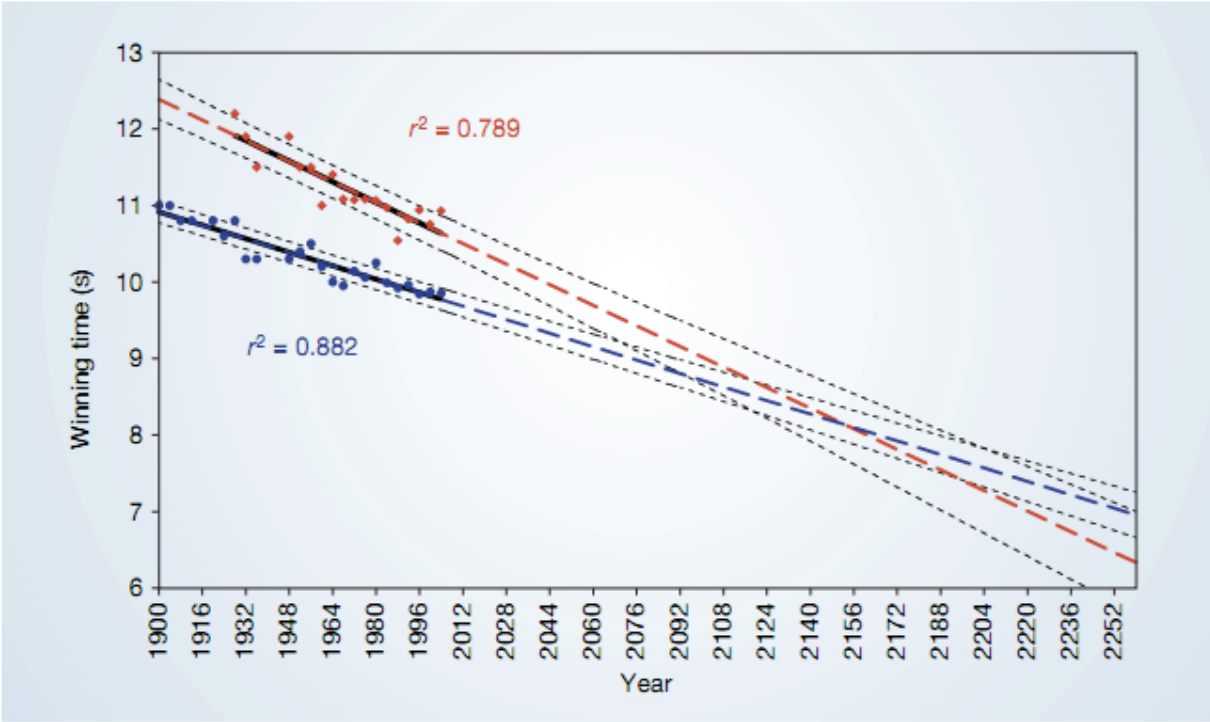


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

R^2

- A força do ajuste de um modelo linear é mais comumente avaliada usando R^2.
- R^2 é calculado como o quadrado do coeficiente de correlação.
- Ele diz-nos a percentagem da variabilidade na variável de resposta explicada pelo modelo.
- O restante da variabilidade é explicado por variáveis não incluídas no modelo ou por aleatoriedade inerente nos dados.
- Para o modelo com que estamos a trabalhar, $R^2 = (-0,62)^2 = 0,38$.

Interpretação de R^2

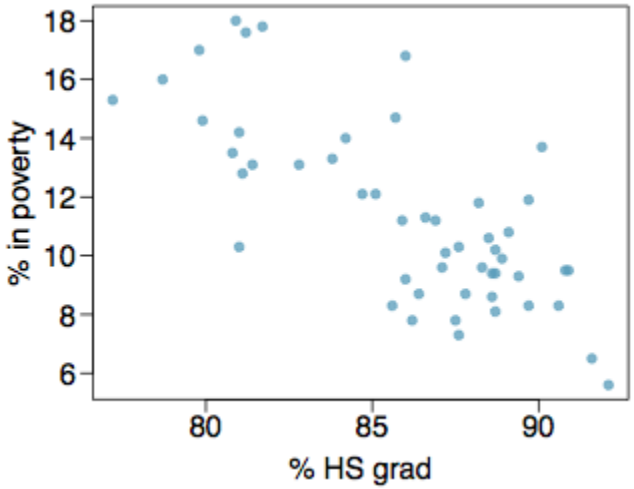
Qual das opções abaixo é a interpretação correta de R=−0.62, R^2=0.38?

a) 38% da variabilidade na percentagem de diplomados do ensino secundário entre os 51 estados é explicada pelo modelo.

b) 38% da variabilidade na percentagem de residentes a viver na pobreza entre os 51 estados é explicada pelo modelo.

c) 38% das vezes, a percentagem de diplomados do ensino secundário prevê corretamente a percentagem de pessoas a viver na pobreza.

d) 62% da variabilidade na percentagem de residentes a viver na pobreza entre os 51 estados é explicada pelo modelo.



Interpretação de R^2

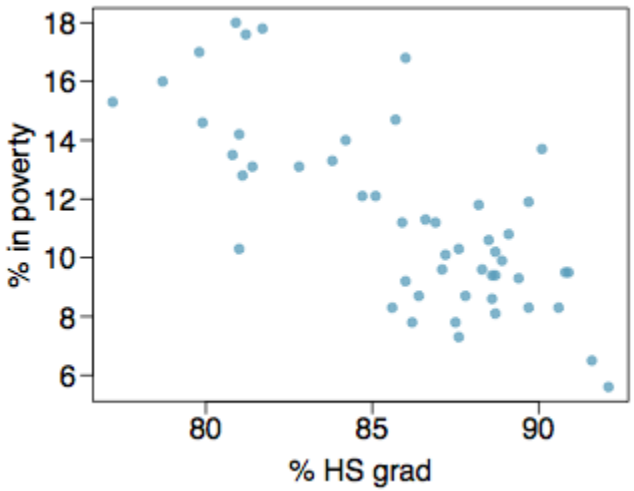
Qual das opções abaixo é a interpretação correta de R=−0.62, R^2=0.38?

a) 38% da variabilidade na percentagem de diplomados do ensino secundário entre os 51 estados é explicada pelo modelo.

b) 38% da variabilidade na percentagem de residentes a viver na pobreza entre os 51 estados é explicada pelo modelo.

c) 38% das vezes, a percentagem de diplomados do ensino secundário prevê corretamente a percentagem de pessoas a viver na pobreza.

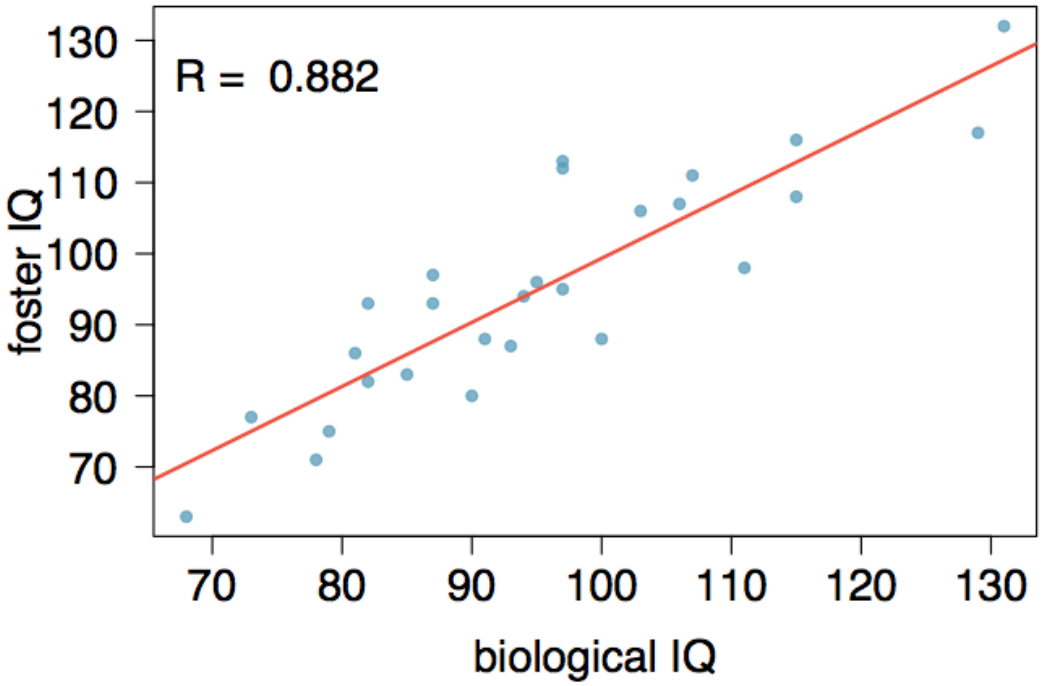
d) 62% da variabilidade na percentagem de residentes a viver na pobreza entre os 51 estados é explicada pelo modelo.



Regressão Linear para Inferência Estatística

Nature or nurture?

Em 1966, Cyril Burt publicou um artigo intitulado"The genetic determination of differences in intelligence: A study of monozygotic twins reared apart?" Os dados consistem em pontuações de QI para [uma amostra aleatória assumida de] 27 gémeos idênticos, sendo que um foi criado por pais adotivos e o outro pelos pais biológicos.



Prática

Qual dos seguintes é falso?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779,     Adjusted R-squared:
0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

(a) Um aumento de 10 pontos no QI do gémeo biológico está associado, em média, a um aumento de 9 pontos no QI do gémeo criado por pais adotivos.

(b) Aproximadamente 78% dos QIs dos gémeos criados por pais adotivos podem ser previstos com precisão pelo modelo.

(c) O modelo linear é $\widehat{fosterIQ} = 9.2 + 0.9 \times bioIQ$

(d) Gémeos criados por pais adotivos com QIs acima da média tendem a ter gémeos biológicos também com QIs acima da média.

Prática

Qual dos seguintes é falso?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

Residual standard error: 7.729 on 25 degrees of freedom
Multiple R-squared: 0.7779,     Adjusted R-squared:
0.769
F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

(a) Um aumento de 10 pontos no QI do gémeo biológico está associado, em média, a um aumento de 9 pontos no QI do gémeo criado por pais adotivos.

(b) Aproximadamente 78% dos QIs dos gémeos criados por pais adotivos podem ser previstos com precisão pelo modelo.

(c) O modelo linear é $\widehat{fosterIQ} = 9.2 + 0.9 \times bioIQ$

(d) Gémeos criados por pais adotivos com QIs acima da média tendem a ter gémeos biológicos também com QIs acima da média.

Testando o declive

Assumindo que estes 27 gémeos constituem uma amostra representativa de todos os gémeos separados à nascença, gostaríamos de testar se estes dados fornecem evidências convincentes de que o QI do gémeo biológico é um preditor significativo do QI do gémeo criado por pais adotivos.

(a) H_0: $b_0 = 0$; H_A: $b_0 \neq 0$
(b) H_0: $\beta_0 = 0$; H_A: $\beta_0 \neq 0$
(c) H_0: $b_1 = 0$; H_A: $b_1 \neq 0$
(d) H_0: $\beta_1 = 0$; H_A: $\beta_1 \neq 0$

Testando o declive

Assumindo que estes 27 gémeos constituem uma amostra representativa de todos os gémeos separados à nascença, gostaríamos de testar se estes dados fornecem evidências convincentes de que o QI do gémeo biológico é um preditor significativo do QI do gémeo criado por pais adotivos.

(a) H_0: $b_0 = 0$; H_A: $b_0 \neq 0$
(b) H_0: $\beta_0 = 0$; H_A: $\beta_0 \neq 0$
(c) H_0: $b_1 = 0$; H_A: $b_1 \neq 0$
(d) H_0: $\beta_1 = 0$; H_A: $\beta_1 \neq 0$

Checking model conditions using graphs

Modeling conditions

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

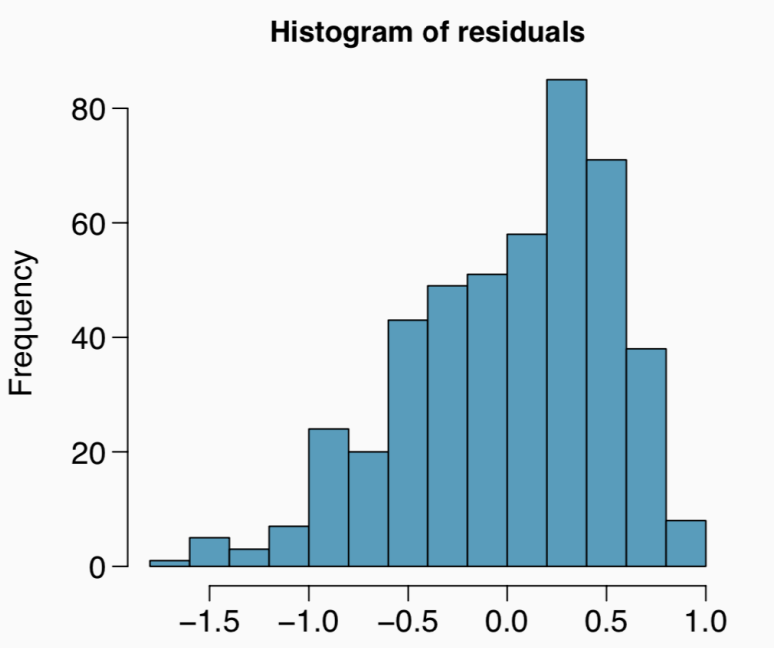
The model depends on the following conditions

1. residuals are nearly normal (less important for larger data sets)
2. residuals have constant variability
3. residuals are independent
4. each variable is linearly related to the outcome

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

(1) nearly normal residuals

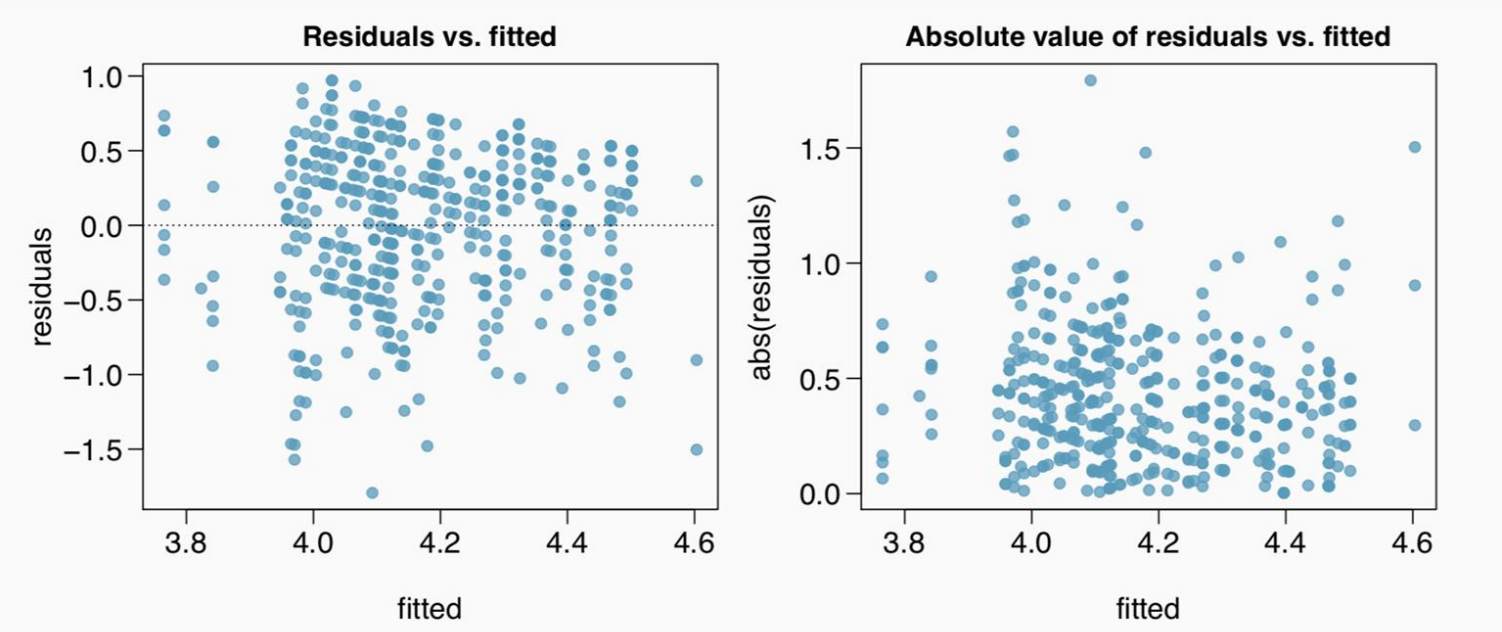
Histogram of the residuals.



Does this condition appear to be satisfied?

(2) constant variability in residuals

Scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted).



Does this condition appear to be satisfied?

Checking constant variance - recap

When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.

With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

Checking constant variance - recap

When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.

With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

(3) independent residuals

Scatterplot of residuals vs. order of data collection.

Does this condition appear to be satisfied?

More on the condition of independent residuals

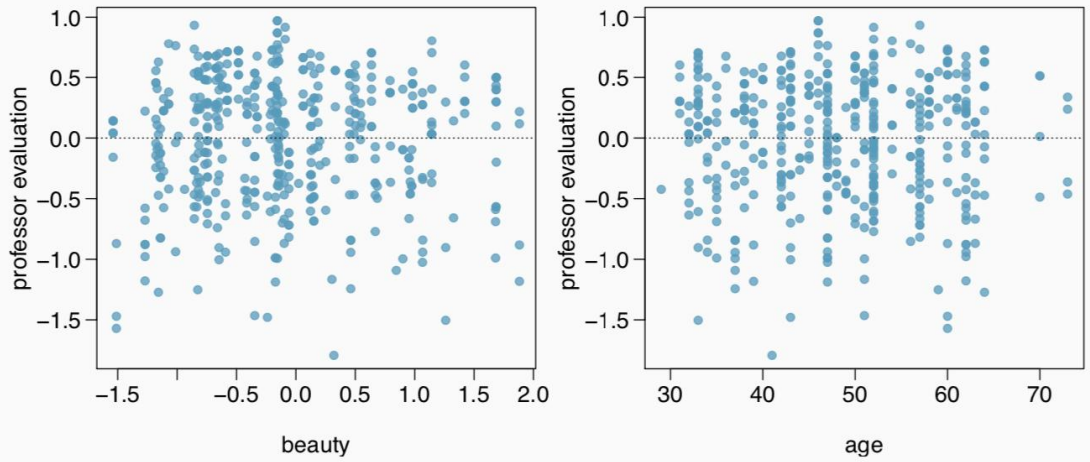
Checking for independent residuals allows us to indirectly check for independent observations.

If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.

This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

(4) linear relationships

Scatterplot of residuals vs. each (numerical) explanatory variable.



Does this condition appear to be satisfied?

Note: We use residuals instead of the predictors on the y-axis so that we can still check for linearity without worrying about other possible violations like collinearity between the predictors.

Several options for improving a model

Transforming variables

Seeking out additional variables to fill model gaps

Using more advanced methods that would account for challenges around inconsistent variability or nonlinear relationships between predictors and the outcome

Transformations

If the concern with the model is non-linear relationships between the explanatory variable(s) and the response variable, transforming the response variable can be helpful.

- Log transformation (log y)
- Square root transformation (sqrt(y))
- Inverse transformation (1/y)
- Truncation (cap the max value possible)

It is also possible to apply transformations to the explanatory variable(s), however such transformations tend to make the model coefficients even harder to interpret.

Models can be wrong, but useful

All models are wrong, but some are useful.

- George Box

No model is perfect, but even imperfect models can be useful, as long as we are clear and report the model's shortcomings.

If conditions are grossly violated, we should not report the model results, but instead consider a new model, even if it means learning more statistical methods or hiring someone who can help.

Logistic regression

Odds

Odds are another way of quantifying the probability of an event, commonly used in gambling (and logistic regression).

For some event E,

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

Similarly, if we are told the odds of E are x to y then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which implies

$$P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

Example - Donner Party

In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested rote to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

Example - Donner Party - Data

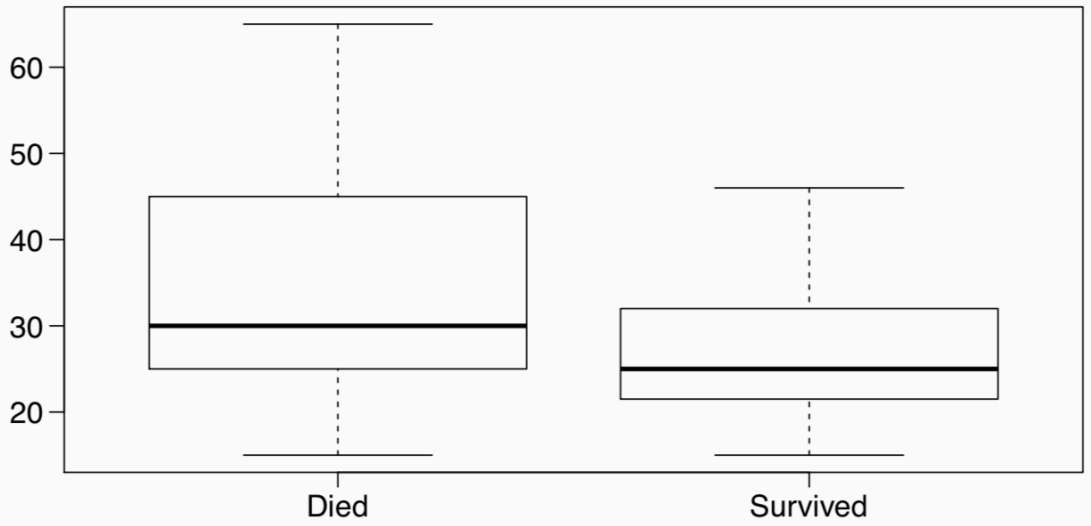
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs Gender

	Male	Female
Died	20	5
Survived	10	10

Status vs Age



Example - Donner Party

It seems clear that both age and gender have an effect on someone's survival, how do we come up with a model that will let us explore this relationship?

Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

One way to think about the problem - we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model: $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$.
3. A link function that relates the linear model to the parameter of the outcome distribution: $g(p) = \eta$ or $p = g^{-1}(\eta)$.

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

Logistic Regression

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p. There are a variety of options but the most commonly used is the logit function.

Logit function

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, 68 more on this later.

The logistic regression model

The three GLM criteria give us:

$$y_i \sim \text{Binom}(p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

$$\text{logit}(p) = \eta$$

From which we arrive at,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_n x_{n,i})}$$

Example - Donner Party - Model

In **R** we fit a GLM in the same was as a linear model except using **glm** instead of **lm** and we must also specify the type of GLM to fit using the **family** argument.

```
summary(glm(Status ~ Age, data=donner, family=binomial))
## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.81852    0.99937   1.820   0.0688 .
## Age         -0.06647    0.03222  -2.063   0.0391 *
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 56.291  on 43  degrees of freedom
## AIC: 60.291
```

Example - Donner Party - Prediction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a newborn (Age=0):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

Example - Donner Party - Prediction

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17$$

$$p = 1.17/2.17 = 0.539$$

Odds / Probability of survival for a 50 year old:

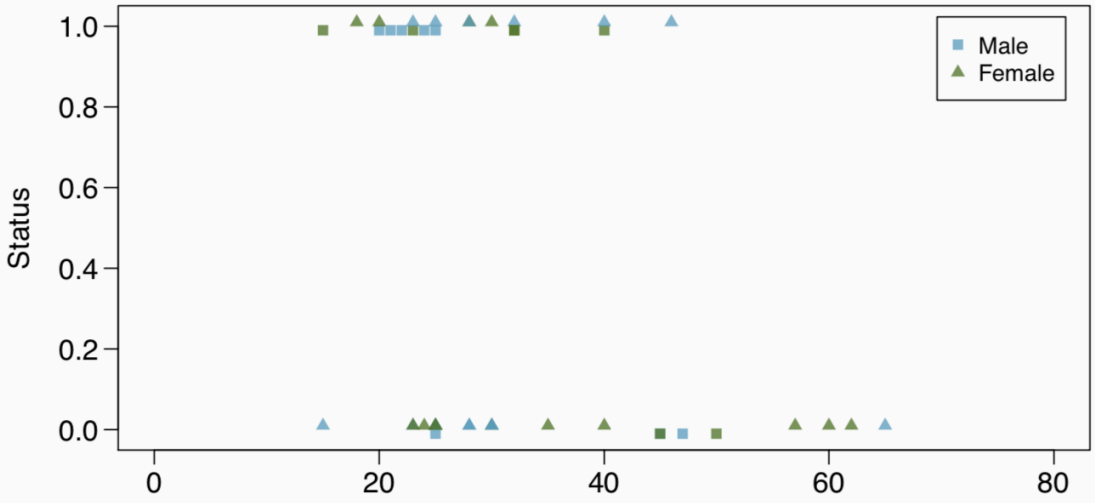
$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50$$

$$\frac{p}{1-p} = \exp(-1.5065) = 0.222$$

$$p = 0.222/1.222 = 0.181$$

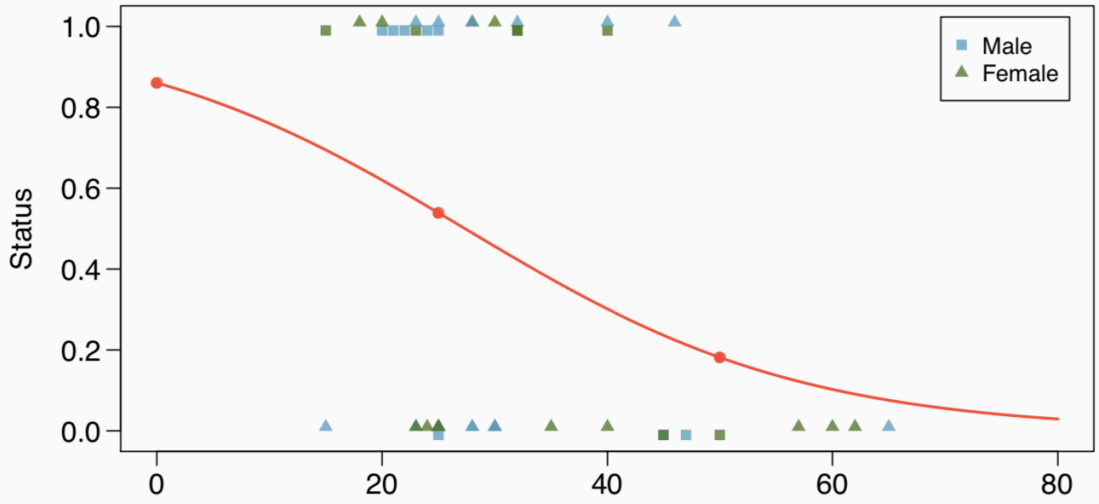
Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Interpretation

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.8185 - 0.0665(x+1)$$

$$= 1.8185 - 0.0665x - 0.0665$$

$$\log\left(\frac{p_2}{1-p_2}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = -0.0665$$

$$\log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}\right) = -0.0665$$

$$\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2} = \exp(-0.0665) = 0.94$$

Example - Donner Party - Age and Gender

```
summary(glm(Status  ~Age + Sex, data=donner, family=binomial))
## Call:
## glm(formula = Status  ~Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

Male model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 0$$

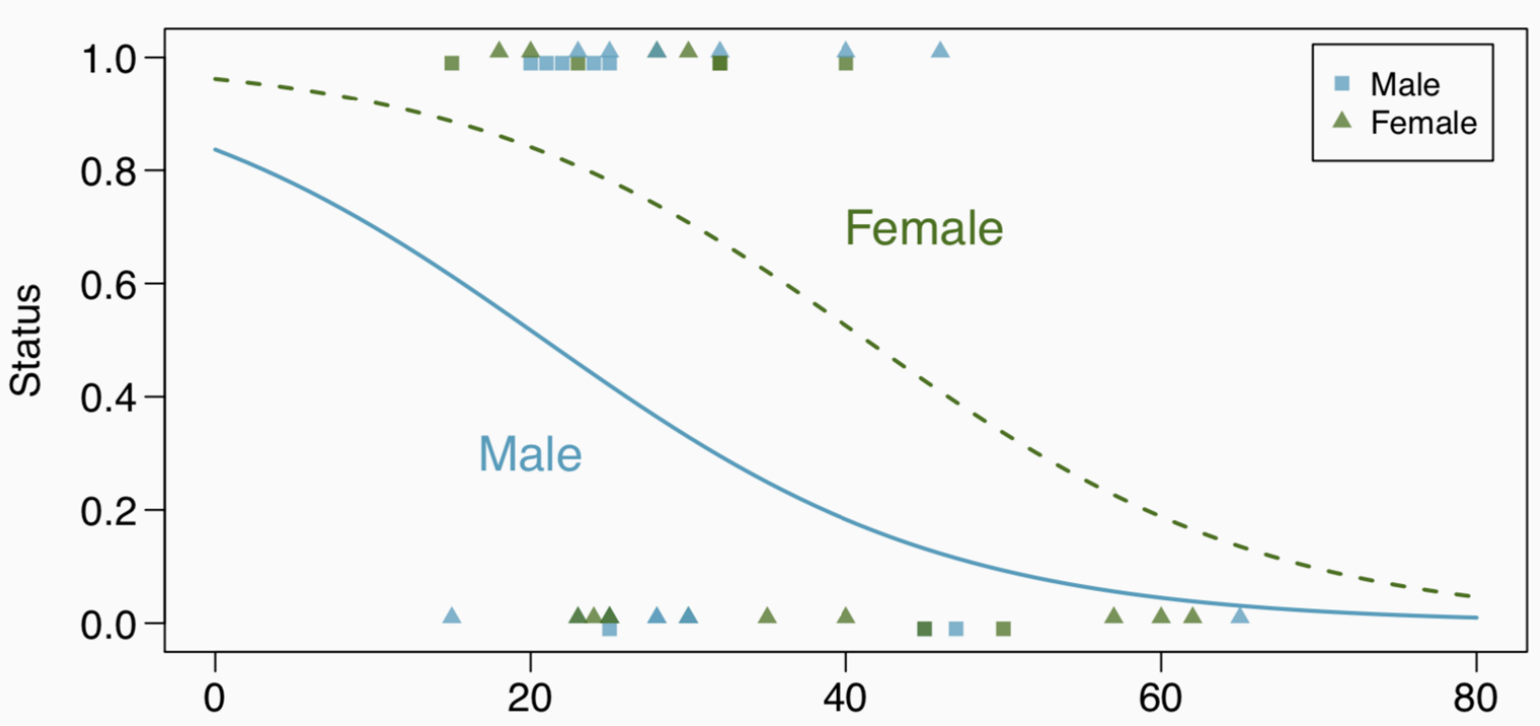
$$= 1.63312 + -0.07820 \times \text{Age}$$

Female model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 1$$

$$= 3.23041 + -0.07820 \times \text{Age}$$

Example - Donner Party - Gender Models (cont.)



Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Hypothesis test for the whole model

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Note: The model output does not include any F-statistic, as a general rule there are not single model hypothesis tests for GLM models.

Hypothesis tests for a coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

We are however still able to perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z-test.

Note: The only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$
$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta_{age}} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\text{p-value} = P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10)$$
$$= 2 \times 0.0178 = 0.0359$$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

Log odds ratio:

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.1513, -0.0051)$$

Odds ratio:

$$\exp(CI) = (\exp -0.1513, \exp -0.0051) = (0.85960.9949)$$

Example - Birdkeeping and Lung Cancer

A 1972 - 1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.

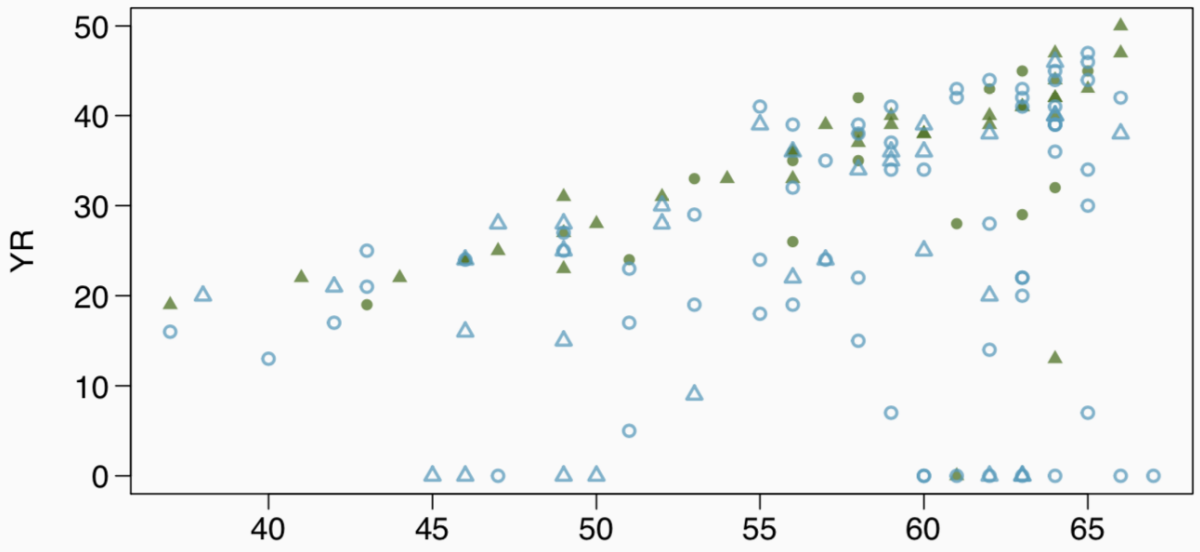
From *Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)*

Example - Birdkeeping and Lung Cancer - Data

	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37.00	19.00	12.00
2	LungCancer	Male	Low	Bird	41.00	22.00	15.00
3	LungCancer	Male	High	NoBird	43.00	19.00	15.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
147	NoCancer	Female	Low	NoBird	65.00	7.00	2.00

- `LC` Whether subject has lung cancer
- `FM` Sex of subject
- `SS` Socioeconomic status
- `BK` Indicator for birdkeeping
- `AG` Age of subject (years)
- `YR` Years of smoking prior to diagnosis or examination
- `CD` Average rate of smoking (cigarettes per day)

Example - Birdkeeping and Lung Cancer - EDA



	Bird	No Bird
Lung Cancer	▲	●
No Lung Cancer	△	○

Example - Birdkeeping and Lung Cancer - Model

```
summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))
## Call:
## glm(formula = LC ~ FM + SS + BK + AG + YR + CD, family = binomial,
##     data = bird)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736  1.80425    -1.074 0.282924
## FMFemale     0.56127  0.53116     1.057 0.290653
## SSHigh       0.10545  0.46885     0.225 0.822050
## BKBird       1.36259  0.41128     3.313 0.000923 ***
## AG          -0.03976  0.03548    -1.120 0.262503
## YR           0.07287  0.02649     2.751 0.005940 **
## CD           0.02602  0.02552     1.019 0.308055
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
```

Example - Birdkeeping and Lung Cancer - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is exp(1.3626) = 3.91.
- The odds ratio of getting lung cancer for an additional year of smoking is exp(0.0729) = 1.08.

What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

Bird keepers are not 4x more likely to develop lung cancer than non-bird keepers.

This is the difference between relative risk and an odds ratio.

$$RR = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

$$OR = \frac{P(\text{disease}|\text{exposed})/[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1 - P(\text{disease}|\text{unexposed})]}$$

Back to the birds

What is probability of lung cancer in a bird keeper if we knew that P(lung cancer|no birds) = 0.05?

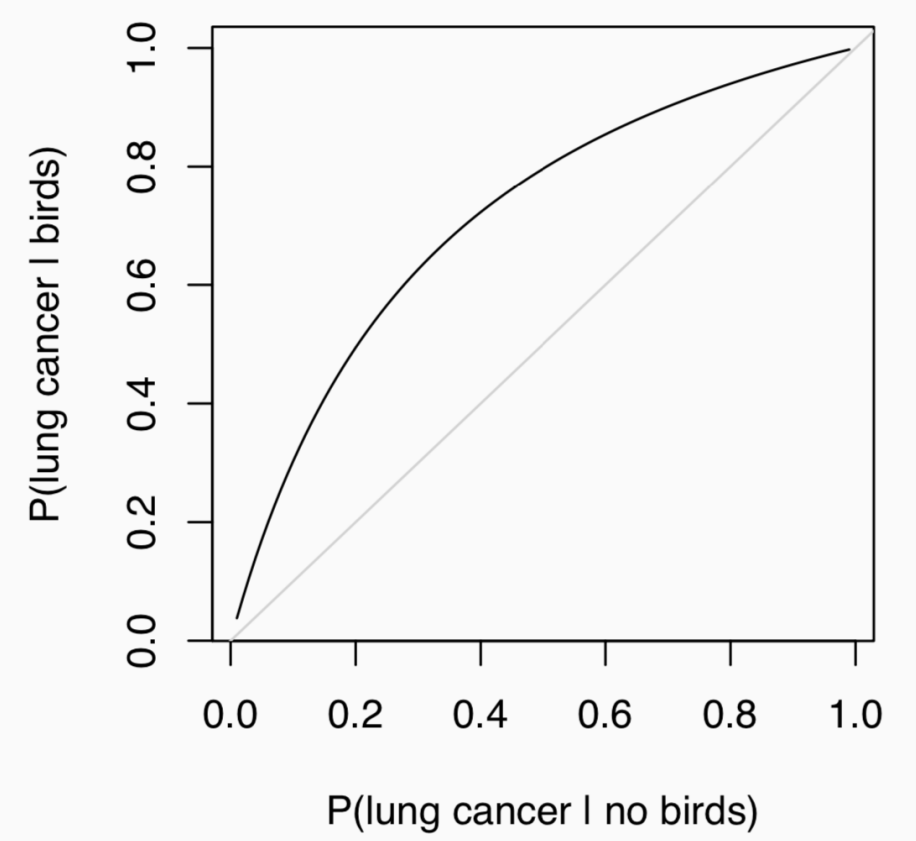
$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 3.91$$

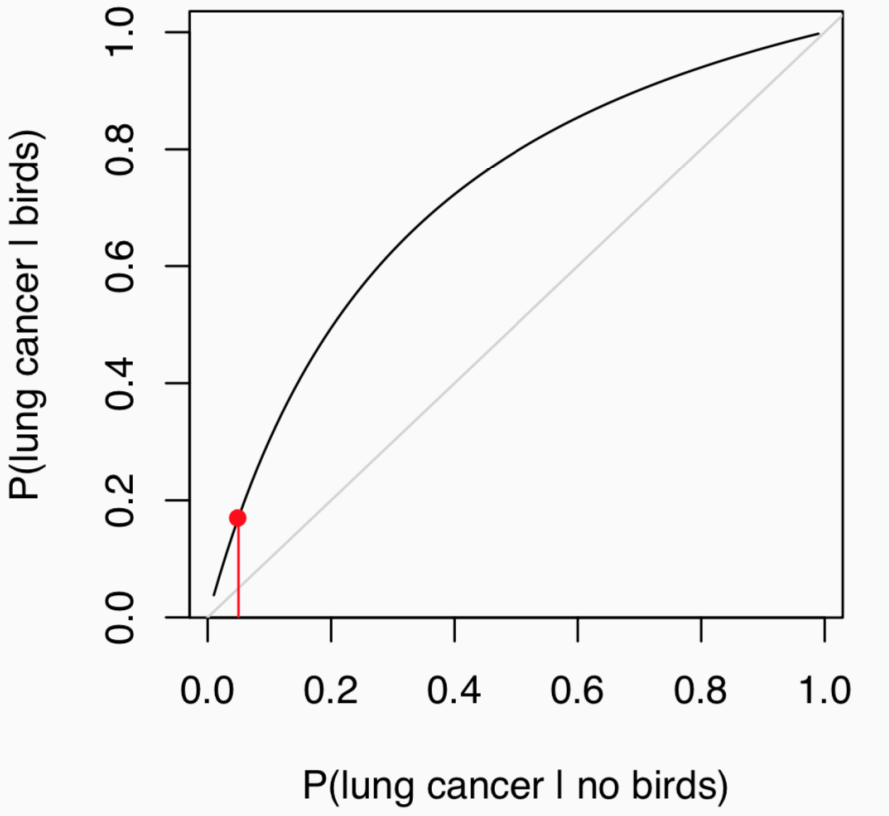
$$P(\text{lung cancer}|\text{birds}) = \frac{3.91 \times \frac{0.05}{0.95}}{1 + 3.91 \times \frac{0.05}{0.95}} = 0.171$$

$$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.171/0.05 = 3.41$$

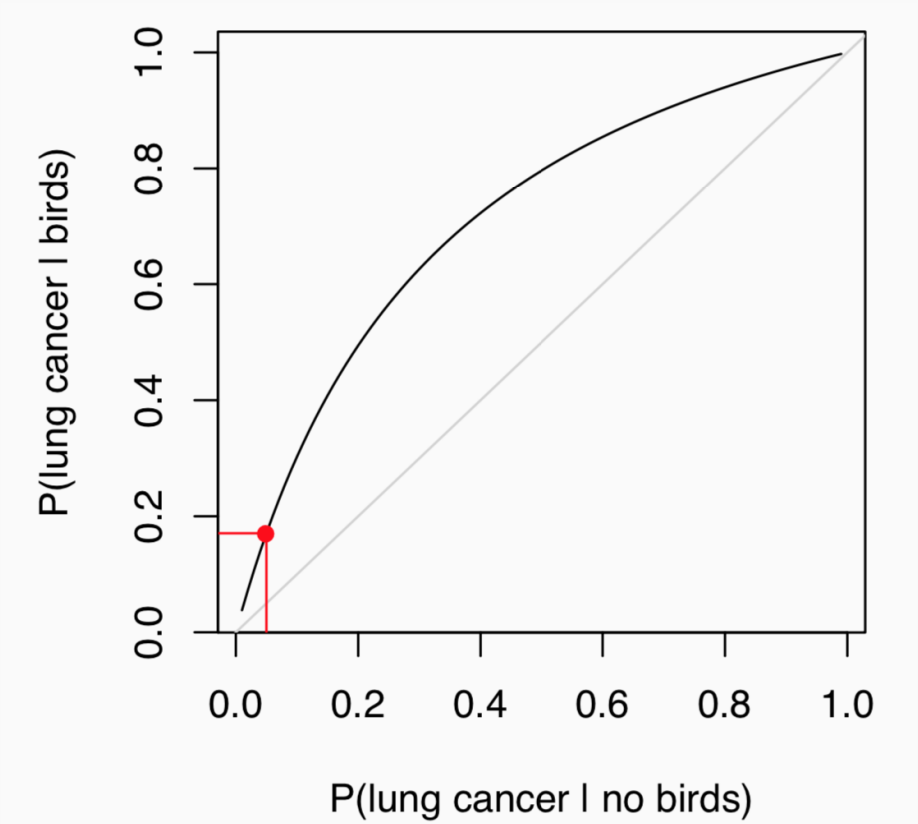
Bird OR Curve



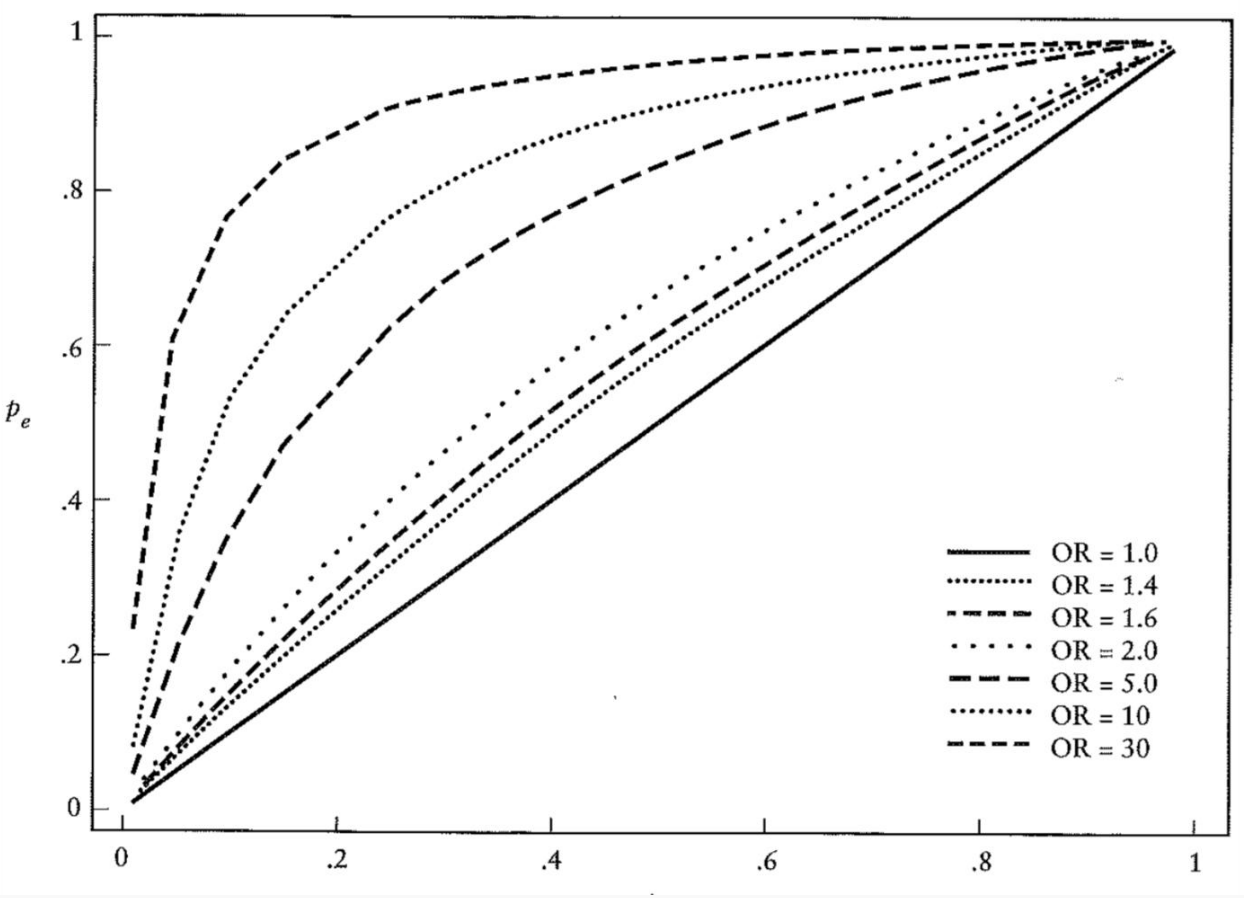
Bird OR Curve



Bird OR Curve



OR Curves



(An old) Example - House

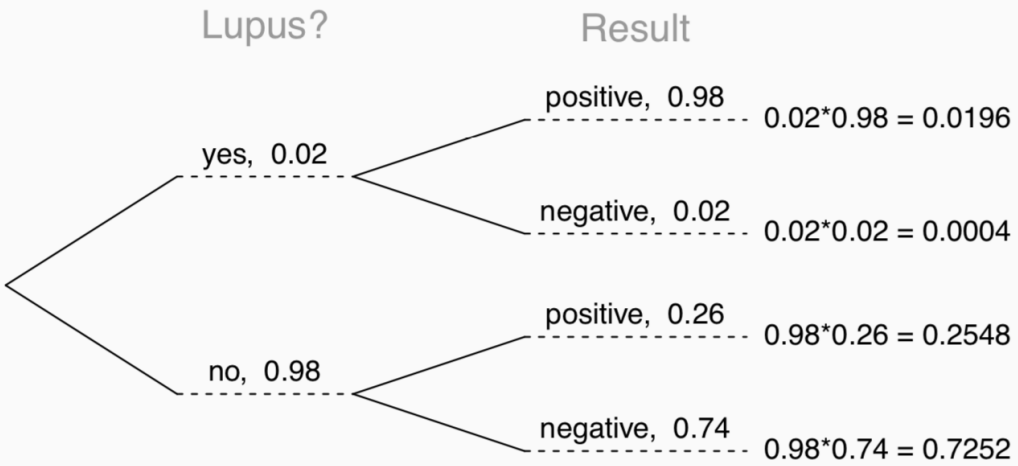
If you've ever watched the TV show House on Fox, you know that Dr. House regularly states, "It's never lupus."

Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease.

The test for lupus is very accurate if the person actually has lupus, however is very inaccurate if the person does not. More specifically, the test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease.

Is Dr. House correct even if someone tests positive for Lupus?

(An old) Example - House



$$P(\text{Lupus}|+) = \frac{P(+, \text{Lupus})}{P(+, \text{Lupus}) + P(+, \text{No Lupus})}$$
$$= \frac{0.0196}{0.0196 + 0.2548} = 0.0714$$

Testing for lupus

It turns out that testing for Lupus is actually quite complicated, a diagnosis usually relies on the outcome of multiple tests, often including: a complete blood count, an erythrocyte sedimentation rate, a kidney and liver assessment, a urinalysis, and or an antinuclear antibody (ANA) test.

It is important to think about what is involved in each of these tests (e.g. deciding if complete blood count is high or low) and how each of the individual tests and related decisions plays a role in the overall decision of diagnosing a patient with lupus.

Testing for lupus

At some level we can view a diagnosis as a binary decision (lupus or no lupus) that involves the complex integration of various explanatory variables.

The example does not give us any information about how a diagnosis is made, but what it does give us is just as important - the sensitivity and the specificity of the test. These values are critical for our understanding of what a positive or negative test result actually means.

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

P(Test + | Conditon +) = P(+|lupus) = 0.98

Specificity - measures a tests ability to identify negative results.

P(Test − | Condition −) = P(−|no lupus) = 0.74

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

$$P(\text{Test} + \mid \text{Conditon} +) = P(+|\text{lupus}) = 0.98$$

Specificity - measures a tests ability to identify negative results.

$$P(\text{Test} - \mid \text{Condition} -) = P(-|\text{no lupus}) = 0.74$$

It is illustrative to think about the extreme cases - what is the sensitivity and specificity of a test that always returns a positive result? What about a test that always returns a negative result?

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$$

$$\textit{False positive rate } (\alpha) = P(\text{Test } + \mid \text{Condition } -) = FP/(FP + TN)$$

$$\textit{Sensitivity} = 1 - \textit{False negative rate} = \text{Power}$$

$$\textit{Specificity} = 1 - \textit{False positive rate}$$

So what?

Clearly it is important to know the Sensitivity and Specificity of test (and or the false positive and false negative rates). Along with the incidence of the disease (e.g. P(lupus)) these values are necessary to calculate important quantities like P(lupus|+).

Additionally, our brief foray into power analysis before the first midterm should also give you an idea about the trade offs that are inherent in minimizing false positive and false negative rates (increasing power required either increasing α or n).

How should we use this information when we are trying to come up with a decision?

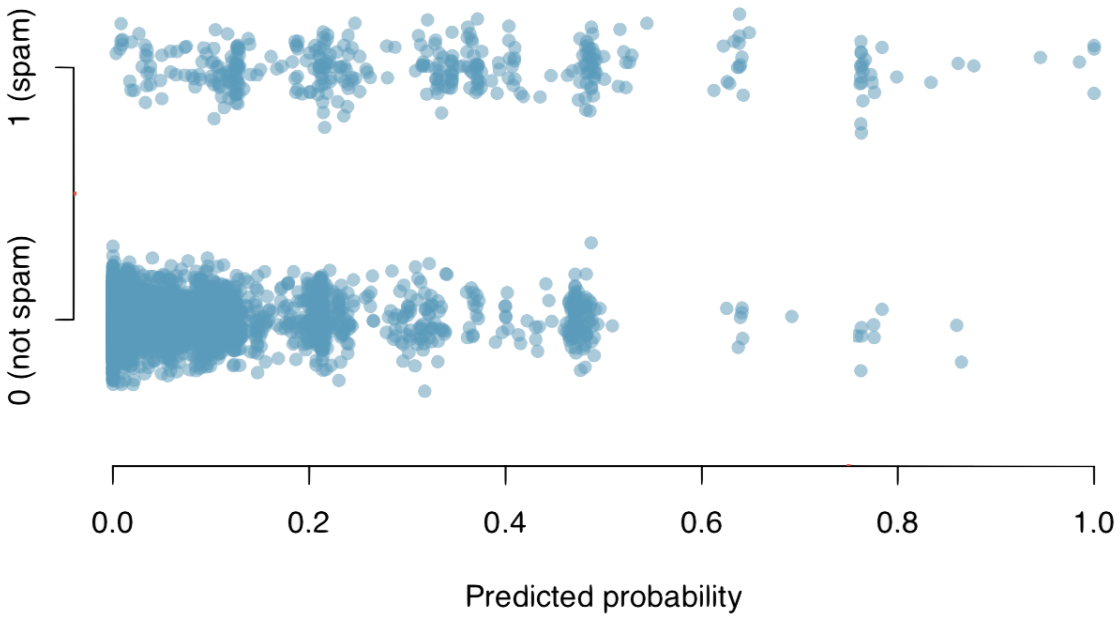
Back to Spam

In lab this week, we examined a data set of emails where we were interesting in identifying the spam messages. We examined different logistic regression models to evaluate how different predictors influenced the probability of a message being spam.

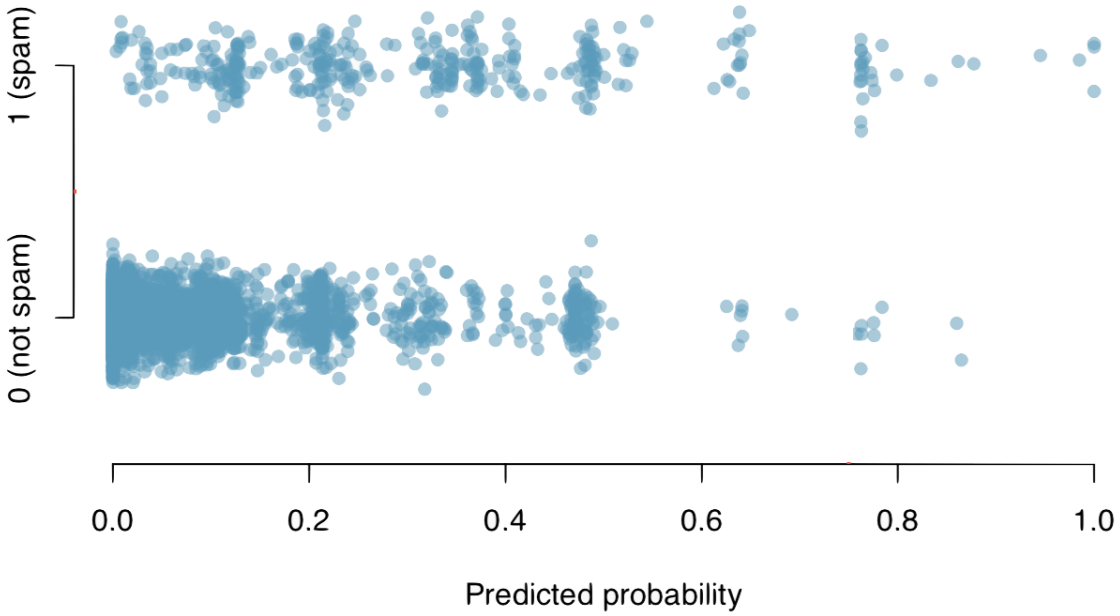
These models can also be used to assign probabilities to incoming messages (this is equivalent to prediction in the case of SLR / MLR). However, if we were designing a spam filter this would only be half of the battle, we would also need to use these probabilities to make a decision about which emails get flagged as spam.

While not the only possible solution, we will consider a simple approach where we choose a threshold probability and any email that exceeds that probability is flagged as spam.

Picking a threshold

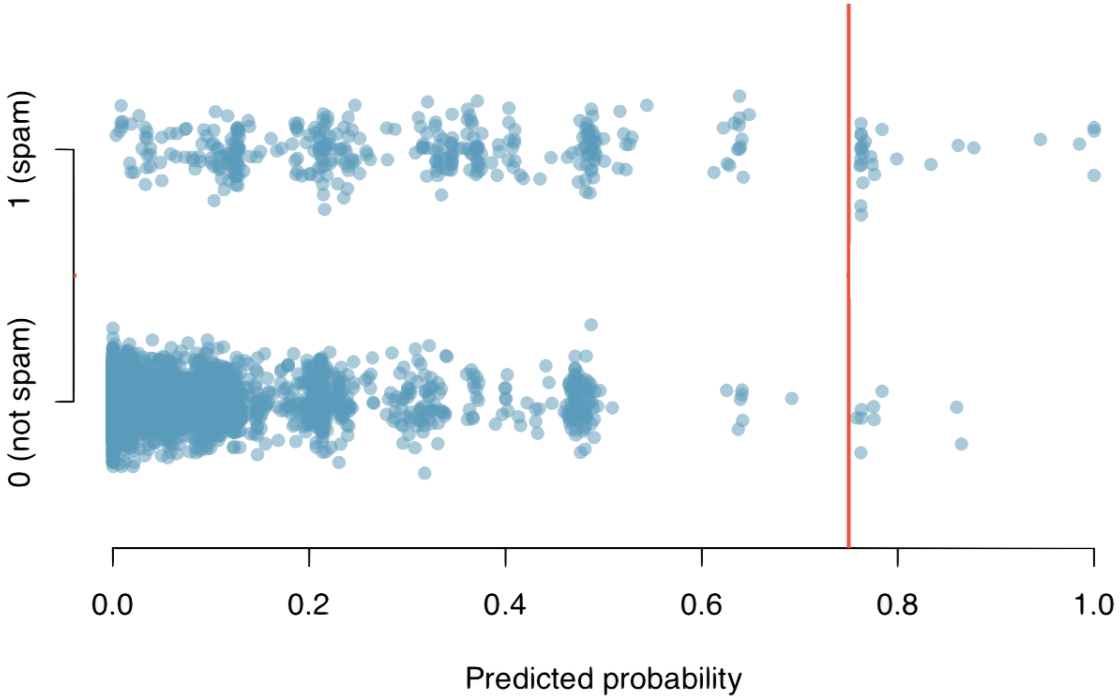


Picking a threshold



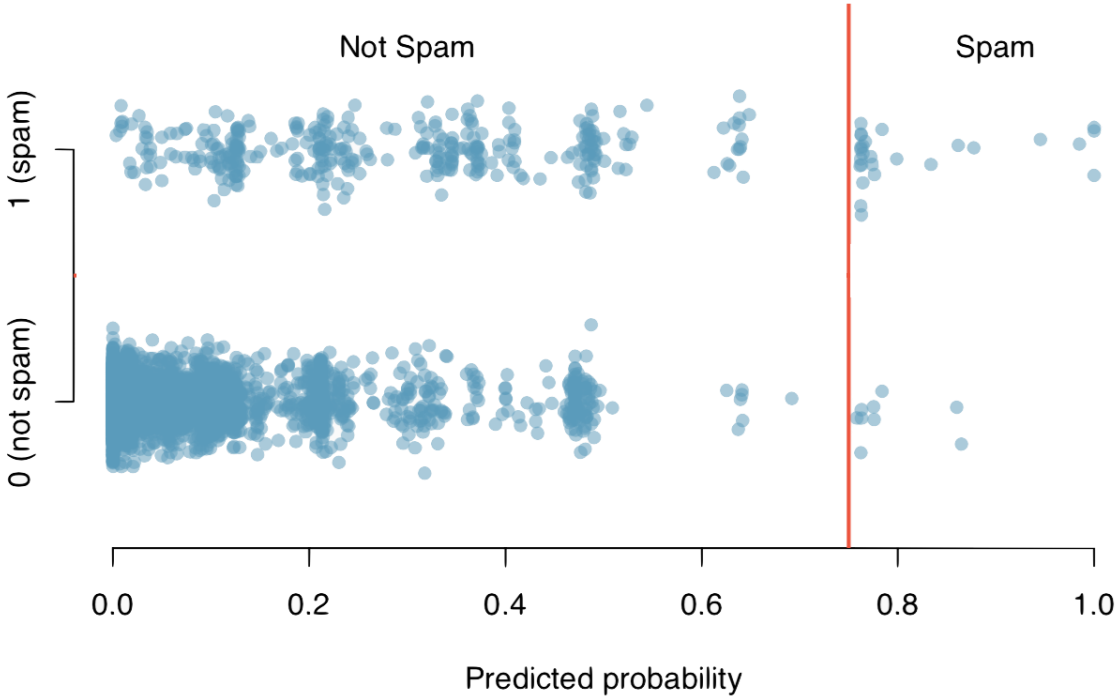
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



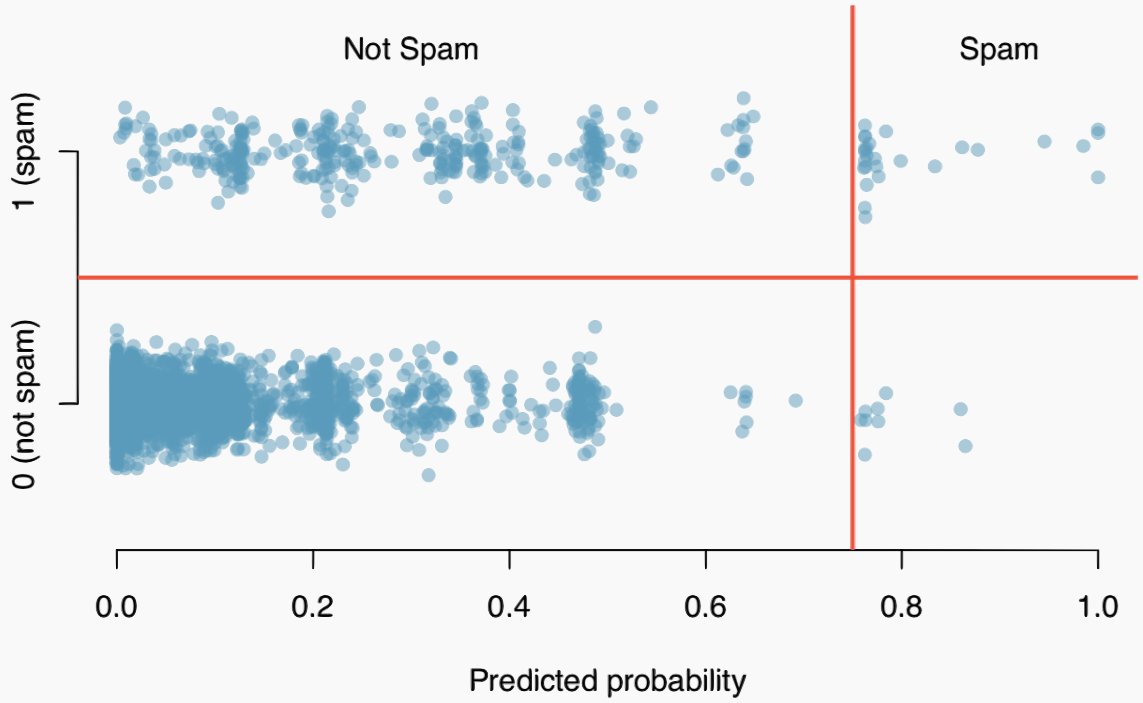
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



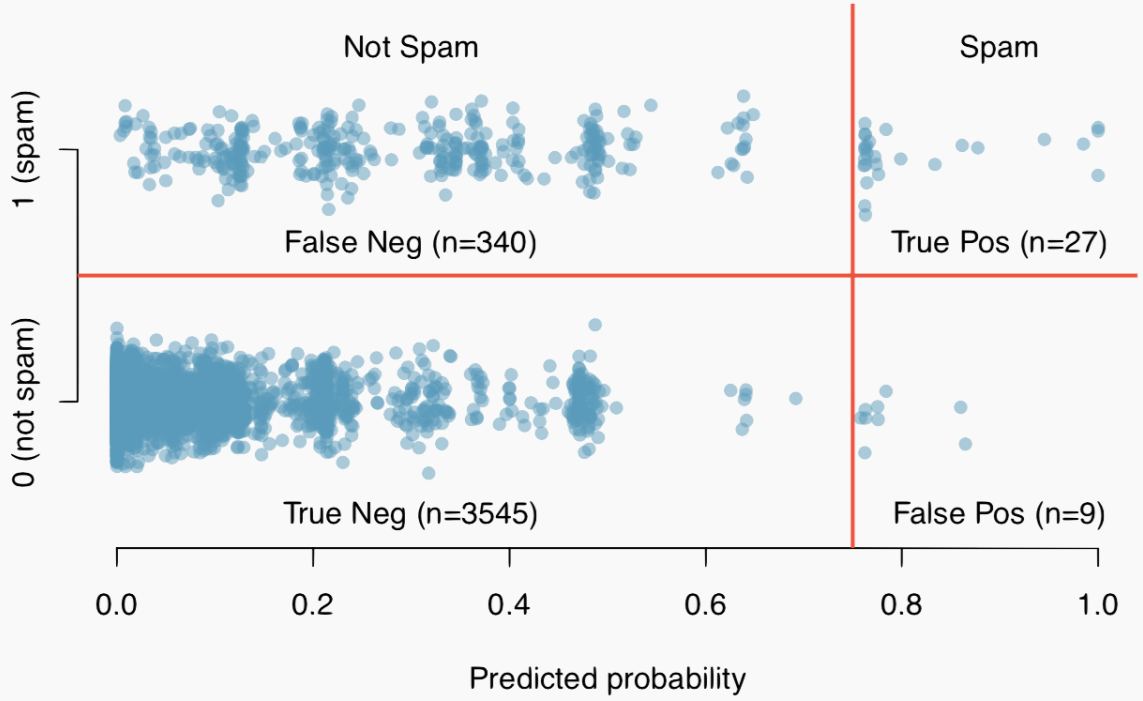
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



Lets see what happens if we pick our threshold to be 0.75.

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

FN = 340	TP = 27
TN = 3545	FP = 9

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

FN = 340	TP = 27
TN = 3545	FP = 9

What are the sensitivity and specificity for this particular decision rule?

Consequences of picking a threshold

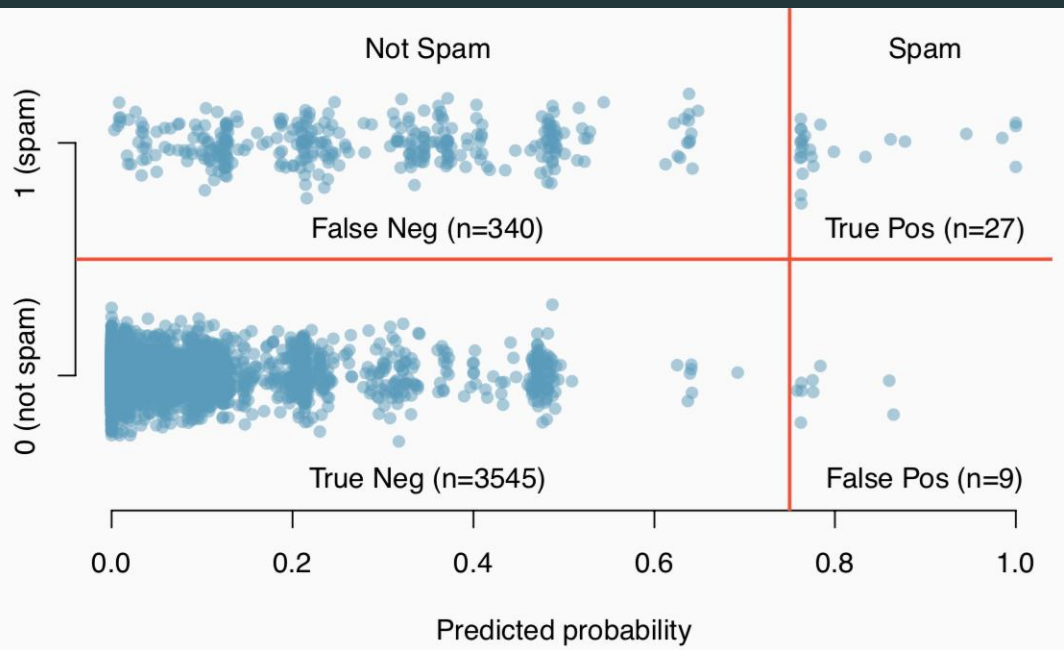
For our data set picking a threshold of 0.75 gives us the following results:

FN = 340	TP = 27
TN = 3545	FP = 9

What are the sensitivity and specificity for this particular decision rule?

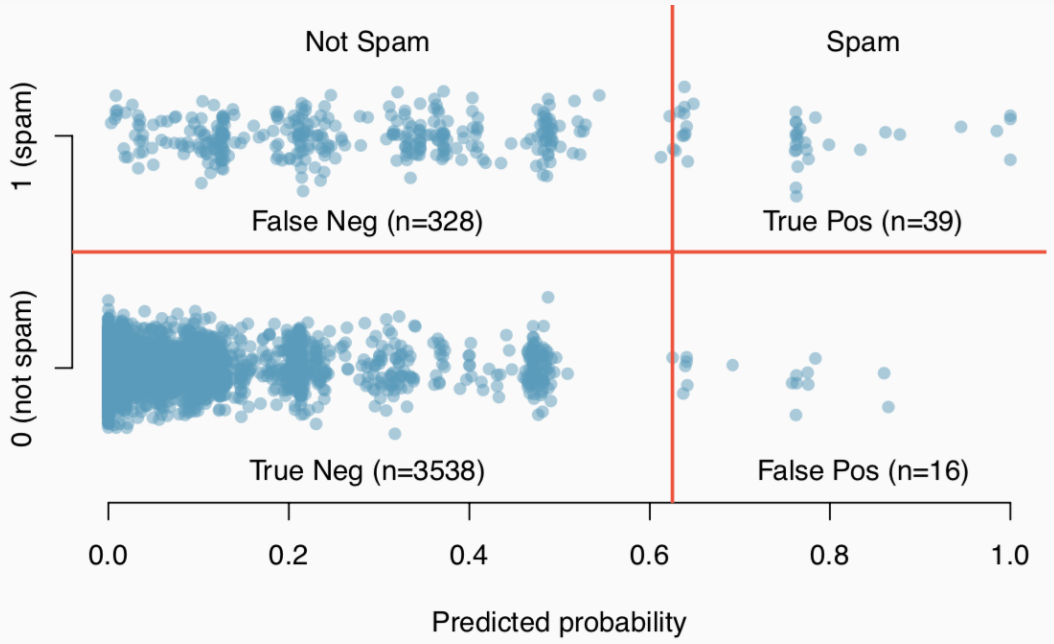
Sensitivity = TP/(TP + FN) = 27/(27 + 340) = 0.073
Specificity = TN/(FP + TN) = 3545/(9 + 3545) = 0.997

Trying other thresholds



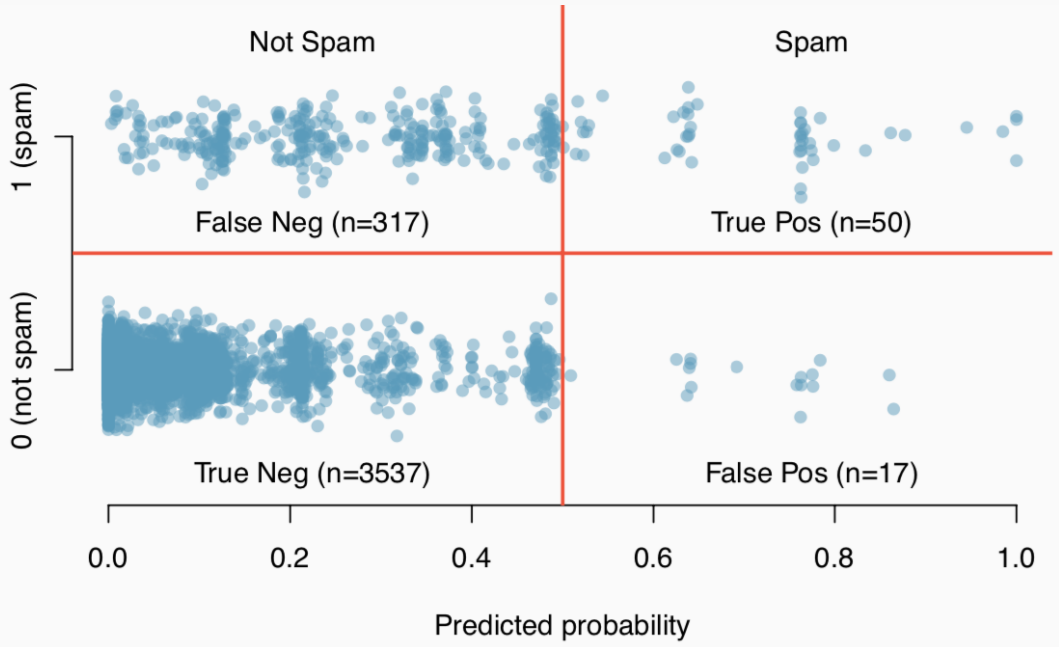
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074				
Specificity	0.997				

Trying other thresholds



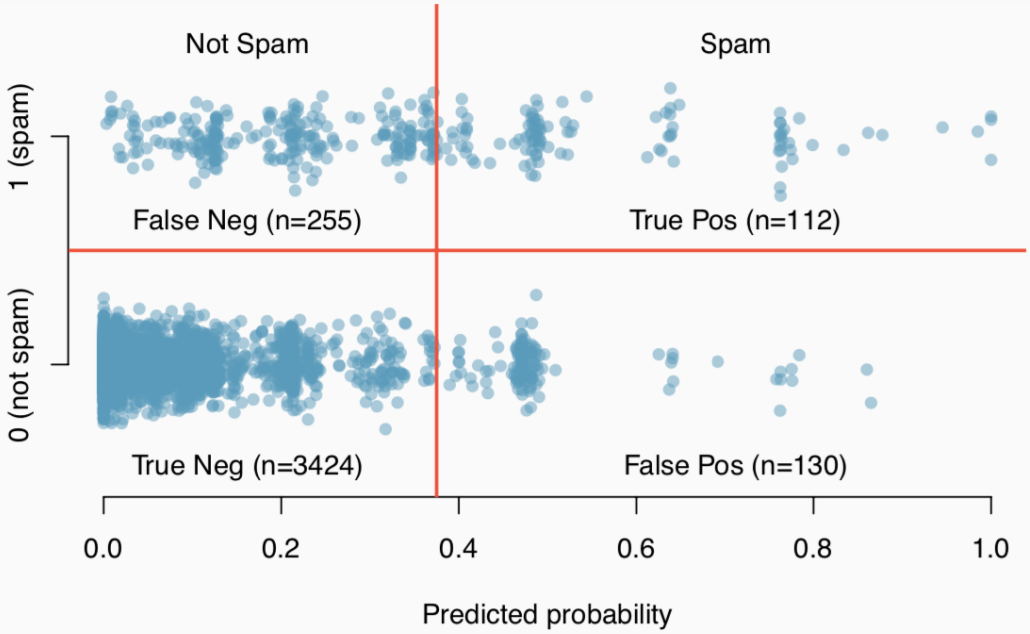
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Trying other thresholds



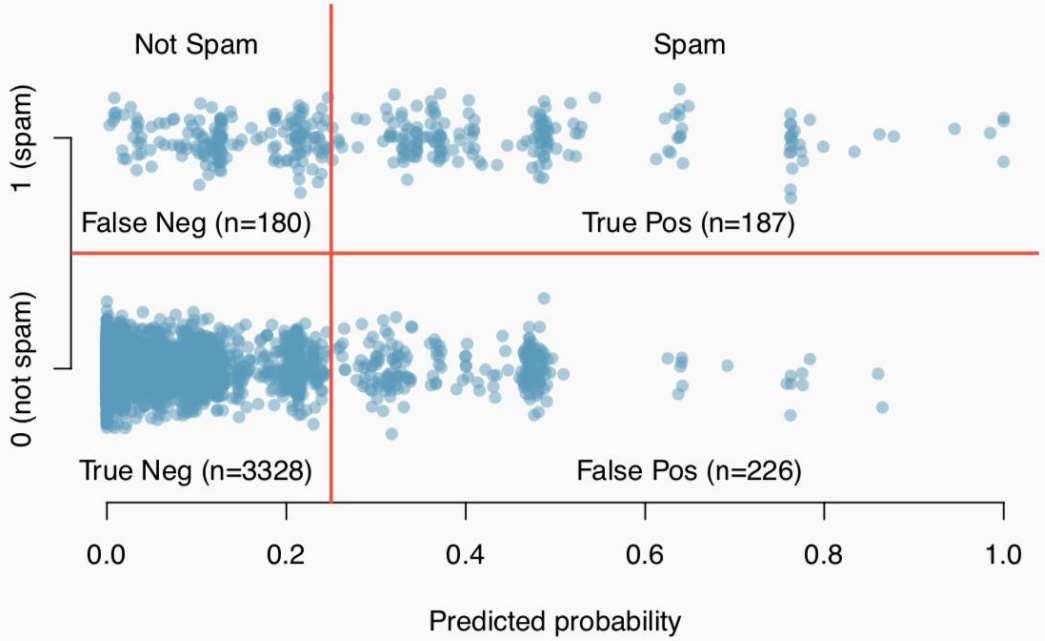
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136		
Specificity	0.997	0.995	0.995		

Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	
Specificity	0.997	0.995	0.995	0.963	

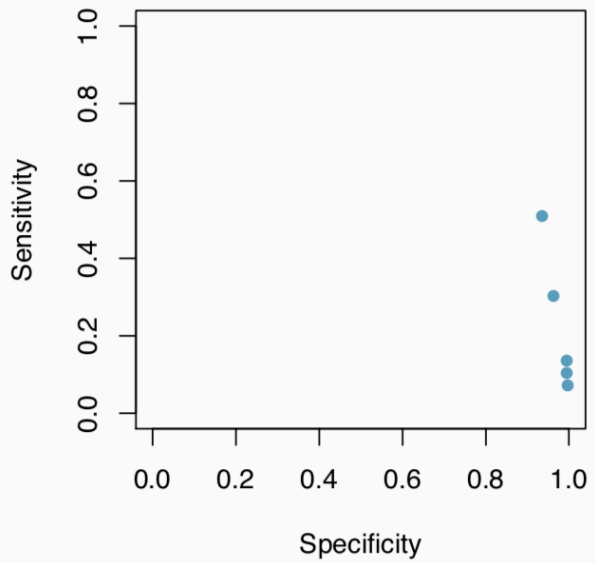
Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

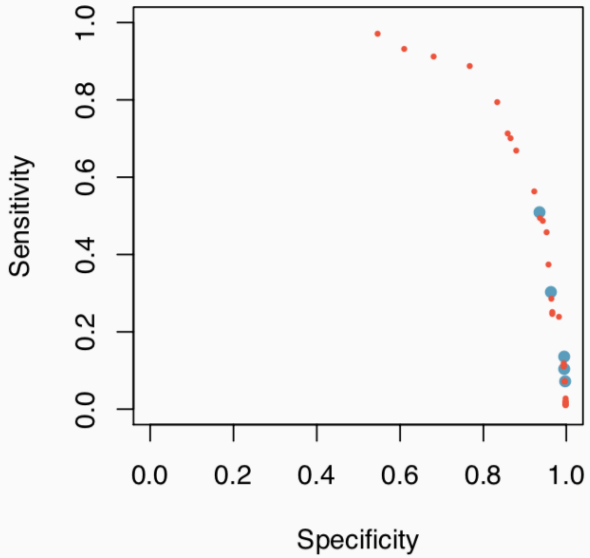
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



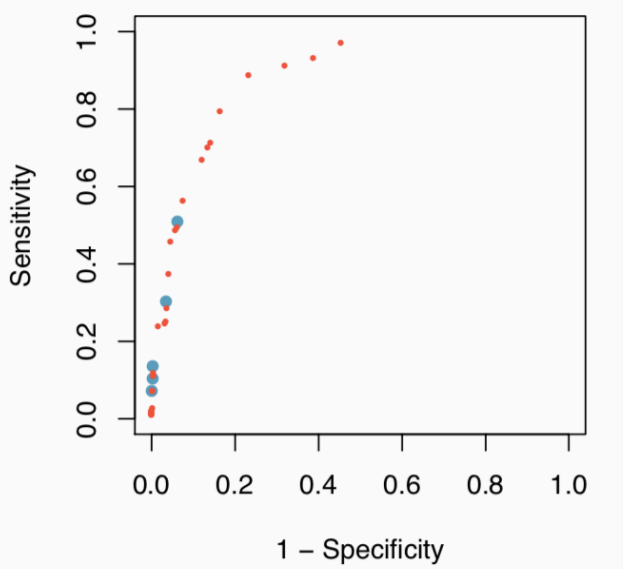
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

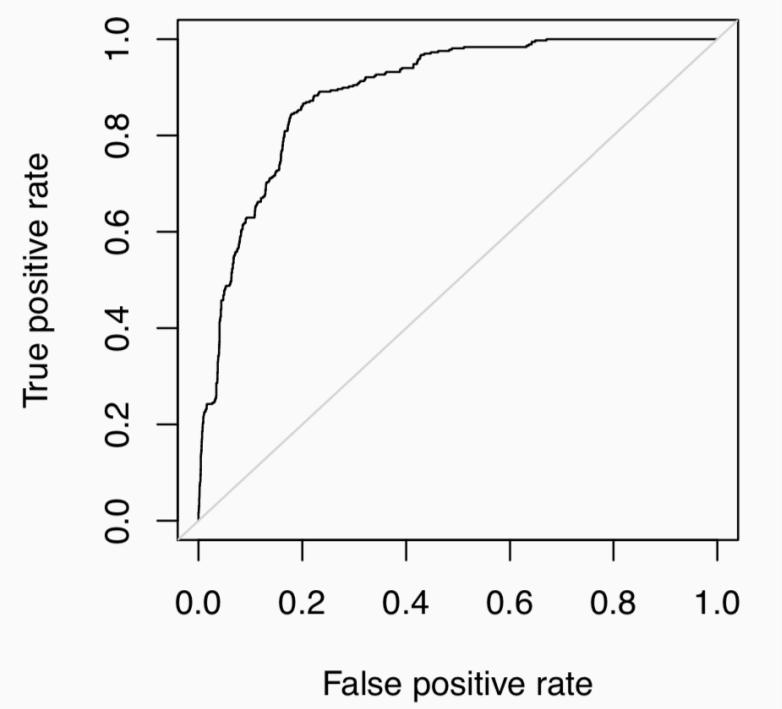


Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



Receiver operating characteristic (ROC) curve



Receiver operating characteristic (ROC) curve (cont.)

Why do we care about ROC curves?

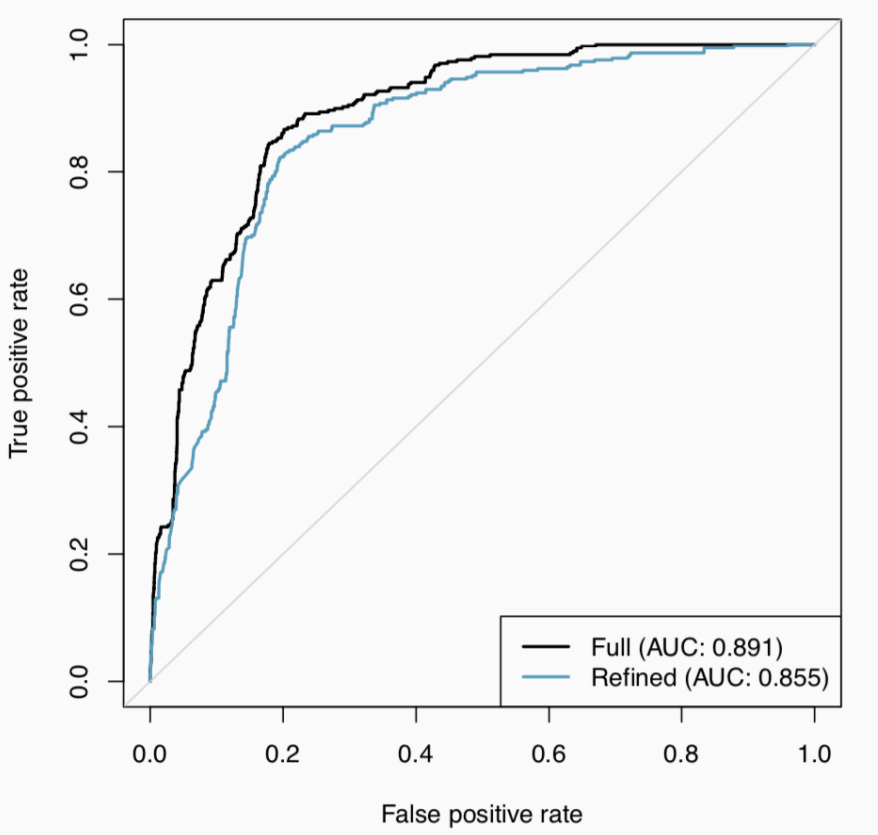
- Shows the trade off in sensitivity and specificity for all possible thresholds.
- Straight forward to compare performance vs. chance.
- Can use the area under the curve (AUC) as an assessment of the predictive ability of a model.

Refining the Spam model

```
g_refined = glm(spam ~ to_multiple+cc+image+attach+winner
                       +password+line_breaks+format+re_subj
                       +urgent_subj+exclaim_mess,
                data=email, family=binomial)
summary(g_refined)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7594	0.1177	-14.94	0.0000
to_multipleyes	-2.7368	0.3156	-8.67	0.0000
ccyes	-0.5358	0.3143	-1.71	0.0882
imageyes	-1.8585	0.7701	-2.41	0.0158
attachyes	1.2002	0.2391	5.02	0.0000
winneryes	2.0433	0.3528	5.79	0.0000
passwordyes	-1.5618	0.5354	-2.92	0.0035
line_breaks	-0.0031	0.0005	-6.33	0.0000
formatPlain	1.0130	0.1380	7.34	0.0000
re_subjyes	-2.9935	0.3778	-7.92	0.0000
urgent_subjyes	3.8830	1.0054	3.86	0.0001
exclaim_mess	0.0093	0.0016	5.71	0.0000

Comparing models



Utility Functions

There are many other reasonable quantitative approaches we can use to decide on what is the “best” threshold.

If you’ve taken an economics course you have probably heard of the idea of utility functions, we can assign costs and benefits to each of the possible outcomes and use those to calculate a utility for each circumstance.

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	
True Negative	
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	-5

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	-5

$$U(p) = TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p)$$

Utility for the 0.75 threshold

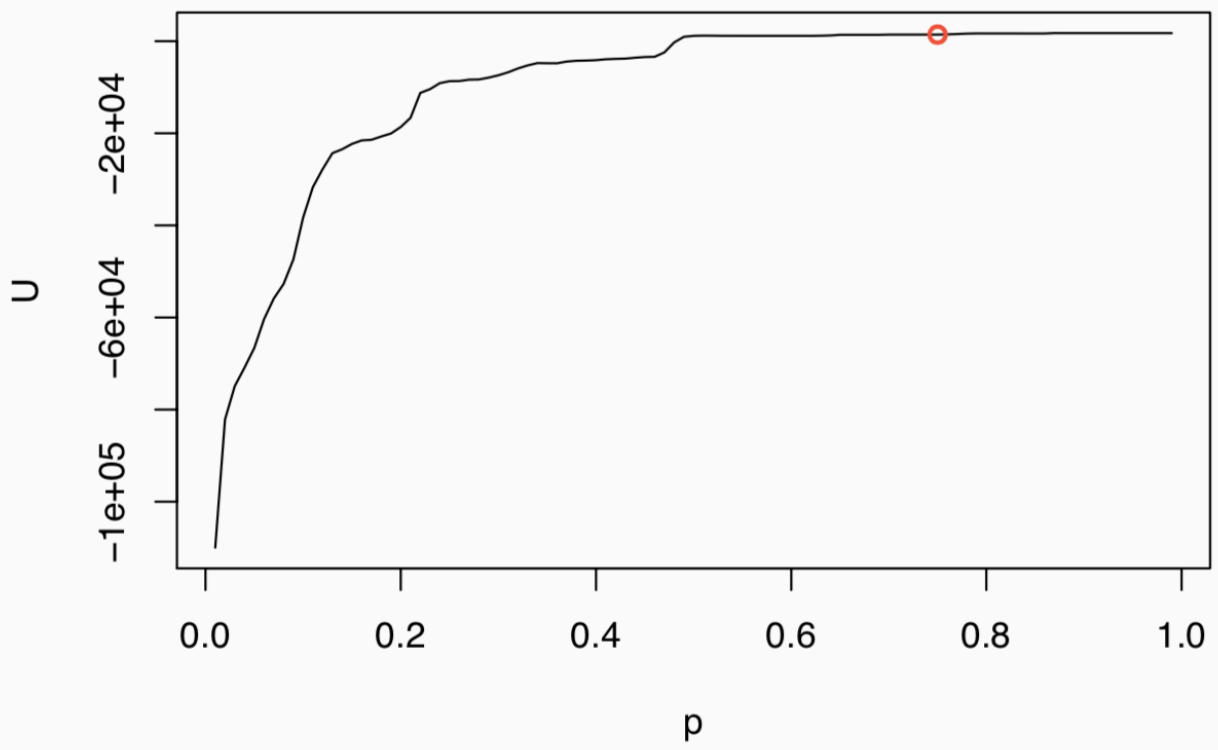
For the email data set picking a threshold of 0.75 gives us the following results:

FN = 340	TP = 27
TN = 3545	FP = 9

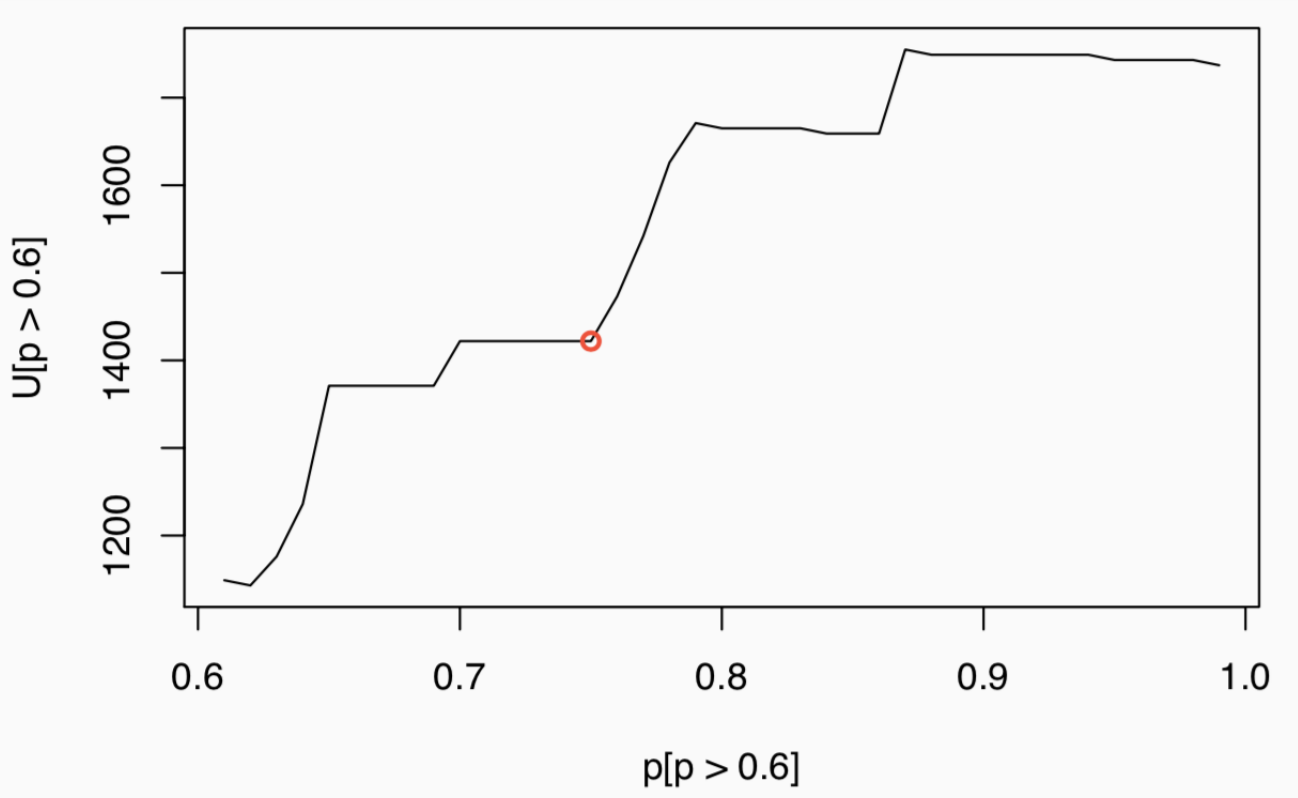
$$U(p) = TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p)$$
$$= 27 + 3545 - 50 \times 9 - 5 \times 340 = 1422$$

Not useful by itself, but allows us to compare with other thresholds.

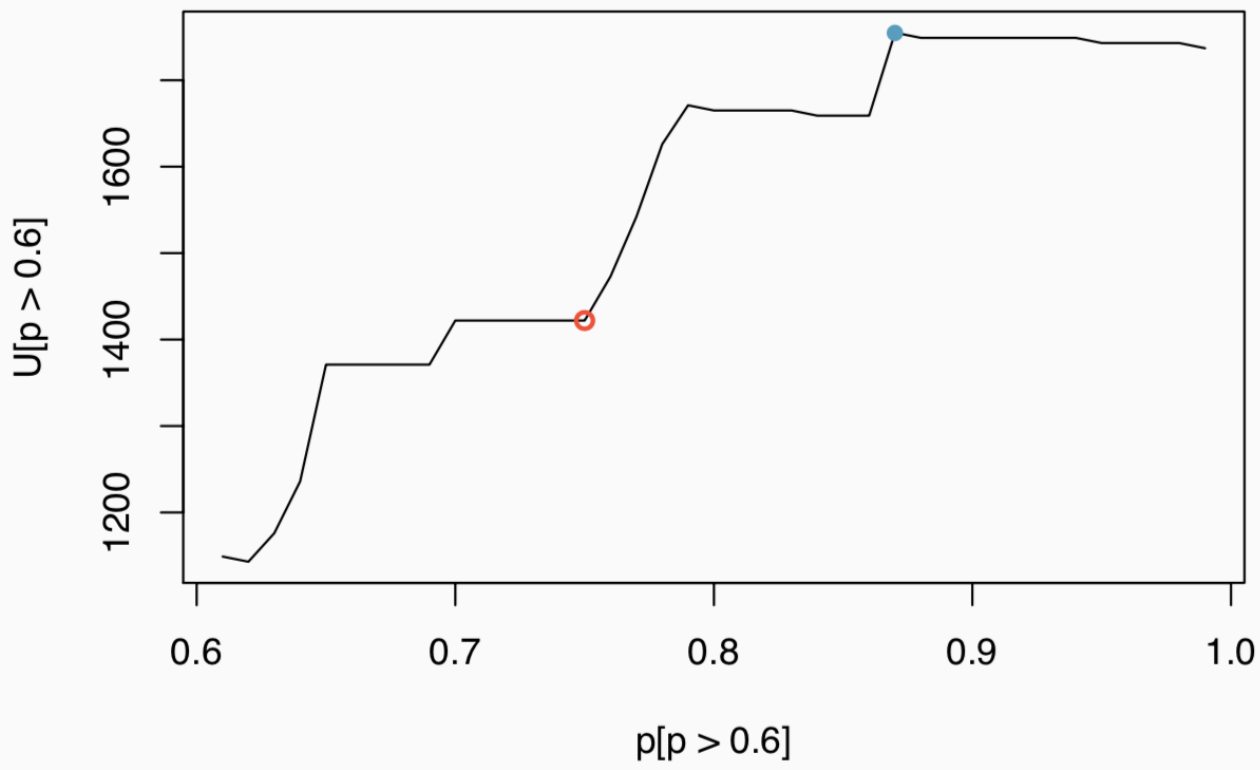
Utility curve



Utility curve (zoom)



Utility curve (zoom)



Maximum Utility