

# MAD-CB



# Regressão Múltipla

# Regressão - O Que Fizemos

- Regressão Simples
- Regressão Polinomial

# O Que Faremos Hoje

- Regressão Múltipla
  - ▶ Regressão com mais que uma variável independente

# Dataset para Hoje - Prestige

- Mede o prestígio com que Canadenses tratam 102 profissões
- Faz parte do pacote car
- Vem do livro **An R Companion to Applied Regression** de John Fox e Sanford Weisberg

```
data("Prestige")
head(Prestige[,c(1:4,6)])
```

##	education	income	women	prestige	type
## gov.administrators	13.11	12351	11.16	68.8	prof
## general.managers	12.26	25879	4.02	69.1	prof
## accountants	12.77	9271	15.70	63.4	prof
## purchasing.officers	11.42	8865	9.11	56.8	prof
## chemists	14.62	8403	11.68	73.5	prof
## physicists	15.64	11030	5.13	77.6	prof

# Variáveis do Dataset

- education: média dos anos de escolaridade para cada profissão
- income: média da renda anual em 1971
- women: porcentagem das mulheres na profissão
- prestige: opinião de prestígio de ocupação baseado numa sondagem
- census: variável não usado em nossa análise
- type: variável classificando as ocupações em wc (colarinho branco), bc (operário), prof (profissional/executivo)

## Passos 1 e 2 — Revisão e Limpeza dos Dados

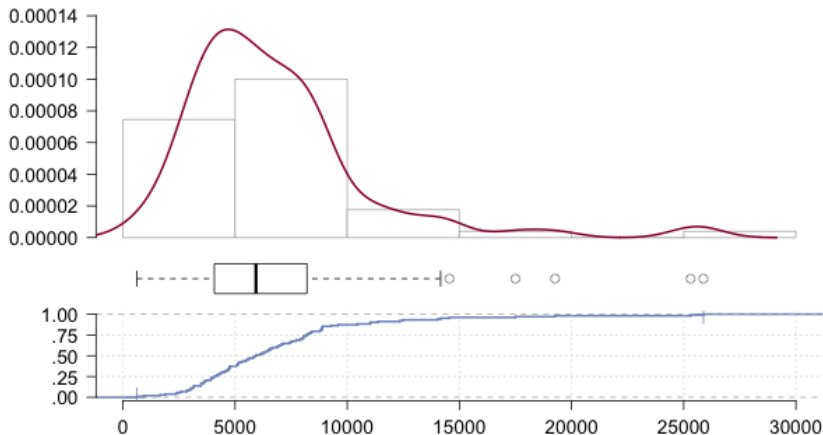
##	education	income	women
##	Min. : 6.380	Min. : 611	Min. : 0.000
##	1st Qu.: 8.445	1st Qu.: 4106	1st Qu.: 3.592
##	Median :10.540	Median : 5930	Median :13.600
##	Mean :10.738	Mean : 6798	Mean :28.979
##	3rd Qu.:12.648	3rd Qu.: 8187	3rd Qu.:52.203
##	Max. :15.970	Max. :25879	Max. :97.510

##	prestige	type
##	Min. :14.80	bc :44
##	1st Qu.:35.23	prof:31
##	Median :43.60	wc :23
##	Mean :46.83	NA's: 4
##	3rd Qu.:59.27	
##	Max. :87.20	



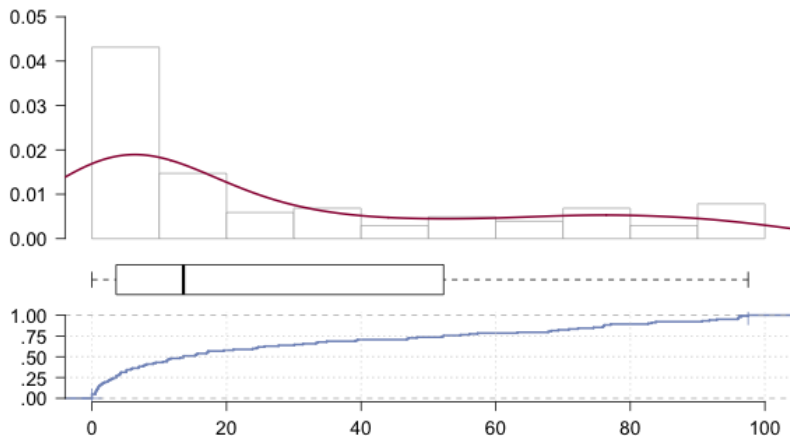
# Gráfico Densidade de income

**Prestige\$income (integer)**

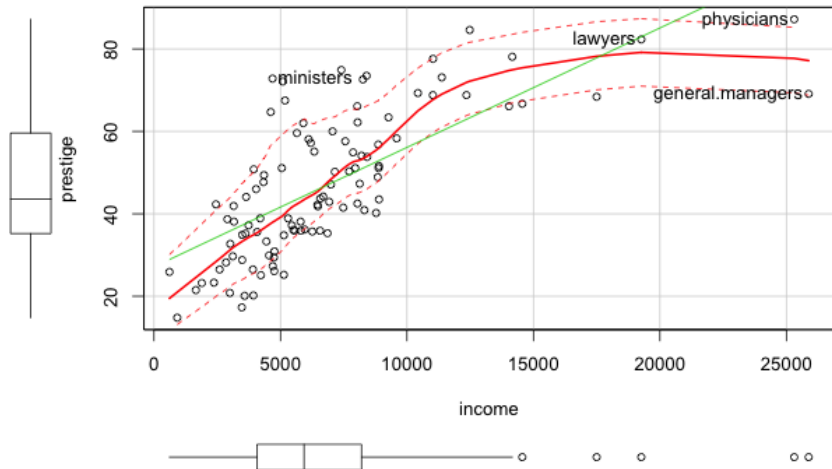


# Gráfico Densidade de women

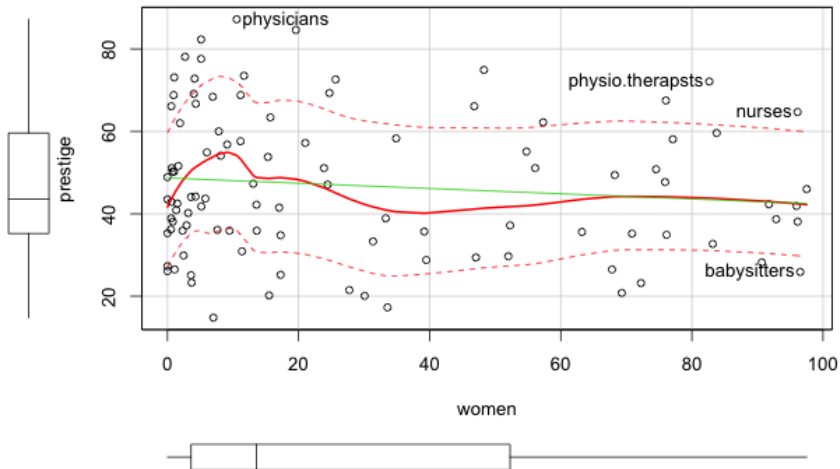
**Prestige\$women (numeric)**



# Scatterplot de prestige vs. income



# Scatterplot de prestige vs. women

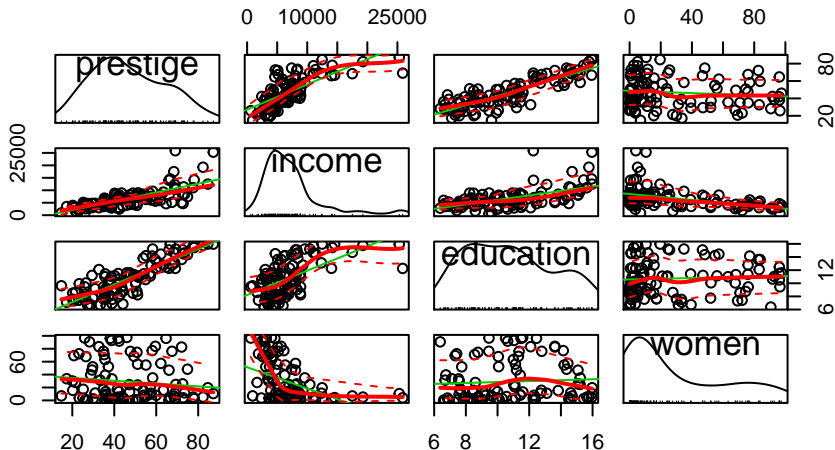


# Correlações entre as Variáveis

##		prestige	education	income	women
##	prestige	1.0000000	0.85017689	0.7149057	-0.11833419
##	education	0.8501769	1.0000000	0.5775802	0.06185286
##	income	0.7149057	0.57758023	1.0000000	-0.44105927
##	women	-0.1183342	0.06185286	-0.4410593	1.0000000

# Matriz de Scatterplots das Variáveis

```
scatterplotMatrix(~ prestige + income + education + women,  
  data = Prestige, span = 0.7)
```



# Transformações em income e women

- Essas variáveis precisam transformação para fazer elas aptas para regressão linear

# Transformações em income e women

- Essas variáveis precisam transformação para fazer elas aptas para regressão linear
- income tem curva de 2º grau no scatterplot



# Transformações em `income` e `women`

- Essas variáveis precisam transformação para fazer elas aptas para regressão linear
- `income` tem curva de 2º grau no scatterplot
- `women` tem assimetria pronunciada a direita (boxplot)

# Transformação da Variável `women`

- 2 candidatos para transformação para restaurar a normalidade

# Transformação da Variável `women`

- 2 candidatos para transformação para restaurar a normalidade
- `log`

# Transformação da Variável `women`

- 2 candidatos para transformação para restaurar a normalidade
- `log`
- `logit`

# Transformação $\log$ para women

- Problema com os 0's

# Transformação *log* para women

- Problema com os 0's
- Qualquer logaritmo não pode calcular o log de 0

# Transformação *log* para women

- Problema com os 0's
- Qualquer logarítmo não pode calcular o log de 0
- Em R,  $\log(0) = -\text{Inf}$

# Profissões Sem Mulheres

```
Prestige[Prestige$women == 0,]
```

##	education	income	women	prestige	type
## firefighters	9.47	8895	0	43.5	bc
## rotary.well.drillers	8.88	6860	0	35.3	bc
## railway.sectionmen	6.67	4696	0	27.3	bc
## train.engineers	8.49	8845	0	48.9	bc
## longshoremen	8.37	4753	0	26.1	bc



# Pode Modificar os Valores para Tirar os 0's

- Pode aumentar todos os valores por um pequeno quantidade sem mudar a distribuição
- Neste caso, vamos usar 0.025

# Transformação *log* para women

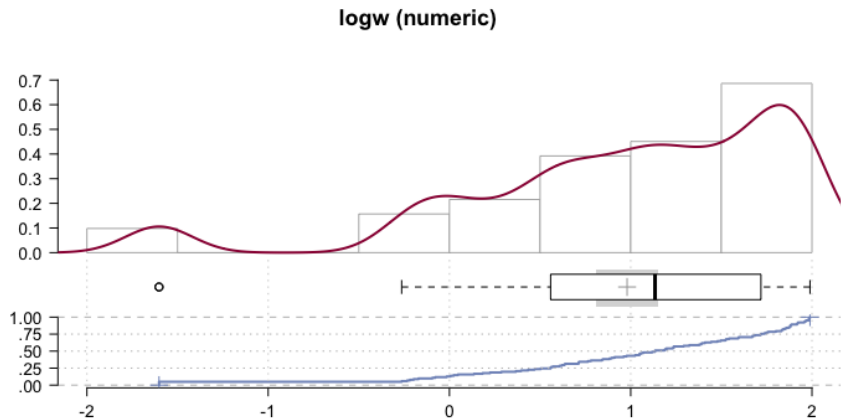
```
summary(log10(Prestige$women))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      -Inf  0.5554  1.1340   -Inf  1.7180  1.9890
```

```
logw <- log10(Prestige$women + 0.025)
summary(logw)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.6020  0.5584  1.1340  0.9802  1.7180  1.9890
```

# Gráfico de Log10(Women)



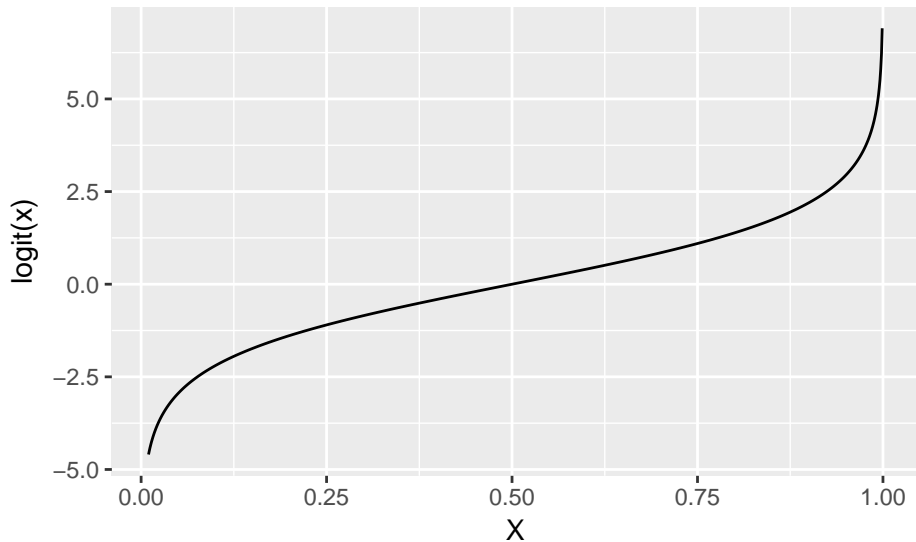
# Transformação *logit*

- *logit* aplica a função seguinte para a variável:

$$\text{logit}(x) = \log_e \frac{x}{1-x}$$

- Chamada a função *logit*
- Precisa ser aplicada às funções no intervalo de 0 e 1
- `women` começou como uma percentagem.
- Podemos dividir por 100 para restaurar a forma decimal sem violar a distribuição da variável

# Curva da Função logit



## Transformação *logit* de women

- Em R, existe a função *logit* no pacote *car*

## Transformação *logit* de women

- Em R, existe a função *logit* no pacote *car*
- Calcula automaticamente o valor e controla para os 0's

# Transformação *logit* de women

- Em R, existe a função *logit* no pacote *car*
- Calcula automaticamente o valor e controla para os 0's
- Quando variável tem 0's, restringe extensão dos valores de 0,025 até 0,975



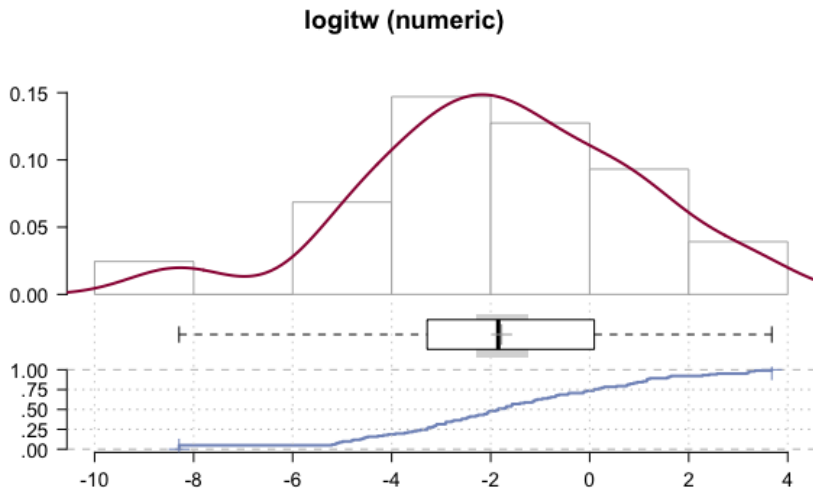
# Transformação *logit* de women

- Em R, existe a função *logit* no pacote *car*
- Calcula automaticamente o valor e controla para os 0's
- Quando variável tem 0's, restringe extensão dos valores de 0,025 até 0,975
- Entretanto, vamos usar a mesmo aumento de 0.025 que usamos na transformação  $\log_{10}$

# Resultado da Transformação

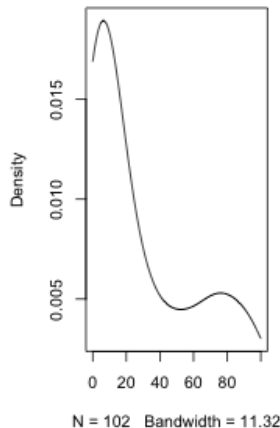
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	-8.29400	-3.28300	-1.84700	-1.77800	0.08916	3.67800

# Nova distribuição Transformada com *logit*

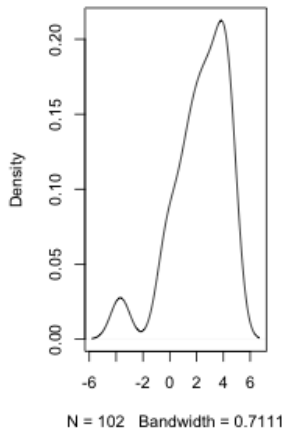


# Resumo das Transformações *log* e *logit*

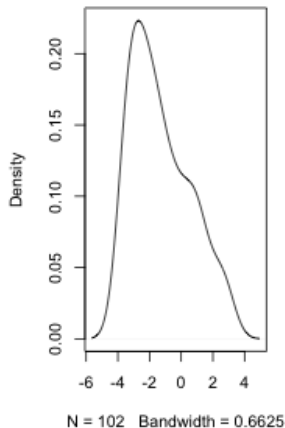
(a) Original



(b) Logarítmico



(c) Logit

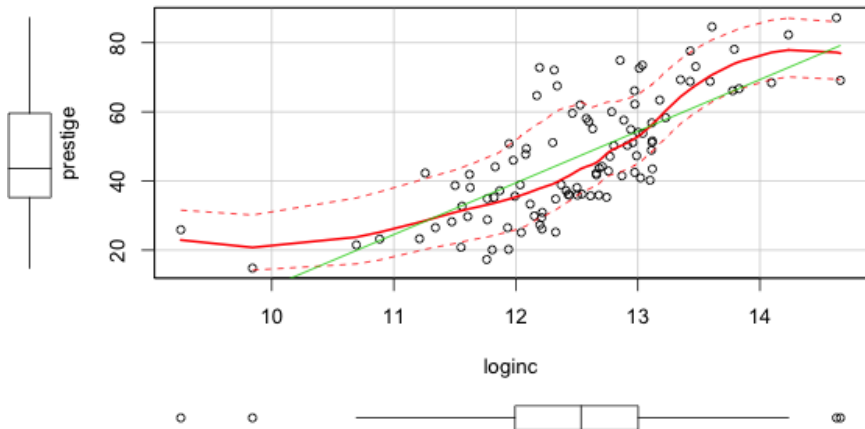


# Transformação da Variável income

- Neste caso, renda normalmente usa a transformação *log*

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	9.255	12.000	12.530	12.490	13.000	14.660

# Scatterplot com transformação da income



# Integrar Transformações no Dataset

- Agora que testamos as transformações, podemos integrar elas no dataset
- Para `women` usaremos a transformação *logit*

```
Prestige <- mutate(Prestige,  
  womenlogit = logit((Prestige$women + 0.025)/100),  
  inclog = log2(income))
```

# Verificação das Correlações

##		prestige	education	womenlogit	inclog
## prestige		1.00000000	0.8501769	-0.03476255	0.7410561
## education		0.85017689	1.0000000	0.16670369	0.5481051
## womenlogit		-0.03476255	0.1667037	1.00000000	-0.4386544
## inclog		0.74105613	0.5481051	-0.43865438	1.0000000



- Primeiro modelo só vai usar os variáveis numéricas
- Outro regressor, `type`, é nominal e precisa tratamento especial
- Usaremos os mesmos símbolos de formula que na regressão simples
  - ▶ “~” (til) para separar as variáveis dependente das independentes
  - ▶ “+” (mais) para separar os termos independentes
- Nas formulas, não precisa especificar o nome de dataframe
  - ▶ Faz isso com o parâmetro `data = Prestige`

# Modelo 1

```
fit1 <- lm(prestige ~ inclog + education + womenlogit,  
          data = Prestige)
```

# Resumo do Modelo 1

Call:

```
lm(formula = prestige ~ inclog + education + womenlogit, data = Prestige)
```

Residuals:

Min	1Q	Median	3Q	Max
-17.2972	-4.1876	0.1766	4.3429	18.4359

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-103.0222	13.3980	-7.689	1.16e-11	***
inclog	8.7821	1.2999	6.756	1.02e-09	***
education	3.7967	0.3705	10.247	< 2e-16	***
womenlogit	0.3609	0.3529	1.023	0.309	

---

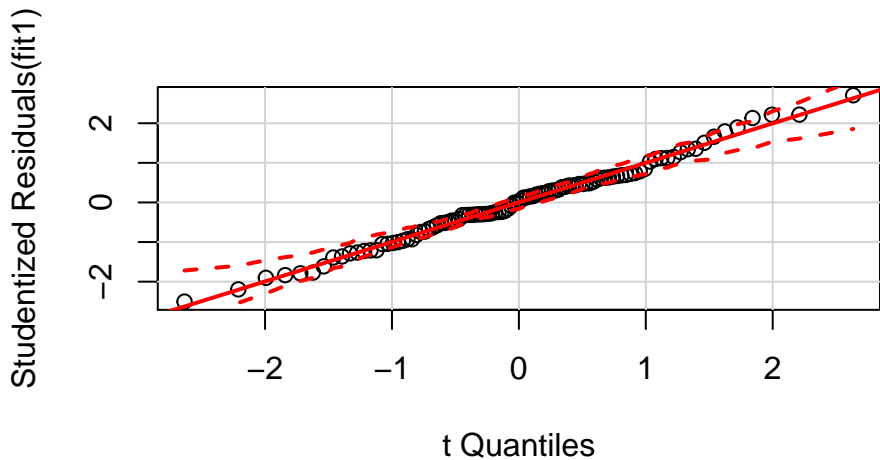
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.143 on 98 degrees of freedom

Multiple R-squared: 0.8327, Adjusted R-squared: 0.8276

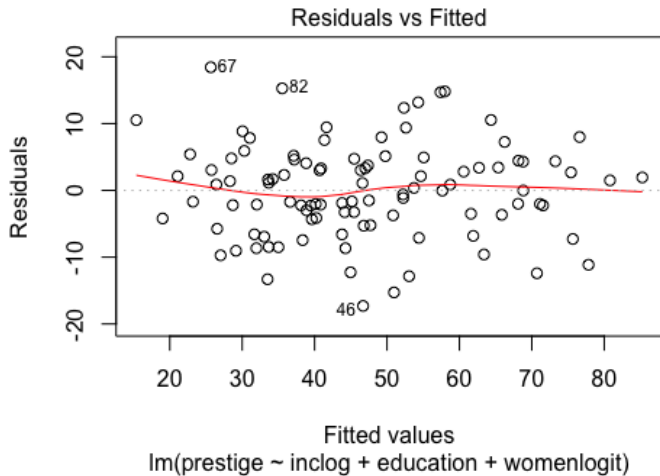
F-statistic: 162.6 on 3 and 98 DF, p-value: < 2.2e-16

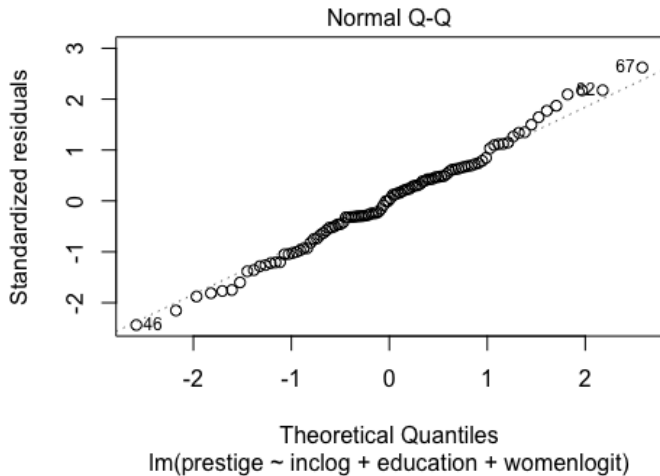
## qqplot do Modelo 1

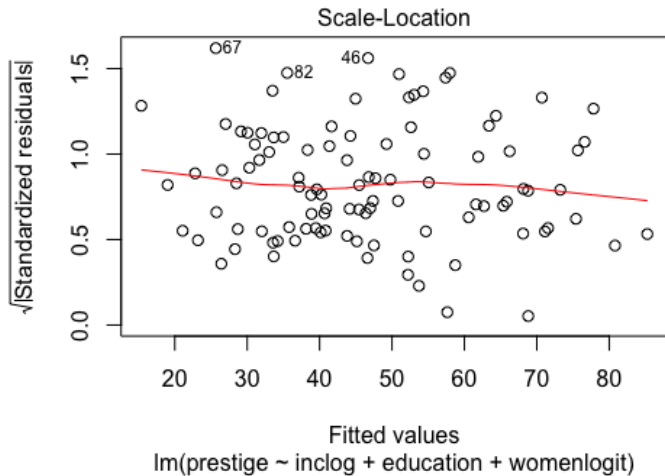


# Plotagens Diagnósticos do Modelo

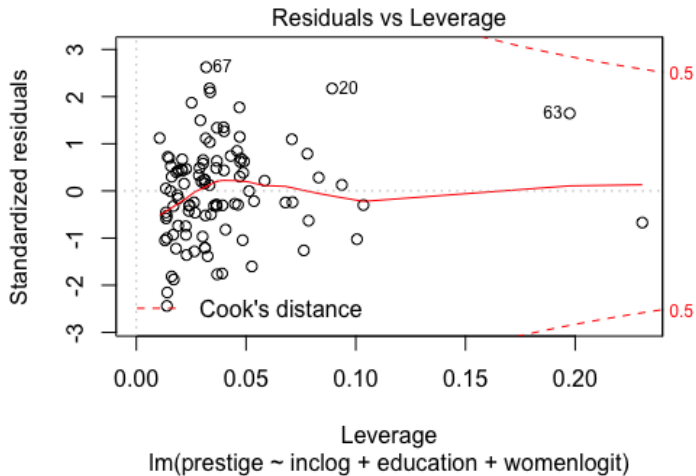
```
par(mfrow=c(2,2))  
plot(fit1)  
par(mfrow=c(1,1))
```











**\*\* NOT BAD\*\***

- O modelo explica 83% da variância

**\*\* NOT BAD\*\***

- O modelo explica 83% da variância
- `women` faz pouco contribuição para modelo

**\*\* NOT BAD\*\***

- O modelo explica 83% da variância
- `women` faz pouca contribuição para modelo
- como prevista nas correlações

## Modelo 2 - Sem women

```
fit2 <- lm(prestige ~ inclog + education,  
          data = Prestige)
```

# Resumo do Modelo 2

Call:

```
lm(formula = prestige ~ inclog + education, data = Prestige)
```

Residuals:

Min	1Q	Median	3Q	Max
-17.0346	-4.5657	-0.1857	4.0577	18.1270

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-95.1940	10.9979	-8.656	9.27e-14 ***
inclog	7.9278	0.9961	7.959	2.94e-12 ***
education	4.0020	0.3115	12.846	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.145 on 99 degrees of freedom

Multiple R-squared: 0.831, Adjusted R-squared: 0.8275

F-statistic: 243.3 on 2 and 99 DF, p-value: < 2.2e-16

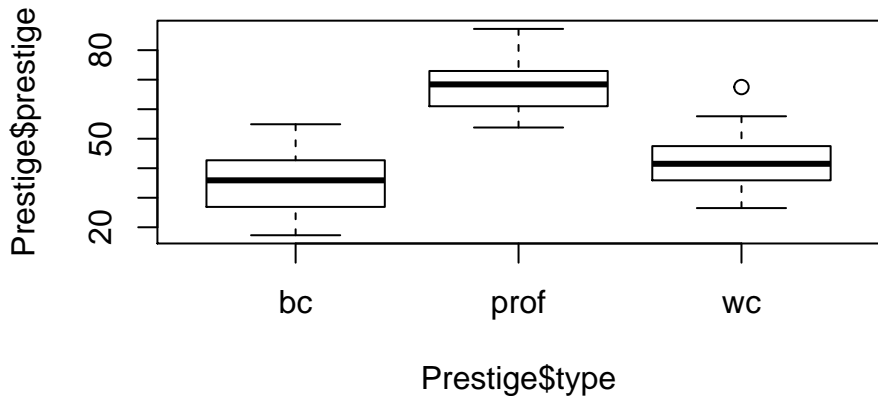
- Como indica os  $R^2$ , o modelo fica o mesmo sem a variável women

# Variável `type` no Modelo

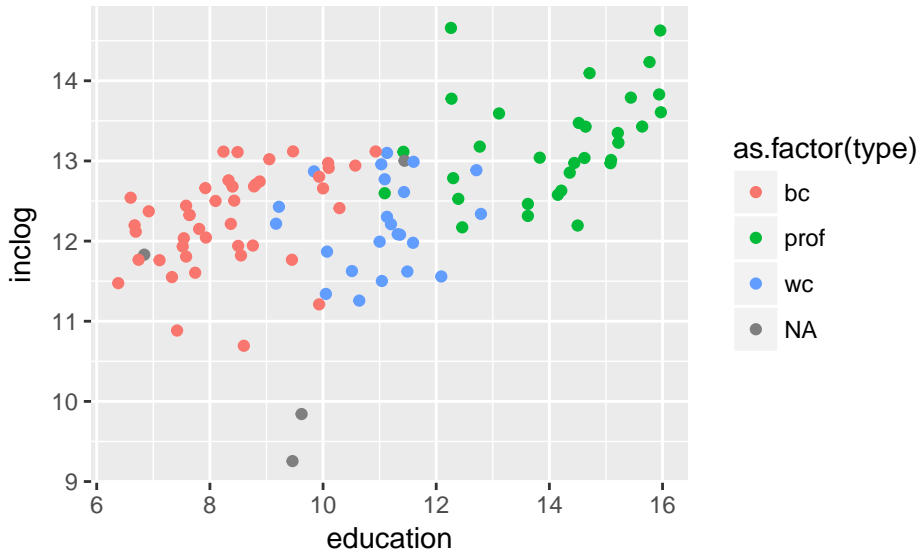
- Variável nominal/qualitativa
- 2 gráficos mostra o efeito da variável no resultado
  - ▶ boxplot mostrando `prestige` contra as categorias de `type`
  - ▶ scatterplot de `education` contra `prestige` com os tipos em cores



# Boxplot



# Scatterplot



# R Cria Variáveis “Dummy” para os Níveis da Variável Nominal

- Pode ver com o `model.matrix` (parcial)

```
with(Prestige, model.matrix(~ type)[c(1:5, 50:54),])
```

##	(Intercept)	typeprof	typewc
## 1	1	1	0
## 2	1	1	0
## 3	1	1	0
## 4	1	1	0
## 5	1	1	0
## 51	1	0	1
## 52	1	0	1
## 54	1	0	0
## 55	1	0	1
## 56	1	0	1

## Modelo com type

```
fit3 <- lm(prestige ~ inclog + education + type,  
          data = Prestige)
```

# Resumo de Modelo 3

Call:

```
lm(formula = prestige ~ inclog + education + type, data = Prestige)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.511	-3.746	1.011	4.356	18.438

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-81.2019	13.7431	-5.909	0.0000000563 ***
inclog	7.2694	1.1900	6.109	0.0000000231 ***
education	3.2845	0.6081	5.401	0.0000005058 ***
typeprof	6.7509	3.6185	1.866	0.0652 .
typewc	-1.4394	2.3780	-0.605	0.5465

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.637 on 93 degrees of freedom

(4 observations deleted due to missingness)

Multiple R-squared: 0.8555, Adjusted R-squared: 0.8493

F-statistic: 137.6 on 4 and 93 DF, p-value: < 2.2e-16

# Interações Entre Variáveis

- Parecem de existir uma relação entre `type` e `income` e `education`.
- `type` não parece de ser realmente independente das outras
- Podemos avaliar esta independência por incluir termos de interação entre as variáveis
- Usamos o símbolo “:” (dois pontos) para indicar interação
- E.g., `education:type`
- Programa vai aumentar mais variáveis “dummy” para tomar conta das interações

## model.matrix com Interação

```
model.matrix(~ type + education + education:type,  
             data = Prestige)[c(1:5, 50:54),]
```

# model.matrix

##	typeprof:education	typewc:education
## gov.administrators	13.11	0.00
## general.managers	12.26	0.00
## accountants	12.77	0.00
## purchasing.officers	11.42	0.00
## chemists	14.62	0.00
## commercial.travellers	0.00	11.13
## sales.clerks	0.00	10.05
## service.station.attendant	0.00	0.00
## insurance.agents	0.00	11.60
## real.estate.salesmen	0.00	11.09
## buyers	0.00	11.03



## Executar o Modelo com Interação

```
fit5 <- lm(prestige ~ inclog + education + type +  
           inclog:type + education:type, data = Prestige)
```

# Resumo do Modelo com Interação

Call:

```
lm(formula = prestige ~ inclog + education + type + inclog:type +  
  education:type, data = Prestige)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.970	-4.124	1.206	3.829	18.059

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-120.0459	20.1576	-5.955	0.0000000507 ***
inclog	11.0782	1.8063	6.133	0.0000000232 ***
education	2.3357	0.9277	2.518	0.01360 *
typeprof	85.1601	31.1810	2.731	0.00761 **
typewc	30.2412	37.9788	0.796	0.42800
inclog:typeprof	-6.5356	2.6167	-2.498	0.01434 *
inclog:typewc	-5.6530	3.0519	-1.852	0.06730 .
education:typeprof	0.6974	1.2895	0.541	0.58998
education:typewc	3.6400	1.7589	2.069	0.04140 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.409 on 89 degrees of freedom

(4 observations deleted due to missingness)

Multiple R-squared: 0.871, Adjusted R-squared: 0.8595

F-statistic: 75.15 on 8 and 89 DF, p-value: < 2.2e-16

- Agora  $\text{inclog}$  tem o maior valor  $t$  e explica mais do prestige que as outras variáveis

# Resultado do Modelo

- Agora `inclog` tem o maior valor  $t$  e explica mais do prestige que as outras variáveis
- `education` recuou em importância por causa da interação com `type`

# Resultado do Modelo

- Agora `inclog` tem o maior valor  $t$  e explica mais do prestige que as outras variáveis
- `education` recuou em importância por causa da interação com `type`
- Modelo agora diz que alguém com uma um trabalho que renda bastante recebe mais prestígio que de alguém numa profissão que precisa alto grau de educação

- Previsão de probabilidades e odds