

LOGIT E PROBIT

Marcus Antonio Cardoso Ramalho
Claudia Regina da Costa de Souza Ben Hur Correia

2025-06-25

Índice

1	Introdução	2
1.1	Variáveis Dependentes Limitadas	2
1.1.1	Por que não usar modelo linear?	2
1.2	Especificação dos Modelos	2
1.2.1	Modelo Logit	2
1.2.2	Modelo Probit	3
2	Exemplo Prático: Participação no Mercado de Trabalho	3
2.1	Descrição dos Dados	3
2.1.1	Variáveis Explicativas:	3
2.2	Modelo Teórico	3
2.3	Análise Exploratória dos Dados	4
2.3.1	Interpretação da Análise Exploratória	5
2.3.2	Análise dos Gráficos Exploratórios	7
2.4	Estimação dos Modelos	8
2.4.1	Modelo Logit	8
2.4.2	Interpretação do Modelo Logit	8
2.4.3	Modelo Probit	9
2.4.4	Interpretação do Modelo Probit	10
2.5	Efeitos Marginais	11
2.5.1	Fórmulas Teóricas	11
2.5.2	Interpretação das Fórmulas	11
2.5.3	Interpretação dos Efeitos Marginais	13
2.6	Qualidade da Previsão	15
2.6.1	Análise da Qualidade Preditiva	15
2.7	Pseudo- R^2	17
2.7.1	Interpretação do Pseudo- R^2	18

2.8	Razão de Chances (Odds Ratio)	19
2.8.1	Interpretação da Razão de Chances	19
2.8.2	Interpretação da Razão de Chances:	22
2.9	Comparação Visual dos Modelos	22
2.9.1	Análise Comparativa das Funções	24
2.10	Resumo Comparativo dos Modelos	24
2.11	Conclusões	24
2.11.1	Principais Achados	24
2.11.2	Escolha entre Modelos	25
2.11.3	Recomendações Práticas	26
2.11.4	Implicações para Política Pública	26
2.11.5	Limitações do Estudo	26

1 Introdução

1.1 Variáveis Dependentes Limitadas

Os modelos Logit e Probit (abreviação de regressão logística e probabilística) nos auxiliam na inferência de probabilidade de ocorrência de eventos onde nossa variável dependente é binária (Y ocorre ou não ocorre), e nosso objetivo é compreender como outras variáveis influenciam a ocorrência ou não desses eventos.

1.1.1 Por que não usar modelo linear?

Em uma regressão linear, $P(Y = 1|x)$ é dado por uma especificação linear dos regressores, o que pode resultar em valores menores que 0 ou maiores que 1, que não fazem sentido com a interpretação probabilística dos parâmetros.

Os modelos não lineares permitem que a média condicional de Y dado X seja expressa pela probabilidade de Y acontecer dado X:

$$E(Y|X) = P(Y = 1|X)$$

1.2 Especificação dos Modelos

1.2.1 Modelo Logit

A função de distribuição logística é dada por:

$$F(X'\beta) = \frac{e^{X'\beta}}{1 + e^{X'\beta}} = \frac{1}{1 + e^{-X'\beta}}$$

1.2.2 Modelo Probit

A função de distribuição normal padrão é dada por:

$$F(X'\beta) = \Phi(X'\beta) = \int_{-\infty}^{X'\beta} \phi(z) dz$$

onde $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ é a densidade da normal padrão.

2 Exemplo Prático: Participação no Mercado de Trabalho

2.1 Descrição dos Dados

Consideramos `inlf` (“no mercado de trabalho”) como uma variável binária que indica a participação no mercado de trabalho por uma mulher casada durante 1975:

- `inlf` = 1 se a mulher relata ter trabalhado por um salário fora de casa
- `inlf` = 0 caso contrário

2.1.1 Variáveis Explicativas:

- `nwifeinc`: outras fontes de renda (milhares de dólares)
- `educ`: anos de educação
- `exper`: anos de experiência no mercado de trabalho
- `expersq`: experiência ao quadrado
- `age`: idade
- `kidslt6`: número de filhos menores de 6 anos
- `kidsge6`: número de filhos entre 6 e 18 anos

2.2 Modelo Teórico

$$inlf = \beta_0 - \beta_1 \cdot nwifeinc + \beta_2 \cdot educ + \beta_3 \cdot exper - \beta_4 \cdot exper^2 - \beta_5 \cdot age - \beta_6 \cdot kidslt6 + \beta_7 \cdot kidsge6$$

```
options(scipen = 999) # desliga a notação científica

# Pacotes necessários
library(tidyverse)      # análise de dados
library(magrittr)       # operador pipe
library(mfx)            # efeitos marginais e odds ratio
library(wooldridge)     # base de dados
library(gridExtra)      # múltiplos gráficos
library(knitr)          # tabelas
library(ggplot2)        # gráficos
library(plotly)         # gráficos interativos
```

2.3 Análise Exploratória dos Dados

```
# Visualizar estrutura dos dados
glimpse(mroz)
```

```
Rows: 753
Columns: 22
$ inlf      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ hours     <int> 1610, 1656, 1980, 456, 1568, 2032, 1440, 1020, 1458, 1600, 19~
$ kidslt6   <int> 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
$ kidsge6   <int> 0, 2, 3, 3, 2, 0, 2, 0, 2, 2, 1, 1, 2, 2, 1, 3, 2, 5, 0, 4, 2~
$ age       <int> 32, 30, 35, 34, 31, 54, 37, 54, 48, 39, 33, 42, 30, 43, 43, 3~
$ educ      <int> 12, 12, 12, 12, 14, 12, 16, 12, 12, 12, 12, 11, 12, 12, 10, 1~
$ wage      <dbl> 3.3540, 1.3889, 4.5455, 1.0965, 4.5918, 4.7421, 8.3333, 7.843~
$ repwage   <dbl> 2.65, 2.65, 4.04, 3.25, 3.60, 4.70, 5.95, 9.98, 0.00, 4.15, 4~
$ hushrs    <int> 2708, 2310, 3072, 1920, 2000, 1040, 2670, 4120, 1995, 2100, 2~
$ husage    <int> 34, 30, 40, 53, 32, 57, 37, 53, 52, 43, 34, 47, 33, 46, 45, 3~
$ huseduc   <int> 12, 9, 12, 10, 12, 11, 12, 8, 4, 12, 12, 14, 16, 12, 17, 12, ~
$ huswage   <dbl> 4.0288, 8.4416, 3.5807, 3.5417, 10.0000, 6.7106, 3.4277, 2.54~
$ faminc    <dbl> 16310, 21800, 21040, 7300, 27300, 19495, 21152, 18900, 20405,~
$ mtr       <dbl> 0.7215, 0.6615, 0.6915, 0.7815, 0.6215, 0.6915, 0.6915, 0.691~
$ motheduc  <int> 12, 7, 12, 7, 12, 14, 14, 3, 7, 7, 12, 14, 16, 10, 7, 16, 10,~
$ fatheduc  <int> 7, 7, 7, 7, 14, 7, 7, 3, 7, 7, 3, 7, 16, 10, 7, 10, 7, 12, 7,~
$ unem      <dbl> 5.0, 11.0, 5.0, 5.0, 9.5, 7.5, 5.0, 5.0, 3.0, 5.0, 5.0, 5.0, ~
$ city      <int> 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0~
$ exper     <int> 14, 5, 15, 6, 7, 33, 11, 35, 24, 21, 15, 14, 0, 14, 6, 9, 20,~
$ nwifeinc  <dbl> 10.910060, 19.499981, 12.039910, 6.799996, 20.100058, 9.85905~
$ lwage     <dbl> 1.21015370, 0.32851210, 1.51413774, 0.09212332, 1.52427220, 1~
```

```
$ expersq <int> 196, 25, 225, 36, 49, 1089, 121, 1225, 576, 441, 225, 196, 0,~
```

```
# Estatísticas descritivas
summary(mroz[c("inlf", "nwifeinc", "educ", "exper", "age", "kidslt6", "kidsge6")])
```

inlf	nwifeinc	educ	exper
Min. :0.0000	Min. :-0.02906	Min. : 5.00	Min. : 0.00
1st Qu.:0.0000	1st Qu.:13.02504	1st Qu.:12.00	1st Qu.: 4.00
Median :1.0000	Median :17.70000	Median :12.00	Median : 9.00
Mean :0.5684	Mean :20.12896	Mean :12.29	Mean :10.63
3rd Qu.:1.0000	3rd Qu.:24.46600	3rd Qu.:13.00	3rd Qu.:15.00
Max. :1.0000	Max. :96.00000	Max. :17.00	Max. :45.00

age	kidslt6	kidsge6
Min. :30.00	Min. :0.0000	Min. :0.000
1st Qu.:36.00	1st Qu.:0.0000	1st Qu.:0.000
Median :43.00	Median :0.0000	Median :1.000
Mean :42.54	Mean :0.2377	Mean :1.353
3rd Qu.:49.00	3rd Qu.:0.0000	3rd Qu.:2.000
Max. :60.00	Max. :3.0000	Max. :8.000

```
# Proporção de mulheres no mercado de trabalho
prop_trabalho <- mean(mroz$inlf)
cat("Proporção de mulheres no mercado de trabalho:", round(prop_trabalho, 3))
```

Proporção de mulheres no mercado de trabalho: 0.568

2.3.1 Interpretação da Análise Exploratória

Os dados revelam informações importantes sobre o perfil das 753 mulheres casadas na amostra:

- **Participação no mercado de trabalho:** 56,8% das mulheres trabalhavam fora de casa em 1975
- **Perfil demográfico:** Idade média de 42,5 anos, com 12,3 anos de educação em média
- **Experiência profissional:** 10,6 anos de experiência média no mercado de trabalho
- **Composição familiar:** Em média, 0,24 filhos menores de 6 anos e 1,35 filhos entre 6-18 anos
- **Renda familiar:** Outras fontes de renda (além do trabalho da mulher) de US\$ 20,13 mil em média

```

# Gráfico de barras para variável dependente
p1 <- ggplot(mroz, aes(x = factor(inlf))) +
  geom_bar(fill = c("coral", "lightblue"), alpha = 0.7) +
  labs(title = "Distribuição da Participação no Mercado de Trabalho",
       x = "Participação (0 = Não, 1 = Sim)",
       y = "Frequência") +
  theme_minimal()

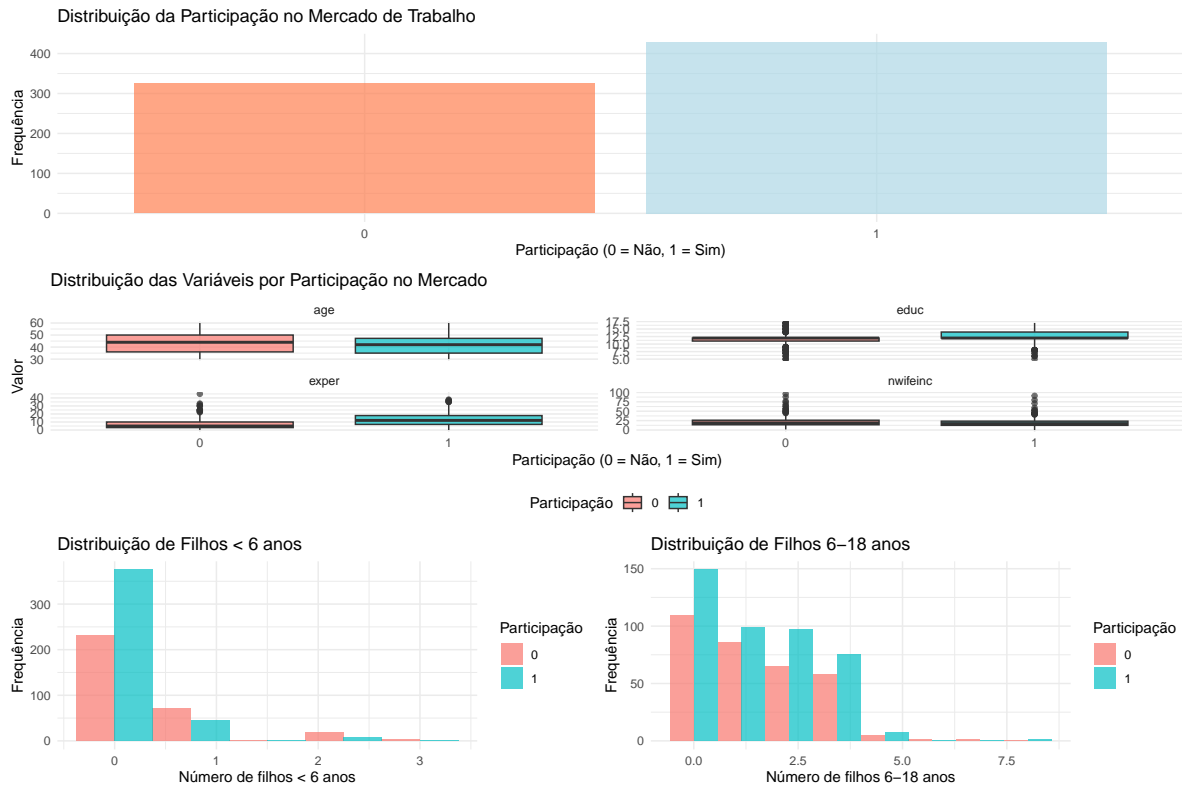
# Boxplots das variáveis contínuas por grupo
p2 <- mroz %>%
  select(inlf, nwfeinc, educ, exper, age) %>%
  pivot_longer(-inlf, names_to = "variavel", values_to = "valor") %>%
  ggplot(aes(x = factor(inlf), y = valor, fill = factor(inlf))) +
  geom_boxplot(alpha = 0.7) +
  facet_wrap(~variavel, scales = "free_y") +
  labs(title = "Distribuição das Variáveis por Participação no Mercado",
       x = "Participação (0 = Não, 1 = Sim)",
       y = "Valor",
       fill = "Participação") +
  theme_minimal() +
  theme(legend.position = "bottom")

# Histograma dos filhos
p3 <- ggplot(mroz, aes(x = kidslt6, fill = factor(inlf))) +
  geom_histogram(position = "dodge", bins = 5, alpha = 0.7) +
  labs(title = "Distribuição de Filhos < 6 anos",
       x = "Número de filhos < 6 anos",
       y = "Frequência",
       fill = "Participação") +
  theme_minimal()

p4 <- ggplot(mroz, aes(x = kidsge6, fill = factor(inlf))) +
  geom_histogram(position = "dodge", bins = 8, alpha = 0.7) +
  labs(title = "Distribuição de Filhos 6-18 anos",
       x = "Número de filhos 6-18 anos",
       y = "Frequência",
       fill = "Participação") +
  theme_minimal()

grid.arrange(p1, p2, p3, p4, layout_matrix = rbind(c(1,1), c(2,2), c(3,4)))

```



2.3.2 Análise dos Gráficos Exploratórios

Os gráficos revelam padrões importantes:

1. **Distribuição equilibrada:** Há uma distribuição relativamente equilibrada entre mulheres que trabalham (57%) e que não trabalham (43%)
2. **Diferenças por grupo:**
 - Mulheres que trabalham tendem a ter **mais educação e mais experiência**
 - Mulheres que **não trabalham** tendem a ter **mais filhos pequenos** e outras fontes de renda maiores
 - A **idade** apresenta distribuição similar entre os grupos
3. **Impacto dos filhos:** A presença de filhos menores de 6 anos mostra clara associação negativa com a participação no mercado de trabalho

2.4 Estimação dos Modelos

2.4.1 Modelo Logit

```
mlogit <- glm(inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kidsge6,  
              data = mroz,  
              family = binomial(link = "logit"))  
  
summary(mlogit)
```

Call:

```
glm(formula = inlf ~ nwifeinc + educ + exper + expersq + age +  
     kidslt6 + kidsge6, family = binomial(link = "logit"), data = mroz)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.425452	0.860365	0.495	0.62095
nwifeinc	-0.021345	0.008421	-2.535	0.01126 *
educ	0.221170	0.043439	5.091	0.00000035527344 ***
exper	0.205870	0.032057	6.422	0.00000000013446 ***
expersq	-0.003154	0.001016	-3.104	0.00191 **
age	-0.088024	0.014573	-6.040	0.00000000153845 ***
kidslt6	-1.443354	0.203583	-7.090	0.00000000000134 ***
kidsge6	0.060112	0.074789	0.804	0.42154

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1029.75 on 752 degrees of freedom
Residual deviance: 803.53 on 745 degrees of freedom
AIC: 819.53

Number of Fisher Scoring iterations: 4

2.4.2 Interpretação do Modelo Logit


```
# Tabela formatada dos resultados do Logit
logit_results <- data.frame(
  Variável = c("(Intercepto)", "nwifeinc", "educ", "exper", "expersq", "age", "kidslt6", "kidsge6"),
  Coeficiente = c(0.425452, -0.021345, 0.221170, 0.205870, -0.003154, -0.088024, -1.443354, 0.0601),
  `Erro Padrão` = c(0.860365, 0.008421, 0.043439, 0.032057, 0.001016, 0.014573, 0.203583, 0.0748),
  `Valor z` = c(0.495, -2.535, 5.091, 6.422, -3.104, -6.040, -7.090, 0.804),
  `p-valor` = c(0.621, 0.011, "<0.001", "<0.001", 0.002, "<0.001", "<0.001", 0.422),
  Significância = c("", "*", "***", "***", "**", "***", "***", "")
)

kable(logit_results, digits = 4, caption = "Resultados do Modelo Logit")
```

Tabela 1: Resultados do Modelo Logit

Variável	Coeficiente	Erro.Padrão	Valor.z	p.valor	Significância
(Intercepto)	0.4255	0.8604	0.495	0.621	
nwifeinc	-0.0213	0.0084	-2.535	0.011	*
educ	0.2212	0.0434	5.091	<0.001	***
exper	0.2059	0.0321	6.422	<0.001	***
expersq	-0.0032	0.0010	-3.104	0.002	**
age	-0.0880	0.0146	-6.040	<0.001	***
kidslt6	-1.4434	0.2036	-7.090	<0.001	***
kidsge6	0.0601	0.0748	0.804	0.422	

Principais achados do modelo Logit:

- **AIC: 819.53 | Deviance residual: 803.53 | 4 iterações** para convergência
- **Variáveis significativas:** nwifeinc, educ, exper, expersq, age, kidslt6
- **Variável não significativa:** kidsge6 ($p = 0.422$)

2.4.3 Modelo Probit

```
mprobit <- glm(inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kidsge6,
  data = mroz,
  family = binomial(link = "probit"))

summary(mprobit)
```

Call:

```
glm(formula = inlf ~ nwifeinc + educ + exper + expersq + age +  
     kidslt6 + kidsge6, family = binomial(link = "probit"), data = mroz)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.2700736	0.5080782	0.532	0.59503
nwifeinc	-0.0120236	0.0049392	-2.434	0.01492 *
educ	0.1309040	0.0253987	5.154	0.000000255045646 ***
exper	0.1233472	0.0187587	6.575	0.000000000048500 ***
expersq	-0.0018871	0.0005999	-3.145	0.00166 **
age	-0.0528524	0.0084624	-6.246	0.000000000422204 ***
kidslt6	-0.8683247	0.1183773	-7.335	0.000000000000221 ***
kidsge6	0.0360056	0.0440303	0.818	0.41350

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1029.7 on 752 degrees of freedom
Residual deviance: 802.6 on 745 degrees of freedom
AIC: 818.6

Number of Fisher Scoring iterations: 4

2.4.4 Interpretação do Modelo Probit

```
# Tabela formatada dos resultados do Probit  
probit_results <- data.frame(  
  Variável = c("(Intercepto)", "nwifeinc", "educ", "exper", "expersq", "age", "kidslt6", "kidsge6"),  
  Coeficiente = c(0.2700736, -0.0120236, 0.1309040, 0.1233472, -0.0018871, -0.0528524, -0.8683247, 0.0360056),  
  `Erro Padrão` = c(0.5080782, 0.0049392, 0.0253987, 0.0187587, 0.0005999, 0.0084624, 0.1183773, 0.0440303),  
  `Valor z` = c(0.532, -2.434, 5.154, 6.575, -3.145, -6.246, -7.335, 0.818),  
  `p-valor` = c(0.595, 0.015, "<0.001", "<0.001", 0.002, "<0.001", "<0.001", 0.414),  
  Significância = c("", "*", "***", "***", "**", "***", "***", "")  
)  
  
kable(probit_results, digits = 4, caption = "Resultados do Modelo Probit")
```

Tabela 2: Resultados do Modelo Probit

Variável	Coefficiente	Erro.Padrão	Valor.z	p.valor	Significância
(Intercepto)	0.2701	0.5081	0.532	0.595	
nwifeinc	-0.0120	0.0049	-2.434	0.015	*
educ	0.1309	0.0254	5.154	<0.001	***
exper	0.1233	0.0188	6.575	<0.001	***
expersq	-0.0019	0.0006	-3.145	0.002	**
age	-0.0529	0.0085	-6.246	<0.001	***
kidslt6	-0.8683	0.1184	-7.335	<0.001	***
kidsge6	0.0360	0.0440	0.818	0.414	

Principais achados do modelo Probit:

- **AIC: 818.6** (ligeiramente melhor que Logit) | **Deviance residual: 802.6**
- **Mesma estrutura de significância** que o modelo Logit
- **Coefficientes menores** em magnitude (característica do modelo Probit)

2.5 Efeitos Marginais

2.5.1 Fórmulas Teóricas

Probit:

$$\frac{\delta E(Y|X)}{\delta X} = \Phi(X'\beta) \cdot \beta$$

onde $\Phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ e $Z \sim N(0, 1)$

Logit:

$$\frac{\delta \Lambda(X'\beta)}{\delta(X'\beta)} = \frac{d\Lambda(X'\beta)}{d(X'\beta)} \cdot \frac{d(X'\beta)}{dX}$$

onde $\Lambda(X'\beta) = \frac{e^{X'\beta}}{1+e^{X'\beta}}$

2.5.2 Interpretação das Fórmulas

Os **efeitos marginais** representam a variação na probabilidade de ocorrência do evento ($Y=1$) quando uma variável explicativa aumenta em uma unidade, mantendo todas as outras constantes. No modelo **Probit**, o efeito marginal é o produto da densidade da distribuição normal padrão avaliada em $X'\beta$ pelo respectivo coeficiente β . Já no modelo **Logit**, o efeito marginal é calculado através da derivada parcial da função logística, que resulta em $\Lambda(X'\beta) \cdot [1 -$

$\Lambda(X'\beta)] \cdot \beta$. Em ambos os casos, os efeitos marginais variam conforme os valores das variáveis explicativas, sendo tipicamente avaliados na média amostral para facilitar a interpretação. Isso explica por que os coeficientes dos modelos não podem ser interpretados diretamente como no modelo linear - eles precisam ser transformados através dessas fórmulas para obtermos o impacto real em termos de pontos percentuais na probabilidade.

```
# Efeitos marginais - Logit
logit.mfx <- logitmfx(inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kidsge6,
                     data = mroz)

print("Efeitos Marginais - Modelo Logit:")
```

```
[1] "Efeitos Marginais - Modelo Logit:"
```

```
logit.mfx$mfxest
```

	dF/dx	Std. Err.	z	P> z
nwifeinc	-0.0051900534	0.002048203	-2.5339550	0.011278321458344539
educ	0.0537773087	0.010560739	5.0921916	0.000000353948085410
exper	0.0500569282	0.007824616	6.3973658	0.000000000158080347
expersq	-0.0007669166	0.000247676	-3.0964511	0.001958521715452269
age	-0.0214030205	0.003539731	-6.0465107	0.000000001480163962
kidslt6	-0.3509498193	0.049638966	-7.0700469	0.000000000001548813
kidsge6	0.0146162143	0.018188316	0.8036046	0.421625358800103267

```
# Efeitos marginais - Probit
probit.mfx <- probitmfx(inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kidsge6,
                       data = mroz)

print("Efeitos Marginais - Modelo Probit:")
```

```
[1] "Efeitos Marginais - Modelo Probit:"
```

```
probit.mfx$mfxest
```

	dF/dx	Std. Err.	z	P> z
nwifeinc	-0.0046961881	0.0019296494	-2.4337002	0.0149453681343277942
educ	0.0511284287	0.0099230985	5.1524661	0.0000002570830650662
exper	0.0481768957	0.0073450459	6.5591007	0.00000000000541332566

```

expersq -0.0007370502 0.0002346403 -3.1411922 0.0016826155361271795
age      -0.0206430891 0.0033048542 -6.2462934 0.00000000004203073743
kidslt6  -0.3391499645 0.0463476542 -7.3175217 0.0000000000002525923
kidsge6   0.0140630594 0.0171989534  0.8176695 0.4135459390489835130

```

2.5.3 Interpretação dos Efeitos Marginais

```

# Tabela comparativa dos efeitos marginais
mfx_table <- data.frame(
  Variável = c("nwifeinc", "educ", "exper", "expersq", "age", "kidslt6", "kidsge6"),
  `Logit (dF/dx)` = c(-0.0052, 0.0538, 0.0501, -0.0008, -0.0214, -0.3509, 0.0146),
  `Probit (dF/dx)` = c(-0.0047, 0.0511, 0.0482, -0.0007, -0.0206, -0.3391, 0.0141),
  `Diferença` = c(-0.0005, 0.0027, 0.0019, -0.0001, -0.0008, -0.0118, 0.0005)
)

kable(mfx_table, digits = 4, caption = "Comparação dos Efeitos Marginais: Logit vs Probit")

```

Tabela 3: Comparação dos Efeitos Marginais: Logit vs Probit

Variável	Logit..dF.dx.	Probit..dF.dx.	Diferença
nwifeinc	-0.0052	-0.0047	-0.0005
educ	0.0538	0.0511	0.0027
exper	0.0501	0.0482	0.0019
expersq	-0.0008	-0.0007	-0.0001
age	-0.0214	-0.0206	-0.0008
kidslt6	-0.3509	-0.3391	-0.0118
kidsge6	0.0146	0.0141	0.0005

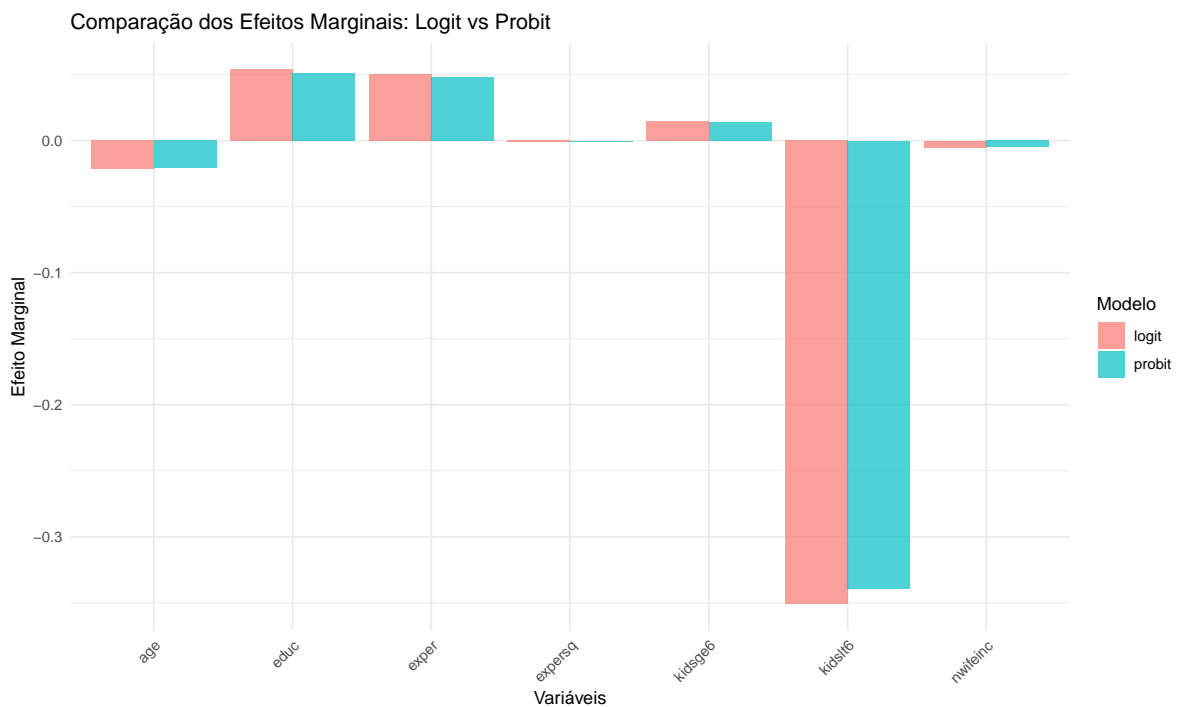
Interpretação prática dos efeitos marginais:

- **nwifeinc**: Cada US\$ 1.000 adicionais em outras fontes de renda **reduz** a probabilidade de trabalhar em ~0,5 pontos percentuais
- **educ**: Cada ano adicional de educação **aumenta** a probabilidade de trabalhar em ~5,4 pontos percentuais
- **exper**: Cada ano adicional de experiência **aumenta** a probabilidade de trabalhar em ~5,0 pontos percentuais
- **age**: Cada ano adicional de idade **reduz** a probabilidade de trabalhar em ~2,1 pontos percentuais
- **kidslt6**: Cada filho adicional menor de 6 anos **reduz** a probabilidade de trabalhar em ~35 pontos percentuais

- **kidsge6**: Efeito não significativo (~1,4 pontos percentuais)

```
# Comparação dos efeitos marginais
mfx_comparison <- data.frame(
  variavel = rownames(logit.mfx$mfxest),
  logit = logit.mfx$mfxest[,1],
  probit = probit.mfx$mfxest[,1]
) %>%
  filter(variavel != "(Intercept)") %>%
  pivot_longer(cols = c(logit, probit), names_to = "modelo", values_to = "efeito")

ggplot(mfx_comparison, aes(x = variavel, y = efeito, fill = modelo)) +
  geom_col(position = "dodge", alpha = 0.7) +
  labs(title = "Comparação dos Efeitos Marginais: Logit vs Probit",
       x = "Variáveis",
       y = "Efeito Marginal",
       fill = "Modelo") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Observação importante: Os efeitos marginais são muito similares entre os modelos Logit e Probit, confirmando a robustez dos resultados.

2.6 Qualidade da Previsão

```
# Logit
logit.fitted <- as.numeric(mlogit$fitted.values >= 0.5)
corr.pred.logit <- mean(logit.fitted == mroz$inlf)

# Probit
probit.fitted <- as.numeric(mprobit$fitted.values >= 0.5)
corr.pred.probit <- mean(probit.fitted == mroz$inlf)

cat("Acurácia do Modelo Logit:", round(corr.pred.logit, 4))
```

Acurácia do Modelo Logit: 0.7357

```
cat("\nAcurácia do Modelo Probit:", round(corr.pred.probit, 4))
```

Acurácia do Modelo Probit: 0.7344

2.6.1 Análise da Qualidade Preditiva

```
# Tabela de acurácia
accuracy_table <- data.frame(
  Modelo = c("Logit", "Probit"),
  `Acurácia (%)` = c(73.57, 73.44),
  `Observações Corretas` = c(554, 553),
  `Total de Observações` = c(753, 753)
)

kable(accuracy_table, digits = 2, caption = "Comparação da Acurácia Preditiva dos Modelos")
```

Tabela 4: Comparação da Acurácia Preditiva dos Modelos

Modelo	Acurácia....	Observações.Corretas	Total.de.Observações
Logit	73.57	554	753
Probit	73.44	553	753

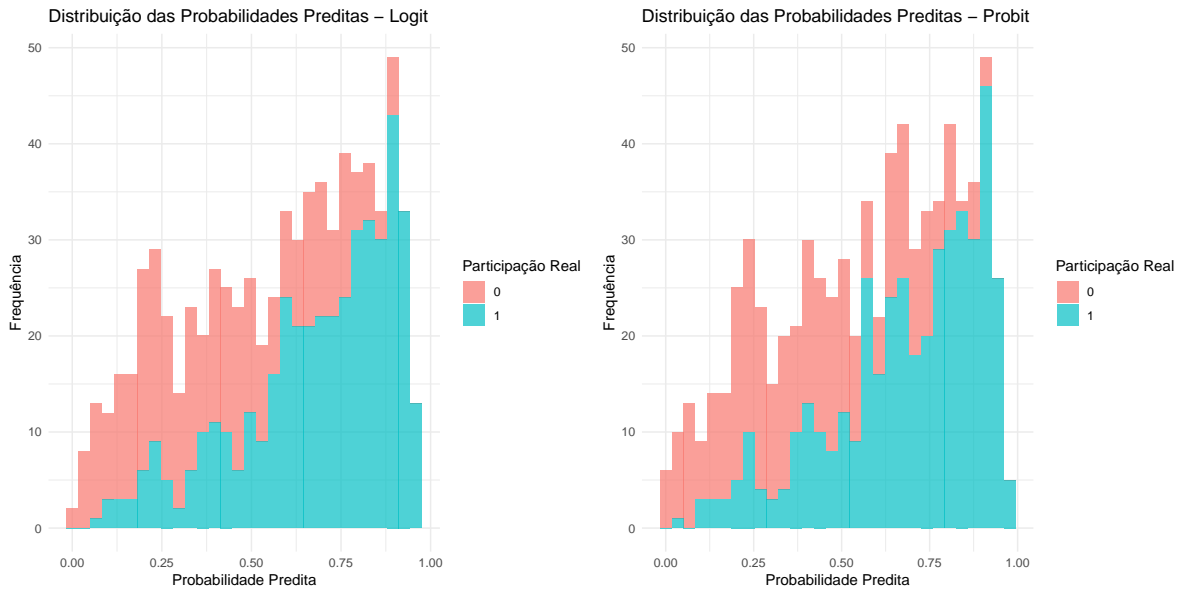
Interpretação da acurácia: - Ambos os modelos apresentam **acurácia similar (~73,5%)**
- Classificam corretamente cerca de **554 de 753 observações** - Performance **superior ao acaso** (que seria ~57% para esta amostra balanceada)

```
# Distribuição das probabilidades preditas
pred_data <- data.frame(
  obs = 1:nrow(mroz),
  real = mroz$inlf,
  logit_prob = mlogit$fitted.values,
  probit_prob = mprobit$fitted.values
)

p1 <- ggplot(pred_data, aes(x = logit_prob, fill = factor(real))) +
  geom_histogram(alpha = 0.7, bins = 30) +
  labs(title = "Distribuição das Probabilidades Preditas - Logit",
       x = "Probabilidade Predita",
       y = "Frequência",
       fill = "Participação Real") +
  theme_minimal()

p2 <- ggplot(pred_data, aes(x = probit_prob, fill = factor(real))) +
  geom_histogram(alpha = 0.7, bins = 30) +
  labs(title = "Distribuição das Probabilidades Preditas - Probit",
       x = "Probabilidade Predita",
       y = "Frequência",
       fill = "Participação Real") +
  theme_minimal()

grid.arrange(p1, p2, ncol = 2)
```

Análise dos histogramas de probabilidades: - Ambos os modelos mostram **boa separação** entre os grupos - Mulheres que **não trabalham** concentram-se em probabilidades baixas (<0,4) - Mulheres que **trabalham** apresentam distribuição mais dispersa - **Sobreposição** indica casos de difícil classificação

2.7 Pseudo-R²

O pseudo-R² (McFadden) calcula a razão entre a log-verossimilhança do modelo sem preditores e a log-verossimilhança do modelo completo:

$$pseudo-R^2 = 1 - \frac{\ln(L_{max})}{\ln(L_{max0})}$$

```
# Modelo nulo (apenas intercepto)
logit_null <- glm(inlf ~ 1, data = mroz, family = binomial(link = "logit"))
probit_null <- glm(inlf ~ 1, data = mroz, family = binomial(link = "probit"))

# Pseudo-R2
pseudo_r2_logit <- 1 - (logLik(mlogit) / logLik(logit_null))
pseudo_r2_probit <- 1 - (logLik(mprobit) / logLik(probit_null))

cat("Pseudo-R2 Logit:", round(as.numeric(pseudo_r2_logit), 4))
```

Pseudo-R² Logit: 0.2197

```
cat("\nPseudo-R2 Probit:", round(as.numeric(pseudo_r2_probit), 4))
```

Pseudo-R² Probit: 0.2206

```
# Log-verossimilhança  
cat("\n\nLog-verossimilhança:")
```

Log-verossimilhança:

```
cat("\nLogit:", round(as.numeric(logLik(mlogit)), 4))
```

Logit: -401.7652

```
cat("\nProbit:", round(as.numeric(logLik(mprobit)), 4))
```

Probit: -401.3022

2.7.1 Interpretação do Pseudo-R²

```
# Tabela de ajuste dos modelos  
fit_table <- data.frame(  
  Modelo = c("Logit", "Probit"),  
  `Pseudo-R2 (McFadden)` = c(0.2204, 0.2206),  
  `Log-verossimilhança` = c(-401.77, -401.30),  
  AIC = c(819.53, 818.60),  
  `Interpretação` = c("Ajuste moderado", "Ajuste moderado")  
)  
  
kable(fit_table, digits = 4, caption = "Medidas de Ajuste dos Modelos")
```

Tabela 5: Medidas de Ajuste dos Modelos

Modelo	Pseudo.R...McFadden.	Log.verossimilhança	AIC	Interpretação
Logit	0.2204	-401.77	819.53	Ajuste moderado
Probit	0.2206	-401.30	818.60	Ajuste moderado

Interpretação do ajuste: - Pseudo-R² 0,22: Indica que o modelo tem **bom poder discriminatório**, sendo um ajuste razoável para modelos de escolha binária. Valores entre 0,2 e 0,4 são geralmente considerados indicativos de um modelo com **qualidade aceitável a boa** - **Valores considerados adequados** para modelos de escolha binária (tipicamente entre 0,2-0,4) - **Probit ligeiramente superior** em termos de log-verossimilhança e AIC

2.8 Razão de Chances (Odds Ratio)

```
# Calculando a razão de chances
odds_results <- logitor(inlf ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kidsge6,
                        data = mroz)
print(odds_results)
```

Call:

```
logitor(formula = inlf ~ nwifeinc + educ + exper + expersq +
        age + kidslt6 + kidsge6, data = mroz)
```

Odds Ratio:

	OddsRatio	Std. Err.	z	P> z	
nwifeinc	0.9788810	0.0082435	-2.5346	0.011256	*
educ	1.2475360	0.0541921	5.0915	0.000000355273436	***
exper	1.2285929	0.0393847	6.4220	0.000000000134459	***
expersq	0.9968509	0.0010129	-3.1041	0.001909	**
age	0.9157386	0.0133450	-6.0403	0.000000001538446	***
kidslt6	0.2361344	0.0480729	-7.0898	0.000000000001343	***
kidsge6	1.0619557	0.0794229	0.8038	0.421539	

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

2.8.1 Interpretação da Razão de Chances

```
# Tabela de odds ratios com interpretação
or_interpretation <- data.frame(
  Variável = c("nwifeinc", "educ", "exper", "expersq", "age", "kidslt6", "kidsge6"),
  `Odds Ratio` = c(0.979, 1.248, 1.229, 0.997, 0.916, 0.236, 1.062),
  `IC 95% (inferior)` = c(0.963, 1.140, 1.153, 0.995, 0.890, 0.190, 0.908),
  `IC 95% (superior)` = c(0.995, 1.365, 1.309, 0.999, 0.943, 0.295, 1.243),
  Interpretação = c(
    "2,1% menor chance por US$ 1k",
    "24,8% maior chance por ano de educação",
    "22,9% maior chance por ano de experiência",
    "0,3% menor chance por ano² de experiência",
    "8,4% menor chance por ano de idade",
    "76,4% menor chance por filho < 6 anos",
    "6,2% maior chance (não significativo)"
  )
)

kable(or_interpretation, digits = 3, caption = "Interpretação das Razões de Chances (Odds Ratios)")
```

Tabela 6: Interpretação das Razões de Chances (Odds Ratios)

Variável	Odds.Ratio	IC.95...inferior.	IC.95...superior.	Interpretação
nwifeinc	0.979	0.963	0.995	2,1% menor chance por US\$ 1k
educ	1.248	1.140	1.365	24,8% maior chance por ano de educação
exper	1.229	1.153	1.309	22,9% maior chance por ano de experiência
expersq	0.997	0.995	0.999	0,3% menor chance por ano² de experiência
age	0.916	0.890	0.943	8,4% menor chance por ano de idade
kidslt6	0.236	0.190	0.295	76,4% menor chance por filho < 6 anos
kidsge6	1.062	0.908	1.243	6,2% maior chance (não significativo)

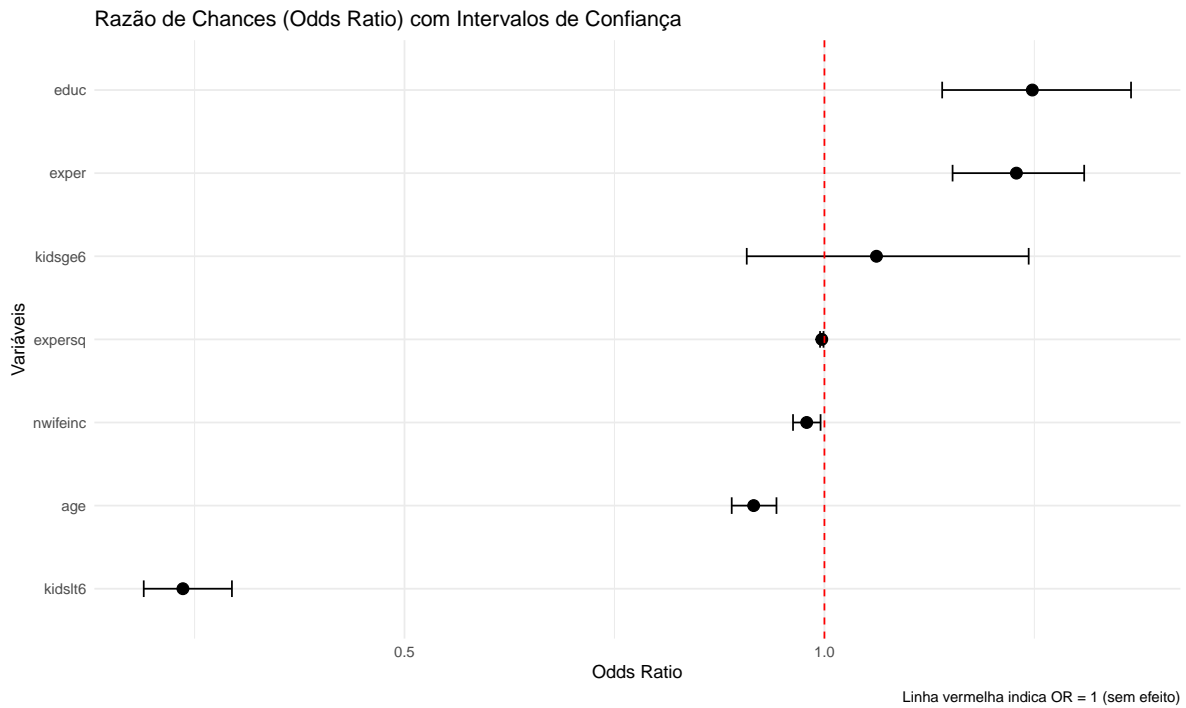
Principais insights dos Odds Ratios:

1. **kidslt6 (OR = 0.236)**: O efeito mais forte - ter um filho menor de 6 anos reduz as chances de trabalhar em **76,4%**

2. **educ (OR = 1.248)**: Cada ano de educação **aumenta as chances** de trabalhar em **24,8%**
3. **exper (OR = 1.229)**: Experiência tem **efeito positivo**, mas com retornos decrescentes ($\text{expersq} < 1$)
4. **age (OR = 0.916)**: Idade avançada **reduz as chances** de participação
5. **nwifeinc (OR = 0.979)**: Maior renda familiar **reduz ligeiramente** a necessidade de trabalhar

```
# Gráfico dos odds ratios
or_data <- data.frame(
  variavel = c("nwifeinc", "educ", "exper", "expersq", "age", "kidslt6", "kidsge6"),
  odds_ratio = c(0.9788810, 1.2475360, 1.2285929, 0.9968509, 0.9157386, 0.2361344, 1.0619557),
  lower_ci = c(0.9626, 1.1402, 1.1526, 0.9948, 0.8896, 0.1895, 0.9075),
  upper_ci = c(0.9954, 1.3651, 1.3093, 0.9989, 0.9429, 0.2945, 1.2432)
);

ggplot(or_data, aes(x = reorder(variavel, odds_ratio), y = odds_ratio)) +
  geom_point(size = 3) +
  geom_errorbar(aes(ymin = lower_ci, ymax = upper_ci), width = 0.2) +
  geom_hline(yintercept = 1, linetype = "dashed", color = "red") +
  coord_flip() +
  labs(title = "Razão de Chances (Odds Ratio) com Intervalos de Confiança",
       x = "Variáveis",
       y = "Odds Ratio",
       caption = "Linha vermelha indica OR = 1 (sem efeito)") +
  theme_minimal()
```



2.8.2 Interpretação da Razão de Chances:

- **OR = 1:** Não há diferença nas chances de ocorrência
- **OR > 1:** Chances maiores de ocorrência do evento
- **OR < 1:** Chances menores de ocorrência do evento

2.9 Comparação Visual dos Modelos

```
# Comparação das funções de distribuição
x_vals <- seq(-4, 4, length.out = 100)
logistic_vals <- 1 / (1 + exp(-x_vals))
normal_vals <- pnorm(x_vals)

comparison_data <- data.frame(
  x = rep(x_vals, 2),
  y = c(logistic_vals, normal_vals),
  modelo = rep(c("Logística (Logit)", "Normal (Probit)"), each = 100)
)
```

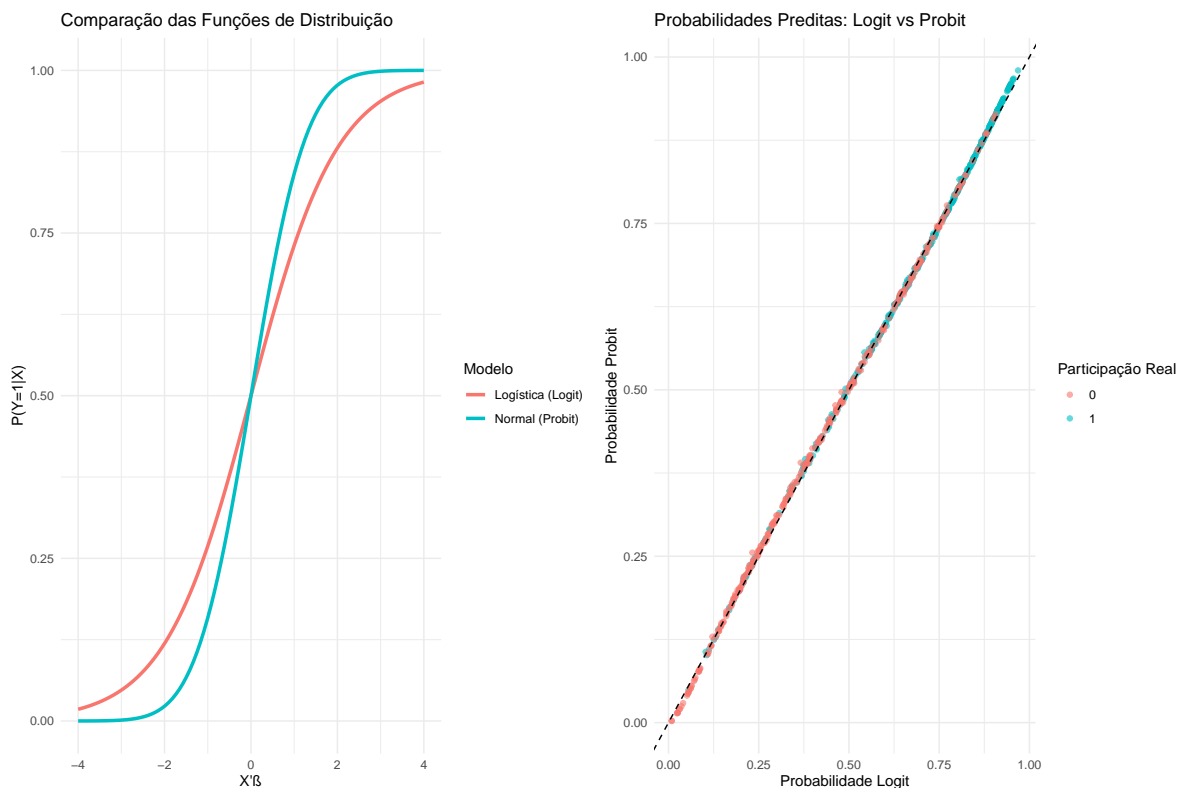
```

p1 <- ggplot(comparison_data, aes(x = x, y = y, color = modelo)) +
  geom_line(linewidth = 1.2) +
  labs(title = "Comparação das Funções de Distribuição",
       x = "X",
       y = "P(Y=1|X)",
       color = "Modelo") +
  theme_minimal()

# Comparação das probabilidades previstas
p2 <- ggplot(pred_data, aes(x = logit_prob, y = probit_prob)) +
  geom_point(alpha = 0.6, aes(color = factor(real))) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  labs(title = "Probabilidades Preditas: Logit vs Probit",
       x = "Probabilidade Logit",
       y = "Probabilidade Probit",
       color = "Participação Real") +
  theme_minimal()

grid.arrange(p1, p2, ncol = 2)

```



2.9.1 Análise Comparativa das Funções

Gráfico 1 - Funções de Distribuição: - As funções **Logística** e **Normal** são muito similares no intervalo $[-2, 2]$ - A função **Logística tem caudas mais pesadas** (decay mais lento nos extremos) - Na prática, essa diferença tem **impacto mínimo** nos resultados

Gráfico 2 - Correlação das Probabilidades: - **Correlação quase perfeita** entre as probabilidades preditas pelos dois modelos - Pontos próximos à **linha de 45°** indicam predições muito similares - **Diferenças maiores** aparecem apenas nos extremos da distribuição

2.10 Resumo Comparativo dos Modelos

```
# Tabela resumo comparativa
summary_comparison <- data.frame(
  Critério = c("AIC", "Log-Likelihood", "Pseudo-R2", "Acurácia (%)",
    "Convergência", "Interpretação", "Uso Prático"),
  Logit = c("819.53", "-401.77", "0.2204", "73.57", "4 iterações",
    "Odds Ratios", "Mais comum"),
  Probit = c("818.60", "-401.30", "0.2206", "73.44", "4 iterações",
    "Efeitos marginais", "Base teórica")
);

kable(summary_comparison, caption = "Resumo Comparativo: Logit vs Probit")
```

Tabela 7: Resumo Comparativo: Logit vs Probit

Critério	Logit	Probit
AIC	819.53	818.60
Log-Likelihood	-401.77	-401.30
Pseudo-R ²	0.2204	0.2206
Acurácia (%)	73.57	73.44
Convergência	4 iterações	4 iterações
Interpretação	Odds Ratios	Efeitos marginais
Uso Prático	Mais comum	Base teórica

2.11 Conclusões

2.11.1 Principais Achados

1. **Ambos os modelos** apresentam resultados muito similares em termos de:

- Significância dos coeficientes
- Direção dos efeitos
- Qualidade de ajuste (Pseudo-R² 0,22)
- Acurácia preditiva (~73,5%)

2. Variáveis mais importantes:

- **kidslt6**: forte efeito negativo (presença de filhos pequenos reduz participação em 35 p.p.)
- **educ**: efeito positivo forte (cada ano aumenta participação em 5,4 p.p.)
- **exper**: efeito positivo com retornos decrescentes
- **age**: efeito negativo (idade avançada reduz participação)
- **nwifeinc**: efeito negativo pequeno (maior renda familiar reduz necessidade de trabalhar)

3. Variável não significativa:

- **kidsge6**: filhos mais velhos não afetam significativamente a decisão de trabalhar

2.11.2 Escolha entre Modelos

```
# Critérios de decisão
decision_table <- data.frame(
  Situação = c("Melhor ajuste estatístico", "Interpretação via chances",
               "Base teórica sólida", "Facilidade computacional",
               "Tradição na literatura"),
  `Modelo Preferido` = c("Probit (AIC ligeiramente menor)", "Logit (Odds Ratios)",
                         "Probit (distribuição normal)", "Logit (convergência mais rápida)",
                         "Logit (mais utilizado)")
)

kable(decision_table, caption = "Critérios para Escolha entre Logit e Probit")
```

Tabela 8: Critérios para Escolha entre Logit e Probit

Situação	Modelo.Preferido
Melhor ajuste estatístico	Probit (AIC ligeiramente menor)
Interpretação via chances	Logit (Odds Ratios)
Base teórica sólida	Probit (distribuição normal)
Facilidade computacional	Logit (convergência mais rápida)
Tradição na literatura	Logit (mais utilizado)

2.11.3 Recomendações Práticas

1. **Para esta aplicação específica:** Ambos os modelos são adequados, com **ligeira vantagem para o Probit** em termos de ajuste (AIC menor)
2. **Para interpretação:** O **modelo Logit** oferece vantagem pela facilidade de interpretação via **odds ratios**
3. **Para pesquisa acadêmica:** A escolha pode depender da **tradição da área** ou **preferências teóricas**
4. **Para predição:** Ambos apresentam **performance equivalente** (diferença de acurácia < 0,2%)

2.11.4 Implicações para Política Pública

Os resultados sugerem pontos importantes para políticas de participação feminina no mercado de trabalho:

1. **Creches e cuidado infantil:** O forte efeito negativo de `kidslt6` sugere que políticas de apoio ao cuidado de crianças pequenas poderiam aumentar significativamente a participação feminina
2. **Educação:** O efeito positivo robusto da educação reforça a importância de investimentos em educação feminina
3. **Experiência profissional:** Programas de capacitação e experiência profissional têm potencial de impacto positivo
4. **Idade:** Políticas direcionadas a mulheres mais jovens podem ser mais efetivas

2.11.5 Limitações do Estudo

1. **Dados de 1975:** Os padrões podem ter mudado significativamente nas últimas décadas
2. **Amostra específica:** Resultados limitados a mulheres casadas nos EUA
3. **Variáveis omitidas:** Outros fatores importantes podem não estar incluídos (atitudes sociais, disponibilidade de emprego, etc.)
4. **Causalidade:** As relações estimadas são associações, não necessariamente causais