

# **Analítica de Datos**

## **Regresión Logística**

**Carlos Cardona Andrade**



# Tabla de contenido

1. Prediciendo variables categóricas
2. Logit y Razón de probabilidad

# **Prediciendo Variables Categóricas**

# Tipos de variables dependientes

## Variable dependiente continua:

- Precio de venta de casas en Bogotá
- **Modelo:** Precio de venta esperado dado el número de cuartos, el tamaño del lote, etc.

## Variable dependiente categorica:

- Riesgo de entrar en default de los clientes con un crédito
- **Modelo:** Probabilidad que un cliente haga default dada su edad, nivel educativo, etc.

# Modelo de probabilidad lineal

```
1 lin_model <- lm(default ~ age + factor(sex_string) + factor(marriage_string), data = default)
2 tidy(lin_model)
```

```
# A tibble: 5 x 5
  term            estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>      <dbl>    <dbl>
1 (Intercept) 0.233      0.0123     18.9  1.73e-79
2 age          -0.000289  0.000296   -0.977 3.29e- 1
3 factor(sex_string)Male 0.0351    0.00494     7.11 1.14e-12
4 factor(marriage_string)Other 0.0253    0.0234      1.08 2.80e- 1
5 factor(marriage_string)Single -0.0290   0.00547     -5.30 1.17e- 7
```

$$Default = \hat{\beta}_1 Age + \hat{\beta}_2 Male + \hat{\beta}_3 Other + \hat{\beta}_4 Single$$

# Modelo de probabilidad lineal

```
1 lin_model <- lm(default ~ age + factor(sex_string) + factor(marriage_string), data = default)
2 tidy(lin_model)
```

```
# A tibble: 5 x 5
  term            estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>      <dbl>    <dbl>
1 (Intercept) 0.233      0.0123     18.9  1.73e-79
2 age          -0.000289  0.000296   -0.977 3.29e- 1
3 factor(sex_string)Male 0.0351    0.00494     7.11 1.14e-12
4 factor(marriage_string)Other 0.0253    0.0234      1.08 2.80e- 1
5 factor(marriage_string)Single -0.0290   0.00547     -5.30 1.17e- 7
```

$$Default = \hat{\beta}_1 Age + \hat{\beta}_2 Male + \hat{\beta}_3 Other + \hat{\beta}_4 Single$$

Cada año adicional en la edad del cliente se asocia, en promedio, con un aumento de 0.02 (0.0002\$\$100) **puntos porcentuales** en la probabilidad de estar en default, manteniendo constantes las demás variables.

# Modelo de probabilidad lineal

```
1 lin_model <- lm(default ~ age + factor(sex_string) + factor(marriage_string), data = default)
2 tidy(lin_model)
```

```
# A tibble: 5 x 5
  term            estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>      <dbl>    <dbl>
1 (Intercept) 0.233      0.0123     18.9  1.73e-79
2 age          -0.000289  0.000296   -0.977 3.29e- 1
3 factor(sex_string)Male 0.0351    0.00494     7.11 1.14e-12
4 factor(marriage_string)Other 0.0253    0.0234      1.08 2.80e- 1
5 factor(marriage_string)Single -0.0290   0.00547     -5.30 1.17e- 7
```

$$Default = \beta_1 Age + \beta_2 Male + \beta_3 Other + \beta_4 Single$$

Ser soltero se asocia, en promedio, con una disminución de 2.9 (0.029\$\$100) **puntos porcentuales** en la probabilidad de estar en default, en comparación con estar casado, manteniendo constantes las demás variables.

# Puntos porcentuales vs Porcentaje

- Un punto porcentual es la unidad para la diferencia aritmética de dos porcentajes.
- Por ejemplo, pasar del 40 % al 44 % es un aumento de 4 puntos porcentuales, pero es un aumento real del 10 % en lo que se está midiendo.
- Al interpretar los coeficientes de un modelo de probabilidad lineal, multiplicamos el  $\beta$  por 100 y su unidad de medida son los puntos porcentuales.

# Modelos para variables categóricas

## Regresión Logística

2 Categorías

1: Si, 0: No

## Regresión Logística Multinomial

3+ Categorías

1: Conservador, 2: Liberal, 3: Independente

# Default Data

Los datos incluyen si el cliente cayó en default y otras características demográficas del cliente.

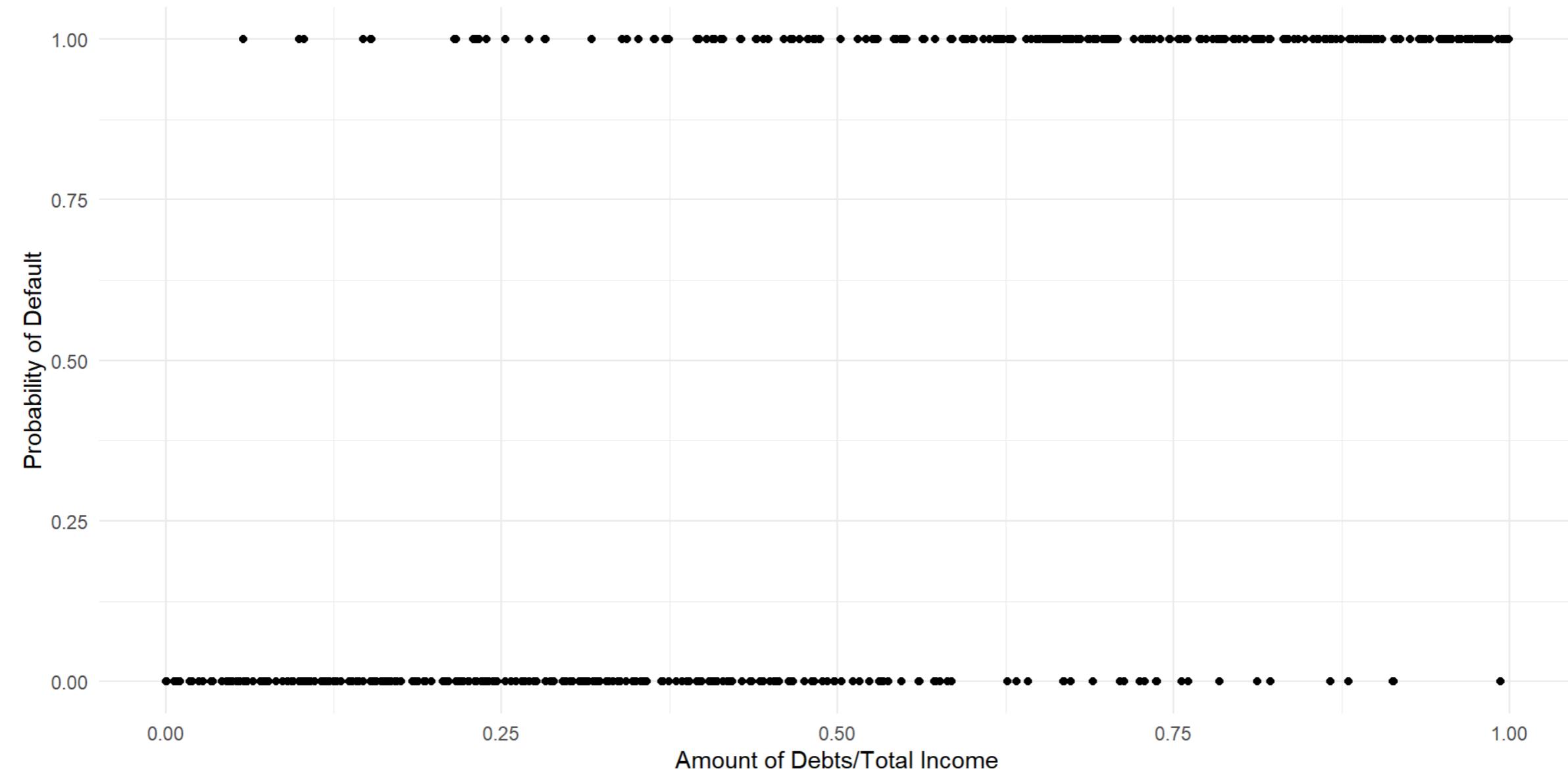
**default**

1: yes

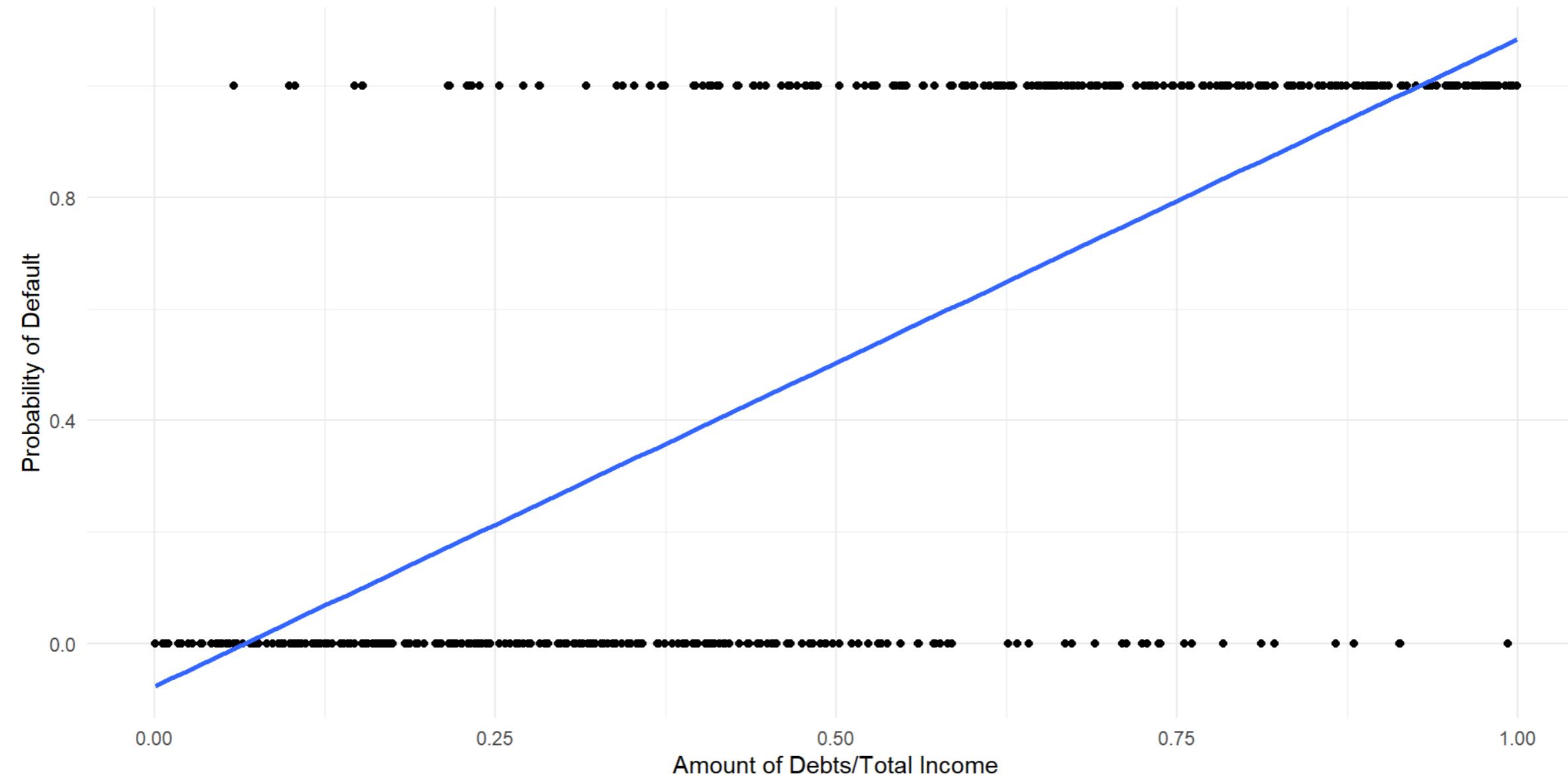
0: no

```
# A tibble: 30,000 x 3
  default    age    sex
  <dbl> <dbl> <dbl>
1     1     24     2
2     1     26     2
3     0     34     2
4     0     37     2
5     0     57     1
6     0     37     1
7     0     29     1
8     0     23     2
9     0     28     2
10    0     35     1
# i 29,990 more rows
```

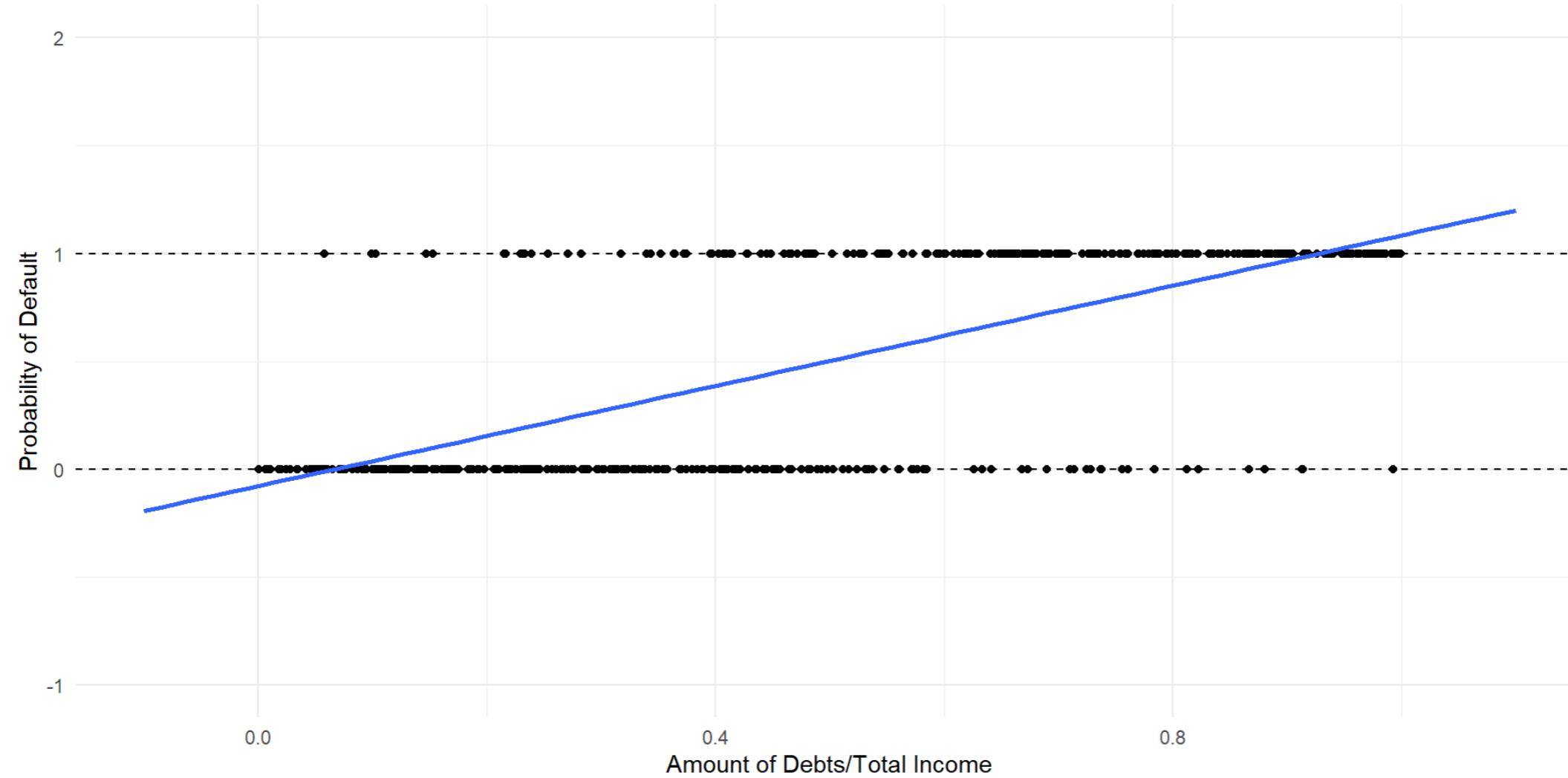
# Miremos los datos



# Regresión Lineal



# Regresión Lineal con Zoom



*Este modelo produce predicciones fuera del intervalo 0 y 1.*

# Intentemos con otro modelo



# Diferentes tipos de modelos

Método	Variable	Modelo
Regresión Lineal	Continua	$Y = \beta_0 + \beta_1 X$
Regresión Lineal (log(Y))	Continua	$\log(Y) = \beta_0 + \beta_1 X$
Regresión Logística	Binaria	$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

# Probabilidad y Razón de probabilidad

# Variable dependiente binaria

- $Y = 1$  : si,  $0$  : no
- $\pi$ : **probabilidad** de que  $Y = 1$ , i.e.,  $P(Y = 1)$
- $\frac{\pi}{1-\pi}$ : **razón de probabilidad (odds)** de que  $Y = 1$
- $\log\left(\frac{\pi}{1-\pi}\right)$ : **log odds**
- Ir de  $\pi$  a  $\log\left(\frac{\pi}{1-\pi}\right)$  usando la **transformación logística**

# Razón de Prob. (Odds)

Supongamos que hay un **70%** de probabilidad de que mañana llueva:

- La probabilidad de que llueva es  $p = 0.7$
- La probabilidad de que no llueva es  $1 - p = 0.3$
- La razón de prob. de que llueva es **7 to 3, 7:3,  $\frac{0.7}{0.3} \approx 2.33$**

# Cuáles son las probabilidades en nuestros datos?

```
1 default %>%
2   count(default) %>%
3   mutate(p = round(n / sum(n), 3))
```

```
# A tibble: 2 x 3
  default     n     p
  <dbl> <int> <dbl>
1      0 23364 0.779
2      1  6636 0.221
```

$$P(\text{default}) = P(Y = 1) = p = 0.221$$

$$P(\text{default}) = P(Y = 0) = 1 - p = 0.779$$

$$P(\text{odds de default}) = \frac{0.221}{0.779} = 0.283$$

# De odds a probabilidades

odds

$$\omega = \frac{\pi}{1 - \pi}$$

probabilidad

$$\pi = \frac{\omega}{1 + \omega}$$

# De odds a probabilidades

1. **Modelo Logístico:**  $\log \text{odds} = \log \left( \frac{\pi}{1-\pi} \right) = \beta_0 + \beta_1 X$

2. **Odds** =  $\exp \left\{ \log \left( \frac{\pi}{1-\pi} \right) \right\} = \frac{\pi}{1-\pi}$

3. Combinando (1) y (2) con lo visto antes

$$\text{probabilidad} = \pi = \frac{\exp\{\beta_0 + \beta_1 X\}}{1 + \exp\{\beta_0 + \beta_1 X\}}$$

# Modelo de Regresión Logística

**Forma Logística:**

$$\log \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 X$$

**Forma en Probabilidad:**

$$\pi = \frac{\exp\{\beta_0 + \beta_1 X\}}{1 + \exp\{\beta_0 + \beta_1 X\}}$$

# Modelo de Regresión Logística

```
1 logit_model <- glm(default ~ age, data = default, family = "binomial")
2
3 tidy(logit_model)

# A tibble: 2 x 5
  term      estimate std.error statistic p.value
  <chr>        <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept) -1.39      0.0552    -25.1  4.13e-139
2 age          0.00361   0.00150     2.41  1.62e- 2
```

$$\log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = -1.39 + 0.0036 \times \text{age}$$

donde  $\hat{\pi}$  es la probabilidad predicha de que un cliente esté en default

# log( odds) predichos

```
1 augment(logit_model)

# A tibble: 30,000 x 8
  default age .fitted .resid      .hat .sigma     .cooksdi .std.resid
  <dbl> <dbl> <dbl> <dbl>     <dbl> <dbl>       <dbl>        <dbl>
1     1    24   -1.30  1.76  0.0000840  1.03  0.000154  1.76
2     1    26   -1.29  1.75  0.0000681  1.03  0.000124  1.75
3     0    34   -1.26 -0.705 0.0000343  1.03  0.00000484 -0.705
4     0    37   -1.25 -0.709 0.0000341  1.03  0.00000487 -0.709
5     0    57   -1.18 -0.732 0.000219   1.03  0.0000336 -0.732
6     0    37   -1.25 -0.709 0.0000341  1.03  0.00000487 -0.709
7     0    29   -1.28 -0.700 0.0000498  1.03  0.00000691 -0.700
8     0    23   -1.30 -0.693 0.0000930  1.03  0.0000126 -0.693
9     0    28   -1.29 -0.699 0.0000552  1.03  0.00000763 -0.699
10    0    35   -1.26 -0.706 0.0000335  1.03  0.00000474 -0.706
# i 29,990 more rows
```

$$\text{predicted odds} = \hat{\omega} = \frac{\hat{\pi}}{1 - \hat{\pi}} = \exp\{-1.30\} = 0.272$$

# log(ods) predichos

```
1 # Get predictions
2 default$predprob <- predict(logit_model, newdata = default, type = "response")
3
4 default %>%
5   select(default, predprob) %>%
6   head (10)

# A tibble: 10 x 2
  default predprob
  <dbl>     <dbl>
1     1     0.214
2     1     0.215
3     0     0.220
4     0     0.222
5     0     0.235
6     0     0.222
7     0     0.217
8     0     0.213
9     0     0.217
10    0     0.221
```

$$\text{predicted probabilities} = \hat{\pi} = \frac{\exp\{-1.30\}}{1 + \exp\{-1.30\}} = 0.214$$

# ¿Cómo predecimos los 0s y 1s?

```
1 # Get predictions
2 default <- default %>%
3   mutate(pred_class = if_else(predprob >= 0.5, 1, 0))
4
5 default %>%
6   select(default, predprob, pred_class) %>%
7   head(10)
```

```
# A tibble: 10 x 3
  default predprob pred_class
  <dbl>     <dbl>      <dbl>
1     1     0.214      0
2     1     0.215      0
3     0     0.220      0
4     0     0.222      0
5     0     0.235      0
6     0     0.222      0
7     0     0.217      0
8     0     0.213      0
9     0     0.217      0
10    0     0.221      0
```

# Matríg de Confusión

```
1 confusion_matrix <- default %>%
2   count(default, pred_class)
3
4 confusion_matrix
```

```
# A tibble: 2 x 3
  default pred_class     n
  <dbl>      <dbl> <int>
1     0          0 23364
2     1          0  6636
```

# Matríg de Confusión

```
1 default <- default %>%
2   mutate(pred_class = if_else(predprob >= 0.3, 1, 0))
3
4 confusion_matrix <- default %>%
5   count(default, pred_class)
6
7 confusion_matrix
```

```
# A tibble: 2 x 3
  default pred_class     n
  <dbl>      <dbl> <int>
1     0          0 23364
2     1          0  6636
```

# Matríg de Confusión

```
1 default <- default %>%
2   mutate(pred_class = if_else(predprob >= 0.2, 1, 0))
3
4 confusion_matrix <- default %>%
5   count(default, pred_class)
6
7 confusion_matrix
```

```
# A tibble: 2 x 3
  default pred_class     n
  <dbl>      <dbl> <int>
1     0          1 23364
2     1          1  6636
```

# Matríg de Confusión

```
1 logit_model2 <- glm(default ~ age + factor(sex) + factor(marriage), data = default, family = "binomial")
2
3 default$predprob <- predict(logit_model2, newdata = default, type = "response")
4
5 default <- default %>%
6   mutate(pred_class = if_else(predprob >= 0.2, 1, 0))
7
8 confusion_matrix <- default %>%
9   count(default, pred_class)
10
11 confusion_matrix
```

```
# A tibble: 4 x 3
  default pred_class     n
  <dbl>      <dbl> <int>
1     0          0  7604
2     0          1 15760
3     1          0  1861
4     1          1 4775
```

# Interpretación de los coeficientes

$$\log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = -1.39 + 0.0036 \times \text{age}$$

Cada año adicional en la edad del cliente se asocia, en promedio, con un aumento de 0.0036 en el log(odd) de caer en default.

**Es útil esa interpretación?**

# Interpretación de los coeficientes

El paquete `mfx` nos permite calcular los efectos marginales sobre la probabilidad directamente

```
1 library(mfx)
2 logitmfx(default ~ age, data = default)
```

Call:

```
logitmfx(formula = default ~ age, data = default)
```

Marginal Effects:

	dF/dx	Std. Err.	z	P> z
age	0.00062123	0.00025821	2.4059	0.01613 *

---

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

Cada año adicional en la edad del cliente se asocia, en promedio, con un aumento de 0.0621 (0.000621\$\$100) **pp** en la probabilidad de estar en default, manteniendo constantes las demás variables.

# Logit vs MLP

```
1 library(mfx)
2 logitmfx(default ~ age, data = default)
```

Call:

```
logitmfx(formula = default ~ age, data = default)
```

Marginal Effects:

	dF/dx	Std. Err.	z	P> z
age	0.00062123	0.00025821	2.4059	0.01613 *

---

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

```
1 lin_model <- lm(default ~ age, data = default)
2 tidy(lin_model)
```

```
# A tibble: 2 x 5
  term      estimate std.error statistic p.value
  <chr>        <dbl>     <dbl>      <dbl>      <dbl>
1 (Intercept) 0.199     0.00953    20.9   3.88e-96
2 age         0.000625   0.000260    2.41  1.61e- 2
```

# Logit vs MLP

- Dado que las interpretaciones de los coeficientes no cambian mucho, el MLP es más usado por que es más flexible
- Sin embargo, el modelo logit tiene un poder predictivo más preciso

# Logit vs MLP

