Analítica de los Negocios

Regresión Logística

Carlos Cardona Andrade

Plan para hoy

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 categoríca:

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

Ejercicio 1

1. Con la función `read_csv()` importen los datos `default_data`:

```r
default_data <- read_csv("default_data.csv")
```

2. Con la función `glimpse()` revisen qué variables incluyen los datos y su tipo.

Ejercicio 1

3. Completen el siguiente código para graficar el total de clientes que caen o no en default:

```
ggplot(default, aes(x = _______________)) +
  geom_bar() +
  labs(title = "Number of Clients by Default Payment",
       x = "",
       y = "Frequency") +
  scale_y_continuous(limits = c(0, 30000)) +
  theme_minimal()
```

4. ¿Qué hecho puntual sobre los datos nos muestra la siguiente gráfica?

```
ggplot(default, aes(fill=sex, x=default_string)) +
  geom_bar(position="fill") +
  labs(title = "Number of Clients by Sex and Default Payment",
       fill='Sex',
       x = "",
       y = "Proportion") +
  theme_minimal()
```

💪 Ejercicio 1

5. Estimen el siguiente modelo:

```
lin_model <- lm(default ~ age + factor(sex) + factor(marriage), data = defa
tidy(lin_model)
```

E interpreten los coeficientes para cada regresión.

Modelo de probabilidad lineal

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

```
# A tibble: 5 × 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

```r
lin_model <- lm(default ~ age + factor(sex_string) + factor(marriage_string
tidy(lin_model)
```

```
# A tibble: 5 × 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 $\times$ 100) **puntos porcentuales** en la probabilidad de estar en default, manteniendo constantes las demás variables.

Modelo de probabilidad lineal

```r
lin_model <- lm(default ~ age + factor(sex_string) + factor(marriage_string
tidy(lin_model)
```

```
# A tibble: 5 × 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
```

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

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 β por 100 y su unidad de medidad 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`

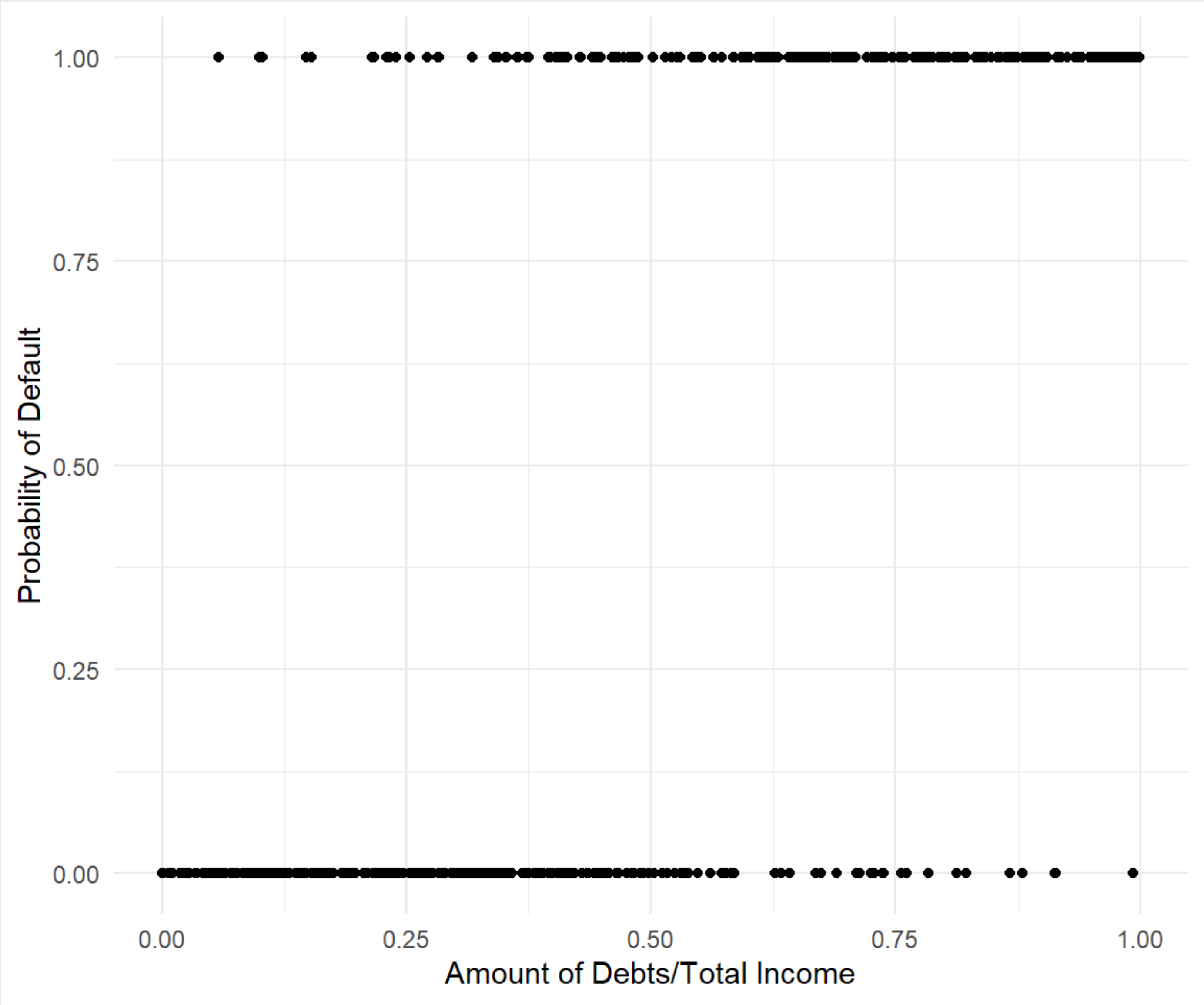
1: yes

0: no

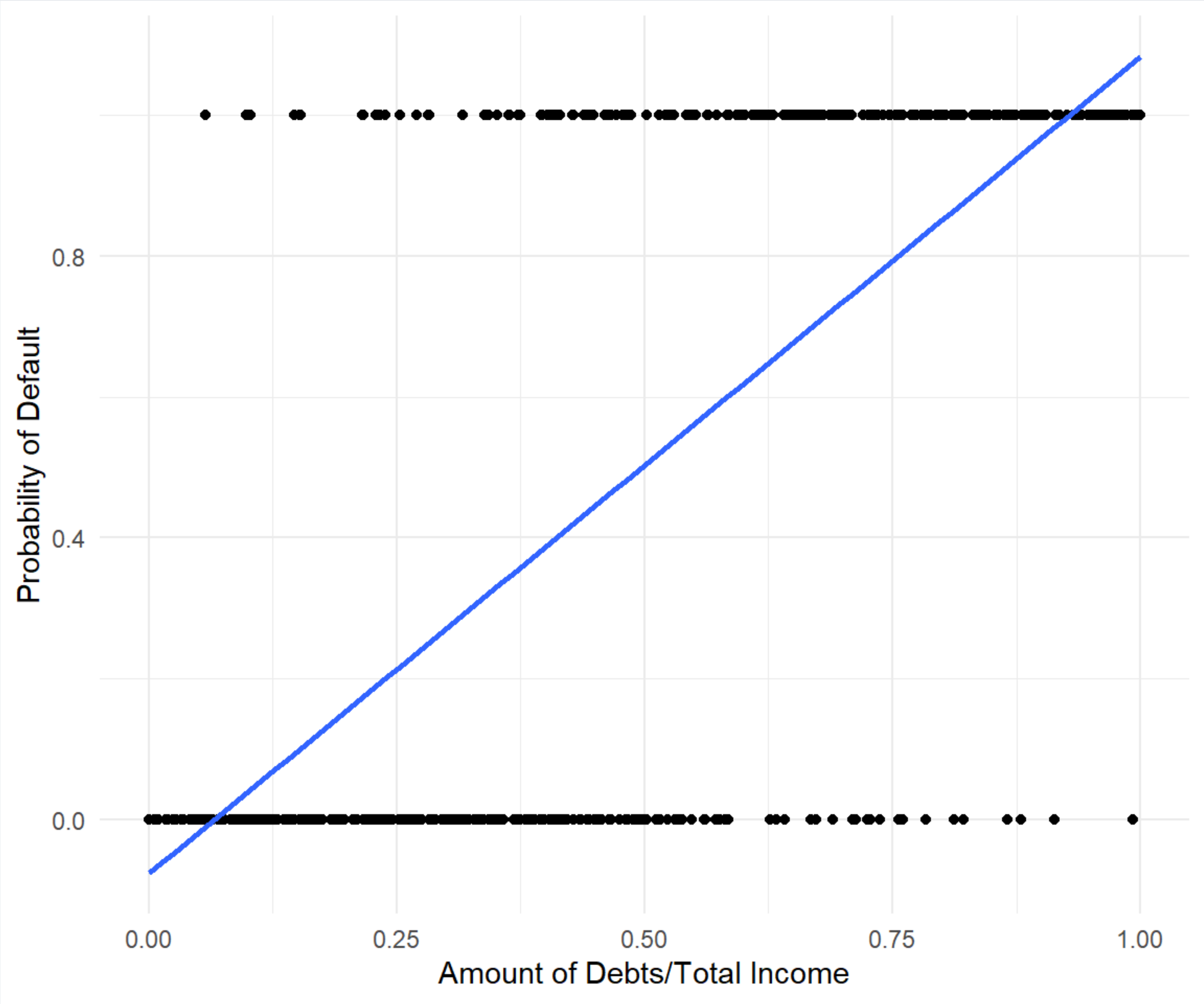
```
default |>
  select(default, age, sex)
```

```
# A tibble: 30,000 × 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
```

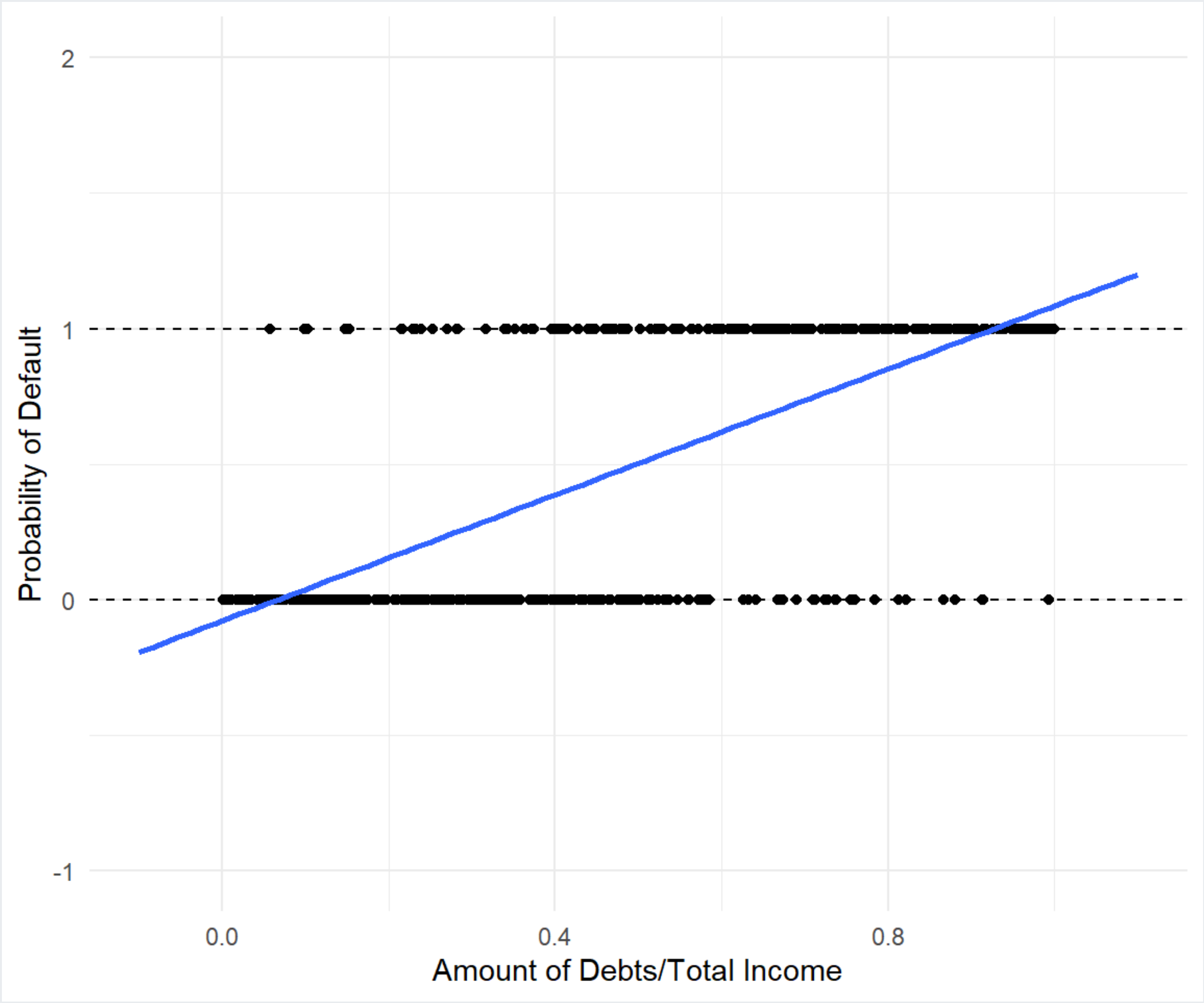
Miremos los datos



Regresión Lineal

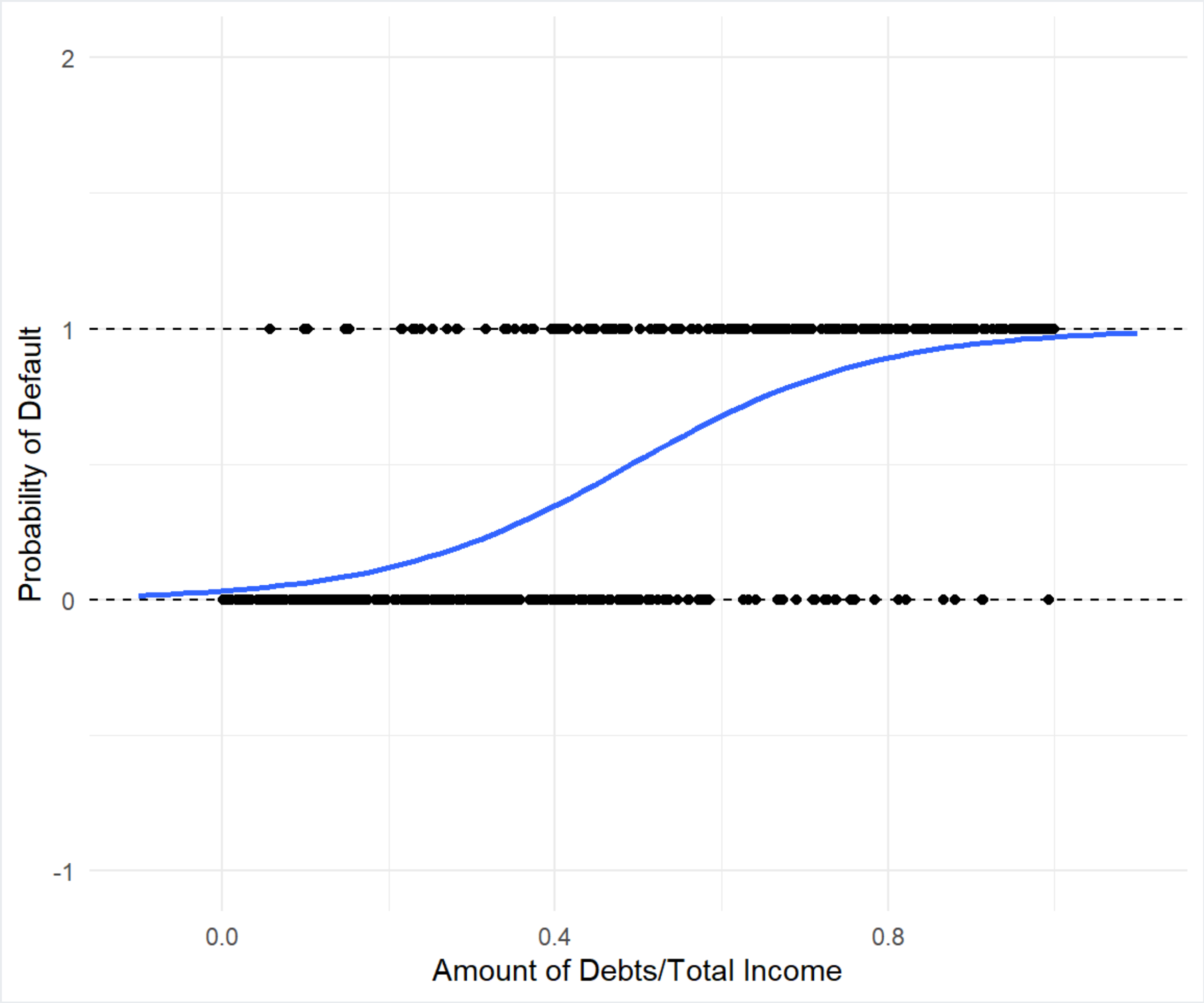


Regresión Lineal con Zoom



Este modelo produce predicciones fuera del intervalo 0 y 1.

Intentemos con otro modelo



✅ Este modelo (llamado **regresión logística**) solo predice en el intervalo 0 y 1

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

- $Y = 1 :\ \text{si}, 0 :\ \text{no}$
- π: **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 π 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 $\mathbf{p = 0.7}$
- La probabilidad de que no llueva es $\mathbf{1 - p = 0.3}$
- La razón de prob. de que llueva es **7 to 3**, **7:3**, $\mathbf{\frac{0.7}{0.3} \approx 2.33}$

Cuáles son las probabilidades en nuestros datos?

```
default |>
  count(default) |>
  mutate(p = round(n / sum(n), 3))
```

```
# A tibble: 2 × 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 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

```r
logit_model <- glm(default ~ age, data = default, family = "binomial")

tidy(logit_model)
```

```
# A tibble: 2 × 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 × 8
   default   age .fitted .resid       .hat .sigma    .cooksd .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
```

$$fitted = \hat{\beta}_0 + \hat{\beta}_1 \ X = -1.3$$

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

log(odds) predichos

```r
# Get predictions
default$predprob <- predict(logit_model, newdata = default, type = "respons

default |>
  select(default, predprob) |>
  head (10)
```

```
# A tibble: 10 × 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?

```
# Get predictions
default <- default |>
  mutate(pred_class = if_else(predprob >= 0.5, 1, 0))

default |>
  select(default, predprob, pred_class) |>
  head (10)
```

```
# A tibble: 10 × 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
```

Matríz de Confusión

```r
confusion_matrix <- default |>
  count(default, pred_class)

confusion_matrix
```

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

Matríz de Confusión

```r
default <- default |>
  mutate(pred_class = if_else(predprob >= 0.3, 1, 0))

confusion_matrix <- default |>
  count(default, pred_class)

confusion_matrix
```

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

Matríz de Confusión

```
default <- default |>
  mutate(pred_class = if_else(predprob >= 0.2, 1, 0))

confusion_matrix <- default |>
  count(default, pred_class)

confusion_matrix
```

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

Ejercicio 2

Estimen el siguiente modelo:

```r
logit_model2 <- glm(default ~ age + factor(sex) + factor(marriage), data =
```

Y construyan la matriz de confusión donde caer en default sea igual a 1 si la probabilidad predicha es mayor a 0,2.

2. Comparen este resultado con el de la diapositiva 32. ¿Por qué son diferentes?

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

```r
library(mfx)
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

```r
library(mfx)
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
```

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

```
# A tibble: 2 × 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