Logistic Regression

Logistic Regression

Introduction to Binary Outcomes

Continuous vs. Categorical Variables

- General linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Independent variables (x's):
  - Continuous: age, income, height → use numerical value.
  - Categorical: gender, city, ethnicity → use dummies.
- Dependent variable (y):
  - Continuous: consumption, time spent → use numerical value.
  - Categorical: yes/no → use dummies.

Examples of Binary Outcomes

- Should a bank give a person a loan or not?
- Is an individual transaction fraudulent or not?
- What determines admittance into a school?
- Which people are more likely to vote against a new law?
- Which customers are more likely to buy a new product?

Representing the Binary Outcomes

- There are two outcomes: Yes and No

- We will create a dummy variable to indicate if an observation is a Yes or a No:
  - $y = 1$ if Yes
  - $y = 0$ if No

- If we code the variable the other way around, our coefficients will have the same magnitudes but opposite signs.

A linear model?

- Aside from being binary, there's really nothing special about our dependent variable (y).
- Its value is higher (from a 0 to a 1) if a customer subscribes, so whatever makes it higher increases the likelihood of subscription.
- We can then run:

$$subscribe = \beta_0 + \beta_1\, age + \varepsilon$$

Result of Linear Model

gretl: model 1

File Edit Tests Save Graphs Analysis LaTeX

```
Model 1: OLS, using observations 1-1000
Dependent variable: subscribe

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -1.70073      0.0638035    -26.66    1.20e-118 ***
  age         0.0645433    0.00178736    36.11    2.52e-183 ***

Mean dependent var   0.573000   S.D. dependent var   0.494890
Sum squared resid    106.0736   S.E. of regression   0.326016
R-squared            0.566464   Adjusted R-squared   0.566030
F(1, 998)            1304.002   P-value(F)           2.5e-183
Log-likelihood      -297.1275   Akaike criterion     598.2550
Schwarz criterion    608.0705   Hannan-Quinn         601.9855
```

$$subscribe = -1.700 + 0.064\ age$$

Interpreting the Result

- If our dependent variable is binary, then we want to see what makes it change from a 0 to 1.

Interpreting the Result

- If our dependent variable is binary, then we want to see what makes it change from a 0 to 1.
- This can also be interpreted as what increases the likelihood of subscription, or $\mathrm{P}(subscribe = 1)$, which we can also simply denote as p.

Interpreting the Result

- If our dependent variable is binary, then we want to see what makes it change from a 0 to 1.
- This can also be interpreted as what increases the likelihood of subscription, or $P(subscribe = 1)$, which we can also simply denote as p.
- The result can be read as:

$$P(subscribe = 1) = p = -1.700 + 0.064\, age$$

Interpreting the Result

- If our dependent variable is binary, then we want to see what makes it change from a 0 to 1.

- This can also be interpreted as what increases the likelihood of subscription, or $P(subscribe = 1)$, which we can also simply denote as p.

- The result can be read as:

$$P(subscribe = 1) = p = -1.700 + 0.064\ age$$

- Every additional year of age increases the probability of subscription by 6.4%.

Problems with the Linear Approach

- Probabilities are bounded whereby $0 \leq p \leq 1$.

Problems with the Linear Approach

- Probabilities are bounded whereby $0 \le p \le 1$.
- The range of age in the data is such that $20 \le age \le 55$.

Problems with the Linear Approach

- Probabilities are bounded whereby $0 \leq p \leq 1$.
- The range of age in the data is such that $20 \leq age \leq 55$.
- The probability that a 35 year-old person subscribes is:

$$p = -1.700 + 0.064 \times 35 = 0.54$$

Problems with the Linear Approach

- Probabilities are bounded whereby $0 \leq p \leq 1$.
- The range of age in the data is such that $20 \leq age \leq 55$.
- The probability that a 35 year-old person subscribes is:

$$p = -1.700 + 0.064 \times 35 = 0.54$$

- What about people with 25 and 45 years of age?

$$p = -1.700 + 0.064 \times 25 = -0.09$$

Problems with the Linear Approach

- Probabilities are bounded whereby $0 \leq p \leq 1$.
- The range of age in the data is such that $20 \leq age \leq 55$.
- The probability that a 35 year-old person subscribes is:

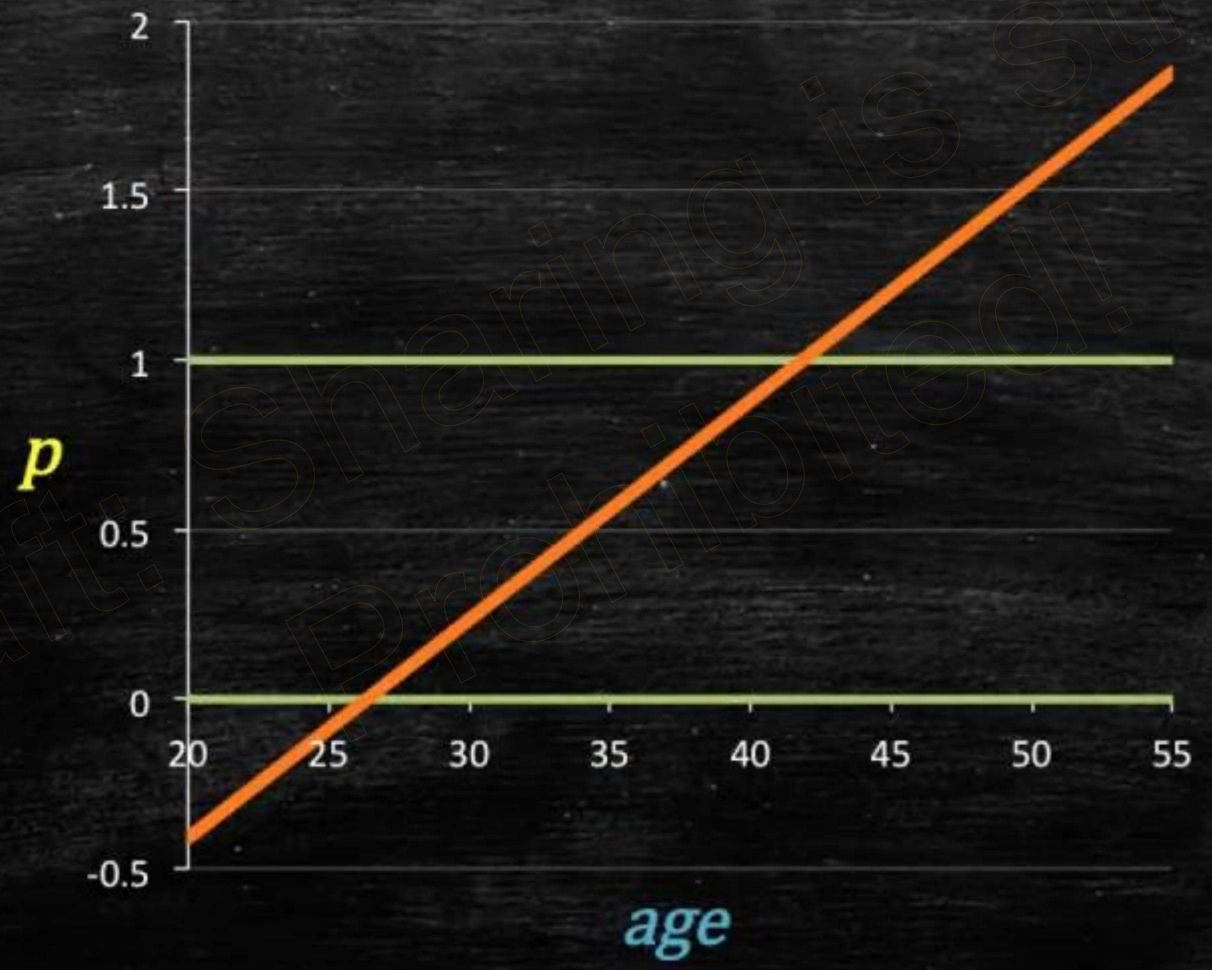
$$p = -1.700 + 0.064 \times 35 = 0.54$$

- What about people with 25 and 45 years of age?

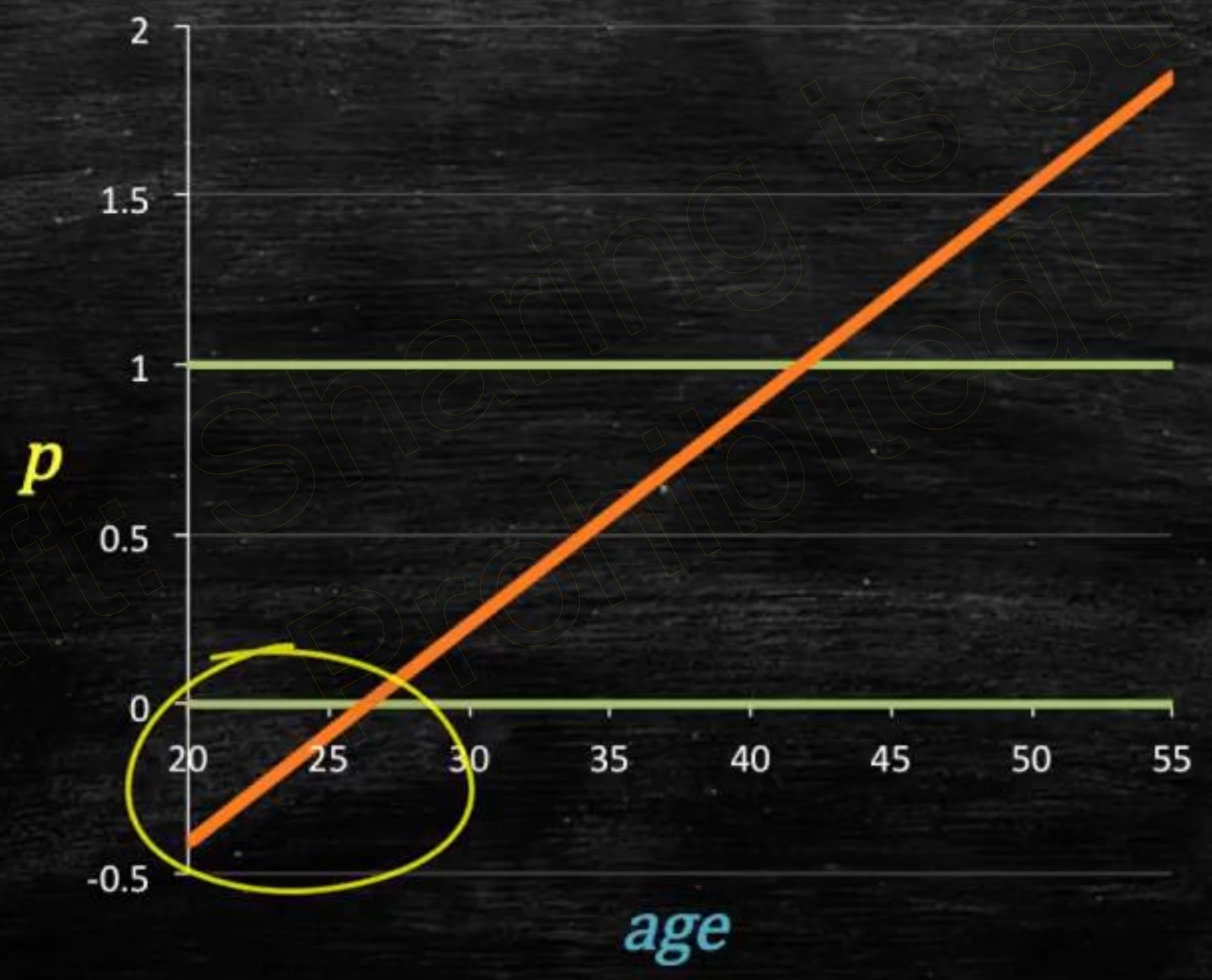
$$p = -1.700 + 0.064 \times 25 = -0.09$$

$$p = -1.700 + 0.064 \times 45 = 1.20$$

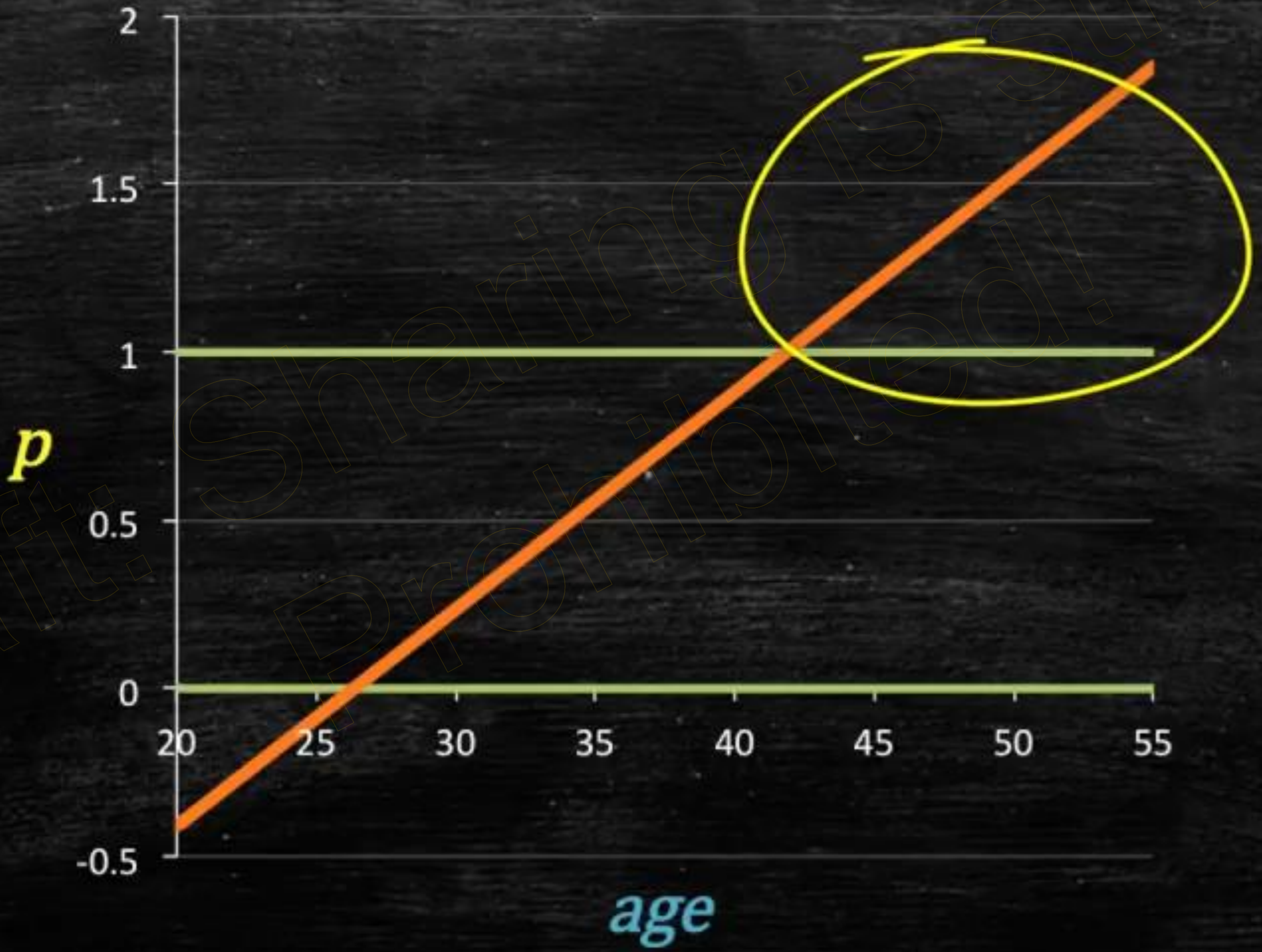
Linear Model Plot



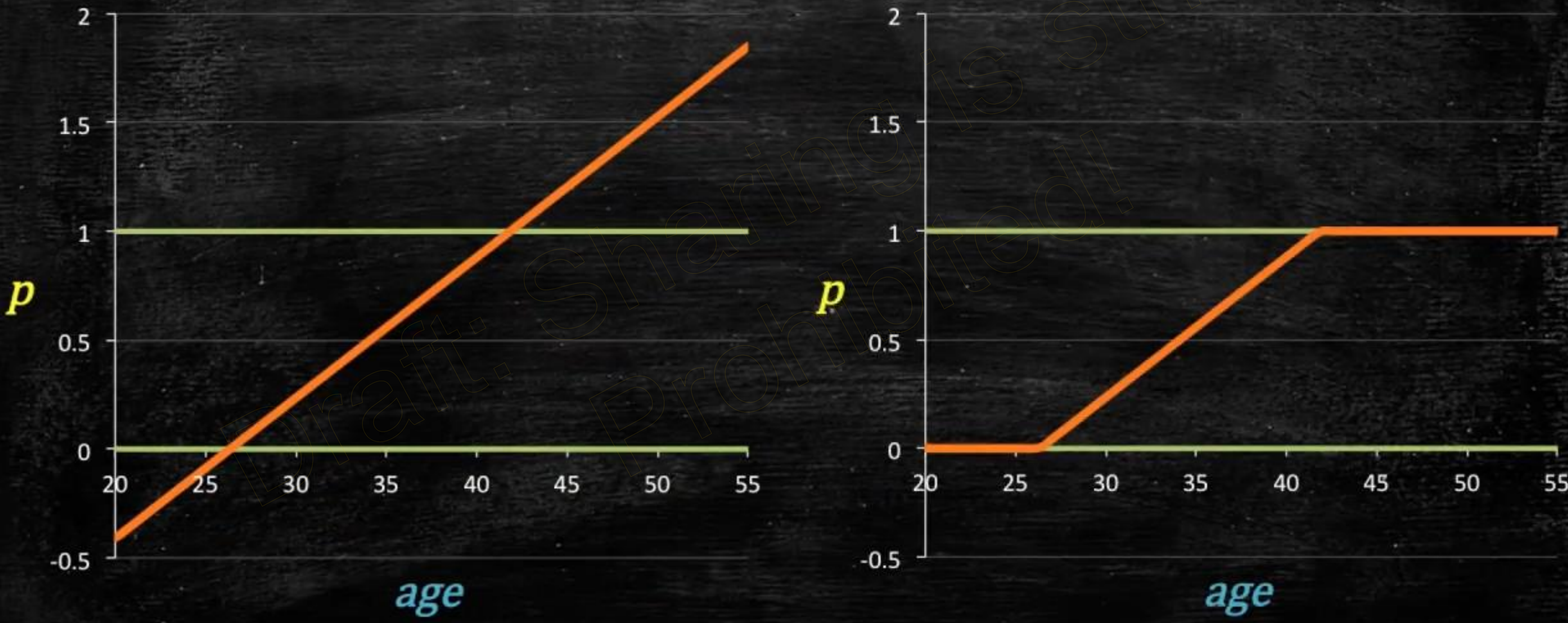
Linear Model Plot



Linear Model Plot



Linear Model Plot



Two Steps!

Two Steps!

1. It must always be positive (since $p \geq 0$)

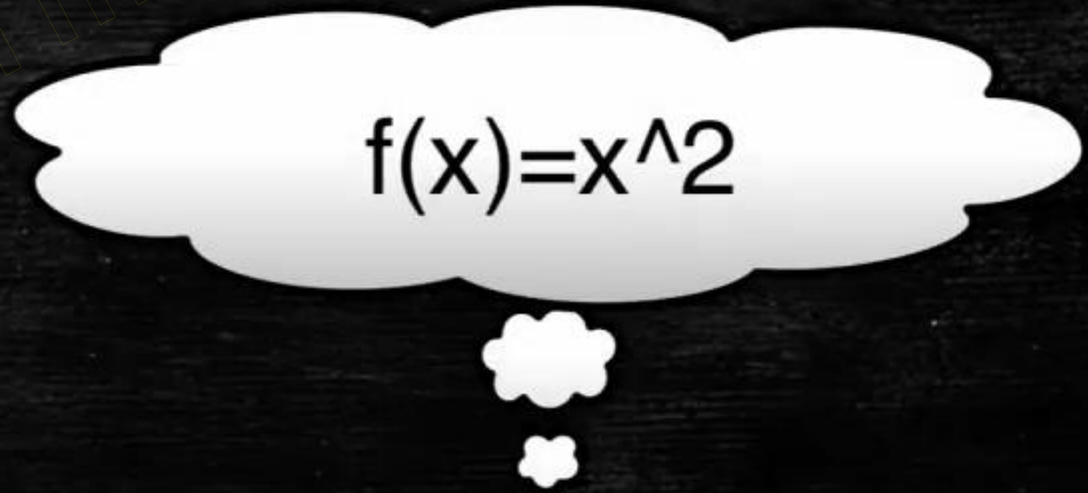
Two Steps!

1. It must always be positive (since $p \geq 0$)

$$f(x)=abs(x)=|x|$$

Two Steps!

1. It must always be positive (since $p \geq 0$)



Two Steps!

1. It must always be positive (since $p \geq 0$)

$$p = \exp(\beta_0 + \beta_1\, age) = e^{\beta_0 + \beta_1\, age}$$

Two Steps!

1. It must always be positive (since $p \geq 0$)

$$p = \exp(\beta_0 + \beta_1\, age) = e^{\beta_0 + \beta_1\, age}$$

2. It must be less than 1 (since $p \leq 1$)

Two Steps!

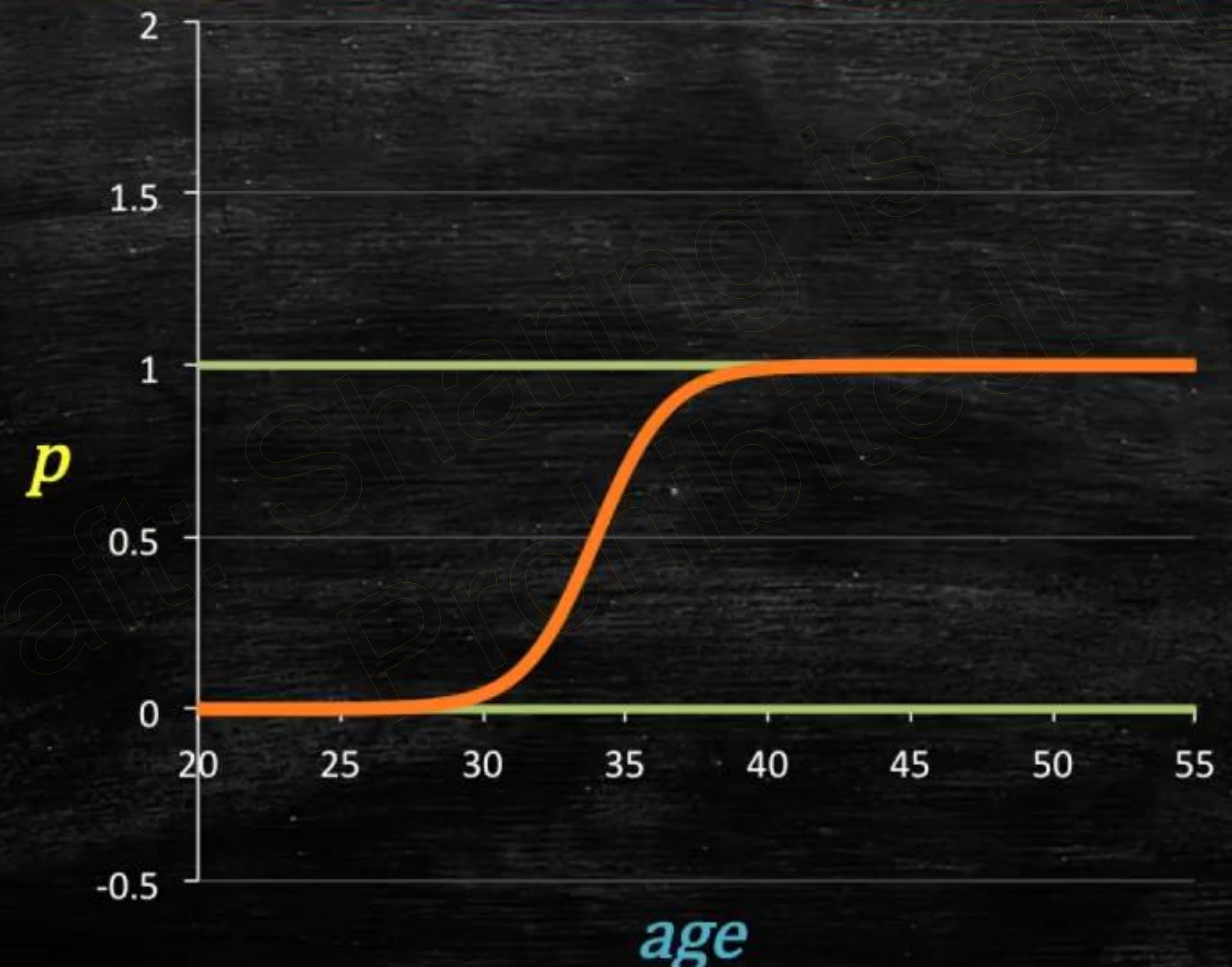
1. It must always be positive (since $p \geq 0$)

$$p = \exp(\beta_0 + \beta_1\, age) = e^{\beta_0 + \beta_1\, age}$$

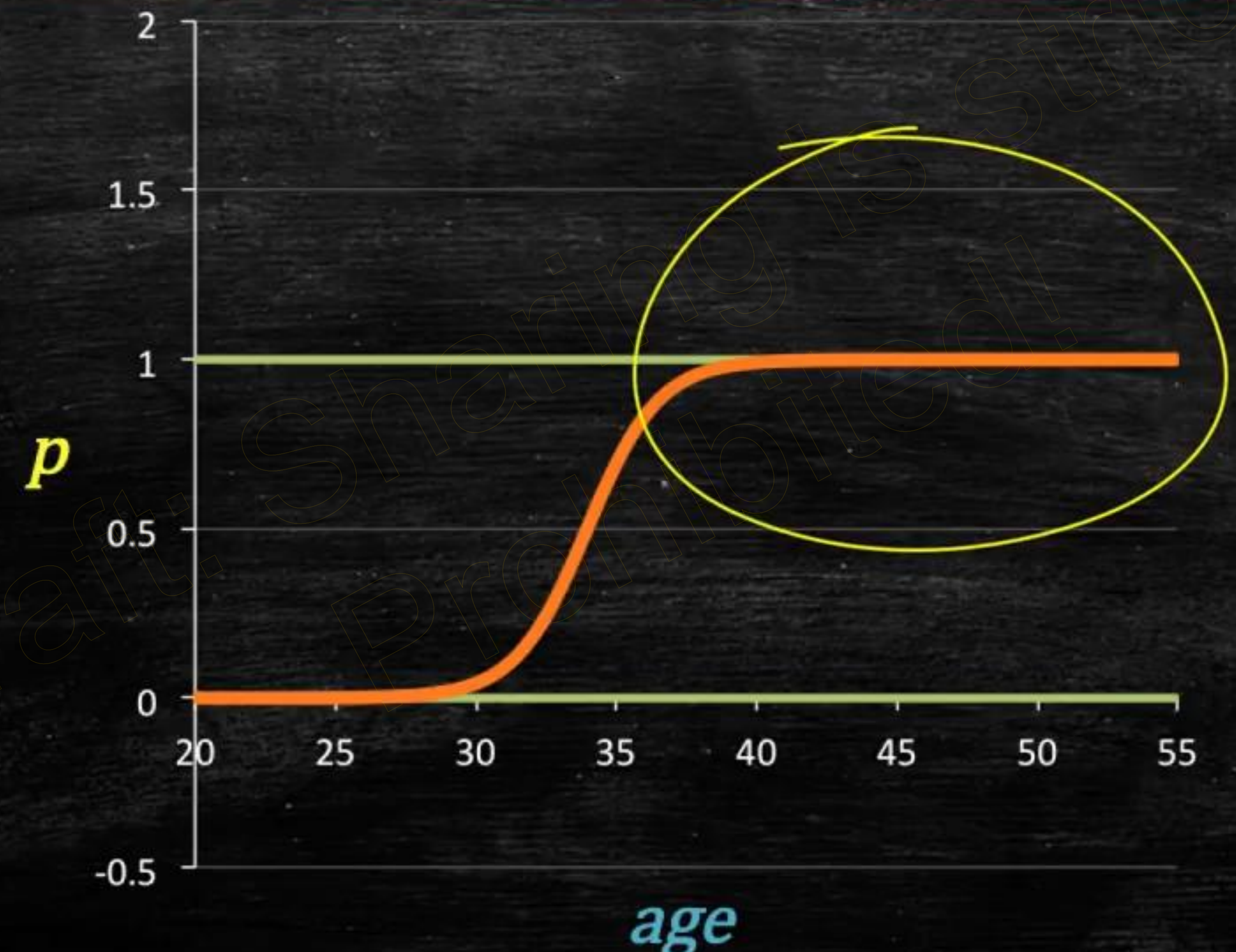
2. It must be less than 1 (since $p \leq 1$)

$$p = \frac{\exp(\beta_0 + \beta_1\, age)}{\exp(\beta_0 + \beta_1\, age) + 1} = \frac{e^{\beta_0 + \beta_1\, age}}{e^{\beta_0 + \beta_1\, age} + 1}$$

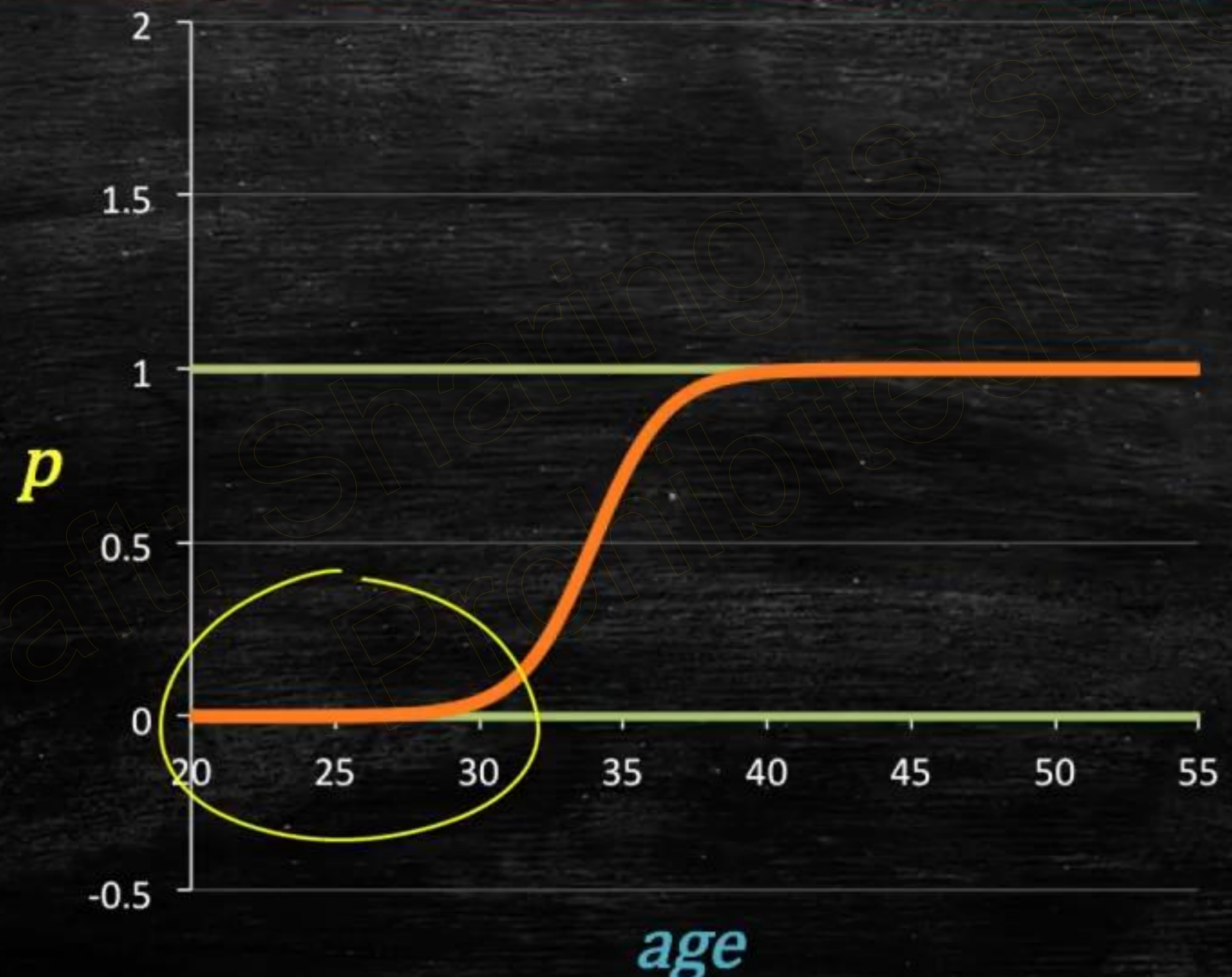
Logistic Model Plot



Logistic Model Plot



Logistic Model Plot



The Linear Thinking is not Completely Gone

The Linear Thinking is not Completely Gone

- The previous expression (by doing some algebra) can be rewritten as:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\, age$$

The Linear Thinking is not Completely Gone

- The previous expression (by doing some algebra) can be rewritten as:

Log Odds

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\, age$$

- Even though the probability of a customer subscribing (p) is not a linear function of age, the simple transformation is a linear function of age.

The Linear Thinking is not Completely Gone

- The previous expression (by doing some algebra) can be rewritten as:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\, age$$

- Even though the probability of a customer subscribing (p) is not a linear function of age, the simple transformation is a linear function of age.

- The above equation is the one used in **logistic regressions**.

Result of Logistic Regression

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

```
Model 2: Logit, using observations 1-1000
Dependent variable: subscribe
Standard errors based on Hessian

             coefficient   std. error      z        slope
  -----------------------------------------------------------
  const      -26.5240      1.82819      -14.51
  age          0.781053    0.0535623     14.58   0.154207

Mean dependent var   0.573000   S.D. dependent var   0.494890
McFadden R-squared   0.636613   Adjusted R-squared   0.633683
Log-likelihood      -247.9937   Akaike criterion     499.9873
Schwarz criterion    509.8028   Hannan-Quinn         503.7179

Number of cases 'correctly predicted' = 884 (88.4%)
f(beta'x) at mean of independent vars = 0.197
Likelihood ratio test: Chi-square(1) = 868.915 [0.0000]

           Predicted
             0      1
  Actual 0  350     77
         1   39    534
```

Result of Logistic Regression

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

```
Model 2: Logit, using observations 1-1000
Dependent variable: subscribe
Standard errors based on Hessian

             coefficient   std. error      z        slope
  ---------------------------------------------------------
  const      -26.5240      1.82819      -14.51
  age          0.781053    0.0535623     14.58    0.154207

Mean dependent var   0.573000   S.D. dependent var   0.494890
McFadden R-squared   0.636613   Adjusted R-squared   0.633683
Log-likelihood      -247.9937   Akaike criterion     499.9873
Schwarz criterion    509.8028   Hannan-Quinn         503.7179

Number of cases 'correctly predicted' = 884 (88.4%)
f(beta'x) at mean of independent vars = 0.197
Likelihood ratio test: Chi-square(1) = 868.915 [0.0000]

           Predicted
              0      1
  Actual 0  350     77
         1   39    534
```

The Estimated Logistic Model

- The estimated model was:

$$\ln\left(\frac{p}{1-p}\right) = -26.52 + 0.78\ age$$

The Estimated Logistic Model

- The estimated model was:

$$\ln\left(\frac{p}{1-p}\right) = -26.52 + 0.78\, age$$

- Or written in terms of the probability p we have:

The Estimated Logistic Model

- The estimated model was:

$$\ln\left(\frac{p}{1-p}\right) = -26.52 + 0.78\, age$$

- Or written in terms of the probability p we have:

$$p = \frac{\exp(-26.52 + 0.78\, age)}{\exp(-26.52 + 0.78\, age) + 1} = \frac{e^{-26.52 + 0.78\, age}}{e^{-26.52 + 0.78\, age} + 1}$$

Logistic Regression

Interpretation of Coefficients and Forecasting

Leveraging the Similarities with Linear Models

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

```
Model 2: Logit, using observations 1-1000
Dependent variable: subscribe
Standard errors based on Hessian

             coefficient   std. error      z        slope
  ---------------------------------------------------------
  const      -26.5240      1.82819      -14.51
  age          0.781053    0.0535623     14.58   0.154207

Mean dependent var   0.573000   S.D. dependent var   0.494890
McFadden R-squared   0.636613   Adjusted R-squared   0.633683
Log-likelihood      -247.9937   Akaike criterion     499.9873
Schwarz criterion    509.8028   Hannan-Quinn         503.7179

Number of cases 'correctly predicted' = 884 (88.4%)
f(beta'x) at mean of independent vars = 0.197
Likelihood ratio test: Chi-square(1) = 868.915 [0.0000]

           Predicted
              0      1
  Actual 0  350     77
         1   39    534
```

Leveraging the Similarities with Linear Models

gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

```
Model 2: Logit, using observations 1-1000
Dependent variable: subscribe
Standard errors based on Hessian

             coefficient   std. error      z        slope
  -----------------------------------------------------------
  const      -26.5240       1.82819      -14.51
  age          0.781053     0.0535623     14.58    0.154207

Mean dependent var    0.573000   S.D. dependent var    0.494890
McFadden R-squared    0.636613   Adjusted R-squared    0.633683
Log-likelihood      -247.9937    Akaike criterion     499.9873
Schwarz criterion    509.8028    Hannan-Quinn         503.7179

Number of cases 'correctly predicted' = 884 (88.4%)
f(beta'x) at mean of independent vars = 0.197
Likelihood ratio test: Chi-square(1) = 868.915 [0.0000]

           Predicted
             0      1
  Actual 0  350     77
         1   39    534
```

Leveraging the Similarities with Linear Models

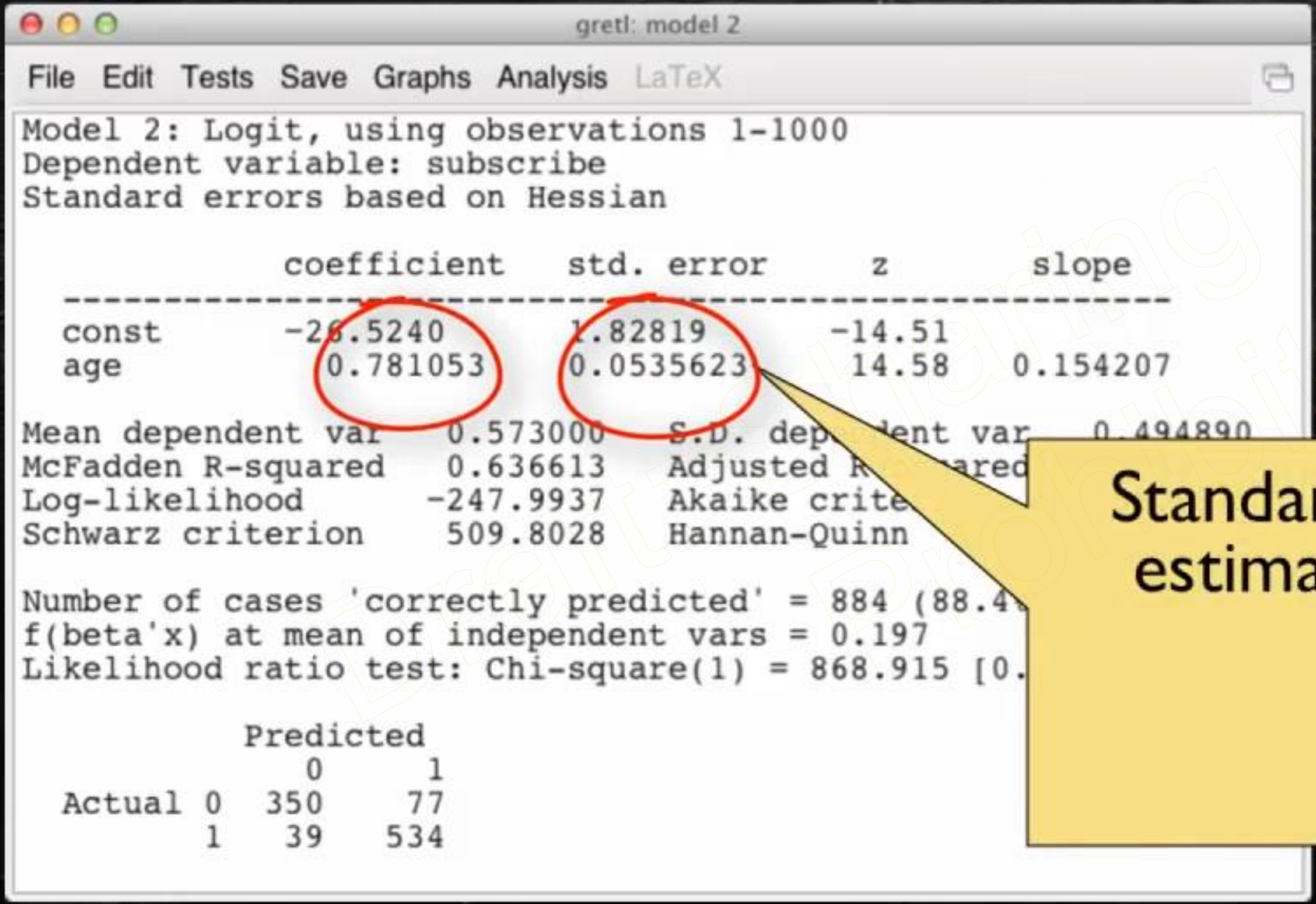
gretl: model 2

File Edit Tests Save Graphs Analysis LaTeX

```
Model 2: Logit, using observations 1-1000
Dependent variable: subscribe
Standard errors based on Hessian

             coefficient   std. error      z
  ---------------------------------------------------------
  const      -26.5240      1.82819       -14.51
  age          0.781053    0.0535623      14.58   0.154207

Mean dependent var   0.573000   S.D. dependent var   0.494890
McFadden R-squared   0.636613   Adjusted R-squared   0.633683
Log-likelihood      -247.9937   Akaike criterion     499.9873
Schwarz criterion    509.8028   Hannan-Quinn         503.7179

Number of cases 'correctly predicted' = 884 (88.4%)
f(beta'x) at mean of independent vars = 0.197
Likelihood ratio test: Chi-square(1) = 868.915 [0.0000]

             Predicted
               0      1
  Actual 0   350     77
         1    39    534
```

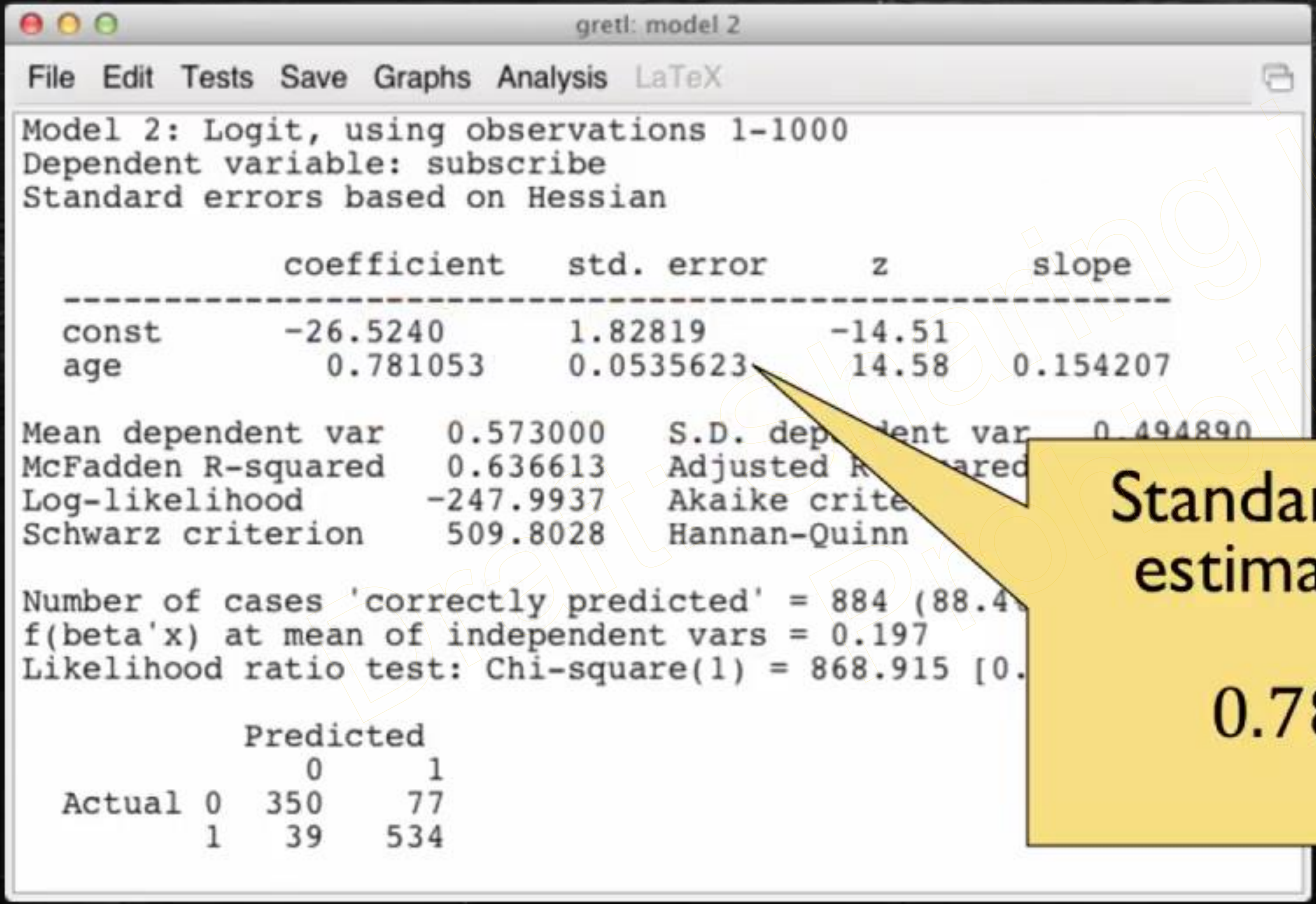
Sign of coefficients still represents a positive or negative influence on dependent variable.

Leveraging the Similarities with Linear Models



Standard errors can be used to estimate confidence intervals:

Leveraging the Similarities with Linear Models



Standard errors can be used to estimate confidence intervals:

$$0.78105 \pm 2 \times 0.05356$$
$$[\,0.674,\ 0.888]$$

What changed?

What changed?

- We can no longer interpret the (magnitude of) the coefficients as we did before.

What changed?

- We can no longer interpret the (magnitude of) the coefficients as we did before.
- What is the meaning of 0.78 in our estimated model?

What changed?

- We can no longer interpret the (magnitude of) the coefficients as we did before.
- What is the meaning of 0.78 in our estimated model?

$$\ln\left(\frac{p}{1-p}\right) = -26.524 + 0.781\ age$$

What changed?

- We can no longer interpret the (magnitude of) the coefficients as we did before.
- What is the meaning of 0.78 in our estimated model?

$$\ln\left(\frac{p}{1-p}\right) = -26.524 + 0.781\ age$$

- For every unit increase of *age*, $\ln\left(\frac{p}{1-p}\right)$ increases 0.78 units.

Increasing ln(odd) is actually increasing probability.

In brief

Logistic Regression

- Supervised learning method for classification.
- "logit" = "log odds"

$$odds = \frac{P(event)}{1 - P(event)}$$

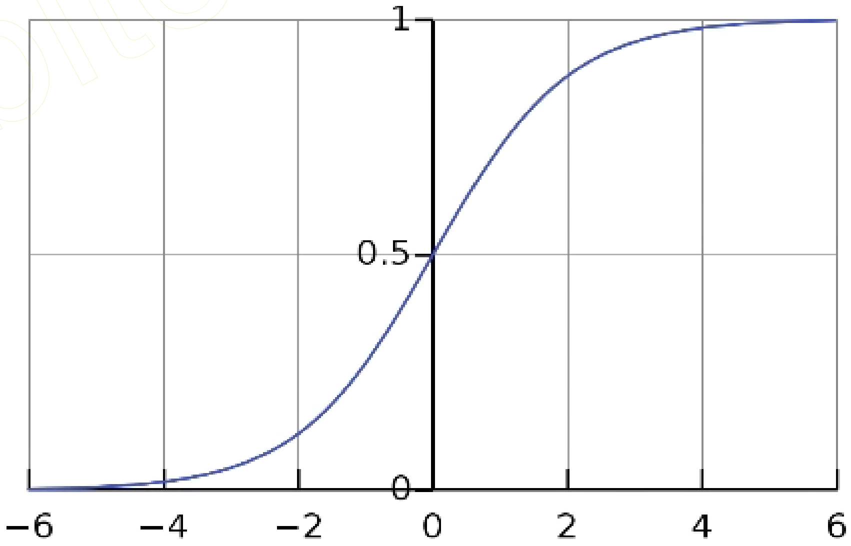
$$X \in \boldsymbol{R}$$

$$p(X) \in [0, 1]$$

- Let $\Pr(y = 1 | X) = p(X)$
- Sigmoid Function: $p(X) = \frac{1}{1 + e^{-\beta X}}$

What is unknown in the sigmoid function?

Estimate that parameter



Parameter Estimation

- Goal of learning is to estimate parameter vector $\widehat{\beta}$

Parameter Estimation

- Goal of learning is to estimate parameter vector $\widehat{\beta}$
- Logistic Regression uses Maximum Likelihood for parameter estimation. How does this work?

Parameter Estimation

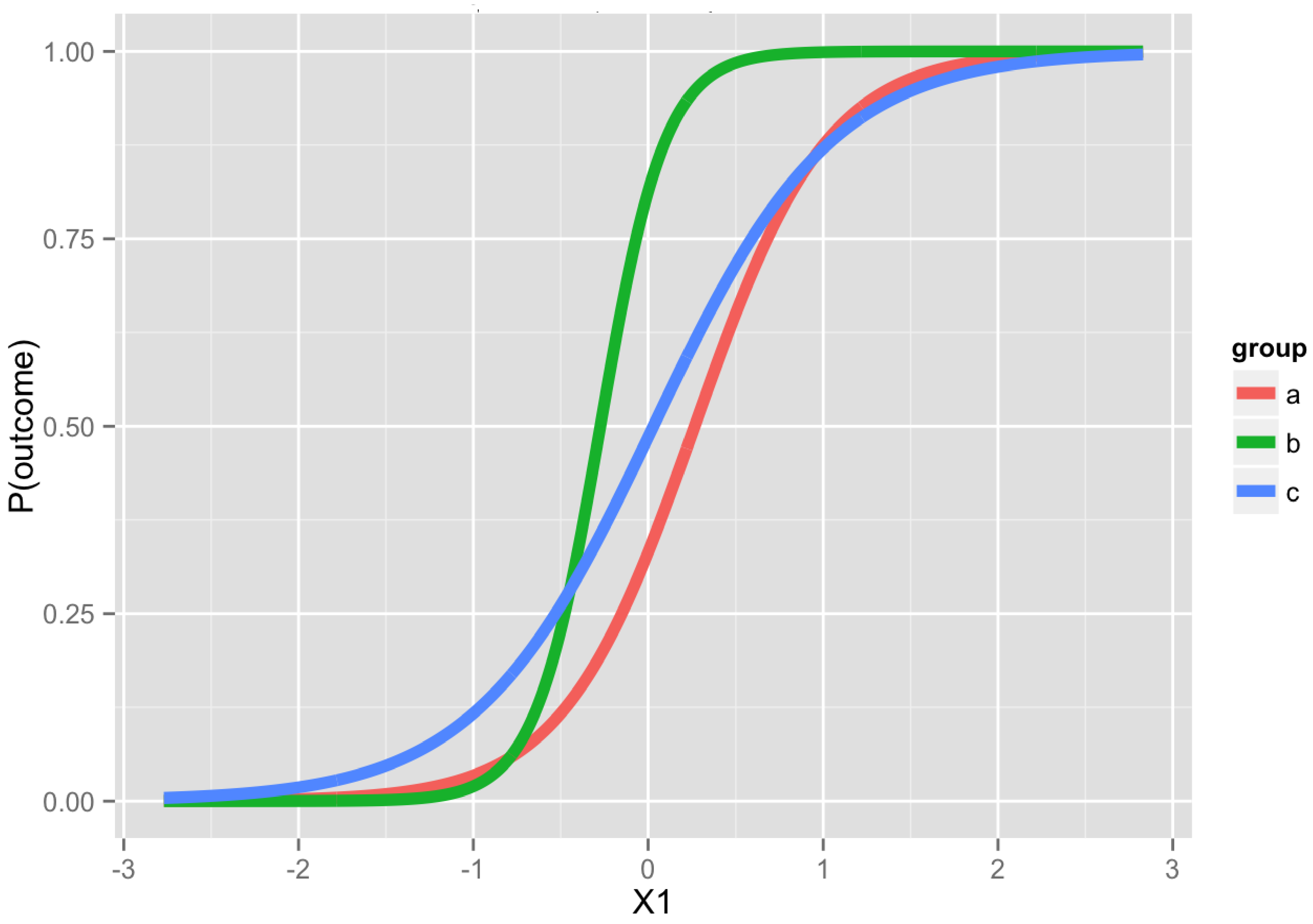
- Goal of learning is to estimate parameter vector $\widehat{\beta}$
- Logistic Regression uses Maximum Likelihood for parameter estimation. How does this work?
  - Consider N samples with labels either 0 or 1

Parameter Estimation

- Goal of learning is to estimate parameter vector $\widehat{\beta}$
- Logistic Regression uses Maximum Likelihood for parameter estimation. How does this work?
  - Consider N samples with labels either 0 or 1
  - For samples labelled "1": Estimate $\widehat{\beta}$ such that $\widehat{p(X)}$ is as close to 1 as possible

Parameter Estimation

- Goal of learning is to estimate parameter vector $\widehat{\beta}$
- Logistic Regression uses Maximum Likelihood for parameter estimation. How does this work?
  - Consider N samples with labels either 0 or 1
  - For samples labelled "1": Estimate $\widehat{\beta}$ such that $\widehat{p(X)}$ is as close to 1 as possible
  - For samples labelled "0": Estimate $\widehat{\beta}$ such that $1 - \widehat{p(X)}$ is as close to 1 as possible



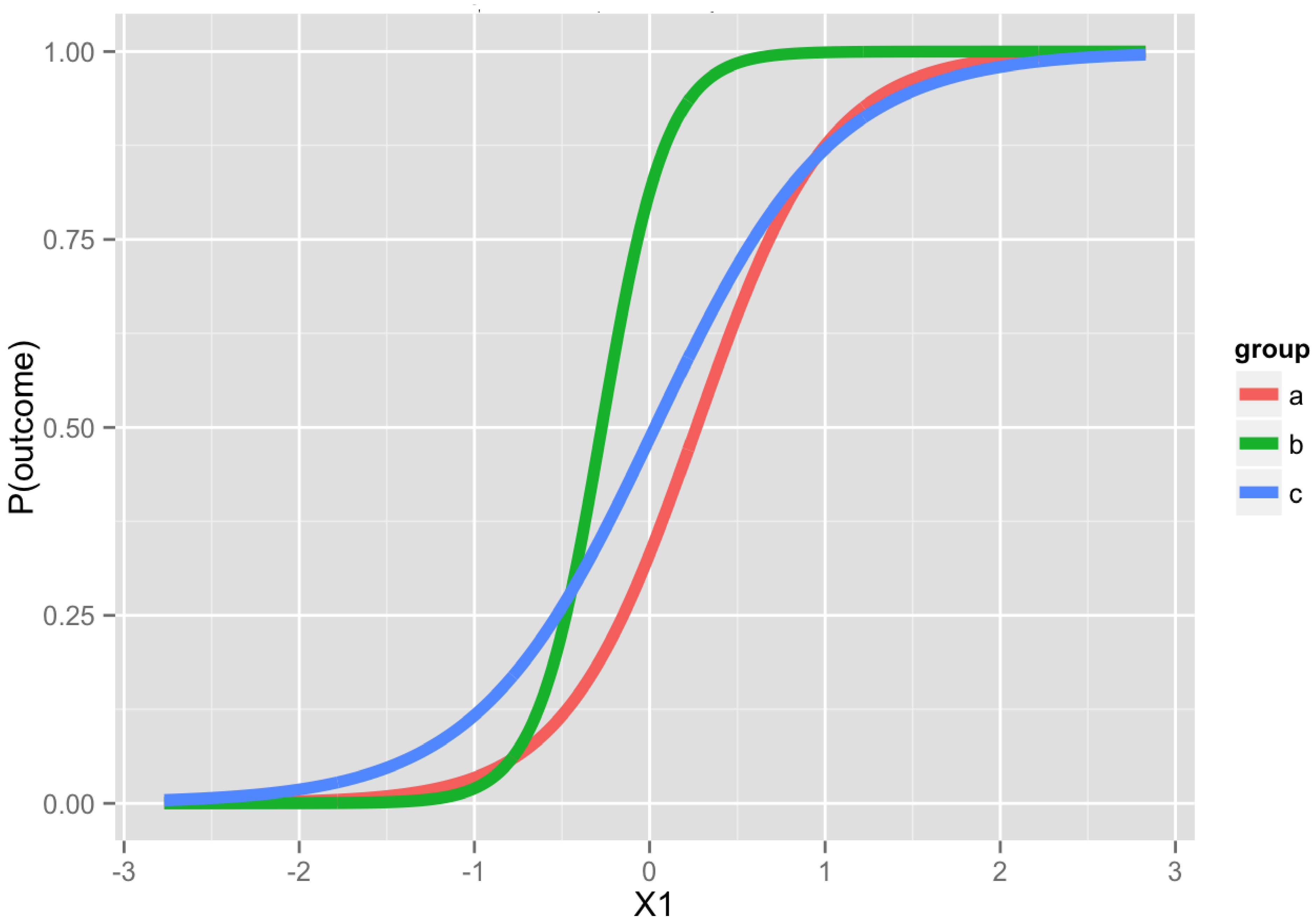
Data:

Students = {A, B, C, D}
A = Pass
B = Fail
C = Fail
D = Pass

M1:
P(A = Pass) = .85
P(B = Pass) = .25
P(C = Pass) = .45
P(D = Pass) = .76

M2:
P(A = Pass) = .94
P(B = Pass) = .23
P(C = Pass) = .10
P(D = Pass) = .91

M3:
P(A = Pass) = .75
P(B = Pass) = .64
P(C = Pass) = .39
P(D = Pass) = .47



Data:

Students = {A, B, C, D}
A = Pass
B = Fail
C = Fail
D = Pass

M1:
P(A = Pass) = .85
P(B = Pass) = .25
P(C = Pass) = .45
P(D = Pass) = .76

M2:
P(A = Pass) = .94
P(B = Pass) = .23 ✓
P(C = Pass) = .10
P(D = Pass) = .91

M3:
P(A = Pass) = .75
P(B = Pass) = .64
P(C = Pass) = .39
P(D = Pass) = .47

Parameter Estimation

- Goal of learning is to estimate parameter vector $\widehat{\beta}$

- Logistic Regression uses Maximum Likelihood for parameter estimation. How does this work?

  - Consider N samples with labels either 0 or 1

  - For samples labelled "1": Estimate $\widehat{\beta}$ such that $\widehat{p(X)}$ is as close to 1 as possible

  - For samples labelled "0": Estimate $\widehat{\beta}$ such that $1 - \widehat{p(X)}$ is as close to 1 as possible

*Note: P(yes, no, no, yes) = p(yes)*p(no)*p(no)*p(yes)*

Parameter Estimation

- Goal of learning is to estimate parameter vector $\hat{\beta}$

- Logistic Regression uses Maximum Likelihood for parameter estimation. How does this work?

  - Consider N samples with labels either 0 or 1

  - For samples labelled "1": Estimate $\hat{\beta}$ such that $\widehat{p(X)}$ is as close to 1 as possible

  - For samples labelled "0": Estimate $\hat{\beta}$ such that $1 - \widehat{p(X)}$ is as close to 1 as possible

Note: P(yes, no, no, yes) = p(yes)*p(no)*p(no)*p(yes)

$$\prod_{s\ in\ yi=1} p(x_i)$$

Parameter Estimation

- Goal of learning is to estimate parameter vector $\widehat{\beta}$

- Logistic Regression uses Maximum Likelihood for parameter estimation. How does this work?

  - Consider N samples with labels either 0 or 1

  - For samples labelled "1": Estimate $\widehat{\beta}$ such that $\widehat{p(X)}$ is as close to 1 as possible

  - For samples labelled "0": Estimate $\widehat{\beta}$ such that $1 - \widehat{p(X)}$ is as close to 1 as possible

Note: P(yes, no, no, yes) = p(yes)*p(no)*p(no)*p(yes)

$$\prod_{s\ in\ yi=1} p(x_i) \qquad \prod_{s\ in\ yi=0} (1 - p(x_i))$$

Parameter Estimation

$$L(\beta) = \prod_{s\ in\ yi=1} p(x_i) \times \prod_{s\ in\ yi=0} (1 - p(x_i))$$

Parameter Estimation

$$L(\beta) = \prod_{s\ in\ yi=1} p(x_i) \times \prod_{s\ in\ yi=0} (1 - p(x_i))$$

$$L(\beta) = \prod_{s} p(x_i)^{y_i} \times \left(1 - p(x_i)\right)^{1-y_i}$$

Parameter Estimation

$$L(\beta) = \prod_{s\ in\ yi=1} p(x_i) \times \prod_{s\ in\ yi=0} (1 - p(x_i))$$

$$L(\beta) = \prod_{s} p(x_i)^{y_i} \times \left(1 - p(x_i)\right)^{1-y_i}$$

$$l(\beta) = \sum_{i=1}^{n} y_i \log\left(p(x_i)\right) + (1 - y_i) \log\left(1 - p(x_i)\right)$$

Parameter Estimation

$$L(\beta) = \prod_{s\ in\ yi=1} p(x_i) \times \prod_{s\ in\ yi=0} (1 - p(x_i))$$

$$L(\beta) = \prod_{s} p(x_i)^{y_i} \times (1 - p(x_i))^{1-y_i}$$

$$l(\beta) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))$$

- A loss function is for a single training example. It is also sometimes called an **error function.**
- A cost function, on the other hand, is the **average loss** over the entire training dataset.
- The optimization strategies aim at minimizing the cost function.

Parameter Estimation

$$L(\beta) = \prod_{s\ in\ yi=1} p(x_i) \times \prod_{s\ in\ yi=0} (1 - p(x_i))$$

$$L(\beta) = \prod_{s} p(x_i)^{y_i} \times (1 - p(x_i))^{1-y_i}$$

$$l(\beta) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))$$

Maximizing $l(\beta)$ is equivalent to minimizing $-l(\beta)$

Gradient Descent

How?

If we expand these equations, we see the parameter β. The job is *to* Find β that minimizes the cost

β

For Linear Regression: $L = (y - f(x))^2$

For Logistic Regression: $L = -y * \log(p) - (1 - y) * \log(1 - p) = \begin{cases} -\log(1-p), & if\ y = 0 \\ -\log(p), & if\ y = 1 \end{cases}$

THANK YOU!