HOCHSCHULE DER MEDIEN

Logistic regression

With multiple predictors

Prof. Dr. Jan Kirenz
HdM Stuttgart

Logistic regression

- **Categorical** response variable
- For example with two levels, e.g., yes and no.

- Note that this is **not** regression
- This process is called **classification**

Discrimination in hiring

first_name	race	sex	first_name	race	sex	first_name	race	sex
Aisha	Black	female	Hakim	Black	male	Laurie	White	female
Allison	White	female	Jamal	Black	male	Leroy	Black	male
Anne	White	female	Jay	White	male	Matthew	White	male
Brad	White	male	Jermaine	Black	male	Meredith	White	female
Brendan	White	male	Jill	White	female	Neil	White	male
Brett	White	male	Kareem	Black	male	Rasheed	Black	male
Carrie	White	female	Keisha	Black	female	Sarah	White	female
Darnell	Black	male	Kenya	Black	female	Tamika	Black	female
Ebony	Black	female	Kristen	White	female	Tanisha	Black	female
Emily	White	female	Lakisha	Black	female	Todd	White	male
Geoffrey	White	male	Latonya	Black	female	Tremayne	Black	male
Greg	White	male	Latoya	Black	female	Tyrone	Black	male

Descriptions of nine variables from the resume dataset

variable	description
`received_callback`	Specifies whether the employer called the applicant following submission of the application for the job.
`job_city`	City where the job was located: Boston or Chicago.
`college_degree`	An indicator for whether the resume listed a college degree.
`years_experience`	Number of years of experience listed on the resume.
`honors`	Indicator for the resume listing some sort of honors, e.g. employee of the month.
`military`	Indicator for if the resume listed any military experience.
`has_email_address`	Indicator for if the resume listed an email address for the applicant.
`race`	Race of the applicant, implied by their first name listed on the resume.
`sex`	Sex of the applicant (limited to only and in this study), implied by the first name listed on the resume.

Modelling the probability of an event

Notation for a logistic regression model

- Outcome variable: Y_i:
  - **Y_i=1** (callback),
  - **Y_i=0** (no callback)
- Outcome probabilities:
  - Y_i=1 with probability **p_i** and
  - Y_i=0 with **1-p_i**

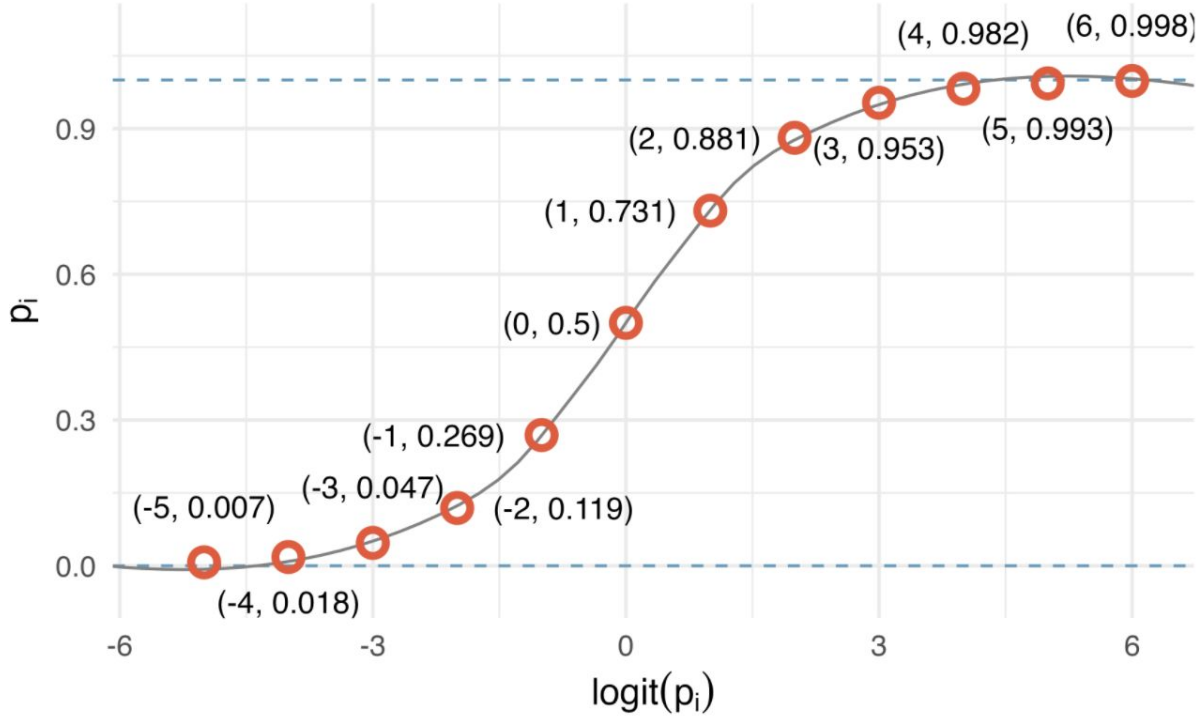
Notation for a logistic regression model

$$transformation(p_i) = b_0 + b_1x_{1,i} + b_2x_{2,i} + \cdots + b_kx_{k,i}$$

$$logit(p_i) = \log_e\left(\frac{p_i}{1-p_i}\right)$$

$$\log_e\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1x_{1,i} + b_2x_{2,i} + \cdots + b_kx_{k,i}$$

Values of p_i against values of $\text{logit}(p_i)$



Convert from values on the logistic regression scale to the probability scale

$$\log_e\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}$$

$$\frac{p_i}{1-p_i} = e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}$$

$$p_i = (1-p_i)\, e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}$$

$$p_i = e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} - p_i \times e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}$$

$$p_i + p_i\, e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}} = e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}$$

$$p_i(1 + e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}) = e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}$$

$$p_i = \frac{e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}{1 + e^{b_0 + b_1 x_{1,i} + \cdots + b_k x_{k,i}}}$$

Fitting a model with a single predictor: honors

$$\log_e\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.4998 + 0.8668 \times \textbf{honors}$$

1. If a resume is randomly selected from the study and it does not have any honors listed, what is the probability it resulted in a callback?
2. What would the probability be if the resume did list some honors?

Fitting a model with a single predictor: honors

1.

$$p_i: \frac{e^{-2.4998}}{1+e^{-2.4998}} = 0.076$$

$$\hat{p}_i = 0.076$$

2.

$$-2.4998 + 0.8668 \times 1 = -1.6330$$

$$\hat{p}_i = 0.163$$

Logistic model with many variables

Summary table for the full logistic regression model for the resume callback example

term	estimate	std.error	statistic	p.value
(Intercept)	-2.66	0.18	-14.64	<0.0001
job_cityChicago	-0.44	0.11	-3.85	1e-04
college_degree1	-0.07	0.12	-0.55	0.5821
years_experience	0.02	0.01	1.96	0.0503
honors1	0.77	0.19	4.14	<0.0001
military1	-0.34	0.22	-1.59	0.1127
has_email_address1	0.22	0.11	1.93	0.0541
raceWhite	0.44	0.11	4.10	<0.0001
sexman	-0.18	0.14	-1.32	0.1863

Model selection with AIC

- We use a statistic called **Akaike information criterion** (AIC)
- **AIC** is a popular model selection method
- Analogous to how we used adjusted R2 in multiple regression.

AIC provides information about the quality of a model **relative** to other models, but does not provide information about the overall quality of a model.

Model selection with AIC

- AIC selects a “best” model by ranking models from best to worst according to their AIC values.
- A penalty is given for including additional variables.
- Attempts to strike a balance between **underfitting** (too few variables in the model) and **overfitting** (too many variables in the model).

- Models with a **lower AIC** value are considered to be “**better**.”

Variable selection has been performed using AIC (college degree is dropped)

term	estimate	std.error	statistic	p.value
(Intercept)	-2.72	0.16	-17.51	<0.0001
job_cityChicago	-0.44	0.11	-3.83	1e-04
years_experience	0.02	0.01	2.02	0.043
honors1	0.76	0.19	4.12	<0.0001
military1	-0.34	0.22	-1.60	0.1105
has_email_address1	0.22	0.11	1.97	0.0494
raceWhite	0.44	0.11	4.10	<0.0001
sexman	-0.20	0.14	-1.45	0.1473

The **race variable** had taken only two levels: Black and White.

Based on the model results, what does the coefficient of this variable say about callback decisions?

Variable selection has been performed using AIC (college degree is dropped)

term	estimate	std.error	statistic	p.value
(Intercept)	-2.72	0.16	-17.51	<0.0001
job_cityChicago	-0.44	0.11	-3.83	1e-04
years_experience	0.02	0.01	2.02	0.043
honors1	0.76	0.19	4.12	<0.0001
military1	-0.34	0.22	-1.60	0.1105
has_email_address1	0.22	0.11	1.97	0.0494
raceWhite	0.44	0.11	4.10	<0.0001
sexman	-0.20	0.14	-1.45	0.1473

The coefficient shown corresponds to the level of White, and it is **positive**.

This positive coefficient reflects a positive gain in callback rate for resumes where the candidate's first name implied they were White.

Estimate the probability of receiving a callback for a job in Chicago where the candidate lists **14 years experience**, **no honors**, **no military experience**, **includes an email address,** and has a first name that implies they are a **white** male.

1

$$
\begin{aligned}
&log_e\left(\frac{p}{1-p}\right) \\
&= -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} \\
&\quad + 0.0206 \times \texttt{years\_experience} \\
&\quad + 0.7634 \times \texttt{honors} - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} \\
&\quad + 0.4429 \times \texttt{race}_{\texttt{White}} - 0.1959 \times \texttt{sex}_{\texttt{man}}
\end{aligned}
$$

Estimate the probability of receiving a callback for a job in Chicago where the candidate lists **14 years experience**, **no honors**, **no military experience**, **includes an email address,** and has a first name that implies they are a **white** male.

1

$$
\begin{aligned}
&log_e\left(\frac{p}{1-p}\right) \\
&= -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} \\
&\quad + 0.0206 \times \texttt{years\_experience} \\
&\quad + 0.7634 \times \texttt{honors} - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} \\
&\quad + 0.4429 \times \texttt{race}_{\texttt{White}} - 0.1959 \times \texttt{sex}_{\texttt{man}}
\end{aligned}
$$

2

$$
\begin{aligned}
&log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) \\
&= -2.7162 - 0.4364 \times 1 + 0.0206 \times 14 \\
&\quad + 0.7634 \times 0 - 0.3443 \times 0 + 0.2221 \times 1 \\
&\quad + 0.4429 \times 1 - 0.1959 \times 1 = -2.3955
\end{aligned}
$$

Estimate the probability of receiving a callback for a job in Chicago where the candidate lists **14 years experience**, **no honors**, **no military experience**, **includes an email address,** and has a first name that implies they are a **white** male.

1

$$
\begin{aligned}
&log_e\left(\frac{p}{1-p}\right) \\
&= -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} \\
&\quad + 0.0206 \times \texttt{years\_experience} \\
&\quad + 0.7634 \times \texttt{honors} - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} \\
&\quad + 0.4429 \times \texttt{race}_{\texttt{White}} - 0.1959 \times \texttt{sex}_{\texttt{man}}
\end{aligned}
$$

2

$$
\begin{aligned}
&log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) \\
&= -2.7162 - 0.4364 \times 1 + 0.0206 \times 14 \\
&\quad + 0.7634 \times 0 - 0.3443 \times 0 + 0.2221 \times 1 \\
&\quad + 0.4429 \times 1 - 0.1959 \times 1 = -2.3955
\end{aligned}
$$

3

$$\frac{e^{-2.3955}}{1+e^{-2.3955}} = 8.35\%.$$

Akaike information criterion

generalized linear model

logistic regression

logit transformation

probability of an event