Logistic regression

Overview

Information on the final project

Brief mention: Visualizing linear models using ggplot

Logistic regression

If there is time: Poisson regression

Final projects!

The final project is a 5-8 page R Markdown report where you analyze your own data to address a question that you find interesting

- It's a chance to practice everything you've learned in class!

The goal of the project is to present a clear and compelling analyses of data showing a few interesting results!

A few sources for data sets are listed on Canvas

- You can use data you collect as well. If you use data for another class your work must be unique for each class.

Final projects!

An R Markdown template describing sections in the project is on the class GitHub site.

- library(SDS230)
- download_any_file("homework/final_project.Rmd")

A challenge is going to be to fit your analyses into 5-8 pages:

- You can include an appendix with additional code that does not count against your 5-8 pages
  - E.g., you can include functions in your appendix and then just call them in the body of your report

Questions?

Project is due at 11pm on Sunday December 10th

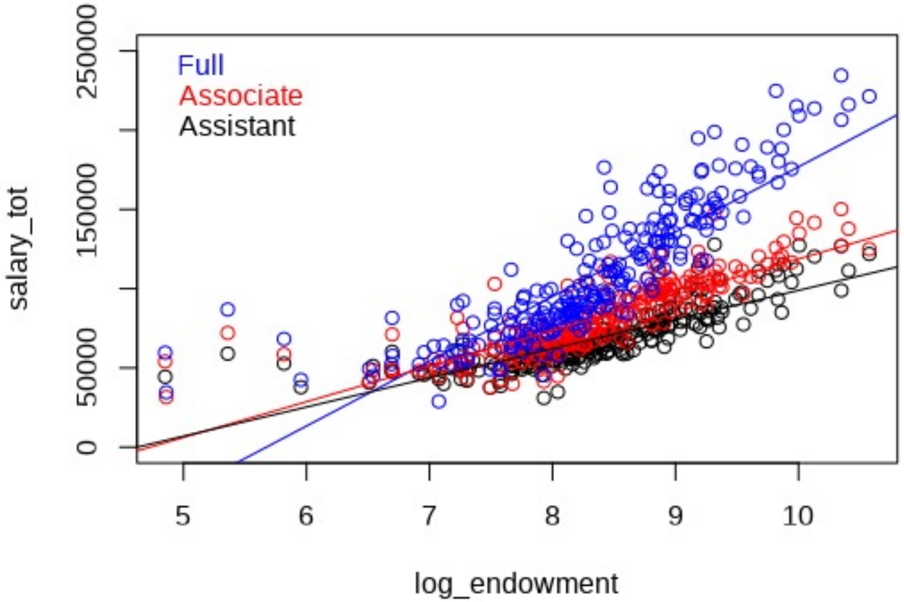
- i.e., the day before the start of reading period

Very quick review of multiple regression

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_3 (x_1 \cdot x_2) + \epsilon$$

There are many uses for multiple regression models:

- To make predictions as accurately as possible
- To understand which predictors (x) are related to the response variable (y)

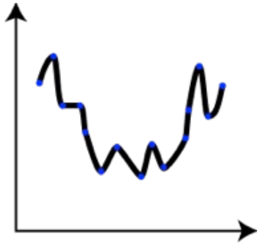


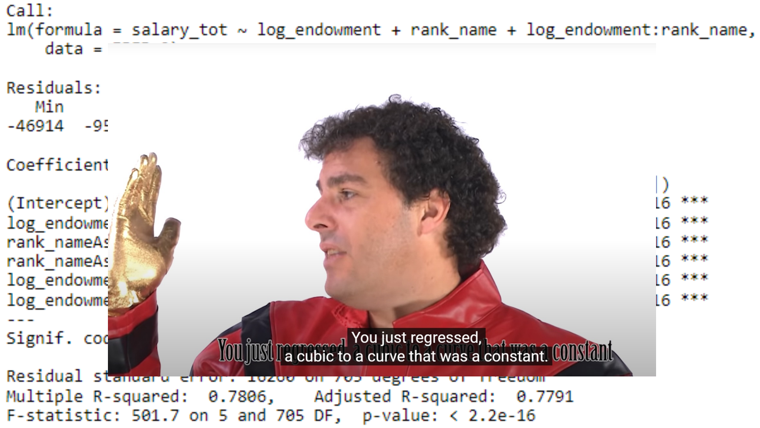
We can have categorical predictors and interactions

We can fit nonlinear functions

There are methods/statistics that help us choose between models

- Adjusted R^2, AIC, BIC, cross-validation



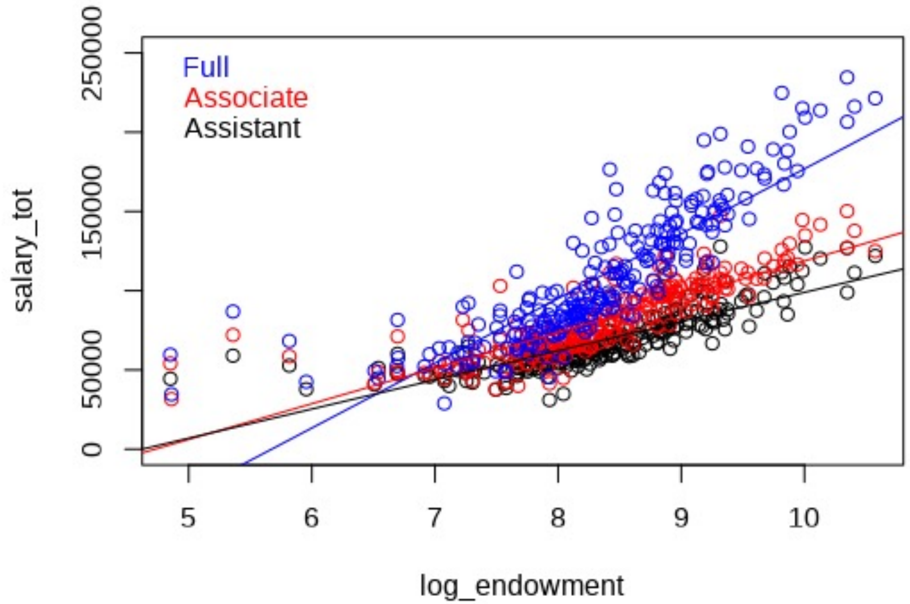


Plotting multiple regression models with ggplot

So far we have plotted our multiple regression models using base R graphics

This was useful for seeing the relationship between how R fits linear models, and what these models represent

However, if you want an easier/prettier way to visualize linear models, we can use ggplot!




```
Call:
lm(formula = salary_tot ~ log_endowment + rank_name + log_endowment:rank_name,
    data = IPED_2)
```

Let’s try it in R!

```
Residual standard error: 16260 on 705 degrees of freedom
Multiple R-squared:  0.7806,    Adjusted R-squared:  0.7791
F-statistic: 501.7 on 5 and 705 DF,  p-value: < 2.2e-16
```

Logistic regression

Logistic regression

In **logistic regression** we try to predict whether a case belongs to one of two categories

- Does a case below to category *1* or category *0*?
- Example: based on the salary level, can we predict if a faculty member is an Assistant of Full professor?

Making predictions for a categorical variable is called **classification**

- The field of Machine Learning has developed many classification methods

In logistic regression we build a conditional probability model:

- P(Class = 1 | x)
- P(Full Professor | salary = $80,000)

Logistic regression

Question: could we use linear regression to make these predictions?

$$P(Y = 1 \mid x_1) = \beta_0 + \beta_1 x_1$$

Problem: we could get negative probabilities and probabilities greater than 1!

Logistic regression

Question: what if we transformed the probability to odds?

$$\frac{P(Y = 1 \mid x_1)}{P(Y = 0 \mid x_1)}$$

Question: what are the range of values odds can take on?

A: 0 to ∞

Logistic regression

Instead, we model the log odds as a linear function of our predictors

$$log(\frac{P(Y=1|x)}{1-P(Y=1|x)}) = \beta_0 + \beta_1 \cdot x$$

log-odds or logit

This scales values in the range of [0 1] to values in the range of (-∞ ∞)

Generalized linear models

Generalized linear models use a linear combinations of predictors to predict ***a function of the mean***

If Y is a binary response variable (Y = 0 or 1)
P(Y = 1|x) is the mean of Y

"Linear predictor"

$$log(\frac{P(Y=1|x)}{1-P(Y=1|x)}) = \beta_0 + \beta_1 \cdot x$$

The logit function (log-odds) is a "link function" that links the mean to the linear predictor

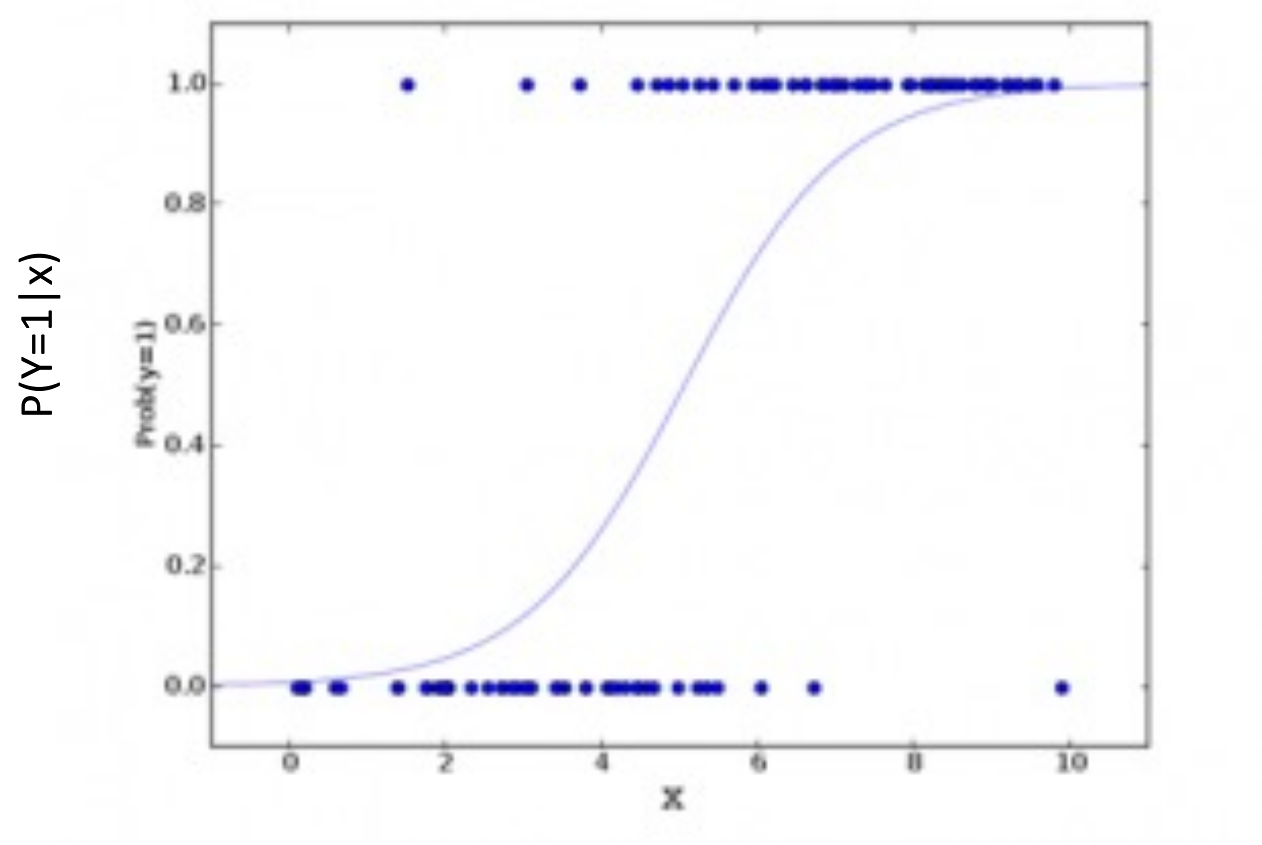
Logistic function

$$log(\frac{P(Y=1|x)}{1-P(Y=1|x)}) = \beta_0 + \beta_1 \cdot x$$

Solving for P(Y = 1| x) we get the "inverse link" function, which in the case of logistic regression is called a ***logistic function***

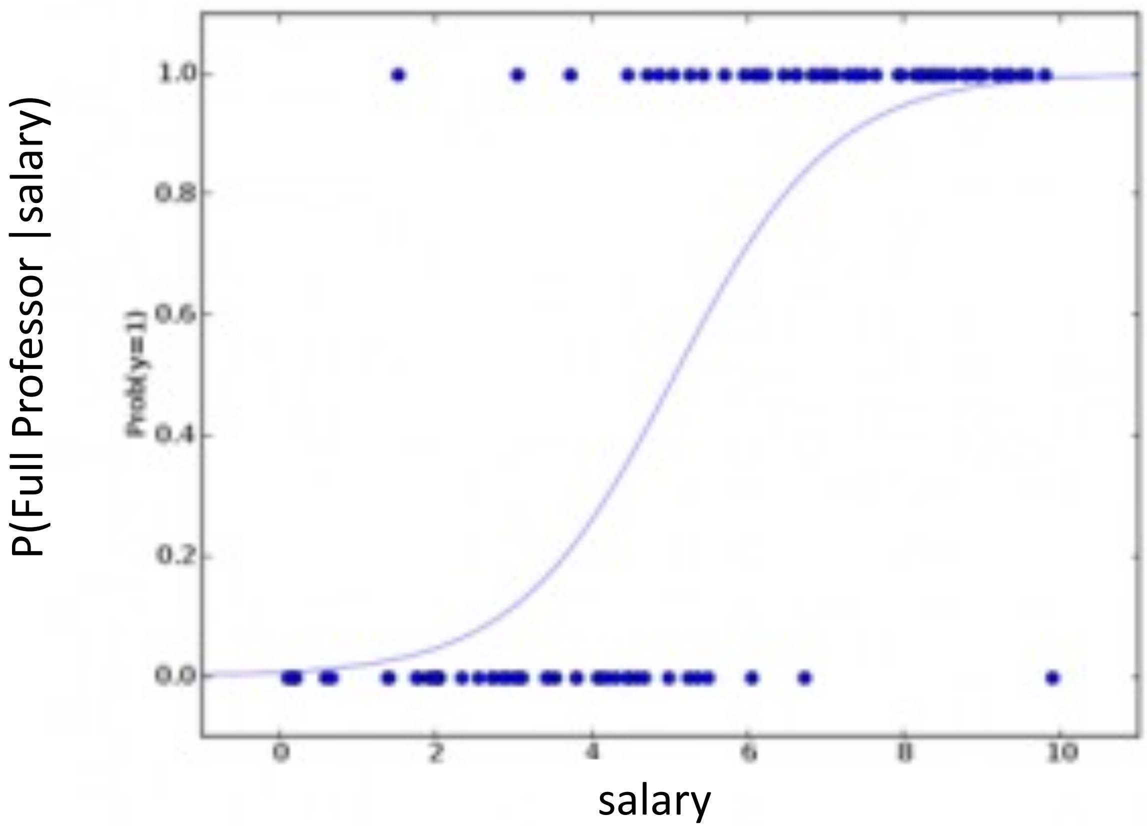
$$P(Y = 1|x) = \frac{\exp(\beta_0+\beta_1 \cdot x_1)}{1+\exp(\beta_0+\beta_1 \cdot x_1)} = \frac{e^{\beta_0+\beta_1 \cdot x_1}}{1+e^{\beta_0+\beta_1 \cdot x_1}}$$

Plotting the logistic function



$$P(Y = 1|x_1) = \frac{e^{\beta_0 + \beta_1 \cdot x_1}}{1 + e^{\beta_0 + \beta_1 \cdot x_1}}$$

Plotting the logistic function



$$P(\text{Full Professor} \mid \text{salary}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{salary}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{salary}}}$$

Multivariate logistic regression

We can easily extend our logistic regression model to include multiple explanatory variables

$$log(\frac{P(Y=1|x)}{1-P(Y=1|x)}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + ... + \hat{\beta}_k \cdot x_k$$

We can also use categorical predictors via dummy variable encoding as we did for regular multiple linear regression

Interpreting categorical predictors

When using a categorical predictor, x_2, in a logistic regression model, the exponential of the regression coefficient $e^{\hat{\beta}_2}$ is the ***odds ratio***

- Tells us how many times greater the odds are when $x_2 = 1$ vs. when $x_2 = 0$

$$log(\frac{P(Y=1|x_1,x_2)}{1-P(Y=1|x_1,x_2)}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2$$

Dummy variable

If $x_2 = 1$ $\quad \frac{P(Y|x_1,x_2=1)}{1-P(Y|x_1,x_2=1)} = e^{\hat{\beta}_0} e^{\hat{\beta}_1 \cdot x_1} e^{\hat{\beta}_2}$

If $x_2 = 0$ $\quad \frac{P(Y|x_1,x_2=0)}{1-P(Y|x_1,x_2=0)} = e^{\hat{\beta}_0} e^{\hat{\beta}_1 \cdot x_1}$

$$\frac{\;\frac{P(Y|x_1,x_2=1)}{1-P(Y|x_1,x_2=1)}\;}{\frac{P(Y|x_1,x_2=0)}{1-P(Y|x_1,x_2=0)}} = \frac{e^{\hat{\beta}_0} e^{\hat{\beta}_1 \cdot x_1} e^{\hat{\beta}_2}}{e^{\hat{\beta}_0} e^{\hat{\beta}_1 \cdot x_1}} = e^{\hat{\beta}_2}$$

Let’s look at this in R…

Poisson regression

Summary of linear regression

We can summarize the linear regression model as:

$$Y_i \;=\; \mu_i \;+\; \varepsilon_i \qquad\qquad \text{where} \quad \varepsilon_i \sim N(0, \sigma_\varepsilon)$$

$$\mu_i \;=\; \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Equivalently, $Y_i \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma_\varepsilon)$

Generalized linear models

We can summarize the linear regression model as:

$$Y_i \;=\; \mu_i \;+\; \varepsilon_i \qquad\qquad \text{where} \quad \varepsilon_i \sim N(0, \sigma_\varepsilon)$$

$$\mu_i \;=\; \beta_0 \;+\; \beta_1 x_1 \;+ \ldots \;+ \beta_k x_k$$

In generalized linear models, we generalize the model to:

$$Y_i \;\sim\; f(y|\theta_i) \qquad\qquad \text{where } f(y|\theta_i) \text{ is some probability distribution}$$

$$\theta_i \;=\; g^{-1}(\beta_0 \;+\; \beta_1 x_1 \;+ \ldots \;+ \beta_k x_k)$$

g^{-1} is called an "inverse link function"
Links "linear predictor" to parameters

We choose a particular "family" of distributions (e.g., Poisson, binomial, etc.)

Example: logistic regression

In logistic regression we model whether a case belongs to one of two categories

- P(Y = 0 | **x**) or P(Y = 1| **x**)

The logit function (log-odds) is a "link function"

"Linear predictor"

$$log(\frac{P(Y=1|x)}{1-P(Y=1|x)}) = \beta_0 + \beta_1 \cdot x$$

Inverse link function (logistic function)

Solving for P(Y = 1 | x)

$$P(Y = 1|x_1) = \frac{e^{\beta_0+\beta_1 \cdot x_1}}{1+e^{\beta_0+\beta_1 \cdot x_1}}$$

Family is Bernoulli distribution (binomial with n = 1)

$$Y_i \sim Bernoulli(P(Y = 1|x))$$

R: `glm_fit <- glm(y ~ x, family = binomial(link = logit))`

Poisson regression

In Poisson regression we model counts

- i.e., integer values: 0, 1, 2, 3, …

The log is the "link function"

"Linear predictor"

$$log(\lambda) = \beta_0 + \beta_1 \cdot x$$

Solving for λ

Inverse link function (exponential function)

$$\lambda = e^{\beta_0 + \beta_1 \cdot x_1}$$

Family is Poisson distributions

$$Y_i \sim Poisson(\lambda))$$

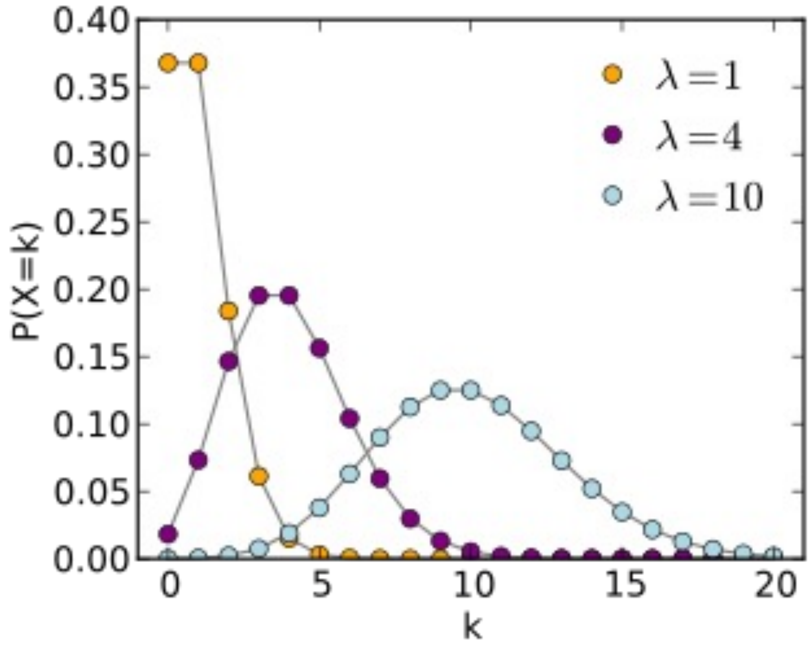
R: glm_fit <- glm(y ~ x, family = Poisson(link = log))

Poisson distributions

A Poisson distribution is a probability distribution over non-negative integers

- i.e., over values 0, 1, 2, 3, …

Poisson distributions have a single parameter λ



$$X \sim Pois(\lambda)$$

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}\ ,\ k = 0, 1, 2, \ldots$$

- Density: dpois()
- Cumulative distribution: ppois()
- Random number: rpois()

Poisson processes

Poisson distributions models the number of outcomes that have occurred from a **Poisson process**

A **Poisson process** is a stochastic process where:

- Events (random outcome) occur at a fixed rate (λ)
- Every event is independent of the other events

Examples of Poisson processes?

Side note: Maximum likelihood estimate (MLE)

When building regression models, we need a way to estimate parameters

The "true" underlying model is:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots \beta_k \cdot x_k + \epsilon$$

We estimate coefficients using a data set to make predictions ŷ

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \ldots + \hat{\beta}_k \cdot x_k$$

For GLMs, the maximum likelihood estimates (MLE) is used to estimate the regression coefficients:

- MLEs find the parameters that make the data as likely as possible
  - (For linear regression with normal errors, MLE is gives the same coefficient estimates as least squares)

Example: Roy Kent saying f#ck

Ted Lasso was a Apple TV+ series that aired from July 2021 to March 2023

One of the main characters on the show was Roy Kent, who tended to say f#ck frequently

In different episodes of the show Roy was:

- A coach
- Dated Keeley Jones

Let's use Poisson regression to assess if Roy said f#ck more when he was coaching and/or when he was dating Keeley

Roy Kent

Keeley Jones

Example from season 2

Let's try it in R...