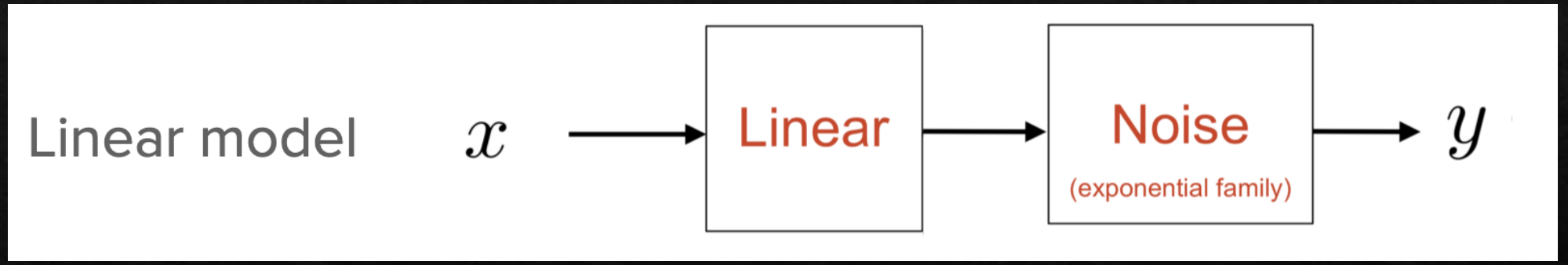
Regression - practicals

BRSM

Agenda

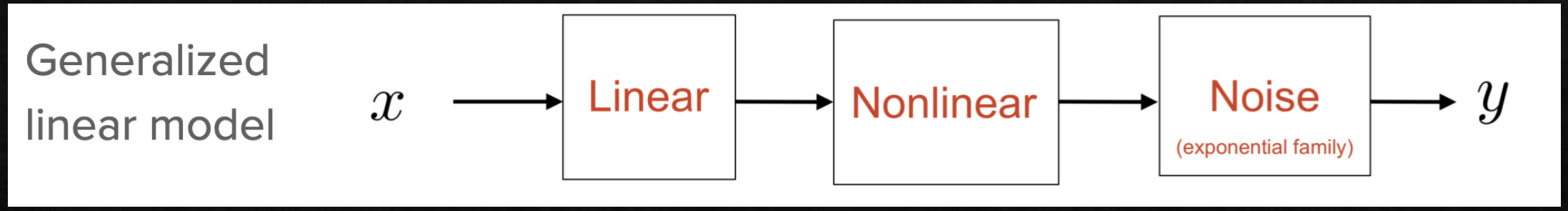
- GLM
- Assignment

Linear Model



Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

Generalized Linear Model



Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

Generalized Linear Model

Example: nonlinear Gaussian model $y = f(\theta x) + \eta$ where $\eta \sim \mathcal{N}(0, \sigma^2)$

nonlinear

f^{-1}: link function

Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

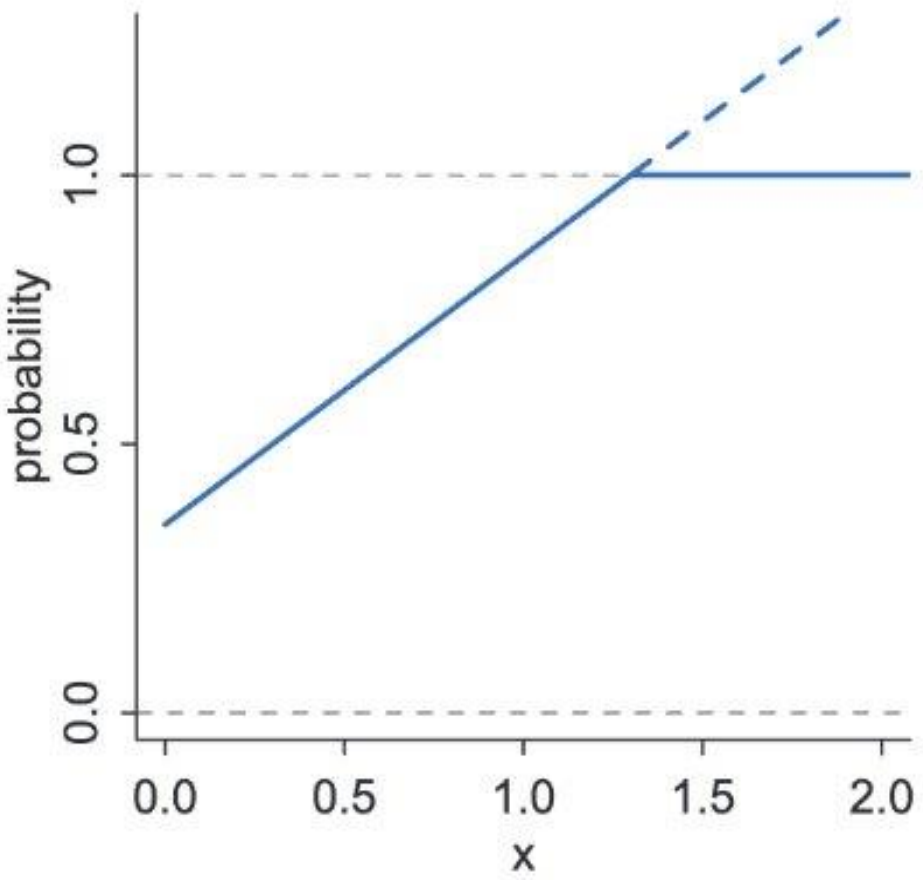


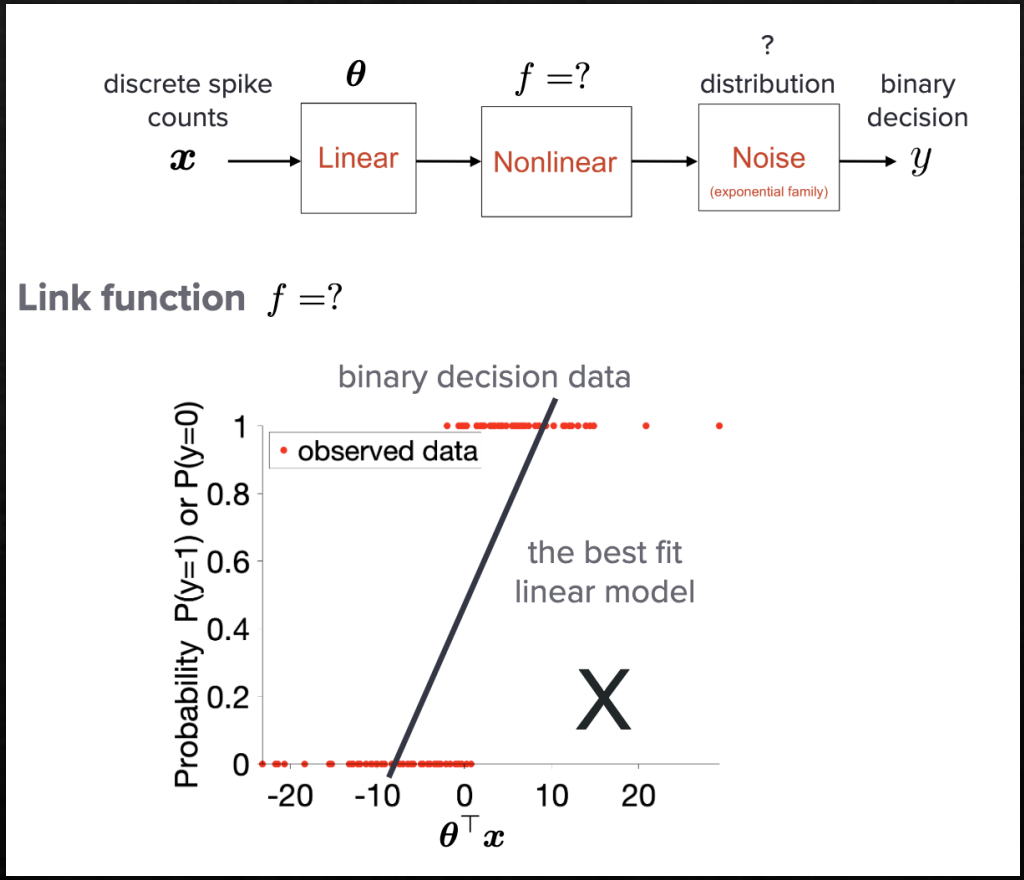
FIGURE 9.5. Why we need link functions. The solid blue line is a linear model of a probability mass. It increases linearly with a predictor, x, on the horizontal axis. But when it reaches the maximum probability mass of 1, at the dashed boundary, it will happily continue upwards, as shown by the dashed blue line. In reality, further increases in x could not further increase probability, as indicated by the horizontal continuation of the solid trend.

https://osf.io/2h6ut

$$y_i \sim \text{Binomial}(n, p_i)$$
$$f(p_i) = \alpha + \beta x_i$$

https://osf.io/2h6ut

Binary Outcome Variable

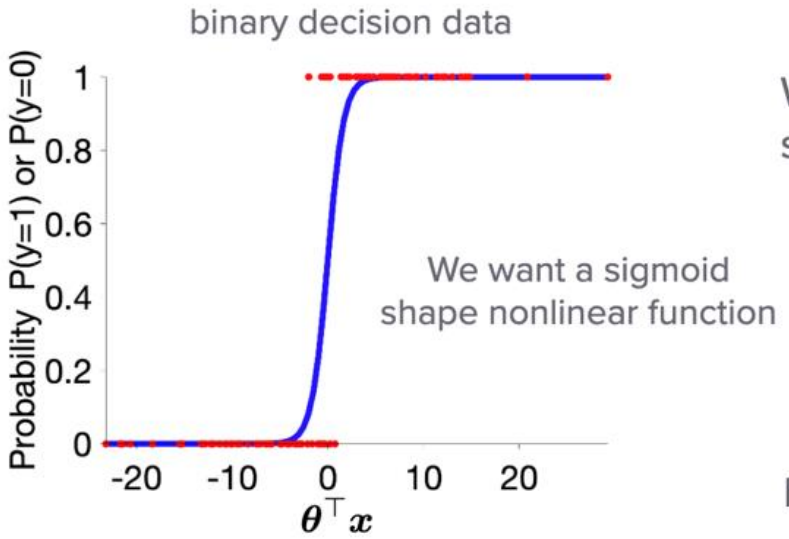


Good mathy introduction: https://www.youtube.com/watch?v=X-ix97pw0xY&ab_channel=MITOpenCourseWare

Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

Binary Outcome Variable

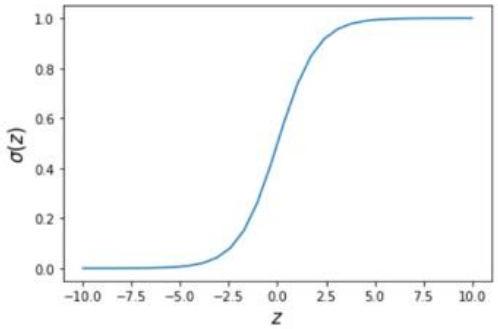
Link function $f = ?$



We want a sigmoid shape nonlinear function

We define $f = \sigma(\cdot)$ which is a "squashing" function called the sigmoid function.

$$\sigma(z) = \frac{1}{\exp(-z) + 1}$$



Notice $0 \leq f(\boldsymbol{\theta}^\top \boldsymbol{x}) = \frac{1}{\exp(-\boldsymbol{\theta}^\top \boldsymbol{x}) + 1} \leq 1$

Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

Binary Outcome Variable

Distribution of the observation noise

$$f(\boldsymbol{\theta}^\top \boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) = \frac{1}{\exp(-\boldsymbol{\theta}^\top \boldsymbol{x}) + 1}$$ only gives us a probability-like value, not a binary decision.

Bernoulli distribution: generate a binary value with some input probability value.

single coin flip

outcome y:	head	tail
probability:	p	1-p

The probability mass function for y is

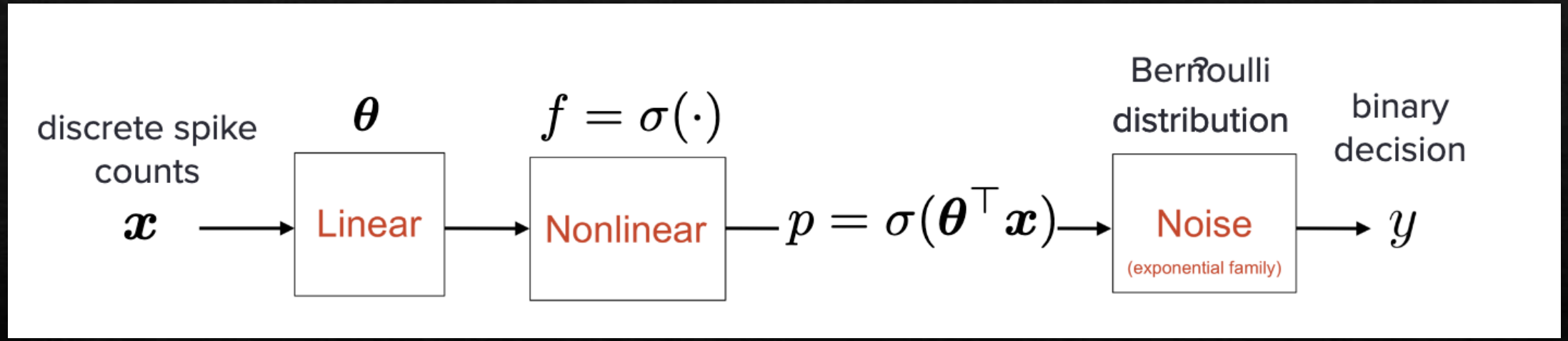
$$P(y|p) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \end{cases}$$

Alternatively,

$$P(y|p) = p^y (1-p)^{1-y}$$

Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

Binary Outcome Variable



Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

Binary Outcome Variable

Bernoulli GLM:

(coin flipping model, y = 0 or 1)

nonlinearity

$$p_t = f(\vec{x}_t \cdot \vec{k})$$

probability of spike at bin t

$$p(y_t = 1|\vec{x}_t) = p_t$$

Logistic regression:

$$f(x) = \frac{1}{1 + e^{-x}}$$

logistic function

- so logistic regression is a special case of a Bernoulli GLM

Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

Binary Outcome Variable: Logistic Regression

$\text{logit}(p) = \beta_0 + \beta_1 * \textbf{female}$

```
Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       3.10
                                                  Prob > chi2     =     0.0781
Log likelihood = -109.80312                       Pseudo R2       =     0.0139

------------------------------------------------------------------------------
         hon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    .5927822   .3414294     1.74   0.083    -.0764072    1.261972
   intercept |   -1.470852   .2689555    -5.47   0.000    -1.997995   -.9437087
------------------------------------------------------------------------------
```

$$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 * \textbf{math} + \beta_2 * \textbf{female} + \beta_3 * \textbf{read}$$

$$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 * \textbf{female} + \beta_2 * \textbf{math} + \beta_3 * \textbf{female} * \textbf{math}$$

Source: http://pillowlab.princeton.edu/teaching/statneuro2018/slides/lec09_GLMs1.pdf

Classic Linear Model to Generalized Linear Model

LM:

1) *Random Component* : Each component of $\underline{Y}$ is independent and normally distributed. The mean μ_i allowed to differ, but all Y_i have common variance σ_e^2
2) *Systematic Component* : The n covariates combine to give the "linear predictor"

$$\underline{\eta} = \underline{\beta}\,\mathbf{X}$$

3) *Link Function* : The relationship between the random and systematic components is specified via a link function. In linear model, link function is identity fnc.

$$E[\underline{Y}] = \underline{\mu} = \underline{\eta}$$

GLM:

1) *Random Component* : Each component of $\underline{Y}$ is independent and from one of the exponential family of distributions
2) *Systematic Component* : The n covariates are combined to give the "linear predictor"

$$\underline{\eta} = \underline{\beta}\,\mathbf{X}$$

3) *Link Function* : The relationship between the random and systematic components is specified via a link function g, that is differentiable and monotonic

$$E[\underline{Y}] = \underline{\mu} = g^{-1}(\underline{\eta})$$

https://www.casact.org/sites/default/files/presentation/rpm_2009_handouts_havlicek.pdf

Poisson Regression

models how the mean of a discrete (count) response variable Y depends on a set of explanatory variables

$$\log \lambda_i = \beta_0 + \beta x_i$$

- **Random component** - The distribution of Y is Poisson with mean λ.
- **Systematic component** - x is the explanatory variable (can be continuous or discrete) and is linear in the parameters. As with the above example, this can be extended to multiple variables of non-linear transformations.
- **Link function** - the log link is used.

Binary Logistic Regression

Binary logistic regression models how the odds of "success" for a binary response variable Y depend on a set of explanatory variables:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

- **Random component** - The distribution of the response variable is assumed to be binomial with a single trial and success probability $E(Y) = \pi$.
- **Systematic component** - x is the explanatory variable (can be continuous or discrete) and is linear in the parameters. As with the above example, this can be extended to multiple variables of non-linear transformations.
- **Link function** - the log-odds or logit link, $\eta = g(\pi) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, is used.

Lab Practice + Assignment

- Housing.csv

longitude

latitude

housing*median*age

total_rooms

total_bedrooms

population

households

median_income

median*house*value

ocean_proximity

Assignment (15 marks)
Part 1 (10 marks)

- Visualize some correlations between variables in the data set (2 marks)
- Pick 2 linear regression models (i.e., sets of predictors) to predict median house value
- Check for collinearity using VIF to remove highly correlated variables from the models (1 mark)
- Plot the distribution of the residuals against the fitted values to check for heteroscedasticity (1 mark)
- Use ncvTest or equivalent to test for heteroscedasticity (1 mark) (https://www.rdocumentation.org/packages/car/versions/3.0-12/topics/ncvTest)
- Test for normality of the residuals (use at least one of Wald test, Q-Q plots, etc). 1 mark
- Compare the 2 models using AIC and pick the best model. 1 mark
- Report the coefficients of the winning model and their statistics (including confidence intervals) and interpret the resulting model coefficients. 3 marks

Part 2 (5 marks)

- Binary.csv
- Predict admission using GRE, GPA, and undergrad institution ranks
- Admission = 1 or 0. Hence use logistic regression (GLM)
- Report the statistics, confidence intervals, etc for the logistic regression and interpret the results (what are the most significant variables that predict whether someone will get admitted? Explain in terms odds by exponentiating the coefficients) - 3 marks
- Can you test an interaction effect? Let's say GPA matters even more if you are from a lower ranked institution (lower GPAs may be tolerated if you are from a higher ranked institution). So include a GPA*rank term in the model and try to interpret the resulting coefficient. - 2 marks

Logistic Regression – Qs with binary or binomial responses

- Are students with poor grades more likely to binge watch Netflix series?
- Is exposure to a particular chemical associated with a cancer diagnosis?
- Are the number of votes for a political candidate associated with the amount of money raised by their party?

- Binomial responses: # of successes in N independent trials, each with probability p of success.

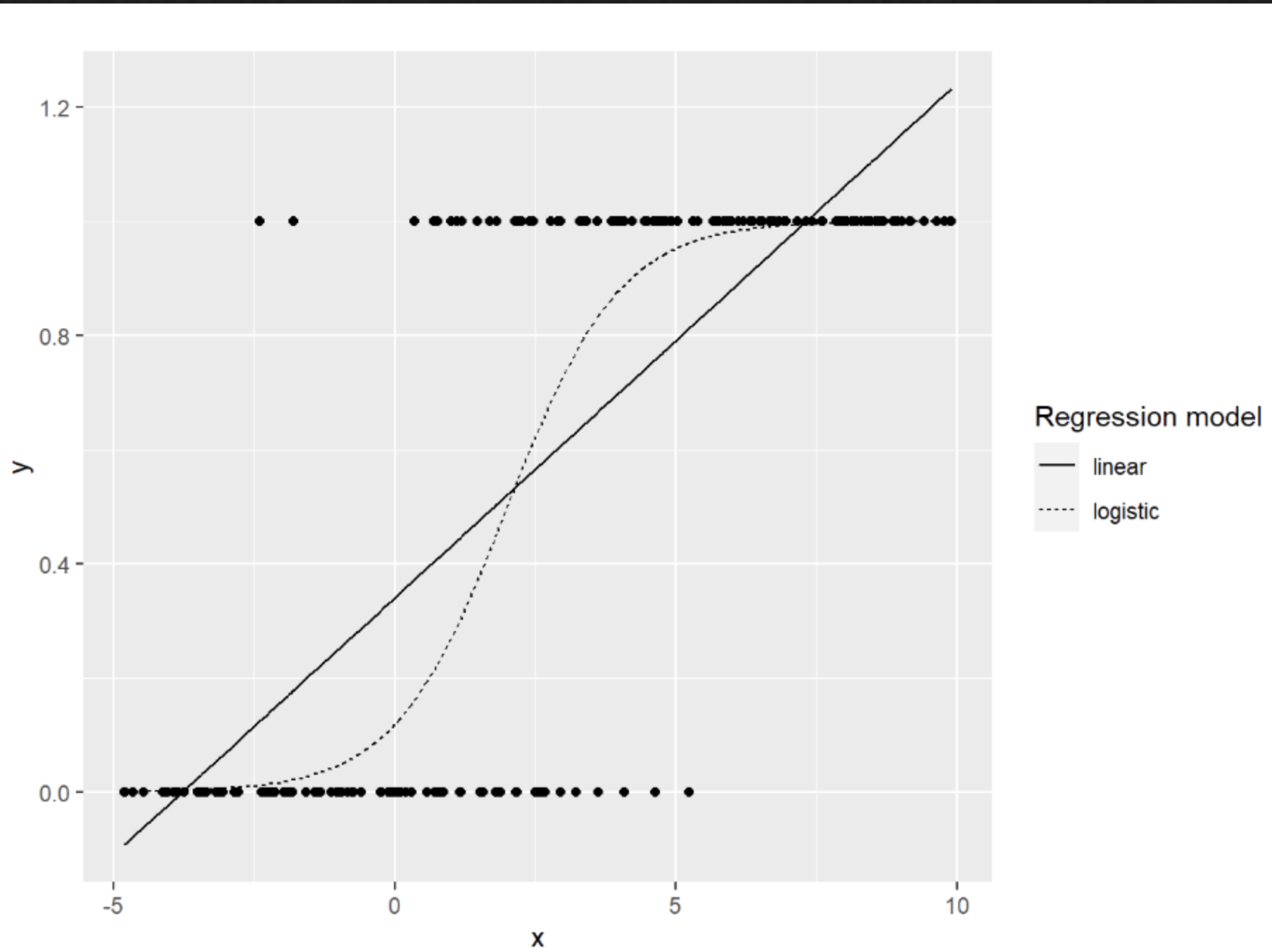
Logistic Regression – Assumptions

Binary Response - The response variable is dichotomous (two possible responses) or the sum of dichotomous responses.

Independence - The observations must be independent of one another.

Variance Structure - By definition, the variance of a binomial random variable is $np(1-p)$, so that variability is highest when $p=.5$.

Linearity - The log of the odds ratio, $\log(p/1-p)$, must be a linear function of x.


1.2
0.8
0.4
0.0
y
-5
0
5
10
x
Regression model
linear
logistic

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_i$$

where the observed values $Y_i \sim$ binomial with $p = p_i$ for a given x_i and $n = 1$ for binary responses.

Do goalkeepers save more goals when their team is behind?

	Saves	Scores	Total
Behind	2	22	24
Not Behind	39	141	180
Total	41	163	204

(Source: Roskes et al. 2011.)

Modeling Odds

- How can we quantify the goalkeeper's performance?
- Odds that he saves a penalty kick when his team is behind = 2/22

$$\text{Odds} = \frac{\#\text{successes}}{\#\text{failures}} = \frac{\#\text{successes}/n}{\#\text{failures}/n} = \frac{p}{1-p}.$$

Modeling Odds

- However, Odds are strictly positive. Cannot model it directly as a linear function since we want to model something that can take values from -inf to +inf.
- So, we will model log(odds).

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

Modeling Odds

- define X=0 for not behind and X=1 for being behind in the game.

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

$$\log\left(\frac{p_0}{1-p_0}\right) = \beta_0,$$

Modeling Odds

- X=0 for not behind

$$\log\left(\frac{p_0}{1-p_0}\right) = \beta_0,$$

Modeling Odds

- X=1 for being behind

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1.$$

Modeling Odds

- What does β1 stand for?
- $e^{\beta 1}$ = odds ratio (ratio of odds of success under one condition and the other condition)

$$\beta_1 = (\beta_0 + \beta_1) - \beta_0 = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) = \log\left(\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right).$$

Logistic Regression: Estimating coefficients

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

$$p_X = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Table 6.1: Soccer goalkeepers' penalty kick saves when their team is and is not behind.

	Saves	Scores	Total
Behind	2	22	24
Not Behind	39	141	180
Total	41	163	204

(Source: Roskes et al. 2011.)

$$\text{Lik}(p_1, p_0) = \binom{24}{22} p_1^{22}(1-p_1)^2 \binom{180}{141} p_0^{141}(1-p_0)^{39}$$

Logistic Regression: Estimating coefficients - MLE

$$\text{Lik}(p_1, p_0) = \binom{24}{22} p_1^{22}(1-p_1)^2 \binom{180}{141} p_0^{141}(1-p_0)^{39}$$

$$\text{Lik}(\beta_0, \beta_1) \propto \left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)^{22} \left(1-\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)^{2} \left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^{141} \left(1-\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)^{39}$$

$$\hat{\beta}_0 = 1.2852 \text{ and } \hat{\beta}_1 = 1.1127.$$

Logistic Regression: Interpreting the coefs

$$\hat{\beta}_0 = 1.2852 \text{ and } \hat{\beta}_1 = 1.1127.$$

Exponentiate **β1 to get odds ratio**
Odds ratio ~ 3.
Three times likely to score when the goalkeeper's team is behind compared to when they're ahead.

CI, p values, model comparisons using AIC/BIC etc as discussed before.

READ: https://bookdown.org/roback/bookdown-BeyondMLR/ch-logreg.html#introduction-to-logistic-regression

Part 2 (5 marks)

- Binary.csv
- Predict admission using GRE, GPA, and undergrad institution ranks
- Admission = 1 or 0. Hence use logistic regression (GLM)
- Report the statistics, confidence intervals, etc for the logistic regression and interpret the results (what are the most significant variables that predict whether someone will get admitted? Explain in terms odds by exponentiating the coefficients) - 3 marks
- Can you test an interaction effect? Let's say GPA matters even more if you are from a lower ranked institution (lower GPAs may be tolerated if you are from a higher ranked institution). So include a GPA*rank term in the model and try to interpret the resulting coefficient. - 2 marks