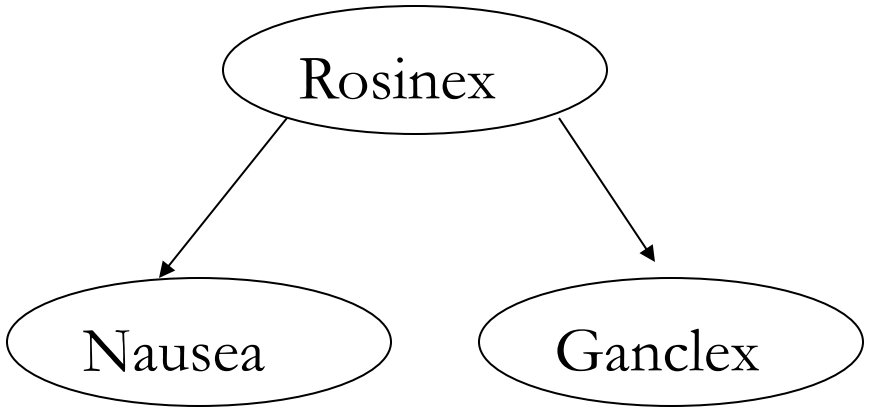
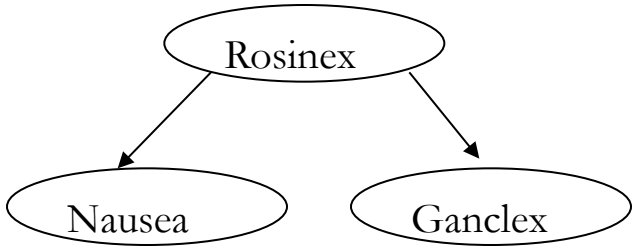
Logistic Regression

Think about this…



		Nausea	*No Nausea*
Rosinex	Ganclex	81	9
Rosinex	No Ganclex	9	1
No Rosinex	Ganclex	1	9
No Rosinex	No Ganclex	90	810

Both Relative Risks are big!



	Nausea	*No Nausea*
Rosinex	90	10
No Rosinex	91	819

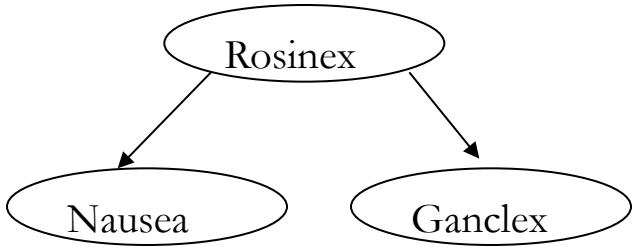
	Nausea	*No Nausea*
Ganclex	82	18
No Ganclex	99	811

RR = (90/100)/(91/910)
= 9.0

RR = (82/100)/(99/910)
= 7.5

		Nausea	*No Nausea*
Rosinex	Ganclex	81	9
Rosinex	No Ganclex	9	1
No Rosinex	Ganclex	1	9
No Rosinex	No Ganclex	90	810

Need a conditional analysis



Rosinex users…

	Nausea	*No Nausea*
Ganclex	81	9
No Ganclex	9	1

RR = (81/90)/(9/10)
= 1.0

Rosinex non-users…

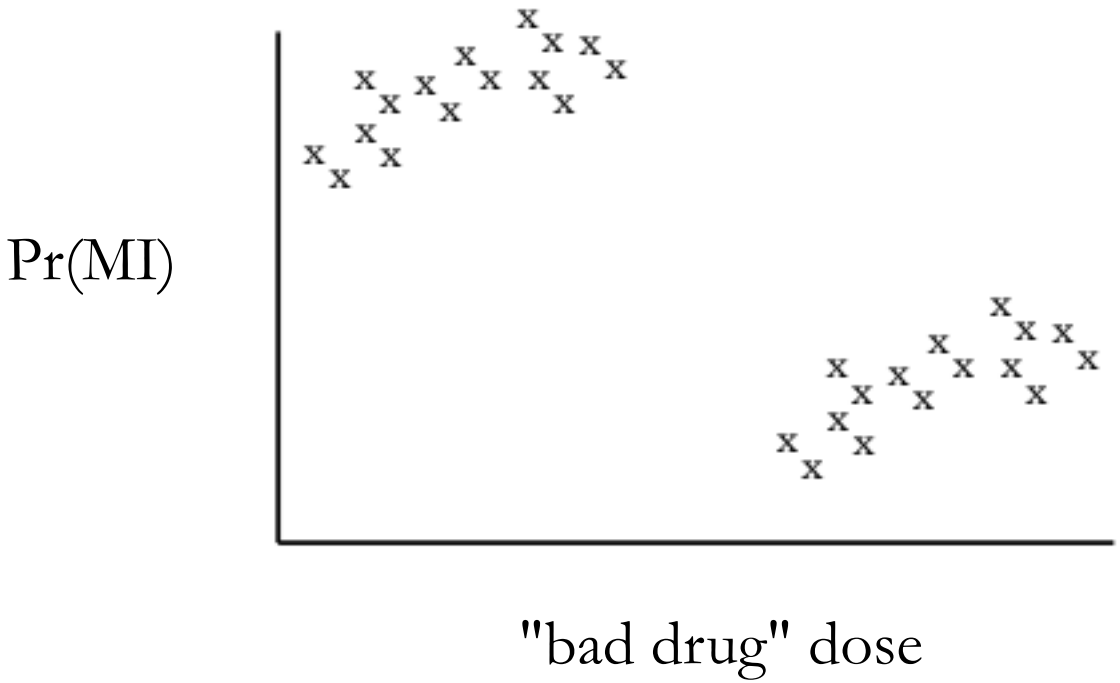
	Nausea	*No Nausea*
Ganclex	1	9
No Ganclex	90	810

RR = (1/10)/(90/900)
= 1.0

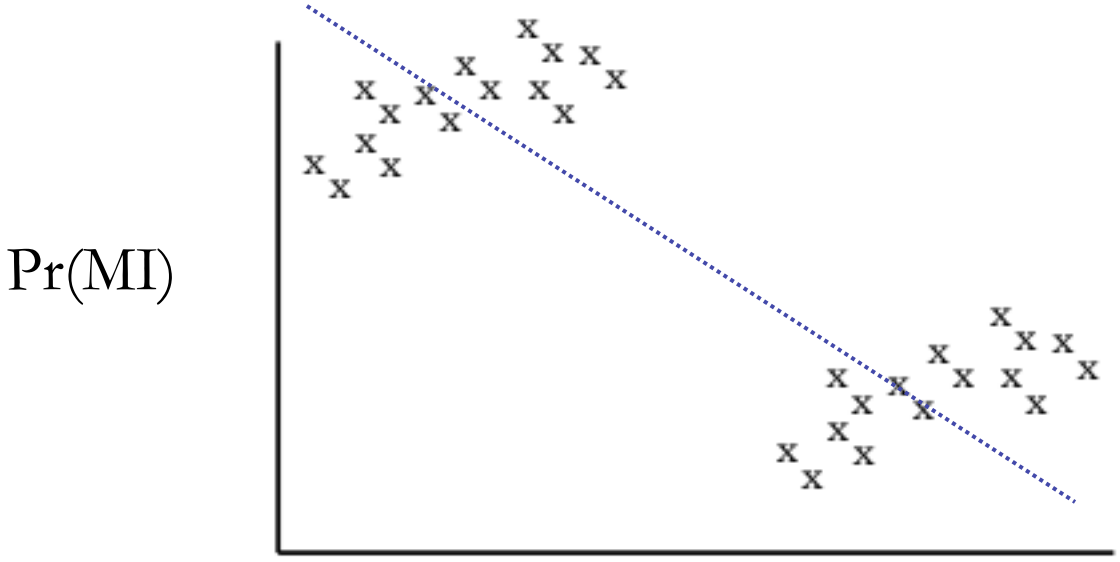
"Holding Rosinex constant, the RR for Ganclex and Nausea is 1"

		Nausea	*No Nausea*
Rosinex	Ganclex	81	9
Rosinex	No Ganclex	9	1
No Rosinex	Ganclex	1	9
No Rosinex	No Ganclex	90	810

Another perspective

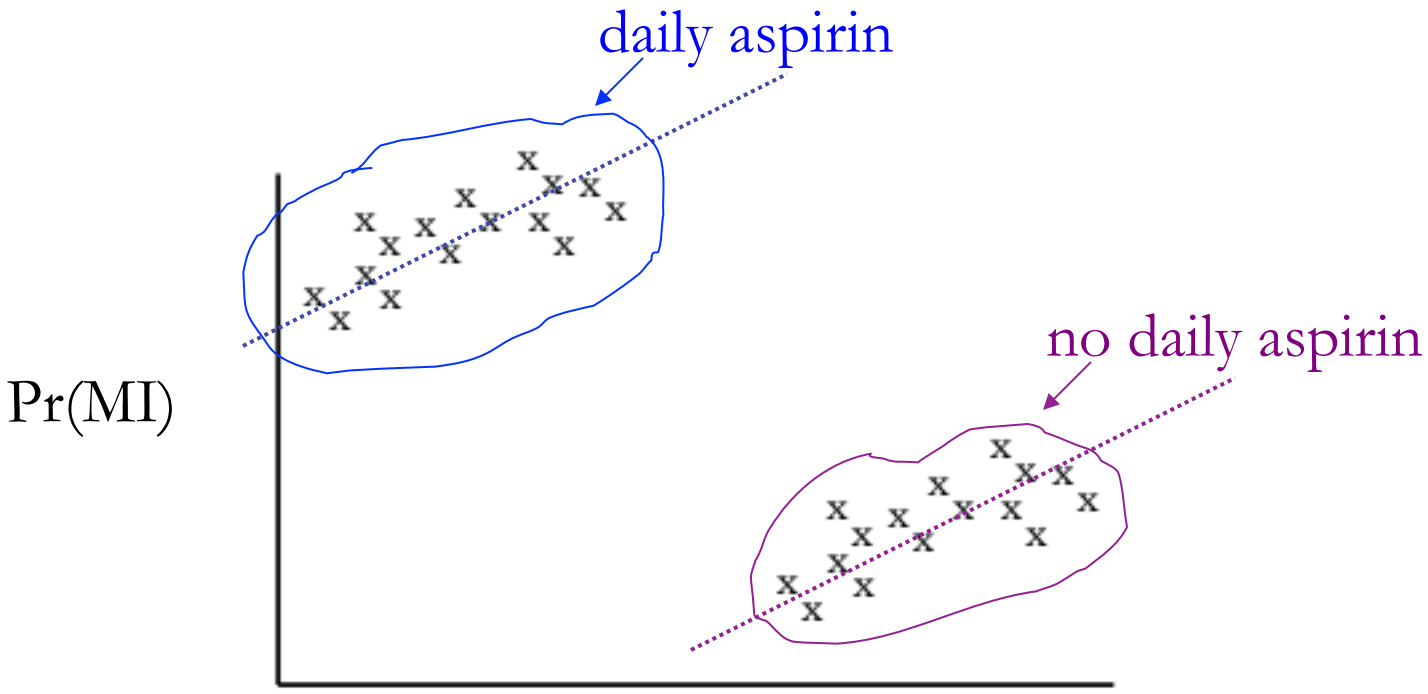


Another perspective



more drug…less chance of MI. Bad drug is good???

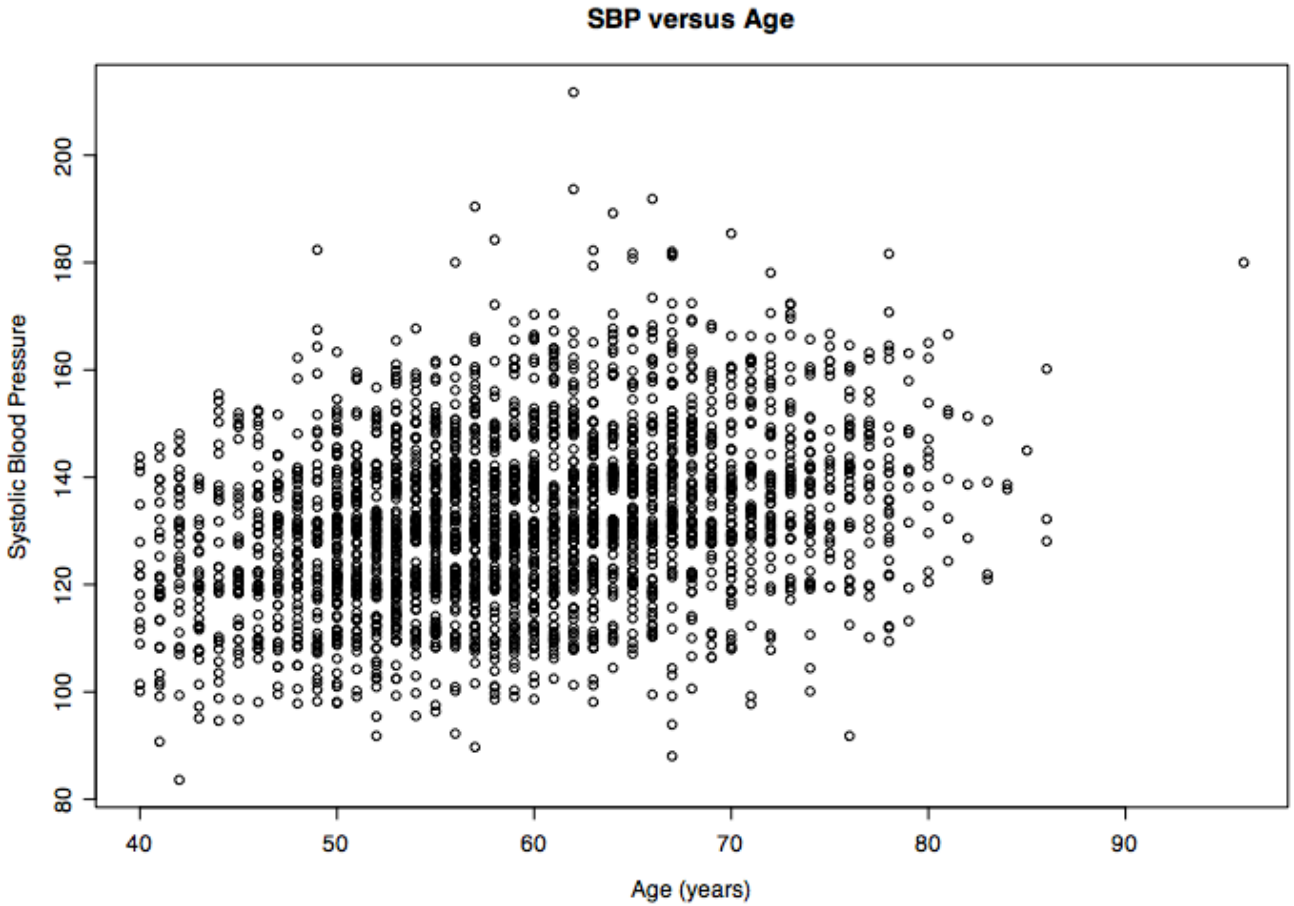
Another perspective



bad for aspirin users, bad for non-users!
Need a conditional analysis

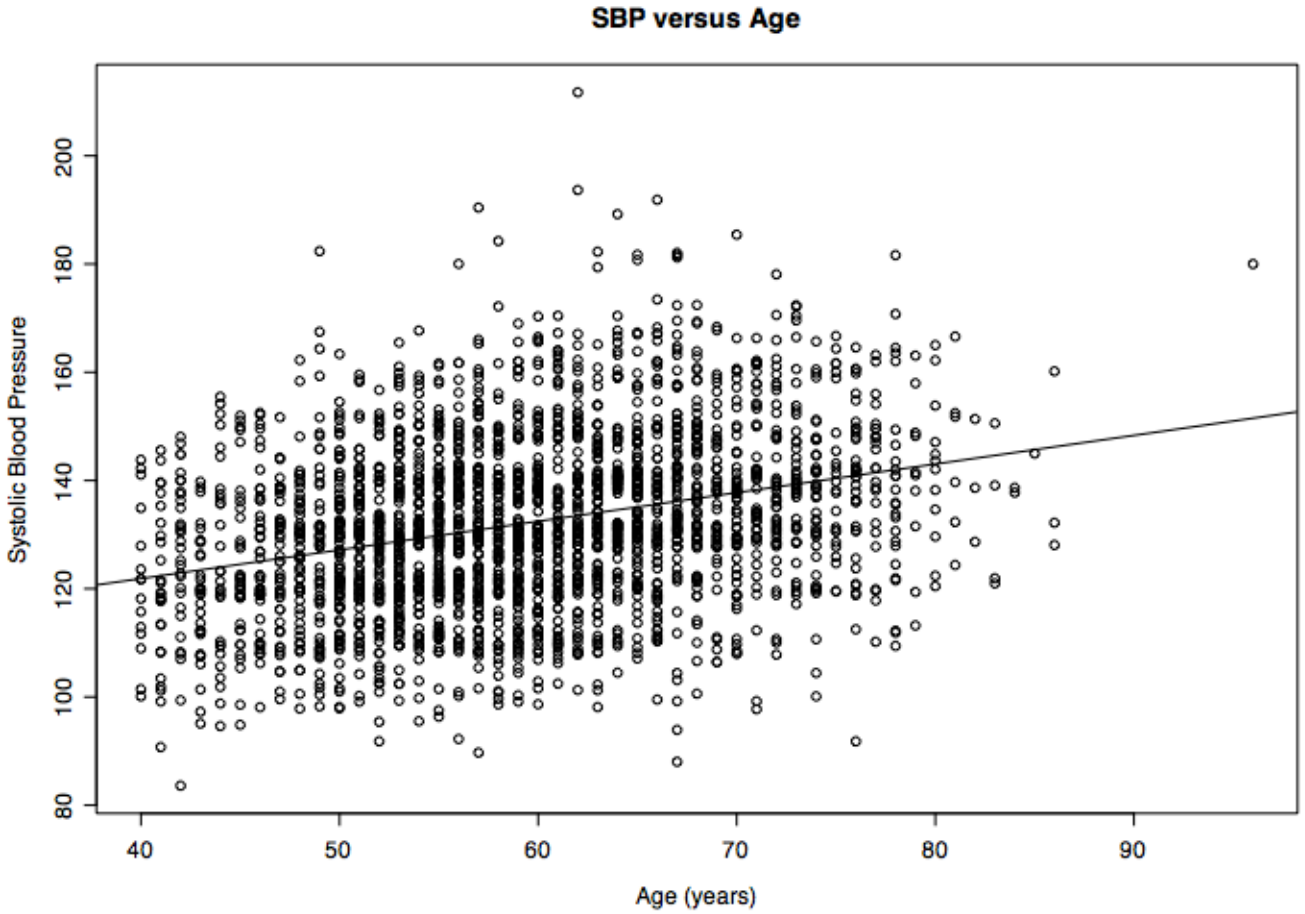
Multiple Regression does this

Start with simple regression…



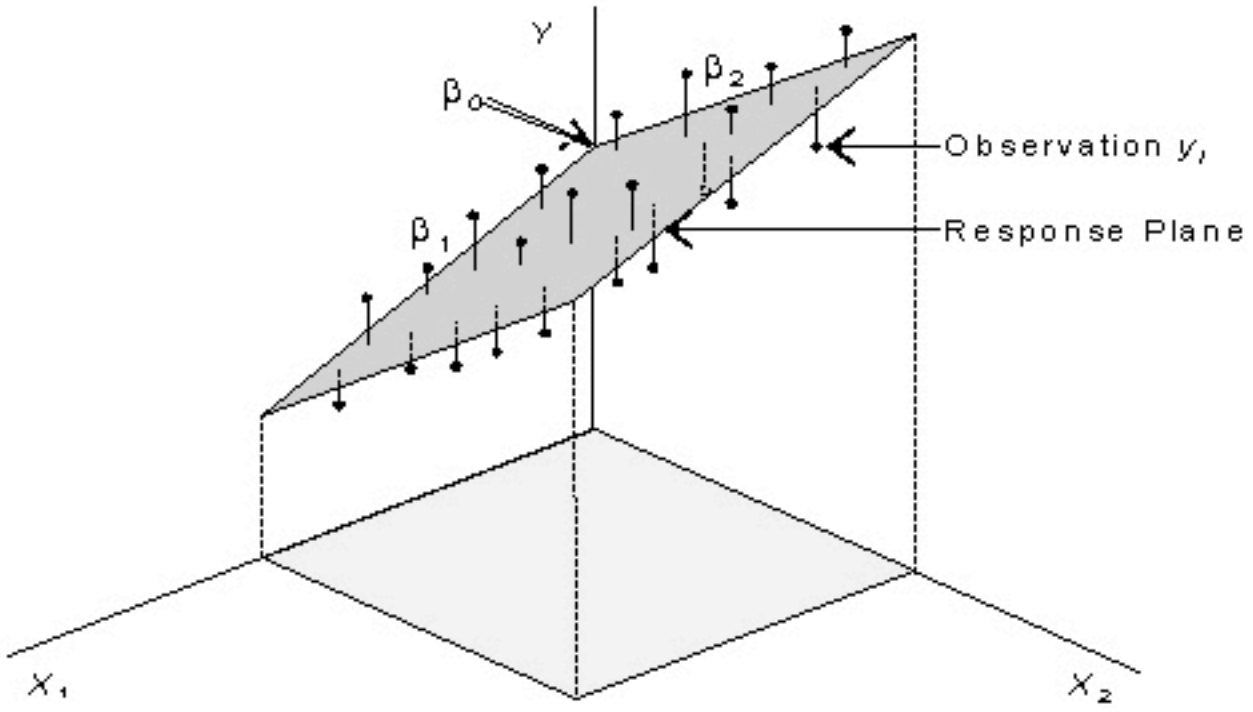
Multiple Regression does this

Start with simple regression…



sbp = 100.7 + 0.53 x age

From Simple to Multiple…



sbp = 83.5 + 0.53 x age + 0.57 x bmi

Multiple Regression Coefficients

sbp = 83.5 + 0.53 x age + 0.57 x bmi

e.g. 46-year old with bmi=25:

83.5 + (0.53 x 46) + (0.57 x 25) = 122.13

46-year old with bmi=26:

83.5 + (0.53 x 46) + (0.57 x 26) = 122.70

Difference = 0.57

"on average, sbp increases 0.57 every time bmi increases by 1, *holding age constant*"

Multiple Regression Coefficients

$$\text{sbp} = 83.5 + 0.53 \times \text{age} + 0.57 \times \text{bmi}$$

e.g. 50-year old with bmi=25:

$$83.5 + (0.53 \times 50) + (0.57 \times 25) = 124.25$$

50-year old with bmi=26:

$$83.5 + (0.53 \times 50) + (0.57 \times 26) = 124.82$$

Difference = 0.57

"on average, sbp increases 0.57 every time bmi increases by 1, *holding age constant*"

(doesn't actually matter which particular age)

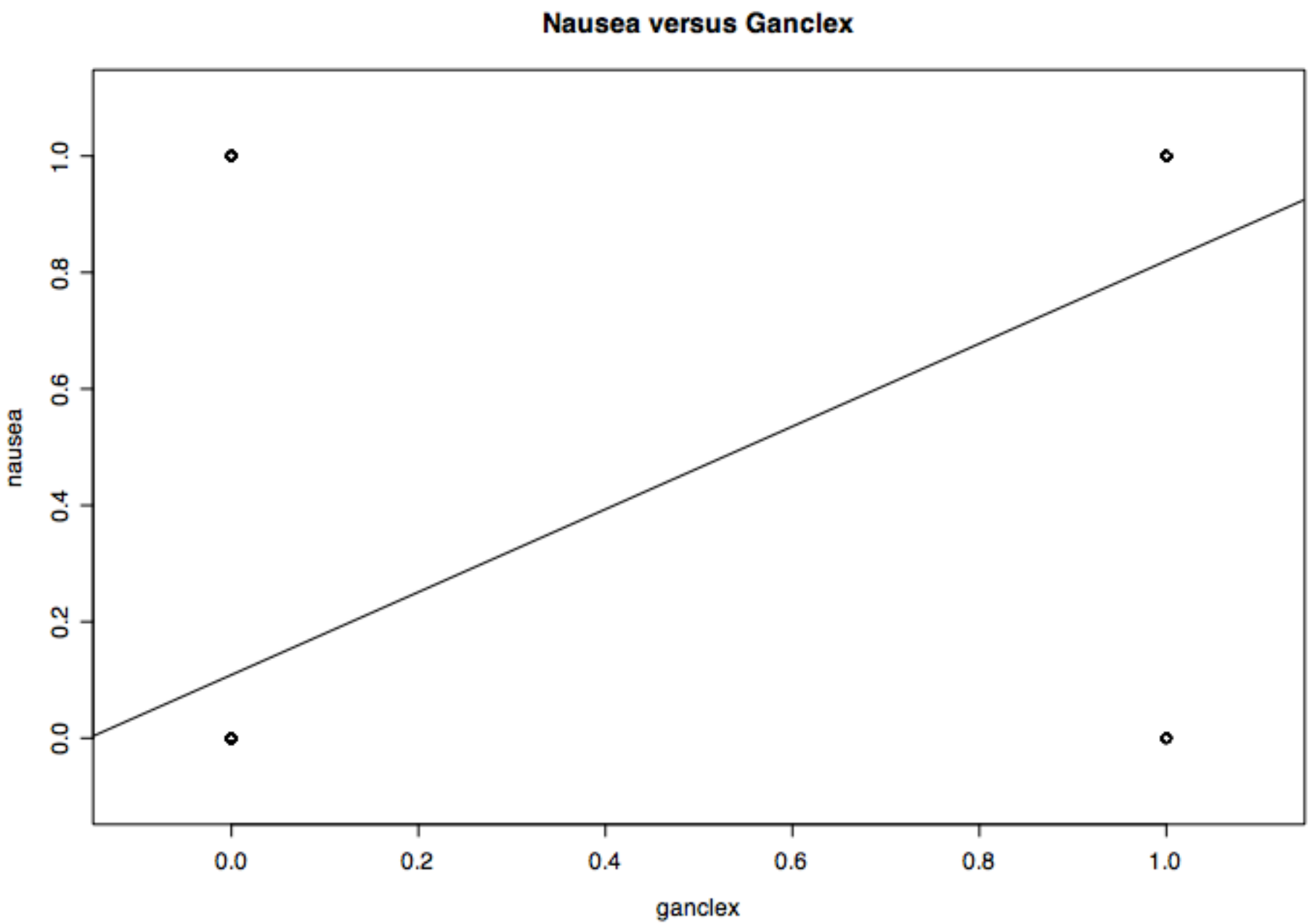
Multiple Regression Coefficients

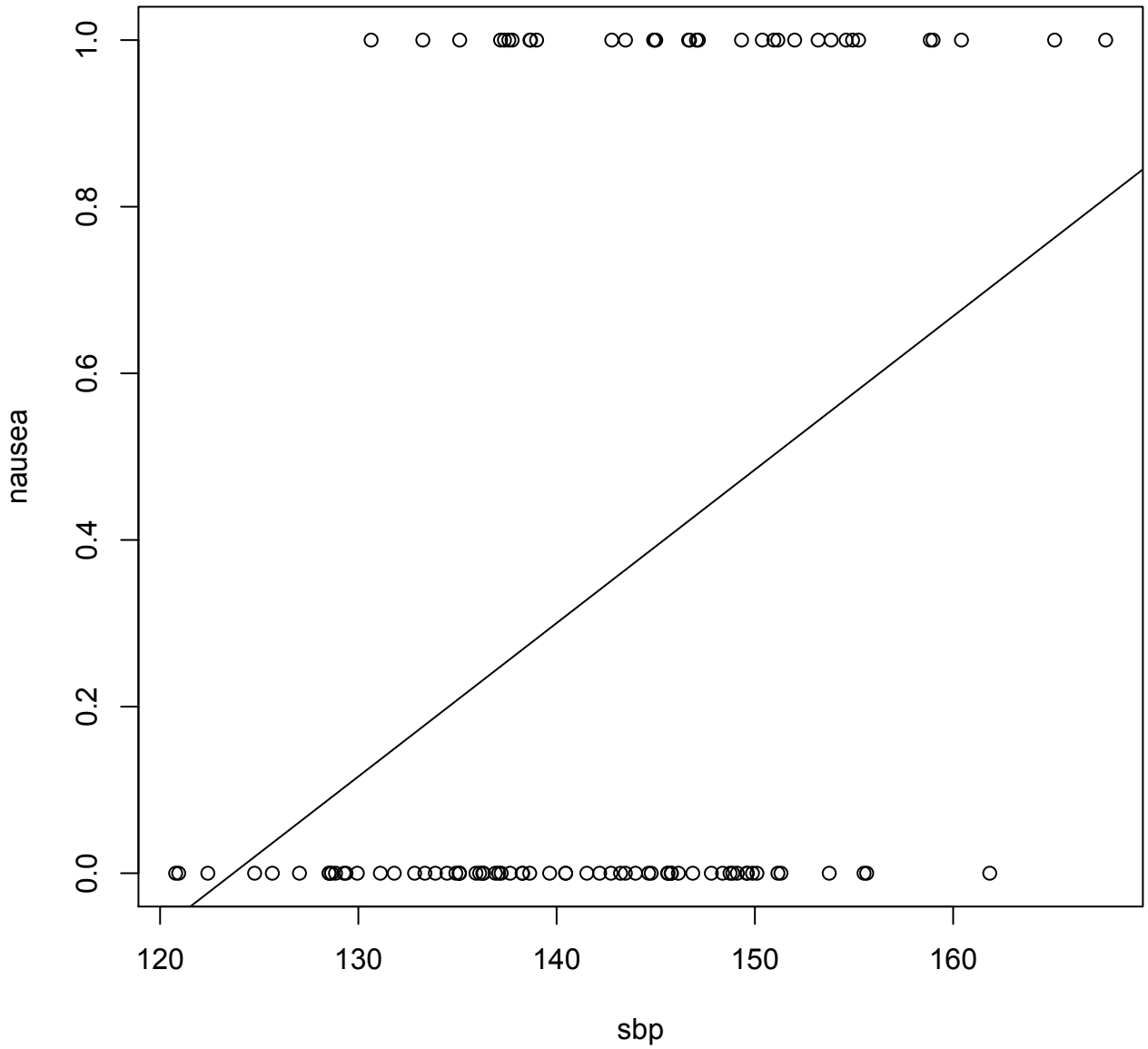
nausea = 0.1 + 0.8 x rosinex + 0.0 x ganclex

effect of rosinex *holding ganclex constant*

effect of ganclex *holding rosinex constant*

- nausea, rosinex, and ganclex are zero-one variables in this analysis
- interactions?


Nausea versus Ganclex
nausea
0.0
0.2
0.4
0.6
0.8
1.0
ganclex
0.0
0.2
0.4
0.6
0.8
1.0


nausea
0.0
0.2
0.4
0.6
0.8
1.0
120
130
140
150
160
sbp

On to Logistic Regression

Could build a model for the probability of nausea…

$$\Pr(\text{nausea}) = 0.1 + 0.8 \times \text{rosinex} + 0.07 \times \text{sbp}$$

…but, in general, the right hand side could be bigger than 1 or negative if some drugs have a protective effect

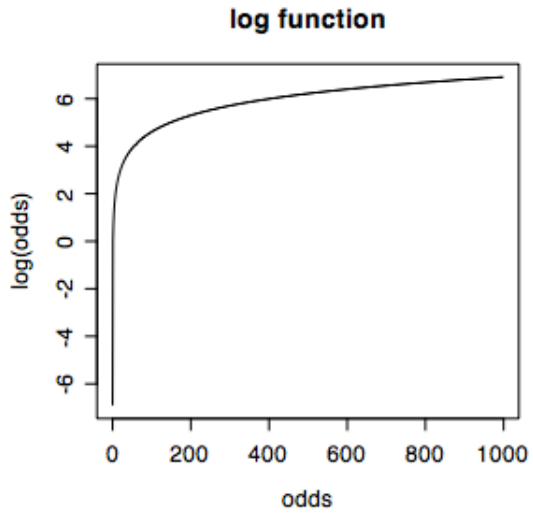
On to Logistic Regression

Could build a model for the odds of nausea…

$$\frac{\Pr(\text{nausea})}{\Pr(\text{no nausea})} = 0.1 + 0.8 \text{ x rosinex} + 0.07 \text{ x sbp}$$

…but, in general, the right hand side could be negative if some drugs have a protective effect

How about log(odds)?



On to Logistic Regression

$$\log \frac{\Pr(\text{nausea})}{\Pr(\text{no nausea})} = -2.2 + 4.4 \text{ x rosinex} + 0.0 \text{ x ganclex}$$

Now the prediction is meaningful no matter what the values of the regression coefficients

But the model no longer predicts nausea - it predicts the log odds of nausea

For someone taking rosinex the predicted log odds of nausea is -2.2 + 4.4 = 2.2

For someone not taking rosinex the predicted log odds of nausea is -2.2

How to Unravel Log Odds

For someone taking rosinex the predicted log odds of nausea is 2.2

$$\log \frac{\Pr(\text{nausea})}{\Pr(\text{no nausea})} = 2.2$$

$$\Rightarrow \quad \frac{\Pr(\text{nausea})}{1-\Pr(\text{nausea})} = \exp(2.2)$$

$$\Rightarrow \quad \Pr(\text{nausea}) = \exp(2.2) - \exp(2.2) \times \Pr(\text{nausea})$$

$$\Rightarrow \quad \Pr(\text{nausea}) = \exp(2.2)/(1+\exp(2.2)) = 0.9$$

Logistic Regression Coefficients

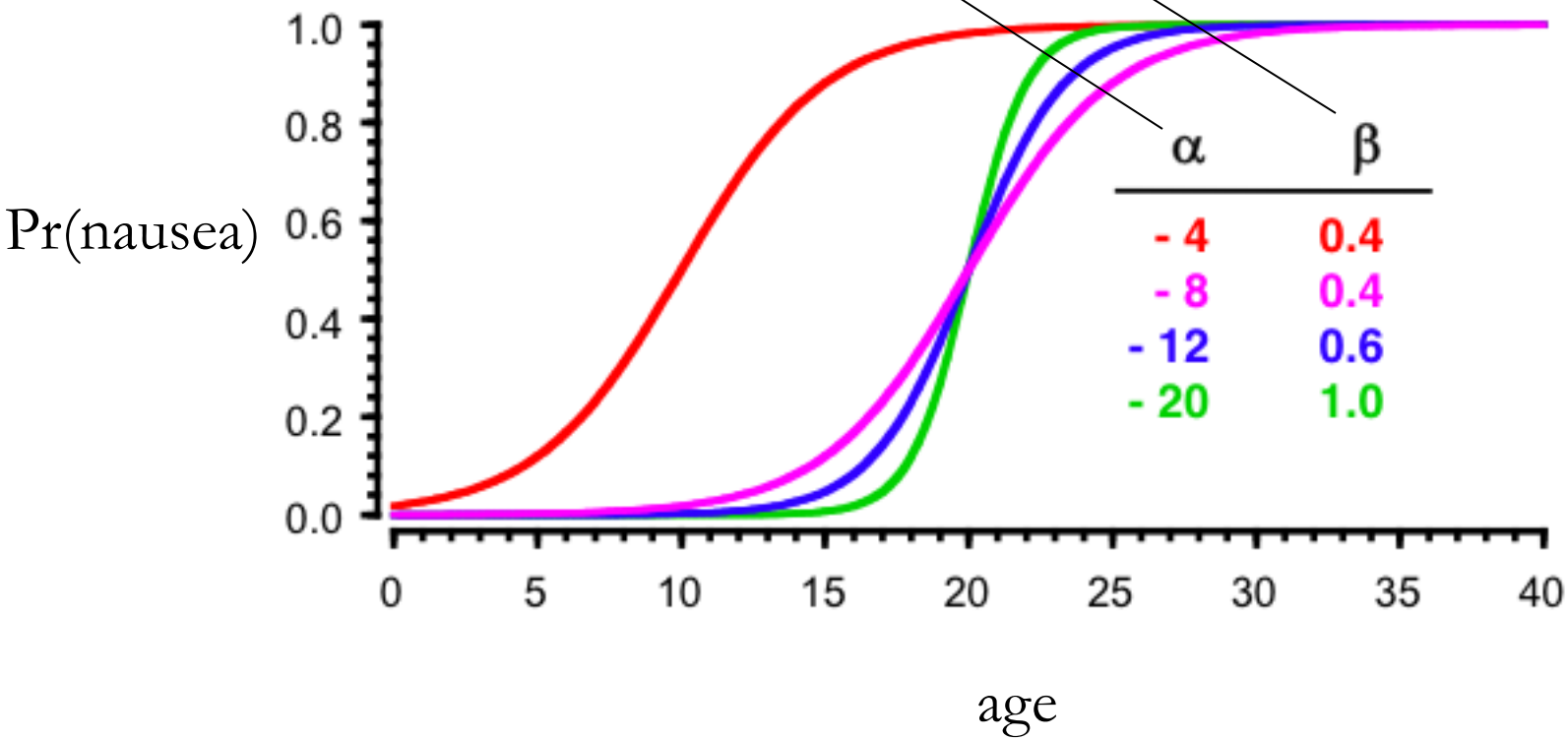
$$\log \frac{\Pr(\text{nausea})}{\Pr(\text{no nausea})} = -2.2 + 4.4 \text{ x rosinex} + 0.0 \text{ x ganclex}$$

"4.4 is the amount by which the log odds of nausea goes up when someone takes ganclex holding everything else constant"

positive coefficient $\Longrightarrow$ odds increases $\Longrightarrow$ probability goes up

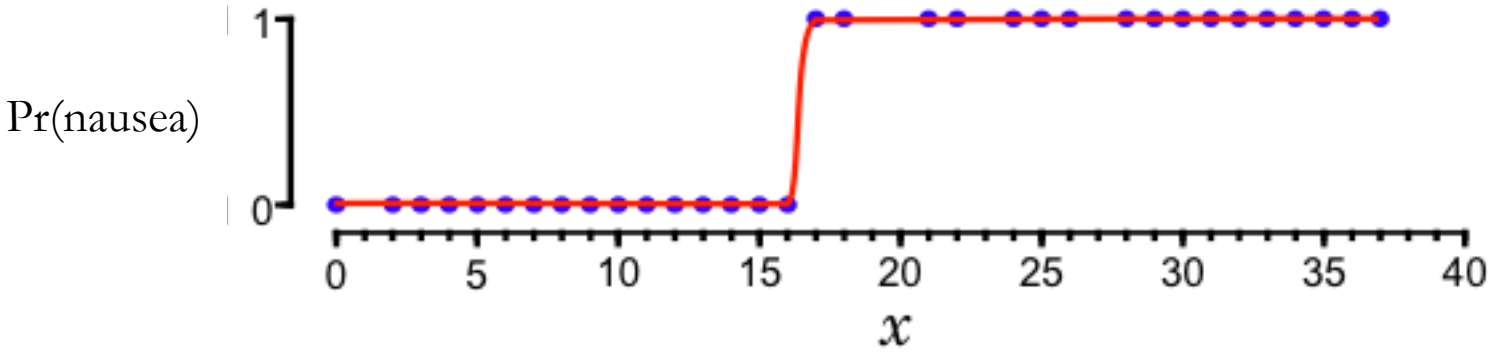
Binary and Continuous Predictors

$$\log \frac{\Pr(\text{nausea})}{\Pr(\text{no nausea})} = -2.2 + 0.3 \times \text{age} + 4.4 \times \text{ganclex}$$

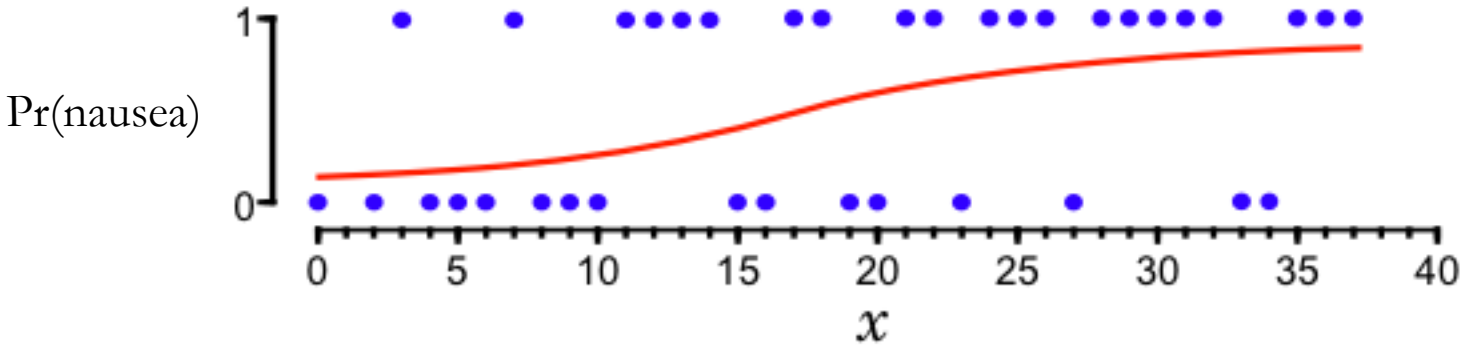


More on Coefficients

coefficient large $\Rightarrow$ predictor strongly discriminates between nausea and no nausea



coefficient small $\Rightarrow$ predictor weakly discriminates between nausea and no nausea



Maximum Likelihood Logistic Regression

Typically estimate the coefficients via "maximum likelihood"

Suppose you want to estimate α and β in this model:

$$\log \frac{\Pr(\text{nausea})}{\Pr(\text{no nausea})} = \alpha + \beta \text{ x rosinex}$$

using these data:

Rosinex	Nausea
0	1
0	0
0	0
1	1
1	0

Maximum Likelihood Logistic Regression

e.g., if $\alpha = -0.22$ and $\beta = 1$ then:

A	B	C	E	F
Rosinex	**Nausea**	**Probability**	-0.22	1
0	1	0.55	ALPHA	BETA
0	0	0.45		
0	0	0.45		
1	1	0.31		
1	0	0.69		
TOTAL PROBABILITY:		**0.0237**		

e.g., if $\alpha = -0.42$ and $\beta = 2.1$ then:

A	B	C	E	F
Rosinex	**Nausea**	**Probability**	-0.42	2.1
0	1	0.60	ALPHA	BETA
0	0	0.40		
0	0	0.40		
1	1	0.16		
1	0	0.84		
TOTAL PROBABILITY:		**0.0126**		

Idea: pick the values of α and β that maximize the probability of the nausea outcomes you actually saw!

Logistic Regression in Practice

- SAS, R, etc. do maximum likelihood logistic regression
- Nice statistical properties; works well in most applications
- Truly large-scale applications with thousands of drugs require "regularized logistic regression" aka lasso and ridge logistic regression
- glm(y~x1+x2, data=foo, family=binomial)

```
Boar <- read.table("/Users/dbm/Documents/W2025/ZuurDataMixedModelling/Boar.txt",header=TRUE)
B1 <- glm(Tb~LengthCT,data=Boar,family=binomial)
summary(B1)
MyData <- data.frame(LengthCT = seq(from = 46.5, to = 165,by = 1))
Pred <- predict(B1, newdata = MyData, type = "response")
plot(x = Boar$LengthCT, y = Boar$Tb)
lines(MyData$LengthCT, Pred)
```

```
ParasiteCod<- read.table("/Users/dbm/Documents/W2025/ZuurDataMixedModelling/ParasiteCod.txt",header=TRUE)
ParasiteCod$fArea <- factor(ParasiteCod$Area)
ParasiteCod$fYear <- factor(ParasiteCod$Year)
Par1 <- glm(Prevalence ~ fArea * fYear + Length, family=binomial, data=ParasiteCod)
summary(Par1)
```

missing values?