Business Statistics

Likelihood and Logistic Regression

Zhanrui Cai

Assistant Professor in Analytics and Innovation

ISLR Chapter 4.1-4.3

Your First Consulting Job

- A billionaire asks you a question:
  - He says: I have a coin. If I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

  - You say: The probability is: **3/5**
  - **He says: Why???**
  - You say: Because…

Bernoulli Distribution

Data, D = 

- P(Head) = θ, P(Tail) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

<u>Choose θ that maximizes the probability of observed data</u>

Maximum Likelihood Estimation

Choose θ that maximizes the probability of observed data

$$P(\theta|D) = \theta(1-\theta)\theta\theta(1-\theta) = \theta^3(1-\theta)^2$$

$$\log P(\theta|D) = 3\log\theta + 2\log(1-\theta)$$

MLE of the probability of head:

$$\hat{\theta} = \frac{3}{3+2} = \frac{3}{5}$$

Maximum Likelihood Estimation: Normal Distribution

- Suppose $y_1, \dots, y_n$ iid Normal with mean μ and standard deviation σ.

- Density function

$$f(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

- Likelihood function

$$f(\mu,\sigma|y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

Normal Mean Problem: Maximum Likelihood

- Given independence, the joint likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}}$$

- Maximum likelihood estimation

$$\hat{\mu}_{MLE} = argmax_{\mu} L(\mu, \sigma^2) = \bar{y}$$

$$\hat{\sigma}^2{}_{MLE} = argmax_{\sigma^2} L(\mu, \sigma^2) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

Classification

- Qualitative variables take values in an unordered set C, such as:

  eye color ∈ {brown, blue, green}

  email ∈ {spam, ham}

- Given a feature vector X and a qualitative response Y taking values in the set C, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y, i.e. $C(X) \in C$

- Often we are more interested in estimating the probabilities that X belongs to each category in C

- For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not


TITANIC

Data

Name	Survived	Passenger Class	Sex	Age	Siblings and Spouses	Parents and Children	Fare	Port	Home / Destination
Allen, Miss. Elisabeth Walton	Yes	1	fem…	29	0	0	211.…	S	St Louis, MO
Allison, Master. Hudson Trevor	Yes	1	male	0.9…	1	2	151.55	S	Montreal, PQ / Chesterville, OI
Allison, Miss. Helen Loraine	No	1	fem…	2	1	2	151.55	S	Montreal, PQ / Chesterville, OI
Allison, Mr. Hudson Joshua …	No	1	male	30	1	2	151.55	S	Montreal, PQ / Chesterville, OI
Allison, Mrs. Hudson J C …	No	1	fem…	25	1	2	151.55	S	Montreal, PQ / Chesterville, OI
Anderson, Mr. Harry	Yes	1	male	48	0	0	26.55	S	New York, NY
Andrews, Miss. Kornelia …	Yes	1	fem…	63	1	0	77.9…	S	Hudson, NY
Andrews, Mr. Thomas Jr	No	1	male	39	0	0	0	S	Belfast, NI
Appleton, Mrs. Edward Dale …	Yes	1	fem…	53	2	0	51.4…	S	Bayside, Queens, NY
Artagaveytia, Mr. Ramon	No	1	male	71	0	0	49.5…	C	Montevideo, Uruguay
Astor, Col. John Jacob	No	1	male	47	1	0	227.…	C	New York, NY
Astor, Mrs. John Jacob …	Yes	1	fem…	18	1	0	227.…	C	New York, NY
Aubart, Mme. Leontine Pauline	Yes	1	fem…	24	0	0	69.3	C	Paris, France
Barber, Miss. Ellen "Nellie"	Yes	1	fem…	26	0	0	78.85	S	
Barkworth, Mr. Algernon …	Yes	1	male	80	0	0	30	S	Hessle, Yorks
Baumann, Mr. John D	No	1	male	•	0	0	25.925	S	New York, NY
Baxter, Mr. Quigg Edmond	No	1	male	24	0	1	247.…	C	Montreal, PQ
Baxter, Mrs. James (Helene …	Yes	1	fem…	50	0	1	247.…	C	Montreal, PQ
Bazzani, Miss. Albina	Yes	1	fem…	32	0	0	76.2…	C	
Beattie, Mr. Thomson	No	1	male	36	0	0	75.2…	C	Winnipeg, MN
Beckwith, Mr. Richard Leonard	Yes	1	male	37	1	1	52.5…	S	New York, NY
Beckwith, Mrs. Richard …	Yes	1	fem…	47	1	1	52.5…	S	New York, NY
Behr, Mr. Karl Howell	Yes	1	male	26	0	0	30	C	New York, NY
Bidois, Miss. Rosalie	Yes	1	fem…	42	0	0	227.…	C	
Bird, Miss. Ellen	Yes	1	fem…	29	0	0	221.…	S	
Birnbaum, Mr. Jakob	No	1	male	25	0	0	26	C	San Francisco, CA
Bishop, Mr. Dickinson H	Yes	1	male	25	1	0	91.0…	C	Dowagiac, MI
Bishop, Mrs. Dickinson H …	Yes	1	fem…	19	1	0	91.0…	C	Dowagiac, MI
Bissette, Miss. Amelia	Yes	1	fem…	35	0	0	135.…	S	
Bjornstrom-Steffansson, Mr. …	Yes	1	male	28	0	0	26.55	S	Stockholm, Sweden / …
Blackwell, Mr. Stephen Weart	No	1	male	45	0	0	35.5	S	Trenton, NJ
Blank, Mr. Henry	Yes	1	male	40	0	0	31	C	Glen Ridge, NJ
Bonnell, Miss. Caroline	Yes	1	fem…	30	0	0	164.…	S	Youngstown, OH
Bonnell, Miss. Elizabeth	Yes	1	fem…	58	0	0	26.55	S	Birkdale, England Cleveland

Jack & Rose: Who Will Survive?

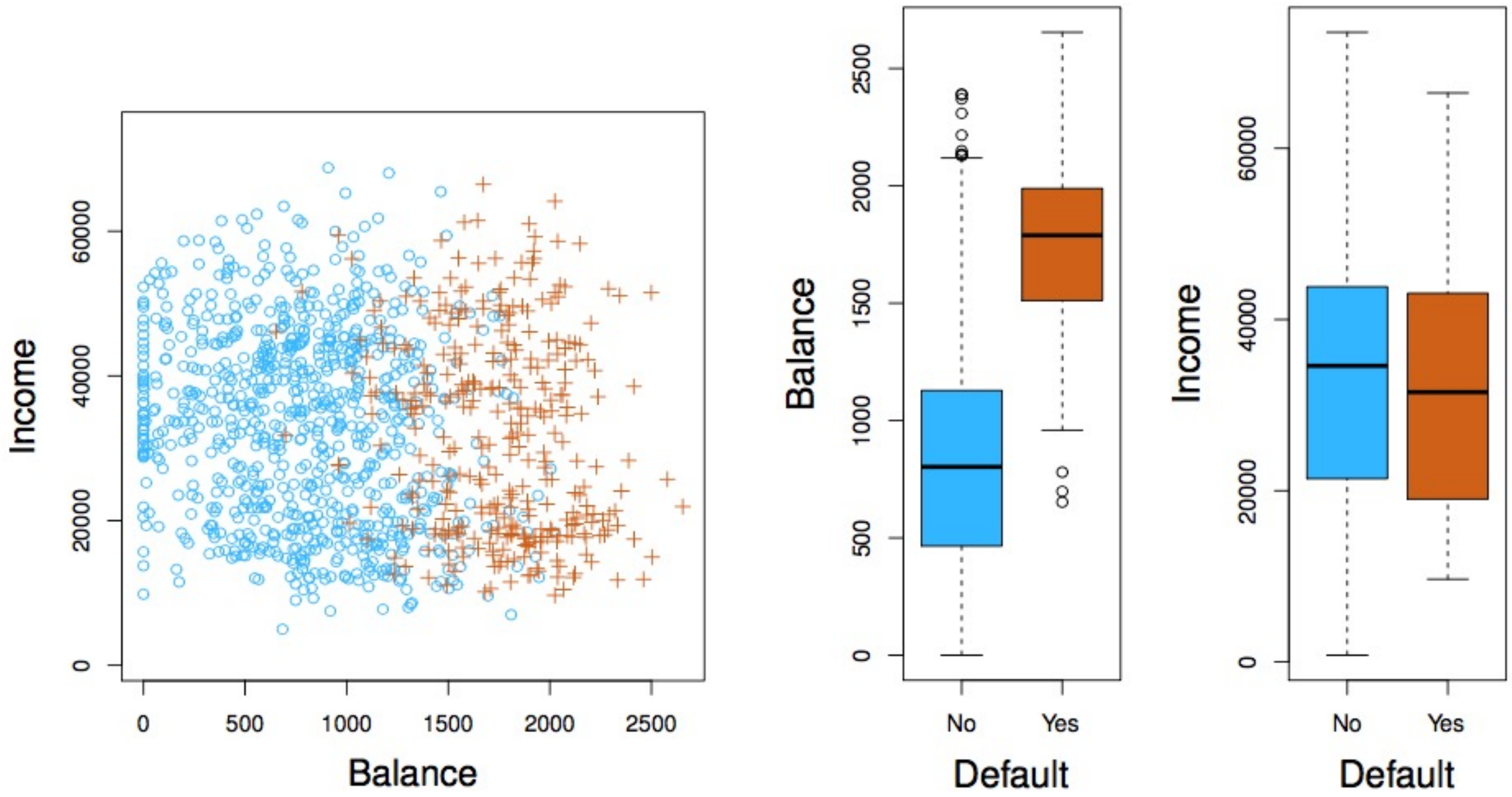
Prob Survival: 15.4%

Class: 3
Sex: Male
Age: 17
Siblings&Spouses: 0
Port: S

Prob Survival: 92.8%

Class: 1
Sex: Female
Age: 20
Siblings&Spouses: 1
Port: S

Example: Credit Card Default



Can we use Linear Regression?

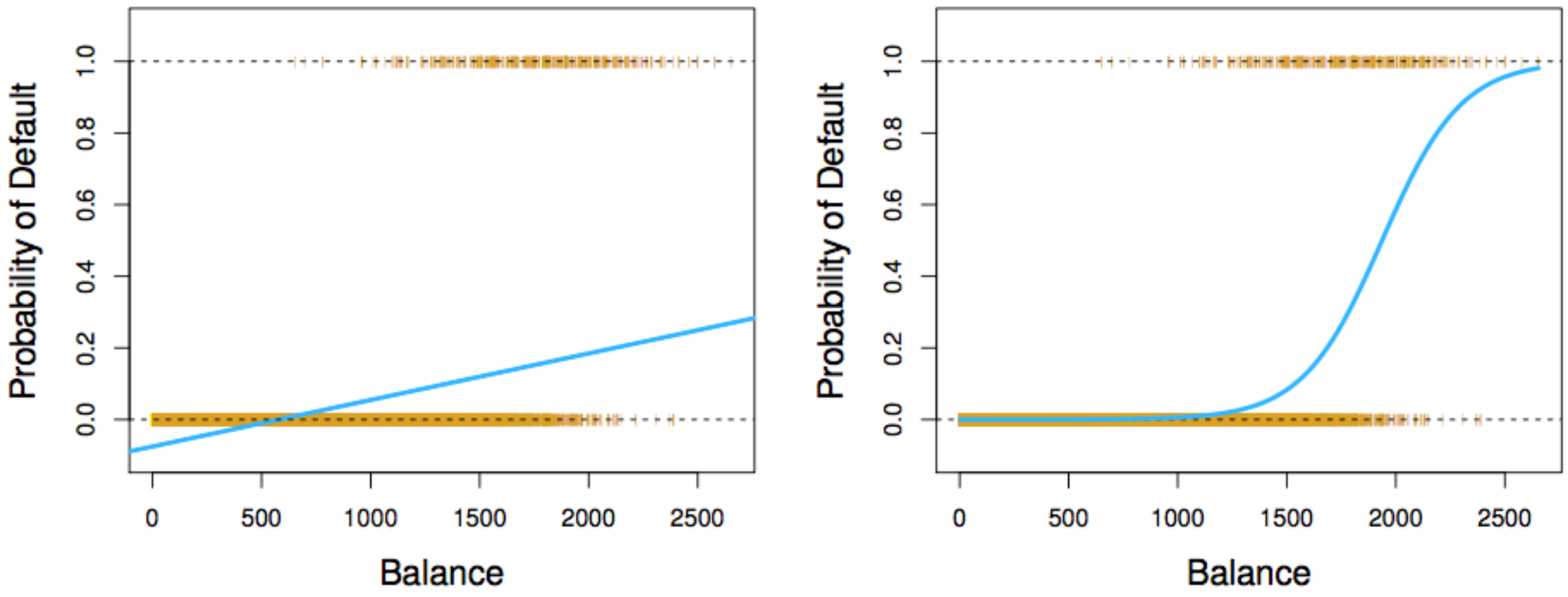
Suppose for the `Default` classification task that we code

$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as `Yes` if $\hat{Y} > 0.5$?

- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to *linear discriminant analysis* which we discuss later.
- Since in the population $E(Y|X = x) = \Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- However, *linear* regression might produce probabilities less than zero or bigger than one. *Logistic regression* is more appropriate.

Linear versus Logistic Regression



The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Linear Regression continued

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

This coding suggests an ordering, and in fact implies that the difference between `stroke` and `drug overdose` is the same as between `drug overdose` and `epileptic seizure`.

Linear regression is not appropriate here.
Multiclass Logistic Regression or *Discriminant Analysis* are more appropriate.

Logistic Regression

Logistic regression is a method for analyzing relative probabilities between discrete outcomes (binary or categorical dependent variables)

- Binary outcome: standard logistic regression
  - ie. Win (1) or loss (0)

- Categorical outcome: multinomial logistic regression
  - ie. Admission with scholarship (1) or admission (2) or waiting list (3) or rejection (4)

- Predictor variables (x_i) can take on *any* form: binary, categorical, and/or continuous

Logistic Regression

- Write $p(X) = \Pr(Y = 1|X)$ for short and consider using balance to predict default. Logistic regression uses the form

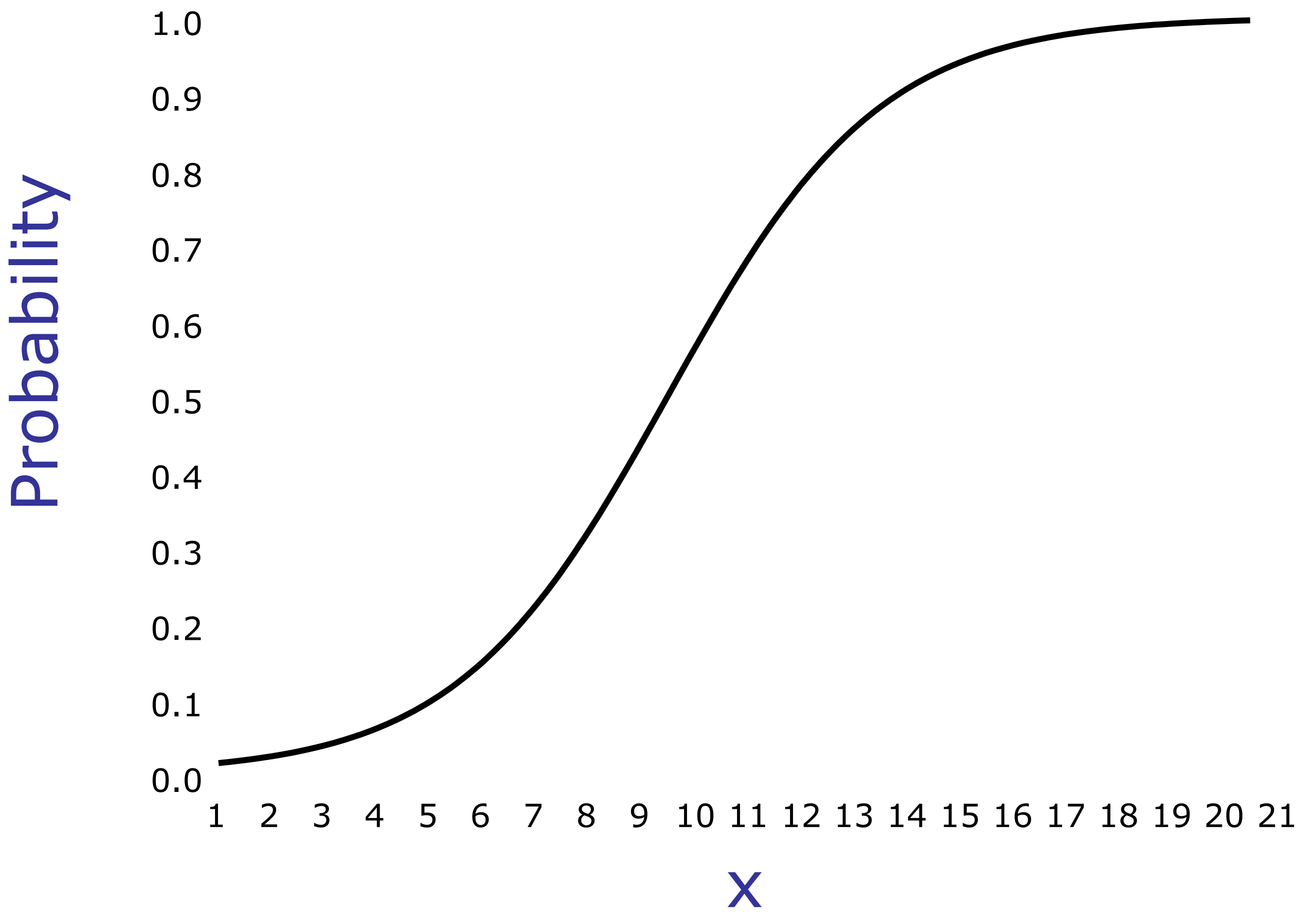
$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

- It is easy to see that $p(X)$ has values between 0 and 1.
- Meanings of β_0 and β_1
  - β_0: The regression constant (moves curve left and right)
  - β_1: The regression slope (steepness of curve)
  - $-\frac{\beta_0}{\beta_1}$: The threshold, where probability of success = .50

Logistic Regression Curve

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$



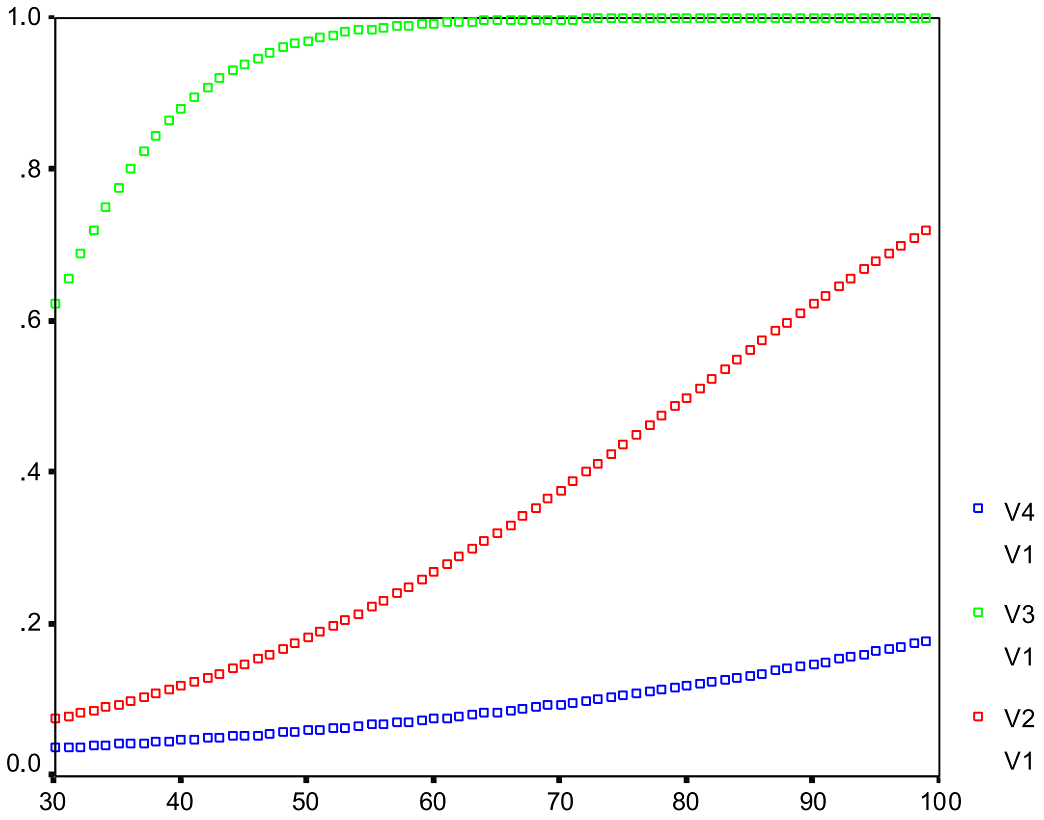
Logistic Function

- Constant intercept
  Different slopes

  – v2: b0 = -4.00
    b1 = 0.05 (middle)

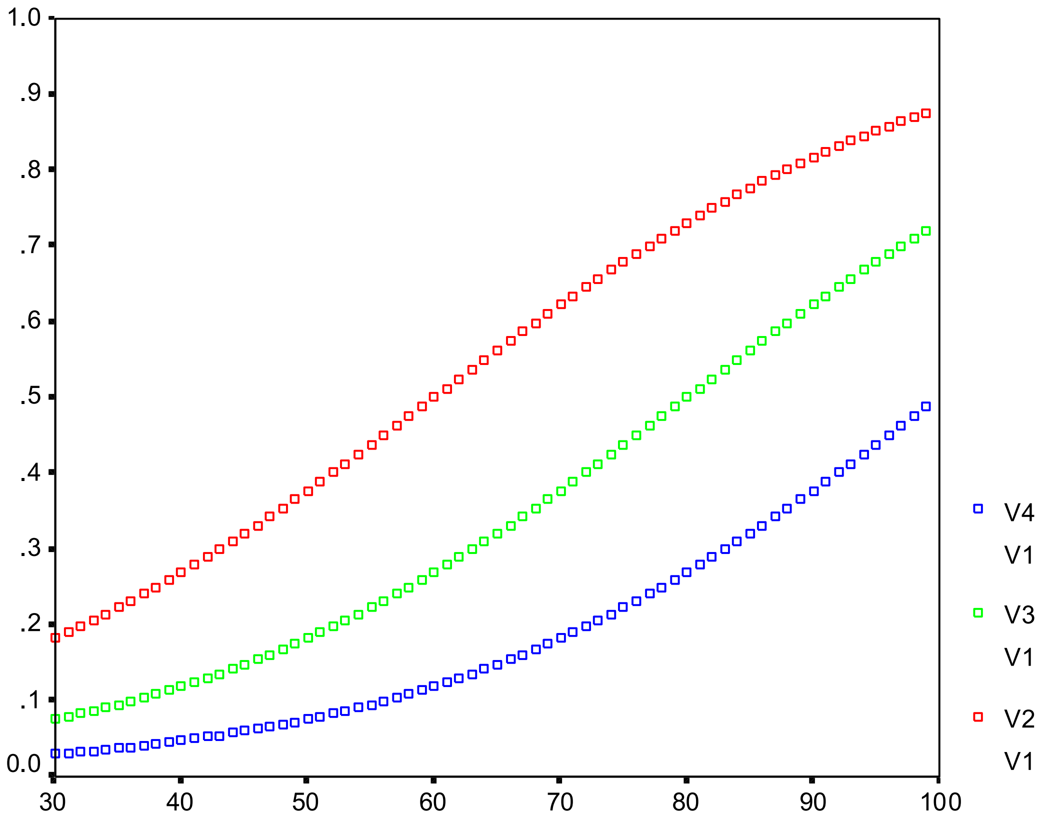
  – v3: b0 = -4.00
    b1 = 0.15 (top)

  – v4: b0 = -4.00
    b1 = 0.025 (bottom)



Logistic Function

- Constant slope
  Different intercept
  - v2: b0 = -3.00
    b1 = 0.05 (top)
  - v3: b0 = -4.00
    b1 = 0.05 (middle)
  - v4: b0 = -5.00
    b1 = 0.05 (bottom)



Odds Ratio

- From $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$, we have $\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} = e^{\beta_0} e^{\beta_1 x}$

- The odds ratio increases **multiplicatively** by e^{β_1} for every 1-unit increase in x
  - The odds at $X = x + 1$ are e^{β_1} times the odds at $X = x$
  - $\frac{odds(x+1)}{odds(x)} = e^{\beta_1}$
- Therefore, e^{β_1} **is an odds ratio!**

- e^{β_1} represents the change in the odds of the outcome (multiplicatively) by increasing x by 1 unit
  - If $\beta_1 > 0$, the odds and probability increase as x increases ($e^{\beta_1} > 1$)
  - If $\beta_1 < 0$, the odds and probability decrease as x increases ($e^{\beta_1} < 1$)
  - If $\beta_1 = 0$, the odds and probability are the same at all x levels ($e^{\beta_1}=1$)

Log Odds or Logit: $\log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$

- The sign ($\pm$) of β_1 determines whether the **log odds** of y is increasing or decreasing *for every 1-unit increase* in x.
  - If $\beta_1 > 0$, there is an increase in the **log odds** of y for every 1-unit increase in x.
  - If $\beta_1 < 0$, there is a decrease in the **log odds** of y for every 1-unit increase in x.
  - If $\beta_1 = 0$ there is *no linear relationship* between the **log odds** and x.

About Logistic Regression

- It uses maximum likelihood estimation rather than the least squares estimation used in linear regression.

- The general form of the distribution is assumed.
  – In this case, the Bernoulli distribution

- The likelihood is then maximized.
  – Only special cases can be solved by hand.
  – Most often via numerical methods, implemented in R.

Maximum Likelihood Estimation

We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

Most statistical packages can fit linear logistic regression models by maximum likelihood. In `R` we use the `glm` function.

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.6513	0.3612	-29.5	< 0.0001
`balance`	0.0055	0.0002	24.9	< 0.0001

Making Prediction

What is our estimated probability of `default` for someone with a balance of $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

With a balance of $2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

With Categorical Predictors

Lets do it again, using student as the predictor.

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Multiple Logistic Regression

- Extension to more than one predictor variable (either numeric or dummy variables).
- With p predictors, the model is written:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

- Adjusted Odds ratio for raising x_i by 1 unit, holding all other predictors constant:

$$OR_i = e^{\beta_i}$$

- odds = (base odds) $OR_1 \; OR_2 \; \ldots \; OR_k$
- Many models have nominal/ordinal predictors, and widely make use of dummy variables

Multivariate Logistic Regression

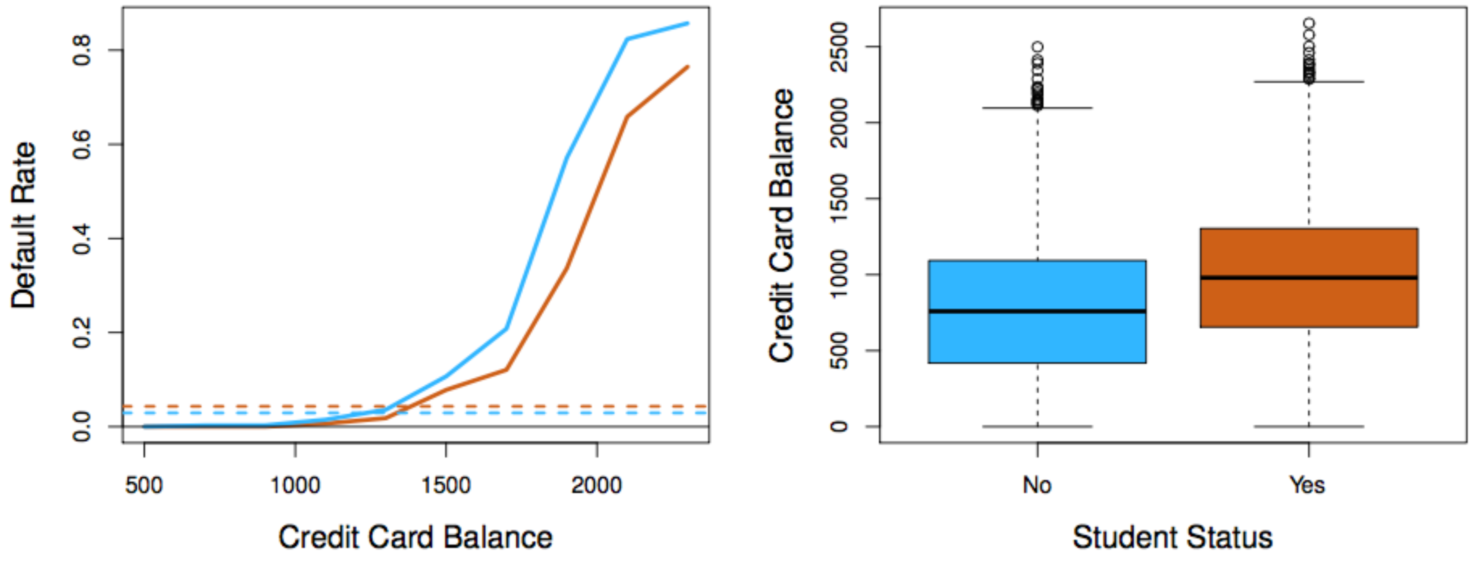
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.8690	0.4923	-22.08	< 0.0001
`balance`	0.0057	0.0002	24.74	< 0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	-0.6468	0.2362	-2.74	0.0062

Why is coefficient for `student` negative, while it was positive before?

Confounding!



- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.

Return to Titanic

```{r}
fit <- glm(Survived~., tit.comp[,-c(1,7:8,10)], family=binomial(logit))
summary(fit)
```

```
Call:
glm(formula = Survived ~ ., family = binomial(logit), data = tit.comp[,
    -c(1, 7:8, 10)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5600  -0.6827  -0.4101   0.6564   2.5314

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.315656   0.382859  11.272  < 2e-16 ***
Passenger.Class2     -1.126422   0.243779  -4.621 3.82e-06 ***
Passenger.Class3     -2.069269   0.238929  -8.661  < 2e-16 ***
Sexmale              -2.632629   0.176375 -14.926  < 2e-16 ***
Age                  -0.038306   0.006712  -5.707 1.15e-08 ***
Siblings.and.Spouses -0.332316   0.103047  -3.225 0.001260 **
PortQ                -1.471228   0.444588  -3.309 0.000936 ***
PortS                -0.668459   0.212694  -3.143 0.001673 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1409.99  on 1042  degrees of freedom
Residual deviance:  954.88  on 1035  degrees of freedom
AIC: 970.88

Number of Fisher Scoring iterations: 5
```

Understanding the Output

- MLEs are approximately normal (for large samples); hence z-value and p-value calculation.

- Deviance: similar to R-Square in linear regression

- Null deviance: total variance in the response (only estimating the mean)

- Residual deviance: residual sum of squares

- Dispersion parameter: fixed at 1 as assuming Binomial distribution; related to quasi-likelihood