Supervised Learning & Logistic Regression

Advanced Data Analytics Applications and Methods

Recap

On Tuesday, we

- Defined predictive analytics
- Saw why it is valuable even when predictions may appear very inaccurate
- Reviewed the most ubiquitous PA technique of linear regression

Today

The moment you have all been waiting for (right?)…

We will move beyond the world of multiple linear regression:

- Supervised learning (AKA Classification)
- Logistic Regression (yes, we're back to regression)

Analytics Map

Predictive Analytics

Linear Regression

Supervised Learning (Classification)

Unsupervised Learning (Clustering)

Machine Learning

Linear Regression versus Classification

Linear Regression predicts *how much* something will happen

- Mathematically the *response* variable is a number

Classification predicts *whether* something will happen

- *response* variable is a *category* or *binary* event
- Ex: “Legitimate” or “Fraudulent” (e.g. credit card transaction)

Classification Example

“Can we find groups of customers who are likely to cancel their service soon after their contracts expire?”

- Clear target: Leaving service

“Can we predict which visitors will click on the advertisement and actually convert?”

- Clear target: Clicking and buying

Classification in Business World

Another condition should be added to classification problems:

- Data must be available on the target!

You want to know whether a customer will stay on for 6 months but data is retained only for 2

- Requires additional investment of acquiring data

Is it classification or regression?

"Will this customer purchase service S1 if given incentive 2?"

- Classification because it has a binary target

"Which service package (S1, S2, or none) will a customer likely purchase if given incentive 2?"

- Classification with a three-valued target

"How much will this customer use the service?"

- Regression because there is a numeric target of the amount of usage per customer

The first supervised learning technique

LOGISTIC REGRESSION

Motivating Example: Beer Preference

- *Hacker Pschorr*
  - One of the oldest beer brewing companies in Munich
  - Collects data on beer preference (light/regular) and demographic information
- Goal: determine demographic factors for preferring light beer

	A	B	C	D	E	F
1	Gender	Married	Income	Income (in $1000)	Age	Preferenc
2	0	0	$31,779	$32	46	Regular
3	1	1	$32,739	$33	50	Regular
4	1	1	$24,302	$24	46	Regular
5	1	1	$64,709	$65	70	Regular
6	1	1	$41,882	$42	54	Regular
7	1	0	$38,990	$39	36	Regular
8	1	0	$22,408	$22	40	Regular
9	1	1	$25,440	$25	51	Regular
10	0	1	$30,784	$31	52	Regular
11	1	0	$31,916	$32	43	Regular
12	1	0	$23,234	$23	31	Regular
13	0	1	$51,094	$51	46	Regular
14	1	0	$38,176	$38	40	Regular
15	1	0	$28,513	$29	34	Regular
16	0	1	$44,955	$45	53	Regular
17	0	1	$42,051	$42	58	Regular

Try Regression

Let's try to use Linear Regression to predict preferences:

- Code preference (the response) as

$$Y = \begin{cases} 1 & \text{if Light} \\ 0 & \text{if Regular} \end{cases}$$

- Fit the model

$$Y = a + b_1\, Gender + b_2\, Married + b_3\, Income + b_4\, Age + \varepsilon$$

Why not regression?

And the result is:

Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	-0.68189	1.930817	0.12472338	0.723967014	0.50566	0.011491	22.25219
Gender	-0.77789	0.716646	1.178209145	0.27772088	0.459376	0.112761	1.871451
Married	0.169661	0.794478	0.04560369	0.830897834	1.184903	0.249702	5.622685
Income	0.000278	6.33E-05	19.32410808	1.10305E-05	1.000278	1.000154	1.000403
Age	-0.22822	0.05239	18.97679446	1.32318E-05	0.795948	0.718275	0.882021

Issue: <u>You absolutely should not trust the p-values in this output.</u> Why? Try going through the standard diagnostic plots → assumptions are violated!

Motivating Example: Beer Preference

What do you predict is the preference for a male (gender=1), who is 25 years old, married with annual household income of $28,000?

Input Variables	Coefficient
Intercept	-0.68189
Gender	-0.77789
Married	0.169661
Income	0.000278
Age	-0.22822

$-0.68 - 0.77*1 + 0.17*1 + 0.0003*28000 - 0.22*28 = 0.81$

What does 0.81 represent?

The probability of preferring light beer

Motivating Example: Beer Preference

What do you predict is the preference for a male (gender=1), who is 25 years old, married with annual household income of $85,000?

Input Variables	Coefficient
Intercept	-0.68189
Gender	-0.77789
Married	0.169661
Income	0.000278
Age	-0.22822

-0.68 – 0.77*1 + 0.17*1 + 0.0003*28000 – 0.22*28 = 16.7

Issue: The predicted value doesn't really make much sense!

We are trying to predict a discrete outcome with a continuous function.

So the two issues:

- Assumptions behind linear regression are violated, so we can’t trust the p-values and other output
- The predictions can be weird

How can you proceed with regression?

Transform the variables

Logistic regression is about transforming the response variable

Logistic Regression Transformation

Instead of

$$\mathrm{Y} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Logistic regression does

$$log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

p represents the *probability* that $Y = 1$.

Logistic Regression Transformation

$$\log(odds) = log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

What is $log\left(\frac{p}{1-p}\right)$?

- The logarithm of the odds that Y=1.
  - If its positive, it is more likely that Y=1
  - If its negative, it is more likely that Y=0
  - If it equals 0, it is equally likely that Y=0 or Y=1

The Coefficients of the Logistic Model

- A positive regression coefficient means that an increase in a predictor increases the probability of the outcome
- A negative regression coefficient means that an increase in a predictor decreases the probability of the outcome
- A large (in absolute terms) regression coefficient means that the predictor strongly influences the probability of the outcome
  - If the predictors are normalized – if not we need to think about the size of the independent variables ($1 vs. $1000)

Odds and Logodds

If p=0.5, what are the odds? logodds?

If p=0.25, what are the odds? logodds?

If p=0.9, what are the odds? logodds?

Interpreting Coefficients

Holding all other variables fixed, a 1 unit increase in x_k changes the log(odds) that Y=1 by β_k, i.e., it makes Y=1 more likely when $\beta_k > 0$

Holding all other variables fixed, a 1 unit increase in x_k <u>multiplies</u> the <u>odds</u> that Y=1 by e^{β_k} , i.e., it makes Y=1 more likely when $e^{\beta_k} > 1$

Interpreting the Coefficients: The Odds Ratio

Logodds:

$$Log(odds) = \beta_0 + \beta_1\, Gender + \beta_2\, Married + \beta_3\, Income + \beta_4\, Age + \varepsilon$$

Odds:

$$odds(x_1, x_2, \ldots, x_k) = exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

Effect of increasing x_2 by one unit on odds, holding all other explanatory variables constant:

$$\text{Odds ratio} \longrightarrow \frac{odds(x_1, x_2+1, \ldots, x_k)}{odds(x_1, x_2, \ldots, x_k)} = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 (x_2+1) + \cdots + \beta_k x_k)}{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)} = \exp(\beta_2)$$

$exp(\beta_2)$ = *multiplicative* factor by which the *odds* (of the event Y=1) increase when the value of X_2 is increased by 1 unit and all other variables are held constant.

The Logistic Regression Model

A **nonlinear** regression model

$$Log(odds) = \beta_0 + \beta_1\, Gender + \beta_2\, Married + \beta_3\, Income + \beta_4\, Age + \varepsilon$$

Exponentiating both sides

$$odds = p/(1-p) = exp\{\beta_0 + \beta_1\, Gender + \beta_2\, Married + \beta_3\, Income + \beta_4\, Age + \varepsilon\}$$

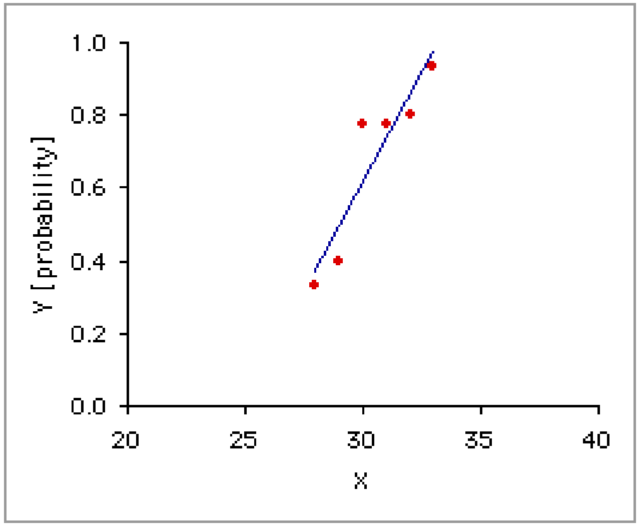
Solving for p

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \mathrm{GENDER} + \beta_2 \mathrm{MARRIED} + \beta_3 \mathrm{INCOME} + \beta_4 \mathrm{AGE} + \varepsilon)}}$$

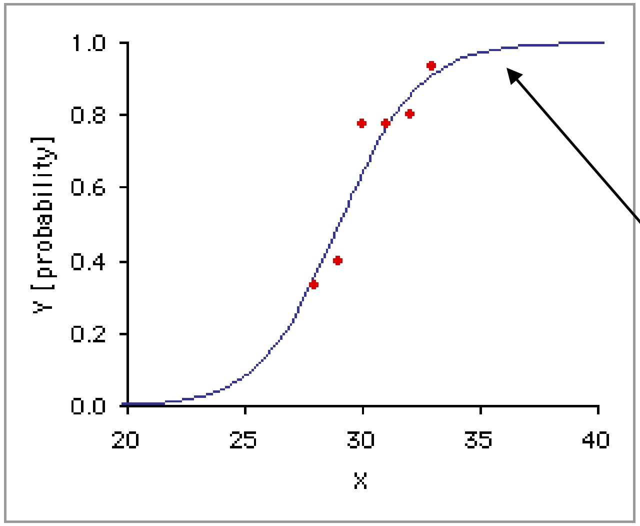
Bottom line: Logistic regression is a nonlinear function that maps any values of the input variables into a probability

Plotting the Logistic Relationship

Schematic for a single predictor:



Linear



Logistic

S-shaped / sigmoidal function

So what?

So the two issues:

1. Assumptions behind linear regression are violated, so we can't trust the p-values and other output
   - Solved! This is essential for policy discussions
2. The predictions can be weird
   - Solved! Predictions from logistic regressions are probabilities, so it always makes sense and are more accurate

The Use of Logistic Regression

Logistic Regression is used for predicting the probability of occurrence of an event

- Can use numerous predictor variables that can be either numerical or categorical

For example:

– the probability that a person accepts a personal loan may be predicted from knowledge of the person's age, sex and annual income

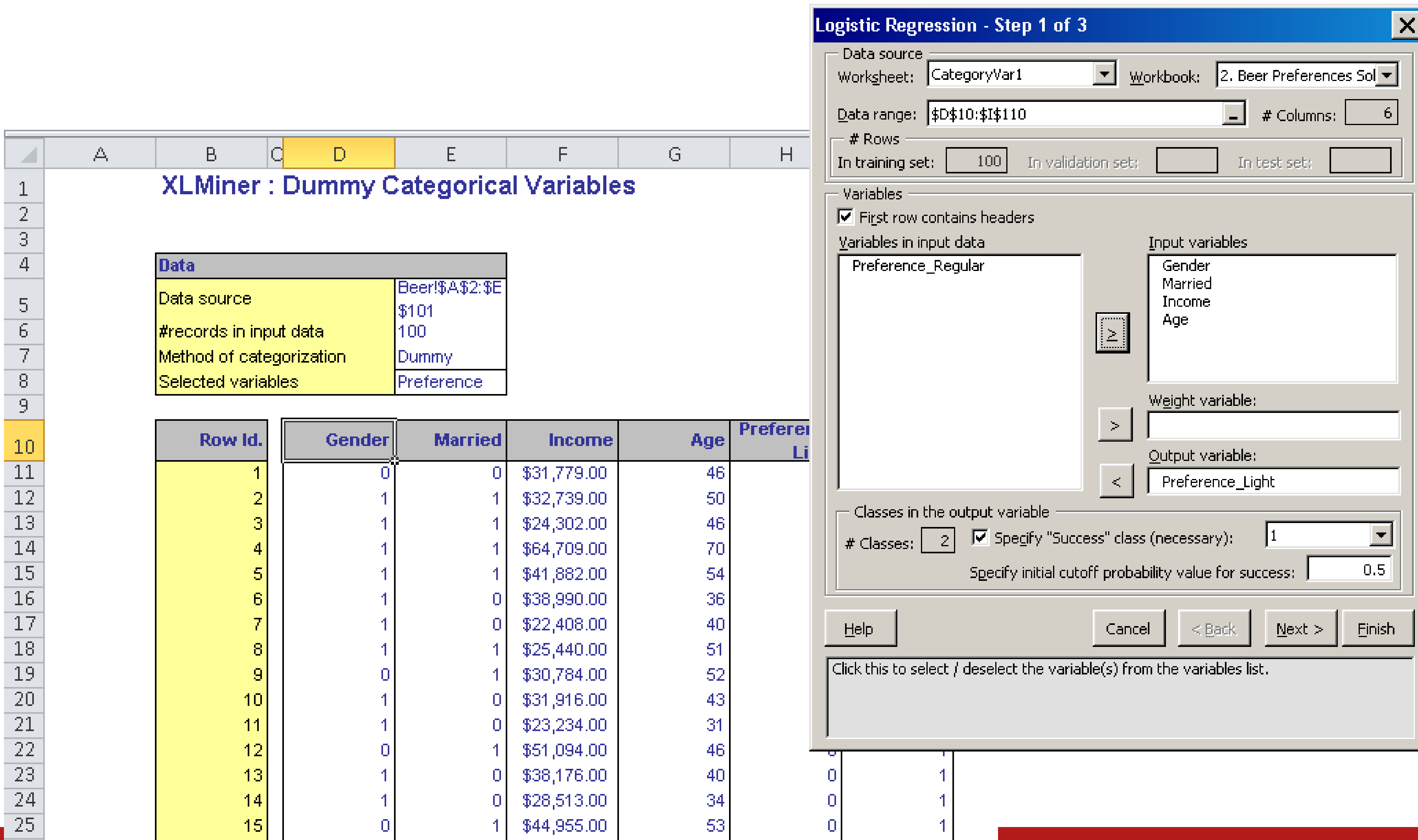
Used <u>extensively</u> in medical sciences and marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription

What do you need to know?

- When to use logistic regression versus linear regression
- How to interpret the coefficients from logistic regression
- How to use the output of a logistic regression model (we are about to discuss this)
- How do we know if the model is doing a good job? (we will discuss this next time)

Running LR in XLMiner

1. Open up the beer dataset posted on Blackboard
2. Create a dummy variable for the beer preference
3. Run the logistic regression using light beer as the output variable


XLMiner : Dummy Categorical Variables
Data
Data source
Beer!$A$2:$E$101
#records in input data
100
Method of categorization
Dummy
Selected variables
Preference
Row Id.
Gender
Married
Income
Age
Logistic Regression - Step 1 of 3
Data source
Worksheet:
CategoryVar1
Workbook:
2. Beer Preferences Sol
Data range:
$D$10:$I$110
Columns:
6
Rows
In training set:
100
In validation set:
In test set:
Variables
First row contains headers
Variables in input data
Preference_Regular
Input variables
Gender
Married
Income
Age
Weight variable:
Output variable:
Preference_Light
Classes in the output variable
Classes:
2
Specify "Success" class (necessary):
1
Specify initial cutoff probability value for success:
0.5
Help
Cancel
< Back
Next >
Finish
Click this to select / deselect the variable(s) from the variables list.

XLMiner Output

This matrix updates automatically in the output when the cut-off is changed

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	44	6
0	6	44

A quick way to see how well the model does for each class separately

Error Report			
Class	# Cases	# Errors	% Error
1	50	6	12.00
0	50	6	12.00
Overall	100	12	12.00

All the information you need about the model coefficients

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.68189073	1.93081641	0.72396708	*
Gender	-0.77788508	0.71664554	0.27772108	0.45937654
Married	0.16966102	0.79447782	0.83089775	1.18490314
Income	0.00027846	0.00006335	0.00001103	1.00027847
Age	-0.22822094	0.05238947	0.00001323	0.79594839

Increased Annual Income is associated with…

1. … higher probability of preferring light beer
2. … lower probability of preferring light beer
3. … we do not have enough information to conclude about the effects of annual income on preferring light beer

Increased Annual Income is associated with…

1. **… higher probability of preferring light beer**
2. … lower probability of preferring light beer
3. … we do not have enough information to conclude about the effects of annual income on preferring light beer

Interpreting Coefficients of Continuous Predictors: Beer Preference Example

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.68189073	1.93081641	0.72396708	*
Gender	-0.77788508	0.71664554	0.27772108	0.45937654
Married	0.16966102	0.79447782	0.83089775	1.18490314
Income	0.00027846	0.00006335	0.00001103	1.00027847
Age	-0.22822094	0.05238947	0.00001323	0.79594839

- Estimated coefficient of Age: b_{Age} =_______, or, *exp*(b_{Age})=________.
- Implies that a 1 year increase in age ___creases the odds of preferring light beer by a factor of ________, *for those with same gender, marital status & income*
- If age increases by 10 years(*but same gender, marital status & income)*, the odds of preferring light beer decreases by a factor of ____

Interpreting Coefficients of Continuous Predictors: Beer Preference Example

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-0.68189073	1.93081641	0.72396708	*
Gender	-0.77788508	0.71664554	0.27772108	0.45937654
Married	0.16966102	0.79447782	0.83089775	1.18490314
Income	0.00027846	0.00006335	0.00001103	1.00027847
Age	-0.22822094	0.05238947	0.00001323	0.79594839

- Estimated coefficient of Age: b_{Age} =__-.228__, or, *exp*(b_{Age})=__0.796__.
- Implies that a 1 year increase in age _de_creases the odds of preferring light beer by a factor of __0.796__, *for those with same gender, marital status & income*
- If age increases by 10 years(*but same gender, marital status & income)*, the odds of preferring light beer decreases by a factor of _*exp(-0.228*10)=0.102 → the odds decreases by 90%!*_

Interpreting Coefficients of Categorical Predictors: Beer Preference Example

- Estimated coefficient for Gender:

  $b_{\text{Gender}} = -0.778$, or,

  $\text{odds}_{\text{Gender}} = exp(b_{\text{Gender}}) = 0.46$.

- Implies that the odds of a **male** customer preferring light beer are 0.46 times the odds of a **female** customer *of the same marital status, age & income* preferring light beer.

Upcoming

- We will discuss how to validate predictive analytics models
- Move onto techniques that are very different from regression
- Transparent methods versus like black box methods

<u>Assignments</u>

- Short regression checkup is due by Fri midnight
- Homework 1 is due after break and focuses on linear regression

Extra Slides

Practical Implication of Nonlinear Regression on Interpretation

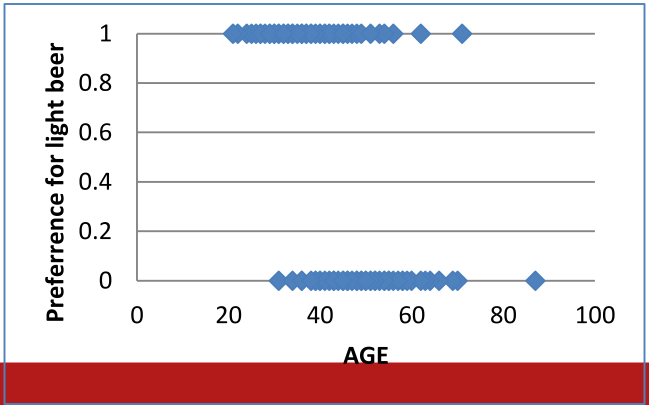
- Probability that a 20-year-old married woman, earning $40,000/year prefers light beer:

$$\hat{p}_{Light} = \frac{1}{1+e^{-6.06}} = 0.99767$$

- What if the same customer was **25 years old**? $\hat{p}_{Light} = \frac{1}{1+e^{-4.92}} = 0.9928$

- What if the same customer was **40 years old**? $\hat{p}_{Light} = \frac{1}{1+e^{-1.50}} = 0.817$

- What if the same customer was **45 years old**? $\hat{p}_{Light} = \frac{1}{1+e^{-0.356}} = 0.588$



Using Model for Classification/ Prediction

What is the probability that a male, 25 year old and married with annual household income of $85,000 prefers light beer?

Solution:

1. Use estimated model to obtain *logit*

2. Estimate p = probability that Y=1

Finding the coefficients

- Logistic Regression: relationship between *Y* and beta parameters is nonlinear.
- Least squares method may not work well
- Hence use maximum likelihood estimation
  - Find estimates that maximize chance of obtaining the data we have.

Finding the coefficients

- Suppose we have 3 data points: $(Y_1 = 0, d_1, t_1), (Y_2 = 1, d_2, t_2), (Y_3 = 0, d_3, t_3)$
- *Given* $\beta_0, \beta_1, \beta_2$, *what is:* $P(Y_1 = 0, Y_2 = 1, Y_3 = 0)$?

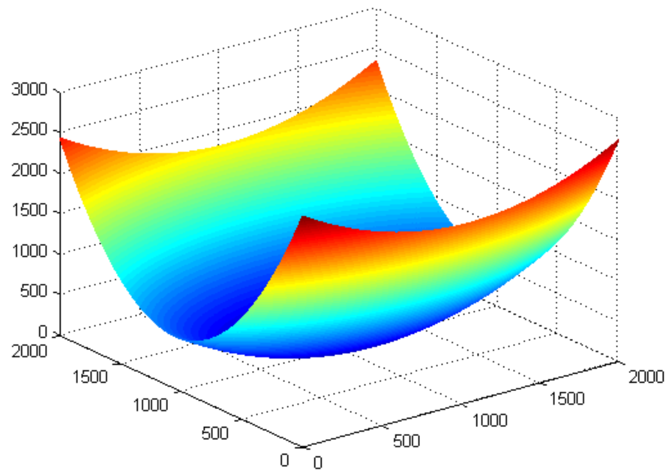
$$P(Y_1 = 0, Y_2 = 1, Y_3 = 0) = P(Y_1 = 0)P(Y_2 = 1)P(Y_3 = 0) =$$

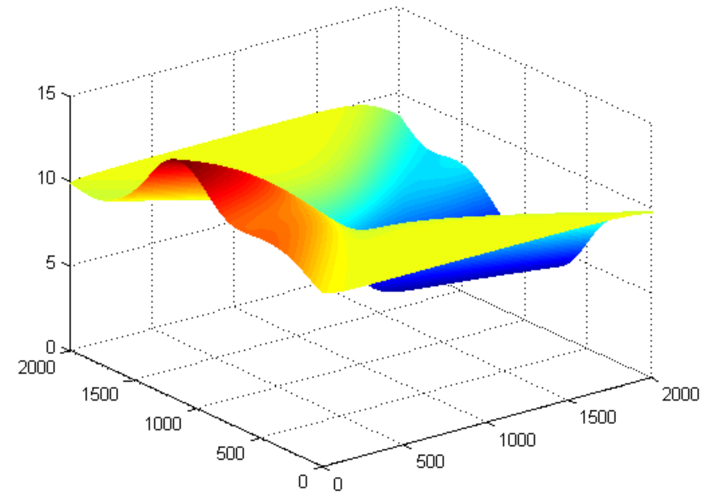
$$= \frac{1}{\left(1+e^{\beta_0+\beta_1 d_1+\beta_2 t_1}\right)} \cdot \frac{e^{\beta_0+\beta_1 d_2+\beta_2 t_2}}{\left(1+e^{\beta_0+\beta_1 d_2+\beta_2 t_2}\right)} \cdot \frac{1}{\left(1+e^{\beta_0+\beta_1 d_3+\beta_2 t_3}\right)} = f\left(\beta_0, \beta_1, \beta_2\right)$$

- *Find* $\beta_0, \beta_1, \beta_2$ *to max* $f(\beta_0, \beta_1, \beta_2)$.
- *This is called maximum likelihood estimation.*

Why Least Squares may not work

- For just about any data, here are the Sum of Squared Errors for linear regression and logistic regression respectively:

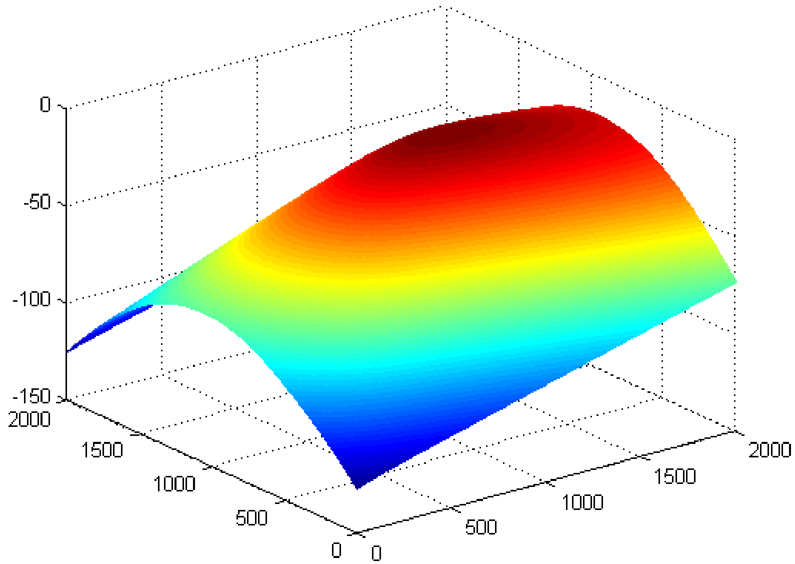




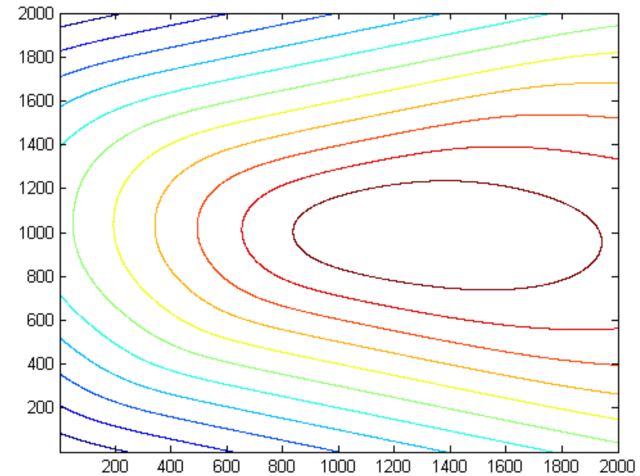
- The former is “bowl-shaped,” the latter is irregular – making minimization difficult

Log-Likelihood of Logistic Regression Model

- To be maximized is the “dome shaped”



Log Likelihood



Contour map of Log Likelihood