



Predicting a Continuous Outcome

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

The Regression Task

- So far we have considered the **classification task**
- In this task the goal is to predict the value of a **categorical outcome**
- In order to do this we have training data that includes the value of both **predictors and outcome**
- Another type of supervised learning is prediction of the value of a **continuous outcome**; this is called the **regression task**
- We will still have training data that includes the value of both **predictors and outcome**

Multiple Linear Regression

Let Y_i be the outcome variable, and x_{ij} be the j th predictor value, for the i th observation. Multiple linear regression does prediction of Y_i as a linear function of the predictors:

Multiple linear regression model with p predictors:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

$$\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

$$i = 1, \dots, n$$

Multiple Linear Regression

The assumptions of multiple linear regression are:



Multiple Linear Regression

The training step for multiple linear regression consists of:



The prediction step consists of:

Example: Cheese Data

Example: 30 samples of cheese; for each we know:

Taste: Subjective taste test score, obtained by combining the scores of several tasters

Acetic: Log of concentration of acetic acid

H2S: Log of concentration of hydrogen sulfide

Lactic: Concentration of lactic acid

Cheese Data

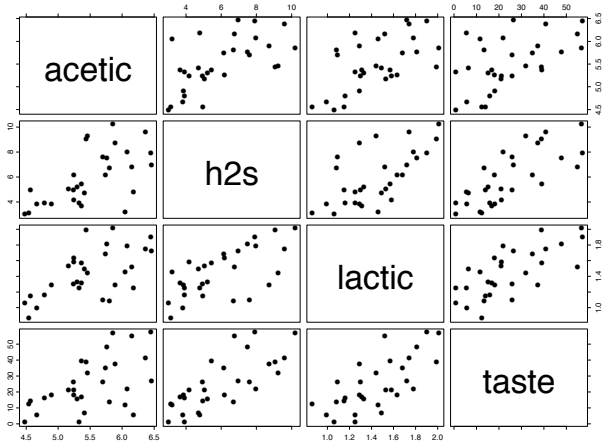
Case	taste	Acetic	H2S	Lactic
1	12.3	4.543	3.135	0.86
2	20.9	5.159	5.043	1.53
3	39	5.366	5.438	1.57
4	47.9	5.759	7.496	1.81
5	5.6	4.663	3.807	0.99
6	25.9	5.697	7.601	1.09
7	37.3	5.892	8.726	1.29
8	21.9	6.078	7.966	1.78
9	18.1	4.898	3.85	1.29
10	21	5.242	4.174	1.58
11	24.0	5.71	6.112	1.60

Cheese Data

- Goal: to **predict taste** based on the chemical composition.
Why?

Cheese Data

Pairwise scatterplots (interpretation?):



Cheese Data

How would you characterize the relationships?



So fit the linear regression model with Y_i equal to the taste score, and the predictors x_{i1} , x_{i2} , and x_{i3} equal to the values of Acetic, H₂S, and Lactic.

Multiple Linear Regression

- R code for TRAINING step on cheese data (estimation of $\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2$ using the 30 cheese samples):

```
cheeseLM = lm(formula = taste ~ acetic + h2s + lactic,  
              data = cheese)
```

```
summary( cheeseLM )
```

R Syntax

Side note about R:

- In R, functions like `lm` have arguments, like `formula` and `data`
- When you call the function you can write:
`myFunction(myArgument1 = "hello", myArgument2 = 5)`
- It also works if you leave out the argument names, but only if all the arguments are in the correct order:
`myFunction("hello", 5)`

Cheese Linear Model

Results for cheese data:

```
Call: lm(formula = taste ~ acetic + h2s + lactic, data = cheese)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-17.39	-6.612	-1.009	4.908	25.45

```
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-28.8768	19.7354	-1.4632	0.1554
acetic	0.3277	4.4598	0.0735	0.9420
h2s	3.9118	1.2484	3.1334	0.0042
lactic	19.6705	8.6291	2.2796	0.0311

```
Residual standard error: 10.13 on 26 degrees of freedom
```

```
Multiple R-Squared: 0.6518
```

```
F-statistic: 16.22 on 3 and 26 degrees of freedom, the p-value is  
3.81e-006
```

Cheese Linear Model

Mark the estimate of β_0

Mark the estimate of β_2 , the regression coefficient for H2S

Mark the estimate of σ

```
Call: lm(formula = taste ~ acetic + h2s + lactic, data = cheese)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-17.39	-6.612	-1.009	4.908	25.45

```
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-28.8768	19.7354	-1.4632	0.1554
acetic	0.3277	4.4598	0.0735	0.9420
h2s	3.9118	1.2484	3.1334	0.0042
lactic	19.6705	8.6291	2.2796	0.0311

```
Residual standard error: 10.13 on 26 degrees of freedom
```

```
Multiple R-Squared: 0.6518
```

```
F-statistic: 16.22 on 3 and 26 degrees of freedom, the p-value is  
3.81e-006
```

Multiple Linear Regression



- Recall that in **naive Bayes** we calculated $Pr(Y|X_1, \dots, X_p)$ from our model in order to predict the value of Y .
- In **linear regression** we are also calculating $Pr(Y|X_1, \dots, X_p)$, the distribution of Y given X_1, \dots, X_p , from our model, to predict the value of Y

Multiple Linear Regression

- For the cheese example, what do we predict the taste rating to be if Acetic = 5.1, H₂S = 9, and Lactic = 1.2?

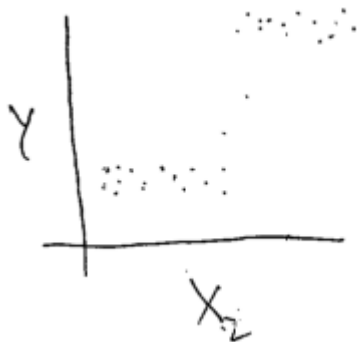
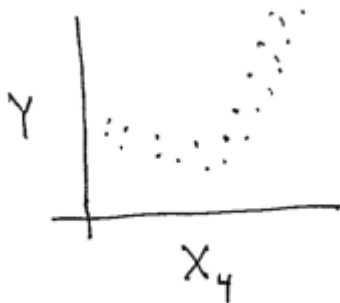


Multiple Linear Regression

When should we transform the (outcome or predictor) variables?

Nonlinearities in a continuous outcome

- What do we do when we're predicting a continuous outcome and the relationship between the predictors and outcome is nonlinear?



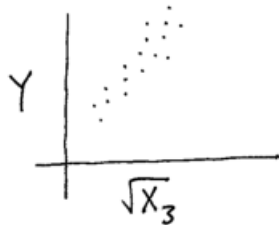
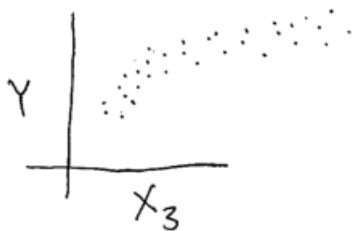
Nonlinearities in a continuous outcome

Option 1:

- Transform the predictor so that the relationship is linear

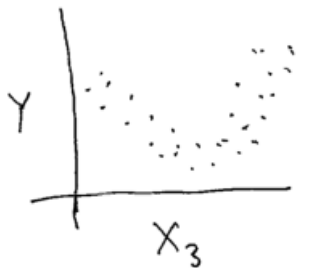
Note: This works sometimes but not always.

example of when this works:



Nonlinearities in a continuous outcome

Example of when this doesn't work:



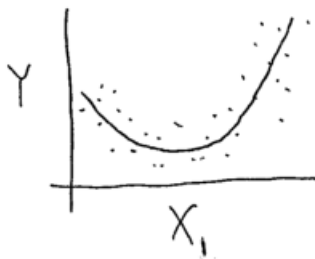
Reason:

Nonlinearities in a continuous outcome

Option 2:

- Include polynomial terms in the regression:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \epsilon_i$$

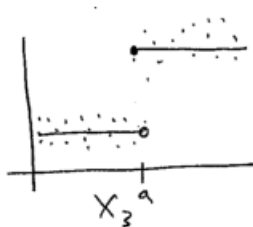
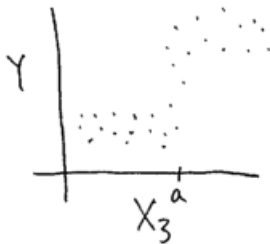


Nonlinearities in a continuous outcome

Option 3:

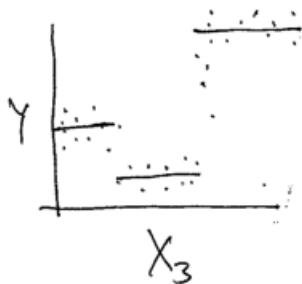
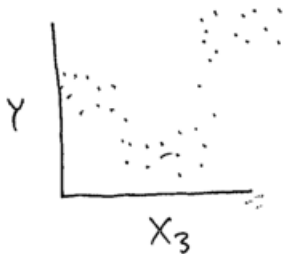
■ Discretize the predictor:

e.g.: cut into $X_3 < a$, $X_3 \geq a$; then the linear regression for this new binary predictor looks like:



Nonlinearities in a continuous outcome

- If you instead discretize into more categories you can get a more flexible model

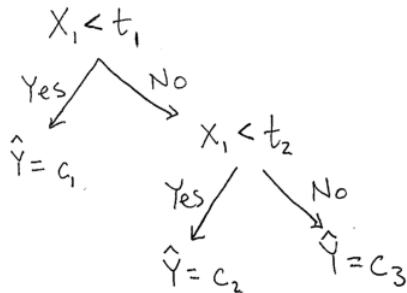
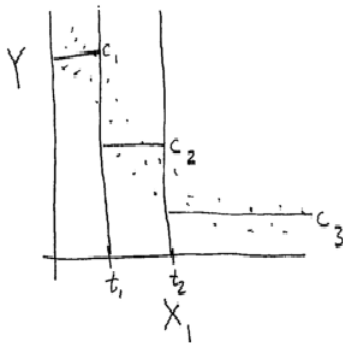


Nonlinearities in a continuous outcome

- A method based on this discretization idea is regression trees.
- The corresponding method for categorical outcomes is called classification trees; we saw an informal version of these in the first week of class.

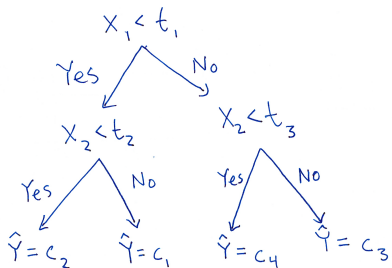
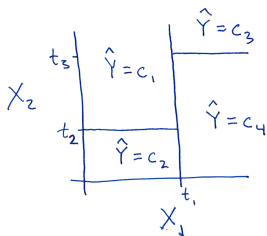
Nonlinearities in a continuous outcome

Ex 1: One predictor X_1



Nonlinearities in a continuous outcome

Ex 2: 2 continuous predictors X_1 , X_2



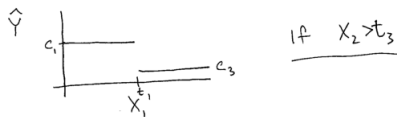
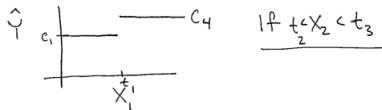
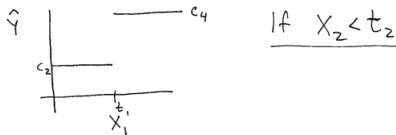
Regression Trees

Regression tree benefits:



Regression Trees

Back to Ex 2:



Notice the **interaction**: the effect of X_1 on \hat{Y} depends on the value of X_2 .