Week 12 Hour 3

Odds Ratio

Logistic regression and the odds ratio

Logistic regression on two variables

(As time permits) Other categorical options

Odds Ratio

As the name suggests, odds ratio is the ratio of odds under two different conditions.

Example: If the odds of having lung cancer by age 70 is 0.15 if you smoke tobacco, and 0.008 if you don't smoke anything, then the odds ratio of getting cancer for smoking vs. non-smoking is…

$$\text{Odds1} / \text{Odds2} = 0.150 / 0.008 = 18.75$$

We can also compute the odds ratio by hand in cases where there is only one explanatory variable, and it is also categorical.

Consider a sample of 20 heart attack patients, in which we know...

- Whether they had a second heart attack within a year (response variable, categorical, 2 levels)

- Whether they attended traditional anger management therapy after their first heart attack (explanatory, categorical, 2 levels)

We can describe the relationship between these two variables as a crosstabulation, or ***crosstab*** for short.

	Anger Management Therapy	
2^{nd} Heart Attack	None	Traditional
No (0)	4	6
Yes (1)	7	3

This means, for example, that 7 of the 20 patients had a second heart attack and did not receive anger management therapy

	Anger Management Therapy	
2nd Heart Attack	None	Traditional
No (0)	4	6
Yes (1)	7	3

We can estimate the odds of a 2nd attack under each condition:

Odds of 2nd attack with no therapy: Pr(event) / Pr(not event),

estimated by # with 2nd attacks / # without 2nd attacks

= 7 / 4

= ***1.75***

	Anger Management Therapy	
2nd Heart Attack	None	Traditional
No (0)	4	6
Yes (1)	7	3

Likewise, we estimate

Odds of 2nd attack WITH therapy: 3 / 6 = ***0.5***

Then we estimate the odds ratio:

Odds without therapy / Odds with therapy = 1.75 / 0.5 = ***3.5***

So the odds of second heart attack are 3.5 times as high without anger management therapy.

That estimate of the true odds ratio comes from a sample, so it is only a ***statistic .***

We only had 20 patients in the sample, so that statistic is going to come with a LOT of uncertainty as well.

$$SE_{log(OR)} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3} + \frac{1}{n_4}}$$

$$SE_{log(OR)} = \sqrt{\frac{1}{3} + \frac{1}{7} + \frac{1}{6} + \frac{1}{4}}$$

You don't need to know this formula, but you do need to know the consequences of it:

1. The standard error gets smaller as the sample ($n_1 + n_2 + n_3 + n_4$) gets larger.

In other words: We become more confident in our results as we collect more information.

2. A very small group (low n) can make the standard error large, no matter how big the other groups are.

In other words: Our results are only as good as our smallest group *(important for rare diseases or cases!!!)*

$$\sqrt{\frac{1}{3}+\frac{1}{500}+\frac{1}{500}+\frac{1}{500}} \approx \sqrt{\frac{1}{3}}$$

Don't let this new information overwhelm you.

The odds ratio is frequently cited in news articles on health:

Red meats and tuna

In the study, Latinas who consumed about 20 grams of processed meat per day (the equivalent of a strip of bacon) were 42 percent more likely to be diagnosed with breast cancer compared to Latinas who ate little or no processed meats, said Andre Kim, lead author and a USC molecular epidemiology doctoral student.

"42 percent more likely to be diagnosed…" means there was a cancer odds ratio between the '20g bacon' and 'no bacon' groups of 1.42.

Source: More meat, more problems: Bacon may increase breast cancer risk in Latinas. U of South Carolina News, Zen Vuong, March 3 2016

The odds ratio is popular because of its tie to logistic regression.

Log(Odds Ratio) = Log(Odds1 / Odds2)

Log(Odds Ratio) = Log(Odds1) – Log(Odds2)

So the difference between two log-odds is the log of the odds ratio.

…and we have a regression system that works in log-odds already.

We've seen the difference between two things in a regression context before.

Specifically, as the coefficient of a dummy variable.

Consider a linear regression with only one categorical variable (broken down into dummy variables in the summary).

The coefficient of each dummy variable is the difference between the mean response in the dummy's group and the baseline group.

Now consider a **logistic** regression with only one categorical variable.

As before, the categories are presented as dummy variables.

The coefficient of each dummy is the difference in the mean log-odds between that group and the baseline group.

In short, it's the ***log of the odds-ratio***.

Example: Say we have a logistic regression model of getting colon-cancer in the next 5 years. The equation for this model is

This model uses one categorical variable: Daily Processed red meat (e.g. bacon) consumption.

A-None

B-Low (0-25g per day)

C-High (26+ g per day)

$$\text{Log-Odds(cancer)} = -4.54 + 0.16(\text{Meat} = \text{Low}) + 0.27(\text{Meat} = \text{High})$$

Log-Odds(cancer)

$$= -4.54 + 0.16(\text{Meat} = \text{Low}) + 0.27(\text{Meat} = \text{High})$$

The odds ratio of cancer of the 'low meat' group vs the 'no meat' group is exp(0.16) = 1.17

So the odds of getting cancer are ***17% higher*** if you consume 1-25g of processed red meat per day.

Likewise, exp(0.27) = 1.31.

So the cancer odds are ***31% higher*** if you consume 26 or more grams per day.

Log-Odds(cancer)

$$= -4.54 + 0.16(\text{Meat} = \text{Low}) + 0.27(\text{Meat} = \text{High})$$

What else does this tell us?

The log odds of developing colon cancer for someone in the 'no bacon' group is ***-4.54***

The cancer odds are exp(-4.54) = ***0.0107***

The cancer probability is 0.0107 / (1 + 0.0107) = ***0.0106***.

So even though an increase in cancer odds of 31% may seen like a lot, that's 31% more than 0.0107.

Cancer odds in the high meat group:

$0.0107 * 1.31 = 0.0140$

Alternatively, we could find the 'high meat' group odds with

$\exp(-4.54 + 0.27) = \exp(-4.27) = 0.0140$

However, if you're not interested in the raw odds or probability, so much as what you can do to improve your cancer risk, then the odds ratio is a lot more useful.

Also, without the p-values, we don't know if these dummy variables are statistically significant. That is, we don't know if the modelled changes to the cancer odds are real or just the result of random variation.

Something to ponder

Log-Odds when controlling for another variable.

In the previous example, there's a lot of objections one could make about the model, especially if you work for the meat industry and don't want the bad publicity.

Maybe colon cancer is caused by age, and it's just that older people tend to eat more red meat.

We can include additional variables in the model to control for these other factors.

Consider another model

Log-Odds(Cancer) = -4.00 + 0.063(Age) + 0.080(Meat = Low) + 0.44(Meat = High)

If this were a linear regression, we would say that the age coefficient is saying "The response is increasing by 0.063 per year, holding sex and meat consumption constant)"

And that the 'Meat=Low' coefficient is saying
"The response is 0.08 higher when Meat=Low vs when Meat=None, holding age constant'

In logistic regression, the interpretations are the same. It's just that 'response' is 'log-odds of cancer'.

From the model, exp(0.08) = 1.083

So if you consume 1-25g of processed red meat per day, your odds of colon cancer are ***8% higher*** than someone ***of the same age*** who eats none.

Likewise, exp(0.44) = 1.56

So the cancer odds are ***56% higher*** if you consume 26 or more grams per day *even when controlling for age*.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.00358    0.55558  -7.206 5.76e-13 ***
meatB.low    0.08024    0.17886   0.449   0.6537
meatC.high   0.44332    0.17447   2.541   0.0111 *
age          0.06321    0.01042   6.067 1.30e-09 ***
```

Now consider the model with the full summary, not just the regression equation.

From the p-values we see that there is strong evidence to say that colon cancer rates increase with age, holding diet constant.

This isn't the sort of information that makes the news though.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.00358    0.55558  -7.206 5.76e-13 ***
meatB.low    0.08024    0.17886   0.449   0.6537
meatC.high   0.44332    0.17447   2.541   0.0111 *
age          0.06321    0.01042   6.067 1.30e-09 ***
```

We also have evidence that those with high-meat diet have a higher rate of colon cancer *than their peers of the same age*.

It seems to be only the high-meat diet with a significant difference. Could this be due to multiple comparisons?
There are 3 pair-comparisons, so the Bonferroni adjusted alpha is...

Pair-wise alpha: 0.05 / 3 = 0.0167 > obtained p, 0.0111

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.00358    0.55558  -7.206 5.76e-13 ***
meatB.low    0.08024    0.17886   0.449   0.6537
meatC.high   0.44332    0.17447   2.541   0.0111 *
age          0.06321    0.01042   6.067 1.30e-09 ***
```

Better still than saying 'the odds-ratio is 1.56, and it is significant', we could produce a confidence interval of this ratio.

Confidence Interval = Estimate +/- (Critical* Standard Error)

= 0.443 +/- (1.96 * 0.175)

= 0.443 +/- 0.343

= ***0.100 to 0.786***

Estimate of the log-odds ratio: 0.443 (OR = 1.56)

Confidence interval of the log-odds ratio: 0.100 to 0.786

Confidence interval of the odds ratio:

exp(0.100) to exp(0.786)

= ***1.105 to 2.195***

So our best estimate is that a high-meat diet increases the odds of cancer by 56%, but by taking a ***confidence interval***, we find that the real increase in odds could be anywhere from a 11% increase to a 120% increase.

Note that everything to do with the confidence interval, modelling, and the hypothesis tests are done in log-odds. It is only when we need to translate these numbers into something a human can read that that we transform back.

Also note that we always use 1.96 for the critical value of a 95% interval with odds ratio.

We are assuming a large enough degrees of freedom that we can use the normal distribution and not the t-distribution.
It's not always a well justified assumption, but it gets around the nightmare of computing degrees of freedom for weighted cases.

Like in a linear regression, logistic regression effects are additive, meaning that an increase in one explanatory variable is taken separately from other variables. Predictions are taken by adding the effects from different variables.

If the logistic model were describing probability, an increase of 4% could turn a response from impossible (Pr = 0) to possible (Pr = 0.04).

That same additive effect could be relatively small (Pr = 0.57 increased to Pr = 0.61).

It could also break the rules of probability. (Pr = 0.99 up to Pr = 1.03) But none of this happens in log-odds.

A more realistic scenario still is to study cancer rates using a large number of variables together.

What if we had a number of other variables that could affect cancer rates, such as sex, body-mass index, exposure to lead, family history, and other factors already known to be associated with cancer?

The odds ratios associated with different levels of red meat in a diet would be a comparison holding all these other known factors constant.

In short, we would be able to isolate the effect of red meat.

..and this is exactly what the World Health Organization did.

Logistic regression isn't the only option for modelling categorical response variables. Other methods include…

Ordinal logistic regression

Clustering

Regression trees

Ordinal logistic regression is an extension of logistic regression to three or more categories.

The ***'Ordinal'*** in implies that the categories are in a ***strict order***.

The following sets of categories are ordinal. The categories move from one extreme to the other.

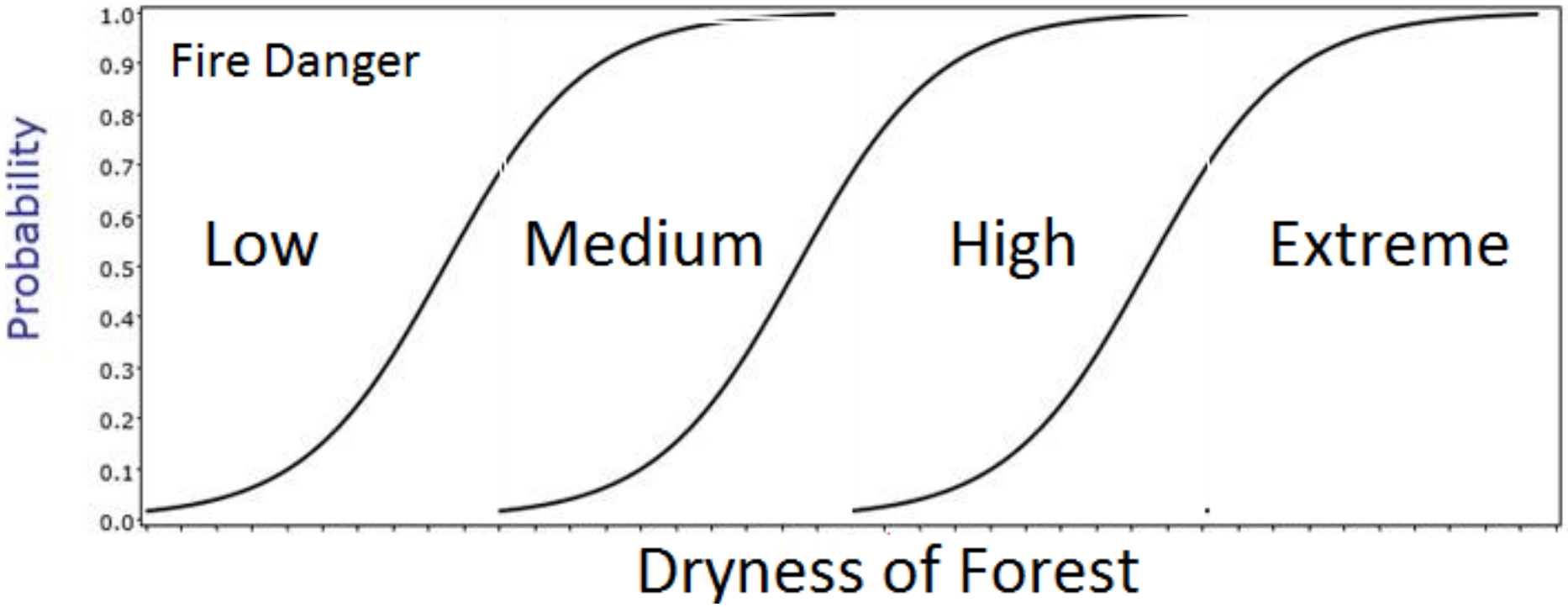
'Low', 'Medium', 'High' and 'Extreme' fire danger.

'Strongly Agree', 'Agree', 'Neutral', 'Disagree', and 'Strongly Disagree' with a given opinion.

The following sets of categories are NOT ordinal. There either isn't any natural order to the categories, or that order is disputable.

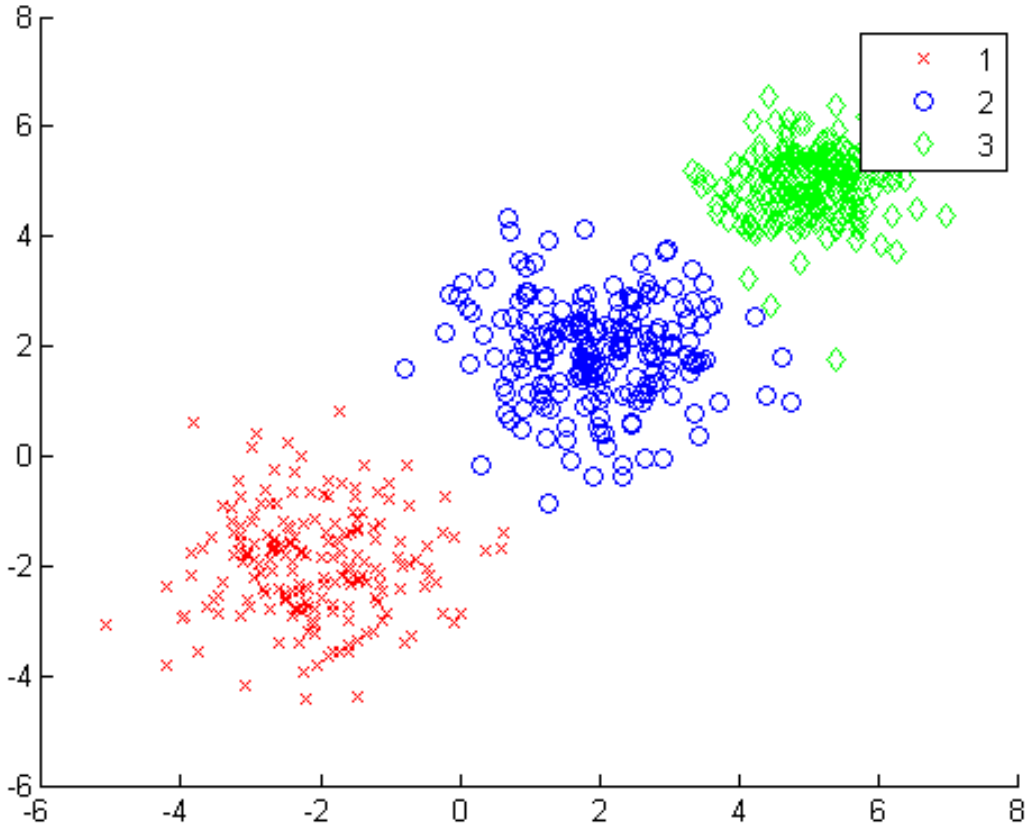
'Cat', 'Dog', 'Dragon', 'Horse', for type of pet. (No ordering)

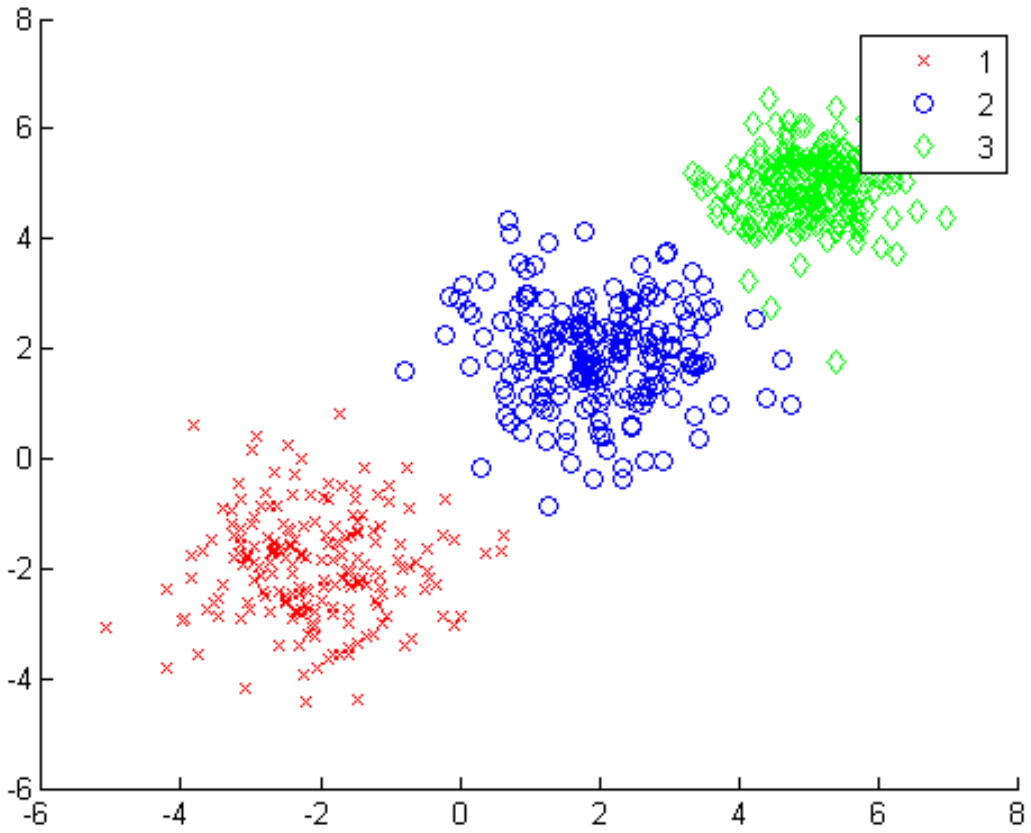
'None', 'Deaf', 'Blind', 'Blind AND Deaf', for disabilities. (Disputable ordering)



An ordinal logistic regression (also called ordered logit) creates logistic regression equations with the same slopes, and uses each one to determine if a response is above or below a certain category.

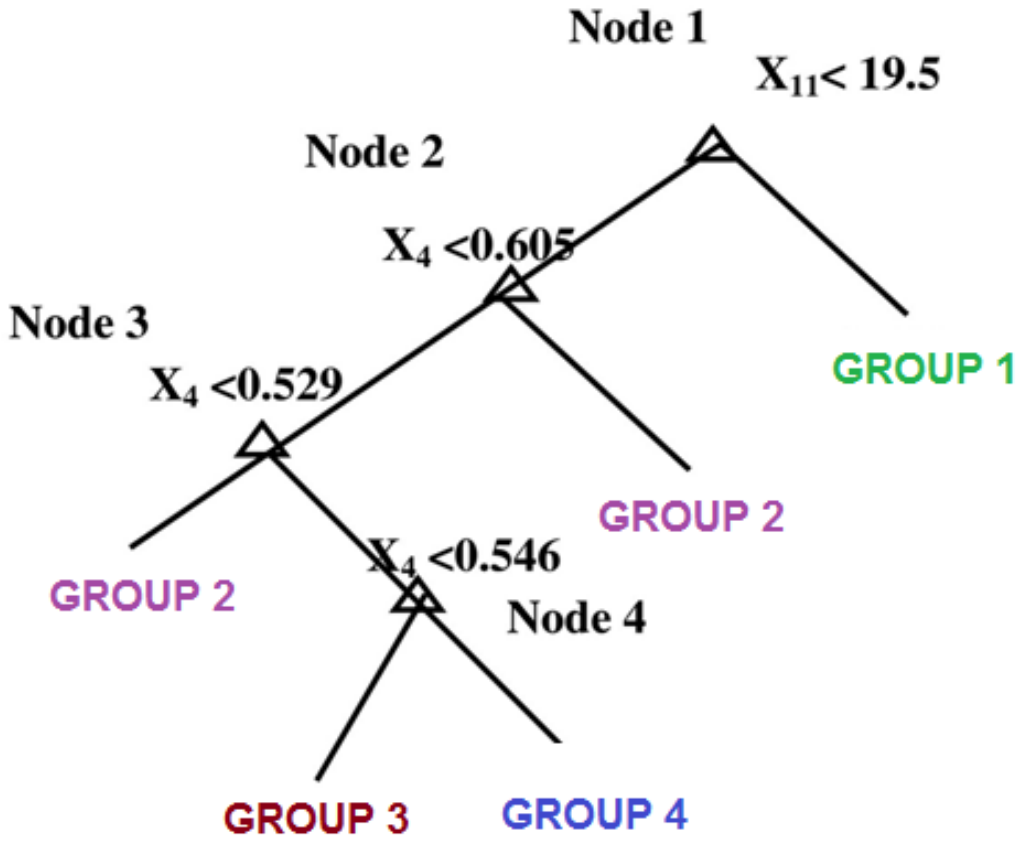
Clustering, or cluster analysis, is a generic term for methods where data points are partitioned into groups, but the groups are unknown.





In this diagram, there are two continuous explanatory variables, shown as position x1 and x2. The response is the group, which is show as colour.

Regression trees were a model selection method that was briefly mentioned in Week 9. It's mentioned here too because it's flexible enough to have numeric OR categorical responses.



Easter bonus image:

Chocolate bunny melted by a blow dryer.