Stat 414 - Day 13
Overview of Logistic Regression

Last Time: Multiple random slopes

- Each random slope adds a slope variance parameter, plus covariances with intercepts and any other random slopes. (The number of correlation terms is equal to the number of unique pairs among Level Two random effects)
  - Try to minimize use of random slopes or model gets very complicated very quickly
  - Can zero out covariances to simplify model but makes sense in context?
- Centering variables can sometimes help with convergence

Example 1: Between 1972-1974 a survey was taken in the Whickham district of the United Kingdom (Appleton et al., 1996; Simonoff, 2003), including information such as smoking status and age. Twenty years later, a follow-up study was conducted, and it was determined whether the interviewee was still alive. First consider the smokers and non-smokers:

	Smokers	Non-smokers	Total
Alive	443	502	945
Died	139	230	369
Total	582	732	1314

(a) What is the response variable? Quantitative or categorical?

still alive or not; categorical - so probably can't use OLS

There are several statistics we could use to compare the likelihood of being alive between the smokers and non-smokers, including

- *difference in conditional proportions* (443/(443+139) - (502)/(502+230)) = 0.075
  The difference in conditional proportions has some limitations, namely if the success probability is small, you will be working with small numbers and so it is difficult to look at the difference and say "that's large" or not.
- relative risk = ratio of conditional proportions (443/(443+139) / (502)/(502+230)) = 1.11
  The relative risk helps you see whether one value is large compared to the other value, but it is problematic to use with "case-control studies" (Find some successes and find some failures ⇒can't turn around and use the data to estimate the probability of success.)
- odds ratio = ratio of *odds* of success where odds is the proportion of successes divided by the proportion of failures, (443/582)/(139/582) / (502/705)/(203/705) = (443/139) / (502/203) = 1.29.
  Odds ratio doesn't have either of these issues, but is more difficult to interpret.

(a) Compute and interpret the odds ratio of being alive for smokers (numerator) compared to non-smokers (denominator). Also report the percentage change (subtract 1 and multiply by 100% and report as a decrease).

(443/139) / (502/203)
= 1.29

29% increase

(502/203)/(443/139)
= 1/1.29 = 0.775

1-.775 = .225 =>
22.5% decrease

You should be more bothered by these data suggesting that smoking is beneficial for your health! So we want to "adjust" for possible confounding variables.

Consider the following data

Age	21	29	39	49	59	69	79
Alive	114	273	209	169	145	35	0
Interviewed	117	281	230	208	236	165	77

(d) Does there appear to be evidence that those who were older when they were first interviewed were less likely to be alive at the follow-up interview? How would you suggest modelling these data? Give some downsides to using a linear model in this case.

older people tended to be less likely to be alive after 20 years

Of course, when we have a relationship we want to fit a line, but that's not appropriate here (and generally for proportions as the response) for two main reasons:

- We can't extend the line much further without predicting probabilities below 0 or above 1
- The relationship is usually not linear.

To solve the second issue, we want to transform the data or use a polynomial model. But remember the transformations we saw before were for "monotonic" relationships. With proportions we tend to see more of an "S-shaped" curve where the "response" values approach zero in one direction and approach one in the other direction. So we will use a different kind of transformation.

Definition The logit transformation is $ln(\pi/(1-\pi))$ which is equivalent to the log odds of success.

The relationship should be more linear and I don't have any problem with the response going off to plus or minus infinity. So the logistic regression model is:

$$expected\ log(\pi/(1-\pi)) = \beta_0 + \beta_1 x$$

We can fit a logistic regression model in R using glm (generalized linear model) rather than lm.

Notice this data file is in "count /frequency format" or "grouped" (already have the counts for each possible explanatory variable combination) so we can think of each row as a binomial random variable where we have the observed number of successes and the sample size for that binomial random variable. Also treating age as a quantitative variable.

```
model1 = glm(cbind(alive, failures)~ ageQ, family=binomial("logit"), data = WhickhamData)

## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.379340   0.401261   18.39   <2e-16 ***
## ageQ        -0.122768   0.006982  -17.58   <2e-16 ***
## ---
##  Null deviance: 641.496  on 13  degrees of freedom
## Residual deviance:  35.654  on 12  degrees of freedom
## AIC: 86.65
```

So the fitted model is $predicted\ log\ odds\ of\ alive = 7.38 - 0.123age$
Clearly age is statistically significant ($z = -17.58$) and with a negative coefficient, which seems to imply the probability of being alive 20 years later is smaller for individuals who were older at the time of the first interview.

(e) To interpret the intercept, what are the predicted log odds of being alive at age = 0? What are the predicted odds of being alive at age = 0?

exp(7.38) = predicted odds = 1603.58

Again, most people don't have good intuition for odds, so you can convert this back to a probability by using the relationship $probability = odds/(1 + odds)$
(f) What is the predicted probability of someone who was a newborn in Whickham UK at the start of the studying being alive 20 years later?

1603.58/1604.58 = .9994 = predicted probability of a newborn (age = 0)

To interpret the slope, we start off as usual with "a one-unit increase in x…" If you do the algebra, this corresponds to an $exp(\hat{\beta})$ predicted *multiplicative* increase in the response.
(g) Estimate the decrease in the odds of survival with each additional year of age. What about a 10 year age difference?

exp(-.123) = .884, each additional predicts a decrease in odds of survival by 12% so a ten-year age difference correspond to .884^10

Now let's go back to the smoking variable

```
model2 = glm(cbind(alive, failures)~ smoking.status, family=binomial("logit"), data = WhickhamData)
##                          Estimate Std. Error    z value Pr(>|z|)
## (Intercept)               0.78052    0.07962   9.803  < 2e-16 ***
## smoking.statussmoker      0.37858    0.12566   3.013  0.00259 **
## Null deviance: 641.5  on 13  degrees of freedom
## Residual deviance: 632.3  on 12  degrees of freedom
## AIC: 683.29
```

Smoking.status is a categorical variable, remember that R creates a 0/1 variable in the model.

(h) Provide an interpretation of the slope coefficient in this model. How does it compare to the odds ratio we computed by hand above?

exp(.37858) = 1.46 => smokers have 1.46 times higher odds of survival compared to nonsmokers (should match the by hand calculation when all have in the model is smoking status)

You saw above that age is related to the response and it turns out the real question is whether there is an association between smoking status and survival status *after adjusting for age*.

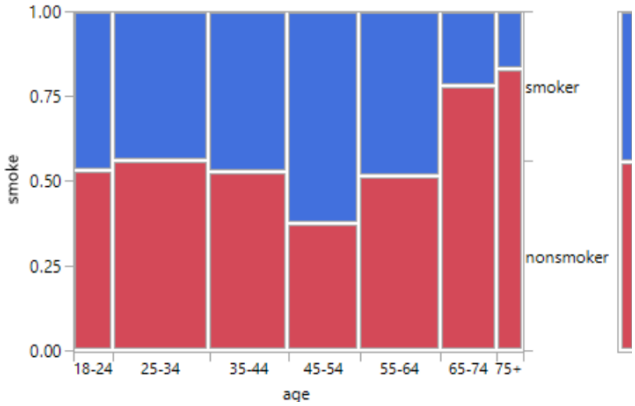
```
model3 = glm(cbind(alive, failures)~ ageQ + smoking.status, family=binomial("logit"),
##                          Estimate       Std. Error z value Pr(>|z|)
## (Intercept)              7.645629       0.444911  17.185   <2e-16 ***
## ageQ                     -0.125365      0.007286 -17.206   <2e-16 ***
## smoking.statussmoker     -0.265067      0.168707  -1.571    0.116
## Null deviance: 641.496  on 13  degrees of freedom
## Residual deviance:  33.163  on 11  degrees of freedom
## AIC: 86.159
```

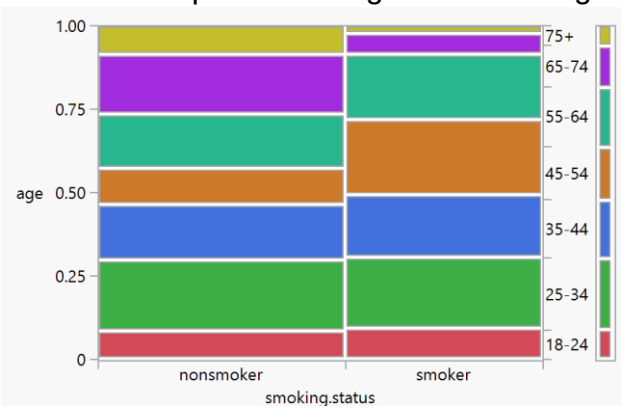
exp(-.2651) = 0.77

(i) What do you learn about the association between smoking.status and probability of being alive, after adjusting for age? Interpret as if to a non-statistician.

So if compare two people of the same age, one is a smoker and one is not, the odds of surviving for 20 years are 0.77 times smaller for the smoker compared to a nonsmoker

To see why this is happening, we can explore the relationship between age and smoking





(j) Explain what you learn and how this relates to the above analyses.

smokers tend to be a little younger than the nonsmokers in this dataset, and so comparing smokers to nonsmokers was comparing younger to older

Summary: Logistic Regression allows us to model the log odds of success for a *categorical response variable* based on any number of quantitative or categorical predictor variables. In general, if x_j is increased by one unit (all other variables fixed), the odds of success, that is the odds that $Y = 1$, are multiplied by $e^{\hat{\beta}_j}$. (And the estimated increase in the odds associated with a change of d units is $exp(d \times \hat{\beta}_j)$.)

With a binary predictor, $exp(\beta)$ is the ratio of the population odds when $x = 1$ to the odds for $x = 0$, more directly the odds ratio between these two groups.

Notes:

- The conclusions are the same no matter which outcome is labeled as success vs. failure.
- A multilevel logistic regression model will look like:

  Random intercepts model: $ln\left(\frac{\pi_j}{(1-\pi_j)}\right) = \beta_0 + u_{0j}$ where $u_{0j} \sim N(0, \tau_0^2)$

  What's missing? Why?