Stat 414 – Day 16
Logistic Regression (Ch. 17)

Last Time: Correlation in residuals (constant variance)
- Correlation matrix = σ^2 **I**
- Compound Symmetry (can model directly or induced by random intercepts)

  $$\begin{pmatrix} 1 & \rho & \rho & \cdots \\ \rho & 1 & \rho & \cdots \\ \rho & \rho & \ddots & \vdots \end{pmatrix}$$
  one new parameter
  "exchangeability"
- AR(1)

  $$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \ddots \\ \rho^2 & \rho & 1 & \ddots \end{pmatrix}$$
  one new parameter
  takes into account distance in time, requires equally spaced timepoints
- More complicated structures can be induced by random slopes
  - Can help explain *causes* of correlation
  - Akin to "unstructured variance" (G-matrix)

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?

(b) Calculate the ***odds*** (proportion of success/proportion of failures) of survival for the smokers. How do these odds compare to the non-smokers? How do you interpret the ***odds ratio***? Does anything bother you about this result? Do you have an explanation?

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

(c) 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.

When working with proportions, you may be able to fit a linear model, but often a linear model is not appropriate. The expected S-shaped curve implies a power transformation is not likely to be helpful either. Instead we can try a ***logit transformation***: $\ln(\pi/(1-\pi))$. This is equivalent to using the *log odds* as the response variable.

(d) Create this transformed variable and consider whether the relationship with age is more linear. Also consider the behavior of this function as the probability approaches 0 or 1. What if the probability equals 0.5?

(e) Fit the logistic model in R. What does the intercept tell us? What does the slope tell us?

(f) Fit the logistic model in R. Use this equation to find the estimated odds of survival for a smoker compared to a non-smoker. How does this compare to your calculation in (a)?

(g) Fit the logistic model with both smoking and age and interpret the coefficient of smoking.status.

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 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, $e^{\beta 1}$ 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 (instead of the multiplicative change in odds).

Note: The conclusions are the same no matter which outcome is labeled as “success” vs. “failure”!

Example 2: Data were collected on 5366 women who recently gave birth in Bangladesh from 361 communities. The main variables of interest to us are:

- antemed = whether or not receive prenatal care at least once during pregnancy
- mage = mother's age at the child's birth in years
- comm = community identifier
- urban = type of residency with 1 = urban and 0 = rural

(a) Create and describe the variation in the proportions who receive prenatal care in the different communities. Are they significantly different?

(b) Write out a "random intercepts" logistic regression model

(c) Fit this random intercepts model to predict the probability of receiving prenatal care while accounting for the multilevel structure of the data. Report and interpret the parameter estimates from the output. How many are there? What's missing? Why?

There are different suggestions for calculating an intraclass correlation coefficient. I'm partial to $\frac{\tau_0^2}{\tau_0^2 + \pi^2/3}$ where $\pi^2/3 \ = 3.29$ comes from the variance of a logistic distribution.

(d) Calculate an intraclass correlation coefficient for the random intercepts model.

(e) Fit the random intercepts model to predict the probability of receiving prenatal care while accounting for the mother's age when the child was born (allowing the odds to vary within each community). Write out the model equation and interpret the slope and the intercept.

(f) Fit the random slopes model for (e) and decide whether the new model is a significant improvement.

(g) How does the intraclass correlation coefficient change? Why?

(h) Predict the probability of care for a 40-year-old mom.

(i) Should *urban* be added to this model? With or without random slopes?

(h) Summarize what you learn from the model.

Example 3: (Example 17.1 – Religiosity across the world)
The response variable is whether or not someone attends religious services at least once a week. The data is from the European Values Survey/World Values Surveys collected in 1990-2001 for 136,611 individuals in 60 countries.

(a) Describe the variation in the proportions who attend religious services regularly in the different countries. Are they significantly different?

(b) Remove Turkey, and (patiently) fit the empty model. What is the estimated “average” probability of religious attendance? Do the Level 2 residuals appear approximately normal?

(c) Try to recreate Table 17.2. What do you conclude?