**GLMs in R: Basic Generalized Linear Regression Modelling**

Once again, the files you need to submit are your **R script and this document in word or PDF by 11:59pm on the due date**. Paste in your code and output into this document where it is specifically requested (or a plot is requested – always paste those in). If directions are given but nothing is asked directly, an “OK” response will suffice. Name your files following this convention (where X is the number of the HW) and email them to Mike:

epid799b\_hwX\_lastname.pdf (Homework document with answers filled in)

epid799b\_hwX\_lastname.r (R script).

**Readability.** A reminder, please, for your sake and the instructor’s, make efforts to comment, create code blocks, and write legible code. However, on answering questions in this document, please write as little as you can – short answers and sentence fragments are encouraged.

**Recoding**. Homeworks 2 through 5 extend on the preceding homework, so please build this onto the foundation build previously (e.g. recoding, etc.). If necessary, update your preceding homework sections with the released answers.

Complete the follow steps in R/RStudio:

# Part 1: Tabular Effect Estimates

1. **Create a summary data.frame/table for crude effects:** Using either dplyr:: functions or table() and prop.table() as we did in class, report the effect of early prenatal care on preterm birth by:
	1. the crude risk difference
	2. the crude risk ratio
	3. the crude odds ratio
2. **Create a summary data.frame/table to report race-ethnicity specific crude RDs**: again by using dplyr or (in this case) a three dimensional table()/prop.table(), report the risk difference effect of prenatal care on preterm birth by race-ethnicity status. Paste that table or text below.
3. **Create a graphic** like that below showing the risk prevalence by exposure status and race-ethnicity.
*Hint: the line (you could also use a ribbon) is created using geom\_hline(). The text is from the annotate() function, which can be used to place any single geometry element directly on the graph without needing a data.frame or an aesthetic mapping.*



# Part 2: Create and explore a crude (no EMM) risk difference model

1. Using glm(), **create a model object**, m\_crude\_rd, that models the crude effect of prenatal care on preterm birth and report the risk difference (the non-intercept coefficient in the crude model when you print or ask a summary() of the model object. For risk differences, you’ll need the binomial family with the identity link. Confirm this risk difference matches your tabular result from part 1.
2. Your risk difference (which the non-intercept coefficient in the crude model represents) should be negative, representing the reduction in preterm birth prevalence with exposure to early prenatal care. If it isn’t, **relevel your factor** to provide the appropriate contrast. If you happen to coded the preterm factor correctly the first time, be sure you know how to relevel your factor to reverse the sign on the coefficient/RD.
3. Describe the **structure of the model** object using str(). The model object can be large, so you may need the max.level parameter (set to 1) to reduce what’s printed. You do not have to describe every element, just the overall structure. Be able to select elements of the model structure using our usual syntax for the data structure type.
4. **Report the coefficients and confident intervals** of this crude risk difference model using coef() and conf.int() functions.
5. Use **broom::tidy** to turn the most commonly used model components into a data.frame. Use cbind() to combine broom::tidy()’s results with conf.int()’s – this is often a useful output structure for graphing.
6. Use ggplot2 to plot your model results (point estimate and confidence interval) and paste in here. Perhaps a geom\_point()+geom\_errorbar()? If you don’t do the below challenge, you’ll just have one estimate and CI. *Hint: I’m using (see below) two geom\_point()s and a geom\_linerange() with color mapped to the standard error to mimic Tufte style boxplots.*


***Optional, but recommended for epis****:
Run additional models besides the crude, then use tools like data.frame(), subset operators (for example, to just pull out the PNC coefficient results) rbind() and cbind() to build a single data.frame of all your results, and plot them using ggplot2. Here is an example, but yours could look different.*

# Optional Part 3: Epi-specific methods and stats

1. **Optional:** Model and report the **risk ratio and odds ratio** with their confidence intervals. For RRs you’ll need binomial with the log link; for ORs you’ll need binomial with logit links. You’ll need to need to exponentiate the estimates. *Hint: Nesting functions can give you what you want: exp(coef(m)).*
2. **Optional**: Compare at least two models using Likelihood Ratio Tests (LRTs). Try anova(model1, model2, test=”LRT”). Note that you’ll need to ensure they are nested. One way to do this is to first subset your data to include just (but all) of the variables you will use in your fullest model, then use na.omit() to drop records missing any of those variables, then pass that same restricted dataset to each of your glm() calls. You may note how mage linear mage performs without magesq, for instance.

*Epi note: per 718 it is best (at least, as we’re taught at UNC!) to reduce a model from the full model through bias-variance trade-off decisions, not LRTs alone, though LRTs can be informative of improved overall fit of the model. As I understand it, this is for a number of reasons, including (1) overall and importantly, we’re interested in change in causal effects, not just model fit and (2) note that LRTs require nesting and the same number of observations. Not including a variable in a model allows you to include observations where only that additional variable is missing, so to use more records, leading potentially to increased precision.*