Refining the Association Among Race, Education, and Health

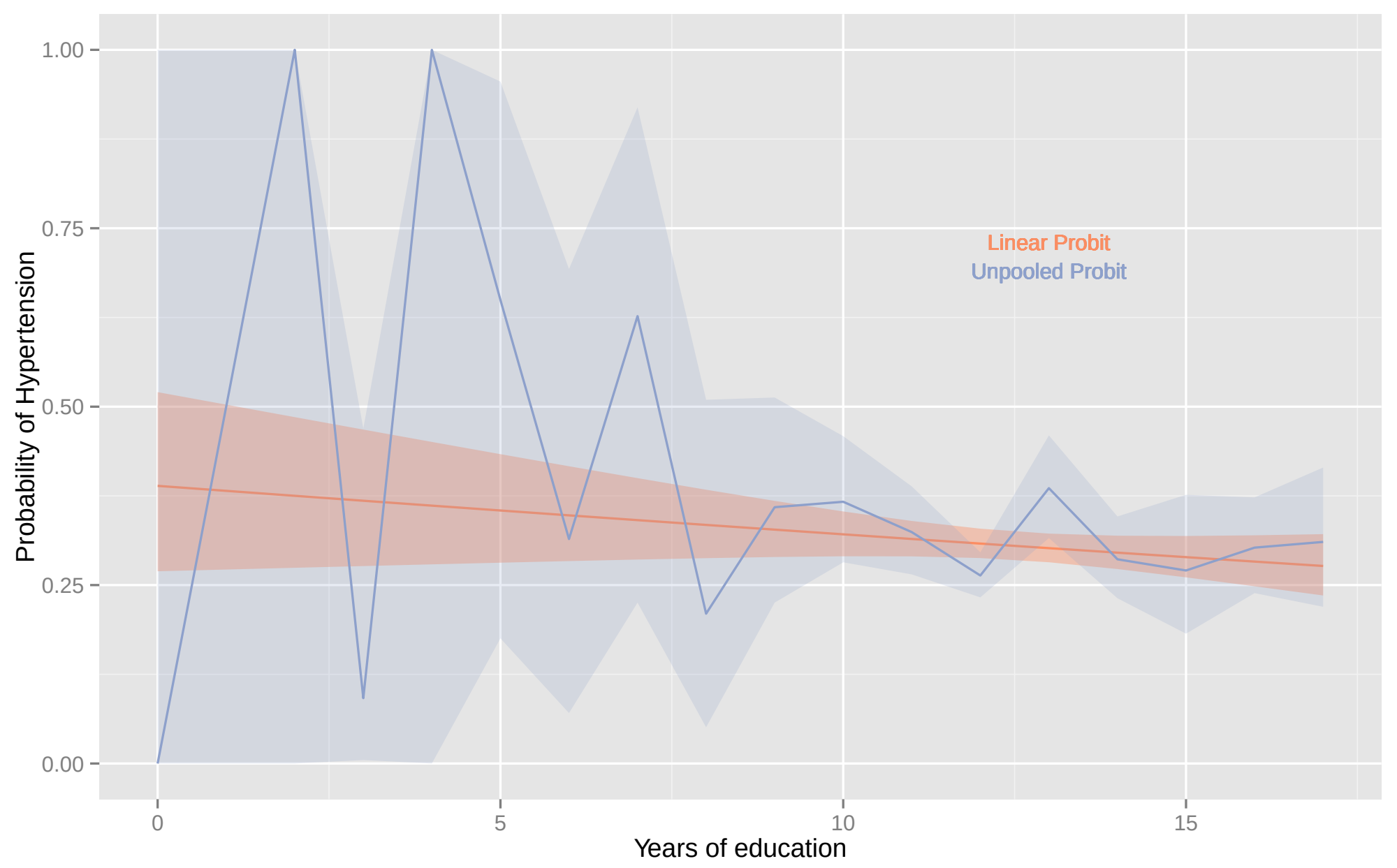
Michael Esposito
University of Washington Sociology

Questions

1. *Does the association between education and health vary across race?*

Background

- In population health, that education and health are associated is well established. It is clear from this research that some educational gradient exists, where greater education generally predicts better health outcomes.
- What's less clear is the exact functional form of this gradient.
  - Some studies find that education has a linear association with health, where each additional year of education yields the same level of health benefit. Others show that the association varies depending on what point the increase in education occurs (Walsemann 2013).
- To complicate matters, a body of research has shown that how health and education are associated is dependent on race. At the moment, the results from this literature are somewhat heterogeneous.
  - Using a fairly simple model, Farmer and Ferraro (2005) show that education has a linear relationship with self rated health for whites where no association exists for blacks...
  - ...but, using a fairly complex model, Kimbro et. al. (2008) give some evidence that education is associated with better health for blacks, but that in a number of cases, this association is weaker than it is for whites (though the lack of representation of uncertainty leaves guessing on where evidence of difference actually exists).
- The heterogeneity in results is likely due in part to the modeling strategies used by the different authors.To better illustrate this point, we can look at the following plot:



- The linear probit gives our predictions when education's relationship with health is assumed to be linear and continuous. This could be underfitting the data by overlooking meaningful nuances in the set of the predictions.
- The unpooled probit predicts the same thing, but allows each level of observed education to predict the outcome itself. This could be overfitting the data by making the predictions overly dependent on the noise in the data. In the case of Kimbro et. al., this concern is magnified by the inclusion of splines and a large set of interaction terms.
- (Perhaps) motivated by these concerns, Montez et. al. (2012) explored how morality risk varied across racial × gender × age groups by testing 13 functional forms that might describe the focal association, and assessing which best balanced complexity and generalizability via BIC comparisons.
- This project is in the spirit of Montez et. al. (2012) with a few distinctions. Perhaps most importantly, this project does not pre-select a set of candidate models for testing. Instead I choose a method that allows the data itself to find the best model for how education, race, gender and age jointly impact health outcomes.

Bayesian Additive Classification and Regression Trees

- This nonparametric Bayesian method (referred to as BART from now on) is well suited for the question at hand:
- BART consist of a sum-of-trees model:

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

where Y is a health outcome, x is a set of criteria, T_j is a binary regression tree with the M_j set of bottom nodes (i.e., μ_{ij}), and $g(\cdot)$ is a function that maps a value of M_j to x. The sum of the values gained from each tree produces an overall predicted value for an individual with covariates x (Chipman et. al. 2010).

Bayesian Additive Classification and Regression Trees cont.

- Also contains a a regularization prior:
  - ▷ Nice 'default' choices for priors that work well in many situations:
    - → $p(T_j)$ with preferences for fewer bottom nodes (and thus less 'complicated trees')
    - → $p(M_j)$ centered around 0, which shrinks values of bottom nodes towards 0
    - → $p(\sigma)$ that puts σ at something a little less than what would be given with least squares.
- The use of many trees allows for the functional form returned to be complex and flexible, while the priors help reduce overfitting.
- Uses backfitting MCMC algorithm to get at posterior:

$$p((T_1, M_1)...(T_m, M_m), \sigma | y)$$

  - On a high level, algorithm is a Gibbs sampler of m successive draws of:

$$(T_j, M_j) | T_{(j)}, M_{(j)}, \sigma, y$$

where (j) indexes the set of all units minus j, and m is the number of trees used then,

$$\sigma | T_1, ..., M_1, ..., M_m, y$$

Analysis

(1) Use 2011 PSID
(2) Model several health outcomes as function of `age, race, gender` and `education`
(3) Check for overfitting via 10 fold cross validation
(4) Find predicted values for various covariate combinations and plot
(5) Compare shapes across racial groups and comment

Results

- In each of the following models, the cross-validated, out of sample metric was very similar to the corresponding in-sample metric, which provided evidence that the following forms generalize well.

Figure 1: *Pr*(Hypertension) for Black Men, White Men, Black Women, and White Women

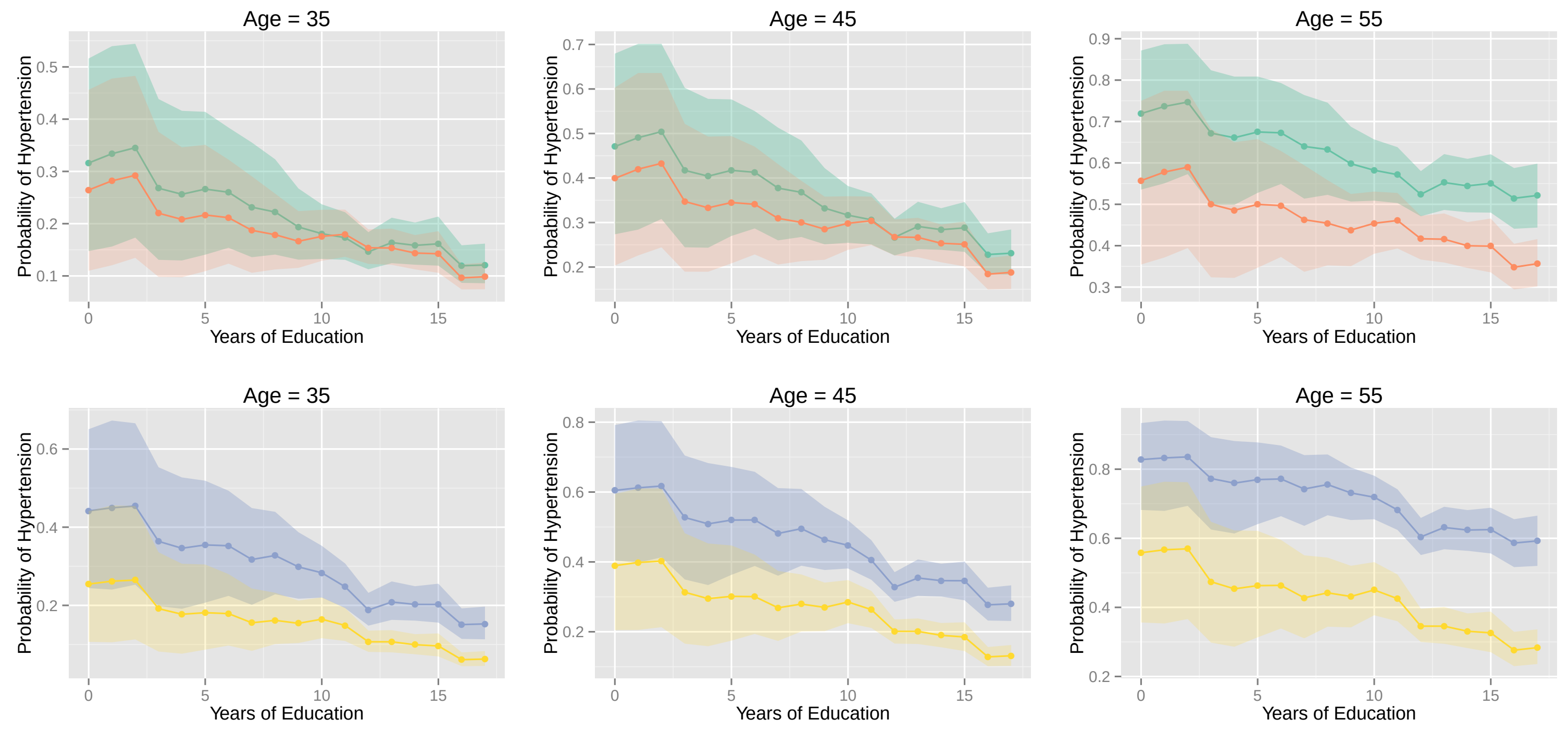


Figure 2: *Pr*(Excellent Health) for Black Men, White Men, Black Women, and White Women

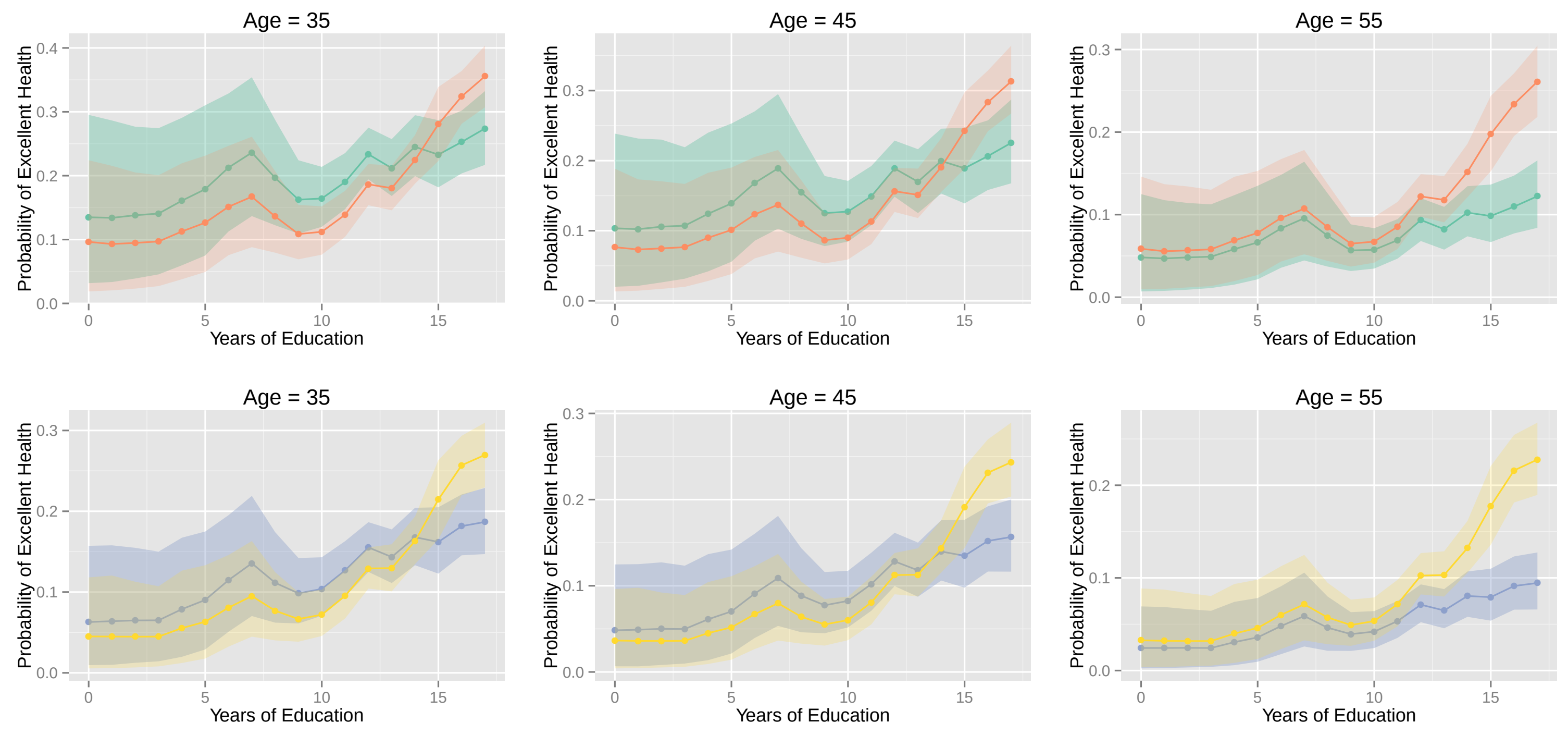


Figure 3: Predicted BMI for Black Men, White Men, Black Women, and White Women

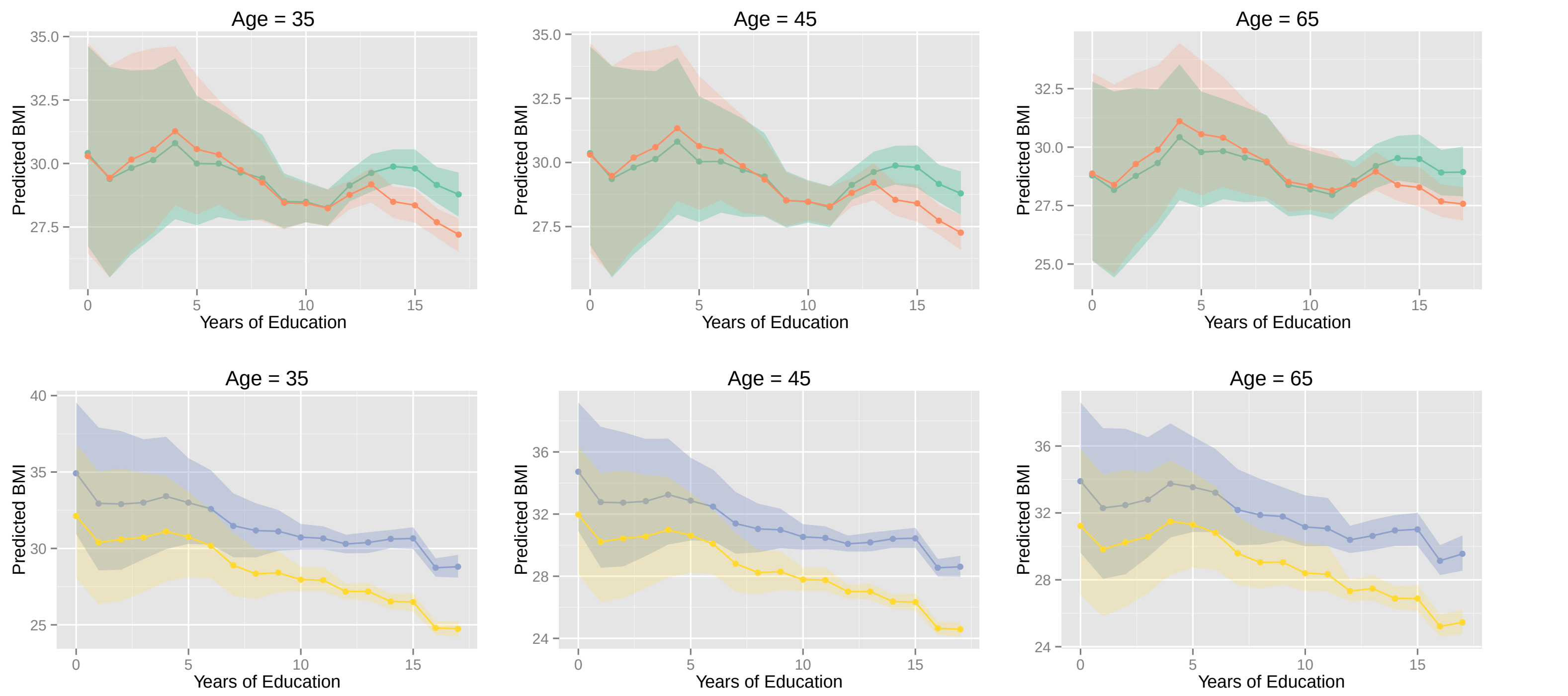


Figure 4: *Pr*(Diabetes) for Black Men, White Men, Black Women, and White Women

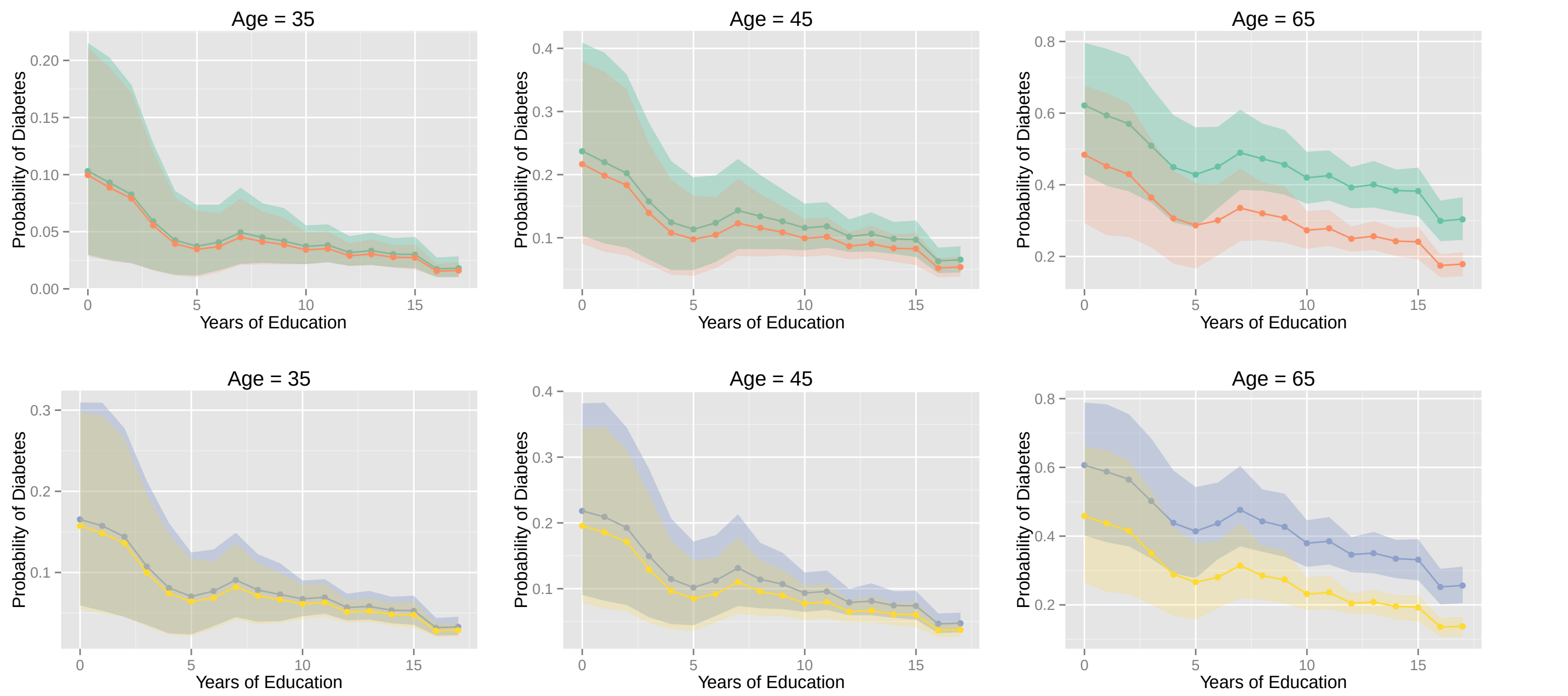
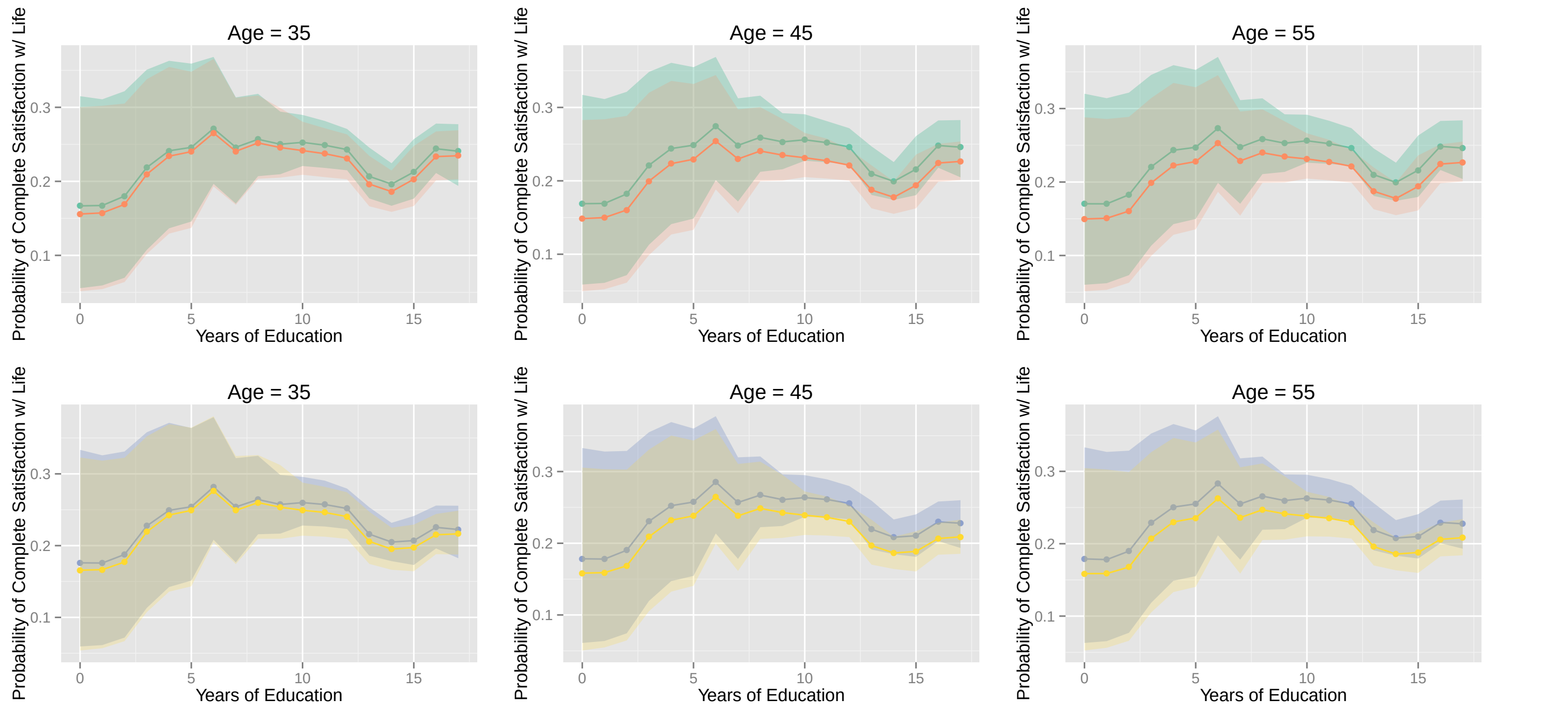


Figure 5: *Pr*(Diabetes) for Black Men, White Men, Black Women, and White Women



Conclusions

- This evidence suggests that the best form for the education-health association is more of a compromise between the existing findings in the literature
  - Not a simple as a linear, monotonic improvement in health per year of education...
  - ...but, not so complex that a fairly monotonic function can't adequately describe the association.
- It also suggests that the *form* of the association is not so dramatically different across racial groups (as previously reported).
- Further, more complex interactions appear to play a role here too
  - E.g., For *Pr*(Hypertension), age appears to to interact with `race` and `education` in a way that produces a wide gap between the predictions for blacks and whites at older ages.

Contact Information

- Email: esposm2@uw.edu
- Phone: +1 (816) 315 3662
- Twitter: @mhEspo