Biost 312: Modern Regression Analysis

Lab 7: Modeling Dose-Response

All problems refer to the salary dataset as found on the class web pages. This is a very large file, so you need to make sure you have sufficient memory available when you start Stata. Also, it is probably most convenient if you code the variables as numbers, and use labels to make them more understandable.

In these problems, you are frequently asked to save the predicted values in order to examine them in a plot in a later question. The fitted values can be obtained following any regression. For instance, you might use Stata commands

```
regress salary female yrdeg startyr if year==95, robust
predict fit
```

You will then have a new variable *fit* that contains the mean values predicted by the model.

We are interested in making inference about the difference in the mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995. In this problem, we focus on the modeling of the continuous variables *yrdeg* and *startyr*.

A) In all parts of this problem, in addition to the year of degree and starting year, you should adjust for the highest degree obtained, field, and administrative duties. What is the best way to model the variables *degree, field,* and *admin*? Briefly justify your answer.

B) In all parts of this problem you should use robust standard error estimates. Briefly explain why inference based on classical linear regression (without robust SE estimates) would be incorrect. Do you think the classical linear regression inference would tend to be conservative or anti-conservative? Justify your answer.

C) Model *yrdeg* and *startyr* as linear continuous variables. Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitC*.

D) Model *yrdeg* and *startyr* as quadratic continuous variables (so linear continuous plus a second order term). Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitD*.

E) Model *yrdeg* and *startyr* as dummy variables for groups defined by earlier than 1966, 1966 – 1975, 1976 – 1985, 1986 – 1995. Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitE*.

F) Model *yrdeg* and *startyr* as splines with knots at years 1965, 1975, and 1985. Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitF*.

G) Model *yrdeg* and *startyr* as dummy variables for groups defined by earlier than 1960, and then each year from 1960 to 1995. Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitG*.

H) What is the difference in the assumed relationship between monthly salary and year of degree in models fit in parts c – g?

I) Examine the agreement between the fitted values derived from the various models in parts c – g. Do you find strong evidence that one or more of the models was superior to the others with respect to its ability to model the relationship between salary and either year of degree or year starting?

J) Examine the agreement between the inference about the adjusted association between monthly salary and sex. Did the inference vary substantially across the various models?

K) In a real situation, how would choose among the models you fit in parts c – g?