Manuscripts for the final presentation of life expectancy project:

To reduce collinearity and reduce dimension, I tried two methods: Partial Least Square and Principle Component Regression.

To balance the interpretability and accuracy, I chose the M equals to 5 in PLS, and 10 in PCR. As we know, PLS and PCR do not have good interpretability. However, they did not give us good test MSEs either. Their test MSEs, which aren’t shown in the slide is around 14~15, the worst among our methods

Next, I tried GAM/Spline model. I picked 10 variables from the Best Subset Selection methods I used. And I fit them one by one for the sake of ANOVA table evaluation. **Here is** the ANOVA table, I found that only model six have less significance, where the Government expenditure on health is added. The result shows that life expectancy is more personal issue, where government can not play a major rule compared to other factors.

I also found some insights from the GAM model. From the first plot, it is very clear that scholars live longer. From the categorical variable Status, we see that developed country people live two years longer than the developing country in average. Also, BMI plot tells that there is a range of healthy BMI – not too low, not too high. The alcohol plot contradict with the current medical studies – the model shows that a little alcohol intake increase life expectancy, happy to know but might due to confounding effect.

This plot, where Diphtheria means the DTP3 vaccine coverage, show in percent, has a positive relationship with life expectancy. And for sure, AIDS is not good for living longer.