STAT111: Mathematical Statistics II **Spring 2024**

Homework 14

David Yang

1. **The countries data set is from the CIA and reports Life Expectancy, Birth Rate, Real GDP per Capita and Education expenditure as a % of GDP. Find a model to predict life expectancy, or some transformation of life expectancy, from some or all of the other variables, or transformations of the other variables (interactions are possible, too). Show evidence that the Normal linear regression assumptions appear to be met. Make a 95% prediction interval for the life expectancy in a country with a birth rate of 12 births per 1000 persons in a year, GDP per capita of $\$70,000$, and with 4% of GDP spent on education (close to the values for the USA).**

   Solution. See attached R Output for specific details.

   In my linear model, I use the birth rate and GDP variables to predict life expectancy (education was left out as it was not significant, and none of the interaction variables with education were significant either). The assumptions for Normal linear regression were satisfied; there appears to be a relatively linear relationship between the variables, we see that the residuals are relatively normal, and satisfy the homoscedasticity and independence conditions.

   The 95% prediction interval I derived for the life expectancy under my model was approximately

   $$\boxed{[68.9, 82.6]}$$

   which seems on par with the data for countries such as the UK and the US. ■

2. **Suppose that the relation of family income to consumption is positive and roughly linear. Of those families in the 90th percentile of income, what proportion would you expect to be at or above the 90th percentile of consumption: (a) exactly 50%, (b) less than 50%, of (c) more than 50%? Justify your answer.**

   Solution. For families in the 90th percentile of income, we expect the proportionto be at or above the 90th percentile of consumption to be $\boxed{\text{(c) more than 50\%.}}$ This follows from the fact that the relation of family income to consumption is positive and roughly linear, so we expect families with higher incomes to have higher levels of consumption.

   ■

3. **The *weighted* regression model assumes**

   $$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \;\; \boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{W}^{-1})$$

   where $\boldsymbol{W}$ is a known, symmetric $n \times n$ positive definite matrix, and $\boldsymbol{X}$ is $n \times p$ and is also known.

(a) **Define $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$. Show that**

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

(b) **Explain why (a) implies that $\hat{\boldsymbol{\beta}}$ is the MLE for $\boldsymbol{\beta}$ but not the least squares estimate.**

(c)