Exercise: **Linear Regression**

Day 3, Part B

1. Load the data set 'data/zmb_mcpa.rds' which contains various maternal and child health indicators measured at the district level in Zambia. (Note: q5 = Under-5 mortality; anc1 = antenatal care coverage; sba = skilled birth attendance coverage; polio = polio vaccine coverage; measles = measles vaccine coverage; dpt3 = DPT3 vaccine coverage; ebf = exclusive breastfeeding coverage; itn = bed net coverage; irs = indoor residual spraying coverage; electricity = prevalence of household electricity; female_edu = mean years of maternal education). Discuss at your table:

   - How many rows and columns are there?
   - What are the classes (variable types) of each column?
   - What is the range of values for the numeric columns?
   - What are the possible values for the factor columns?
   - What does a single row represent?

2. Make a graph (using `ggplot2`) where the `q5` variable is on the y-axis and `year` is the x-axis.

   Discuss at your table:

   - What is the average level of `q5` in 1990? 2010?
   - What is the general trend of `q5` ver time? Is it increasing or decreasing?
   - Approximately how much higher or lower is `q5` each year?
   - Does the relationship between these two variables appear to fit each of the four assumptions I listed during the lecture (slide 31)? (Hint: the answer is "no", but we're going to do it anyway)

3. Fit the model, $q5 = \beta_0 + \beta_1 \cdot year + \epsilon$

   Discuss at your table

   - What is the interpretation of the intercept term in this model?
   - What is the interpretation of the coefficient on `year`? Is it similar to what you estimated "by hand" in question 2?
   - How can you use these two coefficients to estimate the expected value of `q5` in 1990? 2010?
   - Why does the model estimate a different number for 1990 than the average you estimated for 1990's data alone?
   - Is slope term in this model statistically significant? How do you interpret that p-value in lay terms?

4. Using the model from question 3:

   a. Create new columns in the `zmb` data frame for the fitted values, confidence intervals, and residuals from this model.

   b. Discuss at your table:

      - What is the fitted value estimate for 1990? Is it the same for every district?
      - What are the upper and lower bounds of the confidence interval for 1990? How do you interpret these numbers?
      - What is the average residual across the entire dataset? Why is it that number (or so close to that number)?

   c. Make a density plot of the residuals (hint: `geom_density()`).

   d. Discuss at your table:

      - Do these residuals appear to be normally-distributed with mean zero? Is there any skew to this distribution?

e. Make a scatter plot of with fitted values on the y-axis and observed `q5` on the x-axis. Use a separate panel for each province (hint: `facet_wrap()`) and color the points by year. Add an equivalence line (hint: `geom_abline()`) that shows y=x.

f. Discuss at your table:

   - What is the interpretation of this graph? Does the model consistently over-estimate `q5` in certain provinces and under-estimate in others? Certain years?

5. Fit the model, $q5 = \beta_0 + \beta_1 \cdot year + \beta_{2_p} \cdot province + \epsilon$ (Note: the notation β_{2_p} is often used to indicate that β_2 is not just one number, but actually a vector of coefficients, one per province (indexed by p))

   Discuss at your table:

   - What are the interpretation of the coefficients in this model?
   - Which province is the "reference" category (i.e. the one not displayed)?
   - Which provinces are significantly higher than the Central province? Which ones are lower? Which ones are not-significantly higher or lower?

6. Using this new model, re-estimate the fitted values, confidence intervals and residuals as columns in the data frame. Re-make the scatter plot of fitted values vs observed values faceted by province.

   Discuss at your table:

   - How does this graph compare to the previous version you made? Do the fitted values line up with the observed values better by province? Why?

7. Fit the model, $q5 = \beta_0 + \beta_1 \cdot year + \beta_{2_p} \cdot province + \beta_{3_p} \cdot province \cdot year + \epsilon$

   Discuss at your table

   - What are the interpretation of the new coefficients in this model?
   - Which provinces have a steeper negative slope than the Central province? Are any of them significant?
   - Which provinces have a less-steep slope than the Central province? Are any of them significant?

8. Explore the relationship between the variables `electricity` (proportion of households with electricity) and `female_edu` (educational attainment in years among women):

   a. Graph these two variables with `q5` on the y-axis

   b. Discuss at your table:

      - Does this appear to be a linear relationship?
      - What could you do to more it more linear?

   c. Fit the model, $logit(electricity) = \beta_0 + \beta_1 \cdot female_edu + \epsilon$ (Hint: the package `boot` contains the `logit` function. It's a transformation of the form $log(p/(1-p))$, where p is a proportion between 0 and 1. It's often useful to make fractions more normally-distributed)

   d. Discuss at your table:

      - What is the interpretation of these coefficients?

9. Estimate fitted values from the model in question 8 among a **new dataset**.

   a. First, create a new data frame called "prediction_data". This should have only one column in it called "female_edu", which ranges from `min(zmb$female_edu)` to `max(zmb$female_edu)` in increments of one.

   b. Second, create a second column in "prediction_data" that contains fitted values for these levels of "female_edu". (hint: use the `predict()` function)

   c. Third, make a third column that is the inverse logit of the fitted values, to get them back out of "logit space" (hint: use the `inv.logit()` function)

d. Finally, make a graph of `electricity` vs `female_edu`, including the exponentiated fitted values as a line (hint: you will have to use `aes()` twice)

e. Discuss at your table:

- What is the interpretation of this figure?
- How does this best-fit line compare to linear regression without logit transformation?
- What happens if you extend the "female_edu" variable to 20 in "predction_data"?

Bonus:

10. Use the model from question 7 to forecast `q5` to the year 2050. (Hint: you will need to create a prediction data frame like in question 9, but this time it will need two variables and all possible combinations. Check out the `expand.grid` function for an easy way to do this)

    Discuss at your table:

    - Do these values still seem reasonable?
    - What could you do to constrain the values to be positive? (hint: 5q0 is a proportion and the `q5` variable has been multiplied by 1000)