Exercise: **Linear Regression**

Day 3, Part B

```
> library(ggplot2)
```

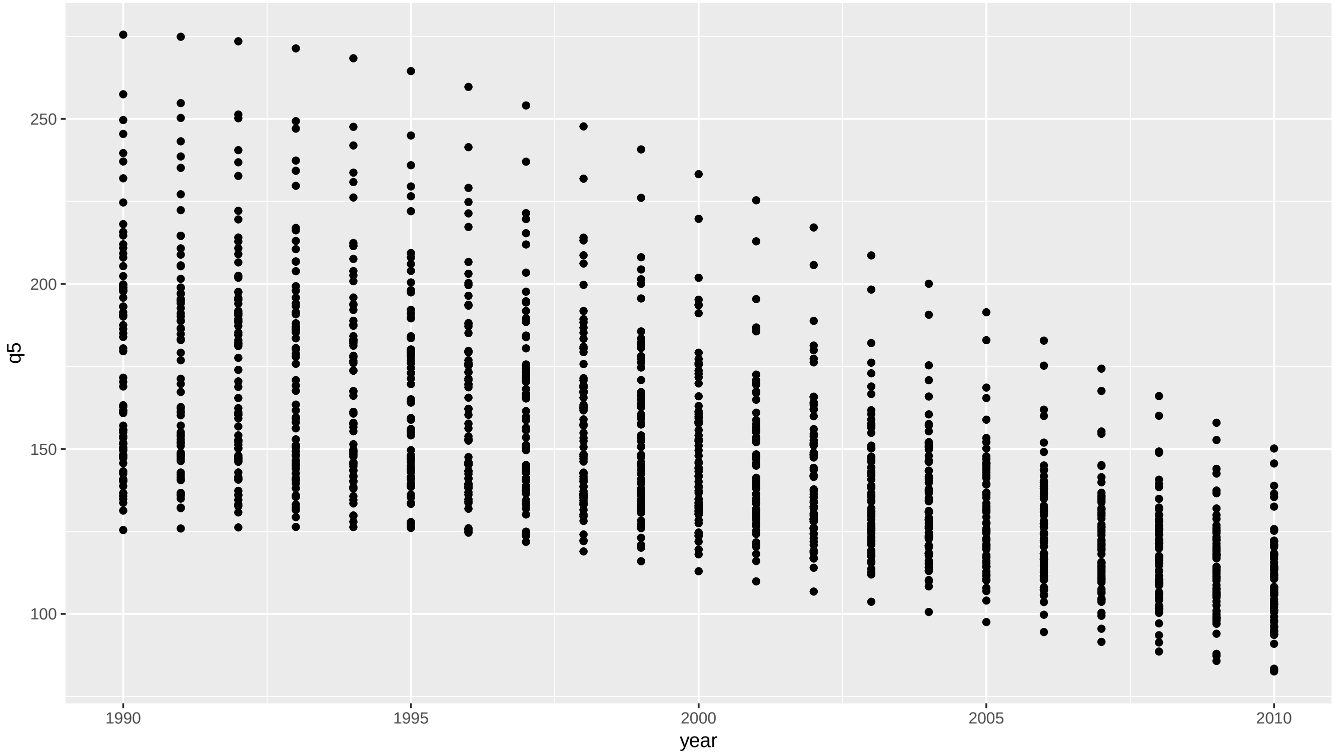
1. Load the data set ‘data/zmb_mcpa.rds’ which contains various maternal and child health indicators measured at the district level in Zambia. (Note: q5 = Under-5 mortality; anc1 = antenatal care coverage; sba = skilled birth attendance coverage; polio = polio vaccine coverage; measles = measles vaccine 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?

    ```
    > main_dir <- "C:/Users/ngraetz/Documents/repos/r_training_penn/"  # CHANGE TO YOUR LOCAL COPY O
    > zmb <- readRDS(paste0(main_dir, "data/zmb_mcpa.rds"))
    ```

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

    ```
    > ggplot(zmb, aes(y = q5, x = year)) + geom_point()
    ```



```
> mean(zmb[zmb$year == 1990, ]$q5)
[1] 178.8833
```

```
> mean(zmb[zmb$year == 2010, ]$q5)
[1] 109.6359
```

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, $q5 = \beta_0 + \beta_1 \cdot year + \epsilon$

```
> mod_a <- lm(q5 ~ year, data = zmb)
> summary(mod_a)
Call:
lm(formula = q5 ~ year, data = zmb)

Residuals:
    Min      1Q  Median      3Q     Max 
-59.817 -17.632  -3.192  12.482  97.730 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7429.1125   214.4251   34.65   <2e-16 ***
year          -3.6401     0.1072  -33.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.24 on 1510 degrees of freedom
Multiple R-squared:  0.4329,	Adjusted R-squared:  0.4326 
F-statistic:  1153 on 1 and 1510 DF,  p-value: < 2.2e-16
```

```
> coef(mod_a)[1] + (1990 * coef(mod_a)[2])
(Intercept) 
   185.2201 
```

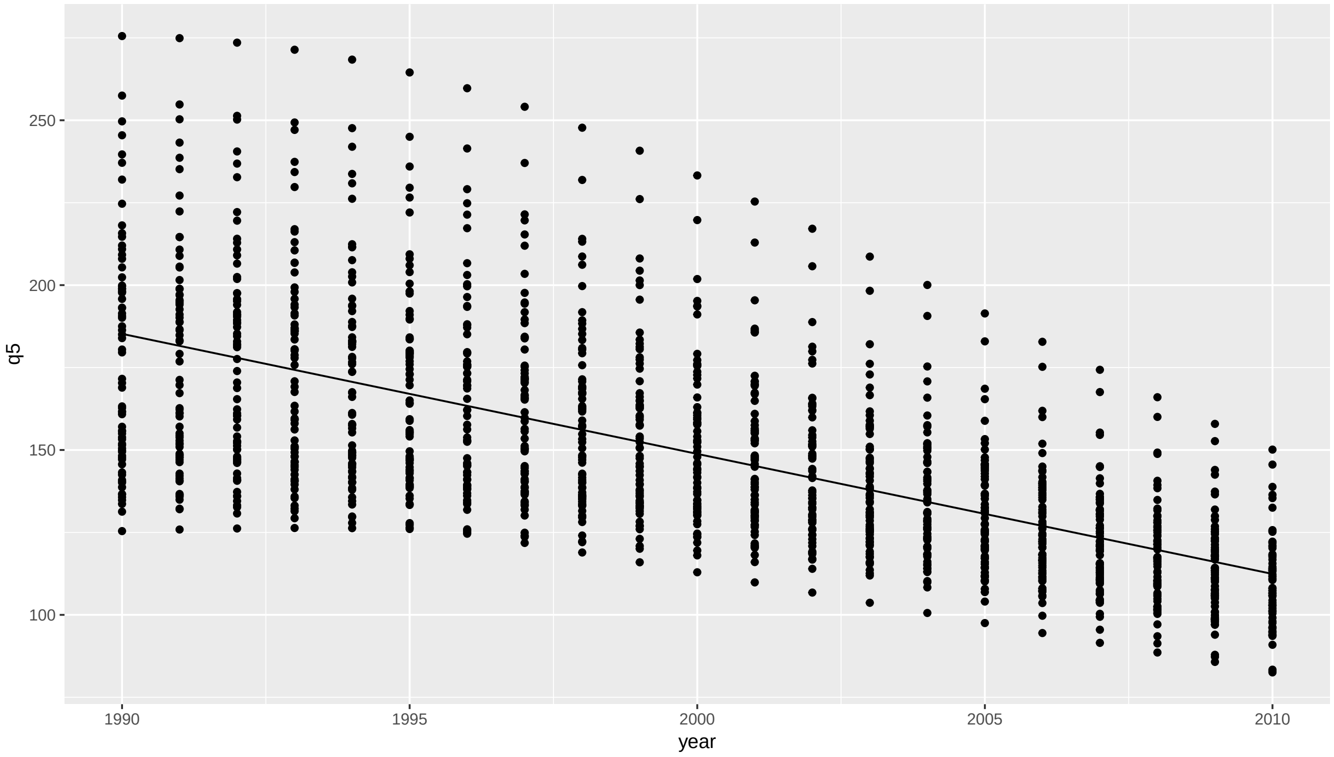
Discuss at your table

- 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.

```
> preds <- predict(mod_a, interval="confidence")
> zmb <- cbind(zmb, preds)
> zmb$resid <- mod_a$residuals
> ggplot(zmb, aes(y=q5, x=year)) +
+   geom_point() +
+   geom_line(aes(y=fit))
```

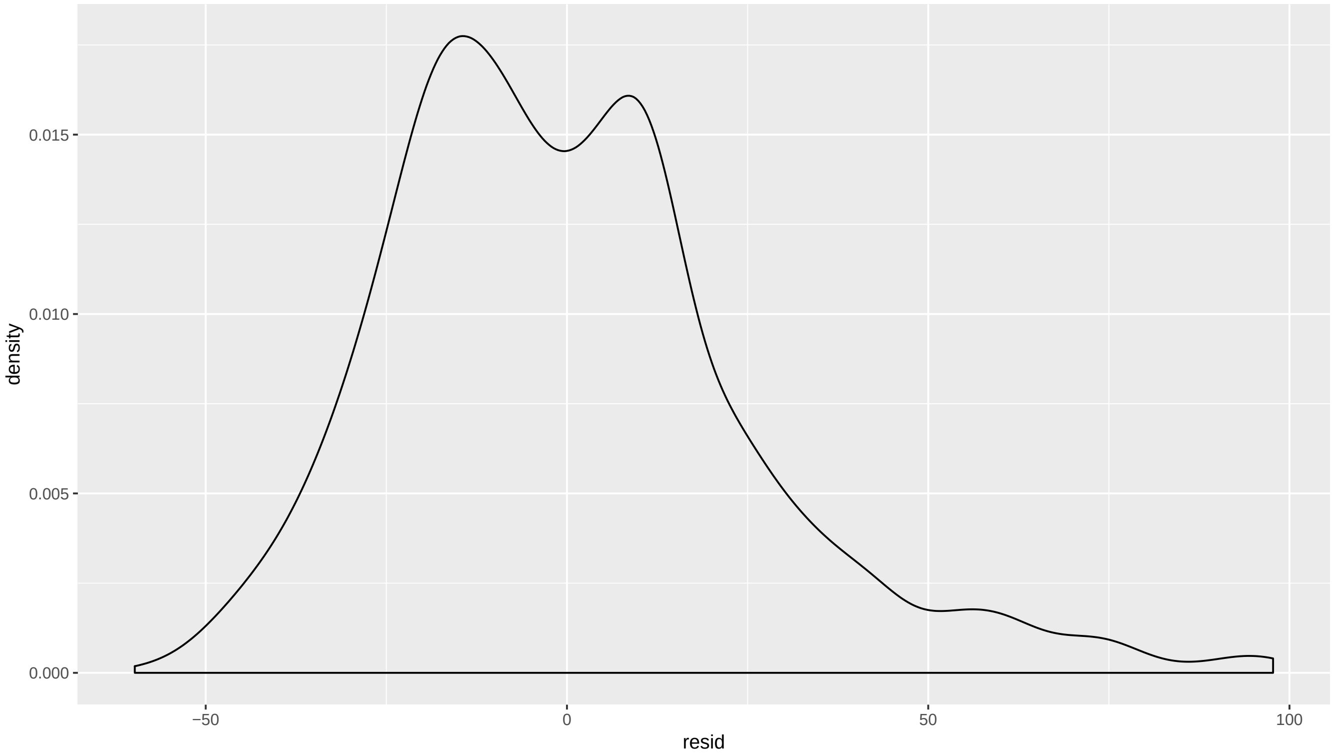


b. Discuss at your table:

- 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()`).

```
> ggplot(zmb, aes(x=resid)) +
+   geom_density()
```

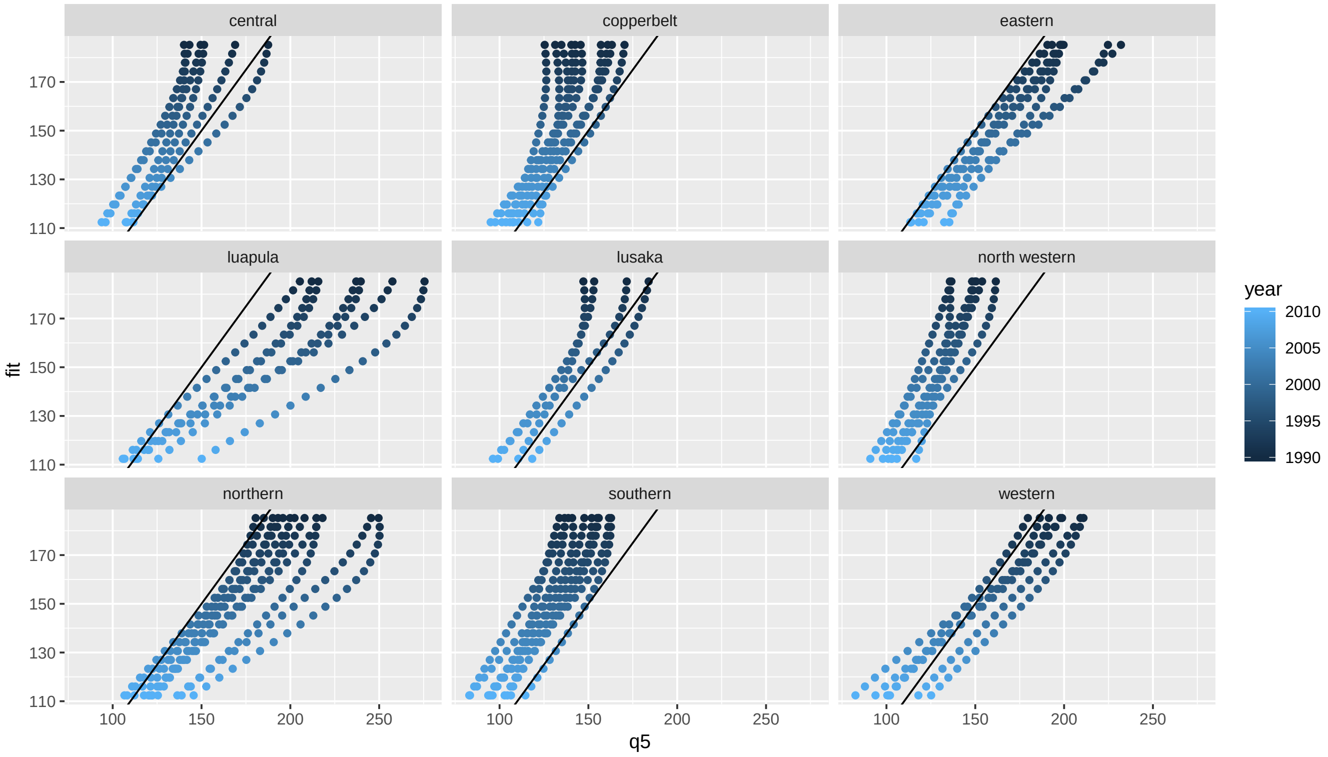


d. Discuss at your table:

   - 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.

```
> ggplot(zmb, aes(x=q5, y=fit, color=year)) +
+   geom_point() +
+   geom_abline(intercept = 0, slope = 1) +
+   facet_wrap(~ province)
```



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))

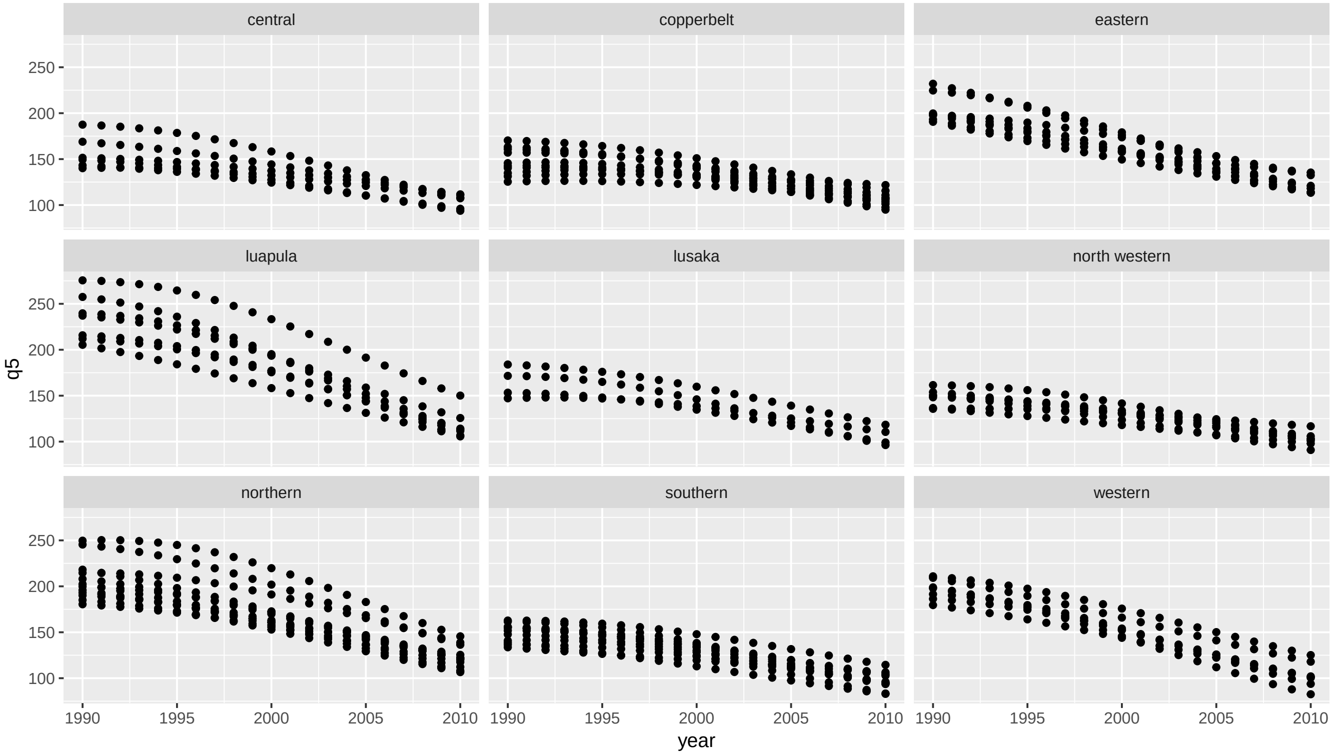
```
> mod_b <- lm(q5 ~ year + factor(province), data = zmb)
> summary(mod_b)
Call:
lm(formula = q5 ~ year + factor(province), data = zmb)

Residuals:
    Min      1Q  Median      3Q     Max
-43.381  -9.659  -0.756   8.714  61.885

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   7415.17381  132.47916  55.972  < 2e-16 ***
year                            -3.64015    0.06624 -54.957  < 2e-16 ***
factor(province)copperbelt      -2.49752    1.75744  -1.421 0.155491
factor(province)eastern         28.71061    1.83797  15.621  < 2e-16 ***
factor(province)luapula         49.78364    1.89340  26.293  < 2e-16 ***
factor(province)lusaka           6.28063    2.19680   2.859 0.004309 **
factor(province)north western   -6.63945    1.89340  -3.507 0.000467 ***
factor(province)northern        35.22938    1.70163  20.703  < 2e-16 ***
factor(province)southern        -7.42970    1.72722  -4.302 1.81e-05 ***
factor(province)western         18.67413    1.89340   9.863  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 15.6 on 1502 degrees of freedom
Multiple R-squared:  0.7847,    Adjusted R-squared:  0.7834
F-statistic: 608.3 on 9 and 1502 DF,  p-value: < 2.2e-16
```

```
> ggplot(zmb, aes(y = q5, x = year)) + geom_point() + facet_wrap(~province)
```

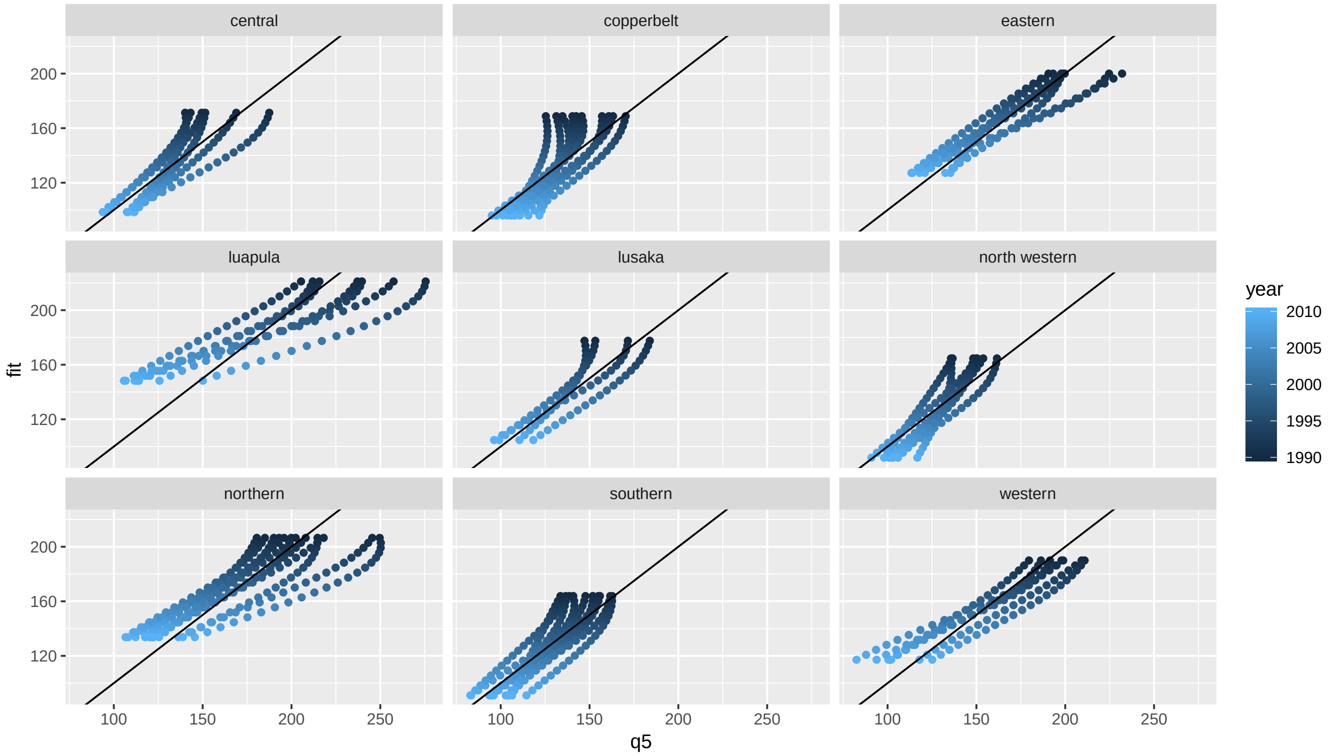


Discuss at your table:

- 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.

```
> preds <- predict(mod_b, interval="confidence")
> zmb <- zmb[, 1:14]
> zmb <- cbind(zmb, preds)
> zmb$resid <- mod_b$residuals
> ggplot(zmb, aes(x=q5, y=fit, color=year)) +
+   geom_point() +
+   geom_abline(intercept = 0, slope = 1) +
+   facet_wrap(~ 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, $q5 = \beta_0 + \beta_1 \cdot year + \beta_{2_p} \cdot province + \beta_{3_p} \cdot province \cdot year + \epsilon$

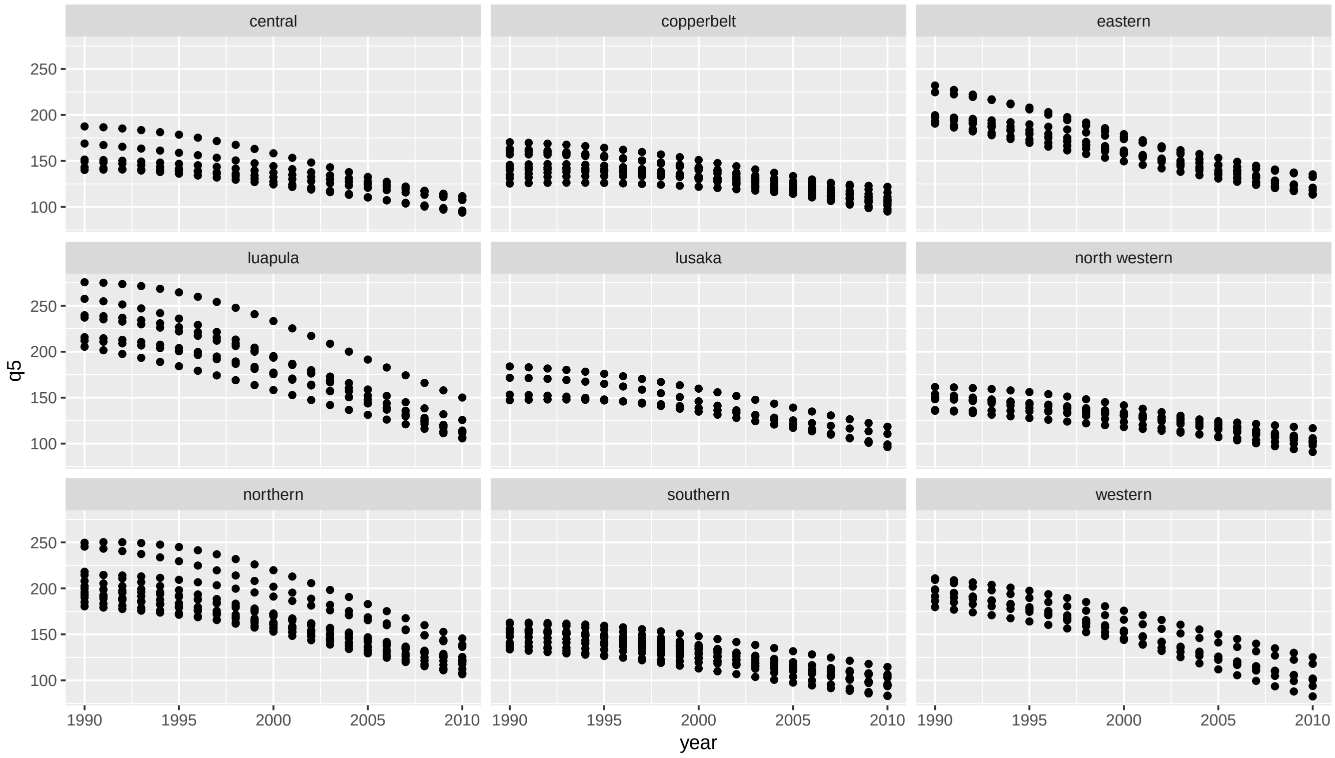
```
> mod_c <- lm(q5 ~ factor(province) * year, data = zmb)
> summary(mod_c)
Call:
lm(formula = q5 ~ factor(province) * year, data = zmb)

Residuals:
    Min      1Q  Median      3Q     Max
-41.309  -8.912  -1.593   7.097  53.756

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         5.617e+03  4.006e+02  14.022  < 2e-16 ***
factor(province)copperbelt         -1.147e+03  5.067e+02  -2.264   0.0237 *
factor(province)eastern             3.186e+03  5.299e+02   6.013 2.29e-09 ***
factor(province)luapula             6.973e+03  5.459e+02  12.774  < 2e-16 ***
factor(province)lusaka              7.016e+02  6.334e+02   1.108   0.2682
factor(province)north western      -7.570e+02  5.459e+02  -1.387   0.1658
factor(province)northern            3.343e+03  4.906e+02   6.814 1.37e-11 ***
factor(province)southern            9.932e+00  4.980e+02   0.020   0.9841
factor(province)western             4.272e+03  5.459e+02   7.826 9.50e-15 ***
year                               -2.741e+00  2.003e-01 -13.686  < 2e-16 ***
factor(province)copperbelt:year     5.723e-01  2.533e-01   2.259   0.0240 *
factor(province)eastern:year       -1.579e+00  2.650e-01  -5.959 3.17e-09 ***
factor(province)luapula:year       -3.462e+00  2.729e-01 -12.683  < 2e-16 ***
factor(province)lusaka:year        -3.477e-01  3.167e-01  -1.098   0.2725
factor(province)north western:year  3.752e-01  2.729e-01   1.374   0.1695
factor(province)northern:year      -1.654e+00  2.453e-01  -6.743 2.22e-11 ***
factor(province)southern:year      -8.681e-03  2.490e-01  -0.035   0.9722
factor(province)western:year       -2.127e+00  2.729e-01  -7.791 1.23e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.61 on 1494 degrees of freedom
Multiple R-squared:  0.8368,    Adjusted R-squared:  0.835
F-statistic: 450.7 on 17 and 1494 DF,  p-value: < 2.2e-16
```

```
> ggplot(zmb, aes(y = q5, x = year)) + geom_point() + facet_wrap(~province)
```



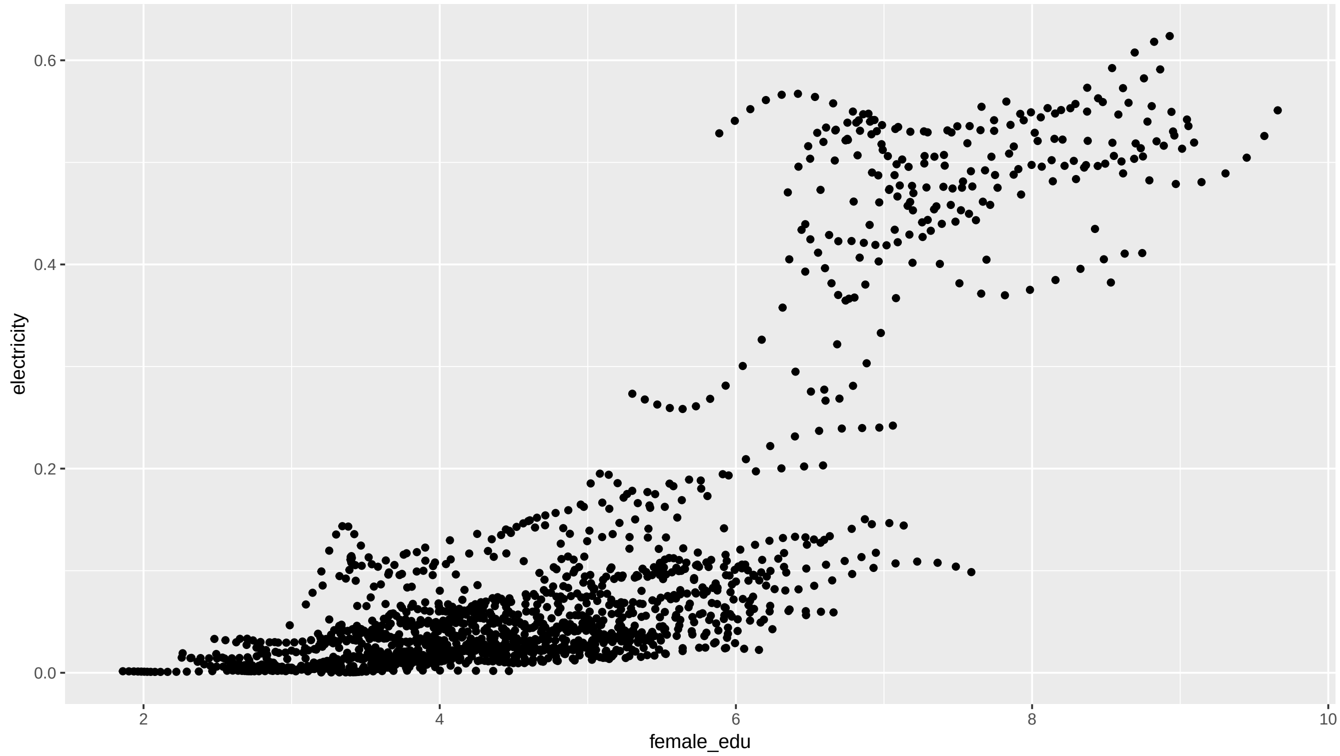
Discuss at your table

- 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

```
> ggplot(zmb, aes(y = electricity, x = female_edu)) + geom_point()
```



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)

```
> library(boot)
> mod_d <- lm(logit(electricity) ~ female_edu, data = zmb)
> summary(mod_d)
Call:
lm(formula = logit(electricity) ~ female_edu, data = zmb)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5289 -0.6387 -0.0003  0.6860  2.4789 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.22209    0.08122  -88.92   <2e-16 ***
female_edu   0.88485    0.01595   55.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9319 on 1510 degrees of freedom
Multiple R-squared:  0.6709,	Adjusted R-squared:  0.6707 
F-statistic:  3079 on 1 and 1510 DF,  p-value: < 2.2e-16
```

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.

   ```
   > prediction_data = data.frame(female_edu = seq(max(zmb$female_edu)))
   ```

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

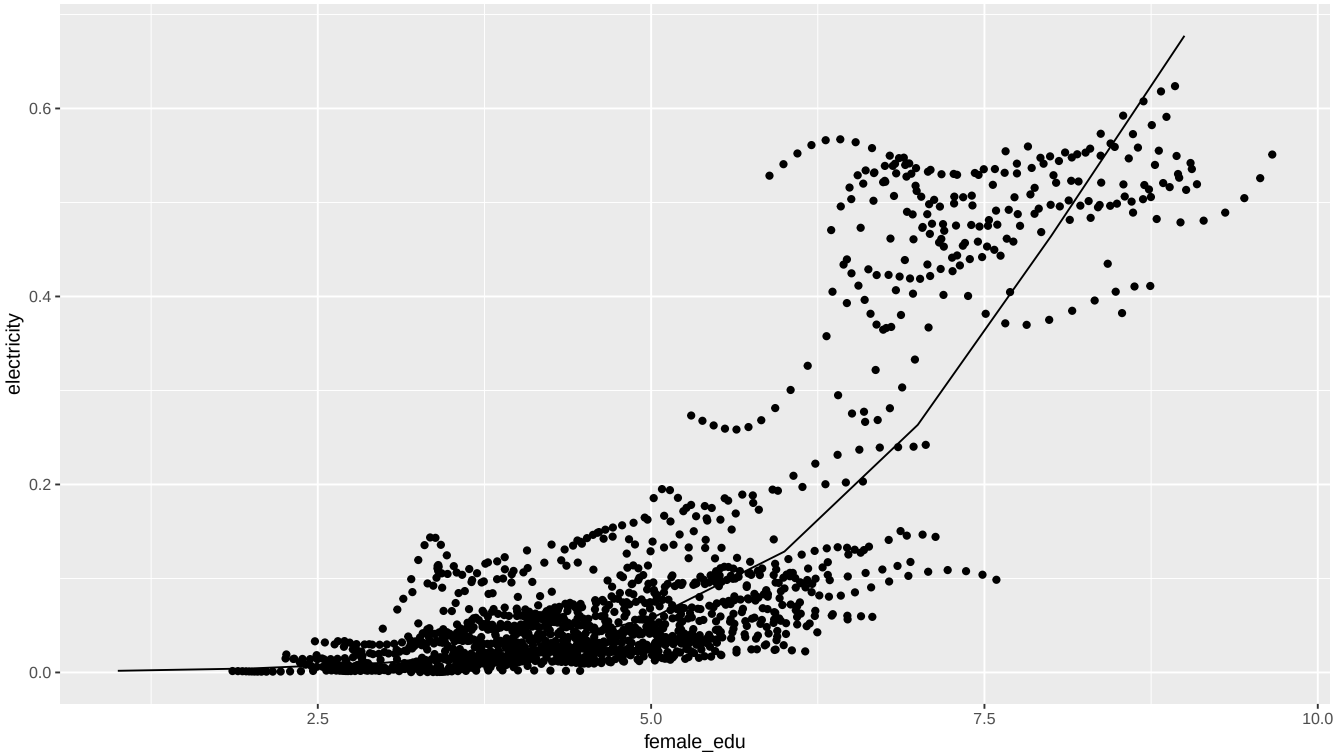
   ```
   > prediction_data$preds <- predict(mod_d, newdata = prediction_data)
   ```

   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)

   ```
   > prediction_data$preds_natural <- inv.logit(prediction_data$preds)
   ```

   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)

   ```
   > ggplot(data = zmb, aes(y = electricity, x = female_edu)) + geom_point() + geom_line(data =
   +     aes(y = preds_natural))
   ```



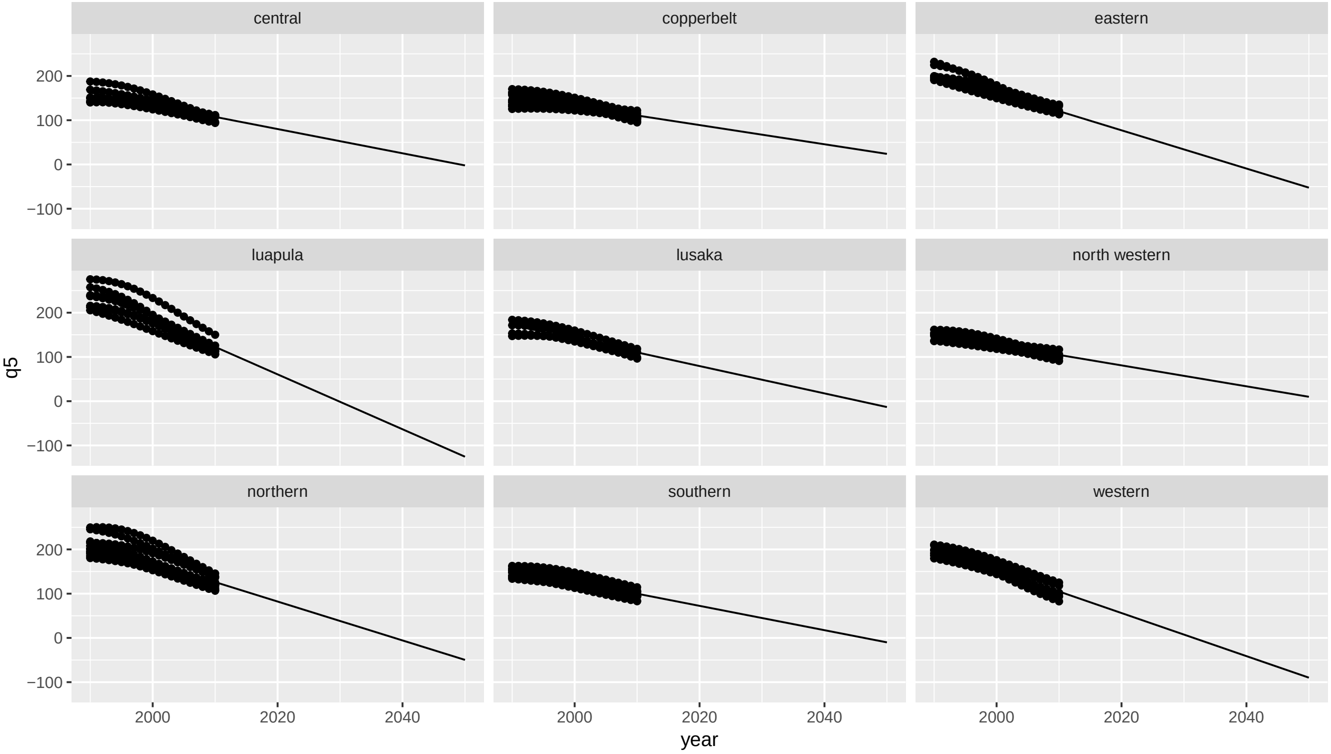
   e. Discuss at your table:

   - 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)

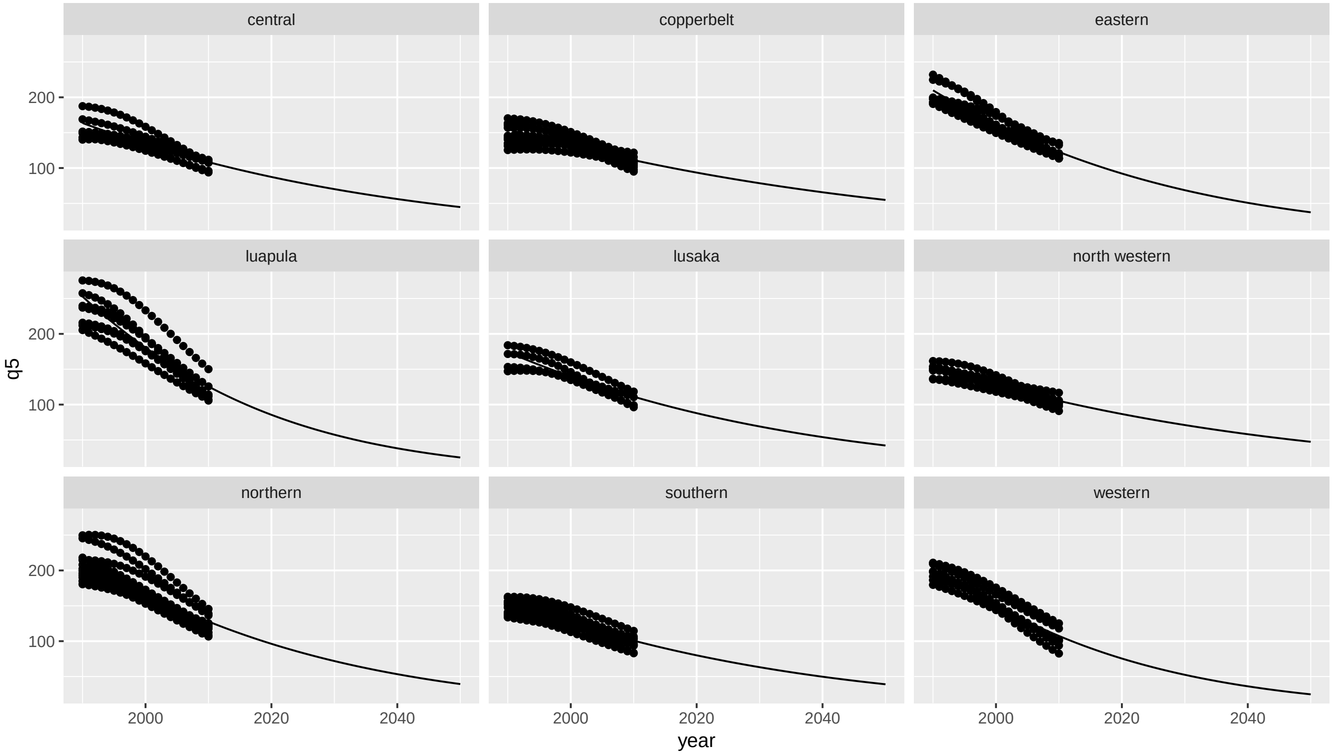
```
> prediction_data = data.frame(expand.grid(province = unique(zmb$province), year = seq(1990,
+     2050)))
> prediction_data$preds <- predict(mod_c, newdata = prediction_data)
> ggplot(data = zmb, aes(y = q5, x = year)) + geom_point() + geom_line(data = prediction_data,
+     aes(y = preds)) + facet_wrap(~province)
```



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)

```
> mod_e <- lm(logit(q5/1000) ~ factor(province) * year, data = zmb)
> prediction_data = data.frame(expand.grid(province = unique(zmb$province), year = seq(1990,
+     2050)))
> prediction_data$preds <- predict(mod_e, newdata = prediction_data)
> prediction_data$preds_natural <- inv.logit(prediction_data$preds)
> ggplot(data = zmb, aes(y = q5, x = year)) + geom_point() + geom_line(data = prediction_data,
+     aes(y = preds_natural * 1000)) + facet_wrap(~province)
```


central
copperbelt
eastern
luapula
lusaka
north western
northern
southern
western
q5
200
100
2000
2020
2040
year