Lecture 2.4 - Model building sample solution

Student

2025-11-07

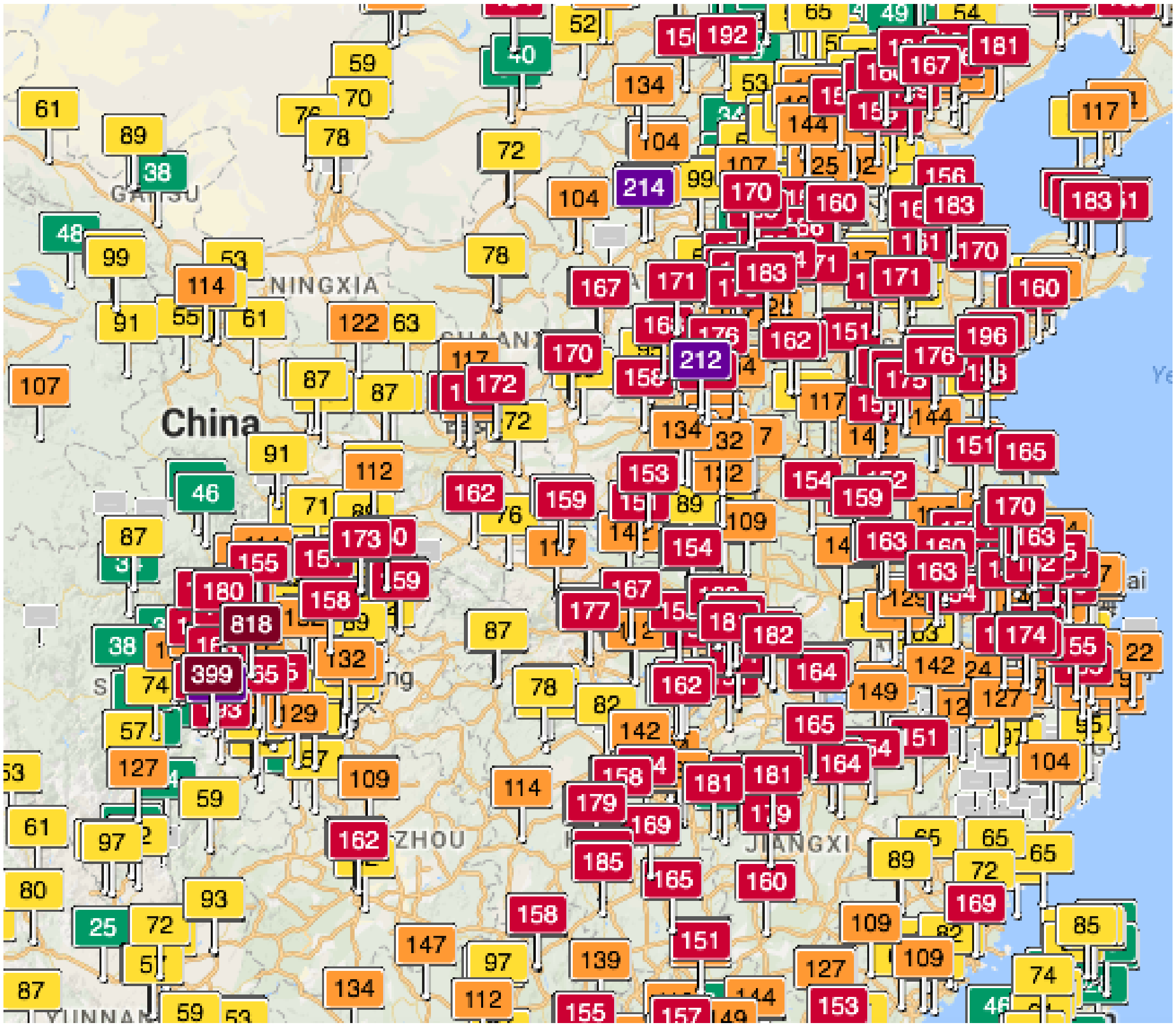


Table 1: Distribution of variable: coastal

Coastal	Count	Proportion
0	243	0.75
1	80	0.25

Interpretation steps

Hypothesis development

Make a list of hypothesized relationships to AQI. Pick four variables that you think are most likely to be associated with AQI. For each variable, list what you expect its relationship to AQI to be and how strong you expect the relationship to be.

- Precipitation: I expect this will have a moderate and negative relationship with AQI
- Incineration amount: I expect this to have a moderate and positive relationship with AQI
- Coastal: I expect this to have a moderate and positive relationship with AQI
- Altitude: I expect this to maybe have a slightly negative relationship with AQI

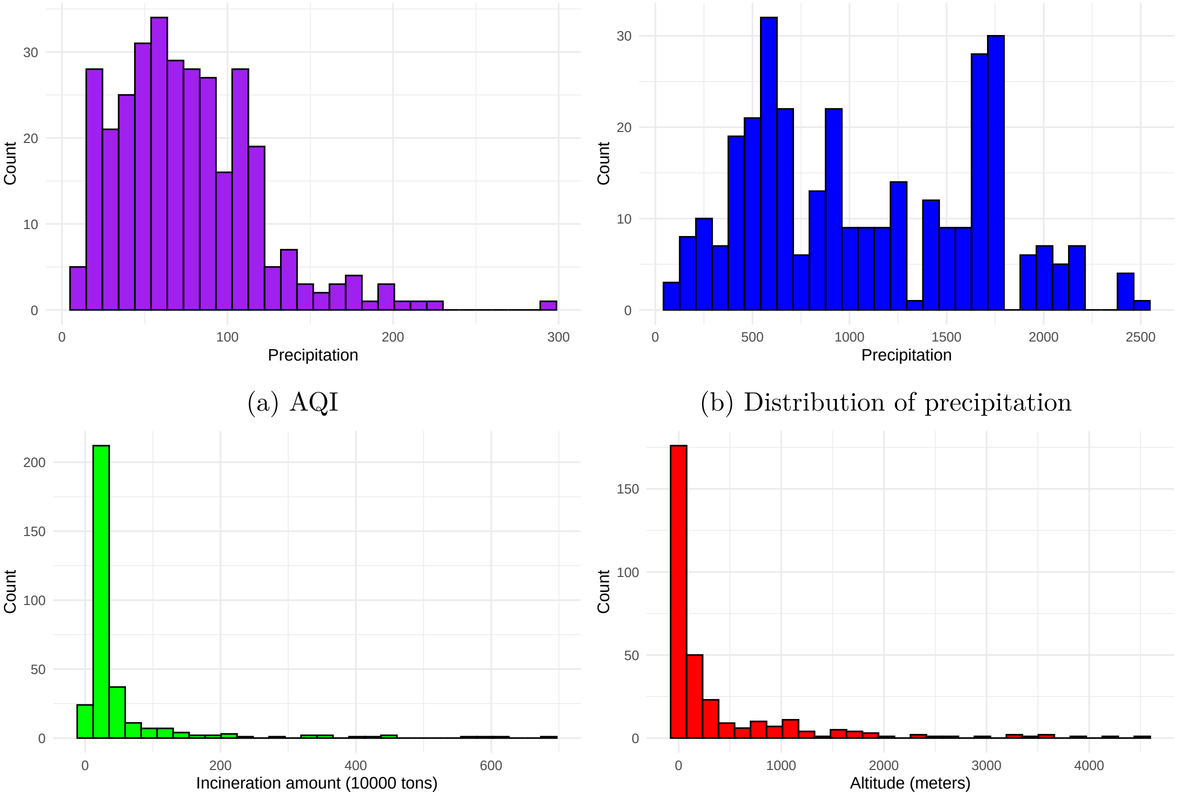
After you have done that, write down the order in which you expect predictor variables to best predict AQI – which do you think are the most important in 'causing' an increase in AQI?

1. Coastal
2. Precipitation
3. Incineration amount
4. Altitude

Variable exploration

For this part of the lab, you should explore the distribution of the four variables via histograms. Make note of any outliers or non-normal distributions that may cause problems for your later statistical test. Also consider if any variables need to be recoded or transformed.

> AQI is slightly right skewed but mostly normally distributed. Probably no transformation is necessary.
>
> Both incineration amount and altitude look highly right skewed, so it is sensible to log transform them.
>
> We can also see that only 25% of the counties are coastal.



(a) AQI

(b) Distribution of precipitation

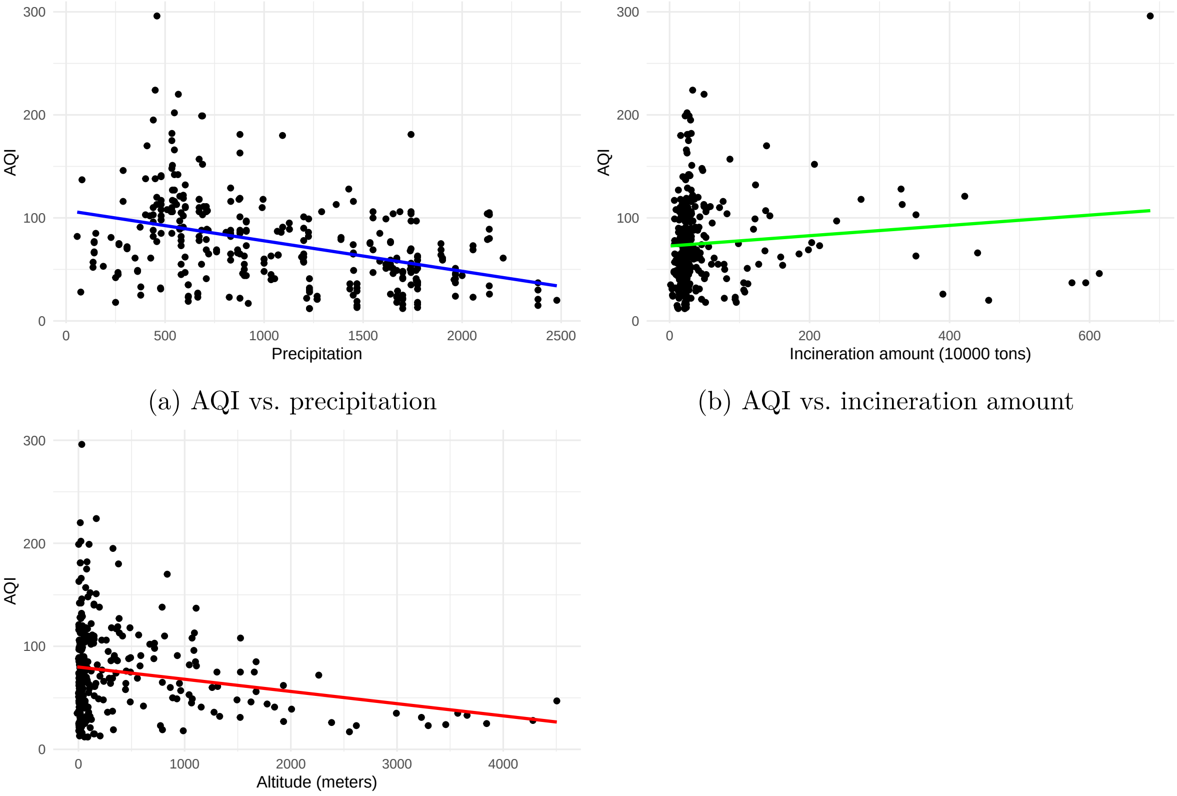
(c) Distribution of incineration amount

(d) Distribution of altitude

Figure 1: Histograms of key variables

Two variable relationship

Based on your hypothesized relationship between each of the predictor variables and AQI, check the two-way (meaning relationships between each individual predictor variable and the response variable) to see your expectations are met or not. You may also want to check the correlations between all variables. The command to create a correlation matrix plot is



(a) AQI vs. precipitation

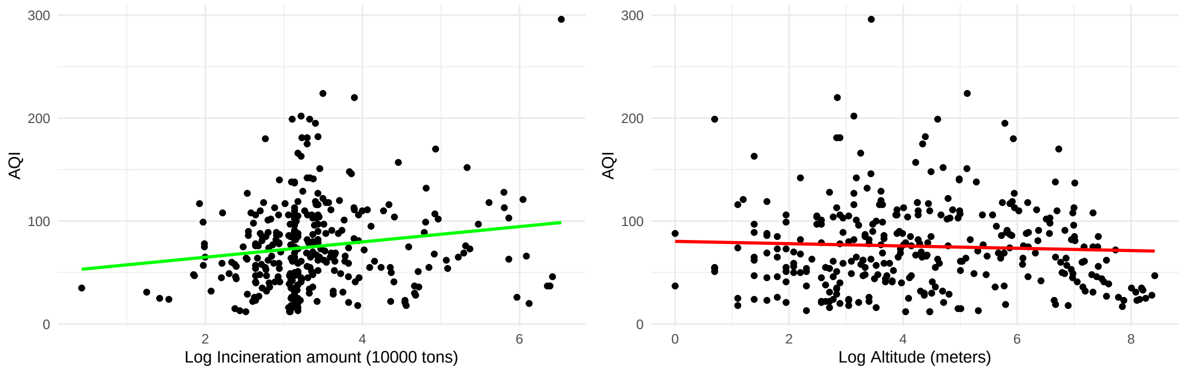
(b) AQI vs. incineration amount

(c) AQI vs. altitude

Figure 2: Two way relationships between AQI and key variables

As we can see, incineration and altitude have very non-linear relationships with AQI. Taking the log improves the relationships though still weak. Note that there are a significant number of zero values the two variables, so a very small amount is added to the values to make them non-zero.

```
china.aqi <- china.aqi %>%
  mutate(adj.Inc.Amount.10.000ton. = Incineration.Amount.10.000ton. + 0.0000001,
         adj.Altitude = Altitude + 0.00000001) %>%
  mutate(Log.Incineration.Amount.10.000ton. = log(adj.Inc.Amount.10.000ton.),
         Log.Altitude = log(adj.Altitude))
```



(a) AQI vs. log incineration amount

(b) AQI vs. log altitude

Figure 3: Two way relationships between AQI and key transformed variables

> Now we can view the correlation data

Did any of the two-way relationships surprise you? Which ones and why?

> As expected, the correlations are in the directions predicted though weaker than expected. Precipitation is the strongest, which is what was expected. The variables do have some correlation between them which may be a problem.

Model building

Now, armed with your hypotheses, work with your lab partner to create the best regression model possible by trying combinations of the most 'important' variables until you arrive at a model you are happy with.

> **Step 1**: Model basic results. As predicted from the correlations, the coefficients on the model are all in the expected direction. Though we will need to consider the units to check how meaningful they are.
>
> **Step 2**: Analyze residuals. We can see the plot thickens condition in the scatterplot and a possible outlier. The histogram has some skew. So we may need to go back and see if there is anything we can revise. While Beijing is an outlier (its actual value is much higher than its predicted value), I don't think there are grounds to remove it.

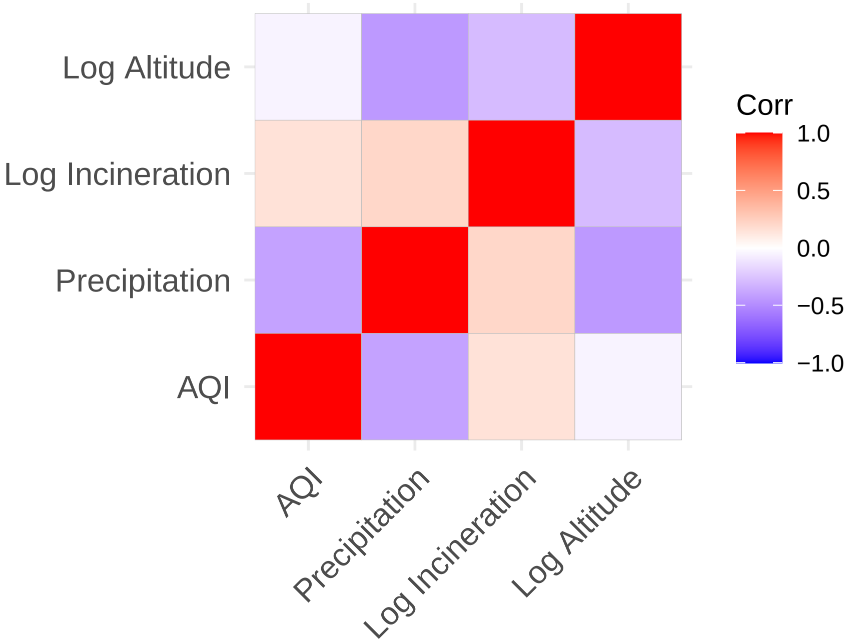


Figure 4: Correlation of key variables

Table 2: Regression model on the response variable: AQI

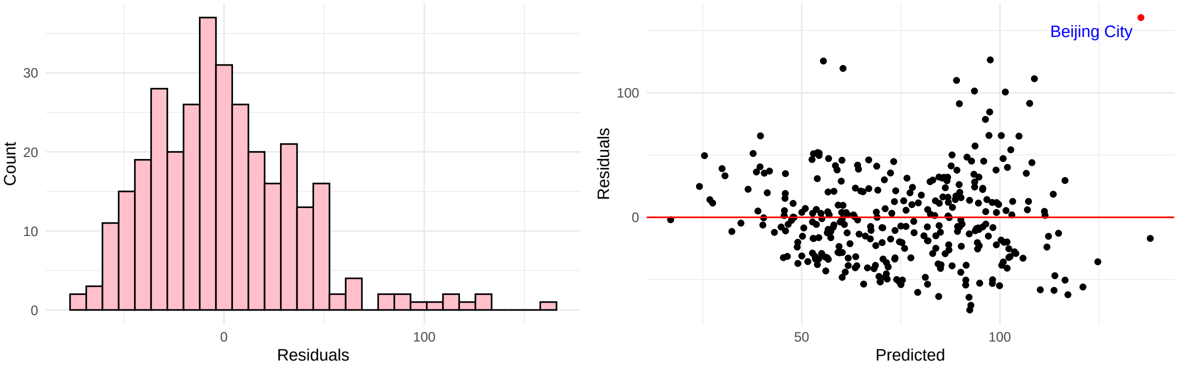
Variable	Estimate	t value	p-value
Intercept	108	8.78	<0.001
Precipitation	-0.04	-10.1	<0.001
Log Incineration	9.7	3.99	<0.001
Log Altitude	-5.2	-4.21	<0.001

$R^2 = 0.260$

Number of cases (n): 322

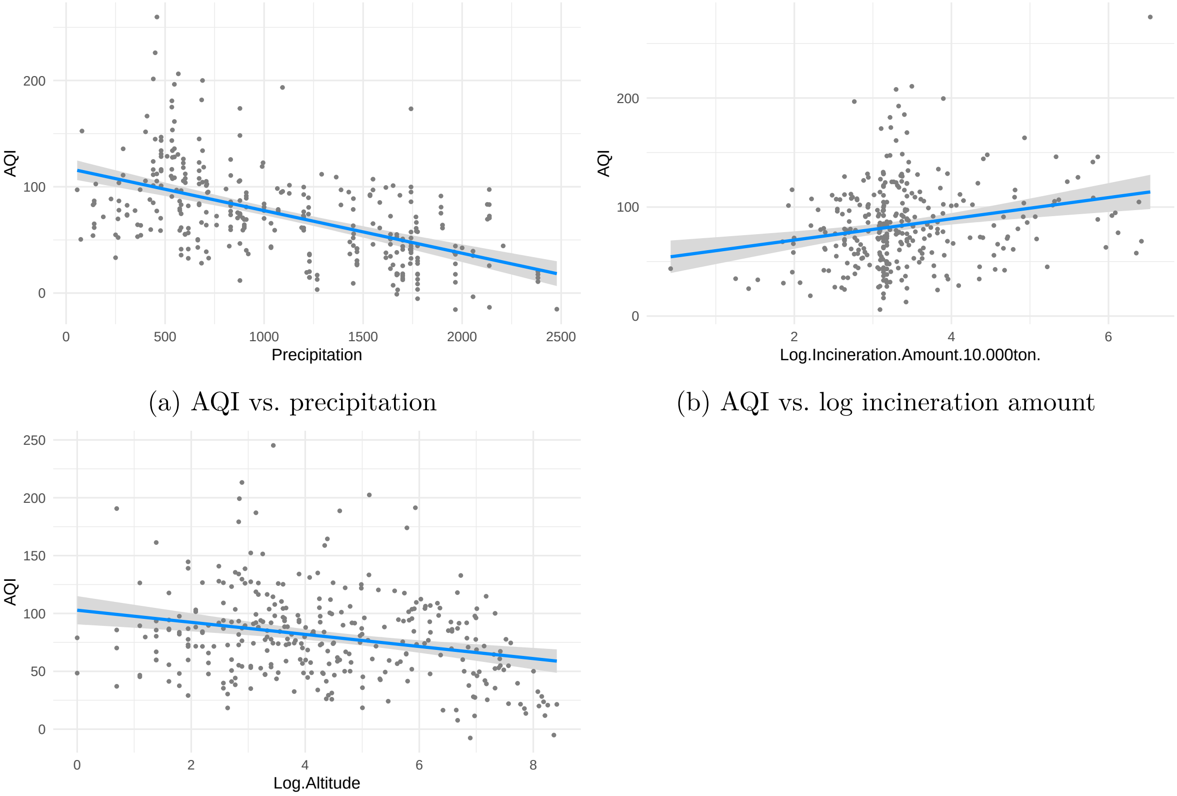
Residuals: Min = -74.39, Q1 = -27.3, Median = -4.19, Q3 = 21.04, Max = 160.42

Residual standard error: 36.99



(a) Histogram of the residuals (b) Residuals vs. predicted

Figure 5: Residual plots of the model



(a) AQI vs. precipitation (b) AQI vs. log incineration amount

(c) AQI vs. log altitude

Figure 6: Partial residual plots

> **Step 3**: Review partial residual plots. The partial residual plots here indicate that the relationship between the variables is about as linear as can be expected, although the only variable that shows a strong relationship is precipitation. I don't see any additional obvious transformation that would improve things.

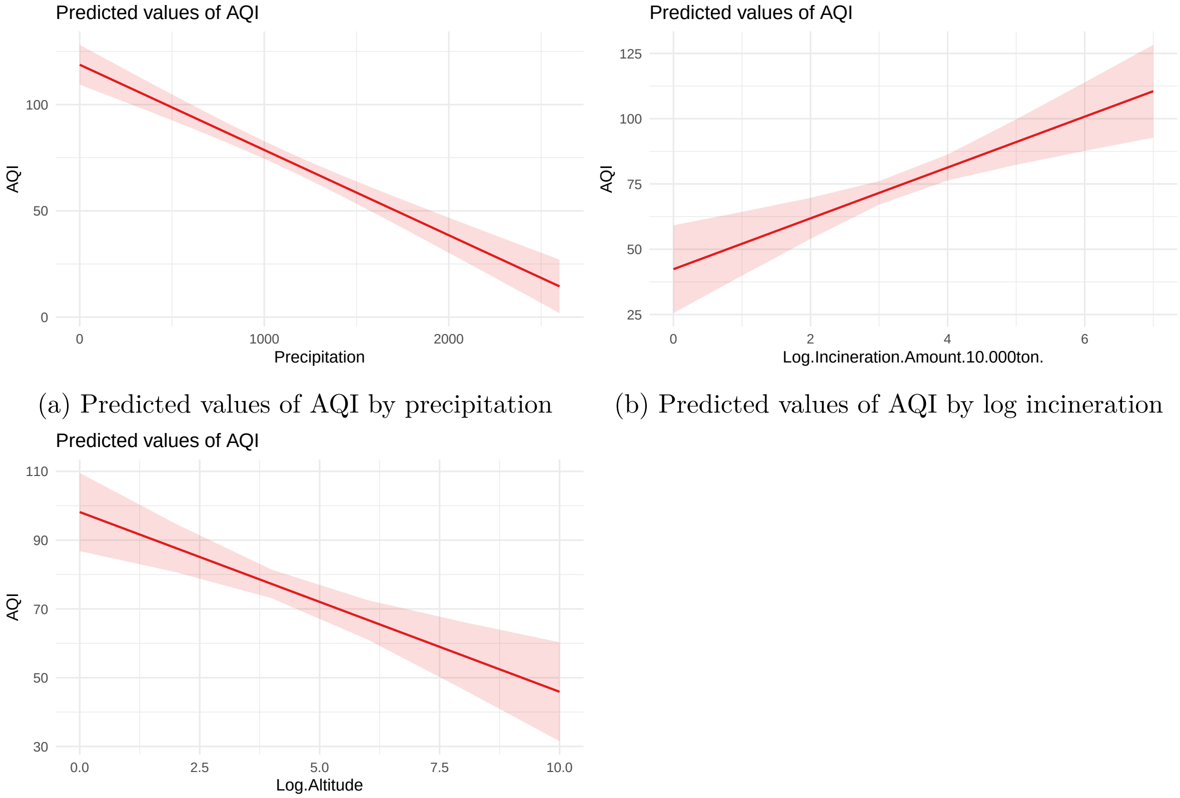
Model interpretation

Next, interpret your coefficients – think about whether your predictor variables are predicted to make a big or a small change in the response variable.

- Some things to consider:
  - What are the units of the slope coefficients?
  - How big of an impact does a one unit change in the predictor variable have on the response variable?
  - Is that change a little or a lot?

Then, evaluate your hypotheses with respect to the outcome of your model. Were you surprised by any of the results? Why or why not?

> **Step 4**: Interpret coefficients. The margin plots here indicate that these variables all have some influence on AQI, with precipitation having perhaps the largest influence. The influence of these variables on AQI is modest, however, which is to be expected given the model's R^2 is only 0.26.



(a) Predicted values of AQI by precipitation

(b) Predicted values of AQI by log incineration

(c) Predicted values of AQI by log altitude

Figure 7: Maginal effect plots