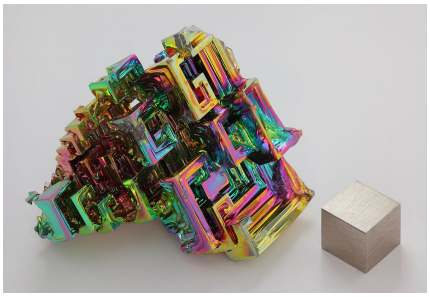
Stat 242 Quiz – Topics Drawn from Chapter 8

What's Your Name? _____

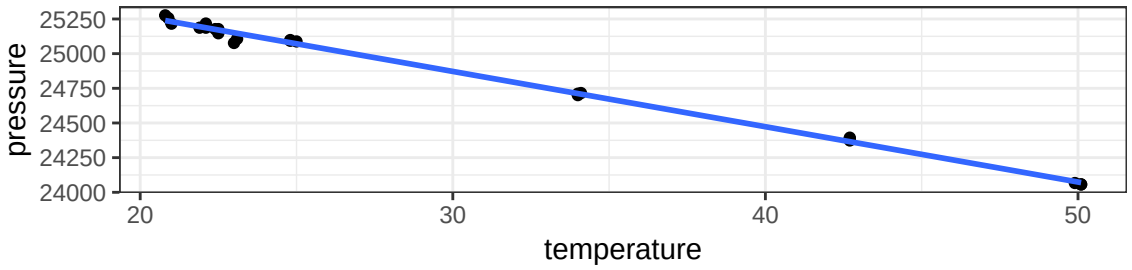


Bismuth is a metal that forms a crystal structure, as illustrated above. The crystal structure of bismuth changes depending on the temperature and pressure it is subjected to; there are four different crystal structures that will be realized at different temperatures and pressures. Houck studied the relationship between the temperature of Bismuth and the pressure at which it undergoes a first change in crystal structure, from Bi-I to Bi-II (Houck, J.C. (1970). Temperature coefficient of the bismuth I-II transition pressure. *J. Res.Nat. Bur. Stand.*, 74 A, 51-54). In this experiment, each sample was maintained at a specified temperature while the pressure was adjusted. The pressure at which each sample changed crystal structure was recorded.

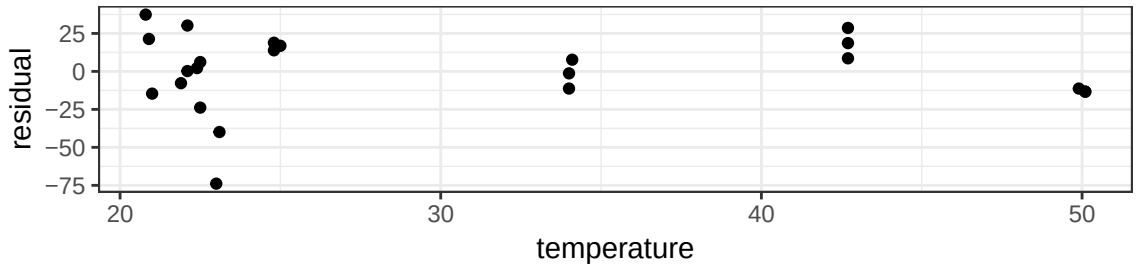
In this experiment, the researchers had experimental equipment that could be set for a specified temperature. To collect the data, they set this equipment to a specified temperature, and then ran several samples at that temperature before setting the equipment to a different temperature.

Here is a scatter plot of the data with a regression line fit overlaid on top, a scatter plot of residuals vs. temperature, and a histogram of the residuals.

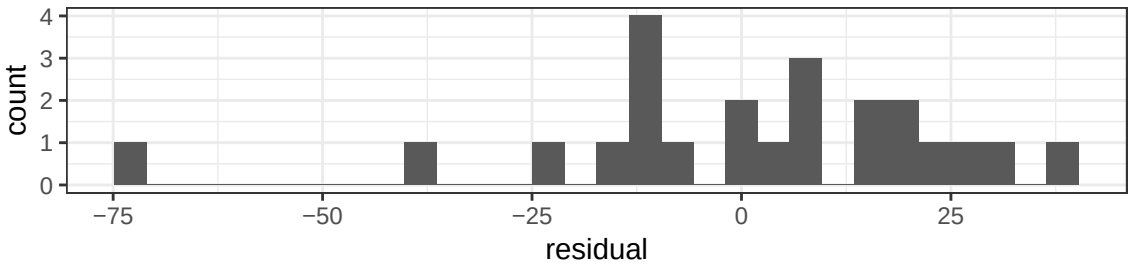
```
ggplot(data = bismuth, mapping = aes(x = temperature, y = pressure)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```




```
ggplot(data = bismuth, mapping = aes(x = temperature, y = residual)) +
  geom_point() +
  theme_bw()
```




```
ggplot(data = bismuth, mapping = aes(x = residual)) +
  geom_histogram() +
  theme_bw()
```



1. Check each of the regression model conditions. For each condition, write a sentence or two describing whether or not the condition is satisfied and why. If your conclusion is based on the plots above, please clearly indicate which plot or plots you are looking at and describe a specific characteristic of that plot that your conclusion is based on. For any conditions that are not satisfied, suggest a possible strategy for addressing the problem.

Linearity: From the scatterplot of the data, the association is approximately linear.

Independence: I have concerns about independence based on the scatter plot of residuals vs. temperature and the experimental design. The residuals at each temperature setting appear to be clustered together, either consistently positive or consistently negative within each group. The researchers set the equipment to a specified temperature, and then ran several samples at that temperature before setting the equipment to a different temperature. Observations at the same temperature setting are not independent; a small miscalibration of the equipment could affect all measurements at that temperature. To address this, we would need to use a different model.

Normally distributed residuals: From the histogram of residuals, the distribution appears to be unimodal and approximately symmetric.

Equal variance of residuals: From the scatter plot of residuals vs temperature, there appears to be more variability (more vertical spread) in the residuals for low temperatures than for high temperatures. We could try a tranformation of the response variable.

Outliers: From the histogram of residuals, there is one potential low outlier with a residual of about -75. From the scatter plot of residuals vs. temperature, it appears that that outlier is consistent with the larger standard deviations at low temperatures (so you could also say this is not an outlier). We could try a transformation of the response variable.

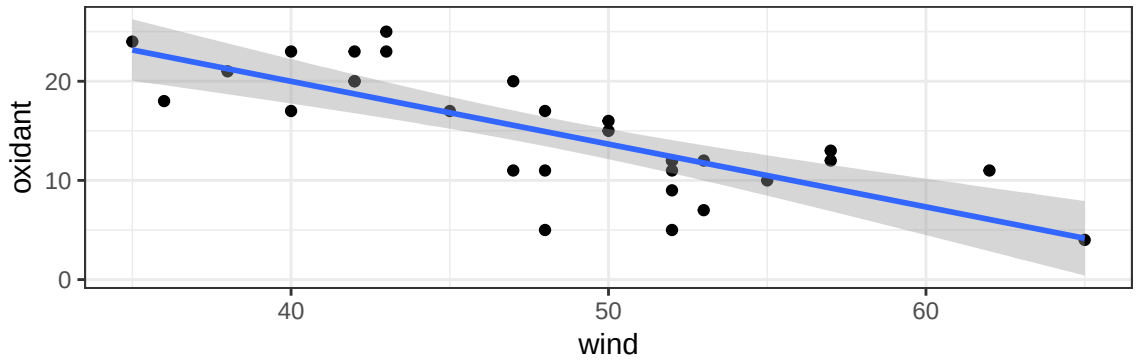
Stat 242 Quiz – Topics Drawn from Chapter 8

What's Your Name? _____

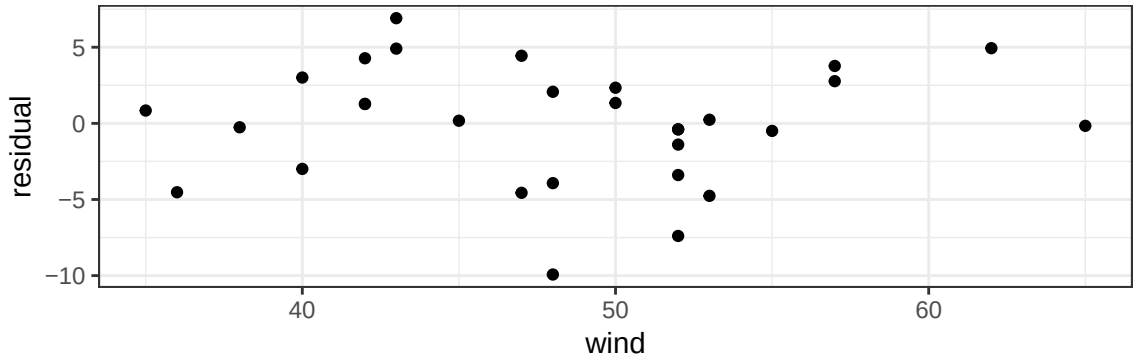
The Los Angeles Pollution Control District records the levels of pollutants and various meteorological conditions, and attempts to construct models to predict pollution levels and to gain a better understanding of the complexities of air pollution. Here we will consider a small subset of this data, with measurements of the maximum level of an oxidant (a photochemical pollutant) and average wind speeds in the mornings of 50 days during one summer.

Here is a scatter plot of the data with a regression line fit overlaid on top, a scatter plot of residuals vs. wind speeds, a scatter plot of residuals vs. day, and a histogram of the residuals.

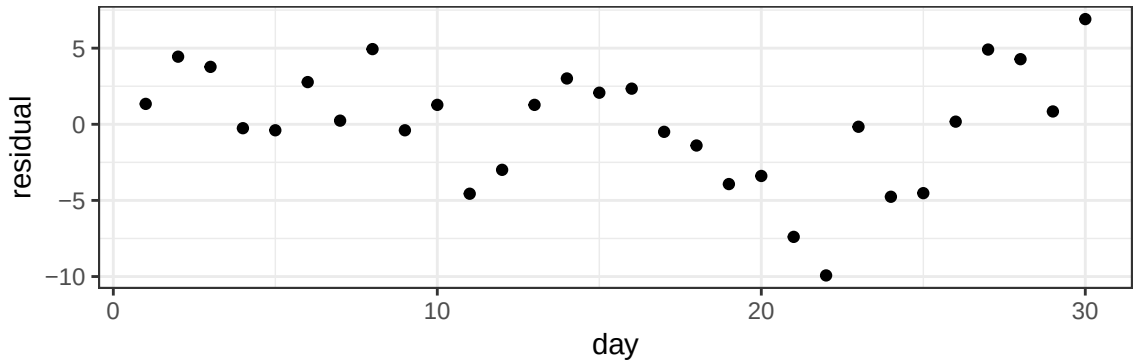
```
ggplot(data = air_pollution, mapping = aes(x = wind, y = oxidant)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()
```



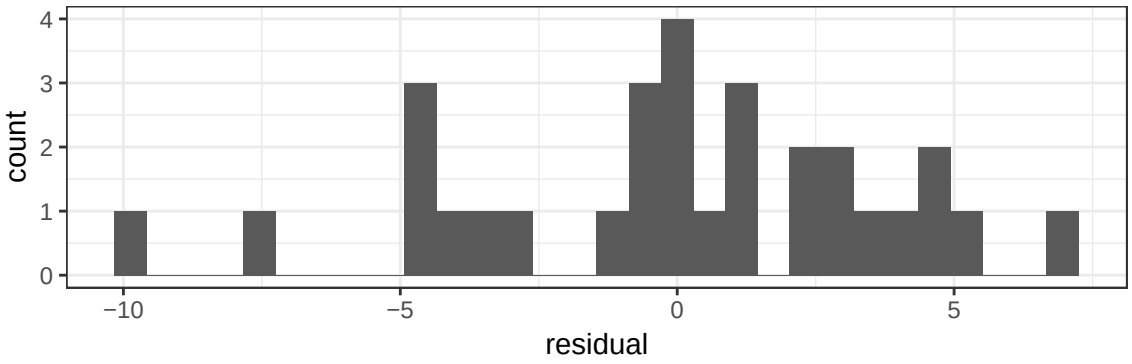

```
ggplot(data = air_pollution, mapping = aes(x = wind, y = residual)) +
  geom_point() +
  theme_bw()
```




```
ggplot(data = air_pollution, mapping = aes(x = day, y = residual)) +
  geom_point() +
  theme_bw()
```



```
ggplot(data = air_pollution, mapping = aes(x = residual)) +
  geom_histogram() +
  theme_bw()
```



1. Check each of the regression model conditions. For each condition, write a sentence or two describing whether or not the condition is satisfied and why. If your conclusion is based on the plots above, please clearly indicate which plot or plots you are looking at and describe a specific characteristic of that plot that your conclusion is based on. For any conditions that are not satisfied, suggest a possible strategy for addressing the problem.

Linearity: It is hard to be sure from the plot of oxidant levels vs wind speed, but the association appears to be approximately linear.

Independence: The plot of residuals vs. day shows that the residuals on consecutive days are similar; independence is not satisfied. We should use a different model.

Normally distributed residuals: The histogram of residuals shows them to be approximately normally distributed.

Equal variance of residuals: The scatter plot of residuals vs wind shows approximately equal vertical spread over the full range of values for wind speed.

Outliers: There are no serious outliers to be concerned about in either scatter plot.

Stat 242 Quiz – Topics Drawn from Chapter 8

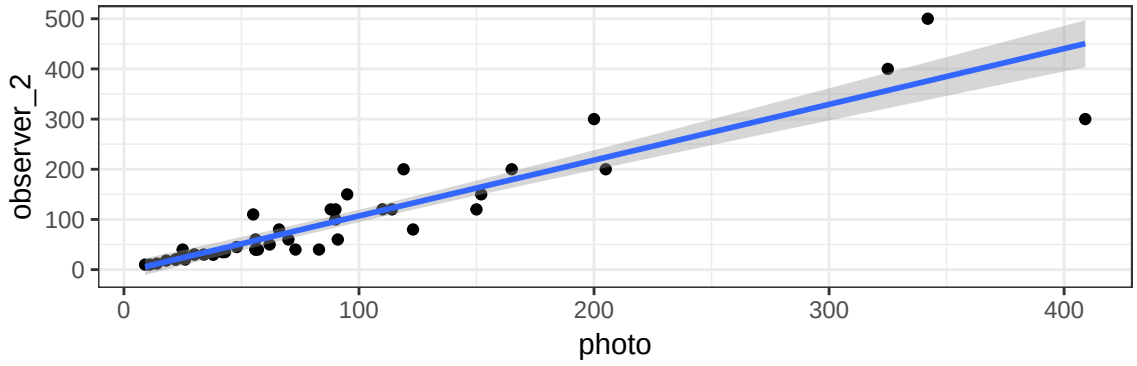
What's Your Name? _____

Aerial survey methods are sued to estimate the number of snow geese in their summer range areas westof Hudson's Bay in Canada. To obtain estimates, small aircraft fly over the range and, when a flock of snow geese is spotted, an experienced observed estimates teh number of geese in teh flock. To investigate the reliability of this method, an experiment in which an airplane carried two observers flew over 45 flocks, and each observer independently estimated the number of geese in the flock. Also, a photograph of the flock was taken so that an exact count of the number of geese in the flock could be obtained (Weisberg, 1985).

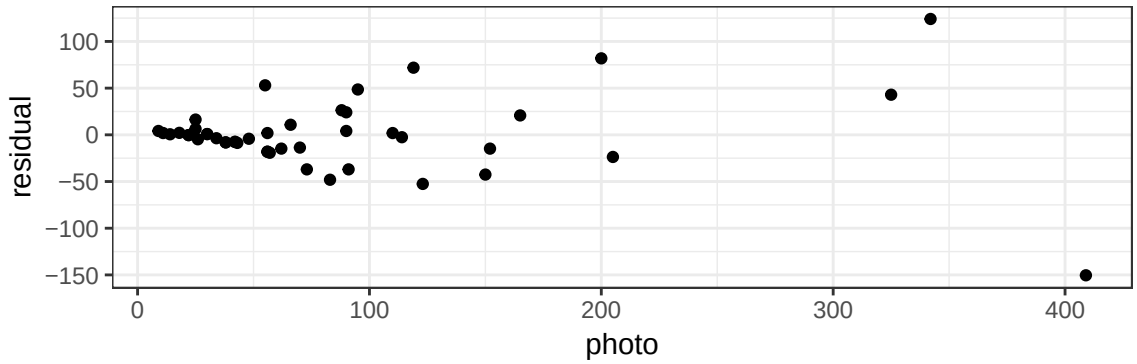
Here we will just look at the estimated counts from the second observer (our response variable), and see how they compare to the exact counts from the photograph (explanatory variable).

Here is a scatter plot of the data with a regression line fit overlaid on top, a scatter plot of residuals vs. the exact count from the photo, and a histogram of the residuals.

```
ggplot(data = geese, mapping = aes(x = photo, y = observer_2)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()
```

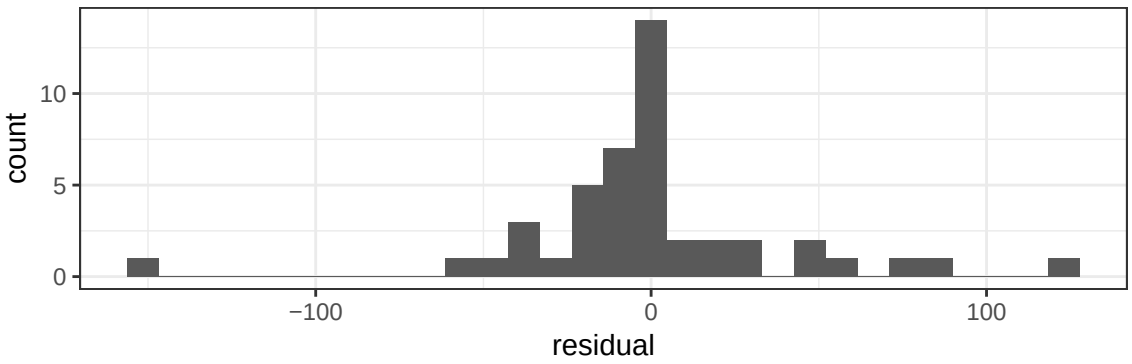



```
ggplot(data = geese, mapping = aes(x = photo, y = residual)) +
  geom_point() +
  theme_bw()
```




```
ggplot(data = geese, mapping = aes(x = residual)) +
  geom_histogram() +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1. Check each of the regression model conditions. For each condition, write a sentence or two describing whether or not the condition is satisfied and why. If your conclusion is based on the plots above, please clearly indicate which plot or plots you are looking at and describe a specific characteristic of that plot that your conclusion is based on. For any conditions that are not satisfied, suggest a possible strategy for addressing the problem.

Linearity: The scatter plot of the data shows an approximately linear association.

Independence: There are no obvious sources of connections between the different observations in our data set; an assumption of independence seems ok.

Normally distributed residuals: The histogram of the residuals shows some residuals that are very large in magnitude, and heavy tails. We should try a data transformation.

Equal variance of residuals: Both scatter plots show increasing standard deviations (more vertical spread in the points) for larger flock sizes. We should try a data transformation, starting with the response variable.

Outliers: I would not describe any points in this data set as outliers, since they are consistent with the general trends in terms of association and standard deviations in the scatter plot of the data. I would also be ok with it if you described the points for very large flocks as outliers.

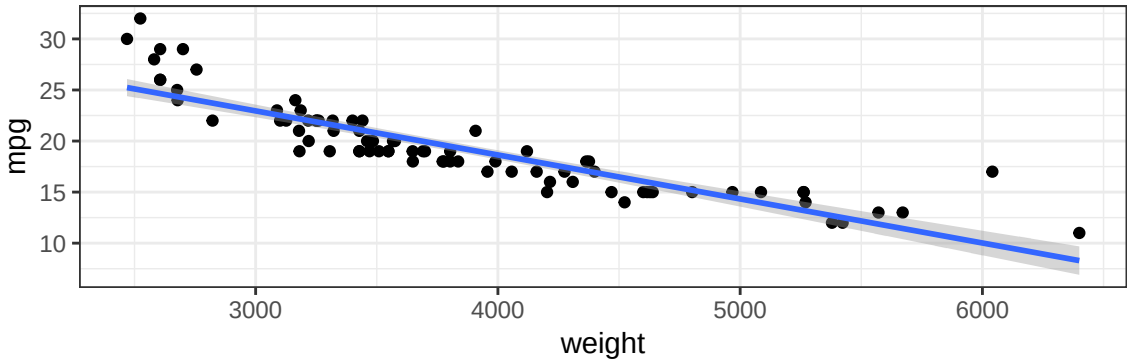
Stat 242 Quiz – Topics Drawn from Chapter 8

What's Your Name? _____

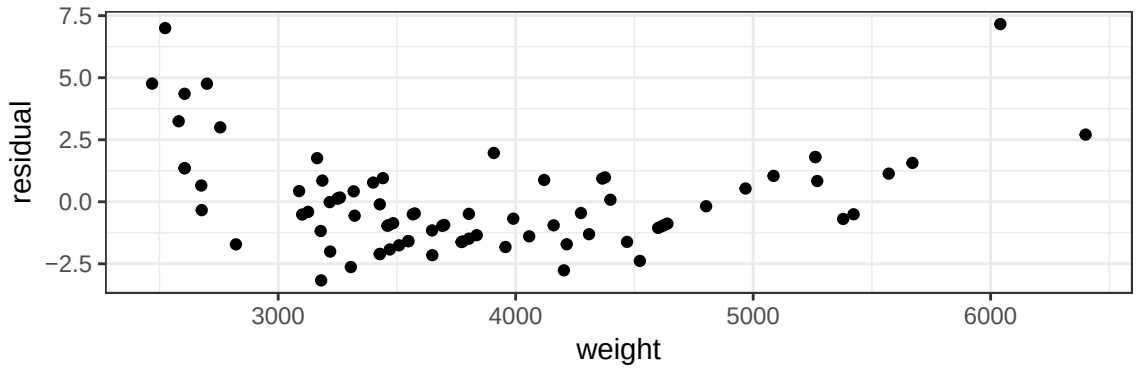
We know that heavier cars need more fuel, but exactly how does a car's weight affect its fuel efficiency? We have a data set with the weight (the explanatory variable) and fuel efficiency (the response variable, in miles per gallon, mpg) for 80 cars.

Here is a scatter plot of the data with a regression line fit overlaid on top, a scatter plot of residuals vs. the exact count from the photo, and a histogram of the residuals.

```
ggplot(data = cars, mapping = aes(x = weight, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()
```

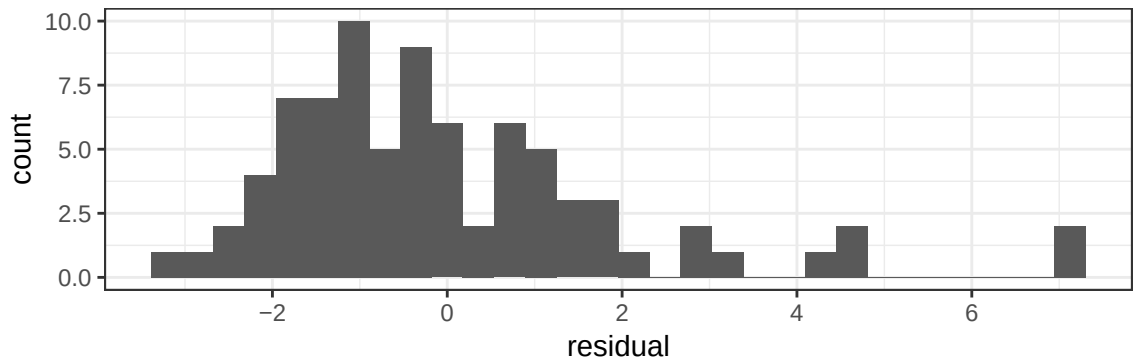



```
ggplot(data = cars, mapping = aes(x = weight, y = residual)) +
  geom_point() +
  theme_bw()
```




```
ggplot(data = cars, mapping = aes(x = residual)) +
  geom_histogram() +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1. Check each of the regression model conditions. For each condition, write a sentence or two describing whether or not the condition is satisfied and why. If your conclusion is based on the plots above, please clearly indicate which plot or plots you are looking at and describe a specific characteristic of that plot that your conclusion is based on. For any conditions that are not satisfied, suggest a possible strategy for addressing the problem.

Linearity: From both scatter plots, there is a non-linear association between weight and fuel efficiency. Points fall above the line for low weights, then below the line for middling weights, and above the line again for large weights.

Independence: A condition of independence could be violated if there are multiple cars in the data set that are based on the same platform. Those cars might have similar fuel efficiencies relative to their weights.

Normally distributed residuals: It's not useful to assess the normality of the residuals when a linear model is fit to a non-linear relationship (there are bigger problems).

Equal variance of residuals: Aside from the outlier (to be discussed next), in the scatter plot of residuals vs weight, there is more vertical spread in the residuals for low weights than for large weights; not satisfied. We could try a transformation of the response.

Outliers: There is one outlier at around 6000 pounds, from the residuals plot. We could see if there was a data entry error, and correct it if so. Otherwise, conduct the analysis both with and without the outlier and report both sets of results.