Regression model

\[ Y = \mu_{y|x} + \varepsilon \]

- \( Y \) represents a value of the response variable.
- \( \mu_{y|x} \) represents the population mean response for a given value of the explanatory variable, \( x \).
- \( \varepsilon \) represents the random error.

Linear Regression Model

\[ Y = \mu_{y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon \]

- \( \beta_0 \) is the Y-intercept parameter.
- \( \beta_1 \) is the slope parameter.

Residual

- (Observed \( Y \) – Fitted \( Y \))
- Fit \( \hat{Y} = \hat{\mu}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x \)
- Residual \( \hat{\varepsilon} = Y - \hat{Y} = Y - (\hat{\beta}_0 + \hat{\beta}_1 x) \)
Conditions
- The relationship is linear.
- The random error term, $\varepsilon$, is
  - Independent
  - Identically distributed
  - Normally distributed with standard deviation, $\sigma$.

Residual vs. Explanatory

Residual vs. Predicted
Interpretation

- Random scatter around the zero line indicates that the linear model is adequate for the relationship between carbon dioxide and temperature.

Patterns

- Over/Under/Over or Under/Over/Under
  - The linear model may not be adequate. We could do better by accounting for curvature with a different model.

Speed and Stopping Distance
Patterns

- Two, or more, groups
  - May require separate regression models for each group.

Gas used vs. Temperature

Checking Conditions

- Independence.
  - Hard to check this but the fact that we obtained the data through a random sample of years assures us that the statistical methods should work.
Checking Conditions

- Identically distributed.
  - Check using an outlier box plot. Unusual points may come from a different distribution
  - Check using a histogram. Bimodal shape could indicate two different distributions.

Checking Conditions

- Normally distributed.
  - Check with a histogram. Symmetric and mounded in the middle.
  - Check with a normal quantile plot. Points falling close to a diagonal line.
Residuals

- Histogram is skewed right and mounded to the left of zero.
- Box plot is skewed right with no unusual points.
- Normal quantile plot has points that do not follow the diagonal, normal model, line very well.

Checking Conditions

- Constant variance.
  - Check the plot of residuals versus the explanatory or predicted.
  - Points should show the same spread for all values of the explanatory variable.

Non-constant variance

- Diagram showing the relationship between residuals and the explanatory variable.
Residual vs. Explanatory

Residual vs. Predicted

Constant Variance

Points show about the same amount of spread for all values of the explanatory variable.
Conclusion

- The independence, identically distributed and common variance conditions appear to be satisfied.
- The normal distribution condition may not be met for these data.

Consequences

- The P-values for tests may not be correct.
  - However, the P-value was so small, there is still strong evidence for a linear relationship between carbon dioxide and temperature.

Consequences

- The stated confidence level may not give the true coverage rate.
  - We have confidence in the intervals, it may not be 95%.