Chapter 7 and 8: Linear Model Appropriateness and Diagnostics

Review

A _____ is the difference between the observed value and the predicted value of the response variable based on linear regression. Recall that:

-
-

Overall, residuals give us a general idea of how well our linear regression fits our data. This brings us to the topic of **Linear Model Appropriateness**. When determining the appropriateness of our linear models, it is useful to ask ourselves the following questions:

1.
    -
2.
    -
    -

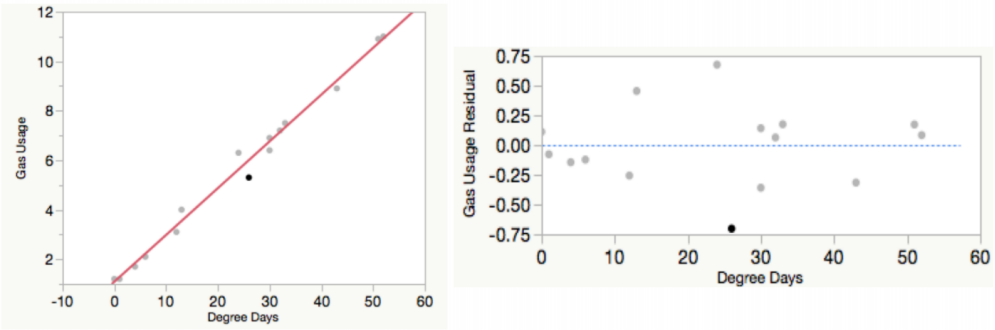
The best way we can answer these questions is by looking at _____

Residual Plots

The goal of a residual plot is to determine if a linear model is appropriate for a relationship between two quantitative variables.

A residual plot is basically a scatterplot with:

-
-
-

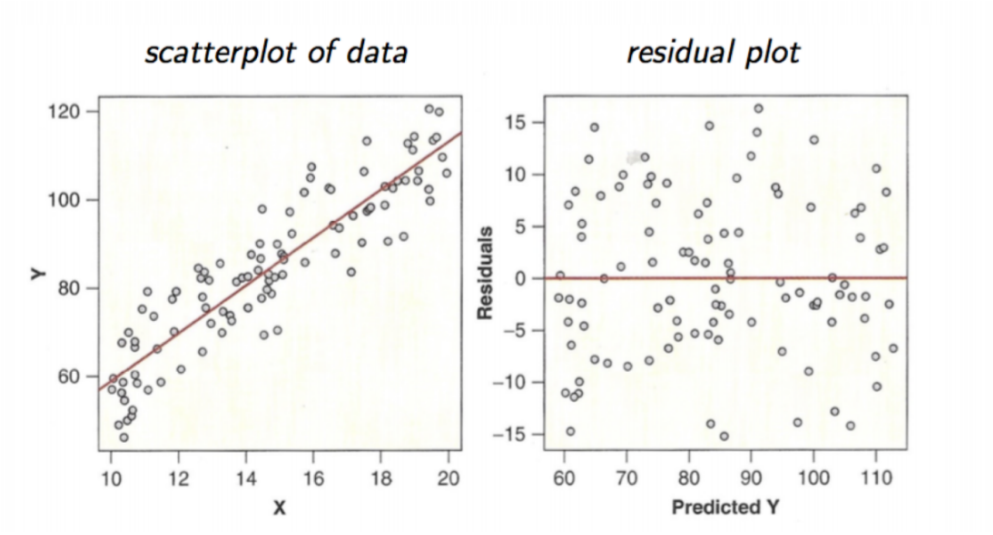


Interpretation

When we encounter a residual plot we look for the four following characteristics:

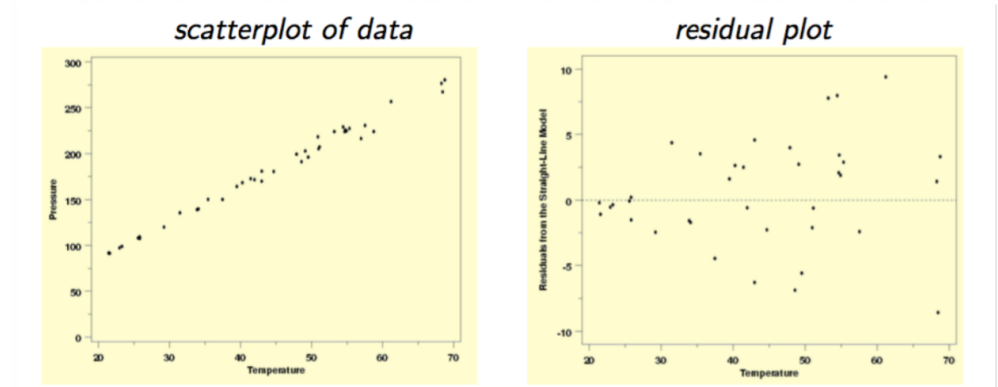
1.
2.
3.
4.

Example *Based on this residual plot we would say this residual plot falls under situation one. Meaning, there is* _____.



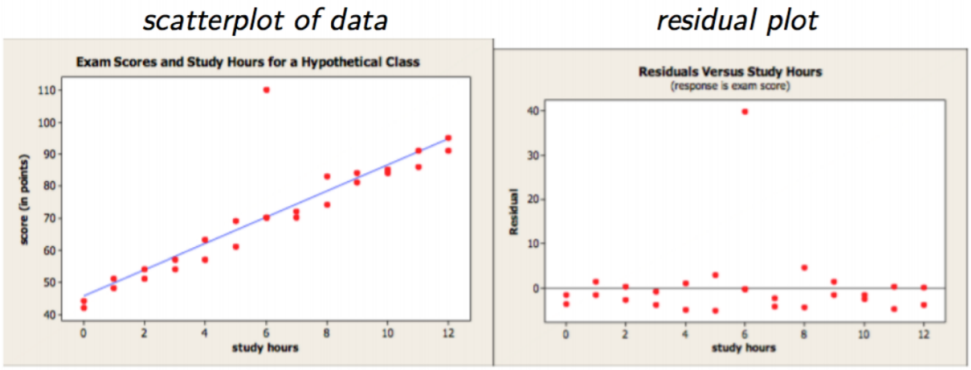
This means a linear model _____

Example *Based on this residual plot we would say this residual plot falls under situation Two. Meaning, there is* _____.



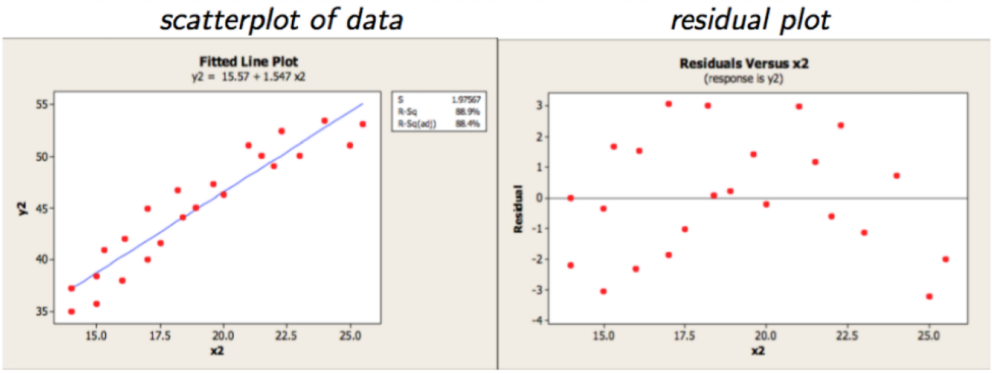
This means a linear model _____ *but caution is needed to continue with analysis since there may be problems with non-constant variance. To learn more, take some higher level stats courses for how to deal with this problem!*

Example *Based on this residual plot we would say this residual plot falls under situation Three. Meaning, there is* _____.



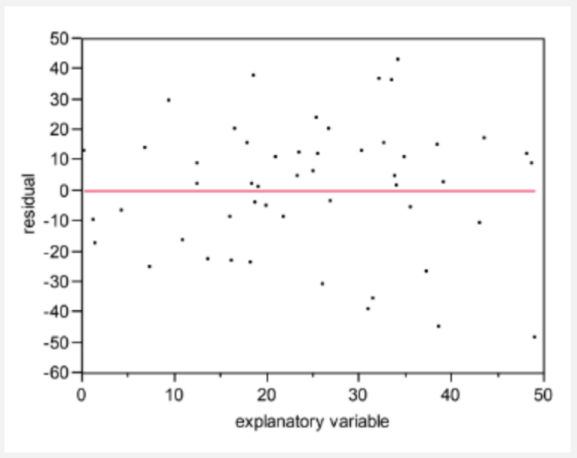
This means a linear model _____ *but there are some special steps we need to take to deal with outliers.*

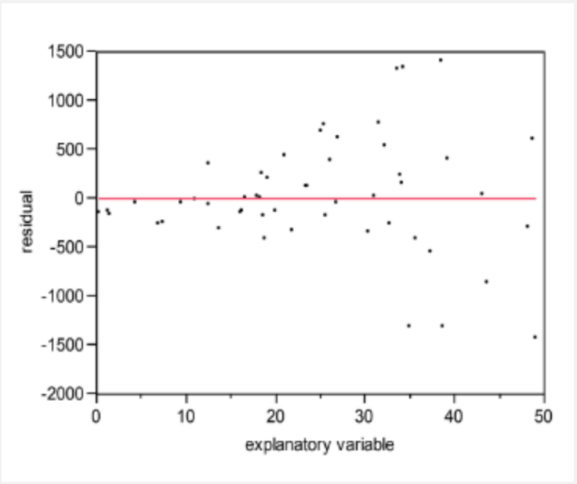
Example *Based on this residual plot we would say this residual plot falls under situation four. Meaning, there is* _____.

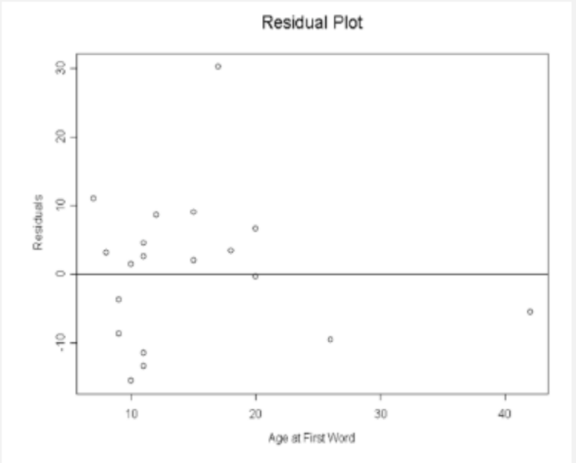


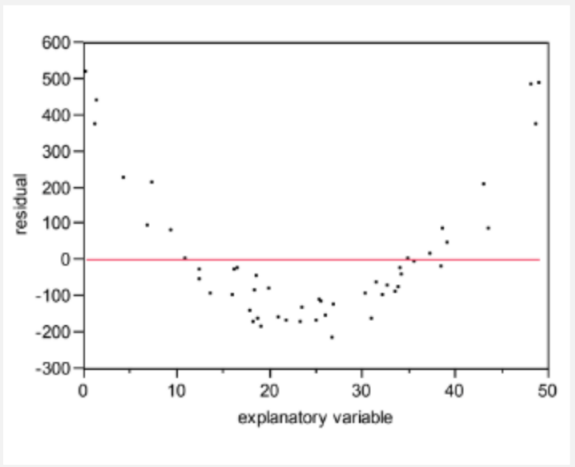
This means a linear model _____

For the following scatterplots, label with each with a "Yes" a "Yes, but" or a "No" for whether each situation is appropriate for a linear model. Then determine which of the four situations the data falls under.









Introduction to R^2

Now that we have covered a method for determining whether a linear model is appropriate, it might also be useful to find out just how much of the variability in our data can be accounted for by our model. First, let's review sources of variability:

	Observed Values (y)	Predicted Values ($\hat{y}$)	Residuals (e)
Mean	$\bar{y}$	$\bar{y}$	0
Standard Deviation	s_y	$s_{\hat{y}}$	s_e
Variance	s_y^2	$s_{\hat{y}}^2$	s_e^2

The following relationship holds for all least squares regression lines:

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

When we compare $s_{\hat{y}}^2$ to s_y^2 we can draw two main conclusions:

-
-

To measure the closeness of $s_{\hat{y}}^2$ and s_y^2 we use the coefficient of determination known as R^2. The coefficient of determination quantifies the variation in the response variable, y, that can be explained by the linear regression model with explanatory variable, x.

To calculate R^2 we use the following equivalencies

$$\begin{aligned} R^2 &= \frac{s_{\hat{y}}}{s_y} \\ &= (r)^2 \\ &= (correlation)^2 \end{aligned}$$

Interpretation:

Properties:

-
-
-
-
-
-