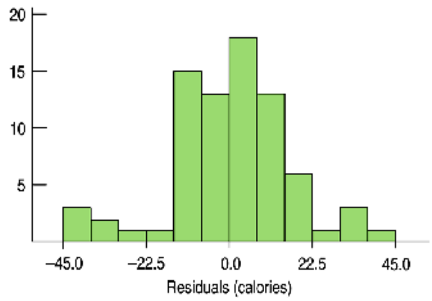
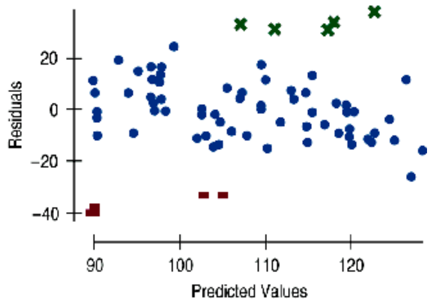
Chapter 9: Regression Wisdom

Residuals

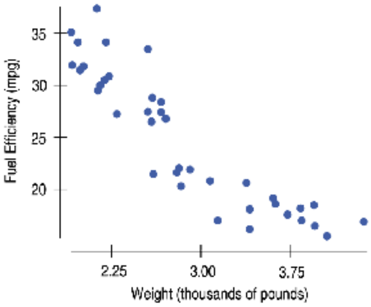
To determine whether a linear model is appropriate, we examine the residual plot. It is a good idea to look at both a histogram of the residuals and a scatterplot of the residuals versus the predicted values. If the histogram of the residuals has multiple modes , that may indicate that there are subgroups within the set of data. If a linear model is appropriate, the histogram should look approximately normal and the scatterplot of residuals should show random scatter .

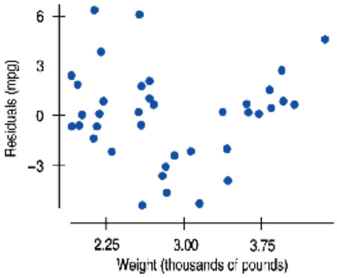




If we see a curved relationship in the residual plot, the linear model is not appropriate.

Another type of residual plot shows the residuals versus the explanatory variable. Note that the scale on the horizontal axis is exactly the same as that of the original scatterplot of *x* and *y*.





When using a model to predict values of the response variable, two types of predictions can be made: interpolation or extrapolation. Making predictions within the given domain of *x*-values is called interpolation. Making predictions outside the given domain of *x*-values is called extrapolation. Only extrapolation is reliable. Interpolation is unreliable, because patterns can change abruptly. In the above example, we would be extrapolating if we use a model to predict the fuel efficiency of cars weighing more than 4.5 thousand pounds or less than 1.5 thousand pounds.

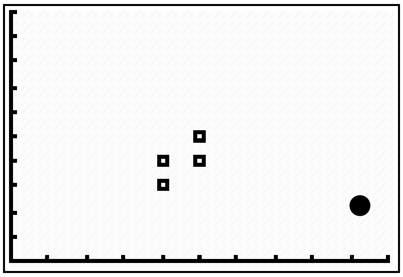
Even if a linear model is appropriate, remember that association does not imply causation. There may be confounding variables that account for the association. Also, if we're plotting summary values (like the mean) instead of individual values, the association may appear stronger than it really is.

Outliers, Leverage, and Influential Points

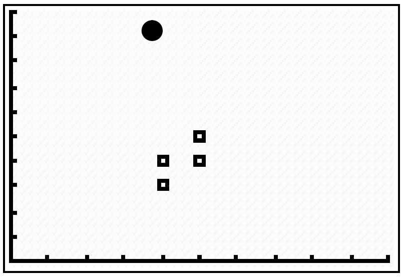
When examining a scatterplot it is important to look for any unusual points. There are three ways a point can be considered unusual.

1. A point is unusual if it stands away from the others. This is called an outlier.
2. A point is unusual if its *x*-value is far from the mean of all the *x*-values. This is called leverage.
3. A point is unusual if, when omitted, it significantly changes the linear model. This is called an influential point.

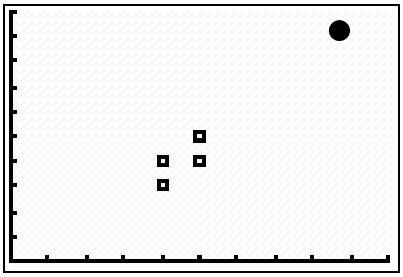
For each scatterplot, add a point with the given characteristics (if possible):



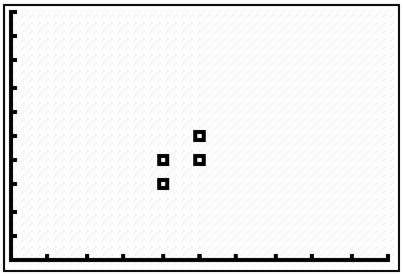
Outlier, influential,
high leverage



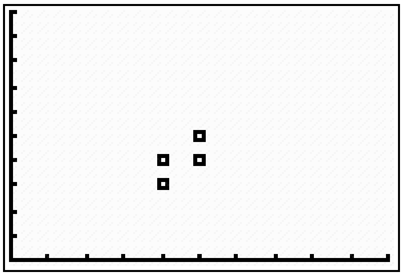
Outlier, influential
low leverage



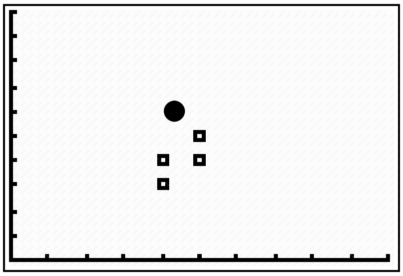
Outlier, not influential
high leverage



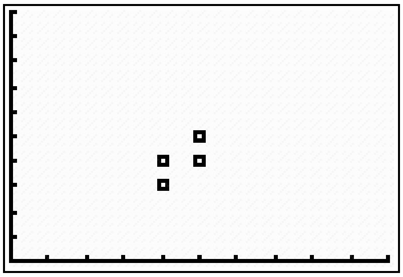
Outlier, not influential,
low leverage (not possible)



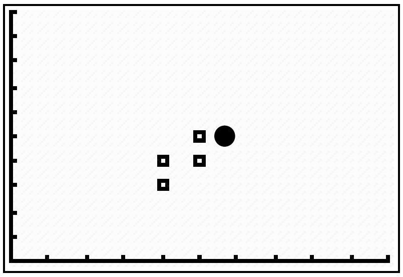
Not an outlier, influential
high leverage (not possible)



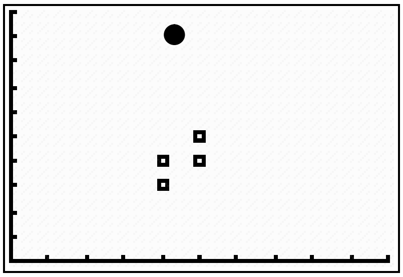
Not an outlier, influential
low leverage



Not an outlier, not influential,
high leverage (not possible)



Not an outlier, not influential
low leverage



Outlier, no leverage

If there is an unusual point in your scatterplot, it must be mentioned. It is then wise to examine two models, one with the point included, and one with the point omitted.