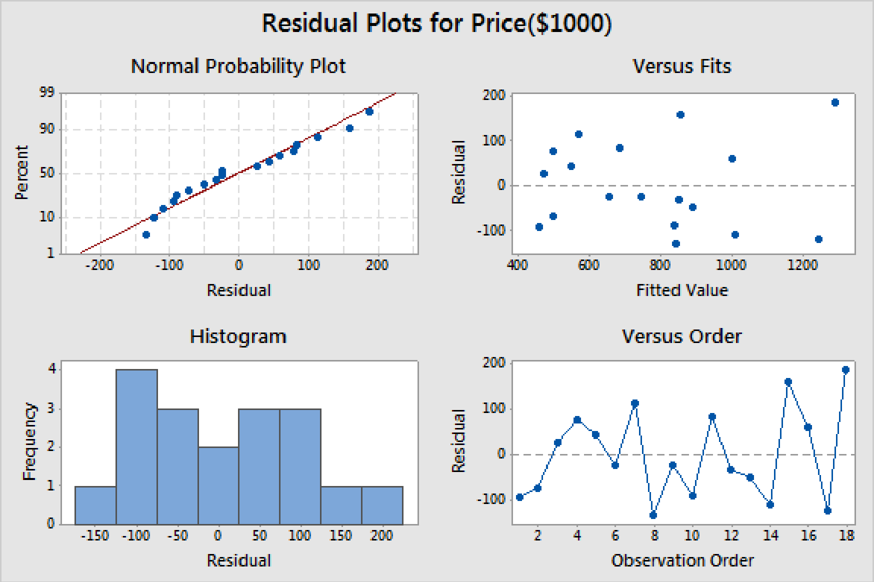
Regression Diagnostics – Solutions
COR1-GB.1305 – Statistics and Data Analysis

Model Assumptions

1. Here are plots of the residuals from the least squares fit to the housing data.



Do the plots indicate any potential violations in assumptions? Specifically, answer the following questions.

(a) Do the residual errors look approximately normal?

> **Solution:** The normal probability plot and the histogram show that the residuals are approximately normal.

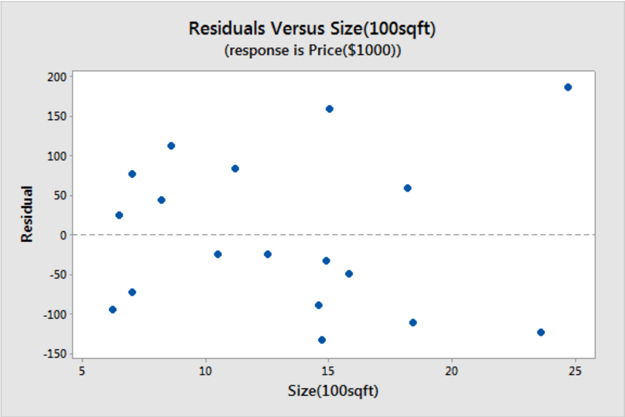
(b) Does the error variance look constant?

> **Solution:** The plot of residuals versus fitted value and residuals versus order hint that the variance of the residuals might be larger when the fitted value is big, but there is not enough data to say for certain.

(c) Is there any apparent dependence in the residuals?

> **Solution:** There is no clear pattern in the plot of residual versus fit or the plot of residual versus observation order. Thus, there is no apparent dependence in the residuals.

2. Here is a plot of the residuals versus Size (x).



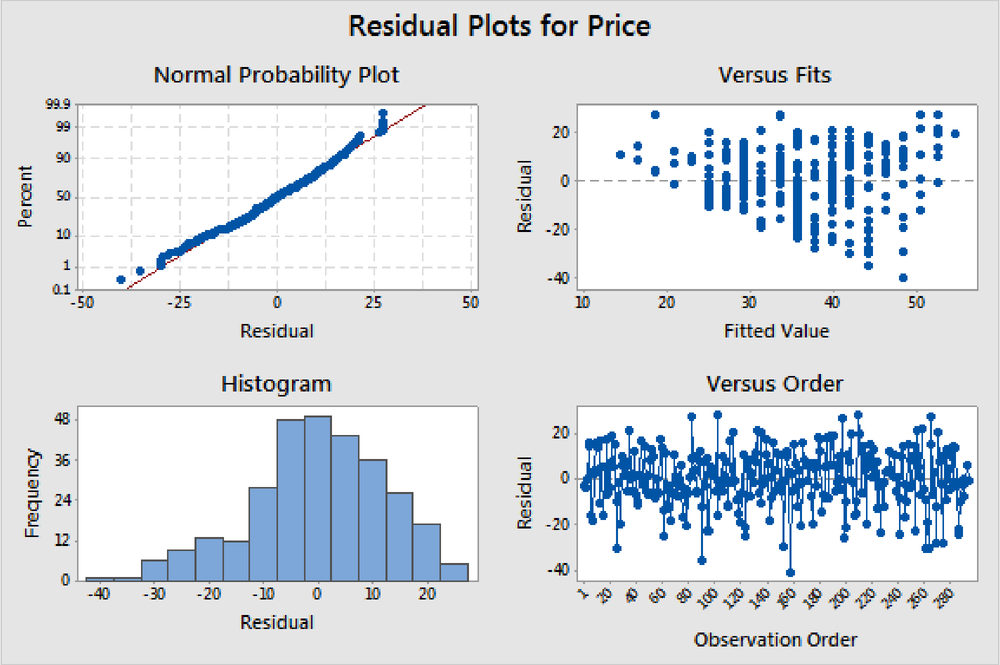
(a) Why is this plot nearly identical to the plot of residuals versus fits?

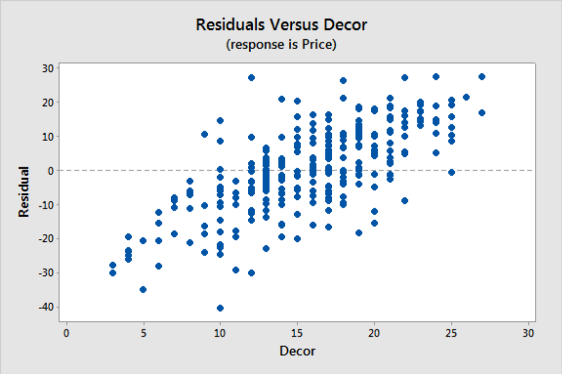
Solution: Both plots have the same Y-axis. The X-axis on the plot of residuals versus size is x_i. The X-axis on the plot of residuals versus fits is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, which is an affine transformation of x_i. Thus, the only difference in the plots is the values on the X-axis scale.

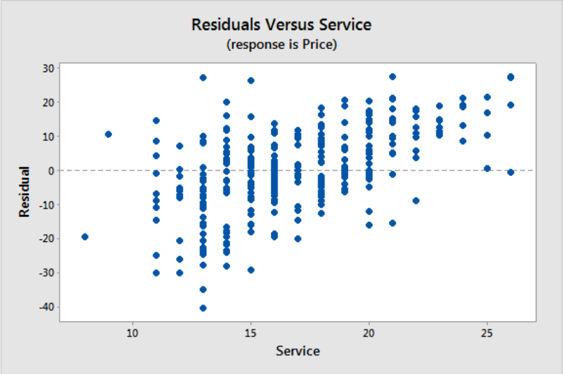
(b) Does the plot of residuals versus fit always look like the plot of residuals versus x?

Solution: No. If $\hat{\beta}_1$ is negative, then the plot is flipped along the horizontal direction.

3. Here are some plots of the residuals from the fit of Price to Food for the Zagat data:







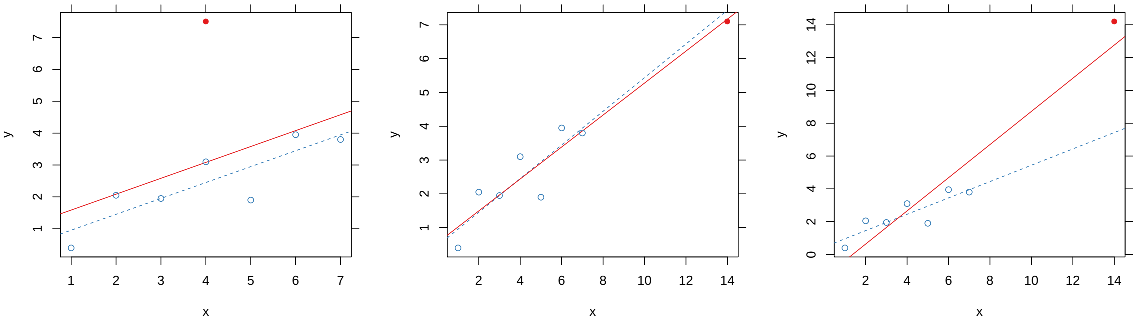
Use the plots to assess whether or not the four regression assumptions hold.

Solution: The normal probability plot and the histogram indicate that the residuals, are approximately normally-distributed. In the Residual verses Fitted Values, it looks like the mean value of the residual is approximately 0. This plot also shows that the error variance tends to increase when the fitted value increases. There is no apparaent pattern in the "Versus Order" plot, but there are clear trends in the "Versus Decor" and the "Versus Service" plots.

In summary, two assumptions are plausible: that the errors are normally distributed, and that the mean value of the error is zero. One assumption is violated, but only mildly so: that the error variance is constant. One assumption is in clear violation: that the errors are independent.

Outliers and Leverage Points

4. Each of the following scatterplots show two regression lines: the solid line is fitted to all of the points, and the dashed line is fitted to just the hollow points.



(a) (b) (c)

(a) In plot (c), including the solid point has a big effect on the regression fit. Why is this?

Solution: The point is a leverage point (its x value is far from the mean of the points) and its y value does not follow the same linear trend as the other points.

(b) Does a leverage point always have a big influence on the regression fit?

Solution: No. For example, the solid point in plot (b) is a leverage point but it does not have a big influence on the fit.

(c) Can a point that is not a leverage point have a big influence on the regression fit?

Solution: Yes. The solid point in (a) is not a leverage point (its x value is close to the mean of the other points), but the point still has a big influence.

(d) In each of the above three cases, should we include the solid point in the regression analysis? If not, what should we do with the point?

Solution: In (a) and (c) we should probably remove the point from the regression analysis. We should discuss this point separately.