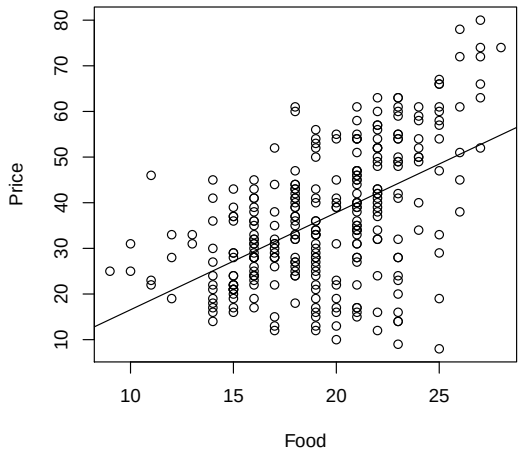
Regression Diagnostics (Solutions)
STAT-UB.0003: Regression and Forecasting Models

Linear regression model

1. Here is the least squares regression fit to the Zagat restaurant data:



Here is the Minitab output from the fit:

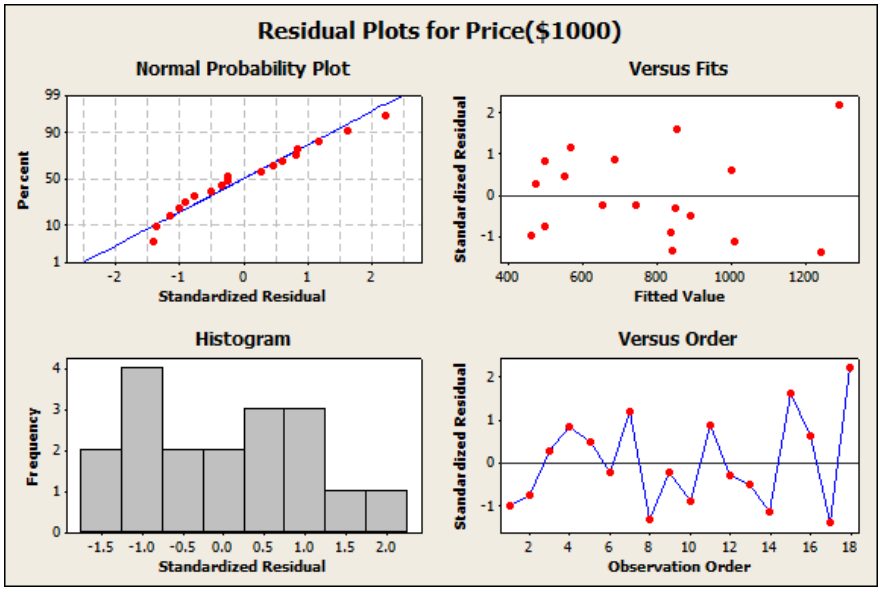
```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
12.5559  27.93%     27.68%      26.86%

Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -4.74     3.95    -1.20    0.232
Food      2.129    0.200    10.64    0.000  1.00

Regression Equation

Price = -4.74 + 2.129 Food
```

What proportion of the variability in the response is explained by the regression model (this is the "coefficient of determination", commonly referred to as the R^2 value)? What is the meaning of this number?

Solution: From the output, $R^2 = 27.93\%$. This is the ratio of the regression sum of squares ($\sum(\hat{y}_i - \bar{y})^2$) to the total sum of squares ($\sum(y_i - \bar{y})^2$).

Model assumptions

2. 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?

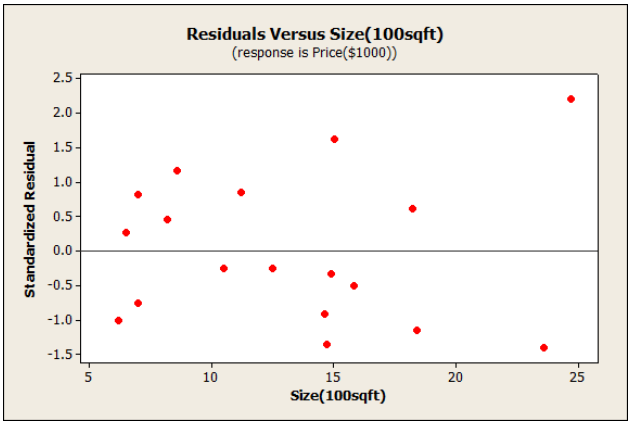
Solution: There is no clear pattern in the plot of residual versus fit or the plot of residual versus observation order. Thus, there is no apparent dependence in the residuals.

3. What is a "standardized" residual? Why is it sometimes easier to interpret a standardized residual than an ordinary residual?

Solution: A "standardized" residual is defined by $\hat{\varepsilon}_i/s$. This is "standardized" so that the sample standard deviation of the standardized residuals is approximately equal to 1.

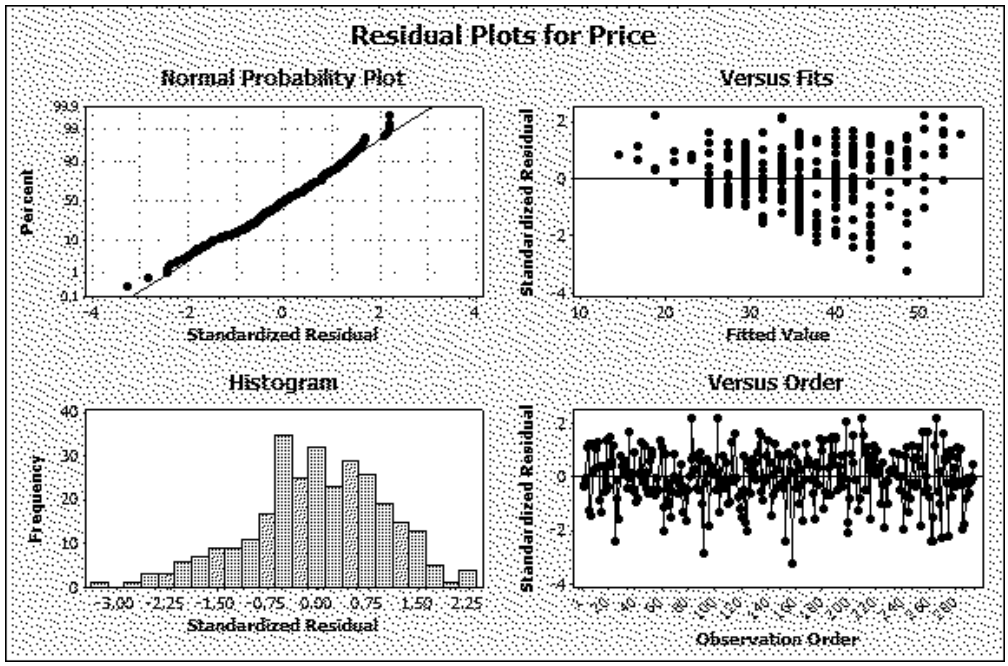
Standardized residuals are sometimes easier to interpret than ordinary residuals because we do not need to know s to know what counts as "large"/"unusual". In particular, 95% of standardized residuals should be in the range $(-2, 2)$. Standardized residuals much larger than this are unusual.

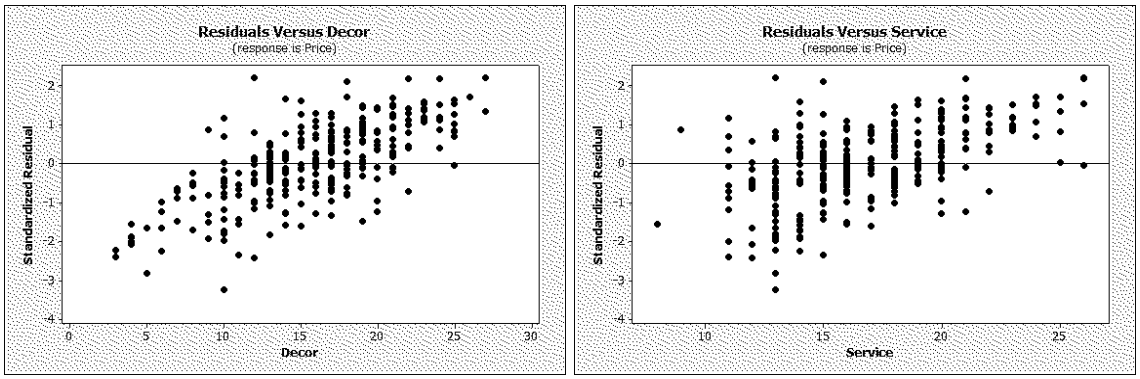
4. Here is a plot of the residuals versus Size (x). Why is this plot nearly identical to the plot of residuals versus fits?



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.

5. 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.