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

Linear Regression (Review)

1. Here is the Minitab output that result from fitting a regression model to the housing data ($n = 18$). Some of the values have been replaced by question marks.

```
Analysis of Variance

Source           DF   Adj SS   Adj MS  F-Value  P-Value
Regression        ?  1087557  ???????   ??????    ?????
  Size(100sqft)   ?  ???????  ???????   ??????    ?????
Error            ??   164431    ?????
  Lack-of-Fit    ??   ??????    ?????     ????    ?????
  Pure Error      ?    ?????    ?????
Total            ??  1251988

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
101.375  86.87%     ??????      ??????

Coefficients

Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant       182.3     62.4     2.92    0.010
Size(100sqft)  44.95     4.37    10.29    0.000  ????

Regression Equation

Price($1000) = 182.3 + 44.95 Size(100sqft)
```

Explain the meanings and uses of all of the non-missing numbers in the regression output.

Solution: The numbers in the Analysis of Variance table are

- SSR = 1087557; the variability in the regression fit, $\hat{y}_i$.
- SSE = 164431; the variability in the residual errors, $y_i - \hat{y}_i$.
- SST = 1251988; the variability in the response, y_i.

Recall that SST = SSR + SSE.

Then numbers in the Model Summary are

- $s = 101.375$; the standard error of the regression. This is an estimate of the standard deviation of the error in the regression model. If the model assumptions are in force, then roughly 95% of the y_i values will be within $2s$ of the regression line. This number is computed as

$$s = \sqrt{\frac{\text{SSE}}{n-k}},$$

where n is the number of data points in the sample and k is the number of predictor variables ($k = 1$ for simple linear regression).

- $R^2 = 86.87\%$; this is the proportion of variability in the response that is explained by the regression fit. It is equal to

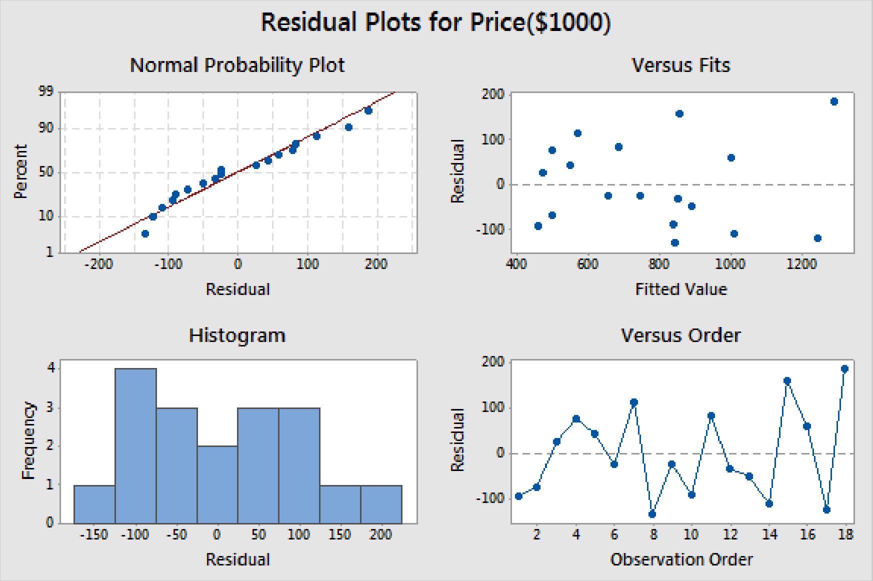
$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

The numbers in the Coefficients table are

- $\hat{\beta}_0 = 182.3$; the estimated intercept.
- $\text{SE}(\hat{\beta}_0) = 62.4$; the standard error of $\hat{\beta}_0$, an estimate of the standard deviation of the random variable associated with $\hat{\beta}_0$.
- $t_0 = 2.92 = \frac{182.3-0}{62.4}$; a test statistic for the null hypothesis $H_0 : \beta_0 = 0$.
- $p_0 = 0.010$ the p-value associated with t_0.
- $\hat{\beta}_1 = 44.95$; the estimated slope.
- $\text{SE}(\hat{\beta}_1) = 4.37$; the standard error of $\hat{\beta}_1$.
- $t_1 = 10.29 = \frac{44.95-0}{4.37}$; a test statistic for the null hypothesis $H_0 : \beta_1 = 0$.
- $p_1 < 0.001$; the p-value associated with t_1. (The p-value is reported as 0.000 in the Coefficients table, but p-values are never exactly equal to 0.)

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.

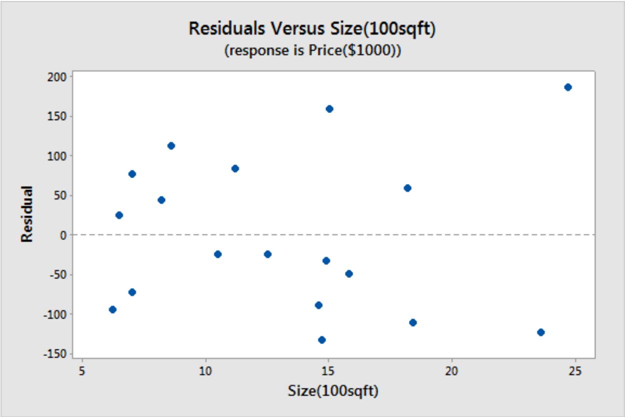
(b) Does the error variance look constant?

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



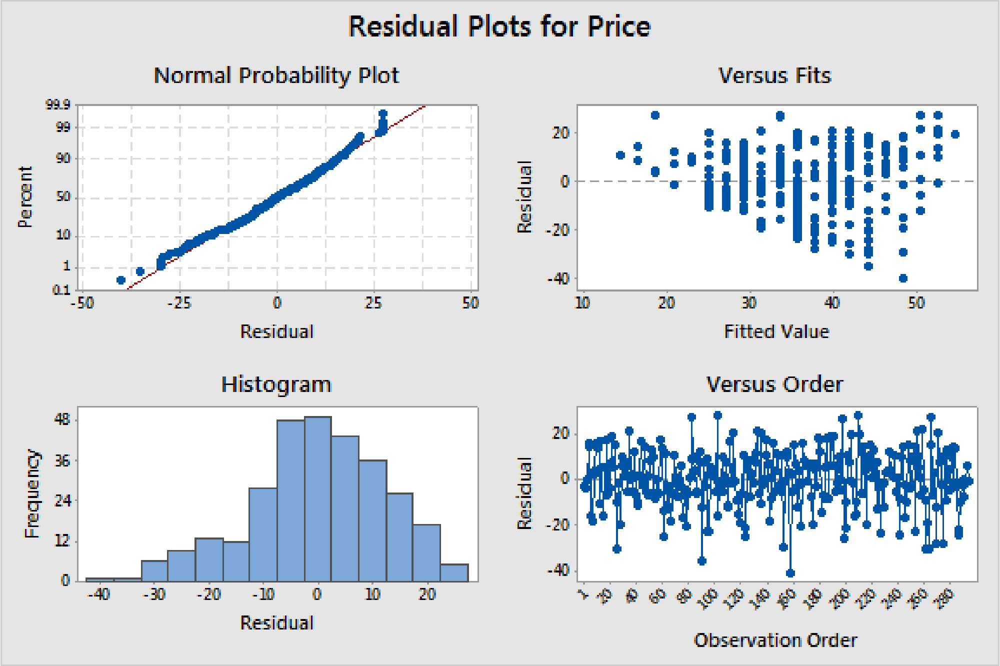
(a) Why is this plot nearly identical to the plot of residuals versus fits?

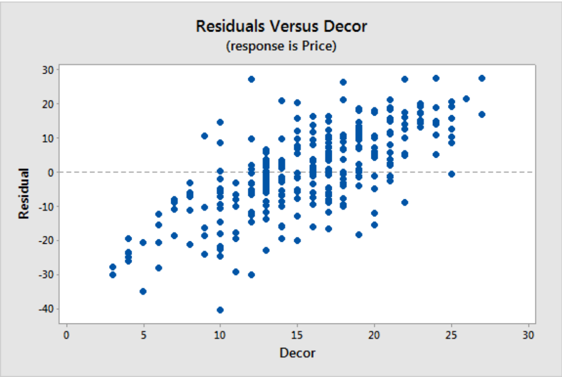
Solution: Both plots have the same Y-axis. The X-axis on the plot of residuals versus size is x_i. The X-axis on the plot of residuals versus fits is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, which is an affine transformation of x_i. Thus, the only difference in the plots is the values on the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

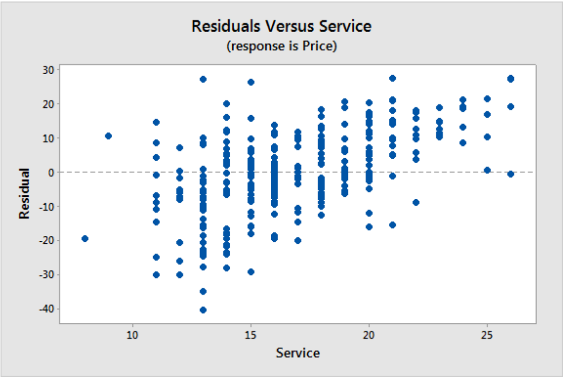
(b) Does the plot of residuals versus fit always look like the plot of residuals versus x?

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

4. 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 apparent 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.