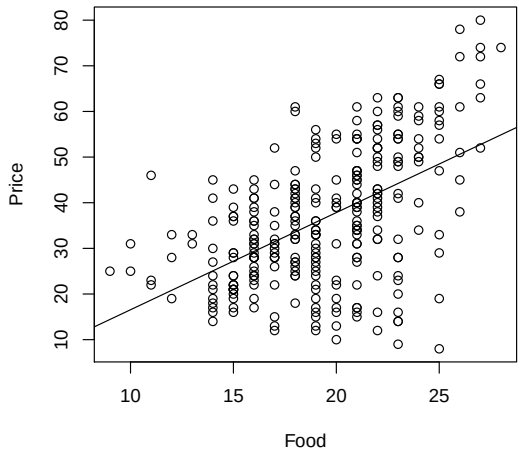
Regression Model Assumptions (Solutions)
STAT-UB.0003: Regression and Forecasting Models

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
```

(a) What are the estimated intercept and slope?

> **Solution:** The estimated intercept is $\hat{\beta}_0 = -4.74$; the estimated slope is $\hat{\beta}_1 = 2.129$.

(b) Use the estimated regression model to estimate the average dinner price of all restaurants with a quality rating of 20.

> **Solution:** If Food $= 20$, then estimated expected price per meal (\$) is $\widehat{\text{Price}} = -4.74 + 2.129(20) = 37.84$.

(c) In the estimated regression model, what is the interpretation of the slope?

Solution: For every 1-point increase in food quality, the expected dinner price goes up by \$2.129.

(d) In the estimated regression model, why doesn't the intercept have a direct interpretation?

Solution: This would be the expected dinner price for a restaurant with a quality of 0. No such restaurant exists (this is outside the range of the data).

2. Refer to the Minitab output from the previous problem, the regression analysis of the Zagat data.

   (a) What is the estimated standard deviation of the error (the "standard error of the regression")? What is the interpretation of this value?

   Solution: The estimated error standard deviation is $s = 12.5559$. Using the empirical rule, the model says that approximately 95% of restaurants have prices within $2s = 25.11$ of the regression line.

   (b) 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 ($\sum(y_i - \bar{y})^2$).

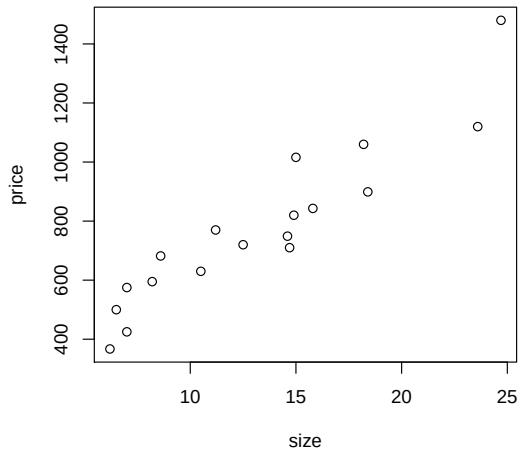
   (c) According to the estimated regression model, what is the range of typical prices for restaurants with quality ratings of 20?

   Solution: $37.84 \pm 25.11 = (12.73, 62.95)$

   (d) According to the estimated regression model, what is the range of typical prices for restaurants with quality ratings of 10?

   Solution: In the estimated regression model, when the quality rating is 10, the expected price is $-4.74 + 2.129(10) = 16.55$; the range of typical prices is $16.56 \pm 25.11 = (-8.5441.66)$. Since price can't be negative, we could just as well report the range as $(0, 41.66)$. Note that since $x = 10$ is at the edge of the range of the data, the values predicted by the model are not very reliable.

3. Here is a scatterplot of the sizes (in 100 ft^2) and prices (in \$1000) for $n = 18$ apartments in the Village.



Here is the Minitab output for the least squares regression fit to the housing data. Some of the entries have been redacted (replaced by question marks).

```
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      44.95     4.37    10.29    0.000  1.00

Regression Equation

Price = 182.3 + 44.95 Size
```

(a) In the fitted regression model, what is the slope? What is the interpretation of this value?

> **Solution:** The slope is 44.95. For every one unit (100 sq. ft) increase in apartment size, expected price increases by 44.95 units (44.95 $\times$ \$1000).

(b) In the fitted regression model, what is the intercept? Does this value have a direct interpretation? If so, what is it?

> **Solution:** The intercept is 182.3. There is no direct interpretation of this value since Size $= 0$ is outside the range of the data.

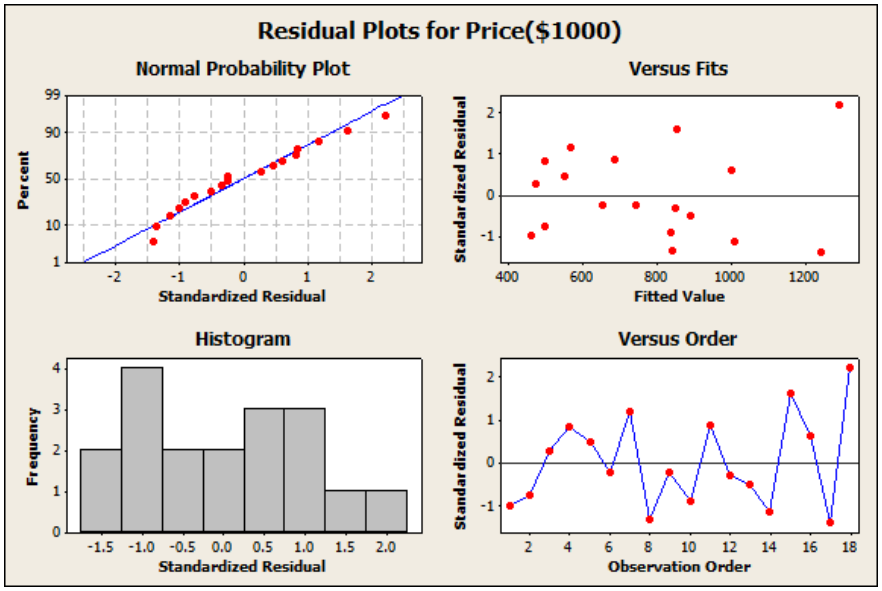
(c) Explain the meaning of the non-redacted values in the "Model Summary" parts of the output.

Solution: The standard error of the regression, $s = 101.375$ is the standard deviation of the regression error in the fitted model. Roughly 95% of the data points should have y values within $2s$ of the regression line.

The proportion of the variability in price explained by the regression model is $R^2 = 86.87\%$. The regression model explains a large proportion of the variability in the response (price).

Model assumptions

4. 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.