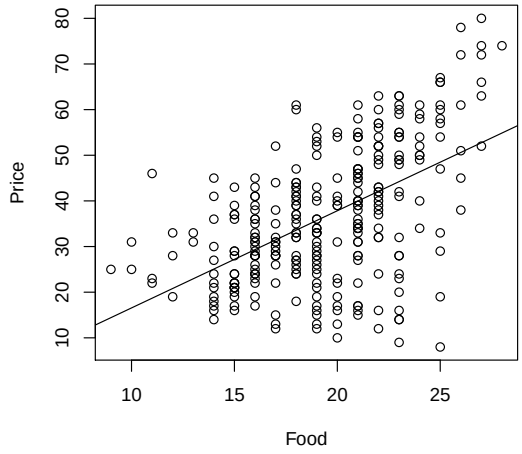
Regression Inference (Solutions)
STAT-UB.0003: Regression and Forecasting Models

Inference

1. Recall the restaurant data: 294 New York City restaurant's from the 2003 Zagat guide. Here is a scatterplot of the data, along with the least squares regression fit:



Here is the Minitab regression output:

```
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 is a reasonable population to go along with this sample?

> **Solution:** All restaurants in New York in 2003; all restaurants in 2003; all restaurants in 2003 and nearby years.

(b) What is the difference between the true regression parameters (β_0 and β_1) and the regression estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$)?

> **Solution:** The true regression parameters apply to the entire population. The regression estimates apply only to the sample.

(c) Construct a 95% confidence interval for β_1, the coefficient of "Food".

Solution: We use

$$\hat{\beta}_1 \pm t_{\alpha/2} SE(\hat{\beta}_1),$$

where $\alpha = .05$ and we have $n - 2 = 292$ degrees of freedom, so $t_{\alpha/2} \approx z_{\alpha/2}$. Using the approximation $z_{.025} \approx 2$, this gives

$$2.1288 \pm 2(0.2001) = 1.1288 \pm 0.4002,$$

or $(1.72, 2.53)$.

(d) What is the meaning of the confidence interval for β_1?

Solution: We are 95% confident that if we increase Food quality by 1 point, then mean dinner price will increase by an amount between \$1.72 and \$2.53.

(e) What is the meaning of a 95% confidence interval for β_0? Is this useful for the restaurant example?

Solution: This would be a confidence interval for the mean dinner price at restaurants with food quality ratings of 0. This is nonsensical (no restaurants have food qualities of 0), and thus not useful.

(f) Perform a hypothesis test at level 5% of whether or not there is a linear relationship between Price and Food.

Solution: We are interested in the following null and alternative hypotheses:

$$H_0 : \beta_1 = 0 \quad \text{(no linear relationship)}$$
$$H_a : \beta_1 \neq 0 \quad \text{(linear relationship)}$$

Based on the minitab output, the p-value for this test is below 0.001. Thus, we reject the null hypothesis at level 5%. There is a statistically significant linear relationship between Price and Food.

We can also do this problem using a rejection region. We reject H_0 at level α if $|T| > t_{\alpha/2}$, where

$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{2.1288}{0.2001} = 10.64$$

For level $\alpha = .05$, we have $t_{\alpha/2} \approx 2$ Since $|T| > 2$, we reject H_0.

2. We used the prices and sizes of 18 apartments in Greenwich Village to fit the model

$$\text{Price} = \beta_0 + \beta_1 \, \text{Size} + \varepsilon,$$

where price is measured in units of \$1000 and size is measured in units of 100 ft^2.

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
101.375  86.87%     86.05%      81.13%

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) What is a reasonable population for this sample?

> **Solution:** The prices and sizes of all apartments in Greenwich Village.

(b) Construct a 95% confidence interval for β_1.

> **Solution:** We use
>
> $$\hat{\beta}_1 \pm t_{\alpha/2} \text{SE}(\hat{\beta}_1),$$
>
> where $\alpha = .05$ and we have $n - 2 = 16$ degrees of freedom. This gives
>
> $$44.95 \pm 2.120(4.37) = 44.95 \pm 9.26,$$
>
> or $(35.69, 54.21)$.

(c) What is the meaning of the confidence interval for β_1?

> **Solution:** We are 95% confident that if we increase size by 100 square feet, then mean price will increase by an amount between \$35.7K and \$54.2K.

(d) What is the meaning of a 95% confidence interval for β_0? In the context of the housing data, is this useful?

> **Solution:** This would be a confidence interval for the mean price of apartments with size 0. This is nonsensical (no apartments have size 0), and thus not useful.

(e) Perform a hypothesis test at level 5% of whether or not there is a linear relationship between Size and Price.

Solution: We are interested in the following null and alternative hypotheses:

$$H_0 : \beta_1 = 0 \quad \text{(no linear relationship)}$$
$$H_a : \beta_1 \neq 0 \quad \text{(linear relationship)}$$

Based on the minitab output, the p-value for this test is below 0.001. Thus, we reject the null hypothesis at level 5%. There is a statistically significant linear relationship between size and mean price.

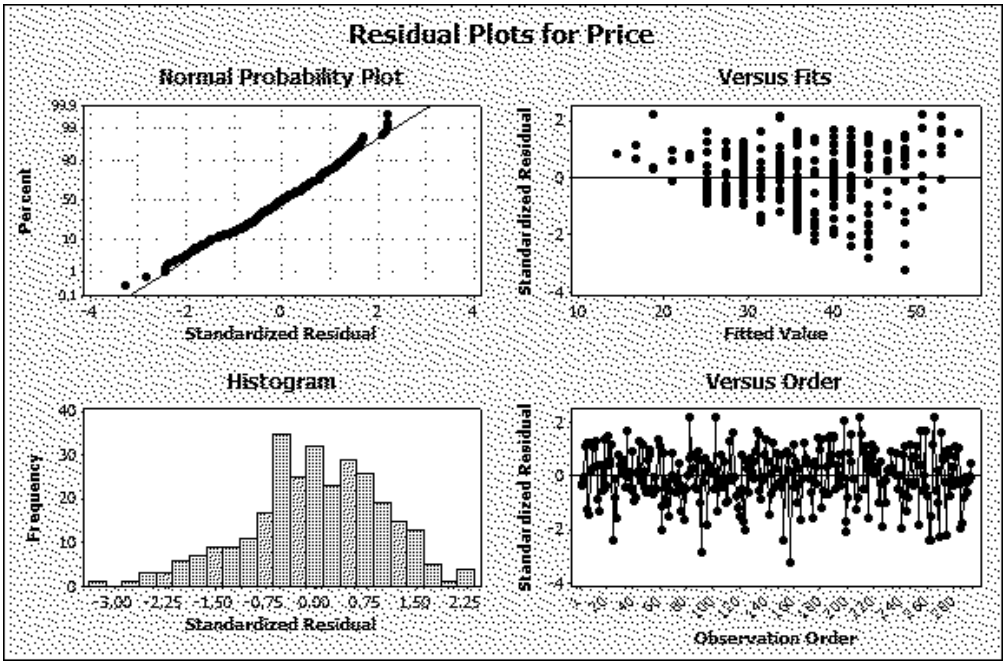
We can also do this problem using a rejection region. We reject H_0 at level α if $|T| > t_{\alpha/2}$, where

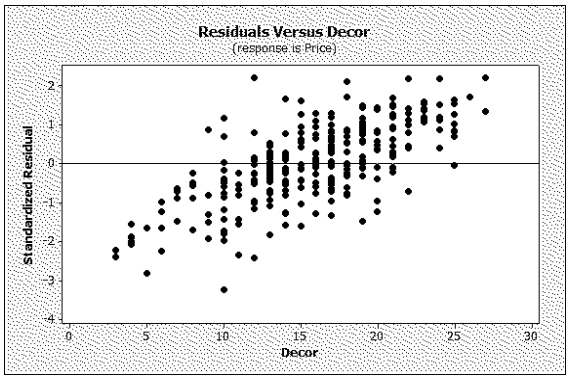
$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{44.95}{4.37} = 10.286$$

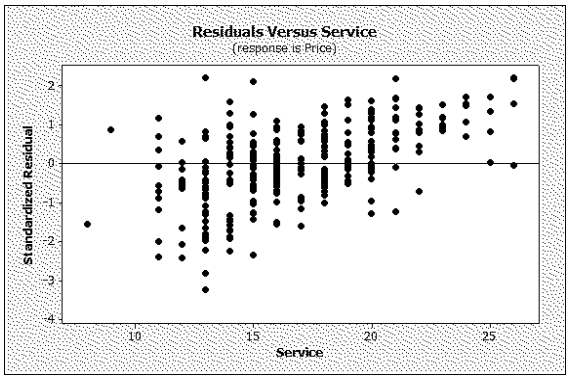
For level $\alpha = .05$, we have $t_{\alpha/2} = t_{.025} = 2.120$ (using $n - 2 = 16$ degrees of freedom). Since $|T| > 2.120$, we reject H_0.

Model assumptions

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.