Introduction to Linear Regression – Solutions
COR1-GB.1305 – Statistics and Data Analysis

Linear Regression

1. In the following scenarios, which would you consider to be predictor (x) and which would you consider to be response (y)?
   (a) Sales revenue; Advertising expenditures
   (b) Starting salary after college; Undergraduate GPA
   (c) The current month's sales; the previous month's sales
   (d) The size of an apartment; the sale price of an apartment.
   (e) A restaurant's Zagat Price rating; a restaurant's Zagat Food rating.

   Solution: This is a little bit subjective, but the following answers make sense: (a) $y =$ sales revenue; (b) $y =$ starting salary; (c) $y =$ current sales; (d) $y =$ sale price; (e) either makes sense.

2. Let y be the payment (in dollars) for a repair which takes x hours. Suppose that

$$y = 25 + 30x.$$

   What is the interpretation of this model?

   Solution: There is a positive linear relationship between y and x. Increasing repair time by one hour increases payment by \$30. There is no interpretation for the intercept since repair time is always positive.

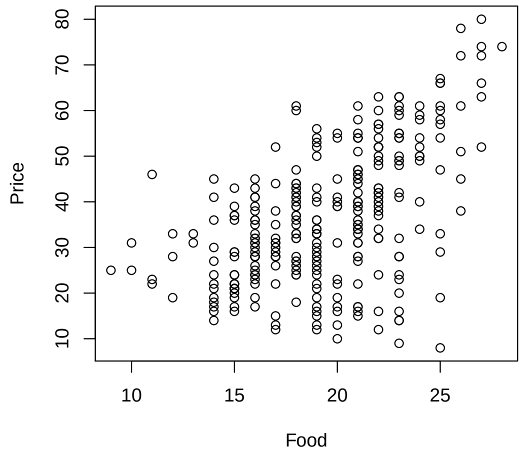
3. Consider two variables measured on 294 restaurants in the 2003 Zagat guide:

$$y = \text{typical dinner price, including one drink and tip (\$)}$$
$$x = \text{Zagat quality rating (0–30).}$$

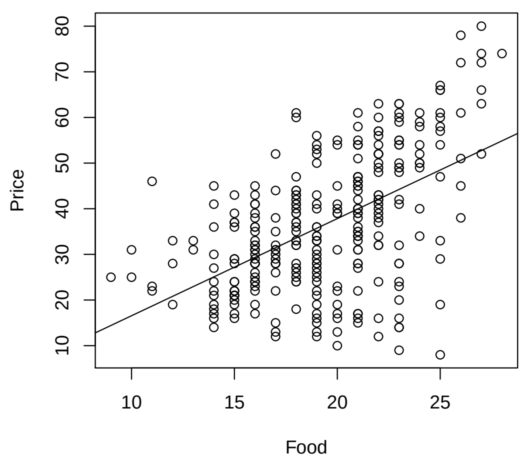
   Here is a scatterplot of y on x:

   Why is an exact linear relationship inappropriate to describe the relationship between y and x?

   Solution: There are no values β_0 and β_1 such that $y = \beta_0 + \beta_1 x$ for all restaurants; no straight line fits the data perfectly.



4. 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).

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

   (a) What is the estimated standard deviation or the error? 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) According to the estimated regression model, what is the range of typical prices for restaurants with quality ratings of 20?

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

   (c) 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 $16.55 \pm 25.11 = (-8.5441.66)$. 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.

The Analysis of Variance Table

6. When we fit the regression model to the Zagat data with response "Price" and predictor "Food", we get the following "Analysis of Variance" table:

```
Analysis of Variance

Source          DF  Adj SS   Adj MS  F-Value  P-Value
Regression       1   17838  17838.4   113.15    0.000
  Food           1   17838  17838.4   113.15    0.000
Error          292   46034    157.7
  Lack-of-Fit   18    5394    299.7     2.02    0.009
  Pure Error   274   40640    148.3
Total          293   63873
```

(a) Find the SSE. Explain how this value is computed.

> **Solution:** We read the sum of squares of the residual errors, SSE, from the `Adj SS` column of the `Error` row: 46034. This quantity is equal to the sum of squares of the residual errors: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

(b) Find the SSR. Explain how this value is computed.

> **Solution:** We read the sum of squares of the regression, SSR, from the `Adj SS` column of the `Regression` row: 17838. This quantity is equal to $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, where $\bar{y}$ is the average value of y_i, i.e. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

(c) Find the SST. Explain how this value is computed.

> **Solution:** We read the total sum of squares, SST, from the `Adj SS` column of the `Total` row: 63873. This quantity is equal to $\sum_{i=1}^{n}(y_i - \bar{y})^2$. Note also that SST $=$ SSR $+$ SST.

(d) Explain how to compute R^2 from the ANOVA table.

> **Solution:** The coefficient of determination, R^2, is equal to the proportion of the variability in the response that is explained by the regression:
>
> $$R^2 = \frac{\text{SSR}}{\text{SST}}.$$

(e) Explain how to compute s from the ANOVA table.

Solution: The standard error of the regression is given by

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

where k is the number of predictors in the model ($k = 1$ for simple linear regression. This is also equal to $\sqrt{\text{MSE}}$, where MSE is given in the `Adj MS` column of the `Error` row. This is an estimate of the standard deviation of the regression error.