Introduction to Linear Regression – Solutions

STAT-UB.0103 – Statistics for Business Control and Regression Models

Covariance (Review)

1. Suppose X and Y are random variables with $\text{var}(X) = 4$, $\text{var}(Y) = 3$, and $\text{cov}(X, Y) = -2$.
   (a) Find $\text{cov}(X, 2X + Y)$.

   Solution:

$$\begin{aligned}
\text{cov}(X, 2X + Y) &= 2\,\text{cov}(X, X) + \text{cov}(X, Y) \\
&= 2\,\text{var}(X) + \text{cov}(X, Y) \\
&= 2(4) + (-2) \\
&= 6.
\end{aligned}$$

   (b) Find $\text{cov}(4X - Y, Y)$.

   Solution:

$$\begin{aligned}
\text{cov}(4X - Y, Y) &= 4\,\text{cov}(X, Y) - \text{cov}(Y, Y) \\
&= 4\,\text{cov}(X, Y) - \text{var}(Y) \\
&= 4(-1) - (3) \\
&= -7
\end{aligned}$$

   (c) Find $\text{cov}(2X + Y, X + 3Y)$.

   Solution:

$$\begin{aligned}
\text{cov}(2X + Y, X + 3Y) &= 2\,\text{cov}(X, X + 3Y) + \text{cov}(Y, X + 3Y) \\
&= 2\big(\,\text{cov}(X, X) + 3\,\text{cov}(X, Y)\big) + \big(\,\text{cov}(Y, X) + 3\,\text{cov}(Y, Y)\big) \\
&= 2\,\text{cov}(X, X) + 7\,\text{cov}(X, Y) + 3\,\text{cov}(Y, Y) \\
&= 2\,\text{var}(X) + 7\,\text{cov}(X, Y) + 3\,\text{var}(Y) \\
&= 2(4) + 7(-2) + 3(3) \\
&= 3.
\end{aligned}$$

Portfolio Optimization (Review)

2. Suppose there are two stocks, X and Y. The annual returns for these stocks can be modeled as *dependent* random variables with correlation ρ. Suppose that the expected returns for the two stocks are both equal to 5%, but the standard deviations for the two stocks are different. Specifically, the standard deviations for stocks X and Y are 1% and 2%, respectively. Consider two strategies for investing \$100 between the two stocks:

$$\begin{aligned} A &= 100X, \\ B &= 50X + 50Y. \end{aligned}$$

When does strategy B have more risk than strategy A?

Solution: We have that $\sigma_X = .01$ and $\sigma_Y = .02$. The variance of the gain from strategy A

$$\begin{aligned} \text{var}(A) &= \text{var}(100X) \\ &= 100^2 \,\text{var}(X) \\ &= 100^2(.01)^2 \\ &= 1. \end{aligned}$$

The variance of the gain from strategy B is

$$\begin{aligned} \text{var}(B) &= \text{var}(50X + 50Y) \\ &= 50^2 \,\text{var}(X) + 2(50)(50)\,\text{cov}(X,Y) + 50^2 \,\text{var}(Y) \\ &= (50)^2(.01)^2 + 2(50)^2(.01)(.02)\rho + (50)^2(.02)^2 \\ &= .25(5 + 4\rho). \end{aligned}$$

Thus, strategy B has more risk than strategy A when

$$\begin{aligned} \text{var}(B) &> \text{var}(A) \\ .25(5 + 4\rho) &> 1 \\ \rho &> -.25. \end{aligned}$$

Linear Regression

3. 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.

4. 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.

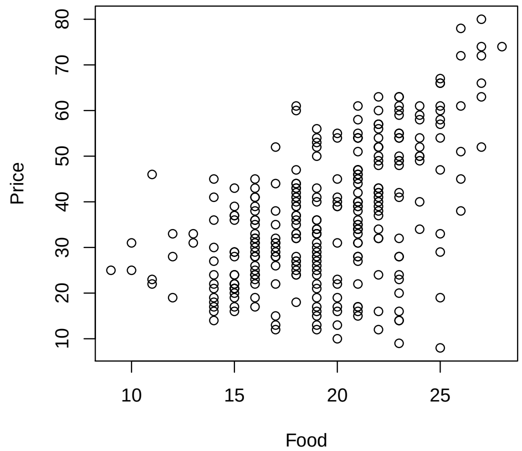
5. Consider two variables measured on 294 restaurants in the 2003 Zagat guide:

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

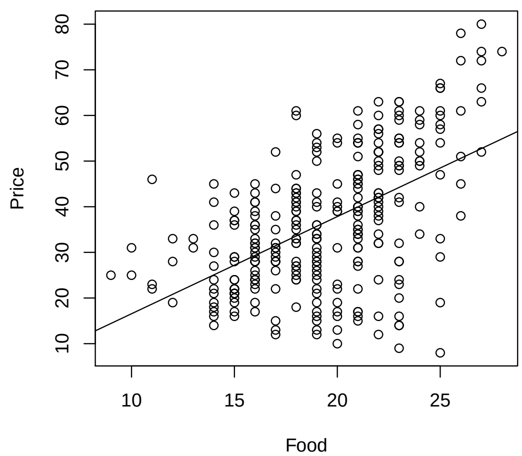
   Here is a scatterplot of y on x:

   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.



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



Here is the Minitab output from the fit:

```
The regression equation is
Price = - 4.74 + 2.13 Food


Predictor     Coef  SE Coef      T      P
Constant    -4.736    3.950  -1.20  0.232
Food        2.1288   0.2001  10.64  0.000


S = 12.5559   R-Sq = 27.9%   R-Sq(adj) = 27.7%
```

(a) What are the estimated intercept and slope?

Solution: The estimated intercept is $\hat{\beta}_0 = -4.74$; the estimated slope is $\hat{\beta}_1 = 2.13$.

(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 $\widehat{\text{Price}} = -4.74 + 2.13(20) = 37.86$.

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

Solution: For every 1-point incrase in food quality, the expected dinner price goes up by $2.13.

(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).

7. 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:** $37.86 \pm 25.11 = (12.75, 62.97)$

   (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 $-4.74 + 2.13 * 10 = 16.56$; the range of typical prices is $16.56 \pm 25.11 = (-8.5541.67)$. Since price can't be negative, we could just as well report the range as $(0, 41.67)$. 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.