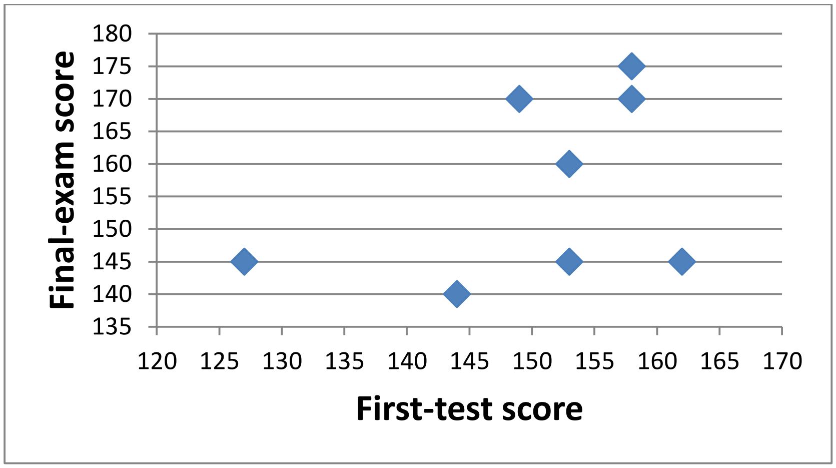
STATISTICS FOR MANAGEMENT (IDS 570)

HOMEWORK 2 SOLUTION

Problem 1

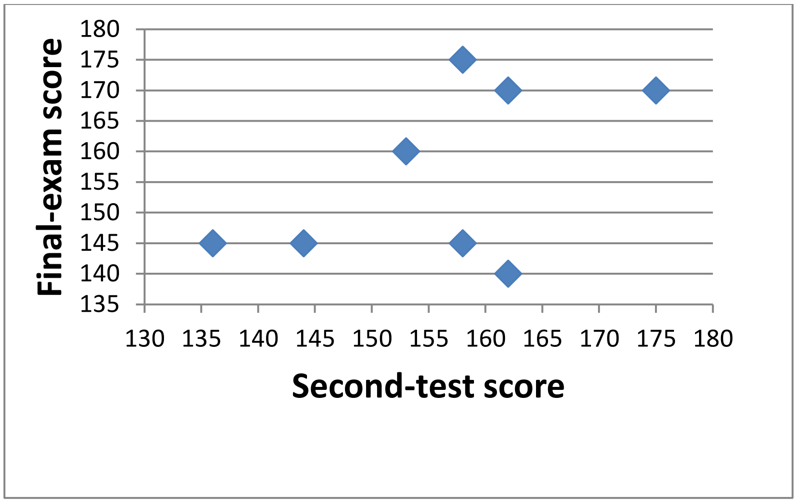
a) The explanatory variable should be the first exam score since it should help explain the variation in the final exam score.

b) r = $\frac{cov(x,y)}{s_x s_y} = 0.408803$



c) The emphasis of the final test is typically quite different with the first exam and the students may not take the first test as serious as the final exam. Therefore, we may not see a very strong relationship between the first-test score and final exam score.

d) A student who scores well on the second exam would demonstrate an understanding of statistical concepts and an ability to think analytically. This type of students would also probably score well on the final exam (and vice versa).

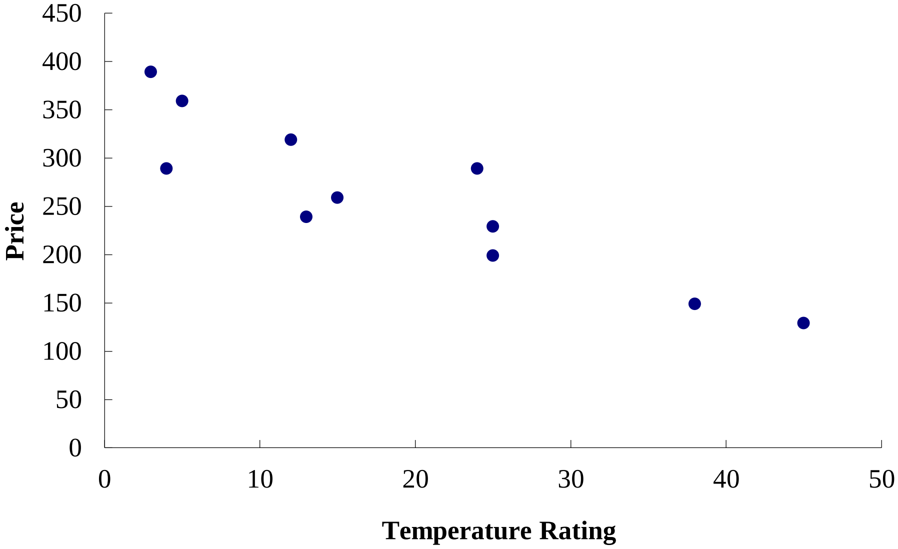


e) r = 0.519

f) Answer may vary. Some possibilities may include additional factors, such as prior knowledge, that might also affect exam 1. By the second exam, students have also figured out the level of studying and effort that they intend to apply to a particular course, and this would likely flow through to the final exam.

Problem 2

a)



b) The scatter diagram and the slope of the estimated regression equation indicate a negative linear relationship between *x* = temperature rating and *y*

= price. Thus, it appears that sleeping bags with a lower temperature rating cost more than sleeping bags with a higher temperature rating. In other words, it costs more to stay warmer.

c) $\bar{x} = \frac{\sum x_i}{n} = \frac{209}{11} = 19, \;\; \bar{y} = \frac{\sum y_i}{n} = \frac{2849}{11} = 259$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = -10,090 \qquad \Sigma(x_i - \bar{x})^2 = 1912$$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{-10,090}{1912} = -5.2772$$

$$b_0 = \bar{y} - b_1\bar{x} = 259 - (-5.2772)(19) = 359.2668$$

$$\hat{y} = 359.2668 - 5.2772x$$

d) $\hat{y} = 359.2668 - 5.2772x$

Thus, the estimate of the price of sleeping bag with a temperature rating of 20, for example, is approximately $254.

Problem 3

a)
$\bar{y} = 74$, SSE = $\sum(y_i - \hat{y}_i)^2$ = 173.88
The total sum of squares is SST = $\sum(y_i - \bar{y})^2 = 756$.

In addition, $\sum(\hat{y}_i - \bar{y})^2$ = 582.12. Therefore, we have r^2 = 582.12/756 = .77

b) The estimated regression equation provided a good fit because 77% of the variability in *y* has been explained by the least squares line.

c) $r = \sqrt{.77} = +.88.$

Problem 4

a) and b)

Let $e_i = y_i - \hat{y}_i$ is the error we make in predicting y from x. In the least squares model, we try to minimize the sum of the squared error of prediction (i.e. the e_i values) across all cases. Mathematically, this quantity can be expressed as

$$SSE = \sum_{i=1}^{n} e_i^2$$

Specifically, we want to find the values of b_0 and b_1 that minimizes the above quantity.

So, how do we do this? The key is to think back to differential calculus and remember how one goes about finding the minimum value of a mathematical function. This involves taking the derivative of that function. As you may recall, if y is some mathematical function of variable x, the derivative of y with respect to x is the amount of change in y that occurs with a tiny change in x. Roughly, it's the instantaneous rate of change in y with respect to changes in x. So, what does this have to do with the minimum of a mathematical function? Well, the derivative of function y with respect to x – the extent to which y changes with a tiny change in x – equals zero when y is at its minimum value. If we find the value of x for which the derivative of y equals zero, then we have found the value of x for which y is neither increasing nor decreasing with respect to x.

Thus, if we want to find the values of b_0 and b_1 that minimize SSE, we need to express SSE in terms of b_0 and b_1, take the derivatives of SSE with respect to b_0 and b_1, set these derivatives to zero, and solve for b_0 and b_1.

By substituting $e_i = y_i - \hat{y}_i$ in the least squares regression model we get $SSE = \sum_{i=1}^{n}(y_i - b_0 - b_1x_i)^2$. The partial derivative of SSE with respect to b_0 and b_1 is

$$\frac{\partial SSE}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1x_i)$$

and

$$\frac{\partial SSE}{\partial b_1} = -2\sum_{i=1}^{n}x_i(y_i - b_0 - b_1x_i)$$

The next step is set each one of these equations to zero:

$$-2\sum_{i=1}^{n}(y_i - b_0 - b_1x_i) = 0 \qquad (1)$$

And

$$-2\sum_{i=1}^{n}x_i(y_i - b_0 - b_1x_i) = 0 \qquad (2)$$

Equations (1) and (2) form a system of equations with two unknowns - b_0 and b_1. By solving these two equations we get

$$-nb_0 + n\bar{y} - nb_1\bar{x} = 0 \Rightarrow b_0 = \bar{y} - b_1\bar{x}$$

And

$$b_1\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_iy_i - b_0(n\bar{x})$$

By substituting $b_0 = \bar{y} - b_1\bar{x}$, in the above expression we get

$$b_1\left(\sum_{i=1}^{n}x_i^2 + n\bar{x}^2\right) = \sum_{i=1}^{n}x_iy_i - n\bar{x}\,\bar{y}$$

Therefore,

$$b_1 = \frac{\sum x_iy_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}} = \frac{\sum x_iy_i - n\bar{x}\bar{y} + n\bar{x}\bar{y} - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x} + n\bar{x} - n\bar{x}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

This indicates that

$$b_1 = r\frac{s_y}{s_x}$$

If we compute $\hat{y} = b_0 + b_1x$ at $x = \bar{x}$, we get $\hat{y} = b_0 + b_1\bar{x} = \bar{y} - b_1\bar{x} + b_1\bar{x} = \bar{y}$. This indicates that the line $\hat{y} = b_0 + b_1x$ passes through the point $(\bar{x}, \bar{y})$.

c) By substituting the b_0 and b_1 that we found in part (a) and (b) in the least squares regression line $\hat{y} = b_0 + b_1x$, we get $\hat{y} = \bar{y} - b_1(x - \bar{x}) = \bar{y} + r\,\frac{s_y}{s_x}(x - \bar{x})$. This equation can be written as

$$\hat{y} - \bar{y} = r\frac{s_y}{s_x}(x - \bar{x}) \Leftrightarrow \frac{\hat{y} - \bar{y}}{s_y} = r\,\frac{(x - \bar{x})}{s_x}$$

d) Suppose x denotes the statistics score variable. Therefore, for the student who has scored 1.5 SD above the mean we have $x = \bar{x} + 1.5\,s_x$. The least squares regression estimate for this student's final exam would be $\hat{y} = \bar{y} + r\frac{s_y}{s_x}(\bar{x} + 1.5\,s_x - \bar{x}) = \bar{y} + 1.5rs_y$. Since $r < 1$ and $s_y > 0$, we have $\hat{y} < \bar{y} + 1.5\,s_y$. Therefore, the regression estimate is fewer than 1.5 SDs above average. For $x = \bar{x} - 1.5\, s_x$, the estimate would be $\hat{y} = \bar{y} + r\frac{s_y}{s_x}(\bar{x} - 1.5\, s_x - \bar{x}) = \bar{y} - 1.5 r s_y$. Since $r < 1$ and $s_y > 0$, we have $\hat{y} > \bar{y} - 1.5\, s_y$. Therefore, the regression estimate is above than 1.5 SDs below average.

Problem 5

a) $\bar{\hat{y}} = \frac{1}{n}\sum_i \hat{y}_i = \frac{1}{n}\sum_i (b_0 + b_1 x_i) = b_0 + b_1\bar{x} = \bar{y}$.

b) $s_{\hat{y}}^2 = \frac{1}{n-1}\sum_i \left(\hat{y}_i - \bar{\hat{y}}\right)^2 = \frac{1}{n-1}\sum_i (\hat{y}_i - \bar{y})^2 = \frac{1}{n-1} r^2 s_y^2 \frac{\sum_i (x_i - \bar{x})^2}{s_x} = r^2 s_y^2$.

$r = 0$ means there is no linear relationship between x and y and the least square regression line would be a horizontal line (i.e. $b_1 = 0, \hat{y} = b_0 = \bar{y}$). Therefore, $\hat{y}$ is a constant for all values of x and there is no deviation from its mean $\bar{\hat{y}}$. Hence $s_{\hat{y}}^2 = 0$. When $r = 1$, the relationship between x and y is perfect and all the data points lie in the least square regression line. Therefore, $\hat{y}_i = y_i$ for all $i = 1, \dots, n$ and $s_{\hat{y}}^2 = s_y^2$.

Problem 6

a) $\bar{e} = \frac{1}{n}\sum_i (y_i - \hat{y}_i) = \frac{1}{n}\sum_i (y_i - b_0 - b_1 x_i) = \bar{y} - b_0 - b_1\bar{x} = \bar{y} - (\bar{y} - -b_1\bar{x}) - b_1\bar{x} = 0$

b) $s_e^2 = \frac{1}{n-1}\sum_i (e_i - \bar{e}_i)^2 = \frac{1}{n-1}\sum_i e_i^2 = \frac{1}{n-1}\sum_i (y_i - \hat{y}_i)^2$. We have $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$. We square each side of the equation- because the sum of deviations about the mean is equal to zero- and sum the result over all n points. Therefore $\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$. Hence $s_e^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2 - \frac{1}{n-1}\sum_i (\hat{y}_i - \bar{y})^2 = s_y^2 - r^2 s_y^2 = (1 - r^2) s_y^2$.

When r is zero, there is no linear relationship between x and y and $\hat{y} = \bar{y}$ is horizontal line. In this case, we have $e_i = y_i - \hat{y}_i = y_i - \bar{y}$. Therefore, $s_e^2 = \frac{1}{n-1}\sum_i e_i^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2 = s_y^2$. When $r = 1$, all the data points are perfectly lied on the regression line and $\hat{y}_i = y_i$. Therefore, $e_i = 0$ for all $i = 1, \dots, n$. This indicates that $s_e^2 = 0$.

c) $s_{xe} = \frac{1}{n}\sum_i (x_i - \bar{x})(e_i - \bar{e}) = \frac{1}{n}\sum_i (x_i - \bar{x}) e_i = \frac{1}{n}\sum_i x_i e_i - \bar{x}\,\bar{e} = \frac{1}{n}\sum_i x_i e_i = \frac{1}{n}\sum_i x_i (y_i - b_0 - b_1 x_i)$. By replacing $b_0 = \bar{y} - b_1\bar{x}$ and $b_1 =$

$r\frac{s_y}{s_x}$ in the last equation, we get $s_{xe} = 0$. This indicates that x and e are uncorrelated.