Review Questions -- Chapter 12 -- The Regression Line
Statistics 1040 -- Dr. McGahagan

* Problem 1. Find the regression equation -- \( E \{ \text{Final} \mid \text{Midterm} \} = a + b \times \text{Midterm} \)

- Average score on midterm = 70; SD on midterm = 10
- Average score on final = 55; SD on final = 20
- Correlation coefficient = 0.6

Slope of regression line = \( \rho \times \frac{SD_y}{SD_x} = 0.6 \times \frac{20}{10} = 1.2 \)

Since regression line passes through the means, we have:
- Final = \( a + b \times \text{Midterm} \)
- Mean(final) = \( a + 1.2 \times \text{Mean(midterm)} \)
- 55 = \( a + 1.2 \times 70 \)
- \( a = 55 - 84 = -29 \)

and the regression equation is: \( E \{ \text{Final} \mid \text{Midterm} \} = -29 + 1.2 \times \text{Midterm} \)

Those who received a score of 100 on the midterm would be on average expected to get a score of
- \( E \{ \text{Final} \mid \text{Midterm} = 100 \} = -29 + 1.2 \times 100 = 91 \).

RMSE for this equation will be \( (\sqrt{1 - 0.36}) \times \text{SDincome} = (\sqrt{0.6}) \times 14,400 \approx 14,109 \), so one should not be very surprised if those who received 100 got a score of 91 - 16 = 75 or 91 + 16 = 107.
Note that if there is an upper bound of 100 to the scores, it would be dangerous to assume normality of the residuals.

* Problem 2. Find the regression equation -- \( E \{ \text{Income} \mid \text{Height} \} = a + b \times \text{Height} \)

- Average height = 70 inches; SD of height = 3 inches
- Average income = $29,800; SD of income = $14,400
- Correlation coefficient = 0.20 [not very impressive; try (scatter .2) to see this]

Slope of regression line = \( \rho \times \frac{SD_y}{SD_x} = 0.2 \times \frac{14,400}{3} = 960 \)

\( E \{ \text{Income} \mid \text{Height} \} = a + 960 \times \text{Height} \)

Since this holds for the means of income and height;
- 29,800 = \( a + 960 \times 70 \)
- 29,800 = \( a + 67,200 \)
- \( a = -37,400 \)

and the regression equation is:
- \( E \{ \text{Income} \mid \text{Height} \} = -37,400 + 960 \times \text{Height} \)

We can predict, for example, that someone 3 feet high would have an income of
- -37,400 + 960 * 36 = -2,840

The slope coefficient could be read to say that every inch of height is associated with another $960 of income; in view of the absence of much plausible connection (were taller people better fed as kids so they grew taller, and are they richer now because their families were well off?), it would be dangerous to interpret this as causation.

The intercept coefficient doesn't mean much either: someone zero inches in height could be predicted to have an income of -37,400 a year.

The RMSE is \( (\sqrt{1 - 0.04}) \times \text{SDincome} = (\sqrt{0.96}) \times 14,400 = 14,109 \), so forecasts on the basis of this equation will have a margin of error of more than 14,000 a third of the time.
Problem 3. Interpreting regression equations.
On the basis of the HANES sample of men 18-24, we can derive the regression equation:
\[ E[\text{height} \mid \text{weight}] = 62.4 \text{ inches} + 0.047 \times \text{weight in lbs.} \]
The prediction that someone who weighs zero pounds is 62.4 inches tall (5 feet 2 inches) is fairly obviously ridiculous.
Does the slope have any real meaning either? If you gain 50 pounds, will you also gain 
\[ 0.047 \times 50 = 2.35 \text{ inches} \]
The slope can be interpreted to mean that on average, individuals who are 50 pounds heavier are also on average, a bit over two inches taller.

Problem 4. RMS error of a horizontal line.
From the graph, errors would seem to average 1 unit.
The line is not the regression line -- there is a weak but definite negative correlation in the graph; the regression line would be downward sloping. The line might be the average value of the Y variable.

* Problem 5. Household income regressions.
Given the regression equation: 
\[ E[\text{wife} \mid \text{husband}] = 12,000 + 0.125 \times \text{husband} \]
where wife = wife's income and husband = husband's income,
does it follow that:
\[ E[\text{husband} \mid \text{wife}] = -96,000 + 8 \times \text{wife} \]
NO. Note that the regression equation should not be written as:
Wife = 12,000 + 0.125 husband, for then the equation would follow as a matter of algebra (multiply through by 8 and rearrange terms)
It is the expected value of the wife's income that we are predicting, and it is NOT the case that this can be treated as a simple algebraic variable.

With the data given:
Avg. husband's income = 32,000; SD = 24,000
Avg. wife's income = 16,000; SD = 15,000; correlation = 0.20

\[ E[\text{husband} \mid \text{wife}] = a + 0.20 \times \frac{24000}{15000} \times \text{wife} = a + 0.32 \times \text{wife} \]
We use the fact that the line must pass through the means to find:
32000 = a + 0.32 * 16000 = a + 5120 or a = 32000 - 5120 = 26,880
\[ E[\text{husband} \mid \text{wife}] = 26,880 + 0.32 \times \text{wife} \]

\[ E[\text{wife} \mid \text{husband}] = a + 0.20 \times \frac{15000}{24000} \times \text{husband} = a + 0.125 \times \text{husband} \]
We use the fact that the line must pass through the means to find:
16000 = a + 0.125 * 32000 = a + 4000, so a = 16000 - 4000 = 12,000
\[ E[\text{wife} \mid \text{husband}] = 12,000 + 0.125 \times \text{husband} \]

Note that with the mistaken equation \[ E[\text{husband} \mid \text{wife}] = -96,000 + 8 \times \text{wife}, \]
when the wife's income is average, we have:
\[ E[\text{husband} \mid \text{wife} = 16,000] = -96,000 + 8 \times 12000 = -96,000 + 96,000 = 0 \text{ zero.} \]
Problem 6. Household income, continued.
I assume that the problem means that average incomes rise by 10 percent but SD does not change:
- Avg. husband's income = 35,200; SD = 24,000
- Avg. wife's income = 17,600; SD = 15,000; correlation = 0.20
the slope of the regression line doesn't change, since it is determined by correlation and SD:
\[ E[\text{wife} | \text{husband}] = a + (0.20 \times 15,000 / 24,000) \times \text{husband} \]
\[ = a + 0.125 \times \text{husband} \]
the intercept is found by substituting the mean values of wife's income and husband's income in the equation:
\[ 17,600 = a + 0.125 \times 35,200 = a + 4400 \]
hence \( a = 17,600 - 4400 = 13,200 \). The intercept term also goes up 10 percent,
and the regression equation is:
\[ E[\text{wife} | \text{husband}] = 13,200 + 0.125 \times \text{husband} \]

* Problem 7. Pizza, beer, and regressions
On average, students drink 4 beers per month with SD = 8 and eat 4 pizzas per month, with SD = 4.
There is a positive association between pizza and beer, and the regression equation is:
\[ E[\text{beer} | \text{pizza}] = 2 + \beta \times \text{pizza} \]
where \( \beta \) is an unknown slope coefficient. What is \( \beta \)? What is the correlation coefficient?
The regression line must pass through the means, so:
\[ 4 = 2 + \beta \times 4 \text{ or } \beta \times 4 = 4 - 2 = 2, \text{ so } \beta = 0.5 \]
We can now also find the correlation coefficient:
\[ 0.5 = \rho \times \text{SDbeer} / \text{SDpizza} = \rho \times 8 / 4 = 2 \times \rho \]
and hence \( \rho = 0.25 \)

Problem 8. IQ and lead levels, with weak positive association (lead makes you smarter???)
Any line can be used, and its RMS error calculated. It might be desirable to use the average IQ
to define a horizontal line, calculate the RMS error and compare it to the RMS error of the regression line.
The RMS error of the regression line is guaranteed to be smaller, but it may not be significantly smaller.
If not, the regression results are not meaningful -- and lead may not make you smarter after all.

* Problem 9. Hypothetical relation of IQ and income.
Data: Average income = $21,000; SD = $15,000
Average IQ = 100; SD = 15 Correlation = 0.50
\[ E[\text{IQ} | \text{income}] = a + (\rho \times \text{SD IQ} / \text{SD income}) \times \text{income} \]
\[ E[\text{IQ} | \text{income}] = 89.5 + (1/2000) \times \text{income} \]
With an income of 41,000, the expected value of IQ is 110

* Problem 10. Income and IQ.
The same data may be used to conclude that
\[ E[\text{Income} | \text{IQ}] = a + (\rho \times \text{SD income} / \text{SD IQ}) \times \text{IQ} \]
\[ E[\text{Income} | \text{IQ}] = -29000 + 500 \times \text{IQ} \]
With an IQ of 110, the expected value of income is $26,000.
Moral: conditional probabilities cannot be manipulated as simply as you might think.
See chapter 10, figure 9 on p. 175 of third edition for a graphical explanation.

Problem 11. Congressional report on income and educational attainment
Average income = $29,300; SD = ??
Average educational attainment = 13.1; SD = ?? Correlation = 0.37
Reported regression: \[ E[\text{education} | \text{income}] = 8.1 + 0.0000617 \times \text{income} \]
Unfortunately, the expected value of education at the average income is NOT the average education:
\[ E[\text{education} | \text{income}] = 8.1 + 0.0000617 \times 29300 = 8.1 + 1.8 = 9.9 \text{ years} \]
There is an error in the computation of the regression equation. We would need the SD values to find the
error.