S520 Homework 12

Enrique Areyan
April 27, 2012

15.7.#3: (X, Y) bivariate normal with parameters: $(\mu_x = 5, \mu_y = 3, \sigma_x^2 = 1, \sigma_y^2 = 4, \rho = 0.5)$

(a) $P(Y > 6) = 1 - pnorm(6, 3, \sqrt{4}) = 0.0668$

(b) $E(Y|X = 6.5) = \hat{y}(6.5) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(6.5 - \mu_x) = 3 + \frac{1}{2}\frac{2}{1}(6.5 - 5) = 3 + 1.5 = 4.5$

(c) $P(Y > 6|X = 6.5)$. We know the distribution of $Y|X = x \sim Normal(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), (1 - \rho^2)\sigma_y^2)$, which results in $Y|X = 6.5 \sim Normal(4.5, 3)$. Thus, $P(Y > 6|X = 6.5) = 1 - pnorm(6, 4.5, \sqrt{3}) = 0.1932$

15.7.#4: For this question, let X be the population of sister and Y the population of brothers. We first need to estimate the five parameters of the bivariate distribution., i.e., $(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$. I used the data from the accompanying website, and the function binorm.estimate to obtain $(64, 69, 6.6, 7.4, 0.5580547)$. Now we can answer the questions.

(a) First note that $5'10'' = 70$ inches. $P(Y \geq 70) \approx 1 - pnorm(70, 69, \sqrt{7.4}) = 0.3566$. So, there are about 35% brothers who are at least $5'10''$

(b) First note that $5'1'' = 61$ inches. We want to predict a value of Y given a value of X, i.e.,

$$E(Y|X = 61) = \bar{y} + r\frac{s_y}{s_x}(61 - \bar{x}) = 69 + 0.5580547\frac{\sqrt{7.4}}{\sqrt{6.6}}(61 - 64) = 67.2273$$

(c) $Y|X = x \sim Normal(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), (1 - \rho^2)\sigma_y^2)$. If $Y = 70$ and $X = 61$, then the distribution is $Normal(67.2273, 5.095455)$. We want to calculate $P(Y \geq 70|X = 61) = 1 - pnorm(70, 67.2273\sqrt{5.095455}) = 0.1097$. So, there are about 10% brothers who are at least $5'10''$ and whose sister is $5'1'$

15.7.#5: The quantities we care about are: $n = 11$, $\bar{x} = 64$, $\bar{y} = 69$, $s_x^2 = 6.6$, $s_y^2 = 7.4$ and $r = 0.5580547$. From this we can compute $SS_T = (n - 1)s_y^2 = 10 * 7.4 = 74$ and $r^2 = 0.5580547^2 = 0.311425$

ANOVA Table for simple linear regression:

Source	SS	df	MS	F	**p**
Regression	23.04545	1	23.04545	4.070472	1-pf(4.070472,1,9) = 0.07441683
Error	50.95455	9	5.661616		
Total	74	10			

(a) The sample coefficient of determination is just the square of the correlation coefficient r, which we calculate in the previous question. Therefore, $r^2 = 0.5580547^2 = 0.311425$

(b) To answer this question we can test the hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. All the necessary calculations are in the above ANOVA Table. We can conclude the following:

$$\mathbf{p} = 0.07441683 > \alpha = 0.05 \Longrightarrow \text{ fail to reject } H_0$$

Because we do not reject H_0, we cannot say that knowing a sister's height (x) helps one predict her brother's height (y).

(c) To construct a 0.90-level confidence interval for β_1, we first compute:

$$\hat{\beta_1} = r\frac{s_y}{s_x} = 0.5580547 \cdot \frac{\sqrt{7.4}}{\sqrt{6.6}} = 0.5909091$$

$g_t = qt(.95, df = 9) = 1.833113$, and

$$\frac{MS_E}{t_{xx}} = \frac{1 - r^2}{n - 2} \cdot \frac{s_y^2}{s_x^2} = \frac{1 - 0.5580547^2}{9} \cdot \frac{7.4}{6.6} = 0.08578206$$

The desired confidence interval is then:

$$\hat{\beta_1} \pm q_t * \sqrt{\frac{MS_E}{t_{xx}}} = 0.5909091 \pm 1.833113 * \sqrt{0.08578206} = (0.05401643, 1.127802)$$

(d) We can derive the length of the as follow:

$$L = 0.1 = \hat{\beta}_1 + q_t \cdot \sqrt{\frac{MS_E}{t_{xx}}} - \hat{\beta}_1 - q_t \cdot \sqrt{\frac{MS_E}{t_{xx}}} = 2 \cdot q_t \cdot \sqrt{\frac{MS_E}{t_{xx}}} \iff \frac{MS_E}{t_{xx}} = (\frac{0.1}{2 * q_t})^2$$

But we can replace the left hand side of the equation as follow:

$$\frac{1 - r^2}{n-2} \cdot \frac{s_y^2}{s_x^2} = (\frac{0.1}{2 * q_t})^2 \iff n - 2 = \frac{4q_t^2(1 - r^2)s_y^2}{0.1^2 s_x^2}$$

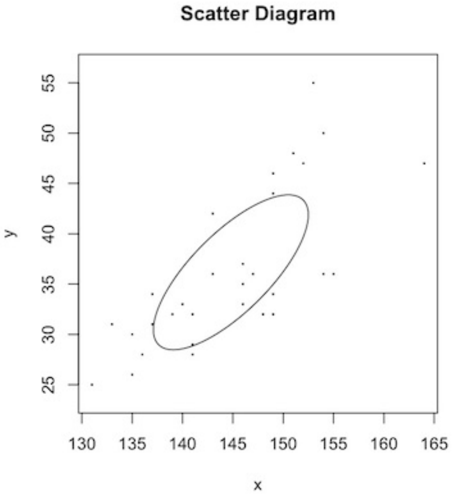
Notice that we want a 0.95-level confidence interval and thus, $q_t = qt(0.975, df = 9) = 2.262157$. Now we have all the information available:

$$n - 2 = \frac{4 * 2.262157^2 * (1 - 0.311425) * 7.4}{0.1^2 * 6.6} \Longrightarrow n = 1578.318$$

We should plan to observe close to 1578 pair of sister-brother.

15.7.#6: I used the data from the accompanying website to construct the answers for this question.

(a) Using the command binorm.scatter(Data), we obtain:



It does seem reasonable to assume that the sample was drawn from a bivariate normal distribution. An idea to see that this is the case, is to generate pseudo random bivariate samples and plot the resulting values. To this end, in R, assuming that in the variable $Data$ I have the measurements from the table 15.2, I ran the following commands several time: $binorm.scatter(binorm.sample(binorm.estimate(Data), 30))$. To keep things simple, I won't include the results here. But one can easily see that repeating this experiment one can obtain diagrams that are very similar to the one obtained with the original data. This is evidence supporting the claim that the data was drawn from a bivariate normal distribution.

(b) We want to estimate the following probability $P(37 \leq Y \leq 42)$. In order to do so, we must estimate the parameters of the Y distribution. We can use the function binorm.estimate, which will give use more information than necessary for this question, but which we still are going to use for the next question. So, $binorm.estimate(Data) = (144.8, 36.1666667, 59.2, 59.2471264, 0.7423653)$. Now, we assume that $Y \sim Normal(36.1666667, 59.2471264)$ and thus, $P(37 \leq Y \leq 42) = P(Y \leq 42) - P(Y \leq 37) =$ $pnorm(42, 36.1666667, \sqrt{59.2471264}) - pnorm(37, 36.1666667, \sqrt{59.2471264}) = 0.2326$. So, there are about 23% girls who weight between 37 and 42 kg.

(c) $P(37 \leq Y \leq 42|X = 150)$. Proceeding in a fashion similar to question 15.7.4 (c), we first need to compute the parameters of the distribution of $Y|X = x \sim Normal(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), (1 - \rho^2)\sigma_y^2)$. In this case, $x = 150$ and $\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = 36.1666667 + 0.7423653 * \frac{\sqrt{59.2471264}}{\sqrt{59.2}}(150 - 144.8) = 40.0285$ and $(1 - \rho^2)\sigma_y^2 = 26.59567$. Thus, $Y|X = x \sim Normal(40.0285, 26.59567)$. Now we can compute:
$P(37 \leq Y \leq 42|X = 150) = P(Y \leq 42|X = 150) - P(Y \leq 37|X = 150) =$
$pnorm(42, 40.0285, \sqrt{26.59567}) - pnorm(37, 40.0285, \sqrt{26.59567}) = 0.3704$. So, there are about 37% girls who weight between 37 and 42 kg and whose height is 150 cm.

15.7.#7:

(a) This is just $r^2 = 0.7423653^2 = 0.5511062$

(b) The quantities we care about are: $n = 30$, $\bar{x} = 144.8$, $\bar{y} = 36.1666667$, $s_x^2 = 59.2$, $s_y^2 = 59.2471264$ and $r = 0.7423653$. From this we can compute $SS_T = (n-1)s_y^2 = 29 * 59.2471264 = 1718.167$ and $r^2 = 0.5580547^2 = 0.311425$

ANOVA Table for simple linear regression:

Source	SS	df	MS	F	**p**
Regression	946.8925	1	946.8925	34.37557	1-pf(34.37557,1,28) = 2.646974e-06
Error	771.2744	28	27.54551		
Total	1718.167	29			

$\mathbf{p} = 2.646974e{-}06 < \alpha = 0.05 \Longrightarrow$ very much reject H_0, and so, there is strong evidence that x help us predict y

(c) To construct a 0.95-level confidence interval for β_1, we first compute:

$$\hat{\beta_1} = r\frac{s_y}{s_x} = 0.7423653 \cdot \frac{\sqrt{59.2471264}}{\sqrt{59.2}} = 0.7426607$$

$g_t = qt(.975, df = 28) = 2.048407$, and

$$\frac{MS_E}{t_{xx}} = \frac{1-r^2}{n-2} \cdot \frac{s_y^2}{s_x^2} = \frac{1-0.7423653^2}{28} \cdot \frac{59.2471264}{59.2} = 0.01604468$$

The desired confidence interval is then:

$$\hat{\beta_1} \pm q_t * \sqrt{\frac{MS_E}{t_{xx}}} = 0.7426607 \pm 2.048407 * \sqrt{0.01604468} = (0.4831939, 1.002127)$$

15.7.#8:

(a) Assuming that grades are drawn from a bivariate normal distribution, where X is the distribution of grades on the first exam and Y is the distribution for test 2, the appropriate parameters are, for $n = 33$: $(75, 64, 10, 12, 0.5)$.
We can calculate what is the probability of Jill's suggestion, i.e., the probability of getting at least 80 points on the second exam given a score of 80 in the first: $P(Y \geq 80|X = 80)$. The distribution of $Y|X = x \sim Normal(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), (1-\rho^2)\sigma_y^2)$, in this case each of the parameters can be computed as follow:

$$\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = 64 + 0.5 * 1.2(80 - 75) = 64 + 0.6 * 5 = 64 + 3 = 67$$

$$(1-\rho^2)\sigma_y^2 = (1 - 0.5^2) * 12 = 0.75 * 12 = 9$$

So, $P(Y \geq 80|X = 80) = 1 - pnorm(80, 67, \sqrt{9} = 3) = 7.343424e - 06$. This probability is very small and thus, unlikely that the score of Jill's second test was 80 or more.
Instead, let us predict her score on the second test given an 80 in the first score. In other words $E(Y|X = 80) = 67$ (we just calculated this same quantity above). One possible suggestion for a grade for Jill on the second test is 67, i.e., what we would expected her score was given that the bivariate normal assumption holds.

(b) Jack's suggestion omits an important detail, namely, that the correlation coefficient is not equal to 1. In other words, we do not have a perfect correlation between the variables and thus, we can not conclude that his score should be one standard deviation above the mean for Test 1. The only way this would be a valid statement is if $\rho = 0$. We are dealing with the regression effect here.
Instead, we can use Corolarry 15.1 to compute a grade for Jack. The Corollary states that if x lies z standard deviations above μ_x, then y lies ρz standard deviations above μ_y The prediction equation is $\hat{y}(x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x)$. We can find what value of x would make the equation to be such that $\hat{y} = 76$. Rewriting the equation:

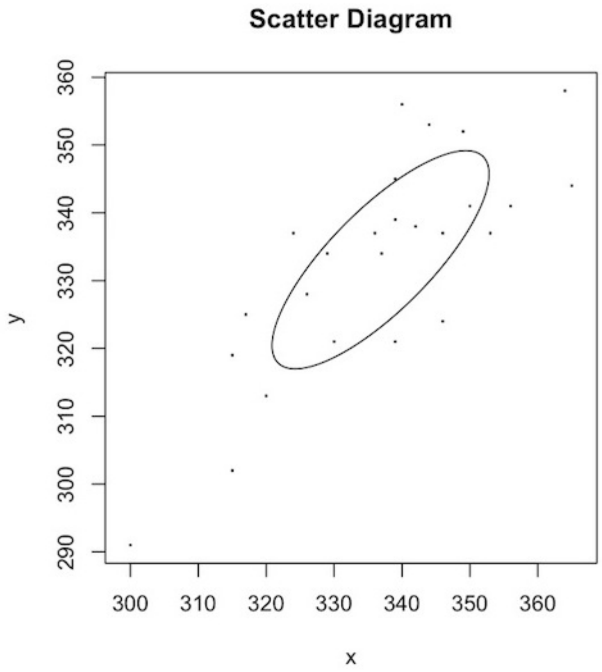
$$x = u_x + \frac{\sigma_x}{\sigma_y}\frac{\hat{y}(x) - \mu_y}{\rho} = 75 + \frac{10}{12}\frac{76 - 64}{0.5} = 75 + 0.83333 * 6 = 80$$

Another way to get this result is from Corollary 15.1. In this case, Jacks grade lies ρz stadndar deviations above the mean, i.e., $0.5 * 10 = 5$ points above mean $75 + 5 = 80$.

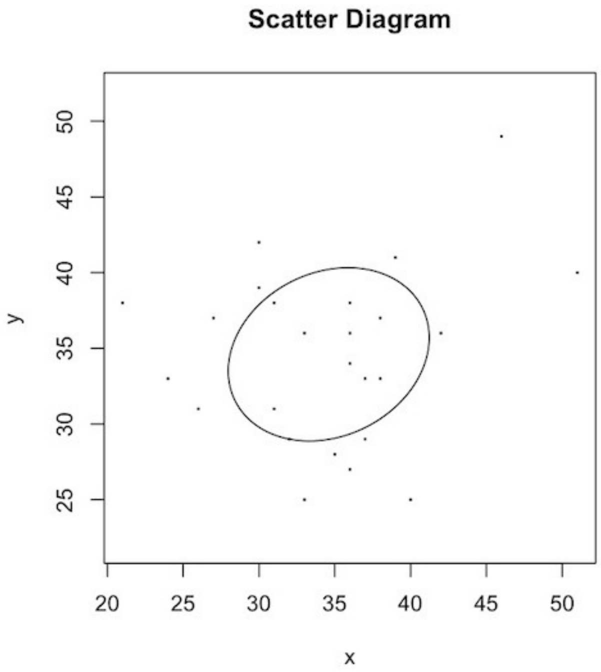
15.7.#11:

(a)

Size



Shape



Both samples appear to be drawn from a bivariate normal distribution, although this assumption seems slighter weaker for the head shapes and very strong for the head sizes data. An analysis of each one of the four samples (size1, size2, shape1, shape2) separately, i.e., qqnorm, boxplot and density graph, as well as IQR/stdev ratio support the evidence that each one was drawn from a normal distribution and thus, that the corresponding (X,Y) are bivariate normal (Note: I do not include all of these graphs here to keep things simple). The only detail worth mentioning are two outliers found in the head shapes data which may conflict with the normality assumption, but I would conclude that it is safe to continue the analysis assuming normality.

In the next two question, we will test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

(b) The data of interest is: $binorm.estimate(size) = (336.84, 333.08, 255.39, 258.91, 0.7860901)$. We can immediately conclude that the sample coefficient of determination is $r^2 = 0.7860901^2 = 0.6179376$. This is the proportion in the second son head size that is explained by variation in first son head size.

ANOVA Table for simple linear regression:

Source	SS	df	MS	F	**p**
Regression	3839.765	1	2374.075	37.19958	1-pf(37.19958, 1, 23) = 3.200658e-06
Error	2374.075	23	103.2207		
Total	6213.84	24			

For any standard alpha level, say $\alpha = 0.05$.

$\mathbf{p} = 3.200658e{-}06 < \alpha = 0.05 \Longrightarrow$ very much reject H_0, and so, there is strong evidence that x help us predict y

(c) The data of interest is: $binorm.estimate(shape) = (34.6, 34.6, 43.9166667, 32.75, 0.1922678)$. We can immediately conclude that the sample coefficient of determination is $r^2 = 0.1922678^2 = 0.03696691$. This is the proportion in the second son head shape that is explained by variation in first son head shape.

ANOVA Table for simple linear regression:

Source	SS	df	MS	F	**p**
Regression	29.05599	1	29.05599	0.8828762	1-pf(0.8828762 ,1, 23) = 0.3571791
Error	756.944	23	32.91061		
Total	786	24			

For any standard aplha level, say $\alpha = 0.05$

$$\mathbf{p} = 0.3571791 > \alpha = 0.05 \Longrightarrow \text{ fail to reject } H_0\text{, there is no evidence that x helps us predict y}$$

(d) length + breadth = 195 + 160 = 355. X denote first son, Y denotes the second. We want to compute,

$$E(Y|X = 355) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = 333.08 + 0.7860901\frac{\sqrt{258.91}}{\sqrt{255.39}}(355 - 336.84) = 347.4534$$

The guess is that the second adult son's head size is going to be approximately 347.4534.