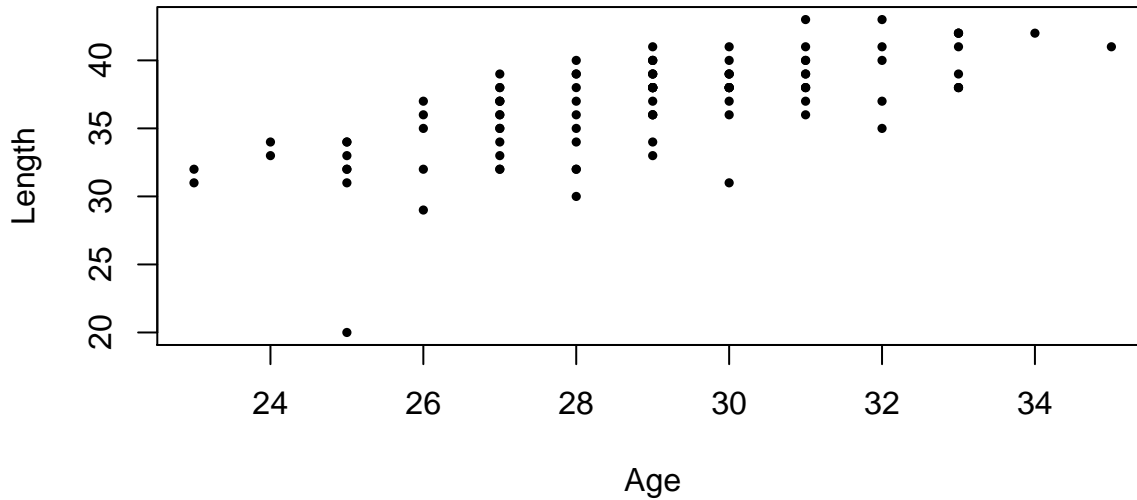


Researchers are interested in examining the relationship between gestational age (in weeks) and length (in cm) for newborn infants from a population of low-birth weight infants, defined to be infants weighing less than 1500 gm.

Here is a scatterplot of a sample of 100 infants from this population.



The R output of fitting a regression line to predict length from age is below.

Call:

```
lm(formula = Length ~ Age)
```

Residuals:

Min	1Q	Median	3Q	Max
-13.1183	-1.1183	0.2931	1.4100	4.1721

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.3282	3.0452	3.063	0.00283 **
Age	0.9516	0.1050	9.062	1.31e-14 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.648 on 98 degrees of freedom

Multiple R-Squared: 0.4559, Adjusted R-squared: 0.4504

F-statistic: 82.13 on 1 and 98 DF, p-value: 1.311e-14

- (a) What feature of the data exhibited in the scatter plot merits further investigation?

Solution: I was looking for students to notice the outlier, the unusually low observation at age 25 weeks. But the question was vague and I accepted many answers.

- (b) Circle the number closest to the correlation coefficient: -0.7 -0.2 0 0.4559 0.7 1.0 2.4.

Solution: Our eye should tell us that r is positive and fairly high, but not super close to 1. The R output indicates that $r^2 = 0.4559$, so $r = \sqrt{0.4559}$ which is about 0.7.

- (c) Write the regression line as a formula in the style of the example below.

$$(\text{height in inches}) = 30.25 + 0.25(\text{age in months})$$

Solution:

$$(\text{length in cm}) = 9.33 + 0.95(\text{gestational age in weeks})$$

- (d) Construct a 95% confidence interval for the slope of the regression line.

Solution: Standardized estimated regression coefficients have t distributions with $n - 2$ degrees of freedom, so we construct confidence intervals with estimates plus or minus a t multiple of the standard error. There are $100 - 2 = 98$ degrees of freedom. (Don't be confused by the fact that there are only 13 unique values for the X 's. This is not an ANOVA where we are testing if the mean value of Y is identical for each value of X for which we have data.) The t multiplier is 1.984, but you will have needed to approximate this from the t table. The SE is read from R output. The confidence interval is then

$$0.95 \pm 1.984(0.105) \quad \text{or} \quad 0.74 < \beta_1 < 1.16$$

- (e) Use the regression line to predict the length of an infant whose gestational age is 25 weeks. Comment on the validity of this prediction.

Solution: Plugging into the regression equation gives $\hat{y} = 9.33 + 0.95(25) = 33.08$ cm. It was acceptable to say that this estimate was valid because it was an interpolation and the linear fit was adequate. It was also good to note that the estimate might be questionable if the linear fit was inadequate, especially in light of the possible effect of the outlier.

- (f) Use the regression line to predict the length of a full-term infant with gestational age 40 weeks. Comment on the validity of this prediction.

Solution: Plugging into the regression equation gives $\hat{y} = 9.33 + 0.95(40) = 47.33$ cm. Here, I wanted you to recognize that this estimate was a large extrapolation, and thus highly untrustworthy. (By the way, gestational age is usually calculated from the first day of the mother's last period rather than the date of conception because the former is often known with more certainty. Due dates are often predicted to be about 40 weeks after the first day of the mother's last period, but any birth within two weeks of that date is considered "full-term".)

- (g) Circle TRUE or FALSE: Suppose that the regression line goes through the point (age=21.7,length=30). Then, the regression line that predicts age based on length would go through the point (length=30,age=21.7).

Solution: The statement is False. Here are two reasons. The regression line minimizes the residual sum of squares, which is the sum of the squared *vertical distances* from points to the line. If we change the roles of X and Y , the regression line would minimize the sum of the squared *horizontal distances* from points to the line, which will be a different line in general.

A second reason is as follows. Suppose that x were z standard deviations from its mean. We would then predict that y would be rz standard deviations from its mean. If we then turned around to predict x from this y , we would predict that x would be $r \times rz = r^2z$ standard deviations from its mean, which is a different value than the original x except in the special cases that $|r| = 1$ or $z = 0$. In this problem, the points do not lie exactly on a line and the point in question is not (\bar{x}, \bar{y}) .

- (h) ***The following questions do not relate to the data set.***

Circle TRUE or FALSE: If the correlation coefficient is $r = -1$, then the points lie exactly on a line with slope -1 .

Solution: The statement is False. If the correlation coefficient is $r = -1$, then the points lie exactly on a line with slope with a negative slope, but the slope need not be -1 .

- (i) Circle TRUE or FALSE: For every simple linear regression, if the X value is 1.7 standard deviations above the mean, then the predicted Y value is also 1.7 standard deviations above the mean.

Solution: The statement is False. In general, the predicted value for Y is rz standard deviations from \bar{y} . We would predict Y to be 1.7 standard deviations above the mean only in the case that $r = 1$.

- (j) Circle TRUE or FALSE: If $r = 0.98$, this implies that a straight line will fit the data much better than a curve.

Solution: The statement is False. As was clear from graphs in lecture, it is possible for X and Y to have a perfect curved relationship that is much better than a linear fit, but still have a high r .