MODULE 10

LINEAR REGRESSION

Objectives:

1. Describe the purposes of regression.
2. Describe the criteria used to determine the best-fit line to a set of bivariate data.
3. Describe the assumptions surrounding the best-fit criteria.
4. Identify the response and explanatory variables.
5. Describe the equation of a line and what the slope and intercept "mean."
6. Make appropriate predictions using the best-fit line.
7. Describe the meaning of the coefficient of determination.

Contents

Linear regression analysis is used to model the relationship between two quantitative variables for two related purposes – (i) explaining variability in the response variable and (ii) predicting future values of the response variable. Examples include predicting ...

- ... the future sales of a product from its price.
- ... family expenditures on recreation from family income.
- ... an animal's food consumption in relation to ambient temperature.
- ... a person's score on a German assessment test based on how many years the person studied German.

◇ **Explaining variability of and predicting future values of response variables are the two goals of regression.**

Exact predictions cannot be made because of natural variability. For example, two people with the same intake of mercury (from consumption of fish) will not have the same level of mercury in their blood stream (e.g., observe the two individuals in Figure 10.1 that had intakes of 580 ug HG/day). Thus, the best that can be accomplished is to predict the average or expected value for a person with a particular intake value. This is accomplished by finding the line that best "fits" the points on a scatterplot of the data and using that line to make predictions. Finding and using the "best-fit" line is the topic of this module.

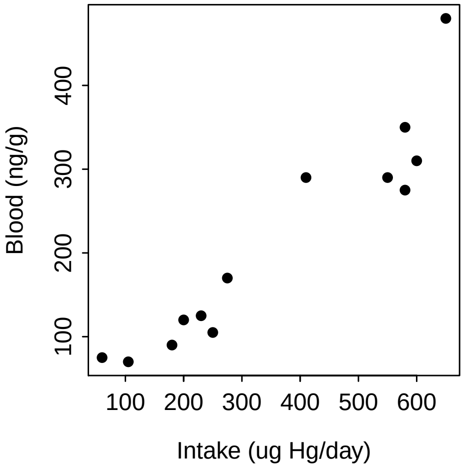


Figure 10.1. Scatterplot of intake of mercury in fish and the mercury in the blood stream.

10.1 Response and Explanatory Variables

Recall from Section 8.1 that the response (or dependent) variable is the variable to be predicted or explained and the explanatoroy (or independent) variale is the variable that will help do the predicting or explaining. In the examples mentioned above, future sales, family expenditures on recreation, the animal's food consumption, and score on the assessment test are response variables and product price, family income, temperature, and years studying German are explanatory variables, respectively. The response variable is on the y-axis and the explanatory variable is on the x-axis of scatterplots.

∆ **Response Variable**: The variable that will be explained or predicted.

△ **Explanatory Variable**: The variable that may explain or be used to predict the response variable.

Review Exercises

10.1 Dudgeon (2000), while describing the features of major tropical Asian rivers, examined the relationship between the length (km) and drainage area (km^2) of 11 waterways. In particular he wanted to determine if a model could be produced that would allow the drainage area of the river to be predicted from the length of the river. Identify the response and explanatory variables. Explain your choices. *Answer*

10.2 Researchers collected data on 56 normal births at a Wellington, New Zealand hospital. They were interested in determining if the weight of the newborn child (labeled as *BirthWt*) could be predicted by knowing the mothers age (labeled as *Age*). Identify the response and explanatory variables. Explain your choices. *Answer*

10.2 Slope and Intercept

The equation of a line is commonly expressed as,

$$y = mx + b$$

where both x and y are variables, m represents the slope of the line, and b represents the y-intercept.[1] It is important that you can look at the equation of a line and identify the response variable, explanatory variable, slope, and intercept. The response variable will always appear on one side of the equation (usually the left) by itself. The value or symbol that is multiplied by the explanatory variable (e.g., x) is the slope, and the value or symbol by itself is the intercept. For example, in

$$blood = 3.501 + 0.579 * intake$$

blood is the response variable, *intake* is the explanatory variable, 0.579 is the slope (it is multiplied by the explanatory variable), and 3.501 is the intercept (it is not multiplied by anything in the equation). The same conclusions would be made if the equation had been written as

$$blood = 0.579 * intake + 3.501$$

◇ **In the equation of a line, the slope is always multiplied by the explanatory variable and the intercept is always by itself.**

In addition to being able to identify the slope and intercept of a line you also need to be able to interpret these values. Most students define the slope as “rise over run” and the intercept as “where the line crosses

[1] Hereafter, simply called the “intercept.”

the y-axis." These "definitions" are very loose geometric representations. For our purposes, the slope and intercept must be more strictly defined.

To define the slope, first think of "plugging" two values of intake into the equation discussed above. For example, if $intake = 100$, then $blood = 3.501 + 0.579 * 100$=61.40 and if $intake$ is one unit larger (i.e., $intake = 101$), then $blood = 3.501 + 0.579 * 101$=61.98.[2] The difference between these two values is 61.98-61.40=0.579. Thus, it is seen that the slope is the change in value of the response variable for a single unit change in the value of the explanatory variable (Figure 10.2). That is, mercury in the blood changes 0.579 units for a single unit change in mercury intake. So, if an individual increases mercury intake by one unit, then mercury in the blood will increase by 0.579 units, ON AVERAGE. Alternatively, if one individual has one more unit of mercury intake than another individual, then the first individual will have, ON AVERAGE, 0.579 more units of mercury in the blood.

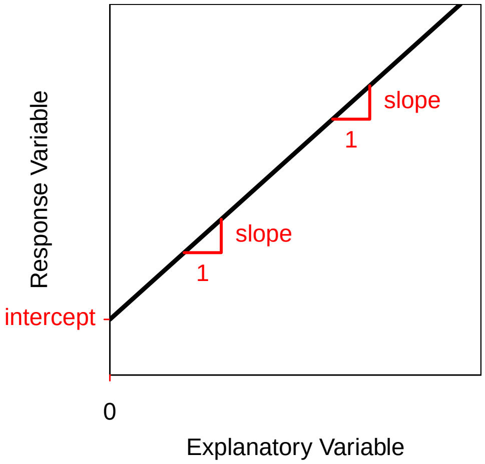


Figure 10.2. Schematic representation of the meaning of the intercept and slope in a linear equation.

To define the intercept, first "plug" $intake = 0$ into the equation discussed above; i.e., $blood = 3.501 + 0.579 * 0$=3.501. Thus, the intercept is the value of the response variable when the explanatory variable is equal to zero (Figure 10.2). In this example, the average mercury in the blood for an individual with no mercury intake is 3.501. Many times, as is true with this example, the interpretation of the intercept will be nonsensical. This is because $x = 0$ will likely be outside the range of the data collected and, perhaps, outside the range of possible data that could be collected.

The equation of the line is a model for the relationship depicted in a scatterplot. Thus, the interpretations for the slope and intercept represent the *average* change or the *average* response variable. Thus, whenever a slope or intercept is being interpreted it must be noted that the result is an *average* or *on average*.

Δ **Slope**: The change in value of the response variable for a unit change in value of the explanatory variable.

Δ **Intercept**: The value of the response variable when the explanatory variable is equal to zero.

[2] For simplicity of exposition, the actual units are not used in this discussion. However, "units" would usually be replaced with the actual units used for the measurements.

Review Exercises

10.3 The research described in Review Exercise 10.1 identified the best-fit line equation as $Area = -159131 + 314.229Length$. Answer

(a) What is the response variable?
(b) Interpret the value of the slope in terms of the variables of this problem.
(c) Interpret the value of the intercept in terms of the variables of this problem.
(d) If one river was 10 km longer than another river, then how much more area would you expect it to drain?

10.4 The research described in Review Exercise 10.2 computed the following regression results: $BirthWt = 2054 + 51.7Age$. Answer

(a) What is the explanatory variable?
(b) Interpret the value of the slope in terms of the variables of this problem.
(c) Interpret the value of the intercept in terms of the variables of this problem.
(d) Assume that a mother had a child when she was 20 and when she was 25. On average, how much more or less would you expect, based on these findings, the second child to weigh compared to the first child?

10.3 Predictions

Once a best-fit line has been identified (criteria for doing so is discussed in Section 10.5), the equation of the line can be used to predict the average value of the response variable for individuals with a particular value of the explanatory variable. For example, the best-fit line for the mercury data shown in Figure 10.1 is

$$blood = 3.501 + 0.579 * intake$$

Thus, the predicated average level of mercury in the blood for an individual that consumed 240 ug HG/day is found with

$$blood = 3.501 + 0.579 * 240 = 142.461$$

Similarly, the predicted average level of mercury in the blood for an individual that consumed 575 ug HG/day is found with

$$blood = 3.501 + 0.579 * 575 = 336.426$$

A prediction may be visualized by finding the value of the explanatory variable on the x-axis, drawing a vertical line until the best-fit line is reached, and then drawing a horizontal line over to the y-axis where the value of the response variable is read (Figure 10.3).

△ **Predicted Value**: The value of y on the best-fit line at the observed value of x; abbreviated as $\hat{y}_i$ for the ith individual.

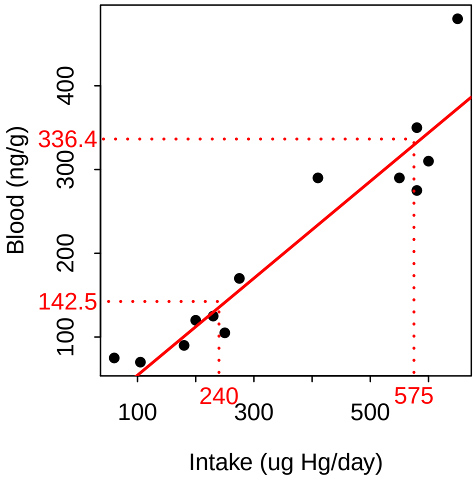


Figure 10.3. Scatterplot between the intake of mercury in fish and the mercury in the blood stream of individuals with superimposed best-fit regression line illustrating predictions for two values of mercury intake.

◇ The predicted value of the response variable at a given value of the explanatory variable is found by "plugging" the value of the explanatory variable into the equation of the line.

When predicting values of the response variable, it is important to not extrapolate beyond the range of the data. In other words, predictions with values outside the range of observed values of the explanatory variable should be made cautiously (if at all). An excellent example would be to consider the height "data" collected during the early parts of a human's life (say the first ten years). During these early years there is likely a good fit between height (the response variable) and age. However, using this relationship to predict an individual's height at age 40 would likely result in a ridiculous answer (i.e., probably over ten feet). The problem here is that the linear relationship only holds for the observed (i.e., the first ten years of life); it is not known if the same linear relationship exists outside that range of years. In fact, with human heights, it is generally known that growth first slows, eventually quits, and may, at very old ages, actually decline. Thus, the linear relationship found early in life does not hold for later years. Critical mistakes can be made when using a linear relationship to extrapolate outside the range of the data.

◇ When making predictions of the response variable, do not extrapolate beyond the range of the data.

10.4 Residuals

The predicted value is a "best-guess" for an individual based on the best-fit line. The actual value for any individual is likely to be different from this predicted value. The **residual** is a measure of how "far off" the prediction is from what is actually observed. Specifically, the residual for an individual is found by subtracting the predicted value (given the individual's observed value of the explanatory variable) from the individual's observed value of the response variable, or

$$\text{residual} = \text{observed response} - \text{predicted response}$$

For example, consider an individual that has an observed intake of 650 and an observed level of mercury in the blood of 480. As shown in the previous section, the predicted level of mercury in the blood for this individual is

$$\widehat{blood} = 3.501 + 0.579 * 650 = 379.851$$

The residual for this individual is then $480 - 379.851 = 100.149$. This positive residual indicates that the observed value is approximately 100 units greater than the average for individuals with an intake of 650.[3] As a second example, consider an individual with an observed intake of 250 and an observed level of mercury in the blood of 105. The predicted value for this individual is

$$\widehat{blood} = 3.501 + 0.579 * 250 = 148.251$$

and the residual is $105 - 148.251 = -43.251$. This negative residual indicates that the observed value is approximately 43 units less than the average for individuals with an intage of 250. A residuals is the vertical distance between an individual's point and the best-fit line (Figure 10.4).

Δ **Residual**: The vertical difference between the observed and predicted values of the response variable for an individual; computed as the difference between the observed and predicted values of the response.

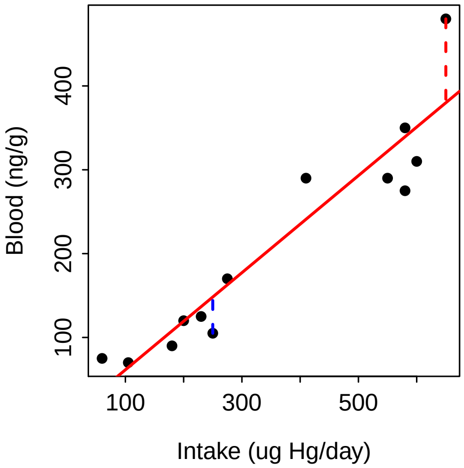


Figure 10.4. Scatterplot between the intake of mercury in fish and the mercury in the blood stream of individuals with superimposed best-fit regression line illustrating the residuals for two individuals.

[3] In other words, the observed value is "above" the line.

Review Exercises

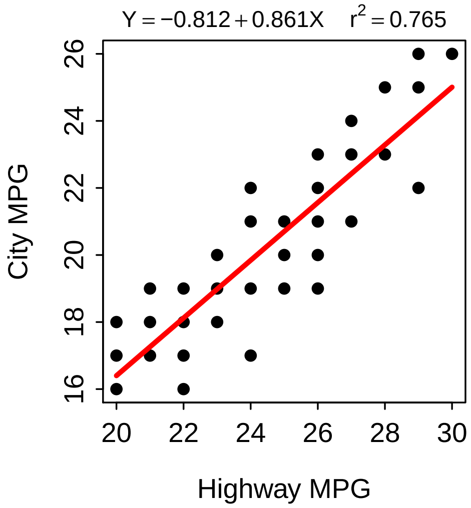
10.5 Use the results described in Review Exercise 10.3 to answer the questions below. Answer

(a) Predict the drainage area for a river 3500 km long.
(b) Calculate the residual if the river above (3500 km length) had a drainage area of 1,000,150 km^2.
(c) Predict the drainage area for a river 7500 km long.

10.6 Use the results described in Review Exercise 10.4 to answer the questions below. Answer

(a) Predict the weight of a child born to a 30-year-old mother.
(b) A 30-year-old mother had a child that weighed 3550 g. Find the residual for that mother.
(c) Predict the weight of a child born to an 18-year-old mother.

10.7 Researchers at Chevrolet attempted to determine the relationship between gas mileage (MPG) of Luminas in the city (CITY) and on the highway (HIGHWAY). Results of their analysis is shown below. Answer



(a) Predict the city mpg for a Lumina that gets 25 mpg on the highway.
(b) Predict the highway mpg for a Lumina that gets 25 mpg in the city.
(c) Predict the city mpg for a Lumina that gets 40 mpg on the highway.
(d) What is the residual for a Lumina that gets 25 mpg on the highway and 20 in the city?

10.5 Best-fit Criteria

An infinite number of lines can be placed on a graph, but many of those lines do not adequately describe the data. In contrast, many of the lines will appear, to our eye, to adequately describe the data. So, how does one find THE best-fit line from all possible lines. The **least-squares** method described below provides a quantifiable and objective measure of which line best "fits" the data.

Residuals are a measure of how far an individual is from a candidate best-fit line. Residuals computed from all individuals in a data set is a measure of how far all individuals are from the candidate best-fit line. Thus, the residuals for all individuals can be used to identify the best-fit line. The residual sum-of-squares (RSS) is the sum of all squared residuals. The least-squares criterion says that the "best-fit" line is the one line out of all possible lines that has the minimum RSS Figure 10.5.

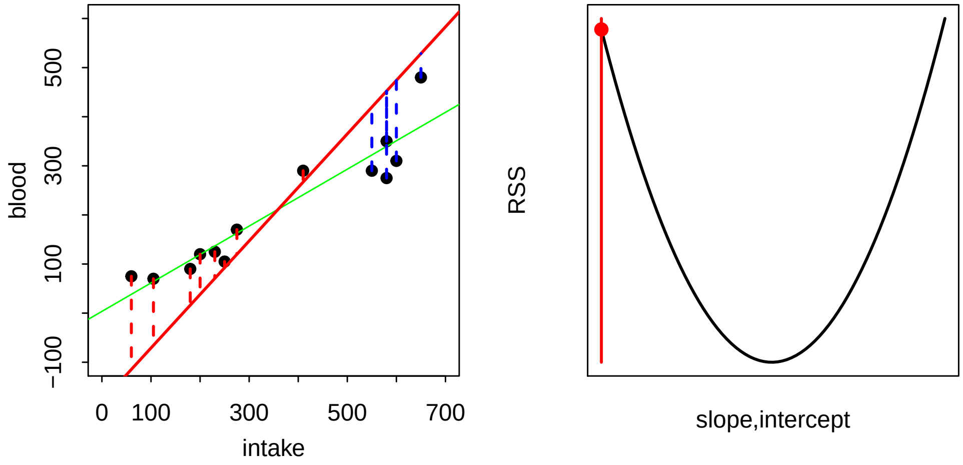


Figure 10.5. An animation illustrating how the residual sum-of-squares (RSS) for a series of candidate lines (red lines) is minimized at the best-fit line (green line).

△ **Residual sum-of-squares**: The sum of all squared residuals; abbreviated as RSS.

◇ **The least-squares criterion is that the "best-fit" line is the line of all possible lines with the minimum RSS.**

The disucssion thusfar implies that all possible lines must be "fit" to the data and the one with the minimum RSS is chosen as the "best-fit" line. As there are an infinite number of possible lines, this would be impossible to do. Theoretical statisticians have shown that the application of the least-squares criterion always produces a best-fit line with a slope given by

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

and an intercept given by

$$intercept = \bar{y} - slope * \bar{x}$$

Thus, using these formulas finds the slope and intercept for the line, out of all possible lines, that minimizes the RSS.

10.6 Assumptions

The least-squares method for finding the best-fit line only works appropriately if each of the following five assumptions about the data has been met.

1. A line describes the data (i.e., a linear form).
2. Homoscedasticity.
3. Normally distributed residuals at a given x.
4. Independent residuals at a given x.
5. The explanatory variable is measured without error.

While all five assumptions of linear regression are important, only the first two are vital when the best-fit line is being used primarily as a descriptive model for data.[4] Description is the primary goal of linear regression used in this course and, thus, only the first two assumptions are considered further.

The linearity assumption appears obvious – if a line does not represent the data, then don't try to fit a line to it! Violations of this assumption are evident by a non-linear or curving form in the scatterplot. The homoscedasticity assumption states that the variability about the line is the same for all values of the explanatory variable. In other words, the dispersion of the data around the line must be the same everywhere along the entire line. Violations of this assumption generally present as a "funnel-shaped" dispersion of points from left-to-right on a scatterplot.

Violations of these assumptions are often evident on "fitted-line plots" – i.e., scatterplots with the best-fit line superimposed (Figure 10.6).[5] If the points look more-or-less like random scatter around the best-fit line, then neither the linearity nor the homoscedasticity assumption has been violated.

In this text, if an assumption has been violated, then one should not continue to interpret the linear regression. However, in many instances, an assumption violation can be "corrected" by transforming one or both variables to a different scale. Transformations are not discussed in this book.

◇ **If the regression assumptions are not met, then the regression results should not be interpreted.**

10.7 Coefficient of Determination

The coefficient of determination, abbreviated as r^2, is the proportion of the total variability in the response variable that is explained away by knowing the value of the explanatory variable and the best-fit model. The r^2 can take values between 0 and 1.[6] In simple linear regression, r^2 is literally the square of r, the correlation coefficient.[7]

Δ **Coefficient of Determination**: The proportion of the total variability in the response variable that is explained away by knowing the value of the explanatory variable and the best-fit model; abbreviated as r^2.

[4] In contrast to using the model to make inferences about a population model.

[5] Residual plots, not discussed in this text, are another plot that often times is used to better assess assumption violations.

[6] It is common for r^2 to be presented as a percentage.

[7] Simple linear regression is the fitting of a model with a single explanatory variable and is the only model considered in this module and this course. See Section 8.4 for a review of the correlation coefficient.

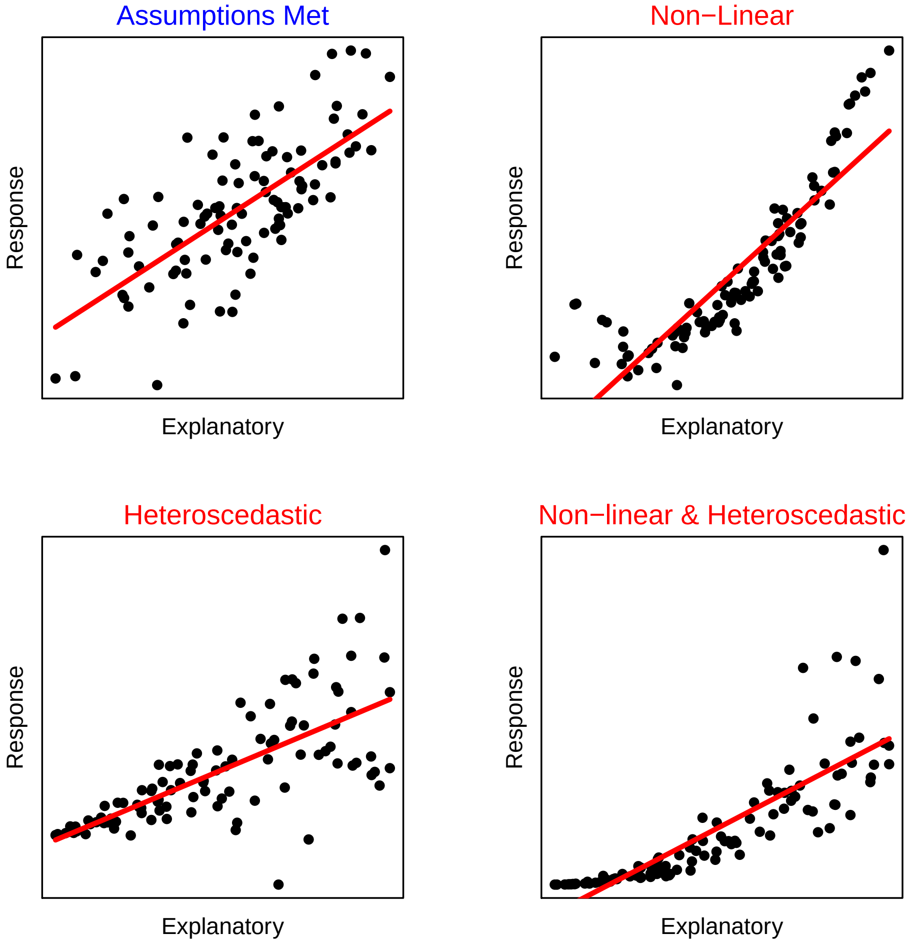


Figure 10.6. Fitted-line plots illustrating when the regression assumptions are met (upper-left) and three common assumption violations.

◇ r^2 can take values between 0 and 1.

The meaning of r^2 can be examined by considering predictions of the response variable with and without knowledge of the value of the explanatory variable. First, consider predicting the value of a particular response variable without any information about the explanatory variable. In this case, the best prediction for the value of the response variable is to use the sample mean of the response variable (represented by the dashed blue horizontal line in Figure 10.7). However, because of natural variability, not all individuals will have this value. Thus, the prediction might be "bracketed" by saying that the individual will be between the observed minimum and maximum values (solid blue horizontal lines). Loosely speaking, this range can be thought of as the "total variability in the response variable" (blue box).

Suppose now that interest is in predicting the value of the response variable for an individual with a known value of the explanatory variable (at the dashed vertical red line in Figure 10.7). The predicted value for this individual is the value of the response variable at the corresponding point on the best-fit line (dashed horizontal red line). Again, because of natural variability, not all individuals with this value of

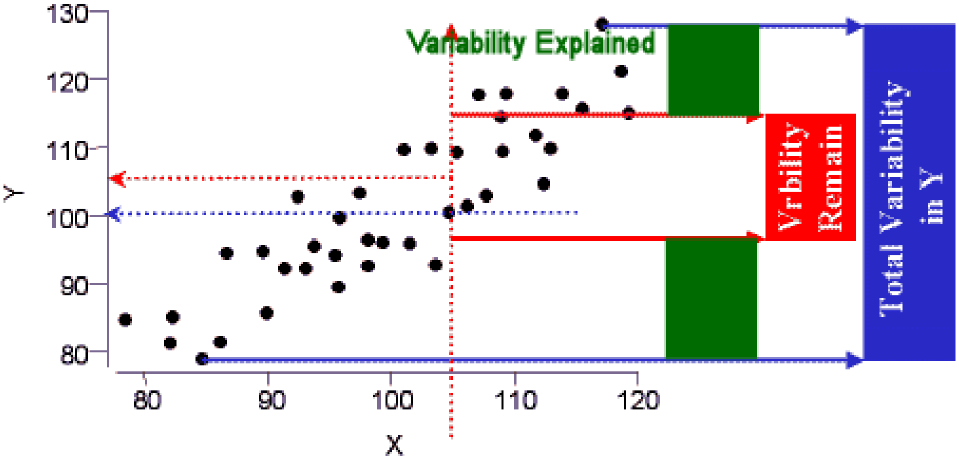


Figure 10.7. Fitted line plot with visual representations of variabilities explained and unexplained. A full explanation is in the text.

the explanatory variable will have this exact value of the response variable. However, the prediction is now "bracketed" by the minimum and maximum value of the response variable **ONLY** for those individuals with the particular value of the explanatory variable (solid red horizontal lines). Loosely speaking, this range can be thought of as the "variability in the response variable remaining after knowing the value of the explanatory variable" (red box). This is the variability in the response variable that remains even after knowing the value of the explanatory variable or the variability in the response variable that cannot be explained away (by the explanatory variable).

The portion of the total variability in the response variable that was explained away consists of all the values of the response variable that would no longer be entertained as possible predictions once the value of the explanatory variable is known (green box in Figure 10.7). Now, by the definition of r^2, the computation of r^2 can be visualized as the area of the green box divided by the area of the blue box. This calculation does not depend on which value of the explanatory variable is chosen as long as the data are evenly distributed around the line (i.e., homoscedastic – see Section 10.6).

If the variability explained away (the green box in Figure 10.7) approaches the total variability in the response variable (the blue box), then r^2 approaches 1. This will happen only if the variability about the line approaches zero. In contrast, the variability explained (the green box) will approach zero if the slope is zero (i.e., there is no relationship between the response and explanatory variables). Thus, values of r^2 also indicate the strength of the relationship; values near 1 are stronger than values near 0. Values near 1 also mean that predictions will be fairly accurate – i.e., there is little variability remaining after knowing the explanatory variable.

◇ A value of r^2 near 1 represents a strong relationship between the response and explanatory variables that will lead to accurate predictions.

10.8 Examples I

There are twelve questions that are commonly asked about linear regression results. These twelve questions are listed below with some hints about things to remember when answering some of the questions. An example of these questions in context is then provided.

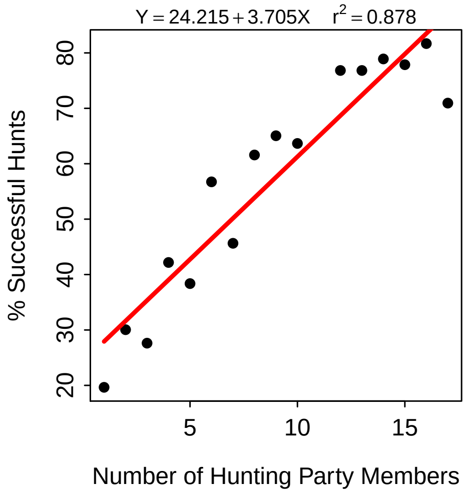
1. What is the response variable? *Identify which variable is to be predicted or explained, which variable is dependent on another variable, which would be hardest to measure, or which is on the y-axis.*
2. What is the explanatory variable? *The remaining variable after identifying the response variable.*
3. Comment on linearity and homoscedasticity. *Examine fitted-line plot for curvature (i.e., non-linearity) or a funnel-shape (i.e., heteroscedasticity).*
4. What is the equation of the best-fit line? *In the generic equation of the line ($y = mx + b$) replace y with the name of the response variable, x with the name of the explanatory variable, m with the value of the slope, and b with the value of the intercept.*
5. Interpret the value of the slope. *Comment on how the response variable changes by slope amount for each one unit change of the explanatory variable, on average.*
6. Interpret the value of the intercept. *Comment on how the response variable equals the intercept, on average, if the explanatory variable is zero.*
7. Make a prediction given a value of the explanatory variable. *Plug the given value of the explanatory variable into the equation of the best-fit line.*
8. Compute a residual given values of both the explanatory and response variables. *Make a prediction (see previous question) and then subtract this value from the observed value of the response.*
9. Identify an extrapolation in the context of a prediction problem. *Examine the x-axis scale on the fitted-line plot and do not make predictions outside of the plotted range.*
10. What is the proportion of variability in the response variable explained by knowing the value of the explanatory variable? *This is r^2.*
11. What is the correlation coefficient? *This is the square root of r^2. Make sure to put a negative sign on the result if the slope is negative.*
12. How much does the response variable change if the explanatory variable changes by X units? *This is an alternative to asking for an interpretation of the slope. If the explanatory variable changes by X units, then the response variable will change by X*slope units, on average.*

All answers should refer to the variables of the problem – thus, “y”, “x”, “response”, or “explanatory” should not be in any part of any answer. The questions about the slope, intercept, and predictions need to explicitly identify that the answer is an “average” or “on average.”

Chimp Hunting Parties

> *Stanford (1996) gathered data to determine if the size of the hunting party (number of individuals hunting together) affected the hunting success of the party (number of hunts that resulted in a kill) for wild chimpanzees (Pan troglodytes) at Gombe. The results of their analysis for 17 hunting parties is shown in the figure below.*[8] *Use these results to answer the questions below.*

[8] These data are in Chimp.csv.



Q: What is the response variable?

A: The response variable is the percent of successful hunts because the authors are attempting to see if success depends on hunting party size. In addition, the percent of successful hunts is shown on the y-axis.

Q: What is the explanatory variable?

A: The explanatory variable is the size of the hunting party.

Q: In terms of the variables of the problem, what is the equation of the best-fit line?

A: The equation of the best-fit line for this problem is % Success of Hunt = 24.215 + 3.705*Number of Hunting Party Members.

Q: Interpret the value of the slope in terms of the variables of the problem.

A: The slope indicates that for every increase of one member to the hunting party the percent of successful hunts increases by 3.705, on average.

Q: Interpret the value of the intercept in terms of the variables of the problem.

A: The intercept indicates that a hunting party with no members will have a percent of successful hunts of 24.215, on average.

Q: What is the predicted hunt success if the hunting party consists of 20 chimpanzees?

A: The predicted hunt success for parties with 20 individuals is an extrapolation, because 20 is outside the range of the number of members observed on the x-axis of the fitted-line plot.

Q: What is the predicted hunt success if the hunting party consists of 12 chimpanzees?

A: The predicted hunt success for parties with 12 individuals is 24.215 + 3.705*12 = 68.7%.

Q: What is the residual if the hunt success for 10 individuals is 50%?

A: The residual in this case is 50-(24.215 + 3.705*10) = 50-61.3 = -11.3. Therefore, it appears that the success of this hunting party is 11.3% lower than average for this size of hunting party.

Q: What proportion of the variability in hunting success is explained by knowing the size of the hunting party?

> **A:** The proportion of the variability in hunting success that is explained by knowing the size of the hunting party is r^2=0.88.

Q: What is the correlation between hunting success and size of hunting party?

> **A:** The correlation between hunting success and size of hunting party is r =0.94.

Q: How much does hunt success decrease, on average, if there are two fewer individuals in the party?

> **A:** If the hunting party has two fewer members, then the hunting success would decrease by 7.4% (i.e., −2*3.705), on average.

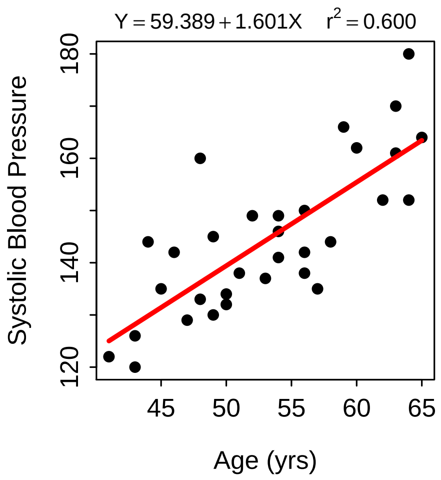
Q: Does any aspect of this regression concern you (i.e., consider the regression assumptions)?

> **A:** The data appear to be very slightly curved but there is no evidence of a funnel-shape. Thus, the data may be slightly non-linear but they appear homoscedastic.

◇ **All interpretations should be "in terms of the variables of the problem" rather than the generic terms of x, y, response variable, and explanatory variable.**

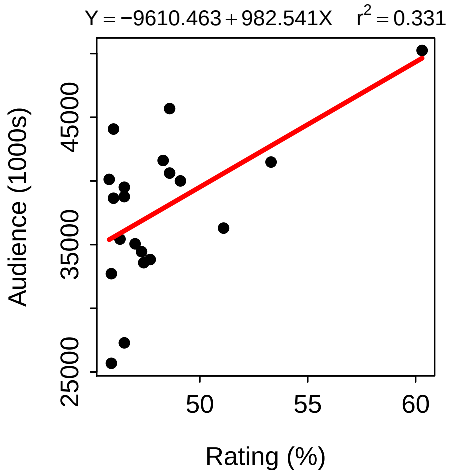
Review Exercises

10.8 The age (in years) and systolic blood pressure were measured for 32 white males over the age of 40. The researchers wanted to determine if systolic blood pressure increased with increasing age. Thus, they computed the regression depicted in the fitted-line plot below. Use these results to answer the questions below. *Answer*



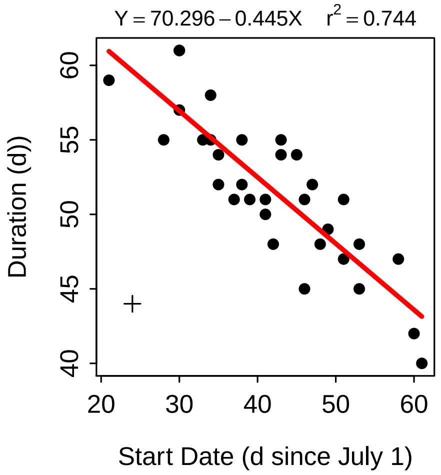
(a) Which is the explanatory variable?
(b) Which is the response variable?
(c) In terms of the variables of this problem, what is the equation of the best-fit line?
(d) In terms of the variables of this problems, interpret the value of the intercept.
(e) In terms of the variables of this problems, interpret the value of the slope.
(f) If male A is 3 years younger than male B, how much difference do you expect to see in their systolic blood pressures?
(g) What is the predicted systolic blood pressure for a 70-year-old male?
(h) What is the residual for a a 50-year-old male with a SBP of 131?
(i) What is the correlation coefficient between Age and SBP?
(j) What proportion of the variability in SBP is explained by knowing the person's AGE?
(k) What is the predicted systolic blood pressure for a 55-year-old male?

10.9 There are at least two ways that special TV programs could be rated, and both are of interest to advertisers – the estimated size of the audience and the percentage of TV-owning households that tuned into the program. Use the results below for the 20 all-time top-rated programs to determine if the estimated size of the audience can be predicted from the percentage of TV-owning households tuned into the program. *Answer*



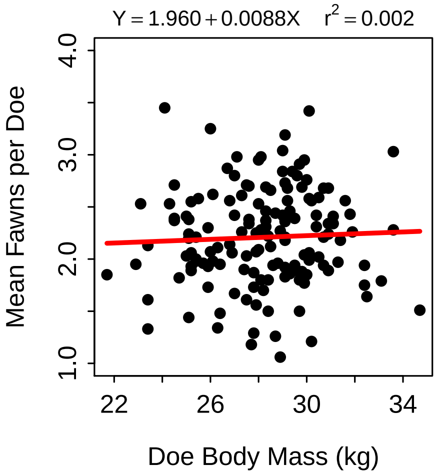
(a) What did the researchers consider the response variable to be?
(b) What is the equation of the best-fit line in terms of the variables of the problem?
(c) Interpret the value of the slope in terms of the variables of the problem.
(d) What is the predicted audience size for a show with a rating of 40.1%?
(e) What is the residual for a show with a rating of 55 and an audience size (1000s) of 40000?
(f) What proportion of the variability in audience size is explained by known the rating percentage?
(g) What is the correlation between audience size and rating percentage?
(h) What are two things that bother you about this analysis as it is presented here? Be specific!

10.10 Vega Rivera *et al.* (1998) examined the relationship between the duration of molt and the date of molt start (measured in days since July 1) for wood thrush (*Hylocichla mustelina*). A recreation of their results is shown below (note that the outlier marked by a "+" in the scatterplot was ignored in the calculation of the best-fit line). Use these results to answer the questions below. *Answer*



(a) What is the explanatory variable?
(b) What is the response variable?
(c) In terms of the variables of this problem, what is the equation of the best-fit line?
(d) In terms of the variables of this problem, interpret the value of the slope.
(e) In terms of the variables of this problem, interpret the value of the intercept.
(f) What is the predicted molt duration if the molt starts on September 10 (71 d since July 1)?
(g) What is the residual if molt duration is 48 d and the start date is Aug. 12 (43 d since July 1)?
(h) What is the correlation between molt duration and molt start date?
(i) What proportion of the variability in molt duration is explained when molt start date is 37?
(j) What proportion of the variability in molt duration is explained when molt start date is 57?
(k) What would happen to the value of the slope if the outlier was NOT ignored?

10.11 Wildlife ecologists in Texas wanted to determine if the number of fawns born to each doe could be explained by the doe's body mass (Ginnett and Young 2000). As part of their study, the researchers recorded the mean number of fawns born to a doe (over a period of time) and the body mass of the doe (kg). Use the results in the following graph to explain the relationship to answer the questions below. Answer



(a) Which is the explanatory variable?
(b) Which is the response variable?
(c) Express the equation of the best-fit line in terms of the variables of the problem.
(d) Interpret the slope of the best-fit line in terms of the variables of the problem.
(e) If a doe weighed 45 kg, how many fawns on average would you expect her to have?
(f) If a doe weighing 32 kg gave birth to an average of 1.9 fawns, what is the residual for this doe?
(g) What is the correlation coefficient between mean number of fawns born and doe body mass?
(h) How much of the variability in the mean number of fawns born is explained by knowing the body mass of does?
(i) If body mass increases by 5 kg, how many more fawns can you expect that doe have?
(j) Do you have any concerns about the strength of this relationship?

10.9 Regression in R

The mercury intake and amount in the blood data is loaded below as an example for this section.

```
> setwd('c:/data/')
> merc <- read.csv("Mercury.csv")
```

```
> str(merc)
'data.frame': 13 obs. of  2 variables:
 $ intake: num  180 200 230 410 600 550 275 580 580 105 ...
 $ blood : num  90 120 125 290 310 290 170 275 350 70 ...
```

The linear regression model is fit to two quantitative variables with `lm()`. The first argument is a formula of the form `response~explanatory`, where `response` contains the response variable and `explanatory` contains the explanatory variable, and the corresponding data.frame is in `data=`. The results of `lm()` should be assigned to an object so that specific results can be extracted.

◇ The same formula is used to make a scatterplot with `plot()` and find the best-fit line with `lm()`.

The regression was fit to the mercury data below. From this it is seen that the intercept is 3.501 and the slope is 0.579.

```
> ( lm1 <- lm(blood˜intake,data=merc) )
Coefficients:
(Intercept)       intake
     3.5007       0.5791
```

A fitted-line plot (Figure 10.8) is constructed by submitting the `lm()` object to `fitPlot()`.

```
> fitPlot(lm1)
```

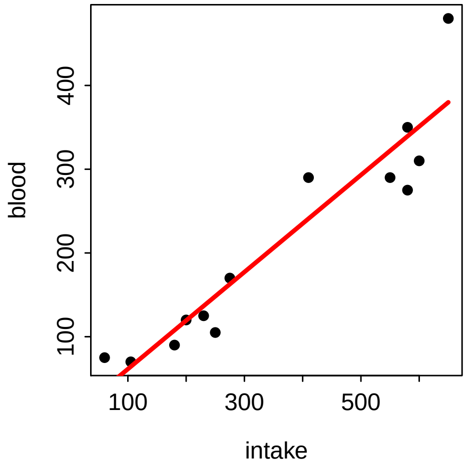


Figure 10.8. Fitted-line plots for the regression of mercury in the blood on mercury intake.

Predicted values from the linear regression are obtained with `predict()`. The `predict()` function requires the saved `lm()` object as its first argument. The second argument is a data.frame constructed with `data.frame()` that contains the **EXACT** name of the explanatory variable as it appeared in `lm()` set equal to the value of the explanatory at which the prediction should be made. For example, the predicted amount of mercury in the blood for an intake of 240 μg per day is 142.5, as obtained below.

```
> predict(lm1,data.frame(intake=240))
       1
142.4895
```

◇ The name of the explanatory variable used in `predict()` must be exactly the same as it appears in the original data frame.

The coefficient of determination is computed by submitting the saved `lm()` object to `rSquared()`. For example, 88.4% of the variability in mercury in the blood is explained by knowing the amount of mercury at intake. [Note the use of `digits=` to control the number of decimals.]

```
> rSquared(lm1,digits=3)
[1] 0.884
```

10.10 Examples II

Car Weight and MPG

In Module 8, an EDA for the relationship between *HMPG* (the highway miles per gallon) and *Weight* (lbs) of 93 cars from the 1993 model year was performed. This relationship will be explored further here as an example of a complete regression analysis. In this analysis, the regression output will be examined within the context of answering the twelve typical questions. These data are read into R below and the linear regression model is fit, coefficients extracted, fitted-line plot constructed, and coefficient of determination extracted.

```
> cars93 <- read.csv("data/93cars.csv")
> ( lm2 <- lm(HMPG~Weight,data=cars93) )
Coefficients:
(Intercept)       Weight
  51.601365    -0.007327
> fitPlot(lm2,ylab="Highway MPG")
> rSquared(lm2,digits=3)
[1] 0.657
```

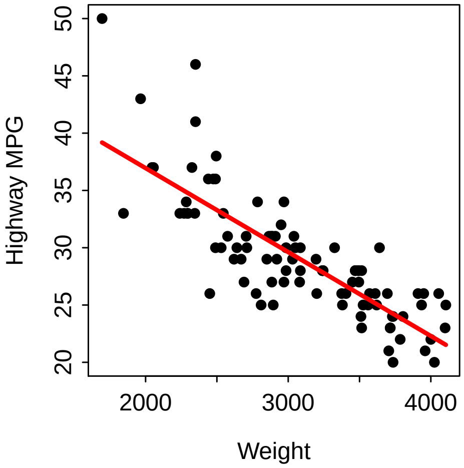


Figure 10.9. Fitted line plot of the regression of highway MPG on weight of 93 cars from 1993.

The simple linear regression model appears to fit the data moderately well as the fitted-line plot (Figure 10.9) shows only a very slight curvature and only very slight heteroscedasticity.[9] The sample slope is -0.0073, the sample intercept is 51.6, and the coefficient of determination is 0.657.

[9] In advanced statistics books, objective measures for determining whether there is significant curvature or heteroscedasticity in the data are used. In this book, we will only be concerned with whether there is strong evidence of curvature or heteroscedasticity. There does not seem to be either here.

Q: What is the response variable?

> **A:** The response variable in this analysis is the highway MPG, because that is the variable that we are trying to learn about or explain the variability of.

Q: What is the explanatory variable?

> **A:** The explanatory variable in this analysis is the weight of the car (by process of elimination).

Q: In terms of the variables of the problem, what is the equation of the best-fit line?

> **A:** The equation of the best-fit line for this problem is HMPG = 51.6 - 0.0073Weight.

Q: Interpret the value of the slope in terms of the variables of the problem.

> **A:** The slope indicates that for every increase of one pound of car weight the highway MPG decreases by -0.0073, on average.

Q: Interpret the value of the intercept in terms of the variables of the problem.

> **A:** The intercept indicates that a car with 0 weight will have a highway MPG value of 51.6, on average.[10]

Q: What is the predicted highway MPG for a car that weighs 3100 lbs?

> **A:** The predicted highway MPG for a car that weighs 3100 lbs is 51.60137 - 0.00733(3100) = 28.9 MPG. Alternatively, this value is computed with

```
> predict(lm2,data.frame(Weight=3100))
       1
28.88748
```

Q: What is the predicted highway MPG for a car that weighs 5100 lbs?

> **A:** The predicted highway MPG for a car that weighs 5100 lbs should not be computed with the results of this regression, because 5100 lbs is outside the domain of the data (Figure 10.9).

Q: What is the residual for a car that weights 3500 lbs and has a highway MPG of 24?

> **A:** The predicted highway MPG for a car that weighs 3500 lbs is 51.60137 - 0.00733(3500) = 26.0. Thus, the residual for this car is 24 - 26.0 = -2.0. Alternatively, this is computed in R with

```
> 24-predict(lm2,data.frame(Weight=3500))
        1
-1.956658
```

> Therefore, it appears that this car gets 2.0 MPG less than an average car with the same weight.

Q: What proportion of the variability in highway MPG is explained by knowing the weight of the car?

> **A:** The proportion of the variability in highway MPG that is explained by knowing the weight of the car is r^2=0.66.

Q: What is the correlation between highway MPG and car weight?

[10] This is the correct interpretation of the intercept. However, it is nonsensical because it is an extrapolation; i.e., no car will weight 0 pounds.

A: The correlation between highway MPG and car weight is r =-0.81.[11]

Q: How much is the highway MPG expected to change if a car is 1000 lbs heavier?

A: If he car was 1000 lbs heavier, you would expect the car's highway MPG to decrease by 7.33 (i.e., 1000 slopes).

Review Exercises

10.12 Wang and Finch (1997) hypothesized that larger willow flycatchers (*Empidonax traillii*) migrated up the Middle Rio Grande River earlier than small willow flycatchers. To test this hypothesis they captured flycatchers on several days during their migration and measured the wing length (mm; an index of overall body size) of each bird. They recorded the date that the bird was captured as a Julian date (days since Jan. 1). The results of their study are found in Flycatcher.csv. Load these data into R and produce results that can be used to answer the questions below. *Answer*

(a) What is the explanatory variable?
(b) What is the response variable?
(c) In terms of the variables of this problem, what is the equation of the best-fit line?
(d) In terms of the variables of this problem, interpret the value of the intercept.
(e) In terms of the variables of this problem, interpret the value of the slope.
(f) How much different do you expect the wing length to be ten days later?
(g) What is the predicted wing length on day 180?
(h) What is the residual for a bird with wing length 66.5 on day 151?
(i) What proportion of the variability in wing length is explained by knowing the date?
(j) What is the correlation coefficient between wing length and date?
(k) Comment on the assumptions of the linear regression.

10.13 Carroll (1975) examined the relationship between per capita consumption of animal fat (g/day; AnimFatI) and age-adjusted death rate from breast cancer (AgeAdjDe) for 39 countries. Her goal was to determine if variability in the breast cancer death rate could be explained by the amount of fat consumed. The data for their study are found in CancerFat.csv. Load these data into R and produce results that can be used to answer the questions below. *Answer*

(a) Which variable is the response variable?
(b) What is an individual in this study?
(c) In terms of the variables of this problem, what is the equation of the best-fit line?
(d) In terms of the variables of this problems, interpret the value of the slope.
(e) If country A consumes 4 g/day less animal fat than country B, how much different will the predicted age adjusted death rate due to breast cancer be for country A?
(f) What is the predicted age adjusted death rate due to breast cancer for a country that consumes 170 g/day of animal fat?
(g) What is the residual for a country that consumes 90 g/d of animal fat and has an age adjusted death rate due to breast cancer of 14.5?
(h) What is the correlation coefficient between the age adjusted death rate and the intake of animal fat?

[11] Put a negative sign in front of your result from taking the square root of r^2, because the relationship between highway MPG and weight is negative.

(i) How much of the variability in a country's age adjusted death rate due to breast cancer is explained by knowing the value of its animal fat intake?
(j) Can it be said that an increase in intake of animal fat is the cause for an increase in the age adjusted death rate due to breast cancer? Why or why not?

10.14 Allen *et al.* (1997) investigated the impact of the density of red-imported fire ants (*Solenopsis invicta*; RIFA) on the recruitment of white-tailed deer (*Odocoileus virginianus*) fawns (an index of does to fawns). A modified version of their results are found in RIFA.csv. Load these data into R and produce results that can be used to answer the questions below. *Answer*

(a) What is the response variable?
(b) What is the explanatory variable?
(c) In terms of the variables of this problem, what is the equation of the best-fit line?
(d) In terms of the variables of this problem, interpret the value of the slope.
(e) If the RIFA index increases by 500, how much different do you expect fawn recruitment to be?
(f) What is the predicted fawn recruitment when the RIFA index is 1700?
(g) What is the residual when the RIFA index is 2700 and fawn recruitment is 0.3?
(h) What is the correlation coefficient between RIFA and fawn recruitment?
(i) What proportion of the variability in fawn recruitment is explained by knowing the RIFA index?
(j) Comment on the assumptions in this regression.

10.15 All incoming freshmen are required to take a math assessment test to determine which math classes they should take. Sometimes pre-registering students will register before taking the assessment. To make the best possible course choices for these students, the adviser would like to predict their assessment score (ASSESS) based on their math ACT scores (ACT). The ACT score and assessment score from 72 freshmen from 2003 are stored in NCAssess.csv. Load these data into R and produce results that can be used to answer the questions below. *Answer*

(a) What is the explanatory variable?
(b) In terms of ACT and Assessment test scores, what does the value of the slope mean?
(c) Mary Lamb had an ACT score of 40. Predict her assessment score.
(d) John Tukey had an ACT score of 19. Predict his assessment score.
(e) John Tukey actually scored a 15 on his assessment test. Calculate his residual?
(f) What proportion of the variability in assessment score is explained by knowing the ACT score?
(g) What are the two most important assumptions in a regression analysis. Are these violated for this data set? Why or why not?
(h) Do you think that these results provide a useful predictor of math assessment scores in cases where those scores are not available but ACT scores are? Explain.

10.16 Suit and Bauer (1990) examined DNA indices obtained from fresh and frozen tissue samples with the goal of determining if fresh values could be predicted from frozen values. The data for their study are found in DNA.csv. Load these data into R and produce results that can be used to answer the questions below. Note that one outlier should be excluded from the analysis. *Answer*

(a) What did the researchers consider as the response variable?
(b) What is the equation of the best-fit line in terms of the variables of the problem?
(c) Interpret the value of the slope in terms of the variables of the problem.
(d) What is the predicted fresh index if the frozen index is 4.05?
(e) What is the residual for a fresh index of 2.1 and a frozen index of 2.2?
(f) What proportion of the variability in the fresh index is explained by knowing the frozen index?

(g) What is the correlation between the fresh and frozen indices?
(h) What are the two major assumptions of regression and do they look like they've been met with these data (be specific!)?

10.17 Wildlife ecologist in Texas wanted to determine if the amount of precipitation could explain some of the variability observed in the number of fawns born to each doe (Ginnett and Young 2000). Because Texas has many different climatic regions, the state was broken down into eight precipitation zones, and the mean precipitation for each zone over a period of five years was calculated. Furthermore, the researchers measured the mean number of fawns born per 100 does for each of these five years. The data for their study are found in deer1.csv. Load these data into R and produce results that can be used to answer the questions below. *Answer*

(a) Express the equation of the best-fit line in terms of the variables of the problem.
(b) Interpret the slope of the best-fit line in terms of the variables.
(c) If the mean precipitation in an area were 1500 mm, how many fawns per 100 does would you expect?
(d) If a precipitation zone has a mean precipitation of 1050 mm and an average of 37 fawns per 100 does, what is the residual of this zone?
(e) What is the correlation coefficient between mean no. of fawns per 100 does and mean precipitation?
(f) What proportion of the variability in the mean number of fawns per 100 does is explained by knowing the mean precipitation?
(g) If the average amount of precipitation increases by 100 mm, how many more fawns per 100 does would you expect to be born?

10.18 It has been said that temperature can be estimated from the number of cricket chirps heard. To determine if this relationship existed, an entomologist recorded the number of chirps in a 15-second interval by crickets held at different temperatures. The data for their study are found in Chirps.csv. Load these data into R and produce results that can be used to answer the questions below. *Answer*

(a) What is the response variable?
(b) What is the explanatory variable?
(c) In terms of the variables of this problem, what is the equation of the best-fit line?
(d) In terms of the variables of this problem, interpret the value of the slope.
(e) If the number of chirps increases by 5, then how much different do you expect temperature to be?
(f) If you hear 18 chirps during the day and 15 chirps at night, then how much different is the temperature, on average?
(g) What is the residual when you hear 12 chirps and the temperature is 65 F?
(h) What is the correlation coefficient between temperature and the number of chirps?
(i) What proportion of the variability in temperature is explained by knowing the number of chirps?
(j) Construct a residual plot and use it to interpret the validity of regression assumptions.

Part III

Inference Concepts