Unit

IV

HIGHLIGHTS AND STUDY

(REFLECTIONS ON CHAPTERS 11 AND 12)

In Chapters 11 and 12, we examined the relationship between a numeric response variable and an explanatory variable that was either categorical with three or more levels (ANOVA) or numeric (regression). In some situations, either tool may be used, such as when the explanatory variable takes on only a few discrete numeric values, as is the case in the following example.

To study the effects of a preservative foliar spray of varying dosages of abscisic acid on potted impatiens plants, researchers randomly sprayed 30 plants with either water (control) or a low dose (300 PPM) or medium dose (600 PPM) of abscisic acid.*,† After the initial treatment, the plants were examined daily to determine if they were in marketable condition. The response variable was then taken to be the total number of marketable days. The results are summarized in Table IV.1 and Figure IV.1.

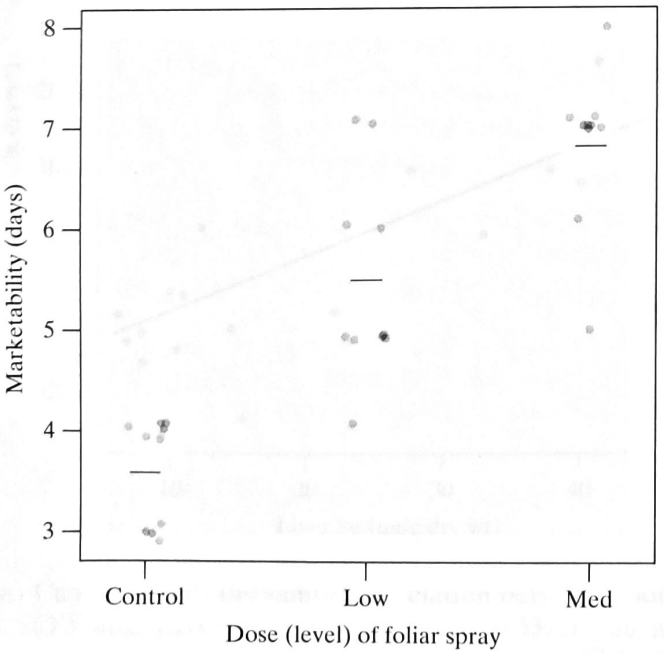


Figure IV.1 The number of marketable days of impatiens plants treated with varying doses of abscisic acid. The lines represent the sample means.

Table IV.1 Marketable days of impatiens plants treated with varying doses of abscisic acid

	Marketability (days)		
Treatment (dose)	Mean	SD	*N*
1. Control (0 PPM)	3.60	0.52	10
2. Low (300 PPM)	5.50	0.97	10
3. Med (600 PPM)	6.80	0.79	10

*Why is it important that the treatments were applied randomly? What might go wrong if the application was nonrandom?

†A high dose was also used, but is not included in this example.

Using analysis of variance and the Global F-test, we can test the null hypothesis that the mean number of marketable days is the same for all three treatments versus the alternative hypothesis that the mean number of marketable days is not the same (i.e., the treatment affects marketability).

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_A: \text{not all } \mu_i\text{'s are equal}$$

Before we begin the analysis, there are two points that are worth noting. The first regards the ability to make a causal claim with the alternative hypothesis. Since the researchers assigned the sprays to plants, this was an experiment, and thus determining cause–effect relationships is possible. The second point regards directionality, or more accurately, the lack of directionality. The researchers in this study most likely wish to show that the spray preserves the plants and increases marketability (a directional claim). At a minimum, they would be interested in showing that the two conditions using the foliar spray preserve the plants longer than the water control. Unfortunately, the Global F-test is an omni-directional test—it will only reveal if there is evidence for differences among the population means but will not indicate the nature of the differences. If the Global F-test yields significant results, we will have to carry out some post-hoc multiple comparisons (e.g., Fisher, Tukey, or Bonferroni intervals) as discussed in the optional Section 11.9.

Statistical software yields the ANOVA table displayed in Table IV.2 for these data.* As noted in Section 11.2, computer software often displays the between-group variability as "treatment" and within-group variability as "error." The ANOVA table in Table IV.2 also displays the P-value as < 0.0001 indicating that differences as extreme as those among the treatment means observed in these data are very unlikely to occur if the treatments have no effect. That is, there is very strong evidence that the treatment means are not all equal—that the foliar spray affects the mean number of marketable days.

Table IV.2 ANOVA table for impatiens preservation data

Source	df	Sum of squares	Mean square	F ratio	Prob > F
Treatment	2	51.800	25.900	42.382	<.0001
Error	27	16.500	0.611		
Total	29	68.300			

Multiple Comparisons (Optional)

Having very strong evidence for a difference among the treatments, it is now natural to investigate which groups differ from which as well as the magnitude of the differences (effect size) as considered in Section 11.9. Table IV.3 shows computer output displaying the conservative experimentwise 95% Bonferroni intervals for all pairwise comparisons (98.33% comparisonwise confidence level).

*It would be a good exercise to review the formulae behind the values listed in the table as well as to confirm their values.

Table IV.3 Experimentwise 95% Bonferroni confidence intervals comparing the mean number of marketable days for the three treatments

Comparison	$d_{ab} = \bar{y}_a - \bar{y}_b$	$SE_{D_{ab}}$	Lower 98.33%	Upper 98.33%
Med − Control: $\mu_3 - \mu_1$	3.20	0.35	2.31	4.09
Low − Control: $\mu_2 - \mu_1$	1.90	0.35	1.01	2.79
Med − Low: $\mu_3 - \mu_2$	1.30	0.35	0.41	2.19

The intervals listed in Table IV.3 provide compelling evidence that the foliar spray is effective when compared to a control, as both the confidence interval comparing the medium level to the control, $\mu_3 - \mu_1$, and low level to control, $\mu_2 - \mu_1$, are entirely positive. Furthermore, there is evidence that the efficacy increases with dosage, since the confidence interval comparing the medium to low level, $\mu_3 - \mu_2$, is also entirely positive. One might summarize these results in a publication in the following tabular (Table IV.4) or graphical forms (Figure IV.2). Because the focus is a comparison of means, the standard error of the mean, rather than the standard deviation of the sample, is listed for the reader's convenience.

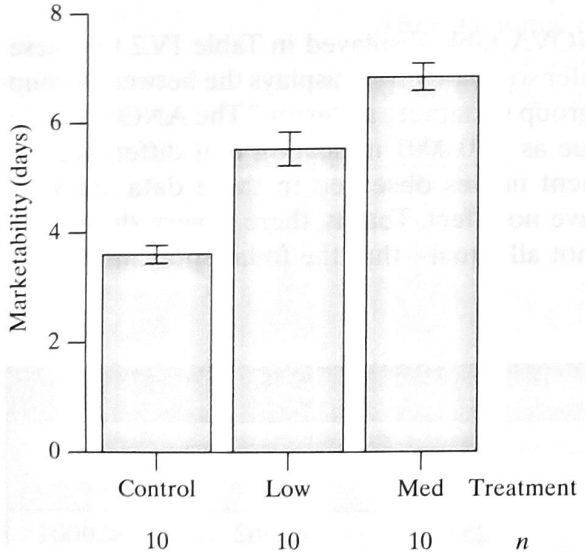


Figure IV.2 Marketable days of impatiens plants treated with varying dosages of abscisic acid. Error bars are ±SE.

Table IV.4 Marketable days of impatiens plants treated with varying dosages of abscisic acid. Means that do not share a common superscript are statistically significantly different

	Marketability (days)		
Treatment	Mean	SE	N
1. Control (0 PPM)	3.60^a	0.16	10
2. Low (300 PPM)	5.50^b	0.31	10
3. Med (600 PPM)	6.80^c	0.25	10

Checking ANOVA Requirements

In order to trust our results, we should investigate whether or not our data meet the conditions for the ANOVA Global F-test as discussed in Section 11.5. We cannot ascertain from the raw data whether or not random sampling was used or if the samples are independent. These conditions must be verified by examining the process that was used to collect the data. In this case we know that the researchers randomly assigned the three treatments to the 30 potted plants; thus, the samples can be regarded as independent from one another. It would be good to know how the initial 30 plants were chosen. While they are probably not from a true random sample, is it reasonable to regard them as a random sample? For example, did the researchers

choose the best plants to apply the sprays to? If so, these plants would not be representative of a larger population of interest.

By examining the residuals from the ANOVA model displayed in Figure IV.3, we can verify (or more accurately rule out gross violations of) the other two conditions: (1) that the several populations being compared are each normally distributed, and (2) that the populations have similar standard deviations. The normal quantile plot in Figure IV.3 shows data that are consistent with normality; the plot is fairly linear.* The horizontal banding is due to the measurements being taken discretely (days of marketability).

The residuals versus fitted value plot in Figure IV.3 (b) suggests that the population standard deviations may not be the same for all three groups (the dots on the left are not as spread out vertically as those in the center or right), and Table IV.1 shows that the sample standard deviation for the low dose is nearly double the standard deviation for the control. We also see mild evidence of the standard deviation growing with the mean, which is a common problem. (We would ordinarily expect to see 30 dots in the plot of the residuals versus fitted values, but because there were only 10 distinct values for the number of marketable days observed there appear to only be 10 observable values. If we were to jitter the points, we would observe that some of the dots on the plot are actually multiple observations.)

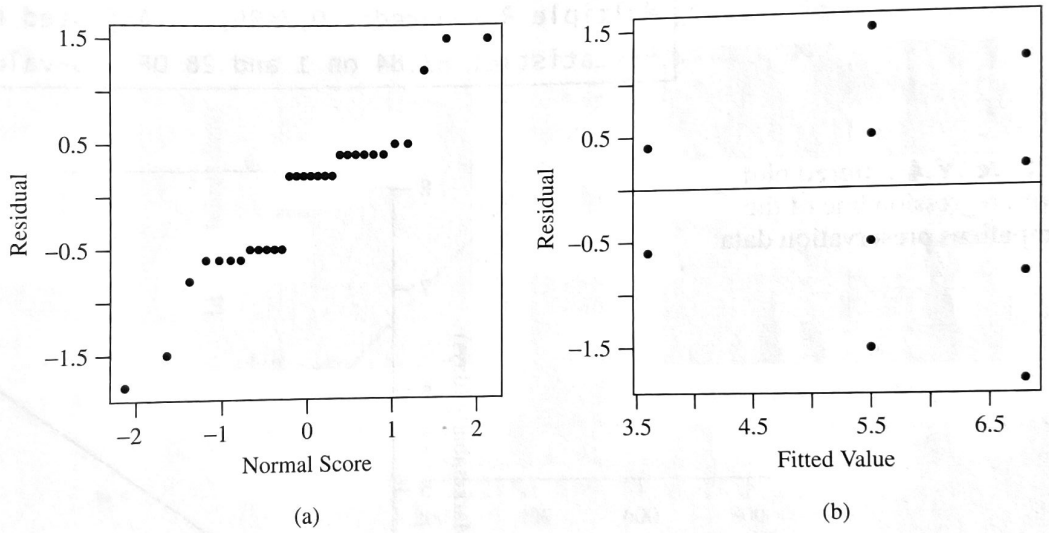


Figure IV.3 (a) Normal quantile plot of ANOVA residuals and (b) ANOVA residuals versus fitted values plot for the impatiens foliar spray study

Regression

Looking back on Figures IV.1 and IV.2, there is a clear upward trend in preservation duration as the dose of abscisic acid is increased in the spray. If the relationship between these numeric variables is linear, we can investigate its statistical significance using regression.† Figure IV.4 shows a plot of days of marketability against dose (with some jittering to better reveal the data since both dose and days of marketability are discretely observed). A regression line is also included in the plot to draw attention to the positive linear trend. Table IV.5 provides numeric summaries of the data for a regression model, and Table IV.6 provides computer output from fitting a regression model.

*Some would argue that the discreteness of the bands indicates nonnormality because the normal distribution characterizes continuous, not discrete, random variables. Although this is true, the model still serves as a useful approximation.

†If the relationship were nonlinear, say quadratic, regression methods could still be used, but more complex models would be required.

Table IV.5 Summary of impatiens preservation data

	X = Dose (PPM)	Y = Marketability (days)
Mean	300.00	5.3000
SD	249.14	1.5345
	$r = 0.8658$	

Table IV.6 Computer output from regressing days of marketability on dose

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.7000000  0.2255945  16.401 6.86e-16 ***
Dose        0.0053333  0.0005825   9.156 6.49e-10 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
Residual standard error: 0.7815 on 28 degrees of freedom
Multiple R-squared:  0.7496,    Adjusted R-squared:  0.7407
F-statistic: 83.84 on 1 and 28 DF,  p-value: 6.492e-10
```

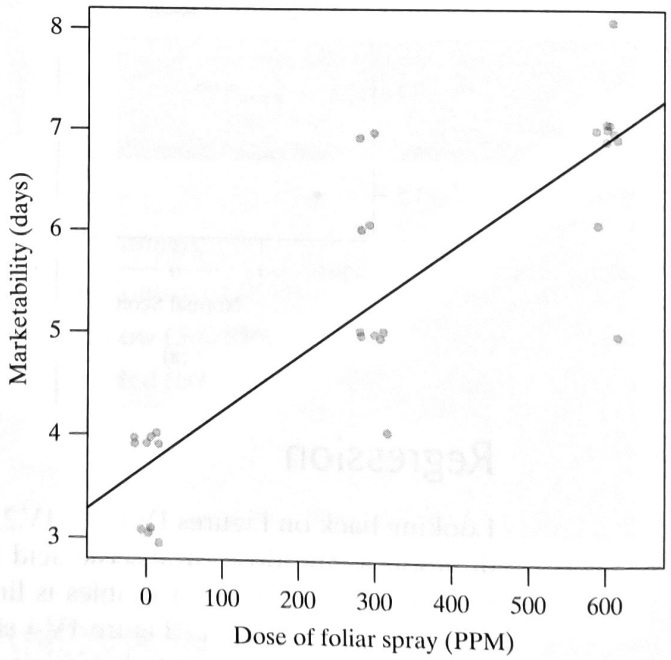


Figure IV.4 Jittered plot with regression line of the impatiens preservation data

From the output in Table IV.6, we can test the null hypothesis that there is no linear relationship between dose and marketability—that the slope of the regression line is zero ($\beta_1 = 0$)—against the alternative (directional) hypothesis that there is a positive linear relationship between dose and marketability ($\beta_1 > 0$). t_s is reported to be 9.156 with a P-value of 6.5×10^{-10}. This small P-value indicates that observing a linear association at least as strong as that observed in these data is extremely unlikely if dose and marketability are unrelated (i.e., results as extreme as these would occur less than 0.01% of the time due to chance). These data provide very strong evidence that increasing the dose provides greater preservation.

Using the output in Table IV.6, we obtain the equation of the regression line: $\hat{y} = 3.70 + 0.00533x$ indicating that for each PPM increase in the dose of abscisic acid, plants are preserved an additional 0.00533 days, on average. Equivalently, for each 100 PPM increase in the dose of abscisic acid, plants are preserved an additional 0.533 days on average. A 95% confidence interval for the slope of the regression line is $0.00533 \pm 2.0484 \times 0.0005825$ or (0.00414, 0.00652) days/PPM.

As discussed in Section 12.4, we must be careful to avoid extrapolating the results of regression models. The predicted increase in preservation is not likely to be valid for doses outside of the range of those studied, nor could we use the model to predict marketability duration for doses much above 600 PPM. If we were to use this model to estimate the mean marketability duration for plants sprayed with a 900 PPM dose, we would predict the plants to last on average $\hat{y} = 3.70 + 0.00533 \times 900 = 8.5$ days. It turns out, however, that for this experiment, the researcher actually did study this dose. At 900 PPM, the mean number of days of marketability was reported to be 6.7 days—much lower than our extrapolated value of 8.5 days. Clearly, the upward linear trend does not continue indefinitely as shown in Figure IV.5.

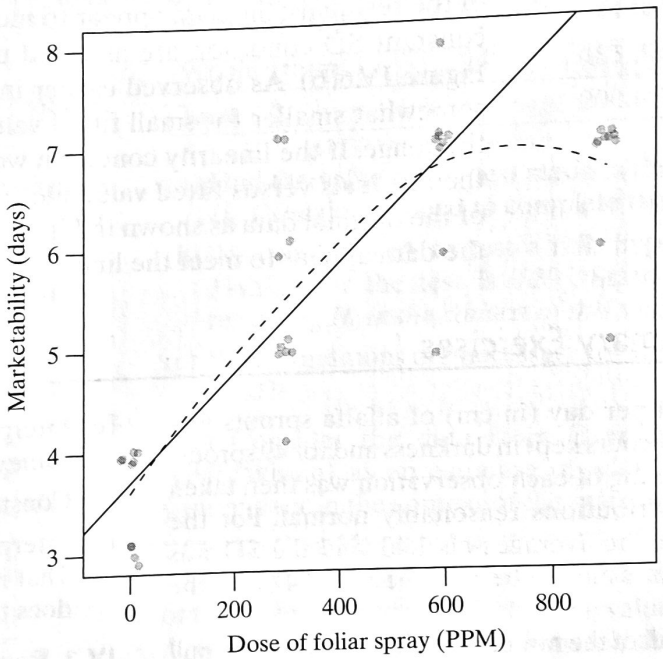


Figure IV.5 Jittered plot of the impatiens preservation data with regression model based only on doses 0 through 600 PPM (solid line) and loess curve based on all doses 0 through 900 PPM (dashed line)

Checking Regression Requirements

The requirements for the linear regression model are very similar to the requirements of the ANOVA model: (1) random sampling (the random subsampling or bivariate random sampling model); (2) normal distribution of Y for each fixed X; and (3) constant standard deviation (σ_e does not depend on X); a further condition for regression (not ANOVA) is (4) linearity: $\mu_{Y|X} = \beta_0 + \beta_1 X$, that the mean value of Y changes linearly with X.

As discussed earlier in the context of ANOVA, we don't know if the 30 plants were truly chosen at random, but we'll regard them as such for this example. Because the researchers set the doses of spray and observed the days of marketability, these data fit the random subsampling model (X fixed, Y random). The normality condition may be evaluated by examining a normal quantile plot of the residuals as shown in Figure IV.6 (a). Because the ANOVA and regression models are slightly different, the ANOVA residuals in Figure IV.3 (a) are slightly different

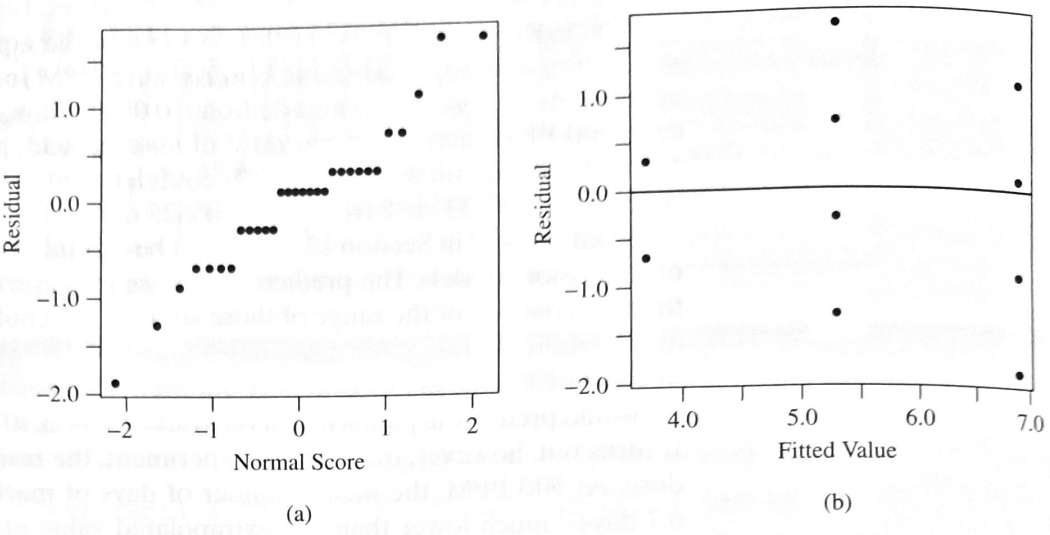


Figure IV.6 (a) Normal quantile plot of regression residuals and (b) regression residuals versus fitted values plot for the impatiens foliar spray study

from those in Figure IV.6(a). The linearity of this plot is consistent with normality of the residuals; the data appear to meet the normality condition. The linearity and constant SD condition are assessed using the residual versus fitted value plot in Figure IV.6(b). As observed earlier in the ANOVA context, the SD appears to be somewhat smaller for small fitted values, but overall the SD is fairly stable across the range. If the linearity condition were not met, we would see curved patterns in the residuals versus fitted value plot (as well as a curved pattern in the scatterplot of the original data as shown in Figure IV.4). No curved patterns are observed; thus, the data appear to meet the linearity condition.

Unit IV Summary Exercises

IV.1 The growth per day (in cm) of alfalfa sprouts was recorded for 50 sprouts kept in darkness and for 49 sprouts kept in light. The log of each observation was then taken to make the distributions reasonably normal. For the darkness sample, the average was 1.40, and the SD was 0.92. For the light sample, the average was 0.47, and the SD was 0.45.

George has all of the raw data and wants to test the null hypothesis that the two population means are equal. He wants to use ANOVA to do this test. Is this a good idea? Discuss (i) whether this is *possible* (saying why or why not, and if not, saying what you would do instead of ANOVA) and (ii) whether this is *wise* (saying why or why not, and if not, saying what you would do instead of ANOVA).

IV.2 Researchers measured initial weight, X, and weight gain, Y, of 15 rats on a high protein diet.[1] All weights are in grams. A scatterplot of the data shows a linear relationship. The fitted regression model is

$$\hat{y} = 54.95 + 1.06x$$

The sample correlation coefficient, r, is 0.489. The SE of b_1 is 0.526. Also, $s_e = 19.3$.

(a) Find r^2 and interpret r^2 in the context of this problem.

(b) Suppose that a rat initially weighs 60 g. What is the predicted weight gain for the rat?

(c) Interpret the value 1.06 from the fitted model *in the context of this problem*. (What does this 1.06 mean?)

(d) Construct a 95% confidence interval for β_1.

(e) Interpret the value of s_e *in the context of this problem*. That is, what does it mean to say that $s_e = 19.3$? How does this relate to your answer to part (b)?

IV.3 Researchers wanted to compare two drugs, formoterol and salbutamol, in aerosol solution, for the treatment of patients who suffer from exercise-induced asthma.[2] Patients were to take a drug, do some exercise, and then have their "forced expiratory volume" measured. There were 30 subjects available.

(a) Explain how to set up a randomized blocks design (RBD) here using age as the blocking variable and five blocks.

(b) How would an RBD be a helpful? That is, what is the main advantage of using a RBD in a setting like this?

IV.4 A confused researcher finds a dime on the sidewalk and wants to test H_0: $p = 0.5$ against H_A: $p \neq 0.5$ where p = Pr[Heads] when tossing the coin. This dime is an ordinary coin for which $p = 0.5$—but she doesn't know that. She tosses the coin 100 times, finds the P-value for a goodness-of-fit test, and compares it to $\alpha = 0.05$. However, if she retains H_0 (because the P-value is large),

then she discards the first sample and gets a new sample by tossing the coin 100 more times and repeating the goodness-of-fit test with the new data. If she retains H_0 for this test, then she discards the data and collects a third sample and does another goodness-of-fit test, after which she stops no matter what. What is the probability that she will make a type I error?

IV.5 A researcher collected data on a random sample of 12 breakfast cereals. He recorded x = fiber (in grams/ounce) and y = price (in cents/ounce). A scatterplot of the data shows a linear relationship. The fitted regression model is

$$\hat{y} = 17.42 + 0.62x$$

The sample correlation coefficient, r, is 0.23. The SE of b_1 is 0.81. Also, $s_e = 3.1$.

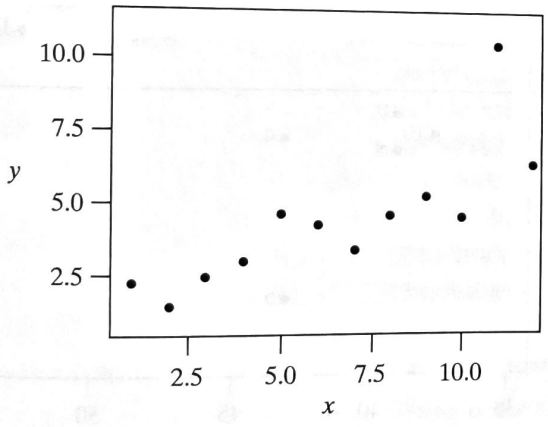
(a) Find r^2 and interpret r^2 in the context of this problem.

(b) Suppose that a cereal has 2.63 grams of fiber/ounce and costs 17.3 cents/ounce. What is the residual for this cereal?

(c) Interpret the value of s_e *in the context of this problem.* That is, what does it mean to say that $s_e = 3.1$?

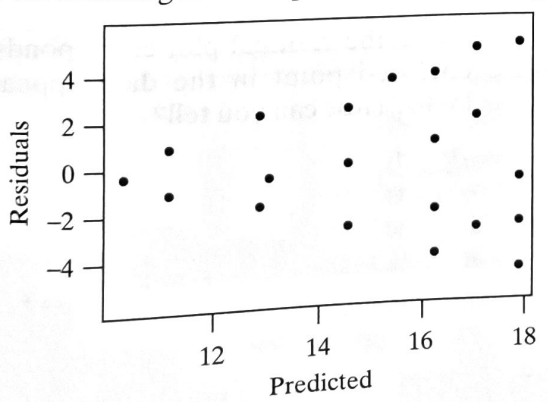
IV.6 Consider a regression setting in which we construct a scatterplot, fit the regression model $\hat{y}_i = b_0 + b_1 x_i$, and generate a residual plot.

(a) Suppose the scatterplot of y versus x is as follows:



Draw a sketch of the resulting residual plot. Label the axes on your graph.

(b) Suppose we fit a regression line to a new set of data and the resulting residual plot is as follows:



Draw a sketch of the scatterplot of y versus x. Label the axes on your graph.

IV.7 A researcher measured the number of tree species per 0.1 hectare plot along the Black, Huron, and Vermilion rivers. The data are summarized in the table below:

	Black	Huron	Vermilion
Mean	9.33	9.89	11.11
Median	10	11	11
SD	3.16	2.42	2.71
n	9	9	9

Here is a partial ANOVA table summarizing the results.

Source	Degree of freedom	Sum of squares	Mean square	F
Between groups		14.889		
Within groups	24	185.778		
Total		200.667		

(a) Find the value of the test statistic that is used to test H_0. (You do not need to complete the test.)

(b) We use ANOVA to test a null hypothesis, H_0. The P-value for the test is 0.397. State the conclusion regarding H_0 *in the context of this setting.*

(c) What conditions are necessary for the ANOVA to be valid?

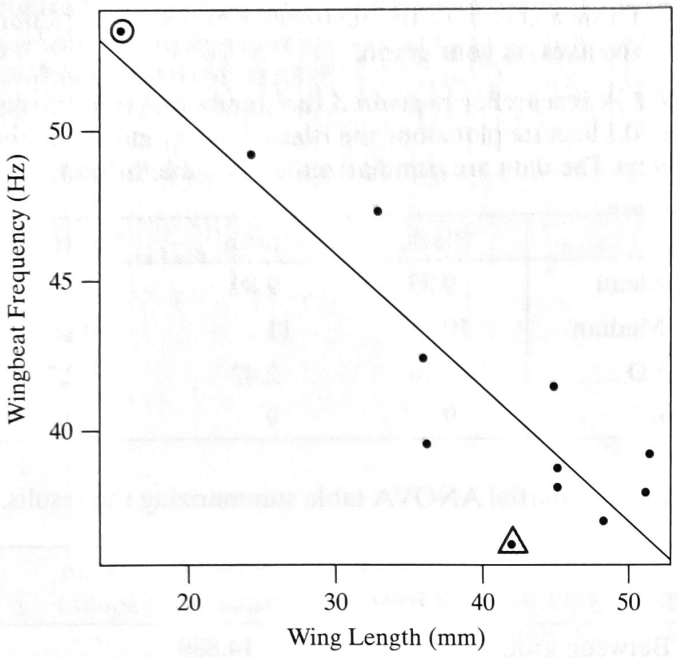
IV.8 Consider the data from Exercise IV.7. We use $\sqrt{\text{MS(Within)}}$ as an estimate of what quantity? Give your answer in the context of the question.

IV.9 (a) Consider the data from Exercise IV.7. Suppose we want to compare the Vermilion River to the average of the other two rivers. Calculate the value of the contrast, L, to measure the difference between the mean number of species (per 0.1 hectare) along the Vermilion River and the mean number of species (per 0.1 hectare) along the other two rivers.

(b) Find the SE of the contrast L from part (a). (Don't make an interval or do a test.)

IV.10 Is there a relationship between wing length (mm) and wingbeat frequency (Hz) among hummingbirds? In one study, researchers measured the wing lengths and wingbeat frequencies of 12 hummingbirds.[3] The following are basic summaries and a plot of the 12 data values.

Variable	Mean	SD
Wing length (mm)	39.4034	11.0090
Wingbeat frequency (Hz)	41.6858	5.4176
$r = -0.9061$		



(a) If the circled point was removed from this dataset, would the value of the sample correlation increase in magnitude, decrease in magnitude, or stay about the same? Briefly explain.

(b) The residual SD for all these data (including the points in the circle and triangle) is $s_e = 2.40$ Hz. If the point in the triangle were removed, would the residual SD increase, decrease, or stay about the same? Briefly explain.

IV.11 Consider the hummingbird data in Exercise IV.10.

(a) What percentage of the variation in wingbeat frequency is explained by the relationship between wing length and wingbeat frequency?

(b) Formally speaking, an important adjective is missing from part (a) that describes the nature of the relationship being described. What is the missing adjective?

IV.12 Consider the hummingbird data and information provided in Exercise IV.10.

(a) Find the equation of the fitted regression line.

(b) Predict the wingbeat frequency for a hummingbird with 30 mm wings.

(c) Predict the mean wingbeat frequency for hummingbirds with 30 mm wings.

(d) Are the predictions of (b) and (c) interpolations or extrapolations? Briefly explain.

(e) Hummingbird B has wings that are 1mm longer than the wings of hummingbird A. How much faster or slower would you expect hummingbird B's wings to beat compared to A's?

IV.13 Consider the hummingbird data and information provided in Exercise IV.10.

(a) Do longer wings tend to beat more slowly (i.e., lower frequency) than shorter wings? In plain English, what are the null and alternative hypotheses to be tested?

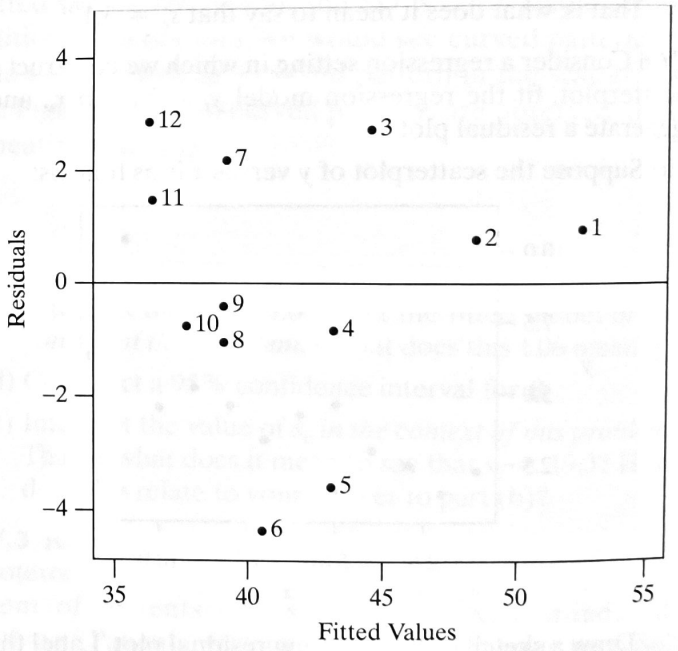
(b) Compute the value of t_s used to test the hypothesis in part (a) for testing H_A: $\rho < 0$.

(c) Compute the value of t_s used to test the hypothesis in part (a) for testing H_A: $\beta < 0$.

(d) Explain why your answers to (b) and (c) agree.

(e) The P-value for the test of H_A: $\rho \neq 0$ is 0.004. Without using a table or computer, what is the P-value for the tests in parts (b) and (c)?

IV.14 The following is a plot of the residuals against the fitted values for the hummingbird data of Exercise IV.10.



(a) Which point in the residual plot corresponds to the circled point in the data appearing in Exercise IV.10? How can you tell?

(b) Which point in the residual plot corresponds to the triangle enclosed point in the data appearing in Exercise IV.10? How can you tell?