

3.3: Chapter 13- Linear Regression

Key Terms

curvilinear models

error

least squares error solution

line of best fit

mean squared error

moderation models

multiple regression

residual

standard error of the estimate

In Chapter 11, we learned about ANOVA, which involves a new way of looking at how our data are structured and the inferences we can draw from that. In Chapter 12, we learned about correlations, which analyze two continuous variables at the same time to see if they systematically relate in a linear fashion. In this chapter, we will combine these two techniques in an analysis called simple linear regression, or regression for short. Regression uses the technique of variance partitioning from ANOVA to more formally assess the types of relationships looked at in correlations. Regression is the most general and most flexible analysis covered in this book, and we will only scratch the surface.

Line of Best Fit

In correlations, we referred to a linear trend in the data. That is, we assumed that there was a straight line we could draw through the middle of our scatter plot that would represent the relationship between our two variables, X and Y. Regression involves solving for the equation of that line, which is called the line of best fit.

The line of best fit can be thought of as the central tendency of our scatter plot. The term best fit means that the line is as close to all points (with each point representing both variables for a single person) in the scatter plot as possible, with a balance of scores above and below the line. This is the same idea as the mean, which has an equal weighting of scores above and below it and is the best singular descriptor of all our data points for a single variable.

We have already seen many scatter plots in Chapter 2 and Chapter 12, so we know by now that no scatter plot has points that form a perfectly straight line. Because of this, when we put a straight line through a scatter plot, it will not touch all of the points, and it may not even touch any! This will result in some distance between the line and each of the points it is supposed to represent, just like a mean has some distance between it and all of the individual scores in the dataset.

The distances between the line of best fit and each individual data point go by two different names that mean the same thing: errors and residuals. The term error in regression is closely aligned with the meaning of error in statistics (think standard error or sampling error); it does not mean that we did anything wrong, it simply means that there was some discrepancy or difference between what our analysis produced and the true value we are trying to get at. The term residual is new to our study of statistics, and it takes on a very similar meaning in regression to what it means in everyday parlance: there is something left over. In regression, what is “left over”—that is, what makes up the residual—is an imperfection in our ability to predict values of the Y variable using our line. This definition brings us to one of the primary purposes of regression and the line of best fit: predicting scores.

Prediction

The goal of regression is the same as the goal of ANOVA: to take what we know about one variable (X) and use it to explain our observed differences in another variable (Y). In ANOVA, we talked about—and tested for—group mean differences, but in regression we do not have groups for our explanatory variable; we have a continuous variable, like in correlation. Because of this, our vocabulary will be a little bit different, but the process, logic, and end result are all the same.

In regression, we most frequently talk about prediction, specifically predicting our outcome variable Y from our explanatory variable X, and we use the line of best fit to make our predictions. Let’s take a look at the equation for the line, which is quite

simple:

$$\hat{Y} = a + bX$$

The terms in the equation are defined as:

\hat{Y} : the predicted value of Y for an individual person

a: the intercept of the line

b: the slope of the line

X: the observed value of X for an individual person

What this shows us is that we will use our known value of X for each person to predict the value of Y for that person. The predicted value, \hat{Y} (“Y-hat”), is our best guess for what a person’s score on the outcome is. Notice also that the form of the equation is very similar to very simple linear equations that you have likely encountered before and has only two parameter estimates: an intercept (where the line crosses the y-axis) and a slope (how steep—and the direction, positive or negative—the line is). These are parameter estimates because, like everything else in statistics, we are interested in approximating the true value of the relationship in the population but can only ever estimate it using sample data. We will soon see that one of these parameters, the slope, is the focus of our hypothesis tests (the intercept is only there to make the math work out properly and is rarely interpretable). The formulas for these parameter estimates use very familiar values:

$$a = M_Y - bM_X$$

$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{SP}{SS_X} = r \left(\frac{s_Y}{s_X} \right)$$

We have seen each of these before. M_Y and M_X are the means of Y and X, respectively; cov_{XY} is the covariance of X and Y we learned about with correlations; and s_X^2 is the variance of X. The formula for the slope is very similar to the formula for a Pearson correlation coefficient; the only difference is that we are dividing by the variance of X instead of the product of the standard deviations of X and Y. Because of this, our slope is scaled to the same scale as our X variable and is no longer constrained to be between 0 and 1 in absolute value. This formula provides a clear definition of the slope of the line of best fit, and just like with correlation, this definitional formula can be simplified into a short computational formula for easier calculations. In this case, we are simply taking the sum of products and dividing by the sum of squares for X.

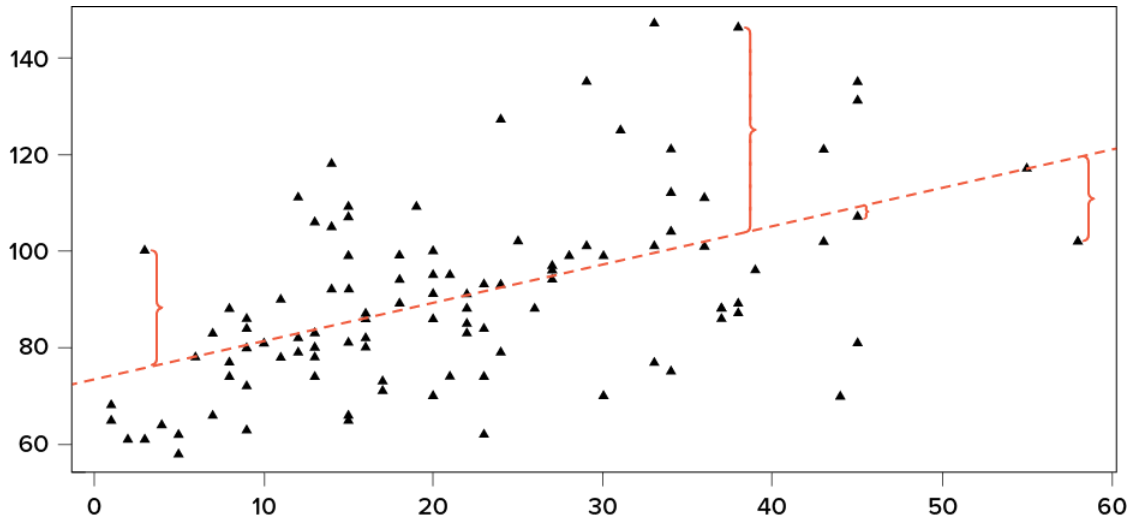
Notice that there is a third formula for the slope of the line that involves the correlation between X and Y. This is because regression and correlation look for the same thing: a straight line through the middle of the data. The only difference between a regression coefficient in simple linear regression and a Pearson correlation coefficient is the scale. So, if you lack raw data but have summary information on the correlation and standard deviations for variables, you can still compute a slope, and therefore an intercept, for a line of best fit.

It is important to point out that the Y values in the equations for a and b are our observed Y values in the dataset, not the predicted Y values (\hat{Y}) from our equation for the line of best fit. Thus, we will have three values for each person: the observed value of X (X), the observed value of Y (Y), and the predicted value of Y (\hat{Y}). You may be asking why we would try to predict Y if we have an observed value of Y, and that is a reasonable question. The answer has two explanations. First, we need to use known values of Y to calculate the parameter estimates in our equation, and we use the difference between our observed values and predicted values ($Y - \hat{Y}$) to see how accurate our equation is. And second, we often use regression to create a predictive model that we can then use to predict values of Y for other people for whom we only have information on X.

Let’s look at this from an applied example. Businesses often have more applicants for a job than they have openings available, so they want to know who among the applicants is most likely to be the best employee. There are many criteria that can be used, but one is a personality test for conscientiousness, with the belief being that more conscientious (more responsible) employees are better than less conscientious employees. A business might give their employees a personality inventory to assess conscientiousness and study existing performance data to look for a relationship. In this example, we have known values of the predictor (X, conscientiousness) and outcome (Y, job performance), so we can estimate an equation for a line of best fit and see how accurately conscientiousness predicts job performance, then use this equation to predict future job performance of applicants based only on their known values of conscientiousness from personality inventories given during the application process.

The key in assessing whether a linear regression works well is the difference between our observed and known Y values and our predicted \hat{Y} values. As mentioned in passing above, we use subtraction to find the difference between them ($Y - \hat{Y}$) in the same way we use subtraction for deviation scores and sums of squares. The value ($Y - \hat{Y}$) is our residual, which, as defined above, is how close our line of best fit is to our actual values. We can visualize residuals to get a better sense of what they are by creating a scatter plot and overlaying a line of best fit on it, as shown in Figure 13.1.

Figure 13.1. Scatter plot with residuals. (“Scatter Plot with Residuals” by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)



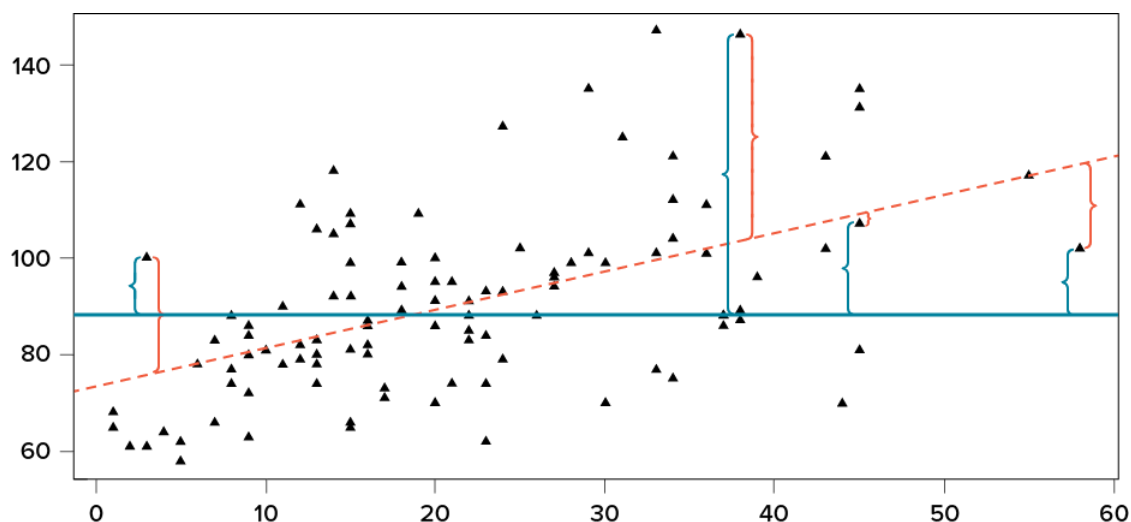
In Figure 13.1, the triangular dots represent observations from each person on both X and Y and the dotted line is the line of best fit estimated by the equation $\hat{Y} = a + bX$. For every person in the dataset, the line represents their predicted score. The brackets between the triangular dots and the predicted scores on the line of best fit are our residuals. (For ease of viewing, they are only drawn for four observations, but in reality there is one for every observation.) You can see that some residuals are positive and some are negative, and that some are very large and some are very small. This means that some predictions are very accurate and some are very inaccurate, and that some predictions overestimate values and some underestimate values. Across the entire dataset, the line of best fit is the one that minimizes the total (sum) value of all residuals. That is, although predictions at an individual level might be somewhat inaccurate, across our full sample and (theoretically) in future samples our total amount of error is as small as possible. We call this property of the line of best fit the least squares error solution. This term means that the solution—or equation—of the line is the one that provides the smallest possible value of the squared errors (squared so that they can be summed, just like in standard deviation) relative to any other straight line we could draw through the data.

Predicting Scores and Explaining Variance

We have now seen that the purpose of regression is twofold: we want to predict scores based on our line and, as stated earlier, explain variance in our observed Y variable just like in ANOVA. These two purposes go hand in hand, and our ability to predict scores is literally our ability to explain variance. That is, if we cannot account for the variance in Y based on X, then we have no reason to use X to predict future values of Y.

We know that the overall variance in Y is a function of each score deviating from the mean of Y (as in our calculation of variance and standard deviation). So, just like the brackets in Figure 13.1 representing residuals, given as $(Y - \hat{Y})$, we can visualize the overall variance as each score's distance from the overall mean of Y, given as $(Y - M_Y)$, our normal deviation score. This is shown in Figure 13.2.

Figure 13.2. Scatter plot with residuals and deviation scores. (“Scatter Plot with Residuals and Deviation Scores” by Judy Schmitt is licensed under CC BY-NC-SA 4.0.)



In Figure 13.2, the solid line is the mean of Y, and the blue brackets are the deviation scores between our observed values of Y and the mean of Y. This represents the overall variance that we are trying to explain. Thus, the residuals and the deviation scores are the same type of idea: the distance between an observed score and a given line, either the line of best fit that gives predictions or the line representing the mean that serves as a baseline. The difference between these two values, which is the distance between the lines themselves, is our model's ability to predict scores above and beyond the baseline mean; that is, it is our model's ability to explain the variance we observe in Y based on values of X. If we have no ability to explain variance, then our line will be flat (the slope will be 0.00) and will be the same as the line representing the mean, and the distance between the lines will be 0.00 as well.

We now have three pieces of information: the distance from the observed score to the mean, the distance from the observed score to the prediction line, and the distance from the prediction line to the mean. These are our three pieces of information needed to test our hypotheses about regression and to calculate effect sizes. They are our three sums of squares, just like in ANOVA. Our distance from the observed score to the mean is the sum of squares total, which we are trying to explain. Our distance from the observed score to the prediction line is our sum of squares error, or residual, which we are trying to minimize. Our distance from the prediction line to the mean is our sum of squares model, which is our observed effect and our ability to explain variance. Each of these will go into the ANOVA table to calculate our test statistic.

ANOVA Table

Our ANOVA table in regression follows the exact same format as it did for ANOVA (hence the name). The top row (Model) is our observed effect, the middle row is our error, and the bottom row is our total. The columns take on the same interpretations as well: from left to right, we have our sums of squares, our degrees of freedom, our mean squares, and our F statistic.

Source	SS	df	MS	F
Model	$\sum(\hat{Y} - M_Y)^2$	1	$\frac{SS_M}{df_M}$	$\frac{MS_M}{MS_E}$
Error	$\sum(Y - \hat{Y})^2$	$N - 2$	$\frac{SS_E}{df_E}$	
Total	$\sum(Y - M_Y)^2$	$N - 1$		

As with ANOVA, getting the values for the SS column is a straightforward but somewhat arduous process. First, you take the raw scores of X and Y and calculate the means, variances, and covariance using the sum of products table introduced in our [chapter on correlations](#). Next, you use the variance of X and the covariance of X and Y to calculate the slope of the line, b. (The [formula for calculating b](#) was provided earlier.) After that, you use the means and the slope to find the intercept, a, which is given alongside b. After that, you use the full prediction equation for the line of best fit to get predicted Y scores (\hat{Y}) for each person. Finally, you use the observed Y scores, predicted Y scores, and mean of Y (M_Y) to find the appropriate deviation scores for each person for each

sum of squares source in the table and sum them to get the sum of squares model, sum of squares error, and sum of squares total. As with ANOVA, you won't be required to compute the SS values by hand, but you will need to know what they represent and how they fit together.

The other columns in the ANOVA table are all familiar. The degrees of freedom column still has $N - 1$ for our total, but now we have $N - 2$ for our error degrees of freedom and 1 for our model degrees of freedom; this is because simple linear regression only has one predictor, so our degrees of freedom for the model is always 1 and does not change. The total degrees of freedom must still be the sum of the other two, so our degrees of freedom error will always be $N - 2$ for simple linear regression. The mean square columns are still the SS column divided by the df column, and the test statistic F is still the ratio of the mean squares. Based on this, it is now explicitly clear that not only do regression and ANOVA have the same goal but they are, in fact, the same analysis entirely. The only difference is the type of data we feed into the predictor side of the equations: continuous for regression and categorical for ANOVA.

Hypothesis Testing in Regression

Regression, like all other analyses, will test a null hypothesis in our data. In regression, we are interested in predicting Y scores and explaining variance using a line, the slope of which is what allows us to get closer to our observed scores than the mean of Y can. Thus, our hypotheses concern the slope of the line, which is estimated in the prediction equation by b (the slope of a population, as opposed to b , which is the slope of a sample). Specifically, we want to test that the slope is not zero:

H_0 : There is no explanatory relationship between our variables

$$H_0: \beta = 0$$

H_A : There is an explanatory relationship between our variables

$$H_A: \beta > 0$$

$$H_A: \beta < 0$$

$$H_A: \beta \neq 0$$

A non-zero slope indicates that we can explain values in Y based on X and therefore predict future values of Y based on X . Our alternative hypotheses are analogous to those in correlation: positive relationships have values above zero, negative relationships have values below zero, and two-tailed tests are possible. Just like ANOVA, we will test the significance of this relationship using the F statistic calculated in our ANOVA table compared to a critical value from the F distribution table. Let's take a look at an example and regression in action.

Example Happiness and Well-being

Researchers are interested in explaining differences in how happy people are based on how healthy people are. They gather data on each of these variables from 18 people and fit a linear regression model to explain the variance. We will follow the four-step hypothesis-testing procedure to see if there is a relationship between these variables that is statistically significant.

Step 1: State the Hypotheses

The null hypothesis in regression states that there is no relationship between our variables. The alternative states that there is a relationship, but because our research description did not explicitly state a direction of the relationship, we will use a non-directional hypothesis.

H_0 : There is no explanatory relationship between health and happiness

$$H_0: \beta = 0$$

H_A : There is an explanatory relationship between health and happiness

$$H_A: \beta \neq 0$$

Step 2: Find the Critical Value

Because regression and ANOVA are the same analysis, our critical value for regression will come from the same place: the F distribution table, which uses two types of degrees of freedom. We saw in the [ANOVA table](#) that the degrees of freedom for our numerator—the Model line—is always 1 in simple linear regression, and that the denominator degrees of freedom—from the Error line—is $N - 2$. In this instance, we have 18 people so our degrees of freedom for the denominator is 16. Going to our F table (a

portion of which is shown in [Table 13.1](#)), we find that the appropriate critical value for 1 and 16 degrees of freedom is $F^* = 4.49$. (The complete F table can be found in [Appendix C](#).)

Table 13.1. Critical Values for F (F Table)

df:De nomi nator (With in)	df:Numerator (Between)														
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28

Step 3: Calculate the Test Statistic and Effect Size

The process of calculating the test statistic for regression first involves computing the parameter estimates for the line of best fit. To do this, we first calculate the means, standard deviations, and sum of products for our X and Y variables, as shown below.

X	(X - MX)	(X - MX) ²	Y	(Y - MY)	(Y - MY) ²	(X - MX)(Y - MY)
17.65	-2.13	4.53	10.36	-7.10	50.37	15.10
16.99	-2.79	7.80	16.38	-1.08	1.16	3.01
18.30	-1.48	2.18	15.23	-2.23	4.97	3.29
18.28	-1.50	2.25	14.26	-3.19	10.18	4.79
21.89	2.11	4.47	17.71	0.26	0.07	0.55
22.61	2.83	8.01	16.47	-0.98	0.97	-2.79
17.42	-2.36	5.57	16.89	-0.56	0.32	1.33
20.35	0.57	0.32	18.74	1.29	1.66	0.73

X	(X - MX)	(X - MX) ²	Y	(Y - MY)	(Y - MY) ²	(X - MX)(Y - MY)
18.89	-0.89	0.79	21.96	4.50	20.26	-4.00
18.63	-1.15	1.32	17.57	0.11	0.01	-0.13
19.67	-0.11	0.01	18.12	0.66	0.44	-0.08
18.39	-1.39	1.94	12.08	-5.37	28.87	7.48
22.48	2.71	7.32	17.11	-0.34	0.12	-0.93
23.25	3.47	12.07	21.66	4.21	17.73	14.63
19.91	0.13	0.02	17.86	0.40	0.16	0.05
18.21	-1.57	2.45	18.49	1.03	1.07	-1.62
23.65	3.87	14.99	22.13	4.67	21.82	18.08
19.45	-0.33	0.11	21.17	3.72	13.82	-1.22
356.02	0.00	76.14	314.18	0.00	173.99	58.29

From the raw data in our X and Y columns, we find that the means are $M_X = 19.78$ and $M_Y = 17.45$. The deviation scores for each variable sum to zero, so all is well there. The sums of squares for X and Y ultimately lead us to standard deviations of $s_X = 2.12$ and $s_Y = 3.20$. Finally, our sum of products is 58.29, which gives us a covariance of $\text{cov}_{XY} = 3.43$, so we know our relationship will be positive. This is all the information we need for our equations for the line of best fit.

First, we must calculate the slope of the line:

$$b = \frac{SP}{SS_X} = \frac{58.29}{76.14} = 0.77$$

This means that as X changes by 1 unit, Y will change by 0.77. In terms of our problem, as health increases by 1, happiness goes up by 0.77, which is a positive relationship. Next, we use the slope, along with the means of each variable, to compute the intercept:

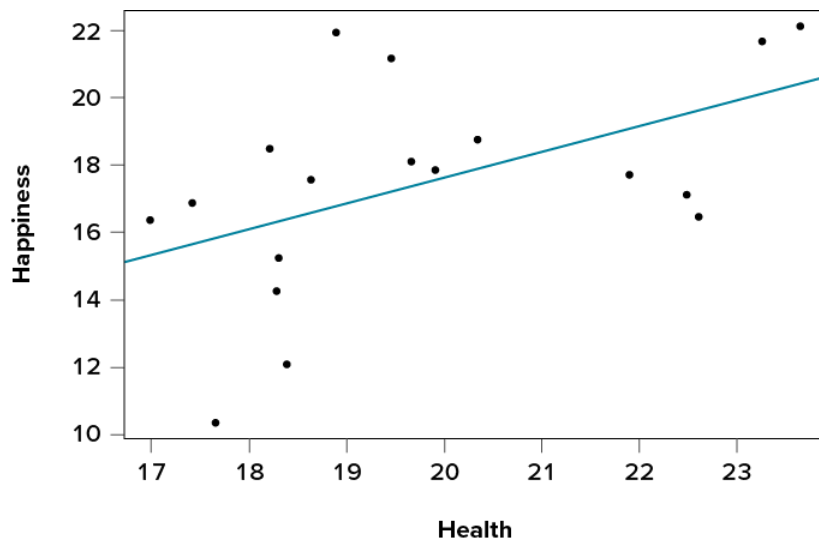
$$\begin{aligned}
 a &= M_Y - bM_X \\
 &= 17.45 - (0.77)(19.78) \\
 &= 17.45 - 15.03 \\
 &= 2.42
 \end{aligned}$$

For this particular problem (and most regressions), the intercept is not an important or interpretable value, so we will not read into it further. Now that we have all of our parameters estimated, we can give the full equation for our line of best fit:

$$\hat{Y} = 2.42 + 0.77X$$

We can plot this relationship in a scatter plot and overlay our line onto it, as shown in [Figure 13.3](#).

Figure 13.3. Health and happiness data and line of best fit. (“[Scatter Plot Health and Happiness](#)” by Judy Schmitt is licensed under [CC BY-NC-SA 4.0.](#))



We can use the line equation to find predicted values for each observation and use them to calculate our sums of squares model and error, but this is tedious to do by hand, so we will let the computer software do the heavy lifting in that column of our ANOVA table:

Source	SS	df	MS	F
Model	44.62			
Error	129.37			
Total				

Now that we have these, we can fill in the rest of the ANOVA table. We already found our degrees of freedom in Step 2:

Source	SS	df	MS	F
Model	44.62	1		
Error	129.37	16		
Total				

Our total line is always the sum of the other two lines, giving us:

Source	SS	df	MS	F
Model	44.62	1		
Error	129.37	16		
Total	173.99	17		

Our mean squares column is only calculated for the model and error lines and is always our SS divided by our df, which is:

Source	SS	df	MS	F
Model	44.62	1	44.62	
Error	129.37	16	8.09	
Total	173.99	17		

Finally, our F statistic is the ratio of the mean squares:

Source	SS	df	MS	F
Model	44.62	1	44.62	5.52
Error	129.37	16	8.09	
Total	173.99	17		

This gives us an obtained F statistic of 5.52, which we will use to test our hypothesis.

Effect Size in Regression

We know that, because our statistic is significant, we should calculate an effect size. In regression, our effect size is variance explained, just like it was in ANOVA. Instead of using h^2 to represent this, we instead use R^2 , as we saw in correlation—yet more evidence that all of these are the same analysis. (Note that in regression analysis, R^2 is typically capitalized, although for simple linear regression it represents the same value as r^2 we used in correlation.) Variance explained is still the ratio of SSM to SST:

$$R^2 = \frac{SS_M}{SS_T} = \frac{44.62}{173.99} = .26$$

We are explaining 26% of the variance in happiness based on health, which is a large effect size. (R^2 uses the same effect size cutoffs as h^2 .)

Step 4: Make the Decision

We now have everything we need to make our final decision. Our obtained test statistic was $F = 5.52$ and our critical value was $F^* = 4.49$. Since our obtained test statistic is greater than our critical value, we can reject the null hypothesis.

Reject H_0 . Based on our sample of 18 people, we can predict levels of happiness based on how healthy someone is, and the effect size was large, $F(1, 16) = 5.52$, $p < .05$, $R^2 < .26$.

Figure 13.4 shows the output from JASP for this example.

Figure 13.4. Output from JASP for the linear regression described in the Happiness and Well-Being example. The output provides the slope (b) of the line in the Coefficients table under Health Unstandardized (0.765), the y-intercept (a) of the line in the Coefficients table under (Intercept) Unstandardized (2.317, note slightly different from hand calculations due to rounding). The Model Summary – Happiness table provides R^2 (.256). The ANOVA table provides the result of the F test statistic. Based on our sample of 18 people, we can predict levels of happiness based on how healthy someone is, $F(1, 16) = 5.5$, $p = .032$. (“JASP linear regression” by Rupa G. Gordon/Judy Schmitt is licensed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/).)

Linear Regression

Model Summary - Happiness

Model	R	R ²	Adjusted R ²	RMSE
H ₀	0.506	0.256	0.210	2.844

Note. Null model includes Health

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H ₀	Regression	44.562	1	44.562	5.509	0.032
	Residual	129.431	16	8.089		
	Total	173.993	17			

Note. Null model includes Health

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H ₀	(Intercept)	2.317	6.485		0.357	0.726
	Health	0.765	0.326	0.506	2.347	0.032

Accuracy in Prediction

We found a large, statistically significant relationship between our variables, which is what we hoped for. However, if we want to use our estimated line of best fit for future prediction, we will also want to know how precise or accurate our predicted values are. What we want to know is the average distance from our predictions to our actual observed values, or the average size of the residual ($Y - \hat{Y}$). The average size of the residual is known by a specific name: the standard error of the estimate ($s_{(Y-\hat{Y})}$), which is given by the formula

$$s_{(Y-\hat{Y})} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N - 2}}$$

This formula is almost identical to our standard deviation formula, and it follows the same logic. We square our residuals, add them up, then divide by the degrees of freedom. Although this sounds like a long process, we already have the sum of the squared residuals in our ANOVA table! In fact, the value under the square root sign is just the SSE divided by the dfE, which is called the mean squared error, or MSE:

$$s_{(Y-\hat{Y})} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N - 2}} = \sqrt{MS_E}$$

For our example:

$$s_{(Y-\hat{Y})} = \sqrt{\frac{129.37}{16}} = \sqrt{8.09} = 2.84$$

So on average, our predictions are just under 3 points away from our actual values. There are no specific cutoffs or guidelines for how big our standard error of the estimate can or should be; it is highly dependent on both our sample size and the scale of our

original Y variable, so expert judgment should be used. In this case, the estimate is not that far off and can be considered reasonably precise.

Multiple Regression and Other Extensions

Simple linear regression as presented here is only a stepping stone toward an entire field of research and application. Regression is an incredibly flexible and powerful tool, and the extensions and variations on it are far beyond the scope of this chapter. (Indeed, even entire books struggle to accommodate all possible applications of the simple principles laid out here.) The next step in regression is to study multiple regression, which uses multiple X variables as predictors for a single Y variable at the same time. The math of multiple regression is very complex but the logic is the same: we are trying to use variables that are statistically significantly related to our outcome to explain the variance we observe in that outcome. Other forms of regression include curvilinear models that can explain curves in the data rather than the straight lines used here, as well as moderation models that change the relationship between two variables based on levels of a third. The possibilities are truly endless and offer a lifetime of discovery.

Exercises

1. How are ANOVA and linear regression similar? How are they different?
2. What is a residual?
3. How are correlation and regression similar? How are they different?
4. What are the two parameters of the line of best fit, and what do they represent?
5. What is our criterion for finding the line of best fit?
6. Fill out the rest of the ANOVA tables below for simple linear regressions:

1.	Source	SS	df	MS	F
	Model	34.21			
	Error				
	Total	66.12	54		

2.	Source	SS	df	MS	F
	Model			6.03	
	Error		16		
	Total	19.98			

7. In Chapter 12, we found a statistically significant correlation between overall performance in class and how much time someone studied. Use the summary statistics calculated in that problem ($M_X = 1.61$, $s_X = 1.12$, $M_Y = 2.95$, $s_Y = 0.99$, $r = .65$) to compute a line of best fit predicting success from study times.
8. Using the line of best fit equation created in Problem 7, predict the scores for how successful people will be based on how much they study:
 1. $X = 1.20$
 2. $X = 3.33$
 3. $X = 0.71$
 4. $X = 4.00$
9. You have become suspicious that the draft rankings of your fantasy football league have no predictive value for how teams place at the end of the season. You go back to historical league data and find rankings of teams after the draft and at the end of the season (shown in the following table) to test for a statistically significant predictive relationship. Assume $SSM = 2.65$ and $SSE = 337.35$.

Draft Projection	Final Rankings
1	14

Draft Projection	Final Rankings
2	6
3	8
4	13
5	2
6	15
7	4
8	10
9	11
10	16
11	9
12	7
13	14
14	12
15	1
16	5

10. You have summary data for two variables: how extroverted someone is (X) and how often someone volunteers (Y). Using these values, calculate the line of best fit predicting volunteering from extroversion, then test for a statistically significant relationship using the hypothesis-testing procedure: $M_X = 12.58$, $s_X = 4.65$, $M_Y = 7.44$, $s_Y = 2.12$, $r = .34$, $N = 67$, $SSM = 19.79$, $SSE = 215.77$.

Answers to Odd-Numbered Exercises

1)

ANOVA and simple linear regression both take the total observed variance and partition it into pieces that we can explain and cannot explain and use the ratio of those pieces to test for significant relationships. They are different in that ANOVA uses a categorical variable as a predictor, whereas linear regression uses a continuous variable.

3)

Correlation and regression both involve taking two continuous variables and finding a linear relationship between them. Correlations find a standardized value describing the direction and magnitude of the relationship, whereas regression finds the line of best fit and uses it to partition and explain variance.

5)

Least squares error solution; the line that minimizes the total amount of residual error in the dataset

7)

$b = r \cdot (s_Y / s_X) = .65 \cdot (0.99 / 1.12) = .57$; $a = M_X - b M_Y = 2.95 - (0.57)(1.61) = 3.87$; $Y = 3.87 + 0.57X$

9)

Step 1: $H_0: b = 0$ "There is no predictive relationship between draft rankings and final rankings in fantasy football," $H_A: b \neq 0$ "There is a predictive relationship between draft rankings and final rankings in fantasy football."

Step 2: Our model will have 1 (based on the number of predictors) and 14 (based on how many observations we have) degrees of freedom, giving us a critical value of $F^* = 4.60$.

Step 3: Using the sum of products table, we find $M_X = 8.50$, $M_Y = 8.50$, $SS_X = 339.86$, $SP = 29.99$, giving us a line of best fit of: $b = 29.99 / 339.86 = 0.09$; $a = 8.50 - (0.09)(8.50) = 7.74$; $Y = 7.74 + 0.09X$. Our given SS values and our df from Step 2 allow us to fill in the ANOVA table:

Source	SS	df	MS	F
Model	2.65	1	2.65	0.11
Error	337.35	14	24.10	
Total	339.86	15		

Step 4: Our obtained value was smaller than our critical value, so we fail to reject the null hypothesis. There is no evidence to suggest that draft rankings have any predictive value for final fantasy football rankings, $F(1, 14) = 0.11$, $p > .05$

This page titled [3.3: Chapter 13- Linear Regression](#) is shared under a [not declared](#) license and was authored, remixed, and/or curated by [Linda R. Cote](#), [Rupa G. Gordon](#), [Chrislyn E. Randell](#), [Judy Schmitt](#), and [Helena Marvin](#).