CHAPTER 4

Linear regression and R

The t-test and its ANOVA extension from last chapter were suitable when the independent variables were categorical. The categorical nature of those variables made it easy to group the records (trials) into groups of identical factors and compare the means. However, such techniques are completely unsuitable when there exist non-categorical independent variables. In such cases, there is no way to separate the records (trials) into a useful number of cases from which we can meaningfully compare means.

There are a couple solutions. The first is to turn the non-categorical data into categorical data and use the previous results. This, however, is very inefficient, as we are discarding some rather important information. The second method takes advantage of the (effectively) continuous aspect of the data. This second method is called linear regression. The standard way of performing linear regression is using the Ordinary Least Squares (OLS) framework.

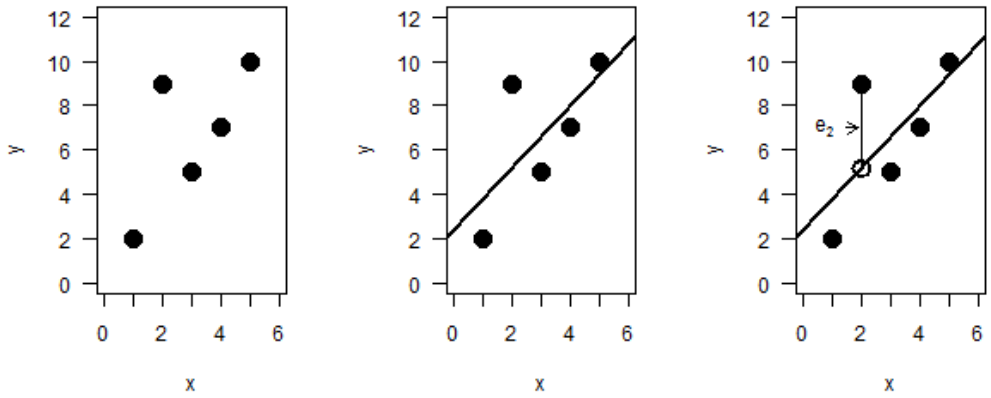

Figure 4.1: *Left Panel: Plot of five points. Center Panel: Line of best fit through the data. Right Panel: Error of estimation for the second point.*

4.1 The method

The method (in the bivariate case) is rather straight-forward to understand. I will use this case to extend it to multiple independent variables.

In linear regression, one is trying to calculate the line of best fit for the data; that is, one tries to calculate a linear function that minimizes the unexplained variance. You probably did something similar at some point in high school.

For our example, let us assume we have five data points: $\{(1,2),(2,9),(3,5),(4,7),(5,10)\}$. One way we can estimate the line of best fit is to plot the point and, by eye, estimate there the line should go. The second way is to set up a series of equations in the square of the error, which is what we wish to minimize. The six equations will be:

$$e_i = \beta_0 + x_i\beta_1 - y_i \tag{4.1}$$

$$t = \sum_{i=1}^{5} e_i^2 \tag{4.2}$$

There are six equations here, since the first line is actually five equations, one for each datum. Now, all we have to do is substitute the five equations in Eqn 4.1 into Eqn 4.2, differentiate the

resulting equation in each of the two parameters (β_0, β_1), set the two equations equal to zero, and solve for the two parameters.

Doing this gives us $\beta_0 = 2.4$ and $\beta_1 = 1.4$. Thus, the equation for the line plotted in the central panel of Figure 4.1 is $y = 1.4x + 2.4$. The expected value of y when $x = 2$ is $\mathbb{E}[y] = 1.4 \times 2 + 2.4 = 5.2$. The actual value of y, when $x = 2$ was 9, therefore, the error is $e_2 = 9 - 5.2 = 3.8$,[1] which is indicated in the right panel of Figure 4.1.

4.1.1 Goodness of fit

We can measure how good this model fits the data by comparing the variance in the data with the variance in the residuals; the residuals measure how much variation remains unaccounted for. This is the famous (or infamous) R^2 measure (Eqn 4.3).

$$R^2 := 1 - \frac{SSR}{SST} \tag{4.3}$$

In equation 4.3, SSR is the sum of the squares of the residuals (or errors); the SST is the sum of the squares of the y values (about its mean). Thus, for this example, we can calculate $SSR = (-1.8)^2 + (3.8)^2 + (-1.6)^2 + (-1.0)^2 + (0.6)^2 = 21.6$; and $SST = (-4.6)^2 + (2.4)^2 + (-1.6)^2 + (0.4)^2 + (3.4)^2 = 41.2$. Thus, for this model, $R^2 = 0.4757$. In the social sciences, this R^2 is acceptable. In the physical and biological sciences, this R^2 is too low for the model to be of interest.

Adjusted R-squared

The strength of the R^2 measure is that it is a 'Proportional Reduction in Error' measure; that is, we can conclude that this model reduces the unexplained error by 47.57%. This is its strength (and its limit).

There is a measure that is frequently (and inappropriately) used to measure the reduction in error. It is called the 'adjusted R-squared.' The adjusted R-squared (Eqn 4.4) adjusts for both the number of independent variables you are using (p) as well as the number of data points in the sample (n).

$$\overline{R}^2 := 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{4.4}$$

[1] The error is also called the residual.

It *should* only be used to determine if the addition or removal of an independent variable is supported.[2] Unfortunately, I have seen far too many use it to quantify the error reduction caused by the model.

※ For the record, the adjusted R-squared helps the researcher determine the appropriate variables to include. It does not directly measure the reduction in error.

Now for a general *caveat*: There is a certain simplicity and attraction to the R^2 measure. however, in my view, too much importance is given to it in social science research. Since we are used to such low values of the R^2, it really tells us little about the model that other statistics do not. I check that the model does not violate the assumptions behind OLS and nothing else. In science, we care more about the effect of our variables than the complete model.

4.1.2 Matrix representation

Extending equations 4.1 and equation 4.2 to handle multiple independent variables introduces a level of complexity that is unnecessary. We can actually write the entire system in matrix form. In matrix form, we need to solve the equation 4.5 for **B**.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{4.5}$$

In the equation, **Y** is the vector of outcome values (values of the dependent variable), **X** is the matrix of independent variables, with the first column all 1's), **E** is the vector of random errors (more on that later), and **B** is the vector of parameter values.

For our example, we have the following:

$$\begin{pmatrix} 2 \\ 9 \\ 5 \\ 7 \\ 10 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}$$

To solve the matrix equation, we make certain assumptions about **E** to eliminate it (more on

[2] To determine if the variable(s) should be included in the model, compare the adjusted R-squared values. If the model has a higher adjusted R-squared with the variable(s) included, then include the variable(s). Otherwise, exclude it (them).

that later) and use some matrix algebra to get Eqn 4.6, where $\mathbf{X^T}$ indicates the transpose of the matrix and $\mathbf{X^{-1}}$ indicates the inverse of the matrix.

$$\mathbf{B} = \left(\mathbf{X^T X}\right)^{-1} \mathbf{X^T Y} \tag{4.6}$$

Solving the matrix equation gives us the coefficients (**B**), the residuals (**E**).

$$\mathbf{E} = \begin{pmatrix} -1.8 \\ 3.8 \\ -1.6 \\ -1.0 \\ 0.6 \end{pmatrix}; \qquad \mathbf{B} = \begin{pmatrix} 2.4 \\ 1.4 \end{pmatrix}$$

The advantage to this form is that it looks just the same no matter how many independent variables we include; the formulas in Eqn 4.1, however, become unwieldy quickly. Not that any of this is important to you and your calculations. That the computer uses the matrix form is between the computer and its operating system.

The *important* things for your understanding are the following items: ※

1. We made one assumption about the **E** vector: The errors are independently, normally distributed with mean zero and variance σ^2. In symbols: $e_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

2. We made two assumptions about the **X** matrix: There is some variation in the independent variables, and the independent variables are not linear combinations of each other.

3. Under these three assumptions, there is an explicit formula for the answer (the coefficients).

Also, under these assumptions, we can calculate an expected value for the dependent variable given values for the independent variables. This is useful when you wish to predict outcomes. However, it is better to interpolate than to extrapolate; it is better to predict within the data range than outside the data range.

4.2 Assumptions

There has been a bit of discussion about just how many assumptions are made when doing linear regression. I have given three assumptions. However, there is a lot of subtlety in those three assumptions.

4.2.1 Normality of errors

We assumed that $e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. This implies several things that may not be met by your data. First, it assumes that the mean error is zero. This is actually not a testable assumption, because the formula self-zeroes the mean using the constant term.

It also implies that the errors are distributed normally. This is testable. One way of testing this assumption is to perform a normality test on the residuals (actual less predicted). If the test says the residuals are normal, then the assumption is met by the data. There are two tests that can be used: the Q-Q Plot and the Shapiro-Wilks Test.[3]

It also implies that the the errors are independent of the independent variables. If not, then this is evidence that your model is misspecified. To test this assumption, calculate the correlations between the residuals and the independent variables.

It also implies that the variance of the residuals is constant (what is called homoskedastic, as opposed to heteroskedastic). To test this assumption, plot the residuals against the dependent variable and visually determine if the variance of the residuals are constant. One can also perform a linear regression of the residuals against the dependent variable.

Violations

Violations of this assumption are common. Solutions to such violations are often worse than the original violation. The nice thing is that OLS is very robust to minor violations of its assumptions. As such, as long as things are approximately alright, there is nothing to worry about.

There are a couple violations that are rather serious, however. The first is correlation in the error terms. The second is heteroskedasticity in the error terms. The first is usually solved using Time Series analysis or Spatial Analysis. The second is often solved using a weighting scheme to reduce the heteroskedasticity.

[3] Actually, there are a slew of normality tests. All give different p-values, but all give about the same interpretation.

As for the normality of the errors, in practice, the errors are rarely normal. However this is of little concern to OLS. This only becomes important when the predictions produce unrealistic values. In such cases, one usually resorts to transforming the dependent variable (see Chapter **??**) or to using Generalized Linear Models (see Chapter **??**).

4.2.2 The independent variables

The first assumption is that there is some variation in the independent variables. This means that one (or) more of the independent variables only takes on a single value in the data. When this happens, we cannot calculate its effect on the dependent variable (its effect cannot be distinguished from the constant term).

When there is little variation in any of the independent variables, the effects of that variable are difficult to estimate. This results in very large standard errors, which implies the variable is not statistically significant. It may be, but because there is little variation, it is indistinguishable from the background noise.

Finally, and relatedly, when an independent variable is a linear combination of others, there is no way to distinguish its effect from the effect of the other variables. And, when correlation is very high, inflated standard errors will result; the variable's effects are indistinguishable from the background noise.

4.3 The R session

For this R session, let us read in data, perform linear regression on it, then check that the assumptions are met. The data we will use will be from the `ssm2.csv` file, which I have in my current working directory (see Section **??**). The `ssm2` data consists of five variables: nominal `stateId`, continuous `yearPassed`, percent `pctWin` (percent vote in favor of the measure), dichotomous `civilBan` (whether the measure also contained a ban on civil unions), and percent `religPct` (the percent of people in the state agreeing that religion is an important part of their lives).

I used this data to quickly predict the outcome of the 2009 ballot measure in Maine to outlaw same sex marriages, without resorting to any published opinion polls. How close did I come? Let us just say I won the bet (although not with what we are doing in this chapter).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4034.1609	715.6774	5.64	0.0000
yearPassed	-2.0095	0.3577	-5.62	0.0000
civilBan	-3.7331	1.9988	-1.87	0.0723
religPct	0.9452	0.1074	8.80	0.0000

Table 4.1: *Results table for the regression of percent support of a generic ballot outlawing same-sex marriage against the three included variables. The R^2 for the model is 0.7801. The probabilities calculated are two-tailed probabilities. The hypotheses were one-tailed hypotheses. As such, all three explanatory variables are statistically significant at the standard level of significance ($\alpha = 0.05$).*

4.3.1 The modeling

Here is the model I am fitting to this data (Eqn 4.8). As we are social scientists and not statisticians, the independent variables actually hold meaning in our models. In fact, the variables are our sole reason for existing as social scientists (alright, people are important, too).

$$pctWin = \beta_0 + \beta_1(yearPassed) + \beta_2(civilBan) + \beta_3(religPct) \tag{4.7}$$

In `R`, performing linear modeling is very straight-forward (as it is in all modern statistical packages). The command is `lm`. As `lm` returns a lot of information, we should store its results in a variable, which I will call `model.1`. Once the computer computes the regression (and all associated information), we can summarize the results in the standard results table (Table 4.1).

Notice that all three variables of interest are statistically significant at the $\alpha = 0.05$ level.[4] Additionally, the model has an R^2 of 0.7801, which is a phenomenal fit in public policy. The direction of the coefficients also agrees with theory: States that are more religious should vote against single-sex marriage at a higher rate; Measures that also ban civil unions should have a harder time passing; Measures passed later should have a more difficult chance of passing, as the young tend to support single-sex marriage, whereas the elderly tend to oppose it.

Thus, the equation for the line of best fit (also known as the prediction line) is approximately

$$pctWin = 4034.16 - 2.01(yearPassed) - 3.73(civilBan) + 0.95(religPct) \tag{4.8}$$

According to this model, what is the expected vote in Maine? To answer this, we need the

[4]This assumes that I specified before hand that theory suggests banning civil unions makes passing the measure more difficult, which I did, but which I did not say above.

information about the Maine ballot measure: yearPassed = 2009, civilBan = 0, religPct = 48. With this information, and under the assumption that the model is correct, we have our prediction that 42.41% of the voters will vote in favor of this ballot measure.

The R code for this part of the analysis is as follows:

```
data <- read.csv("ssm2.csv", header=TRUE)
attach(data)

model.1 <- lm(pctWin ~ yearPassed + civilBan + religPct)
summary(model.1)

predict(model.1, newdata=data.frame(yearPassed=2009, civilBan=0, religPct=48) )
```

Note the inclusion of the `predict()` function, which predicts the dependent variable value given values for each of the independent variables (stick to its syntax closely, or read the help file on `predict()`).

4.3.2 Checking the assumptions

There are basically three assumptions that we need to check—one concerning the residuals and two concerning the independent variables. The residuals need to be independently distributed as $\mathcal{N}(0, \sigma^2)$. The independent variables need to have variation and need to not be linear combinations of each other.

The independent variables

It is quite easy to examine the independent variables to determine if there is variation. The years range from 1998 until 2008, the civilBan dichotomous variable is divided between zero (n=13) and one (n=19), and the religious percent ranges from 51 to 85%. All of these are sufficiently variable to create a model that represents the data.[5]

The level of multi-collinearity among the independent variables can be tested in many ways. The easiest is to note that none of the standard errors in Table 4.1 are extremely large compared to our expectations (the intercept term is so large because the years are measured in thousands and the percents are measured in hundreds). From that, we can conclude that there is no issue with

[5]Note that the lowest level of religiosity in a state was 51, whereas the level in Maine was 48. This means we extrapolated when we predicted the outcome of the Maine election. The 3% is not serious, however. I would have worried if we were trying to predict the outcome of a state whose level of religiosity was 25%.

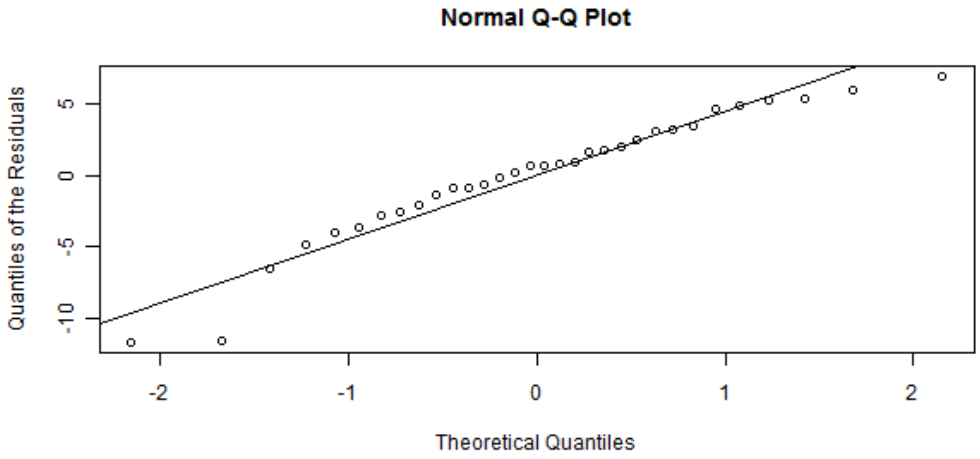

Figure 4.2: *A Q-Q Plot of the residuals of Model 1.*

multi-collinearity.[6]

The residuals

The only thing we need to check is that the error terms (residuals) are independently distributed $\mathcal{N}(0, \sigma^2)$. Actually, this requires several tests. The first test is to check that their means are zero. Before we can check the residuals, let us calculate the residuals. We know that the residuals are defined as the difference between the measured value and the model-predicted value. So, let us define `resids` as:

```
resids <- pctWin - predict(model.1)
```

However, we can do it easier than that; as with all modern statistical programs, there is a function that automatically calculates the residuals: `residuals(model.1)`.

Calculating the mean of the residuals results in a zero value.[7] As such, the residuals pass that test.

The second test is the test of normality. To do that, let us use the Q-Q Plot. Figure 4.2 is a Q-Q Plot of the residuals. If the residuals were perfectly normal, then the residuals would line up

[6] We could have also checked the correlation, `cor(data)` to see if any correlations were close to either 1 or -1.

[7] Actually, the answer given depends on your computer. Most computers will give some number on the order of 10^{-13}, which is the underflow level—the smallest number representable on your computer.

along the diagonal line. Here, the residuals are not perfectly normal. We can quantify how close the residuals are to normality by using either the Kolmogorov-Smirnov test or the Shapiro-Wilk test. The K-S test is extremely sensitive. As such, it is falling into disuse for this purpose. The Shapiro-Wilk test is custom-made for tests of normality. As such, I recommend it. Performing `shapiro.test(resids)` gives us a p-value of just over 0.05. As such, we can conclude that there is not compelling evidence that the residuals are not normally distributed.[8]

The third test is that the residuals are uncorrelated with the independent variables. Using `cor(resids, data)` as a shortcut tells us that there is essentially zero correlation between the residuals and the independent variables.

Finally, the fourth test is a test of homoskedasticity: is the variance of the residuals constant? The typical test of this is the Breusch-Pagan test. The null hypothesis of the test is that the residuals are homoskedastic. To use the Breusch-Pagan test, one needs to load the `lmtest` library. I recommend downloading and installing that library, since it includes several tests that can be used with linear models.

If you do not have the Breusch-Pagan test available in your statistical package (which would be surprising), you can use graphical means: plotting the residuals against the dependent variable and looking for shifts in the variance. This is not as precise as the Breusch-Pagan test, but it does definitively work in some cases; the change in variance is usually not very noticeable. Figure 4.3 is the residual plot for this model. Note that there may or may not be a change in the variance. It is difficult to tell. I would conclude, however, that if there is heteroskedasticity, then it is not too serious.

As the residuals passed all of the tests, we can conclude that this model does not violate the assumptions of Ordinary Least Squares. Here is the R script for the analysis of residuals, including the functions I used to save the plots to my working directory.

```
library(lmtest)

cor(data)
resids <- residuals(model.1)
mean(resids)

png("normalqqplot.png", width=600, height=300)
qqnorm( resids, ylab="Quantiles of the Residuals" )
abline(0,4.5)
dev.off()
```

[8]The null hypothesis in the Shapiro-Wilks test is that of normality.

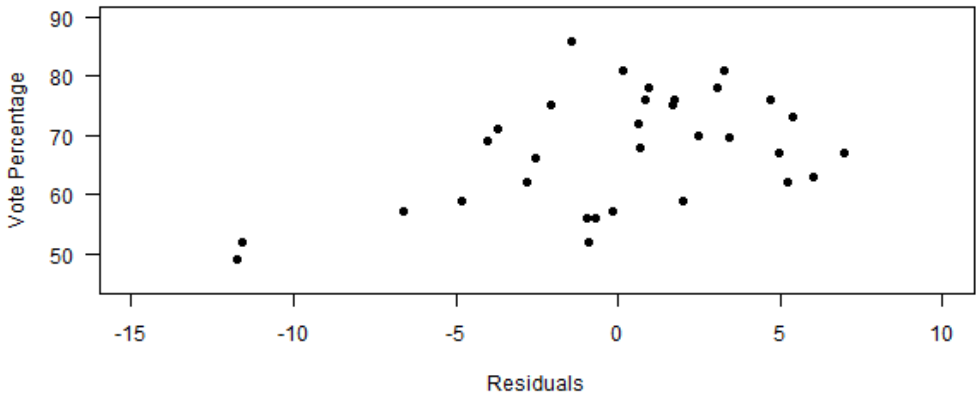

Figure 4.3: *A scatterplot of the vote percentage against the residual. There is no pattern that 'jumps' out at me. The two dots in the lower-left area may simply be outliers—or not.*

```
shapiro.test(resids)
cor(resids,data)
bptest(model.1)

png("residplot.png", width=600, height=300)
plot(resids, pctWin, las=1, pch=16,
    xlab="Residuals", xlim=c(-15, 10),
    ylab="Vote Percentage", ylim=c(45,90)
    )
dev.off()
```

★ ★ ★

Thus, what is our conclusion about the model assumptions? There is no apparent multicollinearity in the independent variables. The residuals are approximately normally distributed. The mean of the residuals is zero. The variance appears to be constant. In short, the model does not appear to violate the assumptions. As such, we can have confidence in the model in that it can be applied to the population, and not just to the sample.

However, there is one thing we overlooked, one thing that all but invalidates this analysis. We will discuss it in the next chapter.

4.4 Conclusion

In this chapter, we started our entry into what is arguably the most important part of statistical analysis in the real world: Regression. The hypotheses we tested in this chapter were rather straight-forward. In the next chapter, we will look at using the same data to answer a slightly different question, and, in doing so, we will be confronted with knowledge that the errors are not normally distributed. In fact, predictions of vote shares less than zero or greater than 100% are possible with this model. Next chapter, we will begin our entry into transformation of the data to force predictions to make sense.

As you leave this chapter, please keep in mind three important things. First, remember that you must know your data before you can analyze it. Second, remember that you must test your assumptions before you can be satisfied with the model (see, for instance, Section 4.2). Finally, remember what the numbers in the regression table actually mean (for instance, Table 4.1). All three of these are extremely important. If you forget any of these, please go back through the chapter.

4.5 Extension

None, yet. Move to the next chapter, as it is an extension of what we did here.

4.6 R commands

predict(model, newdata) As with almost all statistical packages, R has a predict function. It takes two parameters, the model, and a dataframe of the independent values from which you want to predict. If you omit the newdata, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The dataframe must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

residuals(model) R also has a function that calculates the residuals in the data; that is, it calculates the difference between the actual value and the predicted value. This function is extremely useful when doing analyses on the residuals.

qqnorm(model) One of the graphical techniques used to determine normality of residuals is the Q-Q Plot. In R, the `qqplot` function is actually a more general tool that plots the quantiles of any two sets of numbers against each other. The `qqnorm` is dedicated to plotting residuals against the normal distribution to determine normality.

shapiro.test(x) The Shapiro-Wilks test is used to quantify the degree of normality in a group of data. The null hypothesis is that the data is normally distributed. Thus, a p-value greater than $\alpha = 0.05$ signifies that there is not enough evidence to conclude the data is *not* normally distributed.

cor(x,y) This function determines the correlation between two vectors of data (of the same length). You can find the correlations between all of the variables in a data set simply by using `cor(data)`. Apparently, this does not work in Mac OSX if any variable is non-numeric. However, Windows machines will simply report `NA` for such cases.

bptest(model) One of the most important tests in determining homoskedasticity is the Breusch-Pagan test. The null hypothesis of this test is that the residuals are homoskedastic (a good thing). Thus, we look for a p-value greater than $\alpha = 0.05$ to tell us that we cannot reject the null hypothesis of homoskedasticity.