

Recitation 3

Stat 111
Justin Khim
February 5th, 2016

1.1 Introduction

The game plan today is to talk mostly about data in two dimensions. We’ll start out by talking about contingency tables before then going on to the main topic, which is linear regression and the circus that comes with it.

At the beginning of class, you should

1. Check in with me for attendance and turn in your homework.

2. Load the scatterplot data from my website under the “Recitation 03 Data” link.
   By “Load”, I mean that you should download the data, save it to your working directory for R, and then load it into R and call it “grade.data”. The file itself is called “stat101scores.csv”.

3. Talk about the children question, which we didn’t discuss in last class, with a neighbor. This can be found in Section 1.2.

4. Fill in the contingency table in Section 1.3. You should discuss this with a neighbor.

We’ll come back to the last three points and talk about them as a class, although probably not in the order presented.

1.2 Children in a Family Again

We failed to talk about this as a class last time, but it is something you should discuss amongst yourselves. Recall that on the survey, I asked you to report the number of children in your family, yourself included. So, let’s compute a couple of summary statistics.

```
> mean(survey.data$children)
[1] 2.594595
> median(survey.data$children)
[1] 3
> sd(survey.data$children)
[1] 1.157689
```

You could additionally make a plot if you’d like. Here’s a puzzle. the U.S. average for children in a household was 1.86 in 2000 (and is probably similar now, although I didn’t bother looking through the 2010 census data). Why is our number for children so much higher? How could we possibly get our numbers to be more in line with the U.S. average?
1.3 Vietnam War Support

The following comes from [?]. In January 1971, a Gallup poll asked

“A proposal has been made in Congress to require the U.S. government
to bring home all U.S. troops before the end of the year. Would you like to
have your congressman vote for or against this proposal?”

At the time, the question was clearly in reference to the Vietnam War. Your goal is to
guess the percentage of people in each educational category that supported and opposed
withdrawal by filling in Table 1.1.

<table>
<thead>
<tr>
<th>Grade School Education</th>
<th>High School Education</th>
<th>College Education</th>
<th>Total Adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>% for withdrawal of U.S. troops</td>
<td></td>
<td></td>
<td>73 %</td>
</tr>
<tr>
<td>% against withdrawal of U.S. troops</td>
<td></td>
<td></td>
<td>27 %</td>
</tr>
<tr>
<td>Total</td>
<td>100 %</td>
<td>100 %</td>
<td>100 %</td>
</tr>
</tbody>
</table>

Table 1.1 A breakdown of support among adults for U.S. withdrawal from Vietnam by education.

The goal here is twofold. The statistical aspect is to make sure that the numbers
you guess to fill in the blanks are numerically possible. The second goal is to try to
estimate how support varied among different levels of education.

Here’s a hint: assume that there are 100 people of each education type, and redraw
the table with absolute number instead of percentages, e.g. 73 % would become 219, 27
% would become 81, the 100 % in the bottom right would become 300, and each of the
other 100 % would become 100. Then try to fill in the blanks.

1.4 Midterm and Final Grades

As an example of plotting two-dimensional data, let’s use the exam grades from a stat
101 class. We start by loading the data.

```r
g = data = read.csv("stat101scores.csv")
t = which(gdata$Midterm != 0 & gdata$Final != 0)
t = data[took.both, ]
```

Now, we are interested in comparing the midterm and final exam scores. We start by
defining new vectors for these, although we don’t have to do this.

```r
midterm = trimmed.data$Midterm
final = trimmed.data$Final
```

Here’s a version of the plotting with a lot of optional parameters to make the graph
look pretty.

```r
plot(midterm,
    final,
    <-
```
main = "Stat 101 Midterm and Final Grades",
xlab = "Midterm Grades",
ylab = "Final Grades",
xlim = c(40, 100),
ylim = c(40, 100),
cex = 2.0,
cex.axis = 1.5,
cex.lab = 1.5,
cex.main = 2)

The scatter plot can be seen in Figure 1.1

![Scatter plot](image)

**Figure 1.1** The scatter plot looks like this.

**Exercise 1.4.1.** Make a scatter plot of height and weight from the recitation 2 data.

Note that the plot looks a little odd because many of the data points overlap. This is due to the fact that there were many students in the 101 class, and the exams were multiple choice. At this point, let’s plot the line of best fit through the data. We can add the line of best fit with

```r
exam.model = lm(final ~ midterm)
abline(exam.model)
```

I also added the 45 degrees line just because I wanted to for our discussion. At any rate, the other line should now be on the plot, which can be seen in Figure 1.2.

So what do we have to discuss? We want to know why this technique of fitting a line to data like this became known as “linear regression”. The “linear” part is hopefully obvious. The “regression” part can be explained by looking at the exam scores. Consider the points that fall on the 45 degrees line. These are the people who do exactly the same on the midterm and the final. Now, consider the line of best fit. For low midterm scores, the line of best fit is above the 45 degree line, and for high midterm scores, the line of best fit is below the 45 degree line. This is pretty common. What happens is that for low midterm scores, the student is generally both bad and
Figure 1.2 The scatter plot looks like this, but this time we added a line of best fit. I also added a line of 45 degrees to the plot.

unlucky (underprepared, etc.). For high midterm scores, the student is both good and lucky (prepared, whatever).

As a result, a student with a bad midterm score expects to do better on the final, and a student with a good midterm score expects to do worse on the final. This is the “regression” part. Now, the student with the bad midterm score does better than they did before, but not necessarily better than the good student.

By the way, since I didn’t have the final grades for the exams I showed, I made some up using an advanced statistical procedure. I also made a prettier chart and put it in Figure 1.3. The point here is that higher level statistics can do some interesting things. Also, I wanted to remind you that most intro level classes are curved. The curve represented in this figure is probably too generous as far as giving too many As and Bs;
so remember to do well on your exams.

The canonical example of this is your height as the dependent variable and a parent’s height as the independent variable. Tall parents have both good genetics and luck; so they expect their kids to be tall. However, the kids aren’t likely to be as lucky; so they tend to be shorter. The same works in reverse for short parents. They have both bad (like it’s a crime to be short) genetics and bad luck (but seriously, short people make less money, aren’t viewed as natural leaders, etc.). But luckily, their children only necessarily have the bad genetics part, and they probably won’t be as unlucky.

But, let’s return to the grades example. We can get a summary of our model

```r
> summary(exam.model)

Call:  
  lm(formula = final ~ midterm)

Residuals:  
            Min       1Q   Median       3Q      Max  
-28.143 -5.885  1.373  6.453  29.761

Coefficients:  
                  Estimate Std. Error t value Pr(>|t|)  
(Intercept) 26.68970   5.28349   5.052  9.74e-07 ***  
midterm     0.64518   0.06305  10.233  < 2e-16 ***  

Residual standard error: 10.12 on 203 degrees of freedom
Multiple R-squared:  0.3403,  Adjusted R-squared:  0.337
F-statistic: 104.7 on 1 and 203 DF,  p-value: < 2.2e-16
```

The important part here, for the time being, is the coefficient estimates. Here, the intercept is 26.6897, and the slope is 0.64518. So the line of best fit is

\[
\text{final} = 26.6897 + 0.64518 \cdot \text{midterm}.
\]

What is the interpretation of the intercept here? It means that if someone were to get a 0 on the midterm and somehow not drop out of the class, then they would be expected to get a 26.6. This might not be good as a prediction though because we didn’t observe anyone get anything less than a 40.

**Exercise 1.4.2.** If someone got an 80 on the midterm, what would you predict their final exam score to be?

### 1.5 Draw Regression Lines

In this section, you get to draw your own regression lines in a couple of simple cases. Hopefully you will understand the least squares approach better after this exercise. Draw lines of best fit for Figures 1.4, 1.5, and 1.6.
Figure 1.4 A line of best fit for three points (draw a line of best fit).

Figure 1.5 A line of best fit for four points (draw a line of best fit).

1.6 Correlation

The final thing we can talk about is correlation. This is a standardized slope. So, we can do this for the example we’ve already been using.

```r
> cor(midterm, final)
1 0.5833385
```

Thus, we can see that the correlation between midterm and final scores is pretty strong. This is to be expected.

I also computed the correlation for weight and GPA for my two sections of 111. The correlation was \(-0.1396226\). What does this mean, if anything?

The line of best fit was

\[
GPA = 3.814646 - 0.001950 \cdot \text{weight}.
\]

What does the intercept mean in this example?
Exercise 1.6.1. Find the correlation between weight and height in the survey data from recitation 2.