Bookmark this page

Statistics Package Exercise: Finding a Regression Line

Learning Objective: In the special case of linear relationship, use the least squares regression line as a summary of the overall pattern, and use it to make predictions.

In this activity we will:

- find a regression line and plot it on the scatterplot.
- examine the effect of outliers on the regression line.
- use the regression line to make predictions and evaluate how reliable these predictions are.

Background

The modern Olympic Games have changed dramatically since their inception in 1896. For example, many commentators have remarked on the change in the quality of athletic performances from year to year. Regression will allow us to investigate the change in winning times for one event—the 1,500 meter race.

- R | StatCrunch | TI Calculator | Minitab | Excel

R Instructions
To open R with the dataset preloaded, right-click here and choose "Save Target As" to download the file to your computer. Then find the downloaded file and double-click it to open it in R.

The data have been loaded into the data frame

```
olym
```

. Enter the command

```
olym
```

to see the data. The data frame

```
olym
```

has the following variables:

```
Year
```

,

```
Time
```

.

Here is a description of the variables:

```
Year
```

: the year of the Olympic Games, from 1896 to 2012.

```
Time
```

: the winning time for the 1,500 meter race, in seconds.

First, let's explore the relationship between the two quantitative variables:

```
year
```

and

```
time
```

. Produce a scatterplot and use it to verify that year and time are nearly linear in their relationship.

To do this in R, copy the command:

```
plot(olym$Year, olym$Time, xlab="Year of Olympic
Games",ylab="Winning Time of 1500m Race (secs)")
```

Observe that the form of the relationship between the 1,500 meter race's winning time and the year is linear. The least squares regression line is therefore an appropriate way to summarize the relationship and examine the change in winning times over the course of the last century. We will now find the least squares regression line and plot it on a scatterplot.

- R
- StatCrunch
- TI Calculator
- Minitab
- Excel

R Instructions
In order to fit the regression line, we use the command

```
lm()
```

. The

```
lm()
```

command produces a large amount of information, which we will want to extract as we need it, so we save the information to another variable name model.

```
model = lm(olym$Time~olym$Year)
```

To add the regression line to the scatterplot, we can extract the linear equation from model and add the line to the scatterplot. To do this in R, copy the entire command below:

```
plot(olym$Year, olym$Time, xlab="Year of Olympic Games",
ylab="Winning Time of 1500m Race (secs)");abline(model)
```

We can also extract the y-intercept and slope from the model to determine the regression equation. To do this in R, copy the command:

```
coef(model)
```

Learn By Doing (1/1 point)

Give the equation for the least squares regression line, and interpret it in context.

Your Answer:

Time = -0.3527988 * Year + 916.4323092

916 is the base time; then each increment of the year will bring that base time down by 0.35 seconds.

Our Answer:

The equation for the least squares regression line is: Time = 916 - 0.35 * Year. The slope of the line indicates that the winning time for the 1500 meter race decreases by about 0.35 seconds every year, or by about 4 * 0.35 = 1.40 seconds, on average, from one Olympiad to the next. (The fact that the times are decreasing rather than increasing is also indicated by the fact that the value of b is negative.)

Resubmit Reset

- R Instructions
- StatCrunch
- TI Calculator
- Minitab
- Excel

R Instructions

Notice that there is an outlier. Remove the outlier by copying these commands (for this exercise we will not modify x and y-axis labels).To do this copy each line separately and in order, hit enter to see the change:

```
plot(olym$Year[olym$Year!=1896], olym$Time[olym$Year!=1896])
```

```
L =
lm(olym$Time[olym$Year!=1896]~olym$Year[olym$Year!=1896]);
```

```
abline(L);
```

```
cf=coefficients(L);
```

```
legend(1950,240,legend=paste("time =
",round(cf[1],0),round(cf[2],2),"year"))
```

You will now see that the least squares regression line and the values in the equation have changed.

Learn By Doing (1/1 point)

Give the equation for this new line and compare it with the line you found for the whole dataset, commenting on the effect of the outlier.

Your Answer:

Of course, the new line has different coefficients. Instead of a 0.35 difference per year, it's now a 0.3 difference; the base time is also 811 instead of 916.

Our Answer:
Once the outlier for the year 1896 is removed, the equation for the least squares regression line is: Time = 812 + (- 0.30 * Year) Note: Some statistics packages may show the regression line as: Time = 811 + (-0.30 * Year) When the outlier is removed, the line "drops" a bit—the intercept is smaller and the slope is not as negative. Both of these results are quite reasonable, since the original data were pulled upward toward the outlier. Once this outlier is removed, the line drops.

Resubmit Reset

Learn By Doing (1/1 point)

Our least squares regression line associates years as an explanatory variable, with times in the 1,500 meter race as the response variable. Use the least squares regression line you found in question 2 to predict the 1,500 meter time in the 2016 Olympic Games in Rio de Janeiro. Comment on your prediction.

Your Answer:

207 seconds. Seems legit?

Edit: oh, it's extrapolation. I didn't notice. That makes the prediction unreliable.

Our Answer:
812 + (- 0.30 * 2016) = 207.20 seconds. This is an extrapolation. We cannot be sure that the linear dependence of winning times upon years holds past the range of the explanatory variable, which is the year 2012. At some point, the linear dependence must no longer apply, because it would predict impossible winning times.

Resubmit Reset
