Lsn 21 - MA206Y

Clark

Admin

Recall that last lesson we were interested in testing:

One statistic we could use to test this set of hypothesis is the sample correlation coefficient, r. While r tells us how strongly two quantiatiave variables are related, it doesn't give us a feel for what exactly we should expect to happen to y when we change our X values. Remember from 7th grade, the equation of a line is $y = mx + b$. We know that b is the y intercept and m is the slope, but what exactly is a slope? Let's consider two situations, one where $x = 3$ and one where $x = 4$. We can write out the fitted lines as:

So the difference in y is m. That is, the value of m gives us the expected change in y for one unit change in x. Let's consider the study in our book. Are dinner plates getting larger? Why might we care about this question?

Here we can read in the data as:

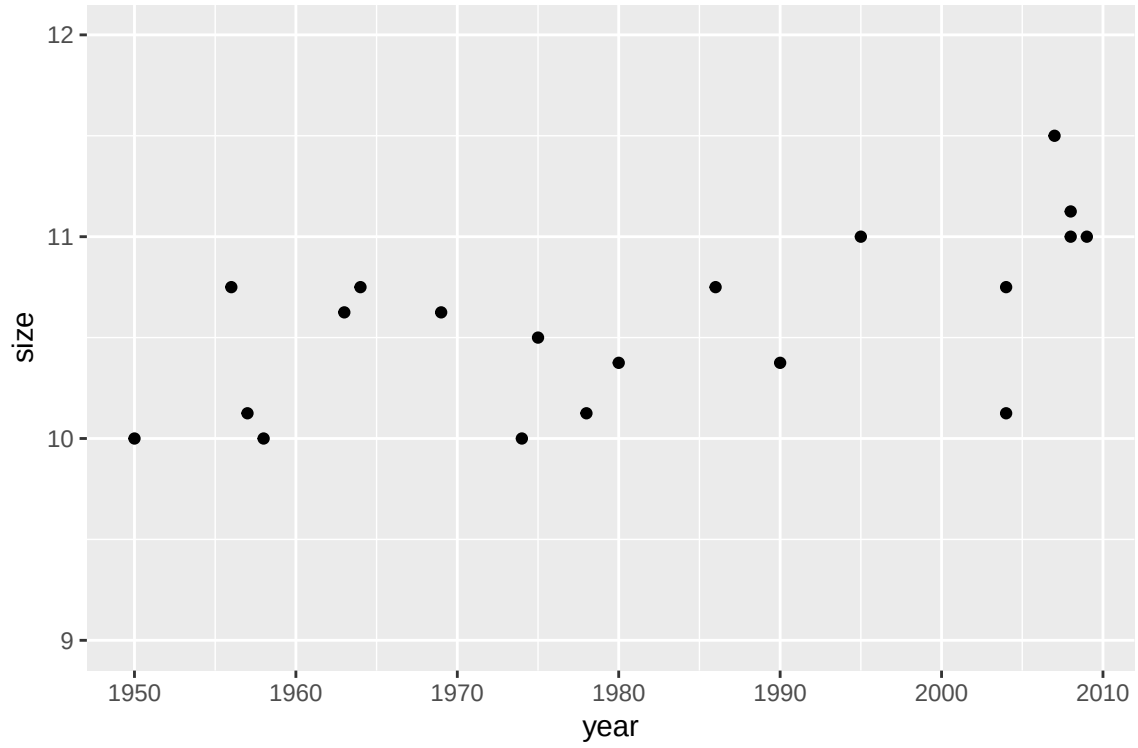
```
plate.dat=read.table("http://www.isi-stats.com/isi/data/chap10/PlateSize.txt",header=T)
```

We can explore the data:

```
plate.dat %>% summarize(mean=mean(size),sd=sd(size))
```

```
##     mean        sd
## 1 10.575 0.4299633
```

```
plate.dat %>% ggplot(aes(x=year,y=size))+geom_point()+
  ylim(c(9,12)) #Note the ylim command here
```



It's kind of hard to tell if there's a relationship, but we can look at r

```
cor(plate.dat$year,plate.dat$size)
```

```
## [1] 0.6037724
```

Which would suggest a pretty strong relationship. To fit a line in R we use the `lm()` function which computes the m and b values such that:

```
my.lm=lm(size~year,data=plate.dat)
summary(my.lm)
```

```
##
## Call:
## lm(formula = size ~ year, data = plate.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73490 -0.20092  0.04375  0.22425  0.60169
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.800332   7.897098  -1.874  0.07724 .
## year          0.012805   0.003985   3.213  0.00482 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3521 on 18 degrees of freedom
## Multiple R-squared:  0.3645, Adjusted R-squared:  0.3292
## F-statistic: 10.33 on 1 and 18 DF,  p-value: 0.004818
```

Whew... What is going on here... The first thing to note is how we input equations into R. R always assumes you have a Y intercept so the form `size~year` is fitting the equation:

To get the estimates we look at the values of `Estimate` in the table. Note what we are doing here, we are not assuming that we KNOW the relationship, but we are using data to ESTIMATE the relationship. Our book doesn't use $y = mx + b$ but rather uses $\hat{y} = a + b(x)$ as the fitted equation.

In terms of our problem a slope of 0.0128 means:

So as year increases we can say that size also increases.

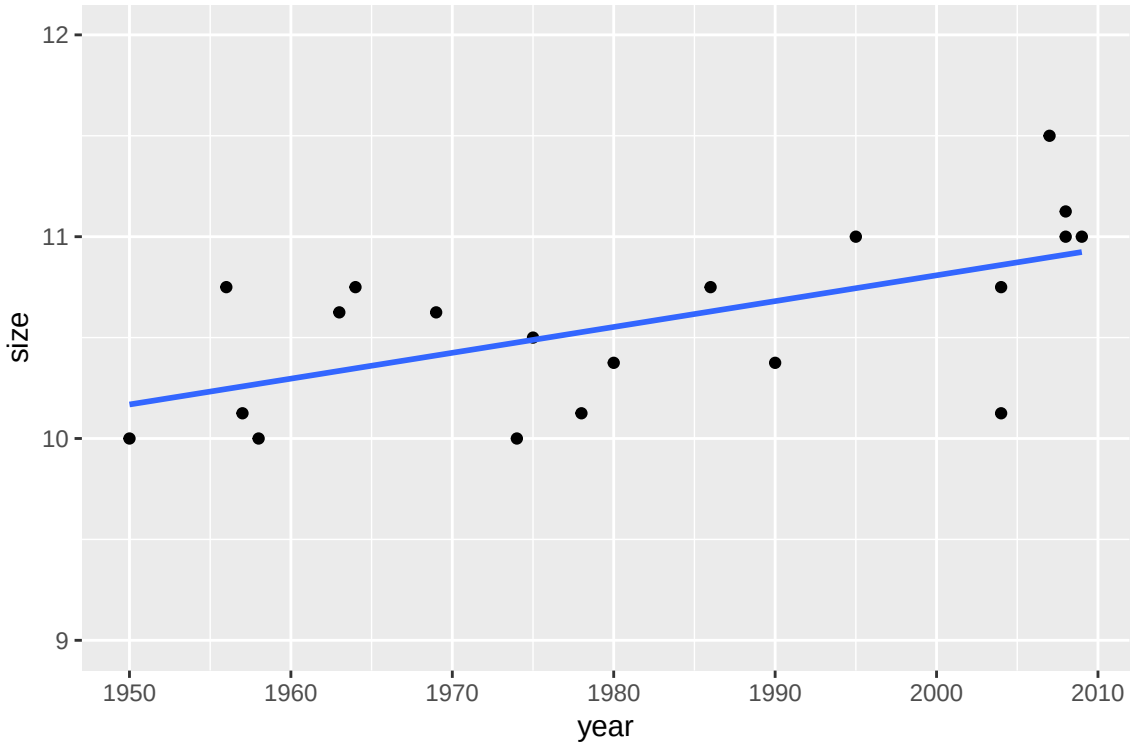
What does it mean for an intercept to be equal to -14.8?

To add the regression line to our data we do:

```
plate.dat %>% summarize(mean=mean(size),sd=sd(size))
```

```
##     mean        sd
## 1 10.575 0.4299633
```

```
plate.dat %>% ggplot(aes(x=year,y=size))+geom_point()+
  ylim(c(9,12)) +stat_smooth(method="lm",se=FALSE)
```



But how good is this fit? Certainly it looks ok for some of the points (1975), but is it good for all of the points? Where does it do the worst?

One way to quantify how good it is is to look at the residuals. Our residuals are found from observed-predicted. Here our residuals can be found from

```
residuals=my.lm$residuals
max(abs(residuals))
```

```
## [1] 0.7349003
```

The maximum absolute residual then gives the point where my data fits the line the worst.

In order to visualize we can look at the applet.

Not what happens to the residuals as we move the regression line.

The final statistic we can use is the R^2 statistic. This statistic gives the proportion of variability in our data that is explained by our model.

That is, if we have an $R^2 = 1$ we have a perfectly linear fit, an $R^2 = 0$ means a no fit at all.

If we go back and look at the `lm` output we see '`Multiple R-squared:`. This is caluclated from teh formula given on page 541. It's not a mistake that we have r as correlation and R^2 as the name of another statistic. In fact here, if we square our correlation coefficient we get `.603774^2=0.364543`

Here we can say that the coefficient of determination is 36.45 %. This means that 36.45% of the variation in plates is due to changes in year.

As we will discover, probably the best statistic to use is the coefficient for slope. Let's look at this a bit more. Recall that our equation we are fitting is:

If our slope is 0 what are we saying about the relationship between X and Y?

Therefore, we can phrase our null and alternative hypothesis as:

Here to test this, we will use, as our test statistic, b

Our realization of our test statistic (or calculated value) is found from minimizing the sums of squared error between:

Let's consider the our IOCT and APFT data again. If we want to use our slope estimate as our test statistic we would calculate:

```
apft.dat=read.csv("APFT.csv")
my.lm=lm(IOCT_Time~APFT_Score,data=apft.dat)
summary(my.lm)
```

```
##
## Call:
## lm(formula = IOCT_Time ~ APFT_Score, data = apft.dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.91 -33.32 -18.89  19.65 425.47
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 287.14155   22.08224  13.003  < 2e-16 ***
## APFT_Score   -0.30404    0.08027  -3.788 0.000177 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.86 on 382 degrees of freedom
## Multiple R-squared:  0.0362, Adjusted R-squared:  0.03367
## F-statistic: 14.35 on 1 and 382 DF,  p-value: 0.0001766
```

Here we see that our estimate of b is -.304. But does this prove that our alternative hypothesis is true?

Using the applet, let's find the p value for the slope.

What is our conclusion?

Now let's try one on your own. Let's look at Exploration 10.4

Write the null and alternative hypothesis in words

Using the data:

```
hand.data=read.table("http://www.isi-stats.com/isi/data/chap10/Handwidth.txt",header=T)
```

Find the correlation between hand with and perceived weight.

Assuming you want to use slope as your test statistic, write out your null and alternative hypothesis.

Using R find the leas square estimates for both slope and y intercept.

Paste the data from the website into the applet and simulate the distrubtion under H_0 and find your p value.

What is your conclusion?