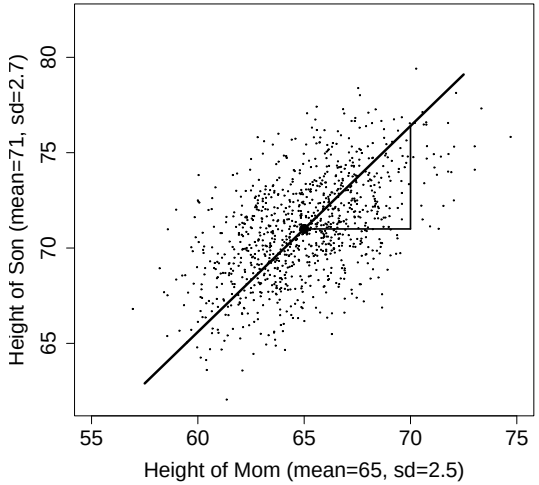
5. REGRESSION (FPP, Ch 10,11,12)
5.1 SCATTER PLOTS AGAIN

• **Remember the Moms and their adult Sons:**



• **Each variable has approximately a normal distribution.**

• **The scatterplot is football-shaped.**

• **The SD line is the axis of the "football".**

• **The points are clustered around the SD line, but not tightly.**

• **In fact, the correlation coefficient in this example is r=0.5.**

Blank page for your notes

5.2 REGRESSION TO THE MEAN

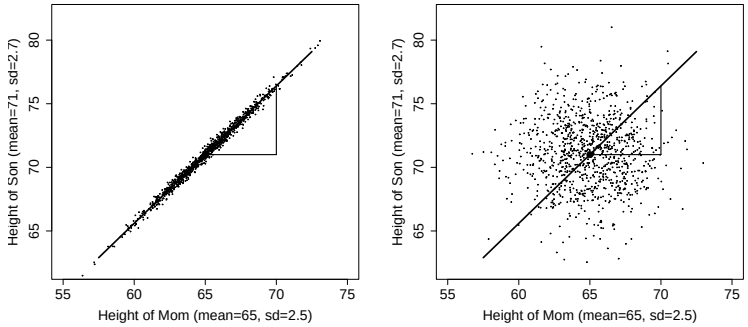
• If mom is right at mean (moms') height, we predict son is at at mean (sons') height.

• Now suppose mom is at 2 SD above mean height:

If association is really strong (r≈1) all points are on the SD line, and we predict son at 2 SD above average.

If there is no association (r=0), we would predict son at the sons' average.

r ≈ 1 **r = 0**



• In fact, on the previous page, r=0.5:
if mom is at 2 SD above average, we predict son at:
(mean + r × 2 SD).

• If mom is 1 SD below average, we predict son at
(r × 1 SD) below average.

• This is regression to the mean.

• It comes from the spread of points: it does NOT mean we are all "getting more average".

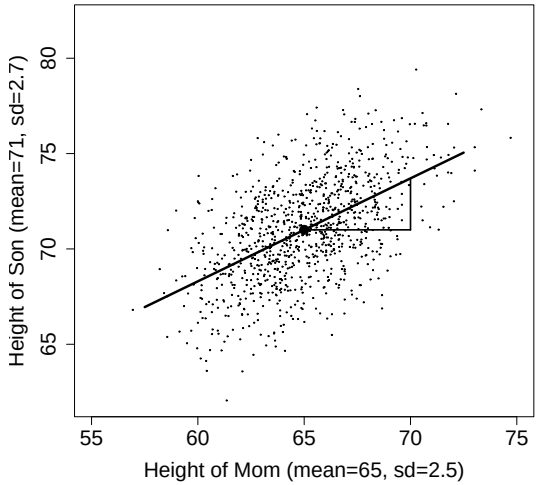
Blank page for your notes

5.3 THE REGRESSION LINE

• The regression line goes through the point of averages.

• The regression of y on x has slope
r × (SD of y)/(SD of x).

• If r is less than 1, the regression line has smaller slope than the SD line.



• The regression of y on x estimates the average y-value corresponding to a given x-value.

Blank page for your notes

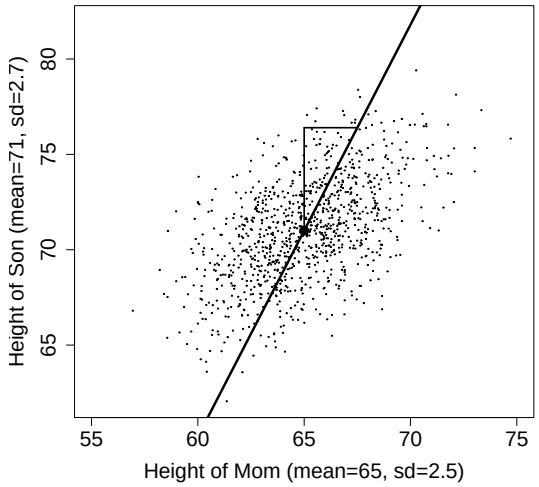
5.4 TWO REGRESSION LINES

• **Which variable are we regressing on which?**
So far we have predicted son's height from his mom's.

• **Suppose we want to predict mom's height from the son's height. Because of regression to the mean we cannot use the same line.**

Now if son is 2 SD above the mean (for men), we predict mom at r $\times$ 2 SD above the women's mean.

• **We could start again, switching the axes, but we do not need to.**



• **This is the line for the regression of x on y:**
it predicts the average value of x corresponding to a given value of y.

• **Again, the regression effect comes from the spread of points.**

Blank page for your notes

5.5 PREDICTIONS FROM A REGRESSION

• The points on the regression line of y on x give the average y-value among subjects of a given x-value.

• So the point on the regression line is the best prediction for the y-value.

• Women have mean height 65", SD = 2.5". Their sons, mean 71" and SD = 2.7". The correlation r=0.5.

• Jane is 70" tall. She is 2 SD above the mean.
We predict her son at (r × 2) SD above the mean
That is, 1 SD above the mean
That is, 71+2.7 = 73.7 inches tall.

• Sarah is 62.5 inches. She is 1 SD below the mean.
We predict her son at (r × 1) SD below the mean
That is, 0.5 SD below the mean
That is, (71- 0.5×2.7) = 69.65 inches tall.

• John is 73.7 inches tall: 1 SD above the mean.
We predict his mom at (r × 1) SD above the mean
That is (65+ 0.5 × 2.5) = 66.26 inches.

• Michelle is at the 90 th percentile for height.
The percent closer to the mean than Michelle is 80%
The z-score for Michelle's height is 1.3 (from A-105)
The predicted z-score for her son's height is (r × 1.3) = 0.65.
So the "middle area" is 48% (from FPP A-105 table)
So he is at (26+48) = 74th percentile for height.

Blank page for your notes

5.6 R.M.S. DEVIATION FROM A REGRESSION

• **Actual values do not fall exactly on the regression line: they vary about this average.**

• **How much do they vary? Typically, how large are the (vertical) distances from the prediction.**

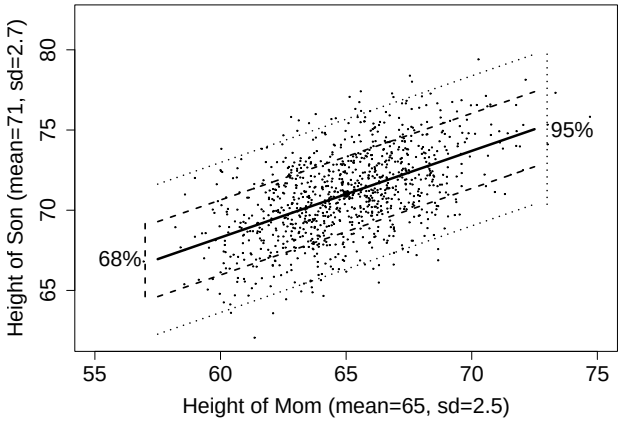
• **The mean difference is 0: some are above, some below.**

• **The SD of these differences is $\sqrt{1-r^2}\times$ (SD of y)**
FPP calls this SD the r.m.s error.

• **If r=1; SD line is regression line, and all points are on the line: r.m.s.error = 0.0.**

• **If r=0; regression line is flat: r.m.s.error = (SD of y).**

• **These prediction errors have a normal distribution.**



• **That is 68% of points are within 1 r.m.s.error of the regression line.**

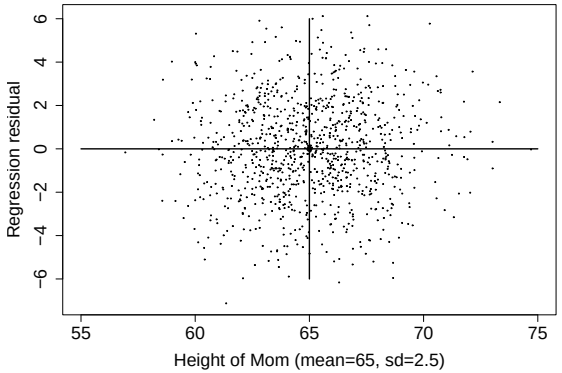
• **... and 95% are within 2 r.m.s.error of the line.**

Blank page for your notes

5.7 RESIDUALS FROM A REGRESSION

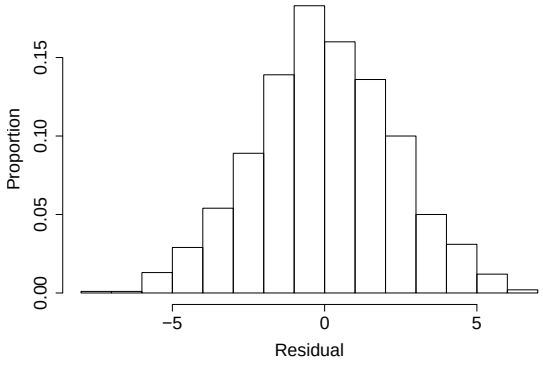
• **The prediction errors are called residuals.**

• **The residual plot plots the prediction error against the corresponding x-value.**



• **Looks good – no pattern of mean or spread with Mom's height.**

• **Now we can try a histogram of all the residuals.**



• **A good bell-shaped curve!! What is the SD?**

• **It should be $\sqrt{(1-r^2)}\times$ (SD of y) or**
$\sqrt{0.74} \times 2.7 \ = \ 2.34$

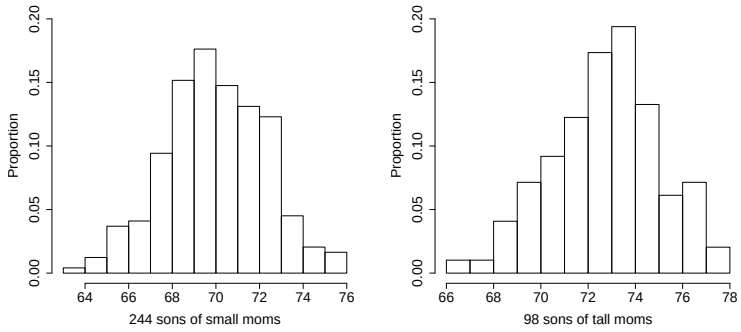
• **2 SD (4.7) should cover 95% – looks good!!**

Rest of this page for your notes

5.8 NORMAL CURVES IN VERTICAL STRIPS

• **Let's look just at the sons of moms 62 to 64 inches tall, and sons of moms 68 to 70 inches tall**

• **Group 1 (n=244): mean = 70.0 inches, SD = 2.29 in.**
Group 2 (n=98): mean = 72.7 inches, SD = 2.29 in.



• **Within a narrow vertical strip, the regression line gives the average prediction (as before).**

• **But now we can also use the normal distribution of residuals, with SD = r.m.s.error, to talk about percentiles etc.**

• **Example: for moms 63 inches tall, what percentage of their sons will be below the (son)-population 20 th percentile? ONE STEP AT A TIME:**
mom's z-score is (63-65)/2.5 =-0.8
mean of these sons = mean - r $\times$ 0.8 $\times$ (SD-sons) =
(71 - 0.5 $\times$ 0.8 $\times$ 2.7) = 69.9 inches
SD of these sons = r.m.s.error = 2.34 inches (above).

Population 20 th percentile: need between-area 60%:
z = 0.85 (from FPP A-105)
Overall 20 th percentile = 71 - 0.85$\times$2.7 = 68.7 inches
For selected sons, this is z-score (68.7-69.9)/2.34 = 0.51
Corresponding between-area is 39% (from FPP A-105)
so the required percentage is 0.5$\times$(100-39) or 30.5%

Rest of this page for your notes

5.9 PREDICTIONS VS INTERVENTIONS

• Weight is associated with height. But if an individual diets to lose weight, he/she does not shrink in height!

• Years of education is associated with income, but giving specific individuals extra years of education will not necessarily increase their income.

• The scatterplot, and regression, describes the population as it is. There are many factors that go in to determining both income and education.

• Intervening – changing one variable for some individuals, makes them atypical members of the population. The regression line does not apply to them.

• Weight is positively associated with metabolic rate. People weighing 110 pounds have a metabolic rate 200 points lower (on average) than those who weigh 150 pounds. But if you lose 40 pounds weight, your body tries to adjust by INCREASING metabolic rate – the opposite of what the population association predicts.

5.10 CAUTIONS

• When the scatterplot is not football shaped – be careful!!

• Remember the shape of the football depends on the SDs of the two variables. It is useful to look at both the SD line and the regression line.

• When the relationship is non-linear, do not use a regression line, either to describe the association, or to predict !

• Even a non-linear relationship can give a substantial value of the correlation, r.

• Consider the pattern of residuals.

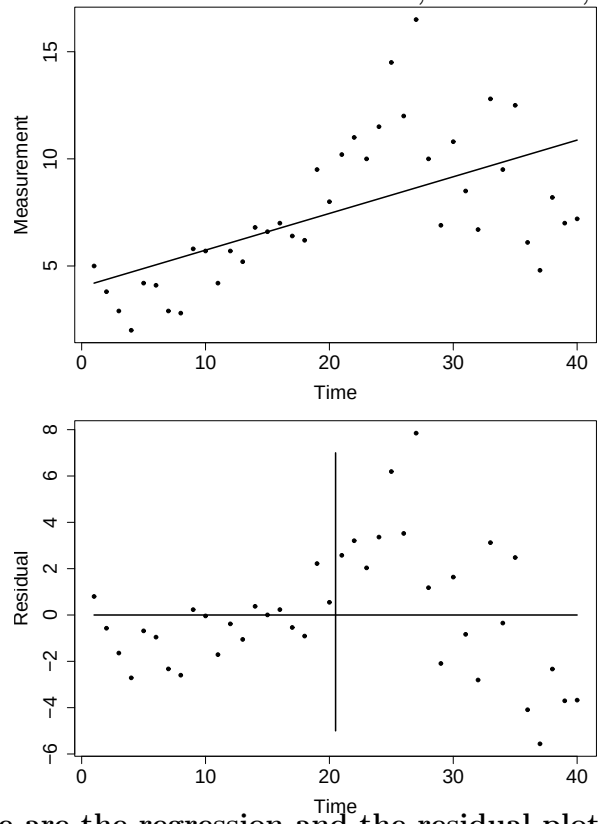
• Heteroscedasticity; If the SD of residuals vary with x, we may be able to predict the average y for a given x (cautiously), but we cannot use the r.m.s.error to predict the spread in a vertical strip,

• Outliers again? : these can have a big effect on a regression line.

• Do not extrapolate:
Beyond the range of values at hand.
Beyond the populations at hand.

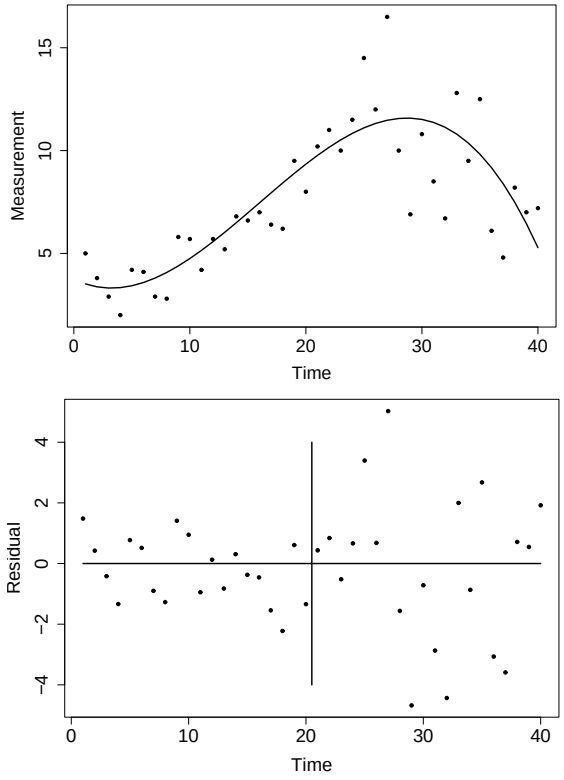
5.11 THE HOMEOPATHY STUDY AGAIN REGRESSION OF LAB1 RESULTS ON TIME

- **The relationship is not linear!**
  The data analyzers fitted a curve not a line.
  But, as example, we will fit a line.
- **For "time": mean = 20.5, SD = 11.69.**
  For measurements: mean = 7.54, SD = 3.40, r = 0.59.



- **Here are the regression and the residual plot.**

- **The reason for fitting a time trend here is not prediction!**
- **Recall there could be a small treatment effect, that might be masked by the time pattern. By taking out the time pattern, we may be able to see smaller effects we could not see in the unadjusted data.**
- **The Study Analysts fitted a curve (a "cubic spline")**



- **The fit is better, but we still have heteroscedasticity**