Outline

Regression

Plotting the Residuals
The Regression line
Predictions for data in a vertical strip
Slope and Intercept
Regression

Plotting the Residuals

Prediction errors are usually called residuals. It is important to explore the graphical properties of residuals to find out about the goodness of the fit by the regression line.

In a residual plot the $x$ coordinates are the same as for the original data. The $y$ coordinates correspond to the values of the residuals. So there is one point for each point in the original scatter diagram.

The average of the residuals is 0 and the regression plot for the residuals is horizontal.
Thus, if everything is OK with the regression line, we expect to see a cloud of points around the zero line in the $y$ axis.
What do you expect in a residual plot?

- We expect to see no trends or clusters in the residuals
- There should be about the same number of positive as negative residuals
- A histogram of the residuals should look symmetric around zero
The following results are taken from a study of about 1,000 families:

- average height of husband $\approx 68$ inches, $SD \approx 2.7$ inches
- average height of wife $\approx 63$ inches, $SD \approx 2.5$, $r \approx 0.25$

Predict the height of a wife when the height of her husband is:
1. 72 inches:

The husband is 4 inches above average height. This is $4/2.7 = 1.5$ SD above the average. So the wife is predicted to have

$$r \times 1.5 = 0.25 \times 1.5 \approx 0.4$$

this corresponds to an increment of $0.4 \times 2.5 = 1$ inch above the average height. This is $63 + 1 = 64$ inches.

2. 68 inches:

In this case the husband is right on the average, so the wife will be right on the average as well.
The Regression line

In section we will try to get a deeper understanding of the regression method. We will explore the geometry of the problem by defining a regression line. We will cover the following topics:

- Predictions on a vertical strip
- Slope and intercept of the regression line
- Least squares
Predictions for data in a vertical strip

Example: A law school finds the following relationship between LSAT scores and first-year scores

average LSAT score = 162, SD = 6
average first-year score = 68, SD = 10, $r=0.60$

Q: About what percentage of the students had first-year scores over 75?

A: We use the normal curve approximation. Converting to standard units

$$\frac{75 - 68}{10} = 0.7$$

this corresponds to a right hand tail of 24% under the normal curve.
Q: Of the students who scored 165 on the LSAT, about what percentage had first-year scores over 75?

A: We first convert to standard units for the $x$ variable:

$$\frac{165 - 162}{6} = .5$$

then convert to standard units for the $y$ variable

$$r \times 0.5 = 0.6 \times 0.5 = 0.3 \text{ SD's}$$

which corresponds to $0.3 \times 10 = 3$ points above average, or $68+3 = 71$. 
Since the data corresponding to a strip are a smaller and more homogeneous sample, the corresponding SD will be smaller.

**How much smaller?**

We expect the dispersion in the $y$ variable to be about the same for each vertical strip. This is given by the RMS error, thus the new SD is

$$\sqrt{1 - r^2} \times \text{SD of } y = \sqrt{1 - 0.6^2} \times 10 = 8 \text{ points}$$

This new SD can be used to convert to standard units

$$\frac{75 - 71}{8} = 0.5$$

and, using the normal curve, we obtain an area of 31% above 0.5. This is the percentage of students scoring more than 75 in the first year among those who scored 165 in the LSAT.
Notice that this percentage is higher than the 24% we obtained before. This is because we have focus on a smaller portion of the sample, obtaining a smaller SD. In summary, when considering data for a vertical strip:

- Convert to standard units in the $x$ variable.
- Obtain the predicted value of the $y$ variable.
- Calculate the SD for the $y$ variable in the strip using RMS error.
- Convert to standard units in the $y$ variable and use the normal curve.
Slope and Intercept

All lines can be determined by a slope and an intercept.
The intercept is the height of the line when $x = 0$.

The slope is the rate at which $y$ increases, per unit increase in $x$. If the slope is negative then $y$ decreases as $x$ increases.
How do you get the slope of a regression line?

Example: A sample of 555 California men age 25-29 in 1993 was surveyed to find out about education and income. The data are summarized by

average education \( \approx 12.5 \) years, SD \( \approx 4 \) years

average income \( \approx \$21,500, \) SD \( \approx \$16,000, \) \( r \approx 0.35 \)
This means that, for every increase of one SD in education, there is an increase of \( r \) SD in income.

Thus, 4 extra years of education are worth an extra \( 0.35 \times 16,000 = 5,600 \) of income. So, each extra year is worth
\[
\frac{0.35 \times 16,000}{4} = 1,400
\]
this, is the slope of the regression line.