REGRESSION DIAGNOSTICS

- After fitting a regression line it is important to do some

  diagnostic checks

to verify that regression fit was OK.

- One aspect of diagnostic checking is to find the

  r.m.s. error

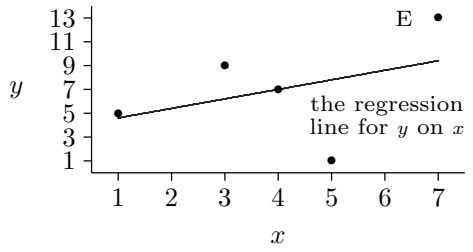
  - This is an overall measure of the "goodness of fit".
  - It measures how close the regression line is to all of the points — simultaneously.
  - The r.m.s. is computed using the residuals from a regression.
- It is also imperative to view residual plots.
- Also, to check another diagnostic:

  homoscedasticity.

RESIDUAL = ACTUAL − PREDICTED

- The following questions refer to this data set:

Point	x	y
A	1	5
B	3	9
C	4	7
D	5	1
E	7	13
Average	4	7
SD	2	4
r	2/5	



- For point E, the actual value of y is _____ .
- For point E, the predicted value of y using the regression method is _____ :
  - 7 is _____ units, or _____ SDs, _____ average for x.
  - Predicted value of y is _____ SDs, or _____ units, _____ average.
- For point E, the error in prediction is _____ .
  - Prediction error = actual value − predicted value.
- The statistical jargon for prediction error is _____ .
- The residual for point E is _____ . What is that saying about the scatter diagram?
  - Point E is _____ vertical units _____ the regression line.
- The residual for point D is -6.8. What is that saying about the scatter diagram?
  - Point D is _____ vertical units _____ the regression line.

THE R.M.S. ERROR OF PREDICTION

• There is a prediction error for each of the points A, B, C, D, and E. How can we measure the overall size of these errors?

  • By the so-called *r.m.s. error of the regression line*, i.e., by the root-mean-square of the residuals:

    • Square the residuals,

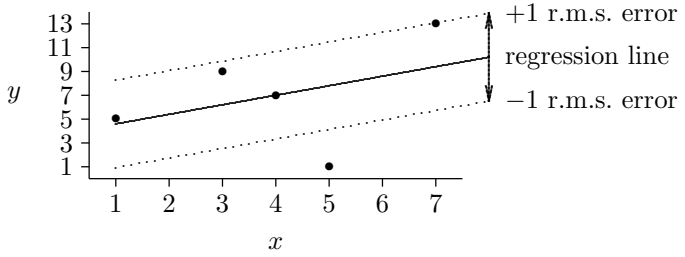
    • Average the squares,

    • Take the square root of the average square.

Point	x	Actual y	Predicted y	Residual	(Residual)^2
A	1	5	4.6		
B	3	9	6.2	2.8	7.84
C	4	7	7.0	0.0	0.00
D	5	1	7.8	−6.8	46.24
E	7	13	9.4	3.6	12.96

$$\text{Average (residual)}^2 = \frac{+7.84 + 0 + 46.24 + 12.96}{5} = 13.44$$

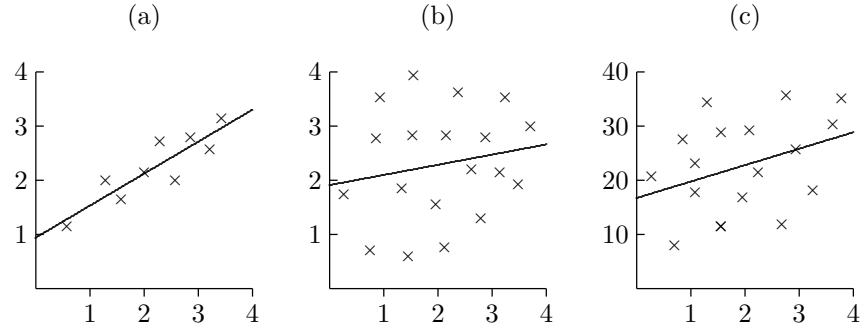
$$\text{r.m.s. error} = \sqrt{} = 3.67$$



• The points on a scatter diagram deviate from the regression line (up or down) by residuals which are similar in size to the r.m.s. error. The r.m.s. error is to the regression line as the SD is to the average.

EXERCISES

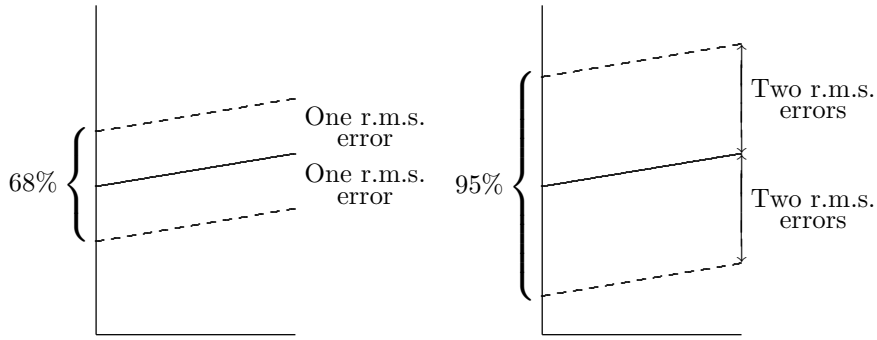
• Below are three scatter diagrams. The regression line has been drawn across each one. In each case, guess whether the r.m.s. error is about 0.2, 1, or 5.



• A regression line for predicting test scores has a r.m.s. error of 8 points.

  • About 68% of the time, the predictions will be right to within _____ points.

  • About 95% of the time, the predictions will be right to within _____ points.



COMPUTING THE R.M.S. ERROR

• How is the r.m.s. error for the regression line (of y on x) related to the SD of y and the correlation coefficient r?

  • r.m.s. error = $\sqrt{1-r^2} \times$ (the SD of y)

Point	x	y
A	1	5
B	3	9
C	4	7
D	5	1
E	7	13
Average	4	7
SD	2	4
r	2/5	

$r =$ _____

$\sqrt{1-r^2} =$ _____

SD of $y =$ _____

r.m.s. error $= \sqrt{1-r^2} \times$ SD of y

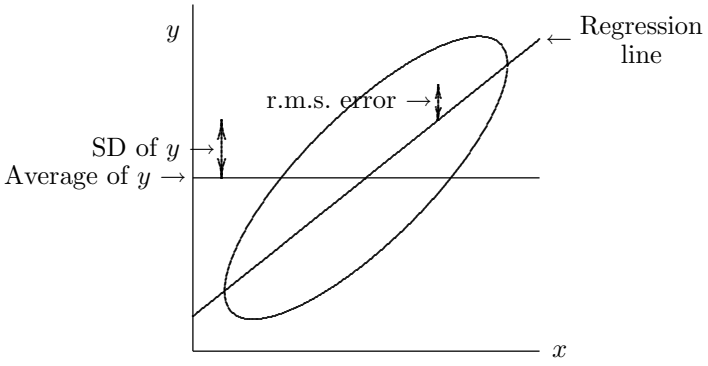
$=$ _____ .

• Does the short-cut formula work for all scatter diagrams?

  • Yes

• What is the interpretation of the quantity $\sqrt{1-r^2}$?

  • Predictions based on the regression line are more accurate than predictions using just the average of y, by a factor of $\sqrt{1-r^2}$.



• What are the units for the r.m.s. error?

  • The same as for ; that's what reminds you to multiply by the SD of y.

TWINS

• In one study of identical male twins, the average height was found to be about 68 inches, with an SD of about 3 inches. The correlation between the heights of the twins was about 0.95, and the scatter diagram was football-shaped.

(a) You have to guess the height of one of these twins, without any further information. What method would you use?

method =

(b) Find the r.m.s. error for the method in (a).

r.m.s. error =

(c) One twin of the pair is standing in front of you. You have to guess the height of the other twin. What method would you use? (For instance, suppose the twin you see is 6 feet 2 inches.)

method =

height of other twin =

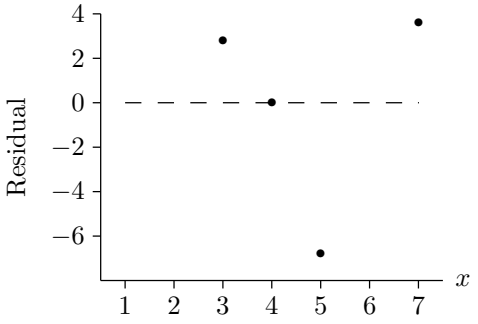
(d) Find the r.m.s. error for the method in (c).

r.m.s. error =

RESIDUAL PLOTS

- What is a residual plot?
  - A plot of the residuals versus x, or perhaps versus some other relevant variable.

Point	x	Residual
A	1	0.4
B	3	2.8
C	4	0.0
D	5	−6.8
E	7	3.6

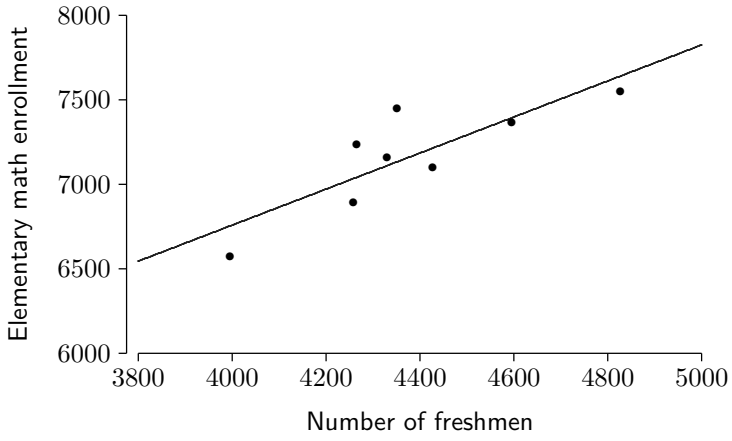


- In general, the residuals average out to _____ , and the regression line for the residual plot is _____ .
  - All the linear trend in the data is accounted for by the regression line for the data.
- What advantage does a residual plot have over the original scatter diagram?
  - A residual plot lets you use a larger vertical scale, which makes departures from linearity stand out more clearly.
- What do you look for in a residual plot?
  - Outliers
  - Unusual patterns
  - Other signs that it would be a mistake to make predictions using a straight line.

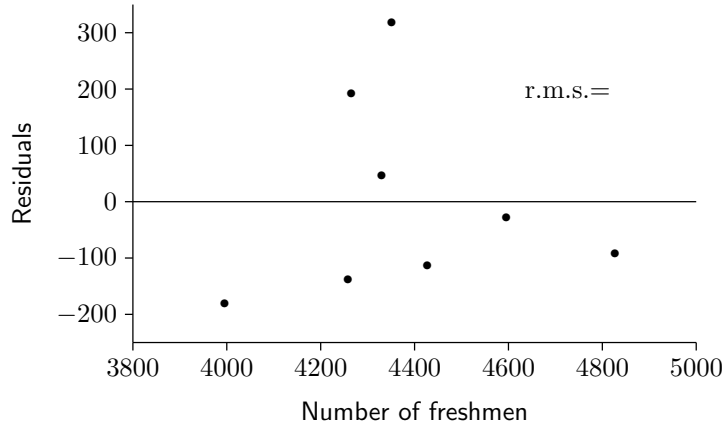
- Example. The math department of a large state university must plan in advance the number of sections and instructors required for elementary courses. The department hopes that the number of students in these courses can be predicted from the number of entering freshmen, which is known before the new students actually choose courses. The table below contains the data for recent years. The explanatory variable x is the number of students in the freshman class, and the response variable y is the number of students who enroll in math courses at the 100 level.

Year	1980	1981	1982	1983	1984	1985	1986	1987
x	4595	4827	4427	4258	3995	4330	4265	4351
y	7364	7547	7099	6894	6572	7156	7232	7450

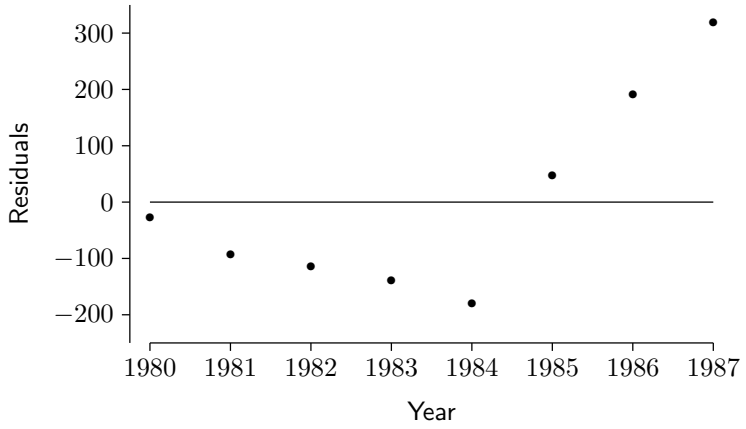
- Here is the scatter plot and the regression line for elementary mathematics enrollment on number of freshmen. How would you describe the nature of the relationship?



• Here is the residual plot for the regression of elementary math enrollment on number of freshmen. Do you see anything strange?

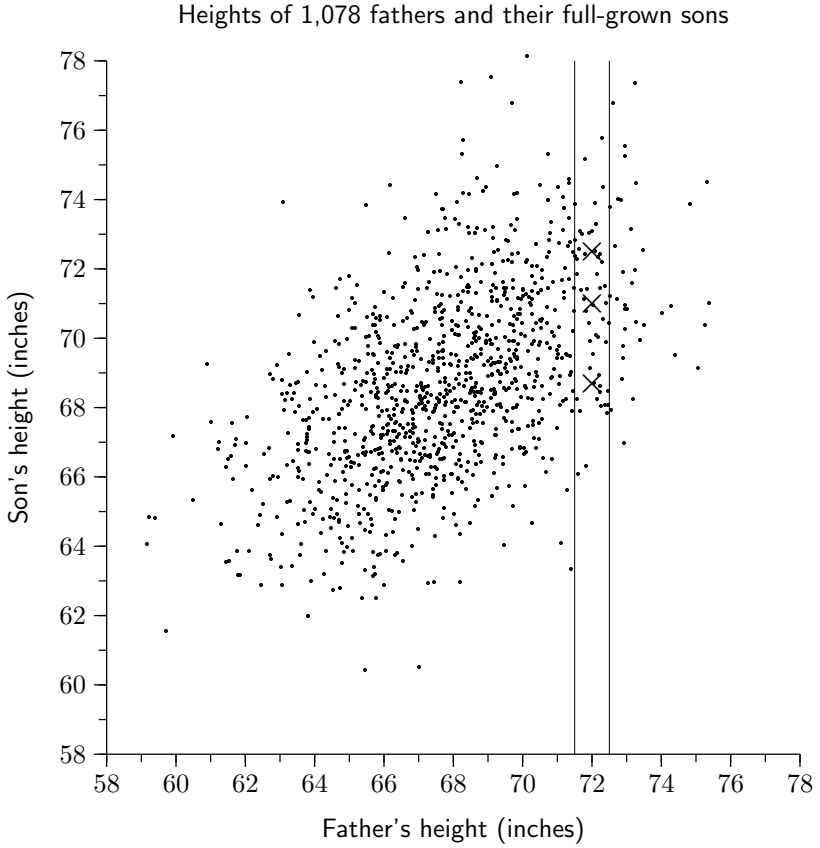


• What does this plot of the residuals against time reveal?



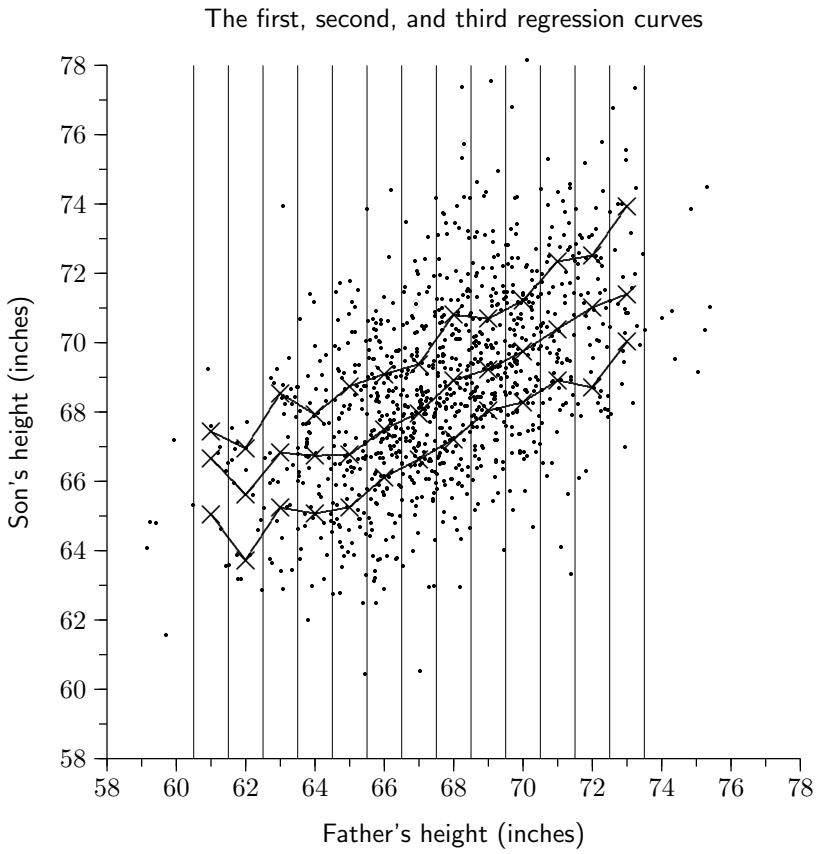
LOOKING AT VERTICAL STRIPS

• Reconsider Pearson's father-son data.



• How does the spread of son's height change as father's height varies?

• The following plot shows how the conditional quartiles of son's height vary with father's height.



• Reading from left to right, the top curve traces out the third quartile of son's height for fathers who are respectively 61, 62, ..., 73 inches tall, to the nearest inch. The middle curve traces out the corresponding medians, and the bottom curve traces out the corresponding first quartiles.

• What are the conclusions?

  • The variation in son's height, as measured by the interquartile range, is about the same in all the vertical strips.

• The interquartile range doesn't vary much from strip to strip, but the (full) range does. Why is that?

  • The full range depends on the extremes, and the extremes depend in part on how many points there are in the strip.

HOMOGENEITY OF VARIANCE

• What does it mean to say that a scatter diagram is *homoscedastic*?

  • All the vertical strips in the diagram show similar amounts of spread. (*Homo* means "same", *scedastic* means "scatter".)

• What bearing does that have on prediction?

  • The prediction errors are similar all along the regression line.

  • The spread in each vertical strip is about the r.m.s. error.

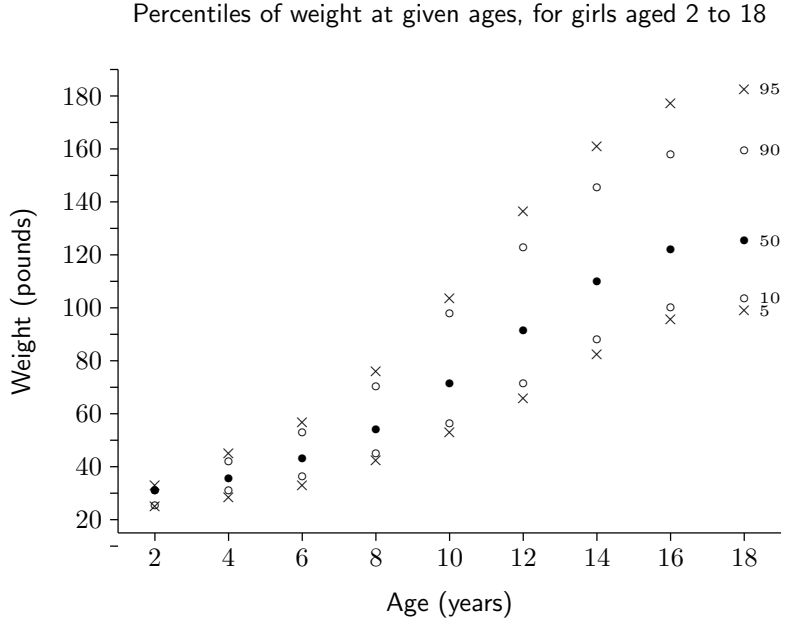
• How do you recognize homoscedasticity?

  • The residual plot should look football-shaped.

  • For a more careful analysis, construct the first and third quartile regression curves and check whether they are roughly parallel. (You could also construct the regression curve for the Average + 1SD, and for the Average − 1SD, and check whether those two curves are roughly parallel.)

HETEROGENEITY OF VARIANCE

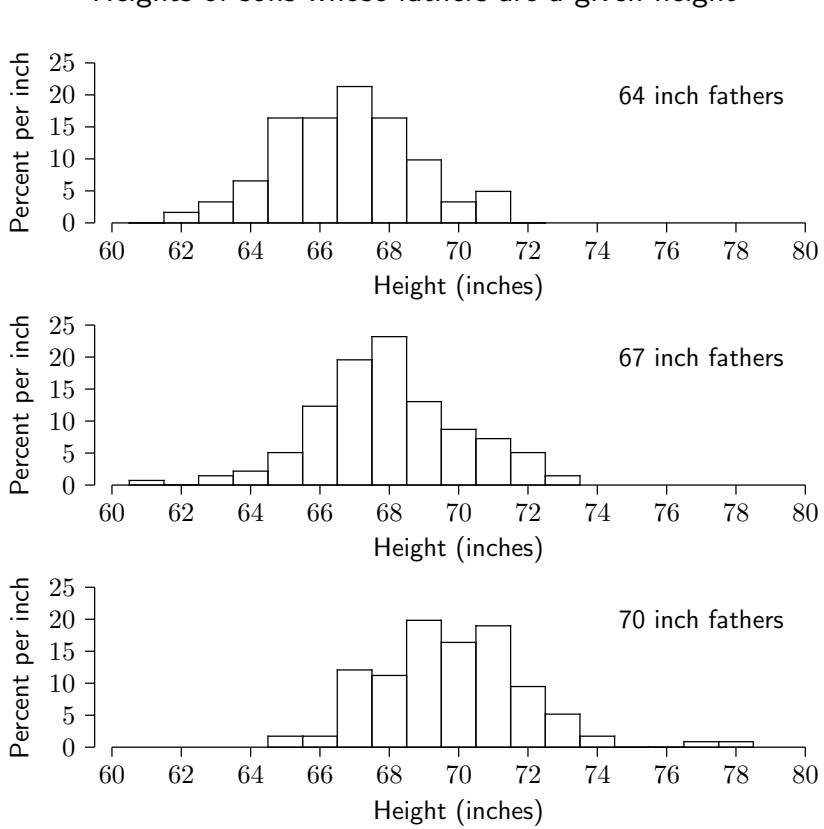
- If you haven't got homoscedasticity, what have you got?
  - Heteroscedasticity. That means that the prediction errors are different in different parts of the scatter diagram. (*hetero* means "different".)
  - In some vertical strips, the spread is greater than the r.m.s. error; in others, the spread is smaller.
- Example



  - How does weight vary with respect to age?
    - In regards to location?
    - In regards to spread?
  - Is the distribution of weight at age 18 normal, or non-normal?

USING THE NORMAL CURVE INSIDE A VERTICAL STRIP

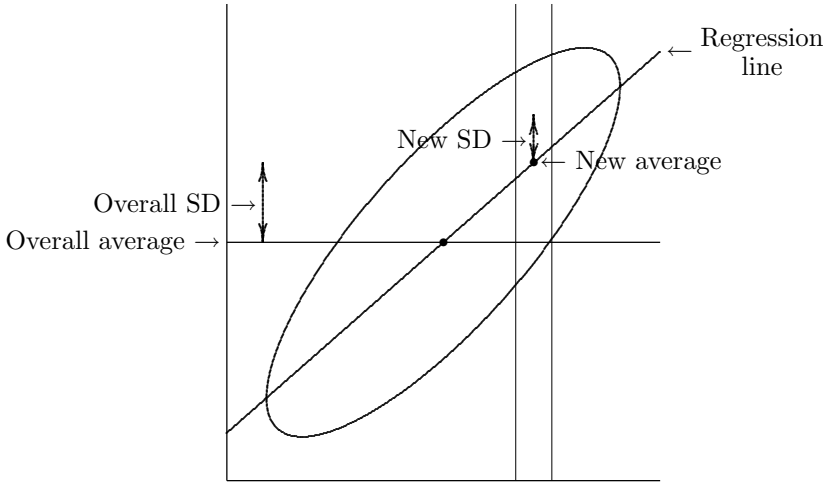
- Reconsider Pearson's father-son data.



- These histograms show the distribution of height among sons whose fathers are respectively 64, 67, and 70 inches tall, to the nearest inch. Note that each of these so-called conditional distributions of sons height is approximately normal. This will be true for any football-shaped scatter diagram.

USING THE NORMAL CURVE INSIDE A VERTICAL STRIP

- The normal approximation can be used on the points inside a narrow vertical strip on a _____ shaped scatter diagram.
  - Think of the y values in this strip as a whole new data set.
  - The new average is estimated by the _____ .
  - The new SD is about equal to the _____ .
  - The normal approximation can be done as usual, based on the _____ average and SD for y, rather than the _____ average and SD for y.



- Statisticians often use the terms *conditional* for "New", and *unconditional* for "Overall".

- Pearson and Lee obtained the following results for about 1,000 families:

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

The scatter diagram is football shaped.

- What percentage of the women were over 5 feet 8 inches?
  - _____ Average = _____
  - _____ SD = _____
  -

• Pearson and Lee obtained the following results for about 1,000 families:

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

The scatter diagram is football shaped.

• Of the women who were married to men of height 6 feet, what percentage were over 5 feet 8 inches?

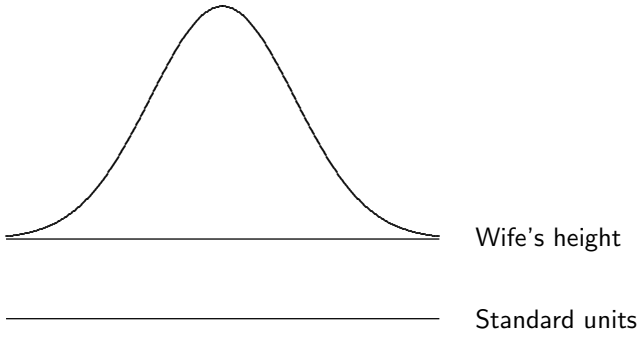
- _____ Average = _____

(These women are somewhat taller than the overall average.)

- _____ SD = _____

(These women are a more homegeneous group, will less variation in their heights.)

- Percentage = _____



SUMMARY

• With a regression line the difference between the actual value and the predicted value is called a *residual*.

• The RMS error measures the "overall size" of the residuals.

• The RMS error for the regression line of y on x can be figured as

$$\sqrt{1-r^2} \times \text{SD of } y$$

• The residual plot shows the residuals! If a pattern appears then the linear regression may not have captured all the structure in the data set.

• Data sets with similar amounts of spread in all vertical strips are called *homoscedastic*. Those that show different amounts of spread are called *heteroscedastic*.

• If the scatter diagram is football-shaped, the normal curve can be used inside a vertical strip, using the conditional mean and SD of y for the given value of x.