

3.2

Influences

- **Correlation r is not resistant.**
 - One unusual point in the scatterplot greatly affects the value of r .
- **Extrapolation is not very reliable.**
- **LSRL also not resistant.**
 - A point extreme in the x direction with no other points near it pulls the line toward itself.
 - This point is influential.

Outliers and Influential Observations in Regression

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line.

Beware correlations based on averages

Correlations based on averages are usually too high when applied to individuals.

- **Example:**

- If we **plot the average height of young children** against their age in months,
 - we will see a very strong positive association with correlation near 1.
- **But individual children of the same age vary a great deal in height.**
 - A **plot of height against age for individual children will show much more scatter and lower correlation** than the plot of average height against age.

Review This Problem: Work Through

Example – Corrosion and Strength

- Consider the following data from the article, “The Carbonation of Concrete Structures in the Tropical Environment of Singapore” (Magazine of Concrete Research (1996):293-300 which discusses how the corrosion of steel (caused by carbonation) is the biggest problem affecting concrete strength:
 - x = carbonation depth in concrete (mm)
 - y = strength of concrete (Mpa)

x	8	20	20	30	35	40	50	55	65
y	22.8	17.1	21.5	16.1	13.4	12.4	11.4	9.7	6.8

- Define the Explanatory and Response Variables.**
- Plot the data and describe the association.**
- Answers must be in context for given problem.**

1) Create a Graph to visualize and describe the Association

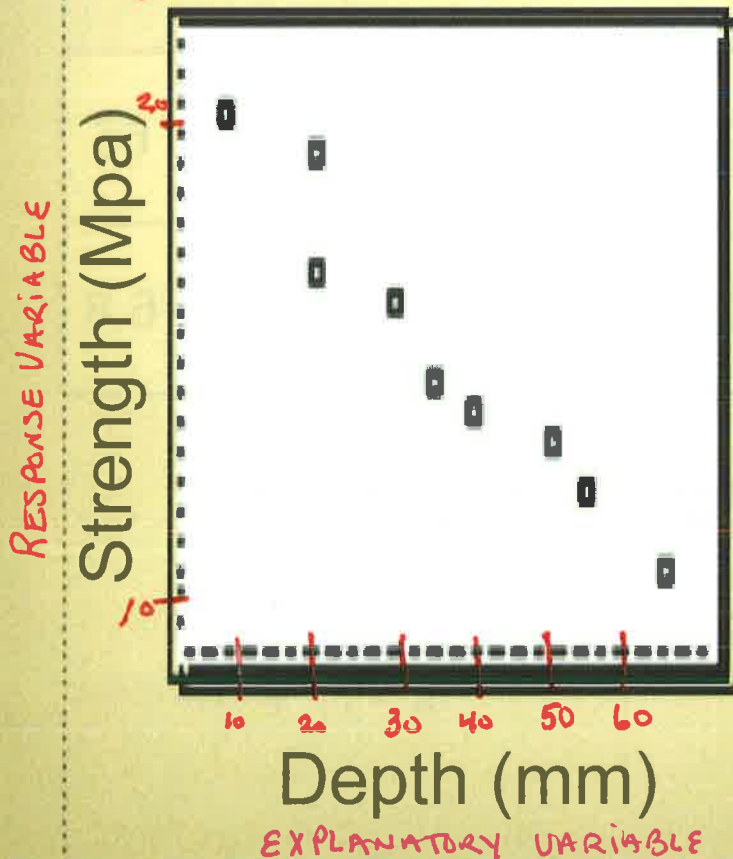
L1	L2	L3	3
8	22.8		
20	17.1		
20	21.5		
30	16.1		
35	13.4		
40	12.4		
50	11.4		

L3(1)=

Plot2 Plot3
Off
Type:
Xlist: L1
Ylist: L2
Mark:

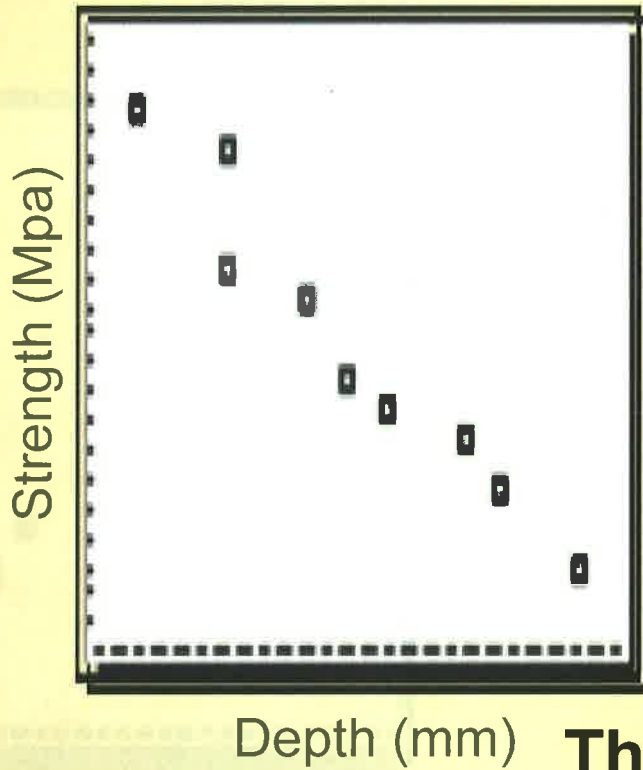
MEMORY
4:ZDecimal
5:ZSquare
6:ZStandard
7:ZTrig
8:ZInteger
9:ZOOMStat
0:ZOOMFit

Remember Labels + Scales



There is a strong, negative, linear relationship between depth of corrosion and concrete strength. As the depth increases, the strength decreases at a constant rate.

2) Find the Means and Standard Deviations for Depth and Strength and describe in context.



```

EDIT  CALC  TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg
5:QuadReg
6:CubicReg
7:QuartReg
      X=35.88888889
      Σx=323
      Σx²=14339
      Sx=18.5300057
      σx=17.47025691
      n=9
    
```

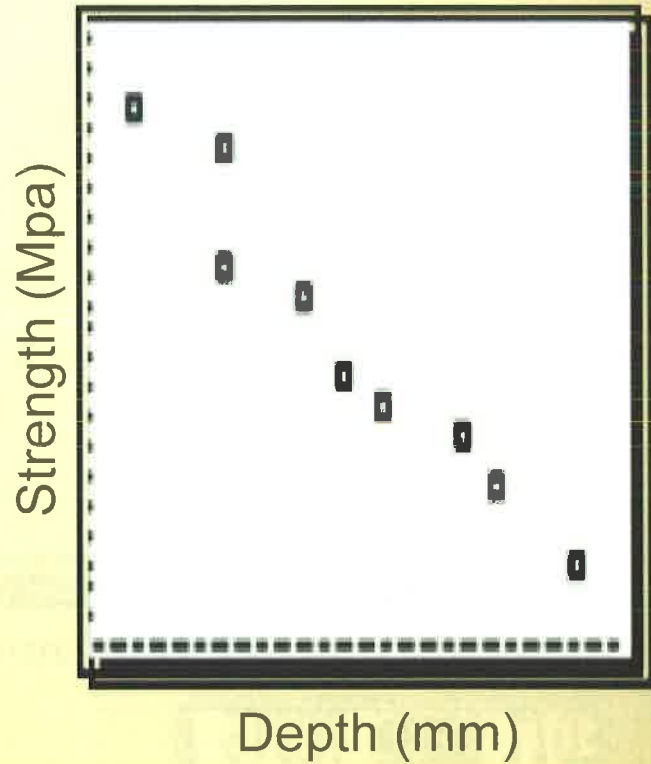
The mean **depth of concrete** is 35.89mm with a standard deviation of 18.53mm.

The mean **strength of concrete** is 14.58 Mpa with a standard deviation of 5.29 Mpa.

3) Find the Correlation Coefficient and describe in context.

4) Find the equation of the Least Squares Regression Line

The correlation coefficient ($r = -.968$) quantifies there is a **Strong, Negative, LINEAR** association between depth of concrete corrosion and strength of concrete.



```

EDIT  $\square$   $\square$  TESTS
4  $\uparrow$  LinReg(ax+b)
5: QuadReg
6: CubicReg
7: QuartReg
8  $\square$  LinReg(a+bx)
9: LinReg(a+bx) L1,
0  $\downarrow$  L2
    
```

```

LinReg
y=a+bx
a=24.51683116
b=-.276939568
r^2=.9375144639
r=-.9682533056
    
```

LSRL Equation:

$$\text{strength} = 24.52 - 0.28(\text{depth})$$

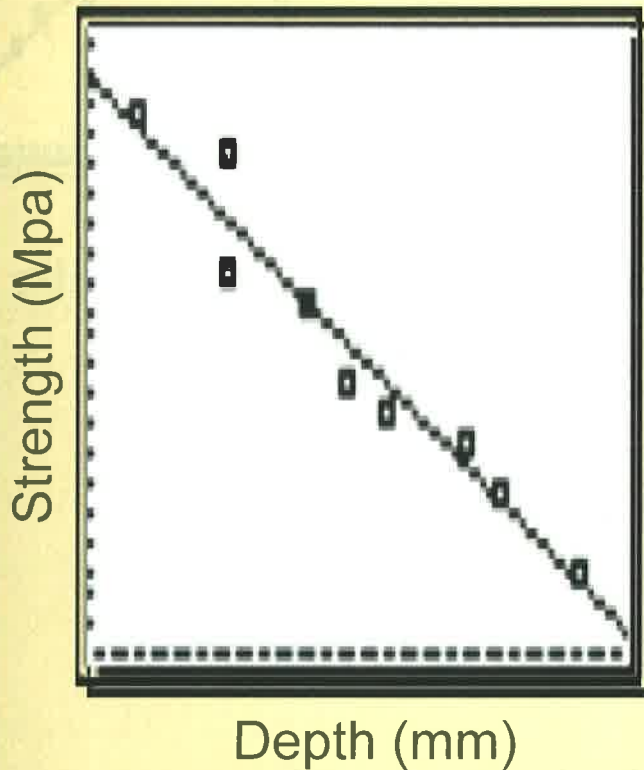


y



x

5) Describe LSRL in Context



LSRL Equation:



$$\text{strength} = 24.52 - 0.28(\text{depth})$$

The slope is $b = -0.28$.
 For every increase of 1mm in depth of concrete corrosion, we predict a 0.28 Mpa decrease in strength of the concrete.

- Use these steps to store the LSRL in Y_1 and overlay it on the scatterplot.

LinReg(a+bx)

VAR

2: Pa

3: Po

4: Or

5: Y1

6: Y2

7: Y3

8: Y4

9: Y5

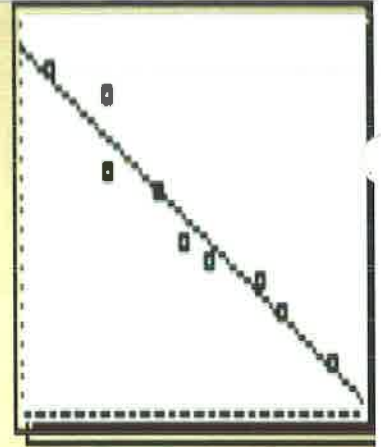
10: Y6

11: Y7

LinReg(a+bx) Y_1

Plot2 Plot3
 $Y_1 = 24.92406703$
 $Y_2 =$
 $Y_3 =$
 $Y_4 =$
 $Y_5 =$
 $Y_6 =$
 $Y_7 =$

6) Use the prediction model (LSRL) to determine the following:



What is the predicted strength of concrete with a corrosion depth of 25mm?

$\text{strength} = 24.52 + (-0.28)\text{depth}$

$\text{strength} = 24.52 + (-0.28)(25)$

The predicted strength is 17.59 Mpa.

What is the predicted strength of concrete with a corrosion depth of 40mm?

$\text{strength} = 24.52 + (-0.28)(40)$

The predicted strength is 13.44 Mpa.

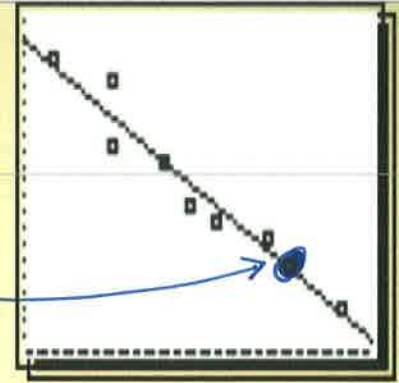
How does this prediction compare with the observed strength at a corrosion depth of 40mm?



RESIDUALS

7) Interpret Residuals (from previous slide)

$(40, 12.4)$
"ACTUAL X+Y
VALUES"



- ☑ We calculated the **predicted strength** when the corrosion depth was 40mm to be **13.44 Mpa**
- ☑ From the given data table, we can find the **observed strength** when corrosion=40mm is to be **12.4mm**
- ☑ The prediction did not match the observation.
 - That is, there is “error” or “residual” between our prediction and the actual observation.
 - RESIDUAL = Observed y - Predicted y
 - The residual when corrosion=40mm is:
 - residual = 12.4 - 13.44
 - residual = -1.04

Assessing the Model

8) Is the model appropriate?

9) What is the strength of the model?

Is the LSRL the most appropriate prediction model for strength?

✓ r suggests it will provide strong predictions...

✓ can we do better?

To determine this, we need to study the residuals generated by the LSRL.

Make a residual plot.

Look for a pattern.

If no pattern exists, the LSRL may be our best bet for predictions.

If a pattern exists, a better prediction model may exist...

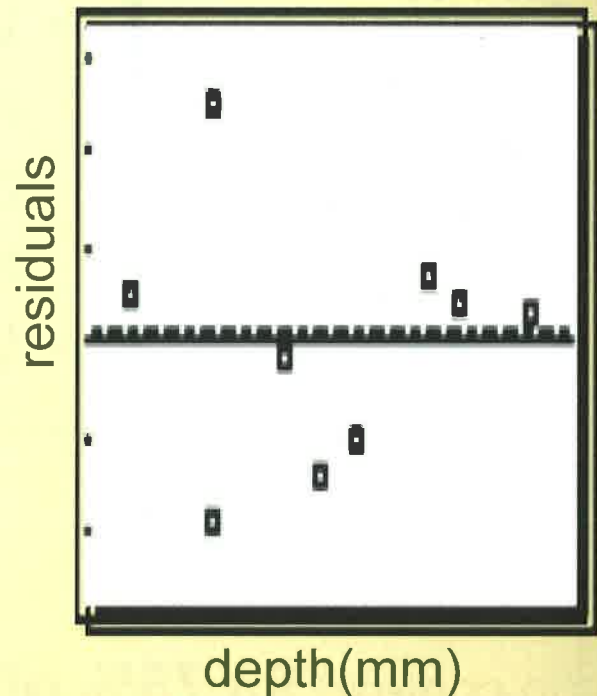
8) Review the Residual Plot to see if our model is appropriate

- ✓ Construct a Residual Plot for the (depth, strength) LSRL.

```
Plot2 Plot3
Off
Type: [ ] [ ] [ ]
      [ ] [ ] [ ]
Xlist:L1
Ylist:Q2
Mark:
```

```
NAME OPS MATH
1:L1
2:L2
3:L3
4:L4
5:L5
6:L6
RESID
```

```
Plot2 Plot3
Off
Type: [ ] [ ] [ ]
      [ ] [ ] [ ]
Xlist:L1
Ylist:RESID
Mark: [ ] + .
```



- ✓ There appears to be **no pattern** to the residual plot...
- ✓ **therefore, the LSRL may be our best prediction model.**

9) Review the Coefficient of Determination (r^2) to assess the strength of our model

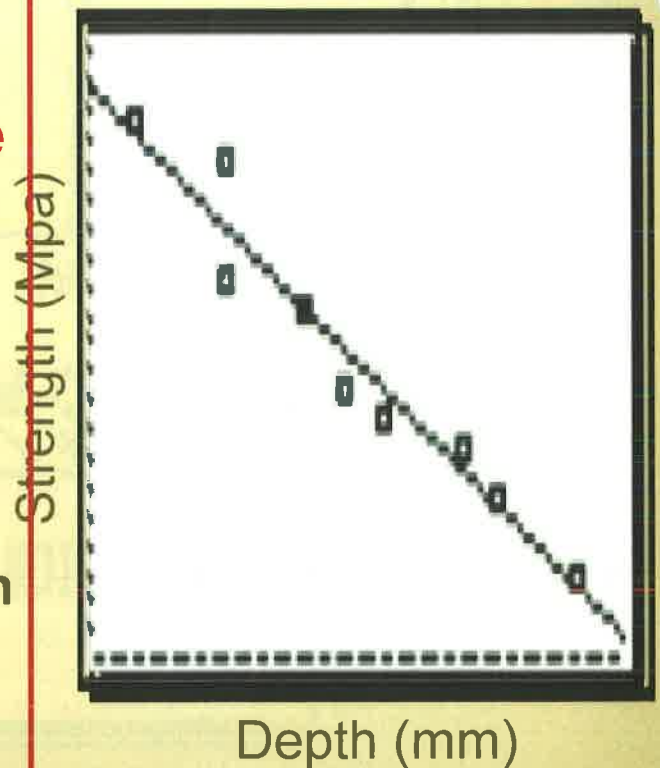
```
LinReg  
y=a+bx  
a=24.51683116  
b=-.276939568  
r2=.9375144639  
r=-.9682533056
```

- We know what “r” tells us about the linear association between depth and strength.
- **What about r^2 ?**

$$r^2 = .9375$$

(in context) **93.75% of the variability in predicted strength can be explained by the LSRL on depth.**

(6.25% of the variability can NOT be explained by our model. This is a very strong model.)



Summary

- ★ When exploring a bivariate relationship:
 - ★ Make and interpret a scatterplot:
 - ★ Strength, Direction, Form
 - ★ Describe x and y:
 - ★ Mean and Standard Deviation in Context
- ★ Find the Least Squares Regression Line.
 - ★ Write in context.
- ★ Construct and Interpret a Residual Plot.
- ★ Interpret r and r^2 in context.
- ★ Use the LSRL to make predictions...

