Today, we will:

- Do some more example problems – regression analysis
- Review the pdf module: **Outlier Points** and do some example problems

Example: Regression Analysis

Given: Twenty data points of (x,y) pairs with lots of scatter (see Excel spreadsheet on website for the raw data).

Data point	x	y
1	0	1.924
2	0.1	2.377
3	0.2	2.088
4	0.3	3.245
5	0.4	3.031
6	0.5	2.779
7	0.6	3.186
8	0.7	4.102
9	0.8	4.278
10	0.9	3.701
11	1	3.654
12	1.1	3.991
13	1.2	3.891
14	1.3	3.790
15	1.4	3.437
16	1.5	4.080
17	1.6	3.204
18	1.7	3.130
19	1.8	2.614
20	1.9	2.066

See EXCEL & MATLAB files posted on the course website ↓

$\boxed{\text{Regression\_nonlinear\_fits.xls} \quad \text{" " .m}}$

To do: Perform regression analysis – linear, quadratic, and cubic – and compare how the fitted curves fit to the data points.

Solution:

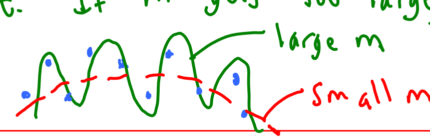
See Excel spreadsheet – I will show in class how to do the regression analysis in Excel.

Conclusions

- As polynomial order $m \uparrow$, mutiple R also $\uparrow$ (correlation coefficient) (i.e. the fit improves with increasing m)
- As polynomial order $m \uparrow$, standard error $S \downarrow$ (better overall agreement with the data points)

$\boxed{\text{But – bigger } m \text{ is not always better!}}$ ⇓

In this example the cubic fit ($m=3$) is the best. If m gets too large the curve fit will have too many "wiggles"



Example: Outliers – single set of data

Given: Four data points are measured.

Data point	x
1	38
2	42
3	44
4	53

38 — min

53 — max

sample mean = 44.25 $= \bar{x}$

sample standard deviation =6.3443 $= S$

[We suspect either the min or the max to possibly be an outlier]

To do: Eliminate any "official" outliers, one at a time.

Solution:

- @ $n=4$, look up Thompson Tau → $\underline{\tau = 1.4250}$ (either from the table or the eqn.)
- @ min x, $\delta = |38 - \underbrace{44.25}_{\bar{x}}| = 6.25$
- @ max x, $\delta = |53 - 44.25| = 8.75$ — This one is biggest, so this is the one we suspect may be an outlier
- Compare δ_{max} & τS → $\left.\begin{array}{l}\delta_{max} = 8.75 \\ \tau \cdot S = (1.4250)(6.3443) = 9.0406\end{array}\right\}$

$\boxed{\text{Since } \delta_{max} < \tau S \ (8.75 < 9.0406), \text{ There are no outliers}}$

Example: Outliers – single set of measurements

Given: Janet takes 12 temperature measurements ranging from 23.0°C (lowest reading) to 25.7°C (highest reading).

- The sample mean of all 12 readings is 24.88°C.
- The sample standard deviation is 1.04°C.

To do: Is there an "official" outlier?

Solution:

- @ $n=12$, look up or calculate τ → $\tau = 1.8290$

  $\tau S = (1.8290)(1.04°C) = \underline{1.902°C}$

- @ minimum reading, $\delta = |23.0 - 24.88| = 1.88°C$ — δ_{max} → The minimum reading is the one we suspect may be an outlier
- @ maximum reading, $\delta = |25.7 - 24.88| = 0.82°C$
- Compare: $\delta_{max} = 1.88°C < \tau S = 1.902°C$

$\boxed{\text{Thus, there are no outliers}}$

Example: Outliers – single set of measurements

Given: Eleven measurements of pump efficiency are taken (listed in increasing order below):

58, 72, 74, 74, 76, 77, 80, 80, 82, 85, and 92%

Point # 1 2 3 4 5 6 7 8 9 10 11

(*a*) To do: Are there any "official" outliers? If so, remove them. How many "good" measurements are left?

(*b*) To do: Based on the measurements that are left, estimate the population mean and its confidence interval to 95% confidence.

(*c*) To do: Based on the measurements that are left, estimate the population standard deviation and its confidence interval to 95% confidence.

Solution:

(a). First calculate sample mean & sample standard deviation: $\bar{x} = 77.273$, $S = 8.580$

- At x_{min}, $\delta_1 = |58 - 77.273| = 19.27$ → Point x_1 is the suspected outlier since $\delta_{max} = \delta_1 = 19.27$

Point #1

- At x_{max}, $\delta_{11} = |92 - 77.273| = 14.73$

Point #11

- TABLE → @ $n = 11$, $\tau = 1.8153$ → $\tau S = (1.8153)(8.580) = 15.575 = \tau S$
- Compare: $\delta_{max} = \delta_1 = 19.27 > \tau S = 15.575$ → $\boxed{\text{So, Point 1 is an outlier}}$ ✩

NOTE: Other data points may or may not also have $\delta > \tau S$ but it does not matter. $\boxed{\text{WE CAN REMOVE ONLY ONE OUTLIER AT A TIME}}$ ✩

- Remove point 1 & start all over, now with 10 points instead of 11 pts.

✩ NOTE: Since we removed a data point, we have to re-calculate $\bar{x}$ & S ✩

ROUND 2: • New $\bar{x} = 79.20$ (sample mean of points $x_2, x_3, \ldots x_{11}$)

$S = 6.0332$ (sample standard deviation of points $x_2, x_3, \ldots x_{11}$)

- τ is also different, since now $n = 10$ instead of 11

TABLE @ $n = 10$, $\tau = 1.7984$ → $\tau S = 10.850$

- The maximum deviation turns out to be point 11:

$$\delta_{max} = \delta_{11} = |92 - 79.20| = 12.80 = \delta_{max}$$

- Compare: $\delta_{max} = \delta_{11} = 12.80 > \tau S = 10.850$ → $\boxed{\text{Point 11 is also an outlier !}}$

ROUND 3 → Remove Point 11 & we are left with only 9 data points: (Points 2 through 10)

x →	~~58~~	72	74	74	76	77	80	80	82	85	~~92~~
Point #	1	2	3	4	5	6	7	8	9	10	11

· New $\bar{x}, S, \tau$ → $\bar{x} = 77.778$, $S = 4.2655$, $\tau = 1.7770$ (@ $n = 9$)

→ $\tau S = 7.5798$

· Must calculate the deviations again, based on the revised value of $\bar{x}$

min x → $\delta_2 = |72 - 77.778| = 5.778$

max x → $\delta_{10} = |85 - 77.778| = 7.222$ → $= \delta_{max} = \delta_{10}$ this time

· Compare: $\delta_{max} = \delta_{10} = 7.222 < \tau S = 7.5798$ → **NO MORE OUTLIERS**

Bottom line: We removed two official outliers (points 1 & 11), and are left with the remaining 9 (points 2 through 10)

(b) Estimate μ based on the remaining data points, to 95% confidence

Solution: This is review → $\bar{x} = 77.778$ $\quad df = n-1 = 9-1 = 8$

$S = 4.2655$ $\quad t_{\alpha/2} = 2.3060$ (TABLE)

So, we predict $\mu = \bar{x} \pm t_{\alpha/2} \dfrac{S}{\sqrt{n}} = 77.778 \pm \dfrac{(2.306)(4.2655)}{\sqrt{9}}$

$$\boxed{\mu = 77.78 \pm 3.28 \ \% \text{ to } 95\% \text{ confidence}}$$

(c) Similarly, for review estimate σ to 95% confidence

* DO THIS ON YOUR OWN FOR PRACTICE

ANSWER: $\boxed{2.88 \le \sigma \le 8.17 \text{ to } 95\% \text{ confidence}}$