27th March '2024

Last we left off, we learned to

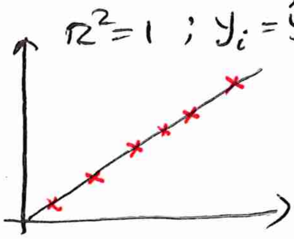
- fit a Straight Line Model in R `lm(y ~ x)`
- interpret the output of `lm( )`
  - Stating fitted model
  - Finding sample s.d.
  - Testing hypothesis on α and β
  - R^2 to assess model fit ★ Continue with this today.
- Also revisited the anatomy of Boxplots.

R^2 = Coefficient of Determination.

- Denotes the proportion of variation in y (response variable) explained by the linear model $(\hat{\alpha} + \hat{\beta} x)$

$R^2 \in [0, 1]$

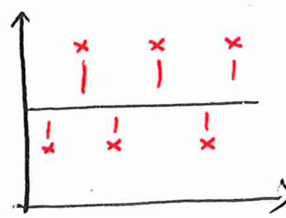
$R^2 = 1$; $y_i = \hat{y}_i, \forall i$



"perfect fit"

$R^2 = 0$

$\hat{y}_i = \bar{y}$



"perfect unfit"

⊛ Highlights the importance of plotting the data.

After we fit the model, R^2 value can be useful to assess the quality of the fit, when our MODEL ASSUMPTIONS ARE SATISFIED!

[*] Pre-fitting : Summaries and plots of data.

- Scatterplot y vs x – assess a linear trend
- Histogram of y – assess Normality of y.

⊠ Model fit : R^2, hypothesis tests and estimates of α and β.

⊠ Post-fitting : Residual Analysis ($\hat{\varepsilon}_i$) ; $\varepsilon_i \overset{iid}{\sim} N(0, \delta^2)$

$\varepsilon_i = y_i - \alpha - \beta x_i$

⊙ Residual vs fitted plot

↳ The constant variance assumptions

↳ The linear model assumptions.

↳ Outliers : points with large residuals.

↳ Influential points : points that have a large impact on the regression/model fitting.

⊙ Q-Q plots

↳ Normality assumptions of $\hat{\varepsilon}_i$'s

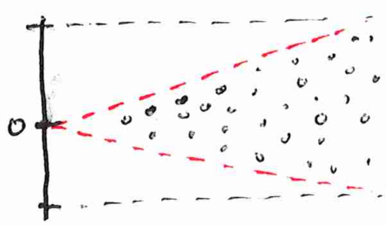
↳ Outliers.

☆☆ Deriving R^2 is not going to be tested.

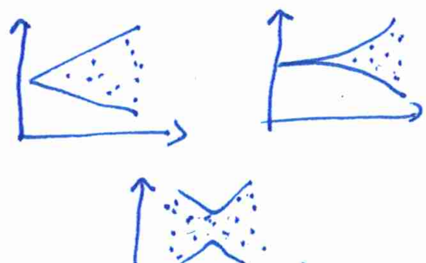
☆☆ ANOVA is not going to be tested.

Anscombe's Data : showcase how crucial residual analysis is,

Constant variance assumption, when violated, results in residuals vs fitted plots with non-horizontal bands.



In this case, to stabilize the variance, we instead fit $\log(y) \sim X$.



Look For Funnel shapes, bowtie shapes

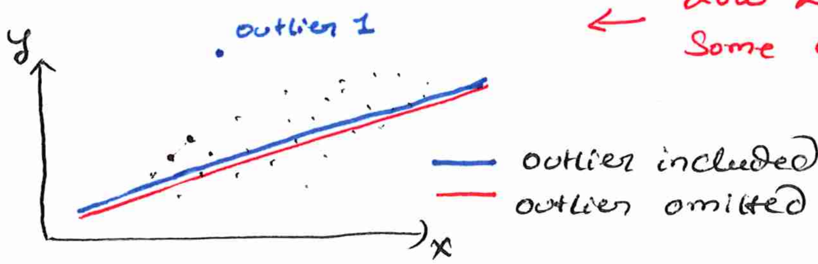
↳ Indicators of non-constant variances

heteroskedasticity.

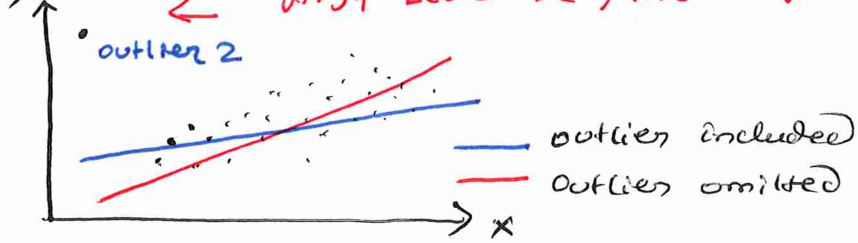
Outliers and Influential points :-

- An outlier is a point that falls far from the other data points.
- If, the parameter estimates change a great deal when a point is removed from the calculations, the point is said to be "Influential"

Low Leverage
Some influence.



High Leverage ; High influence.

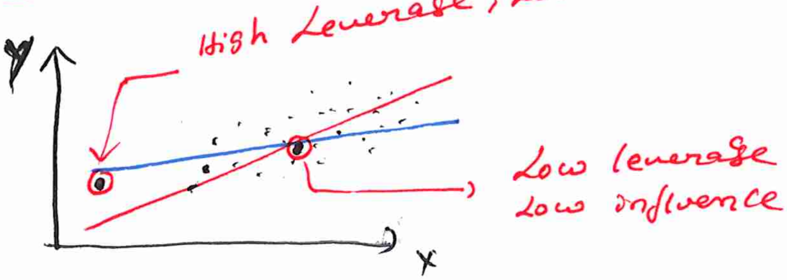


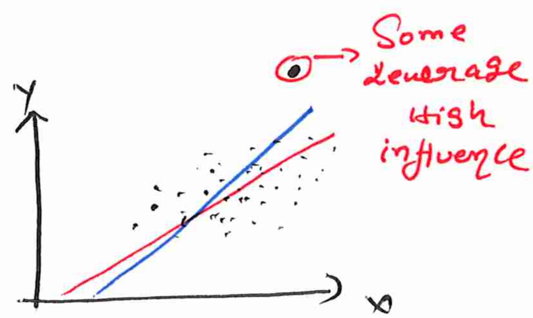
Extreme in X, it's falling away out of the other value of X

- Points with extreme values of X, are said to have high leverage.
- High Leverage points have a greater ability to move the line.
- If these points fall outside the overall pattern, they can be influential.

Just FYI !-

High Leverage ; Low influence.





```
9   12 12 12  8 10.84 9.13  8.15  5.56
10   7  7  7  8  4.82 7.26  6.42  7.91
11   5  5  5  8  5.68 4.74  5.73  6.89
```

Below is the output from R for the fits of the linear models, $Y = \alpha + \beta X + \epsilon$, for each of the four pairs of x and y. Yes, they all have the SAME fit; SAME R^2, SAME coefficients estimates, SAME everything, so I only included one version. The graphs of the data pairs are quite different however.

```
> summary(ans.lm1)

Call:
lm(formula = y1 ~ x1, data = anscombe)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

From this output, we can find

$\hat{y}_1 = 3 + 0.5x_1$

$s^2 = 1.237^2$

In practice, R^2 = 0.6 is very good for a model fit.

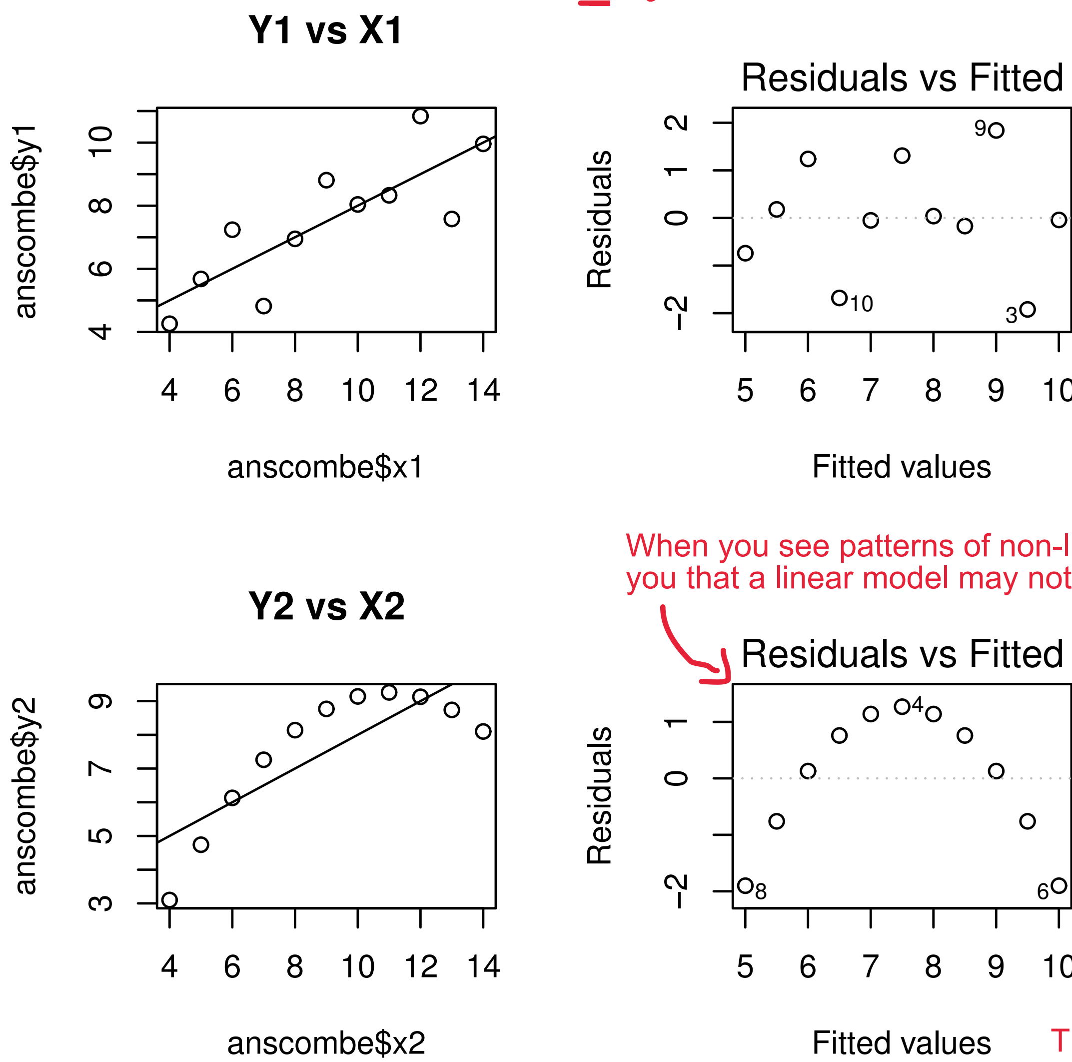
Figure 6.8: R output, fit of linear model of Y on X

Figures 6.9 and 6.10 below show the scatterplots of the pairs of Anscombe's data together with the fitted linear model in the first columns. Plots of residuals versus fitted values from linear model fits are shown in the second columns. A linear model seems appropriate for the first pair, $(x1, y1)$. The second pair, $(x2, y2)$ require a quadratic model. The third pair has an outlier which raises the regression line. The fourth pair has an influential point which totally determines the line. Thus, although their R^2 values are all the same, we see that the linear model fits are all very different for the four pairs.

Want to see random scattering of points within a horizontal band around 0.

Specially, we want the points to be within

$\pm 3s$



When you see patterns of non-linearity, this tells you that a linear model may not be appropriate.

Try fitting a Quadratic model

Figure 6.9: Anscombe pairs 1 and 2, Scatterplots; Residual plots

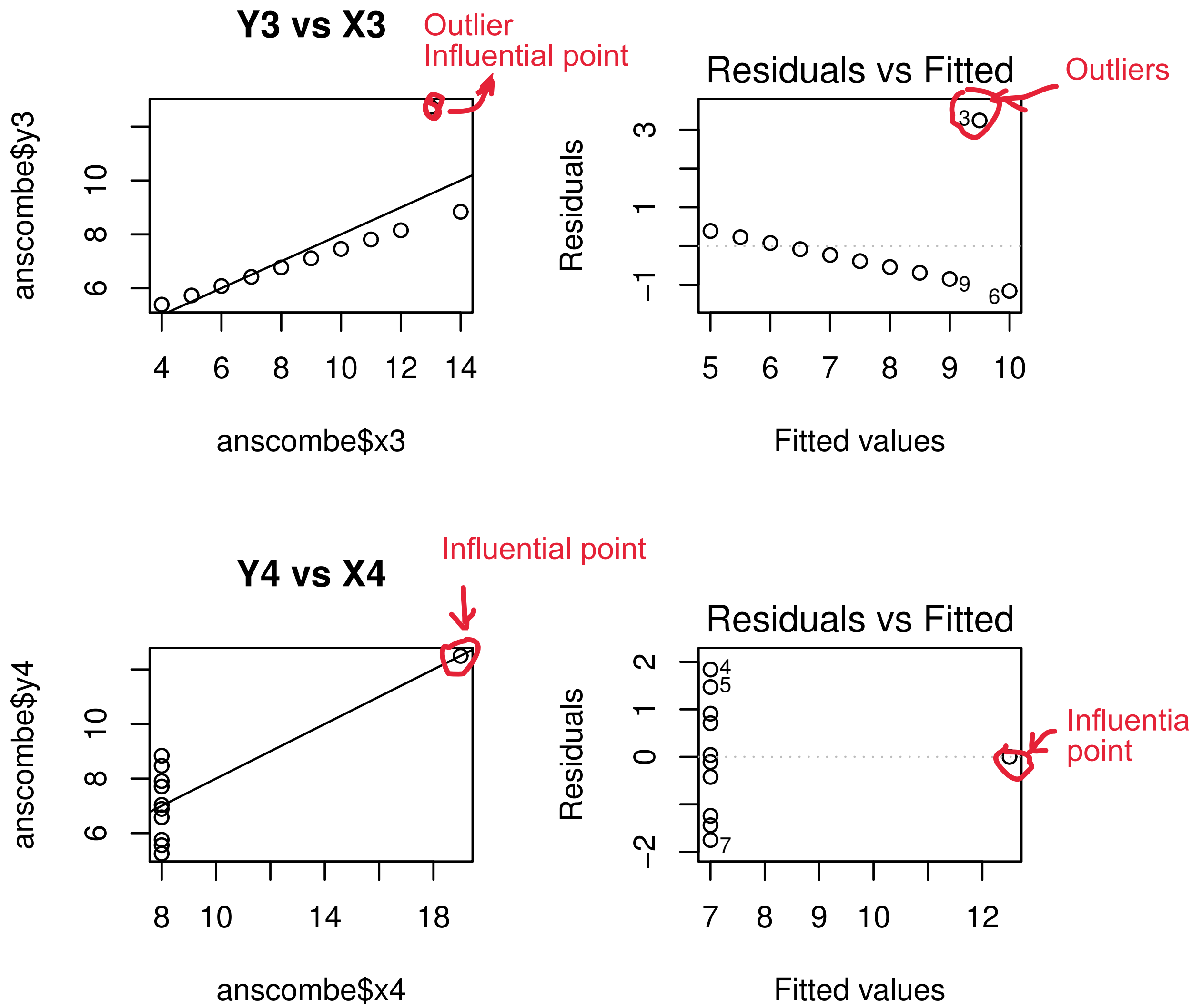


Figure 6.10: Anscombe pairs 3 and 4, Scatterplots; Residual plots

R Code for Anscombe analyses:

```
anscombe
plot(anscombe$x1, anscombe$y1, main='Y1 vs X1')
ans.lm1<-lm(y1~x1, data=anscombe)
abline(ans.lm1)
```