THE UNIVERSITY OF WESTERN AUSTRALIA

STAT2401 Analysis of Experiments

Lecture Week 6 **Dr Darfiana Nur**

Aims of Lecture Week 6

- **Aim 1 Critically Assessing the Regression Model**

(Sheather Ch 3.1, Moore et al Ch 10)

- **Aim 2 Diagnostic Checking**

(Sheather Ch 3.2, Moore et al Ch 10)

- **2.1 Residual Analysis**
- **2.2 Outliers and Leverage**

- **Aim 3 Transformation**

(Sheather Ch 3.3, Moore et al Ch 10)

3.1 The detail

Recap: SLR – Model and assumptions

- Simple linear regression model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- To complete the specification of the model, we assume
  1. $E(\epsilon_i) = 0$, for all i (zero means residuals)
  2. $\text{var}(\epsilon_i) = \sigma^2$, for all i (constant variance residuals)
  3. ϵ_i and ϵ_j are independent for all $i \neq j$ (independence residuals)
  4. $\epsilon_i \sim N(0, \sigma^2)$ if we wish to make inferences about the regression model (normality of residuals)

- The assumptions imply that

  $E(Y \mid X = x) = \beta_0 + \beta_1 x$ and
  $\text{var}(Y \mid X = x) = \sigma^2$

  and hence that if we have repeated observations at different values of x, the scatter about the true line will be Normally distributed with constant variance σ^2

Recap: Checking the Conditions for Regression Inference

- You can fit a least-squares line to any set of explanatory-response data when both variables are quantitative. If the scatterplot does not show a roughly linear pattern, the fitted line may be almost useless.
- Before you can trust the results of inference, you must check the conditions for inference one by one.

> ✓ **The relationship is linear in the population.**
> ✓ **The response varies Normally about the population regression line.**
> ✓ **Observations are independent.**
> ✓ **The standard deviation of the responses is the same for all values of *x*.**

You can check all of the conditions for regression inference by looking at graphs of the residuals or **residual plots.**

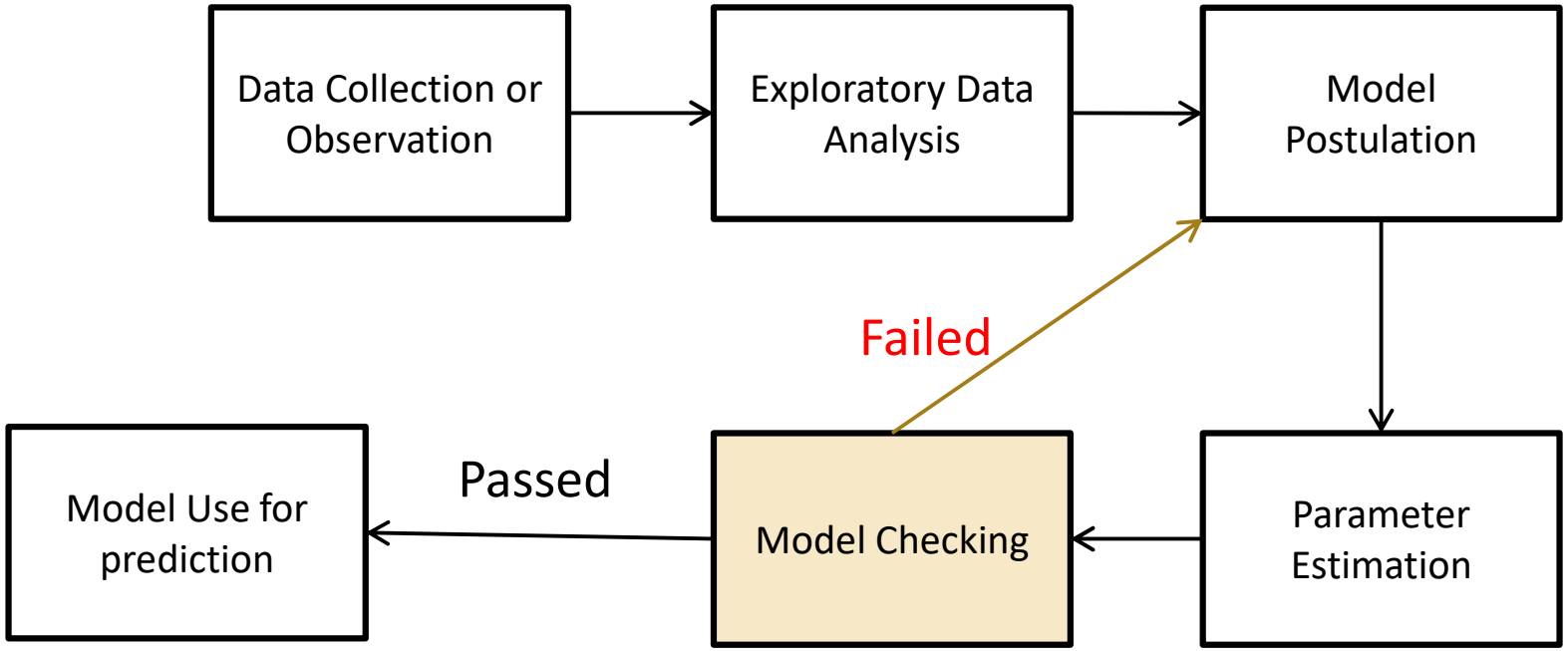
Aim 1 Critically assessing the regression model

- Always draw (generate) a scatterplot. Does a straight line capture the basic pattern of the data?
- Is there a statistically significant linear relationship between x and y? (Test H_0: $\beta_1 = 0$)
- Does the model explain the variation in y sufficiently? (Check the coefficient of determination R^2)
- Check the residuals.

Aim 2 Diagnostic checking – introduction

- The conditions for inference involve the population (true, but unknown) regression line and deviations from this line
- We can't observe this line, but the least-squares regression is our best estimate of this line, and the residuals $\hat{e}_i$ estimate the deviations from this population line
- We can check the conditions for regression inference by looking at graphs of the standardized residuals
  - (For now, we look at simple standardization, $\hat{e}_i/s$)

Role of diagnostics



Checking model validity

Sheather (2009), p. 50 & 51

1. Determine whether the proposed regression model is a valid model (i.e., determine whether it provides an adequate fit to the data). The main tools we will use to validate regression assumptions are plots of standardized residuals.
2. The plots enable us to assess visually whether the assumptions are being violated and point to what should be done to overcome these violations. Determine which (if any) of the data points have x-values that have an unusually large effect on the estimated regression model (such points are called leverage points).
3. Determine which (if any) of the data points are outliers, that is, points which do not follow the pattern set by the bulk of the data, when one takes into account the given model.

Checking model validity

Sheather (2009), p. 50 & 51

4. If leverage points exist, determine whether each is a bad leverage point. If a bad leverage point exists we shall assess its influence on the fitted model.

5. Examine whether the assumption of constant variance of the errors is reasonable. If not, we shall look at how to overcome this problem.

6. If the data are collected over time, examine whether the data are correlated over time.

7. If the sample size is small or prediction intervals are of interest, examine whether the assumption that the errors are normally distributed is reasonable

Aim 2.1 Residual (scatter) Analysis

To check the appropriateness of the Least Squares line (model)

- Calculate predicted value, $\hat{y}$ for each case (x, y) in the data set
- Calculate the residuals ⇒ *residual* = $y - \hat{y}$
- Plot residuals against the x values

Residual Plot

- A scatterplot that helps assess the fit of the least squares regression line.

- Properties:

1. Mean of the residuals equals zero (0).
2. Residuals should have no pattern.
3. Residuals should be Normally distributed with constant variance.

Residual Analysis

- The **linear** model is appropriate IF:

1. **Pattern-less Residuals**: The residuals should be **RANDOMLY distributed around the horizontal line through zero**.

   If the residuals show any **systematic pattern** (e.g. **curvature**), then the model is **not** a good summary of the data.

Residual Analysis (continued)

The **linear** model is appropriate if:

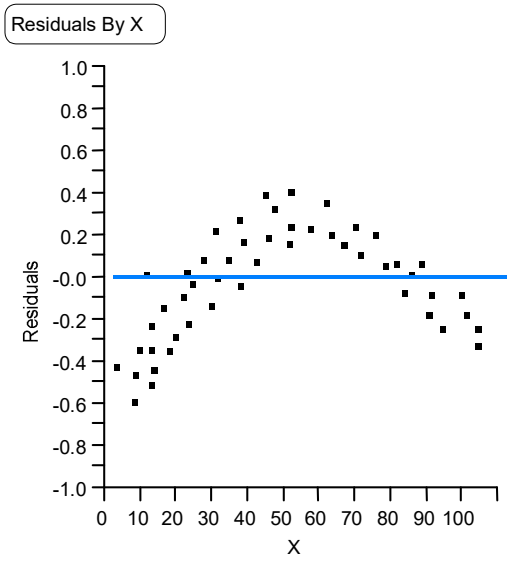
2. **Constant variance**: If the residuals are **more spread out at one end than the other,** then the model is **not** a good summary of the data.

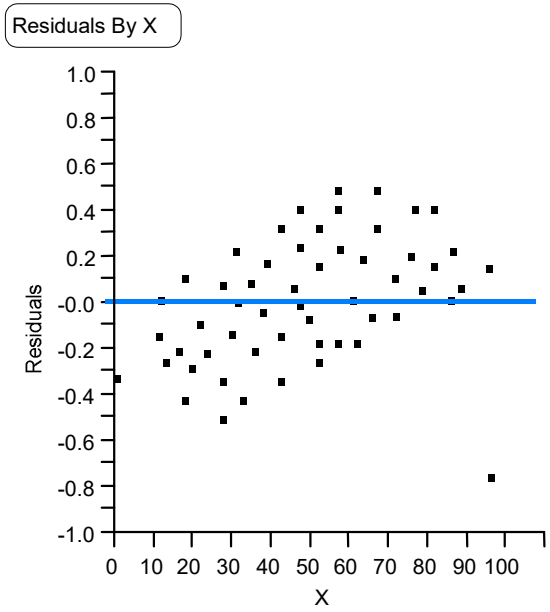
3. **Residuals** follow **Normal distribution** (approx.)

Check 1. The observations are independent

- Use information about how the data were collected to determine whether observations are independent
- Do the residuals appear correlated when plotted against time? For example, do positive residuals tend to clump together?

Residual Plots

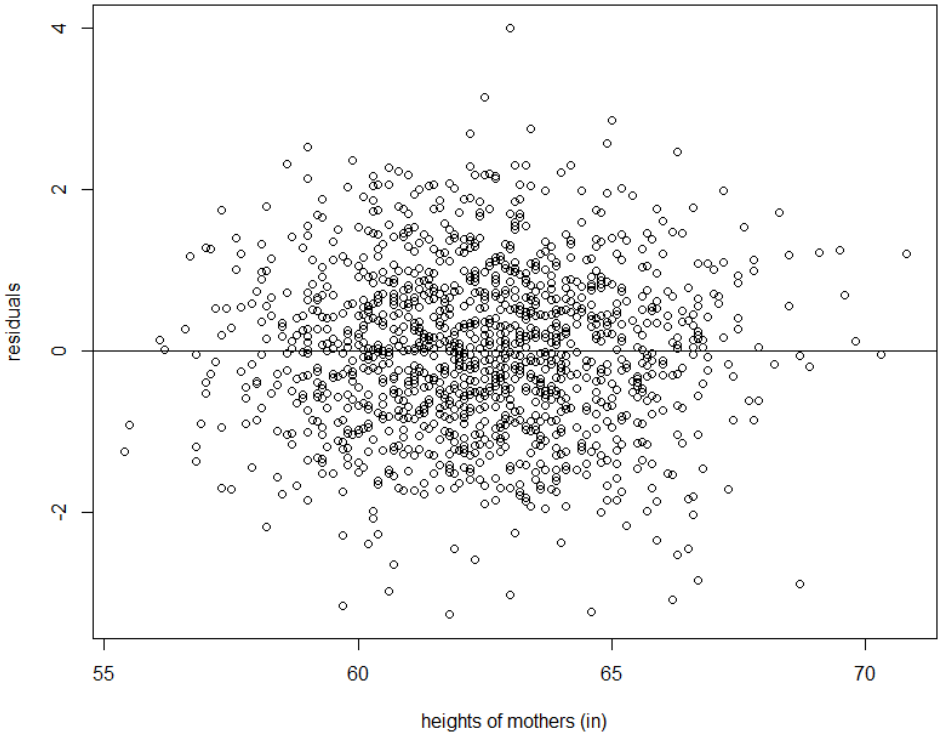




Systematic patterns in the above residual plots indicate that fitting a **straight line** to the data was **not appropriate**.

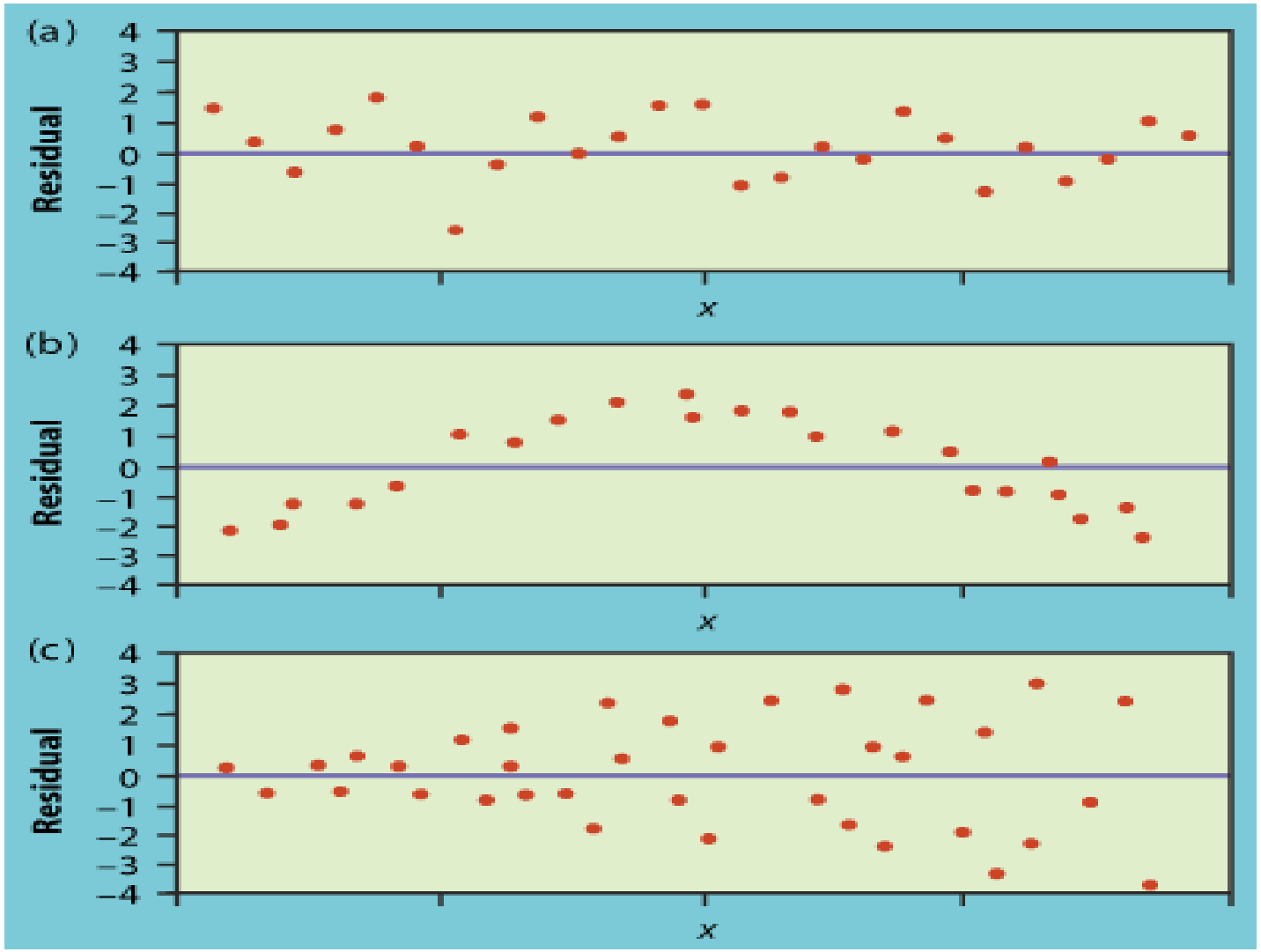
Check 2. The standard deviation of the responses is the same for all values of x

- Look at a scatter plot of the residuals plotted against the explanatory variable x
- The scatter should be roughly the same from one end to the other but be aware of the number of observations at each value of x




```
plot(heights$Mheight, stdres, xlab = "heights of mothers (in)", ylab = "residuals")
abline(h = 0)
```

Residual Plots - patterns



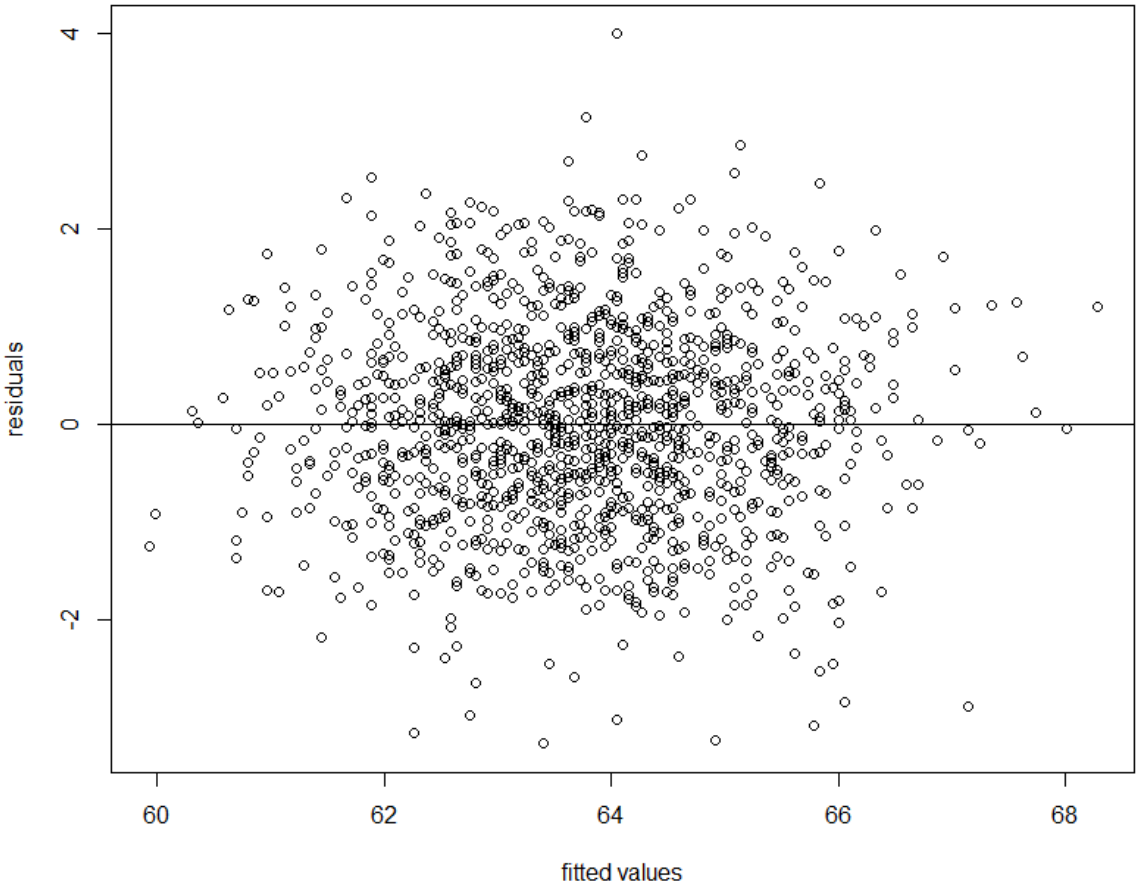
No pattern (random), fairly constant spread (variance)

The residual plot has a curve pattern

The spread (variance) increases as x increases

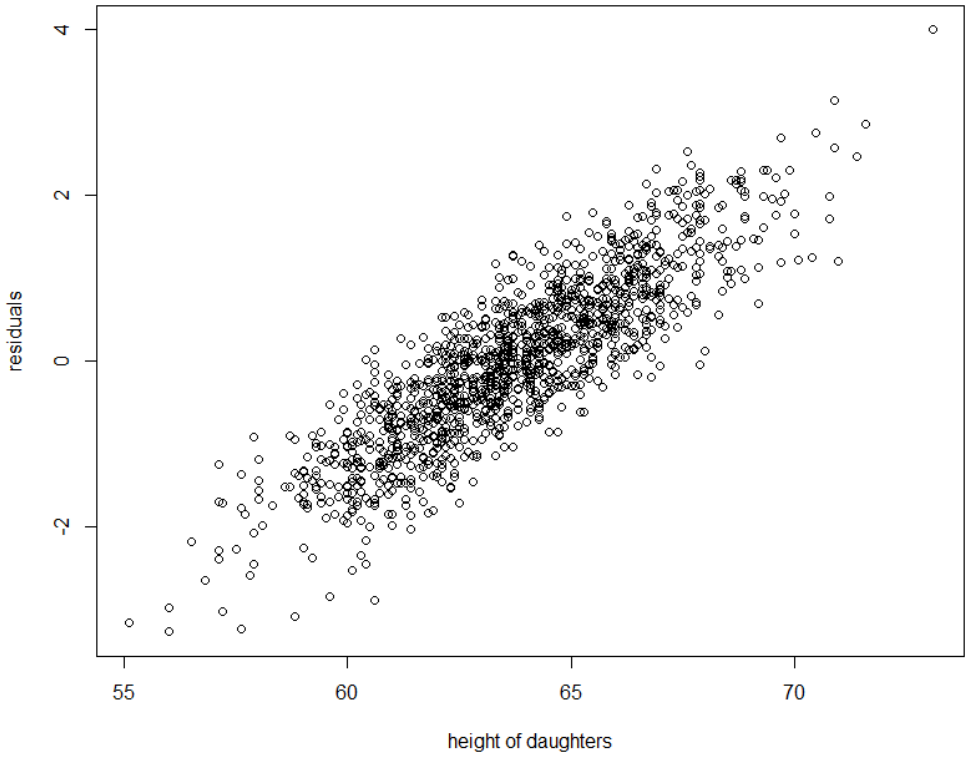
The relationship is linear in the population

- Look at a scatter plot of the residuals plotted against the explanatory variable x or fitted values $\hat{y}$ if there are any curved patterns or other departures
- Can use the scatterplot of y and x, but residual plots magnify any departures
- Same as previous plot for single x

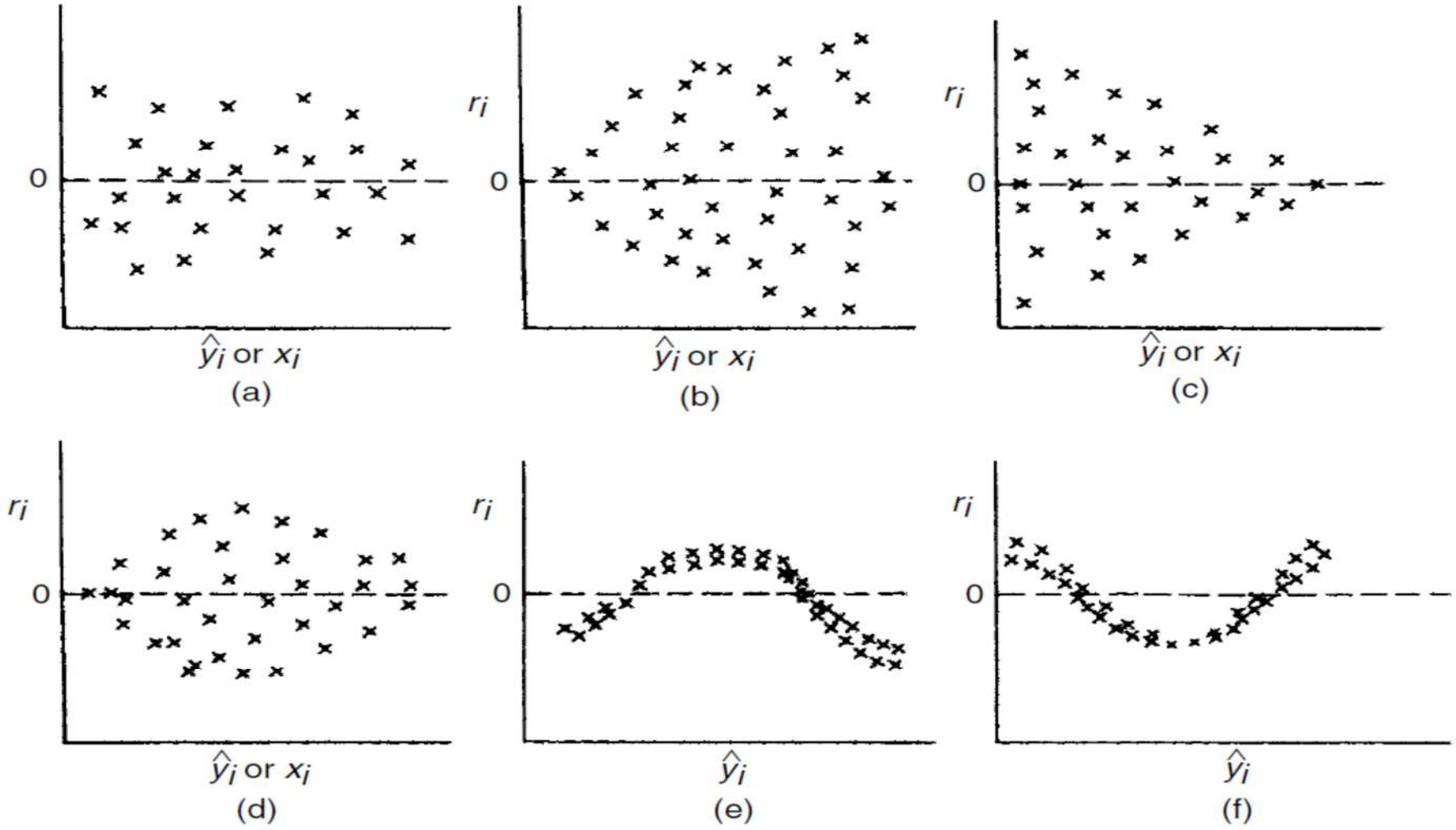


Residual plots – a question

- Why do we plot residuals $\hat{e}_i$ against $\hat{y}_i$ or x_i, but not against y_i?
  - Because the $\hat{e}_i$ and the y_i are correlated, whereas the $\hat{e}_i$ and $\hat{y}_i$ and x_i are not

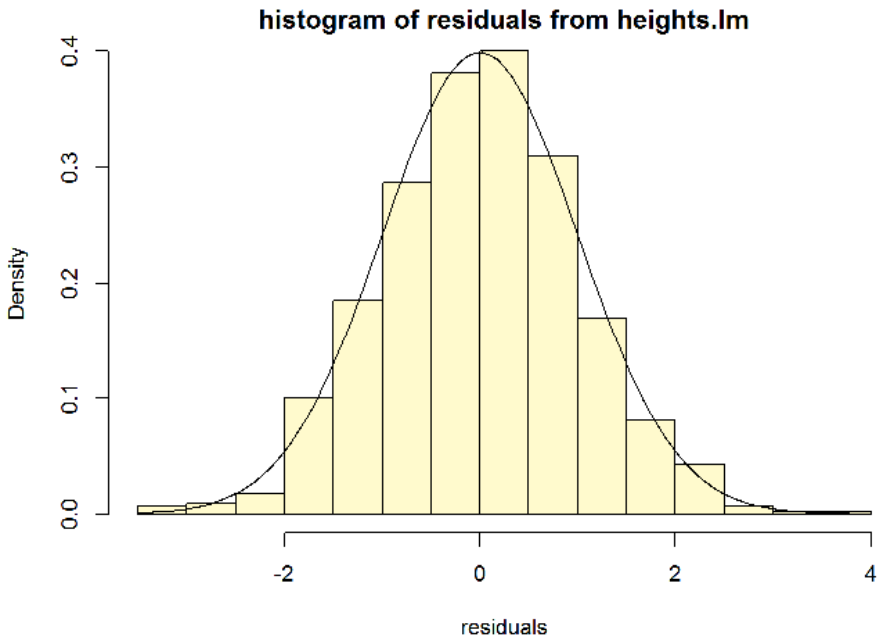


Residual patterns indicating problems



Check 3. Response varies Normally about the population regression line

- The deviations from the population line – estimated by the residuals – must be Normally (or nearly so) distributed
- Check a histogram of the residuals – is it roughly symmetric or are there major departures from Normality?



Visually assessing Normality – *Q-Q* plots

- Motivation: if we have standardized $\hat{e}_i$ as $\hat{e}_i/s$, then $\hat{e}_i/s$ should be distributed (approximately*) as $N(0,1)$
- Given an ordered sample of observations

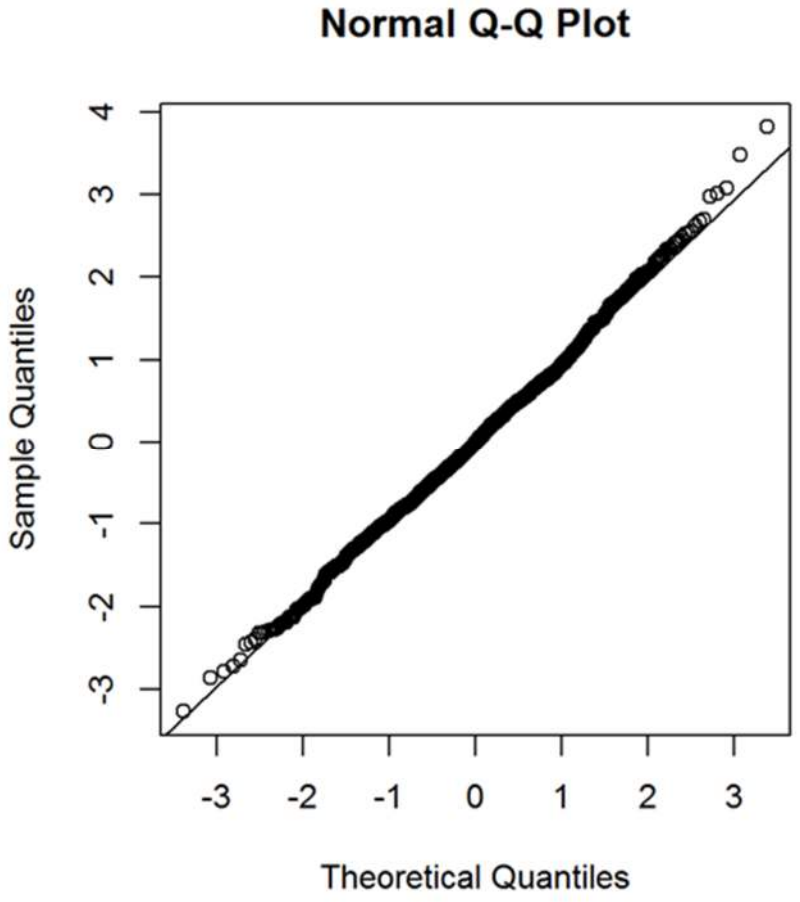
$$z_{(1)} \leq z_{(2)} \leq \cdots \leq z_n$$

which we believe arises from a Normal distribution with distribution function F, a quantile-quantile (*Q-Q*) plot consists of the points

$$F^{-1}\left(\frac{i}{n+1}\right), z_i, \qquad i = 1, 2, \ldots, n$$

(*We'll see a better way of standardizing residuals later)

- We expect that the points lie on a straight line as a reference for normality.



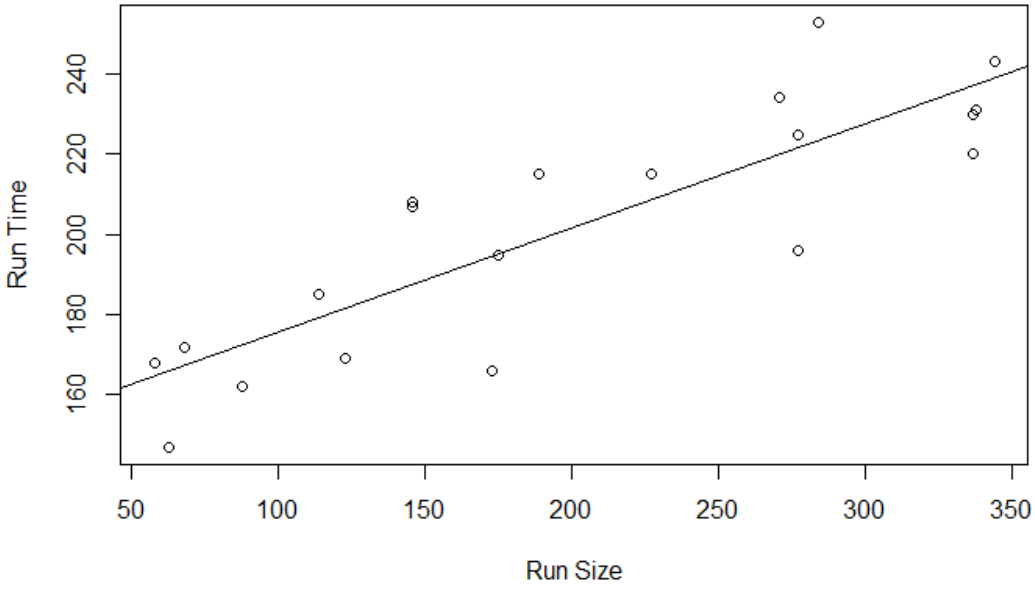
q-q plot – 1375 random Normal deviates

Example 1: *Timing of production runs*

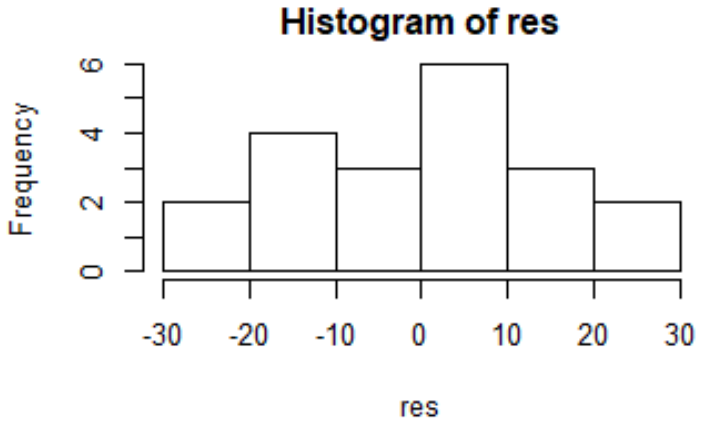
(Shearer (2009))- prod.lm

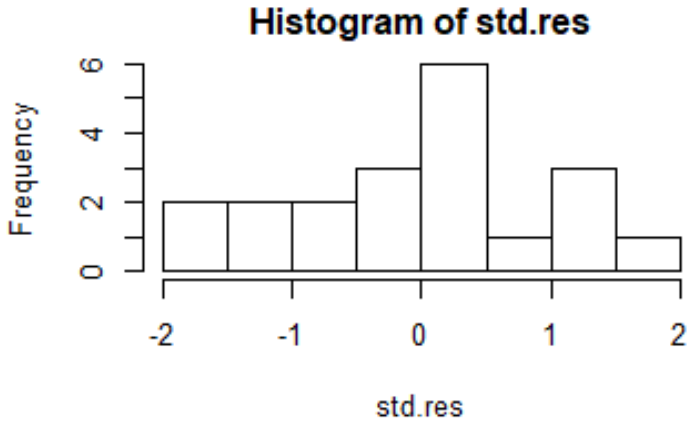
- The data are in the form of the time taken (in minutes) for a production run, Y, and the number of items produced,X, for 20 randomly selected orders as supervised by three managers. At this stage we shall only consider the data for one of the managers.
- The equation of the least squares line of best fit is **y = 149.75 + 0.26x**

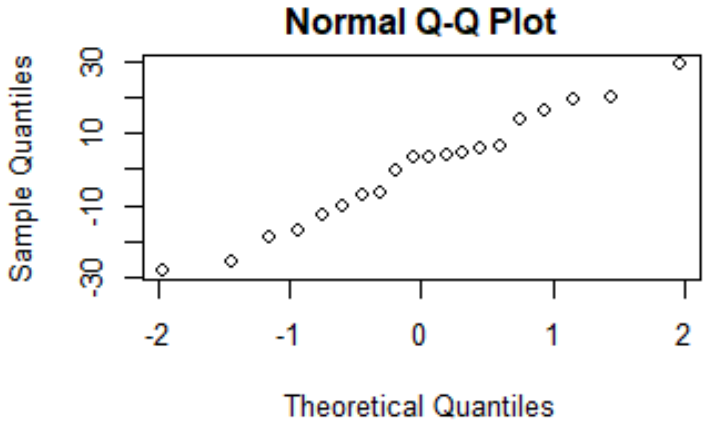
```
lm(formula = RunTime ~ RunSize, data = production)

Residuals:
   Min      1Q  Median      3Q     Max
-28.597 -11.079   3.329   8.302  29.627

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.74770    8.32815   17.98 6.00e-13 ***
RunSize       0.25924    0.03714    6.98 1.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-squared:  0.7302,     Adjusted R-squared:  0.7152
F-statistic: 48.72 on 1 and 18 DF,  p-value: 1.615e-06
```

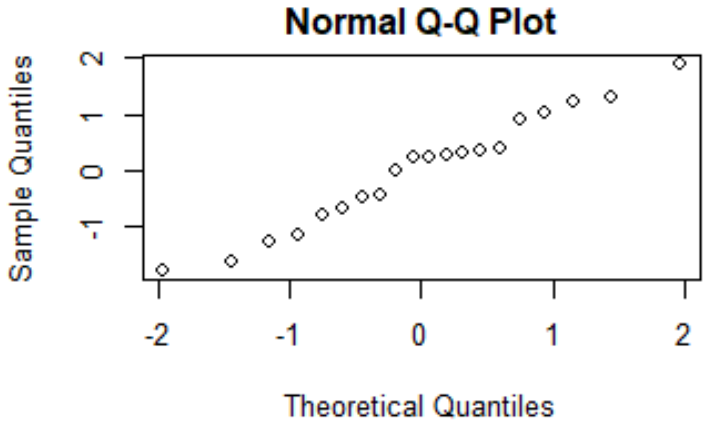


Example 1: Normality Checking



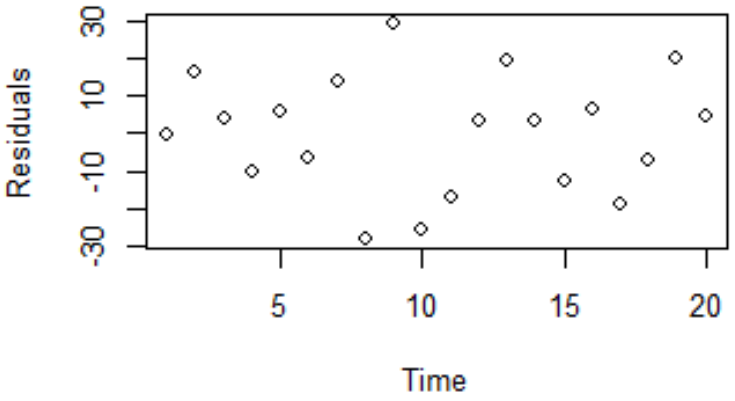


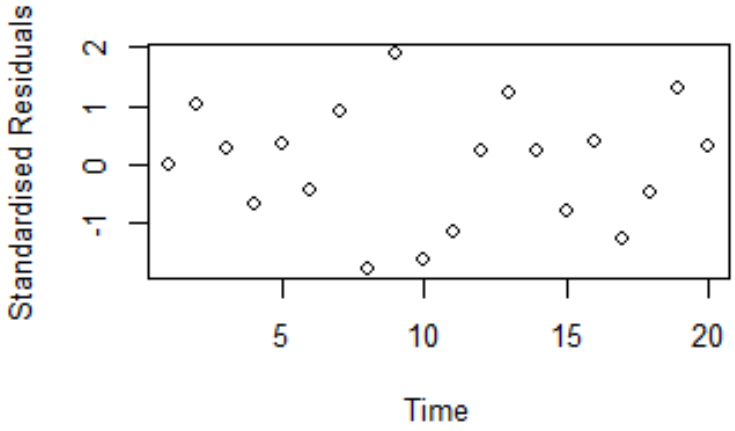


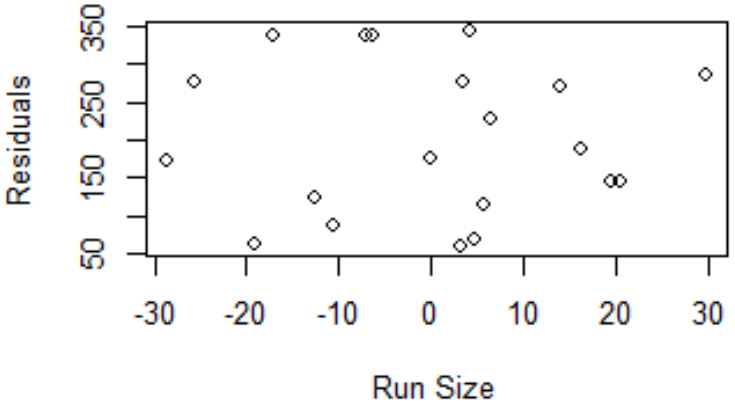


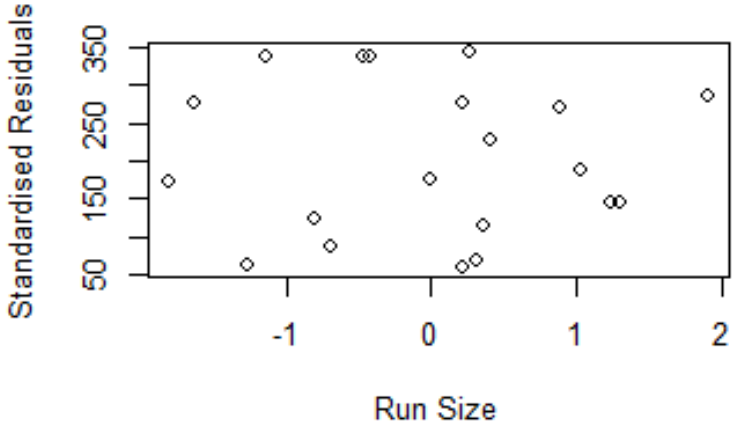
Normality is satisfied

Example 1: Randomness, independence, constant variance





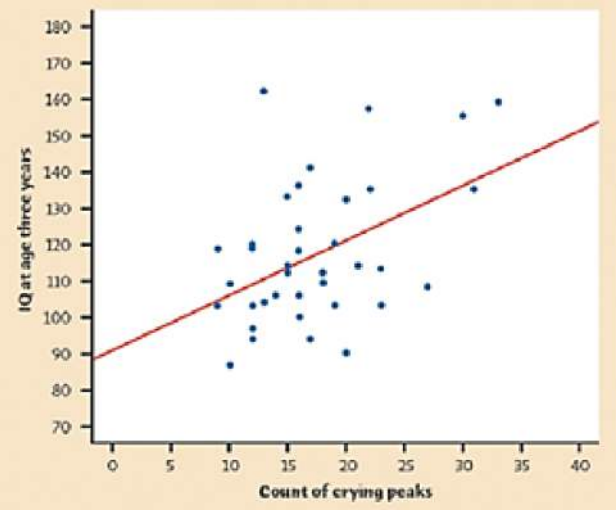




No pattern (random), fairly constant spread (variance)

Example 2 (Moore et al (2017))

- Infants who cry easily may be more easily stimulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants 4 to 10 days old and their later IQ test scores.
- A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured its intensity by the number of peaks in the most active 20 seconds.
- They later measured the children's IQ at age 3 years using the Stanford-Binet IQ test. A scatterplot and Minitab output for the data from a random sample of 38 infants is below.



Regression Analysis: IQ versus Crycount

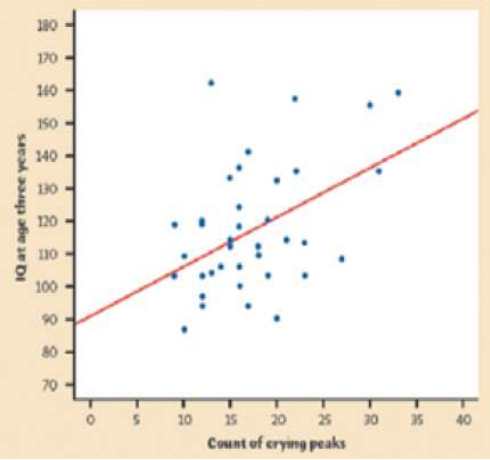
Predictor	Coef	SE Coef	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004

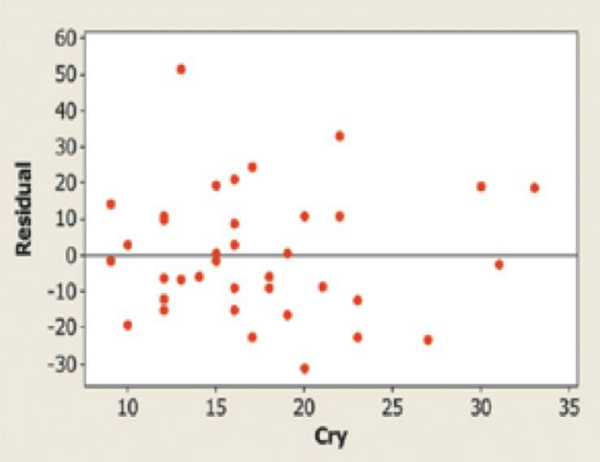
S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%

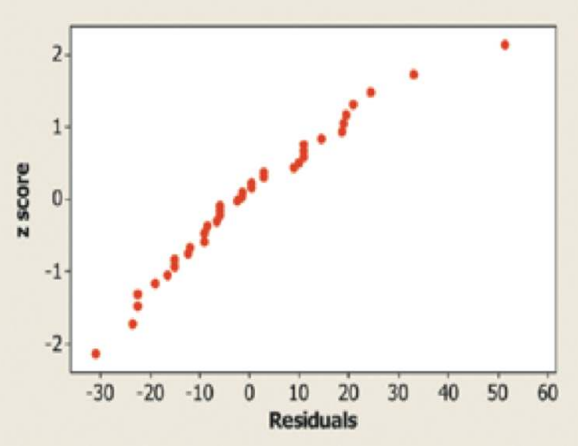
Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants?

Example 2

- The scatterplot suggests a moderately positive linear relationship between crying peaks and IQ. The residual plot shows a random scatter of residuals about the line $y = 0$.







- IQ scores of individual infants should be independent.
- The Normal probability plot of the residuals shows a slight curvature, which suggests that the responses may not be Normally distributed about the line at each x-value. With such a large sample size ($n = 38$), however, the t procedures are robust against departures from Normality.
- The residual plot shows a fairly equal amount of scatter around the horizontal line at 0 for all x-values.

Aim 2.2 Outlying observations

- **Outliers** are observations that are "far away" from the rest of the data
  - Shouldn't automatically eliminate outliers - try to determine why an observation might be an outlier – mistake?
  - Data may have characteristics that cannot be summarized by the postulated linear relationship
- **(Bad) Leverage points** are outliers that have the potential to influence the fitted line

Influence analysis

- Aims to determine observations that have influential effect on the fitted model
- Potentially influential points become candidate for removal from the model
- Criteria used are
  - The hat matrix elements hi (we use this one in SLR)
  - The Studentized deleted residuals t_i*
  - Cook's distance statistic Di (we use this one in SLR)
- All three criteria are complementary
- Only when all three criteria provide consistent result should an observation be removed

The Hat Matrix Element h_i

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

If $h_i > 4/n$, X_i is a leverage point

X_i may be considered a candidate for removal from the model if it is a bad leverage point.

Cook's Distance Statistic D_i

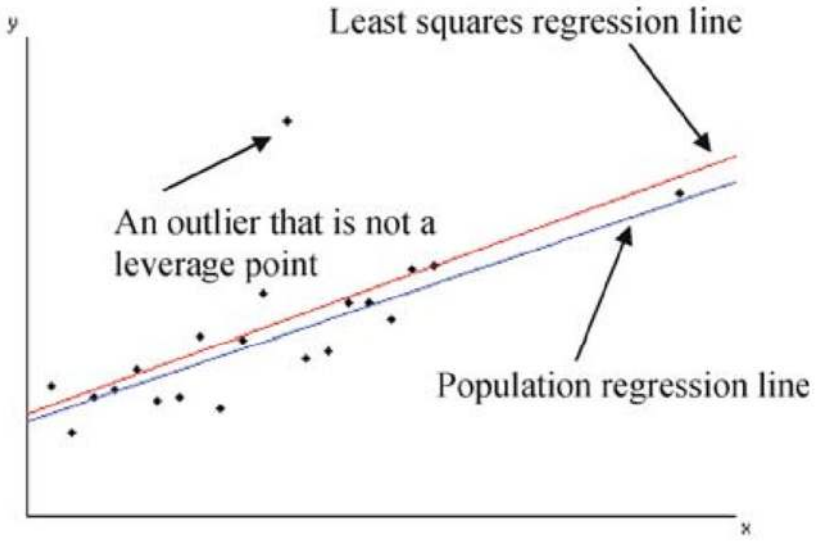
$$D_i = \frac{SR_i^2 h_i}{2(1-h_i)}$$

$$SR_i = \frac{e_i}{S_{YX}\sqrt{1-h_i}}$$

If $D_i > 4/(n-2)$

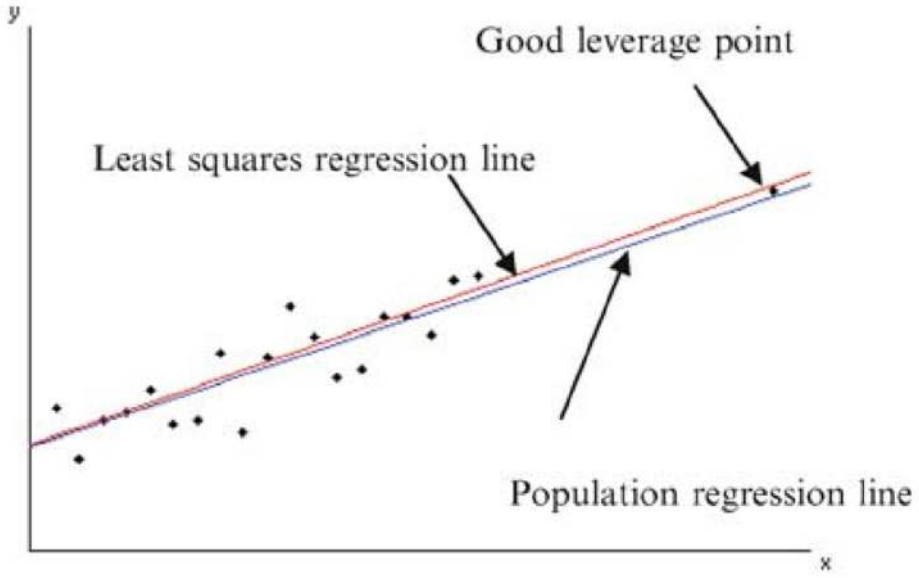
an observation is considered influential

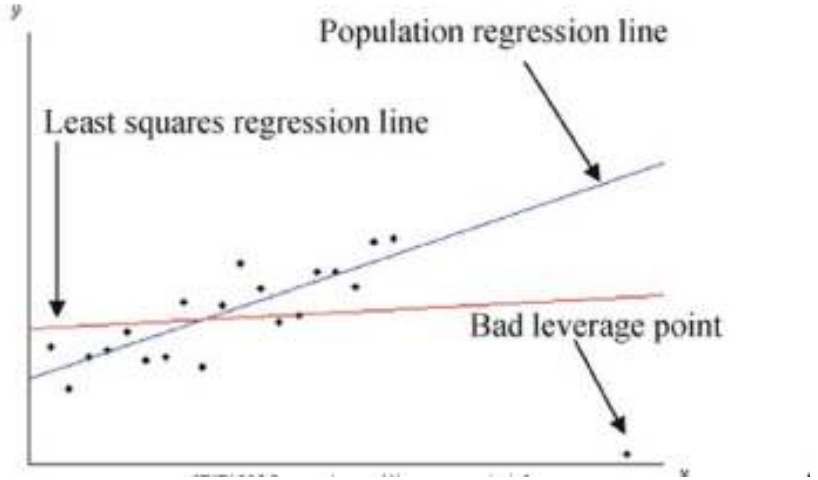
Use the function `influence.measures` to explore measures of leverage and Cook's distance in R.



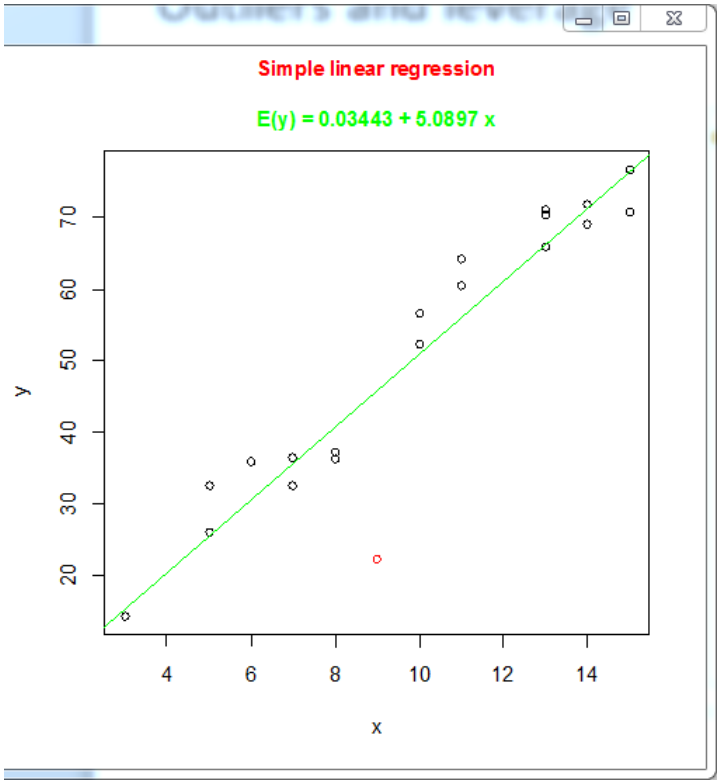
The least squares regression line has been levered down by single point. Hence we call this point a leverage point.

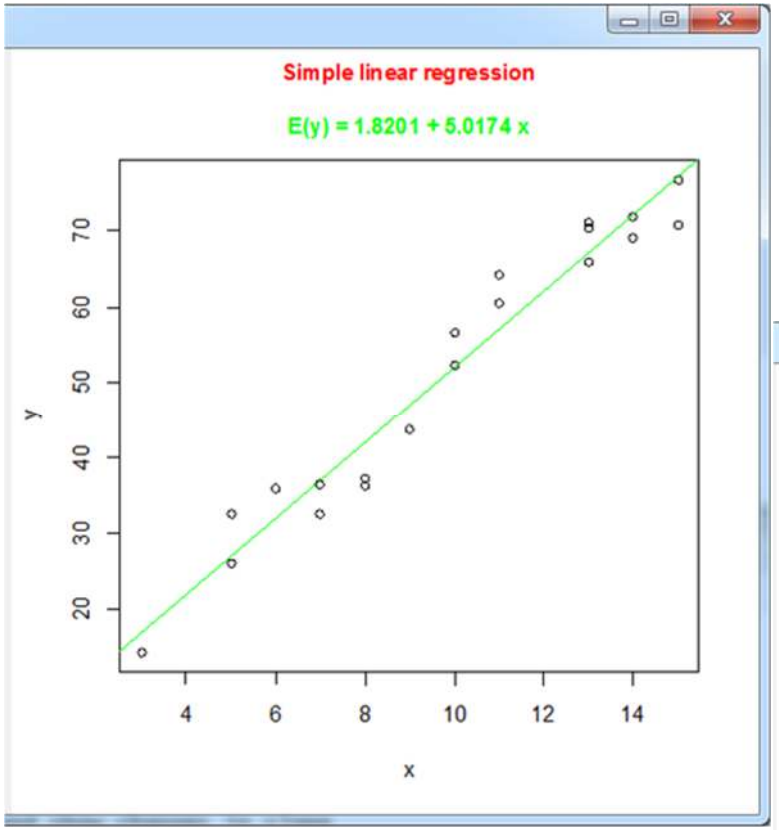
It is a bad leverage point since its Y -value does not follow
the pattern set by the other 19 points.

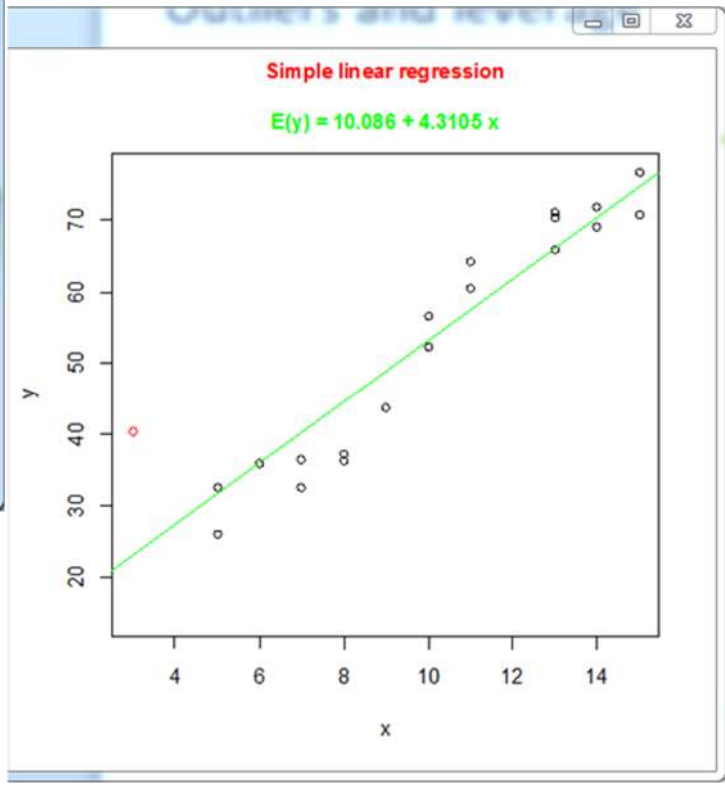




Outliers and leverage



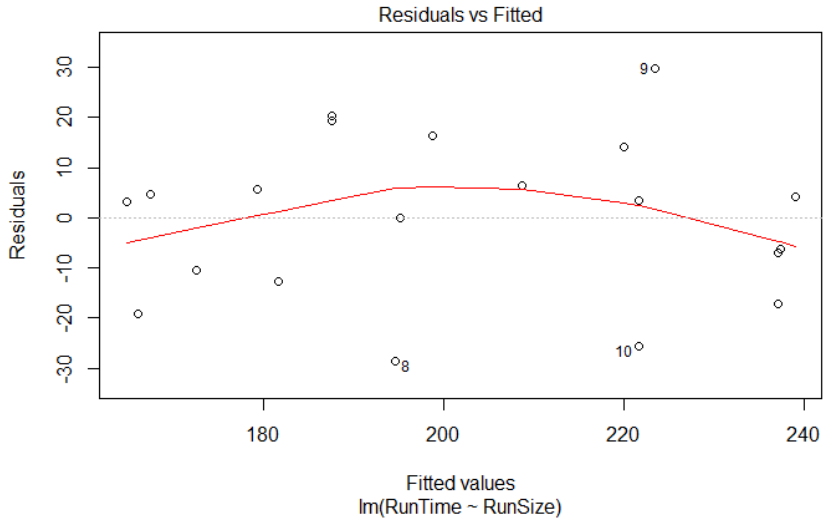


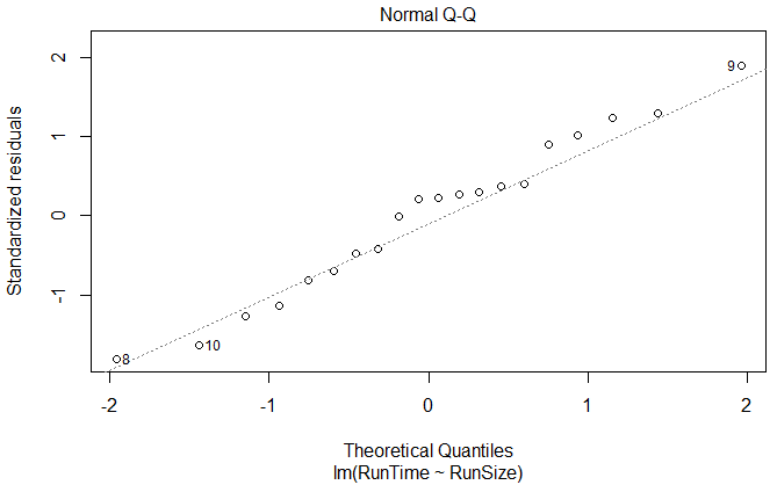


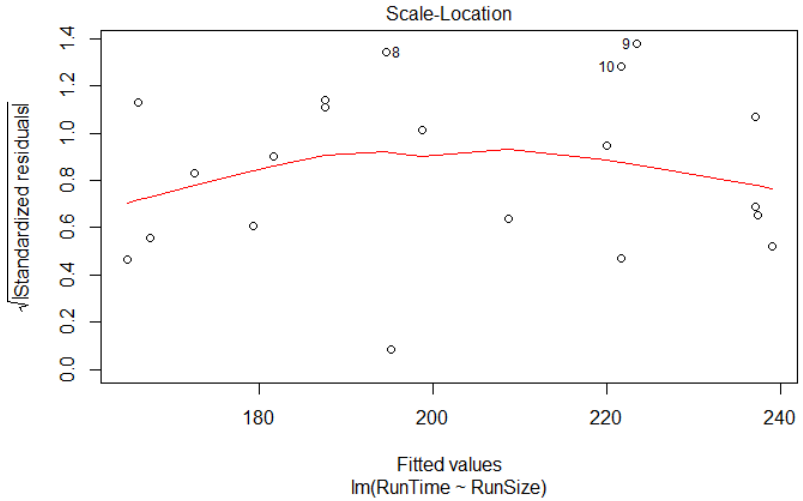
Outliers and leverage

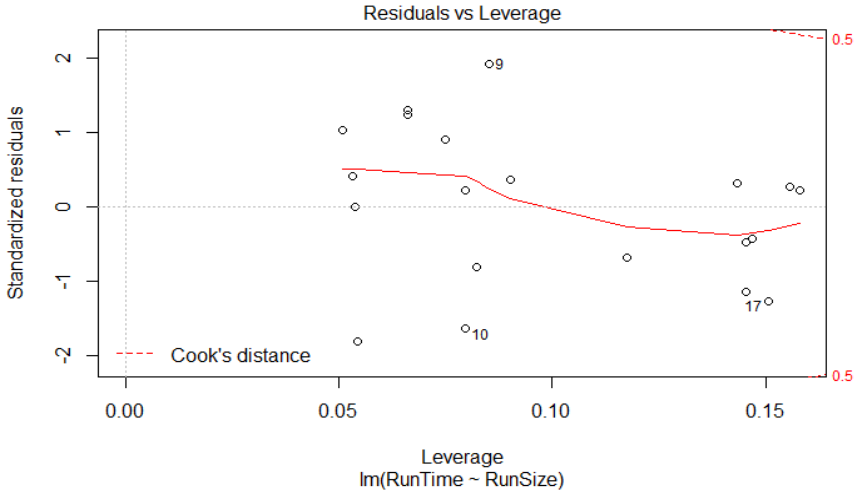
- An **outlier** is a point whose **standardized residual falls outside**
  - the interval from –2 to 2 for small to moderate sample size
  - the interval from –4 to 4 for large sample size
- A **bad leverage point** is a leverage point whose **standardized residual falls outside** the interval from –2 to 2 for small to moderate sample size.
- A **good leverage point** is a leverage point whose standardized residual falls inside the interval from –2 to 2 for small to moderate sample size.

Revisiting Example 1: "plot(prod.lm)" in R









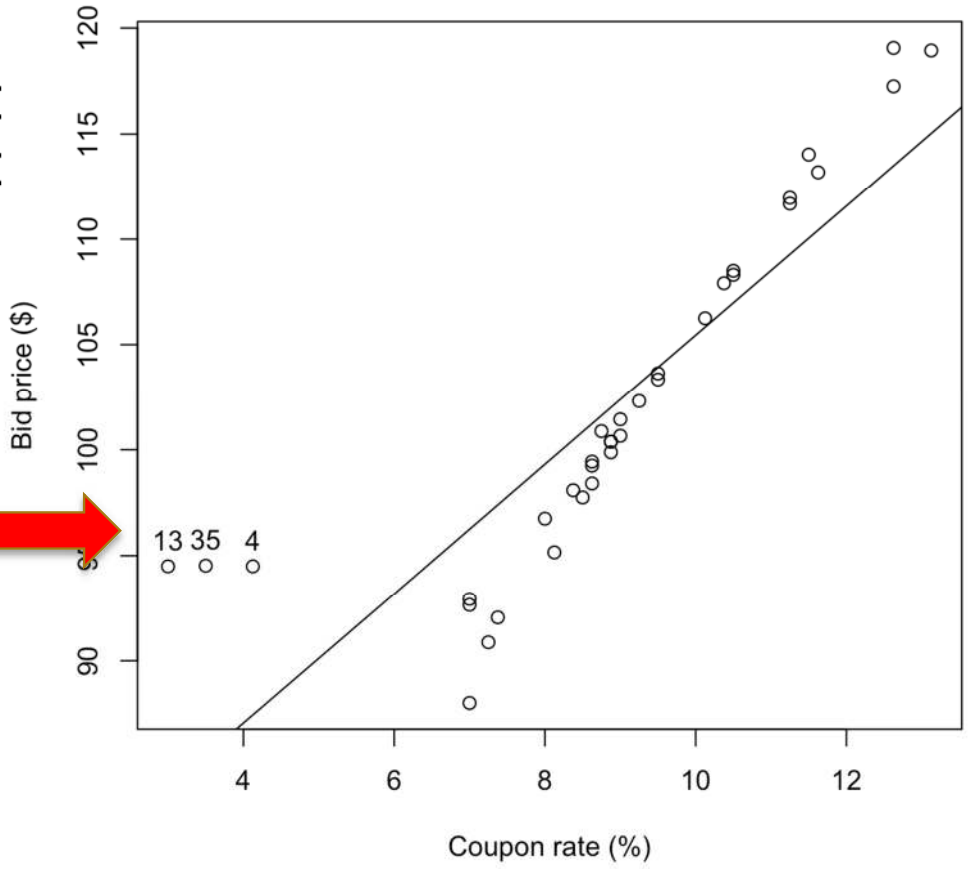
- The smoothing red curves to help identifying patterns
- No pattern (random), fairly constant spread (variance),
- Normality is satisfied.
- No leverage points

Example 3 : US Treasury bond prices

- US Treasury bonds are seen as a safe investment
- Data set consists of coupon rate (size c payment twice a year) and the current selling price
- If coupon rate is 7%, payment is $3.50 every six months until maturity, at which time it pays an additional $100

```
summary(bonds.lm)
```

```
Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

Residuals:
    Min      1Q  Median      3Q     Max
-8.249 -2.470 -0.838   2.550 10.515

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```



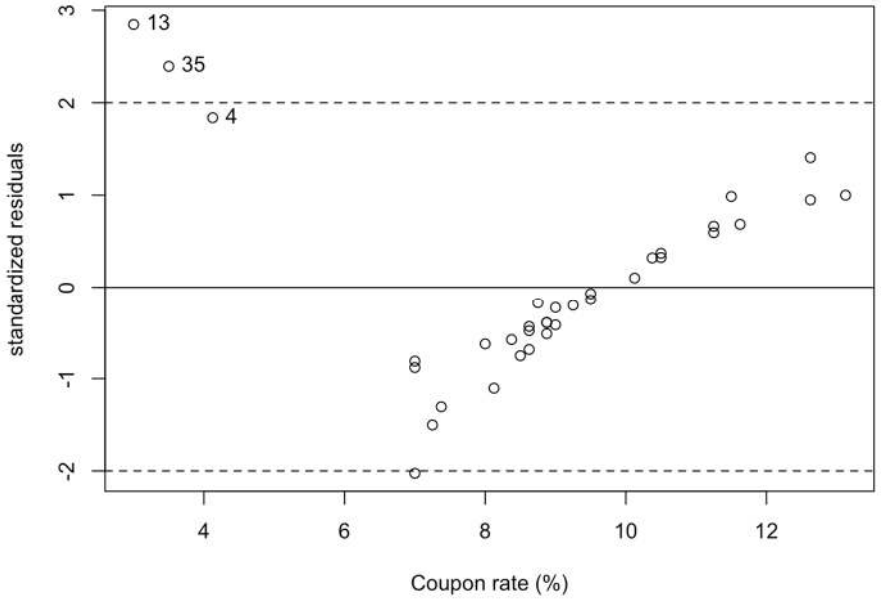
Example 3: US Treasury bond prices - comments

- Clearly the fitted model does not describe the data well
- Three points on the left drag the least squares line away from the bulk of the points
- If the model were appropriate, the interpretation of the slope would be that for every unit increase in the coupon rate, the **mean** bid price increases by about $3.07.
- A 95% confidence for the slope is given by (2.44, 3.69)

Example 3: US Treasury bond -residuals

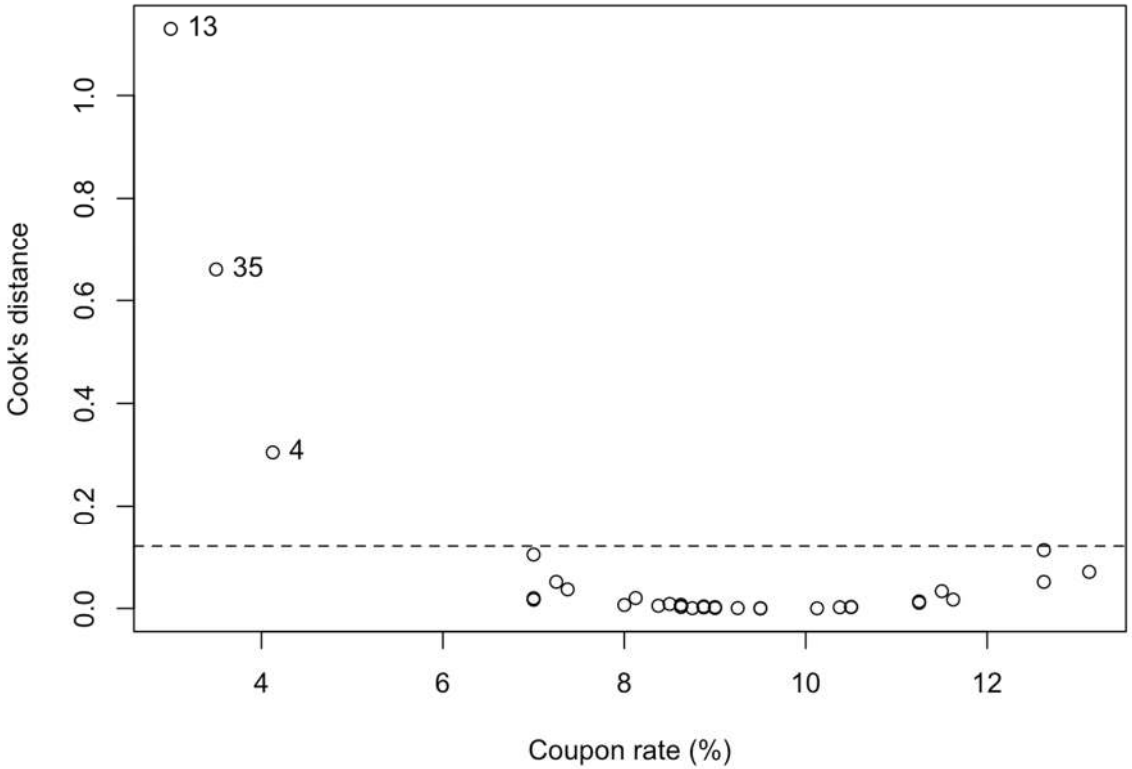
- Even though the outlying observations were clear in the plot of the data, they are even more easily identifiable in a plot of the standardized residuals
- Clearly not well fitted by the model and stand out from the other points, which appear to follow a linear pattern
- Intuitively, we can see that points which have large residuals (outliers that are poorly fitted) and that have high leverage (away from the bulk of the points) are likely to be influential
- These points are for bonds that have different tax advantages compared to the other bonds, and that bid price is higher than what would be expected for other bonds having the same coupon rate; so, remove, and re-fit model (or add dummy variables)

```
stdres <- rstandard(bonds.lm)
plot(stdres ~ CouponRate, data = bonds,
xlab = "Coupon rate (%)", ylab = "standardized
residuals")
abline(h = c(-2, 2), lty = 2)
abline(h = 0)
```



Example 3: US Treasury bond prices - Cook's distance

- Cook's distance is one measure of influence: for each point, it combines the size of its residual along with a measure of the leverage
- An approximate cutoff is $4(n-2)^{-1}$, but in practice it is important to look for gaps in values of Cook's distance instead of just whether or not the values exceed the cutoff

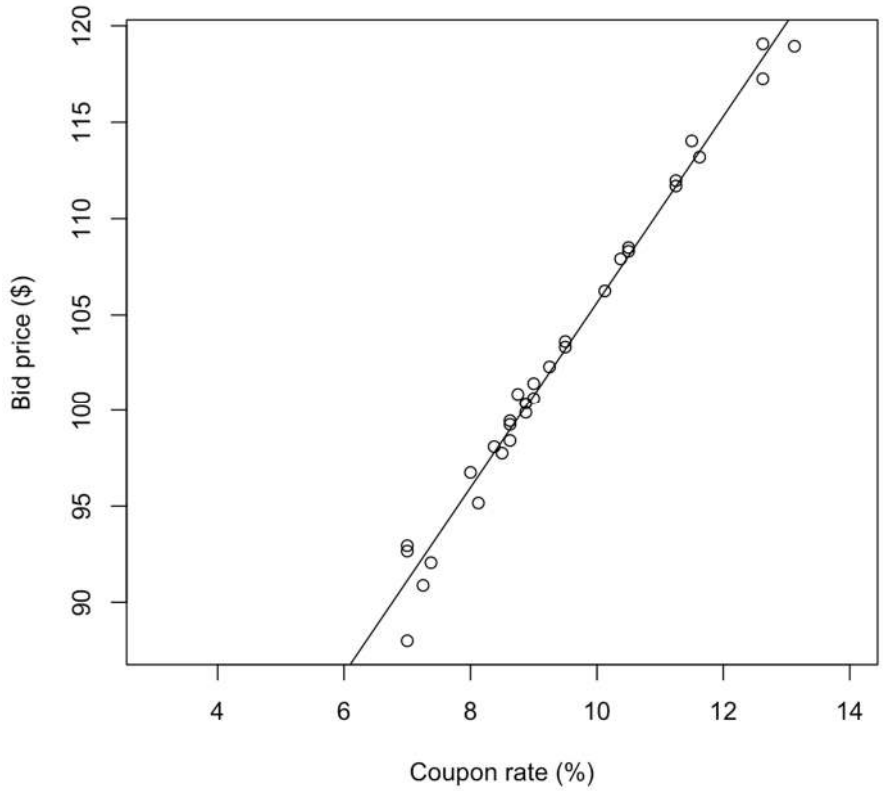



```
CooksD <- cooks.distance(bonds.lm)

plot(CooksD ~ CouponRate, data = bonds, xlab = "Coupon rate (%)", ylab = "Cook's
distance")
abline(h = 4/(nrow(bonds) - 2), lty = 2)
```

Example 3: US Treasury bond prices - after removing outliers

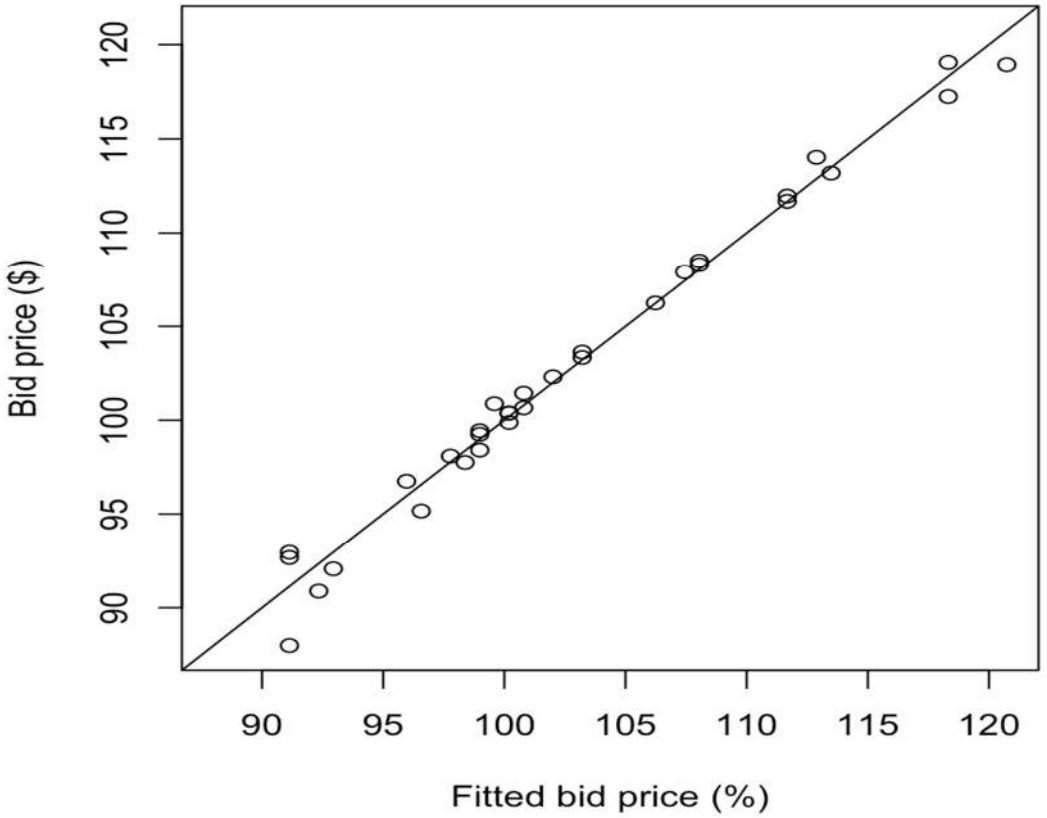
```
summary(bonds.lm2)
```

```
Call:
lm(formula = BidPrice ~ CouponRate, data = bonds, subset = -c(4,
    13, 35))

Residuals:
    Min      1Q  Median      3Q     Max
-3.1301 -0.3789  0.2240  0.4576  1.8099

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.2932     1.0358   55.31   <2e-16 ***
CouponRate    4.8338     0.1082   44.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.024 on 30 degrees of freedom
Multiple R-squared:  0.9852,    Adjusted R-squared:  0.9847
F-statistic:  1996 on 1 and 30 DF,  p-value: < 2.2e-16
```



Example 3: US Treasury bond prices- after removing outliers

- After removing influential points, residual standard error has decreased and R^2 has increased.
- A 95% confidence interval for the slope is (4.61, 5.05), and now for every percent increase in coupon rate, **mean** bid price increases by $4.83.
- Originally, 4.83 is not in the 95% CI when all data points are used
- Also useful to plot the fitted values against the actual values



Handling outliers and leverage points

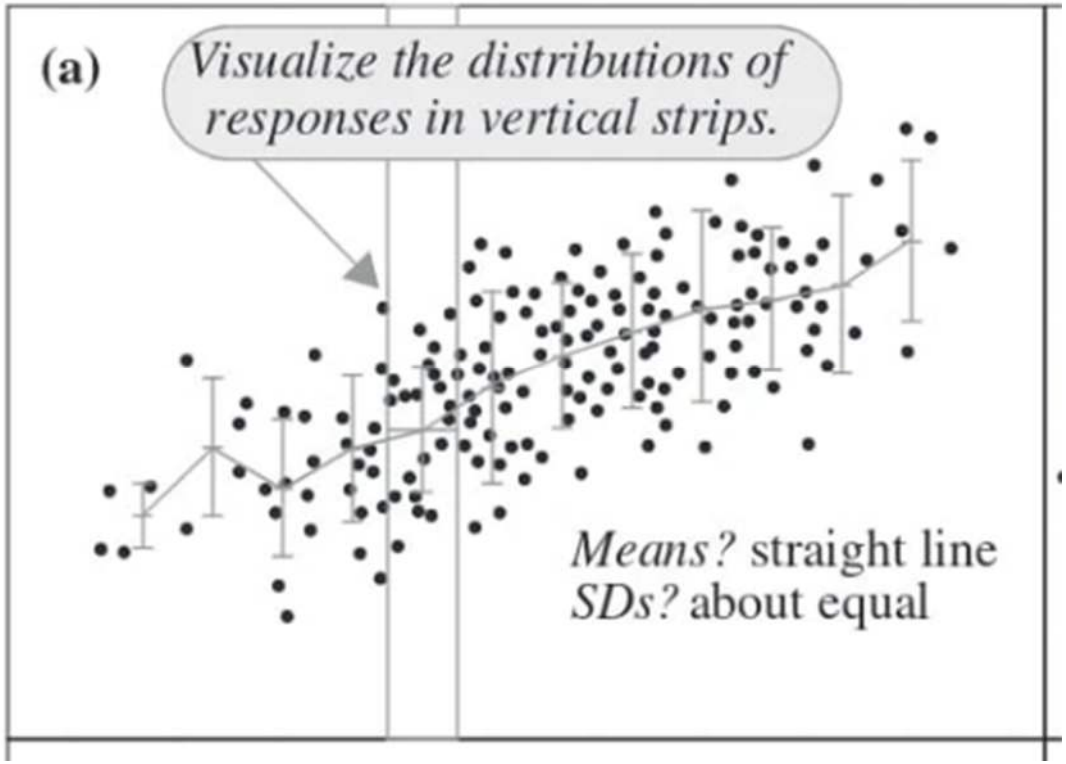
Sheather (2009), p. 67 & 68

- Points should not be routinely deleted from an analysis just because they do not fit the model. Outliers and bad leverage points are signals, flagging potential problems with the model.
- Outliers often point out an important feature of the problem not considered before. They may point to an alternative model in which the points are not an outlier. In this case it is then worth considering fitting an alternative model.
- Including one or more dummy variables in the regression model is one way of coping with outliers that point to an important feature.

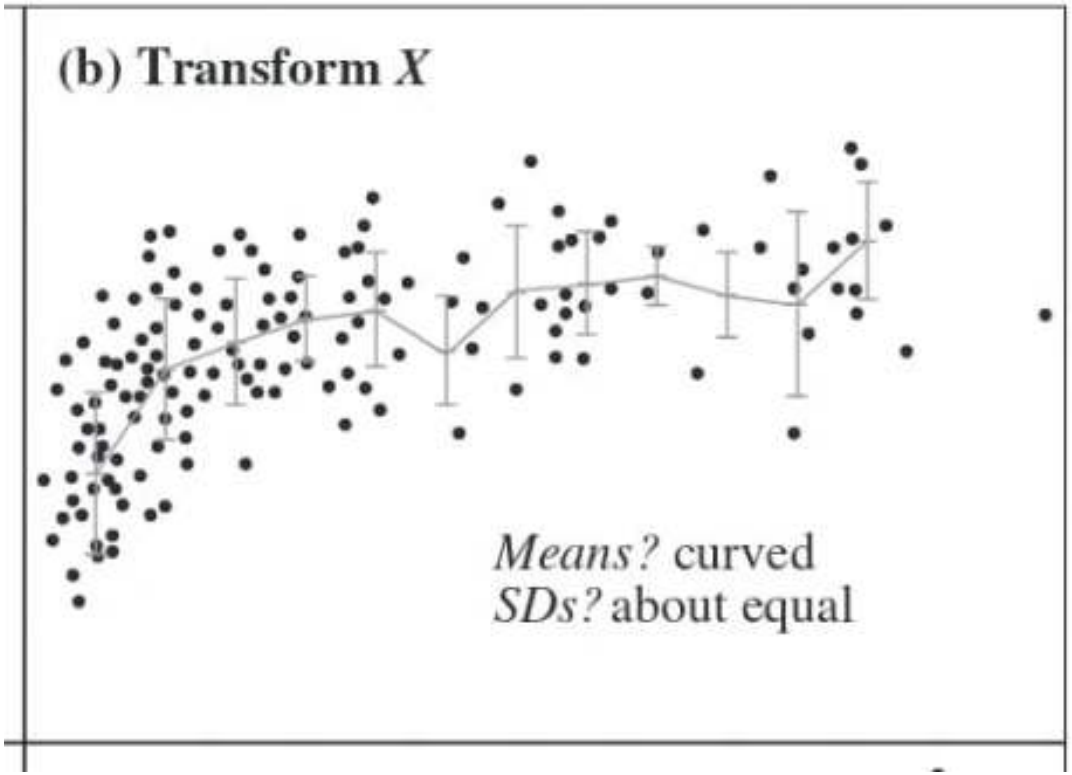
Aim 3 Transformation

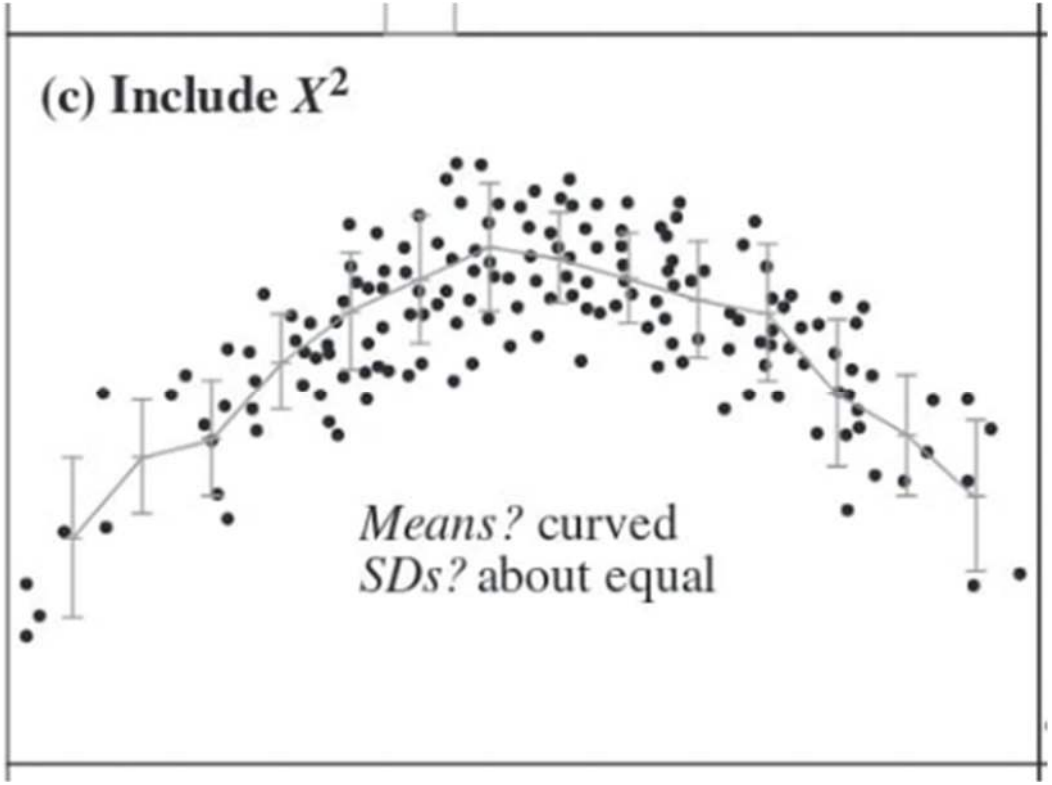
- Transformations can be used to
  - Overcome problems due to nonconstant variance
  - Estimate percentage effects
  - Overcome problems due to nonlinearity

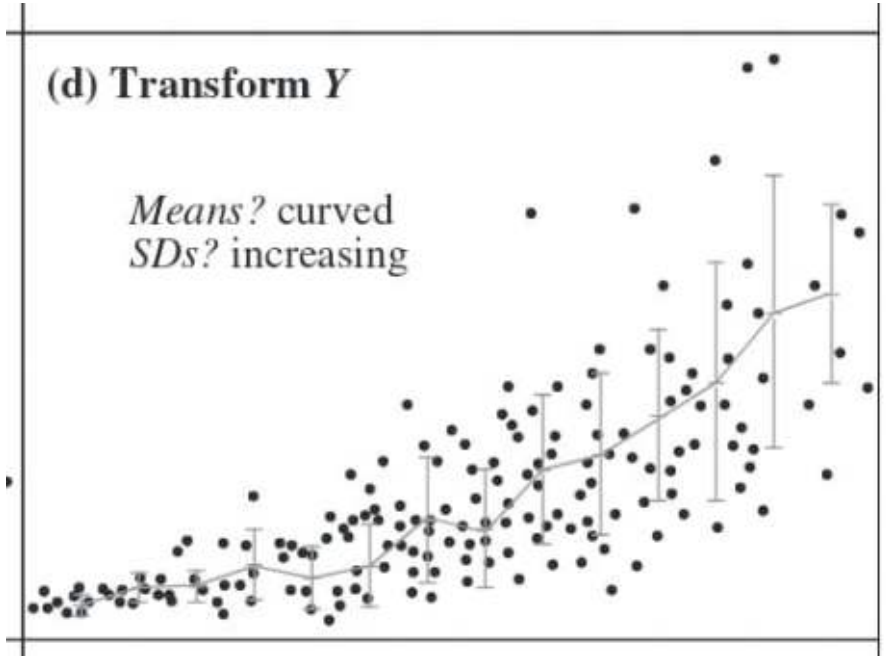
The "ideal" Plot

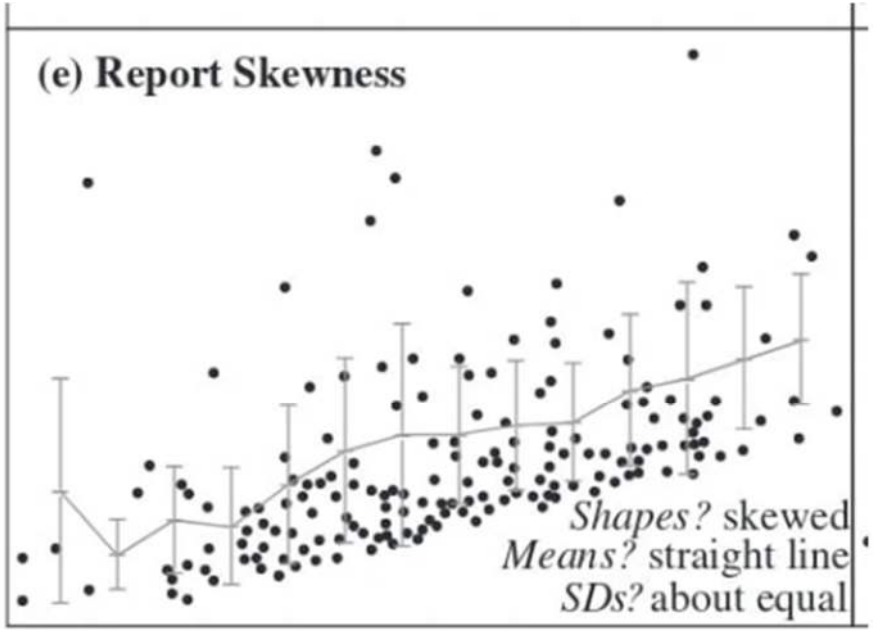


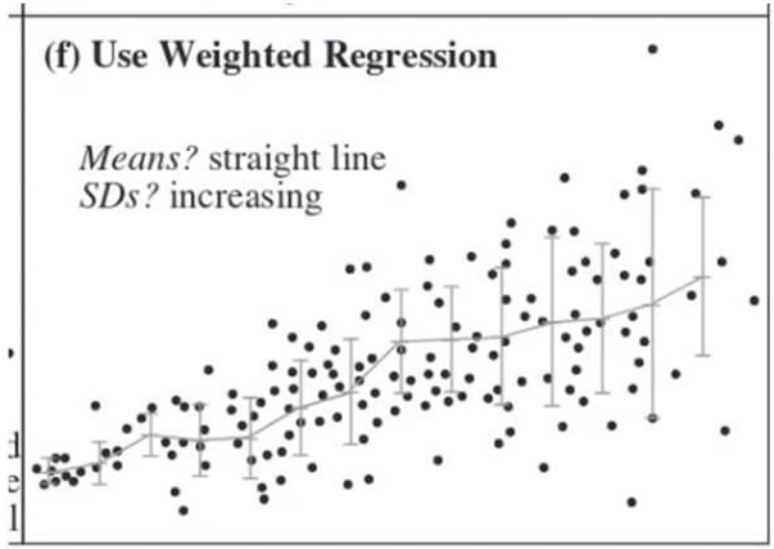
From "The Statistical Sleuth"


(b) Transform X
Means? curved
SDs? about equal


(c) Include X^2
Means? curved
SDs? about equal







(e) The regression is a straight line, the variability is roughly constant, but the distribution of Y about the regression line is skewed. Remedies are unnecessary, and transformations will create other problems. Use simple linear regression, but report the skewness.

(f) The regression is a straight line but the variability increases as the mean of Y increases. Simple linear regression gives unbiased estimates of the straight line relationship, but better estimates are available using *weighted regression*, as

Example 4 Developing a bid on contract cleaning (Sheather,2009)

- A building maintenance company is planning to submit a bid on a contract to clean Corporate offices scattered throughout an office complex. The costs incurred by the maintenance company are proportional to the number of cleaning crews needed for this task.
- Recent data are available for the number of rooms that were cleaned by varying numbers of crews. For a sample of 53 days, records were kept of the number of crews used and the number of rooms that were cleaned by those crews.
- The first model fit to the data was a SLR:

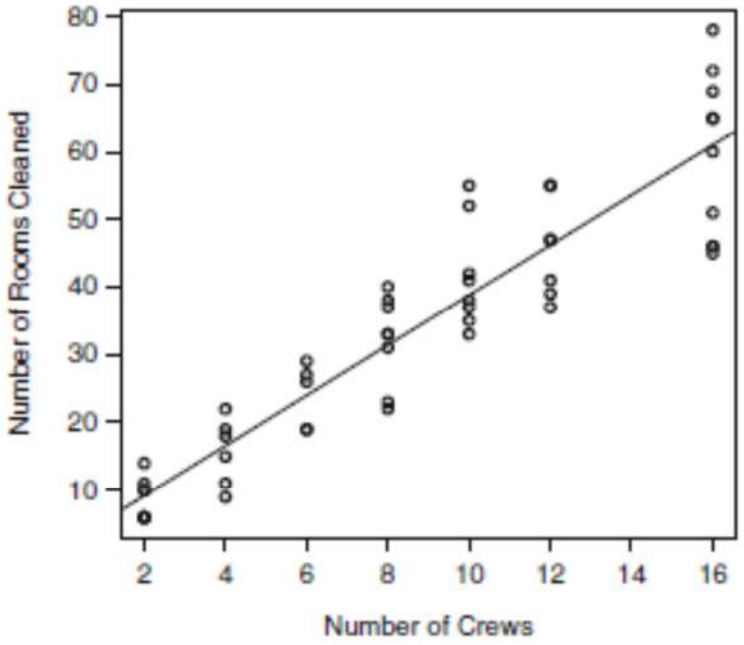
Rooms= 1.7847 + 3.7009 Crews

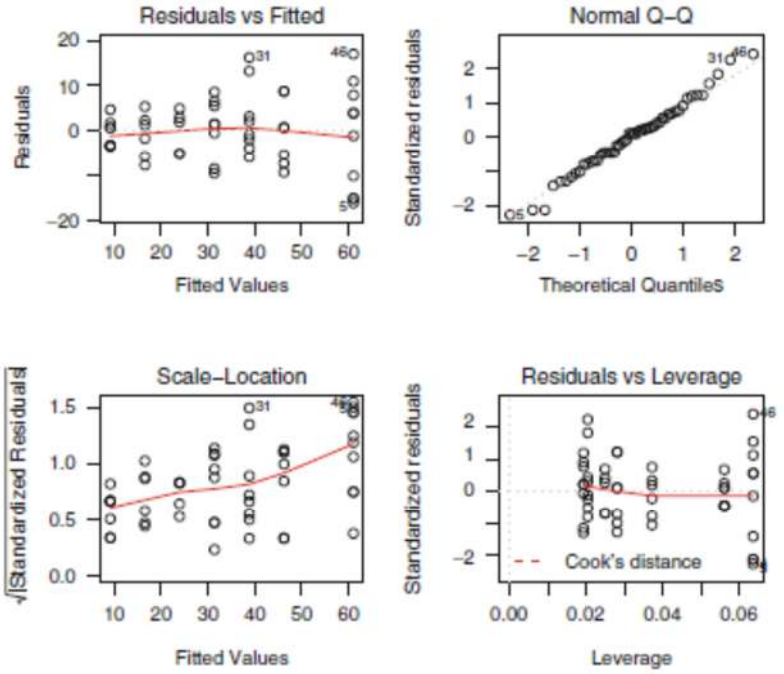
Regression output from R

```
Call:
lm(formula = Rooms ~ Crews)
Coefficients:
             Estimate   Std. Error   t value  Pr(>|t|)
(Intercept)    1.7847       2.0965     0.851     0.399
Crews          3.7009       0.2118    17.472    <2e-16 ***
---
Residual standard error: 7.336 on 51 degrees of freedom
Multiple R-Squared: 0.8569, Adjusted R-squared: 0.854
F-statistic: 305.3 on 1 and 51 DF, p-value: < 2.2e-16
```

Example 4 Developing a bid on contract cleaning (Sheather,2009)

- The first model fit to the data was a SLR: **Rooms= 1.7847 + 3.7009 Crews**



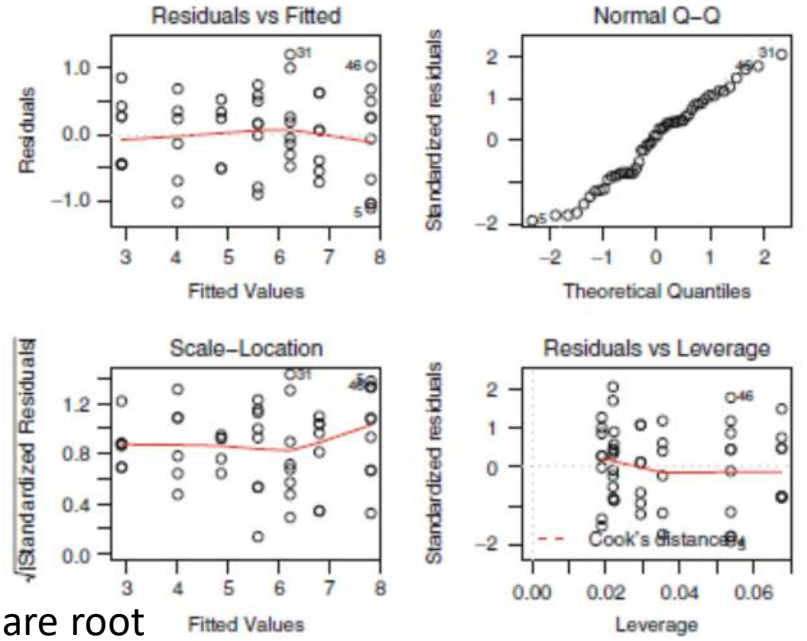


- The bottom left-hand plot is a plot of the square root of the absolute value of the standardized residuals against fitted values.
- There is clear evidence of an increasing trend, which implies that there is evidence that the variance of the errors increases with x .

Example 4 Developing a bid on contract cleaning (Sheather,2009)

- In this case, since the data on each axis are in the form of counts , we shall try the square root transformation of both the predictor variable and the response variable.
- When both Y and X are measured in the same units then it is often natural to consider the same transformation for both X and Y.
- *Y* = the **square root** of the number of rooms cleaned
- *x* = the **square root** of the number of cleaning crews

```
Call:
lm(formula = sqrtrooms ~ sqrtcrews)
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)    0.2001       0.2758     0.726       0.471
sqrtcrews      1.9016       0.0936    20.316      <2e-16 ***
---
Residual standard error: 0.594 on 51 degrees of freedom
Multiple R-Squared: 0.89,  Adjusted R-squared: 0.8879
F-statistic: 412.7 on 1 and 51 DF,  p-value: < 2.2e-16
```



The bottom left-hand plot further demonstrates the benefit of the square root transformation in terms of stabilizing the error term. Thus, taking the square root of both the x and the y variables has stabilized the variance of the random errors and hence produced a valid model.