Linear Regression

Aims

- Understand linear regression with one continuous predictor
- Understand how we assess the fit of a regression model
  - Total sum of squares
  - Model sum of squares
  - Residual sum of squares
  - F
  - R^2
- Know how to do regression using **R**
- Interpret a regression model

What is Regression?

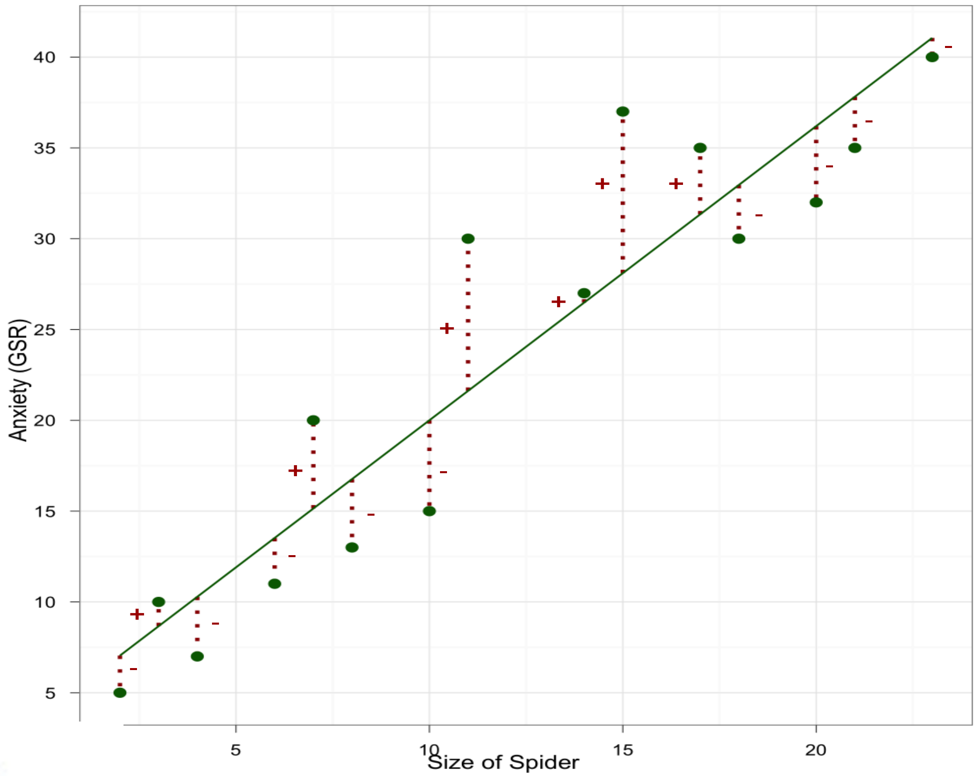
- A way of predicting the value of one variable from another.
  - It is a hypothetical model of the relationship between two variables.
  - The model used is a linear one.
  - Therefore, we describe the relationship using the equation of a straight line.

Describing a Straight Line

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

- b_1
  - Regression coefficient for the predictor
  - Gradient (slope) of the regression line
  - Direction/strength of relationship
- b_0
  - Intercept (value of *Y* when *X* = 0)
  - Point at which the regression line crosses the *Y*-axis (ordinate)

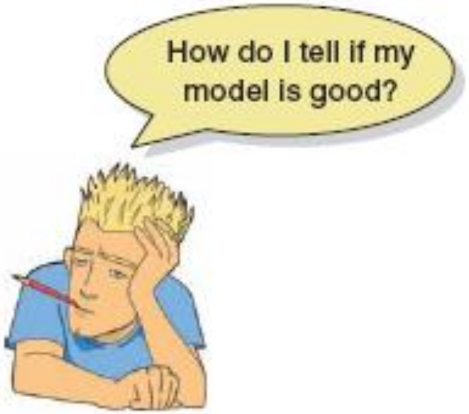
The Method of Least Squares



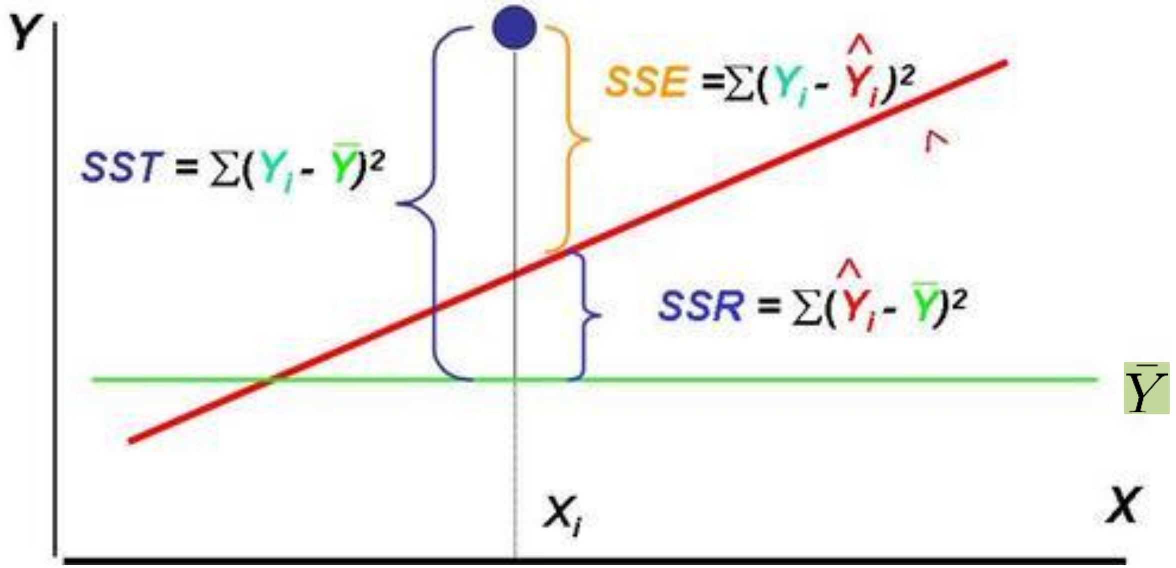
This graph shows a scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

How Good Is the Model?

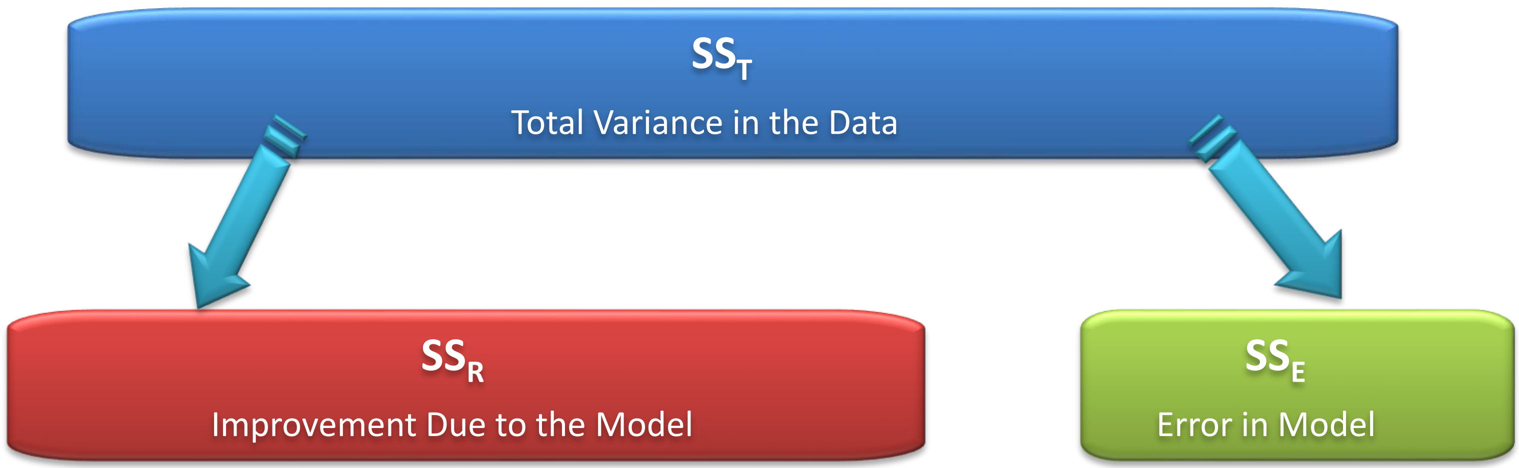
- The regression line is only a model based on the data.
- This model might not reflect reality.
  - We need some way of testing how well the model fits the observed data.
  - How?



Sums of Squares



Testing the Model



- If the model results in better prediction than using the mean, then we expect SS_R to be much greater than SS_E

Testing the Model: R^2

- R^2
  - The proportion of variance accounted for by the regression model.
  - The Pearson Correlation Coefficient Squared

$$R^2 = SS_R / SS_T$$

Testing the Model

- Mean squared error
  - Sums of squares are total values.
  - They can be expressed as averages.
  - These are called mean squares, MS.

$F = MS_R / MS_E$
with MSR=SSR/1
MSE=SSE/(n-2)

Assessing individual predictors

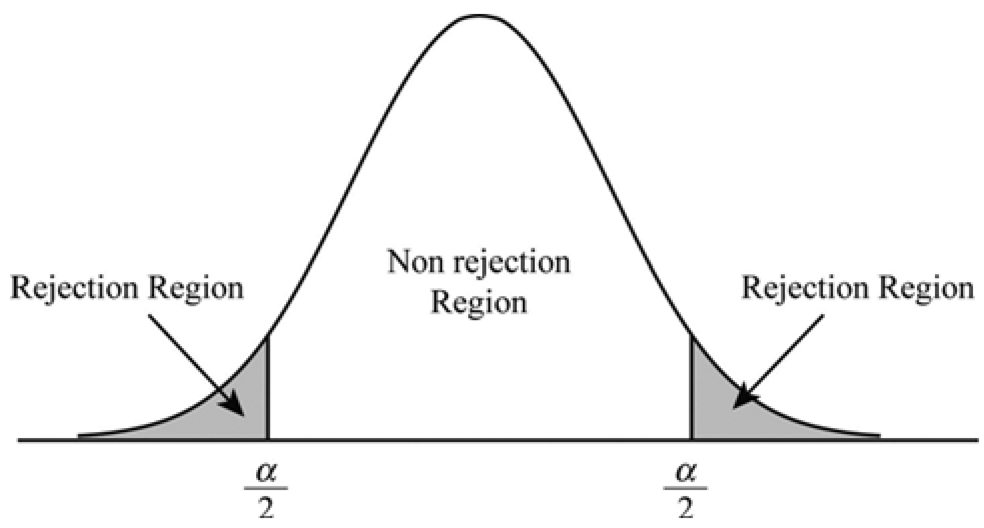
- Interpretation b_1
  - Change in average predicted outcome resulting from a unit change in the predictor
- Significance of b_1
  - H_0: $b_1=0$, tested with t-test
  - $t_{df=N-p-1} = \frac{b_{1observed} - b_{1expected}}{SE_{b1}} = \frac{b_{1observed}}{SE_{b1}}$

(p is the number of predictors in the model, thus p=1)

Test Statistic

- Known numerical summary of a data-set that reduces the data to one value.
- Used to perform the hypothesis test.
- The test statistic compares the data with what is expected under H_0.
- It is used to calculate the p-value.

Hypothesis testing



t_c = qt(0.025,N-p-1) t_c = qt(0.975,N-p-1)

What Does Statistical Significance Tell Us?

- p is the probability, given that H_0 is true, that the test statistic takes a value as extreme or more extreme than the one observed by chance

  significance depends on sample size.

  look also at the effect

- $P < 0.05$:
  - There is enough evidence to reject H_0
- $P > 0.05$:
  - There is not enough evidence to reject H_0
  - This does not mean that H_0 is true

Regression: An Example

- A record company boss was interested in predicting record sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and downloads) in the week after release
- Predictor variable:
  - The amount (in units of £1000) spent promoting the record before release.

workflow

- Summary statistics
- Scatterplots
- Statistical model
- Testing the assumptions

Regression in R: lm{stats}

- We run a regression analysis using the *lm()* function – lm stands for 'linear model'. This function takes the general form:

newModel <- lm(**outcome** ~ **predictor**, data = **dataFrame**, na.action = an action))

na.action=na.exclude will exclude the cases that have missing data on any variable in the model, known as casewise deletion

Regression in R

```
> albumSales.1 <- lm(sales ~ adverts, data = album1)
```

Output of a Simple Regression

- We have created an object called *albumSales.1* that contains the results of our analysis. We can show the object by executing:

summary(albumSales.1)

>Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.341e+02	7.537e+00	17.799	<2e-16 ***
adverts	9.612e-02	9.632e-03	9.979	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.99 on 198 degrees of freedom

Multiple R-squared: 0.3346, Adjusted R-squared: 0.3313

F-statistic: 99.59 on 1 and 198 DF, p-value: < 2.2e-16

Using the Model

$$\begin{aligned}\text{Record Sales}_i &= b_0 + b_1\text{Advertising Budget}_i \\ &= 134.14 + \left(0.09612 \times \text{Advertising Budget}_i\right)\end{aligned}$$

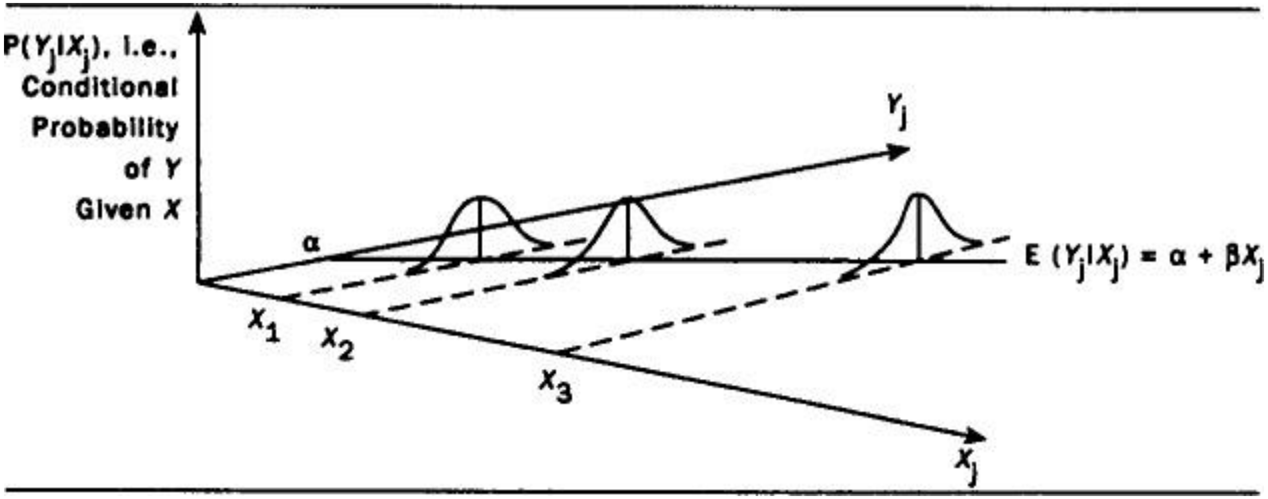
$$\begin{aligned}\text{Record Sales}_i &= 134.14 + \left(0.09612 \times \text{Advertising Budget}_i\right) \\ &= 134.14 + \left(0.09612 \times 100\right) \\ &= 143.75\end{aligned}$$

Checking Assumptions

- Variable type:
  - Pred cont or cat with non-zero variance
  - Outcome var: continuous
- Linearity:
  - Linear in the parameters
- Independent observations

Checking Assumptions

- Y|X identical normal distributions
- Homoscedasticity:
  - At each level of the predictor(s), the variance of the residuals should be the same



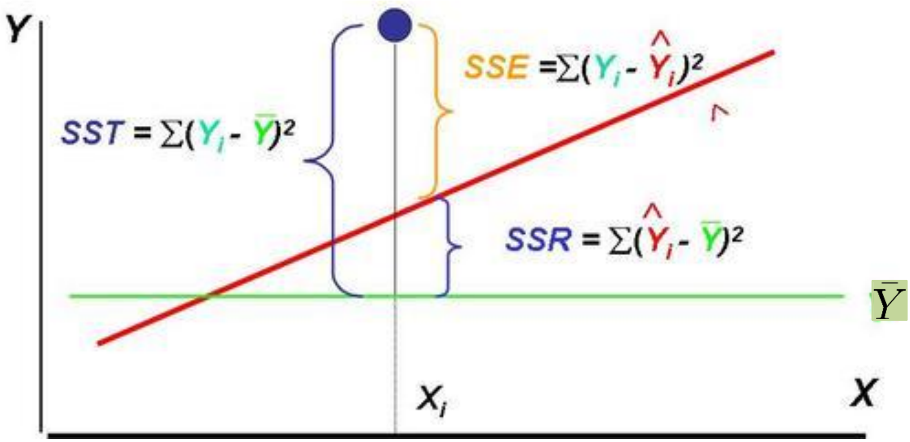
Checking Assumptions

$$\epsilon_i \sim i.i.d. N(0, \sigma^2)$$

Fitted Values and Residuals

- Fitted values are the estimates of Y as determined by the regression equation.
- Residuals are the differences between each observed value and the corresponding fitted value.

$$y_i = b_0 + b_1 x_i + e_i$$
$$\hat{y}_i = b_0 + b_1 x_i$$
$$y_i = \hat{y}_i + e_i$$
$$y_i - \hat{y}_i = e_i$$

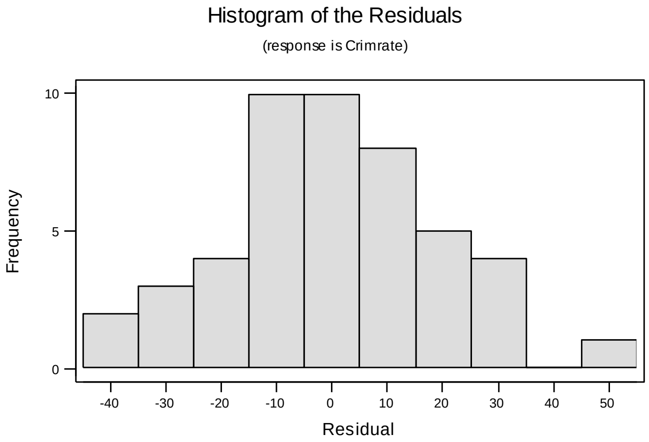


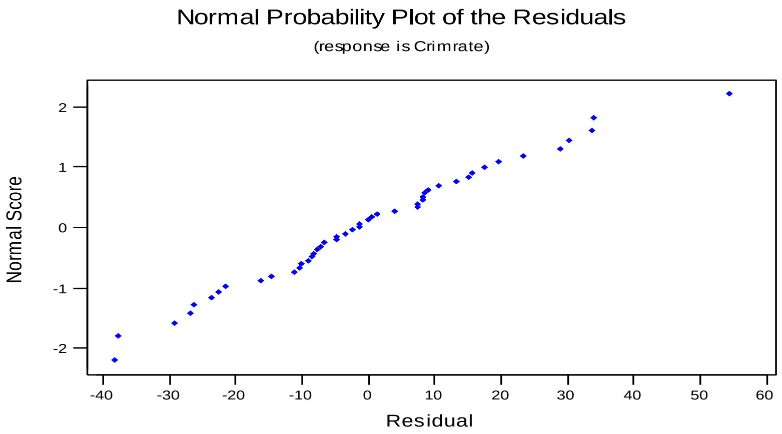
Diagnosis of Violation of Assumptions

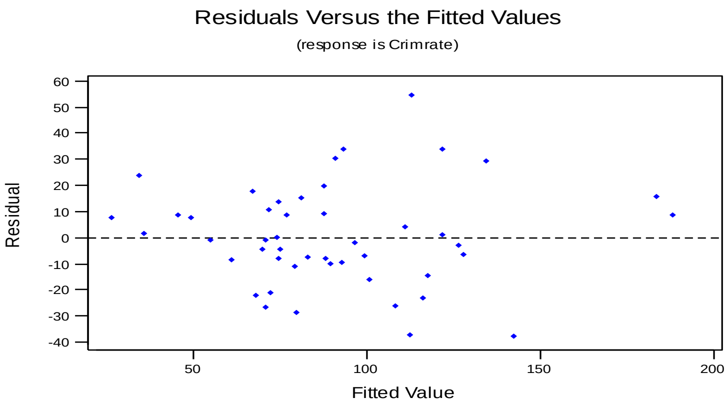
Residual Plots are used to check for:

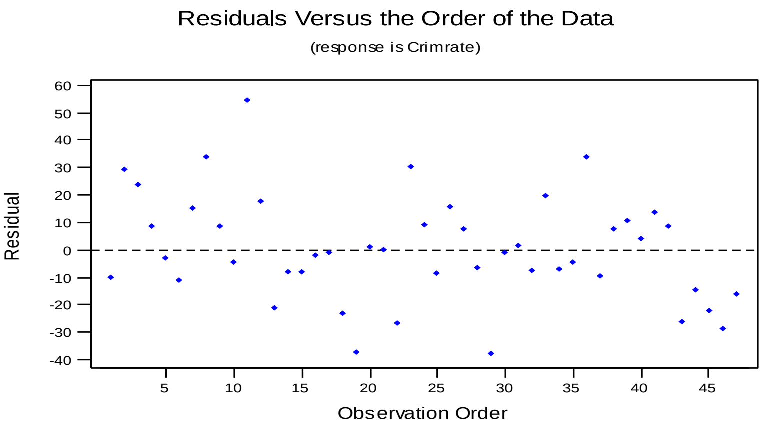
- Variance not being constant across the explanatory variables.
- Fitted relationship not being linear.
- Random variation not having a Normal distribution.

Residual Plots





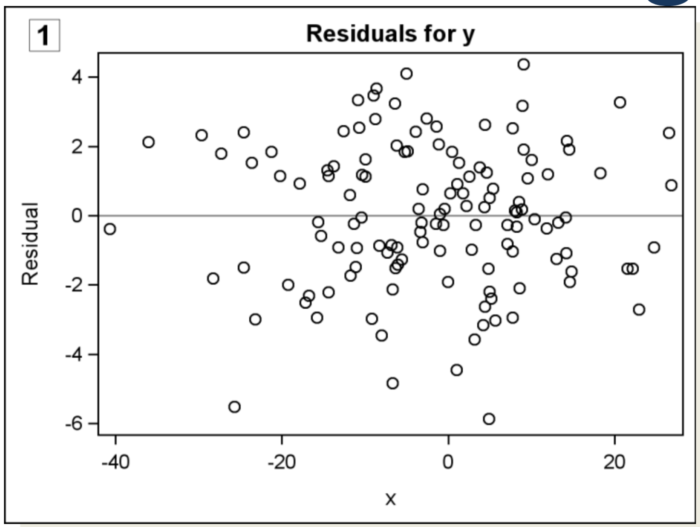


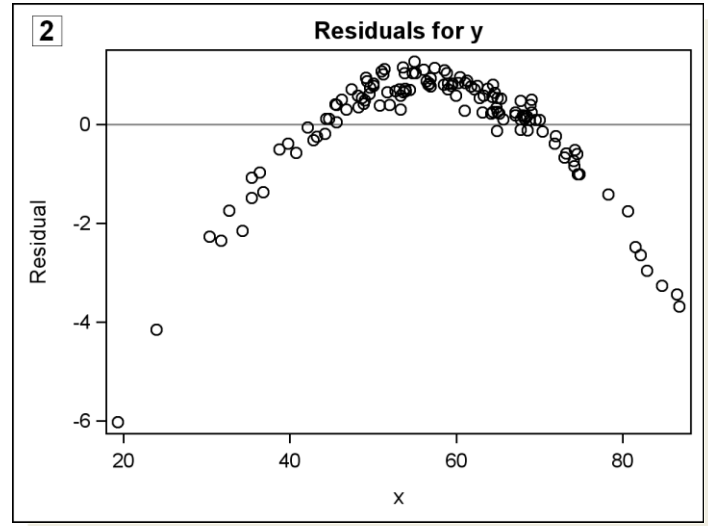


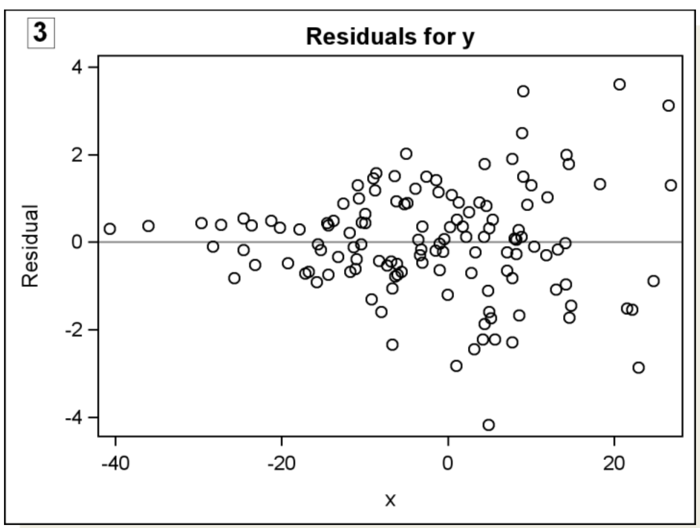
Studentized Residuals

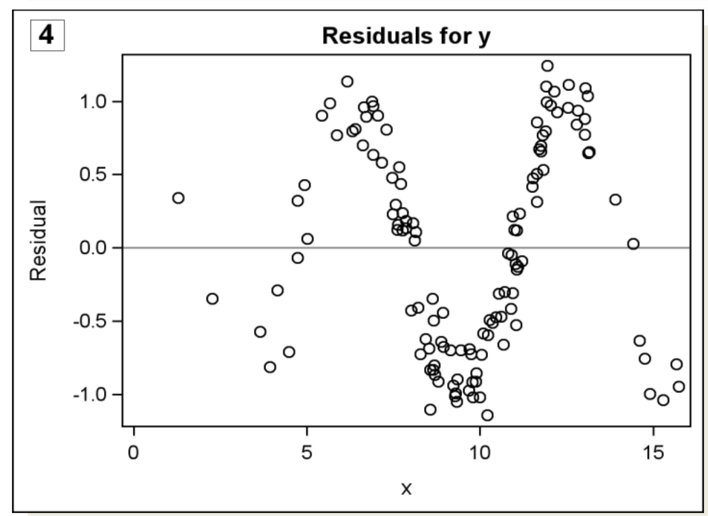
- Disadvantage raw residuals:
  - In same unit as observation (what is small/large?)
- Studentized residual:
  - Residual divided by the estimated standard deviation of the residuals (95% within -2 and +2)
- Suggested cutoffs are as follows:
  - |SR| > 2 for data sets with a relatively small number of observations
  - |SR| > 3 for data sets with a relatively large number of observations

Examining Residual Plots





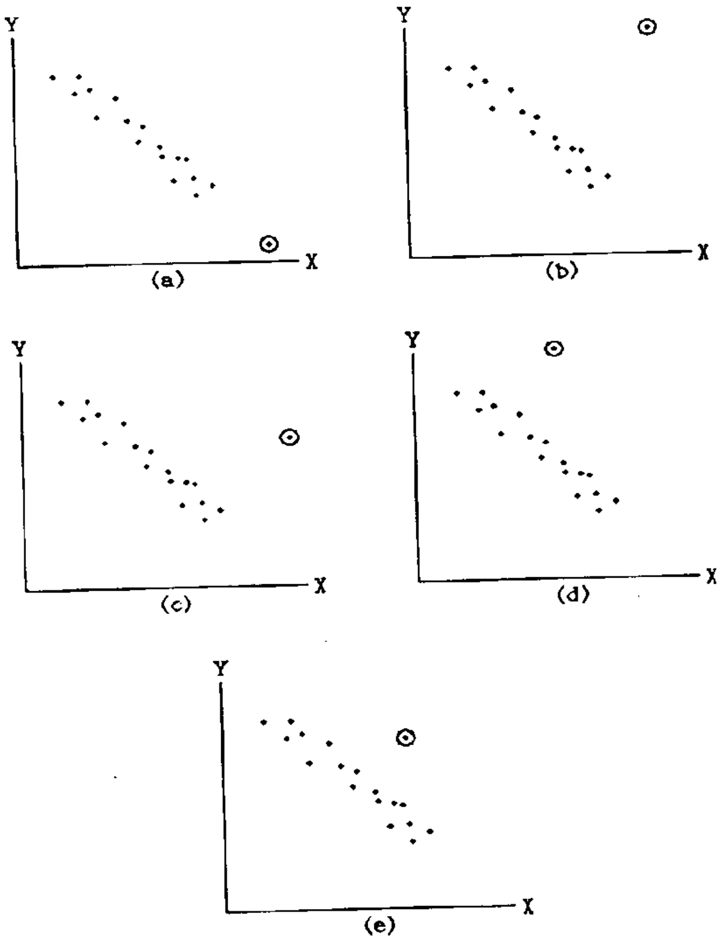




Checking for independence

- Index plot
- Durban-Watson test (d)
  - Tests for 1st order autocorrelations ρ between adjacent errors
  - Value of 2: residuals uncorrelated
  - Values < 2 : positive correlation
  - Values > 2: negative correlation

Influential observations



- An observation is influential if the estimates change substantially when the point is omitted.

Diagnostic Statistics

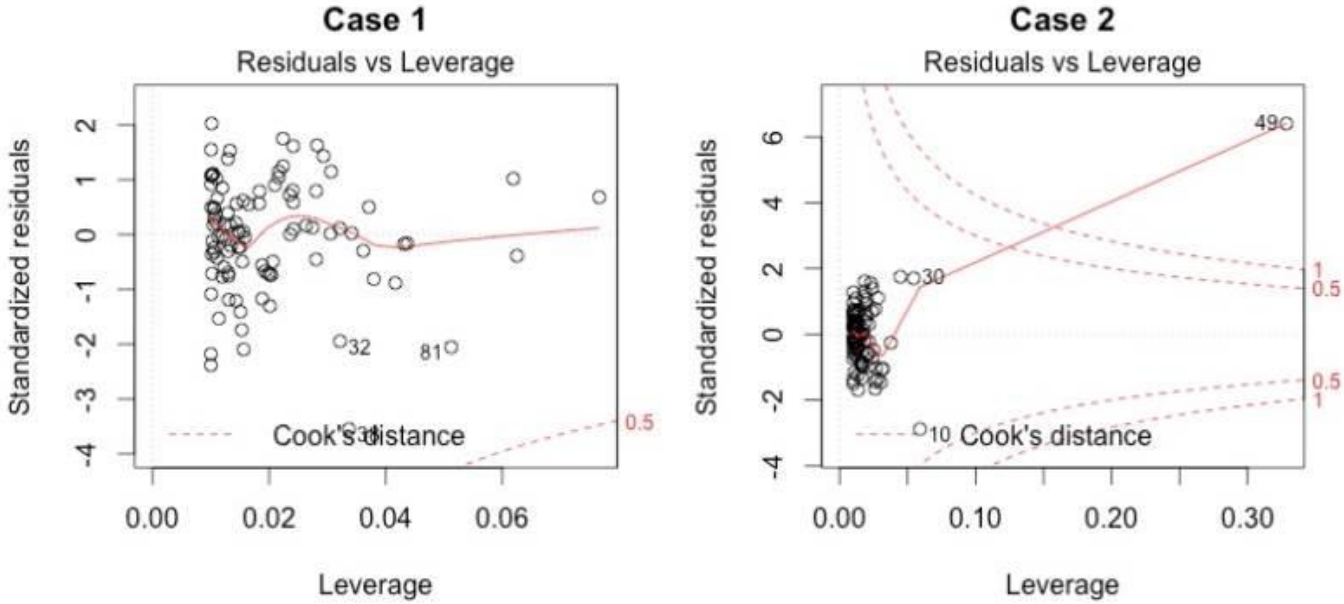
- Statistics that help identify influential observations are the following:
  - Studentized residuals
  - Cook's Distance
  - Leverage

Cook's *D* Statistic

- Cook's *D* statistic is a measure of the simultaneous change in the parameter estimates when the i^{th} observation is deleted from the analysis.

Leverage

- Check for outliers long distance away from the rest of the data. They exercise **leverage**, they can influence the regression line



How to Handle Influential Observations

1. Recheck the data to ensure that no transcription or data entry errors occurred.
2. If the data are valid, one possible explanation is that the model is not adequate.
3. Report the results both with influential observations included and with influential observations deleted.

- A model with higher-order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well.

final comment on diagnostics

- Diagnostics are tools to see how good or bad a model is
- They are **NOT** a way of justifying the removal of data points

DEMO REGRESSION

Open the program Ch7_regression.R

Exercises

- Task 1

  Read in the file “ANCEX.csv”. It has 2 variables XVAR and YVAR

  a. Calculate summary statistics
  b. Draw a scatter plot with XVAR and YVAR
  c. Perform a linear regression with XVAR predicting YVAR
  d. Write down the equation of the regression line. Interpret the results
  e. Assess the fit of the model
  f. Extract the confidence intervals for the parameters using the command ***confint***
  g. Test the assumptions
  h. Perform a linear regression with YVAR predicting XVAR and compare the slopes

Exercises

- Task 2

  load the **bodyfat2.txt** data (tab delimited)

  a. Calculate summary statistics
  b. Draw a scatter plot with PctBodyFat2 and Weight
  c. Perform a simple linear regression model with PctBodyFat2 as the response variable and Weight as the predictor variable.
  d. Write down the equation of the regression line. Interpret the slope
  e. What is the p- value of the F statistic. How would you interpret this with regards to the null hypothesis?
  f. What is the value of R^2. How would you interpret this?
  g. Assess the fit of the model.
  h. Extract the confidence intervals for the parameters.
  i. Test the assumptions.

DFFITS

- DFFITS_i measures the impact that the i^{th} observation has on the predicted value.
- A suggested cutoff for influence is shown below:

$$|\,\mathbf{DFFITS_i}\,| > 2\sqrt{\frac{p}{n}}$$

DFBETAS

- Measure of change in the j^{th} parameter estimate with deletion of the i^{th} observation
- One DFBETA per parameter per observation
- Helpful in explaining on which parameter coefficient the influence most lies
- A suggested cutoff for influence is shown below:

$$|\mathbf{DFBETA}_{\mathbf{ij}}| > 2\sqrt{\frac{1}{n}}$$

Leverage

- Check for outliers long distance away from the rest of the data. They exercise leverage, which is checked by "h_i". It is considered large if more than (p+1)/n (p=number of predictors including the constant).

$$\mathbf{y} = \mathbf{Xb} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y} = \mathbf{Hy}$$