Linear Regression

Robert Gentleman

Linear Regression

- Similar to correlation analysis, simple linear regression can be used to explore the nature of the relationship between two continuous random variables
- One difference is that regression looks at the change in one variable that corresponds to a given change in the other, we think of one as the response, and the other variable is often selected because we believe it affects the response, or is the result of an experimental design
- A more important distinction is that we can use a number of covariates to build a model
- One objective is to predict or estimate the value of the response associated with a fixed value of the explanatory variable
- Correlation analysis does not distinguish between the two variables

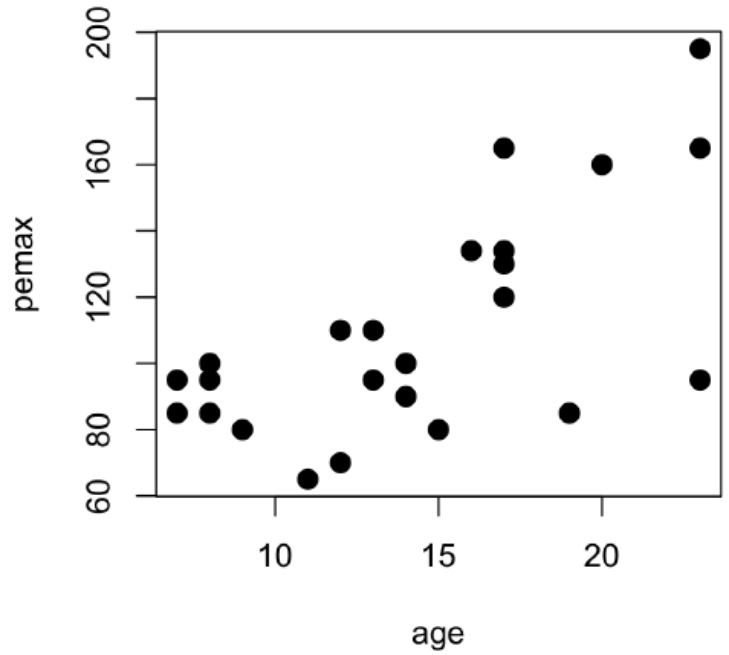
Example: Lung Function in CF patients

- A study on lung function in patients with cystic fibrosis
- PEmax (maximal static expiratory pressure, cm H_20) is the response variable
- A potential list of explanatory variables relate to body size or lung function: age, sex, height, weight, BMP (body mass as a percentage of the age-specific median), FEV1 (forced expiratory volume in 1 second), RV (residual volume), FRC (functional residual capacity), TLC (total lung capacity)
- For now, let's consider age alone
- Quantify this relationship by postulating a model of the form

$$y = \alpha + \beta x + e \ , \ \ e \sim N\left(0, \sigma^2\right)$$

Example: Lung Function in CF patients

- Plot PEmax vs age



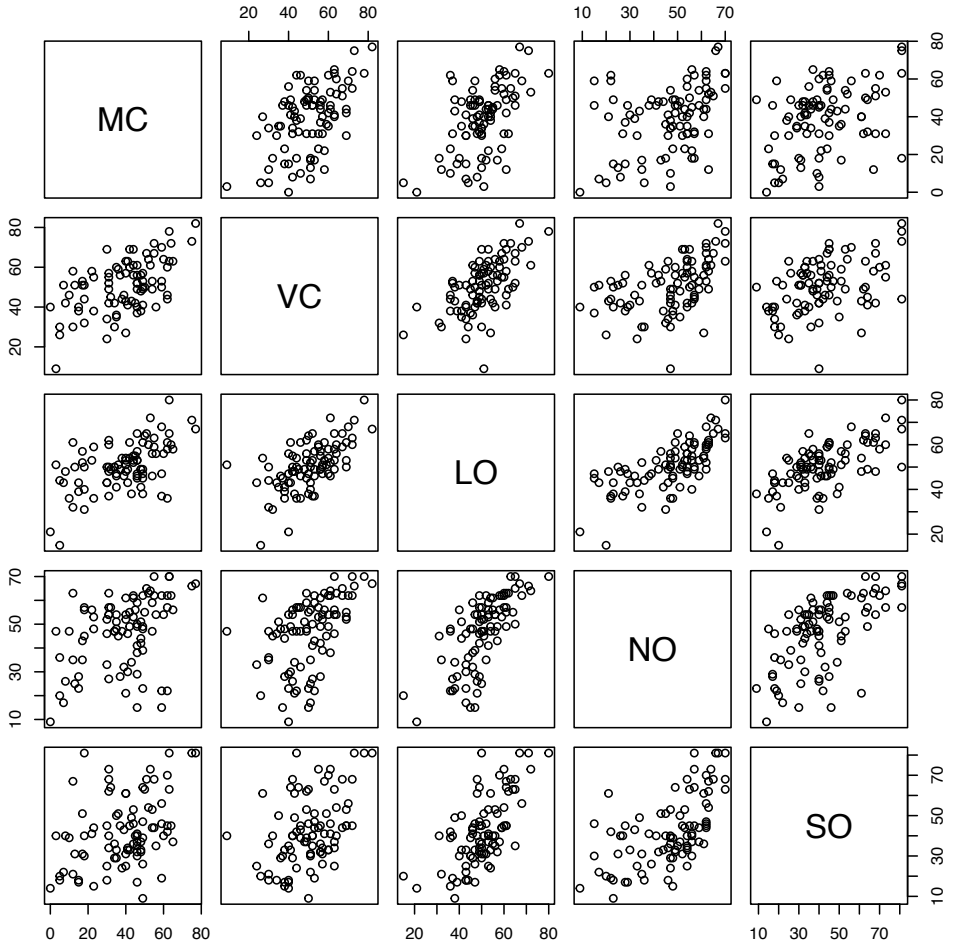
- Despite the scatter, it appears that PEmax tends to increase as age increases
- Data (O'Neill et al, Am Rev Respir Dis. 1983) available from ISwR package ("Introductory statistics with R" book by Dalgaard)

You can find out more about it from R

```
cystfibr                package:ISwR                R Documentation

Cystic fibrosis lung function data

Description:

     The 'cystfibr' data frame has 25 rows and 10 columns.  It contains
     lung function data for cystic fibrosis patients (7-23 years old).

Usage:

     cystfibr

Format:

     This data frame contains the following columns:

     'age' a numeric vector, age in years.

     'sex' a numeric vector code, 0: male, 1:female.

     'height' a numeric vector, height (cm).

     'weight' a numeric vector, weight (kg).

     'bmp' a numeric vector, body mass (% of normal).

     'fev1' a numeric vector, forced expiratory volume.

     'rv' a numeric vector, residual volume.

     'frc' a numeric vector, functional residual capacity.

     'tlc' a numeric vector, total lung capacity.

     'pemax' a numeric vector, maximum expiratory pressure.

Source:

     D.G. Altman (1991), _Practical Statistics for Medical Research_,
:
```

A multivariate data set

- For data sets with multiple variables using the **pairs** function plots all by all



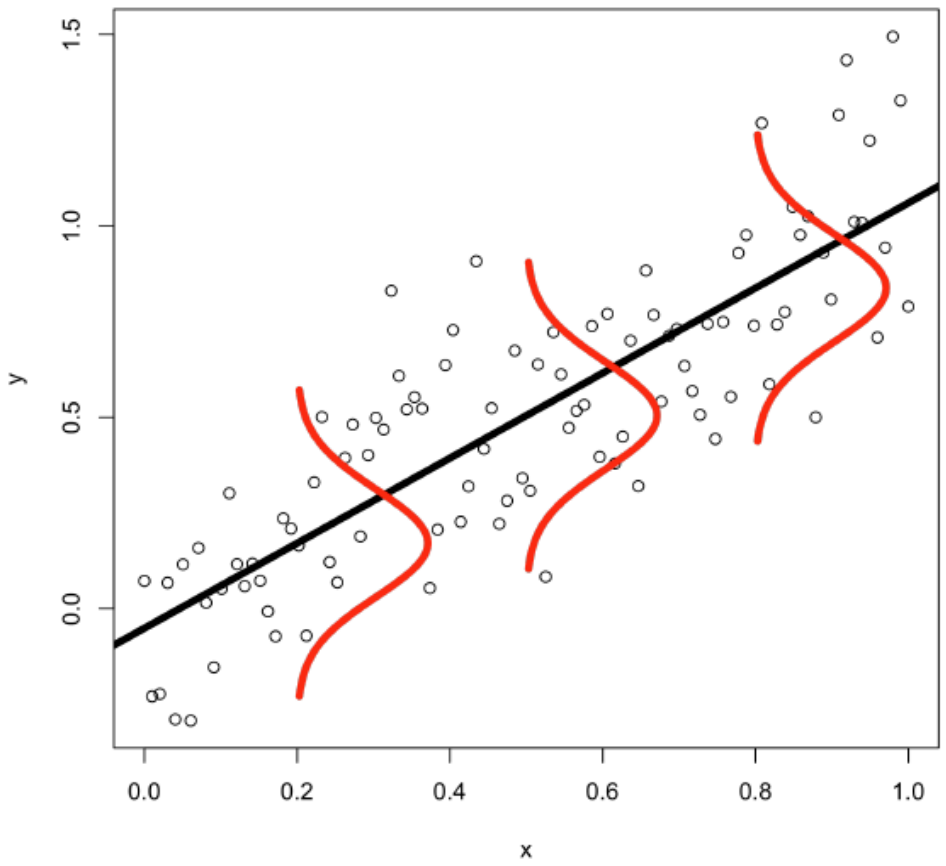
Linear Regression Model

$$y = \alpha + \beta x + e \ , \ \ e \sim N\left(0, \sigma^2\right)$$

- y: dependent/response/outcome variable
- x: independent/explanatory/predictor variable
- e: error term
- α, β: coefficients/regression coefficients/model parameter
  - α: intercept
  - β: slope, describes the magnitude of association between X and Y
- For any given x, y = constant + normal random variable

Assumptions

- For a specified value of x, the distribution of the y values is normal with mean $y = \alpha + \beta x$ and standard deviation σ



- For any specified value of x, σ is constant
- This assumption of constant variability across all values of x is known as **homoscedasticity**

Residuals

- Use the data from the sample to estimate α and β, the coefficients of the regression line

$$y = \alpha + \beta x + e \ , \ e \sim N(0, \sigma^2)$$

- Call the estimators *a* and *b*

$$\hat{y} = a + bx$$

- The discrepancies between the observed and fitted values are called residuals

$$d = y - \hat{y}$$
$$= y - a - bx$$

Fitting the Model: well described in MSMB

- One mathematical technique for fitting a straight line to a set of points is known as the method of least squares
- To apply this method, note that each data point (x_i, y_i) lies some vertical distance from d_i from an arbitrary line (d_i is measured parallel to the vertical axis)
- Ideally, all residuals would be equal to 0
- Since this is impossible, we choose another criterion: we minimize the sum of squared residuals

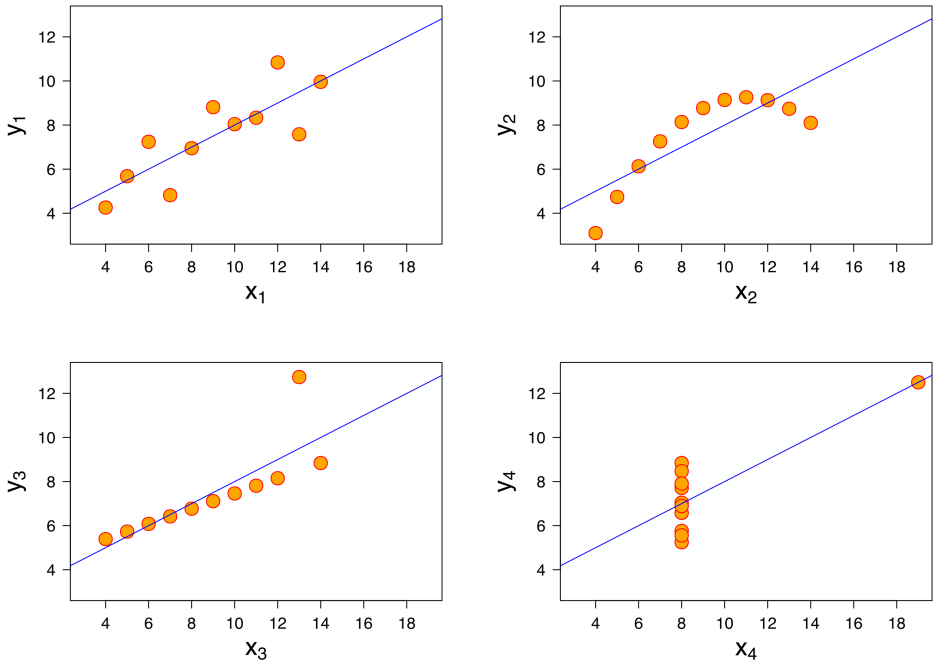
$$S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Goodness of Fit

- After estimating the model parameters, we need to evaluate how well the model fits the data
- In general we use graphical methods to assess GoF - summary statistics are usually not sufficient (see homework 5 Q2)
- Graphics are essential
  - Residual plots and other tools are essential
- You can get some sense of how the model fits the data by looking at
  - Inference about beta
  - R^2
  - But any interpretation of them assumes that the model is correct - and so they cannot inform you about GoF

Why Graphics are Essential

- Anscombe's plots - the summary statistics are identical



For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

In R the data are available as **anscombe**, plot them, fit models, plot residuals etc

https://en.m.wikipedia.org/wiki/Anscombe%27s_quartet

Inference about β

- Because the parameter β describes the relationship between X and Y, inference about β tells us about the strength of the linear relationship.
- After estimating the model parameters, we can do hypothesis testing and build confidence intervals for β
- The standard error of b in a sample linear regression is estimated as

$$\hat{s.e.}(b) = \sqrt{\frac{\left(\frac{1}{n-2}\right)\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

Inference about β

- To test the hypothesis H_0:β=0, we calculate the test statistic

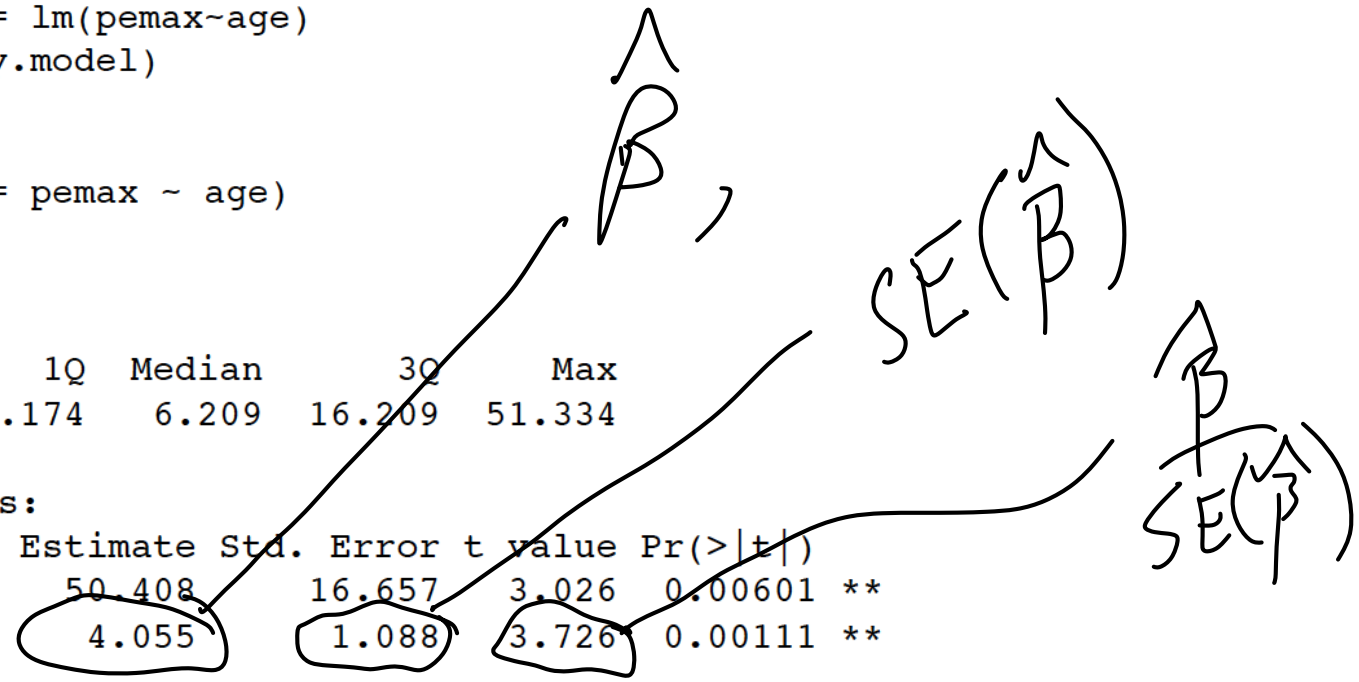
$$t = \frac{b}{\hat{s.e.}(b)}$$

- Under H_0, this has a t distribution with n-2 df (for a simple model with one covariate- with more covariates you need to adjust)
- If the true population slope is equal to 0, there is no linear relationship between x and y; x is of no value in predicting y
- 100(1-α) CI for β

$$b \pm t_{n-2,\, 1-\frac{\alpha}{2}} \hat{s.e.}(b)$$

Example of Cystic Fibrosis Patients

```
> install.packages("ISwR")
> library(ISwR)
> data(cystfibr)
> attach(cystfibr)

> my.model = lm(pemax~age)
> summary(my.model)

Call:
lm(formula = pemax ~ age)


Residuals:
    Min      1Q  Median      3Q     Max
-48.666 -17.174   6.209  16.209  51.334

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.408     16.657   3.026  0.00601 **
age            4.055      1.088   3.726  0.00111 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.97 on 23 degrees of freedom
Multiple R-squared: 0.3764,   Adjusted R-squared: 0.3492
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109
```



Interpretation

```
Residuals:
    Min      1Q  Median      3Q     Max
-48.666 -17.174   6.209  16.209  51.334

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.408     16.657   3.026  0.00601 **
age            4.055      1.088   3.726  0.00111 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.97 on 23 degrees of freedom
Multiple R-squared: 0.3764,   Adjusted R-squared: 0.3492
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109
```

- The multiple R-squared is called the coefficient of determination and reflects the amount of variation in the y’s explained by the model
- It must increase if a new variable is added.
- The F-test at the bottom tests H_0 that a null model, with only a mean vs H_1 that the covariates improve fit

Example: CF

- We reject H_0 and conclude that the population slope is not equal to 0. PEmax increases as age increases.
- Check:

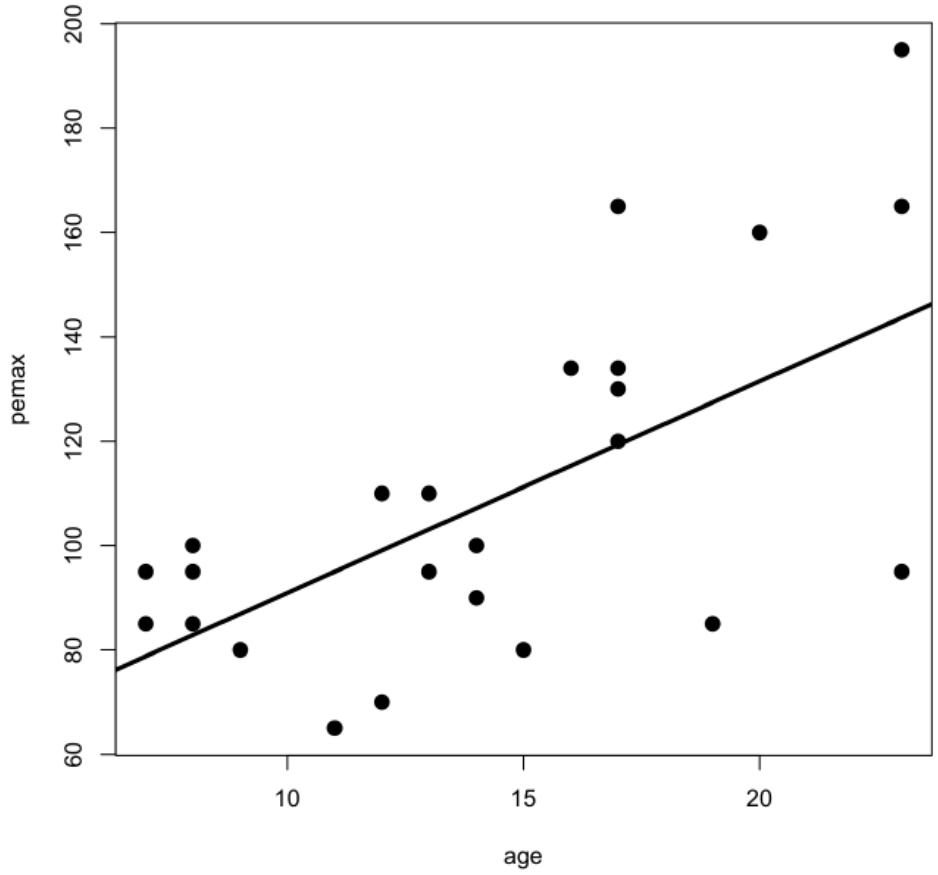
```
4.055/1.088 = 3.727
(1-pt(3.726,23))*2 = 0.0011
```

- A 95% confidence interval for beta is

```
4.055 +/- 2.069*1.088 = (1.80, 6.31)
qt(.975,23) = 2.07
```

Plotting the Regression Line

```
plot(age,pemax,cex=2,pch=20)
names(my.model)
abline(my.model$coeff[1],my.model$coeff[2],lw=3)
```



R^2

- R^2, sometimes called the coefficient of determination:

$$R^2 = \frac{\text{Reg SS}}{\text{Total SS}} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

- **This is the proportion of variation explained by the model**
- Higher values of R^2 indicate that more of the variability in Y is explained by the covariate(s)
- Adding new covariates to a model necessarily increases R^2 so we need to have statistical methods to ascertain whether the increase is sufficiently large to merit inclusion of the covariate

Residual Plots

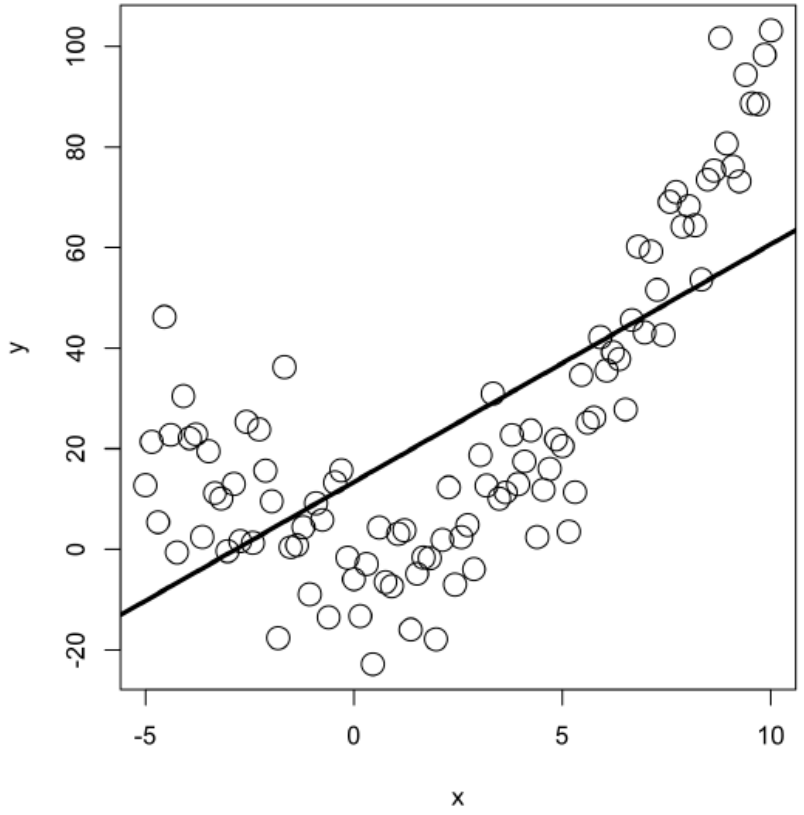
- We've been assuming that the association between *X* and *Y* in the population is truly linear.
- Even if the association is nonlinear, these methods may still fit a line without detecting a problem. In this case, inferences from the model will not be correct.
- Previously we defined a point's **residual**:

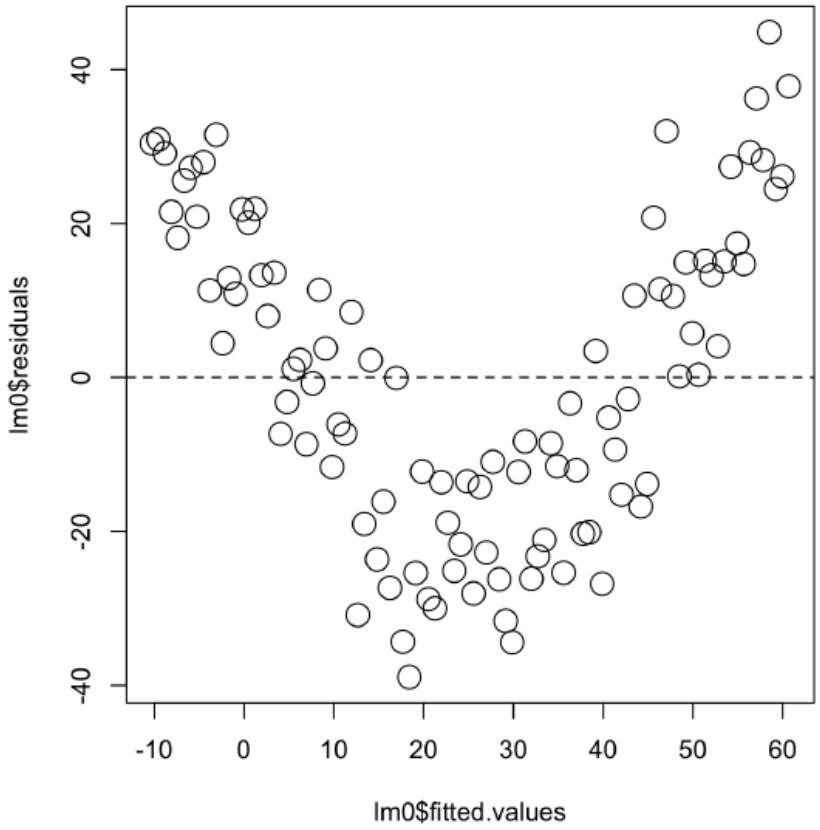
$$d_i = y_i - \hat{y}_i = y_i - a - bx_i$$

- Because of the assumptions of linear regression, we expect all the residuals to be normally distributed with the same mean (0) and the same variance.
- Violations of the linear regression assumptions can often be detected on a residual plot

Residual Plots

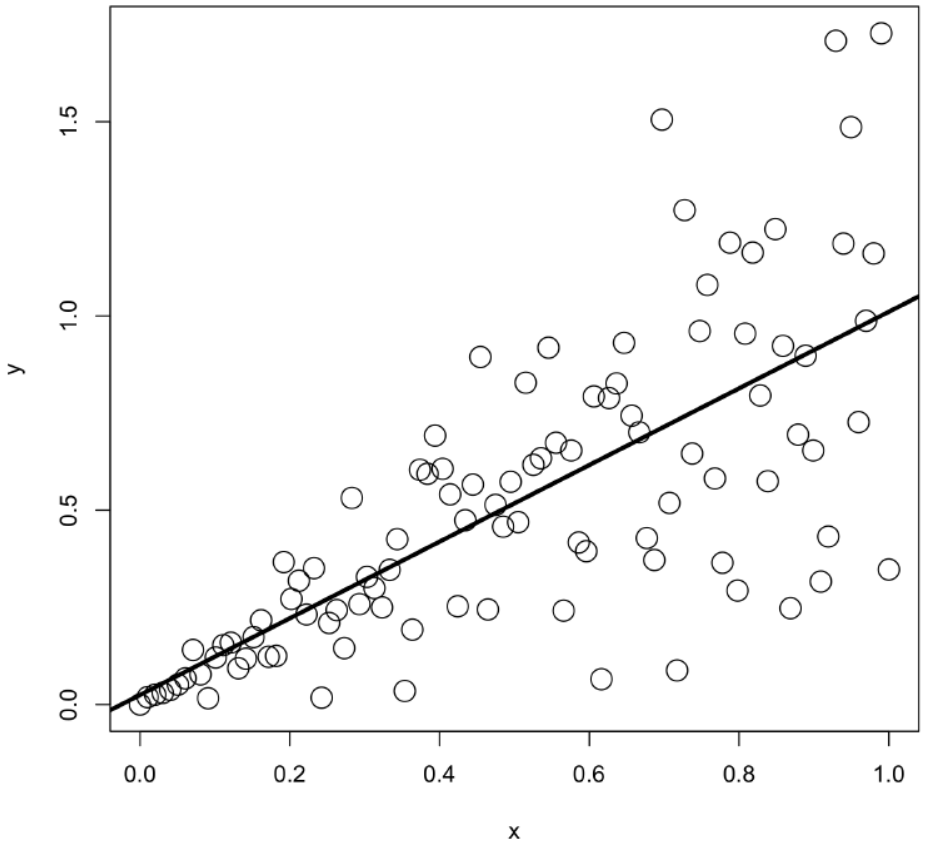
- Plot the fitted (predicted) *y-values* on the *x-axis* and the residuals on the *y-axis - right? On the left we see y vs x.*
- Are the residuals normally distributed with constant variance?

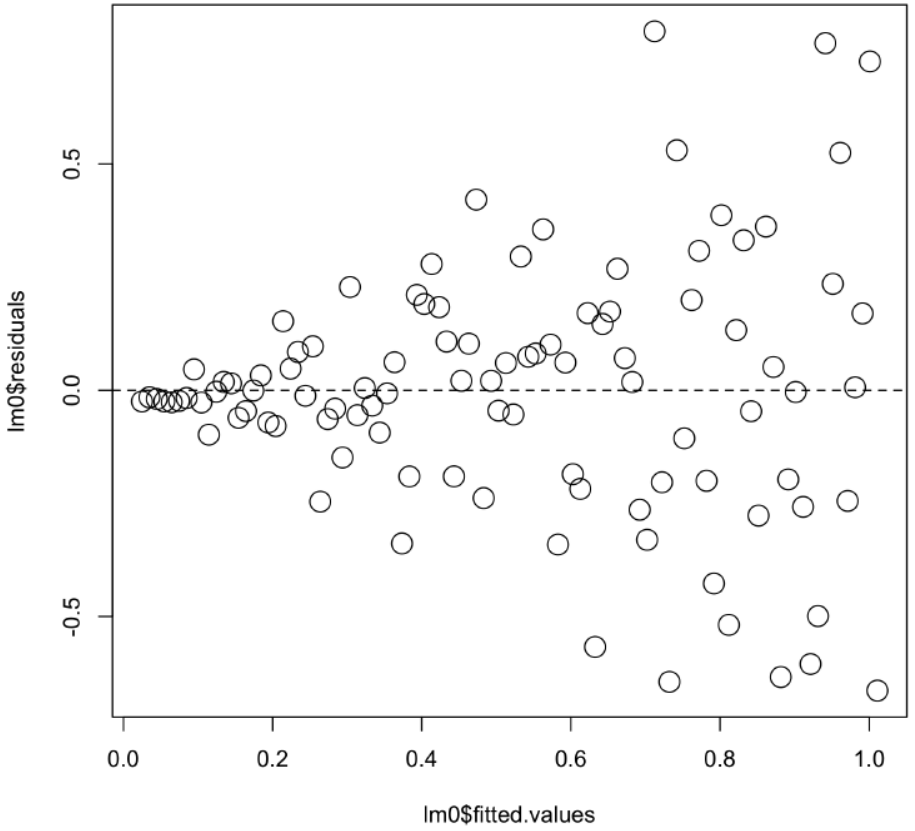




Residual Plots

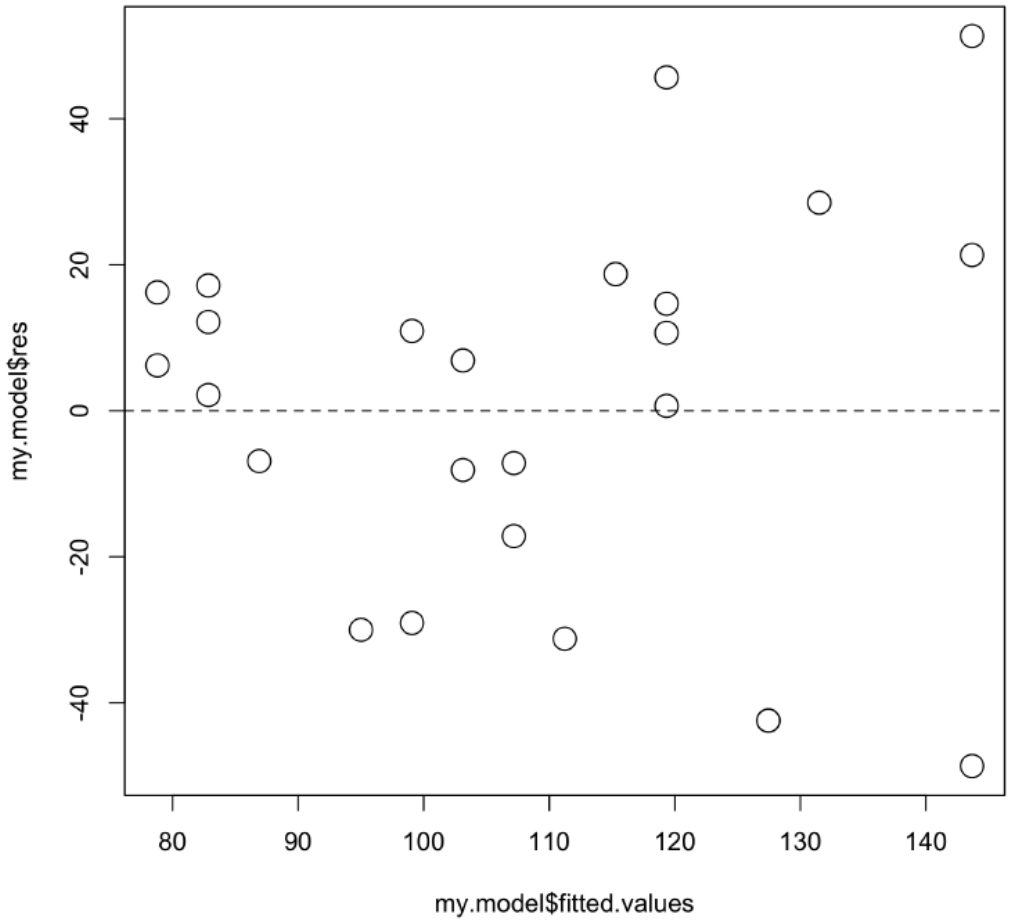
- A different example (not the same data set):





Example: Cystic Fibrosis Patients

- Does this model violate the assumption for constant variance?



Linear Regression

- A linear regression equation is ***linear in the parameters***.
- Which models are 'linear'?
  - $y = a + bx$
  - $y = bx$
  - $y = a + b_1x_1 + b_2x_2$
  - $y = a + b\ x_1^2$
  - $\log(y) = a + bx$
- In fact, linear regression is not so restrictive
- And we often want to transform both y (eg log(y)) or some of the covariates in order to improve the assumptions

Summary: Simple Linear Regression

- Linear model

$$y = \alpha + \beta x + e \;,\; e \sim N(0, \sigma^2)$$

- Method of Least Squares

$$S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- Testing for significance of coefficients

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{s.e.}(b) = \sqrt{\frac{\left(\frac{1}{n-2}\right)\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$t = \frac{b}{\hat{s.e.}(b)}$$

Multiple Linear Regression

- If knowing the value of a single explanatory variable improves our ability to predict a continuous response, we might suspect that information about additional variables could also be used to our advantage
- To investigate the more complicated relationship among a number of different variables, we use multiple linear regression analysis

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$
$$e \sim N(0, \sigma^2)$$

Multiple Linear Regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$
$$e \sim N(0, \sigma^2)$$

- The intercept α is the mean value of the response when all k explanatory variables are equal to 0
- The slope β_j is the change in y that corresponds to a one-unit increase in x_j, given that all other explanatory variables remain constant
- The model is no longer a simple but something multidimensional

MSMB discusses diff^t ways to fit a line to data.

Least Squares

- Again, we define the "best" line by minimization of the sum of squared residuals

$$S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(y_i - \left[a + b_1 x_{i1} + \cdots + b_k x_{ik} \right] \right)^2$$

- Unfortunately, there is no simple formulas for the coefficients
- There is an elegant solution but this requires more mathematical notations
- Hypothesis testing for the coefficients is done the same way

Visualizing Data

- Before performing any analysis, it is good to view the data

```
> plot(cystfibr)
pairs(cystfibr,gap=0)
```

- You can see the close relationship between age and height and weight

A Single Predictor Model

```
> my.model = lm(pemax ~ age)
> summary(my.model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.408     16.657   3.026  0.00601 **
age            4.055      1.088   3.726  0.00111 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.97 on 23 degrees of freedom
Multiple R-squared: 0.3764,       Adjusted R-squared: 0.3492
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109
```

- Age is a significant predictor of PEmax
- PEmax = 50.4 + 4.06 * age

A Two-Predictor Model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

```
> my.model = lm(pemax ~ age + height)
> summary(my.model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.8600    68.2493   0.262    0.796
age           2.7178     2.9325   0.927    0.364
height        0.3397     0.6900   0.492    0.627

Residual standard error: 27.43 on 22 degrees of freedom
Multiple R-squared: 0.3831,       Adjusted R-squared: 0.3271
F-statistic: 6.832 on 2 and 22 DF,  p-value: 0.00492
```

- PEmax = 17.9 + 2.72 * age + 0.40 * height
- How to interpret the coefficients?
- Which terms are significant here?

Inference for Coefficients

- We test the following hypothesis:
  $H_0 : \beta_j = 0$ (and all other βs $\neq 0$)
  $H_1 : \beta_j \neq 0$ (and all other βs $\neq 0$)

- The test statistic

$$t = \frac{b_j}{\hat{s}.e.(b_j)}$$

  follows a *t*-distribution with (*n-k-1*) df under the null
- *k* is the number of explanatory variables

Adjusted R2

```
> my.model = lm(pemax ~ age)
Multiple R-squared: 0.3764,       Adjusted R-squared: 0.3492
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109

> my.model = lm(pemax ~ age + height)
Multiple R-squared: 0.3831,       Adjusted R-squared: 0.3271
F-statistic: 6.832 on 2 and 22 DF,  p-value: 0.00492
```

- Age explained 37.6% of the variability in PEmax (about its mean)
- Age and height explained 38.3% of the variability in PEmax
- The inclusion of an additional variable in a regression model can never cause R2 to decrease
- To get around this problem, we use the adjusted R2 to penalize for the added complexity of the model
- Here, adjusted R2 decreased. We conclude that this model is not an improvement over the age-only model
- Note that the F-statistic remains significant

F-test

- We perform inference about them together to determine whether the model demonstrates a statistically significant relationship between any predictor variable and the outcome variable

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$
$$e \sim N(0, \sigma^2)$$

- $\mathbf{H_0}$: $\beta_1 = \beta_2 \ldots = \beta_k = 0$ vs $\mathbf{H_1}$: at least one $\beta_i \neq 0$
- We use the F-test to test this hypothesis

F-test

$$y_i = a + b_1 x_1 + \ldots + b_K x_K$$

- *Total sum of squares can be decomposed into* ***Regression*** *sum of squares (part explained by the model) and* ***Residual*** *sum of squares (remaining part)*

$$\text{Total SS} = \text{Reg SS} + \text{Res SS}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- *We normalize by the degrees of freedom to get regression and residual mean sum of squares. The ratio of these two values follows an F-distribution with (k, n-k-1) df.*

$$\text{Reg MS} = \frac{\text{Reg SS}}{k}$$

$$\text{Res MS} = \frac{\text{Res SS}}{n-k-1}$$

$$F = \frac{\text{Reg MS}}{\text{Res MS}}$$