Lecture 8: Linear Regression

- Model
- Inferences on regression coefficients
- R^2
- Residual plots
- Handling categorical variables
- Adjusted R^2
- Model selection
- Forward/Backward/Stepwise

BMI 713
November 14, 2017
Peter J Park

Linear Regression

- Like correlation analysis, simple linear regression can be used to explore the nature of the relationship between two continuous random variables
- The main difference is that regression looks at the change in one variable that corresponds to a given change in the other
- The 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

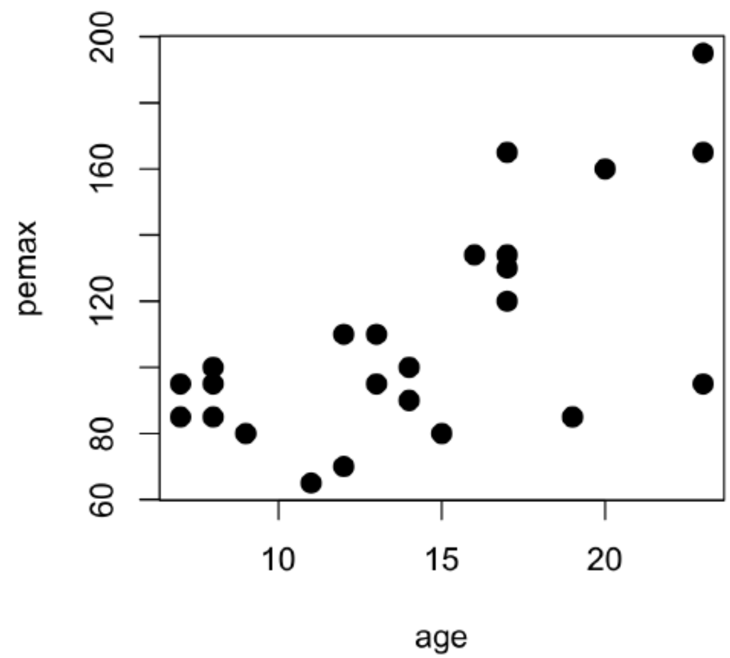
Example: Lung Function in CF patients

- A study on lung function in patients with cystic fibrosis
- PEmax (maximal static expiratory pressure, cm $H_2 0$) 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

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

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)

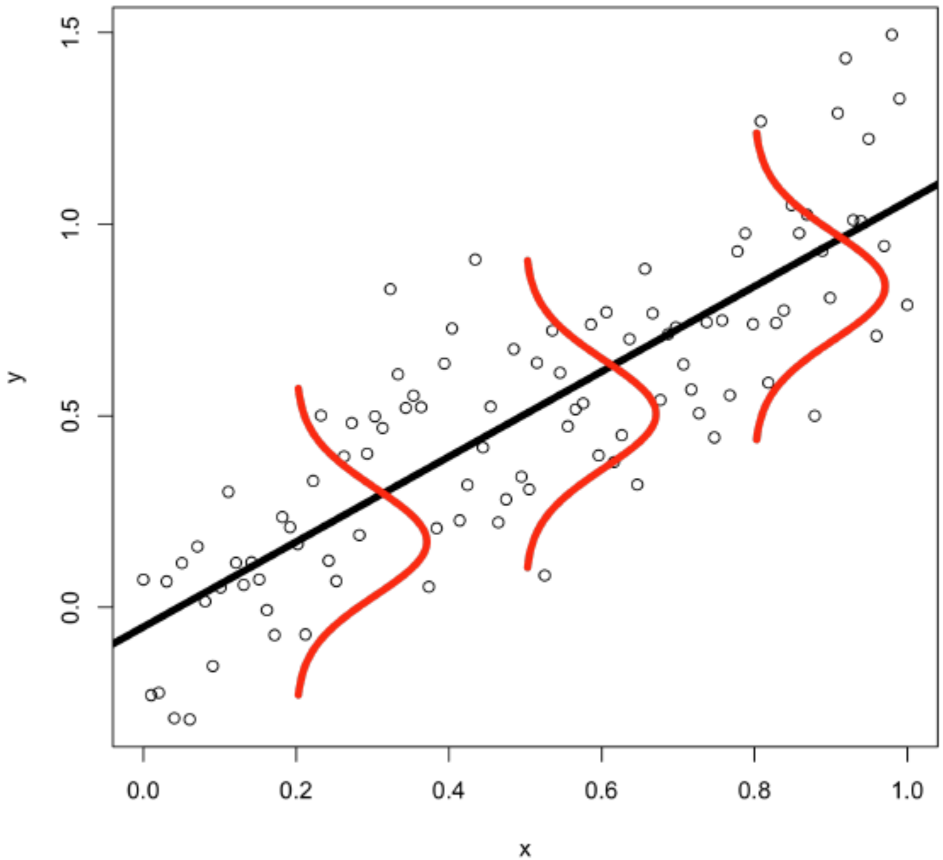
Linear Regression Model

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

- 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
- The values x are considered to be measured without error

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

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

- Call the estimators *a* and *b*

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

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

$$\begin{aligned} d &= y - \hat{y} \\ &= y - a - bx \end{aligned}$$

Fitting the Model

- 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

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

Fitting the Model

- The resulting line is the **least squares line**
- Using calculus, it can be shown that

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

$$a = \bar{y} - b\bar{x}$$

- Once a and *b* are known, we can substitute various values of x into the regression and compute *y*.

Goodness of Fit

- After estimating the model parameters, we need to evaluate how well the model fits the data
- Three criteria:
  - Inference about beta
  - R^2
  - Residual plots
- These concepts will hold for more complex cases, such as multiple regression, logistic regression, and Cox regression

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

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

Inference about β

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

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

- Under H_0, this has a t distribution with n-2 df
- 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

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

- We can also carry out a similar procedure for α

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
```

Example: CF

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

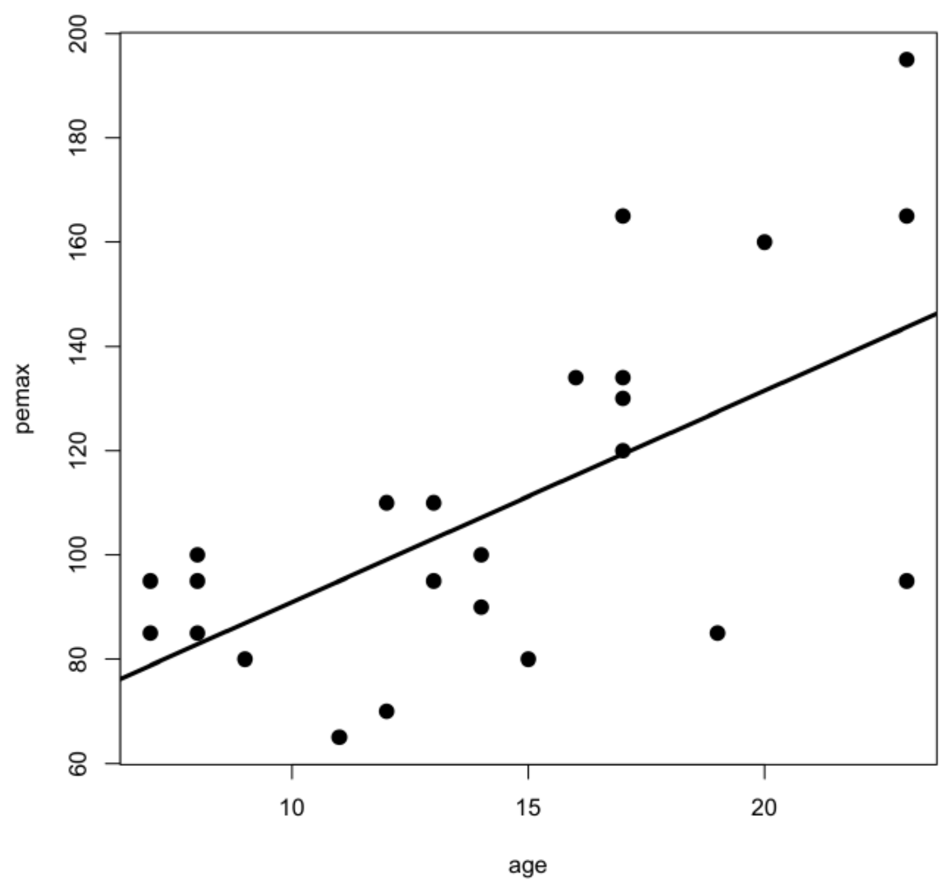
```
50.408/16.657=3.026
(1-pt(3.026,23))*2 = 0.00601
```

- A 95% confidence interval for beta is

```
50.408 +/- 2.069*(16.657)          (qt(.975,23)=2.06866)
(15.9, 84.9)
```

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

- Another measure is 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**
- It is also the square of Pearson's correlation coefficient

```
> cor(pemax,age)^2
[1] 0.3763505
```

Residual Plots

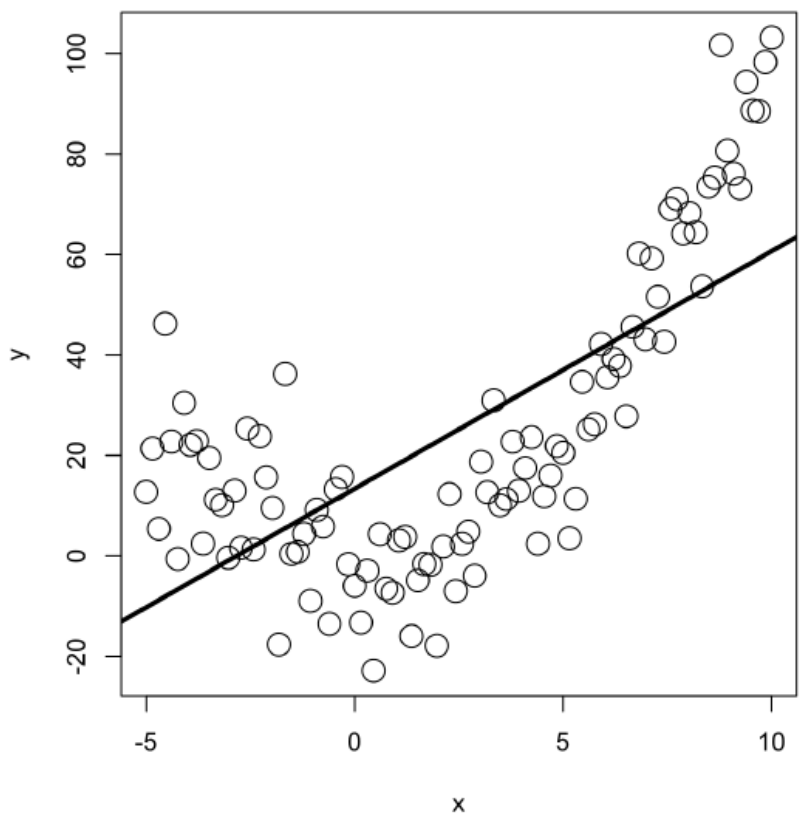
- We've been assuming that the association between *X* and *Y* in the population is truly linear.
- Even if the association is nonlinear, these methods may still fit a line without detecting a problem. In this case, inferences from the model will not be correct.
- Previously we defined a point's **residual**:

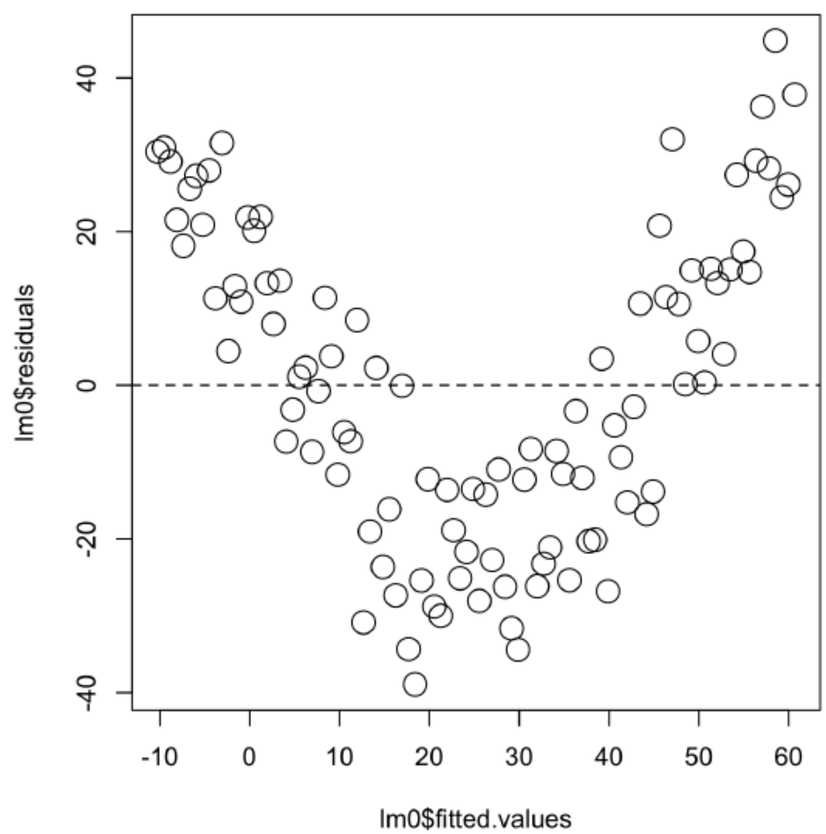
$$d_i = y_i - \hat{y}_i = y_i - a - bx_i$$

- 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

Residual Plots

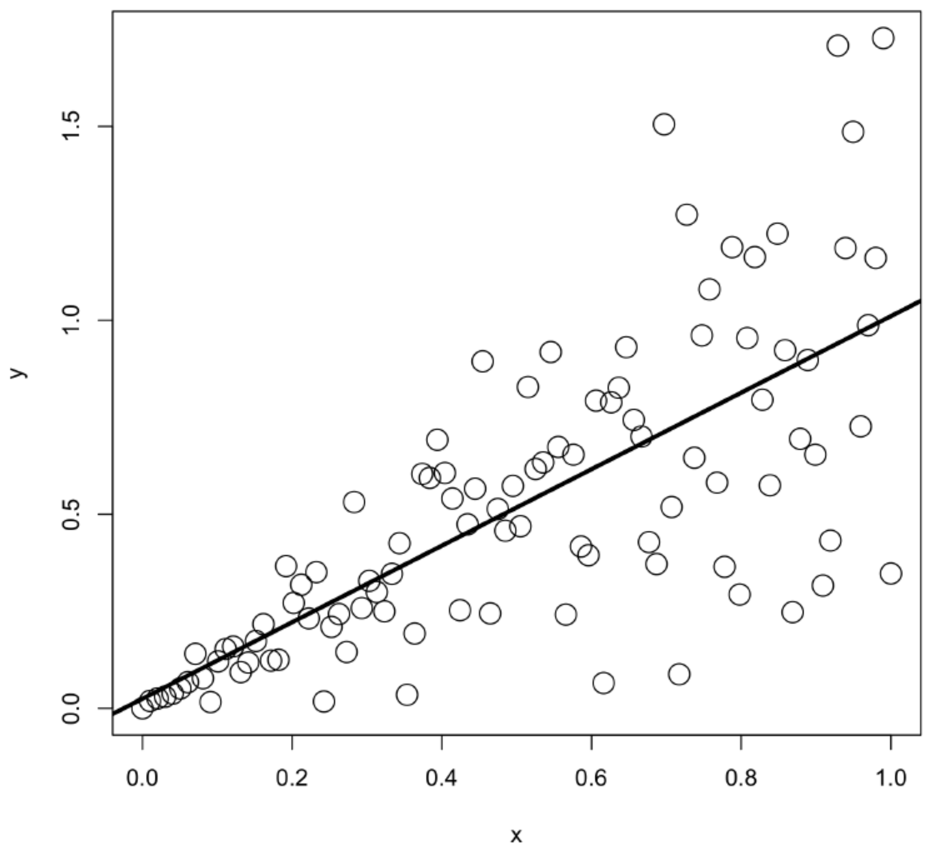
- Plot the predicted *y-values* on the *x-axis* and the residuals on the *y-axis*
- Are the residuals normally distributed with constant variance?

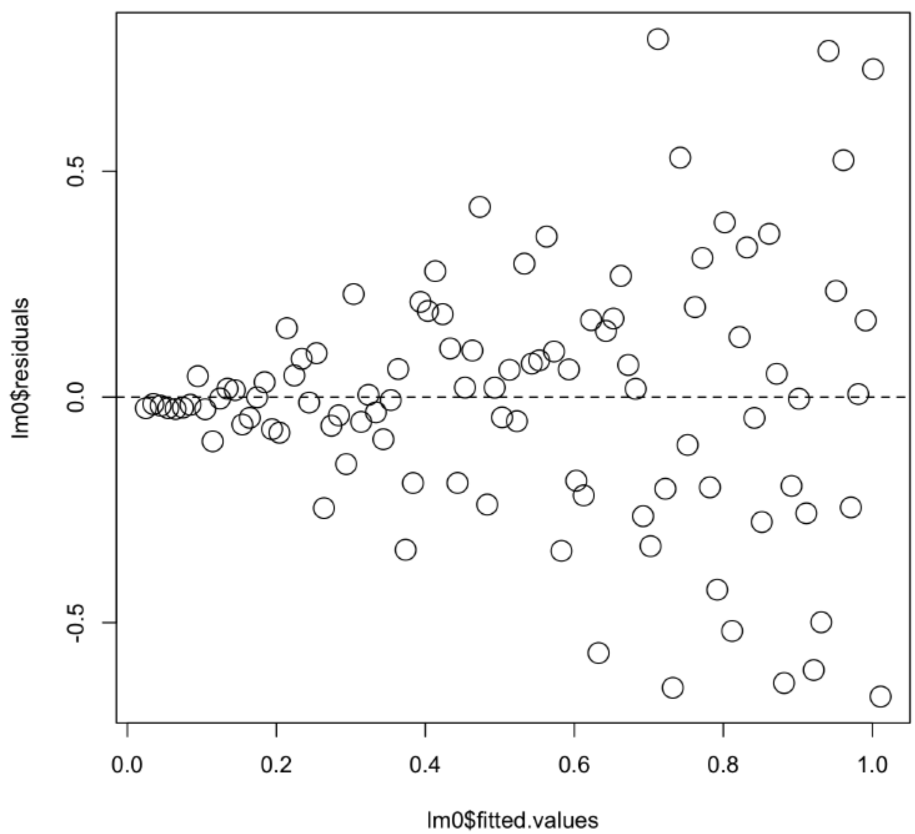




Residual Plots

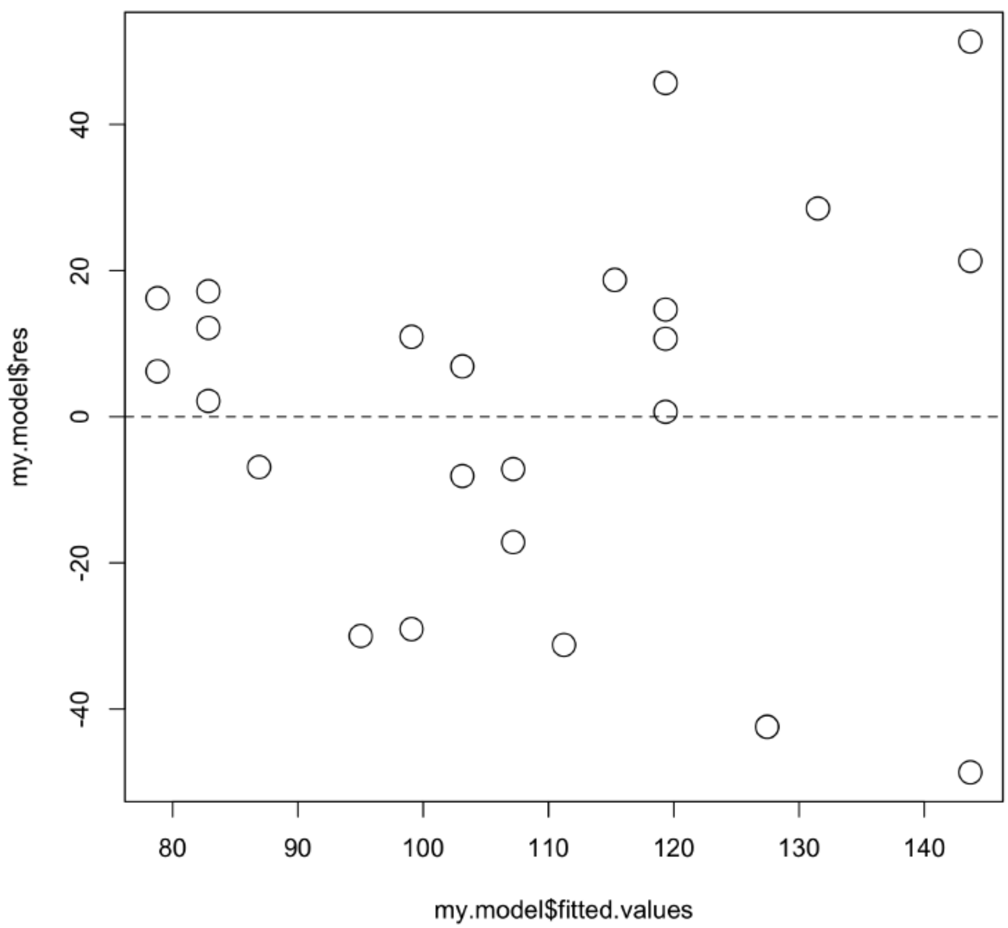
- Another example:





Example: Cystic Fibrosis Patients

- Does this model violate the assumption for constant variance?



Linear Regression

- Which models are 'linear'?
  - $y = a + bx$
  - $y = bx$
  - $y = a + b_1x_1 + b_2x_2$
  - $y = a + b\ x_1{}^2$s
  - $\log(y) = a + bx$
- In fact, linear regression is not so restrictive

Summary: Simple Linear Regression

- Linear model

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

- Method of Least Squares

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

- Testing for significance of coefficients

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

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

$$\boxed{\begin{gathered} y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e \\ e \sim N\left(0, \sigma^2\right) \end{gathered}}$$

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

Least Squares

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

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

- 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

$$\boxed{y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e}$$

```
> 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 β's $\neq 0$
  $H_1 : \beta_j = 0$ and all other β's $\neq 0$

- The test statistic

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

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

Adjusted R^2

```
> 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
- Age and height explained 38.3% of the variability in PEmax
- The inclusion of an additional variable in a regression model can never cause R^2 to decrease
- To get around this problem, we use the **adjusted R^2** to penalize for the added complexity of the model
- Here, adjusted R^2 decreased. We conclude that this model is not an improvement over the age-only model

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

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

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

F-test

- 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.

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

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

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