Lecture 3

Goodness of Fit (R^2), Residual Plots

STAT 432, Spring 2021

Simple Linear Regression (SLR) Model

Let $\{(x_i, y_i) : i = 1, \cdots, n\}$ be a collection of n data points. A **simple linear regression model** expressing the relationship between y_i and x_i is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i response variable
- x_i explanatory variable
- β_0 intercept parameter
- β_1 slope parameter
- ϵ_i is the random error term; assume $\epsilon_i \sim N(0, \sigma^2)$

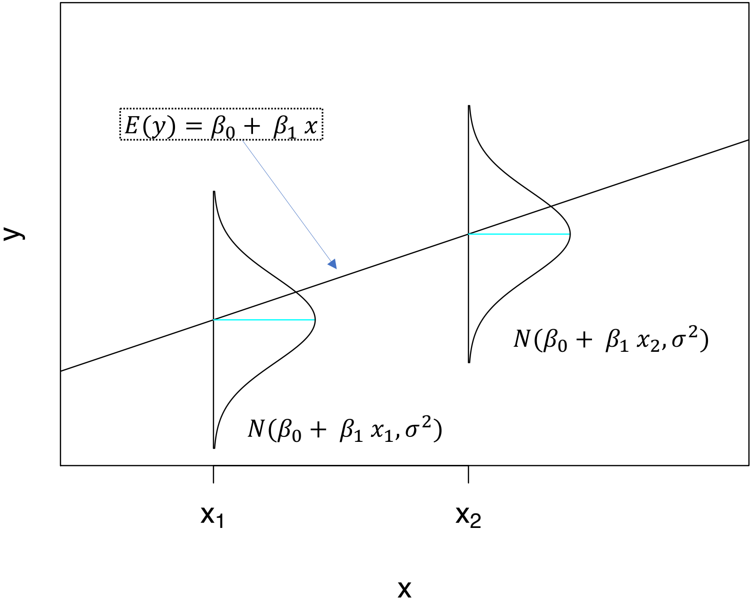
Expectation and Variance

What is $E(y_i)$?

Expectation and Variance

What is $Var(y_i)$?

What distribution does y_i follow?


$E(y) = \beta_0 + \beta_1 x$
$N(\beta_0 + \beta_1 x_2, \sigma^2)$
$N(\beta_0 + \beta_1 x_1, \sigma^2)$
y
x_1
x_2
x

Question

What is $\hat{e}_i$ and how is it different than ϵ_i?

Estimating σ^2

Estimate of $Var(\epsilon_i) = \sigma^2$:

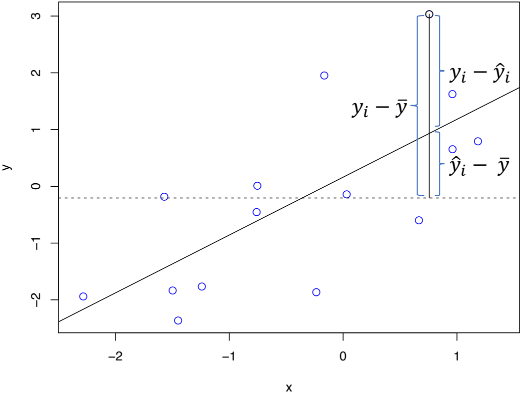
$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Remarks:

- $\sum_{i=1}^{n} \hat{e}_i = 0$
- $\hat{\sigma} = \sqrt{RSS/(n-2)}$ is called the **residual standard error**
- The divisor is $n-2$ since two parameters β_0 and β_1 were estimated

Partitioning Variability

Graphical description that $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$



Partitioning Variability

It can be shown that

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

$$\text{TSS} = \text{RSS} + \text{RegSS}$$

- TSS is the total sum of squares (total variability in the response variable)
- RSS is the residual sum of squares (unexplained variability)
- RegSS is the regression sum of squares (variability in the response explained by the model)

Coefficient of Determination

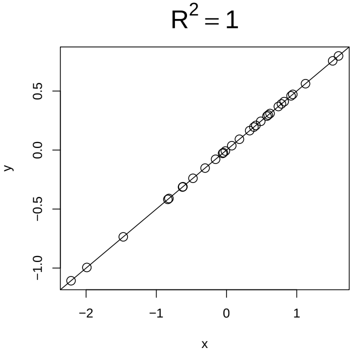
The **coefficient of determination** (R^2) is a measure of how well the linear regression model fits the data.

$$R^2 = \frac{\text{RegSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

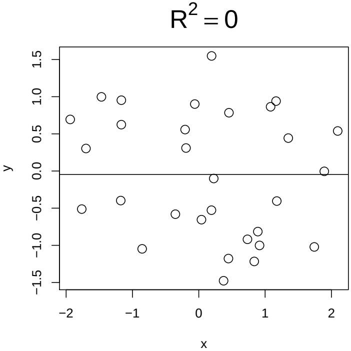
- R^2 can be interpreted as the proportion of variability in the response variable y that is explained by x (i.e., the regression model).
- $0 \leq R^2 \leq 1$; the closer R^2 is to 1, the better the linear regression model fits the data.
- R^2 can also be computed as the correlation coefficient r squared.

Limiting cases:

- $R^2 = 1$ when all points fall on the regression line (RSS=0)
- $R^2 = 0$ when $\hat{y} = \bar{y}$, which implies RSS=TSS.



$R^2 = 1$



$R^2 = 0$

Example

Recall our example of fitting a linear regression model for predicting weight (kg) from height (cm) using data collected from 247 physically active men.

```
> library(openintro)
> bdims_males <- subset(bdims, sex == 1)
> lm1 <- lm(wgt ~ hgt, data=bdims_males)
> summary(lm1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -60.95336   14.05436  -4.337 2.11e-05 ***
hgt           0.78257    0.07901   9.905  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.902 on 245 degrees of freedom
Multiple R-squared:  0.2859,Adjusted R-squared:  0.283
F-statistic: 98.11 on 1 and 245 DF,  p-value: < 2.2e-16
```

- Based on the summary output $R^2 = 0.2859$ (see `Multiple R-squared`). Therefore, 28.6% of the variability in weight can be explained by height.

- Alternatively, we can compute R^2 by taking the sample correlation (using the `cor()` function) and then squaring it.

  ```
  > cor(bdims_males$wgt, bdims_males$hgt)^2
  [1] 0.2859487
  ```

- The residual standard error is 8.902, which can also be found in the summary output.

Conditions for Simple Linear Regression

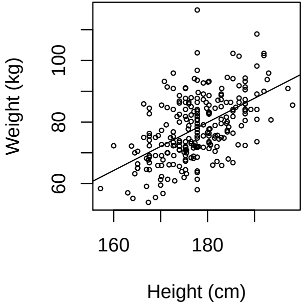
- **Linearity**. The data should follow a linear trend.
- **Constant variability**. The variability of the points around the least squares line remains roughly constant.
- **Normality**. The residuals should be approximately normally distributed with mean 0.
- **Independence**. Values of the response variable are independent of each other.

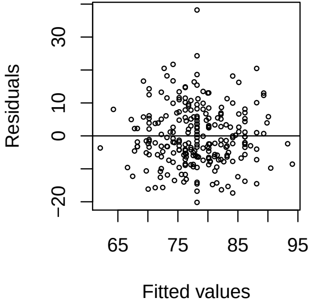
Residual Plots

- One useful way to check the conditions is to look at a plot of the residuals $\hat{e}_i = y_i - \hat{y}_i$ versus the fitted values $\hat{y}_i$, for $i = 1, \cdots, n$. It is also common to plot the residuals $\hat{e}_i$ versus the predictor x_i.
- One purpose of residual plots is to identify characteristics or patterns still apparent in the data after fitting the model.
- Residual plots are especially useful for checking linearity and constant variability.
- Ideally, the residual plot should show no obvious pattern, and the points are randomly scattered around 0.

Residual Plots

For the simple linear regression model between male height and weight, the points in the residual plot look randomly scattered and show no obvious patterns, indicating that the conditions are reasonably satisfied. Although, there is one potential outlier.





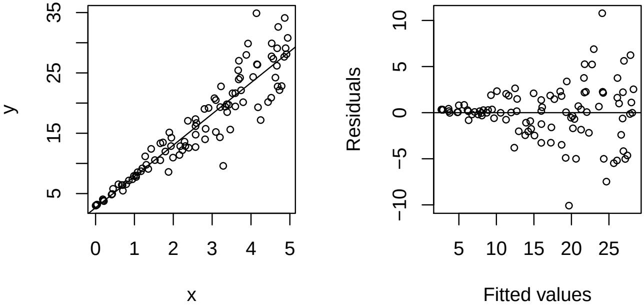
Here is the code used to create the last plot.

```
# scatter plot with least squares line
> plot(wgt ~ hgt, data=bdims_males,
       xlab = 'Height (cm)' , ylab = 'Weight (kg)')
> abline(lm1)

# residual plot
> plot(predict(lm1), resid(lm1), xlab='Fitted values', ylab='Residual')
> abline(h=0)
```

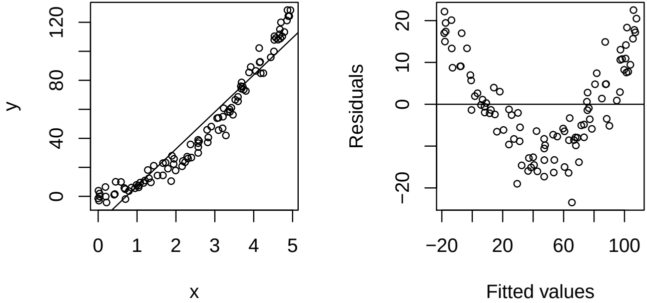
Residual Plots

An example of nonconstant variability, also called **heteroscedasticity**. The residual plot below shows a fan pattern.



Residual Plots

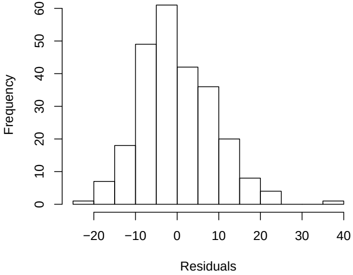
An example of nonlinearity.



Normality of Residuals

One way to check whether the residuals follow a normal distribution is to make a histogram. The histogram should be symmetric around 0, and have an approximate bell-curve shape. For the example below, the residuals look normally distributed, except for one potential outlier.

```
> hist(resid(lm1), main='', xlab='Residuals')
```


Frequency
Residuals