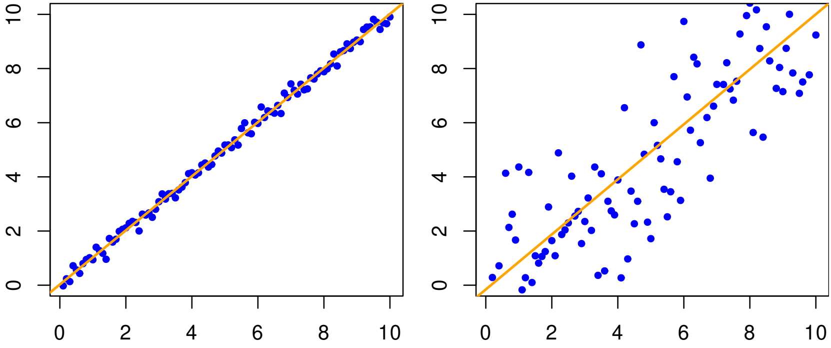
Abschätzung der Genauigkeit des Modells: R^2

- Nullhypothese verworfen: *In welchem Ausmass passt das Modell zu den Daten?*

- Abbildung:



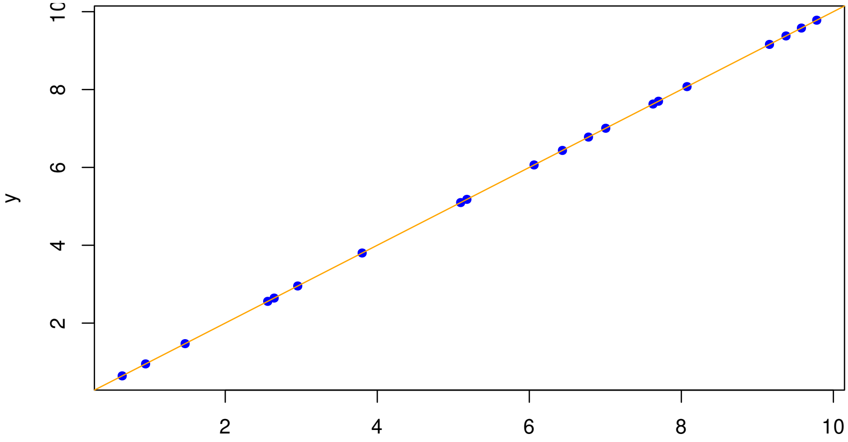
  - Links: Steigende Gerade passt sehr gut zu Punkten
  - Rechts: Steigende Gerade passt *nicht* gut zu Punkten

- Qualität einer linearen Regression abgeschätzt durch den *residual standard error* (RSE) und die R^2-Statistik
- R^2 wichtiger
- R^2-Statistik: Wert zwischen 0 und 1
- Sie gibt an, welcher Anteil der Variabilität in Y mit Hilfe des Modells durch X erklärt werden
- Wert nahe bei 1: ein grosser Anteil der Variabiliät wird durch die Regression erklärt. Das Modell beschreibt also die Daten sehr gut.
- Wert nahe bei 0: Regression erklärt die Variabilität der Zielvariablen nicht
- Graphische „Herleitung"

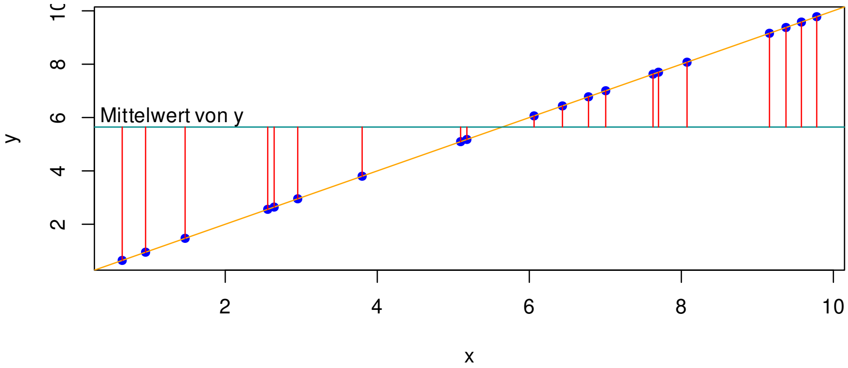
Punkte folgen linearem Modell

- Abbildung:

```
x <- runif(min = 0, max = 10, n = 20)
y <- x
plot(x, y, col = "blue", pch = 16)
abline(lm(y ~ x), col = "orange")
```



- Abbildung Varianz:



- Varianz: „Mittelwert" der quadrierten Unterschiede der y-Werte der Datenpunkte zu $\overline{y}$

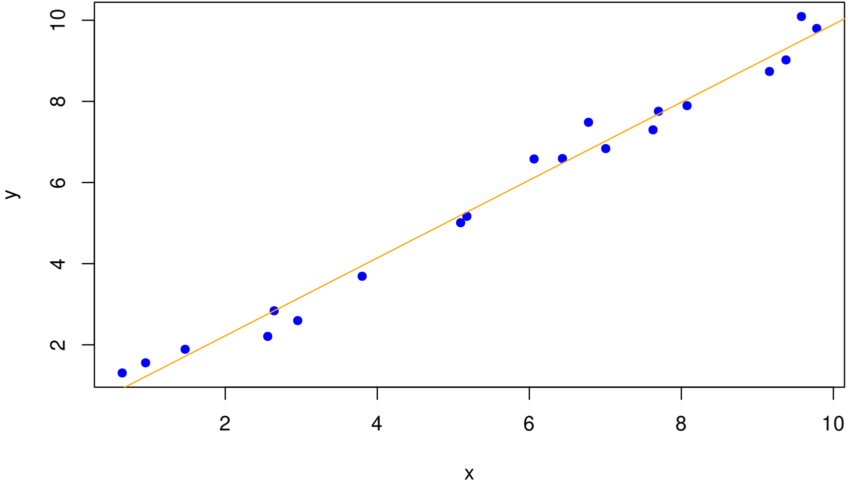
- Varianz:

```
var(y)
## [1] 8.998626
```

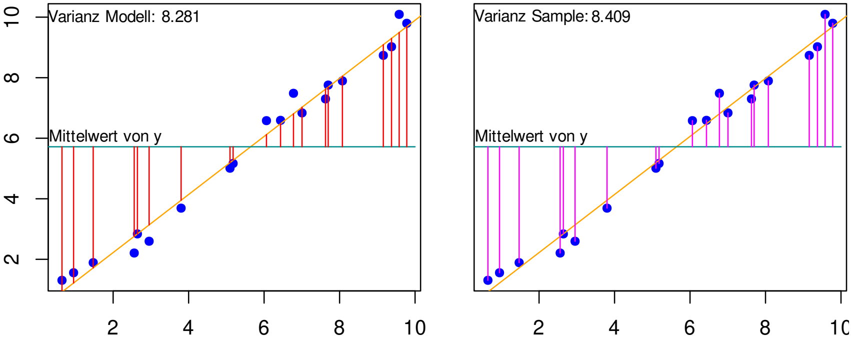
Punkte folgen mehr oder weniger linearem Modell

- Abbildung:

```
y <- x + rnorm(n = 20, mean = 0, sd = 0.3)
```



- Abbildung:



- „Durchschnitt" der Quadrate der roten Linien: Varianz des Modelles
- „Durchschnitt" der Quadrate der pinken Linien: Varianz der Daten

- Definition R^2:

$$R^2 = \frac{\text{Varianz Modell}}{\text{Varianz Sample}}$$

- Beispiel:

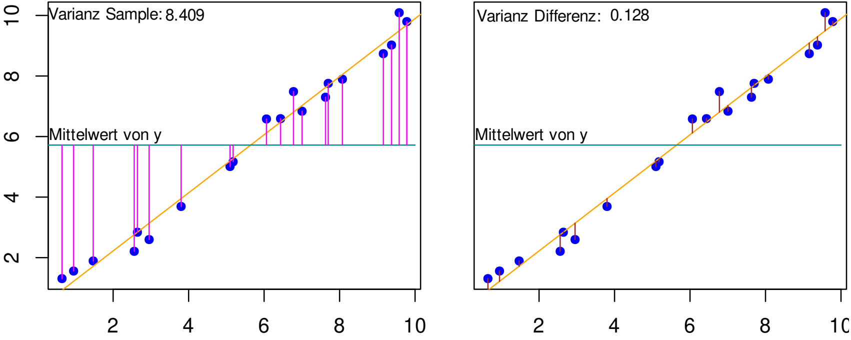
$$R^2 = \frac{8.281}{8.409} = 0.985$$

- Code:

```
summary(lm(y ~ x))$r.squared
## [1] 0.9848312
```

Andere Betrachtungsweise für Interpretation von R^2

- Abbildung:



- „Durchschnitt" der Quadrate der pinken Linien: Varianz der Daten
- „Durchschnitt" der Quadrate der brauen Linien (rechts): Varianz Unterschied zum Modell

- Alternative Definition von R^2:

$$R^2 = 1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$$

- Bedeutung:
  - Varianz Differenz:
    Varianz des Samples, dass *nicht* durch das Modell erklärt wird
  - $\frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$:
    Anteil der Varianz vom Sample, der *nicht* vom Modell erklärt wird
  - $1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$:
    Anteil der Varianz vom Sample, der vom Modell erklärt wird
  - R^2: Anteil der Varianz vom Sample, der vom Modell erklärt wird

- Output:
  - R^2:

```r
summary(lm(y ~ x))$r.squared
## [1] 0.9848312
```

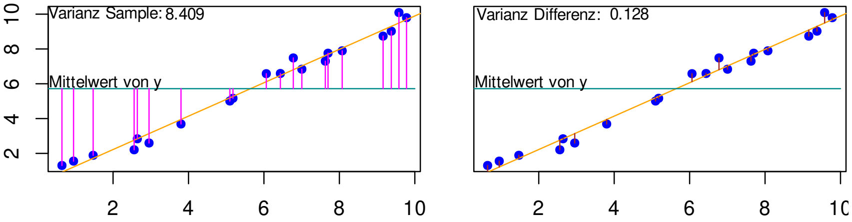
  - Varianz:

```r
var(y)
## [1] 8.40886
```

  - 98.48% der Varianz von 8.41 wird durch das Modell erklärt

Interpretation: Beispiel

- Nochmals Abbildung:

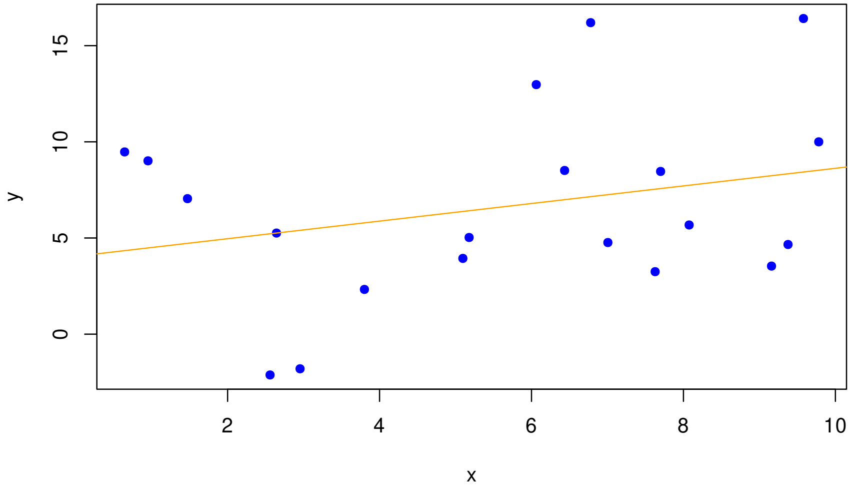


- Bedeutung:
  - Varianz Differenz: Hier nahe bei 0
  - $\frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Nahe bei 0
  - $1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Nahe bei 1
  - R^2: Passen die Punkte gut zum Modell, dann ist R^2 angenähert 1

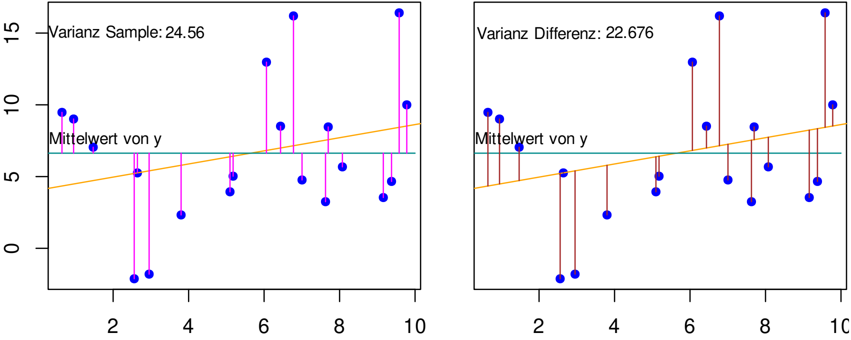
Punkte folgen dem linearen Modell nicht

- Abbildung:

```
y <- x + rnorm(n = 20, mean = 0, sd = 4)
```



- Varianzen:



- Berechnung R^2:

$$R^2 = 1 - \frac{22.676}{24.56} = 0.07671$$

- Output:

```
summary(lm(y ~ x))$r.squared
## [1] 0.07670148
```

- Bedeutung:
  - Varianz Differenz: Ähnlich Varianz Sample
  - $\frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Nahe bei 1
  - $1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Nahe bei 0
  - R^2: Passen die Punkte nicht gut zum Modell, dann ist R^2 angenähert 0
- Varianz Sample:

```
var(y)
## [1] 24.55976
```

- Interpretation: Nur 7.67% der Varianz von 24.56 wird durch das Modell erklärt
- Punkte passen nicht gut zum Modell

- Output:
  - Korrelation:

```
cor(x, y)
## [1] 0.2769503
```

  - R^2:

```
summary(lm(y ~ x))$r.squared
## [1] 0.07670148
```

  - Varianz:

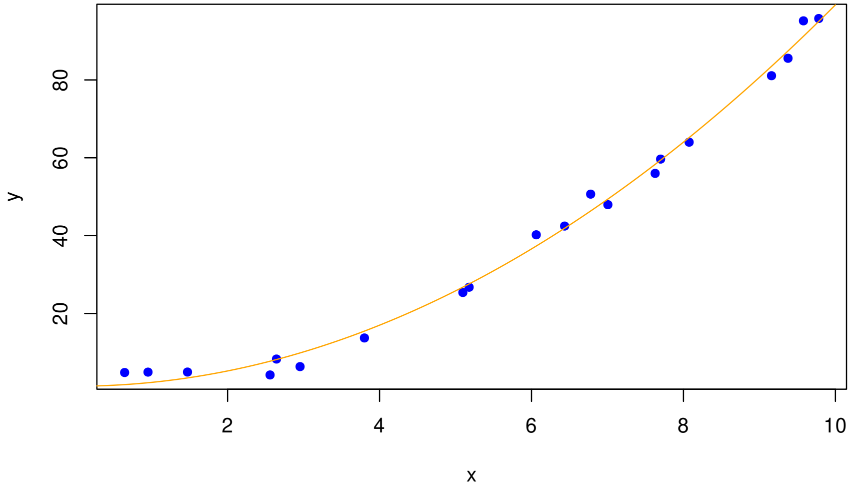
```
var(y)
## [1] 24.55976
```

  - 7.67% der Varianz von 24.56 wird durch das Modell erklärt

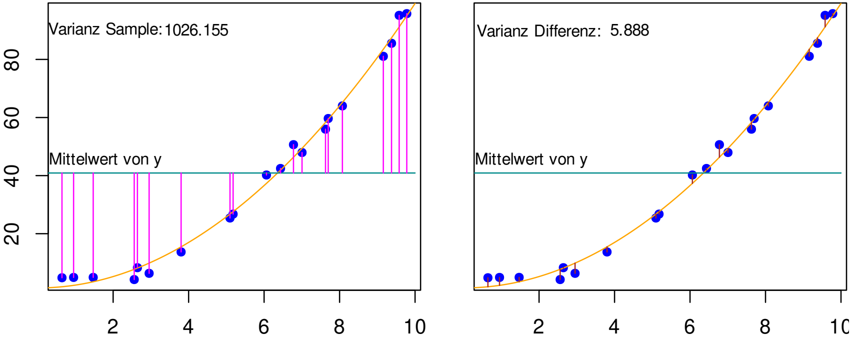
Quadratisches Modell

- Punkte folgen mehr oder weniger dem Modell:

```
y <- x^2 + rnorm(n = 20, mean = 0, sd = 2)
```



- Varianzen:



- Berechnung R^2:

$$R^2 = 1 - \frac{5.888}{1026.155} = 0.994262$$

- Output:
  - R^2:

```r
summary(lm(y ~ I(x^2)))$r.squared
## [1] 0.9942619
```

  - Varianz:

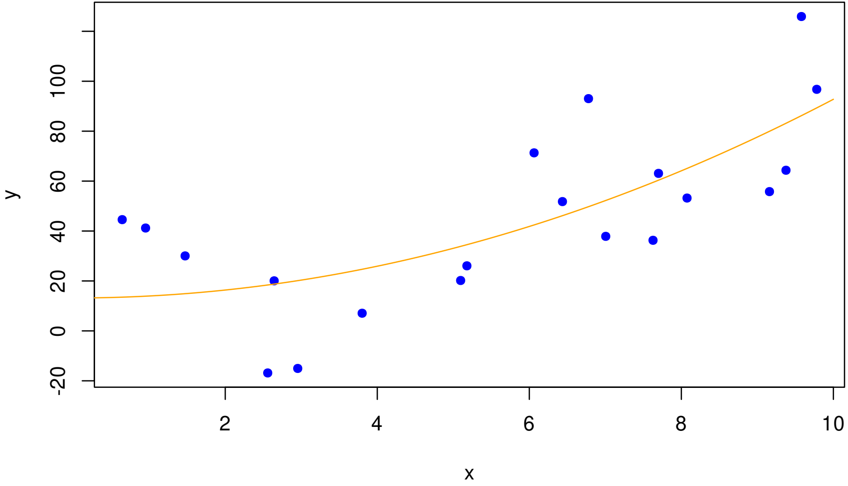
```r
var(y)
## [1] 1026.155
```

  - 99.43% der Varianz von 1026.15 wird durch das Modell erklärt
  - R^2-Wert nahe bei 1 und somit passen die Daten gut zum Modell

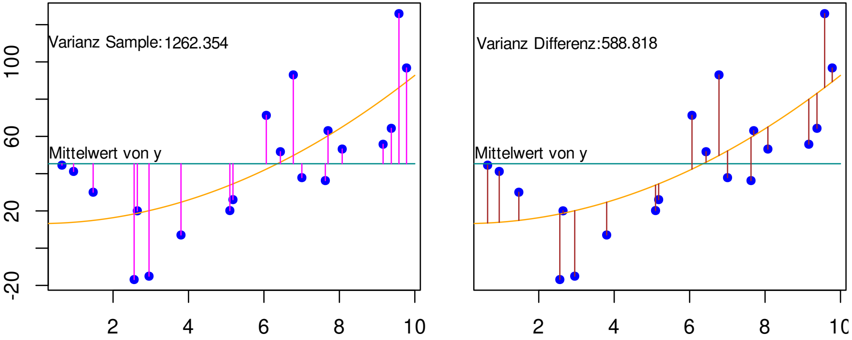
Quadratisches Modell

- Punkte folgen dem Modell nicht gut:

```
y <- x^2 + rnorm(n = 20, mean = 0, sd = 20)
```



- Varianzen:



- Berechnung R^2:

$$R^2 = 1 - \frac{588.818}{1262.354} = 0.533556$$

- Output:
  - R^2:

```r
summary(lm(y ~ I(x^2)))$r.squared
## [1] 0.5335559
```

  - Varianz:

```r
var(y)
## [1] 1262.354
```

  - 53.36% der Varianz von 1262.35 wird durch das Modell erklärt
  - R^2-Wert nicht so nahe bei 1 und somit passen die Daten nicht so gut zum Modell

Bemerkungen:

- Empirische Korrelation gibt nur die Güte einer *linearen* Regression an
- R^2 kann für jede Regression angewendet werden