Residuals and residual variation

Brian Caffo, Jeff Leek and Roger Peng
Johns Hopkins Bloomberg School of Public Health

Residuals

die beta0 und beta1 sind die "richtigen" Parameter fuer das Model.

Die haben wir aber nicht. Wir haben nur die estimated, also beta^0, beta^1.

- Model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
- Observed outcome i is Y_i at predictor value X_i
- Predicted outcome i is $\hat{Y}_i$ at predictor valuve X_i is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

der Hut (^) heisst einfach 'estimated', also estimated interceptor, estimated slope.

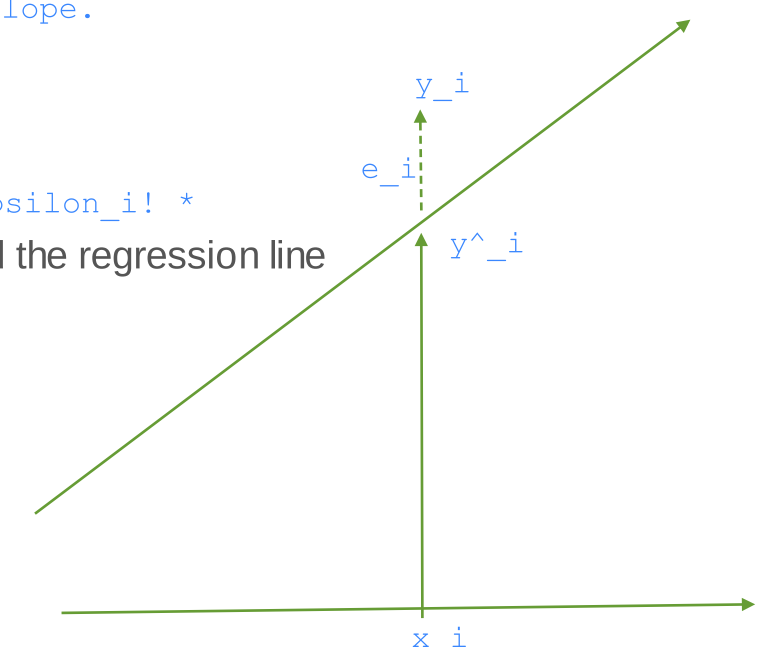
- Residual, the between the observed and predicted outcome

$$e_i = Y_i - \hat{Y}_i$$

Achtung: e_i ist /nicht/ epsilon_i! *

  - The vertical distance between the observed data point and the regression line

- Least squares minimizes $\sum_{i=1}^n e_i^2$
- The e_i can be thought of as estimates of the ϵ_i.



* epsilon_i ist der "richtige" Fehler,
e_i die beobachtete Distanz zwischen
unserer Vorhersage Y^_i und dem richtigen Y_i.

Properties of the residuals

- $E[e_i] = 0$.

  dh die e_is koennen nicht unabhaengig voneinander sein, da sie sich zu 0 aufsummieren. Wenn man n-1 eins kennt, kann man das nte ausrechnen.

- If an intercept is included, $\sum_{i=1}^{n} e_i = 0$

- If a regressor variable, X_i, is included in the model $\sum_{i=1}^{n} e_i X_i = 0$.

  Falls X=1, sei diese 3. Aussage equivalent zur zweiten, sagt er.
  X, X_i, x, x_i ??

- Residuals are useful for investigating poor model fit.
- Positive residuals are above the line, negative residuals are below.
- Residuals can be thought of as the outcome (Y) with the linear association of the predictor (X) removed. = "Y adjusted for X"
- One differentiates residual variation (variation after removing the predictor) from systematic variation (variation explained by the regression model).
- Residual plots highlight poor model fit.

Code

```
data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
e <- resid(fit)
yhat <- predict(fit)
max(abs(e -(y - yhat)))
```

```
[1] 9.486e-13
```

residuals = predicted - observed

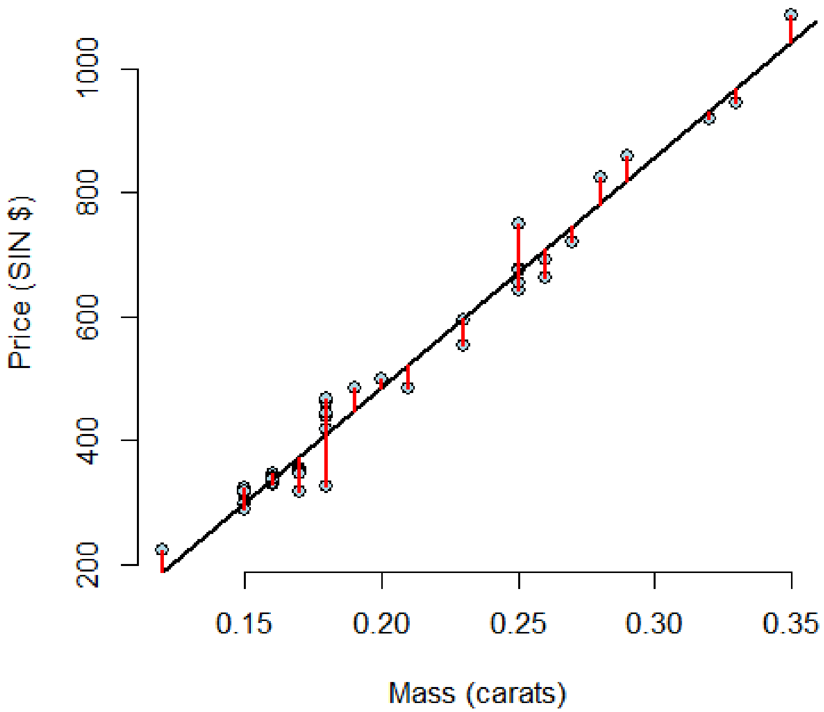
beta^0 beta^1

```
max(abs(e - (y - coef(fit)[1] - coef(fit)[2] * x)))
```

dasselbe:

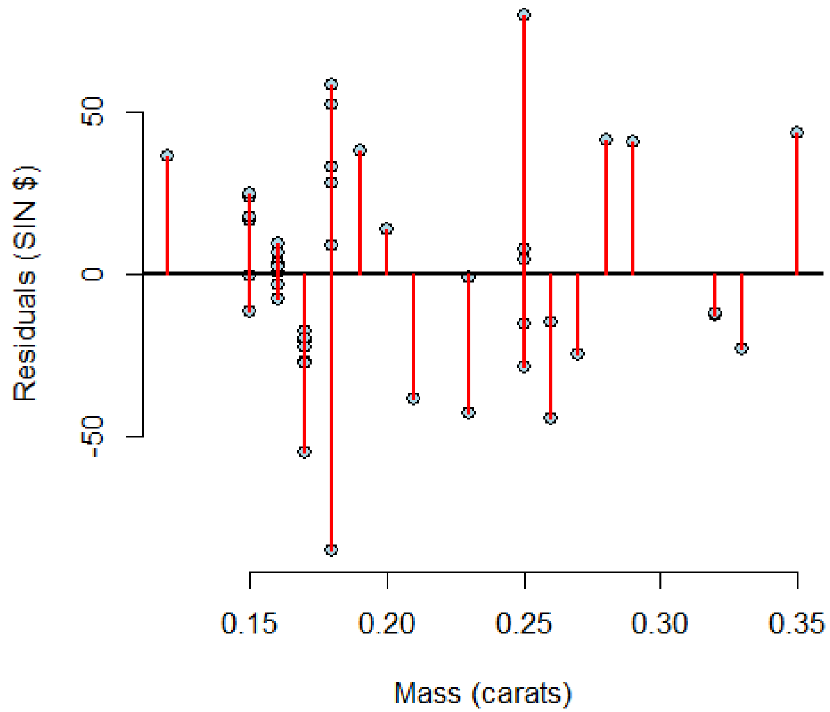
```
[1] 9.486e-13
```

Residuals are the signed length of the red lines



Residuals versus X

Wichtig: die Residuals sollten keinen Zusammenhang haben mit dem Predictor. Keine Muster.

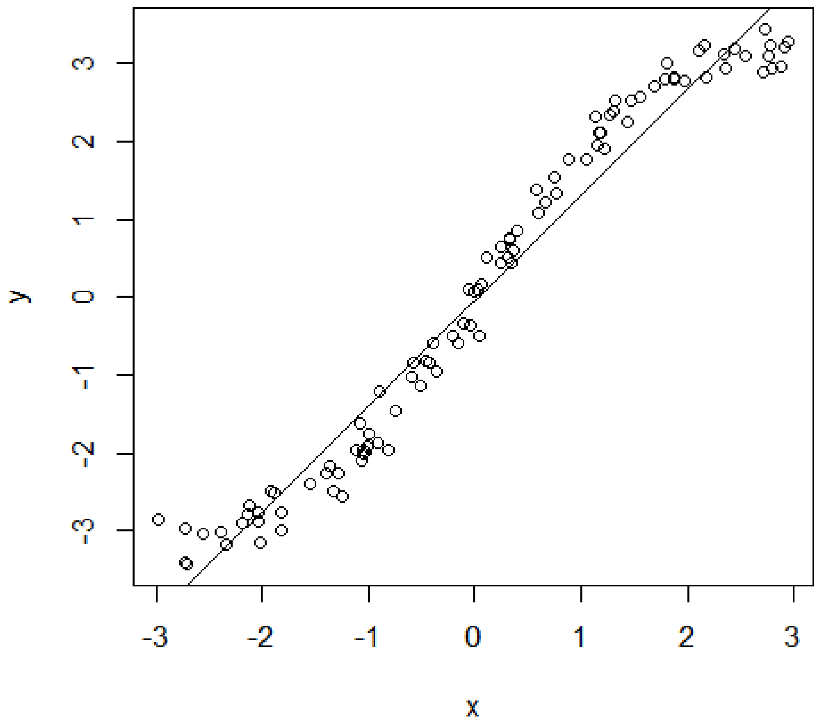


Non-linear data

Daten simulieren: X ist random uniform numbers. Y ist X plus eine Sinusfkt plus etwas Zufall:

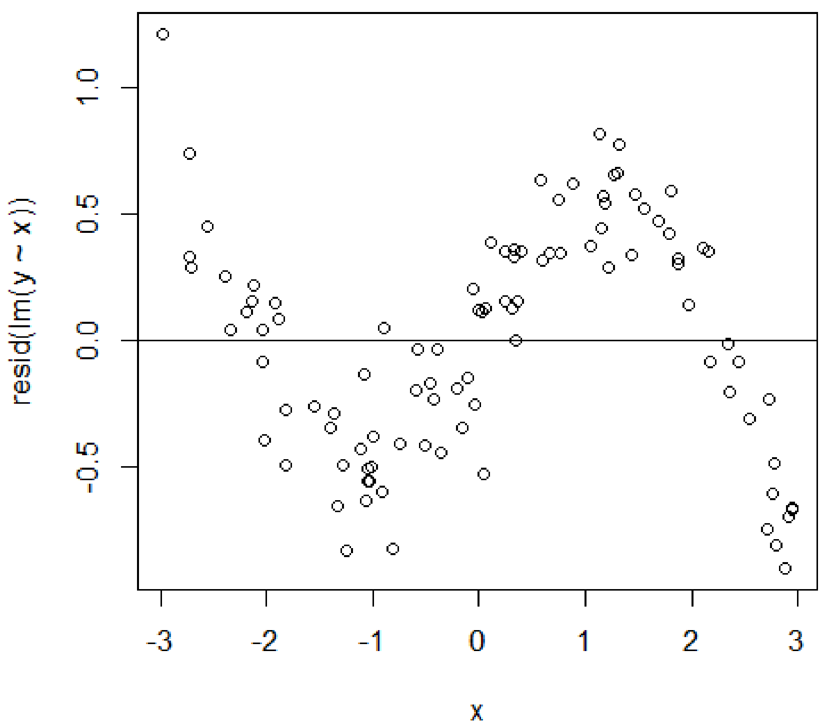
```
x <- runif(100, -3, 3); y <- x + sin(x) + rnorm(100, sd = .2);
plot(x, y); abline(lm(y ~ x))
```

Und diese Fkt (den Synus) sieht man im Plot:



Aber noch viel deutlicher im Residualplot!

```
plot(x, resid(lm(y ~ x)));
abline(h = 0)
```

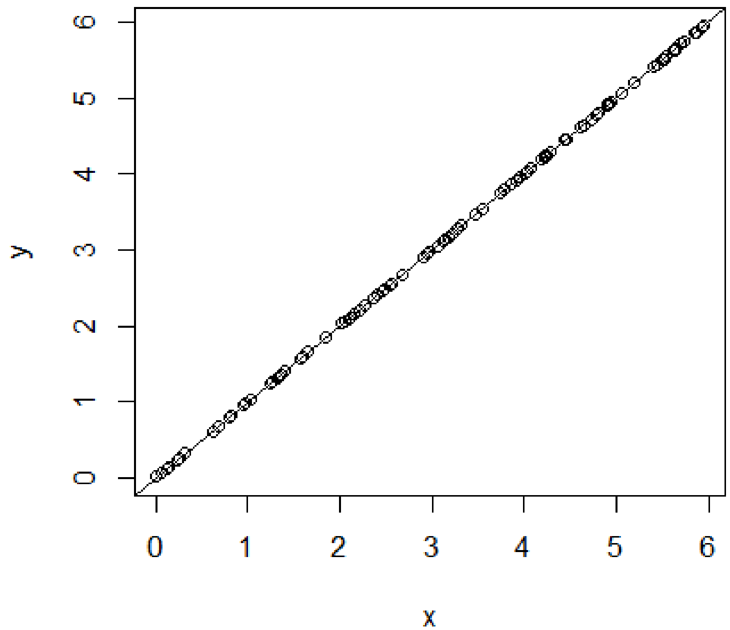


= Varianz nicht ueberall gleich!

Heteroskedasticity

Simulierte Daten: Varianz abhaengig von x!

```
x <- runif(100, 0, 6); y <- x + rnorm(100,  mean = 0, sd = .001 * x);
plot(x, y); abline(lm(y ~ x))
```

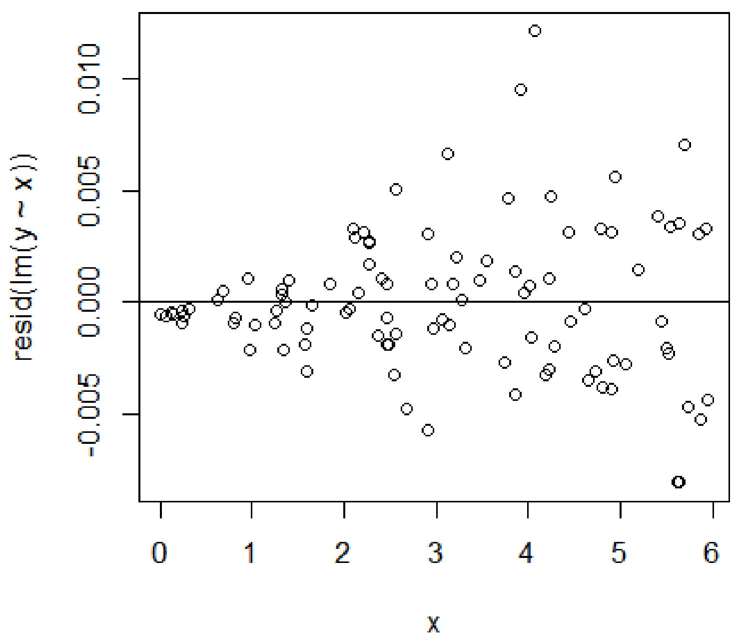


sieht eigentlich
aus
wie die perfekte

…jedoch im Residualplot sieht man, dass die Residuen nicht zufaellig sind!

Getting rid of the blank space can be helpful

```
plot(x, resid(lm(y ~ x)));
abline(h = 0)
```



Leider ist so ein Residualplot nur mit 1 Variable moeglich :-(

Estimating residual variation

- Model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.

Residual variation ist sigma^2.

- The ML estimate of σ^2 is $\frac{1}{n}\sum_{i=1}^n e_i^2$, the average squared residual.

Max.Likelihood

- Most people use

$$\hat\sigma^2 = \frac{1}{n-2}\sum_{i=1}^n e_i^2.$$

-2: zwei Freiheitsgrade weniger:
eins fuer estimate of beta0,
eins fuer estimate of beta1

- The $n-2$ instead of n is so that $E[\hat\sigma^2] = \sigma^2$

-> unbiased

between friends: the difference is rather small, so forget about it…

Diamond example

```
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
summary(fit)$sigma
```

```
[1] 31.84
```

```
sqrt(sum(resid(fit)^2) / (n - 2))
```

```
[1] 31.84
```

…same

Summarizing variation

Trick: Y^_i addieren und subtrahieren

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

dann Quadrat ausmultiplizieren

$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\boxed{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})} + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Scratch work

Y^_i = beta^0 + beta^1*X_i
und beta^0 = Y_ + beta^1*X_. Einsetzen:

Hier immer benuetzt:
sum(X_i - X_) = 0,
sum(Y_i - Y_) = 0

$$(Y_i - \hat{Y}_i) = \{Y_i - (\bar{Y} - \hat{\beta}_1\bar{X}) - \hat{\beta}_1 X_i\} = (Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})$$

$$(\hat{Y}_i - \bar{Y}) = (\bar{Y} - \hat{\beta}_1\bar{X} - \hat{\beta}_1 X_i - \bar{Y}) = \hat{\beta}_1(X_i - \bar{X})$$

derselbe Trick bei Y^_i - Y_

$$\textstyle\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n}\{(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X}))\}\{\hat{\beta}_1(X_i - \bar{X})\}$$

Nun diese beiden vereinfachten Terme hier einsetzen.

$$= \hat{\beta}_1 \textstyle\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= \hat{\beta}_1^2 \textstyle\sum_{i=1}^{n}(X_i - \bar{X})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 = 0$$

zweimal dasselbe, voneinander subtrahiert - gibt 0!

Summarizing variation

$$\boxed{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \boxed{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}$$

Totale Var. = Residual Var + "Model-Variation"

Or

Total Variation = Residual Variation + Regression Variation

weil mittlerer Term oben 0 wird!
(siehe Folie 13)

Define the percent of total varation described by the model as

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Model-Variation ist immer kleiner als die totale Variation.

Relation between R^2 and r (the corrrelation)

Recall that $(\hat{Y}_i - \bar{Y}) = \hat{\beta}_1 (X_i - \bar{X})$ so that

2 slides ago

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \frac{\sum_{i=1}^{n} (X_i - \bar{X})}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \mathrm{Cor}(Y, X)^2$$

Since, recall,

$$\hat{\beta}_1 = \mathrm{Cor}(Y, X) \frac{\mathrm{Sd}(Y)}{\mathrm{Sd}(X)}$$

So, R^2 is literally r squared.

Some facts about R^2

- R^2 is the percentage of variation explained by the regression model.
- $0 \leq R^2 \leq 1$
- R^2 is the sample correlation squared.
- R^2 can be a misleading summary of model fit.
  - Deleting data can inflate R^2.
  - (For later.) Adding terms to a regression model always increases R^2.
- Do `example(anscombe)` to see the following data. `illustrates some issues with R^2`
  - Basically same mean and variance of X and Y.
  - Identical correlations (hence same R^2).
  - Same linear regression relationship.

data(anscombe);example(anscombe)

Die haben alle
dasselbe R^2!