Course 02402 Introduction to Statistics Lecture 8:

Simple linear regression

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

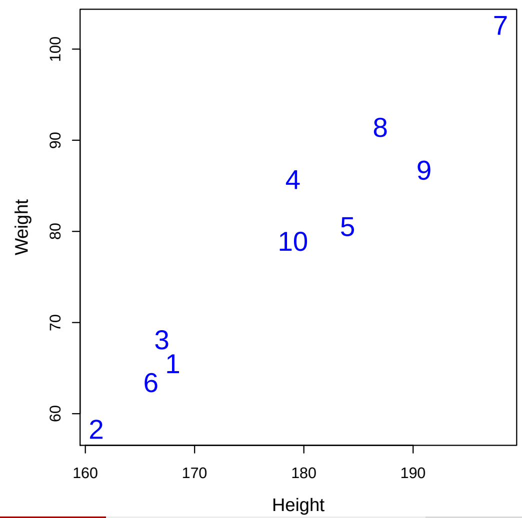
1. Example: Height-Weight
2. Linear regression model
3. Least Squares Method
4. Statistics and linear regression??
5. Hypothesis tests and confidence intervals for β_0 and β_1
6. Confidence and prediction interval for the line
7. Summary of summary(lm(y∼x))
8. Correlation
9. Residual Analysis: Model control

Oversigt

1. Example: Height-Weight
2. Linear regression model
3. Least Squares Method
4. Statistics and linear regression??
5. Hypothesis tests and confidence intervals for β_0 and β_1
6. Confidence and prediction interval for the line
7. Summary of summary(lm(y∼x))
8. Correlation
9. Residual Analysis: Model control

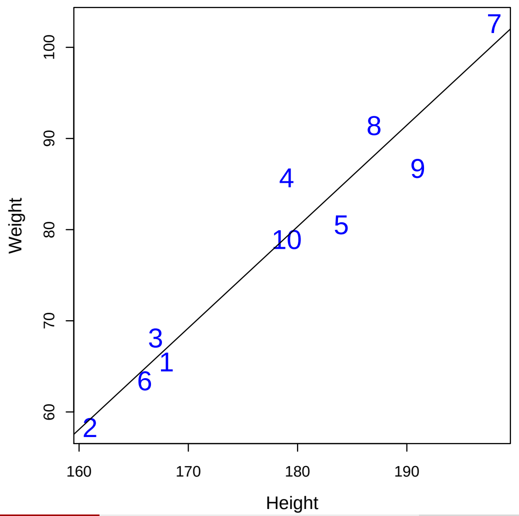
Example: Height-Weight

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Example: Height-Weight

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

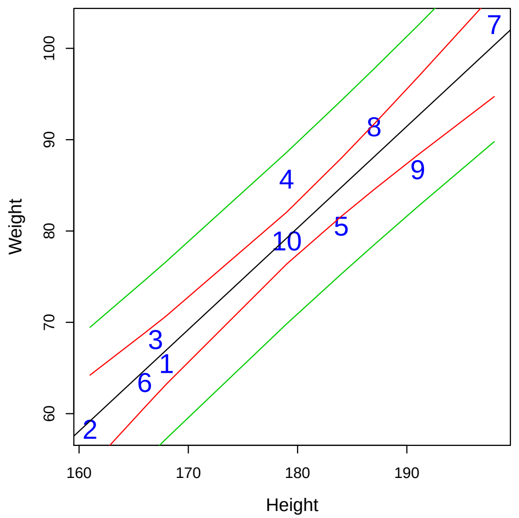


Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

```
summary(lm(y ~ x))

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.876 -1.451 -0.608  2.234  6.477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -119.958     18.897   -6.35  0.00022 ***
## x              1.113      0.106   10.50  5.9e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.9 on 8 degrees of freedom
## Multiple R-squared:  0.932,Adjusted R-squared:  0.924
## F-statistic:  110 on 1 and 8 DF,  p-value: 5.87e-06
```

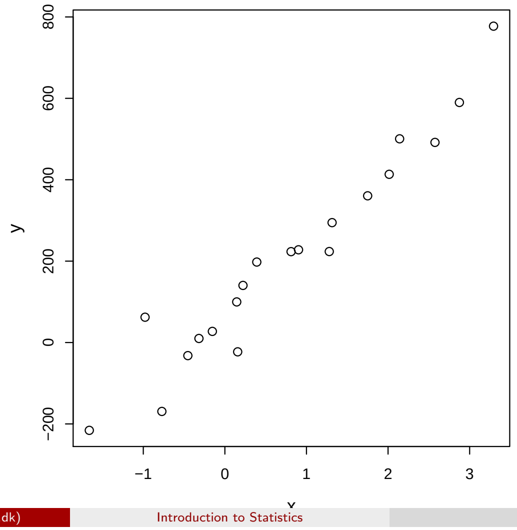

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Oversigt

A scatter plot of some data

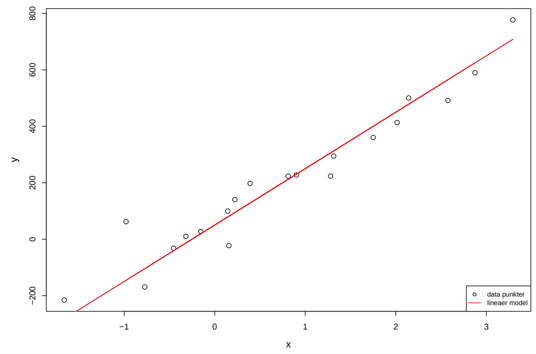
- We have n pairs of data points (x_i, y_i)



Express a linear model

- Express a linear model

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



but something is missing in the desrcription of the *random variation*!

Express a linear regression model

- Express the *linear regression model*

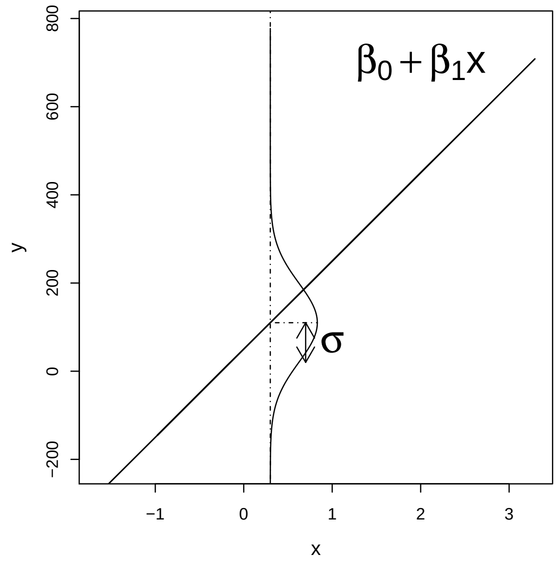
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Y_i is the *dependent variabel*. A random variable.
- x_i er en *explanatory variable*. Given numbers.
- ε_i is the deviation (error). A random variable.

and we assume

ε_i is independent and identically distributed (i.i.d.) and $N(0, \sigma^2)$

Model illustration



Oversigt

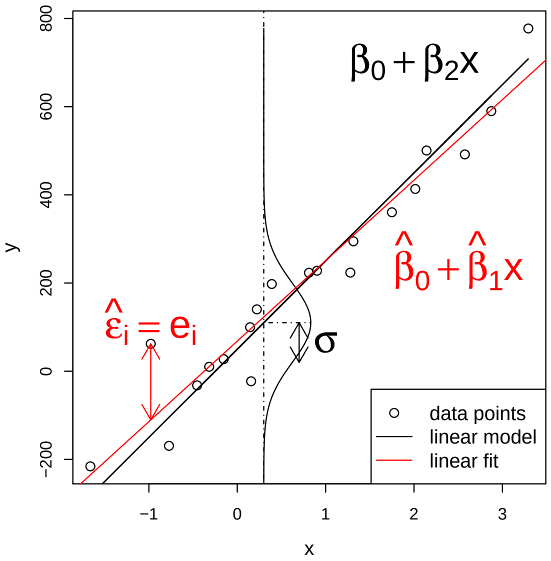
Least Squares Method

- How can we estimate the parameters β_0 and β_1?
- Good idea: Minimize the variance σ^2 of the residuals. It is in almost any way the best choice in this setup.
- But how!?
- Minimize the sum of the Residual Sum of Squares (RSS))

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ minimizes RSS

Illustration of model, data and fit



Least squares estimator

Theorem 5.4 (here as estimators as in the book)

The least squares estimators of β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Least squares estimates

Theorem 5.4 (here as estimates)

The least squares estimatates of β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Don't think too much about this for now!

R example

```
 ## Simulate a  linear model with  normally distrubuted
 ## errors and estimate the parameters

## FiRST MAKE DATA:
## Generates x
x <- runif(n=20, min=-2, max=4)
## Simulate y
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

## FROM HERE: as  real data analysis, we have th data in x and y:
## A scatter plot of x and y
plot(x, y)

## Find the  least squares estimates, use Theorem 5.4
(beta1hat <- sum( (y-mean(y))*(x-mean(x)) ) / sum( (x-mean(x))^2 ))
(beta0hat <- mean(y) - beta1hat*mean(x))

## Use lm() to find the estimates
lm(y ~ x)

## Plot the fitted  line
abline(lm(y ~ x), col="red")
```


Oversigt

The parameter estimates are random variables

What if we took a new sample?

Would the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ be the same?

No, they are random variables!

If we took a new sample we would get another realisation.

What is the (sampling) distribution of the parameter estimators?

in a linear regression model (given normal distributed errors)?

Try to simulate to have a look at this...

Let's go to R!!

- What is the (sampling) distribution of the parameter estimates in a linear regression model (given normal distributed errors)?
- Answer: They are normally distributed (for $n < 30$ use the t-distribution) and their variance can be estimated:

Theorem 5.7 (first part)

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}$$
$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$$
$$Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{S_{xx}}$$

- The Covariance $Cov[\hat{\beta}_0, \hat{\beta}_1]$ we do not use for anything for now..

Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$

Theorem 5.7 (second part)

Where σ^2 is usually replaced by its estimate ($\hat{\sigma}^2$). The central estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of σ^2 is used the variances also become estimates and we'll refer to them as $\hat{\sigma}^2_{\beta_0}$ and $\hat{\sigma}^2_{\beta_1}$.

Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ (equations 5-41 and 5-42)

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}; \quad \hat{\sigma}_{\beta_1} = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Oversigt

Hypothesis tests for the parameters

- We can carry out hypothesis tests for the parameters in a linear regression model:

$$H_{0,i}: \quad \beta_i = \beta_{0,i}$$
$$H_{1,i}: \quad \beta_i \neq \beta_{1,i}$$

- We use the t-distributed statistics:

Theorem 5.11

Under the null-hypothesis ($\beta_0 = \beta_{0,0}$ and $\beta_1 = \beta_{0,1}$) the statistics

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}; \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

are t-distributed with $n-2$ degrees of freedom, and inference should be based on this distribution.

- See Example 5.12 for example of hypothesis test.
- Test if the parameters are signifikantly different from 0

$$H_{0,i}: \quad \beta_i = 0$$
$$H_{1,i}: \quad \beta_i \neq 0$$

- See the resultats in R

```
## Hypothesis tests om signifikante parametre

## Generate x
x <- runif(n=20, min=-2, max=4)
## Simulate Y
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

## Use lm() to find the estimates
fit <- lm(y ~ x)

## See summary - what we need
summary(fit)
```


Confidence intervals for the parameters

Method 5.14

$(1-\alpha)$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t_{1-\alpha/2}\,\hat{\sigma}_{\beta_0}$$
$$\hat{\beta}_1 \pm t_{1-\alpha/2}\,\hat{\sigma}_{\beta_1}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a t-distribution with $n-2$ degrees of freedom.

- remember that $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ are found from equations 5-41 and 5-42
- in R we can read off $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ udner "Std. Error" from "summary(fit)"

Simulation illustration of CIs

```
## Make confidence intervals for the parameters

## number of repeats
nRepeat <- 100

## Did we catch the correct parameter
TrueValInCI <- logical(nRepeat)

## Repeat the simulation and estimation nRepeat times:
for(i in 1:nRepeat){
  ## Generate x
  x <- runif(n=20, min=-2, max=4)
  ## Simulate y
  beta0=50; beta1=200; sigma=90
  y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

  ## Use lm() to find the estimates
  fit <- lm(y ~ x)

  ## Luckily  R can compute the  confidence interval (level=1-alpha)
  (ci <- confint(fit, "(Intercept)", level=0.95))

  ## Was the correct parameter value "caught" by the interval? (covered)
  (TrueValInCI[i] <-  ci[1] < beta0  &  beta0 < ci[2])
}

## How often did this happen?
sum(TrueValInCI) / nRepeat
```


Oversigt

Method 5.17 Confidence interval for $\beta_0 + \beta_1 x_0$

- The confidence interval for $\beta_0 + \beta_1 x_0$ corresponds to a confidence interval for the line in the point x_0
- Is computed by

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- The confidence interval will in $100(1-\alpha)\%$ of the times contain the correct line, that is $\beta_0 + \beta_1 x_0$

Method 5.17 Prediction interval for $\beta_0 + \beta_1 x_0 + \varepsilon_0$

- The prediction interval for Y_0 is found using a value x_0
- This is done *before* Y_0 is observed with

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- The prediction intervallet will in $100(1-\alpha)\%$ of the times contain the observered y_0
- A prediction intervalis wider than a confidence interval for a given α

Example of confidence interval for the line

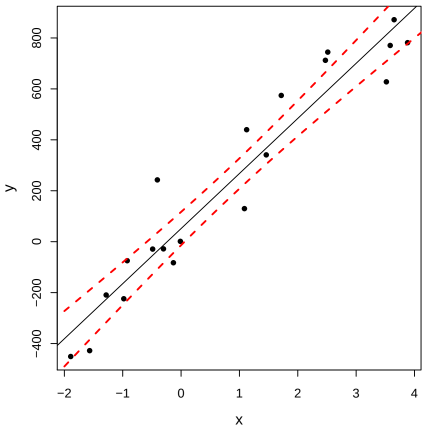
```r
## Example of confidence interval for the line

## Make a sequence of x values
xval <- seq(from=-2, to=6, length.out=100)

## Use the  predict function
CI <- predict(fit, newdata=data.frame(x=xval),
interval="confidence",
level=.95)

## Check what we got
head(CI)

## Plot the data, model fit and intervals
plot(x, y, pch=20)
abline(fit)
lines(xval, CI[, "lwr"], lty=2, col="red", lwd=2)
lines(xval, CI[, "upr"], lty=2, col="red", lwd=2)
```



Example of prediction interval

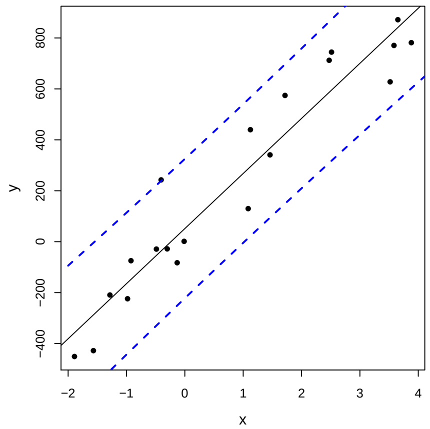
```r
## Example with prediction interval

## Make a sequence of x values
xval <- seq(from=-2, to=6, length.out=100)

## Use the  predict function
PI <- predict(fit, newdata=data.frame(x=xval),
interval="prediction",
level=.95)

## Check what we got
head(CI)

## Plot the data, model fit and intervals
plot(x, y, pch=20)
abline(fit)
lines(xval, PI[, "lwr"], lty=2, col="blue", lwd=2)
lines(xval, PI[, "upr"], lty=2, col="blue", lwd=2)
```



Oversigt

What more do we get from summary?

```
summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -184.7  -96.4  -20.3   86.6  279.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     51.5       31.1    1.66     0.12
## x              216.3       15.2   14.22  3.1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 126 on 18 degrees of freedom
## Multiple R-squared:  0.918,Adjusted R-squared:  0.914
## F-statistic:  202 on 1 and 18 DF,  p-value: 3.14e-11
```


summary(lm(y~x)) wrap up

- `Residuals:      Min       1Q    Median       3Q       Max:`
  The residuals': Minimum, 1. quartile, Median, 3. quartile, Maximum
- `Coefficients:`
  `Estimate Std. Error t value Pr(>|t|) "stars"`
  The coefficients':
  Estimate $\hat{\sigma}_{\beta_i}$ t_{obs} p-value
  - The test is $H_{0,i}: \beta_i = 0$ vs. $H_{1,i}: \beta_i \neq 0$
  - The stars is showing the size categories of the p-value
- `Residual standard error: XXX on XXX degrees of freedom`
  $\varepsilon_i \sim N(0, \sigma^2)$ printed is $\hat{\sigma}$ and ν degrees of freedom (used for hypothesis test)
- `Multiple R-squared:  XXX`
  Explained variation r^2
- The rest we do not use in this course

Oversigt

Explained variation and correlation

- Explained variation in a model is r^2, in summary "Multiple R-squared"
- Found as

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- The proportion of the total variability explained by the model

Explained variation and correlation

- The correlationen ρ is a measure of *linear relation* between two random variables
- Estimated (i.e. empirical) correlation

$$\hat{\rho} = r = \sqrt{r^2}\, sgn(\hat{\beta}_1)$$

where $sgn(\hat{\beta}_1)$ er: -1 for $\hat{\beta}_1 \leq 0$ and 1 for $\hat{\beta}_1 > 0$

- Hence:
  - Positive correlation when positive slope
  - Negative correlation when negative slope

Test for significance of correlation

- Test for significance of correlation (linear relation) between two variables

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

is equivalent to

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

where $\hat{\beta}_1$ is the estimated slope in a simple linear regression model

R Illustration

```
## Generates x
x <- runif(n=20, min=-2, max=4)
## Simulate y
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

## Scatter plot
plot(x,y)

## Use lm() to find the estimates
fit <- lm(y ~ x)

## The "true"  line
abline(beta0, beta1)
## Plot of fit
abline(fit, col="red")

## See summary
summary(fit)

## Correlation between  x and y
cor(x,y)

## Squared  becomes the "Multiple R-squared" from summary(fit)
cor(x,y)^2
```

Oversigt

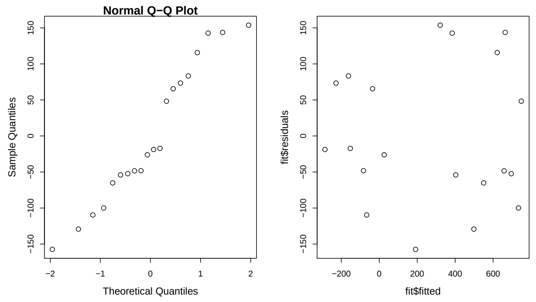
Residual Analysis

Method 5.26

- Check normality assumption with qq-plot.
- Check (non)systematic behavior by plotting the residuals e_i as a function of fitted values $\hat{y}_i$

Residual Analysis in R

```
fit <- lm(y ~ x)
par(mfrow = c(1, 2))
qqnorm(fit$residuals)
plot(fit$fitted, fit$residuals)
```



OR: Wally plot again!

Agenda