Goodness of fit tests for linear and logistic regression

Janez Stare, Rok Blagus

IBMI, Faculty of Medicine, Ljubljana, Slovenia

2018

Have you ever used goodness of fit test for linear regression?

Have you ever used goodness of fit test for linear regression?

- Probably not.

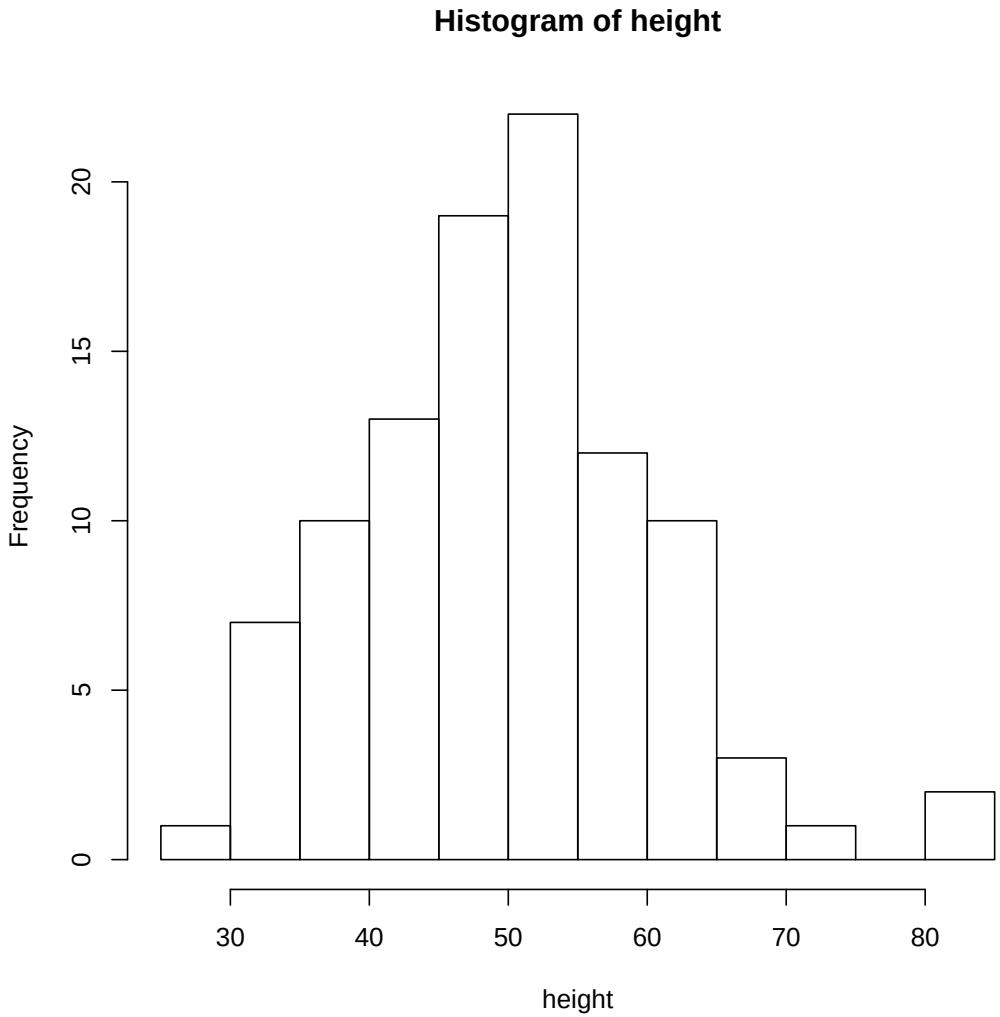
Have you ever used goodness of fit test for linear regression?

- Probably not.
- You were just **looking** at residuals.

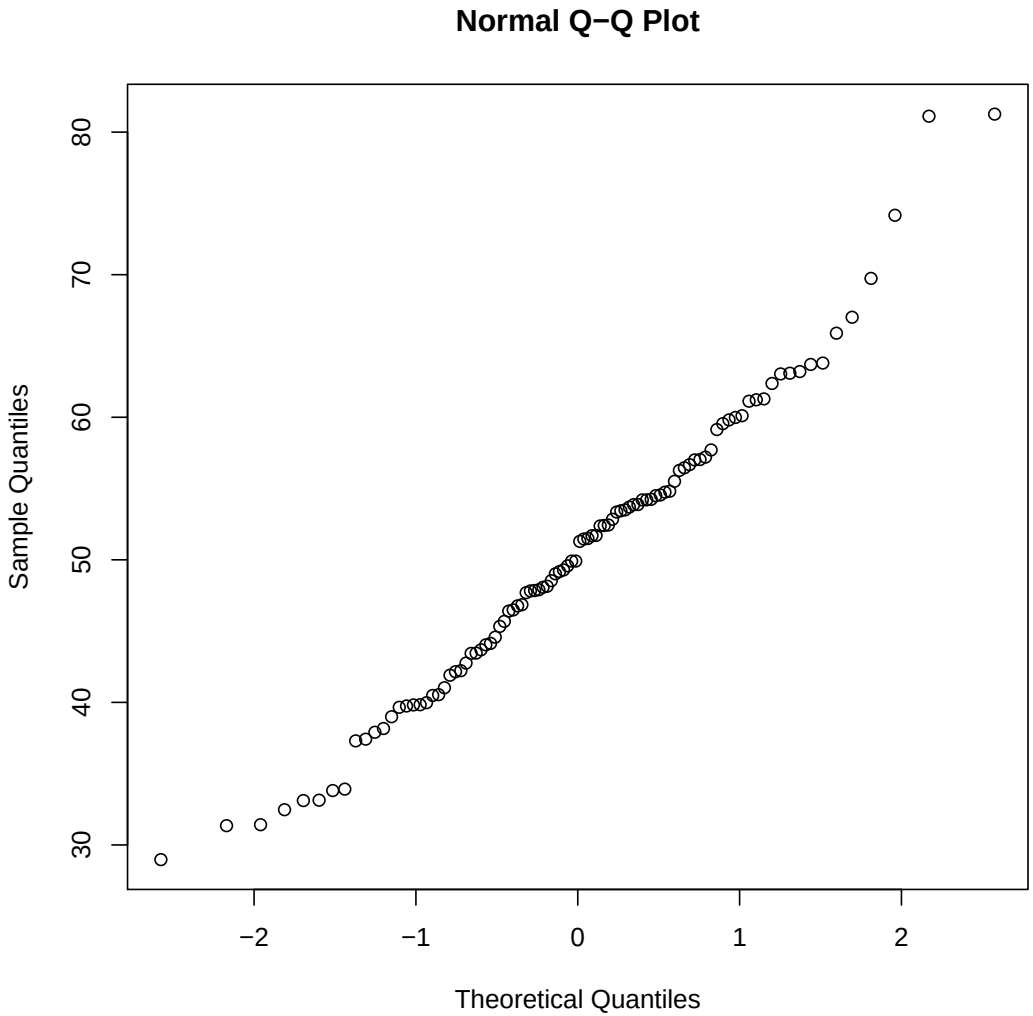
Have you ever used goodness of fit test for linear regression?

- Probably not.
- You were just **looking** at residuals.
- And you may have tested their normality.

Testing normality - enough to look at the histogram?



Or maybe at qq plot?



Testing normality

Testing normality

- testing means testing

Testing normality

- testing means testing
- Shapiro Wilk test gives $p - value = 0.1887$

Testing normality

- testing means testing
- Shapiro Wilk test gives $p - value = 0.1887$
- But: this test is just a test for the **marginal** distribution of residuals!

Testing normality

- testing means testing
- Shapiro Wilk test gives $p - value = 0.1887$
- But: this test is just a test for the **marginal** distribution of residuals!
- It has very little to do with the goodness of fit test for the model.

Is R^2 measuring fit?

7

Analysis of Model Fit

Model fit has traditionally been assessed by means of summary test statistics or by residual analysis. Since the analysis of fit is so important to the modeling process, we shall discuss four categories of fit assessment:

- Traditional fit tests
- Hosmer–Lemeshow GOF test
- Information criteria tests
- Residual analysis

7.1 Traditional Fit Tests for Logistic Regression

Fit statistics were first developed for the **generalized linear models** (GLM) parameterization of logistic regression. They were first integrated into the GLIM statistical package, which was written in the early 1970s. Recall from earlier discussion that the GLM approach uses an **iteratively re-weighted least squares** (IRLS) algorithm to estimate parameters, standard errors, and other model statistics. The results are identical to the full maximum likelihood method of estimation. GLIM was the first statistical software to incorporate a full set of fit statistics and residual analysis for the various GLM family models.

The traditional GOF statistics that we shall discuss are

1. R^2 and Pseudo-R^2 statistics
2. Deviance statistic
3. Likelihood ratio statistic

7.1.1 R^2 and Pseudo-R^2 Statistics

The standard **goodness-of-fit** (GOF) statistic for ordinary least squares regression, or Gaussian regression, is the R^2 statistic, also called the **coefficient of determination.** The value represents the percentage variation in the data accounted for by the model. Ranging from 0 to 1, the higher the R^2 value, the better the fit of the model.

243

Is R^2 measuring fit?

7

Analysis of Model Fit

Model fit has traditionally been assessed by means of summary test statistics or by residual analysis. Since the analysis of fit is so important to the modeling process, we shall discuss four categories of fit assessment:

- Traditional fit tests
- Hosmer–Lemeshow GOF test
- Information criteria tests
- Residual analysis

7.1 Traditional Fit Tests for Logistic Regression

Fit statistics were first developed for the **generalized linear models** (GLM) parameterization of logistic regression. They were first integrated into the GLIM statistical package, which was written in the early 1970s. Recall from earlier discussion that the GLM approach uses an **iteratively re-weighted least squares** (IRLS) algorithm to estimate parameters, standard errors, and other model statistics. The results are identical to the full maximum likelihood method of estimation. GLIM was the first statistical software to incorporate a full set of fit statistics and residual analysis for the various GLM family models.

The traditional GOF statistics that we shall discuss are

1. R^2 and Pseudo-R^2 statistics
2. Deviance statistic
3. Likelihood ratio statistic

7.1.1 R^2 and Pseudo-R^2 Statistics

The standard **goodness-of-fit** (GOF) statistic for ordinary least squares regression, or Gaussian regression, is the R^2 statistic, also called the **coefficient of determination.** The value represents the percentage variation in the data accounted for by the model. Ranging from 0 to 1, the higher the R^2 value, the better the fit of the model.

243

Is R^2 measuring fit?

7

Analysis of Model Fit

Model fit has traditionally been assessed by means of summary test statistics or by residual analysis. Since the analysis of fit is so important to the modeling process, we shall discuss four categories of fit assessment:

- Traditional fit tests
- Hosmer–Lemeshow GOF test
- Information criteria tests
- Residual analysis

7.1 Traditional Fit Tests for Logistic Regression

Fit statistics were first developed for the **generalized linear models** (GLM) parameterization of logistic regression. They were first integrated into the GLIM statistical package, which was written in the early 1970s. Recall from earlier discussion that the GLM approach uses an **iteratively re-weighted least squares** (IRLS) algorithm to estimate parameters, standard errors, and other model statistics. The results are identical to the full maximum likelihood method of estimation. GLIM was the first statistical software to incorporate a full set of fit statistics and residual analysis for the various GLM family models.

The traditional GOF statistics that we shall discuss are

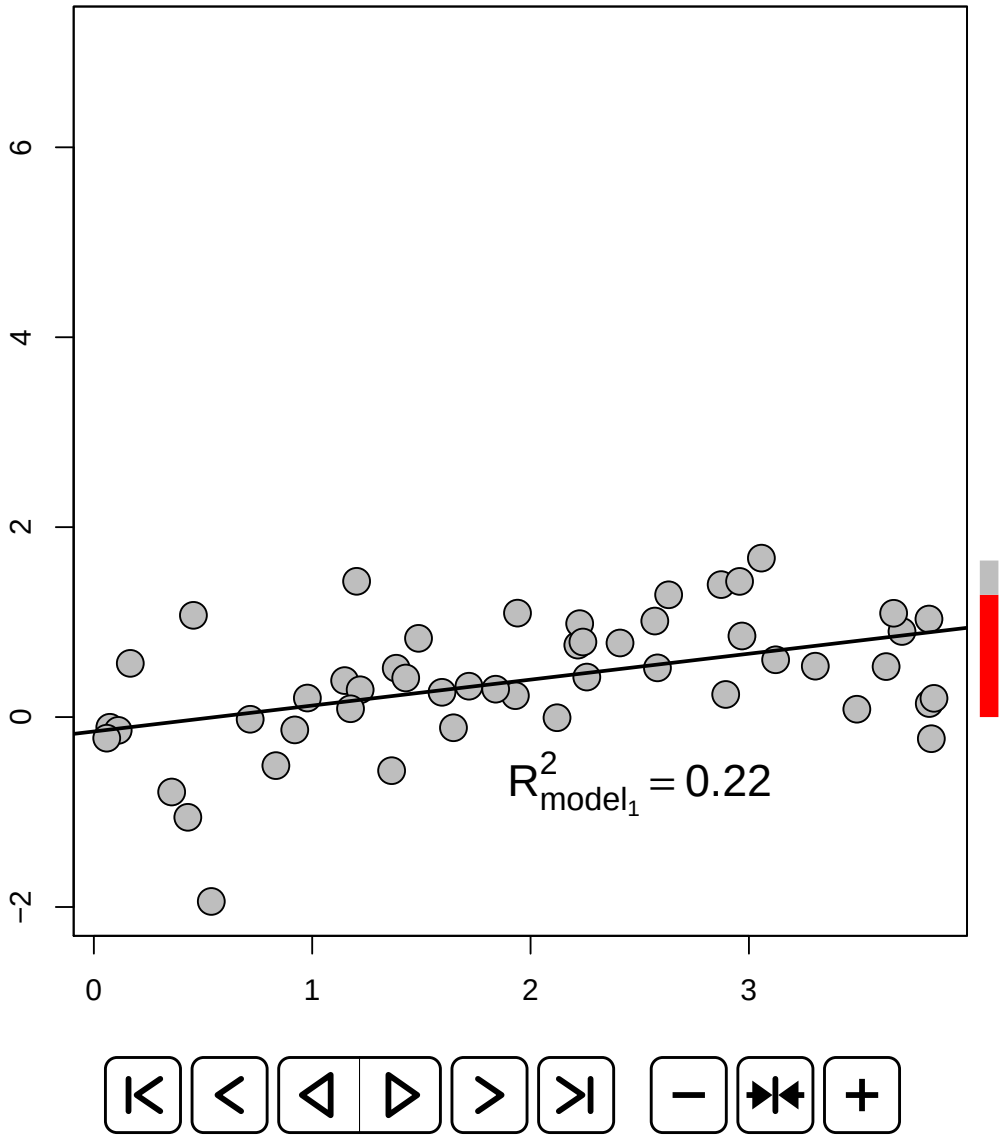
1. R^2 and Pseudo-R^2 statistics
2. Deviance statistic
3. Likelihood ratio statistic

7.1.1 R^2 and Pseudo-R^2 Statistics

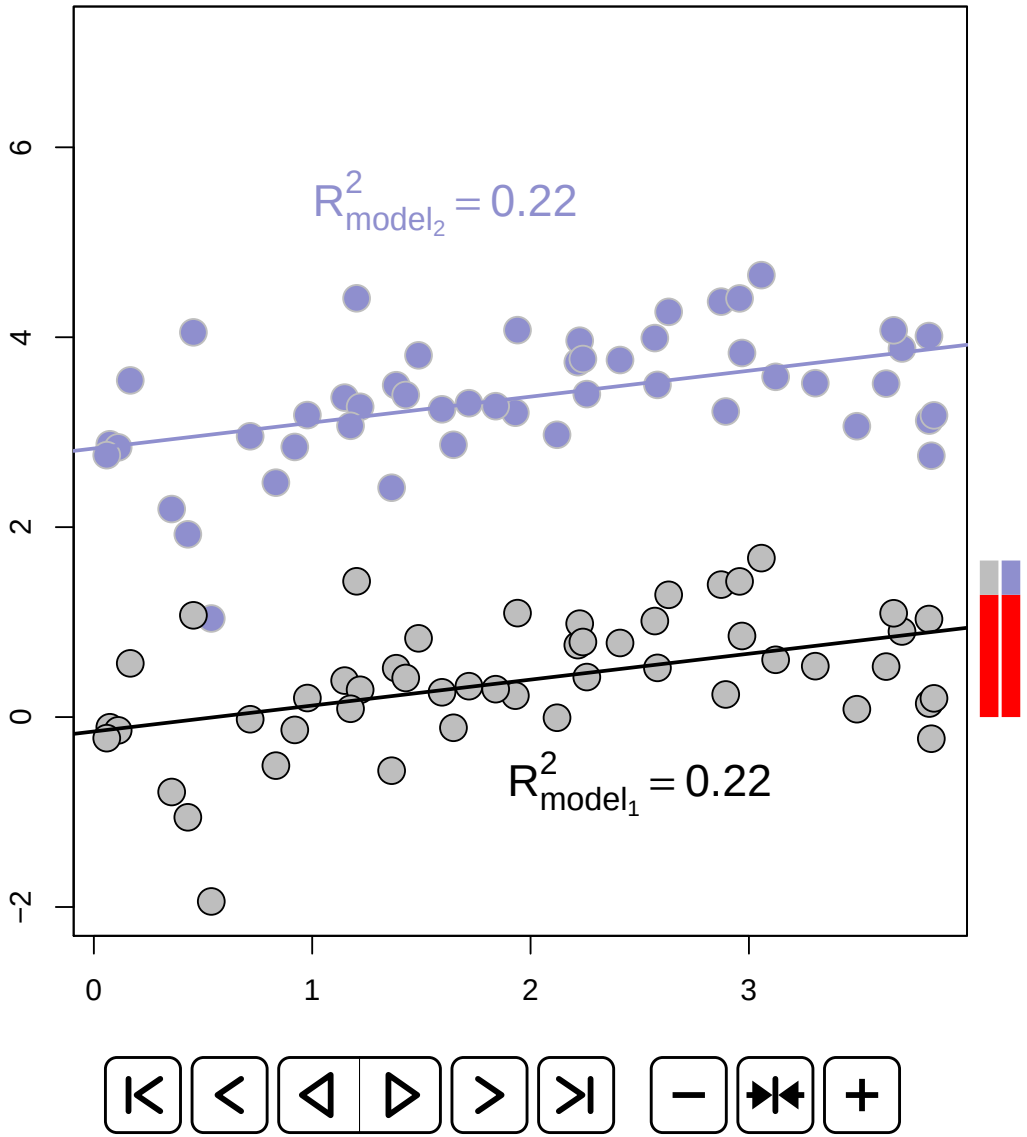
The standard **goodness-of-fit** (GOF) statistic for ordinary least squares regression, or Gaussian regression, is the R^2 statistic, also called the **coefficient of determination.** The value represents the percentage variation in the data accounted for by the model. Ranging from 0 to 1, the higher the R^2 value, the better the fit of the model.

243

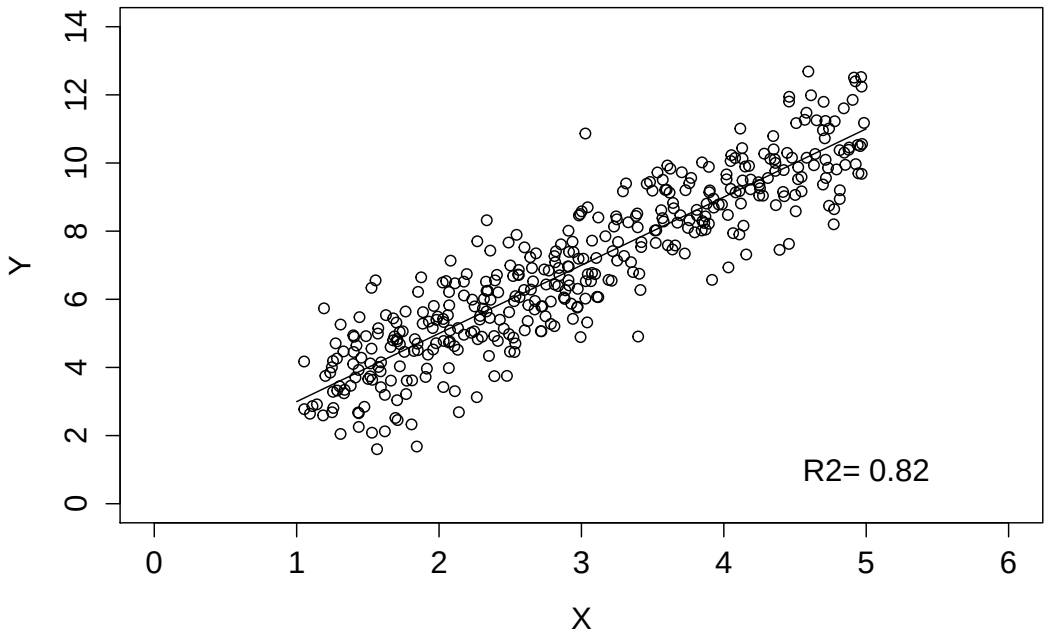
Changing intercept



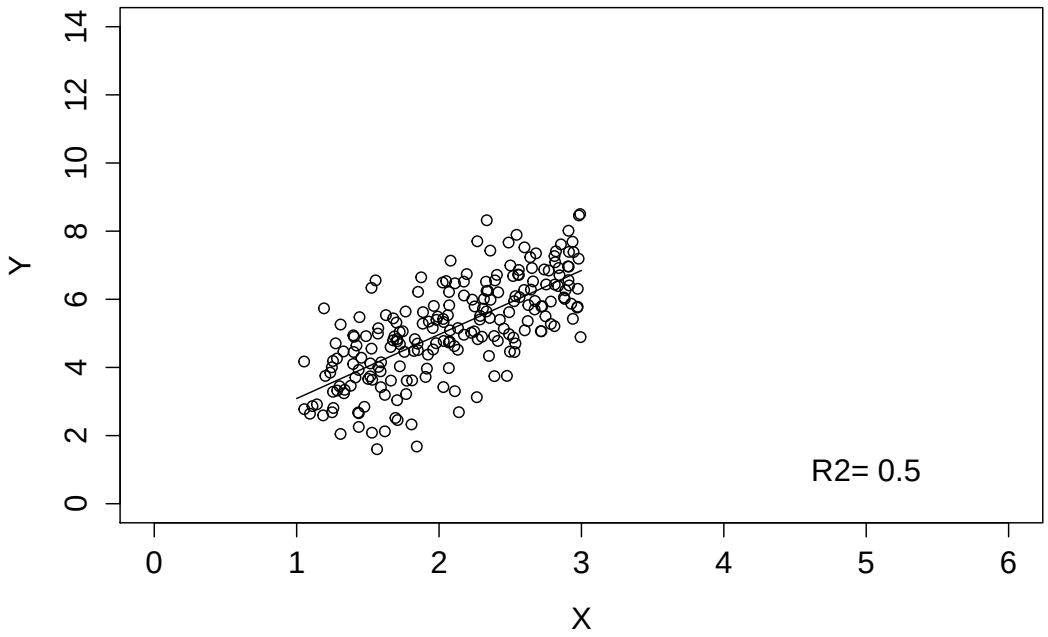
Changing slope



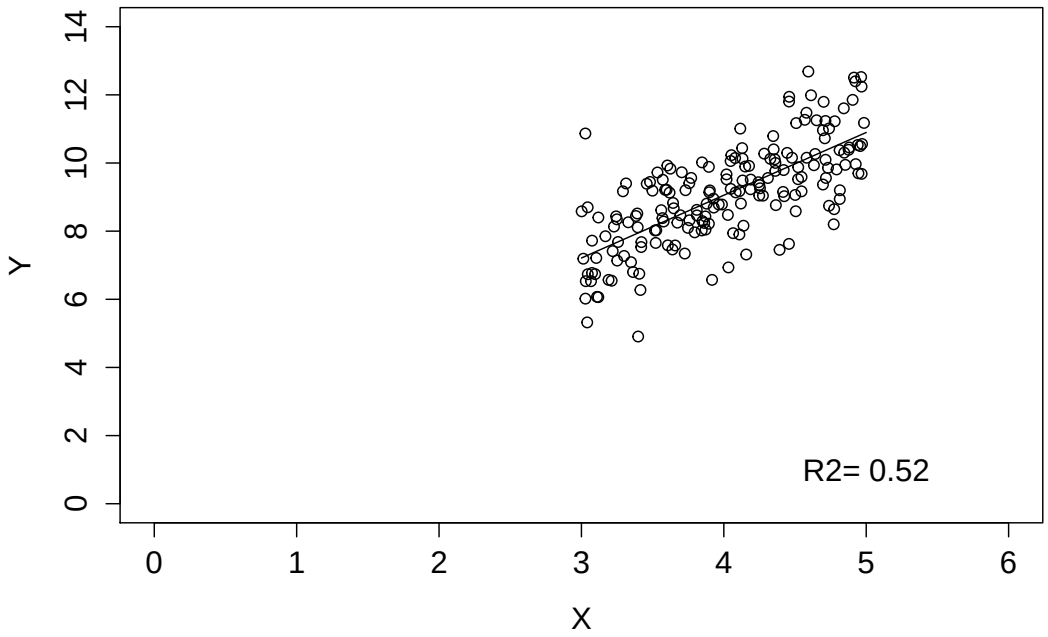
Dependence on distribution of X



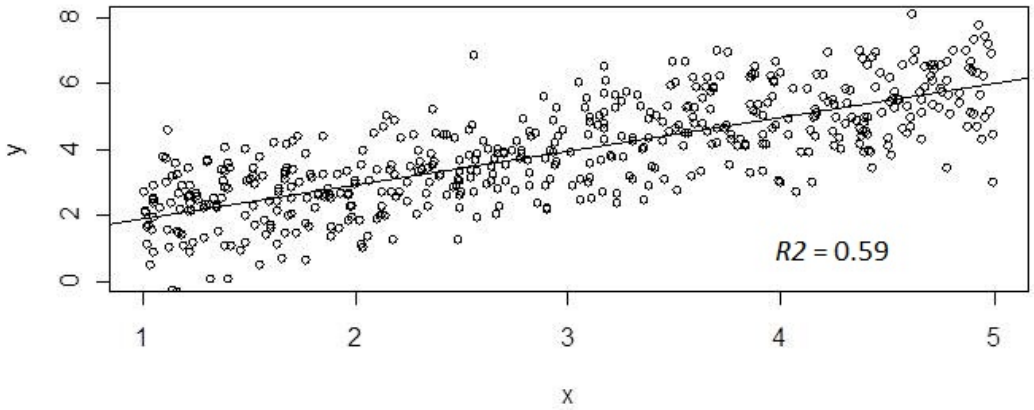
Dependence on distribution of X

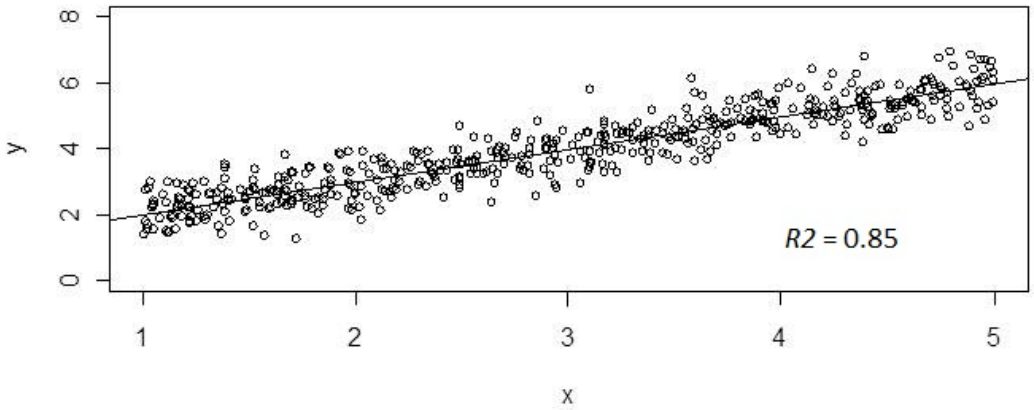


Dependence on distribution of X

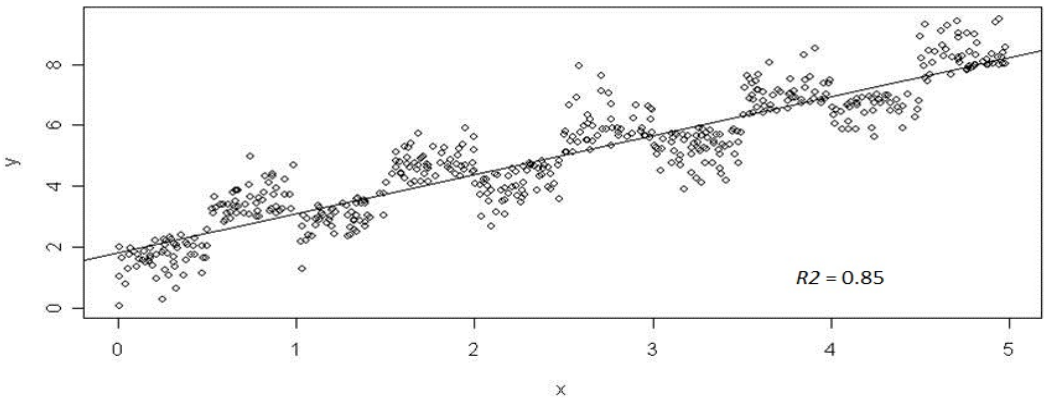


Misunderstanding - R2 is NOT a measure of fit





Misunderstanding - R2 is NOT a measure of fit



Goodness of fit

Answer to the question: **Could our model have generated the data**?

Karl Pearson (1900). On the Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen From Random Sampling. Philosophical Magazine.

Goodness of fit

Let X be a random variable with unknown distribution function $F(x)$ and $x_1, \ldots, x_n$ independent observations of X. The problem of testing

$$H_0 : F(x) = F_0(x)$$

is called a **goodness of fit problem**.

Goodness of fit in linear regression

In linear regression we are testing the hypothesis

$$H_0 : F(Y|x) = \mathcal{N}(\alpha + \beta x, \sigma^2)$$

An obvious idea - look at the sum of the residuals!

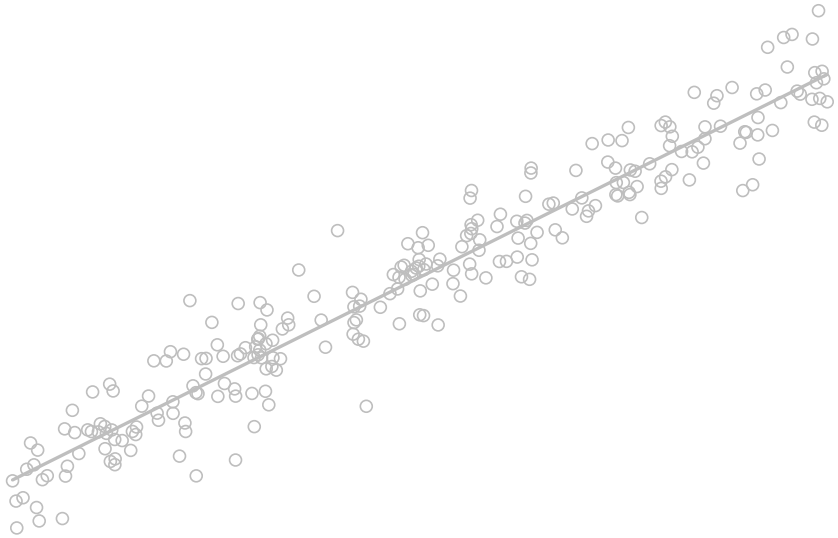
$$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$

An obvious idea - look at the sum of the residuals!

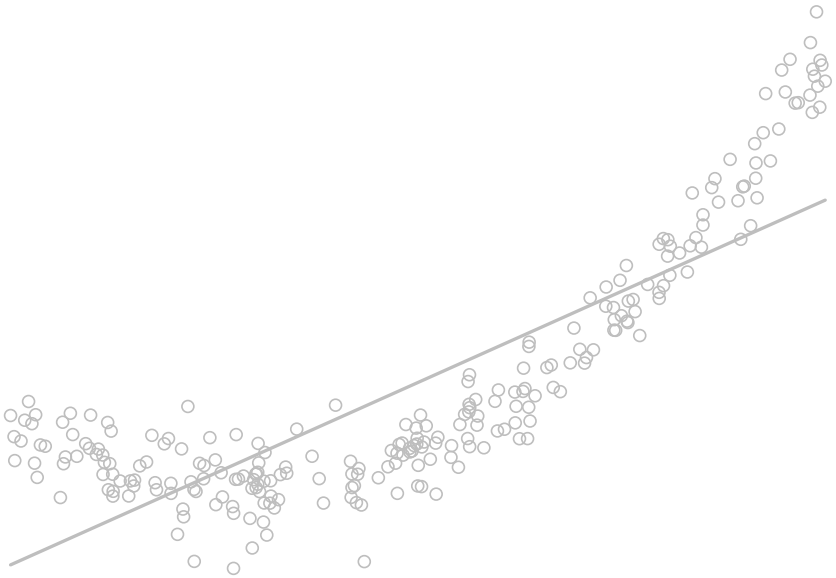
$$W_k = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) I(i \leq k)$$

where the ordering is based on predicted values $\hat{y}_i$.

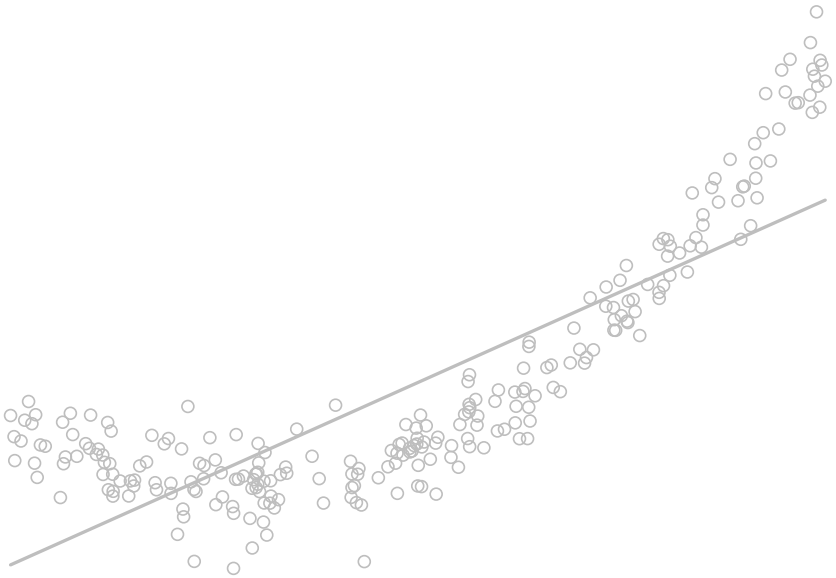
Goodness of fit - good model



Goodness of fit - bad model



Goodness of fit - bad model



If we had a Brownian bridge B_t

If we had a Brownian bridge B_t

we could define

$$K = \sup_t |B_t| \text{ and } C = \int_t B_t^2 dt$$

and use the fact that distributions of K and C are known.

If we had a Brownian bridge B_t

we could define

$$K = \sup_t |B_t| \text{ and } C = \int_t B_t^2 dt$$

and use the fact that distributions of K and C are known.

Unfortunately, we are far from having a Brownian bridge!

Variances of the empirical process

Data generated from

$$Y \sim \mathcal{N}(0.2 + 0.4X_1 + 0.2X_2, 1)$$

where X_1 and X_2 are independent and both distributed as $\mathcal{N}(0, 1)$.

Process is rescaled to $[0, 1]$

Variances of the empirical process

Data generated from

$$Y \sim \mathcal{N}(0.2 + 0.4X_1 + 0.2X_2, 1)$$

where X_1 and X_2 are independent and both distributed as $\mathcal{N}(0, 1)$.

Process is rescaled to $[0, 1]$

n	0.25	0.5	0.75
5000	0.086	0.089	0.086
1000	0.086	0.092	0.086
500	0.086	0.094	0.086
100	0.085	0.091	0.085

Variances of the empirical process

Data generated from

$$Y \sim \mathcal{N}(0.2 + 0.4X_1 + 0.2X_2, 1)$$

where X_1 and X_2 are independent and both distributed as $\mathcal{N}(0, 1)$.

Process is rescaled to $[0, 1]$

n	0.25	0.5	0.75
5000	0.086	0.089	0.086
1000	0.086	0.092	0.086
500	0.086	0.094	0.086
100	0.085	0.091	0.085
theoretical	0.185	0.250	0.185

So how do we test?

So how do we test?

- we decide on the test statistic (like max distance from 0)

So how do we test?

- we decide on the test statistic (like max distance from 0)
- we need its distribution under the null hypothesis

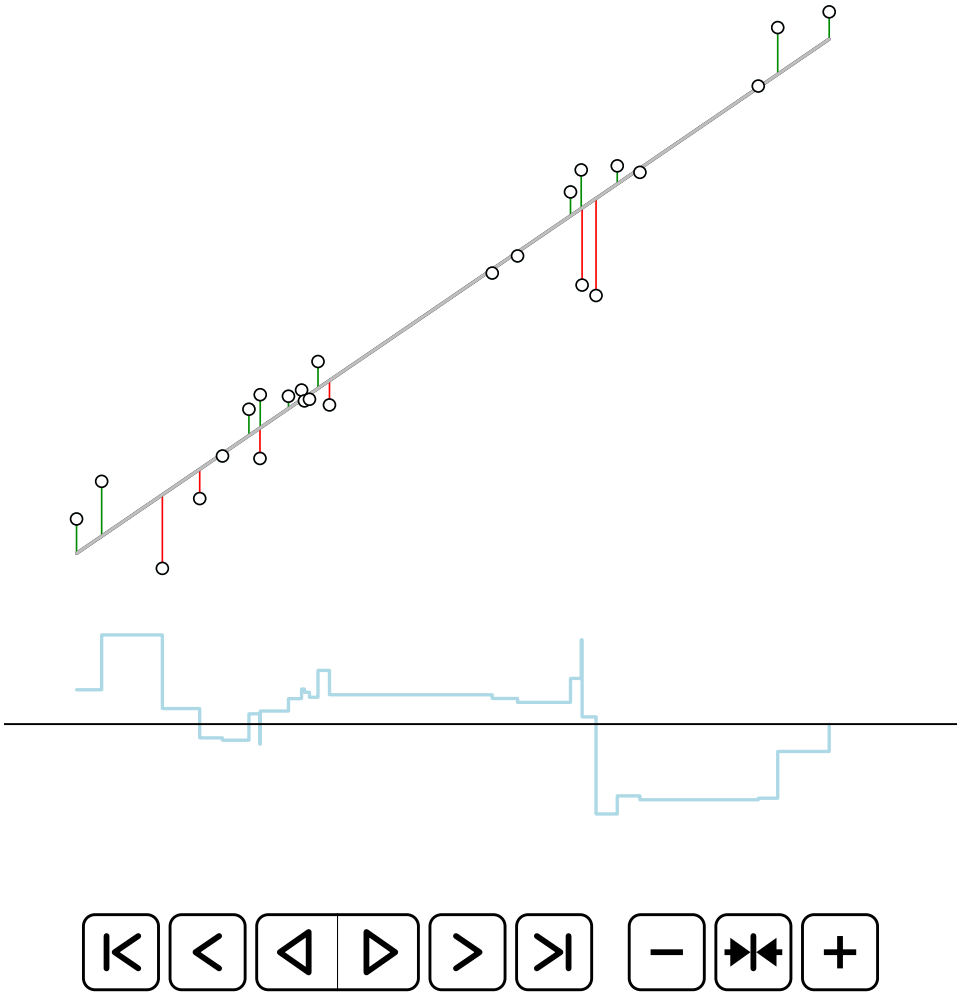
So how do we test?

- we decide on the test statistic (like max distance from 0)
- we need its distribution under the null hypothesis
- this is difficult!

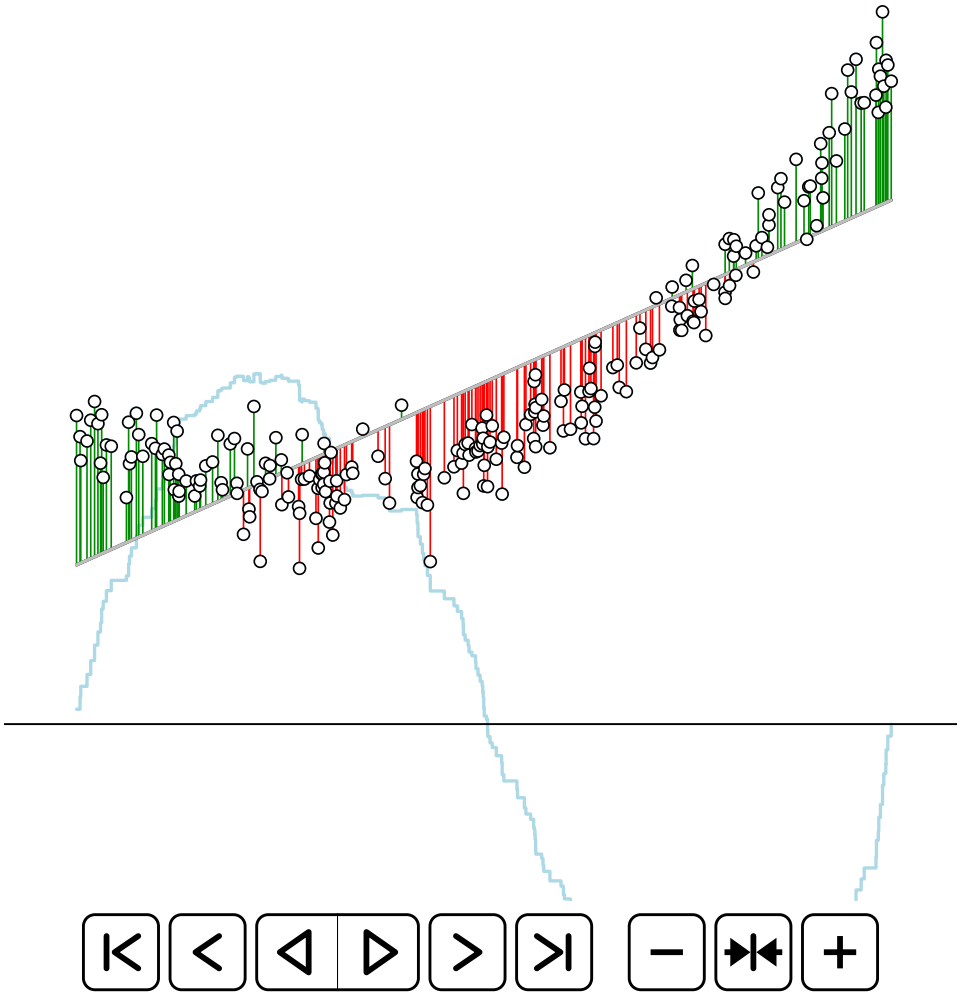
So how do we test?

- we decide on the test statistic (like max distance from 0)
- we need its distribution under the null hypothesis
- this is difficult!
- but possible!

Goodness of fit - permutation under a good model.



Goodness of fit - permutations under a bad model.



The proposed GOF test

- we order the residuals by predicted values $\hat{y}_i$

The proposed GOF test

- we order the residuals by predicted values $\hat{y}_i$
- define,

$$\widehat{W}_k = \frac{1}{\sqrt{n}\hat{\sigma}} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) I(i \leq k) \quad (1)$$

The proposed GOF test

- we order the residuals by predicted values $\hat{y}_i$
- define,

$$\widehat{W}_k = \frac{1}{\sqrt{n}\hat{\sigma}} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) I(i \leq k) \quad (1)$$

where $\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ (residuals are **standardized**).

The proposed GOF test

- we order the residuals by predicted values $\hat{y}_i$
- define,

$$\widehat{W}_k = \frac{1}{\sqrt{n}\hat{\sigma}} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) I(i \leq k) \tag{1}$$

  where $\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ (residuals are **standardized**).
- Test statistics:

$$K = \max |\widehat{W}_n(t)| \text{ and } C = \sum \widehat{W}_n(t)^2,$$

- The null distributions of the proposed test statistics are obtained with permutations.

Why is it necessary to refit the model for each permutation?

In short: to make sure that correlation of residuals is kept.

Why is it necessary to standardize the residuals?

It is obvious that under permutation of residuals,

$$\sum_{i=1}^{n}(e_i^g)^2 \leq \sum_{i=1}^{n} e_i^2, g = 1, ..., G$$

for any permutation g among all G permutations.

Does it work?

Requirements for a test

Does it work?

Requirements for a test

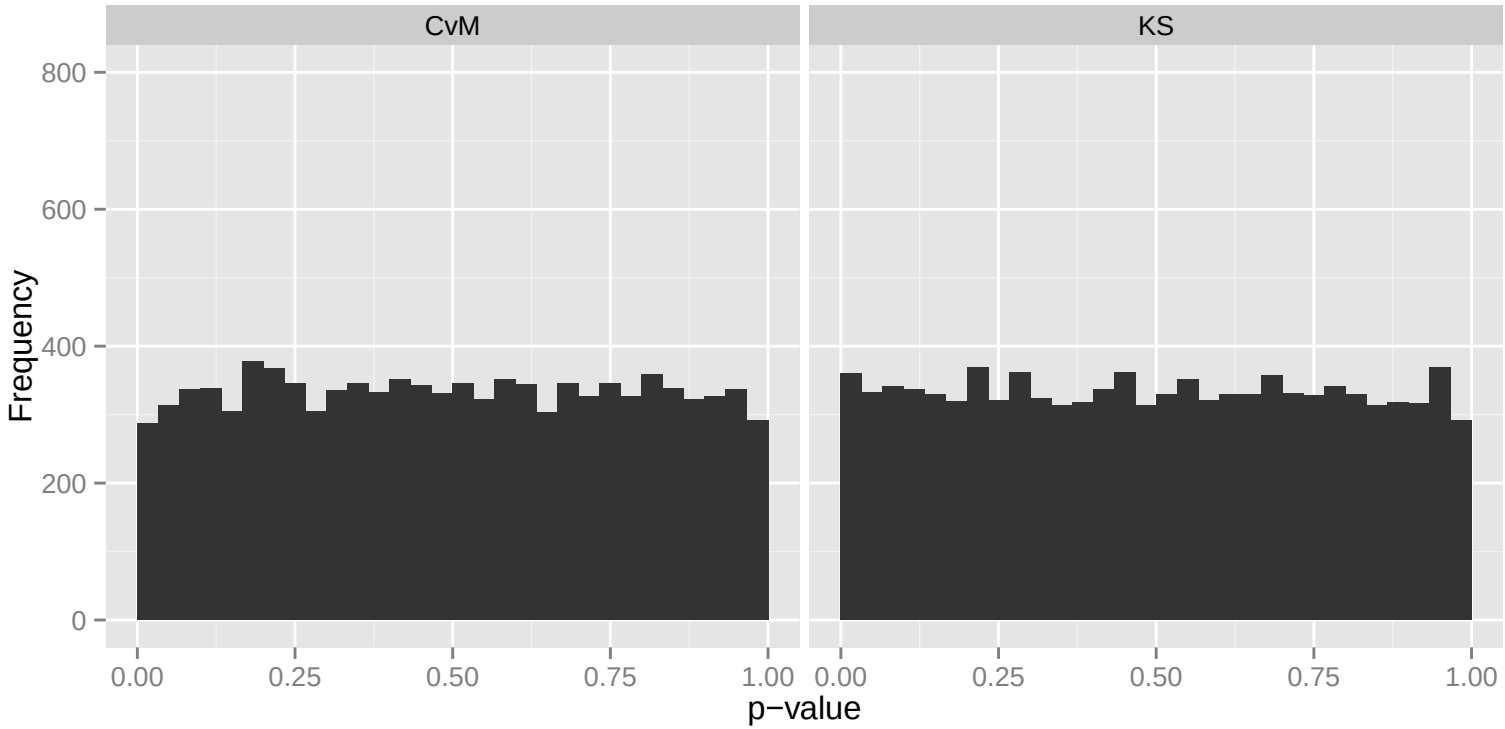
- *p*-values **must** have a uniform distribution under the null (**must** have a correct size)

Does it work?

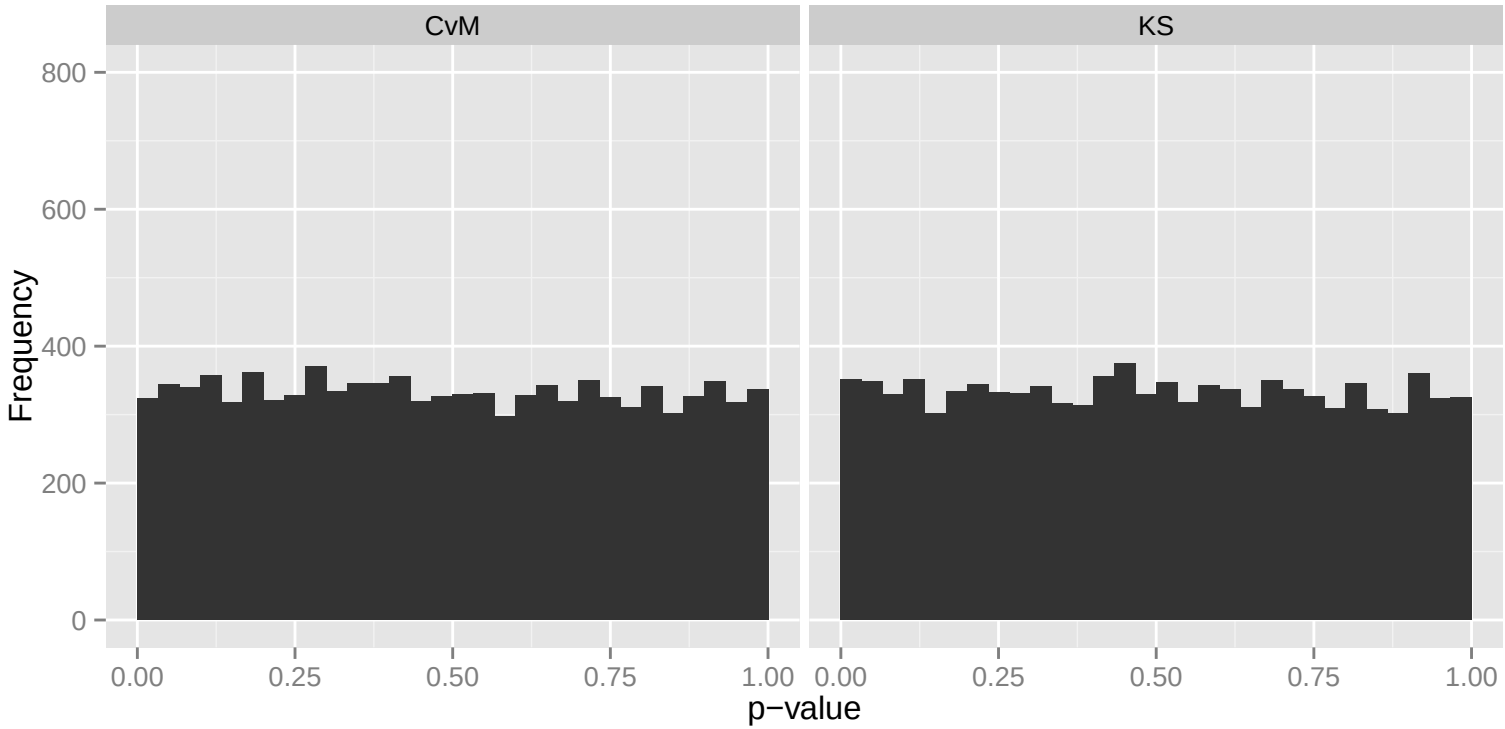
Requirements for a test

- *p*-values **must** have a uniform distribution under the null (**must** have a correct size)
- it is nice if it has good power

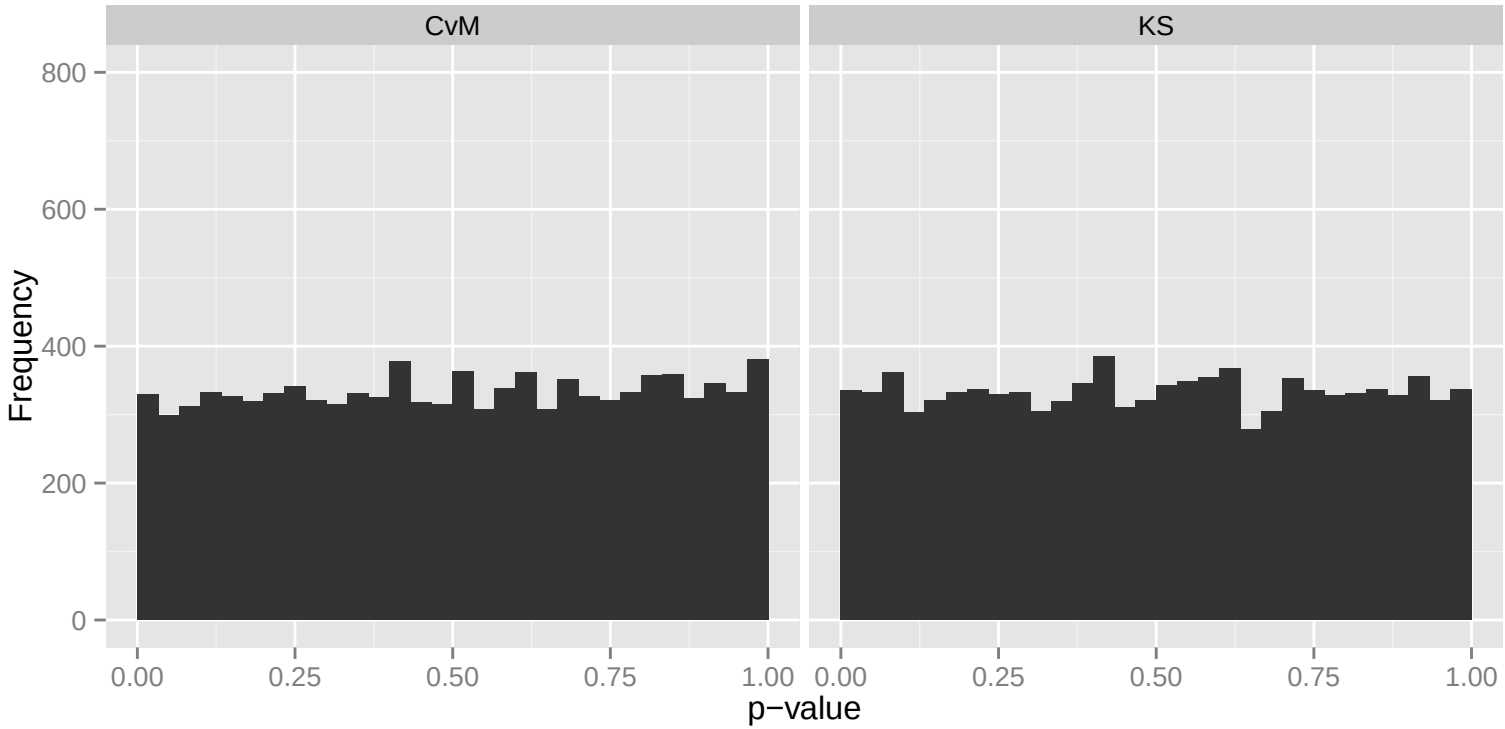
Distribution of p-values under the null for $n = 10$



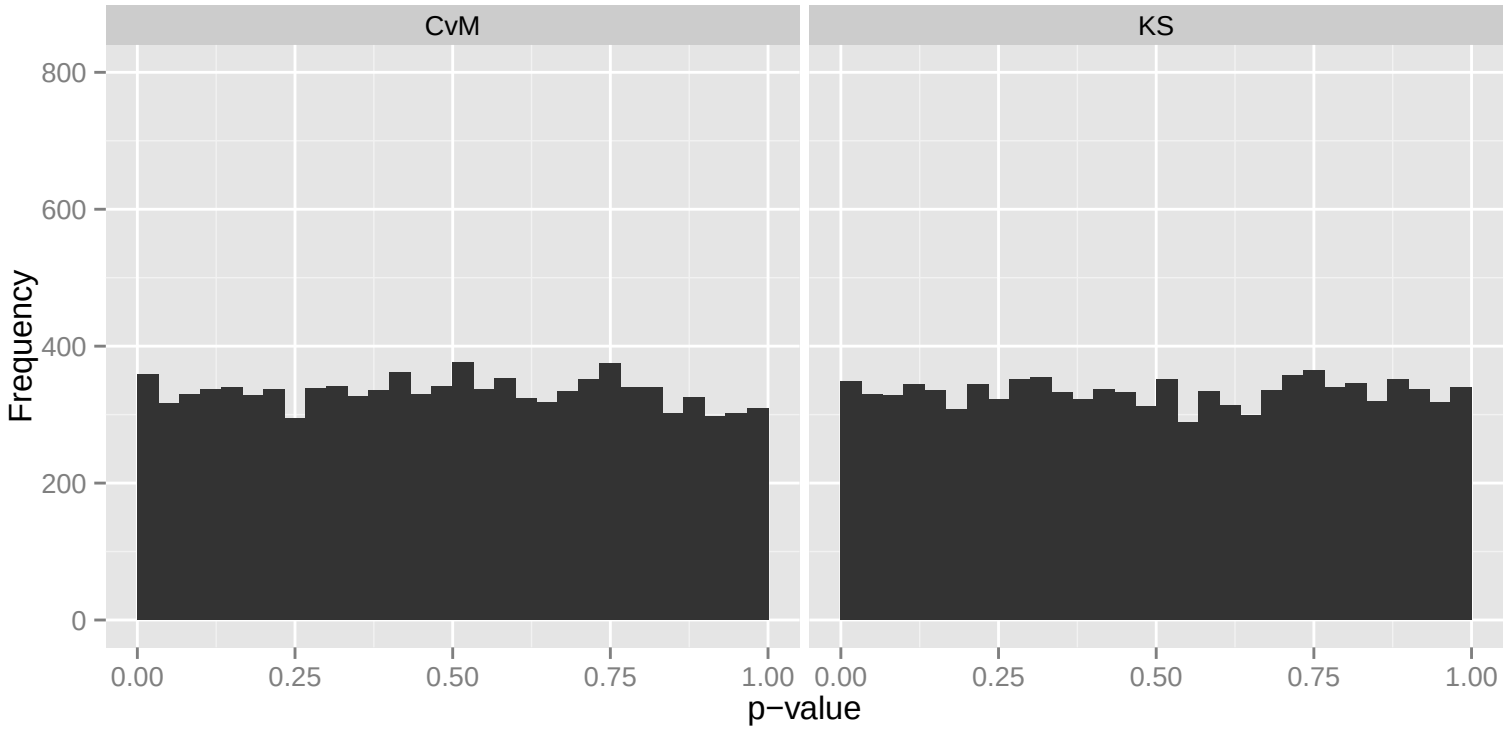
Distribution of p-values under the null for $n = 50$



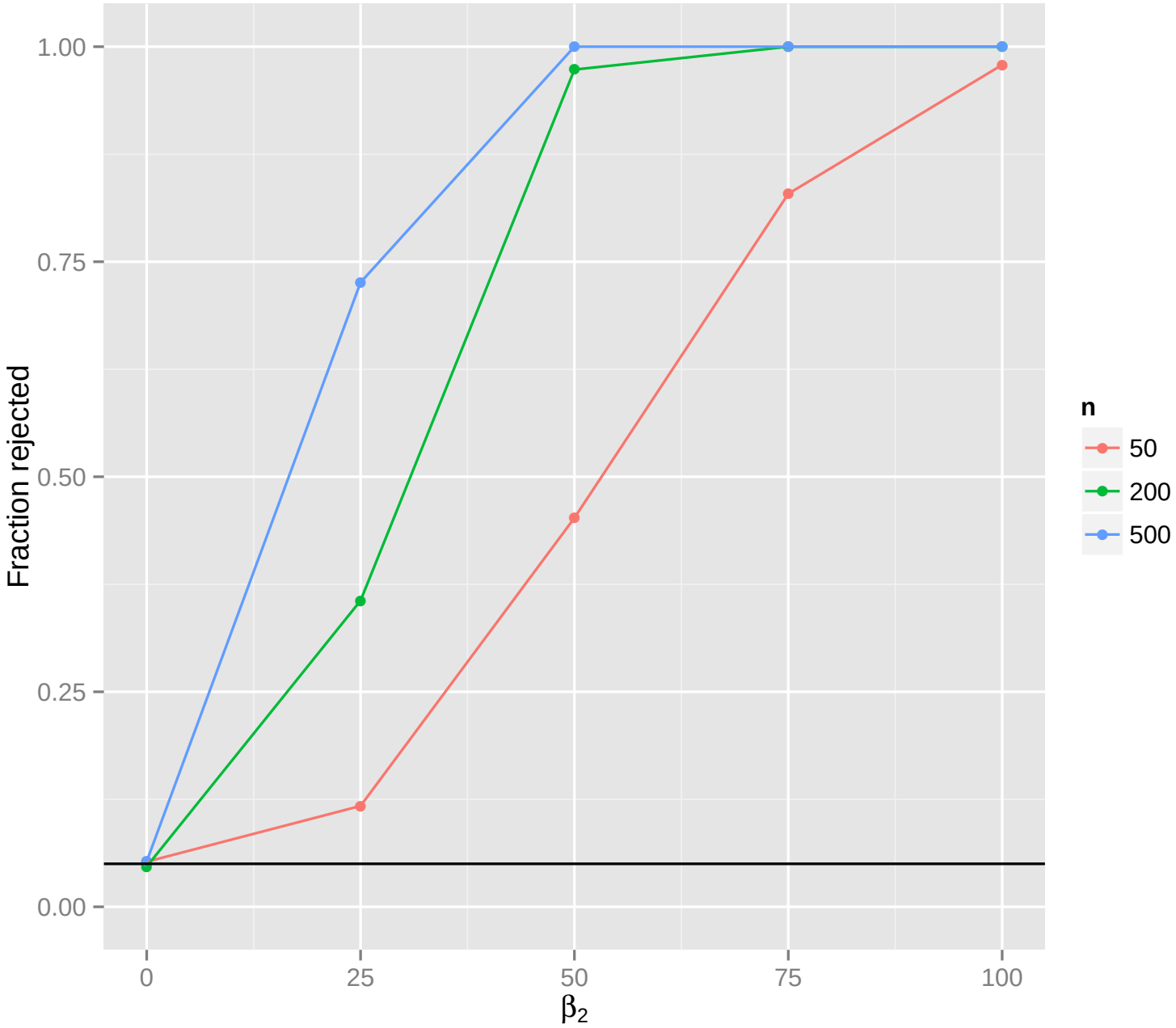
Distribution of p-values under the null for $n = 200$



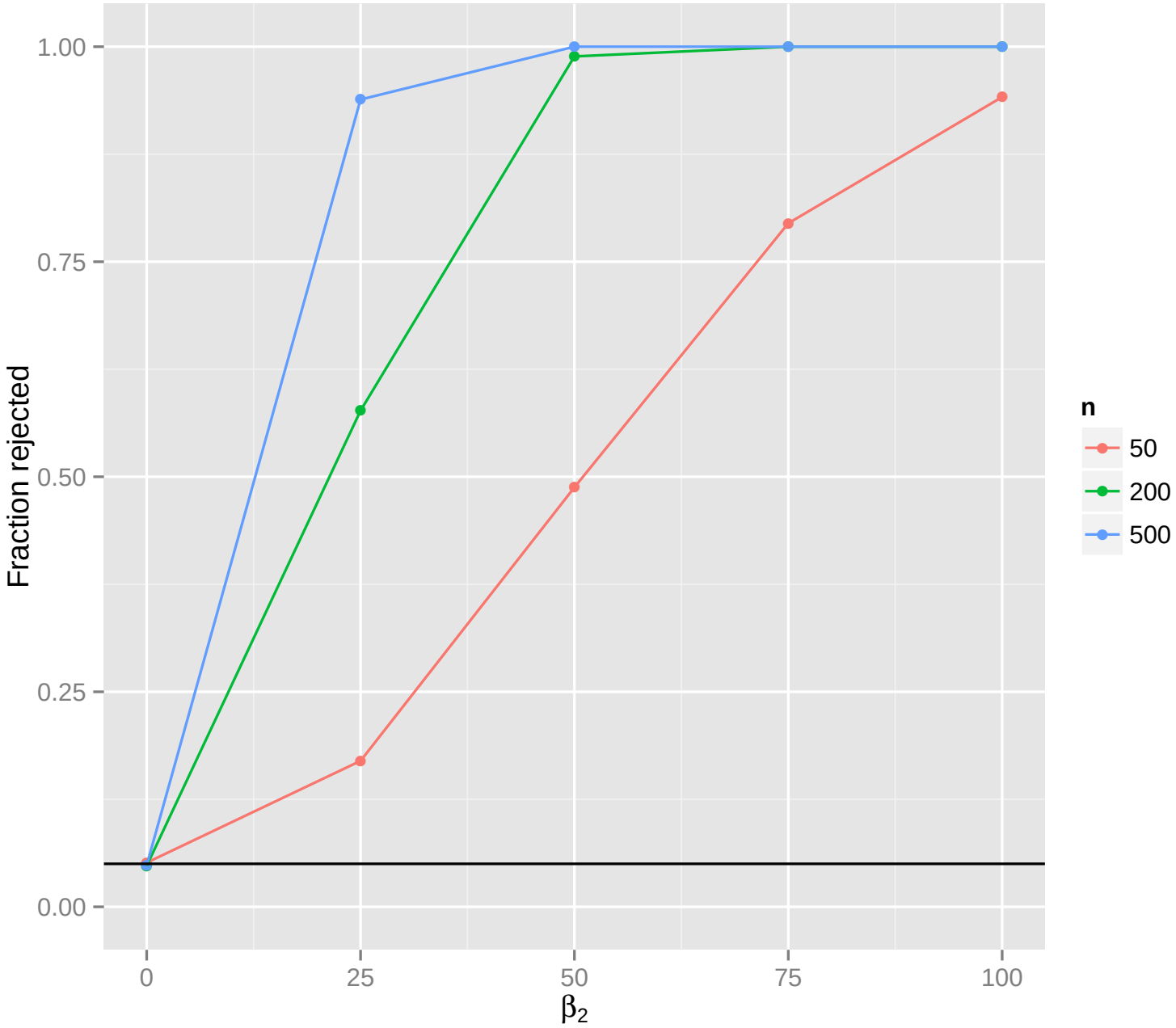
Distribution of p-values under the null $n = 500$



Power: alternative - squared term missing

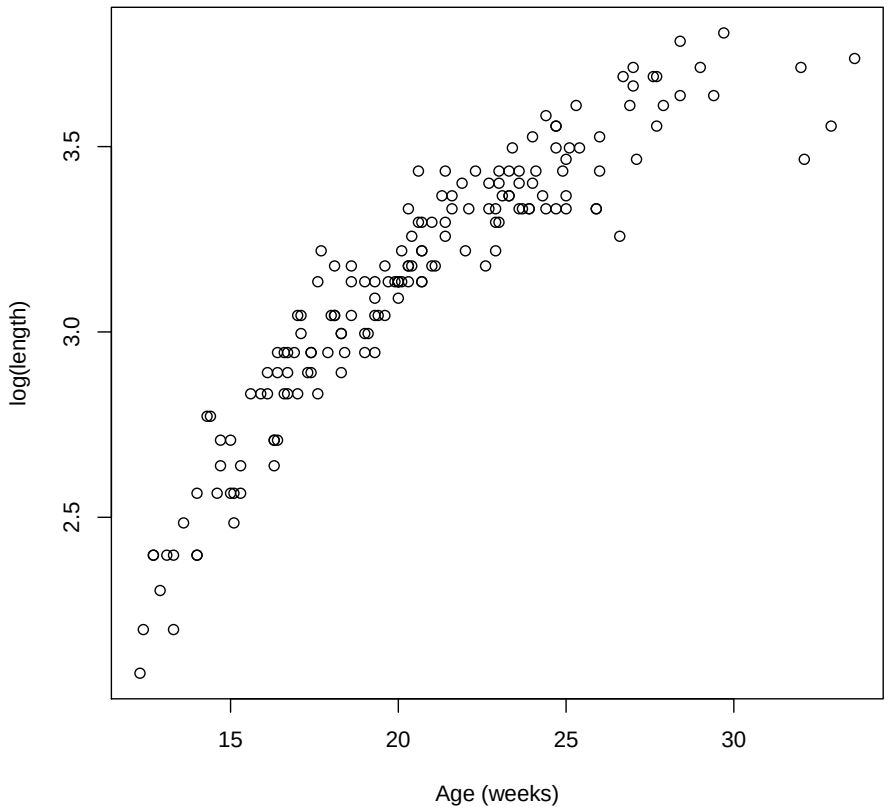


Power: alternative - interaction missing



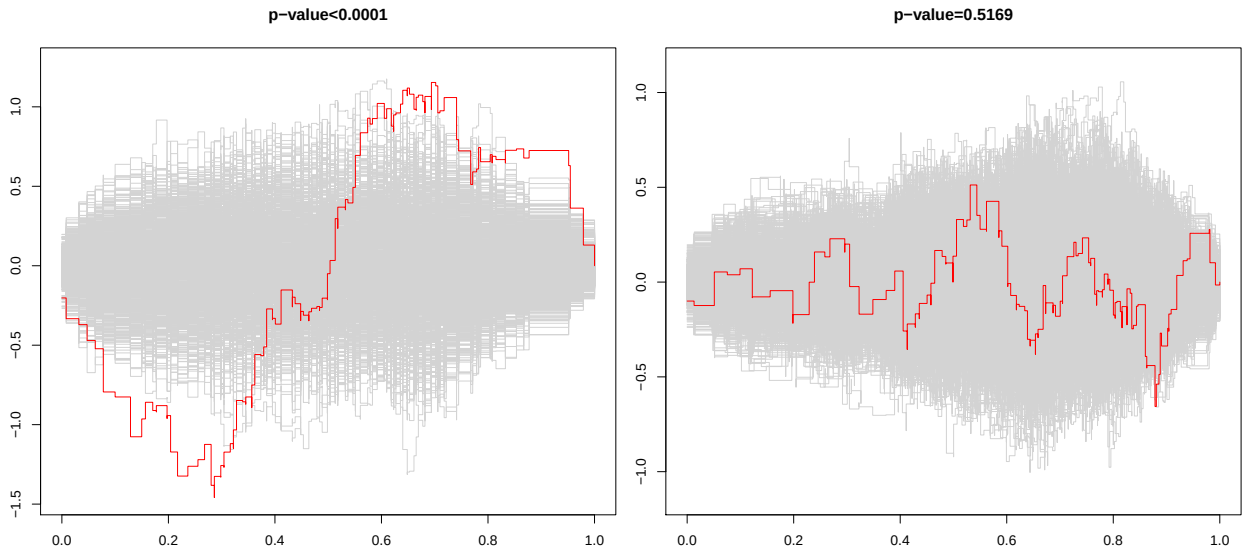
Example

- Data from a study of fetal mandible length
- Measurements of mandible length (log transformed) and gestational age in 158 fetuses.

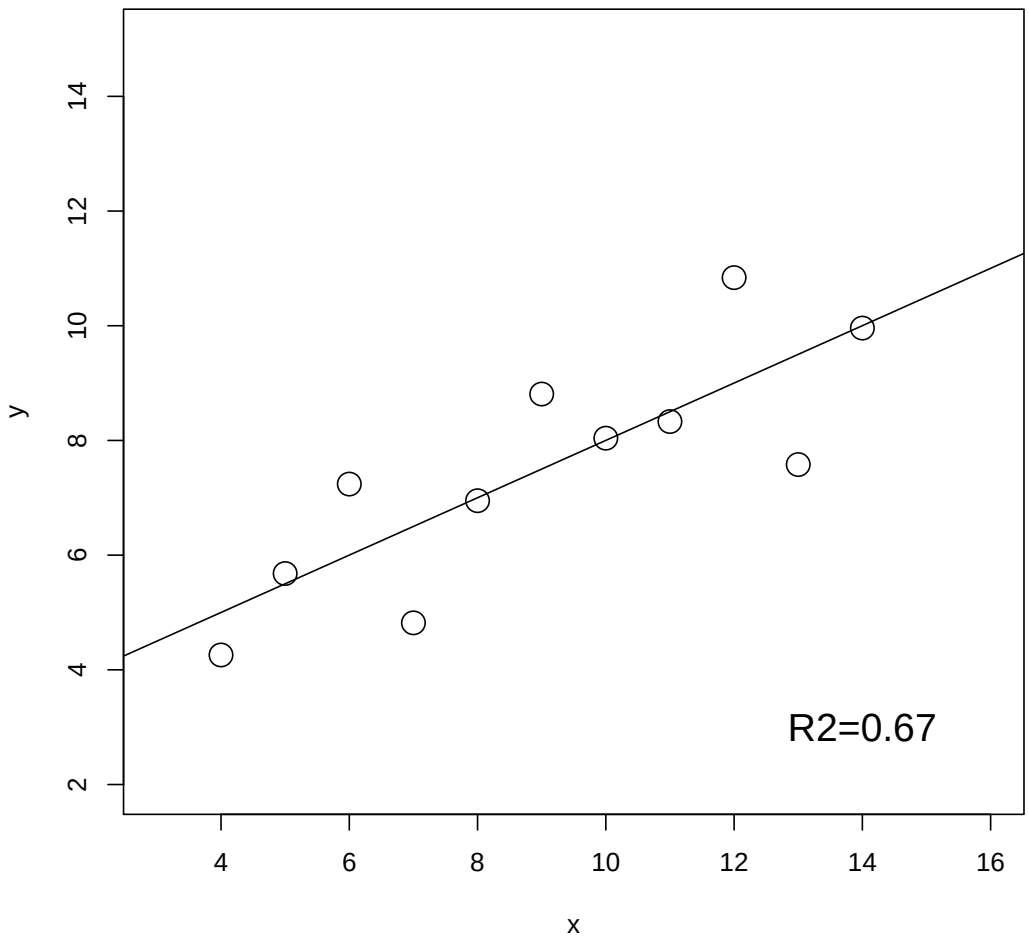


Example (revisited)

The process on the data (red curves) and processes obtained after permutations with Cramer-von Mieses type test statistic. Left panel is for the model including only age, the right panel is for the model including also square of age.

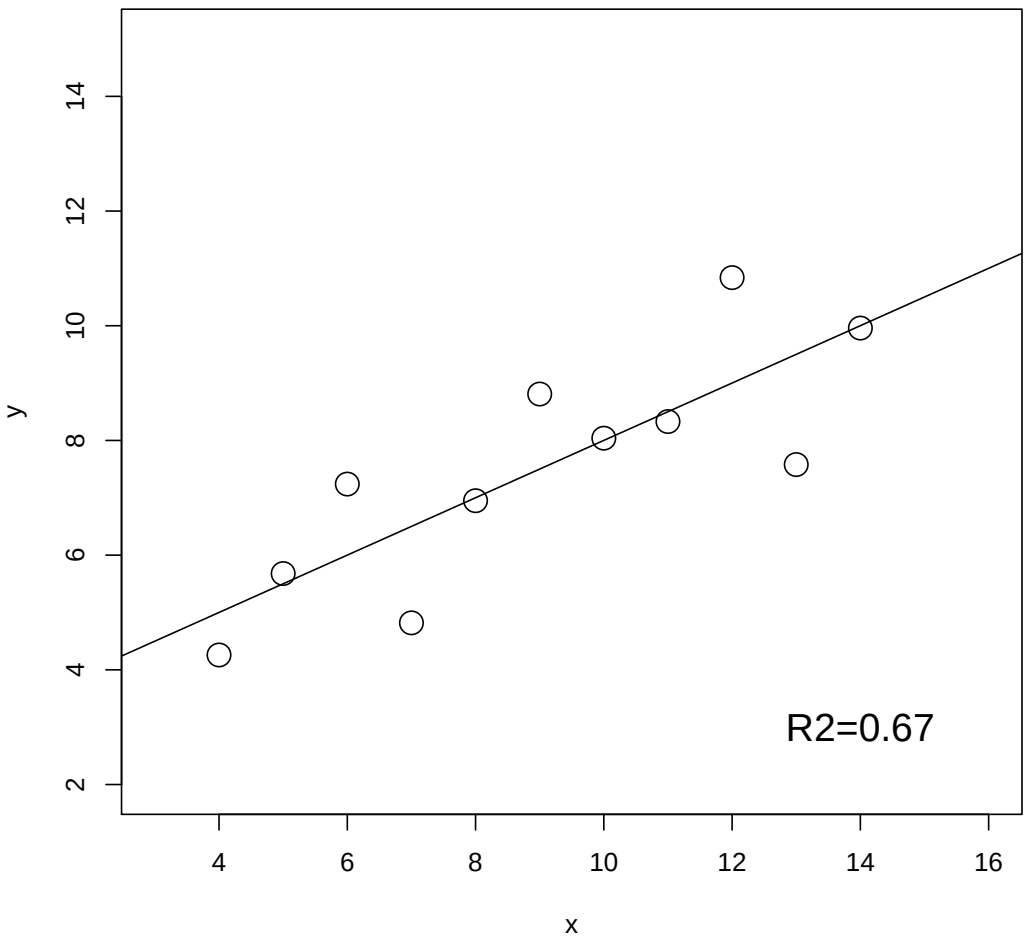


Goodness of fit - Anscombe data with *p* values

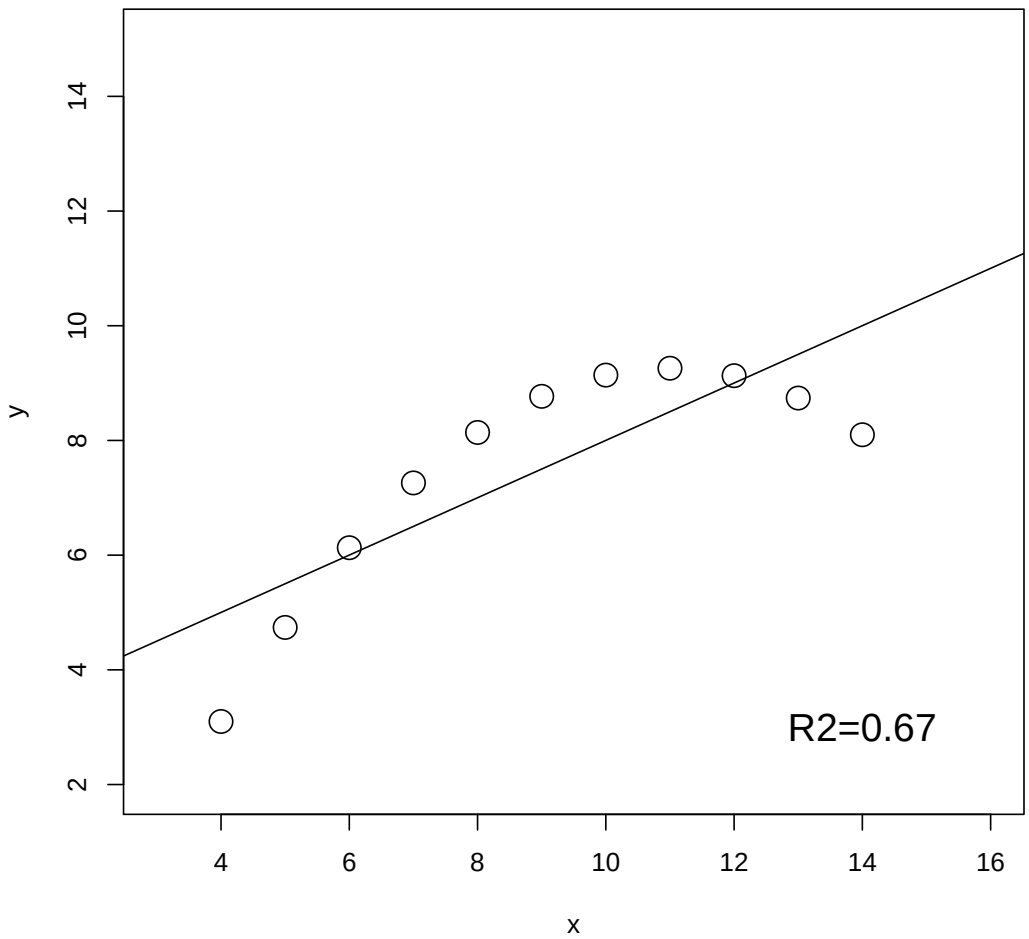


Goodness of fit - Anscombe data with *p* values

$KS = 0.44928$ $CvM = 0.80572$

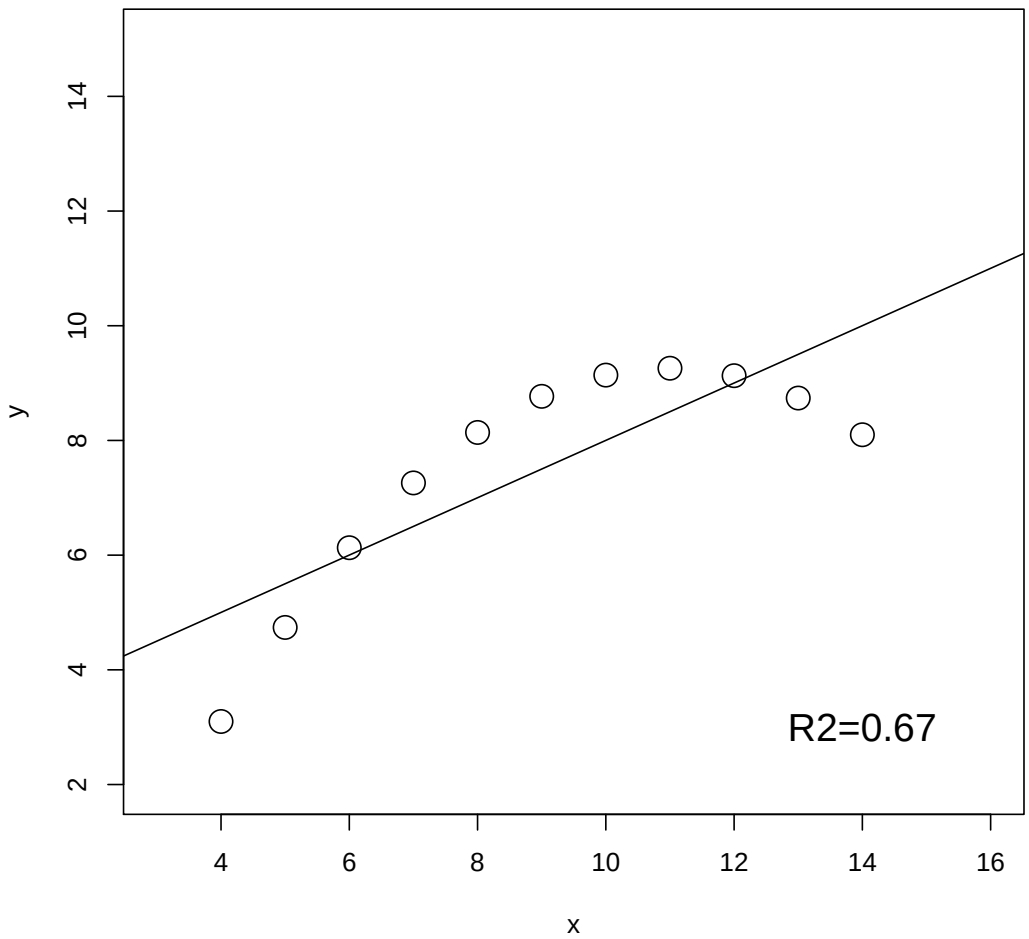


Goodness of fit - Anscombe data with *p* values

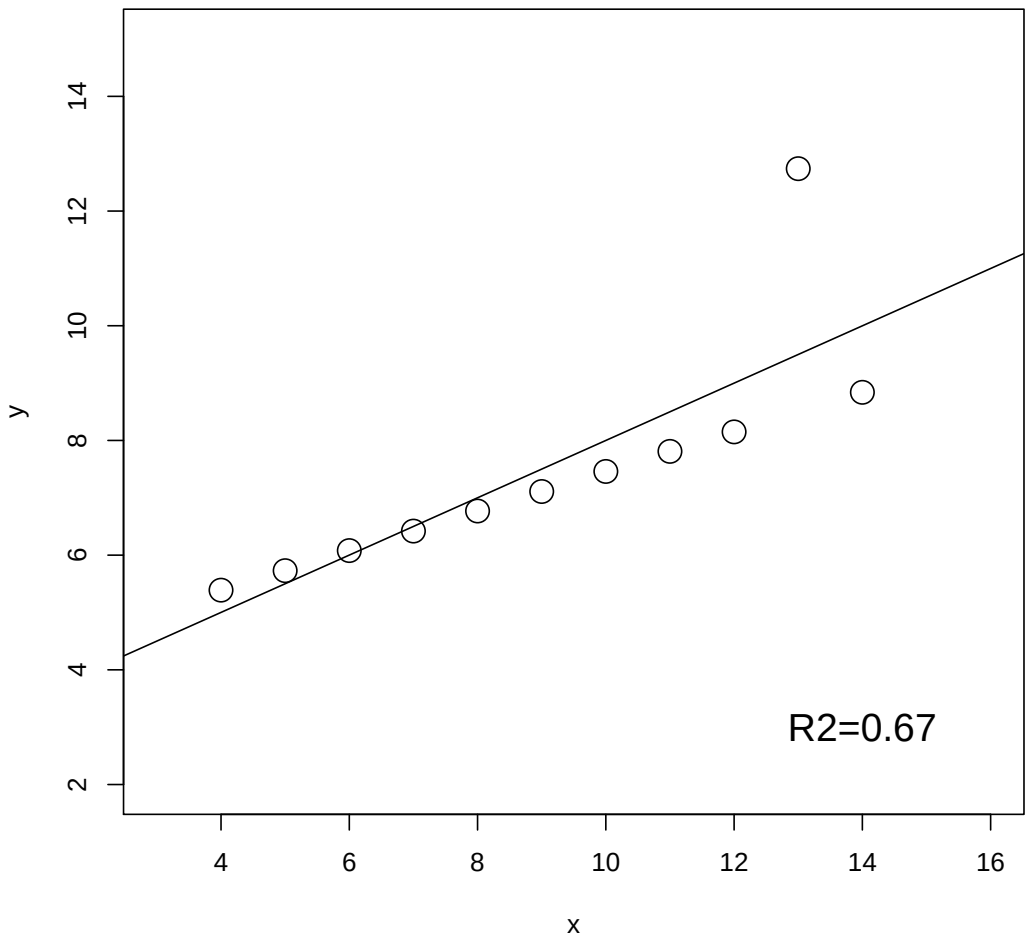


Goodness of fit - Anscombe data with p values

$KS = 0.05890$ $CvM = 0$

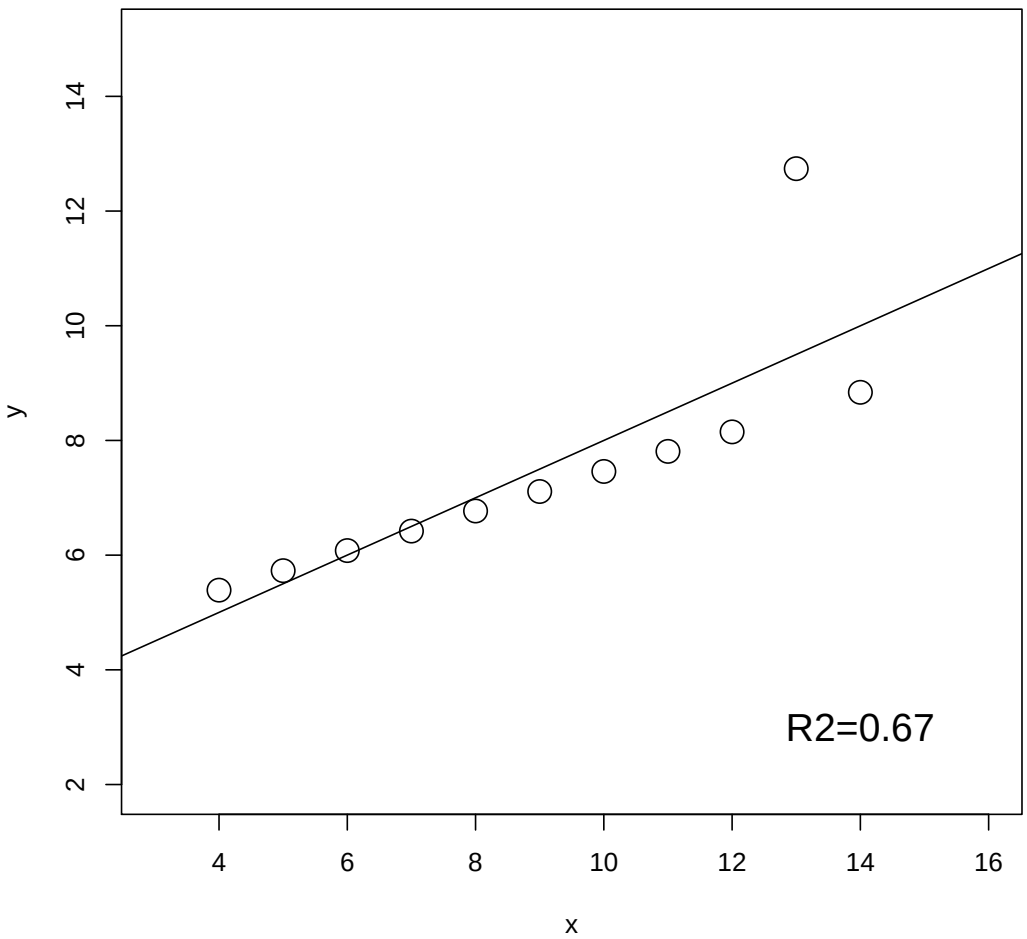


Goodness of fit - Anscombe data with *p* values

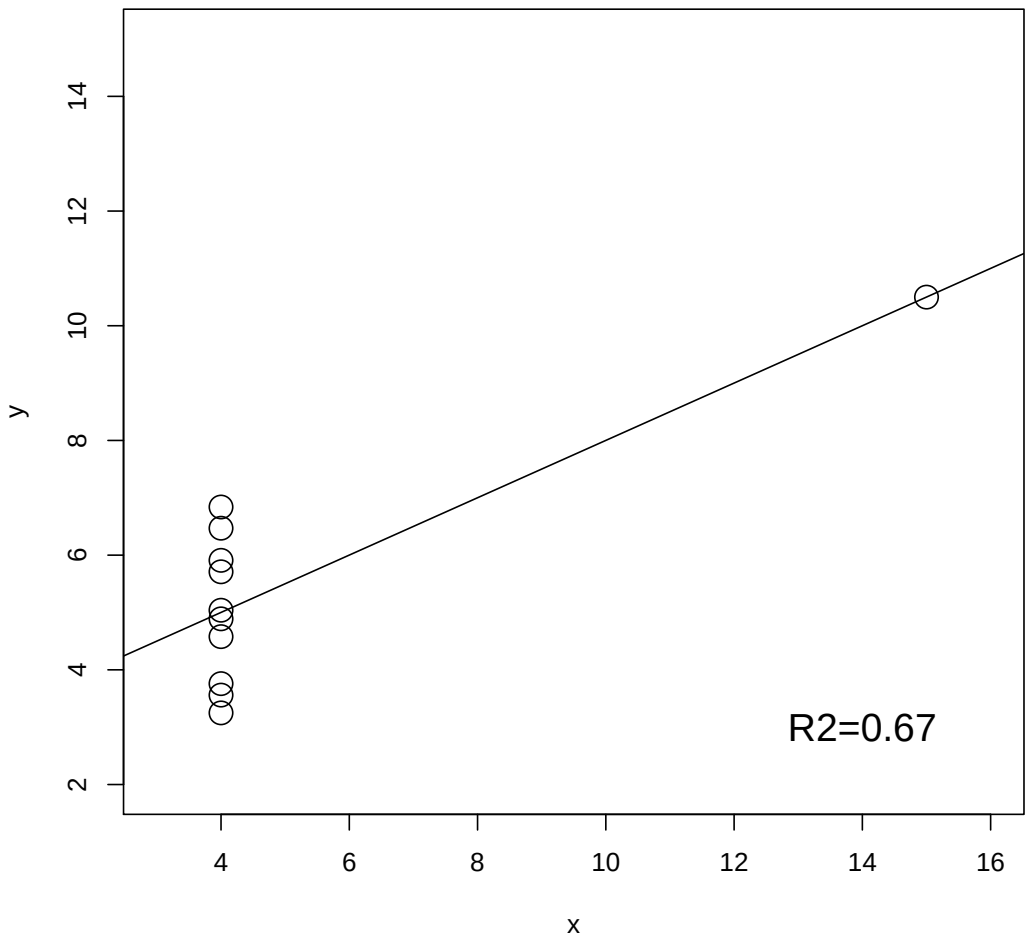


Goodness of fit - Anscombe data with *p* values

$KS = 0.41412$ $CvM = 0.71304$

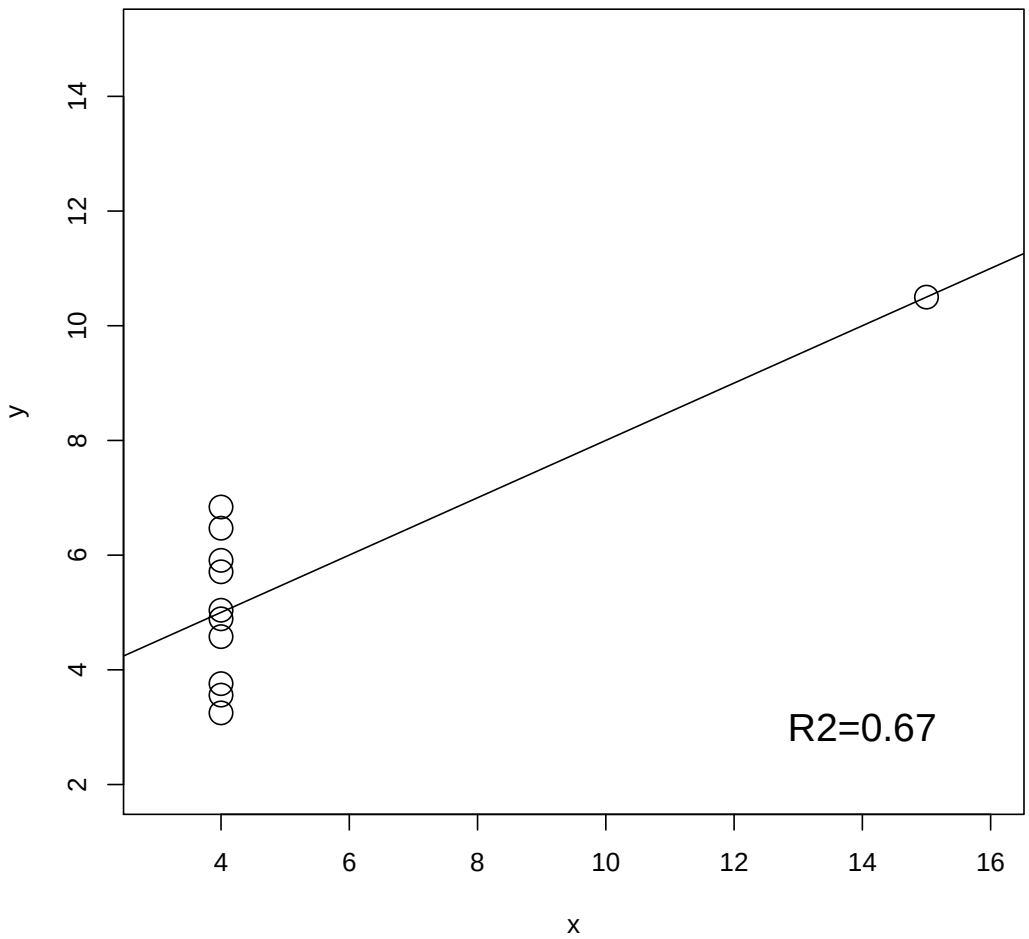


Goodness of fit - Anscombe data with *p* values



Goodness of fit - Anscombe data with *p* values

$KS = 1$ $\quad$ $CvM = 1$



Goodness of fit for binary logistic regression models

Goodness of fit for binary logistic regression models

- the outcome of interest is a binary variable Y

Goodness of fit for binary logistic regression models

- the outcome of interest is a binary variable Y
- we are interested in *probability* of seeing a given outcome (say 1)

Goodness of fit for binary logistic regression models

- the outcome of interest is a binary variable Y
- we are interested in *probability* of seeing a given outcome (say 1)
- the model is

$$p_i := P(Y_i = 1|x_i) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

Goodness of fit for binary logistic regression models

- the outcome of interest is a binary variable Y
- we are interested in *probability* of seeing a given outcome (say 1)
- the model is

$$p_i := P(Y_i = 1|x_i) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

- where x_i are values of an explanatory variable (or vector of such variables), and α and β are coefficients to be estimated from the data

Goodness of fit for binary logistic regression models

- the outcome of interest is a binary variable Y
- we are interested in *probability* of seeing a given outcome (say 1)
- the model is

$$p_i := P(Y_i = 1|x_i) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

- where x_i are values of an explanatory variable (or vector of such variables), and α and β are coefficients to be estimated from the data
- result of the fit are estimated probabilities $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_n$

Logistic regression

- our data are $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- fitted probabilities $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_n$,

The usual test in packages is the Hosmer Lemeshow test

- order the predicted probabilities
- group them in 10 (or so) groups
- count the number of 1s in each group (observed frequencies) and compare them to the sum of predicted probabilities (expected frequencies)
- compare observed to predicted using the chi square test

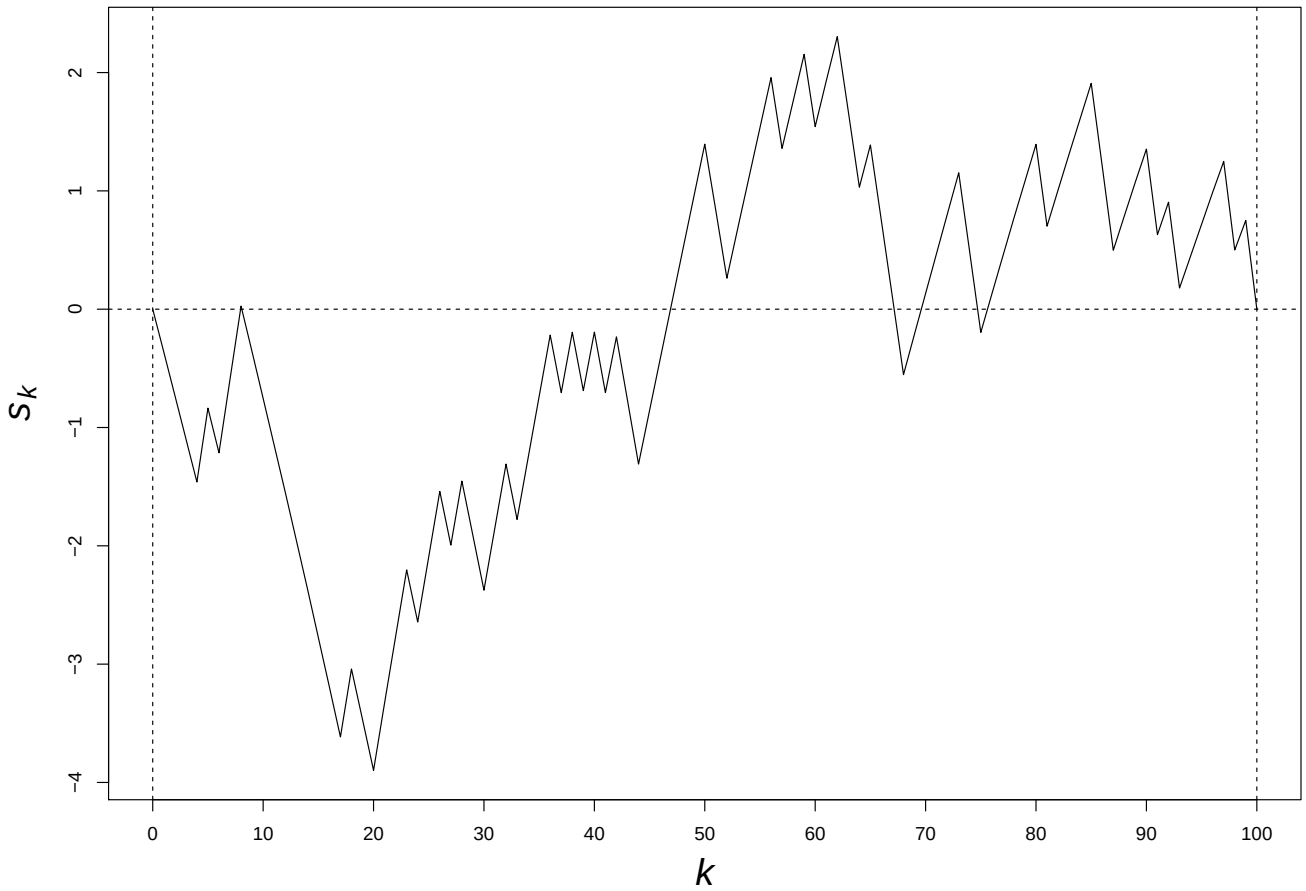
A typical output

$$\chi - squared = 10.714, df = 8, p - value = 0.2184$$

Obvious idea

Since we have $\sum_{i=1}^{n}(y_i - \widehat{p}_i) = 0$,

let's look at $s_k = \sum_{i=1}^{k}(y_i - \widehat{p}_i)$, $k \in \{1, 2, \ldots, n\}$.



Slightly changed idea

we're looking at

$$s_k = \sum_{i=1}^{k} \frac{(y_i - \widehat{p}_i)}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}, k \in \{1, 2, \ldots, n\}.$$

Slightly changed idea

we're looking at

$$s_k = \sum_{i=1}^{k} \frac{(y_i - \widehat{p}_i)}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}, k \in \{1, 2, \ldots, n\}.$$

First results say we're doing slightly better than the existing test, but it is too early to make claims.

Conclusions

- Several GOF tests for linear regression which are based on cumsum processes have been proposed
- Our proposed test for linear regression has better size under homoscedastic (even non-normal) errors (but can be (slightly) too liberal with heteroscedastic errors)
- Our proposed test is at least as powerful as the other existing alternatives
- As for logistic regression: one should stop using Hosmer-Lemeshow test :)