Statistics 191: Introduction to Applied Statistics

Diagnostics for simple linear regression

Jonathan Taylor
Department of Statistics
Stanford University

February 22, 2010

Outline

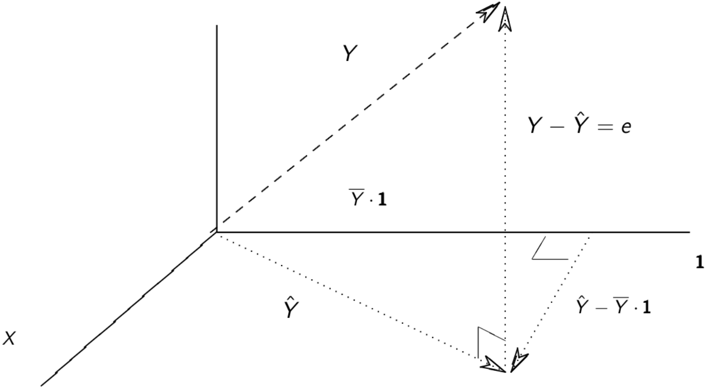
Diagnostics for simple regression

- Goodness of fit of regression: analysis of variance.
- F-statistics.
- Residuals.
- Diagnostic plots.

Geometry of Least Squares

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Goodness of fit

Jonathan Taylor
Department of Statistics
Stanford University

Sums of squares

$$SSE = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

$$SSR = \sum_{i=1}^{n} (\overline{Y} - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (\overline{Y} - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = SSE + SSR$$

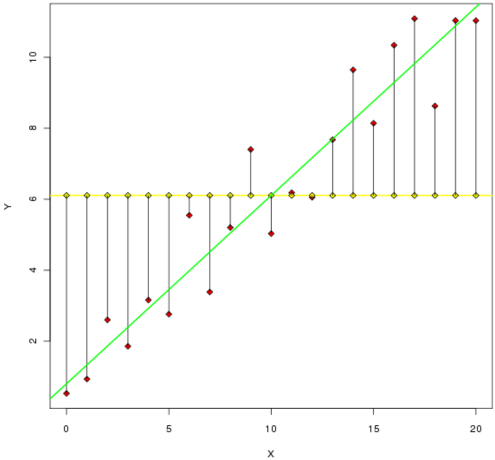
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \widehat{Cor}(\boldsymbol{X}, \boldsymbol{Y})^2.$$

Basic idea: if R^2 is large: a lot of the variability in $\boldsymbol{Y}$ is explained by $\boldsymbol{X}$.

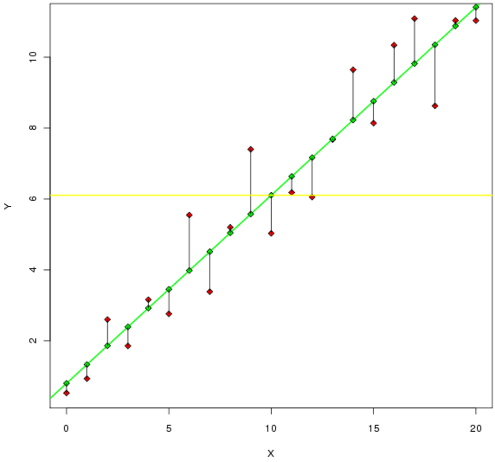
Total sum of squares

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



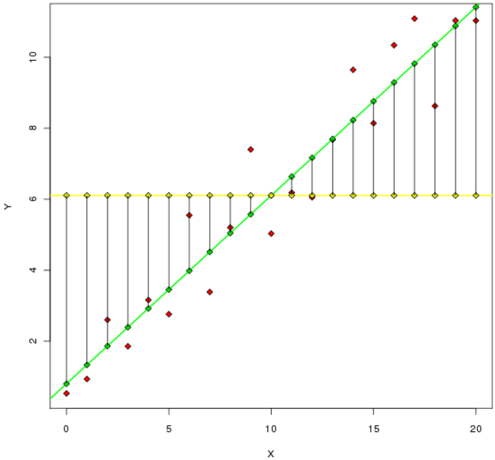
Error sum of squares



Regression sum of squares

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



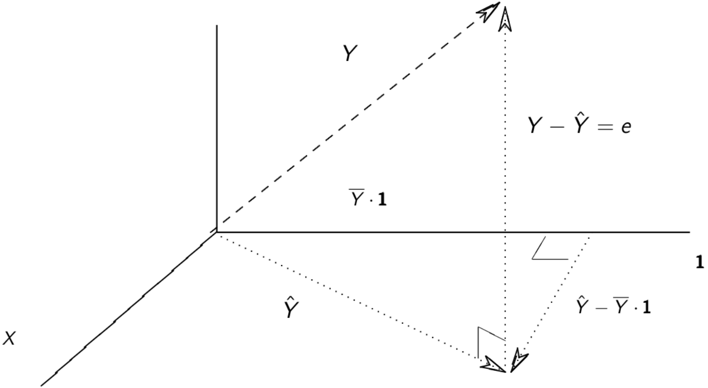
F-statistics

Jonathan Taylor
Department of Statistics
Stanford University

What is an F-statistic?

- An F-statistic is a ratio of "*sample variances (mean squares)*": it has a numerator, N, and a denominator, D that are independent.
- Let
$$N \sim \frac{\chi^2_{\rm num}}{df_{\rm num}}, \qquad D \sim \frac{\chi^2_{\rm den}}{df_{\rm den}}$$
and define
$$F = \frac{N}{D}.$$
- We say F has an F distribution with parameters $df_{\rm num}, df_{\rm den}$ and write $F \sim F_{df_{\rm num}, df_{\rm den}}$.

Geometry of Least Squares



F-statistic in simple linear regression

Goodness of fit F-statistic

- The ratio
$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$
can be thought of as a *ratio of "variances"*.
- In fact, under $H_0 : \beta_1 = 0$,
$$F \sim F_{1,n-2}$$
because
$$SSR = \|\widehat{\boldsymbol{Y}} - \overline{Y} \cdot \mathbf{1}\|^2$$
$$SSE = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2$$
and from our picture, these vectors are orthogonal.

F and t statistics

Relation between F and t

- If $T \sim t_\nu$, then

$$T^2 \sim \frac{N(0,1)^2}{\chi^2_\nu/\nu} \sim \frac{\chi^2_1/1}{\chi^2_\nu/\nu}.$$

- In other words, the square of a t-statistic is an F-statistic. Because it is always positive, an F-statistic has no *direction* ($\pm$) associated with it.
- In fact, (see R code)

$$F = \frac{MSR}{MSE} = \frac{\widehat{\beta}_1^2}{SE(\widehat{\beta}_1)^2}.$$

F-statistics in regression models

Jonathan Taylor
Department of Statistics
Stanford University

Interpretation of an F-statistic

- In regression, the numerator is usually a difference in "goodness" of fit of two (nested) models.
- The denominator is $\widehat{\sigma}^2$ – an estimate of σ^2.
- Our example today: the bigger model is the simple linear regression model, the smaller is the model with constant mean (one sample model).
- If the F is large, it says that the "bigger" model explains a lot more variability in $\boldsymbol{Y}$ (relative to σ^2) than the smaller one.

F-test in simple linear regression

Example in more detail

- *Full (bigger) model :*

$$FM: \qquad Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- *Reduced (smaller) model:*

$$RM: \qquad Y_i = \beta_0 + \varepsilon_i$$

- The F-statistic has the form

$$F = \frac{(SSE(RM) - SSE(FM))/(df_{RM} - df_{FM})}{SSE(FM)/df_{FM}}.$$

- Reject $H_0 : RM$ is correct, if $F > F_{1-\alpha,1,n-2}$.

Diagnostics

What are the assumptions

- $$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
- Errors ε_i are assumed independent $N(0, \sigma^2)$.

Diagnostics

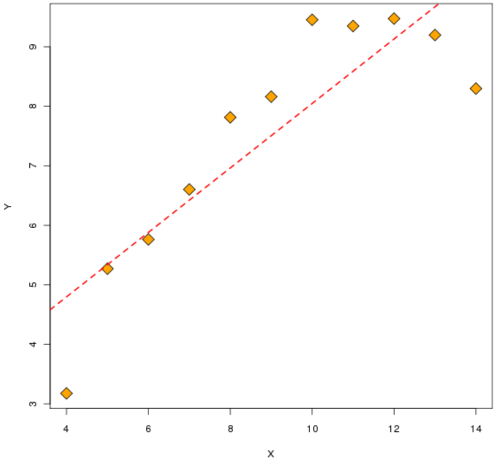
What can go wrong?

- Regression function can be wrong: maybe regression function should be quadratic (see R code).
- Model for the errors may be incorrect:
  - may not be normally distributed.
  - may not be independent.
  - may not have the same variance.
- Detecting problems is more *art* then *science*, i.e. we cannot *test* for all possible problems in a regression model.
- Basic idea of diagnostic measures: if model is correct then residuals $e_i = Y_i - \widehat{Y}_i, 1 \leq i \leq n$ should look like a sample of (not quite independent) $N(0, \sigma^2)$ random variables.

A bad simple linear regression model

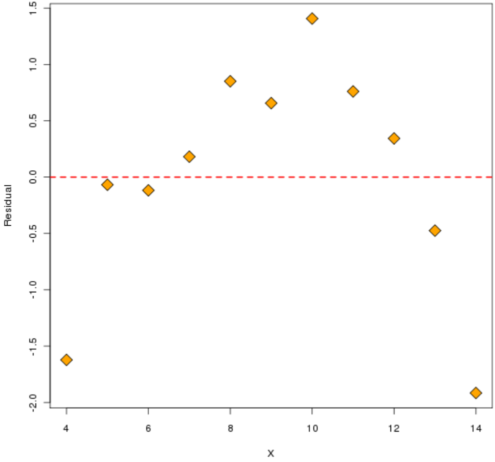
Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Residuals from linear model

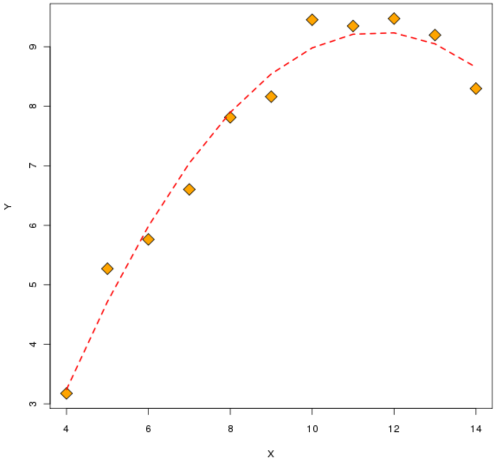
Jonathan Taylor
Department of Statistics
Stanford University



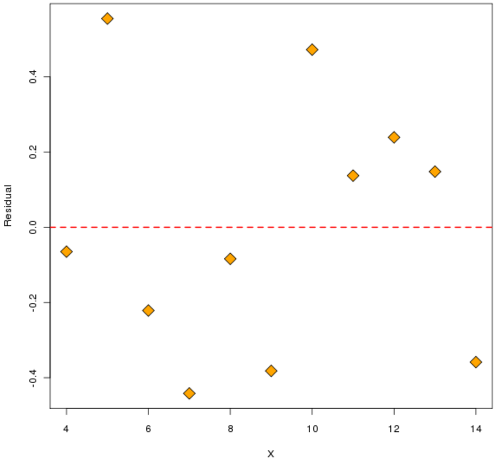
Quadratic model

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University



Residuals from quadratic model



Problems with the errors

Possible problems & diagnostic checks

- Errors may not be normally distributed or may not have the same variance – `qqnorm` can help with this.
- Variance may not be constant. Can also be addressed in a plot of $\boldsymbol{X}$ vs. $\boldsymbol{e}$: *fan shape* or other trend indicate non-constant variance.
- Outliers: points where the model really does not fit! Possibly mistakes in data transcription, lab errors, who knows? Should be recognized and (hopefully) explained.

Non-normality

Jonathan Taylor
Department of Statistics
Stanford University

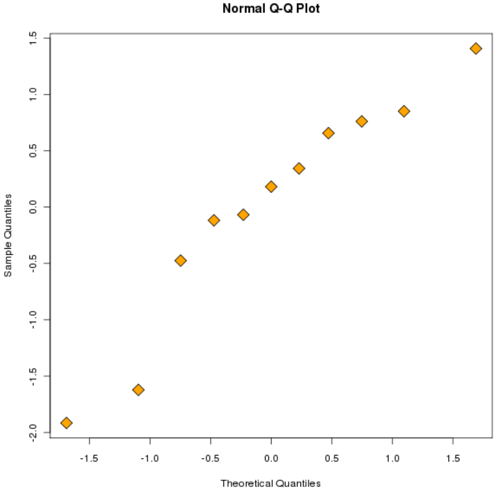
`qqnorm`

- If $e_i, 1 \leq i \leq n$ were really a sample of $N(0, \sigma^2)$ then their sample quantiles should be close to the sample quantiles of the $N(0, \sigma^2)$ distribution.
- Plot:

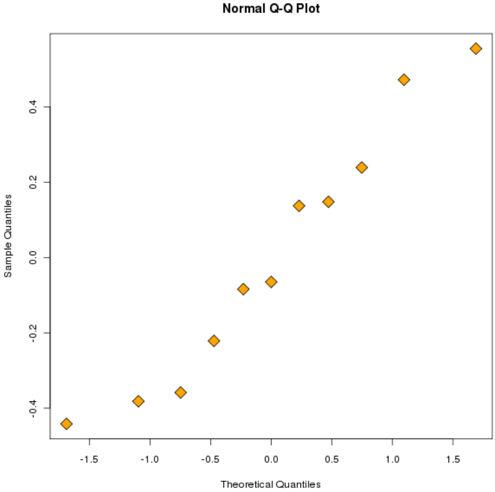
$$e_{(i)} \text{ vs. } \mathbb{E}(\varepsilon_{(i)}), \qquad 1 \leq i \leq n.$$

  where $e_{(i)}$ is the i-th smallest residual (order statistic) and $\mathbb{E}(\varepsilon_{(i)})$ is the expected value for independent ε_i's $\sim N(0, \sigma^2)$.

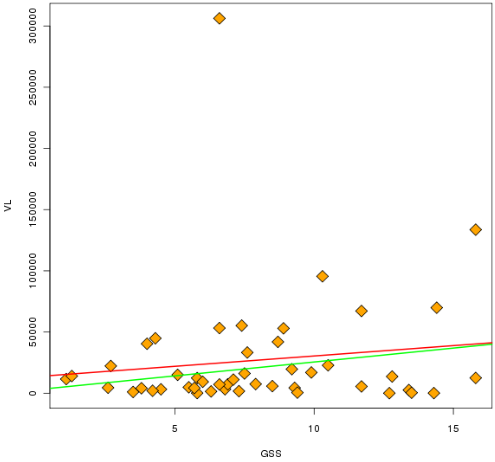
QQplot of residuals from linear model



QQplot of residuals from quadratic model



Outlier and nonconstant variance



Outlier and nonconstant variance