Lecture 8: Diagnostics for Simple linear regression

Pratheepa Jeganathan

10/09/2019

Recap

- What is a regression model?
- Descriptive statistics – graphical
- Descriptive statistics – numerical
- Inference about a population mean
- Difference between two population means
- Some tips on R

Recap

- Simple linear regression (covariance, correlation, estimation, geometry of least squares)
- Inference on simple linear regression model

Diagnostics for simple linear regression

- Goodness of fit of regression: analysis of variance.
- F-statistics.
- Residuals.
- Diagnostic plots for simple linear regression (graphical methods).

The full model

- The full regression model is

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- The β_0 coefficient represents the intercept.
- The β_1 coefficient represents the slope.
- The vector $\hat{\boldsymbol{Y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix}$ is the vector of fitted values in the above model.

The reduced model

- The reduced regression model

$$Y = \beta_0 + \epsilon.$$

- The β_0 coefficient represents the intercept.
- Since $\beta_1 = 0$, we have assumed there is no slope.
- The vector $\bar{\boldsymbol{Y}} = \begin{pmatrix} \bar{Y} \\ \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}$ is the vector of fitted values in the above model.

Goodness of fit

- The closer $\hat{\boldsymbol{Y}}$ is to the $\bar{\boldsymbol{Y}}$, the less "variation" there is along the X.
- The closeness can be measured by the length of the vector $\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}$.
- The square of a vector's length is the sum of its elements squared. These quantities are usually referred to as *sums of squares*.

Sums of squares

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

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

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

- The quantity SSE, or *error sum of squares*, is the squared length of the vector $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$.
- The quantity SSR, or *regression sum of squares*, is the length of the vector $\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}$.
- The quantity SST, or *total sum of squares*, is the length of the vector $\boldsymbol{Y} - \bar{\boldsymbol{Y}}$.

Coefficient of determination R^2

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

- The quantity R^2 is a measure of the goodness of fit of the simple linear regression model. Values near 1 indicate much of the total variability in Y is explained by the regression model.

Adjusted R^2

- SSR increases with p so R^2 can be artificially large.
- $$\text{Adjusted } R_a^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)},$$

  where p is the number of predictors.
- For simple linear regression $p = 1$ so

$$\text{Adjusted } R_a^2 = 1 - \frac{\text{SSE}/(n-2)}{\text{SST}/(n-1)}.$$

- R_a^2 cannot be interpreted as proportion of total variation in Y accounted for by the predictors.
- R_a^2 is sometimes used to compare models with different predictor variables.
- R_a^2 can decrease as p increases.

Mean squares

- Each sum of squares gets an extra bit of information associated to them, called their *degrees of freedom*.
- Roughly speaking, the *degrees of freedom* can be determined by dimension counting.
- The SSE has $n - 2$ degrees of freedom.
- The SST has $n - 1$ degrees of freedom.
- The SSR has 1 degree of freedom.

Mean squares

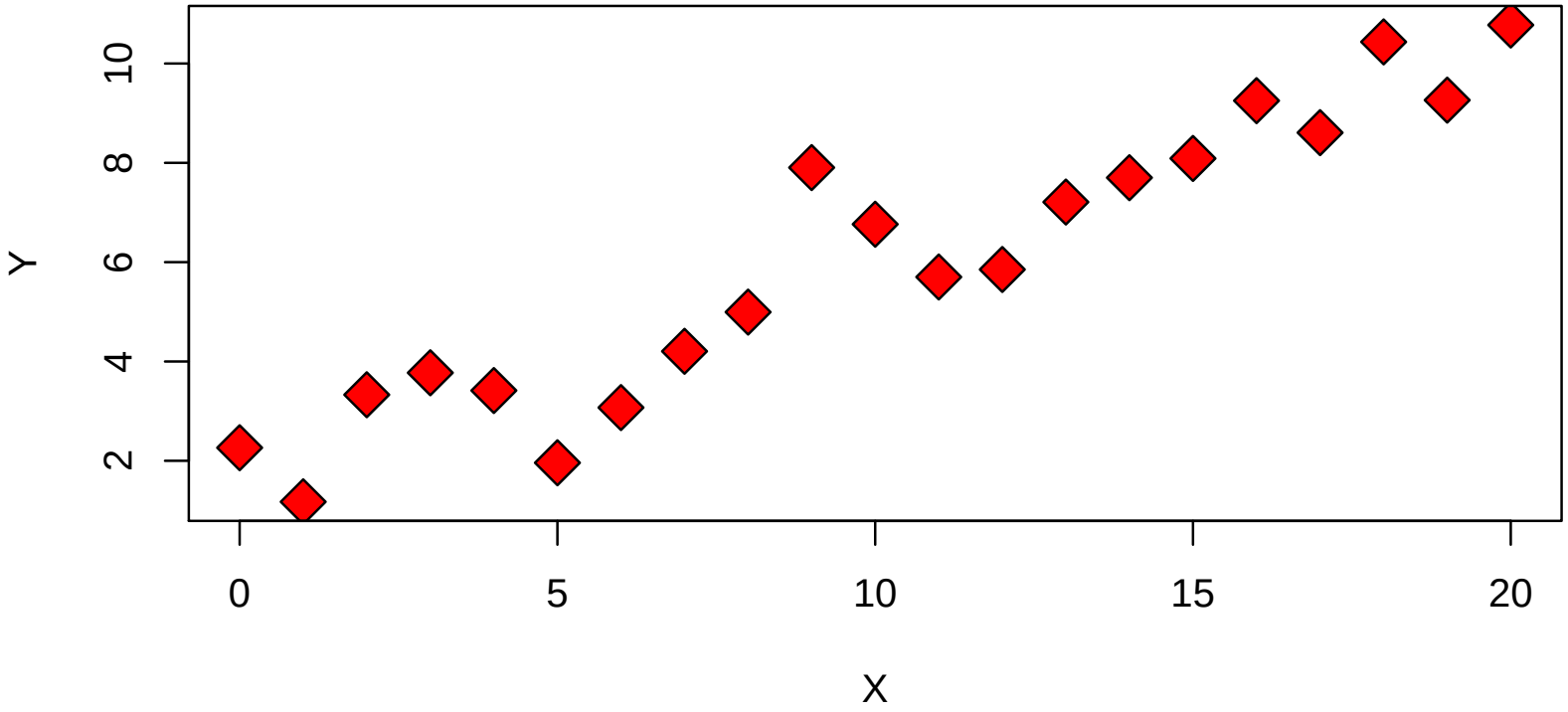
$$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

$$MSR = \sum_{i=1}^{n}(\overline{Y} - \widehat{Y}_i)^2$$

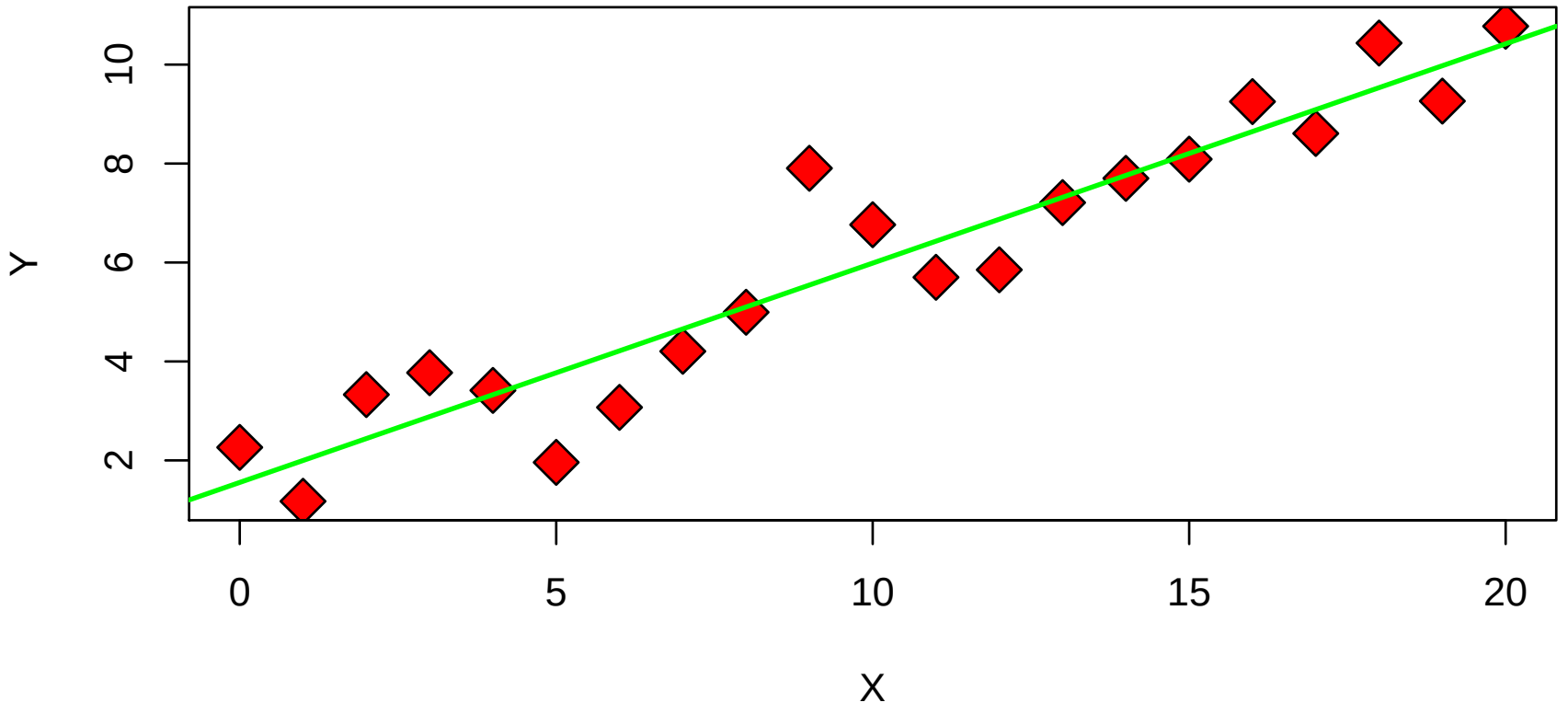
$$MST = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Visualization of sum of squares

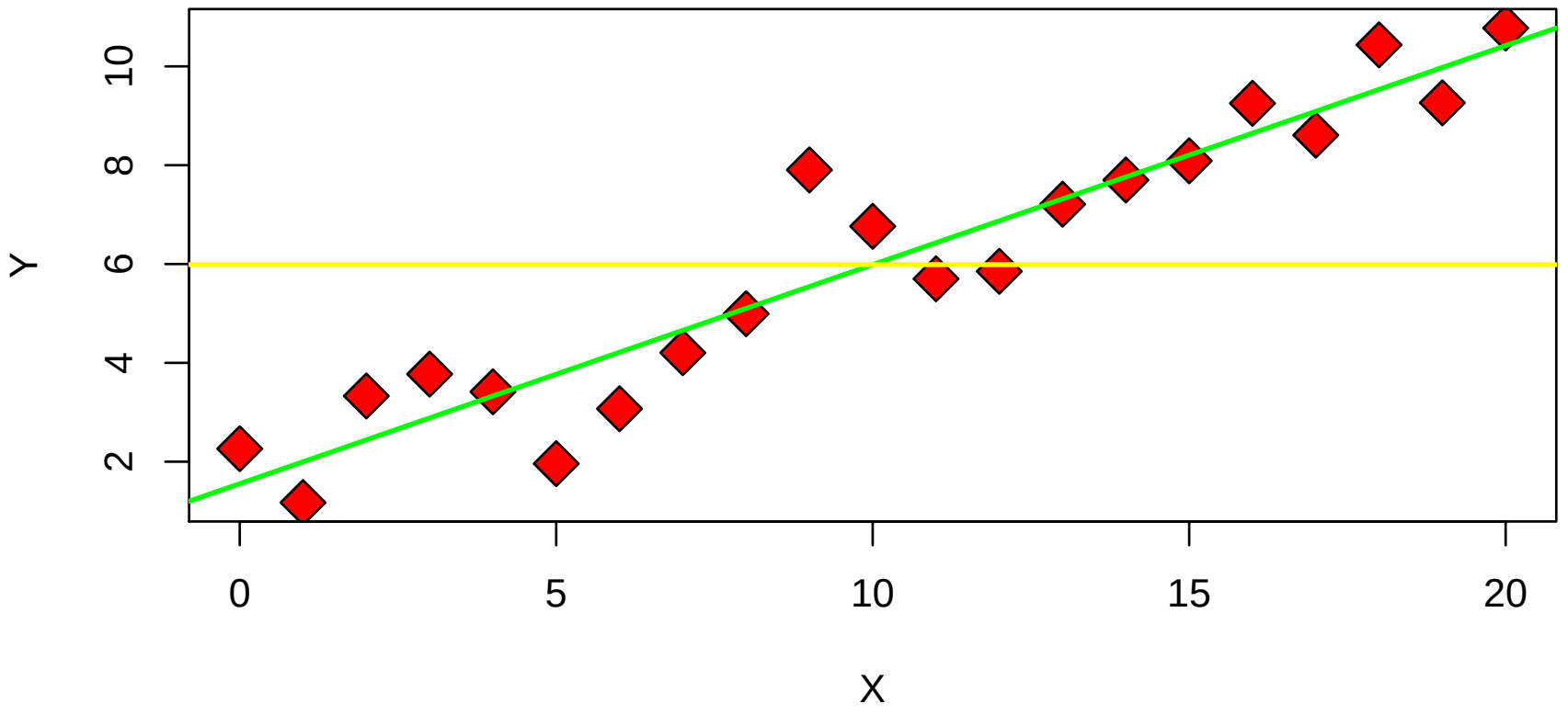
- These sums of squares can be visualized as follows:
- We will illustrate with a synthetic data set.



Fit the model

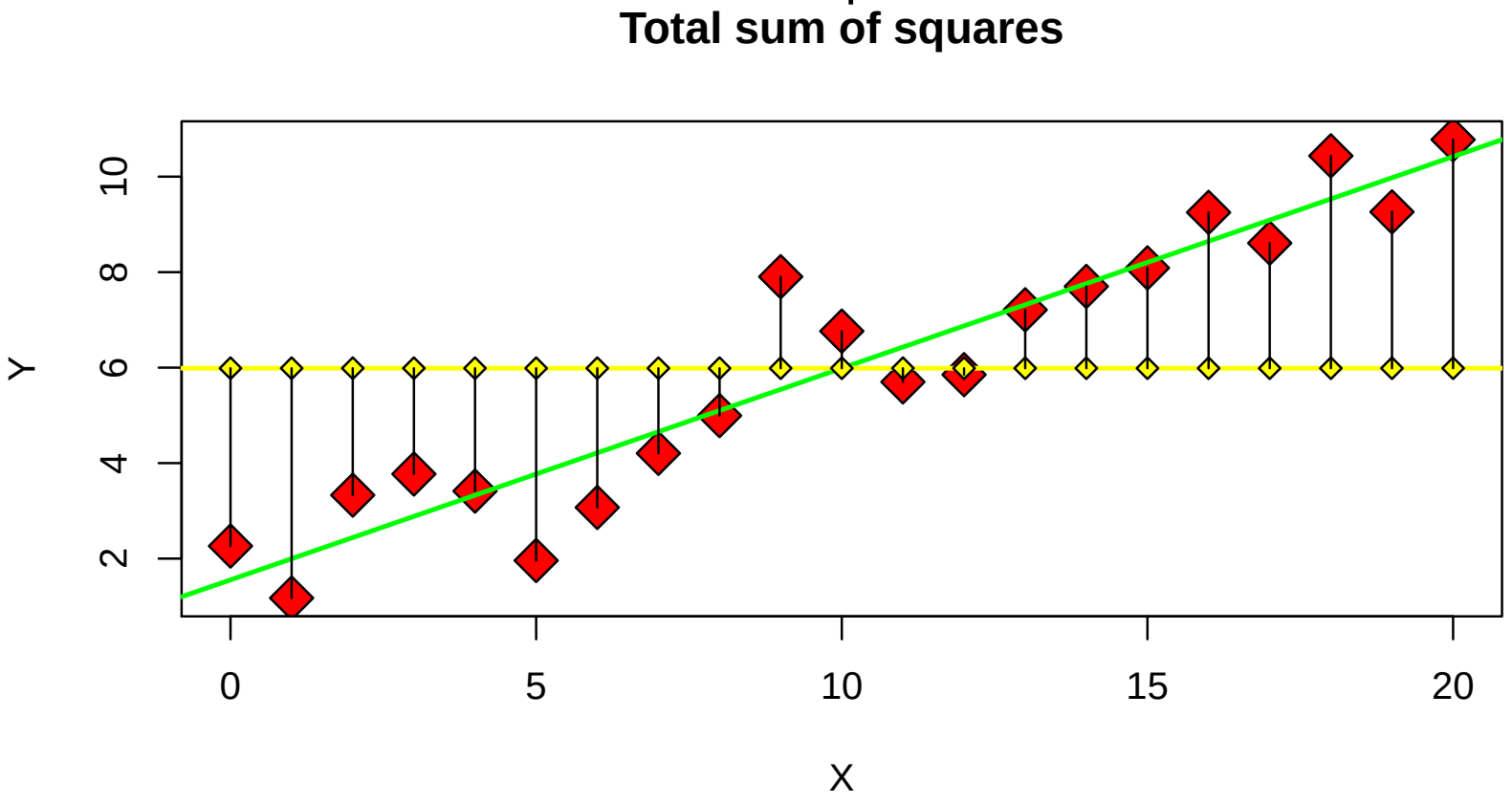


Add mean of Y to the plot



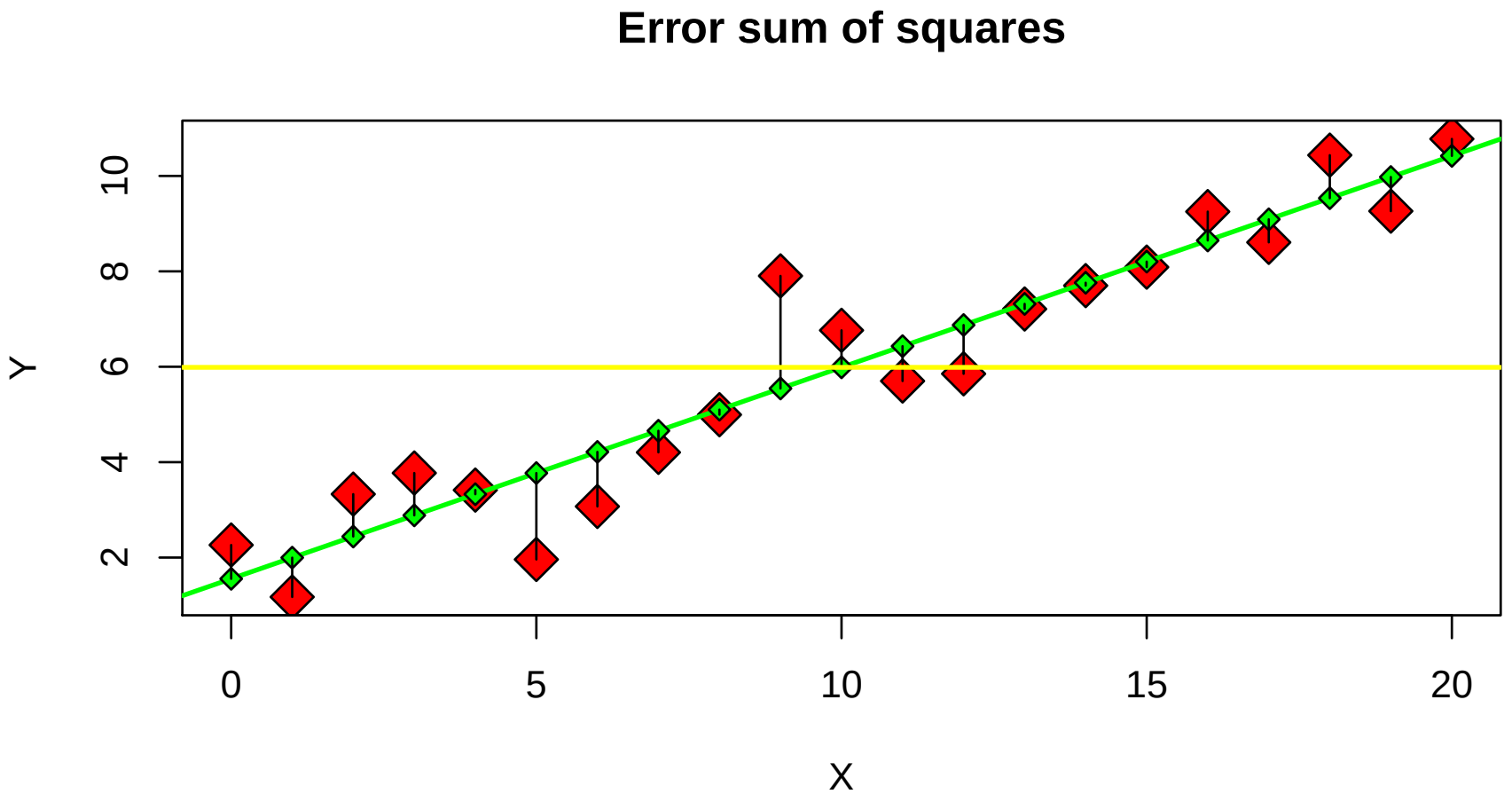
SST

- The total sum of squares, SST: sum of the squared differences between the Y values and the sample mean of the Y values.



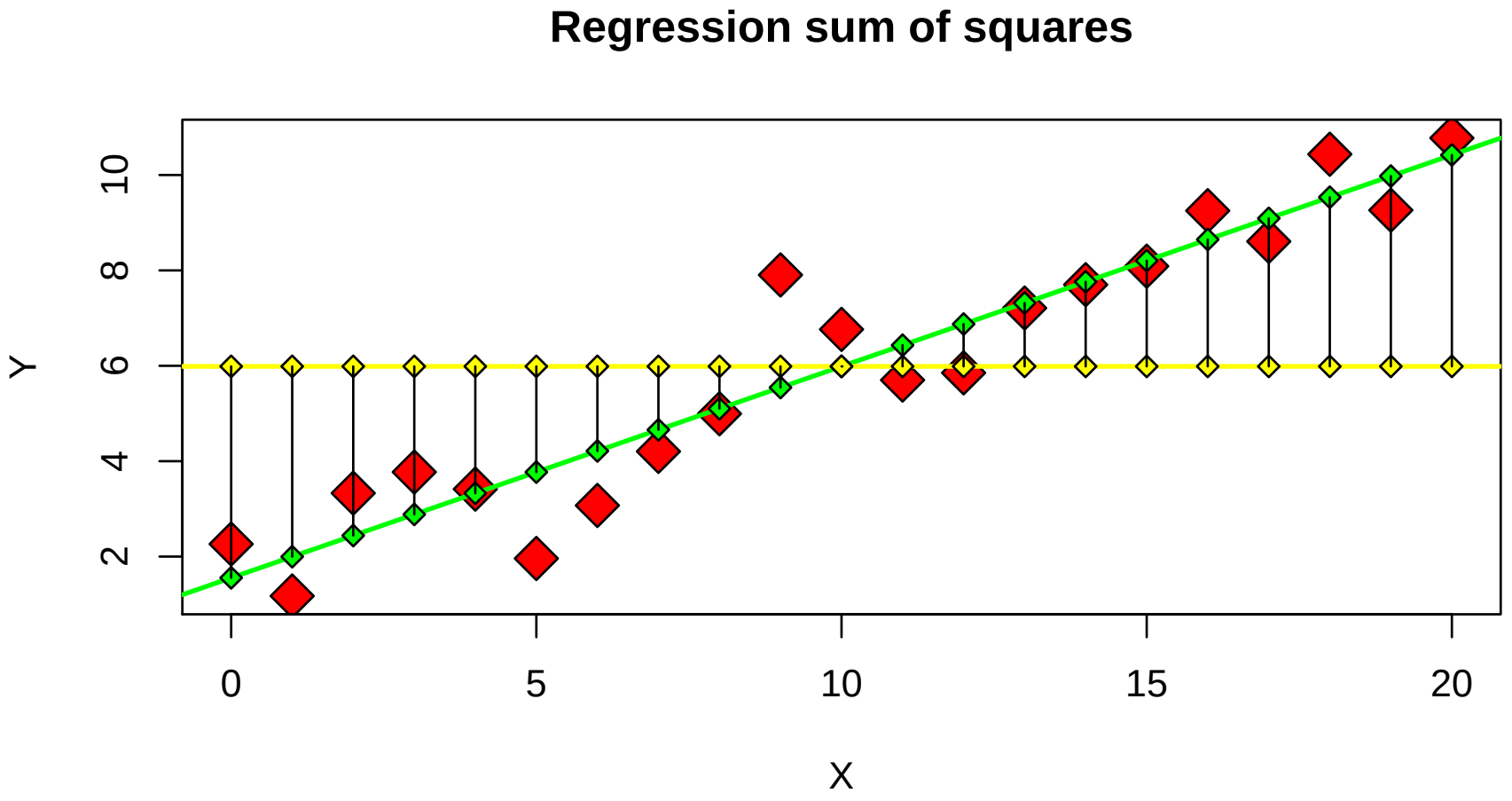
SSE

- The error sum of squares, SSE: sum of the squared differences between the Y values and the $\hat{Y}$ values, i.e. the fitted values of the regression model.



SSR

- Finally, the regression sum of squares, SSR: sum of the squared differences between the $\hat{Y}$ values and the sample mean of the Y values.



Definition of R^2

- As noted above, if the regression model fits very well, then SSR will be large relative to SST.
- The R^2 score is just the ratio of these sums of squares.
- We'll verify this on the `wages` data.

```
url = "http://stats191.stanford.edu/data/wage.csv"
wages = read.table(url, sep = ",", header = T)
wages.lm = lm(logwage ~ education, data=wages)
```

Let's verify our claim SST = SSE + SSR:

```
SSE = sum(resid(wages.lm)^2)
SST = sum((wages$logwage - mean(wages$logwage))^2)
SSR = sum((mean(wages$logwage) - predict(wages.lm))^2)
data.frame(SST, SSE + SSR)
```

```
##         SST SSE...SSR
## 1 410.2148  410.2148
```

F-statistics

```
##
## Call:
## lm(formula = logwage ~ education, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78239 -0.25265  0.01636  0.27965  1.61101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.239194   0.054974   22.54   <2e-16 ***
## education   0.078600   0.004262   18.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.4038 on 2176 degrees of freed
## Multiple R-squared:  0.1351, Adjusted R-squared:  0.1347
## F-statistic:   340 on 1 and 2176 DF,  p-value: < 2.2e-16
```

- The R^2 is also closely related to the F statistic reported as the goodness of fit in *summary* of *lm*.

```
F = (SSR / 1) / (SSE / wages.lm$df)
print(F)
```

```
## [1] 340.0297
```

- In other words, for simple linear regression that `F-statistic` is

$$F = \frac{(n-2) \cdot R^2}{1 - R^2}$$

where $n - 2$ is `wages.lm$df`.

```
(nrow(wages)-2)*0.1351 / (1 - 0.1351)
```

```
## [1] 339.8978
```

- Finally, $R = \sqrt{R^2}$ is called the (absolute) *correlation coefficient* because it is equal to the absolute value of sample correlation coefficient of X and Y.

```
round(cor(wages$education, wages$logwage)^2, digits = 2)
```

```
## [1] 0.14
```

F-statistics

- After a t-statistic, the next most commonly encountered statistic is a χ^2 statistic, or its closely related cousin, the F statistic.
- Roughly speaking, an F-statistic is a ratio of two scaled sum of squares: it has a numerator, N, and a denominator, D that are independent.
- Let

$$N \sim \frac{\chi^2_{df_{\text{num}}}}{df_{\text{num}}}, \qquad D \sim \frac{\chi^2_{df_{\text{den}}}}{df_{\text{den}}}$$

  and define

$$F = \frac{N}{D}.$$

- We say F has an F distribution with parameters $df_{\text{num}}, df_{\text{den}}$ and write $F \sim F_{df_{\text{num}}, df_{\text{den}}}$

F statistic for simple linear regression

- The ratio

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{(SST - SSE)/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

  can be thought of as a *ratio of a difference in sums of squares normalized by our "best estimate" of variance* .

- In fact, under $H_0 : \beta_1 = 0$,

$$F \sim F_{1,n-2}$$

  because SSR has 1 degrees of freedom and SST has $n-2$ degrees of freedom.

- The null hypothesis $H_0 : \beta_1 = 0$ implies that $SSR \sim \chi^2_1 \cdot \sigma^2$.

Relation between F and t statistics.

- 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* associated with it.
- Let's check this in our example.

```r
summary(wages.lm)
```

```
##
## Call:
## lm(formula = logwage ~ education, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78239 -0.25265  0.01636  0.27965  1.61101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.239194   0.054974   22.54   <2e-16 ***
## education   0.078600   0.004262   18.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.4038 on 2176 degrees of freed
## Multiple R-squared:  0.1351, Adjusted R-squared:  0.1347
## F-statistic:   340 on 1 and 2176 DF,  p-value: < 2.2e-16
```

- The t statistic for *education* is the t-statistic for the parameter β_1 under $H_0 : \beta_1 = 0$.
- Its value is 18.44 above. If we square it, we should get about the same as the *F-statistic*.

```
18.44^2
```

```
## [1] 340.0336
```

Interpretation of an F-statistic

- In regression, the numerator is usually a difference in *goodness of fit* of two (nested) models.
- The denominator is $\hat{\sigma}^2$: an estimate of σ^2.
- In 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 Y (relative to σ^2) than the smaller one.

Analysis of variance

- The equation
$$SST = SSE + SSR$$
is a *decomposition* of the total variability into separate pieces.
- This decomposition is often referred to as an **analysis of variance (ANOVA)**.

The F-statistic for simple linear regression revisited

- The F statistic should compare two models. What are these models?
  - The *full model* would be

$$(FM) \qquad Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

  - The *reduced model* would be

$$(RM) \qquad Y_i = \beta_0 + \varepsilon_i$$

- The F-statistic then has the form

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

The F-statistic for simple linear regression revisited

- The *null hypothesis* is

$$H_0 : \text{reduced model (RM) is correct.}$$

- The usual α rejection rule would be to reject H_0 if the F_{obs} the observed F statistic is greater than $F_{1,n-2,1-\alpha}$.
- In our case, the observed F was 340, $n - 2 = 2176$ and the appropriate 5% threshold is computed below to be 3.85.

```
qf(0.95, 1, 2176)
```

```
## [1] 3.845736
```

- Therefore, we strongly reject H_0.

Diagnostics for simple linear regression

Diagnostics

- While we have used a particular model for our data, it may not be correct. It is important that we have some tools that help us determine whether the model is reasonable or not.

What can go wrong?

- Our model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{V}[\epsilon_i] = \sigma^2$, where $i = 1, 2, \cdots, n$.
  - ϵ_i are independent and identical.
  - For inferences we assume that $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$.

What can go wrong?

- Using a linear regression function can be wrong: maybe regression function should be quadratic.
- We assumed independent Gaussian errors with the same variance. This may be incorrect.
  - The errors may not be normally distributed.
  - The errors may not be independent.
  - The errors may not have the same variance.
- Detecting problems is more *art* than *science*, i.e. we cannot *test* for all possible problems in a regression model.

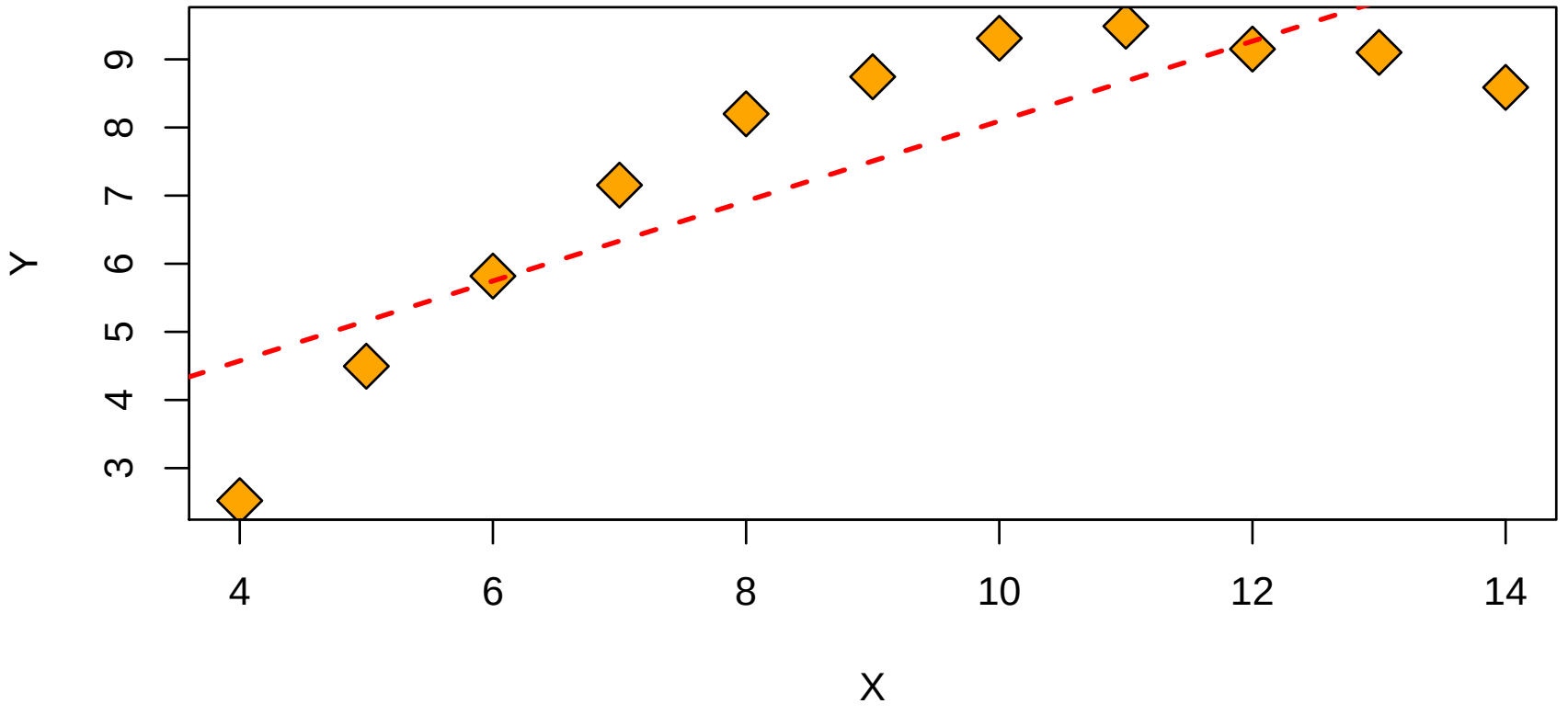
The basic idea of most diagnostic measures is the following. *If the model is correct then residuals* $e_i = Y_i - \widehat{Y}_i, 1 \leq i \leq n$ *should look like a sample of (not quite independent)* $N(0, \sigma^2)$ *random variables.*

A poorly fitting model

- Here is an example of a poorly fitting model.
- It will turn out that there is a simple fix for this data set: a model that includes a quadratic term for X will turn out to have a much better fit.
- Finding this fix in practice can be difficult.

```
y = anscombe$y2 + rnorm(length(anscombe$y2)) * 0.45
x = anscombe$x2
```

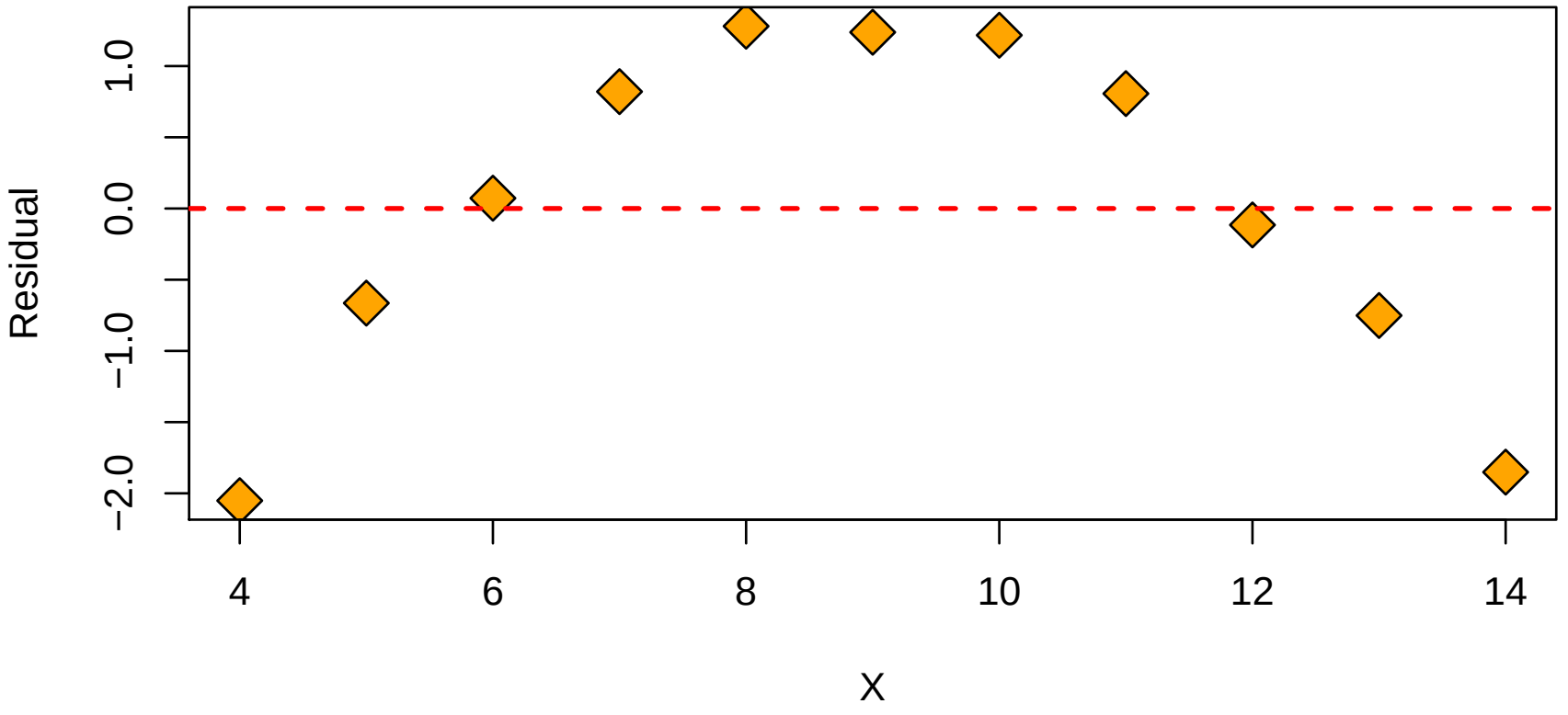
A poorly fitting model



A poorly fitting model (residuals)

- Let's take a look at the residuals from this model.
- Patterns in these residual plots may suggest something like a quadratic effect is missing, but they can also suggest some sort of serial dependence in the random errors.
- We will discuss this later, when we discuss correlated-errors.

A poorly fitting model (residuals)

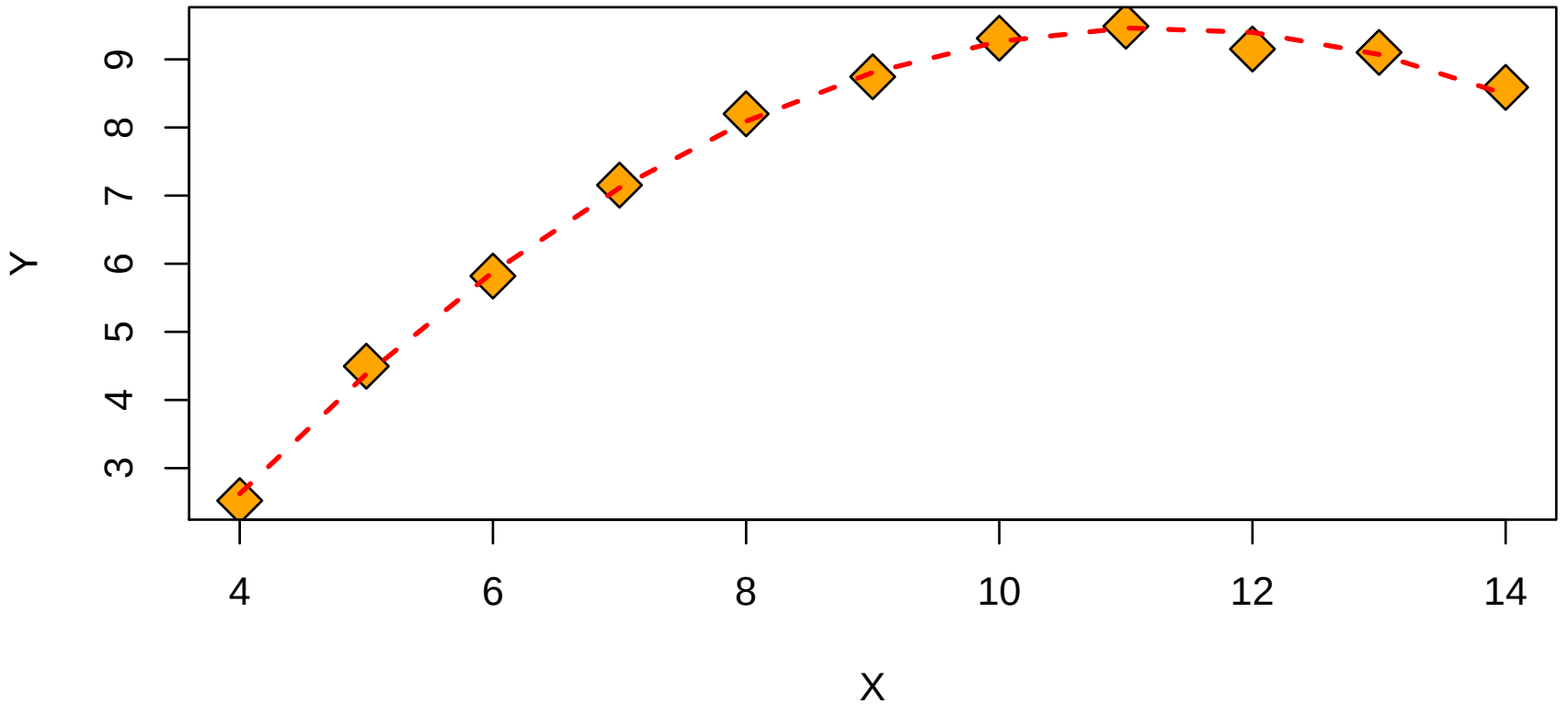


A poorly fitting model (add a quadratic term)

- We will add a quadratic term to our model. This is our first example of a *multiple linear regression model*.

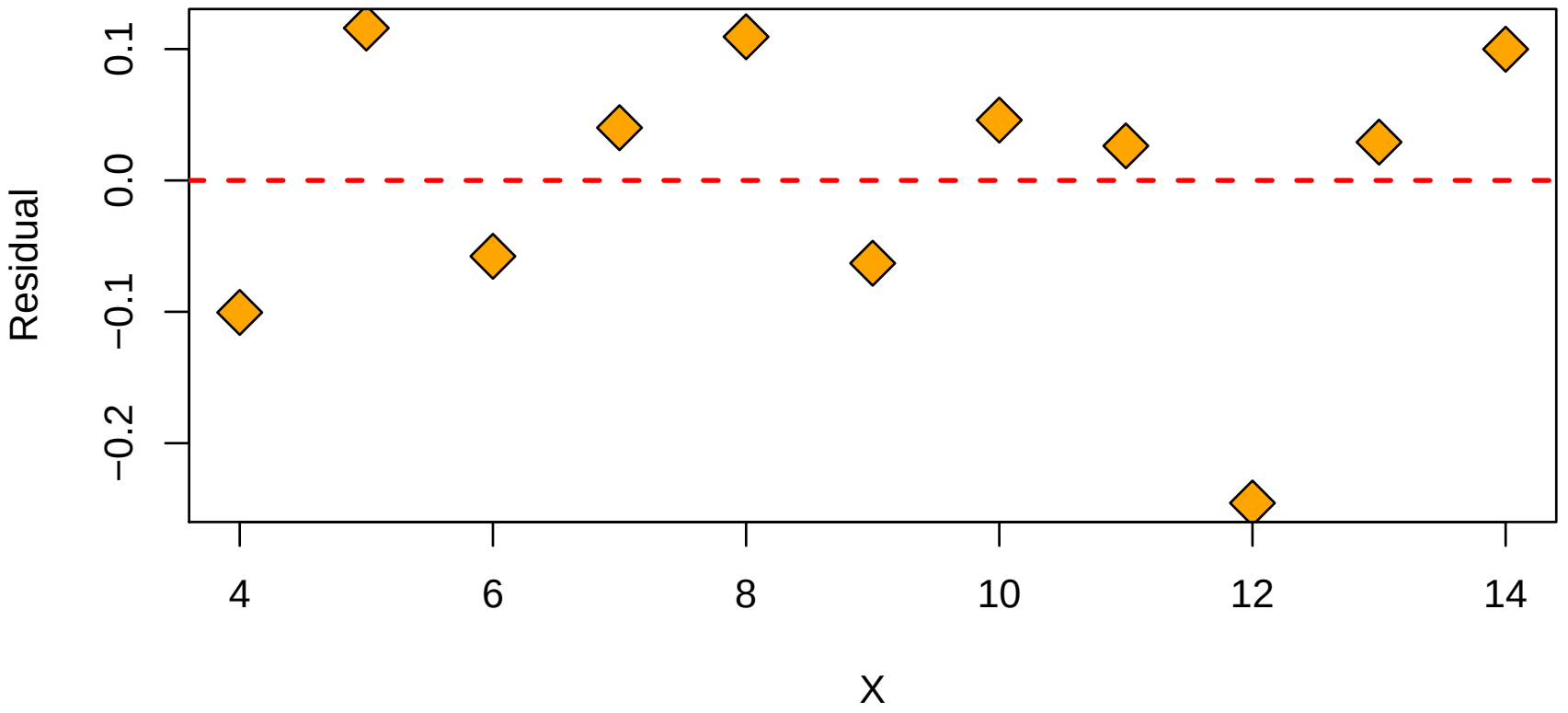
```
quadratic.lm = lm(y ~ poly(x, 2))
Xsort = sort(x)
```

A poorly fitting model (add a quadratic term)



A poorly fitting model (residuals)

- The residuals of the quadratic model have no apparent pattern in them, suggesting this is a better fit than the simple linear regression model.



Assessing normality of errors

- Another common diagnostic plot is the *qqplot* where *qq* stands for *Quantile-Quantile*.
  - Roughly speaking, a qqplot is designed to see if the quantiles of two distributions match.
- The function *qqnorm* can be used to ascertain if a sample of numbers are roughly normally distributed.
  - If the points lie on the diagonal line, this is evidence that the sample is normally distributed.
  - Various departures from the diagonal indicate skewness, asymmetry, etc.
- 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.

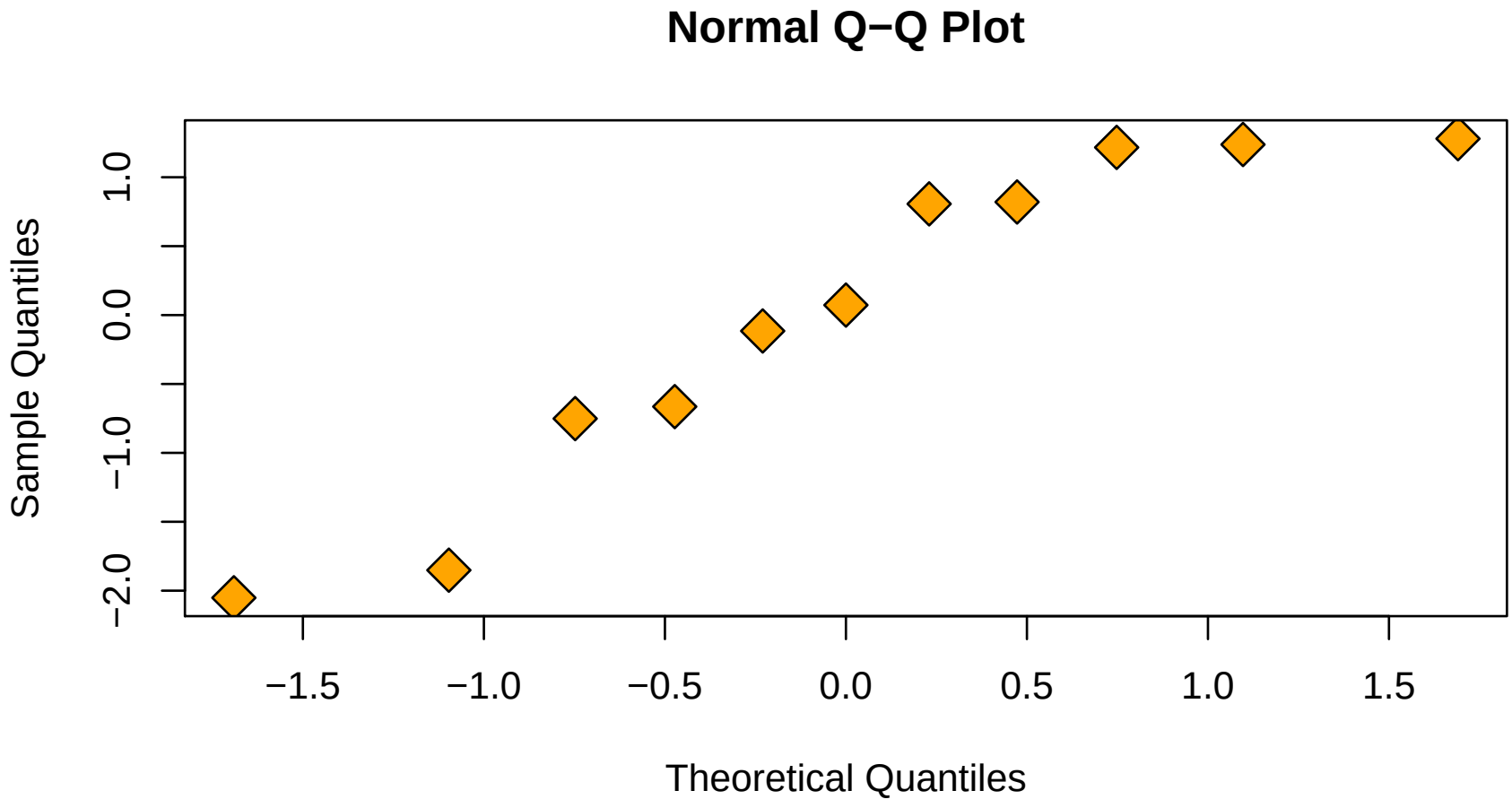
Assessing normality of errors

- The *qqnorm* plot is a plot of

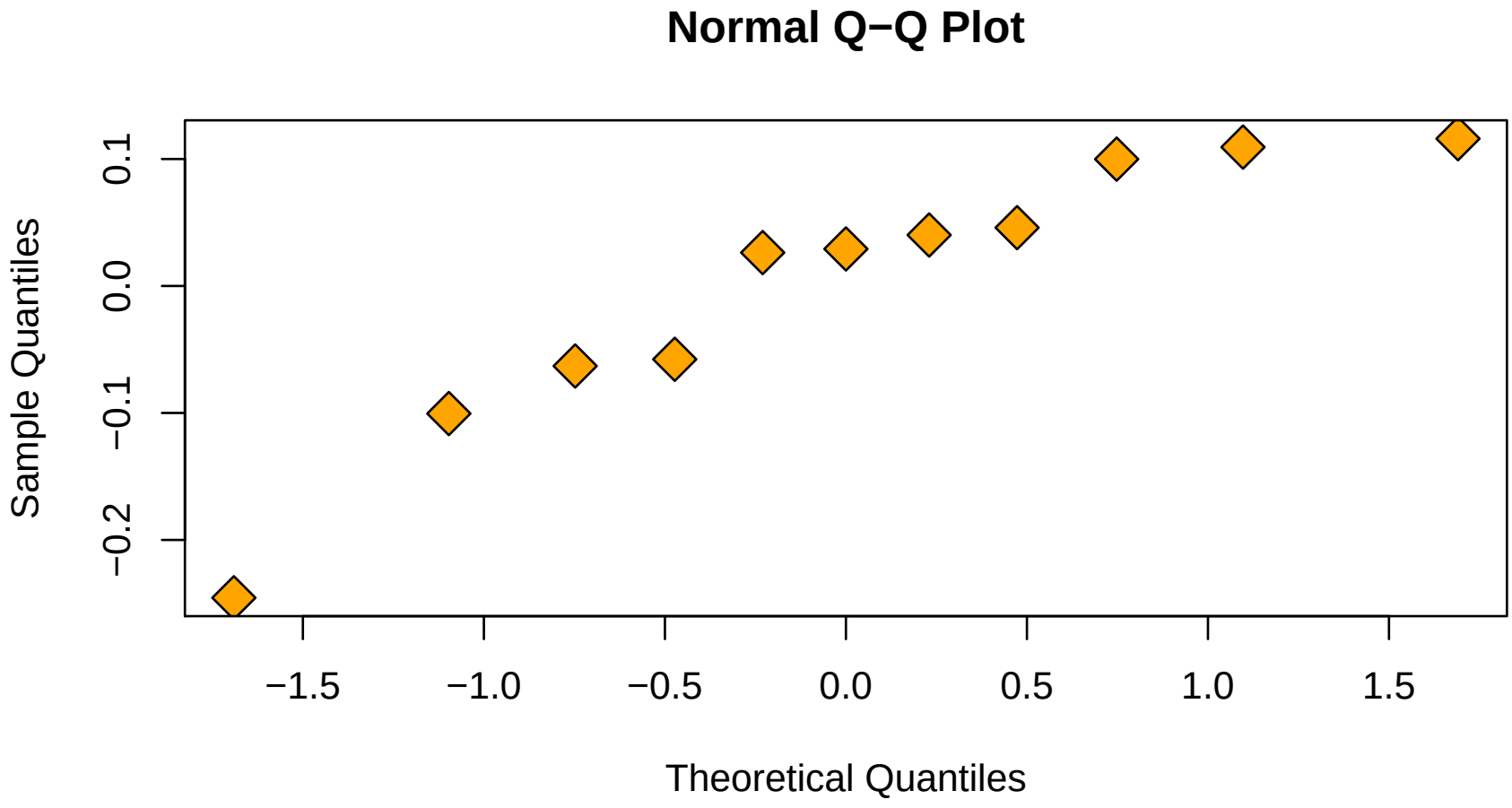
$$e_{(i)} \text{ vs. } \mathbb{E}(\varepsilon_{(i)}), \qquad 1 \leq i \leq n.$$

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 $\sim N(0, \sigma^2)$.

```r
qqnorm(resid(simple.lm), pch = 23,
  bg = "orange", cex = 2)
```



```
qqnorm(resid(quadratic.lm), pch = 23,
  bg = "orange", cex = 2)
```



- In these two examples, the qqplot does not seem vastly different, even though we know the simple model is incorrect in this case.
- This indicates that several diagnostic tools can be useful in assessing a model.

Assessing constant variance assumption

- One plot that is sometimes used to determine whether the variance is constant or not is a plot of X against $e = Y - \hat{Y}$.
- If there is a pattern to the spread in this plot, it may indicate that the variance changes as a function of X.
- In our earlier plots, we noticed a trend in this plot, not necessarily evidence of changing variance.

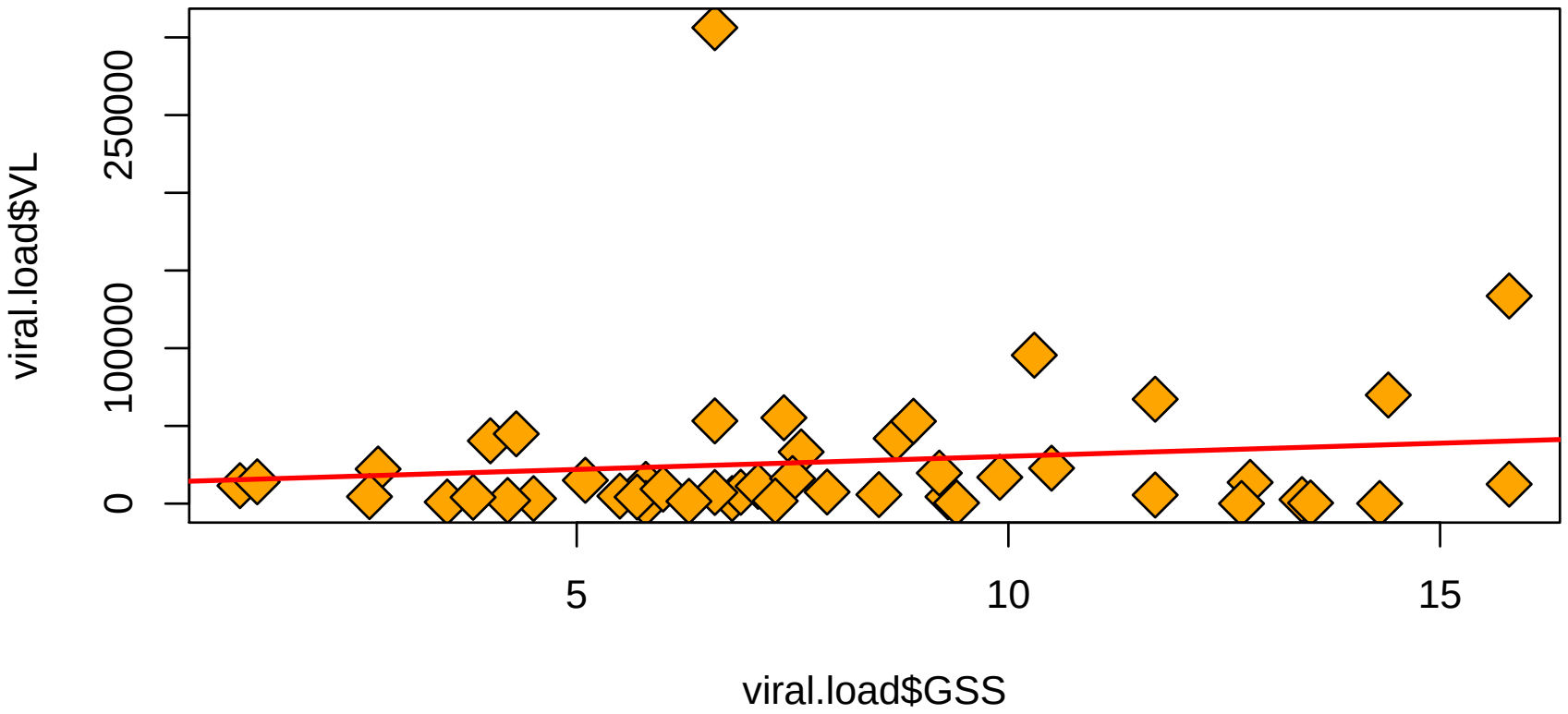
Assessing constant variance assumption

- The data set below, taken from some work done with Dr. Robert Shafer here at Stanford http://hivdb.stanford.edu, plots HIV virus load against a score related to the the genetic makeup of a patient's virus shows clear non-constant variance.
- It also provides a clear example of an outlier, or a point that is a clear departure from the model.

```
url = 'http://stats191.stanford.edu/data/HIV.VL.table'
viral.load = read.table(url, header=T)
```

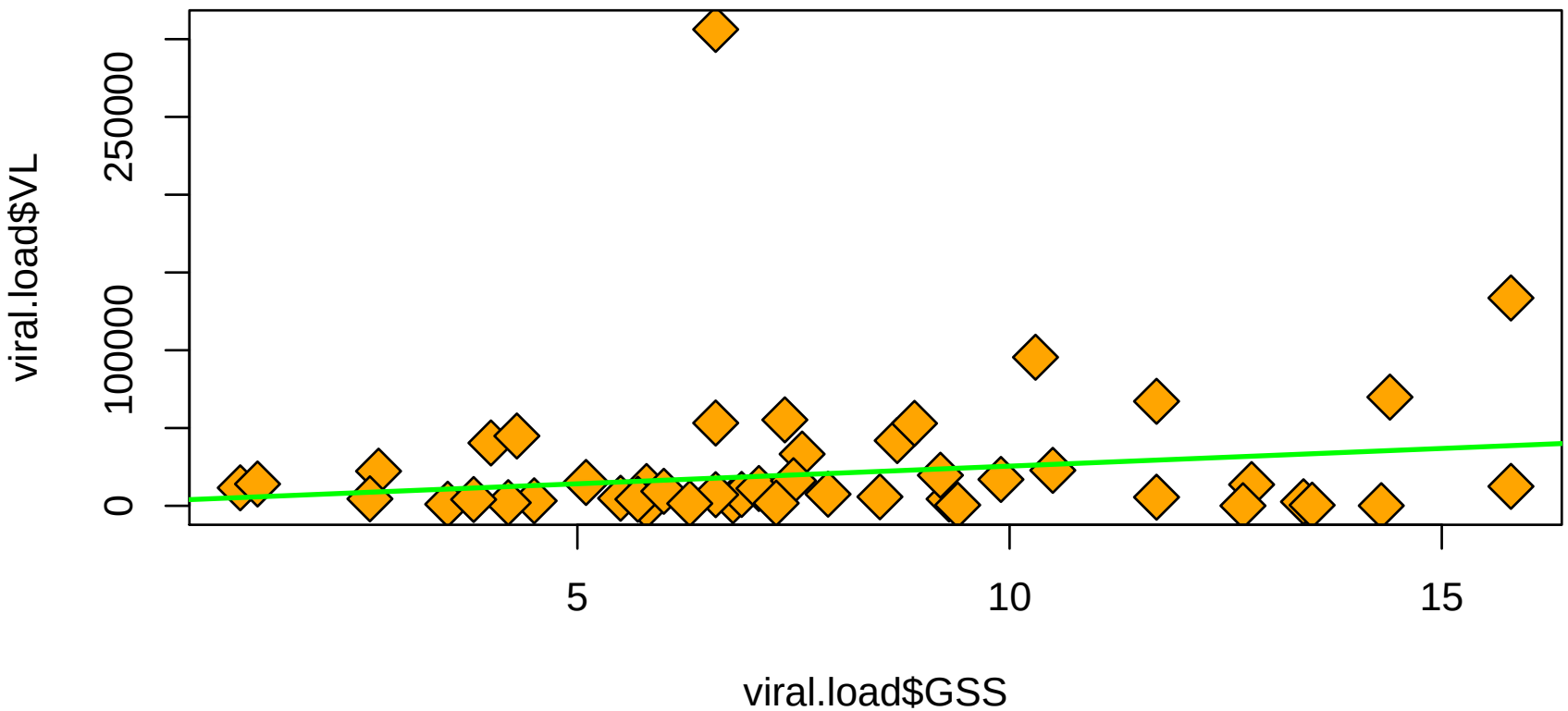
Assessing constant variance assumption

```
plot(viral.load$GSS, viral.load$VL,
  pch=23, bg='orange', cex=2)
viral.lm = lm(VL ~ GSS,
  data = viral.load)
abline(viral.lm, col='red', lwd=2)
```



Assessing constant variance assumption

```
good = (viral.load$VL < 200000)
plot(viral.load$GSS, viral.load$VL,
  pch=23, bg='orange', cex=2)
viral.lm.good = lm(VL ~ GSS,
  data = viral.load, subset=good)
abline(viral.lm.good, col='green', lwd=2)
```

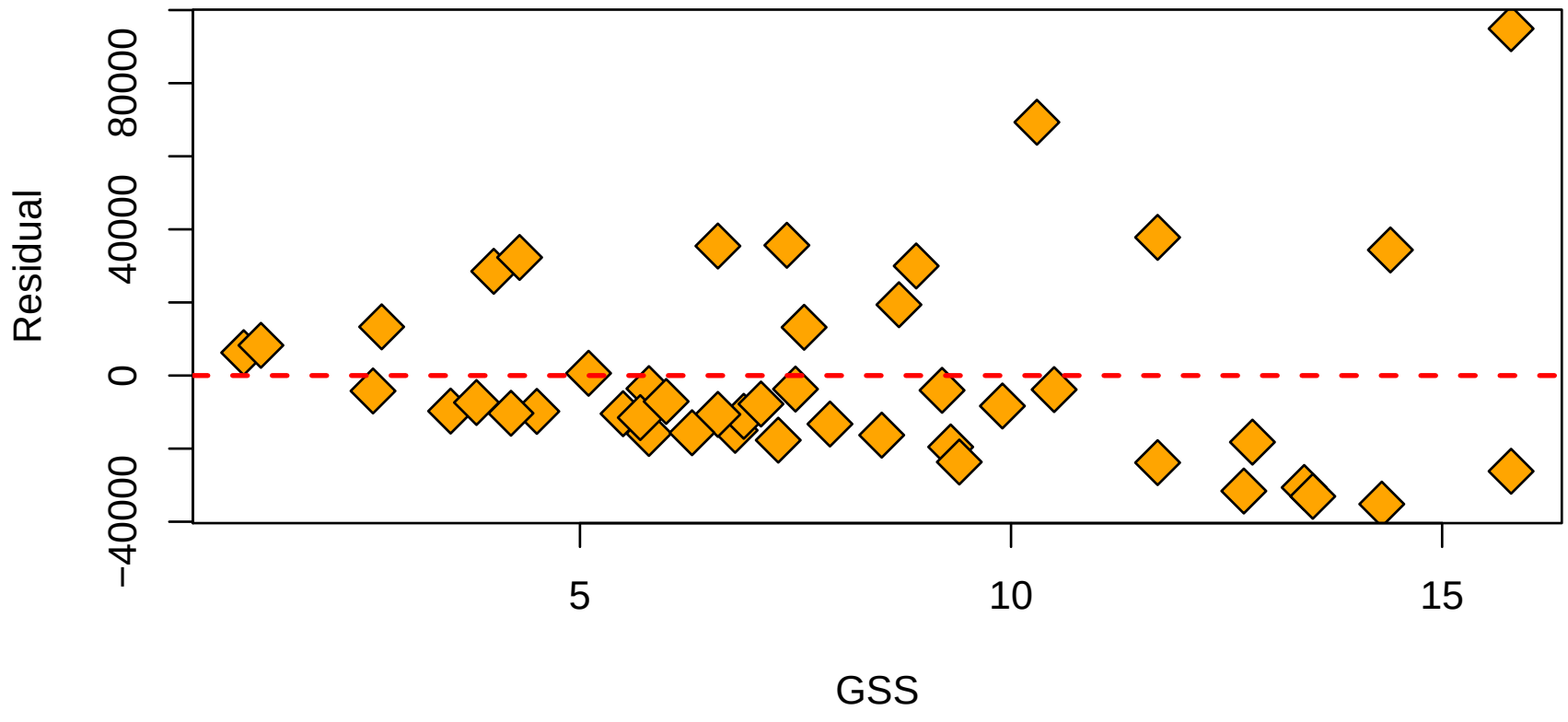


Residual plot

- When we plot the residuals against the fitted values for this model (even with the outlier removed) we see that the variance clearly depends on GSS.
- They also do not seem symmetric around 0 so perhaps the Gaussian model is not appropriate.

Residual plot

```
plot(viral.load$GSS[good], resid(viral.lm.good), pch=23,
     bg='orange', cex=2, xlab='GSS', ylab='Residual')
abline(h=0, lwd=2, col='red', lty=2)
```



Outliers

- Outliers can be obvious to spot (or not) but very difficult to define rigorously.
- Roughly speaking, they points where the model really does not fit.
- They might correspond to mistakes in data transcription, lab errors, who knows? If possible, they should be identified and (hopefully) explained.
- Later, we'll talk about some formal ways to detect outliers.

References for this lecture

- Based on the lecture notes of Jonathan Taylor .