Week 05: Model diagnostics

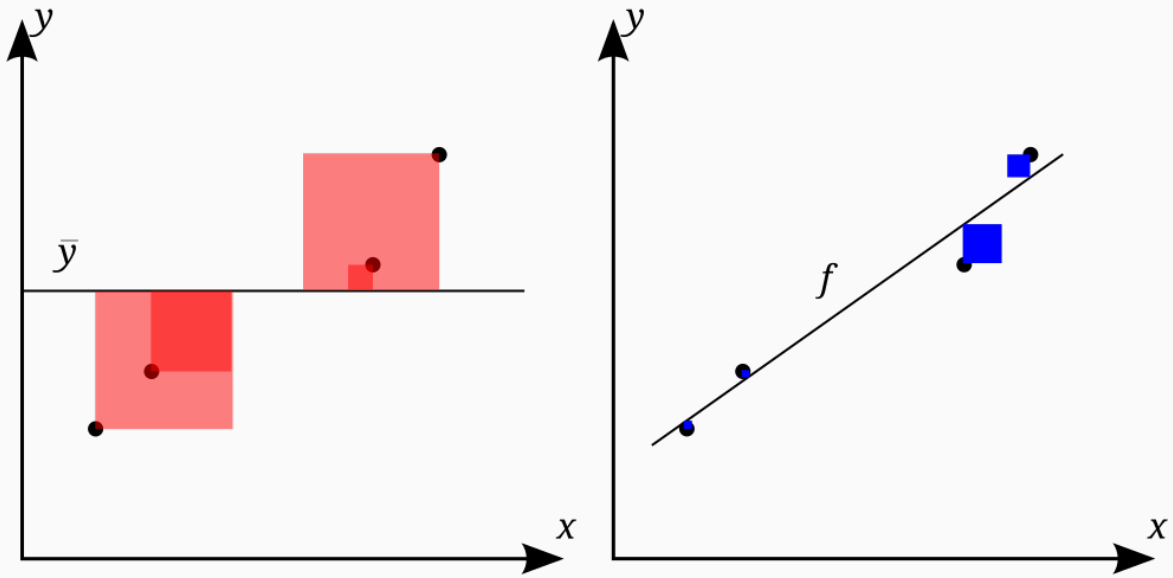
Assumptions, residuals, outliers, influence (Cook's distance)

Bartosz Maćkiewicz

The Accuracy of Prediction: R^2

In statistics, the coefficient of determination denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$



Example: `Ginzberg` dataset

The `Ginzberg` data frame in `carData` library has 82 rows and 6 columns. The data are for psychiatric patients hospitalized for depression. We will be interested in investigating the relationship between subjects' need to see the world in black and white (`simplicity` column) and their symptoms of depression (`depression` column).

Assumptions of linear regression and correlation

Regression

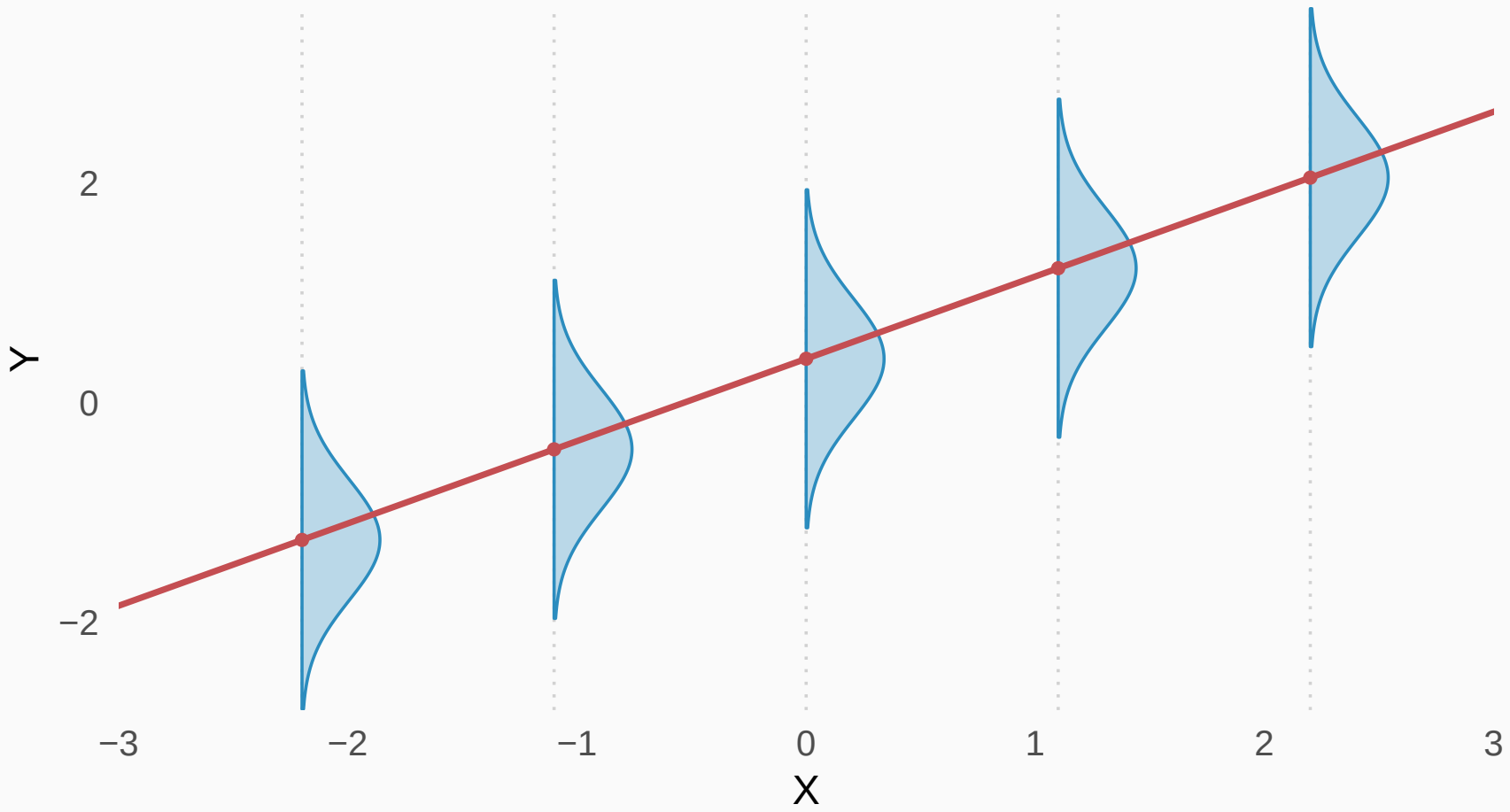
- homogeneity of variance (in arrays!)
    - variance of Y for each value of X is constant in the population
- normality (in arrays!)
    - in the population, the values of Y corresponding to any specified value of X are normally distributed

Correlation

- bivariate normal distribution

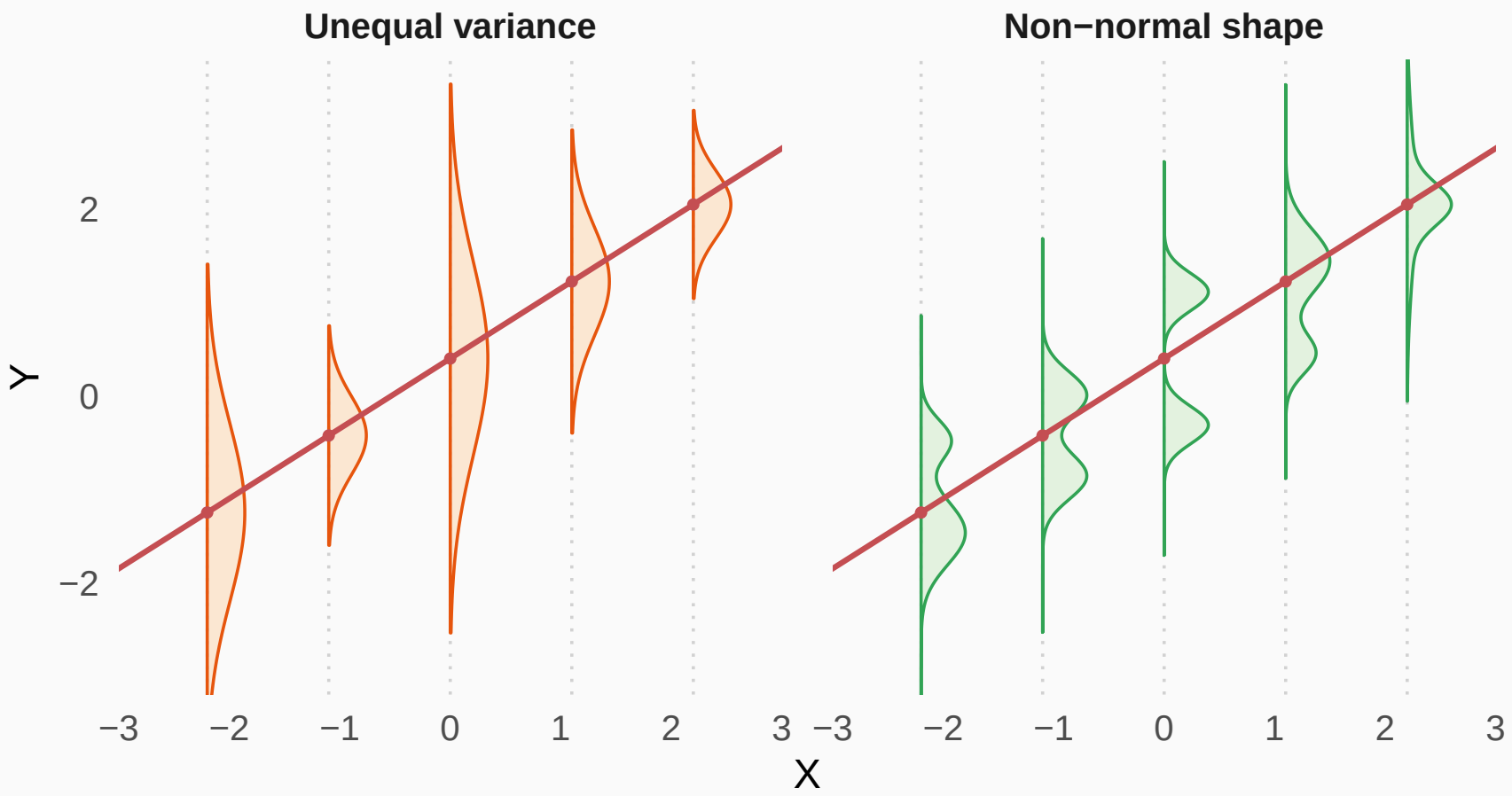
Regression assumptions: slices of $Y \mid X$

At several values of X, the conditional distribution of Y should be centered on the regression line, have similar spread, and look roughly bell-shaped.



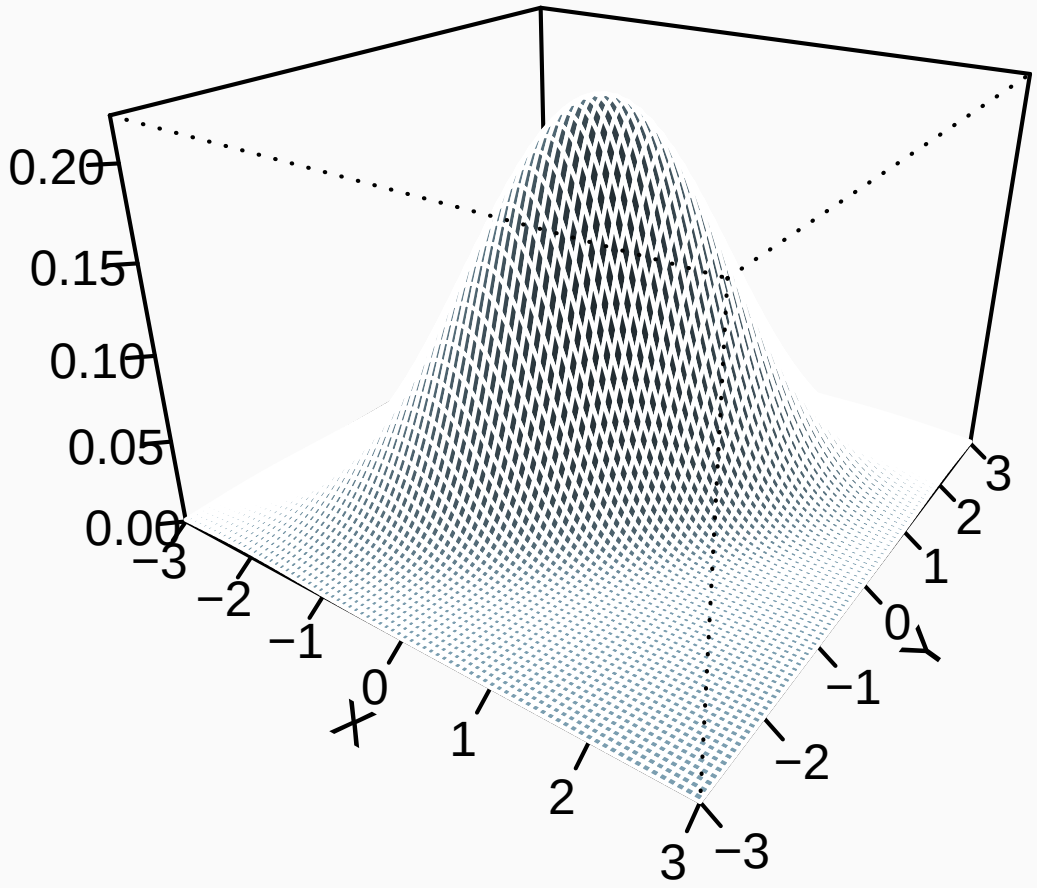
Regression assumptions: ways they fail

Violations happen when the slices fan out or when the shape stops looking approximately Gaussian.



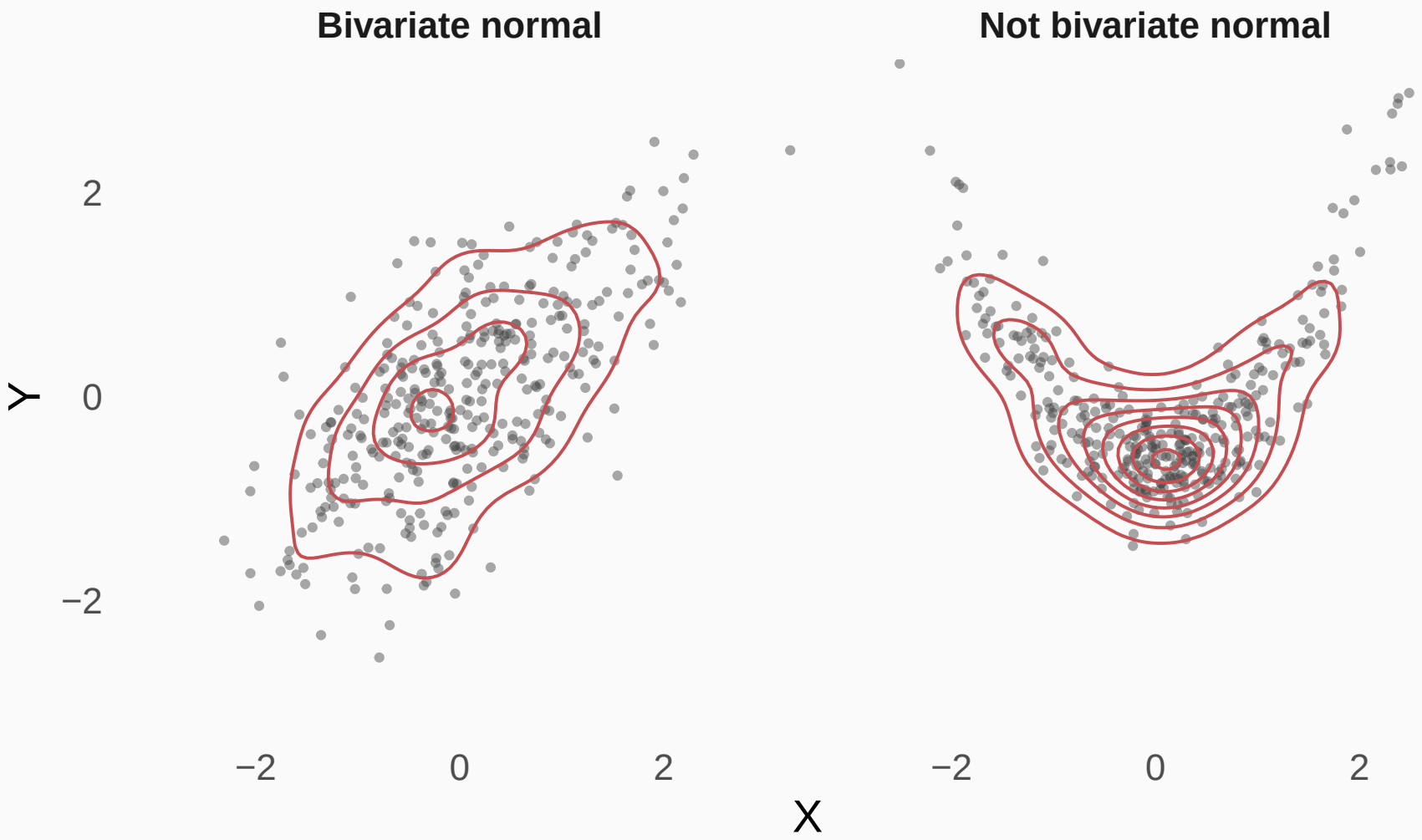
Bivariate normal: a 2D bell shape

If X and Y are jointly normal, their distribution is a smooth bell-shaped surface over the (X, Y) plane. Height shows density.



Correlation assumption: bivariate normality

If X and Y are jointly normal, the cloud of points should look roughly elliptical, not curved or split into clusters.



Q-Q plots

Quantile-quantile plots are a way to check whether the observations came from the normal distribution.

The idea behind Q-Q plots is quite simple. Suppose that we have a sample of 100 observations that is perfectly normally distributed with mean = 0 and standard deviation = 1.

With that distribution, we can easily calculate what value would cut off, for example, the lowest 1% of the distribution. We could make this calculation for every value of $0.00 < p < 1.00$, and we could name the results the *expected quantiles* of a normal distribution.

Now we go to the data we actually have. Because we have specified that they are perfectly normally distributed and that there are $n = 100$ observations, the lowest score will be the lowest 1%. Similarly, the second lowest value would cut off 2% of the distribution. We will call these the *obtained quantiles* because they were calculated directly from the data.

Q-Q plots

For a perfectly normal distribution, the two sets of quantiles should agree exactly.

The value that forms the 15th percentile of the obtained distribution should be exactly the value that the normal distribution would have, given the population mean and standard deviation.

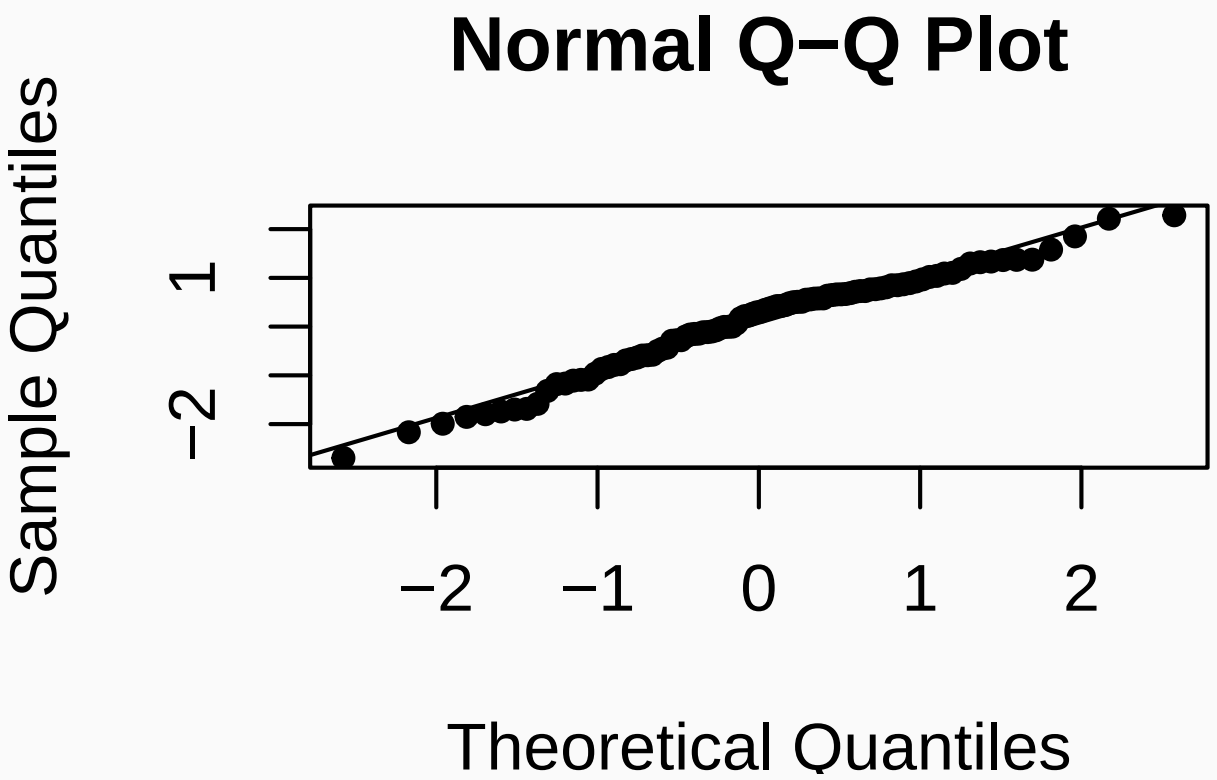
But how do we measure agreement? The easiest way is to plot the two sets of quantiles against each other, putting the expected quantiles on the Y axis and the obtained quantiles on the X axis. If the distribution is normal the plot should form a straight line running at a 45-degree angle.

Q-Q plots: `qqnorm()` by hand

```r
qqnorm_by_hand <- function(x) {
  observed <- sort(x)
  p <- ((1:length(observed)) - 0.5) / length(observed)
  expected <- qnorm(p)
  plot(expected, observed, pch = 20,
       xlab = "Expected", ylab = "Observed")
  abline(0, 1, lty = 2)
}
```

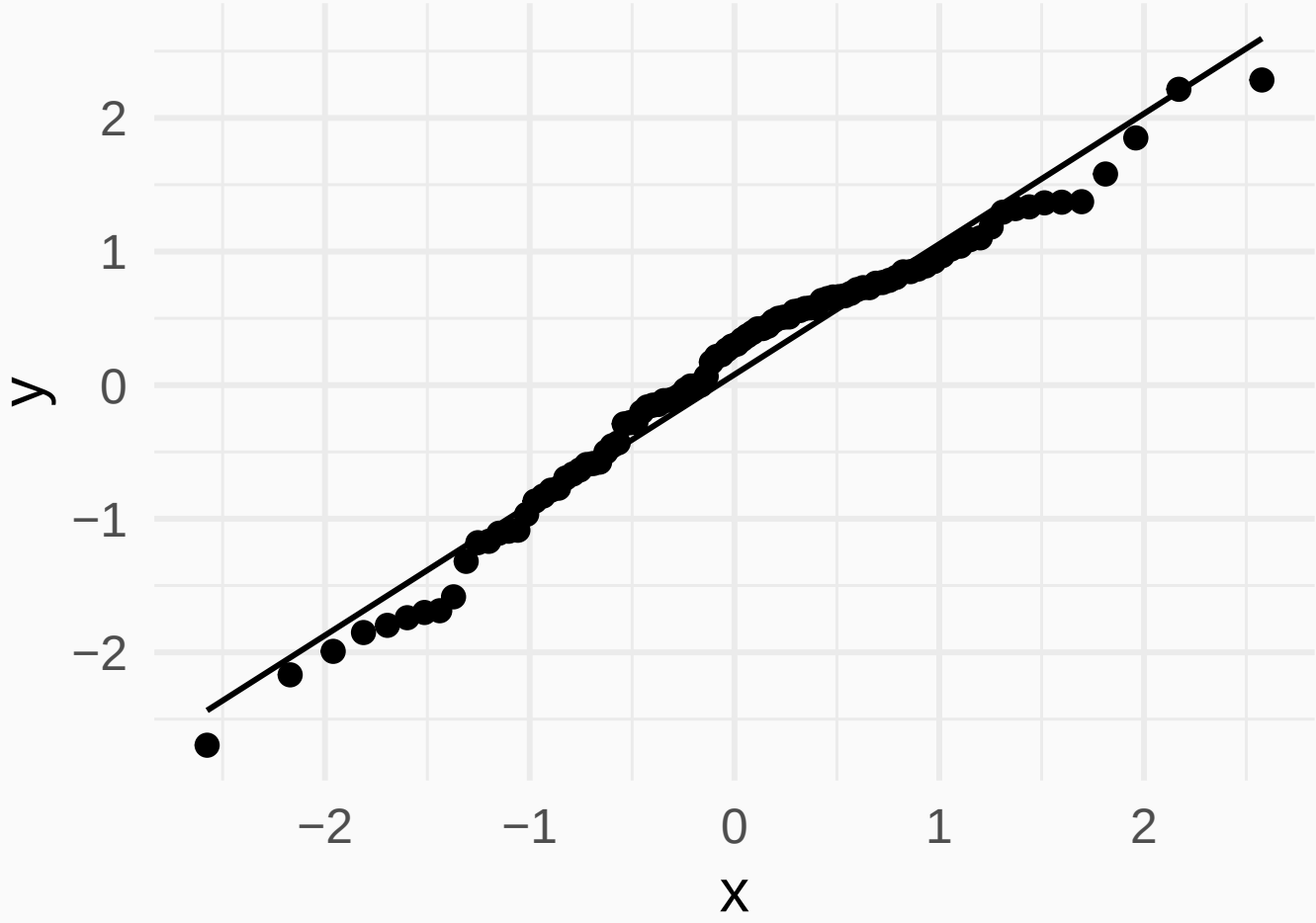
Q-Q plots

```r
x <- rnorm(100)
qqnorm(x, pch = 20)
qqline(x)
```



Q-Q plots

```
library(ggplot2)
ggplot(data.frame(x)) + aes(sample = x) + geom_qq() + geom_qq_line()
```



Regression diagnostics

If you call the `plot` function on a `lm` object you will get the following plots:

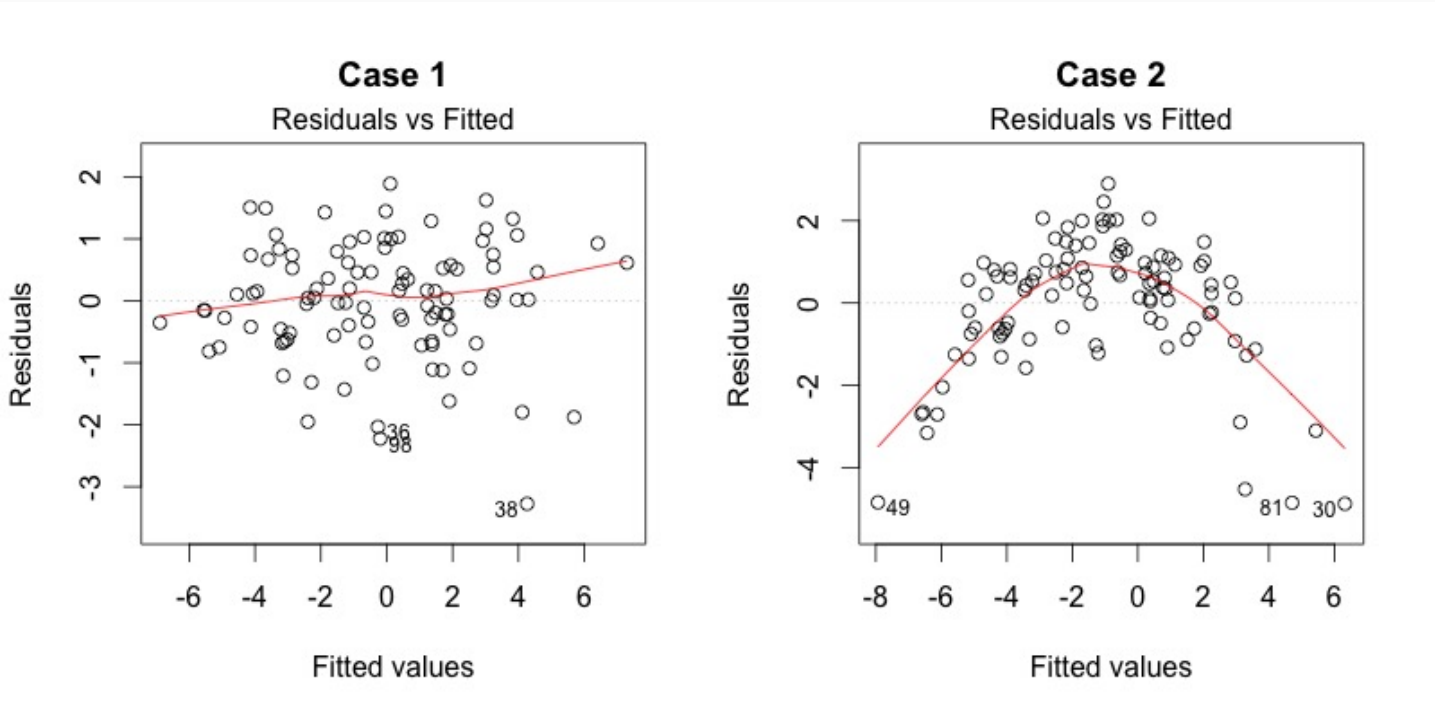
1. Residuals vs Fitted
2. Normal Q-Q
3. Scale-Location
4. Residuals vs Leverage

See https://data.library.virginia.edu/diagnostic-plots/ for more details

Regression diagnostics: Residuals vs Fitted

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.

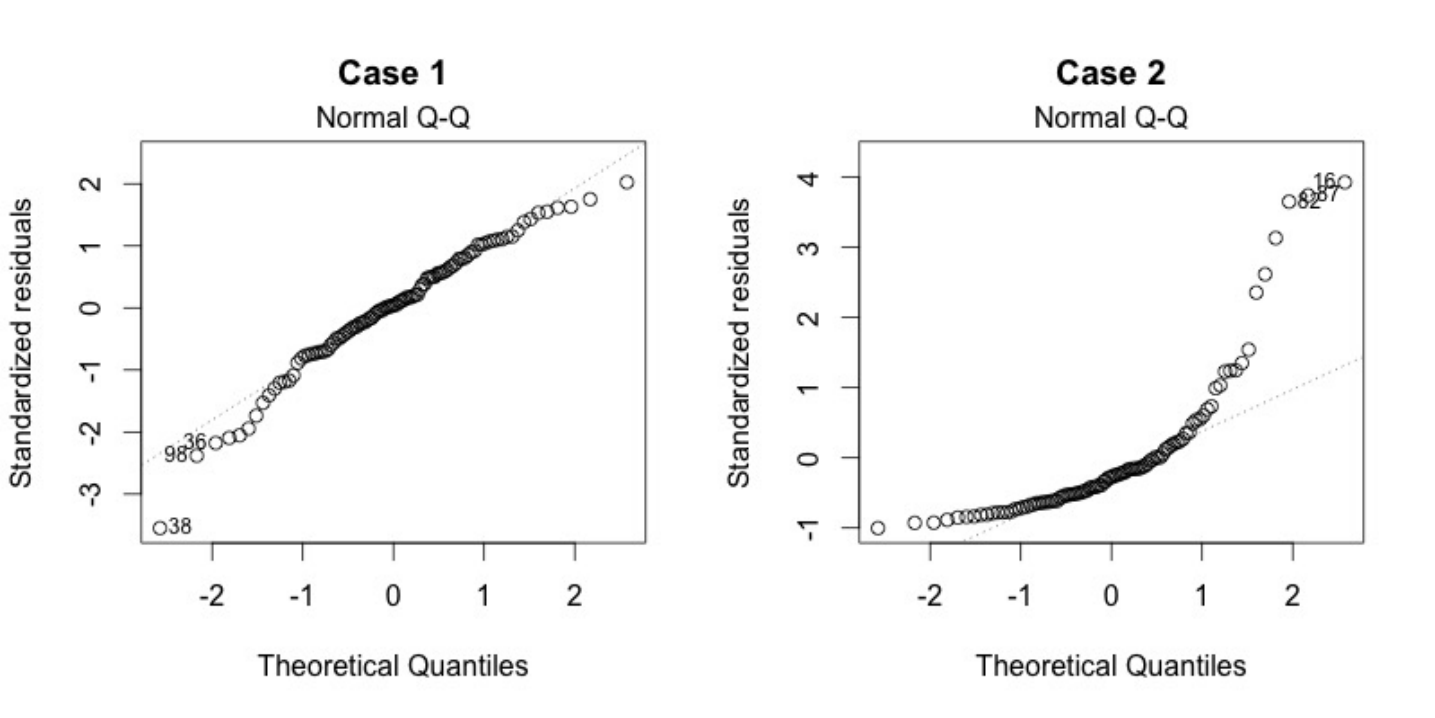
The good model data (Case 1) are simulated in a way that meets the regression assumptions very well, while the bad model data (Case 2) are not.



Regression diagnostics: Normal Q-Q

This plot shows whether residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.

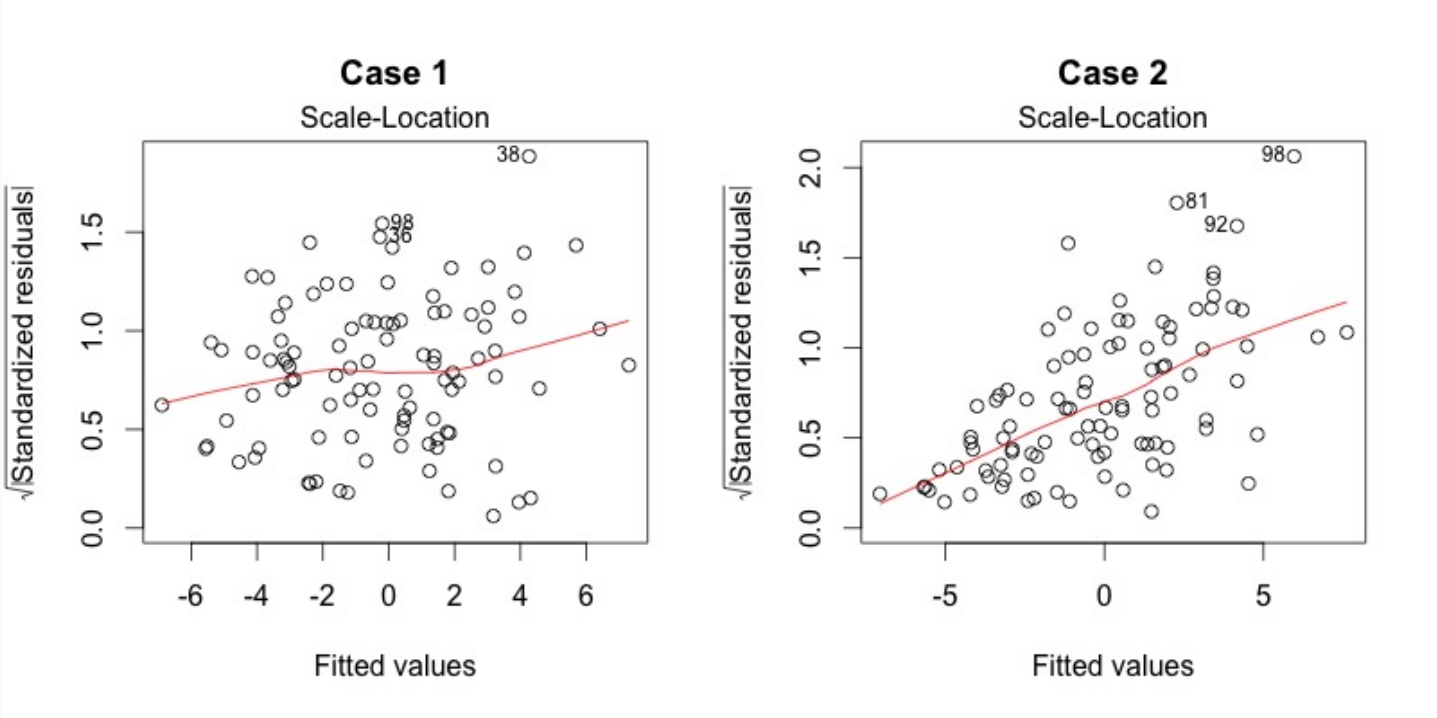
Of course, in practice they wouldn't form a perfect straight line and this will be your call. Case 2 is definitely concerning. Observation numbered 38 in Case 1 looks a little off. We will look at the next plot while keeping in mind that #38 might be a potential problem.



Regression diagnostics: Scale-Location

This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points.

In Case 1, the residuals appear randomly spread. Whereas, in Case 2, the residuals begin to spread wider along the x-axis. Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle in Case 2.



Regression diagnostics: Residuals vs Leverage

This plot helps us to find influential cases (i.e., subjects) if any. Not all outliers are influential in linear regression analysis (whatever outliers mean). Even though data have extreme values, they might not be influential as far as determining a regression line is concerned. That means the results wouldn't be much different if we either include or exclude them from the analysis. They follow the trend in the majority of cases and they don't really matter; they are not influential.

On the other hand, some cases could be very influential even if they look to be within a reasonable range of values. They could be extreme cases against a regression line and can alter the results if we exclude them from the analysis.

Another way to put it is that they don't get along with the trend in the majority of the cases.

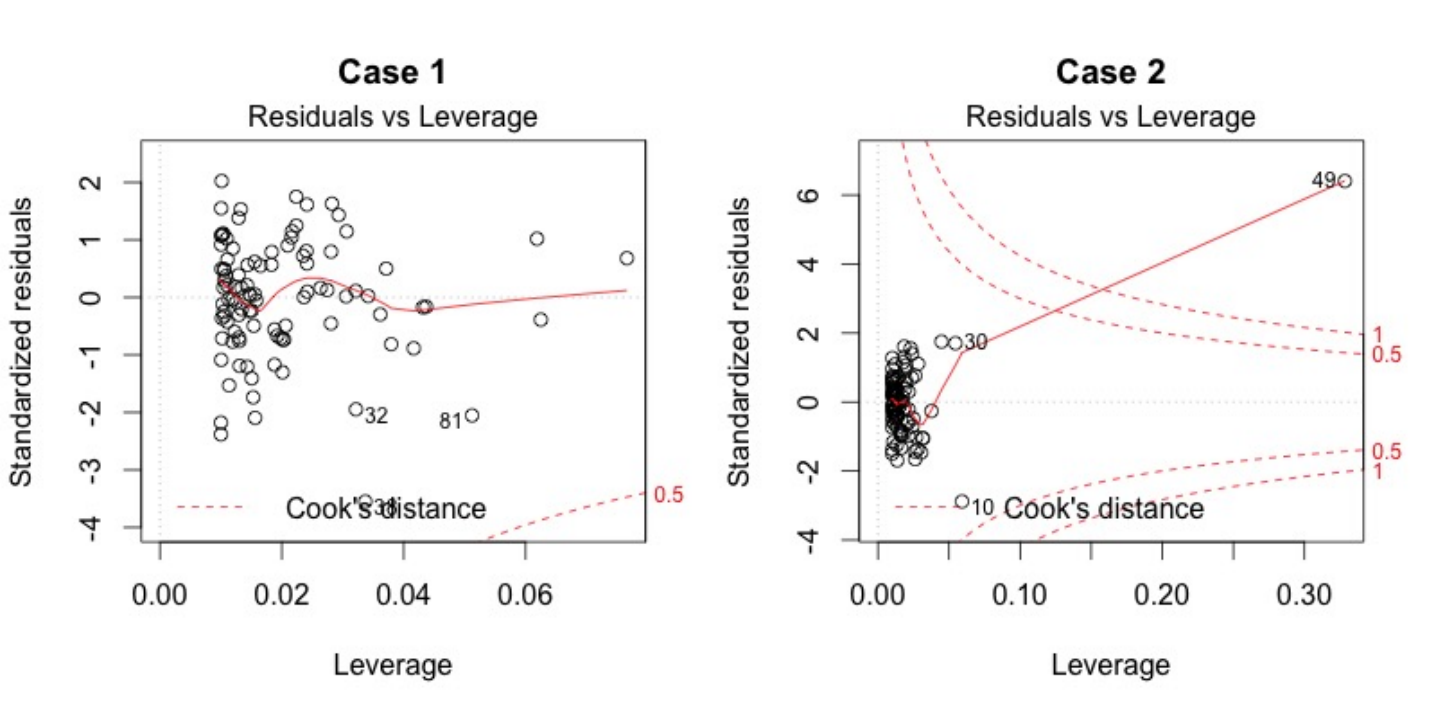
Regression diagnostics: Residuals vs Leverage

Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook's distance.

When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.

Regression diagnostics: Residuals vs Leverage

Case 1 is the typical look when there is no influential case, or cases. In Case 2, a case is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot). The plot identified the influential observation as #49.



Why diagnostics?

A fitted model is never “the truth”. It is a structured approximation.

Diagnostics are about a very practical question:

Can we trust the model-based conclusions enough to report them?

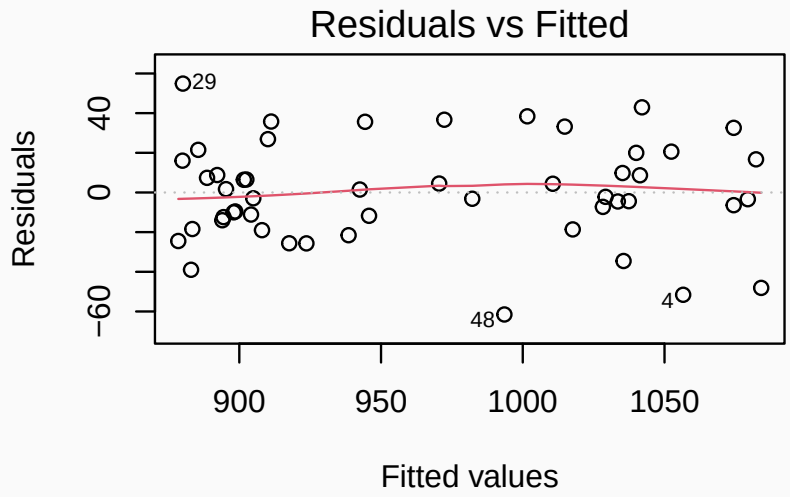
In this course, diagnostics are a repeatable workflow:

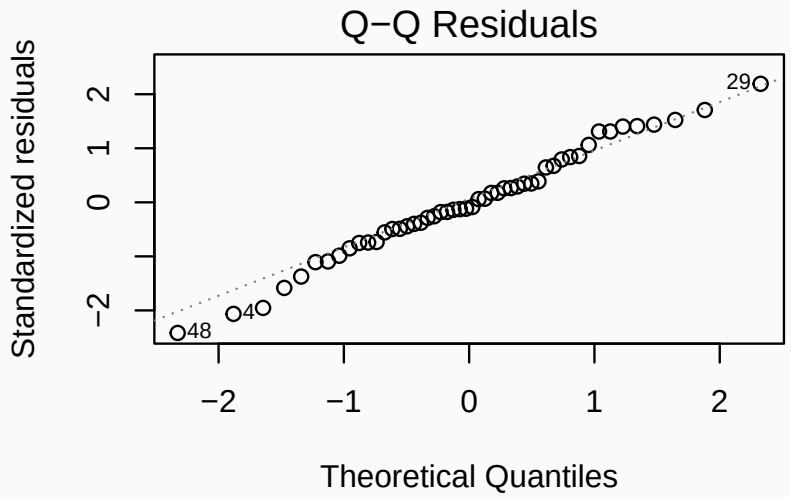
1. Fit a model.
2. Check assumptions and influence (plots, leverage, Cook’s distance).
3. Decide what to do, and document the decision.

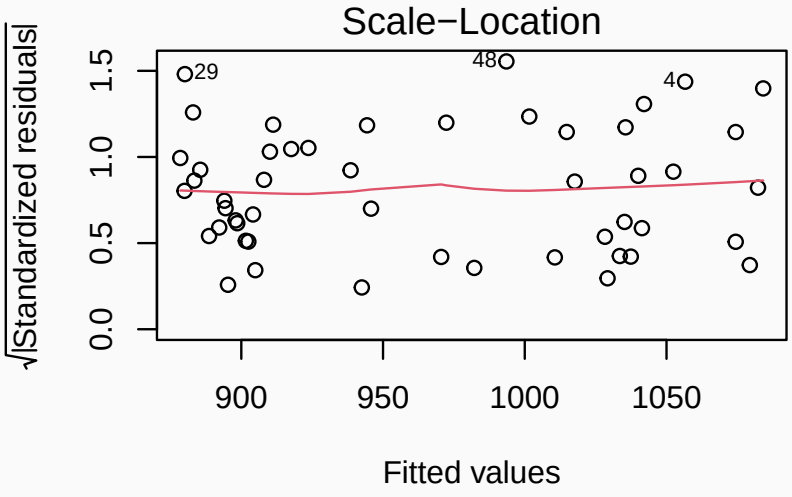
Example: education spending and SAT (`guber1999.csv`)

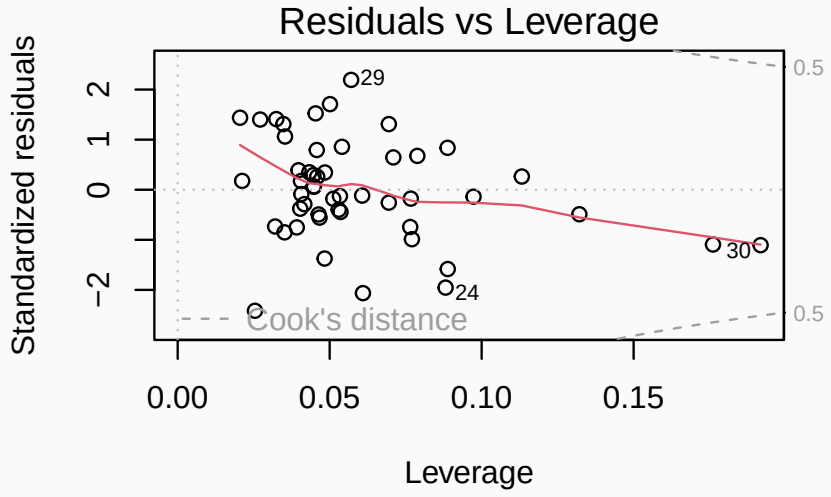
We will fit a multiple regression model and diagnose it.

The 4 standard diagnostic plots









Outliers vs leverage vs influence (important distinction)

- **Outlier**: unusual outcome value given predictors (large residual)
  - look at (studentized) residuals
- **High leverage**: unusual predictor values (far in X space)
  - look at hat values (`hatvalues()`)
- **Influential point**: changes the fitted model noticeably if removed
  - look at Cook's distance and influence diagnostics

Do not delete points automatically. Use them as a signal to:

- verify data quality,
- check if the model is misspecified,
- run sensitivity analysis.

Cook's distance: identify influential points

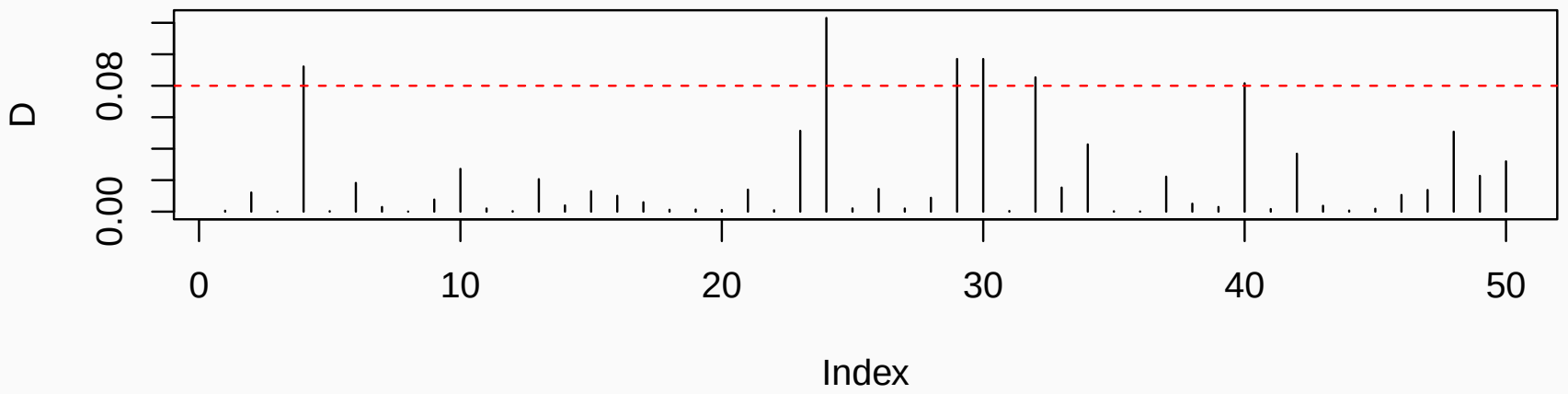
Rule-of-thumb threshold:

$$D_i > \frac{4}{n}$$

row	cooks_d
24	0.1231
29	0.0970
30	0.0970
4	0.0922
32	0.0853

Cook's distance plot

```r
cd <- cooks.distance(fit)
plot(cd, type = "h", ylab = "D")
abline(h = 4 / nobs(fit), lty = 2, col = "red")
```

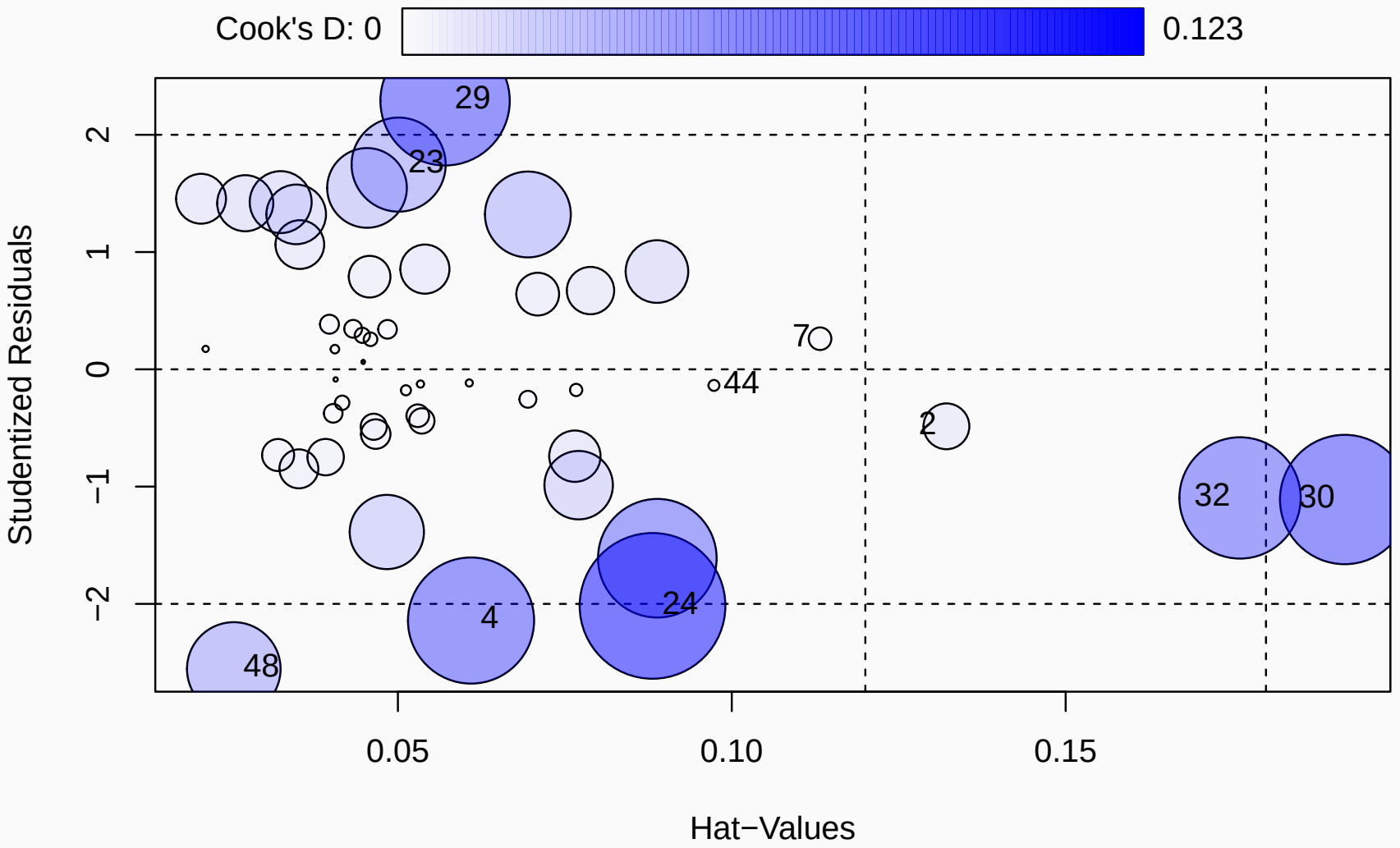


Influence plot (leverage + residual + Cook's): code

```
car::influencePlot(fit, id = list(n = 5))
```

- X-axis: hat-values, so unusual predictor values (high leverage)
- Y-axis: studentized residuals, so unusual outcome values given the model
- bubble size and color: Cook's distance, so overall influence on the fitted model
- `id = list(n = 5)` labels the union of the top 5 cases by residual, leverage, and Cook's distance, so not 5 total points

Influence plot (leverage + residual + Cook's)



Sensitivity analysis: quick check vs broader check

- Quick check: refit after omitting the single most influential case, case 24.
- Broader check: refit after omitting all cases with Cook's D $>$ 4/n = 0.08: 4, 24, 29, 30, 32, 40.
- If the coefficients stay similar across these choices, the conclusion is less dependent on a few influential cases.

model	term	estimate	conf.low	conf.high
all data	Expend	11.13	4.56	17.70
all data	LogPctSAT	-78.20	-87.20	-69.21
omit case 24	Expend	10.74	4.36	17.11
omit case 24	LogPctSAT	-79.83	-88.70	-70.96
omit D > 4/n (6 cases)	Expend	13.05	6.07	20.02
omit D > 4/n (6 cases)	LogPctSAT	-82.48	-90.79	-74.17

What to do when diagnostics look bad?

There is no single “correct” fix. Common actions include:

- check for data errors (coding, units, duplicates)
- revise the model (add terms, nonlinearity, interactions)
- consider transformations (Week 6)
- consider robust / resampling approaches (Week 6)
- report sensitivity analysis (with vs without influential points)

Takeaways

- Diagnostics are part of the analysis pipeline, not an afterthought.
- Q-Q plots are a practical tool for thinking about normality.
- Learn the difference: outlier vs leverage vs influence.
- Cook's distance is a practical tool for **influence** checks.
- Always document decisions (why a point was kept/dropped, if ever).