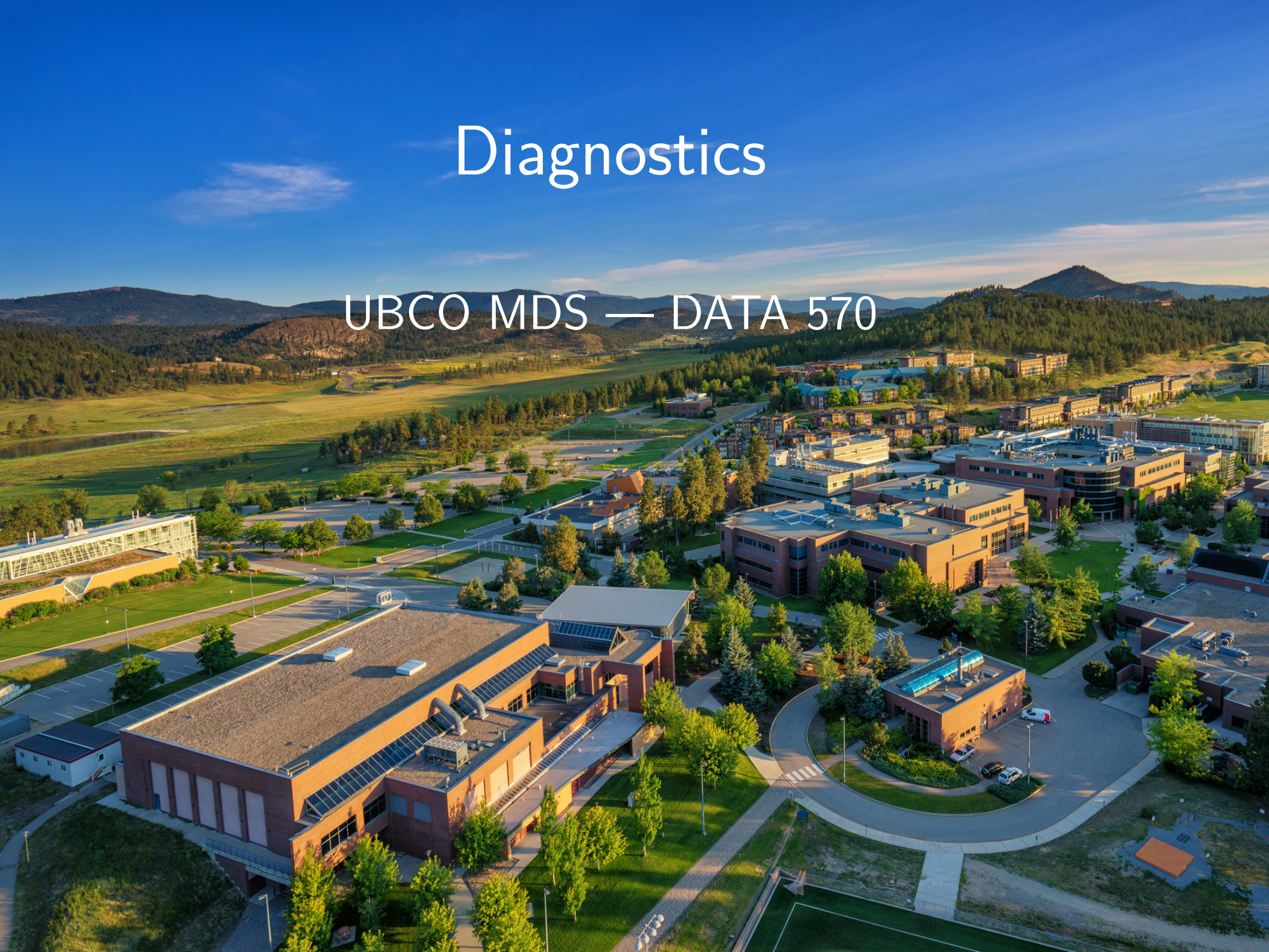
Diagnostics

UBCO MDS — DATA 570



Potential Problems

- When we fit a linear regression model to a particular data set, a number of problems may occur.
- Today we will go through the most common among these.

Non-linearity of the Data

The SLM assumes that the relationship between the predictors and the response fall ($\approx$) on a straight line.

When the true relationship deviates markedly from linear, then all conclusions/predictions derived from this fit are dubious at best.

Non-linearity of the Data

Residual plots are a useful graphical tool for identifying non-linearity.

SLR: plot the residuals, $e_i = y_i - \hat{y}_i$, versus the predictor x_i.

MLR: plot the residuals versus the predicted (or fitted) values $\hat{y}_i$.

Non-linearity of the Data

- Ideally, the residual plot will show no discernible pattern.
- The presence of a pattern may indicate a problem with some aspect of the linear model.
- A "good" plot should produce randomly dispersed residuals around the horizontal axis

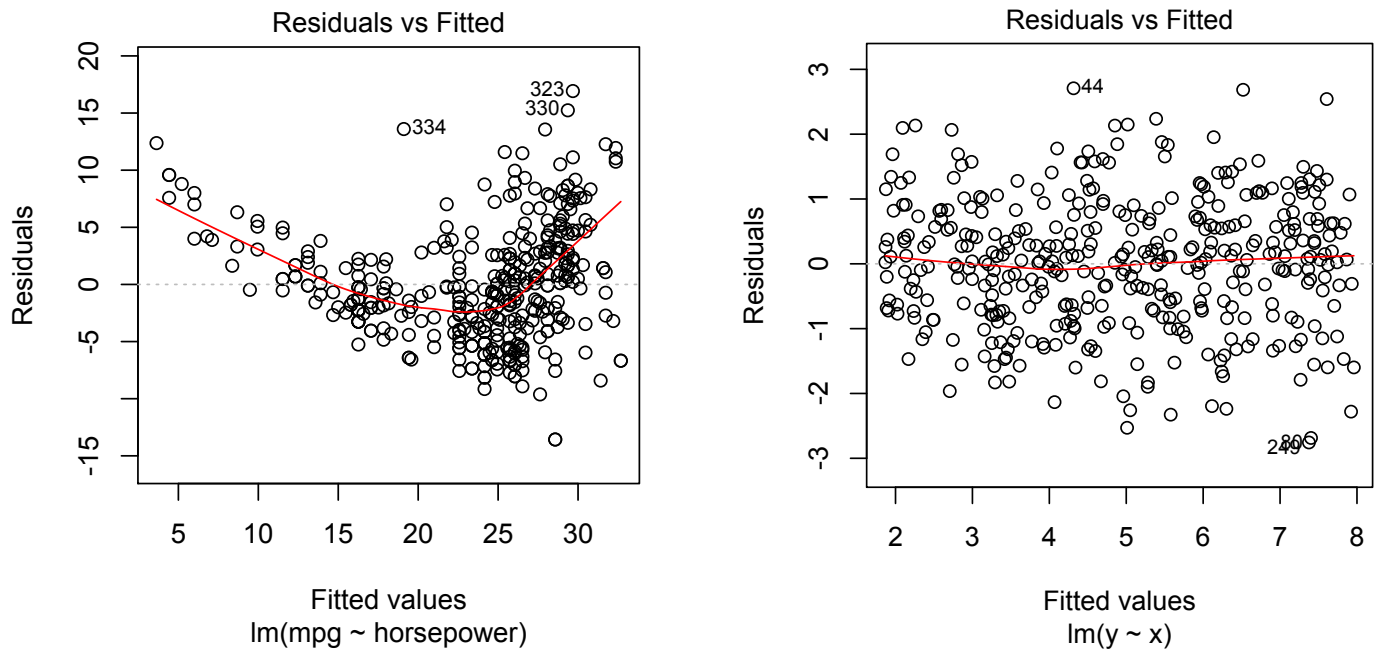


Figure: “Bad” (left) and “good” (right) residual plots.

Residual Plots

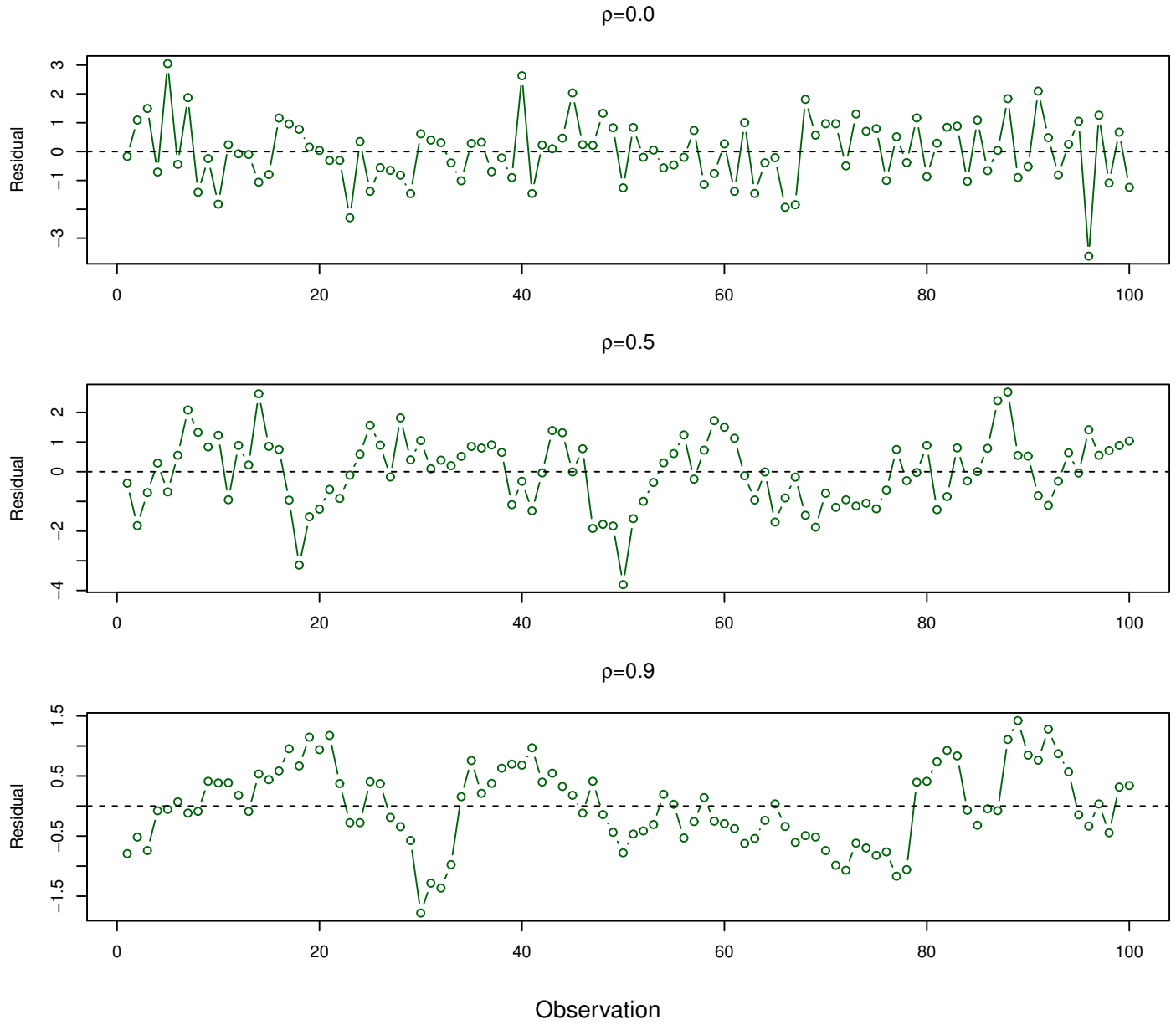
- To help us identify any trends, a red smooth line is fit to the residuals
- In the ideal scenario, this smooth line should be a horizontal line at 0.
- In presence of non-linear associations, we could try and fit a model with *transformed* predictors, eg. $\log(X)$, $\sqrt{X}$, and X^2, (as discussed in Lecture 5)

Correlation of error terms

- As discussed in Lecture 2, the linear regression model assumes that the error terms, $e_1, e_2, \ldots, e_n$, are independent.
- If error terms are correlated, then the estimated standard errors will tend to underestimate the true standard errors.
- In other words, if the error terms are correlated, we may have an unwarranted sense of confidence in our model.

Correlation of error terms

- Correlated errors can occur when multiple measurements are taken on the same subject over time (i.e *time series* data)
- To investigate this assumption violation, we plot the residuals from our model as a function of time.
- Other examples can be when our data contain
  - members of the same family
  - people exposed to the same environmental factors

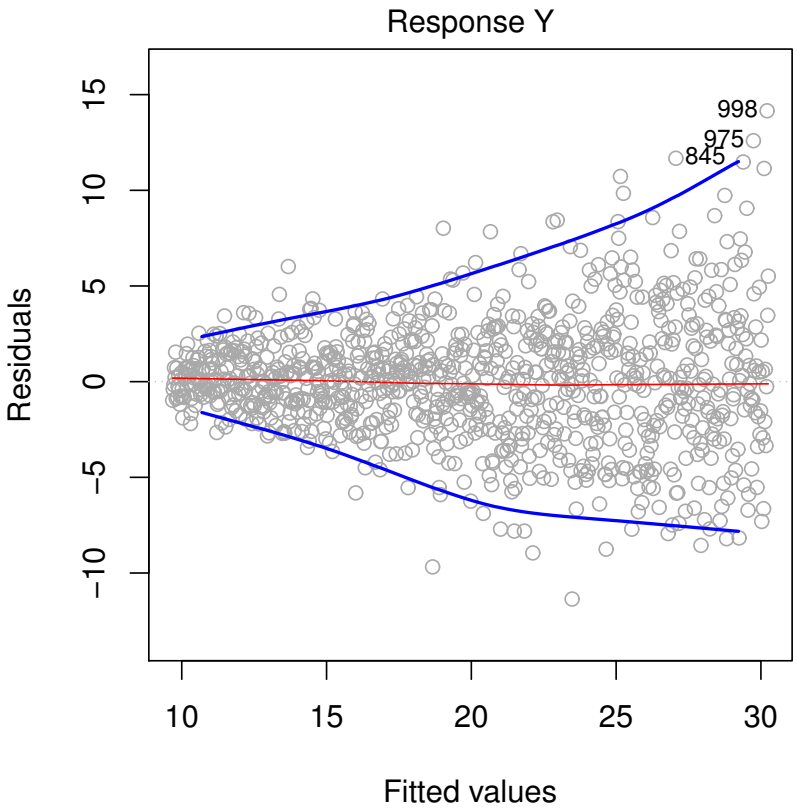

ρ=0.0
Residual
3
2
1
0
−1
−3
0
20
40
60
80
100
ρ=0.5
Residual
2
1
0
−2
−4
0
20
40
60
80
100
ρ=0.9
Residual
1.5
0.5
−0.5
−1.5
0
20
40
60
80
100
Observation

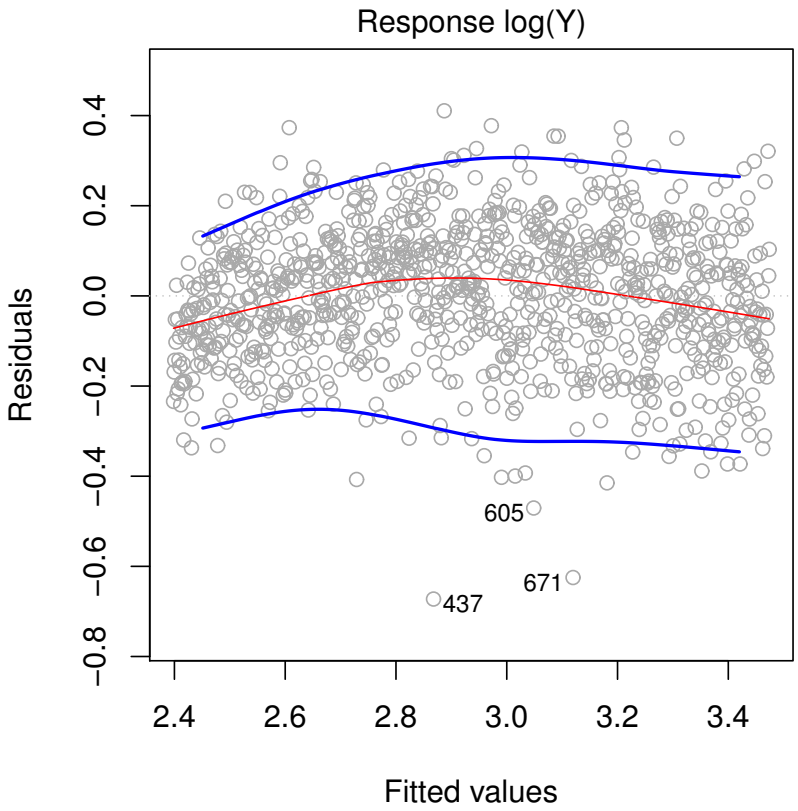
Non-constant variance

- Non-constant variance of error terms, i.e. $Var(e_i) = \sigma^2$, is another important assumption of the linear regression model.
- This assumption is often violated, eg. the variances of the error terms may increase with the value of the response.

Non-constant variance

- Non-constant variance, i.e. *heteroscedasticity* can be identified through the residual plot.
- Often this violation takes shape as a funnel/fan shape in the residual plot.
- In presence of non-constant variance, we could try transforming the *response* Y using a concave function eg. $\log(Y)$, $\sqrt{Y}$.


Response Y
Residuals
Fitted values
998
975
845
10
15
20
25
30
−10
−5
0
5
10
15


Response log(Y)
Residuals
Fitted values
605
671
437
2.4
2.6
2.8
3.0
3.2
3.4
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4

Outliers

- An outlier is a point for which y_i is far from the predicted value $\hat{y}_i$ (see next slide for example)
- Outliers can be genuine or be a result of, for example, incorrect reading/recording.
- If an outlier is believed to be a result of an error, then one might simply remove that observation.
- However, genuine outliers may indicate a deficiency with the model, eg. a missing predictor.

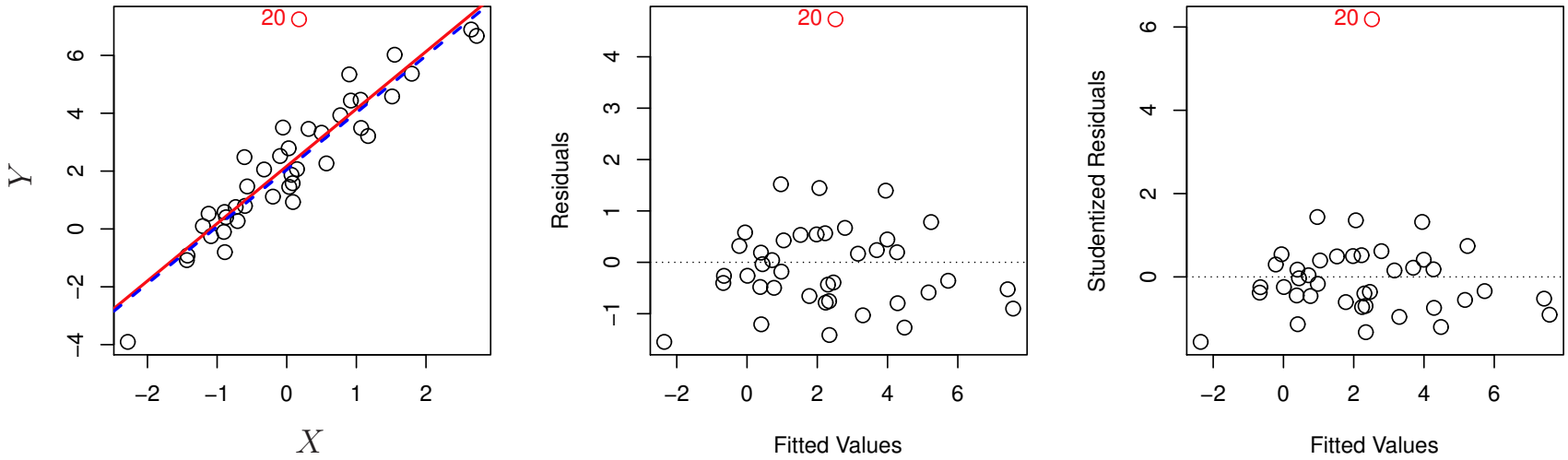


Figure: The red/blue line shows the fit with the outlier included/excluded.

Outliers

- As depicted in the previous slide, it can happen that outliers do not greatly affect on the least squares regression line.
- It can however affect things like RSE, p-values, and R^2 values.
- Again, residual plots can be useful in identifying outliers.

Outliers

- To avoid arbitrary cut-offs in a residual plot, sometimes we turn to the so-called *studentized residuals*
- These scaled residuals are produced by dividing each residual e_i by its estimated standard studentized error, $se(e_i) = \sqrt{MSE(1 - h_i)}$
- Observations whose studentized residuals are greater than 3 are then flagged as potential outliers.

High-leverage points

- While outliers produce unusual y_i values, observations with high leverage have unusual x_i values.
- The inclusion/exclusion of high leverage points tend to have a higher impact on estimated regression line
- It is very important to identify such points as they have the potential to invalidate the entire fit.

High-leverage points

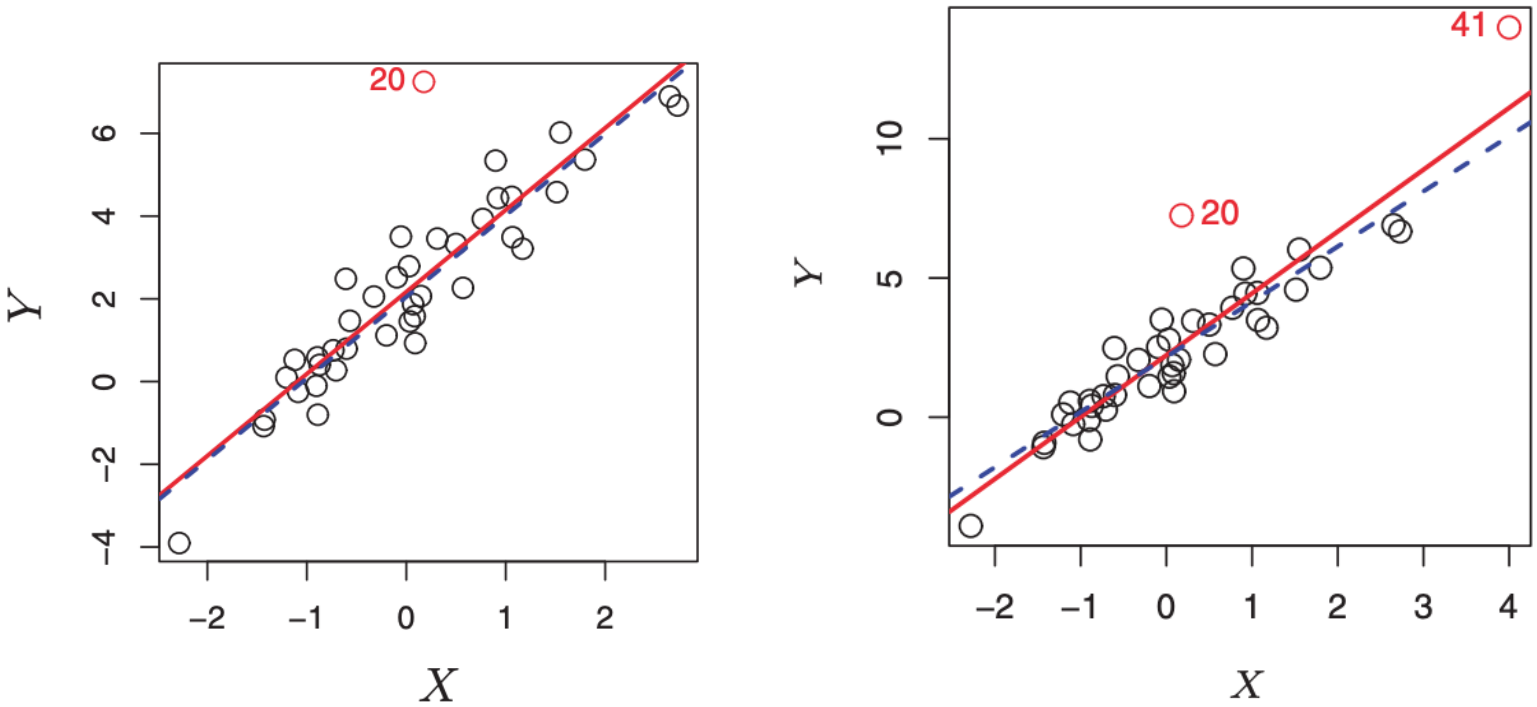


Figure: Left: example of outlier Right: example of high-leverage. The red (blue) line shows the fit with the outlier/high-leverage included (excluded)

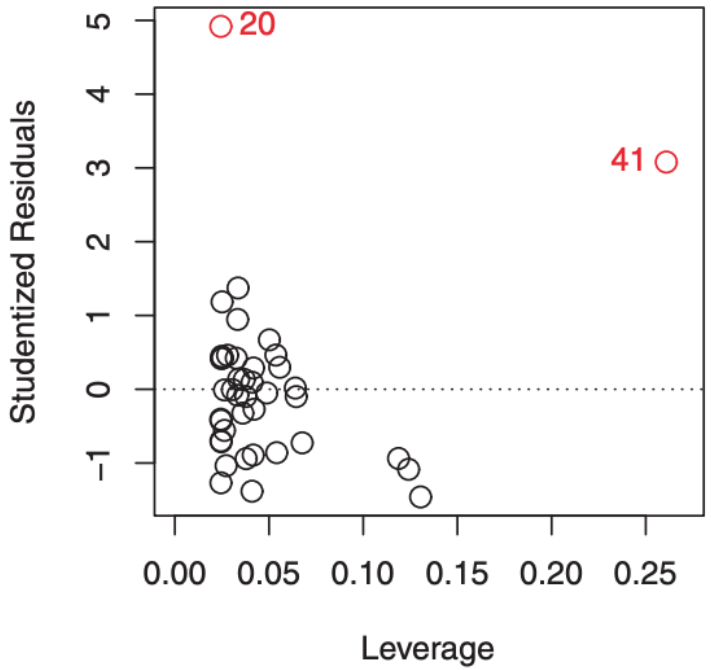
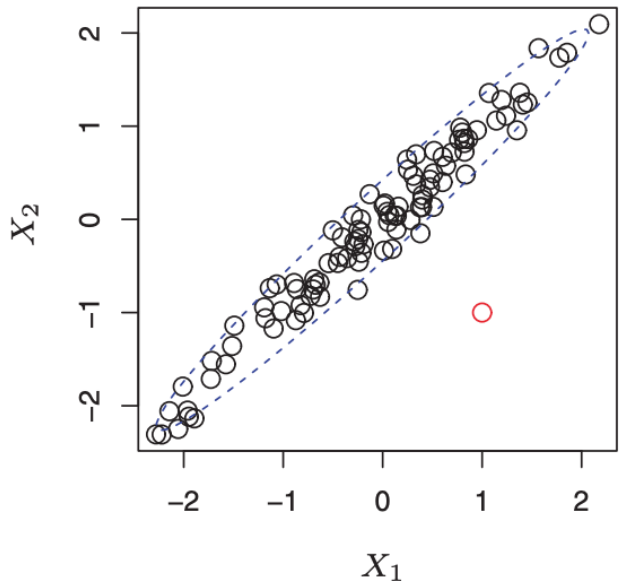


Figure: Plotting the studentized residuals versus h_i flags observation 41 is an outlier as well as a high-leverage point, while observation 20 is an outlier with low leverage.

- For SLR high leverage can be easily identified by looking for observations having predicted values outside of the "normal" range
- With MLR, these points can be harder to identify.



High-leverage observations will have larger values of the so-called *leverage statistic*. For SLR:

$$h_i = \frac{1}{n} + \frac{x_i - \overline{x}}{\sum_j^n (x_j - \overline{x})^2} \tag{1}$$

As the distance between x_i and $\overline{x}$ increase, so to does the h_i.

For MLR, leverage is given the ith diagonal element of the *hat* matrix:

$$H = X(X^T \cdot X)^{-1}X^T \tag{2}$$

In words: how far the vector $(X_{i1}, X_{i2} \ldots X_{ip})$ is from $(\overline{X}_1, \overline{X}_2 \ldots \overline{X}_p)$, with distance measured in standard deviation units.

A general guideline is to use $h_i > 2(p+1)/n$ as an indicator for high leverage (note: $1/n \leq h_i \leq 1$)

Influential points

- Observations having a relatively large effect on the regression model's predictions are called influential observations
- A high leverage point is not necessarily an influential point.
- An influential point is not necessarily a high leverage point.

Influential points

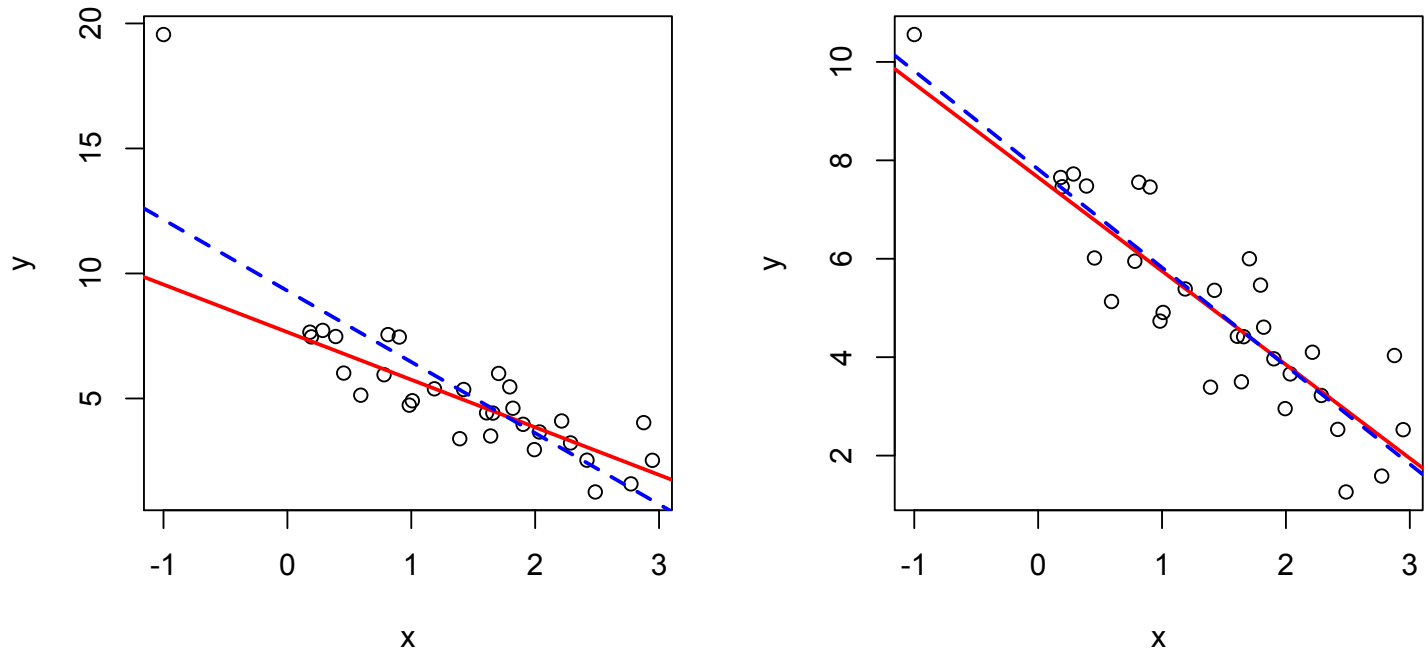


Figure: Left: Influential high leverage outlier Right: High leverage

Influential points

- A typical measure of influence is the Cook's D-statistic.
- The Cook's distance statistics, for observation i:

$$D_i = \frac{e_i^2}{MSE \cdot d}\left[\frac{h_i}{(1-h_i)^2}\right] \tag{3}$$

where d is the dimension of your data (ie. $X_{n\times d}$)

- Influential observations will have high Cook's distance score.

Collinearity

- Collinearity occurs when two or more predictor variables are closely related to one another.
- Simple bivariate scatterplots can show us correlations between predictors
- However, one variable may be correlated with some *linear combination* of two or more other variables (multicollinearity)

Collinearity

- Collinearity reduces the accuracy of the estimates of the regression coefficients, ie. increase $SE(\hat{\beta}_j)$.
- Consequently, the *power* of the hypothesis test—the probability of correctly power detecting a non-zero coefficient—is reduced by collinearity

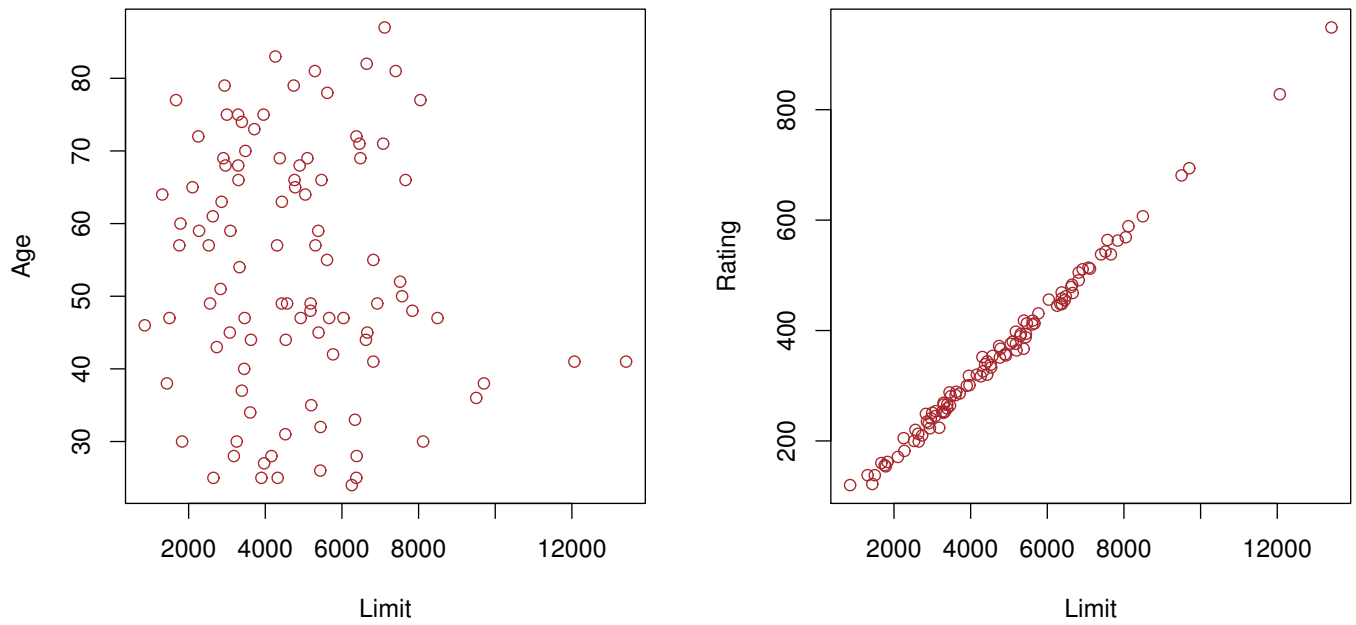


Figure: **ISLR Figure 3.14:** Scatterplots of the observations from the `Credit` data set. Left: A plot of `age` versus `limit`. These two variables are not collinear. Right: A plot of `rating` versus `limit`. There is high collinearity

		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* `Credit` *data set are shown. Model 1 is a regression of* `balance` *on* `age` *and* `limit`*, and Model 2 a regression of* `balance` *on* `rating` *and* `limit`*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

Collinearity

- The most straightforward measure of collinearity is called the Variance Inflation Factor (VIF).
- VIF measures the ratio of the variance of $\hat{\beta}_j$ in the full model and the variance of $\hat{\beta}_j$ if it were fit on its own (i.e SLR).
- A VIF=1 (smallest possible value) indicates the complete absence of collinearity

Collinearity

- The VIF is calculated as follows:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 for the regression of X_j on all the other X_s.

- If R_j^2 is close to one, then collinearity is present, and so the VIF will be large.

- Typically values exceeding 10 indicate a problem.

Collinearity

- In practice small amount of collinearity among the predictors is expected
- VIFs do not tell how many collinearities there are, or which variables are included in them.
- There are other more sophisticated measures of collinearity (eg. based on eigenvalues and eigenvectors of the matrix of Xs) but those fall outside the scope of this module.

Collinearity

In the face of collinearity, we may decide to:

- eliminate one of the problematic variables our model
- combine the collinear variables together into a single predictor

Diagnostic Plots

If we apply the `plot()` function fo the output from a `lm()` (see `?plot.lm` for details), four diagnostic plots are produced:

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

Normal Q-Q plot

- A normal Q-Q plot shows if residuals are normally distributed.
- More generally, a Q-Q (quantile-quantile) plot plots two sets of quantiles against one another.
- Quantiles coming from the same distribution should roughly form points along a straight line

Normal Q-Q plot for normal data

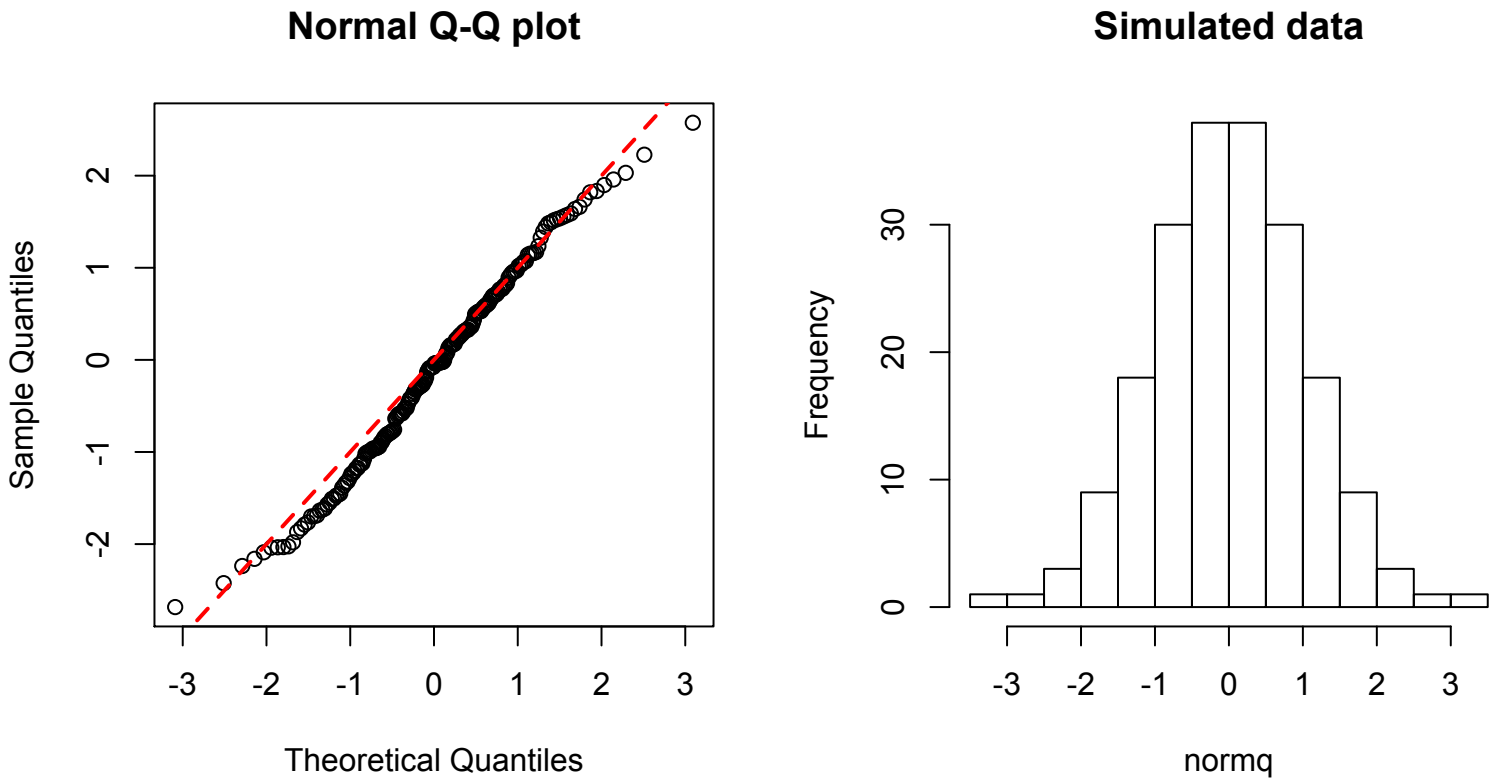


Figure: Normal data will tend to produce points on the dashed line of a Normal Q-Q plot.

Normal Q-Q plot

- While visual checks of this sort are subjective, it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.
- This is not a formal test unlike other statistical approaches
- If points fall very far from a straight line, that is cause for concern.

Normal Q-Q plot for skewed data

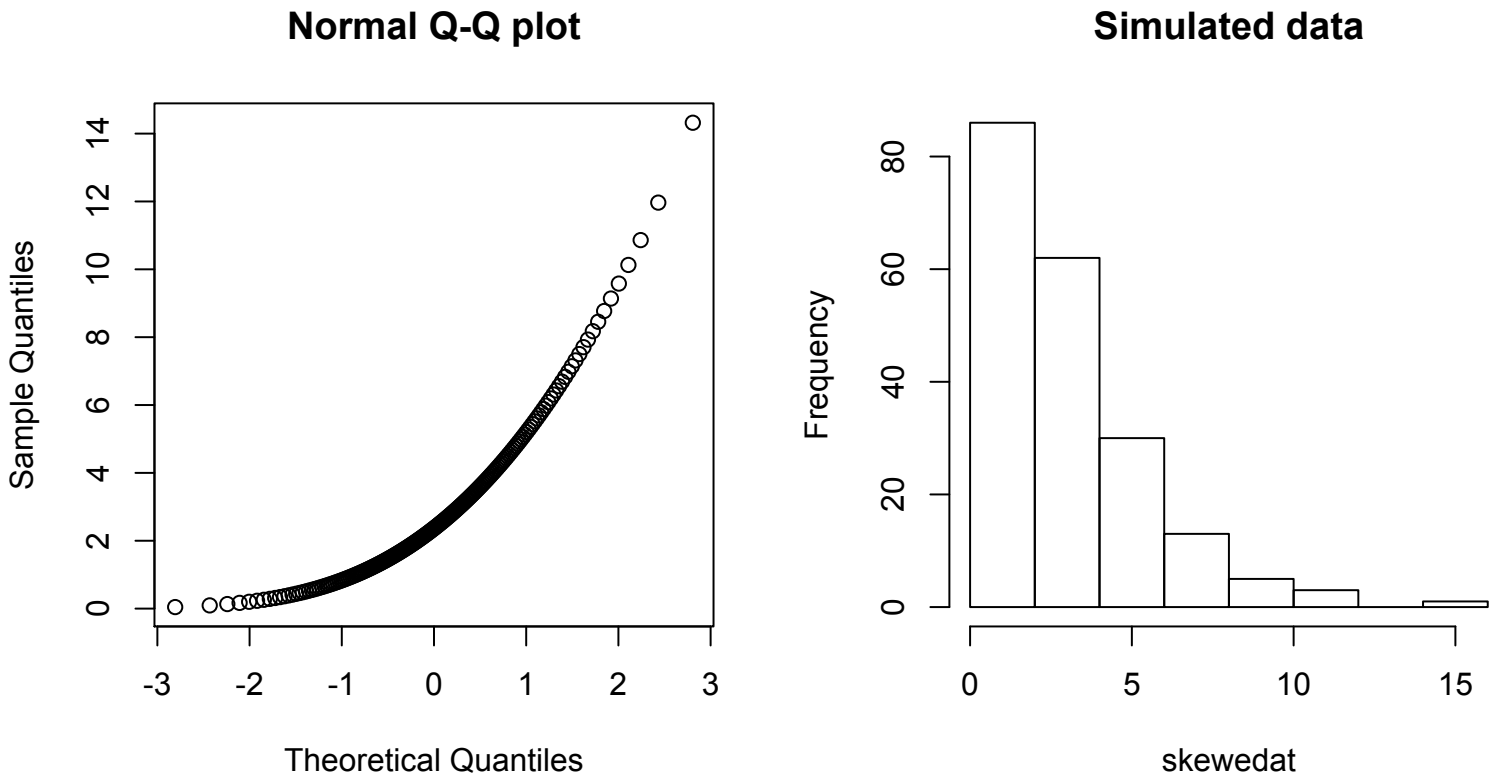


Figure: Curved Normal Q-Q plots may be an indication that your data are skewed.

Normal Q-Q plot for heavy-tailed data

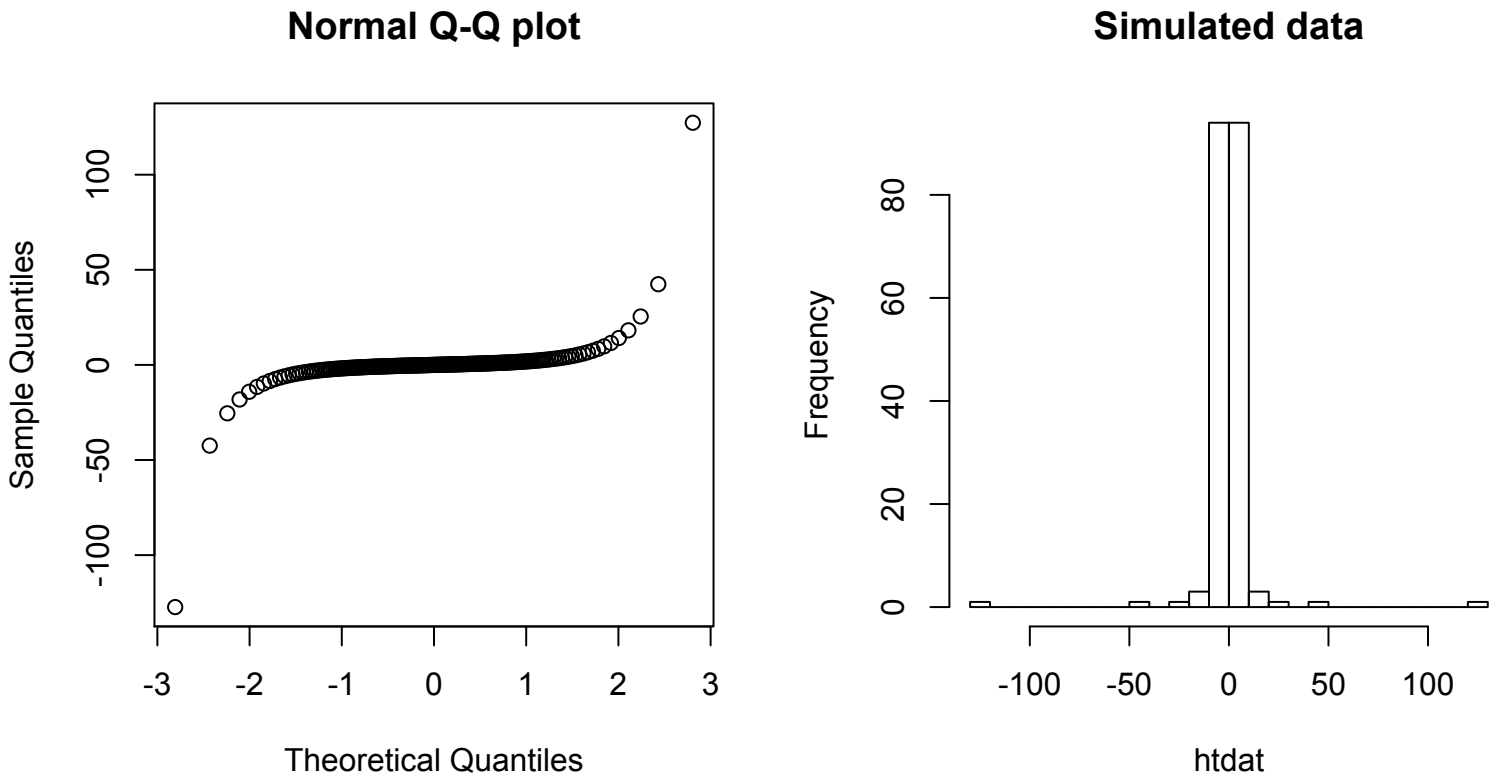


Figure: Normal Q-Q plots that are flat in the middle and highly curved at extremities may indicate heavy-tailed data.

Conclusion

- While many of the steps towards fitting a linear regression model are algorithmic, model building is more an art than a science.
- These diagnostics tools are meant to guide you through making a decision, but the decision is ultimately yours.