Knowledge and solutions for a changing world

Be boundless

Advancing data-intensive discovery in all fields

MULTIPLE LINEAR REGRESSION

UW DIRECT

(Data Intensive Research Enabling Cutting-edge Tech)

https://uwdirect.github.io

Stéphanie Valleau

Chemical Engineering

Multiple linear regression

Concept: independently assess the variation in *Y* with different input features *X*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

- Coefficients are determined by setting the partial derivatives to zero and solving the resultant *p+1* linear equations
- There is an exact solution (see e.g. Wikipedia)

Assumptions

Key assumptions when using a linear regression model

- Errors are **uncorrelated** and normally distributed
- The variance of the error (in Y) is **independent of where we are in X**
- **Linear relationship** between X and Y (the predictor-response relationship)
- Individual contributions of your X's are **piecewise additive** to the response

Questions we looked at

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response? (**F-statistic**)
2. Do all the predictors help to explain *Y*, or is only a subset of the predictors useful? (**Step wise feature** selection)
3. How well does the model fit the data? (**R^2 score and RSE**)
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

4 Q directly copied from ISL page 75!

Possible problems

We just saw **non-linearity of the data,** other potential issues include

- Correlation **in error terms**
- **Non-constant variance** in error terms
- **Outliers** – here you can find them by looking at the residuals and remove within a set threshold

Correlation in error

The most common way that error becomes correlated is with time series data

residuals

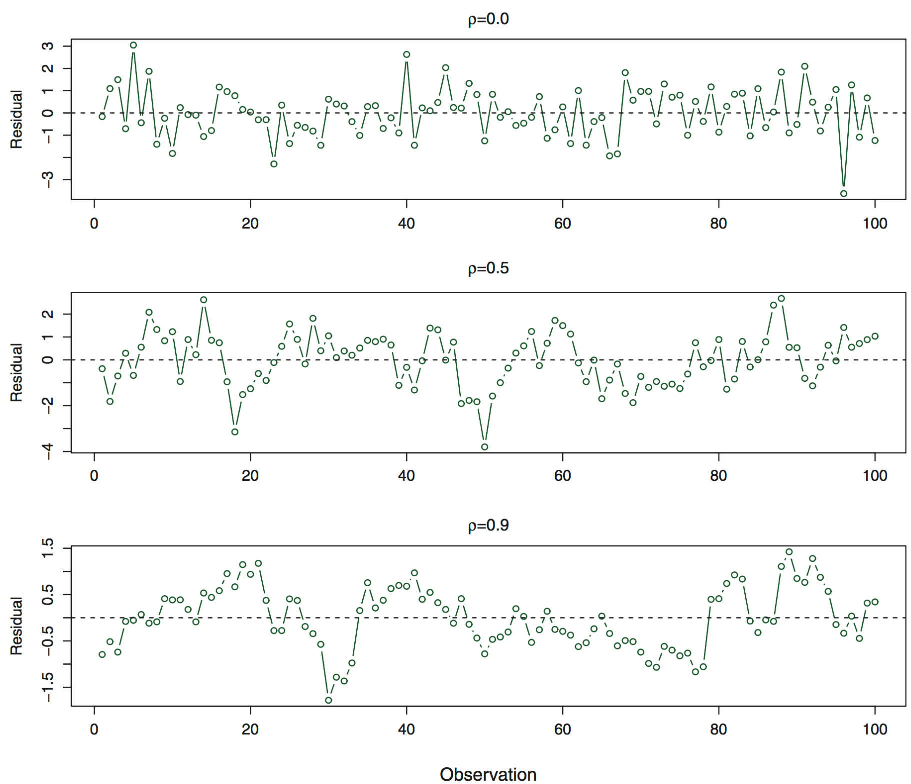


FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

Identifying correlation in error

Plot the residual vs. observation – see if there is some correlation

$$\rho_{xy} = \mathrm{Cor}(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

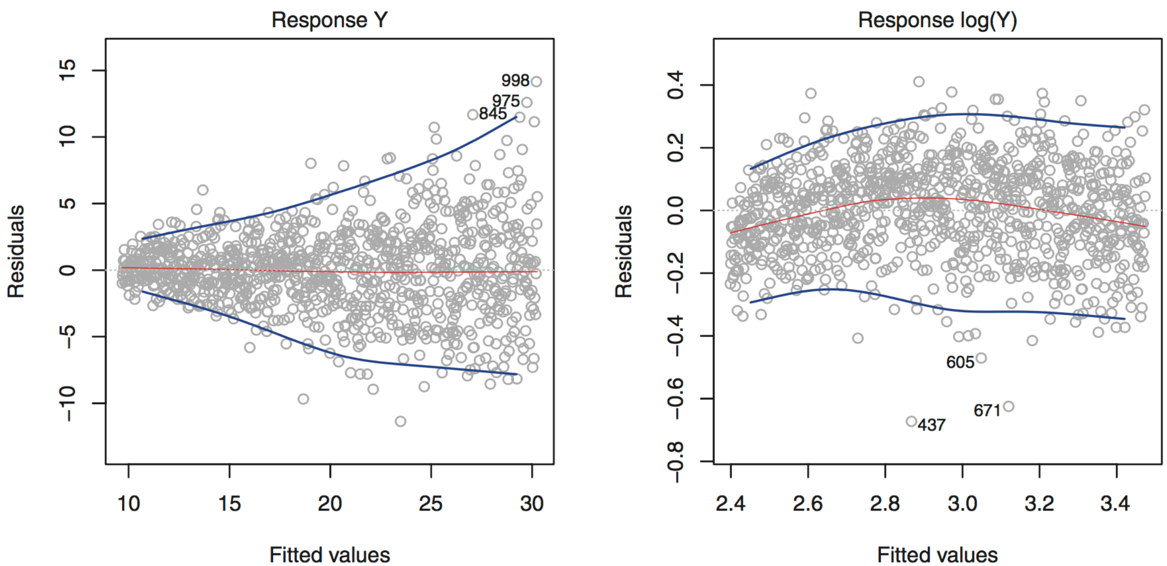
Introduction and practical implementation of methods to deal w/correlated errors is beyond scope of this class. See e.g. https://online.stat.psu.edu/stat462/node/189/

Non-constant variance

$$\mathrm{Var}(\epsilon_i) = \sigma^2 \neq \mathrm{const}$$

This phenomena is known as **heteroscedasticity**

- One solution is to **transform the response** data Y
- Another is to **use weighted least squares** – weights proportional to the variance

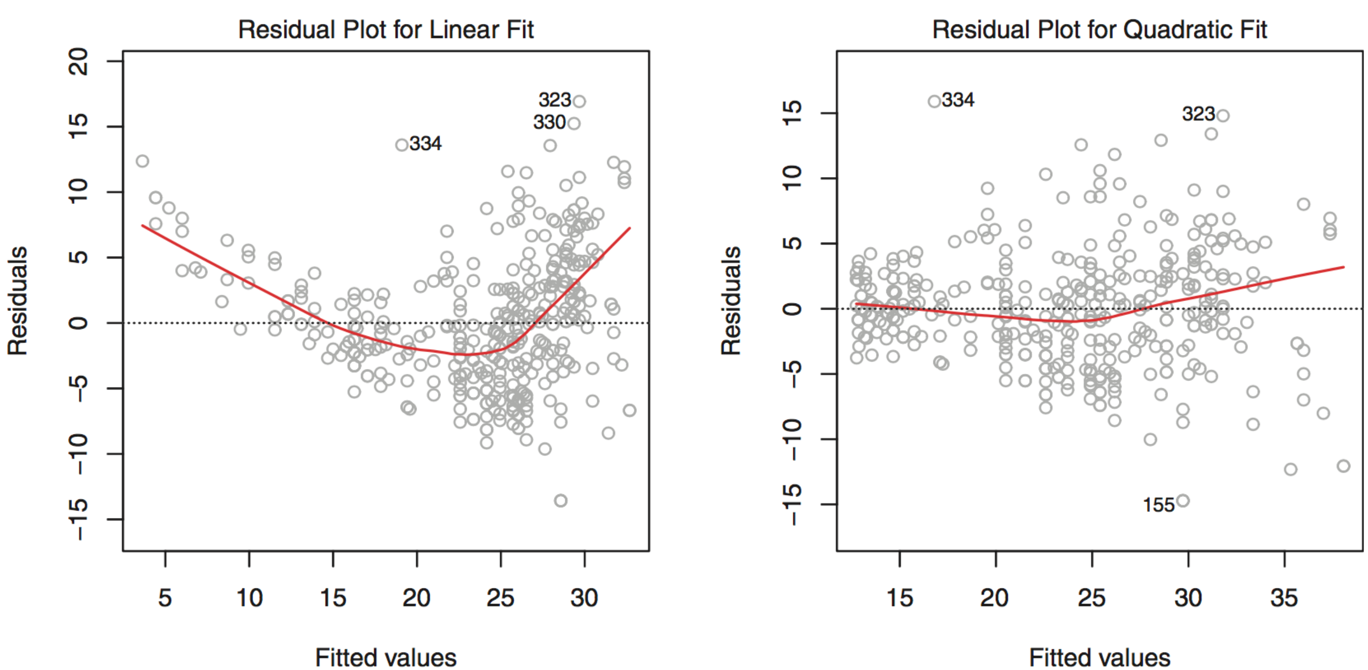


Beyond linear regression

Sometimes your variables have a clear

non-linear dependence on the response

We saw this in part in the notebook – also here Fig. 3.9 from the textbook



Simple nonlinear regression – polynomial

Note that in the case of some simple polynomial regressions, the model is still linear ...

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \ldots + \beta_p X_1^p + \epsilon$$

All we need to do is define

$$X_2 = X_1^2$$
$$X_3 = X_1^3$$
$$\ldots$$
$$X_p = X_1^p$$

For more examples, see Section 3.3.2 of the textbook

Other topics / suggestions

Chapter 3 of ISL is strongly suggested to read carefully (maybe multiple times)

Additional topics we didn't cover

- Outliers and high leverage points in your training set
- Collinearity
- More about nonlinear regression
- AND MANY MORE!