COMP9033

DATA ANALYTICS

6/12

LINEAR REGRESSION

DR. DONAGH HORGAN

DEPARTMENT OF COMPUTER SCIENCE
CORK INSTITUTE OF TECHNOLOGY

2017.03.08

Overview

0.1 / LAST WEEK

1. Data modelling:
   - Rule-based models
   - Statistical models.
   - Machine learning.
2. Sources of model error:
   - Underfitting and overfitting.
   - Bias, variance and irreducible error.
   - The bias-variance trade off.
3. Cross validation:
   - Split size.
   - Exhaustive vs. non-exhaustive.
   - Cross validation techniques.
   - Stratification.
   - Model selection.

0.2 / THIS WEEK

1. Linear regression:
    - What it is.
    - How it works.
    - Measuring model error.
2. The least squares technique:
    - The residual sum of squares.
    - The least squares solution.
    - Performance considerations.
3. Shrinkage methods:
    - Problems with least squares.
    - Ridge regression.
    - Hyperparameters.
4. Subset selection:
    - Best subset selection.
    - Forward stepwise selection.
    - Backward stepwise selection.
    - Hybrid methods.

Linear regression

1.1 / LINEAR REGRESSION

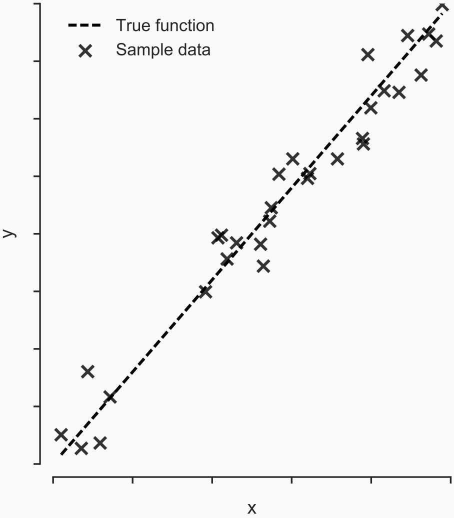
- The term *linear regression* describes a class of mathematical models that can be used to describe *quantitative* data.
- Generally, machine learning algorithms are used to build linear regression models, *e.g*.
    - The least squares technique.
    - Ridge regression.
    - The lasso technique.
    - Best subset selection.
- These are *supervised* machine learning algorithms, *i.e*. they learn from labelled data using both statistics and heuristics.

1.2 / EXAMPLE: LINE FITTING

- One familiar example of linear regression is the fitting of a straight line, *i.e.* the estimation of m and c in the equation

$$y = mx + c. \tag{6.1}$$

- For instance, the figure opposite shows some data that has been noisily sampled from the function $y = 2x + 1$.

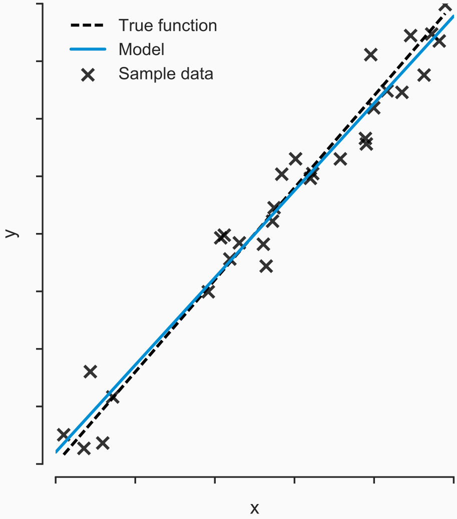


1.3 / EXAMPLE: LINE FITTING

- Using linear regression, we can create a model of a line that fits this data, *i.e.*

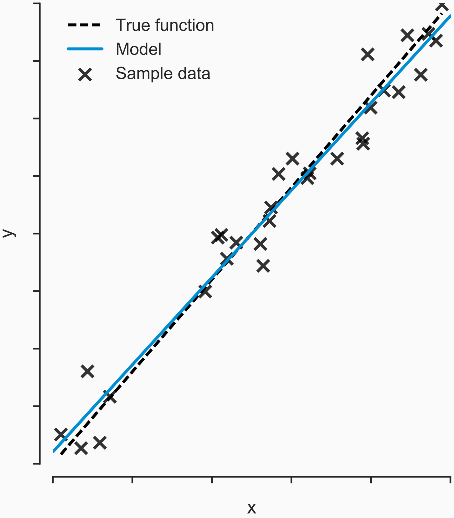
$$\hat{y} = 1.9x + 1.05,$$

  where $\hat{y}$ represents a *prediction* of the true value y, *i.e.* $\hat{y} \approx y$.
- In this case, we say that x is a predictor of y, *i.e.* if we know x then we can compute y.



1.4 / EXAMPLE: LINE FITTING

- More generally, we would like to fit coefficients to *many* predictors, *e.g.*
  - Predict temperature based on atmospheric pressure *and* wind speed (*two* predictors).
  - Predict server CPU load based on the number of active users, network throughput and disk I/O (*three* predictors).
  - Predict the price of Apple stocks based on the prices of other stocks (an *arbitrary* number of predictors).



1.5 / LINEAR REGRESSION

- Linear regression makes predictions based on multiple inputs, known as *predictors*.
- Typically, we have p predictors and we want to estimate some *quantitative* output function y, also known as the target.
- Linear regression does this by fitting an intercept (β_0) and p coefficients ($\beta_1, \beta_2, \ldots, \beta_p$) to the data[1] as

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \tag{6.2}$$

where $\hat{y}$ is the predicted value of y.

[1] In fact, Equation 6.1 is a just special case of Equation 6.2 with $p = 1$, $\beta_0 = c$ and $\beta_1 = m$.

- Predictors can take a number of forms, but must be *quantitative*:
    - An arbitrary quantitative input variable, *e.g.* temperature.
    - A polynomial transformation of an input variable, *e.g.* $x_2 = \sqrt{x_1}$.
    - An interaction between input variables, *e.g.* $x_3 = x_1 \cdot x_2$.
    - A dummy indicator variable encoding the values of some categoric input, *e.g.* $\{\text{True}, \text{False}\} \rightarrow \{1, 0\}$.
- Polynomial transformations and interactions can be used to account for non-linear relationships between the predictors and the target.
- Predictors should be centred, so that they have zero mean, before generating polynomial transformation or interaction terms.
- Dummy variables are used to account for the relationship between a categoric input variable and the target or may be the result of *one hot encoding*.

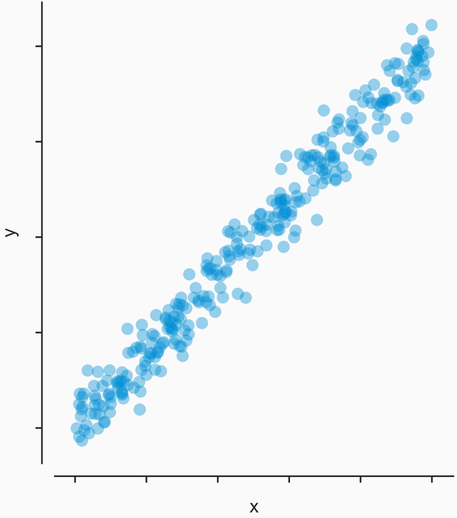
1.7 / DEALING WITH NON-LINEAR PREDICTORS

- Linear regression works well when we are dealing with *linear* systems, *i.e.* systems where there is a linear dependency between y and x, *e.g.*

$$y = 10x,$$

$$y = 2x_1 + 4x_2,$$

$$y = 3.5x_1 + 2x_2 + 4x_3 + 17.$$



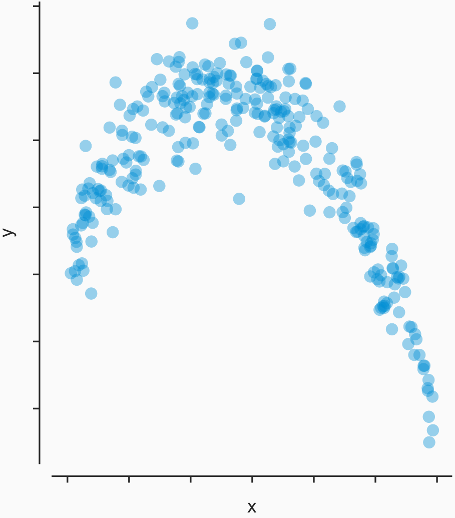
1.8 / DEALING WITH NON-LINEAR PREDICTORS

- However, things get a little trickier if we are dealing with *non-linear* systems, *i.e.* systems where y is a function of *powers* of x variables, *e.g.*

$$y = 10x^2,$$

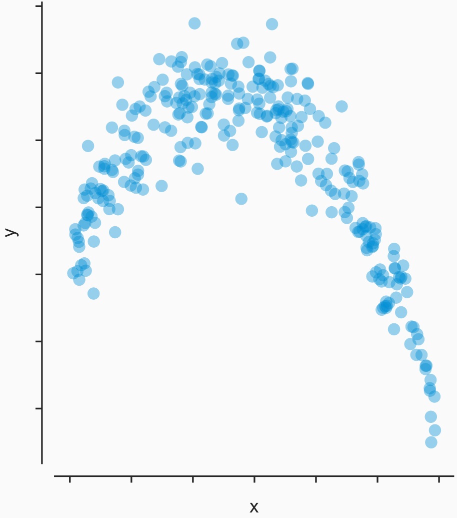
$$y = 2x_1^3 + 4x_2,$$

$$y = 3.5x_1^5 + 2x_2^3 + 4\sqrt{x_3} + 17.$$



1.9 / DEALING WITH NON-LINEAR PREDICTORS

- This is because Equation 6.2 computes predictions based on *linear* combinations of the predictors.
- If this assumption is violated, *e.g.* the relationship between *y* and its predictors is non-linear, Equation 6.2 will not produce a reasonable prediction.



1.10 / DEALING WITH NON-LINEAR PREDICTORS

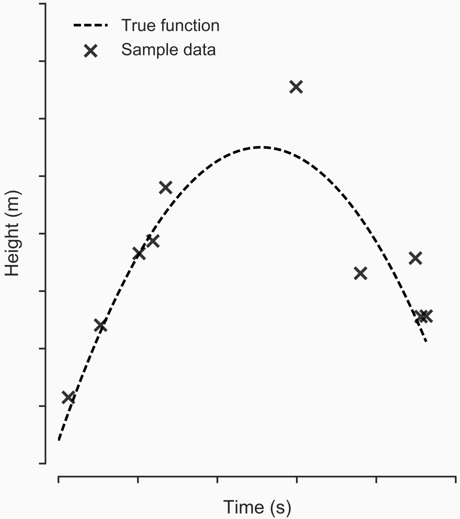
- If we know (or suspect) that there is a non-linear relationship between y and one of its predictors, then we can use *feature generation* to create a new predictor that adequately describes the relationship.
- For instance, if $y \sim x_i^n$, we can create a new feature $x_{p+1} = x_i^n$.
- If we include this new predictor in our regression, Equation 6.2 becomes

  $$\hat{y} = \beta_0 + \sum_{j=1}^{p+1} \beta_j x_j = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \beta_{p+1} x_{p+1} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \beta_{p+1} x_i^n,$$

  i.e. we have included the non-linear relationship in our prediction.
- Of course, we can also *replace* the original feature with the new feature, or even use cross validation to determine whether we should do so or not.

1.11 / EXAMPLE: NON-LINEAR PREDICTORS

- In an experiment, a rocket is projected into the air at a speed of 50 m s^{-1} from a height of 10 m.
- As the rocket travels through the air, its height above ground level is measured.
- What height is the rocket at after an arbitrary period of time, *t*?

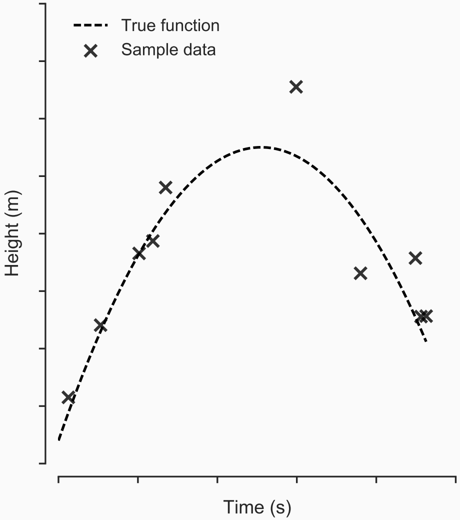


1.12 / EXAMPLE: NON-LINEAR PREDICTORS

- The true relationship between the height (h) and time (t) is well-known from the laws of physics:

$$h = -4.9t^2 + 50t + 10.$$

- However, we can approximate it using linear regression and feature selection.

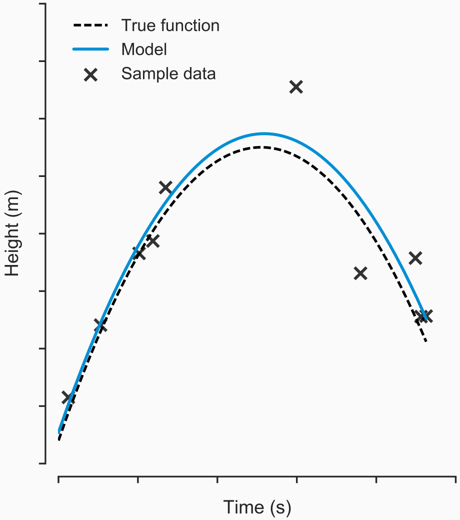


1.13 / EXAMPLE: NON-LINEAR PREDICTORS

- As we know that there is a non-linear (in this case, quadratic) relationship between h and t, we cannot use linear regression without feature generation.
- If we define the features $x_1 = t$ and $x_2 = t^2$ (*i.e.* generate it from x_1), then we can use linear regression to build a model:

$$\hat{h} = -4.8t^2 + 50.2t + 13.5,$$

which, as can be seen, is a reasonably accurate approximation.

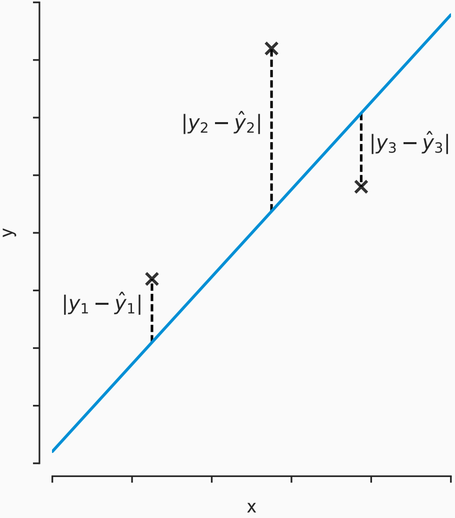


1.14 / MEASURING MODEL ERROR

- If we use a linear regression model to predict a single value $\hat{y}_i \approx y_i$, then the prediction error, ϵ_i, is given by:

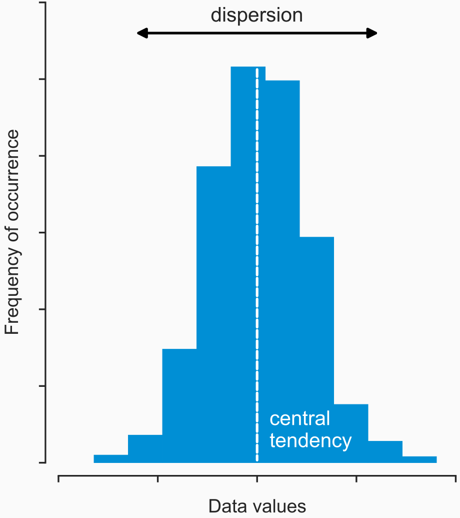
$$\epsilon_i = y_i - \hat{y}_i. \tag{6.3}$$

- Using Equation 6.3, we can infer that:
  - If $\epsilon_i < 0$, then we have *overestimated* the real value.
  - If $\epsilon_i > 0$, then we have *underestimated* the real value.
  - If $\epsilon_i = 0$, then our prediction was completely accurate.



1.15 / MEASURING MODEL ERROR

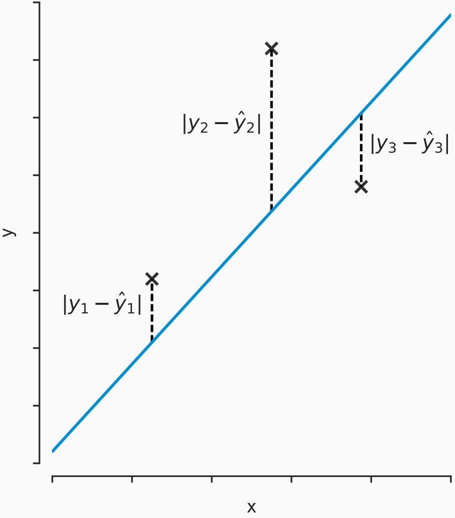
- Equation 6.3 is useful when dealing with single sample values, but we will usually generate many sample predictions in order to validate our model, *e.g.* if we use a test set in cross validation.
- We could simply average the errors, but this doesn't take their variation into account (*e.g.* low bias, high variance).
- Need some way to measure the *magnitude* of the errors.



1.16 / MEASURING MODEL ERROR: MEAN ABSOLUTE ERROR

- The *mean absolute error* (MAE) is one way to do this: it simply averages the absolute values of the sample errors, *i.e.*

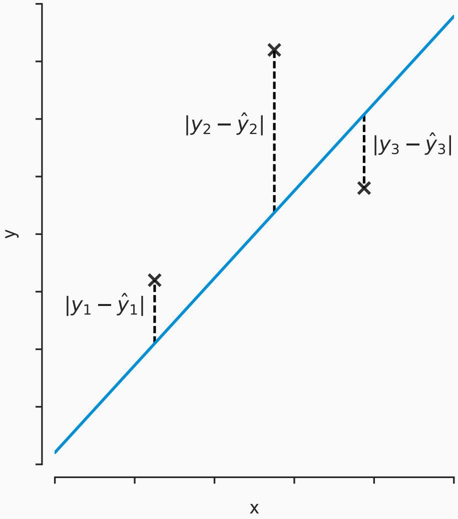
$$\begin{aligned}\text{MAE} &= \frac{1}{n}\sum_{i=1}^{n}|\epsilon_i| \\ &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \qquad (6.4)\end{aligned}$$



1.17 / MEASURING MODEL ERROR: ROOT MEAN SQUARE ERROR

- The *root mean square error* (RMSE) is another way to do this: it is the square root of the average of the squared errors, *i.e.*

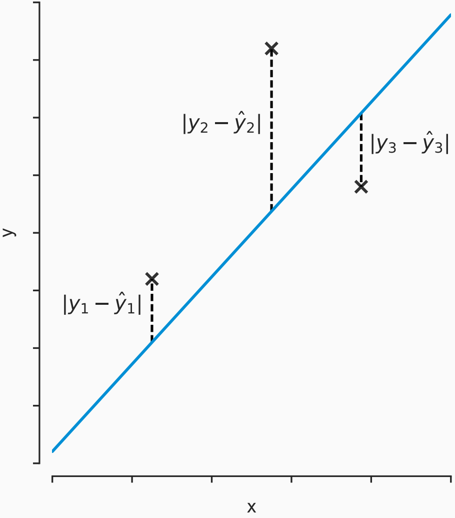
$$\begin{aligned} \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \end{aligned} \tag{6.5}$$



1.18 / MEASURING MODEL ERROR: MAE VS. RMSE

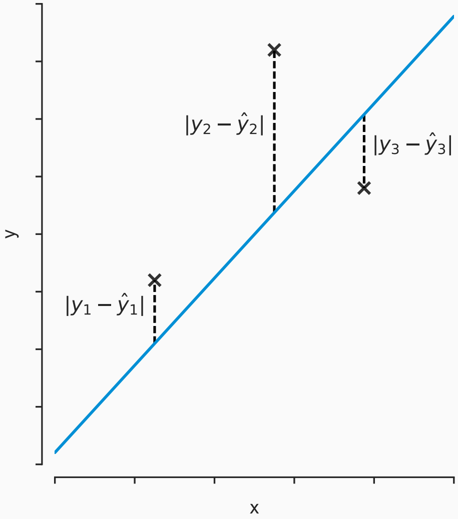
- MAE relies on the absolute value of the errors ($|\epsilon_i|$), whereas RMSE relies on the squares of the errors (ϵ_i^2).
- This is an important distinction as

$$\begin{aligned} \epsilon_i < 1 &\implies \epsilon_i^2 << 1, \\ \epsilon_i = 1 &\implies \epsilon_i^2 = 1, \\ \epsilon_i > 1 &\implies \epsilon_i^2 >> 1. \end{aligned}$$



1.19 / MEASURING MODEL ERROR: MAE VS. RMSE

- Consequently, if a sample contains a small number of large errors, then its RMSE tends to be larger than its MAE.
- This can be exploited in model selection to choose a model that produces fewer larger errors, at the cost of producing more smaller ones.



The least squares technique

- The *least squares technique* is a popular and widely used method for estimating the intercept and model coefficients in Equation 6.2.
- It does this by attempting to minimise a quantity known as the residual sum of squares (RSS), *i.e.*

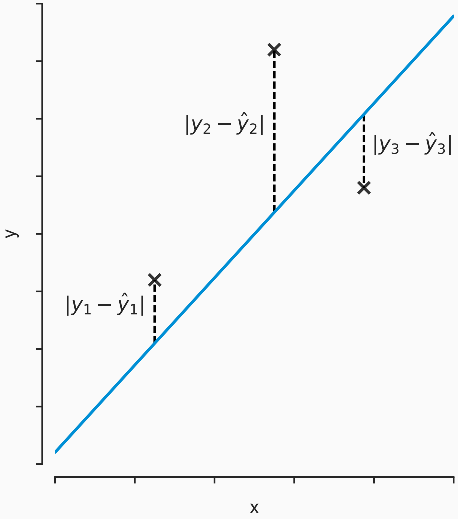
$$\mathrm{RSS} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} \epsilon_i^2. \tag{6.6}$$

- Noting the definition in Equation 6.5, the RSS is related to RMSE as

$$\mathrm{RSS} = n\mathrm{RMSE}^2. \tag{6.7}$$

2.2 / THE LEAST SQUARES TECHNIQUE

- The RSS is simply the sum of the squared distances between a candidate fit line and the sample data.
- Consequently, if we can choose our intercept and model coefficients well, then we can minimise our RSS and have a good chance of finding a line that fits the data well.
- Effectively, least squares reverses this process: by minimising the RSS, we hope to determine the intercept and coefficients that give the best fit line.



- Typically, we build linear regression models with p predictors, where each predictor has n values (*e.g.* time series).
- We can represent these predictors as a matrix, $\boldsymbol{X}$, where each row corresponds to a predictor and the first column is all ones, *i.e.*

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix}, \tag{6.8}$$

where $x_{i,j}$ represents the i^{th} data point of the j^{th} predictor and $x_{0,j} = 1 \; \forall \; j$ is a dummy variable that represents the intercept.

- In this case, Equation 6.2 becomes

$$\hat{\boldsymbol{y}} = X\boldsymbol{\beta}, \tag{6.9}$$

where

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

- Using Equation 6.9, it can be shown that the value of $\boldsymbol{\beta}$ that minimises the RSS is given by

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{6.10}$$

  where $\boldsymbol{X}^T$ is the transpose of the matrix $\boldsymbol{X}$ and $\boldsymbol{X}^{-1}$ represents the inverse of the matrix $\boldsymbol{X}$.

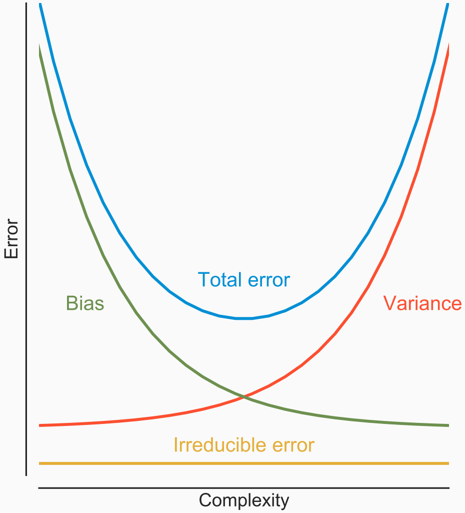
- Equation 6.10 is known as the *least squares* solution and is the *optimal* choice of $\boldsymbol{\beta}$ among all *unbiased* estimates.

2.6 / PERFORMANCE CONSIDERATIONS

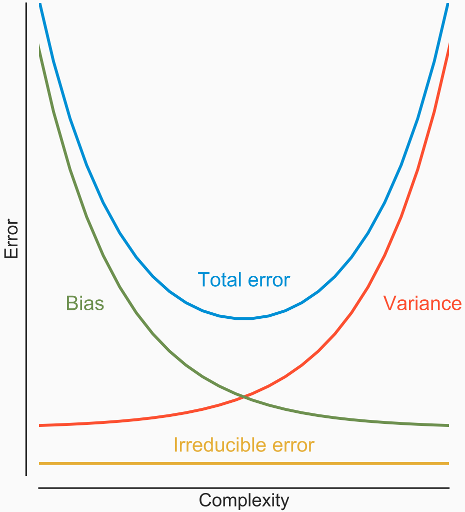
- In practice, n is often large and p may also be large, and so computing Equation 6.10 can become computationally expensive!
- If you are dealing with a large number of samples or predictors, then it’s best to use an optimised matrix multiplication library.
    - Typically, variants of BLAS and ATLAS are used - some are free/open source, some are proprietary.
    - Generally, the free/open source solutions are quite good.
- Alternatively, numerical methods can be used to find an *approximate* solution:
    - Stochastic gradient descent is often used.
    - Much faster than directly solving Equation 6.10 when n and/or p are large.
    - Can yield results close to the true value of $\boldsymbol{\beta}$, but sometimes yields a local optimum rather than a global one.

Shrinkage methods

- Least squares models tend to have low bias error but high variance error, *i.e.* they often overfit the data.
- If we use an overfitted model, then we are more likely to make mistakes when we evaluate new data.
  - Overfitted models tend to be unstable.
  - Small deviations in the magnitude of the input can produce large deviations in the magnitude of the output.



- Typically, this flaw is overcome by deliberately *biasing* the regression.
    - By building a model with increased bias error, we *should* lower our variance error.
    - If the model is overfit, then this will result in a model with lower overall error.
- Biasing typically has the effect of reducing the number of predictors (*i.e.* the complexity, see Equation 6.2) of a linear regression model, and vice-versa.



- Shrinkage methods aim to reduce the magnitudes of the linear regression coefficients, so that they have less of an effect on the final prediction:
    - If a predictor is unimportant, then its coefficient may shrink to zero, in which case it is eliminated (see Equation 6.2).
    - If a predictor is not *very* important, but still adds *some* value, then the magnitude of its coefficient is reduced, and so it has less of an effect on the predicted value, but does still have an effect.
    - If a predictor is important, then the magnitude of its coefficient should remain more or less unchanged.
- Examples include ridge regression and the lasso technique.

- Ridge regression is a shrinkage method for linear regression, where the regression coefficients are computed as

$$\boldsymbol{\beta} = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}, \tag{6.11}$$

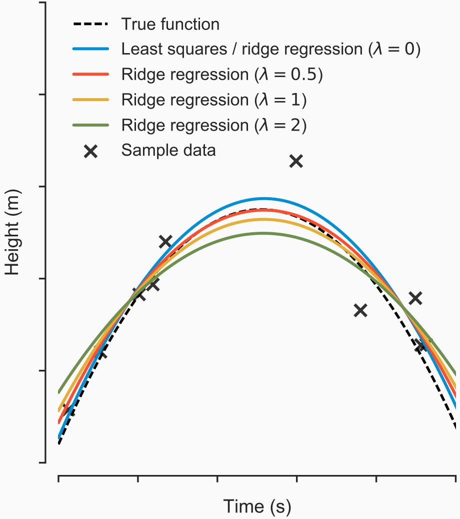
  where I denotes an *identity matrix* of size $p+1$ and $\lambda \in [0, \infty)$ is known as the ridge parameter.
- By adjusting the value of λ, we can directly introduce bias error into the model which, ideally, will move it closer to having minimum total error:
  - The *smaller* the value of λ, the *lesser* the amount of bias/shrinkage.
  - The *larger* the value of λ, the *greater* the amount of bias/shrinkage.
  - When $\lambda = 0$, ridge regression is equivalent to the least squares technique.

3.5 / HYPERPARAMETERS

- The ridge parameter is an example of a *hyperparameter*, *i.e.* it is an adjustable parameter that controls some aspect of the model building process.
- Generally, adjusting the hyperparameters of a model building algorithm affects the quality of model that is produced.
- Using model selection via cross validation, we can determine the optimum value of a particular hyperparameter or set of hyperparameters and, therefore, choose the optimal model.

3.6 / EXAMPLE: RIDGE REGRESSION

- If we apply ridge regression to our earlier rocket trajectory problem, we can produce a number of variations on the least squares fit.
- As λ increases, the fit becomes more biased.
- Can use model selection to pick the best model.



Subset selection

4.1 / SUBSET SELECTION

- Subset selection is an alternative set of methods for generating linear regression models with lower variance error than standard least squares models.
- Generally, subset selection works by building a least squares model using only a subset of the available predictors.
    - The *smaller* the subset, the *smaller* the complexity of the model.
    - The *larger* the subset, the *larger* the complexity of the model.
- In this way, the complexity of the linear regression model is reduced which should, in turn, reduce its variance error.
- If the regression is not computationally expensive to compute, then model selection can be used to determine the optimal subset size.

4.2 / SUBSET SELECTION: BEST SUBSET SELECTION

- The best subset selection algorithm determines the set of k best predictors from the total available set of p, where $k \leq p$.
- It's a similar idea to model selection via cross validation, but the selection is done internally in the algorithm without the aid of a test set.
- The main advantage of the technique is that the predictors selected by the algorithm are the *optimal* set of predictors of size k.
    - For instance, if we pick $k = 4$, the algorithm will find the best four predictors for our target variable.
- However, there are also disadvantages to the technique:
    - Determining the best predictors is an exhaustive process, and so the algorithm does not scale well when the number of candidate predictors becomes large.
    - There is no way to pick the optimal value for k (cross validation becomes less feasible as p becomes very large), and so we must typically rely on a heuristic approach.

4.3 / SUBSET SELECTION: FORWARD STEPWISE SELECTION

- Forward stepwise selection is a widely used alternative to best subset selection. It takes a *top down* approach, as follows:
  1. Compute a value for the intercept (β_0) by building a model with no predictors.
  2. Create a test model by adding a predictor (X_j) to the current model and determining its coefficient (β_j).
  3. Repeat Step 2 for each of the remaining predictors and evaluate the error resulting from each of the generated models. Choose the model with the lowest error to be the current model.
  4. Repeat Steps 2 and 3 until the fit ceases to improve or a desired number of predictors (k) is reached.
- The main advantage of this approach is that a much smaller number of evaluations are carried out than in best subset selection.
- However, as the algorithm is heuristic, the final result is not guaranteed to be optimal.

4.4 / SUBSET SELECTION: BACKWARD STEPWISE SELECTION

- Backward stepwise selection is a further alternative to best subset selection. It takes a *bottom up* approach, as follows:
    1. Build a linear regression model using all p predictors.
    2. Remove the weakest predictor in model. Typically, this is determined by some user-defined test.
    3. Repeat Step 2 until none of the remaining predictors fail the test or a desired number of predictors (k) is reached.
- As with forward stepwise selection, a much smaller number of evaluations are carried out than in best subset selection, although the final result is not guaranteed to be optimal.
- However, unlike forward subset selection, backward stepwise selection works back from the full model and so never misses a valuable predictor.
- One major disadvantage is that backward stepwise selection can only be used when the number of samples is greater than the number of predictors ($n > p$).

4.5 / HYBRID METHODS

- Subset selection excludes entire predictors from a model:
  - Reduces complexity, but can drop valuable predictors.
  - If too many predictors are dropped, there is a risk of underfitting (high bias).
- Shrinkage methods reduce the effect of predictors in a model:
  - Reduces complexity, but can include poor predictors if coefficients aren't shrunk to zero.
  - If too many predictors are retained, there is a risk of overfitting (high variance).
- Subset selection can be combined with shrinkage methods to get the benefits of both, *e.g.* forward stepwise selection with ridge regression.
- Again, can use model selection to optimise hyperparameters (ridge parameter, subset size) as required.

Summary

X.1 / SUMMARY

- Linear regression:
    - Least squares: straightforward, but tends to overfit (no hyperparameters).
    - Shrinkage methods: reduce the effect of weak predictors.
    - Subset selection: eliminate weak predictors entirely.
    - Hybrid methods: combine benefits of subset selection and shrinkage methods.
- This week's lab:
    - Build a linear regression model using the least squares method.
    - Detect anomalies in time series data.
- Next week: k nearest neighbours.

X.2 / REFERENCES

1. Hastie et al. *The elements of statistical learning: data mining, inference and prediction.* 2nd edition, February 2009. (stanford.io/1dLkiAv)
2. Ullman et al. *Mining of massive data sets.* Cambridge University Press, 2014. (stanford.io/1qtgAYh)