Unit Schedule: Modules

Module	**Week**	**Content**	**Ross**
1.	1	introduction to modelling	1,2
2.	2	probability refresher	3
	3	random vars & expected values	4
	4	special distributions	5
3.	5	statistical inference	6&7
	6	confidence intervals	7
	7	hypothesis testing	8
4.	8	**dependence & linear regression**	9
	9	classification, clustering & mixtures	
5.	10	random numbers & simulation	15(bits)
	11	basic machine learning	
6.	12	modelling, validation & review	

Revision at *https://flux.qa/43FMK4*

FIT5197 Statistical Data Modelling

Module 4

Linear Regression

2020 Lecture 8

Monash University

Regression (ePub sections 4.1, Ross 9.1, 9.2, 9.6, 9.7, 9.9, 9.10)

Outline

Regression Example

Regression

Least Squares Fitting

Multiple Linear Regression

Maximum Likelihood Estimation

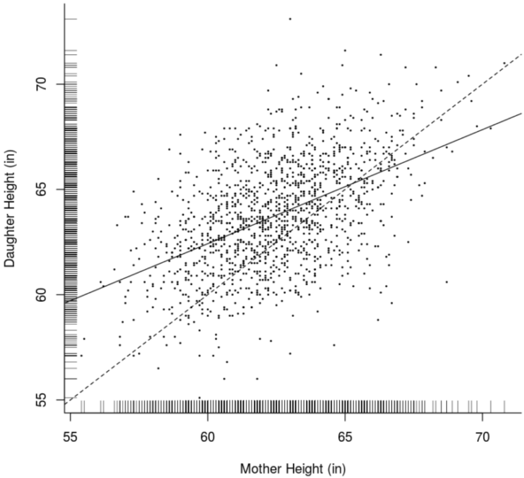
Model Selection

Inheritance of Height?

In the late nineteenth century, a statistician named Pearson collected heights of 1375 mothers and one of their adult daughters:

- are mother and daughter heights correlated?
- are mothers or daughters taller on average?
- what is the quantitative relationship?

Inheritance of Height



(dotted line)
Daughter's Height =
0 + 1 × Mother's Height

(solid line)
Daughter's Height =
29.92 + 0.54 × Mother's Height

NB. note regression to the mean

Inheritance of Height?

Questions:

- are mother and daughter heights correlated?
  - moderately
- are mothers or daughters taller on average?
  - daughters slightly taller
- what is the quantitative relationship?
  - Daughter's Height = 29.92 + 0.54 × Mother's Height

Supervised Learning

- learning under “supervision”
- a form of predictive analytics
- one variable is predicted from a set of other variables
    - training requires that the “correct” values are provided for that one variable in the training set
- one of the most common statistical tasks

Supervised Learning

- learning under “supervision”
- a form of predictive analytics
- one variable is predicted from a set of other variables
  - training requires that the “correct” values are provided for that one variable in the training set
- one of the most common statistical tasks

- if predicting a **numerical** variable, called regression

  Example: predicting the quality of a wine from chemical and seasonal information
- if predicting a **categorical** variable, called classification

  Example: predicting if someone has diabetes from medical measurements

Outline

Regression Example

Regression

Least Squares Fitting

Multiple Linear Regression

Maximum Likelihood Estimation

Model Selection

Regression

- The variable predicted is designated the “y” variable
  - we have $(y_1, \ldots, y_n)$
- This variable is often called the:
  - target; response;
  - output;
  - outcome.

Regression

- The variable predicted is designated the “y” variable
  - we have $(y_1, \ldots, y_n)$
- This variable is often called the:
  - target; response;
  - output;
  - outcome.

- The other variables used as inputs for prediction are usually designated “X” variables
  - we have $(x_{i,1}, \ldots, x_{i,p})$ for $i = 1, \ldots, n$
- These variables are often called the
  - explanatory variables;
  - predictors; covariates;
  - inputs; attributes;
  - exposures.

Regression

- The variable predicted is designated the “y” variable
  - we have $(y_1, \ldots, y_n)$
- This variable is often called the:
  - target; response;
  - output;
  - outcome.

- The other variables used as inputs for prediction are usually designated “X” variables
  - we have $(x_{i,1}, \ldots, x_{i,p})$ for $i = 1, \ldots, n$
- These variables are often called the
  - explanatory variables;
  - predictors; covariates;
  - inputs; attributes;
  - exposures.
- Usually we assume the targets are random variables and the predictors are known without error

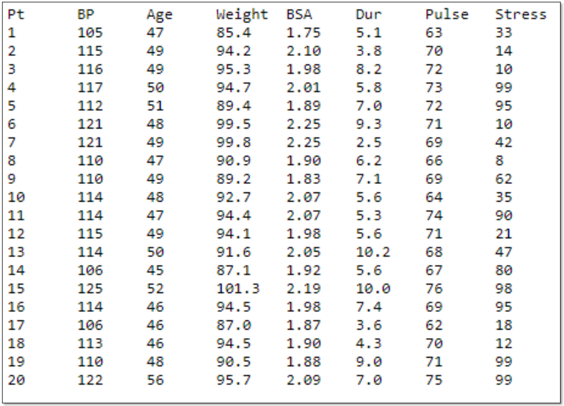
Linear Regression

- Linear regression is a special type of supervised learning
- In this case, we construct a function that relates the predictors to the target as being linear

- One of the most important models in statistics
  - The resulting model is highly interpretable
  - It is very flexible. It can even handle nonlinear relationships!
  - It is computationally efficient to fit, even for very large p
- Enormous area of research and work
  $\Longrightarrow$ we will get acquainted with the basics

Example Data

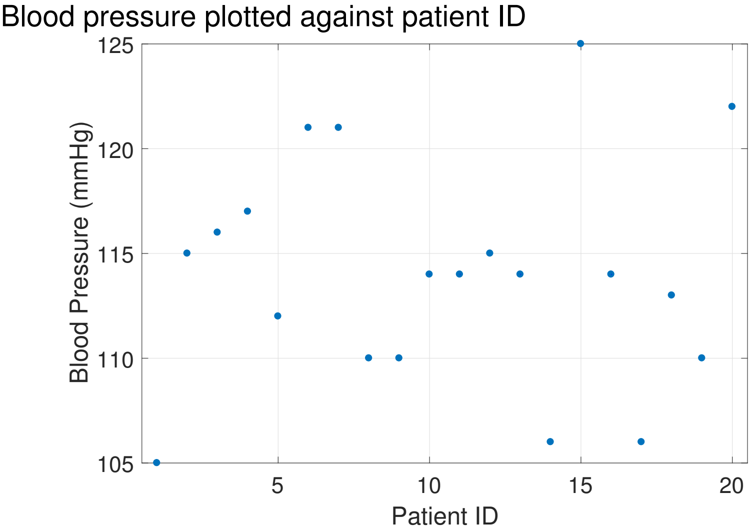
Consider the following dataset:



Pt	BP	Age	Weight	BSA	Dur	Pulse	Stress
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.10	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7.0	72	95
6	121	48	99.5	2.25	9.3	71	10
7	121	49	99.8	2.25	2.5	69	42
8	110	47	90.9	1.90	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90
12	115	49	94.1	1.98	5.6	71	21
13	114	50	91.6	2.05	10.2	68	47
14	106	45	87.1	1.92	5.6	67	80
15	125	52	101.3	2.19	10.0	76	98
16	114	46	94.5	1.98	7.4	69	95
17	106	46	87.0	1.87	3.6	62	18
18	113	46	94.5	1.90	4.3	70	12
19	110	48	90.5	1.88	9.0	71	99
20	122	56	95.7	2.09	7.0	75	99

Imagine we want to model blood pressure

Blood Pressure



Modelling Blood Pressure

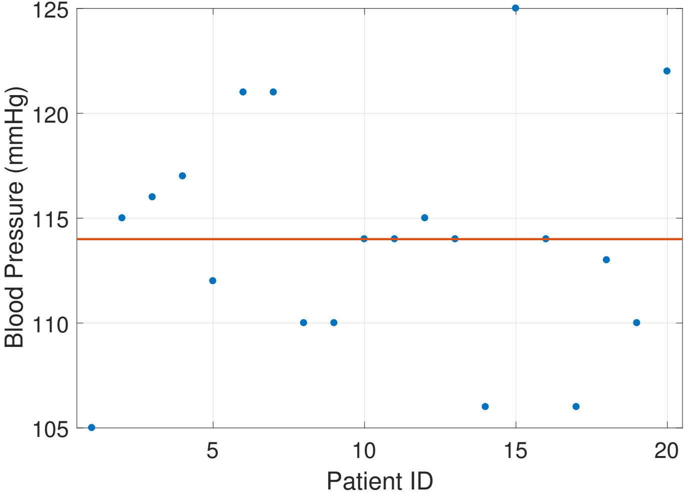
- Our blood pressure variable $\mathrm{BP}_1, \ldots, \mathrm{BP}_{20}$ is continuous $\Longrightarrow$ we choose to model it using a normal distribution
- The maximum likelihood estimate of the mean μ is

$$\hat{\mu} = \frac{1}{20} \sum_{i=1}^{n} y_i = 114$$

  which is equivalent to the sample mean
- We have a new person from the population this sample was drawn from and we want to predict their blood pressure
- Using our simple model our best guess of this persons blood pressure is 114, i.e., the estimated mean $\hat{\mu}$

Modelling Blood Pressure

Prediction of BP using the mean



Modelling Blood Pressure

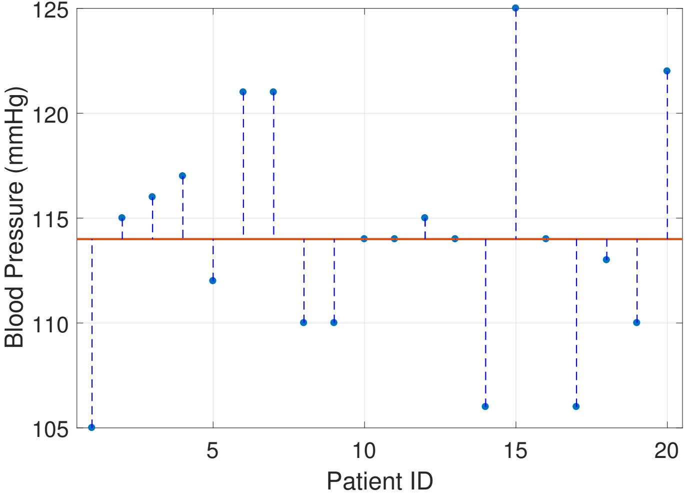
- How good is our model at predicting?
- One way we could measure this is through prediction error
- We don't know future data, but we can look to see how well it predicts the data we have
- Let $\hat{y}_i$ denote the prediction of sample y using a model; then

  $$e_i = \hat{y}_i - y_i$$

  are the errors between our model predictions $\hat{y}_i$ and the observed data y_i
  $\Longrightarrow$ often called residual error, or just residuals
- A good fit would lead to overall small errors

Predicting Blood Pressure

Prediction of BP using the mean, showing errors/residuals



Predicting Blood Pressure

- We can summarise the total error of fit of our model by

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2$$

which is called the residual sum-of-squared errors (RSS).

- For our simple mean model $\mathrm{RSS} = 560$

Predicting Blood Pressure

- We can summarise the total error of fit of our model by

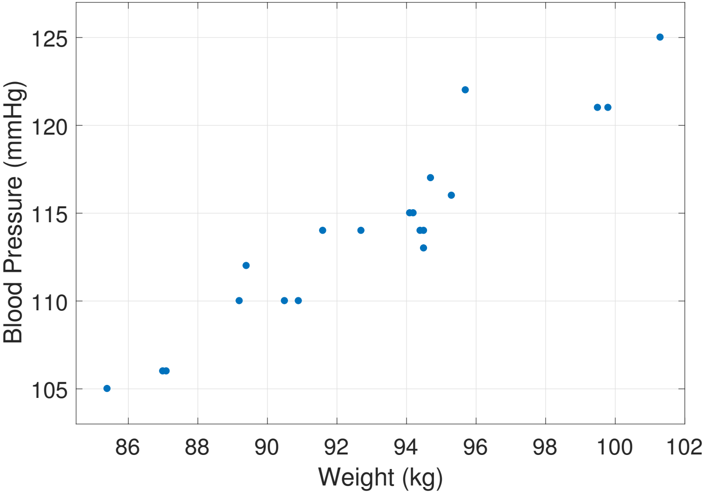
$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2$$

which is called the residual sum-of-squared errors (RSS).

- For our simple mean model $\mathrm{RSS} = 560$
- **Can we do better** (smaller error) if we use one of the other measured variables to help predict blood pressure?
- For example, if we took a persons weight into account, could we build a better predictor of their blood pressure?
- To get an idea if there is scope for improvement we can plot blood pressure vs weight

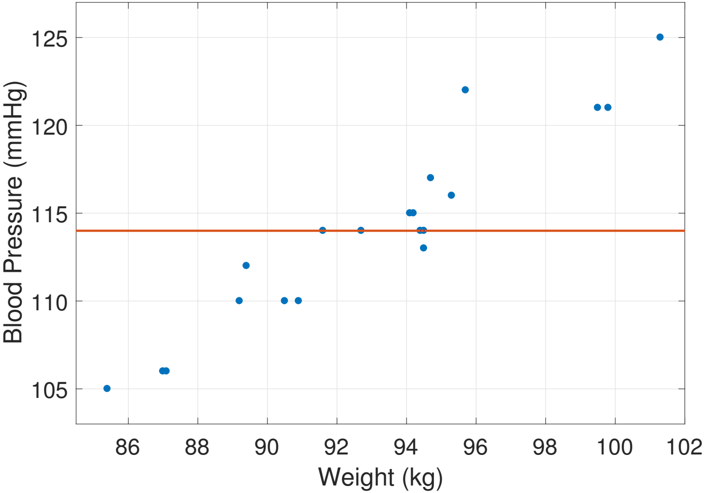
Blood Pressure vs. Weight

Blood pressure vs weight – BP appears to increase with weight



Simple Linear Regression

Our simple mean model is clearly not a good fit



Simple Linear Regression

- Our simple mean model predicts blood pressure by

$$\mathbb{E}\left[\mathrm{BP}_i\right] = \mu$$

  irrespective of any other data on individual *i*

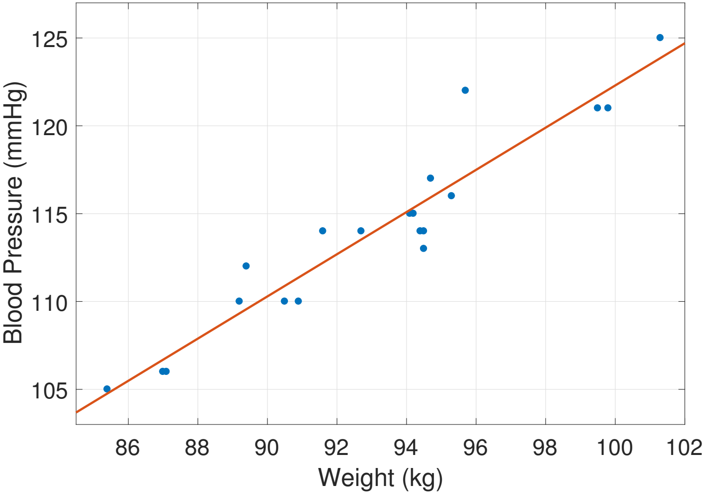
- Let $(\mathrm{Weight}_1, \ldots, \mathrm{Weight}_{20})$ be the weights of our 20 individuals
- We assume the mean is a linear function of weight, i.e.,

$$\mathbb{E}\left[\mathrm{BP}_i \,|\, \mathrm{Weight}_i\right] = \beta_0 + \beta_1 \mathrm{Weight}_i$$

- This says that the conditional mean of blood pressure BP_i for individual *i*, given the individual's weight Weight_i, is equal to β_0 plus β_1 times the weight Weight_i
- Note our simple mean model is a linear model with $\beta_1 = 0$

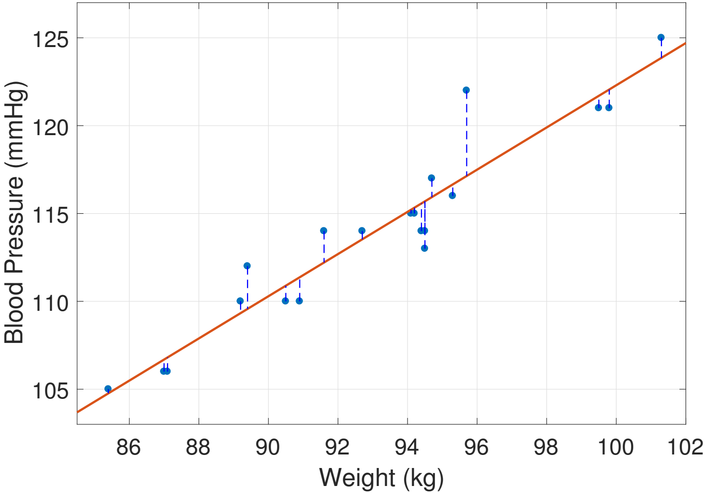
Simple Linear Regression

The linear model $\mathbb{E}\left[\mathrm{BP}_i \,|\, \mathrm{Weight}_i\right] = 2.2053 + 1.2009\,\mathrm{Weight}_i$



Simple Linear Regression

Residuals; $e_i = \mathrm{BP}_i - 2.2053 - 1.2009\,\mathrm{Weight}_i$ $(\mathrm{RSS} = 120)$



Simple Linear Regression

- A linear model of the form

$$\mathbb{E}\left[Y_i \,|\, x_i\right] = \hat{y}_i = \beta_0 + \beta_1 x_i$$

  is called a simple linear regression.

- It has two free regression parameters
  - β_0 is the intercept; it is the value of the predicted value $\hat{y}_i$ when the predictor $x_i = 0$
  - β_1 is a regression coefficient; it is the amount the predicted value $\hat{y}_i$ changes with one unit change of the predictor x_i

Simple LR Result

- In our example y_i is blood pressure and x_i weight;

  $$\hat{y}_i = 2.2053 + 1.2009x_i = (2.2053 + \hat{X}) + 1.2009(x_i - \hat{X})$$

  so

  - For every additional kilogram a person weighs, their blood pressure increases by $1.2009 mmHg$
  - For a person who weighs zero kilograms, the predicted blood pressure is $2.2053 mmHg$
- The predictions might not make sense outside of sensible ranges of the predictors!

Outline

Regression Example

Regression

Least Squares Fitting

Multiple Linear Regression

Maximum Likelihood Estimation

Model Selection

Fitting Linear Regressions

- How did we arrive at $\hat{\beta}_0 = 2.2053$ and $\hat{\beta}_1 = 1.2009$ in our blood pressure vs weight example?
- Measure fit of a model by its RSS (called SS_R in Ross)

$$\begin{aligned}\mathrm{RSS} &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} e_i^2\end{aligned}$$

- Smaller error = better fit

Using Least Squares

- So least-squares principle says we choose (estimate) β_0, β_1 to minimise the RSS
- Formally

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

- These are often called least-squares (LS) estimates.
- There are alternative measures of error; for example least sum of absolute errors.
- Least squares is popular due to simplicity, computational efficiency and connections to normal models

Working Least Squares

(optional)

- The RSS is a function of β_0, β_1, i.e.,

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- The least-squares estimates are the solutions to the equations

$$\begin{aligned} \frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned}$$

where we use the chain rule.

Least Squares Solution

The solution for β_0 is

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} y_i x_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and the solution for β_1 is

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{n} y_i x_i\right) - \hat{\beta}_0 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

Least Squares Solution

- With some rearrangement ...
- The solution for $\hat{\beta}_0, \hat{\beta}_1$ is

$$\hat{\beta}_0 = \frac{\overline{Y}\,\overline{X^2} - \overline{XY}\,\overline{X}}{\overline{X^2} - \overline{X}^2} \qquad \hat{\beta}_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}$$

- Can also represent it using the following:

$$\begin{aligned} \mathrm{SS}_{XX} &= \sum_{i=1}^{n} (x_i - \overline{X})^2 = n\left(\overline{X^2} - \overline{X}^2\right) \\ \mathrm{SS}_{XY} &= \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}) = n\left(\overline{XY} - \overline{X}\,\overline{Y}\right) \\ \hat{\beta}_1 &= \frac{\mathrm{SS}_{XY}}{\mathrm{SS}_{XX}} \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \end{aligned}$$

- Diagnostics for β_0 and β_1 done later.

Fitting Linear Regressions

- Given LS estimates $\hat{\beta}_0$, $\hat{\beta}_1$ we can find the predictions for our data

  $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  and residuals

  $$e_i = y_i - \hat{y}_i$$

Fitting Linear Regressions

- Given LS estimates $\hat{\beta}_0$, $\hat{\beta}_1$ we can find the predictions for our data

  $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  and residuals

  $$e_i = y_i - \hat{y}_i$$

- The vector of residuals $\mathbf{e} = (e_1, \ldots, e_n)$ has the properties

  $$\sum_{i=1}^{n} e_i = 0 \text{ and } \mathrm{corr}\left(\mathbf{x}, \mathbf{e}\right) = 0$$

  where $\mathbf{x} = (x_1, \ldots, x_n)$ is our predictor variable.

- This means least-squares fits a line such that the mean of the resulting residuals is zero, and the residuals are uncorrelated with the predictor.

Outline

Regression Example

Regression

Least Squares Fitting

Multiple Linear Regression

Maximum Likelihood Estimation

Model Selection

Multiple Linear Regression

- We have used one explanatory variable in our linear model
- A great strength of linear models is that they easily handle multiple variables
- Let $x_{i,j}$ denote the variable j for individual i, where $j = 1, \ldots, p$; i.e., we have p explanatory variables. Then

$$\mathbb{E}\left[y_i \,|\, x_{i,1}, \ldots, x_{i,p}\right] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$$

  - ▶ The intercept is now the expected value of the target when $x_{i,1} = x_{i,2} = \cdots = x_{i,p} = 0$
  - ▶ The coefficient β_j is the increase in the expected value of the target per unit change in explanatory variable j

Multiple Linear Regression

- Fit a multiple linear regression using least-squares $\Longrightarrow$ assume $p < n$, otherwise solution is non-unique
- Given coefficients β_0, β_1,...,β_p the RSS is

$$\mathrm{RSS}(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2$$

- Now we have to solve

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p = \underset{\beta_0, \beta_1, \ldots, \beta_p}{\arg\min} \left\{ \mathrm{RSS}(\beta_0, \beta_1, \ldots, \beta_p) \right\}$$

- Efficient algorithms exist to find these estimates
- Usually done with linear algebra packages

Multiple LR Setup

- Matrix algebra can simplify linear regression equations
  - We have a vector of targets $\mathbf{y} = (y_1, \ldots, y_n)$
  - We have a vector of coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$
  - We can treat each variable as a vector $\mathbf{x}_j = (1, x_{1,j}, \ldots, x_{n,j})$
- Note we add "1" as an extra attribute to account for the intercept β_0
- Arrange these vectors into a matrix $\mathbf{X}$ of predictors:

$$\mathbf{X} = (\mathbf{1}, \mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_p') = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix},$$

- We call this the design matrix
  $\Longrightarrow$ has $p+1$ columns (predictors) and n rows (individuals)

Multiple LR Setup

- We can form our predictions and residuals using

$$\hat{\mathbf{y}} = \mathbf{X}\beta \text{ and } \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}.$$

- We can then write our RSS very compactly as

$$\mathrm{RSS}(\boldsymbol{\beta}) = \mathbf{e}'\mathbf{e}$$

- If $\hat{\beta}$ are least-squares estimates, then

$$\mathrm{corr}(\mathbf{x}_j, \mathbf{e}) = 0 \text{ for all } j$$

- That is, the residuals of the least-squares solution are uncorrelated with all predictors in the model

MLR Solution (Ross 9.10)

(optional)

- The least squares and MLE solution for this is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- Moreover, this is an unbiased estimate:

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}$$

- The covariance is

$$\mathrm{cov}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Moreover, $\frac{1}{\sigma^2}\mathrm{RSS}(\boldsymbol{\beta}) \sim \chi^2_{n-p-1}$
- Thus an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2_u = \frac{1}{n-p-1}\mathrm{RSS}(\boldsymbol{\beta})$$

Computing RSS (Simple LR)

Ross 9.3 gives a simple expression for the RSS

$$\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1) = \frac{\mathrm{SS}_{XX}\mathrm{TSS} - \mathrm{SS}_{XY}^2}{\mathrm{SS}_{XX}}$$

where *TSS* is called the total sum of squares given by

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

which is the residual sum-of-squares obtained by fitting the intercept only (the "mean model").

R-squared (R^2)

- Residual sum-of-squares tells us how well we fit the data
- But the scale is arbitrary – what does an RSS of 2.352 *mean?*
- Instead, we define the RSS relative to some reference point, the total sum-of-squares as the reference:

R-squared (R^2)

- The R^2 value is then defined as

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

- For simple linear regression:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{SS}_{XY}^2}{\mathrm{SS}_{XX}\mathsf{TSS}} = r_{XY}^2$$

- r_{XY} is the correlation between predictor and target.
- In multiple regression $R^2 = r_{\hat{\mathbf{y}}\mathbf{y}}^2$
- R^2 is also called the coefficient-of-determination

R-squared (R^2)

- We say that R^2 is the proportion of the variance of the target variable explained by the model
- R^2 is strictly between 0 (model has no explanatory power) and 1 (model completely explains the data)
- The higher the R^2 the better the fit to the data
- Adding an extra predictor **always** increases R^2
  $\Longrightarrow$ predictors that greatly increase R^2 are potentially important

Example: Multiple LR and R^2

- Let us revisit our blood pressure data
- The residual sum-of-squares of our mean model was 560 $\Longrightarrow$ this is our reference model (total sum-of-squares)
- Regression of blood pressure (BP) onto weight gave us

$$\mathbb{E}\left[\mathrm{BP} \mid \mathrm{Weight}\right] = 2.20 + 1.2\,\mathrm{Weight}$$

which had an RSS of 54.52 $\Rightarrow R^2 \approx 0.9$

Example: Multiple LR and R^2

- In our data we also have an individual's age
- We fit a multiple linear regression of BP onto weight and age

$$\mathbb{E}\left[\mathrm{BP} \,|\, \mathrm{Weight}, \mathrm{Age}\right] = -16.57 + 1.03\,\mathrm{Weight} + 0.71\,\mathrm{Age}$$

- This says that:
  - for every kilogram, a person's bloodpressure rises by $1.03 mmHg$;
  - for every year, a person's bloodpressure rises by $0.71 mmHg$;
- This model has an RSS of 4.82 $\Rightarrow R^2 = 0.99$
- So including age seems to increase our fit substantially

Another Example: Exam like problem

- First lets do some revision of last weeks lecture at *https://flux.qa/43FMK4*
- Example simple linear regression problem at *https://stattrek.com/regression/regression-example.aspx*
- During the exam relevant formulas can be found in the *formula sheet* provided on the 'Exam' page in the unit's Moodle site

Handling Categoricals

- Sometimes our predictors are categorical variables
- This means the numerical values they take are on just codes for different categories
- Makes no sense to "add" or "multiply" them
- Examples: nationality, sex, treatment group, ...
- Instead we turn them into $K - 1$ new predictors (if K is the number of categories)
- These predictors take on a one when an individual is in a particular category, and zero otherwise
- They are called indicator variables.
- Example: convert the variable *type* taking four values "a", "b", "c" and "d" into three Booleans denoted $1_{\mathit{type}=\text{“b”}}$, $1_{\mathit{type}=\text{“c”}}$, $1_{\mathit{type}=\text{“d”}}$ and "a" is treated as default

Handling Categoricals

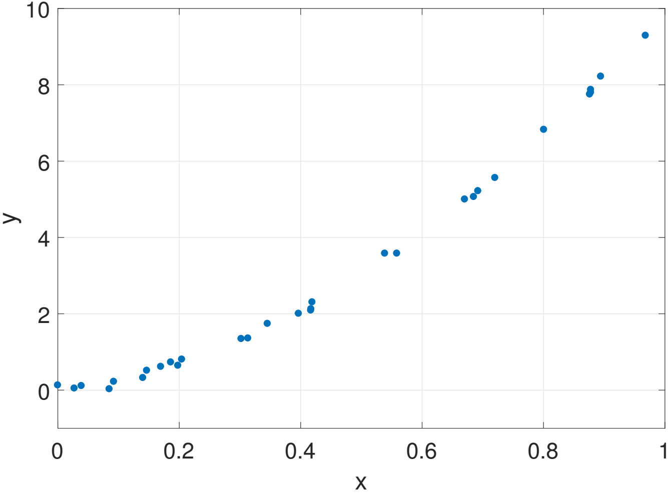
- Example variable with four values coded as *a*, *b*, *c* and *d*

$$\begin{pmatrix} a \\ b \\ a \\ c \\ d \\ b \\ c \\ b \\ c \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

- We do not build indicators for first category
- Regression coefficients for other categories are increases in target *relative* to being in the first category

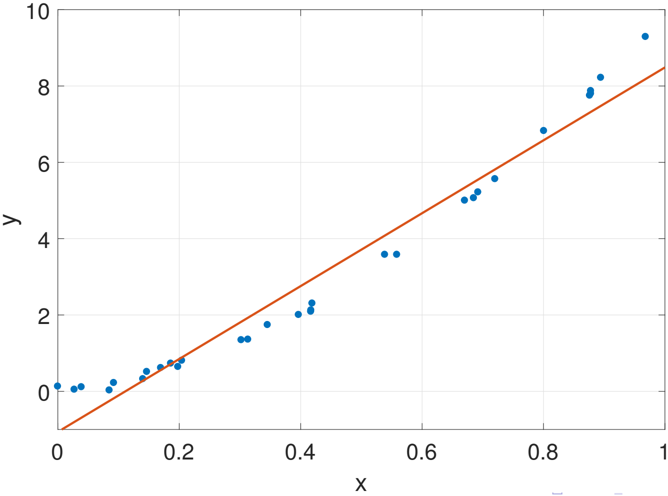
Nonlinear Effects

- What if the data is non-linear?



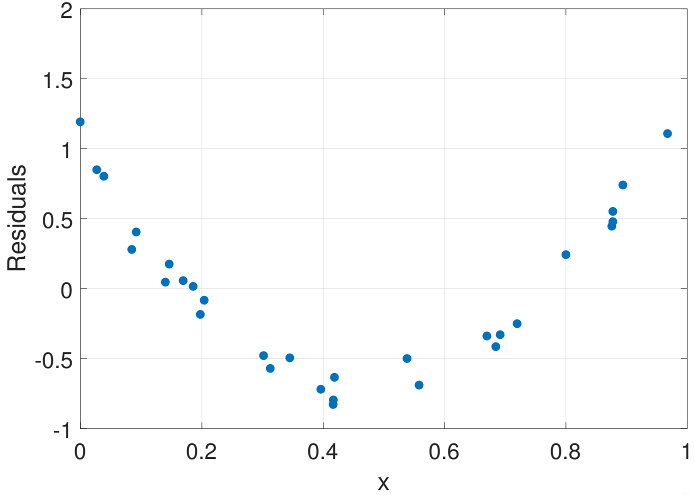
Nonlinear Effects: Linear Fit

- Suppose we do a linear regression
- Fitted model: $\hat{y} = -1.07 + 9.55x$; $R^2 = 0.95$



Nonlinear Effects: Residuals

- Now lets plot the residuals
- residuals exhibit clear nonlinear trend



Nonlinear Effects

- A pattern in the residuals is a sign of a problem with the model (why?)

Nonlinear Effects

- A pattern in the residuals is a sign of a problem with the model (why?)
- Because it means we can predict when the model will overestimate or underestimate the target
  $\Longrightarrow$ we can do better!

Nonlinear Effects

- A pattern in the residuals is a sign of a problem with the model (why?)
- Because it means we can predict when the model will overestimate or underestimate the target
  $\Longrightarrow$ we can do better!
- In our example the residuals show a nonlinear pattern.
- We can still use linear regression to fit nonlinear models by transforming the predictors

Transformations

- There are several common transformations
- A logarithmic transformation can be used if we expect a percentage increase in predictor x to be associated with a constant increase in target y

$$x_{i,j} \;\Rightarrow\; \log x_{i,j}$$

Can only be used if all $x_i > 0$

Transformations

- There are several common transformations
- A logarithmic transformation can be used if we expect a percentage increase in predictor x to be associated with a constant increase in target y

  $$x_{i,j} \;\Rightarrow\; \log x_{i,j}$$

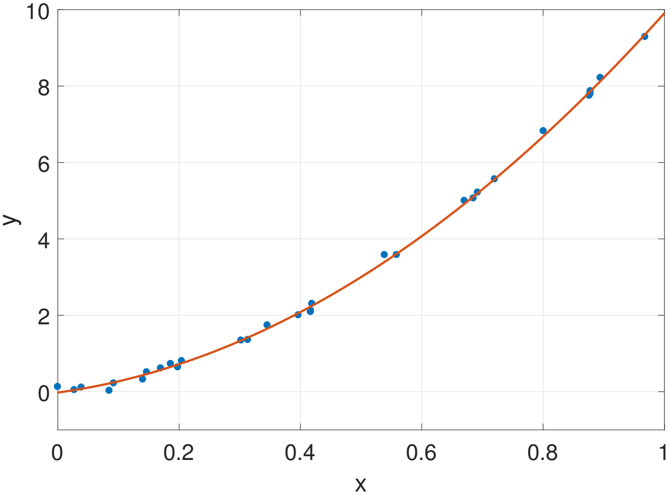
  Can only be used if all $x_i > 0$
- Polynomial transformations offer general purpose nonlinear fits
- We turn our variable into q new variables of the form:

  $$x_{i,j} \;\Rightarrow\; x_{i,j},\, x_{i,j}^2,\, x_{i,j}^3,\, \ldots,\, x_{i,j}^q$$

- The higher the q the more nonlinear the fit can become, but at risk of overfitting

Nonlinear Effects: Quadratic

New model: $\hat{y} = -0.02 + 2.16x + 7.77x^2$, $R^2 = 0.999$



Outline

Regression Example

Regression

Least Squares Fitting

Multiple Linear Regression

Maximum Likelihood Estimation

Model Selection

Connecting LS to MLE

- To show this, let our targets $Y_1, \ldots, Y_n$ be RVs
- Write the linear regression model as

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i$$

where ε_i is a random, unobserved "error"

- Now assume that $\varepsilon_i \sim N(0, \sigma^2)$
- We also assume ε_i is independent of $x_{1,j}$
- Therefore:

$$Y_i \,|\, x_{i,1}, \ldots, x_{i,p} \sim N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}, \sigma^2\right)$$

Connecting LS to MLE

(optional)

- Each Y_i is independent
- Given target data $\mathbf{y}$ the likelihood function can be written

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp \left(- \frac{\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2}{2\sigma^2} \right)$$

- Noting $e^{-a} e^{-b} = e^{-a-b}$ this simplifes to

$$\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left(- \frac{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2}{2\sigma^2} \right)$$

where we can see term in the numerator in the $\exp(\cdot)$ is the residual sum-of-squares.

Connecting LS to MLE

- Taking the negative-logarithm of this yields

$$L(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{\mathrm{RSS}(\boldsymbol{\beta})}{2\sigma^2}$$

- As the value of σ^2 scales the RSS term, it is easy to see that the values of $\boldsymbol{\beta}$ that minimise the negative log-likelihood are the least-squares estimates $\hat{\boldsymbol{\beta}}$
- LS estimates are same as the maximum likelihood estimates assuming the random "errors" ε_i are normally distributed
- Our residuals

$$e_i = y_i - \hat{y}_i$$

can be viewed as our estimates of the errors ε_i.

Connecting LS to MLE

- How to estimate the error variance σ^2?
- The maximum likelihood estimate is:

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}})}{n}$$

but this tends to underestimate the actual variance.

- A better estimate is the unbiased estimate

$$\hat{\sigma}^2_{\mathrm{u}} = \frac{\mathrm{RSS}(\hat{\boldsymbol{\beta}})}{n - p - 1}$$

where p is the number of predictors used to fit the model.

Simple LR Theory (Ross 9.3)

(optional)

- $\hat{\beta}_0$ and $\hat{\beta}_1$ found as for least squares
- The estimates are unbiased:

$$\mathbb{E}\left[\hat{\beta}_0\right] = \beta_0 \qquad \mathbb{E}\left[\hat{\beta}_1\right] = \beta_1$$

- Their covariance is

$$\operatorname{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{\sigma^2}{n\left(\overline{X^2} - \overline{X}^2\right)} \begin{pmatrix} 1 & -\overline{X} \\ -\overline{X} & \overline{X^2} \end{pmatrix}$$

- Moreover, $\frac{1}{\sigma^2}\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_1) \sim \chi^2_{n-2}$
- From this one can show $\hat{\sigma}^2_u$ is unbiased.
- Moreover one can get a confidence interval for σ^2.

Fitting Simple LR Theory

- From Ross 9.4 we get that confidence intervals and hypothesis testing can be done about β_0 using

$$\frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{\overline{X^2}}{\overline{X^2}-\overline{X}^2}}}(\hat{\beta}_0-\beta_0) \sim \mathrm{Student\text{–}t}(n-2)$$

- Similarly, confidence intervals and hypothesis testing can be done about β_1 using

$$\frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{1}{\overline{X^2}-\overline{X}^2}}}(\hat{\beta}_1-\beta_1) \sim \mathrm{Student\text{–}t}(n-2)$$

- the square-root terms are the corresponding standard errors (i.e., error in estimating β_0,β_1),
  - as reported by R diagnostics

Making Predictions

- Given estimates $\hat{\beta}$ can make predictions about new data
- To estimate value of target for some **new** predictor values $x'_1, x'_2, \ldots, x'_p$

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x'_j$$

- Using normal model of residuals, we can also get probability distribution over future data:

$$Y \sim N\left(\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x'_j,\ \sigma^2\right)$$

- By changing predictors we can see how target changes
  - Example: seeing how weight and age effect blood pressure
- Careful using predictions outside of sensible predictors values!

Outline

Regression Example

Regression

Least Squares Fitting

Multiple Linear Regression

Maximum Likelihood Estimation

Model Selection

Underfitting/Overfitting

- We often have many measured predictors
  - In our blood pressure example, we have weight, body surface area, age, pulse rate and a measure of stress
- Should we use them all, and if not, why not?
- The R^2 always improves as we include more predictors
  $\Longrightarrow$ so model always fits the data we have better
- But prediction on new, unseen data might be worse
- We call this generalisation

Under/Overfitting Example

- Example: we observe x and y data and want to build a prediction model for y using x
  - Data looks nonlinear so we use polynomial regression
  - We take $x, x^2, x^3, \ldots, x^{20} \Rightarrow$ very flexible model
  - How many terms to include?
- For example, do we use

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

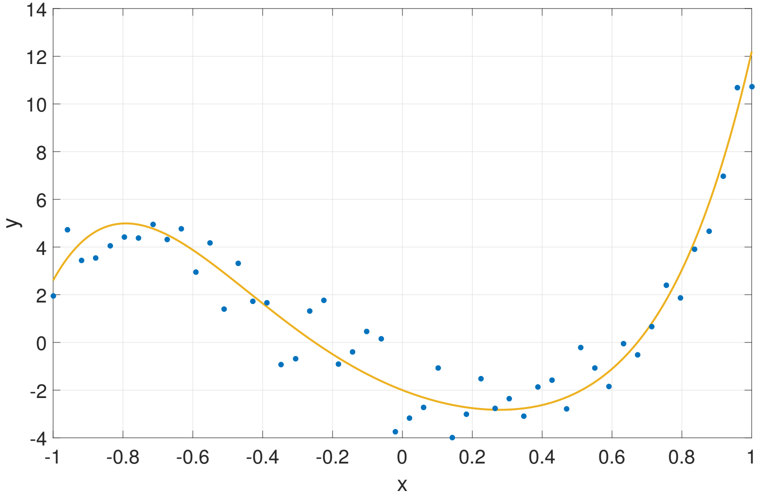
or

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \varepsilon$$

or another model with some other number of polynomial terms.

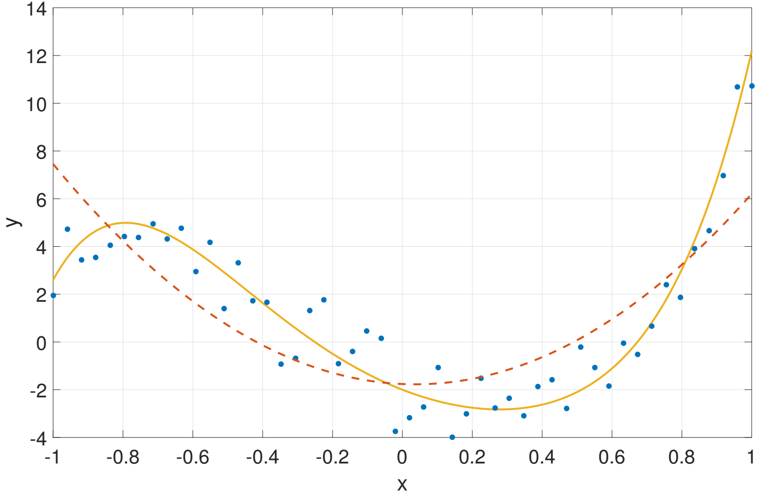
Under/Overfitting Example

Example dataset of 50 samples



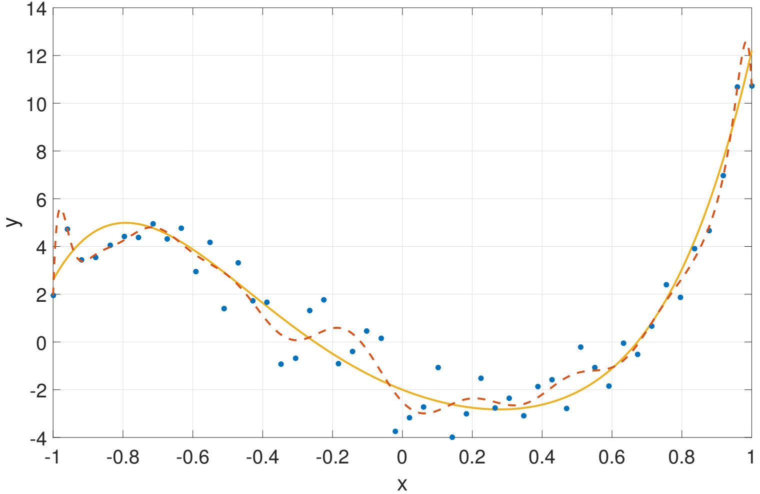
Underfitting Example

Use (x, x^2), too simple – underfitting



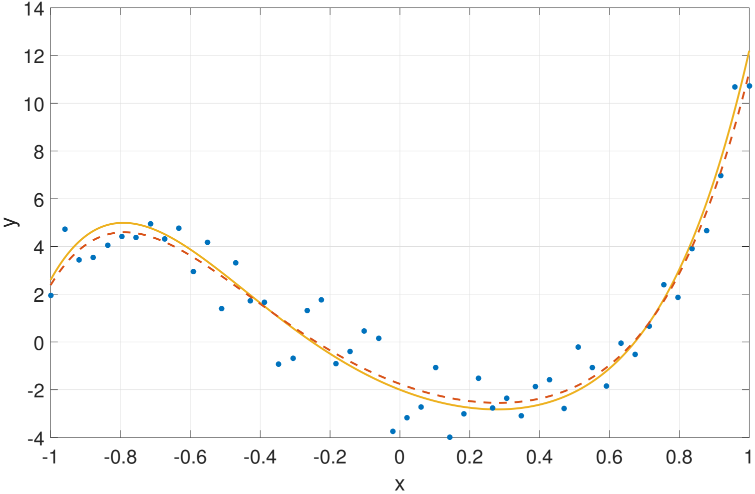
Overfitting Example

Use $(x, x^2, \ldots, x^{20})$, too complex – overfitting



Fitting Example

$(x, x^2, \ldots, x^6)$ seems “just right”. But how to find this model?



Underfitting/Overfitting

- Risks of including/excluding predictors
- Omitting important predictors:
  - Called underfitting
  - Leads to systematic error, bias in predicting the target
  - means **high bias**
- Including spurious predictors:
  - Called overfitting
  - Leads our model to “learn” noise and random variation
  - Poorer ability to predict to new, unseen data from our population
  - means **high variance**

Using Hypothesis Testing

- One approach is to use hypothesis testing
- We know that a predictor *j* is unimportant if $\beta_j = 0$
- So we can test the hypothesis:

$$\begin{aligned} H_0 \quad &: \quad \beta_j = 0 \\ &\text{vs} \\ H_A \quad &: \quad \beta_j \neq 0 \end{aligned}$$

  which, in this setting is a variant of the *t*-test (see Ross, 9.4)
- Strengths: easy to apply, easy to understand
- Weaknesses: difficult to directly compare two different models, will depend on what other predictors are included

Model Selection, Example

- A different approach is through model selection
- In the context of linear regression, we define a model by specifying which predictors are included in the linear regression
- For example, in our blood pressure example:
  - $\{\text{Weight}\}$
  - $\{\text{Weight}, \text{Age}\}$
  - $\{\text{Age}, \text{Stress}\}$
  - $\{\text{Age}, \text{Stress}, \text{Pulse}\}$

  are some of the possible models we could build
- Given a model, we can estimate the associated linear regression coefficients using least-squares/maximum likelihood
- The question then becomes how to choose a good model

Model Selection with MLE?

- We optimise maximum likelihood or RSS to choose the parameters:
    - Remember, this means we adjust the parameters of our distribution until we find the ones that maximise the probability of seeing the data **y** we have observed
- Can we use this to select a model as well as parameters?

Model Selection with MLE?

- We optimise maximum likelihood or RSS to choose the parameters:
  - Remember, this means we adjust the parameters of our distribution until we find the ones that maximise the probability of seeing the data **y** we have observed
- Can we use this to select a model as well as parameters?
- The maximum likelihood always increases and the RSS always decreases as we add more predictors to our model $\Longrightarrow$ cannot be used to select models, only to fit parameters

Model Selection
(optional)

- Let $\mathcal{M}$ denote a model (set of predictors to use)
- Let $L(\mathbf{y} \mid \hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\mathrm{ML}}, \mathcal{M})$ denote minimised negative log-likelihood for the model $\mathcal{M}$
- We can select a model by minimising an information criterion

$$L(\mathbf{y} \mid \hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\mathrm{ML}}, \mathcal{M}) + \alpha(n, k_{\mathcal{M}})$$

  where
  - $\alpha(\cdot)$ is a model complexity penalty;
  - $k_{\mathcal{M}}$ is the number of predictors in model $\mathcal{M}$;
  - n is the size of our data sample.
- This is a form of penalized likelihood estimation $\Longrightarrow$ a model is penalized by its complexity (ability to fit data)

Model Selection, cont.

(optional)

- How to measure complexity, i.e., choose $\alpha(\cdot)$?
- Akaike Information Criterion (AIC)

$$\alpha(n, k_{\mathcal{M}}) = k_{\mathcal{M}}$$

- Bayesian Information Criterion (BIC)

$$\alpha(n, k_{\mathcal{M}}) = \frac{k_{\mathcal{M}}}{2} \log n$$

- AIC penalty smaller than BIC; increased chance of overfitting
- BIC penalty bigger than AIC; increased chance of underfitting
- Differences in scores of ≥ 3 or more are considered significant
- both `AIC()` and `BIC()` supported in R

Finding a Good Model

- Most obvious approach is to try all possible combinations of predictors, and choose one that has smallest information criterion score
- Called the all subsets approach
- If we have p predictors then we have 2^p models to try
- For $p = 50$, $2^p \approx 1.2 \times 10^{15}$!
- So this method is computationally intractable for moderate p

Naive Forward Selection

- a simple, crude approach for smaller numbers of attributes
  1. for each attribute
     1.1 build a simple linear regression model for the attribute
     1.2 view p-value in R diagnostics of the model to see if we reject H_0 that the coefficient is equal to zero
  2. keep attributes with rejected H_0, and rebuild one model with just them

Naive Backward Selection

- a simple, crude approach for smaller numbers of attributes
  1. build one full model using all attributes
  2. view p-value for each attribute in R diagnostics to see if we reject H_0 that the coefficient is equal to zero
  3. keep attributes with rejected H_0, and rebuild one model with just them

Stepwise Model Selection
(optional)

- An alternative is to search through the model space
- Forward selection algorithm:
  1. Start with the empty model;
  2. Find the predictor that reduces info criterion by most
  3. If no predictor improves model, end.
  4. Add this predictor to the model
  5. Return to Step 2
- Backwards selection is a related algorithm
  - Start with the full model and remove predictors
- Is computationally tractable for large *p*, but may miss important predictors
- implemented in the `step()` routine in R with either AIC or BIC

End of Week 8