FIT2086 Lecture 6
Linear Regression

Daniel F. Schmidt

Faculty of Information Technology, Monash University

August 30, 2022

Outline

1. Linear Regression Models
   - Supervised Learning
   - Linear Regression Models
2. Model Selection for Linear Regression
   - Under and Overfitting
   - Model Selection Methods

Revision from last week

- Hypothesis testing; test null hypothesis vs alternative

$$H_0 \quad : \quad \text{null hypothesis}$$

$$\text{vs}$$

$$H_A \quad : \quad \text{alternative hypothesis}$$

- A test-statistic measures how different our observed sample is from the null hypothesis
- A p-value quantifies the evidence against the null hypothesis
- A p-value is the probability of seeing a sample that results in a test statistic as extreme, or more extreme, than the one we observed, just by chance if the null was true.

Today's Relevant Figure (N/A)

Adrien-Marie Legendre (1752 - 1833). Born in Paris, France, to a wealthy family. Studied mathematics and physics at the École Militaire, and lost his fortune during the French Revolution, but under Napoleon he was a well respected scientific figure. Published extensively on applied and pure mathematics. Developed the method of least-squares for curve fitting.

Outline

1. Linear Regression Models
   - Supervised Learning
   - Linear Regression Models
2. Model Selection for Linear Regression
   - Under and Overfitting
   - Model Selection Methods

Supervised Learning (1)

- Over the last three weeks we have looked at parameter inference
- In week 3 we examined point estimation using maximum likelihood
  - Selecting our "best guess" at a single value of the parameter
- In week 4 we examined interval estimation using confidence intervals
  - Give a range of plausible values for the unknown population parameter
- In week 5 we examined hypothesis testing
  - Quantify statistical evidence against a given hypothesis

Supervised Learning (2)

- Now we will start to see how these tools can be used to build more complex models
- Over the next four weeks we will look at supervised learning
- In particular, we we will look at linear regression
- But first, what is supervised learning?

Supervised Learning (3)

- Imagine we have measured $p + 1$ variables on n individuals (people, objects, things)
- We would like to predict one of the variables using the remaining p variables

- If the variable we are predicting is categorical, we are performing classification
  - Example: predicting if someone has diabetes from medical measurements.

- If the variable we are predicting is numerical, we are performing regression
  - Example: Predicting the quality of a wine from chemical and seasonal information.

Supervised Learning (4)

- The variable we are predicting is designated the "y" variable
  - We have $(y_1, \ldots, y_n)$
- This variable is often called the:
  - target;
  - response;
  - outcome.
- The other variables 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;
  - exposures.
- Usually we assume the targets are random variables and the predictors are known without error

Supervised Learning (4)

- Supervised learning: find a relationship between the targets y_i and associated predictors $x_{i,1}, \ldots, x_{i,p}$.
- That is, learn a function $f(\cdot)$ such that

$$y_i = f(x_{i,1}, \ldots, x_{i,p})$$

- Usually error in measuring y_i so that no $f(\cdot)$ fits perfectly
  $\Rightarrow$ we model y_i as realisation of RV Y_i
- So instead, find an $f(\cdot)$ that is "close" to $y_1, \ldots, y_n$

- It is "supervised" because we have examples to learn from
- Supervised learning model depends on form of $f(\cdot)$

Linear Regression

- Linear regression is a special type of supervised learning
- In this case, we take the function $f(\cdot)$ 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 and can even handle nonlinear relationships
  - It is computationally efficient to fit, even for very large p
- Enormous area of research and work
  $\Rightarrow$ we will get acquainted with the basics

Simple Linear Regression (1)

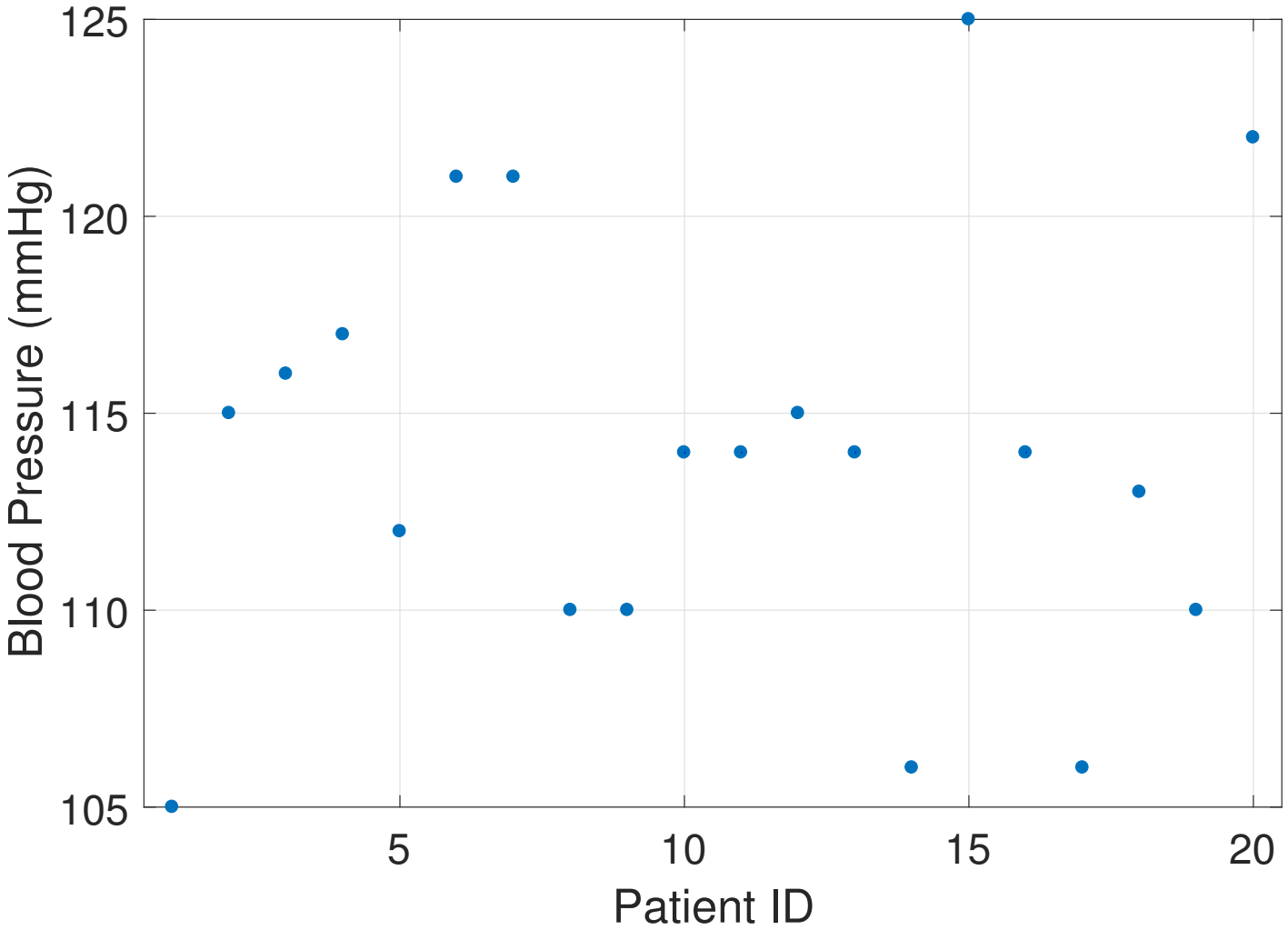
- Consider the following dataset (we examined in Studio 5):

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

Simple Linear Regression (2)

- Blood pressure plotted against patient ID



Simple Linear Regression (3)

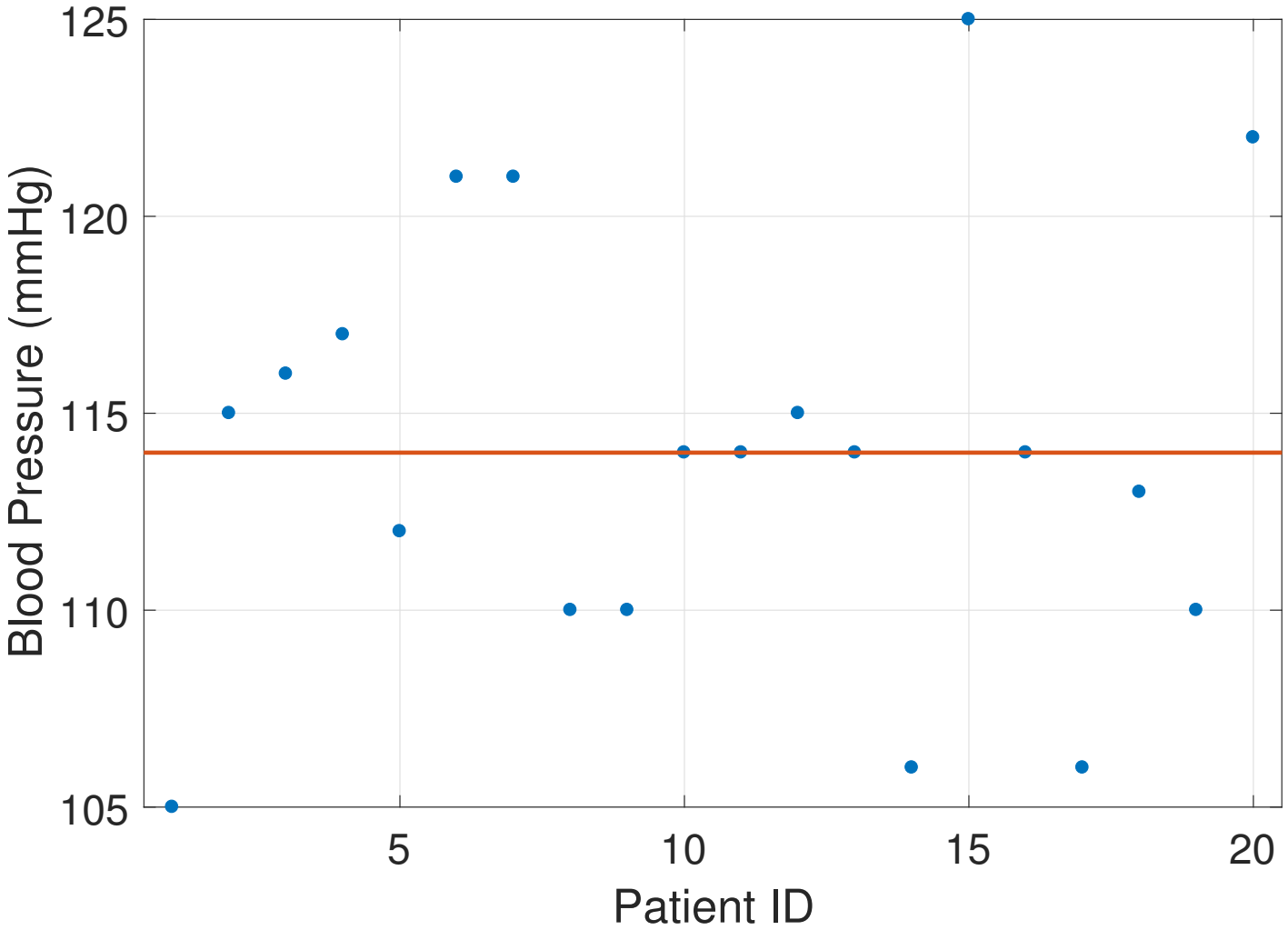
- Our blood pressure variable $\mathrm{BP}_1, \ldots, \mathrm{BP}_{20}$ is continuous
  $\Rightarrow$ we choose to model it using a normal distribution
- The maximum likelihood estimate of the mean μ is

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

  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 $\hat{\mu}$

Simple Linear Regression (4)

- Prediction of BP using the mean



Simple Linear Regression (5)

- 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

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

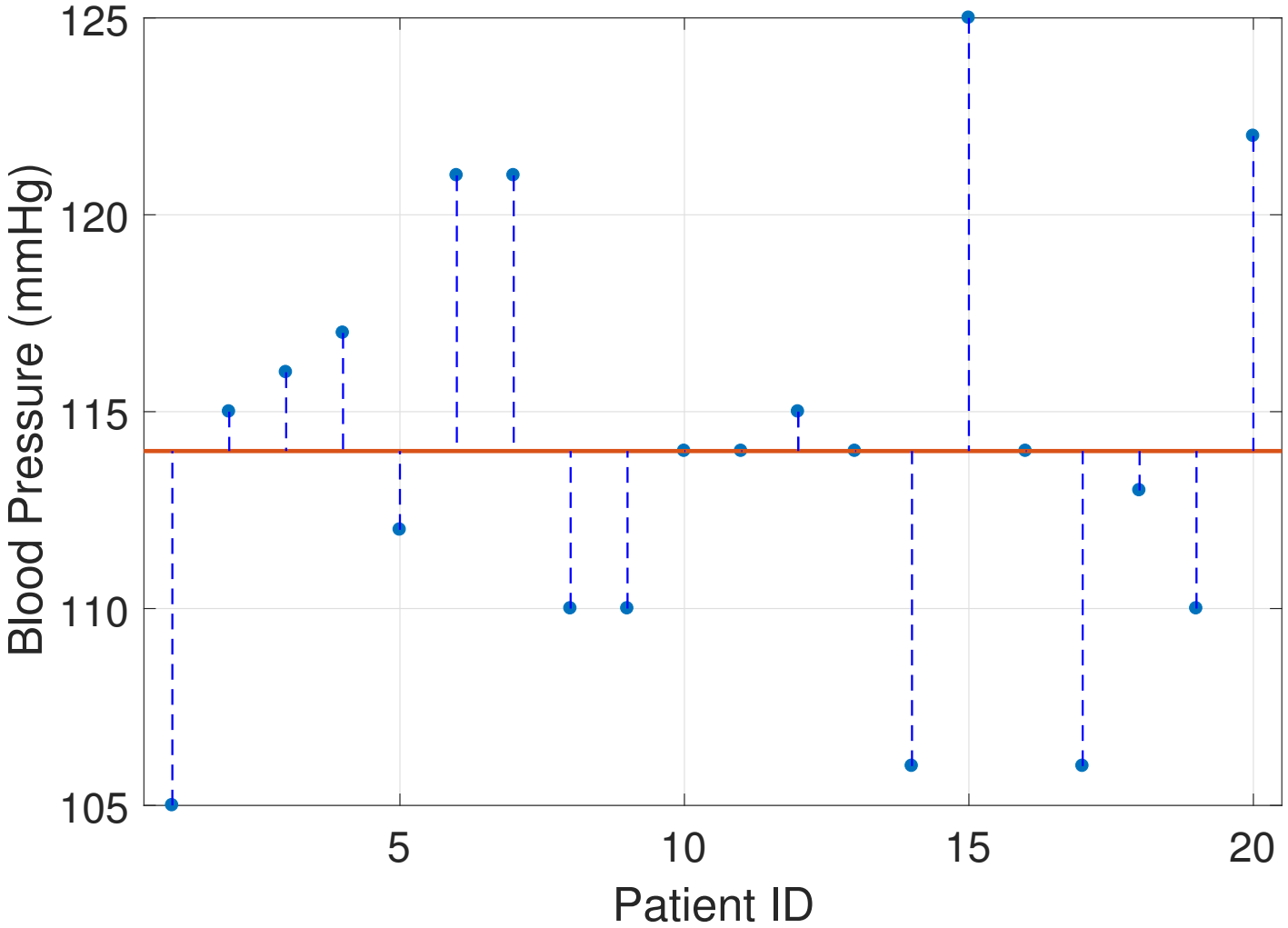
are the errors between our model predictions $\hat{y}_i$ and the observed data y_i

$\Rightarrow$ often called residual error, or just residuals

- A good fit would lead to overall small errors

Simple Linear Regression (6)

- Prediction of BP using the mean, showing errors/residuals



Simple Linear Regression (7)

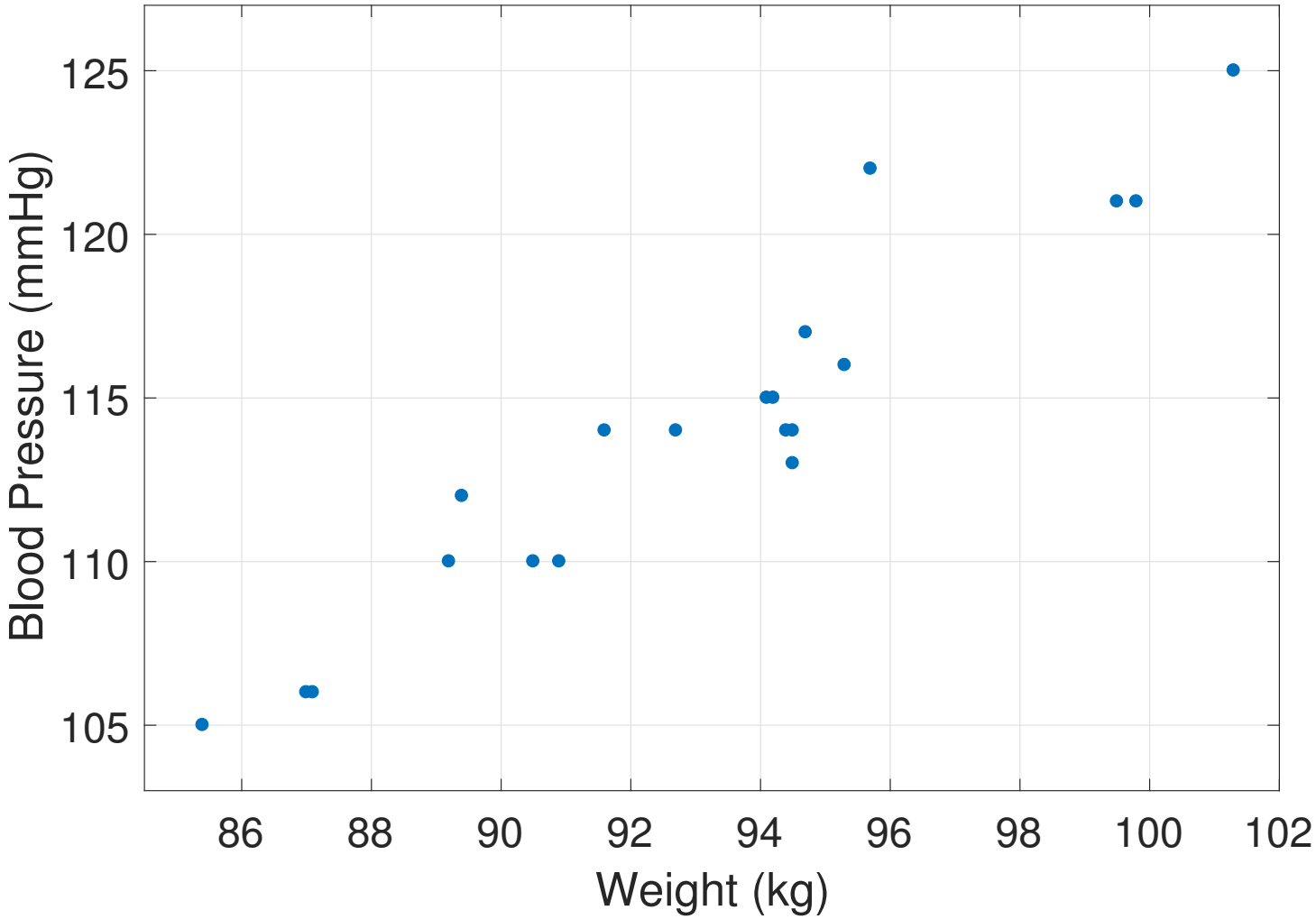
- We can summarise the total error of fit of our model by

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

  which is called the residual sum-of-squared errors.
- 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

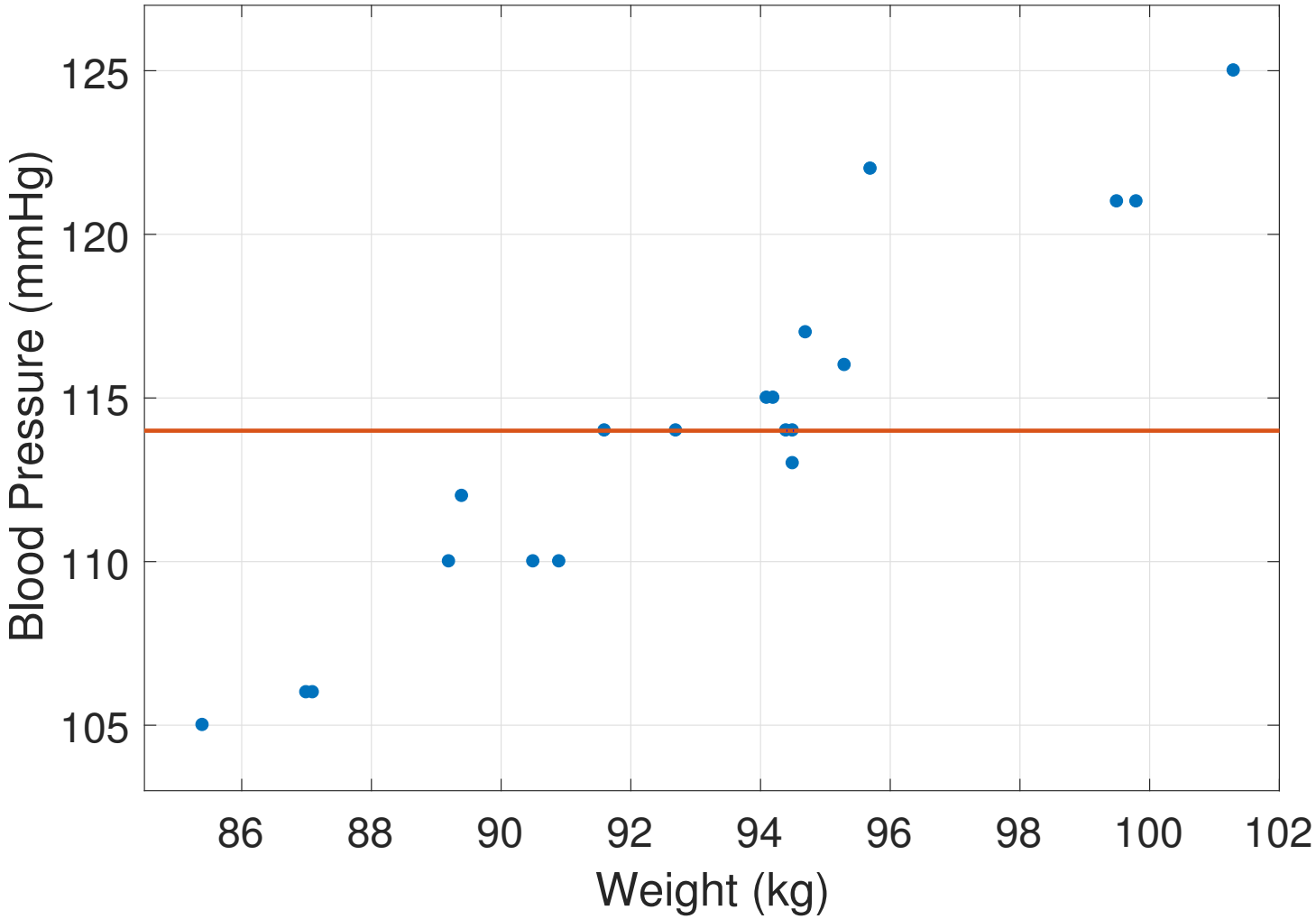
Simple Linear Regression (8)

- Blood pressure vs weight – BP appears to increase with weight



Simple Linear Regression (9)

- Our simple mean model is clearly not a good fit



Simple Linear Regression (10)

- Our simple mean model predicts blood pressure by

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

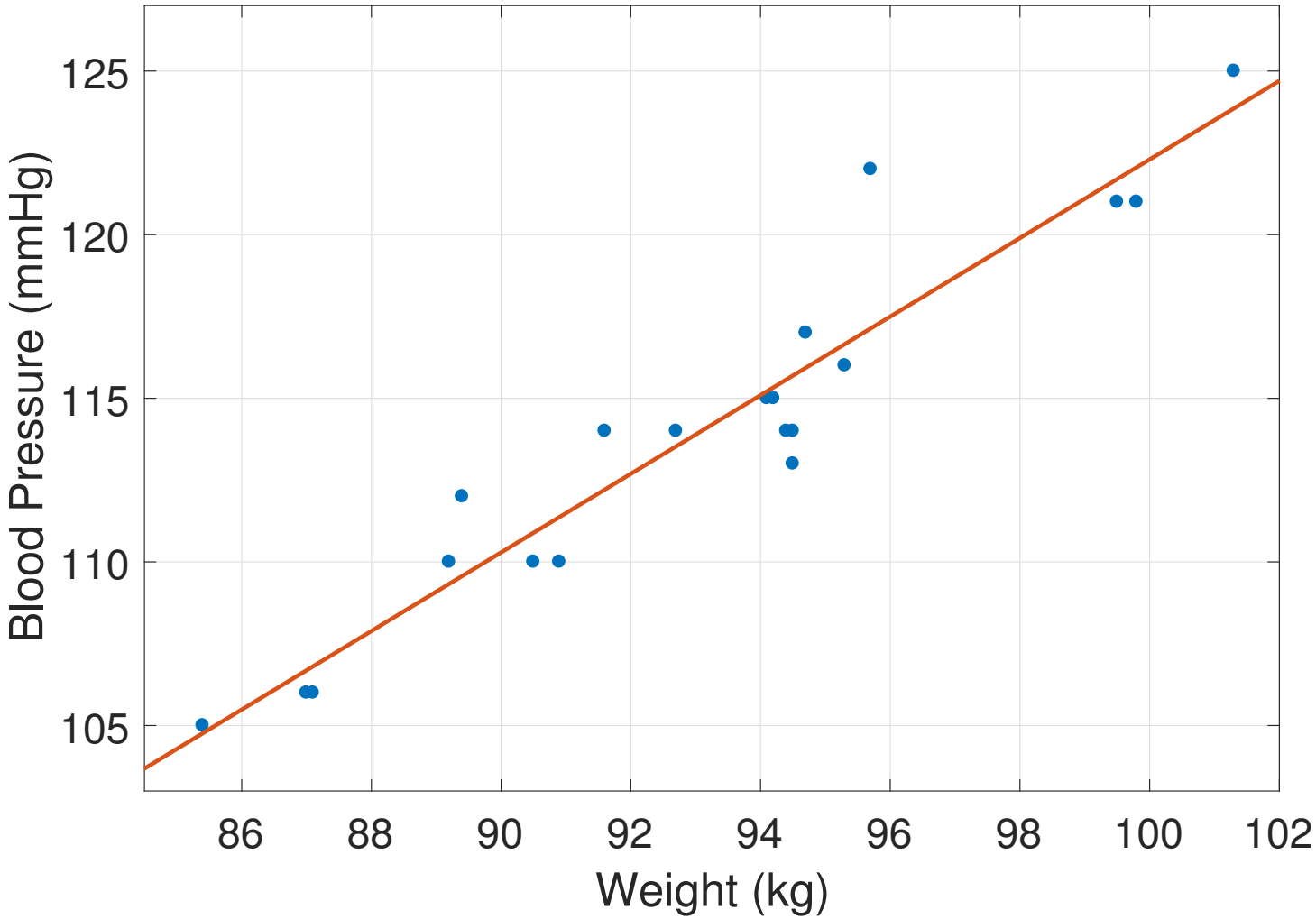
  irrespective of any other data on individual i
- Let $(\mathrm{Weight}_1, \ldots, \mathrm{Weight}_{20})$ be the weights of our 20 individuals
- We can let the mean vary as a linear function of weight, i.e.,

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

- 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 $\beta_1 = 0$

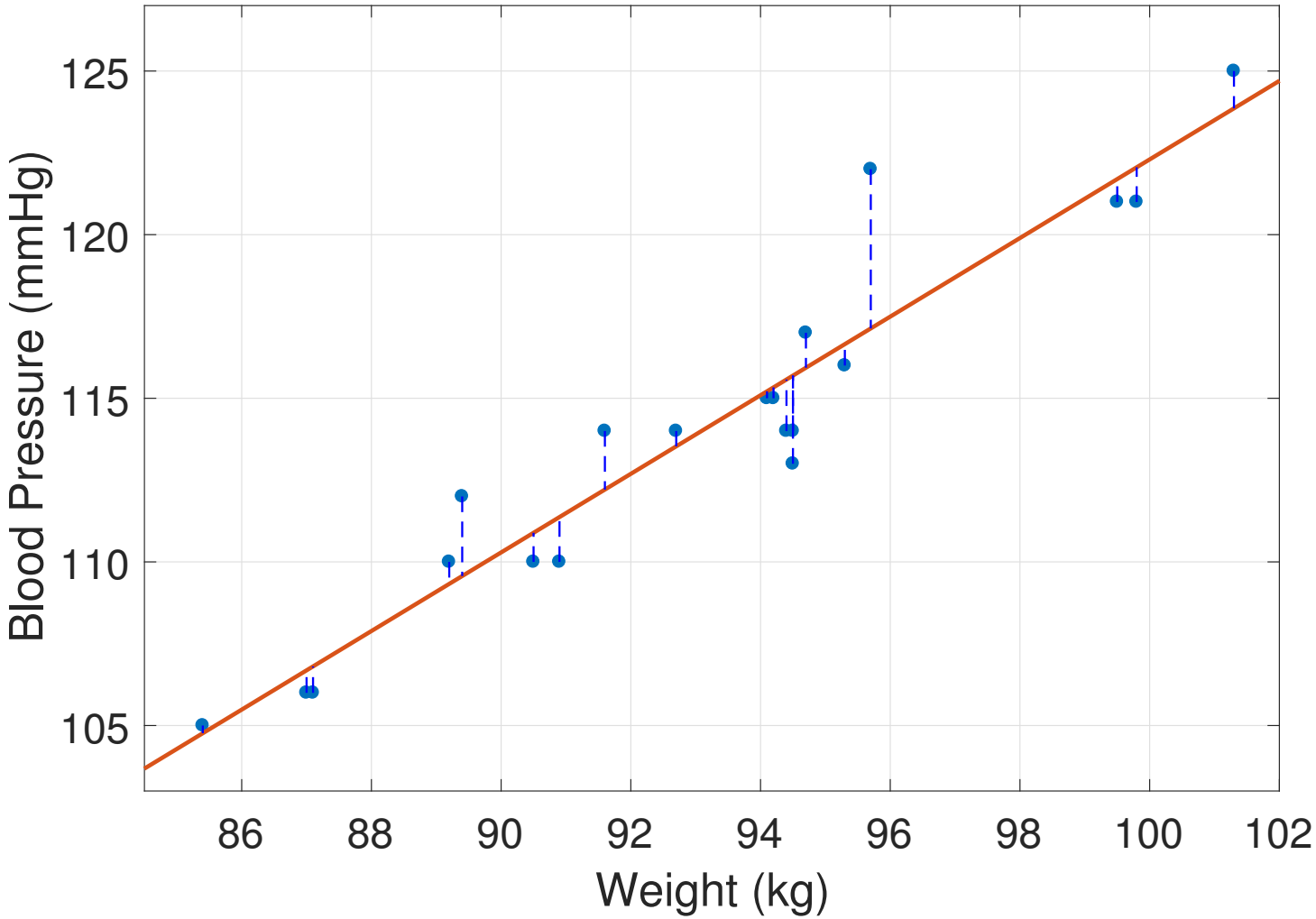
Simple Linear Regression (11)

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



Simple Linear Regression (12)

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



Simple Linear Regression (13) – Key Slide

- 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 by in one unit change of the predictor x_i

Simple Linear Regression (14)

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

$$\hat{y}_i = 2.2053 + 1.2009x_i$$

so

  - For every additional kilogram a person weighs, their average blood pressure increases by $1.2009mmHg$
  - For a person who weighs zero kilograms, the predicted average blood pressure is $2.2053mmHg$

- The predictions might not make sense outside of sensible ranges of the predictors!

Fitting Simple Linear Regressions (1)

- 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

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

- Smaller error = better fit

Fitting Simple Linear Regressions (2)

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

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

- 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

Fitting Simple Linear Regressions (3)

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

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

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

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

where we use the chain rule.

Fitting Simple Linear Regressions (4)

- 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

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

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.

Multiple Linear Regression (1) – Key Slide

- 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 (2) – Key Slide

- Fit a multiple linear regression using least-squares
  $\Rightarrow$ 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

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

- Efficient algorithms exist to find these estimates

R-squared (R^2) (1)

- 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
- We use the total sum-of-squares as the reference:

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

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

R-squared (R^2) (2) – Key Slide

- The R^2 value is then defined as

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

  which is also called the coefficient-of-determination
- 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
  $\Rightarrow$ predictors that greatly increase R^2 are potentially important

Example: Multiple regression and R^2 (1)

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

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

which had an RSS of $120 \Rightarrow R^2 = 1 - 120/560 \approx 0.78$

Example: Multiple regression and R^2 (2)

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

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

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

Break: Questions/Exercise

- Question 1: If our model is

$$\hat{y} = 2.2053 + 1.2009x$$

what is the predicted value if $x = 2$?

- Question 2: If we had two predictors and the least-squares estimates of their coefficients were $\hat{\beta}_1 = 2.5$ and $\hat{\beta}_2 = -4$ respectively, which one is more strongly associated with our target y?

- Question 3: If one of our predictors was categorical variable (say, level of education: high school, undergrad, postgrad) can we use it in a linear regression?

Handling Categorical Predictors (1)

- 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
- 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.

Handling Categorical Predictors (2) – Key Slide

- Example variable with four categories coded as 1, 2, 3 and 4

$$\begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 4 \\ 2 \\ 3 \\ 2 \\ 4 \end{pmatrix} \Longrightarrow \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

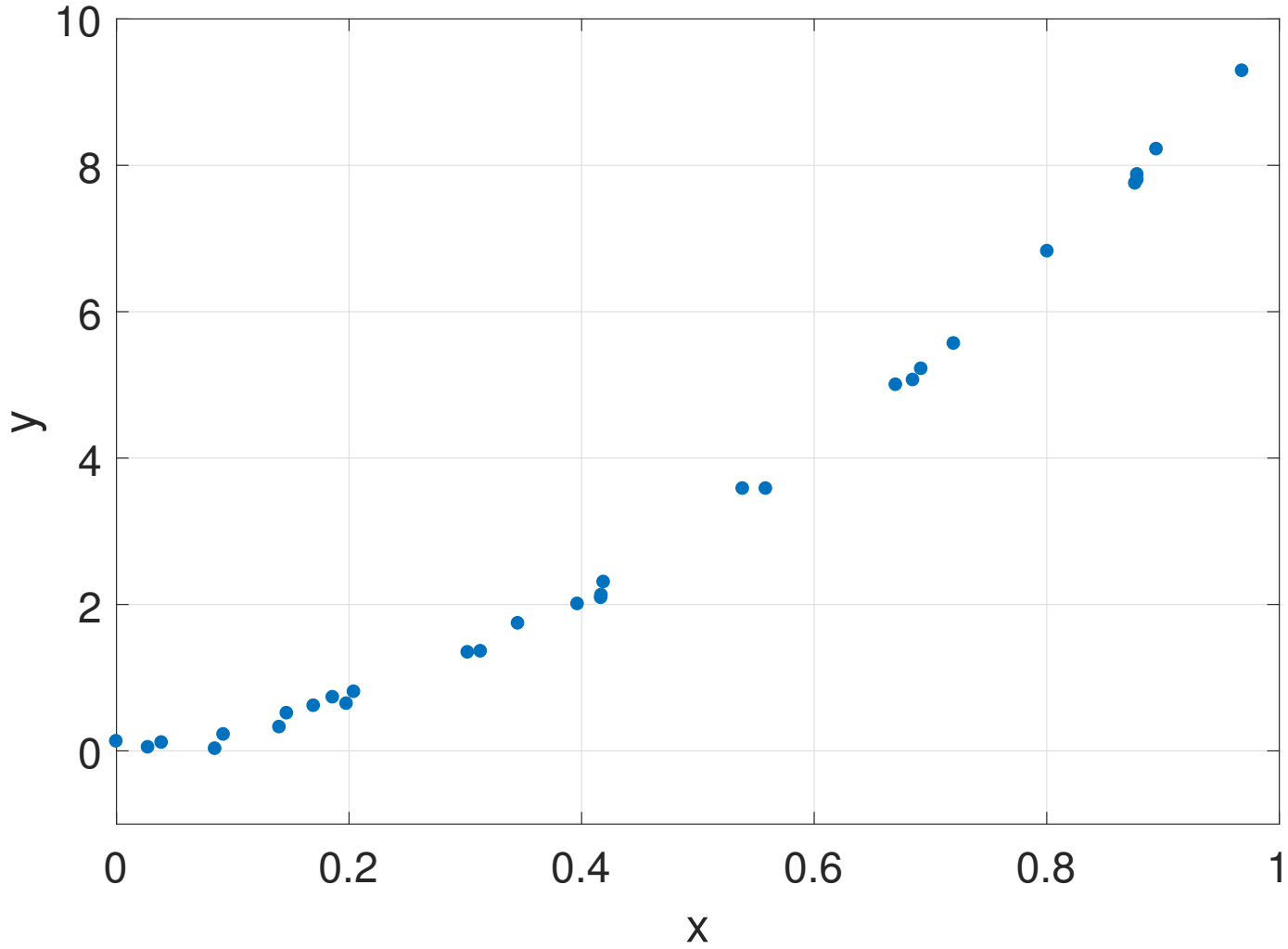
- We do not build indicators for first category
- Regression coefficients for other categories are increases in target relative to being in the first category
  - See Page 7-8 of the Lecture Notes for more detail

Nonlinear effects (1)

- Sometimes predictors are related to the target in a nonlinear fashion
- We can still use linear models by *transforming* the predictors
- If the transformed predictors are linearly related to the target, regression will work well
- We can often detect this by plotting the residuals against a variable – if they exhibit a nonlinear trend or curve it is sign that a transformation might be needed

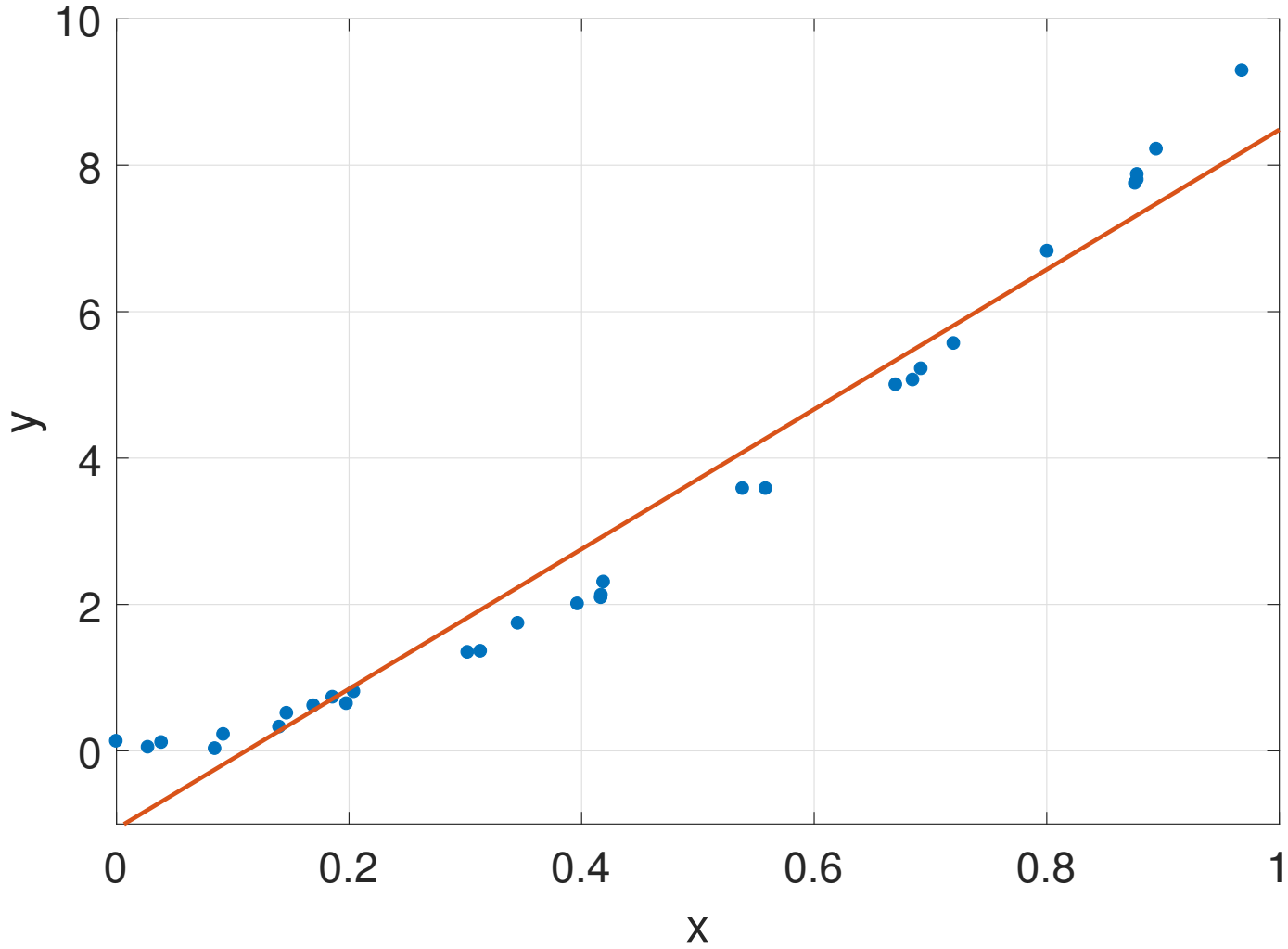
Nonlinear effects (2)

- Example dataset



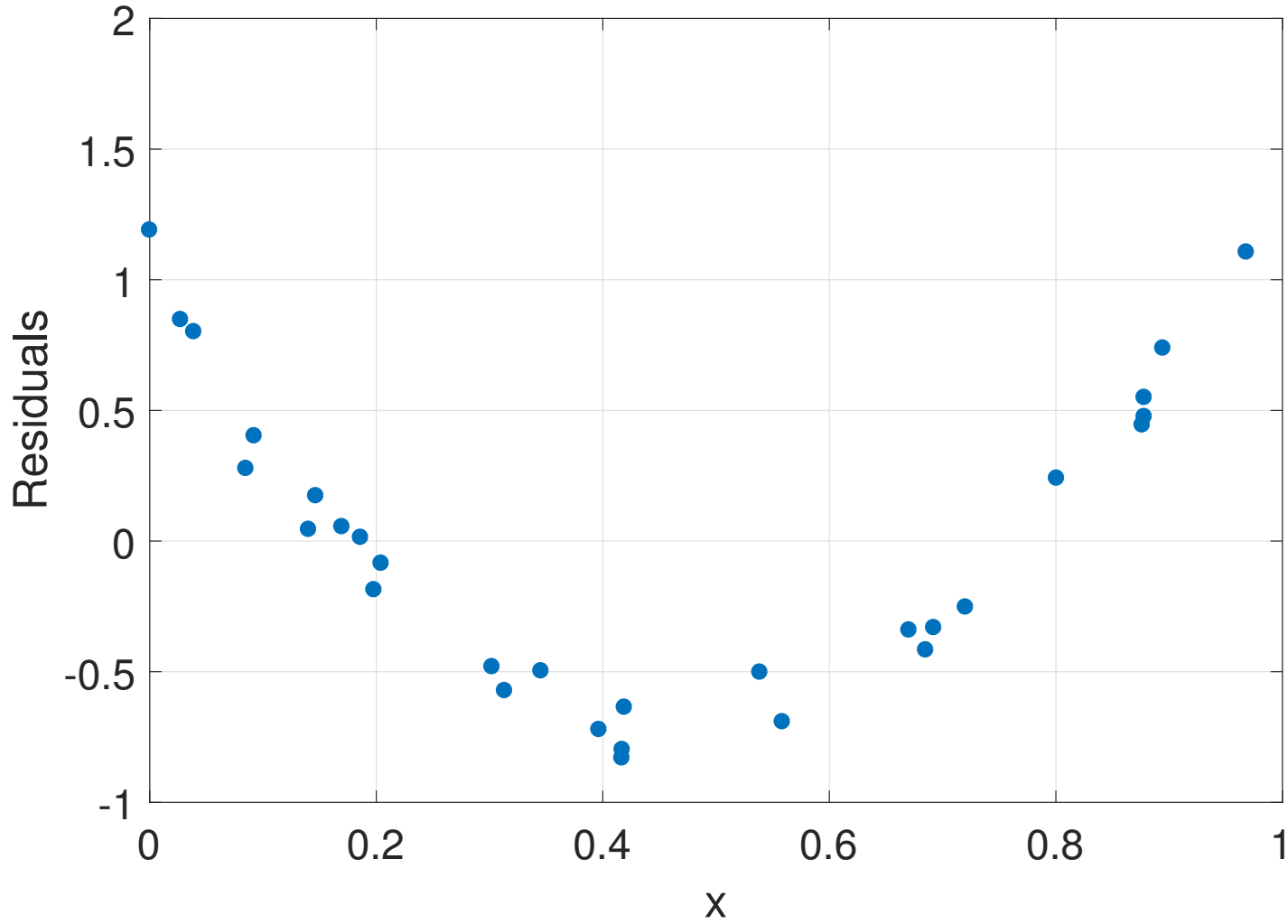
Nonlinear effects (3)

- Fitted model: $\hat{y} = -1.07 + 9.55x$; $R^2 = 0.95$



Nonlinear effects (4)

- Example data: residuals exhibit clear nonlinear trend



Nonlinear effects (5)

- There are several common transformations
- A logarithmic transformation can be used if the predictor seems to be more variable for larger values of the predictor

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

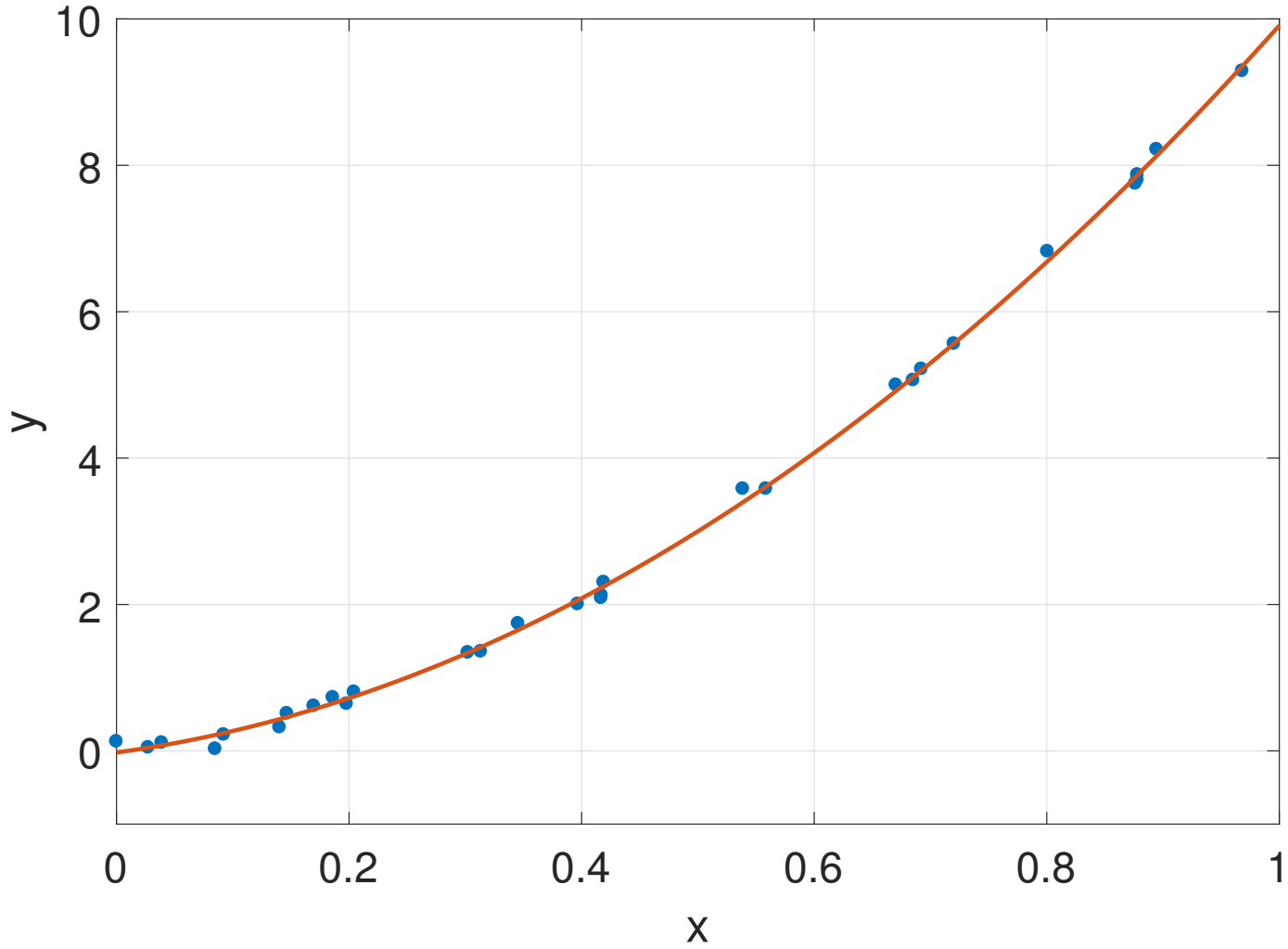
  Can only be used if all $x_i > 0$
- Polynomial transformations offer general purpose nonlinear fits
- We turn our variable into q new variables of the form:

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

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

Nonlinear effects (6)

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



Connecting LS to ML (1)

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

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

where ε_i is a random, unobserved "error"

- Now assume that $\varepsilon_i \sim N(0, \sigma^2)$
- This is equivalent to saying that

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

so each Y_i is normally distributed with variance σ^2 and a mean that depends on the values of the associated predictors

Connecting LS to ML (2)

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

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

- 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 ML (2) – Key Slide

- Taking the negative-logarithm of this yields

$$L(\mathbf{y} \mid \beta_0, \boldsymbol{\beta}, \sigma^2) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{\text{RSS}(\beta_0, \boldsymbol{\beta})}{2\sigma^2}$$

- As the value of σ^2 scales the RSS term, it is easy to see that the values of β_0 and $\boldsymbol{\beta}$ that minimise the negative log-likelihood are the least-squares estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$
- 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 ML (3)

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

$$\hat{\sigma}^2_{\text{ML}} = \frac{\text{RSS}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})}{n}$$

  but this tends to underestimate the actual variance.
- A better estimate is the unbiased estimate

$$\hat{\sigma}^2_{\text{u}} = \frac{\text{RSS}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})}{n - p - 1}$$

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

Making predictions with a linear model

- Given estimates $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$ 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:

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

- 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

1. Linear Regression Models
   - Supervised Learning
   - Linear Regression Models

2. Model Selection for Linear Regression
   - Under and Overfitting
   - Model Selection Methods

Underfitting/Overfitting (1)

- 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
  $\Rightarrow$ so model always fits the data we have better
- But prediction on new, unseen data might be worse
- We call this generalisation

Underfitting/Overfitting (2) – Key Slide

- Risks of including/excluding predictors
- Omitting important predictors
  - Called underfitting
  - Leads to systematic error, bias in predicting the target
- 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

Underfitting/Overfitting Example (1)

- 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

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

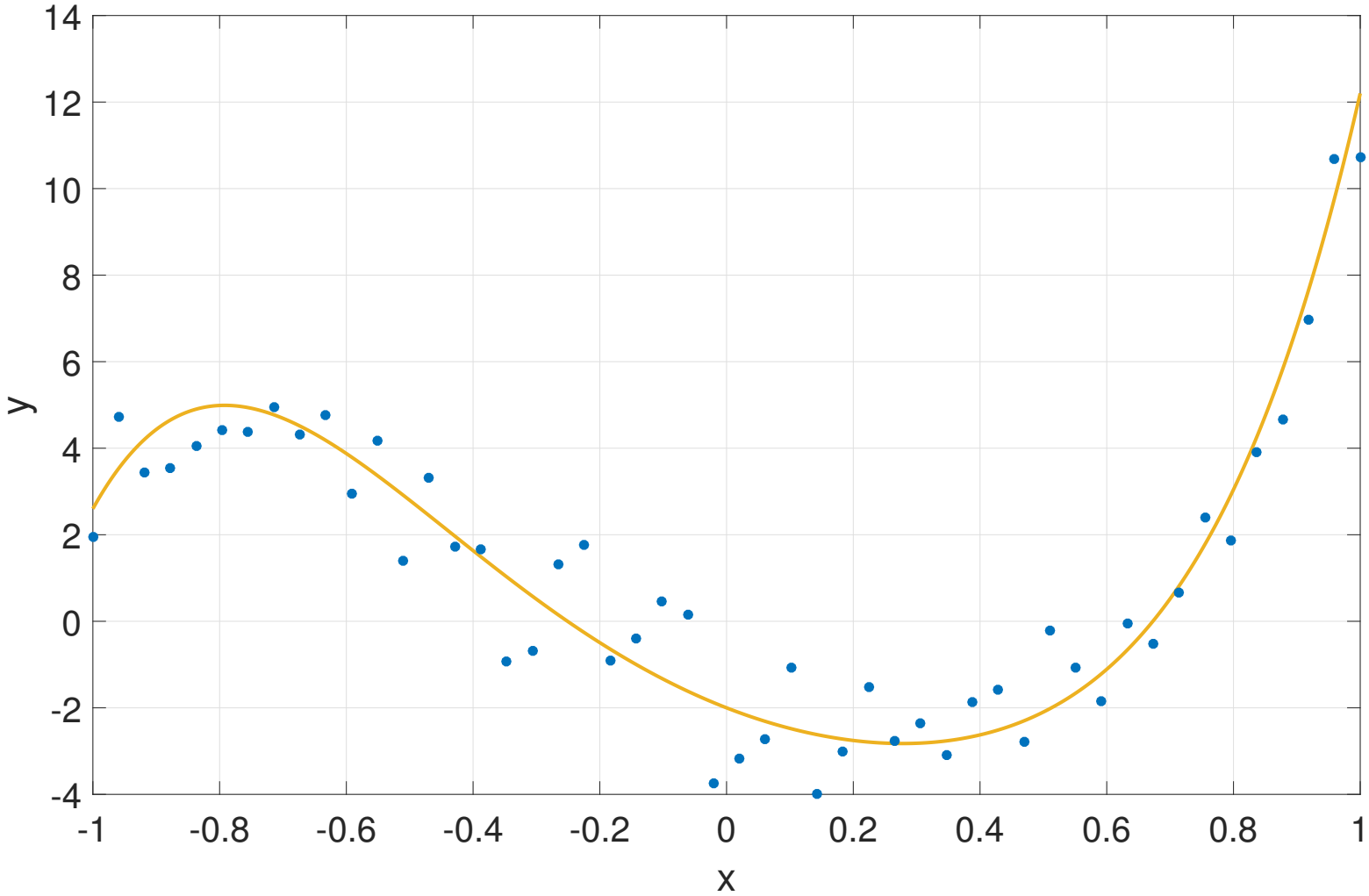
or

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

or another model with some other number of polynomial terms.

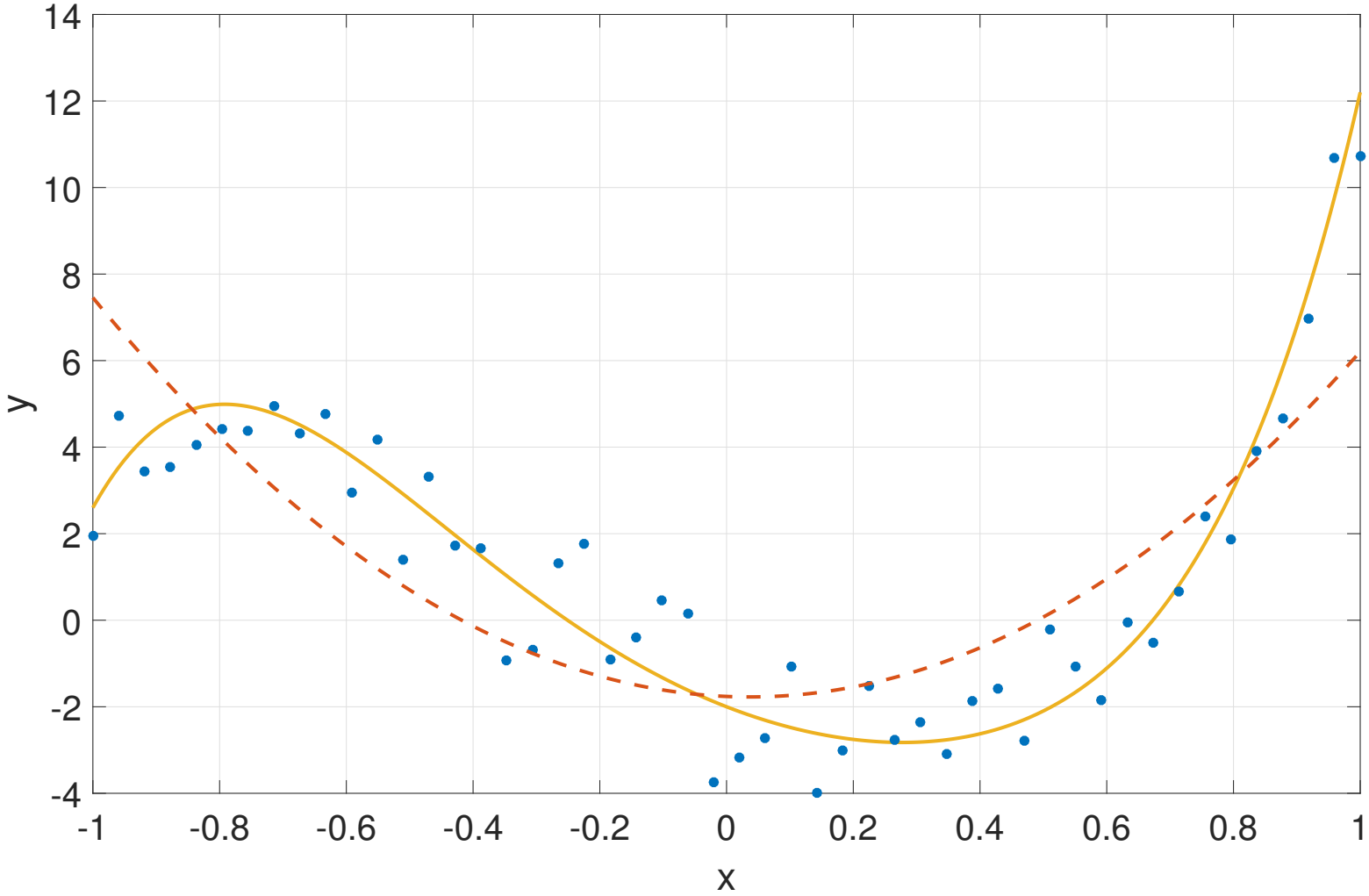
Underfitting/Overfitting Example (2)

- Example dataset of 50 samples



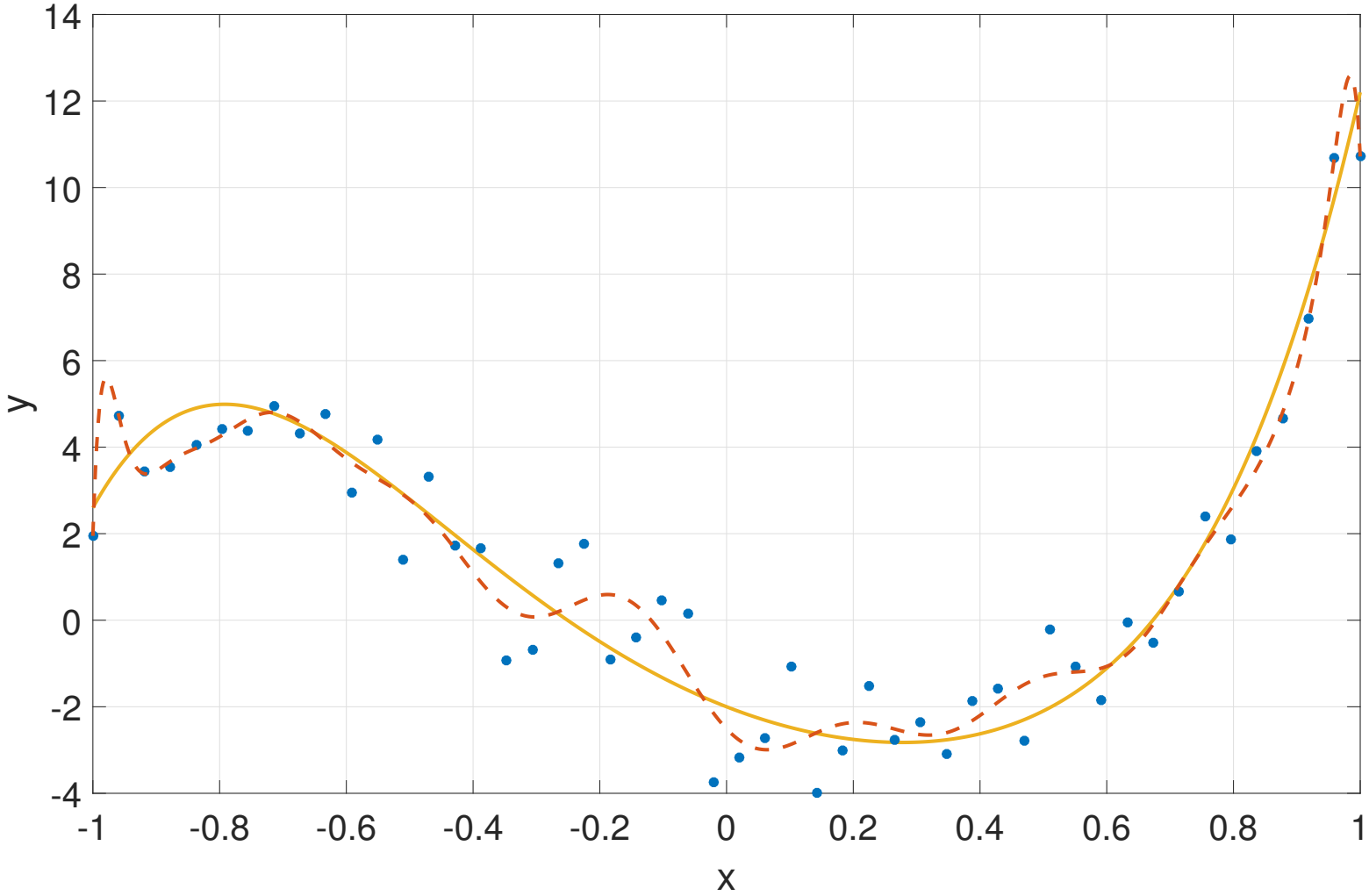
Underfitting/Overfitting Example (3)

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



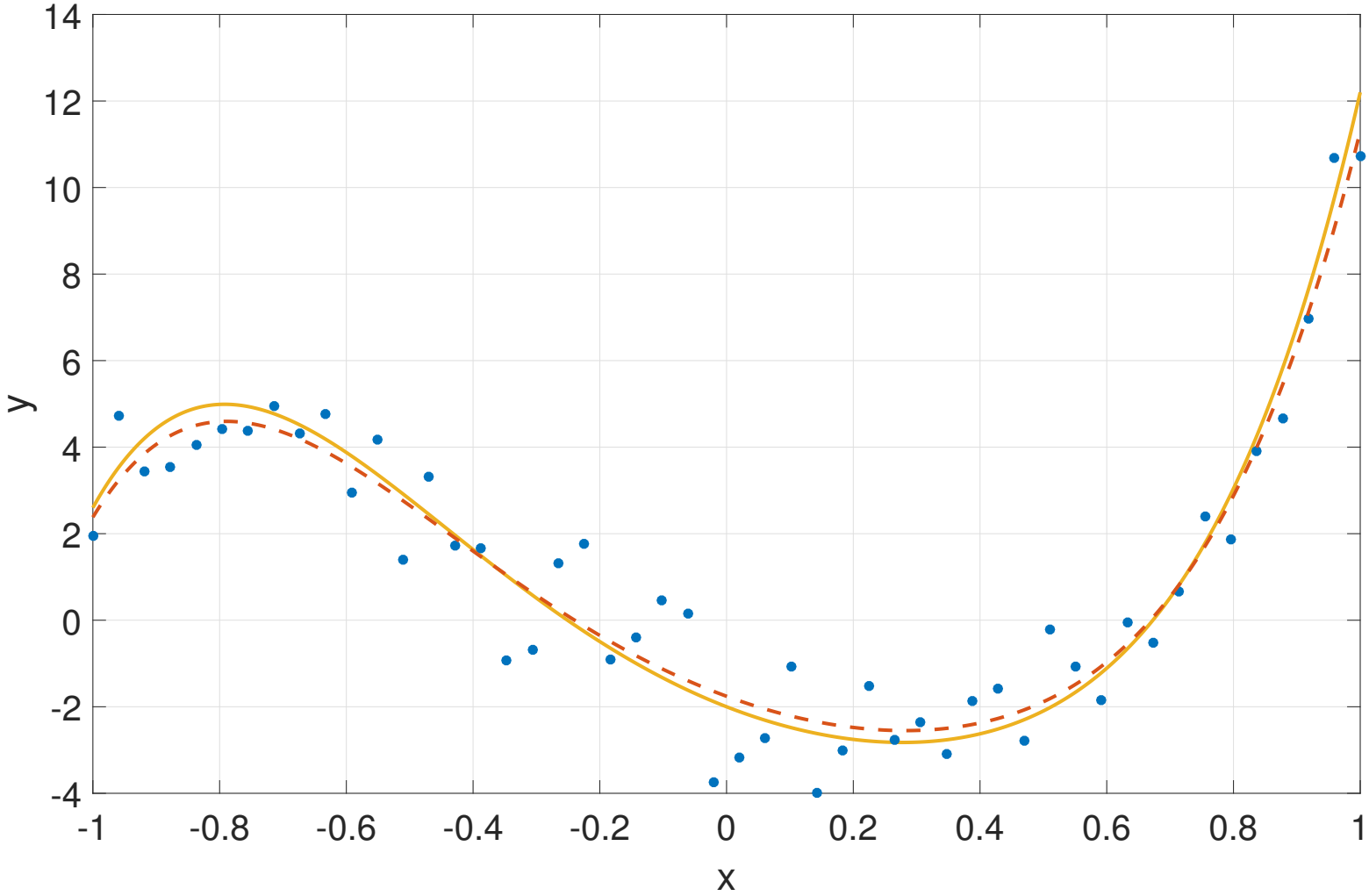
Underfitting/Overfitting Example (4)

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



Underfitting/Overfitting Example (5)

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



Measuring Strength of Importance

- How could we say a predictor is more important than another?
- Could we use the magnitude of the estimated coefficient $|\hat{\beta}_j|$?
  - This depends on the unit of measurement of the predictor
- Instead we could use the t-score:

$$t_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$$

where $\text{se}(\hat{\beta}_j)$ is the standard error of the coefficient
$\implies$ The larger the $|t_j|$, the more important the variable

- Standardises by how variable the estimate is (and removes scale)

Using Hypothesis Testing – Key Slide

- To decide if a variable is associated, we could use hypothesis testing
- We know that a predictor j is unimportant if $\beta_j = 0$
- So we can test the hypothesis:

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

  which, in this setting is a variant of the t-test (see Ross, Chapter 9 and Studio 6) based on the t-score t_j on previous slide
- Strengths: easy to apply, easy to understand
- Weaknesses: difficult to directly compare two different models

Model Selection (1)

- 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 (2)

- We use maximum likelihood 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 $\mathbf{y}$ we have observed
- Can we use this to select a model as well as parameters?
- Assume normal distribution for our regression errors
- The minimised negative log-likelihood (i.e., the negative log-likelihood evaluated at the maximum likelihood estimates $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2_{\mathrm{ML}}$) behaves similar to the R^2 as we add more predictors
- It always decreases as we add more predictors to our model
  $\Rightarrow$ cannot be used to select models, only parameters

Model Selection (3) – Key Slide

- Let $\mathcal{M}$ denote a model (set of predictors to use)
- Let $L(\mathbf{y} \mid \hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\mathrm{ML}}, \mathcal{M})$ denote minimised negative log-likelihood for the model $\mathcal{M}$
- We can select a model by minimising an information criterion

$$L(\mathbf{y} \mid \hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{\mathrm{ML}}, \mathcal{M}) + \alpha(n, k_{\mathcal{M}})$$

  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
  $\Rightarrow$ a model is penalized by its complexity (ability to fit data)

Model Selection (4)

- 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 (for $n > 7$); so relative to each other
  - AIC more likely to overfit (include spurious predictors), but less likely to underfit (exclude important predictors)
  - BIC more likely to underfit, but less likely to overfit
- Differences in scores of ≥ 3 or more are considered "significant"

Finding a Good Model (1)

- 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

Finding a Good Model (2)

- 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 related algorithm
  - Start with the full model and remove predictors
- Is computationally tractable for large p, but may miss important predictors

Reading/Terms to Revise

- Reading for this week: Chapter 9 of Ross.
- Terms you should know:
  - Target, predictor, explanatory variable;
  - Intercept, coefficient;
  - R^2 value;
  - Categorical predictors;
  - Polynomial regression;
  - Model, model selection;
  - Overfitting, underfitting
  - Information Criteria;
- This week we looked at supervised learning for continuous targets; next week we will examine supervised learning for categorical targets (classification).