ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 9: Linear Regression and Prediction

William Marsh

Outline

- Aim: Understand the use of linear regression models

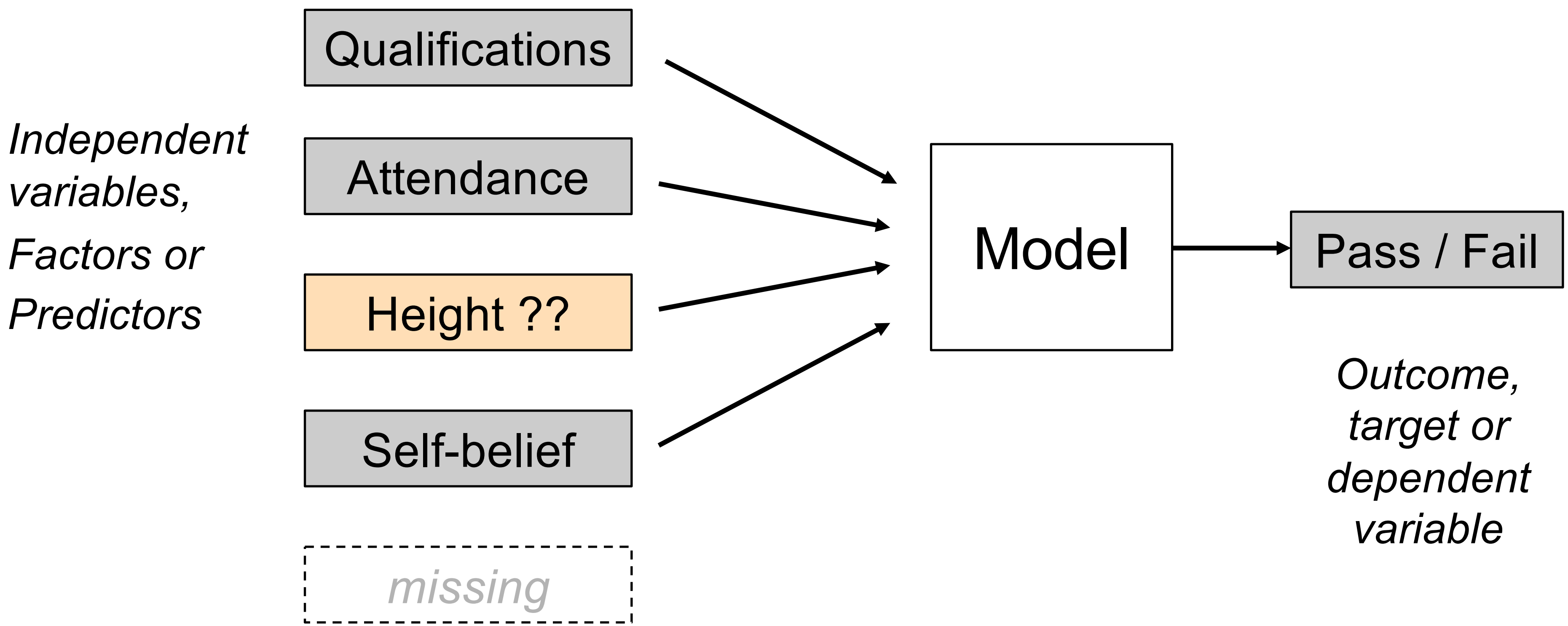
For code: see separate notebook on ‘regression’

- Statistical modelling
- Linear regression
- Fitting the regression line
  - Best fit using least squares
  - Goodness of fit
- How many Predictors?
- Including categorical variables

Statistical Modelling Principles

Statistical Modelling

- Model: one (or some) variables determined from other
- Example: why do some students fail?
    - Statistics: what factor explain failure
    - ML: Can we predict failure?



Statistical Modelling

- Model: one (or some) variables determined from other
- Example: why do some students fail?
  - Statistics: what factor explain failure(in a data set)
  - ML: Can we predict failure (given a data set)?

Statistics

- Aim is explanation
  - Which variables?
  - Contribution of variables
- Goodness of fit
- Population

(Supervised) Machine Learning

- Aim is prediction
  - Which variables?
  - Which algorithm?
- Prediction accuracy
- Individual

Linear Regression

Continuous Variables

Linear Regression

- Simplest form of statistical model
- Very widely used

- Many extensions
  - Logistic regression (dependent variable is binary)
  - Multi-level regression
  - Poisson regression
  - Non-linear regression
  - Generalised linear models
  - …

Equation of a Line (1 Independent Variable)

- Two parameters
  - Intercept: Y when X = 0
  - Slope: increase in Y when X increases by 1

$$Y = 2.5 * X + 3$$

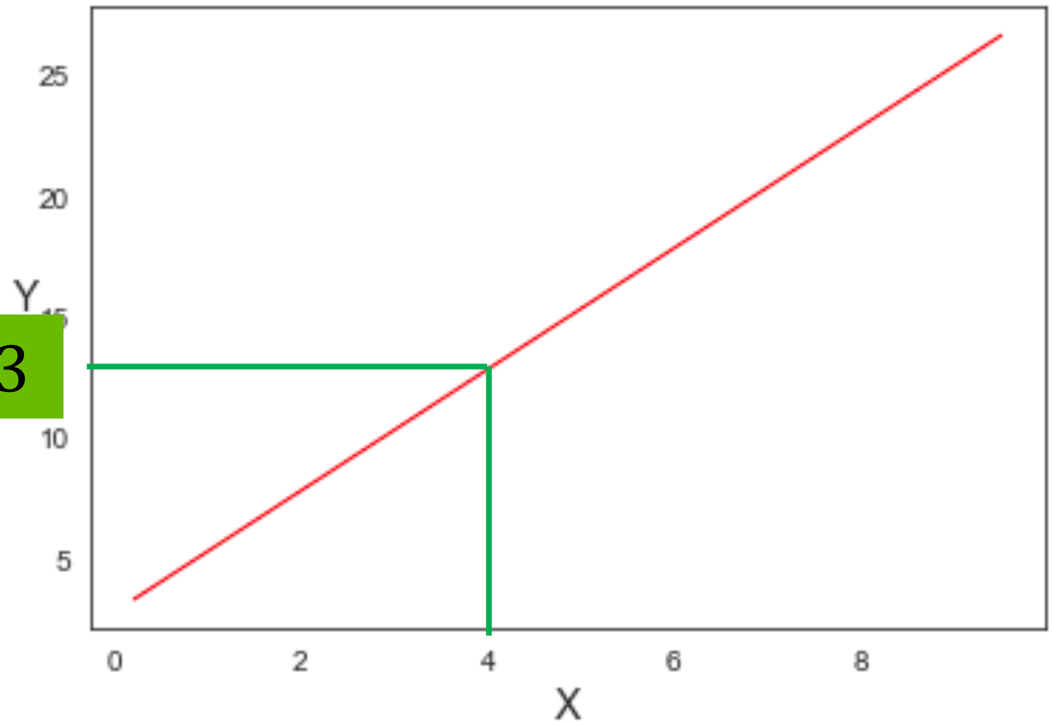
Dependent

Slope

Independent

Intercept

$$Y = (2.5 * 4) + 3 = 13$$

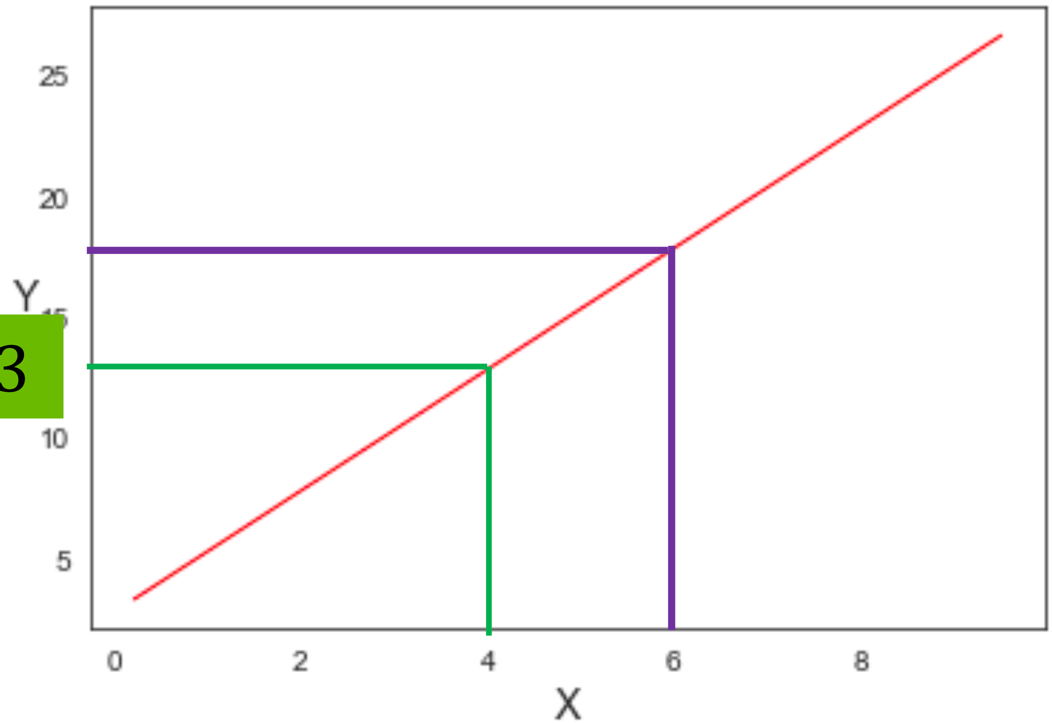


Equation of a Line

- Intercept = 3 (*Y when X = 0*)
- Slope = 2.5 (*increase in Y when X increases by 1*)

$Y = ?$ when X = 6

$Y = (2.5 * 4) + 3 = 13$



Linear Regression Assumptions

- Can have multiple independent variables (predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- Each independent variable X_i
  – Adds or subtracts to Y in independently of other X_j
  – Has it's own 'coefficient' β_i
  – Linear: the same change in X_i gives same change in Y

- Cannot be true if X_i and X_j are correlated

Explaining using Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- Each β shows the importance of its predictor
- β can be +ve or –ve
- If β very 'small' then predictor not important
  - Size is relative to other predictors
  - Standardise range of Xs

- *What about missing predictors?*

Quiz 1

Coursework Review

From Web Page - MCQ

- Multi-choice questionnaires (MCQs) - 40% in total
- MCQ1 (10%)
  - Examines the concepts delivered in the sprint week.
  - Set on Thursday in week 1.
- MCQ2 (15%)
  - Examines further concepts.
  - Set on Thursday in week 5.
- MCQ3 (15%)
  - Examines all concepts delivered on the module.
  - Set on Thursday in week 7.

How MCQs Work I

The MCQ is open book and you have an extended time to complete the quiz. You cannot see the questions until the quiz opens; you can start anytime after the quiz opens and work for any length of time until the deadline. The time the quiz opens and the deadline are visible in advance.

You do not have to complete the MCQ in one sitting: you can save you answers and continue work later. You do not have to answer the questions in order. However, you must submit before the deadline and once submitted you cannot make any further changes. You do not get feedback until the deadline has passed.

Late submission is not possible.

How MCQs Work II

Each question in the MCQ has 5 statements. Each statement is either true or false. You must choose either true or false, or you may choose neither if you do not know whether the statement is true or false. **Incorrect answers in the MCQs have negative marks.** Overall, the marking is as follows:

- The correctly choose either true or false: +1
- You choose not to answer: 0
- You choose true or false incorrectly: -1 (however, the minimum mark on any question is zero).

Negative marking is used to encourage you to tell the difference between what you know and what you do not yet understand. The effect of the negative marks is to discourage guessing; without this, random answers would gain a mark of around 50%.

QM+ does not calculate the negative marks, so the mark displayed (when the deadline passes) needs to be adjusted. You can do this but counting the number of incorrect responses you have made and subtracting this from the score shown.

From Web Page – Notebook Submission

- Data analysis exercises - 30% in total
  - Course work 1 (15%). Released before the start of week 1; submission Thursday of week 2.
  - Course work 2 (15%). Released during week 2; submission Thursday of week 4.

- Statistical exercises, submitted as notebooks - 30% in total
  - Course work 3 (15%): Released during week 4; submission Thursday of week 6.
  - Course work 4 (15%): Released during week 5; submission Thursday of week 8.

Feedback and Issues

- Feedback has the following forms
  - Sample answer
  - Individual grades of r sub-parts and overall marks

- Sample answer released after last submission deadline
  - Late submission penalties (1 week)
  - Possible extensions

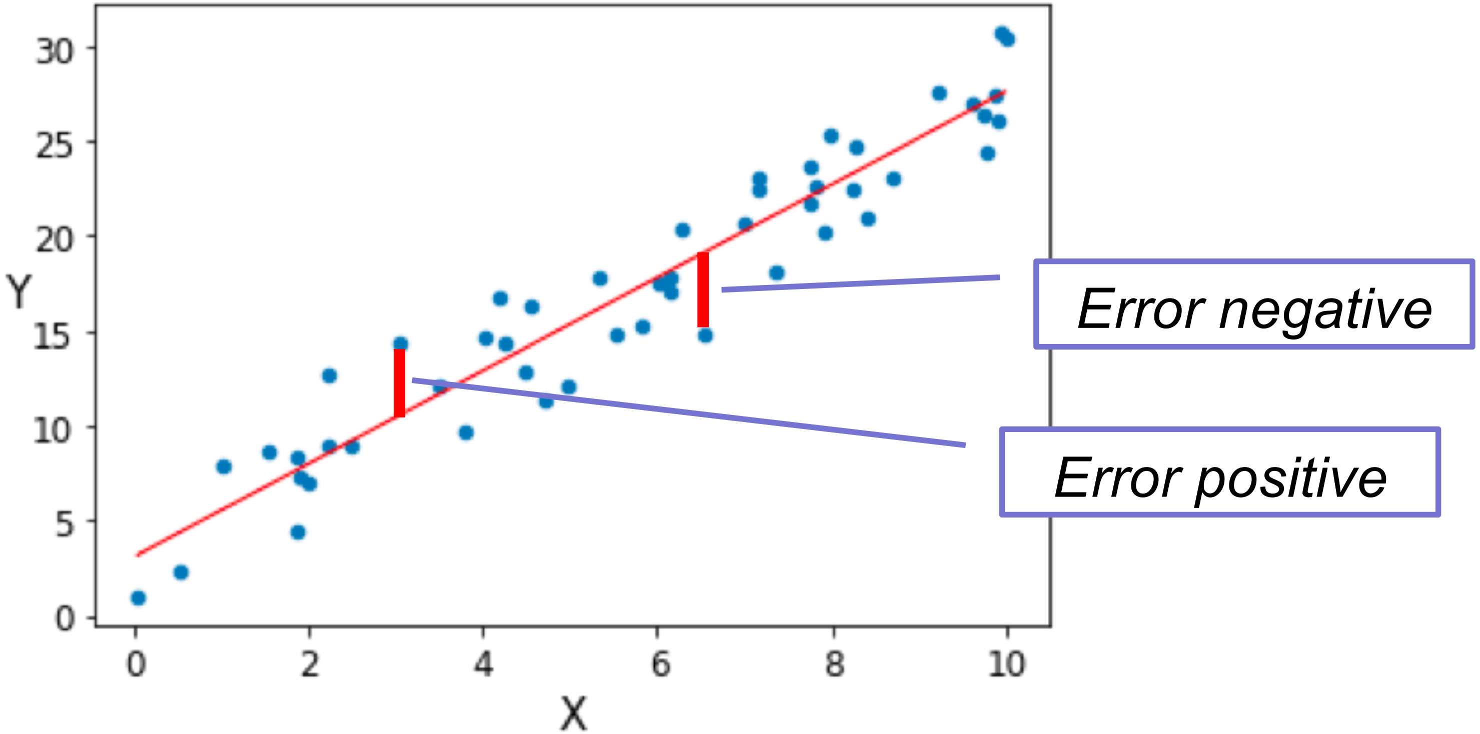
Fitting the Regression Line

Regression Line for Data Points

- Points are not exactly on a line

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

error



Residuals (Errors)

Prediction – if the point on the line

$$\hat{y}_{i_i} = \beta_0 + \beta_1 x_{1i}$$

Actual – off the line by an error

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

Residuals (errors)

$$e_i = y_i - \hat{y}_{i_i}$$

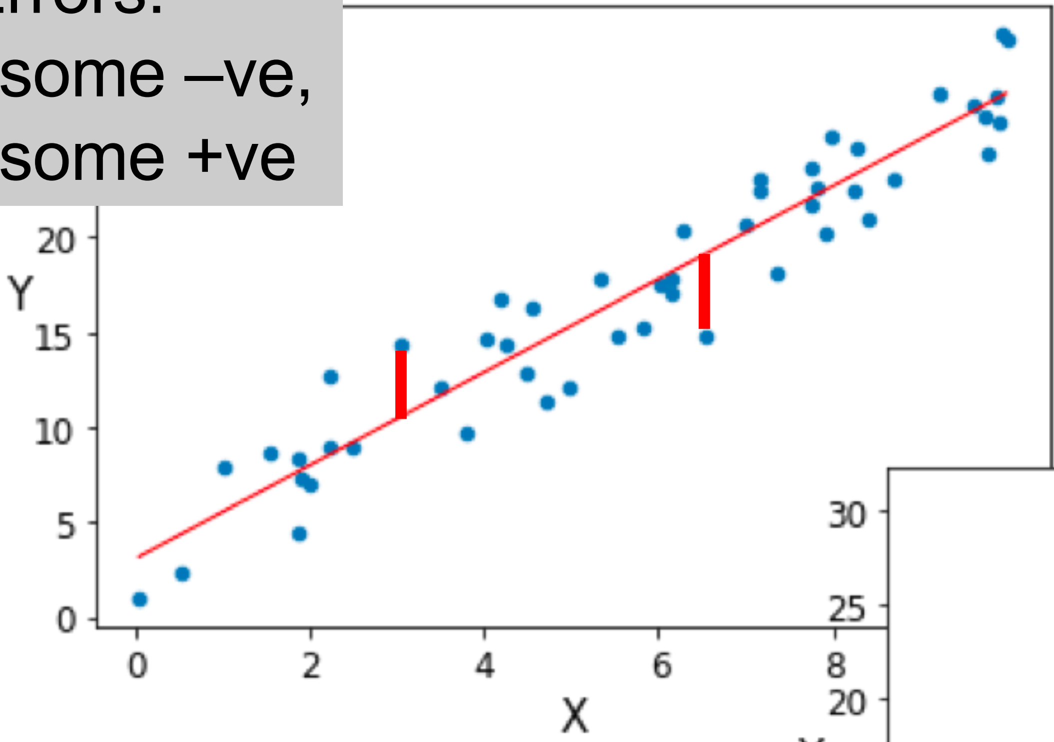
Best Fit Regression Line

- Give a set of data points (and chosen predictors)
  - Choose the 'best' values of the parameters
  - *Measure whether it is a good fit*

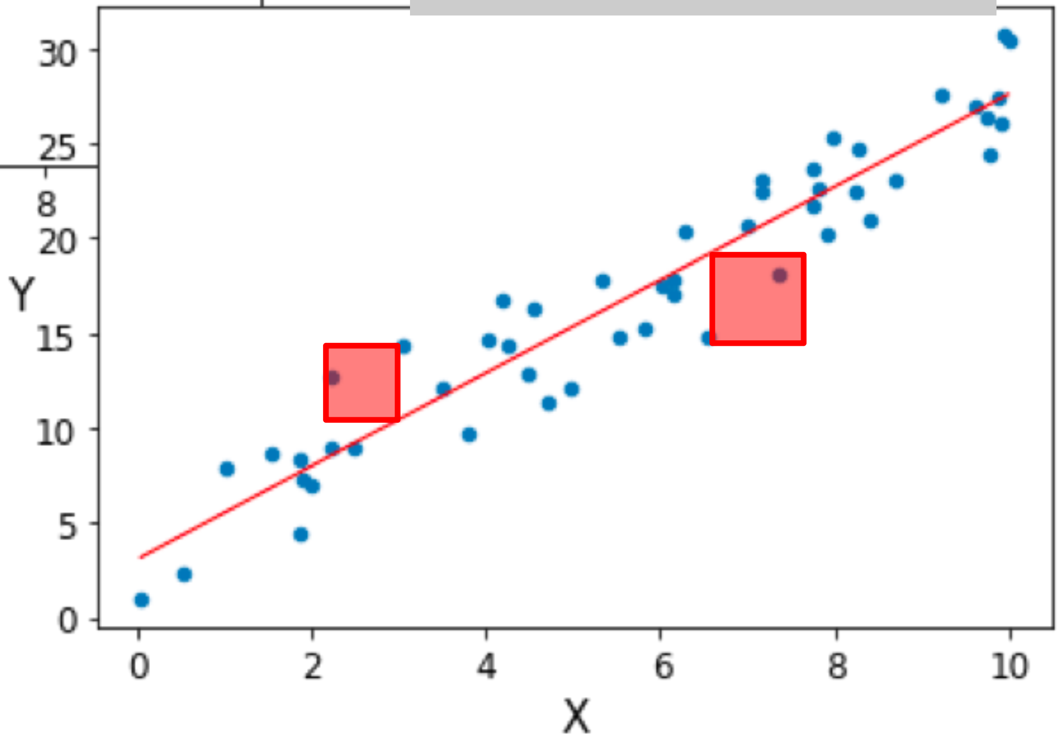
- 'Best' parameters $\widehat{\beta_0}$ and $\widehat{\beta_1}$
  - Minimise 'residual sum of squares' RSS
  - RSS = $\sum {e_i}^2 = \sum\left(y_i - \hat{y}_{i_i}\right)^2$
  - Idea: balance the errors
  - 'Ordinary least squares'

Understanding RSS

Errors:
some –ve,
some +ve



Errors squared:
All +ve
Min => balance



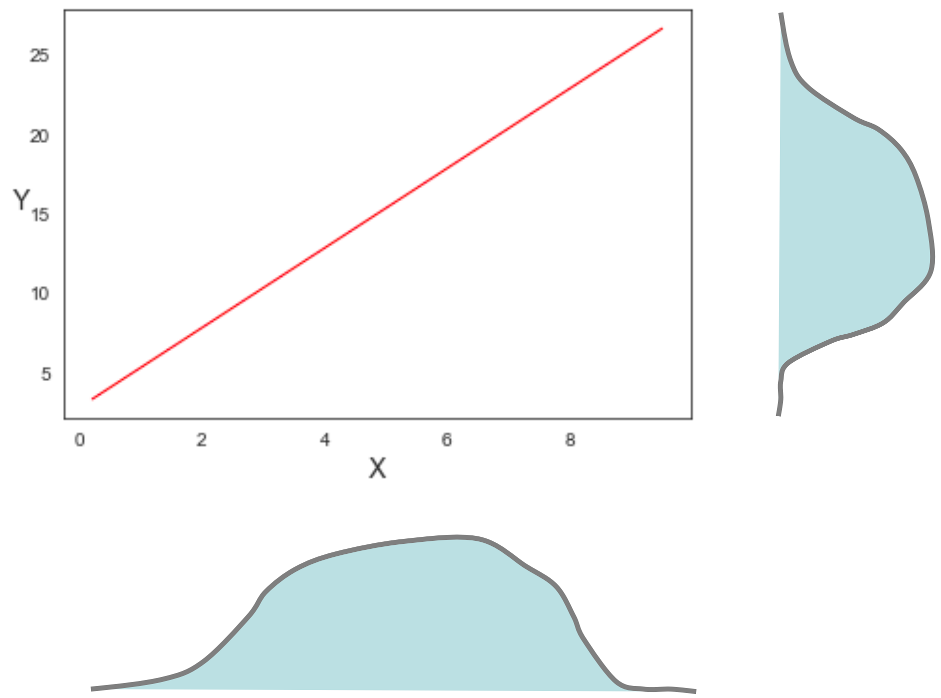
Warning: remove outliers

‘Best’ Fit and Distribution of Errors

- Theory assumes that distribution of residuals (errors) is normal
- You can check this
  - Plot the distribution
  - QQplot for normality
- If distribution of residuals skewed, then the parameters may not be ‘best’

Goodness of Fit: R^2

- R^2 is popular: *coefficient of determination*
- Range 0 to 1
- Proportion of the variance of Y that is predictable from X
  - Rest of the variance due to errors
  - i.e. missing predictors



Goodness of Fit: R^2

- Proportion of the variance Y that is predictable from X

$$R^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- Perfect prediction the $R^2 = 1$
- If we always predict $\bar{y}$ then $R^2 = 0$

- *Note: this is not the most general definition, but it applies in linear regression*

Connection between R^2 and $\rho_{x,y}$

- Recap: The correlation coefficient
  - Range -1 to 1
  - True value $\rho_{x,y}$ and sample $\mathrm{r}_{x,y}$
- Linear regression
  - Dependent variable y_i and predicted value $\widehat{y_i}$
  - Goodness of fit $R^2 = {r_{y,\hat{y}}}^2$
- Linear regression with a single predictor X
  - Goodness of fit $R^2 = {r_{x,y}}^2$

So 'correlation coefficient' actually about linear relationship

Goodness of Fit: RMSE

- RMSE: root mean squared error
- RMSE = $\sqrt{\frac{1}{N}\sum_i {e_i}^2} = \sqrt{\frac{1}{N}\sum_i \left(y_i - \widehat{y_i}_i\right)^2}$
  - Instead of N, sometimes N – p – 1 (for p predictors) as number of degrees of freedom
- More common in ML
  - Accuracy of predictor for continuous variable

Quiz 2

How Many Predictors?

How Many Predictors

- Issue 1: Enough Data?
  - Each has β to be estimated from data
  - A statistical model can be too complex for the data
  - Most statistical models have more parameters

- Issue 2: co-linearity
  - Remember assumption: predictor independent
  - Always check correlation of predictors

- Stepwise regression
  - Algorithm for choosing best set of predictors

Is Everything Linear?

- No

- Log transformation of predictor with positive skew
  - Best if supported by some theory
- Interaction
  - Predictors act together, not independently
- Non-linear relationship – e.g. $y = x^2$

Including Categorical Variables

Categorical Variables in Regression

- Recall: we cannot add or multiply categories
- Create binary dummy variables
  - False or 0 – no contribution to target variable
  - True or 1 – contribution determined by β
- Reference coding

- Example: ‘ChestPain’ variable in heart data
  - Typical
  - Atypical
  - Non_anginal
  - Asymptomatic

Reference Coding

Drop first

Patient	Age	ChestPain
1	46	Typical
2	53	Atypical
3	90	Typical
4	51	Asymptomatic
5	75	Non_anginal
6	67	Typical

Use ‘Typical’ as reference value

Patient	Age	Atypical	Asymptomatic	Non_anginal
1	46	0	0	0
2	53	**1**	0	0
3	90	0	0	0
4	51	0	**1**	0
5	75	0	0	**1**
6	67	0	0	0

Summary

- Regression is a fundamental technique
  - First type of statistical model
  - Many elaborations
- Finding parameters for ‘best fit’
- Measuring ‘goodness of fit’
- Including categorical variables as predictors
- Understand the assumption