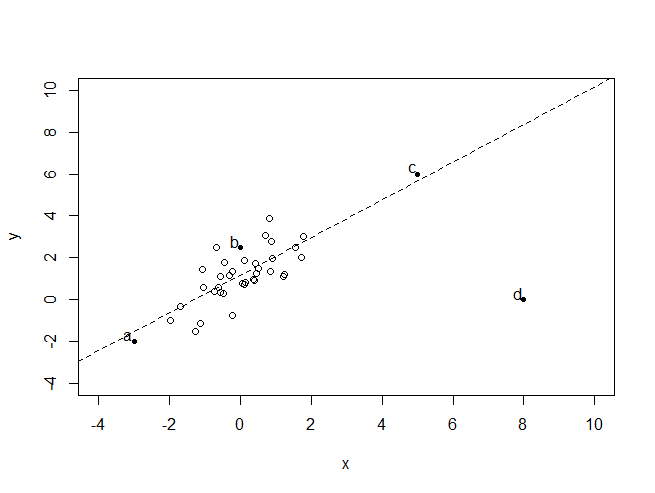
## Week 1 – Simple Linear Regression

1. Assume you fit a regression model to predict house prices from square feet based on a training data set consisting of houses with square feet in the range of 1000 to 2000. In which interval would we expect predictions to do best?
   1. [1000, 2000]
2. In a simple regression model, if you increase the input value by 1, then you expect the output to change by:
   1. The value of the *slope* parameter.
3. Two people present you with fits of their simple regression model for predicting house prices from square feet. You discover that the estimated intercept and slopes are exactly the same. This necessarily implies that these two people fit their models on exactly the same data set.
   1. False
4. You have a dataset consisting of the sales prices of houses in your neighbourhood, with each sale time-stamped by the month and year in which the house sold. You want to predict the average value of houses in your neighbourhood over time, so you fit a simple regression model with average house price as the output and the time index (in months) as the input. Based on 10 months of data, the estimated intercept is $4,569 and the estimated slope is 143 ($/month). If you extrapolate this trend forward in time, at which time index (in months) do you predict that your neighbourhood’s value will have doubled relative to the value at month (index) 10? (Assume months are 0-indexed).
   1. 52
5. Your friend in the US gives you a simple regression fit for predicting house prices from square feet. The estimated intercept is -44,850 and the estimated slop is 280.76. You believe that your housing market behaves very similarly, but houses are measured in square metres. To make predictions for inputs in square metres, what intercept must you use?
   1. -44,850
6. To make predictions for inputs in square meters, what slope must you use?
7. Consider the following dataset and the regression line fitted on this data:



Which bold/labelled point, if removed, will have the largest effect on the fitted regression line?

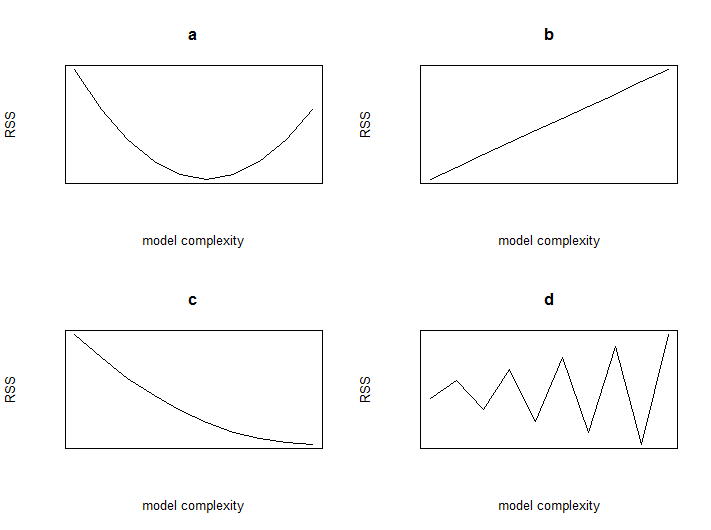
* 1. D.

## Week 2 – Multiple Regression

1. Which of the following is NOT a linear regression model.
2. Your estimated model for predicting house prices has a large positive weight on 'square feet living'. This implies that if we remove the feature 'square feet living' and refit the model, the new predictive performance will be worse than before.
   1. False.
3. Your estimated model for predicting house prices has a positive weight on 'square feet living'. You then add 'lot size' to the model and re-estimate the feature weights. The new weight on 'square feet living' [\_\_\_\_\_\_\_\_\_] be positive.
   1. Might.
4. If you double the value of a given feature (i.e. a specific column of the feature matrix), what happens to the least-squares estimated coefficients for every other feature? (Assume you have no other feature that depends on the doubled feature i.e. no interaction terms).
   1. They stay the same.
5. Gradient descent/ascent is:
   1. An algorithm for maximising/minimising a function.
6. Gradient descent/ascent allows us to:
   1. Estimate model parameters from data.
7. Which of the following statements about step-size in gradient descent are true?
   1. If the step-size is too large, gradient descent may not converge.
   2. If the step-size is too small (but not zero), gradient descent may take a very long time to converge.
8. Let's analyse how many computations are required to fit a multiple linear regression model using the closed-form solution based on a data set with 50 observations and 10 features. In the videos, we said that computing the inverse of the matrix was on the order of operations. Let's focus on forming this matrix prior to inversion. How many multiplications are required to form the matrix ?
9. More generally, if you have features and observations what is the total complexity of computing ?

## Week 3 – Assessing Performance

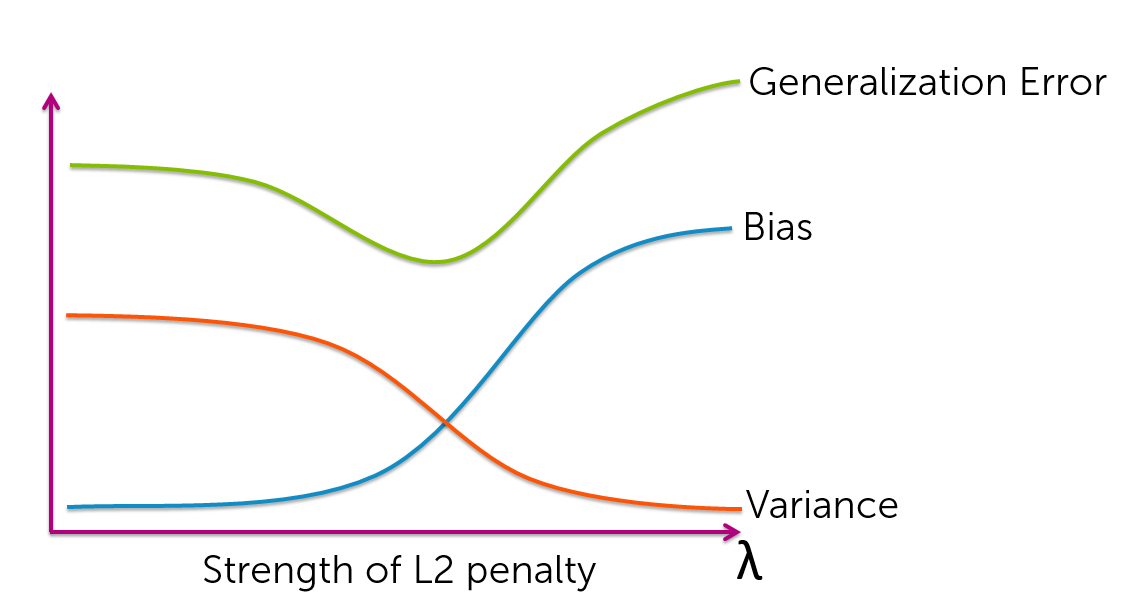
1. If the features of Model 1 are a strict subset of Model 2, the training error of the two models can never be the same.
   1. False.
2. If the features of Model 1 are a strict subset of Model 2, which model will usually have the lowest training error?
   1. Model 2.
3. If the feature of Model 1 are a strict subset of Model 2, which model will usually have the lowest test error?
   1. It’s impossible to tell with only this information.
4. If the features of Model 1 are a strict subset of those in Model 2, which model will usually have lower bias?
   1. Model 2.
5. Which of the following plots of model complexity vs RSS is most likely from the training data?



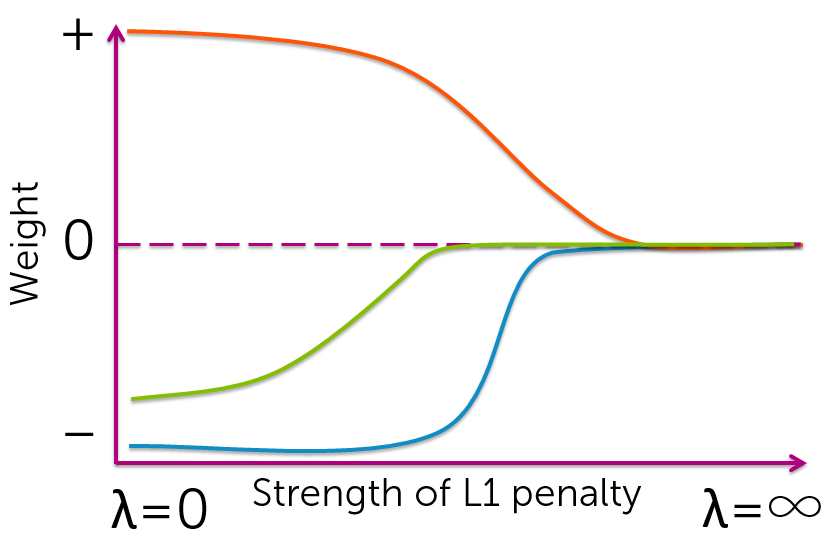
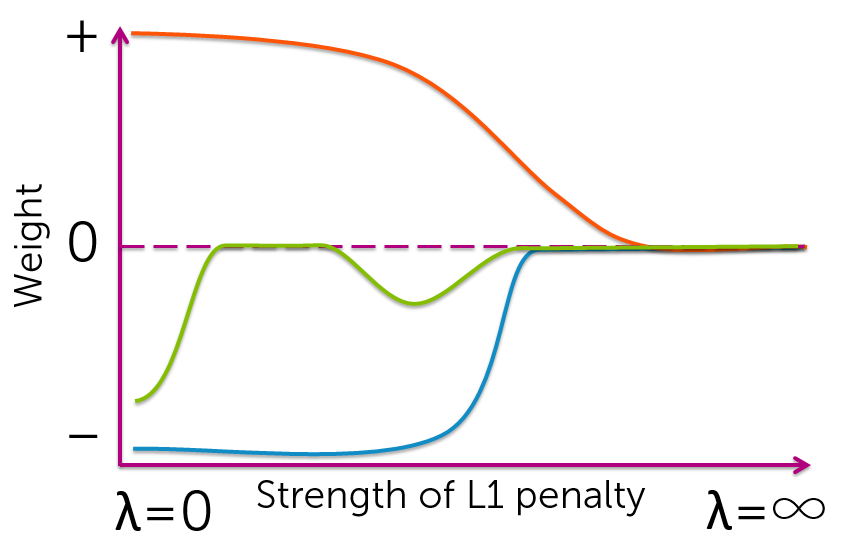
* 1. C

1. Which of the plots of model complexity vs RSS is most likely from the test data?
   1. A
2. It is always optimal to add more features to a regression model?
   1. False
3. A simple model with few parameters is most likely to suffer from:
   1. High bias.
4. A complex model with many parameters is most likely to suffer from:
   1. High variance.
5. A model with many parameters that fits training data very well, but does poorly on test data is considered to be:
   1. Overfitted.
6. A common process for selecting a parameter like the optimal polynomial degree is:
   1. Minimising validation error.
7. Selecting model complexity on test data:
   1. Provides an overly optimistic assessment of performance of the resulting model.
   2. Should never be done.
8. For a fixed model complexity, in the limit of an infinite amount of training data:
   1. Variance goes to 0.

## Week 4 – Ridge Regression

1. Which of the following is NOT a valid measure of overfitting?
   1. Sum of parameters
2. In ridge regression, choosing a large penalty strength tends to lead to a model with:
   1. High bias and low variance.
3. Which of the following plots best characterize the trend of bias, variance, and generalization error (all plotted over )?
   1. 
4. In ridge regression using unnormalized features, if you double the value of a given feature (i.e., a specific column of the feature matrix), what happens to the estimated coefficients for every other feature? They:
   1. Impossible to tell from the information provided.
5. If we only have a small number of observations, K-fold cross validation provides a better estimate of the generalization error than the validation set method.
   1. True.
6. 10-fold cross validation is more computationally intensive than leave-one-out (LOO) cross validation.
   1. False
7. Assume you have a training dataset consisting of observations and features. You use the closed-form solution to fit a multiple linear regression model using ridge regression. To choose the penalty strength , you run leave-one-out (LOO) cross validation searching over values of . Let be the computational cost of running ridge regression with data points and features. Assume the prediction cost is negligible compared to the computational cost of training the model. Which of the following represents the computational cost of your LOO cross validation procedure?
8. Assume you have a training dataset consisting of 1 million observations. Suppose running the closed-form solution to fit a multiple linear regression model using ridge regression on this data takes 1 second. Suppose you want to choose the penalty strength by searching over 100 possible values. How long will it take to run leave-one-out (LOO) cross-validation for this selection task?
   1. About 3 years.
9. If you only want to spend about 1 hour to select , what value of should you use for k-fold cross-validation?
   1. .

## Week 5 – Feature Selection and Lasso

1. The best fit model of size 5 (i.e. with five features) always contains the set of features from the best fit model of size 4.
   1. False
2. Given 20 potential features, how many models do you have to evaluate in the all-subsets algorithm?
3. Given 20 potential features, how many models do you have to evaluate if you are running the forward stepwise greedy algorithm? Assume you run the algorithm all the way to the full feature set.
4. Which plot/s could correspond to a lasso coefficient path?  
   Hint: notice in the bottom right of the plots. How should coefficients behave eventually as goes to infinity?
   1. 
   2. 
5. Which statements about coordinate descent are true?
   1. To test the convergence of coordinate descent, look at the size of the maximum step you take as you cycle through coordinates.
6. Using normalised features, the OLS coordinate descent update for feature j has the form (where ):
7. Using normalised features, the ridge regression coordinate descent update for feature j has the form:

## Week 6 – Nearest Neighbours and Kernel Regression

1. Which of the following datasets is best suited to nearest neighbour or kernel regression?
   1. A dataset with two features whose observations are evenly scattered throughout the input space.
   2. A dataset with many observations.
2. What is the most significant advantage of k-nearest neighbour regression (for k>1) over 1-nearest neighbour regression?
   1. Better copes with noise in the data.
3. To obtain a fit with low variance using kernel regression, we should choose the kernel to have:
   1. Large bandwidth
4. In k-nearest neighbour regression and kernel regression, the complexity of functions that can be represented grows as we get more data.
   1. True
5. Parametric regression and 1-nearest neighbour regression will converge to the same solution as we collect more and more noiseless observations.
   1. False.
6. Y**ou are creating a website to help shoppers pick houses. Every time a user of your website visits the webpage for a specific house, you want to compute a prediction of the house value. You are using 1-NN to make the prediction and have 100,000 houses in the dataset, with each house having 100 features. Computing the distance between two houses using all the features takes about 10 microseconds. Assuming the cost of all other operations involved (e.g., fetching data, etc.) is negligible, about how long will it take to make a prediction using the brute-force method described in the videos?**
7. You learn that you need predictions within 50 milliseconds. To accomplish this, you decide to reduce the number of features in your nearest neighbour comparisons. How many features can you use?