ML: Lecture 2

Regression Analysis

Ioanna Miliou, PhD

Today

- What is regression?
- What is linear regression?
- What is regularization?
- What are the different regularization techniques or shrinkage methods?
- Which are the evaluation metrics for regression?
- What is logistic regression?

Examples of Supervised Learning

- Predict the price of a stock in 6 months from now
  - based on company performance measures, economic data and other exogenous variables
- Predict whether a patient, hospitalized due to heart attack, will have a second heart attack
  - based on demographics, diet, clinical measurements, and clinical history
- Estimate the amount of glucose in the blood of a diabetic person
  - based on the infrared absorption spectrum of that person's blood
- Predict whether an email is spam or not
  - based on the words and punctuation marks in the email message

Examples of Supervised Learning

- Predict the price of nths from now
  - based on company easures, economic data and other exogenous variables
- Predict whether a patient, hospitalized due to heart attack, will have a se tack
  - based on demogra ical measurements, and clinical history
- Estimate the amou e in the blood of a diabetic person
  - based on the infrared absorption spectrum of that person's blood
- Predict whether an or not
  - based on the wor tion marks in the email message

quantitative

categorical

quantitative

categorical

Formulating the problem

In a typical scenario, we have an outcome measurement, usually quantitative (such as a stock price) or categorical (such as heart attack/no heart attack), that we wish to predict based on a set of features.

- Regression: when we predict quantitative outputs
- Classification: when we predict categorical (qualitative) outputs

Regression

- Statistical measure that attempts to determine the relationship between the dependent variable Y and a set of independent variables X

- Widely used for predicting the next value or values of the dependent variable Y from the values of the independent variables X

Dependent and independent variables

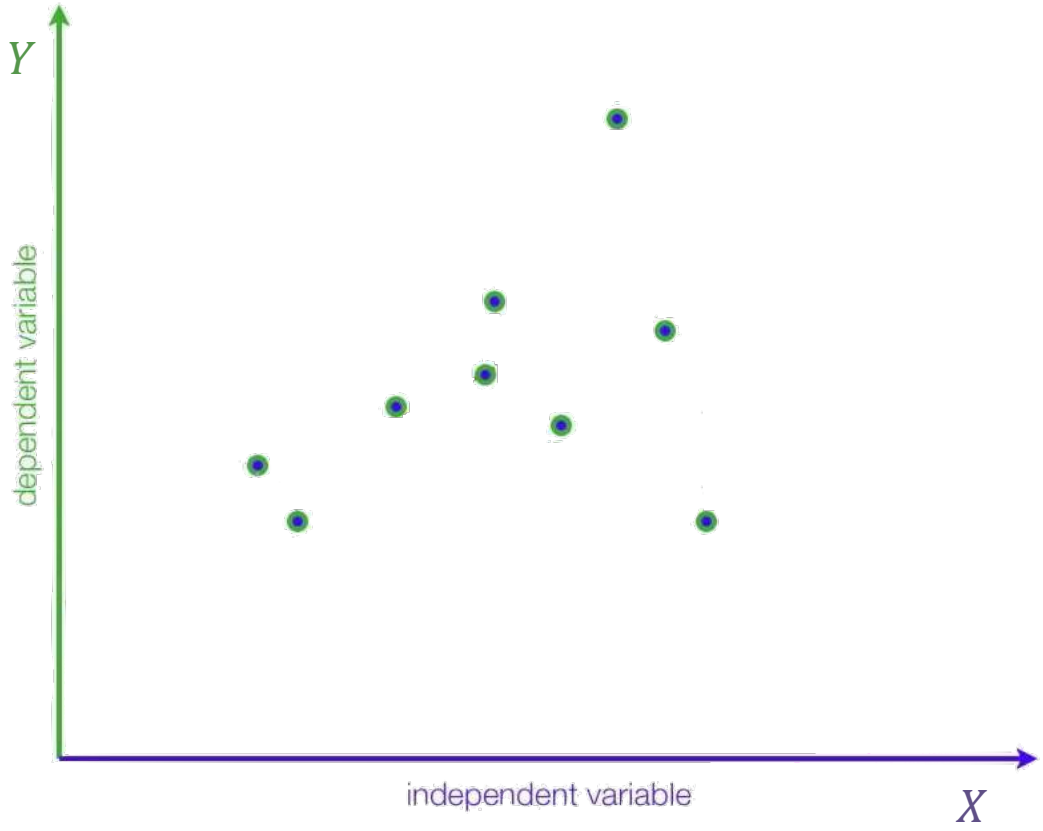
- Dependent variable – Y – is the variable whose values change as a consequence of changes in other values in the system
  - Also called response variable

- Independent variables – $X_1, X_2, \ldots, Xp$ – are regarded as inputs to the system and may take on different values freely
  - Also called predictors or explanatory variables

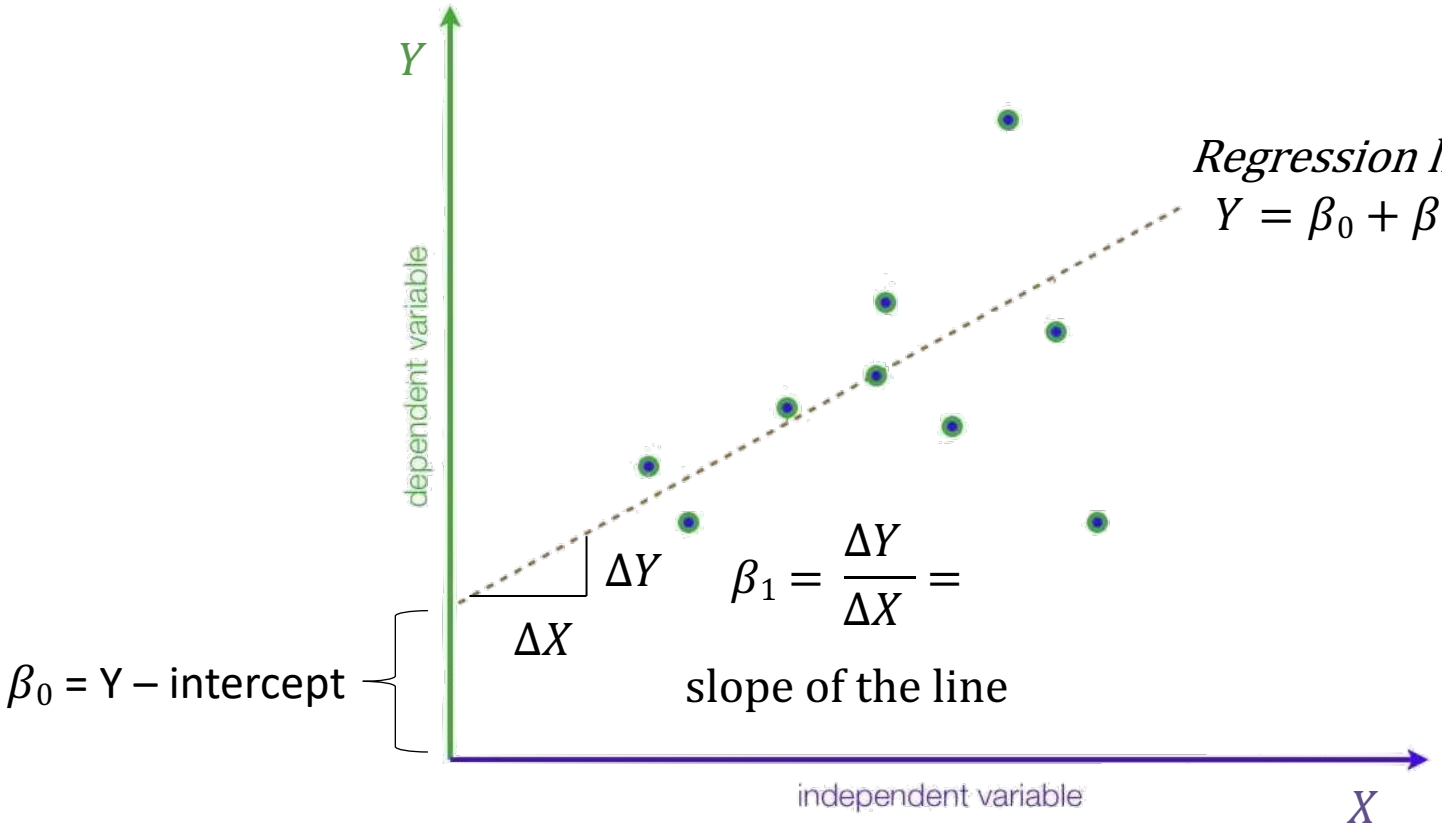
Simple Linear regression

- The simplest mathematical relationship between the variables X and Y is a linear relationship
- In a cause-and-effect relationship, the independent variable X is the cause, and the dependent variable Y is the effect
- It attempts to model the relationship between the two variables by fitting a linear equation to the observed data

Linear Model

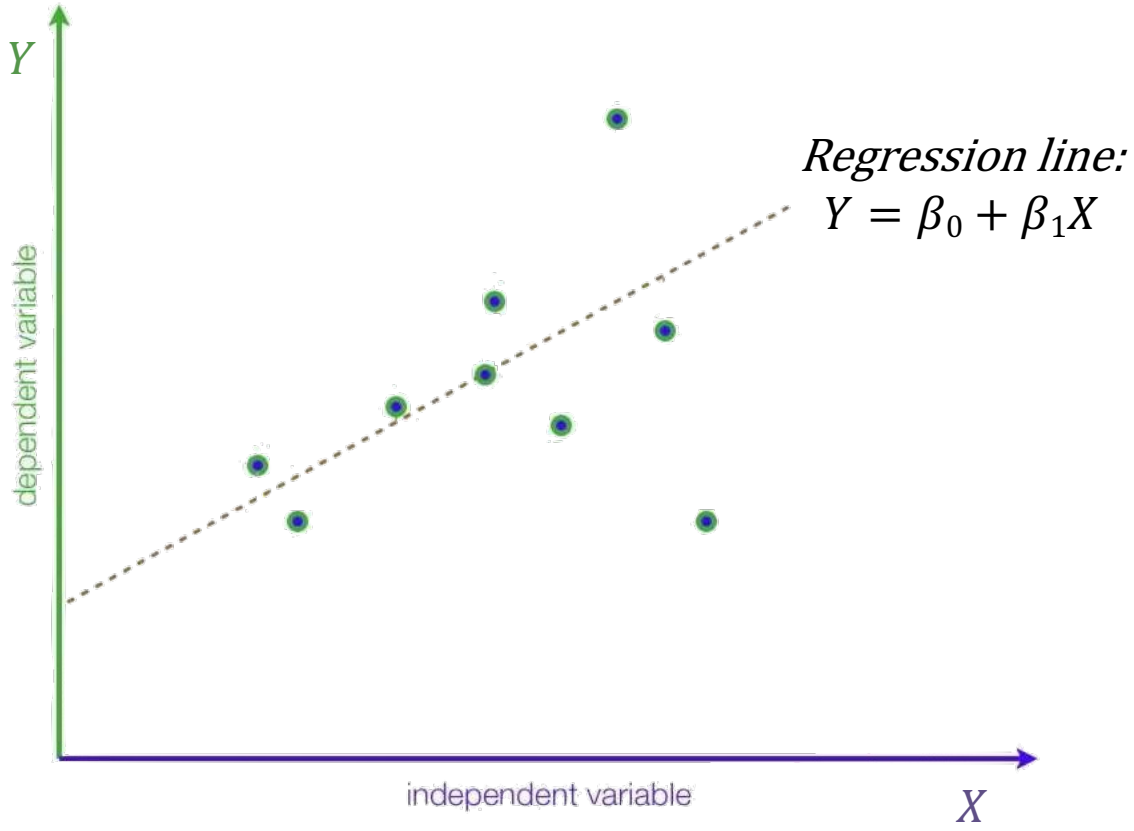


Linear Model

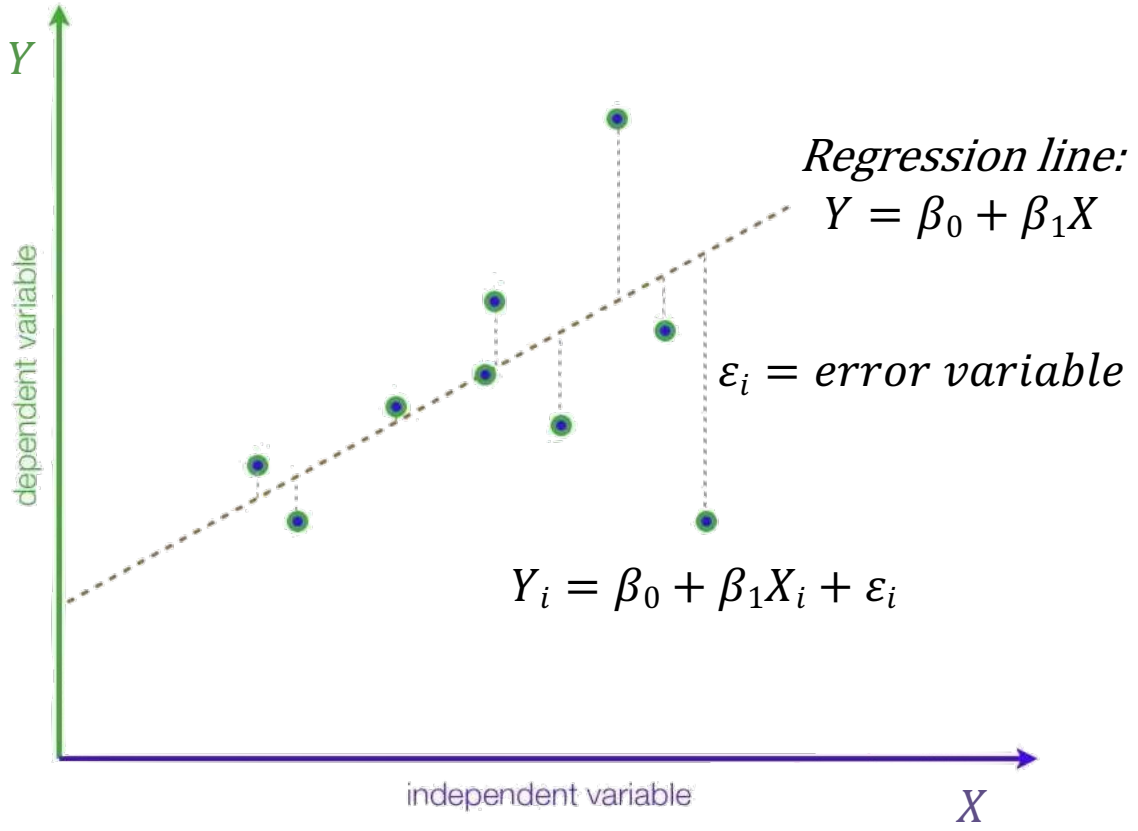


β_0, β_1 : *model coefficients or parameters*

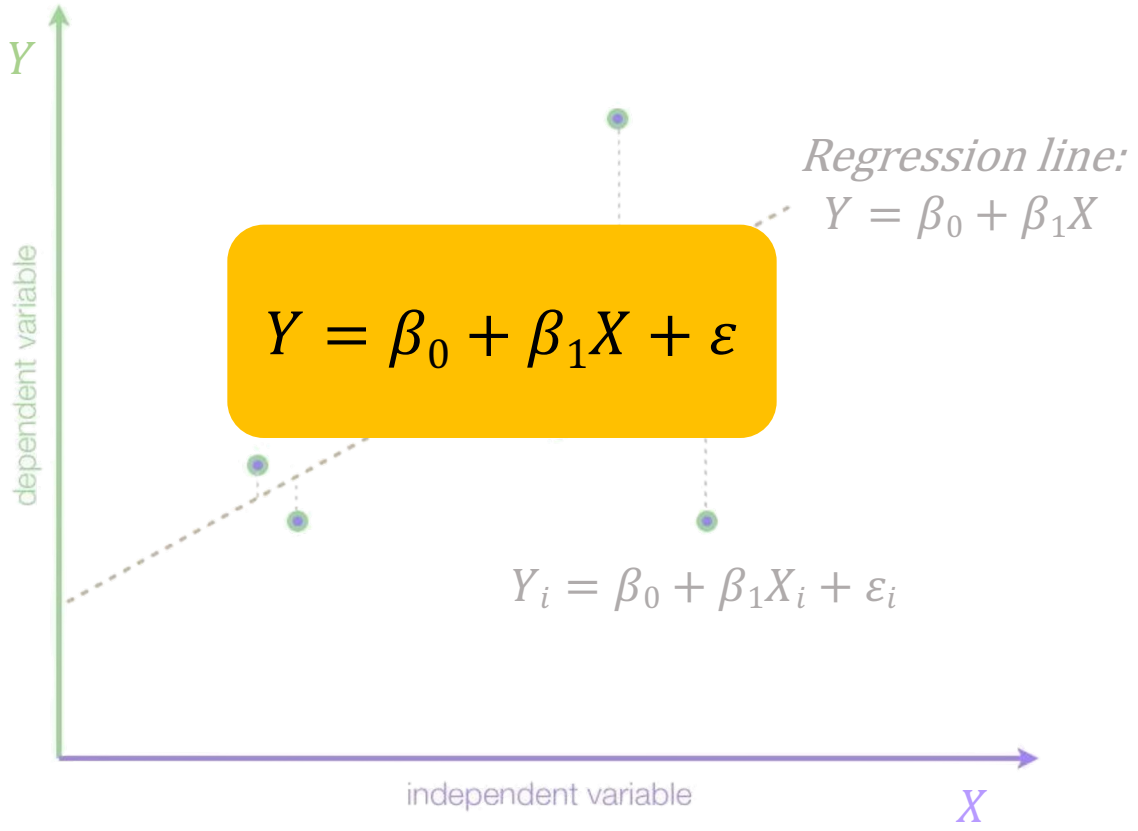
Linear Model



Linear Model



Linear Model



Goal

- To find the line that best describes the data

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- To find estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters β_0 and β_1 which would provide the "best fit" for the training data

How do we fit the linear model?

- We attempt to find the "best fit line" by minimizing the difference between the actual and predicted Y values
- But positive differences could offset negative ones
- That's why we take the squared differences

Ordinary Least Squares

Ordinary Least Squares (OLS)

- Most common method for fitting a regression line
- It minimizes the residual sum of squares (RSS) of observed values to the straight-line:

$$RSS = \sum_{i=1}^{n} (Yi - \hat{Y}_i)^2$$

$$= \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$= \sum_{i=1}^{n} \hat{\varepsilon}_i{}^2$$

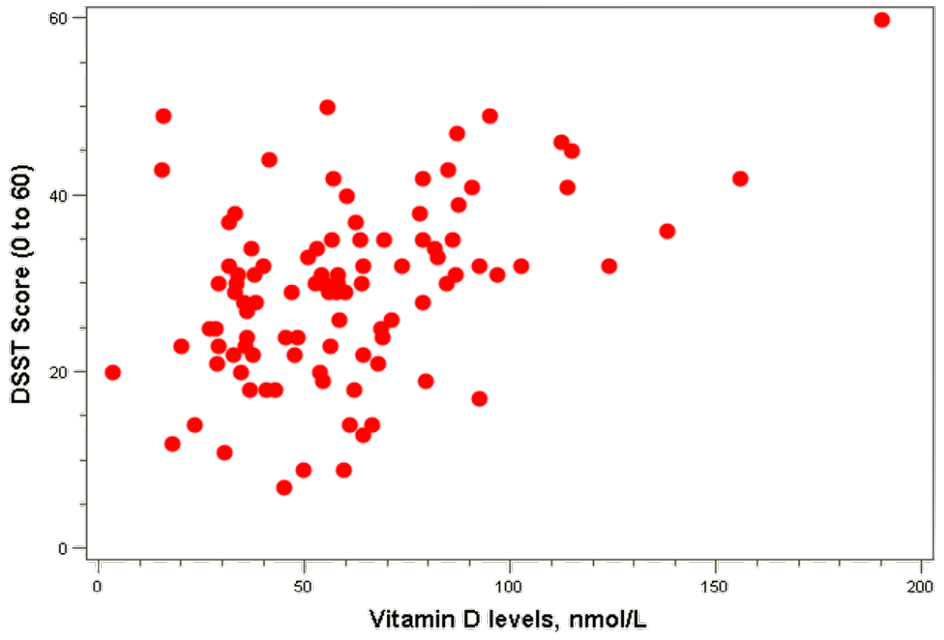
Coefficient Equations

- Y-intercept: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- Slope: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- Prediction equation: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

where $\bar{Y}, \bar{X}$ are the average of the Y_i, X_i *respectively*

Example: Cognitive function

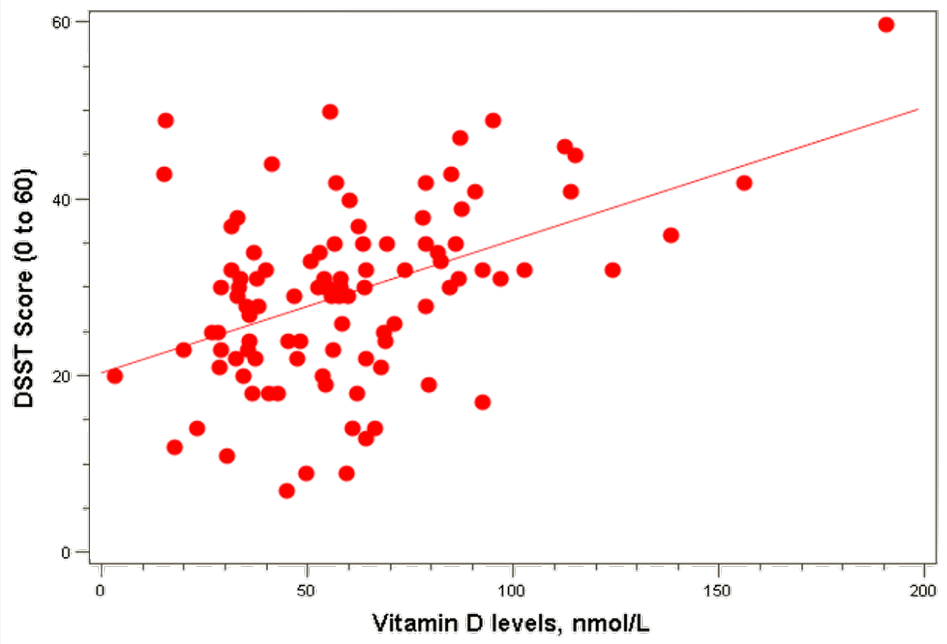
- Cognitive function is measured by the Digit Symbol Substitution Test (DSST)
- Study of the association between vitamin D levels and DSST
- Hypothetical data loosely based on [1]; cross-sectional study of 100 middle-aged and older European men



1. Lee DM, Tajar A, Ulubaev A, et al. *Association between 25-hydroxyvitamin D levels and cognitive performance in middle-aged and older European men*. J Neurol Neurosurg Psychiatry. 2009 Jul;80(7):722-9.

Example: Cognitive function

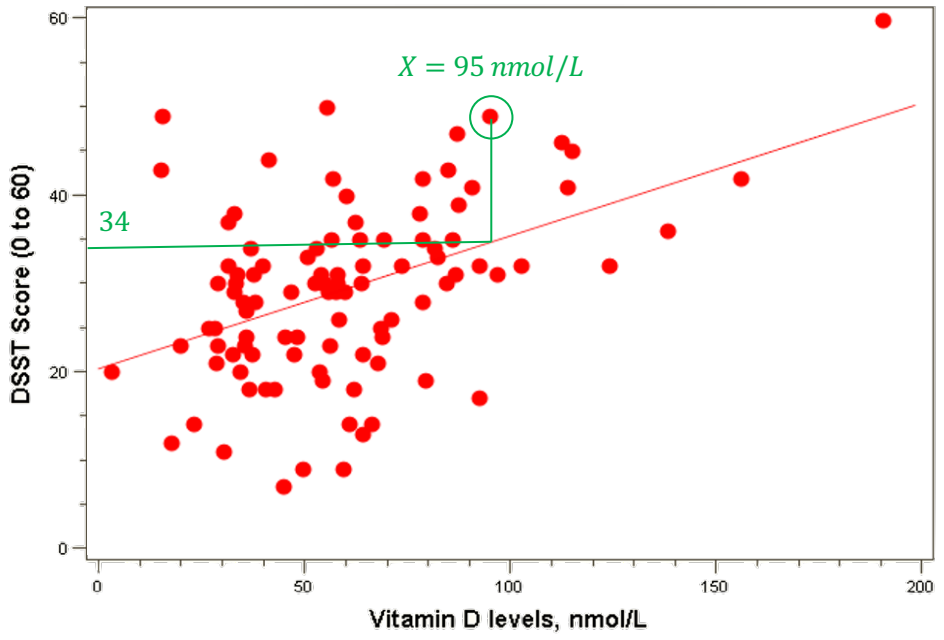
- $\hat{\beta}_0 = 20$
- $\hat{\beta}_1 = 1.5\ points$ $per\ 10\ nmol/L$
- $\hat{Y}_i = 20 + 1.5\ vitDi$



Example: Cognitive function

- $\hat{\beta}_0 = 20$
- $\hat{\beta}_1 = 1.5\ points$ $per\ 10\ nmol/L$
- $\hat{Y}_i = 20 + 1.5\ vitDi$
- *For vitD = 95 nmol/L (or 9.5 in 10 nmol/L):*

  $\hat{Y}_i = 20 + 1.5\ (9.5) = 34$



Multiple Linear Regression

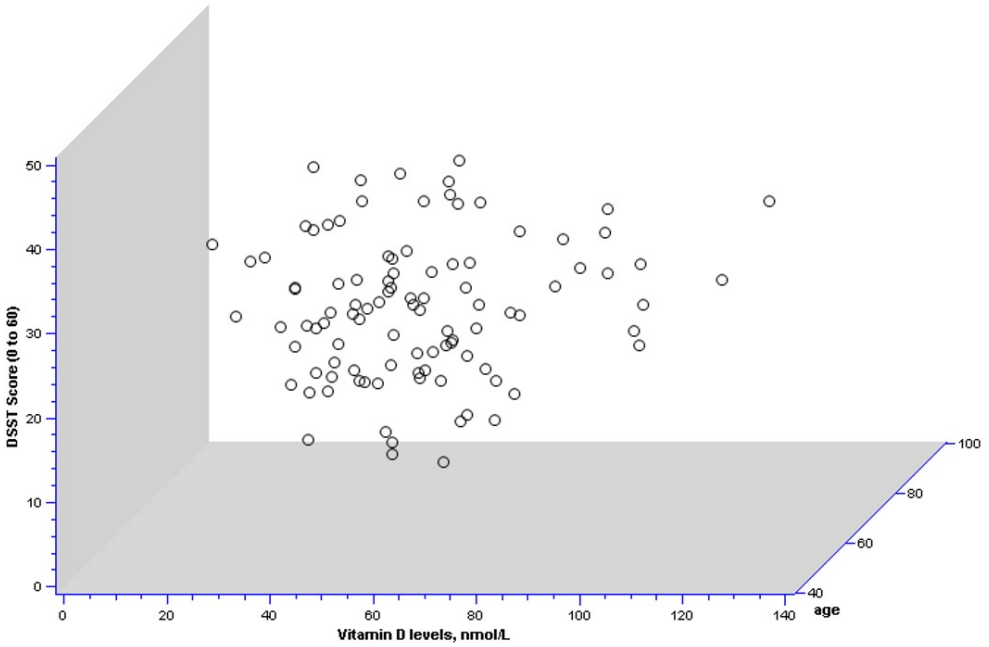
- Regression is not limited to two variables as we could have two or more variables showing a relationship
- When there is a single input variable, the regression is referred to as Simple Linear Regression
- When there are multiple input variables, $X^T = (X_1, X_2, \ldots, Xp)$, the regression is referred to as Multiple Linear Regression

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j$$

Example: Cognitive function

What if age is a confounder here?

- Older men have lower vitamin D
- Older men have poorer cognition



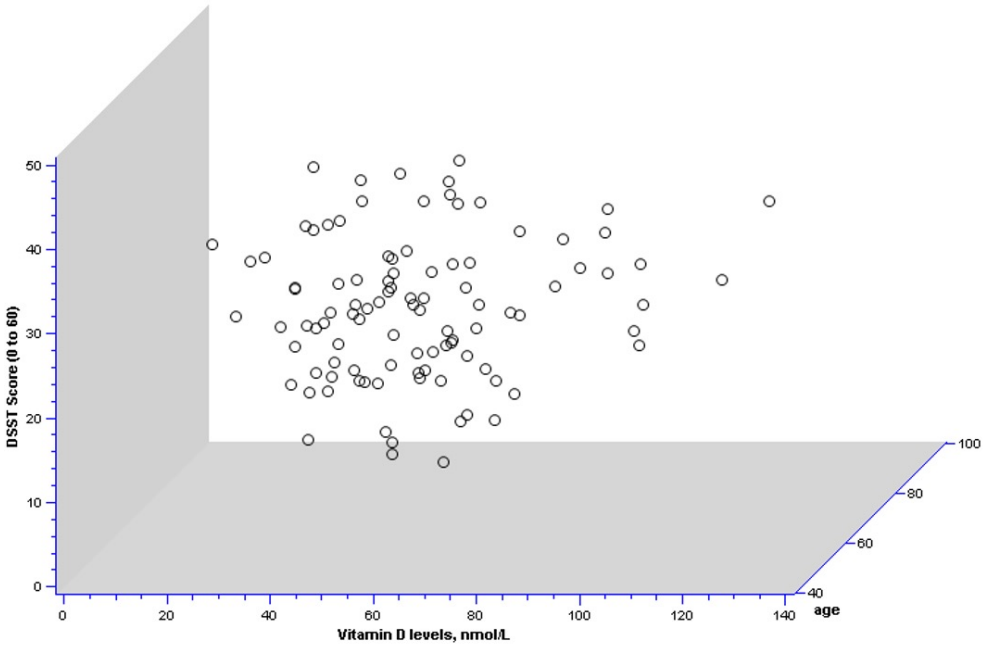
"Adjust" for age by putting age in the model:

$$DSST\ score = \hat{\beta}_0 + \hat{\beta}_1\ vitD + \hat{\beta}_2\ age$$

Example: Cognitive function

What if age is a confounder here?

- Older men have lower vitamin D
- Older men have poorer cognition



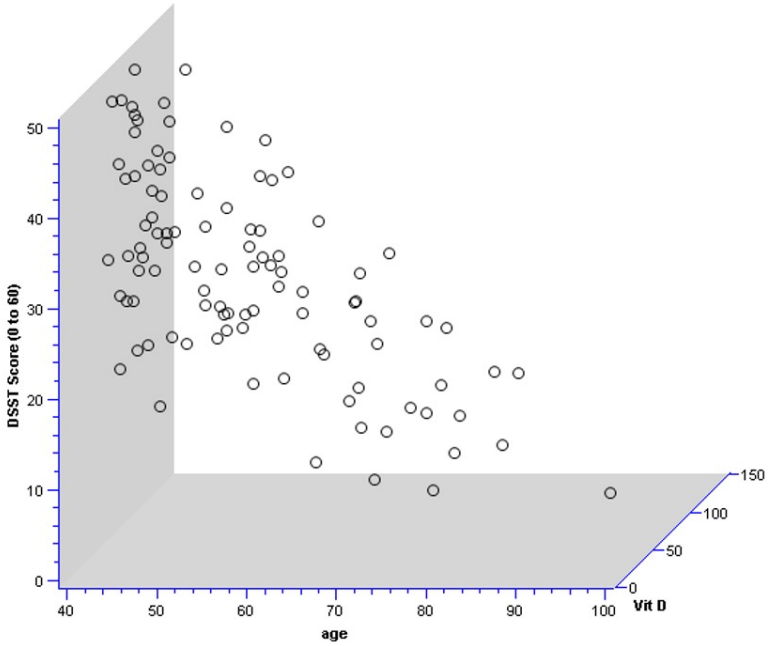
"Adjust" for age by putting age in the model:

$$DSST\ score = 53 + 0.039\ vitD\ (in\ 10nmol/L) - 0.46\ age$$

Example: Cognitive function

What if age is a confounder here?

- Older men have lower vitamin D
- Older men have poorer cognition



Thus, relationship with vitamin D was due to confounding by age!

Assumptions

- The observations are independent (random sampling)
- The relationship of Y with X and the error term is linear
- Y is normally distributed at each value of X
- The error term is normally distributed with mean zero and constant variance
- The X variables are independent – no multicollinearity (applies only to Multiple Linear Regression)

Gauss-Markov theorem: when all assumptions hold, OLS will produce better coefficient estimates compared to all other linear model estimation methods

Gradient Descent

- It's an optimization algorithm that finds the linear regression coefficients iteratively
- The model uses Gradient Descent to update the coefficient values in order to reduce cost function J (minimizing RSS) and achieving the best fit line
- A good way to ensure that Gradient Descent is working correctly is to make sure that the error decreases for each iteration

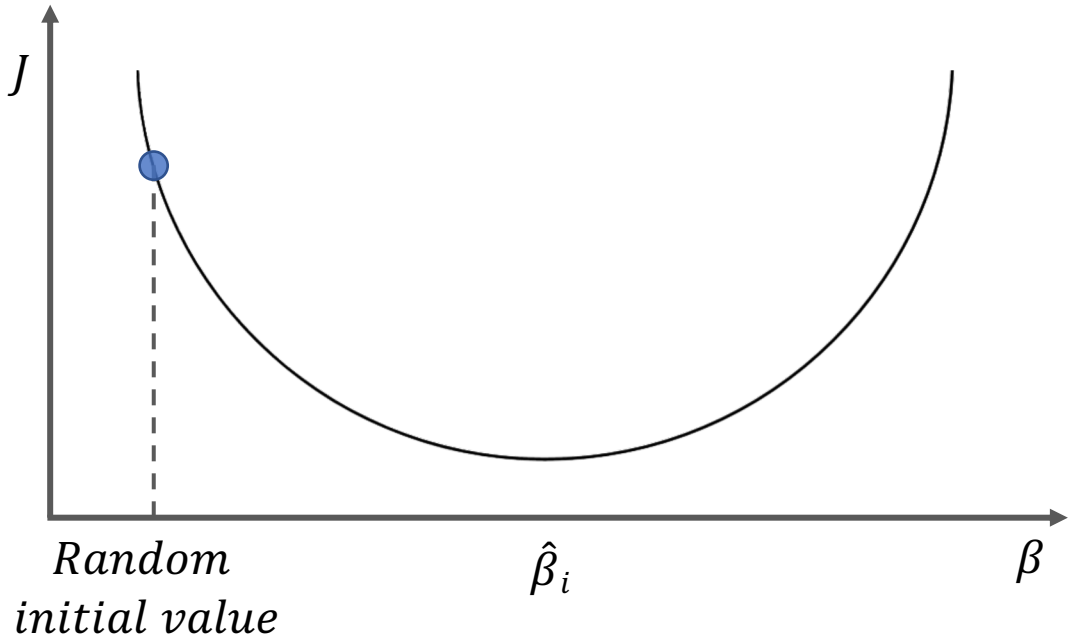
Gradient Descent

$$\hat{Y} = \hat{\beta}_0^{(k)} + \hat{\beta}_1^{(k)} X_i \qquad (k) = Iteration$$

$$J\left(\beta^{(k)}\right) = Err(k) = \sum_{i=1}^{n} (Y_i^{(k)} - \hat{\beta}_0^{(k)} - \hat{\beta}_1^{(k)} X_i)^2$$

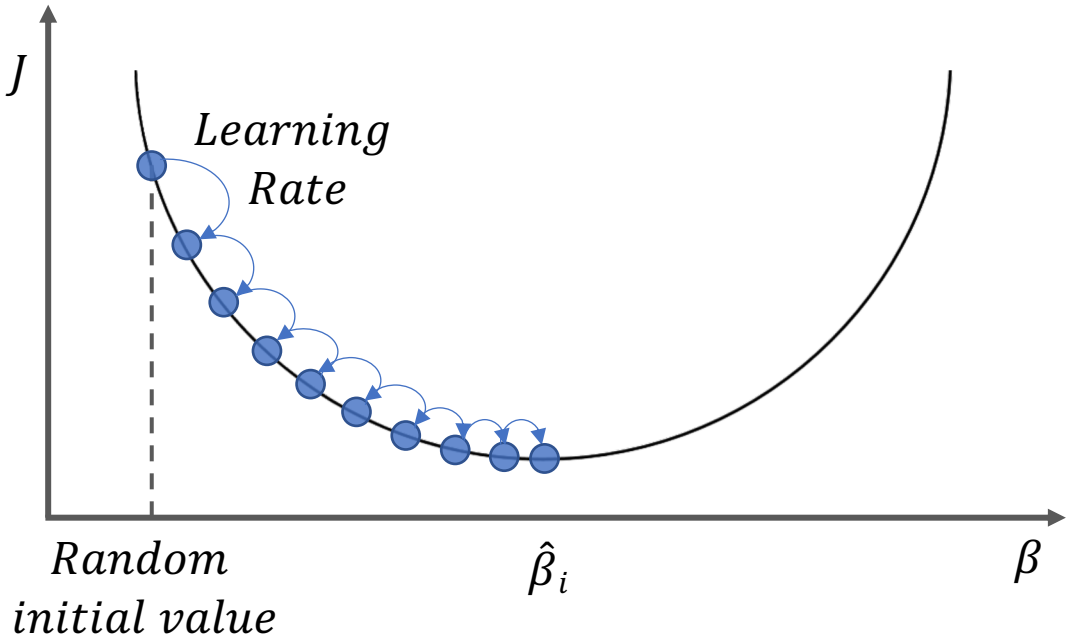
- Gradient Descent, at every iteration, step-downs the cost function in the direction of the steepest descent
- The size of each step is determined by parameter α known as Learning Rate

Gradient Descent



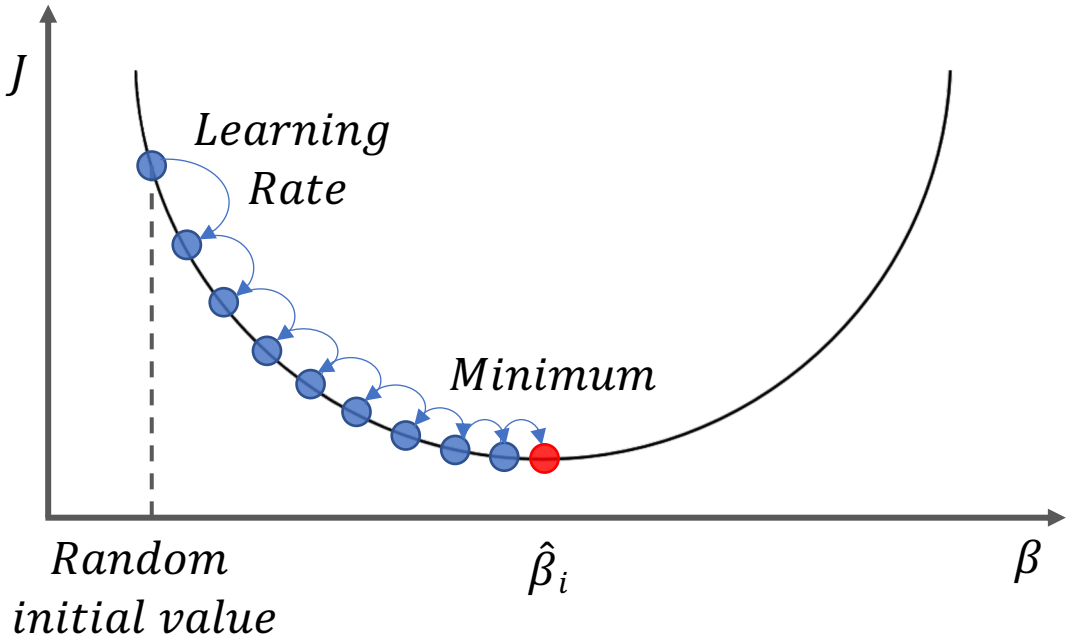
1. It assignes an initial set of random values to the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$

Gradient Descent



1. It assignes an initial set of random values to the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$
2. It iteratively updates those values proportional to the negative of the gradient of the function

Gradient Descent



1. It assignes an initial set of random values to the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$
2. It iteratively updates those values proportional to the negative of the gradient of the function
3. At the end, it finds a local (or global) minimum for the loss function J

How good is the fit?

The R^2 measure or Coefficient of Determination can be used to determine how well the model fits the training data:

$$R^2 = \frac{Variance\ explained\ by\ the\ model}{Total\ variance}$$

$$= 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

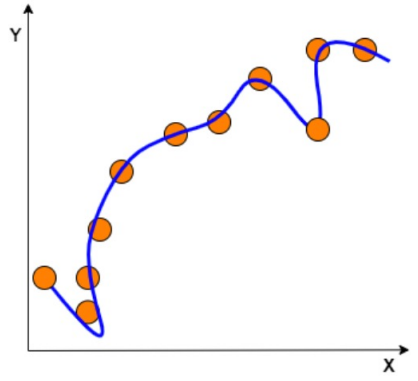
$$= 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

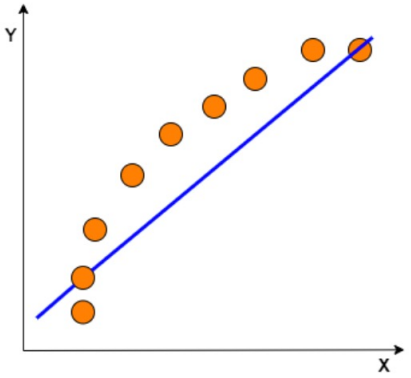
Residual Sum of Squared Errors, the difference between actual_y and predicted_y, squared.

Total Sum of Squared Errors, the difference between actual_y and the mean of y, squared.

Problems with fitting the data

- Over-fitting: the model models the training data too well
  - The model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data as it looses its ability to generalize
  - It happens often if we have large number of features and the test score is worst than the training score
- Under-fitting: model that can neither model the training data nor generalize to new data
  - It happens if we have very few features on the dataset and the score is poor for both training and test set





Regularization

- There are two reasons why we are often not satisfied with the Ordinary Least Squares:
  - Low performance (such as over-fitting): it could be improved by shrinking or setting some coefficients to zero
  - Interpretation: to get the "bigger picture" we would like to determine a smaller subset of predictors that exhibit the strongest effects, even if we are sacrificing some small details

- Solution: Regularizing the coefficient estimates (Shrinkage Methods)
  - Ridge Regression
  - Lasso Regression

Let's remember first..

- Multiple Linear Regression:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j$$

- Cost function:

$$RSS = \sum_{i=1}^{n} (Yi - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j X_{ij})^2$$

where n instances in the dataset and p predictors

Ridge Regression

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size
- The ridge coefficients minimize a penalized residual sum of squares:

$$\sum_{i=1}^{n}(Yi - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Yi - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j X_{ij})^2 + \lambda \sum_{j=1}^{p}\hat{\beta}_j{}^2$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage

Ridge Regression

- An equivalent way to write the ridge problem is:

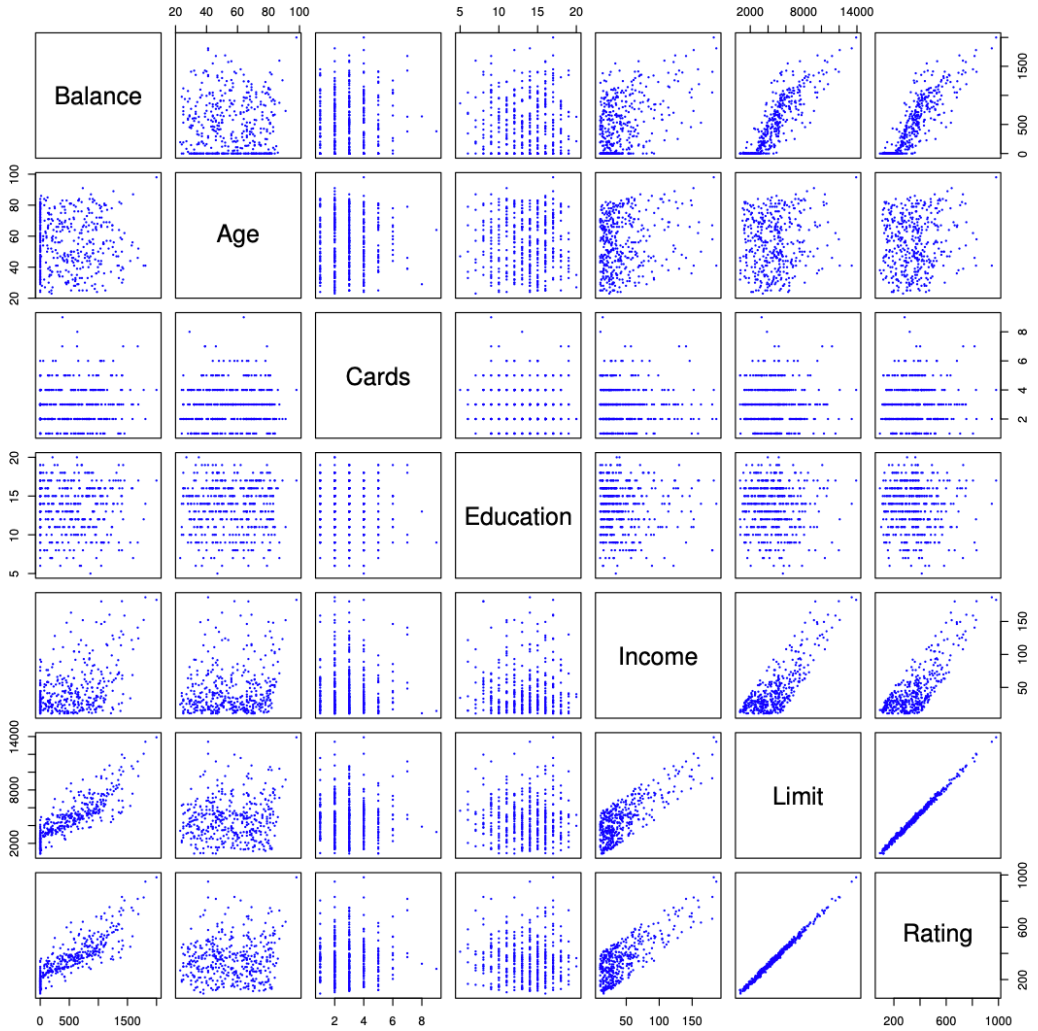
$$\sum_{i=1}^{n}(Yi - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j X_{ij})^2, \qquad subject\ to \sum_{j=1}^{p}\hat{\beta}_j{}^2 \leq t$$

  which makes explicit the size constraint on the coefficients

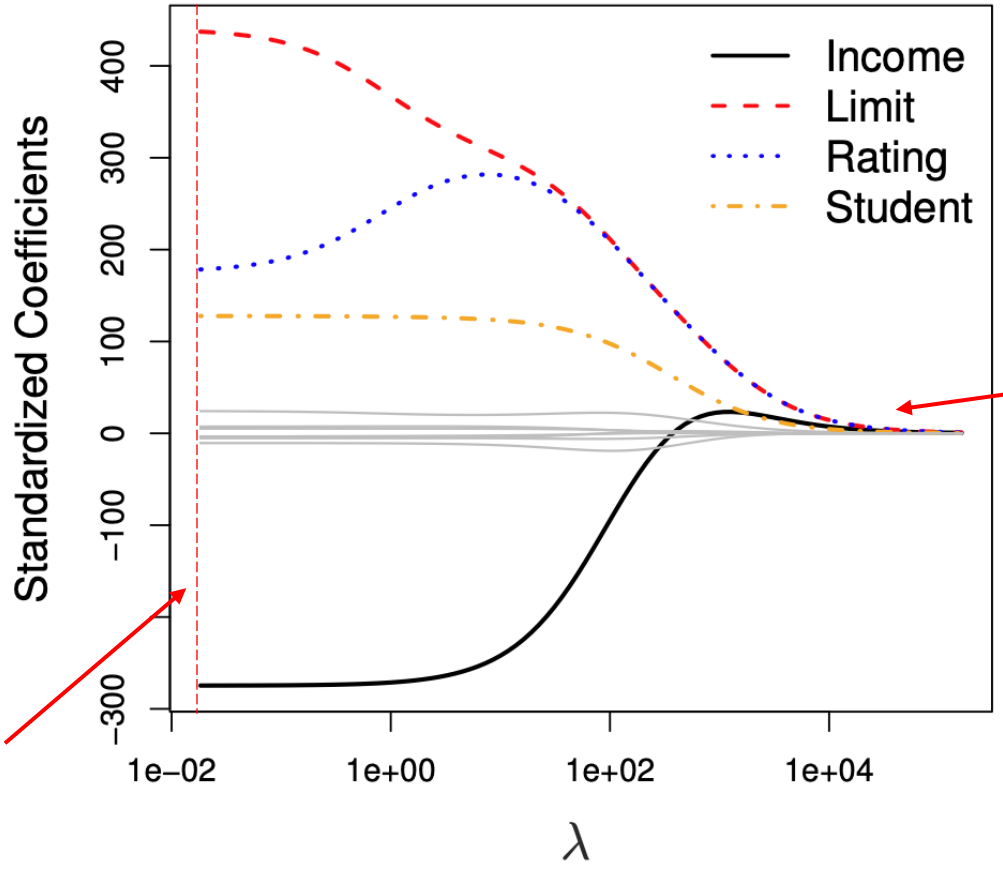
- Ridge Regression shrinks the coefficients toward zero (L2 ridge penalty) and it helps to reduce the model complexity and multi-collinearity
- When $\lambda \rightarrow 0$, the cost function becomes similar to the linear regression cost function

Example: Credit Card Debt

- Average Credit Card Debt - Balance - is measured for each individual
- We have information on several predictors:
  - age
  - number of cards
  - income
  - credit rating
  - credit limit
  - …



Ridge Regression Example: Credit Card Debt



$\lambda \approx 0$: *Ridge coefficients same as the OLS estimates*

λ *increases*: *Ridge coefficients estimates shrink towards zero*

Lasso Regression

- Lasso regression is a shrinkage method like ridge. The only difference is instead of taking the square of the coefficients, magnitudes are taken into account:

$$\sum_{i=1}^{n}(Yi - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Yi - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j X_{ij})^2 + \lambda \sum_{j=1}^{p}|\hat{\beta}_j|$$

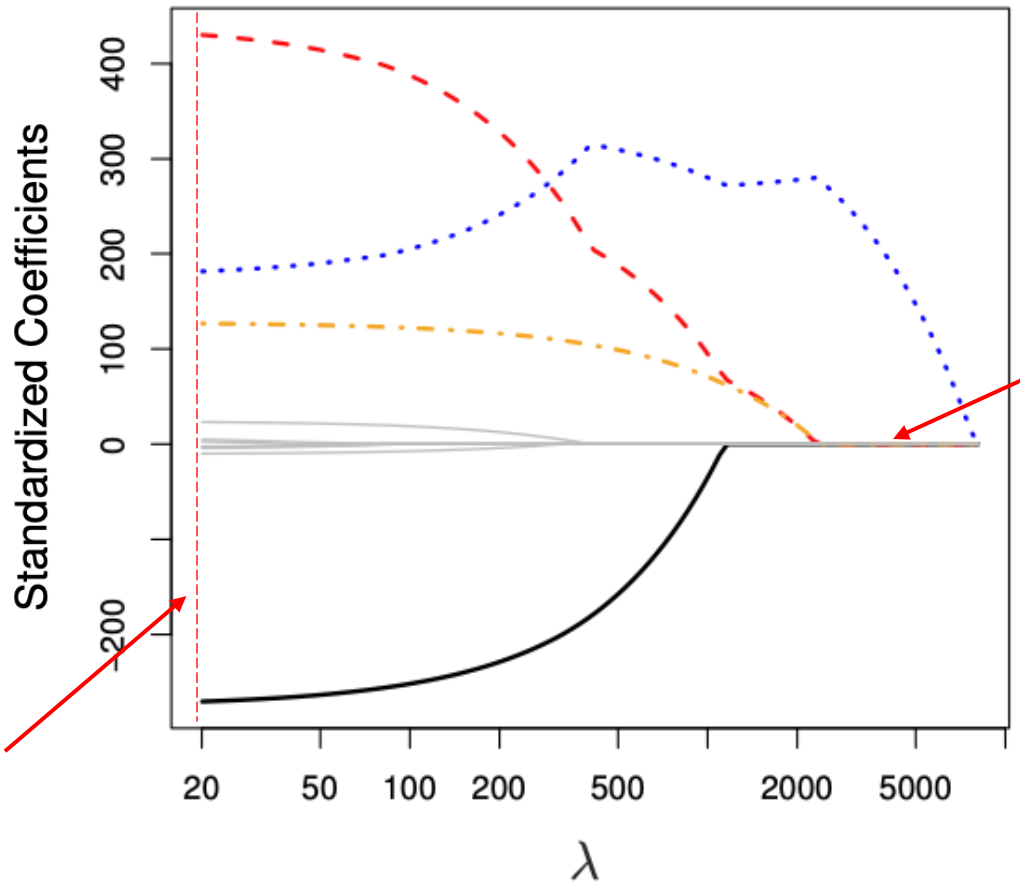
- An equivalent way to write the lasso problem is:

$$\sum_{i=1}^{n}(Yi - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j X_{ij})^2, \qquad subject\ to \sum_{j=1}^{p}|\hat{\beta}_j| \leq t$$

Lasso Regression

- This type of regularization (L1) can lead to zero coefficients i.e., some of the features are completely neglected for the evaluation of output
- Lasso Regression performs feature selection
- Like Ridge Regression, the λ parameter can be controlled
  - When $\lambda \rightarrow 0$, the cost function becomes similar to the linear regression cost function
  - When λ sufficiently large, some Lasso coefficients will equal zero

Lasso Regression Example: Credit Card Debt



λ sufficiently large: Lasso coefficients estimates equal zero

$\lambda \approx 0$: Lasso coefficients same as the OLS estimates

Evaluation Metrics

- We cannot calculate accuracy for a linear regression model. The performance of the model must be reported as an error for the predictions
- There are four error metrics that are commonly used for evaluating and reporting the performance of a linear regression model:
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)
  - Mean Absolute Error (MAE)
  - Mean Absolute Percentage Error (MAPE)

Mean Squared Error

- MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}_i)^2$$

- The difference between these two values is squared, which has the effect of removing the sign, resulting in a positive error value

- The squaring also has the effect of inflating or magnifying large errors; this has the effect of "*punishing*" models by inflating the average error score when used as a metric

Root Mean Squared Error

- RMSE is calculated as the square root of the MSE, which means that the units of the error are the same as the units of the target value that is being predicted:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}_i)^2}$$

- It may be common to use MSE loss to train a regression predictive model, and to use RMSE to evaluate and report its performance
- MSE uses the square operation to remove the sign of each error value and to punish large errors; the square root reverses this operation, although it ensures that the result remains positive

Mean Absolute Error

- MAE is calculated as the average of the absolute error values, and like RMSE, the units of the error score match the units of the target value that is being predicted:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

- Unlike the RMSE, the changes in MAE are linear and therefore intuitive; the MAE does not give more or less weight to different types of errors (like MSE and RMSE) and instead the scores increase linearly with increases in error

Mean Absolute Percentage Error

- MAPE is the percentage equivalent of MAE. Each residual $(Y_i - \widehat{Y}_i)$ is scaled against the actual value:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{\left|Y_i - \widehat{Y}_i\right|}{Y_i}$$

- MAPE has a clear interpretation since percentages are easier for people to conceptualize
- Both MAPE and MAE are robust to the effects of outliers thanks to the use of absolute value
- Since everything is scaled by the actual value, MAPE is undefined for data points where the value is 0; similarly, the MAPE can grow unexpectedly large if the actual values are exceptionally small

Evaluation Metrics

Acronym	Full Name	Residual Operation?	Robust to Outliers?
MSE	Mean Squared Error	Square	No
RMSE	Root Mean Squared Error	Square	No
MAE	Mean Absolute Error	Absolute Value	Yes
MAPE	Mean Absolute Percentage Error	Absolute Value	Yes

Classification

- The linear regression model assumes that the response variable Y is quantitative
  - But as we saw in the beginning, in many situations, the response variable is categorical, and the process of predicting categorical responses is called Classification
- There are many possible classification techniques
  - Amongst the most widely-used classifiers there is:

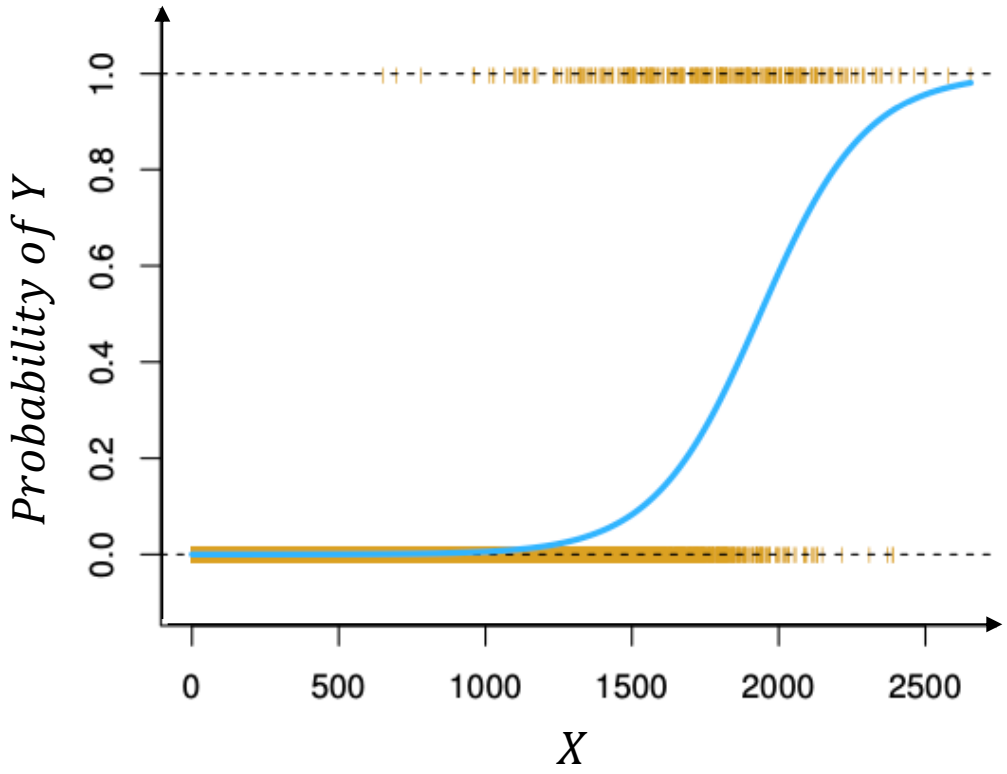
Logistic Regression

Logistic regression

- Statistical model that uses a logistic function to model a binary dependent variable
- It models the probability that Y belongs to a particular category/class
- The output of the Logistic Regression problem can be only between the 0 and 1 (unlike Linear Regression that is unbounded)

Logistic Model

- We want to model the relationship between $p(X) = \Pr(Y = 1|X)$ and X:



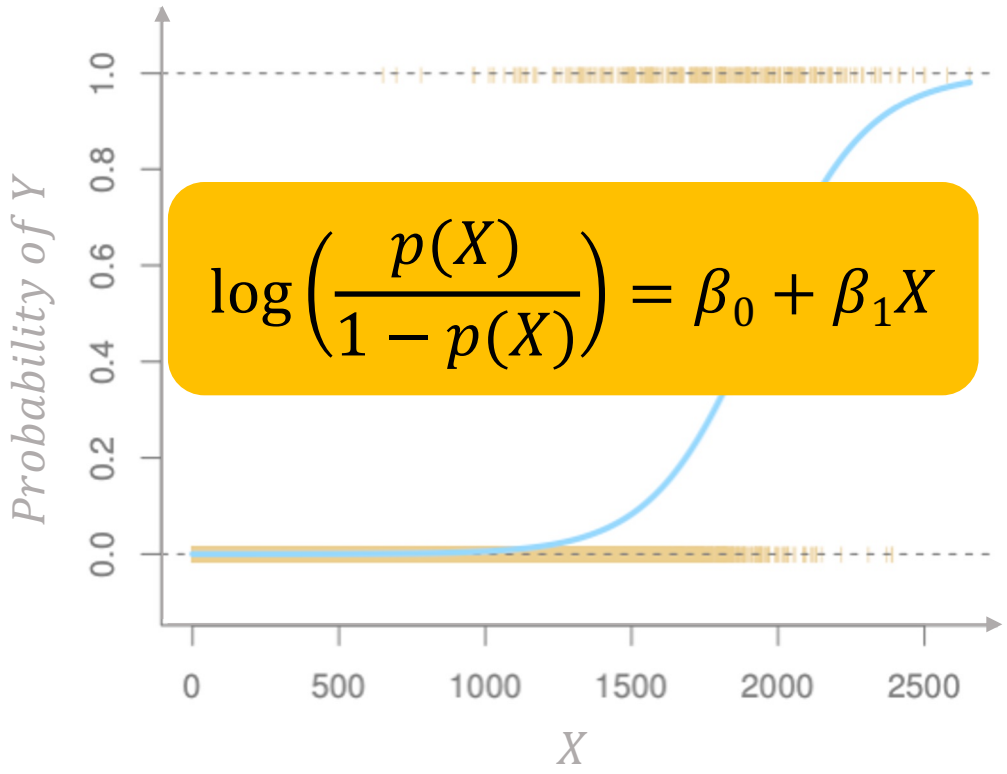
Logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

$$= Se(\beta_0 + \beta_1 X)$$

Logistic Model

- We want to model the relationship between $p(X) = \Pr(Y = 1|X)$ and X:



$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$$= Se(\beta_0 + \beta_1 X)$$

Goal

- To find the S-curve by which we can classify the data

$$\log\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- To find estimated $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the predicted probability $\hat{p}(X)$ for each Y corresponds as closely as possibly to the observed data

How do we fit the logistic model?

- We use the method called:

Maximum likelihood

- It maximizes the likelihood function:

$$\ell\left(\hat{\beta}_0, \hat{\beta}_1\right) = \prod_{i:yi=1} p(Xi) = \prod_{i':y_{i'}=0} (1 - p(X_{i'}))$$

Example: Default on credit card payment

- We are interested in predicting whether an individual will default on his or her credit card payment, on the basis of the monthly credit card balance
- $\hat{\beta}_0 = -10.6513$
- $\hat{\beta}_1 = 0.0055$
- *For* $balance = \$1,000$*:*

$$\hat{p}(X) = \frac{e^{-10.6513+0.0055*1,000}}{1+e^{-10.6513+0.0055*1,000}} = 0.00576 < 1\%$$

Types of Logistic Regression

- Binary Logistic Regression: the categorical response has only two 2 possible outcomes
  - Example: Spam or Not
- Multinomial Logistic Regression: three or more categories without ordering
  - Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)
- Ordinal Logistic Regression: three or more categories with ordering
  - Example: Movie rating from 1 to 5

TODOs

- **Reading:**
  - Main course book (ESL): Chapters 3 and 4