Regression

Dr Muhammad Atif Tahir
Professor
NUCES Fast

Regression versus Classification

- Classification: the output variable takes class labels

- Regression: the output variable takes continuous values

Examples

- Predicting House Value
  - Actual Price: £100,000
  - Predicted 1: £99,950 (Very Good Prediction)
  - Predicted 2: £50,000 (Very Bad Prediction)

- Predicting Car Premium
  - Using Location, Age, History etc

Regression Techniques

- Linear Regression
- Ridge Regression
- Lasso Regression
- And many more

Linear Regression

- Theoretically well motivated algorithm: developed from Statistical Learning Theory
- Empirically good performance: successful applications in many fields (stock prices, insurance etc)

Given examples $(x_i, y_i)_{i=1...n}$

Predict y_{n+1} given a new point x_{n+1}

Formula

$$Y = a + bX$$

where

$$b = r \frac{SDy}{SDx}$$

$$a = \overline{Y} - b\overline{X}$$

Another formula for Slope:

Slope = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

Where,

b = The slope of the regression line

a = The intercept point of the regression line and the y axis.

$\overline{X}$ = Mean of x values

$\overline{Y}$ = Mean of y values

SD_x = Standard Deviation of x

SD_y = Standard Deviation of y

Example

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

Find Y if X = 64

To Find,

Least Square Regression Line Equation

Solution :

Step 1 :

Count the number of given x values.
N = 5

Step 2 :

Find XY, X^2 for the given values.
See the below table

X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 =186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Step 3 :

Now, Find ΣX, ΣY, ΣXY, ΣX^2 for the values

$\Sigma X = 311$

$\Sigma Y = 18.6$

$\Sigma XY = 1159.7$

$\Sigma X^2 = 19359$

Step 4

Substitute the values in the above slope formula given.

$\text{Slope}(b) = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

$= ((5)*(1159.7)-(311)*(18.6))/((5)*(19359)-(311)^2)$

$= (5798.5 - 5784.6)/(96795 - 96721)$

$= 0.18783783783783292$

Step 5 :

Now, again substitute in the above intercept formula given.
Intercept(a) = (ΣY - b(ΣX)) / N
= (18.6 - 0.187837837837783292(311))/5
= -7.964

Step 6 :

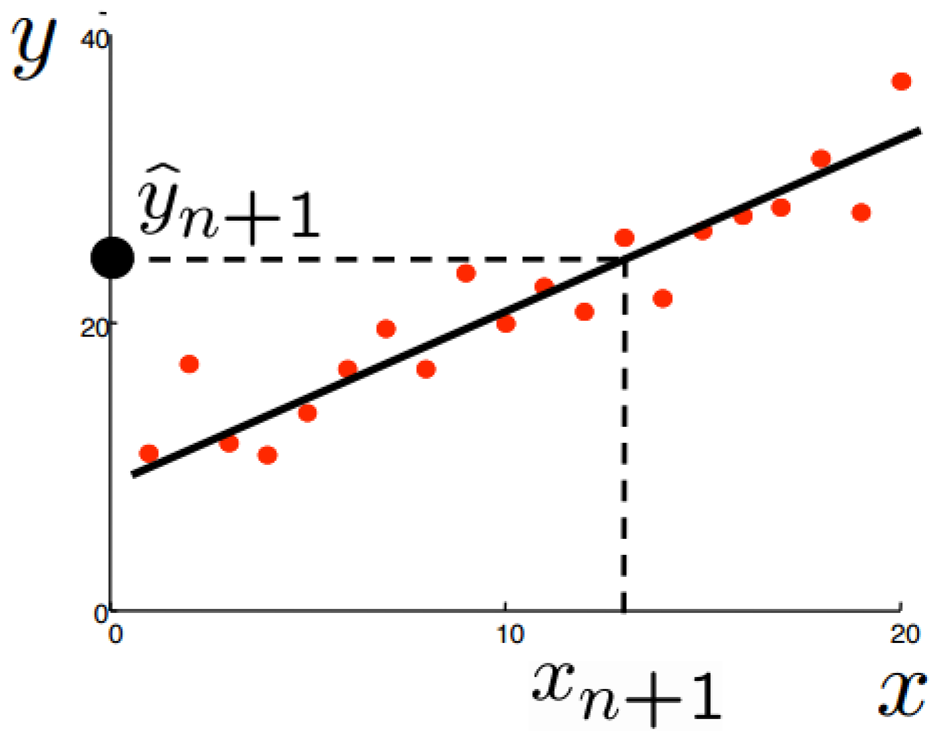
Then substitute these values in regression equation formula
Regression Equation(y) = a + bx
= -7.964 + 0.188x
Suppose if we want to calculate the approximate y value for the variable x = 64 then, we can substitute the value in the above equation
Regression Equation(y) = a + bx
= -7.964 + 0.188(64)
= 4.068

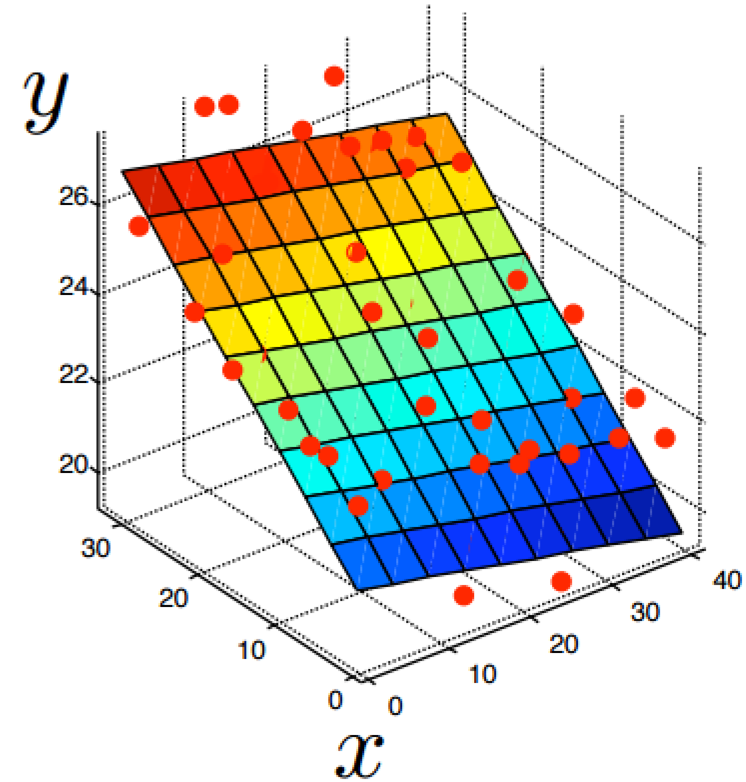
Linear regression

We wish to estimate $\hat{y}$ by a linear function of our data x:

$$
\begin{aligned}
\hat{y}_{n+1} &= w_0 + w_1 x_{n+1,1} + w_2 x_{n+1,2} \\
&= w^\top x_{n+1}
\end{aligned}
$$

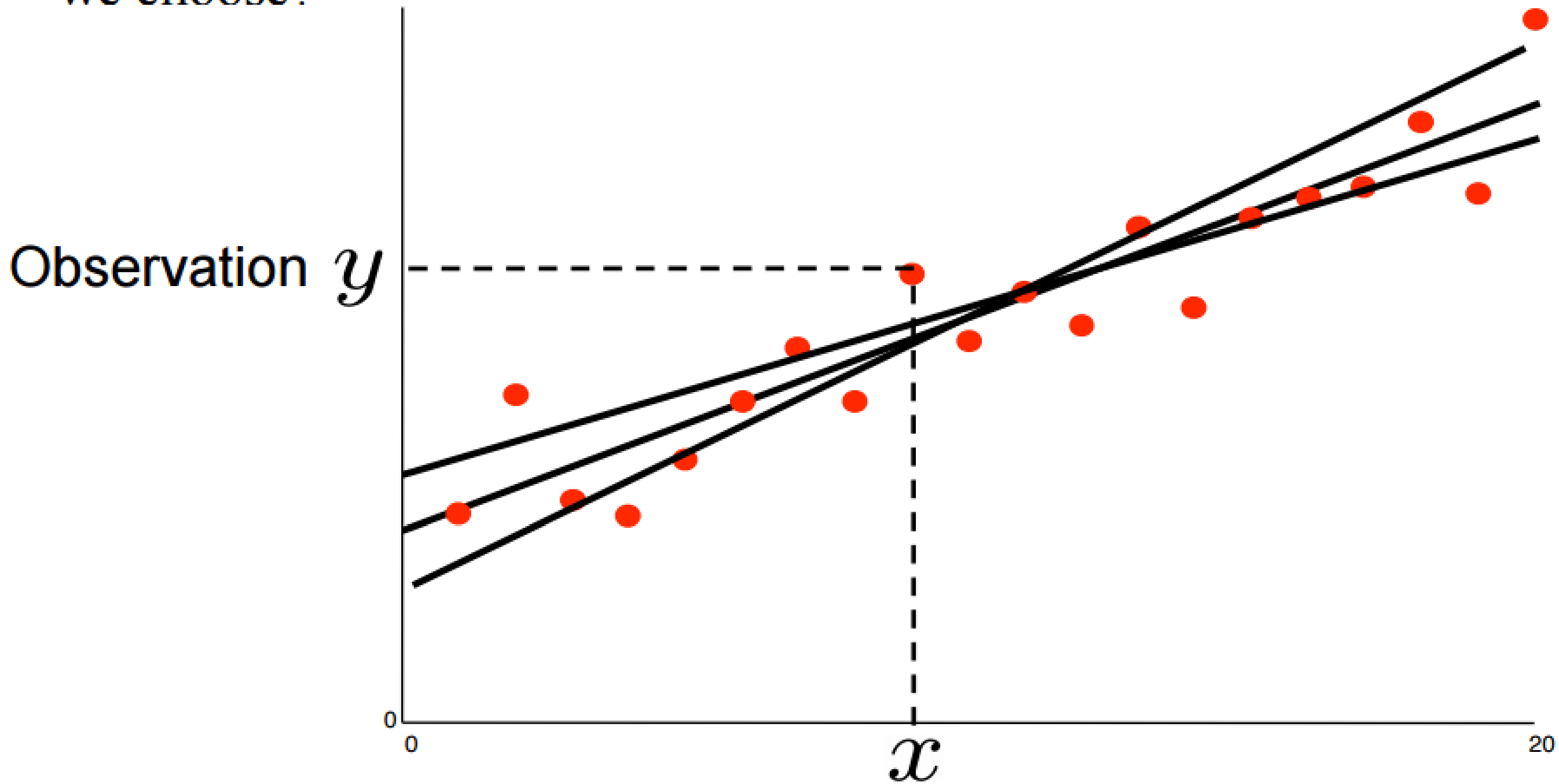
where w is a parameter to be estimated and we have used the standard convention of letting the first component of x be 1.





Choosing the regressor

Of the many regression fits that approximate the data, which should we choose?



Evaluation Measure

- Mean Squared Error

Actual (Y)	Predicted (Y')	Y'-Y	Square (Y'-Y)
41	43.6	2.6	6.76
45	44.4	-0.6	0.36
49	45.2	-3.8	14.44
47	46	-1	1
44	46.8	2.8	7.84

Sum of Error = 30.4 /5 = 6.08

Ordinary Least Square

Ordinary least squares, or linear least squares, estimates the parameters in a regression model by minimizing the sum of the squared residuals

This method draws a line through the data points that minimizes the sum of the squared differences between the observed values and the corresponding fitted values

$$\hat{\beta} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{y}$$

$\hat{\beta}$ = ordinary least squares estimator

X = matrix regressor variable X

$\top$ = matrix transpose

y = vector of the value of the response variable

Improving the Linear Model

- We may want to improve the simple linear model by replacing OLS estimation with some alternative fitting procedure.

- Why use an alternative fitting procedure?
  - Prediction Accuracy
  - Model Interpretability

Prediction Accuracy

- The OLS estimates have relatively <u>low bias</u> and <u>low variability</u> especially when the relationship between the response and predictors is linear and $n >> p$.

- If n is not much larger than p, then the OLS fit can have high variance and may result in over fitting and poor estimates on unseen observations.

- If $p > n$, then the variability of the OLS fit increases dramatically, and the variance of these estimates in infinite.

Model Interpretability

- When we have a large number of predictors in the model, there will generally be many that have little or no effect on the response.

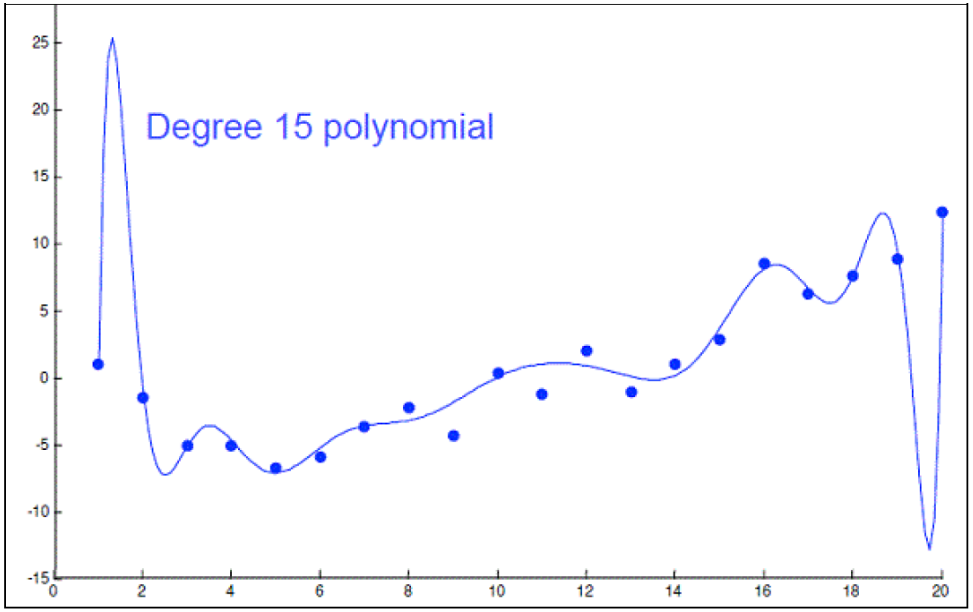
- Including such irrelevant variable leads to unnecessary complexity.

- Leaving these variables in the model makes it harder to see the effect of the important variables.

- The model would be easier to interpret by removing (i.e. setting the coefficients to zero) the unimportant variables.

Feature/Variable Selection

- Carefully selected features can improve model accuracy, but adding too many can lead to overfitting.
  - Overfitted models describe random error or noise instead of any underlying relationship.
  - They generally have poor predictive performance on test data.



- For instance, we can use a 15-degree polynomial function to fit the following data so that the fitted curve goes nicely through the data points.
- However, a brand new dataset collected from the same population may not fit this particular curve well at all.

Feature/Variable Selection (cont.)

- Subset Selection
  - Identify a subset of the *p* predictors that we believe to be related to the response; then, fit a model using OLS on the reduced set.
  - Methods: best subset selection, stepwise selection

- Shrinkage (Regularization)
  - Involves shrinking the estimated coefficients toward zero relative to the OLS estimates; has the effect of reducing variance and performs variable selection.
  - Methods: ridge regression, lasso

- Dimension Reduction
  - Involves projecting the *p* predictors into a *M*-dimensional subspace, where $M < p$, and fit the linear regression model using the *M* projections as predictors.
  - Methods: principal components regression, partial least squares

Ridge Regression

- Recall that the OLS fitting procedure estimates the beta coefficients using the values that minimize:

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

- Ridge regression is similar to OLS, except that the coefficients are estimated by minimizing a slightly different quantity:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

Ridge Regression (cont.)

- Note that $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
- The idea of penalizing by the sum-of-squares of the parameters is also used in neural networks, where it is known as *weight decay*.
- An equivalent way to write the ridge problem is:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,$$

Ridge Regression (cont.)

- The effect of this equation is to add a shrinkage penalty of the form

$$\lambda \sum_{j=1}^{p} \beta_j^2,$$

where the tuning parameter λ is a positive value.

- This has the effect of shrinking the estimated beta coefficients towards zero. It turns out that such a constraint should improve the fit, because shrinking the coefficients can significantly reduce their variance.

- Note that when λ = 0, the penalty term as no effect, and ridge regression will procedure the OLS estimates. Thus, selecting a good value for λ is critical (can use cross-validation for this).

Ridge Regression (cont.)

Computational Advantages of Ridge Regression

- If p is large, then using the best subset selection approach requires searching through enormous numbers of possible models.

- With ridge regression, for any given λ we only need to fit one model and the computations turn out to be very simple.

- Ridge regression can even be used when $p > n$, a situation where OLS fails completely (i.e. OLS estimates do not even have a unique solution).

Ridge Regression (cont.)

- In matrix form:

$$\mathrm{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta,$$

the ridge regression solutions are easily seen to be

$$\hat{\beta}^{\mathrm{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

- The solution adds a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$ before inversion (making the problem non-singular).

- The *singular value decomposition* (SVD) of the centered matrix **X** gives us some additional insight into the nature of ridge regression.

The Lasso

- One significant problem of ridge regression is that the penalty term will never force any of the coefficients to be exactly zero.

- Thus, the final model will include all *p* predictors, which creates a challenge in model interpretation

- A more modern machine learning alternative is the *lasso*.

- The lasso works in a similar way to ridge regression, except it uses a different penalty term that shrinks some of the coefficients exactly to zero.

The Lasso (cont.)

- The lasso coefficients minimize the quantity:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

- The key difference from ridge regression is that the lasso uses an ℓ_1 penalty instead of an ℓ_2, which has the effect of forcing some of the coefficients to be exactly equal to zero when the tuning parameter λ is sufficiently large.

- Thus, the lasso performs variable/feature selection.

The Lasso (cont.)

- One can show that the lasso and ridge regression coefficient estimates solves the problems:

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

and

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Lasso vs. Ridge Regression

- The lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involved only a subset of predictors.
- The lasso leads to qualitatively similar behavior to ridge regression, in that as λ increases, the variance decreases and the bias increases.
- The lasso can generate more accurate predictions compared to ridge regression.
- Cross-validation can be used in order to determine which approach is better on a particular data set.

Selecting the Tuning Parameter λ

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration in best; thus, we required a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint *s*.

- Select a grid of potential values; use cross-validation to estimate the error rate on test data (for each value of λ) and select the value that gives the smallest error rate.

- Finally, the model is re-fit using all of the variable observations and the selected value of the tuning parameter λ.

References

- https://people.eecs.berkeley.edu/~jordan/courses/294-fall09/lectures/regression/slides.pdf
- https://www.easycalculation.com/analytical/learn-least-square-regression.php

Questions!