Polynomial Regression - Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

Spring 2021

Multiple Linear Regression (MLR)

- MLR can be used to describe the relationships between a set of explanatory or independent variables $(x_1, x_2, \ldots, x_k)$ and a dependent variable (y)
- We are interested in the relationship between **each independent variable** and the dependent variable **after accounting for remaining independent variables**.
- MLR allows **quantifying the relationship** between our response variable and our explanatory variables as well as providing a tool for **predicting the response of a new observation** for a given set of values for $x_1, x_2, \ldots,$ and x_k.

Multiple Linear Regression (MLR)

The equation for the simple linear regression line is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

- y is the response or dependent variable
- $x_1, x_2, \ldots, x_k$ are the explanatory or independent variables
- β_0 is the intercept (the value of y when the $x_1, x_2, \ldots, x_k$ are set to 0)
- β_1 is the slope (the expected change in y for each one-unit change in x_1 after adjusting for $x_2, \ldots, x_k$)
- β_k is the slope (the expected change in y for each one-unit change in x_k after adjusting for $x_1, x_2, \ldots, x_{k-1}$)
- e is the random error which we assume is normally distributed with **a mean of 0 and a variance of σ^2**

Regression Equation

The equation for the multiple linear regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

where

- $\hat{y}$ (read y hat) is the expected or predicted value of y for a given values of $x_1, x_2, \ldots, x_k$
- $\hat{\beta}_0$ is the least-squares estimates of (the intercept)
- $\hat{\beta}_1, \hat{\beta}_2, \ldots$ and $\hat{\beta}_k$ are the least-squares estimates of $\beta_1, \beta_2, \ldots$ and β_k respectively

Regression Equation

The equation for the multiple linear regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

In the least-squares regression, the estimates are selected in such a way that the following quantity is **minimized**:

$$(y - \hat{y})^2 = \left[y - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k\right)\right]^2$$

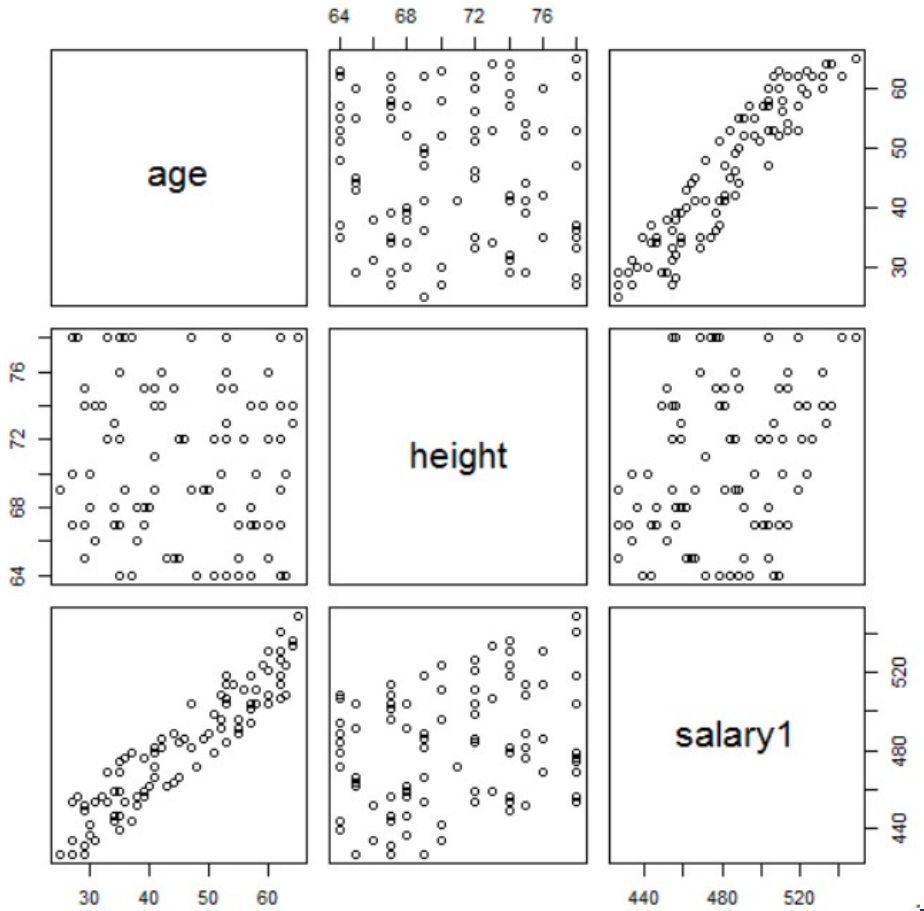
A MLR Example

In the book Blink by Malcolm Gladwell, Gladwell states that a study of CEOs of Fortune 500 companies found that these individuals tend to be taller than the average US population.

In order to study this phenomenon in more detail and to see if height is associated with increased success in business (as measured by salary), 100 men between the ages of 25 and 65 were polled for their heights (in inches) and annual salaries.

A MLR Example R Commands - Scatterplot Matrix

```
# Scatterplot Matrix
> data <- read.csv("CEO_salary.csv")
> attach(data)
> salary1 <- salary/1000
> data1 <- data.frame(age, height, salary1)
> cor(data1)
> pairs(data1)
```



Inference Global F test

- We use the ANOVA table and are interested in testing the **alternative hypothesis that at least one of the slope parameters is different than 0.**
- If this test confirms that there is at least one slope parameter that is different than 0, then **subsequent F-tests or t-tests** can be used to assess whether each individual slope parameter is different than 0.
- In MLR, the F-test for the model is referred to as the **global test.**

Inference Global F test

- We use the ANOVA table and are interested in testing the **alternative hypothesis that at least one of the slope parameters is different than 0.**
- If this test confirms that there is at least one slope parameter that is different than 0, then **subsequent F-tests or t-tests** can be used to assess whether each individual slope parameter is different than 0.
- In MLR, the F-test for the model is referred to as the **global test.**

Difference to SLR is that, here $k > 1$. The exact value of k depends on the number of variables in the model.

	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic	p-value
Regression	Reg SS	Reg df$= k$	Reg MS = Reg SS/Reg df	F=Reg MS/Res MS	$P(F_{\text{Reg df,Res df},\alpha} > F)$
Residual	Res SS	Res df$= n-k-1$	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

Inference - Anova Table Components

	SS (Sum of Squares)	df (degrees of freedom)	MS (Mean Square)	F-statistic	p-value
Regression	Reg SS	Reg df= k	Reg MS = Reg SS/Reg df	F=Reg MS/Res MS	$P(F_{\text{Reg df},\text{Res df},\alpha} > F)$
Residual	Res SS	Res df= $n-k-1$	Res MS = Res SS/Res df		
Total	Total SS = Reg SS + Res SS				

- Reg SS=$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, Res SS=$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, Total SS=$\sum_{i=1}^{n}(y_i - \bar{y})^2$
- **Reg *df* = *k* equals to the number of predictors in the model.**
- **Res *df* = *n* – *k* – 1** equals to sample size minus the number of predictors in the model minus 1.
- Reg MS = Reg SS/Reg df (the regression mean square)
- Res MS = Res SS/Res df (the residual mean square)
- F = Reg MS/Res MS
- p-value = the probability that the observed value of test statistic or a more extreme value could have been observed by chance

Global F-Test

In MLR, the first formal tests of hypotheses is for the overall model.

They test the null hypothesis that

- all slope coefficients are equal to 0
  $(H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0)$
  The null hypothesis is the same as asserting that there is **no linear relationship between the response and explanatory variables.**
- alternative that at least one of the slope coefficients is different from zero
  $(H_1 : \beta_i \neq 0$ **for at least one** $i)$.

The null hypothesis is rejected if there is at least one that is sufficiently far from 0 or (equivalently) a large majority of the total sum of squares is explained by the regression.

F-Test for MLR

$$F = \frac{\text{MS Reg}}{\text{MS Res}}$$

F-distribution with k and $n-k-1$ degrees of freedom under H_0.

The decision rule for a level α test is:

Reject $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ if $F \geq F_{k,n-k-1,\alpha}$

Otherwise, do not reject H_0

where **$F_{k,n-k-1,\alpha}$ is the value from the F-distribution** with

- k degree of freedom (numerator) and
- $n-k-1$ degrees of freedom (denominator) and
- associated with a right-hand tail probability of α.

MLR Inference t-test

If the overall model is significant, then the significance could be attributed to any one of the independent variables.

Perform testing on each individual parameter to identify the relative contribution of each independent variable.

In order to test each if $\beta_i = 0$ after controlling for the other independent variables in the model, we use a t statistic:

$$t = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

where

$SE_{\hat{\beta}_i}$ the standard error of the estimate of (in the regression model with the other independent variables included) which follows a t-distribution with n-k-1 degrees of freedom under H_0.

MLR Inference t-test

The decision rule for a two-sided level α test is:

- **Reject $H_0 : \beta_i = 0$ if $|t| \geq t_{n-k-1,\alpha/2}$**
- Otherwise, do not reject $H_0 : \beta_i = 0$

where

$$t_{n-k-1,\alpha/2}$$

is the value from the t-distribution with $n - k - 1$ degrees of freedom and associated with a right hand tail probability of $\alpha/2$.

MLR Inference t-test Confidence Interval

We can calculate the two-sided $100\%(1-\alpha)$ confidence interval for using the following formula:

$$\hat{\beta}_i \pm t_{n-k-1,\frac{\alpha}{2}} \cdot SE_{\hat{\beta}_i}$$

Non-Linear & Polynomial Regression

Multi Linear Regression assumption

- Parameters are assumed no correlation
- How to we deal with **interaction** between parameters?
  - Example: increasing spending on radio advertisement increases effectiveness of TV advertisement?
  - Example: administering combined medicine for an illness.

Extending Linear Model

- Not assuming additive assumption
- Instead assuming interactions between parameters
- Example: Drug 1 and Drug 2 impact on managing blood pressure

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Blood pressure as a function of Drug 1 (X_1) and Drug 2 (X_2)

Example of two Drugs

- Assume changing one drug changes of the other drug
- In other words, there is an interaction between parameters
- In this case we want to predict blood pressure after using both drugs

Capturing interactions

- We add a product term, which can be seen as adding one parameter as coefficient of the other parameter, like …
- Interaction can be assessed separately with its own p-value and R squared.

Interaction between Qualitative & Quantitative Parameters

- Following the same logic, multiplication of parameters will be added, as follows

- **Example**: predicting risk of heart attack based on blood pressure and race (simplifying to two class, e.g Latino vs not Latino)
  - X1 is blood pressure (quantitative)
  - X2 is race (qualitative)

$$R = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$$

$$R = \beta_0 + \beta_1 X_1 + \begin{cases} \beta_2 + \beta_3 X_1 \\ 0 \end{cases}$$

Note: qualitative parameter would change the intercept of the multi linear regression line

$$R = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$or$$

$$R = \beta_0 + \beta_1 X_1 + \begin{cases} \beta_2 \\ 0 \end{cases}$$

While with interaction element not only the intersection, but also the slope of the line changes.

$$R = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_1$$

$$Or \quad R = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2$$

$$R = \beta_0 + \beta_1 X_1 + \begin{cases} \beta_2 + \beta_3 X_1 \\ 0 \end{cases}$$

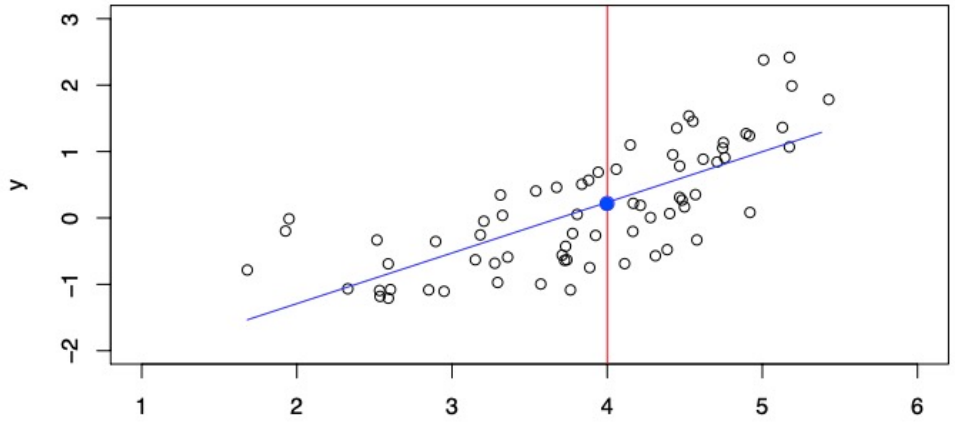
Hierarchy Principle

- What if interaction has a small p-value, but the main parameter don't?

- If we include the interaction, we have to include the main parameters as well – this is called "Hierarchy principle"
  - Intuition behind this: it will help to interpret the model. Otherwise, the interaction term will also capture the effect of the main parameter.

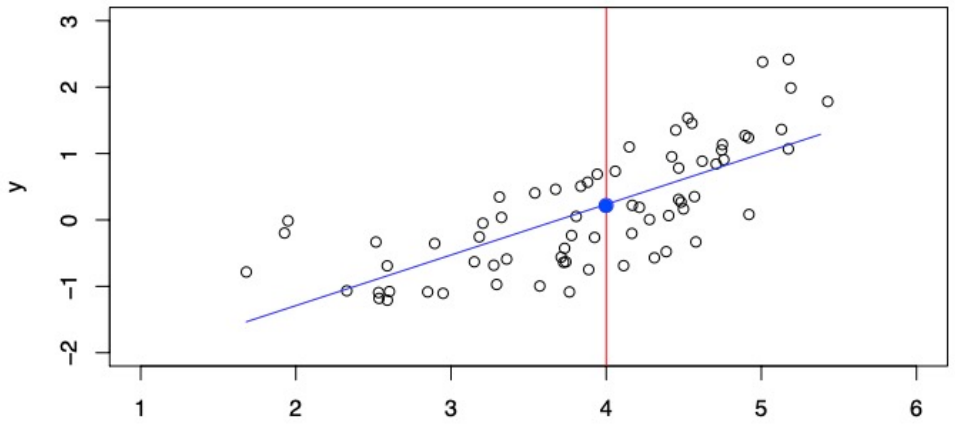
Polynomial Regression

- A linear model $f(x) = \beta_0 + \beta_1 X$ is a good fit here

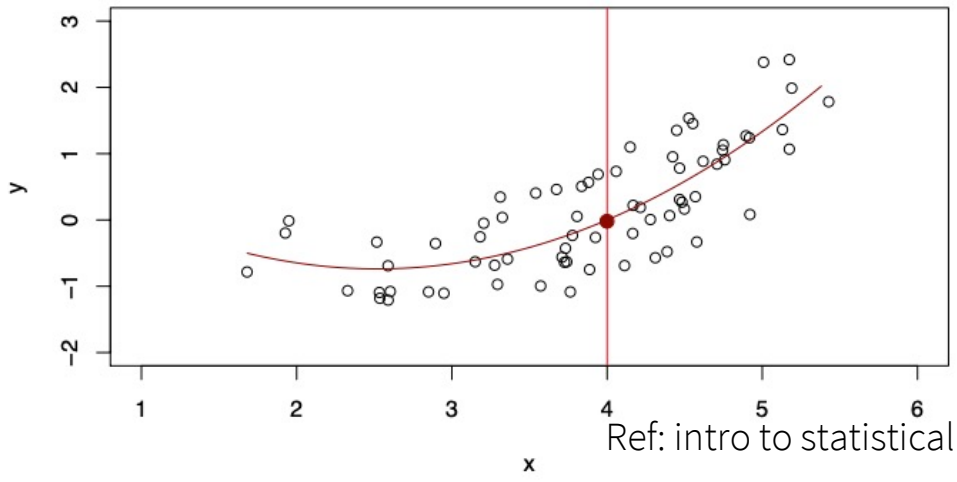


Ref: intro to statistical learning, G. James, et. al.

- A linear model $f(x) = \beta_0 + \beta_1 X$ is a good fit here



- A quadratic model $f(x) = \beta_0 + \beta_1 X + \beta_2 X^2$ fits even better



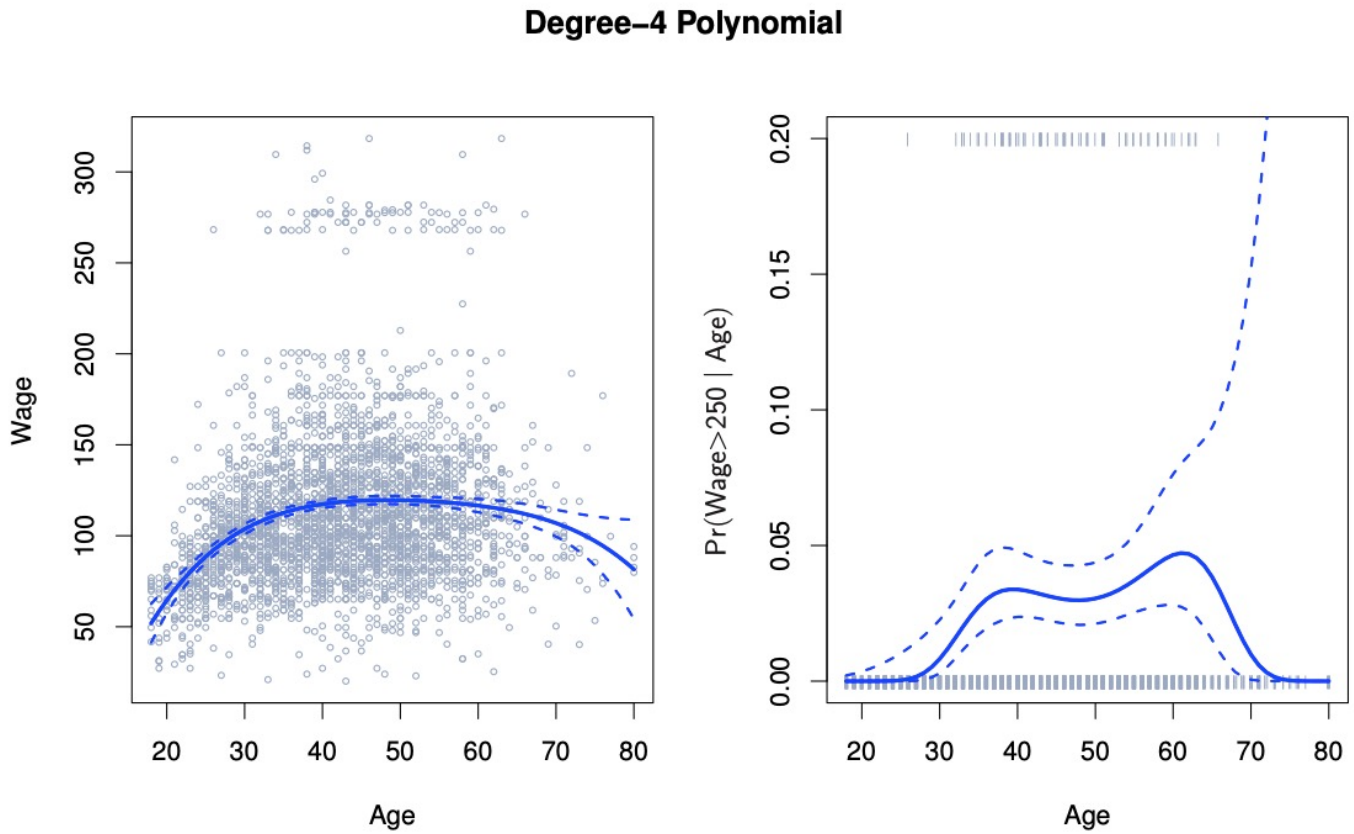
Ref: intro to statistical learning, G. James, et. al.

Why Non-Linear Model?

- Real data is never linear, but linear is a good approximation
- If not, polynomial regression is a good extension
  - Flexible
  - Easy
  - Interpretable

Example of Polynomial Regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$



Ref: intro to statistical learning, G. James, et. al.

Polynomial Regression

- Same as multi-linear regression, except higher degree of parameters also get included.

$$f(x) = \beta_0 + \beta_1 X + \cdots + \beta_n X^n$$

- Since $f(x)$ is a linear combination of β values, we can compute pointwise variance of the function – $\mathbf{Var}[f(x_0)]$ and also find confidence interval for prediction at point x_0

What Degree is good?

- What degree should be used? We mostly don't go above the power of 3.

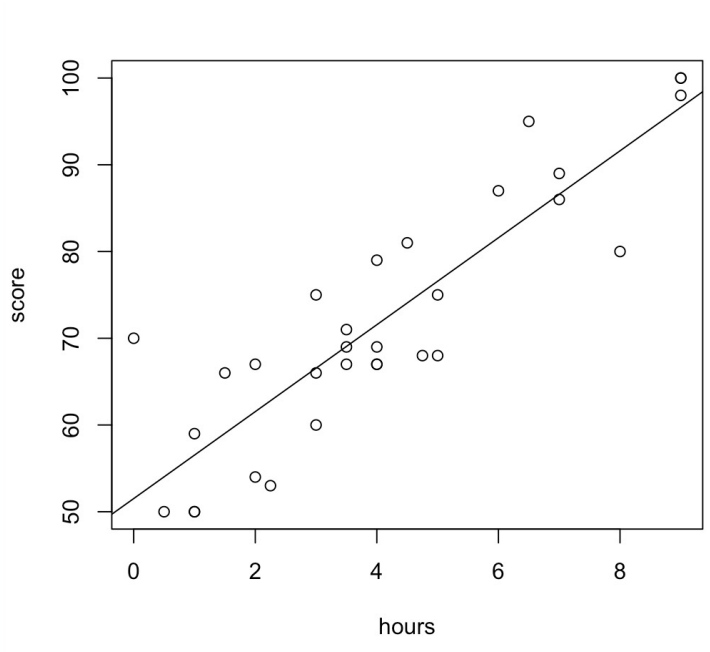
Draw backs of Polynomial Regression

- Overfitting
- Polynomial regression is very bad for extrapolation
  - Note that Polynomial regression might behave widely at the end of the range, since mostly the data is thin at the end and the model tries to match the data in other places.

Example of Overfitting

Data: number of hours studied and score of 31 students.

Linear regression estimate



	name	hours	score
1	Allen	8.0	80
2	Brown	2.0	67
3	Cole	9.0	98
4	Collins	9.0	100
5	Cooper	7.0	86
6	Cox	6.5	95

Learning From the Entire Data

- Perfect training with zero error

  $R^2 = 1!$ PERFECT!

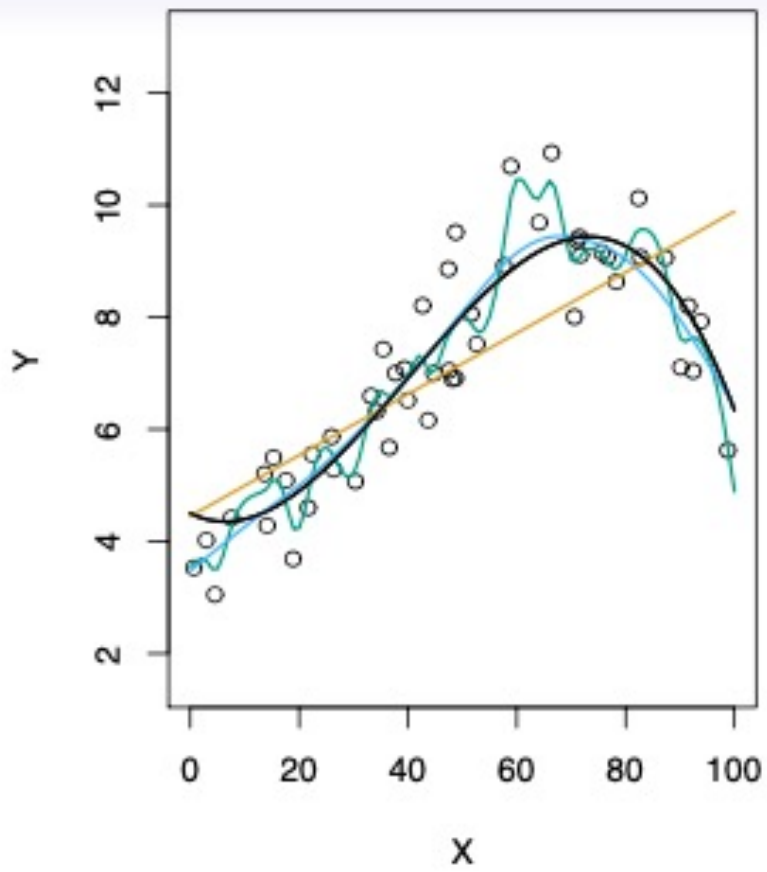
```
Residuals:
ALL 31 residuals are 0: no residual degrees of freedom!

Coefficients: (2 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)        80         NA      NA       NA
nameBrown         -13         NA      NA       NA
nameCole           18         NA      NA       NA
nameCollins        20         NA      NA       NA
nameCooper          6         NA      NA       NA
nameCox            15         NA      NA       NA
nameHall          -13         NA      NA       NA
nameHans          -21         NA      NA       NA
nameHoward          9         NA      NA       NA
nameJeffers       -10         NA      NA       NA
nameJohnson       -30         NA      NA       NA
nameJones         -14         NA      NA       NA
nameKing            1         NA      NA       NA
nameKnight        -11         NA      NA       NA
nameLee            -9         NA      NA       NA
nameMartin        -14         NA      NA       NA
nameMiller         -5         NA      NA       NA
nameMoore         -20         NA      NA       NA
nameMorris        -13         NA      NA       NA
nameMurphy          7         NA      NA       NA
nameReed           -5         NA      NA       NA
nameSmith         -30         NA      NA       NA
nameStewart       -12         NA      NA       NA
nameTaylor        -27         NA      NA       NA
nameThomas        -26         NA      NA       NA
nameThompson      -11         NA      NA       NA
nameWalker        -13         NA      NA       NA
nameWard           20         NA      NA       NA
nameWilliams      -30         NA      NA       NA
nameWright        -12         NA      NA       NA
nameYoung          -1         NA      NA       NA
hours              NA         NA      NA       NA
id                 NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
```

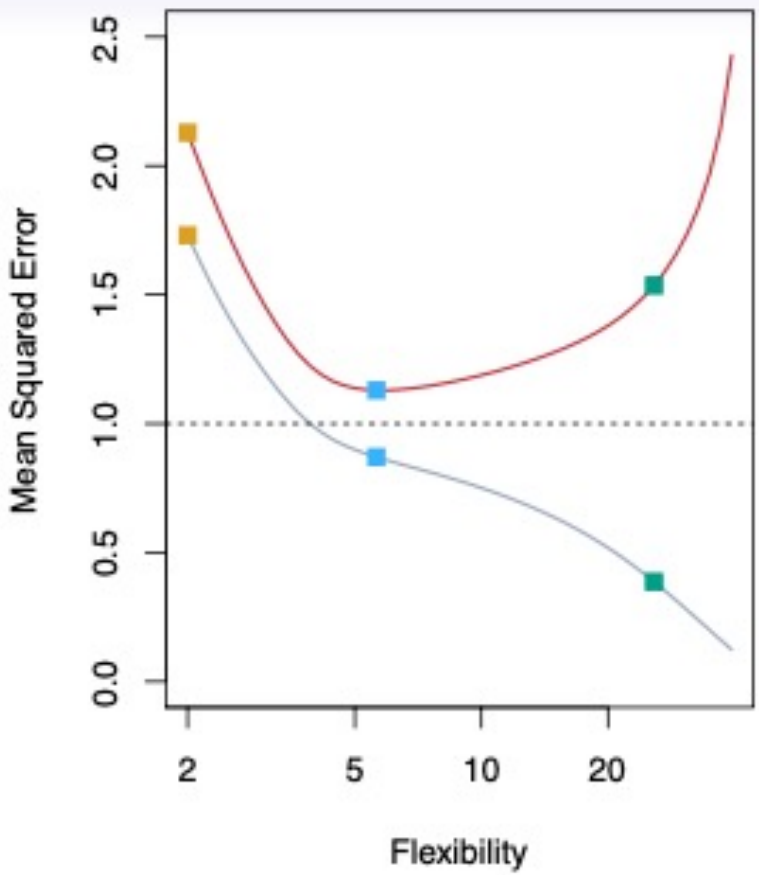
Overfitting - Training vs Test Data

- The approach minimizes MSE on a training set, e.g. with polynomial regression finds a curve that minimizes MSE.

- But the goal is having a model for the entire world!
  - Or how well the model works on a new data - "**Test Data**".

- We know - the smallest training MSE does not guarantee the smallest test data MSE!





Reference: An Intro to Statistical Learning, G. James

- Black curve – left: actual model
- Gray curve – right: training error
- Red curve – right: testing error
- Blue square: Linear & Green square: Polynomial Reg

Trade-offs in the models

- Prediction accuracy versus interpretability.
  - Linear models are easy to interpret; more complex models are not.
- Good fit versus over-fit or even under-fit.
  - How do we know when the fit is just right?
- Parsimony versus black-box.
  - We often prefer a simpler model involving fewer variables over a black-box predictor involving all of them.

Bias Variance Trade-off

- The examples of test versus training MSE's shows an important tradeoff of choosing a model.

Bias vs Variance

  - Variance is variability of the model over different training set
  - Bias is the expected difference between the model and actual value or $E[\hat{f}(x_0)] - f(x_0)$
- These are two competing choices

Bias of Learning Methods

- Bias definition - the expected difference beween the model and the actual value

$$E[\hat{f}(x_0)] - f(x_0)$$

  - E.g. linear regression assumes linear relationship between Y and X, which is mostly not correct.

The more complex a method, the more flexible and less bias, generally.

Model Variance

- Variance is variability of the model over different training set

The more complex a method, the more variance it has, generally.

- Suppose a model $\hat{f}(x)$ is fit to a training data, and let (x_0, y_0) be a test observation point.
- If the true model is $y = f(x) + \varepsilon$ with

$$f(x) = E[Y|X = x]$$

- Then the expected MSE is

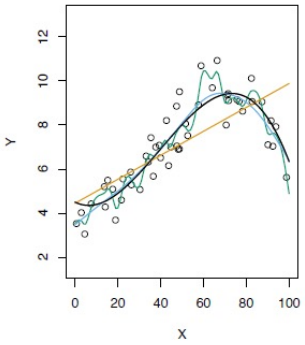
$$E\left[\left(y_0 - f(x_0)\right)^2\right] = bias^2 + var\left[\hat{f}(x_0)\right] + var(noise)$$

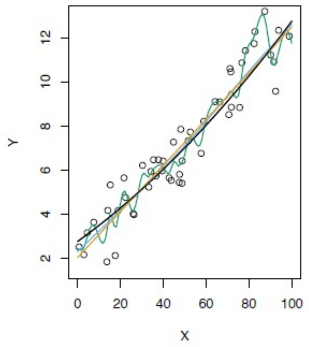
- Where

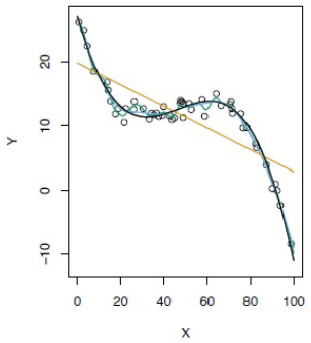
$$bias\ \hat{f}(x_0) = E[\hat{f}(x_0)] - f(x_0)$$

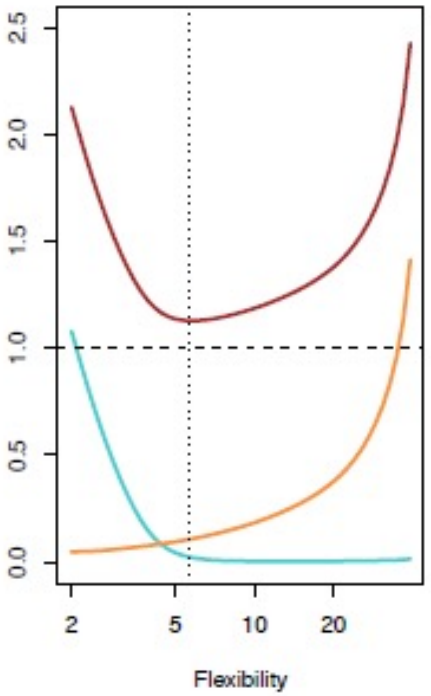
- Typically as the flexibility of the model increases, its variance increases, and its bias decreases.

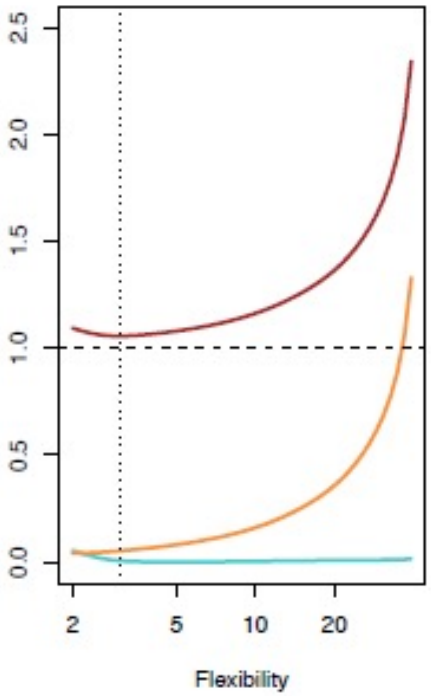
Test MSE, Bias & Variance

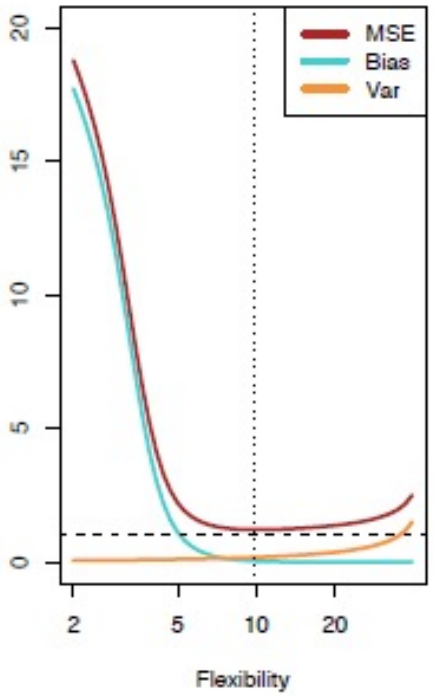












Ref: In-depth introduction to machine learning, by T. Hastie and R. Tibshirani

Classification

Classification

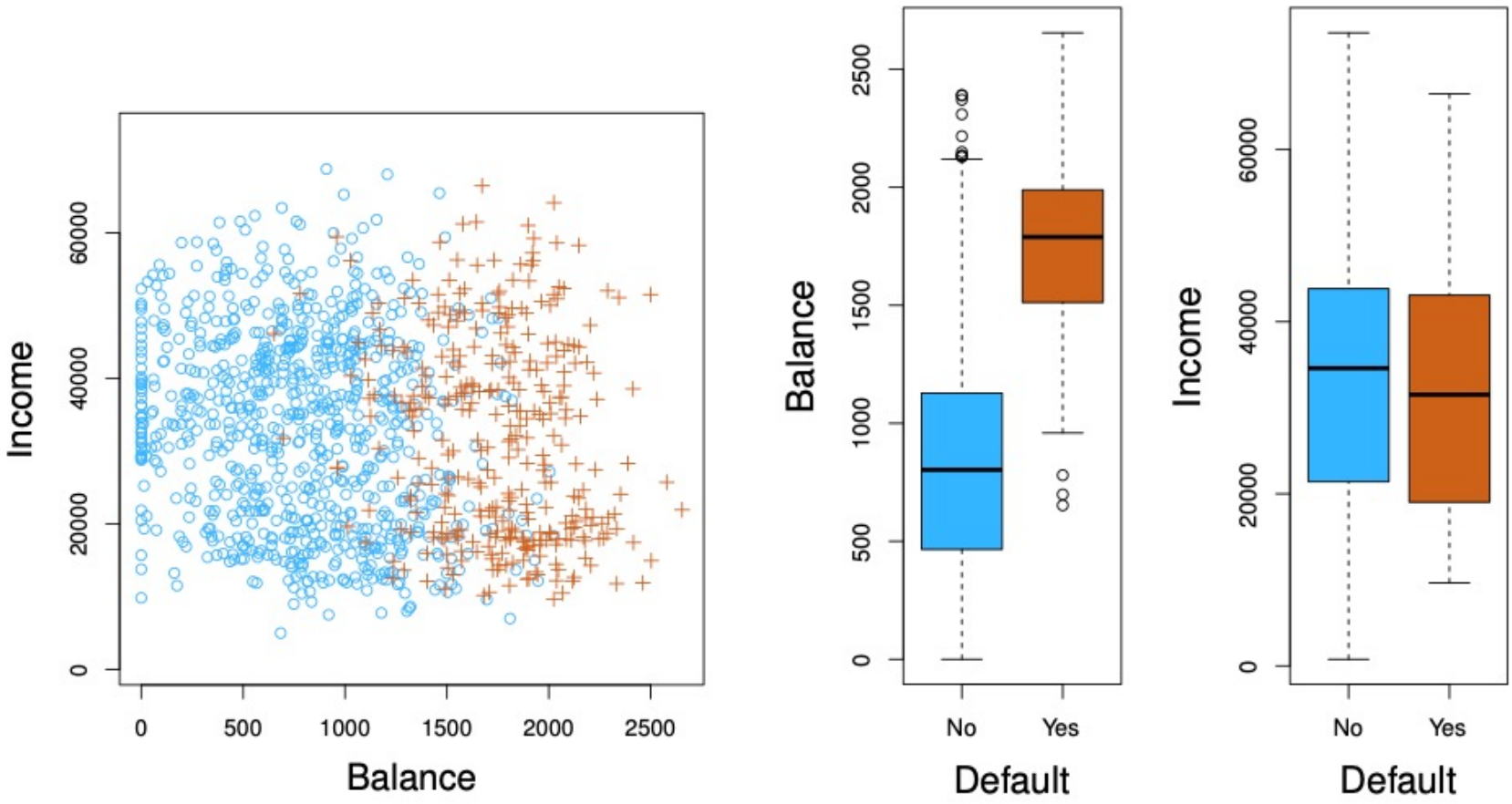
Classification is

- Predicting a qualitative output from
- A set of quantitative inputs/parameters, X
- And most of the time, also estimating quality of the estimate

For example

- Detecting spam email
- Detecting credit card fraud
- Diagnosing a patient's illness
- Object detection
- Face recognition
- News article classification
- Recommending products to customers

Example of Classification



Ref: **In-depth introduction to machine learning, by T. Hastie and R. Tibshirani**

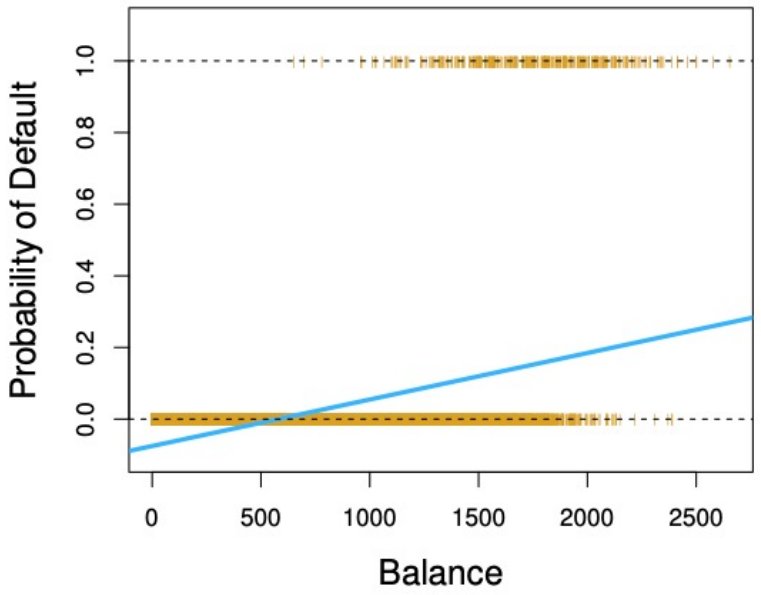
Linear Regression as classification

- Can linear regression solve classification with output

$$y = \begin{cases} 0 \\ 1 \end{cases} \ \& \ Threshold = \ 0.5$$

- Short answer is YES – linear regression does a good job, BUT
  - Linear regression can produce an output outside [0,1] !!
  - It can give a probability much higher than one for class-1
  - Also a probability much lower than zero for class-0
  - Many assumption of linear regression are not met

Linear Regression



Ref: **In-depth introduction to machine learning, by T. Hastie and R. Tibshirani**

Why Logistic Function

- The goal is
  - Finding probability of a success event, which is p (Diagnose is C).
  - Using linear combination of parameters to estimate p
- Note p changes between *[0,1]*, but linear combination of parameters change between $(-\infty, \infty)$ – Big discrepancy

- So probability of no-success is (1-p)
- The "odds ratio" is defined as $(\frac{p}{1-p})$, which mean how odd is having a success
- The odds ratio varies between $[0, \infty)$, closer to linear output
- Easy way to map a range of real positive to real numbers is *log* function
- So, we have it!

Logistic Regression

- So the output y will be found as

$$y = \frac{e^{\beta 0 + \beta 1 X 1}}{1 + e^{\beta 0 + \beta 1 X 1}}$$

- e (~ 2.71828) is constant. It is called Euler's number or "natural number"
- Note $(0 \leq y \leq 1)$

Another way of looking at logistic Function

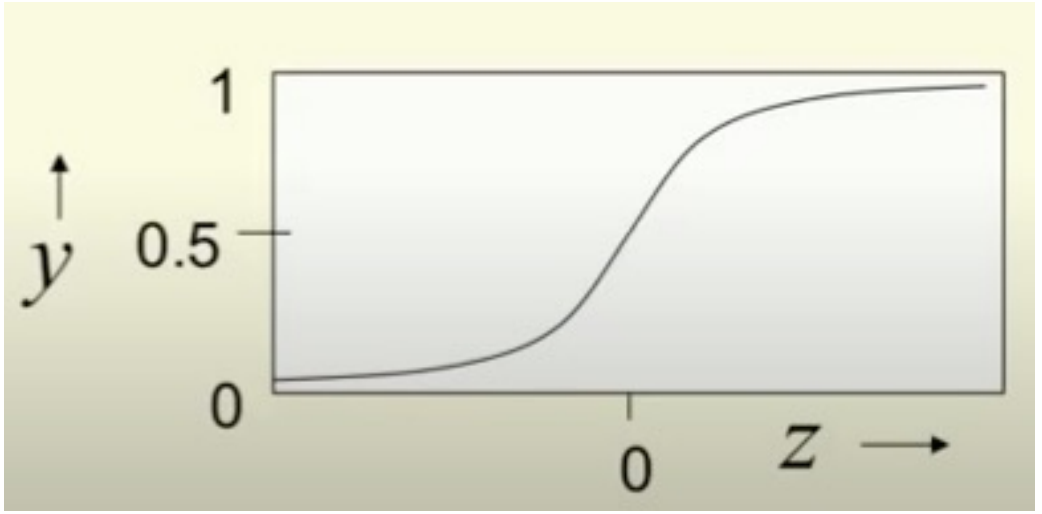
$$Z = \beta_0 + \beta_1 X_1$$

$$y = \frac{e^z}{1+ez}$$

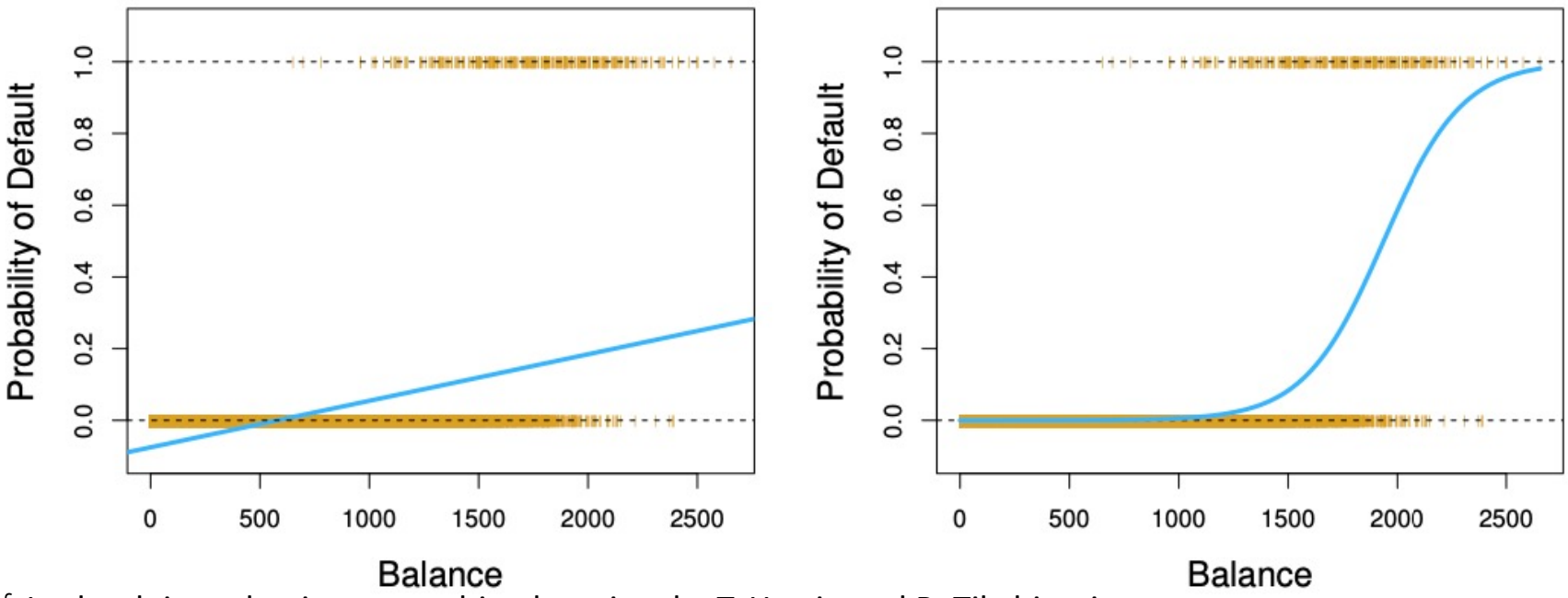
Logistic Function Is a Good Fit

- Logistic Function:

$$y = \frac{e^z}{1+ez}$$



Linear vs Logistic Regression



Ref: In-depth introduction to machine learning, by T. Hastie and R. Tibshirani

- The above example is a yes and no answers. Orange dots are marking the answers.

Example – Prediction of Default

- Predicting default as a function of balance

```
> glm(default ~ balance , data=Default , family=binomial)

Call:  glm(formula = default ~ balance, family = binomial, data = Default)

Coefficients:
(Intercept)      balance
 -10.651331     0.005499

Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
Null Deviance:      2921
Residual Deviance: 1596         AIC: 1600
```

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

Example – Prediction of Default Examples

- Probability of a person with $1000 balance default

$$y = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}} = \frac{e^{-10{\cdot}6 + 0{\cdot}0055 \times 1000}}{1 + e^{-10{\cdot}6 + 0{\cdot}0055 \times 1000}} = 0.006$$

What if balance is $2000

$$y = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}} = \frac{e^{-10{\cdot}6 + 0{\cdot}0055 \times 2000}}{1 + e^{-10{\cdot}6 + 0{\cdot}0055 \times 2000}} = 0.586$$

Multivariable Logistic Regression

- So the output y will be found as

$$y = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

Note that $(0 \leq y \leq 1)$

And

$$y = \begin{cases} 0, & Default \\ 1, & No\ default \end{cases}$$

Logistic Regression In R

- Install.packages("ISLR")
- library(ISLR)
- Function
  - glm(y ~ X, data=DataName ,family = binomial)
  - glm(y ~ X+Z, data=DataName ,family = binomial)
  - glm(y ~ ., data=DataName ,family = binomial)
- Part of the package ISLR data "Default"
  - Default as a function of Balance
  - Default as a function of Student
    - Default$studentBinFlag = ifelse(student=="Yes", 1, 0)
  - Default as a function of everything

Example –Default for Students

- What about probability of default for students

```
> glm(default ~ student , data=Default , family=binomial)

Call:  glm(formula = default ~ student, family = binomial, data = Default)

Coefficients:
(Intercept)   studentYes
    -3.5041       0.4049

Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
Null Deviance:      2921
Residual Deviance: 2909         AIC: 2913
```

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

- Predicting default as a function of “Student” being an student.

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Example – Default vs Several Variables

- Predicting default as a function of "Student", income, and balance is as follows.

```
> glm(default ~ . , data=Default , family=binomial)

Call:  glm(formula = default ~ ., family = binomial, data = Default)

Coefficients:
(Intercept)   studentYes      balance       income
 -1.087e+01   -6.468e-01    5.737e-03    3.033e-06

Degrees of Freedom: 9999 Total (i.e. Null);  9996 Residual
Null Deviance:      2921
Residual Deviance: 1572         AIC: 1580
```

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	-0.6468	0.2362	-2.74	0.0062

Multivariable Logistic Regression

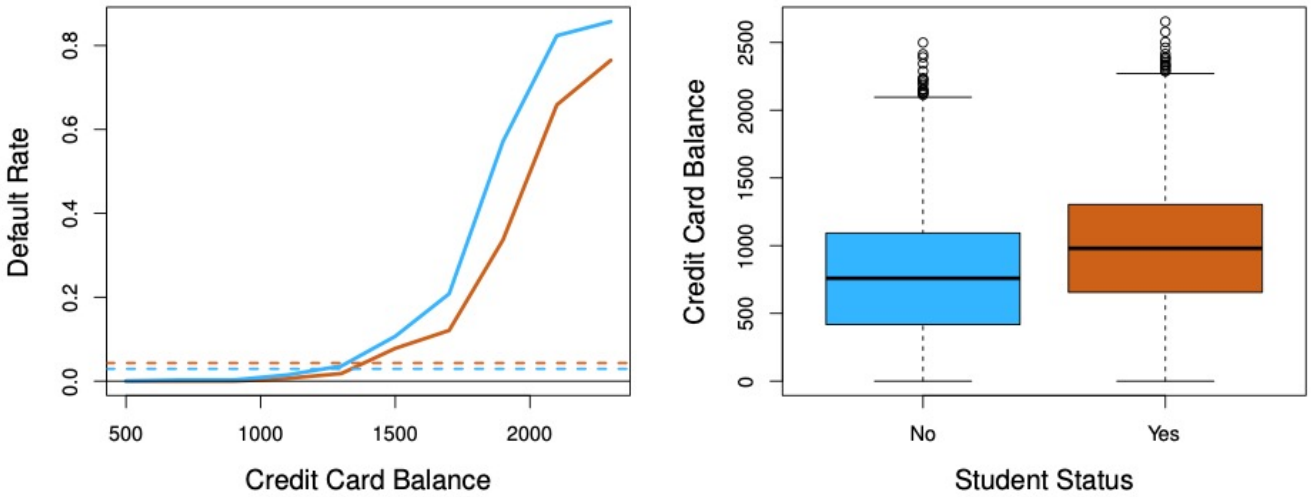
- Why student became negative here ?!

Note correlation between variables in multi variable logistic regression can make inference hard.

```
> glm(default ~ . , data=Default , family=binomial)

Call:  glm(formula = default ~ ., family = binomial, data = Default)

Coefficients:
(Intercept)   studentYes      balance       income
 -1.087e+01   -6.468e-01    5.737e-03    3.033e-06

Degrees of Freedom: 9999 Total (i.e. Null);  9996 Residual
Null Deviance:      2921
Residual Deviance: 1572         AIC: 1580
```

Logic of the Results for Students

- Students have higher balance, so it is more likely for students to default
- But for a given balance, students default is lower



Ref: **In-depth introduction to machine learning, by T. Hastie and R. Tibshirani**

Logistic Regression Function in R

- Logistic regression function in R provides
  - Coefficients of the model
  - p-value of the parameters same as linear regression
  - Also, “Z-statistics”, which is coefficient for normalized parameters

```
> tmp = glm(default ~ . , data=Default , family=binomial)
> summary(tmp)

Call:
glm(formula = default ~ ., family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression & Case Control Sampling

- When Logistic regression is used for rare events, the model will be trained for a none proportional ratio of samples
  - Like training with a set with 40% of rare event sample
- As a result the model will calculate probabilities wrong!

Solution:

- Case control sampling
  - Regression parameters β_i are accurate, and only the intercept β_0 is not, which gets corrected by

$$\beta_0{}^{*} = \beta_0 + \log\left(\frac{p_{rare}}{1 - prare}\right) - \log(\frac{p_{set}}{1 - pset})$$

- P_{rare}: actual probability of the rare event
- P_{set}: probability of the rare event in the training set

Control vs Case Sample Size

- Control to case ratio: In order to have smaller variance in the coefficients it is good to have more control samples.
- Question is how much more?
  - Rule of thumb is five to six times is sufficient

What about Multi Classification

- For example
  - Classifying news articles to sport, politics, family, kids, etc.
  - Classifying different people in a picture

Multiclass Logistic Regression or Multinomial Regression

- Logistic regression can be easily extended to more than two class prediction.

$$\Pr(y = k|X) = \frac{e^{\beta 0k + \beta 1kX1 + \cdots + \beta nkXn}}{\sum_{j=1}^{K} e^{\beta 0j + \beta 1jX1 + \cdots + \beta njXn}}$$

K : capital K is total number of classes

k: small K is one of the classes

- Select the class with the highest probability

- This is also called “softmax” function
- Note - Linear regression cannot solve this problem

Interesting Math about Logistic Regression

- Logistic function has interesting mathematical properties

1. $\ln\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 X_1 + \cdots + BnXn$

   - This is called *log odds* or *logit* transformation of y

2. $\frac{\partial y}{\partial z} = y(1-y)$ - derivative is important for numerical solutions

Important note:
Logistic Regression is not Stable for fully separable classes.
In this case, other classification methods like Discriminant analysis has to be used.