BROWN
Computer Science

CS1951A: Data Science

Lecture 10: Linear Regression

Lorenzo De Stefani

Spring 2022

Outline

- Single variable linear regression
- Minimizing least squares error
- Understanding the results
- Multiple variables regression
- Dummy variables
- Non-linear relationships
- Using statsmodel

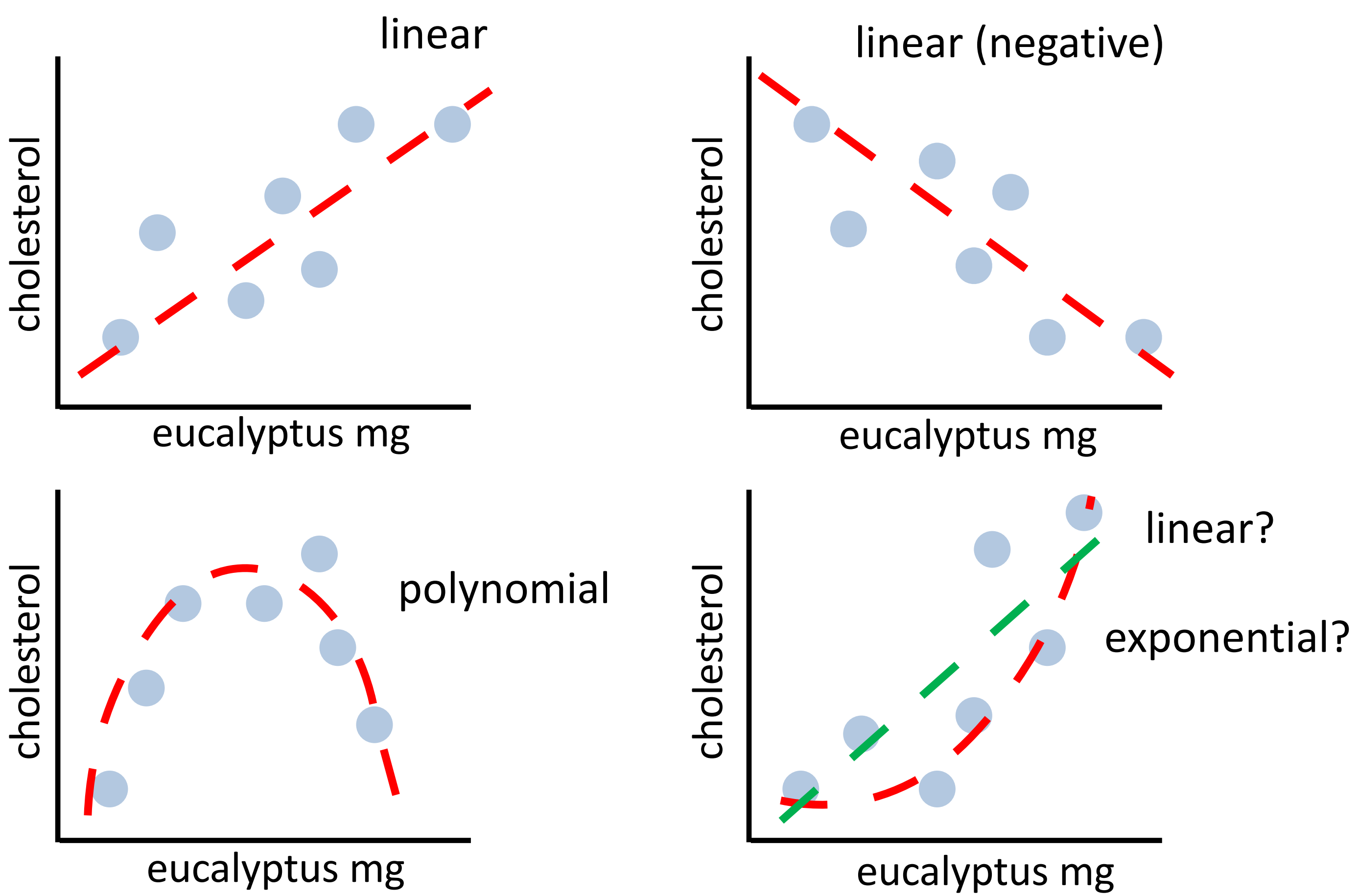
Regression

- In many cases, we are not interested just in determining if a quantity x is correlated with another y
- We would like to model the dependence between the two quantities

$$y = f(x)$$

Regression

```
cholesterol = f(mg eucalyptus oil)
```



Linear Regression

intercept expected cholesterol
when eucalyptus = 0

independent
variable
(mg eucalyptus oil)

$$y = mx + b + e$$

dependent
variable
(cholesterol)

random (hopefully) error

slope (co-efficient)
expected delta cholesterol for 1mg increase in
eucalyptus oil

Linear Regression

$$\begin{matrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{matrix} = m \begin{matrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{matrix} + b + \begin{matrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{matrix}$$

Linear Regression

$$\begin{matrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{matrix} = m \begin{matrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{matrix} + b + \begin{matrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{matrix}$$

observed values

Linear Regression

$$\begin{matrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{matrix} = m \begin{matrix} x_1 \\ x_2 \\ x_3 \\ \ldots \\ x_n \end{matrix} + b + \begin{matrix} e_1 \\ e_2 \\ e_3 \\ \ldots \\ e_n \end{matrix}$$

parameters to be estimated

Linear Regression

$$
\begin{matrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{matrix}
= \color{red}{m}
\begin{matrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{matrix}
+ \color{red}{b} +
\begin{matrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{matrix}
$$

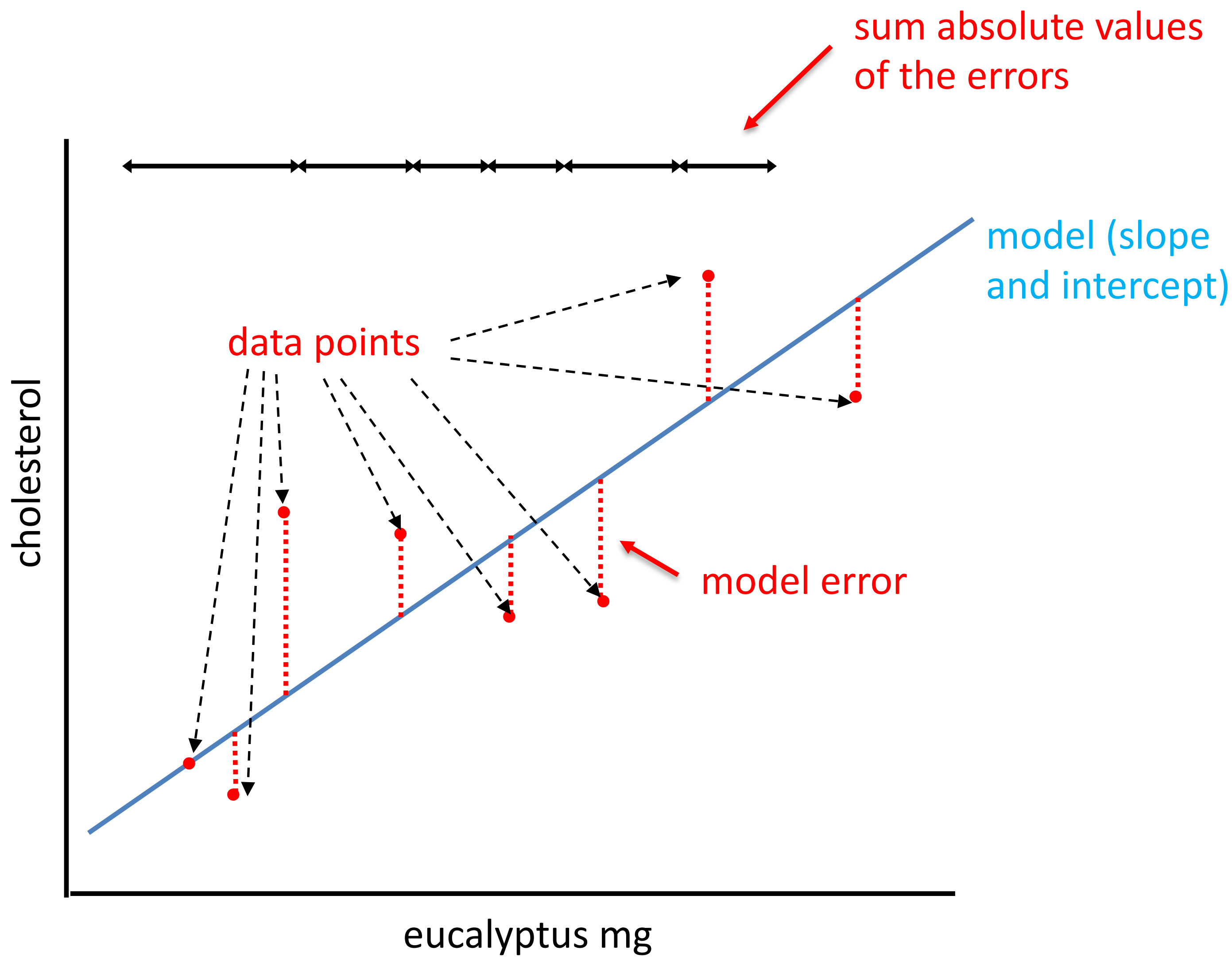
assumed to be shared
across the population

Linear Regression

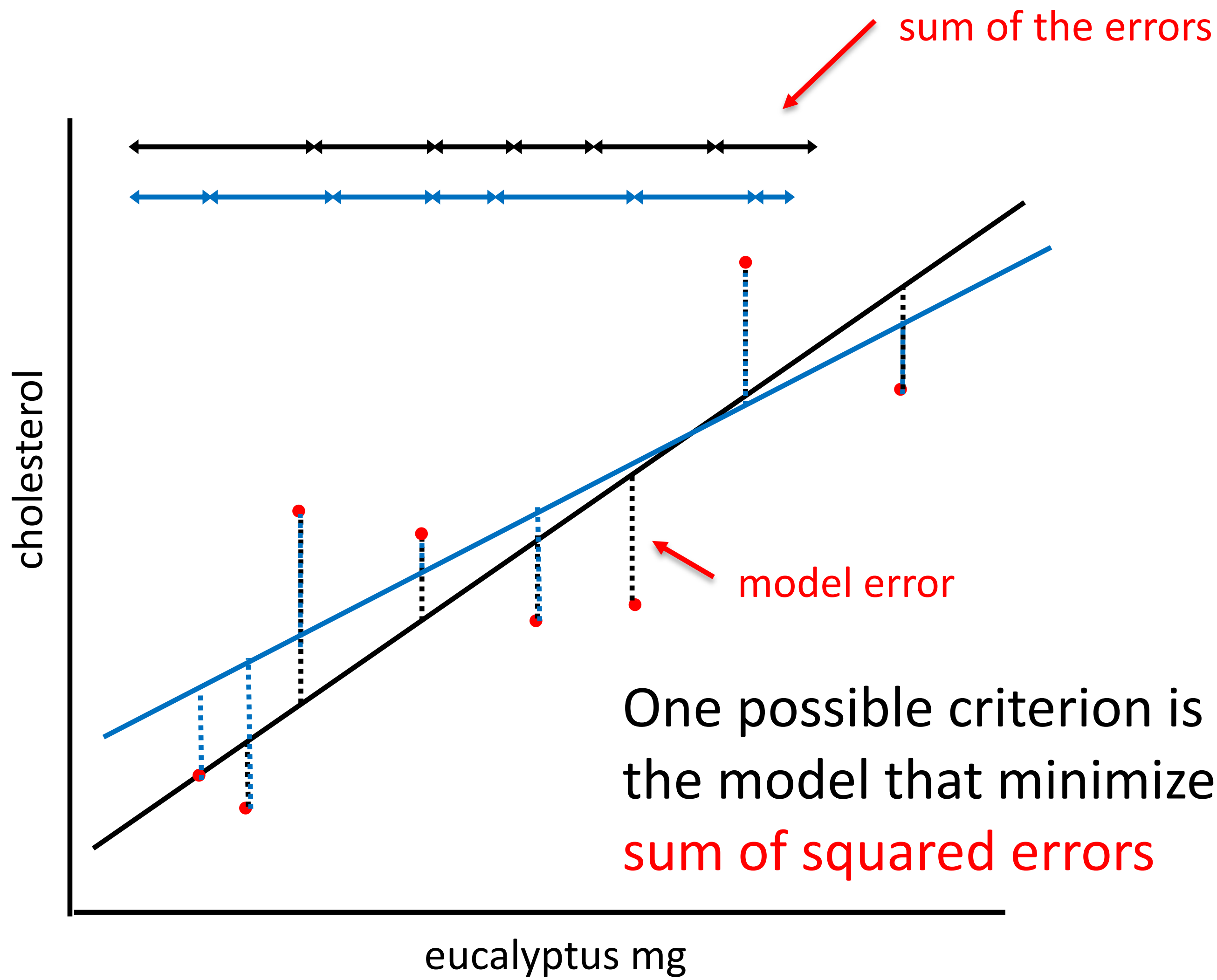
$$
\begin{matrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{matrix} = m \begin{matrix} x_1 \\ x_2 \\ x_3 \\ \ldots \\ x_n \end{matrix} + b + \begin{matrix} e_1 \\ e_2 \\ e_3 \\ \ldots \\ e_n \end{matrix}
$$

- we want to minimize the impact of the error in our model
- The more significant the impact of the error, the less accurate/useful our model is

Linear Regression



Linear Regression



One possible criterion is to pick the model that minimizes the sum of squared errors

Linear Regression with Least Squared Error

We want to find a pair (m,b) that minimizes the sum of squared errors - Linear least squares (LLS)

$$Q = \sum_{i=1}^{n} (Y_i - (mX_i + b))^2$$

Linear Regression

$$Q = \sum_{i=1}^{n} (Y_i - (mX_i + b))^2$$

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2(Y_i - mX_i - b) = 0$$

First: Derivative in b
Find value b^* which minimizes Q

$$\frac{\partial Q}{\partial m} = \sum_{i=1}^{n} -2X_i(Y_i - b^* - mX_i) = 0$$

Then: Derivative in m
Substiute b $= b^*$ and find value of the slope which minimizes Q

Linear Regression

The slope and intercept m^*, b^* minimizing the sum of squared errors can be computed exactly form data

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Sample averages
- Estimates of $E[X], E[Y]$

$$\widetilde{Var}[\bar{X}] = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

- Empirical variance of $\bar{X}$

$$\widetilde{Cov}(\bar{X}, \bar{Y}) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

- Empirical covariance of $\bar{X}$ and $\bar{Y}$

$$m^* = \frac{\tilde{Cov}[\bar{X}, \bar{Y}]}{\tilde{Var}\bar{X}} \qquad b^* = \bar{Y} - m^* \bar{X}$$

Linear Regression

The slope and intercept $\mathrm{m}^*, \mathrm{b}^*$ minimizing the sum of squared errors can be computed exactly form data

$$m^* = \frac{\tilde{Cov}[\bar{X}, \bar{Y}]}{\tilde{Var}\bar{X}} \qquad b^* = \bar{Y} - m^*\bar{X}$$

- $\widetilde{Var}[\bar{X}]$ the empirical variance accounts for variation in Y which can be attributed to variation in the values of X itself
- $\widetilde{Cov}[\bar{X}, \bar{Y}]$ captures the relative change of Y given X
- The intercept is often not as significant as the slope but still necessary for the model

Covariance

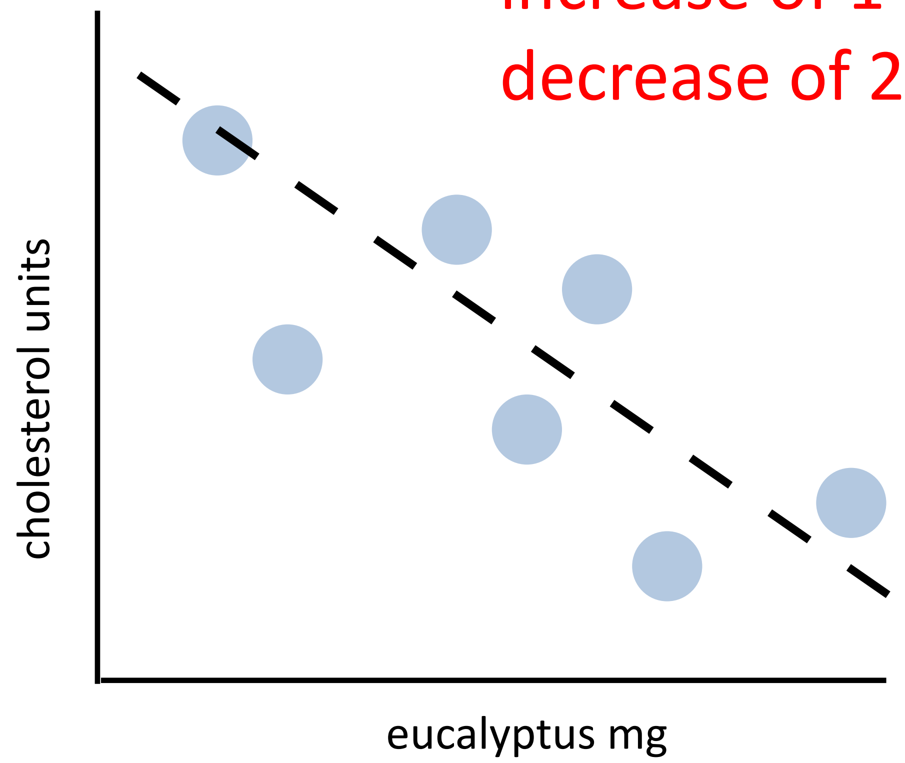
$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

- The covariance is a measure of the joint variability of two random variables
- The sign of the covariance shows the tendency in the linear relationship between the variables
  - If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values the covariance is positive
  - In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other the covariance is negative
  - If the two variables are independent then $Cov(X,Y) = 0$

Linear Regression

```
cholesterol = m(eucalyptus) + b
m = -2.4
```

increase of 1 mg eucalyptus oil -> decrease of 2.4 cholesterol units



Discussion + Question time!

What can we say of the observed relationship between use of eucalyptus oil and cholesterol levels?

(a) There probably actually is a relationship. Linear regression is a legitimate method, so we should trust the result.
(b) There is probably no actual relationship. We are confusing correlation with causation.
(c) There is probably no actual relationship. We are measuring eucalyptus oil in the wrong units, so it just appears correlated.
(d) There is probably no actual relationship. We are failing to capture other relevant variables.

Omitted Variable Bias

- By construction, we assume that the dependent variable can be predicted from the explanatory variables only
- We assume changes in the dependent variable that are correlated with the explanatory variable are because of the explanatory variable
- We assume that changes in the dependent variable that are not explained by the explanatory variables is "noise"

Multiple Linear Regression

```
Y = m1X1 + m2X2 + m3X3 + m4X4
```

```
Y: cholesterol level
X1: eucalyptus mg consumed
X2: cholesterol meds consumed
X3: had breakfast? (Y/N)
X4: constant term
```

Multiple Linear Regression

```
Y = m1X1 + m2X2 + m3X3 + m4X4
```

```
Y: cholesterol level
X1: eucalyptus mg consumed
X2: cholesterol meds consumed
X3: had breakfast? (Y/N)
X4: constant term
```

"intercept"

think of it at as a "correction"
independent of any other variable

Multiple Linear Regression

```
Y = m1X1 + m2X2 + m3X3 + m4X4
```

```
Y: cholesterol level
X1: eucalyptus mg consumed
X2: cholesterol meds consumed
X3: had breakfast? (Y/N)
X4: constant term
```

slopes / coefficients / effects

capture the dependence between observed variables

Multiple Linear Regression

```
Y = m1X1 + m2X2 + m3X3 + m4X4
```

$$Q = \sum_{i=1}^{n} (Y_i - (m_1 X_{1i} + m_2 X_{2i} + m_3 X_{3i} + m_4 X_{4i}))^2$$

$$\frac{\partial Q}{\partial m_1} = f(X_1, X_2, X_3, X_4, Y)$$

Multiple Linear Regression

Y = m1X1 + m2X2 + m3X3 + m4X4

$$Q = \sum_{i=1}^{n} (Y_i - (m_1 X_{1i} + m_2 X_{2i} + m_3 X_{3i} + m_4 X_{4i}))^2$$

$$\frac{\partial Q}{\partial m_1} = f(X_1, X_2, X_3, X_4, Y)$$

Variation on Y may depend on other explanatory variables

Multiple Linear Regression

Y = m1X1 + m2X2 + m3X3 + m4X4

$$Q = \sum_{i=1}^{n} (Y_i - (m_1 X_{1i} + m_2 X_{2i} + m_3 X_{3i} + m_4 X_{4i}))^2$$

$$\frac{\partial Q}{\partial m_1} = f(X_1, X_2, X_3, X_4, Y)$$

change in cholesterol (Y) associated with an increase of eucalyptus oil (X1), holding other variables constant

Multiple Linear Regression

```
Y = m1X1 + m2X2 + m3X3 + m4X4
```

$$\mathbf{Y} = \mathbf{X}\beta$$

Matrices of observations

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Vector of coefficients

Multiple Linear Regression

Y = m1X1 + m2X2 + m3X3 + m4X4

$$\mathbf{Y} = \mathbf{X}\beta$$

Matrices of observations

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Vector of coefficients

$$X = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad X' = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

Matrix transposition

Multiple Linear Regression

Y = m1X1 + m2X2 + m3X3 + m4X4

$$\mathbf{Y} = \mathbf{X}\beta$$

Matrices of observations

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Vector of coefficients

$$\mathbf{X} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad X^{-1} = \frac{1}{Det[X]}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Matrix inversion

Multiple Linear Regression

Y = m1X1 + m2X2 + m3X3 + m4X4

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 2 & 6 \\ 4 & 9 & 12 \end{bmatrix}$$

The values of tow vectors of observed variables could be linearly dependent/collinear

- I.e., one is the scaled version of the other
- But then matrix **X** would not be full rank
  - Determinant is 0
  - Not invertible

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

If a matrix **A** cannot be inverted, we can use a generalization known as the "Pseudoinverse" (Moore–Penrose inverse) $\mathbf{A}^{+}$

- Always exists!
- Computation can be tricky unless the matrix has linearly independent columns in which case

$$\mathbf{A}^{+} = (\mathbf{A}^{*}\mathbf{A})^{-1}\mathbf{A}^{*}$$

Multiple Linear Regression

$$Y = m1X1 + m2X2 + m3X3 + m4X4$$

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 2 & 6 \\ 4 & 9 & 12 \end{bmatrix}$$

The values of tow vectors of observed variables could be linearly dependent/collinear

- I.e., one is the scaled version of the other
- But then matrix **X** would not be full rank
  - Determinant is 0
  - Not invertible

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Given a $n \times m$ matrix **A**, its conjugate transpose (or Hermitian transponse) **A*** is an $\mathrm{m} \times n$ matrix obtained by first taking the transpose $\mathbf{A}^T$ of **A**, and then the complex conjugate of each entry

- If $\mathbf{A}^T$ is real-valued, then $\mathbf{A}^T = \mathbf{A}^*$

$$\mathbf{A}^{+} = (\mathbf{A}^{*}\mathbf{A})^{-1}\,\mathbf{A}^{*}$$

Dummy Variables

```
Y = m1X1 + m2X2 + m3X3 + m4X4
```

```
Y: cholesterol level
X1: eucalyptus
X2: cholesterol meds
X3: had breakfast
X4: constant term
```

A dummy variable is a numeric variable that represents categorical data ,(i.e., membership in a class/category)

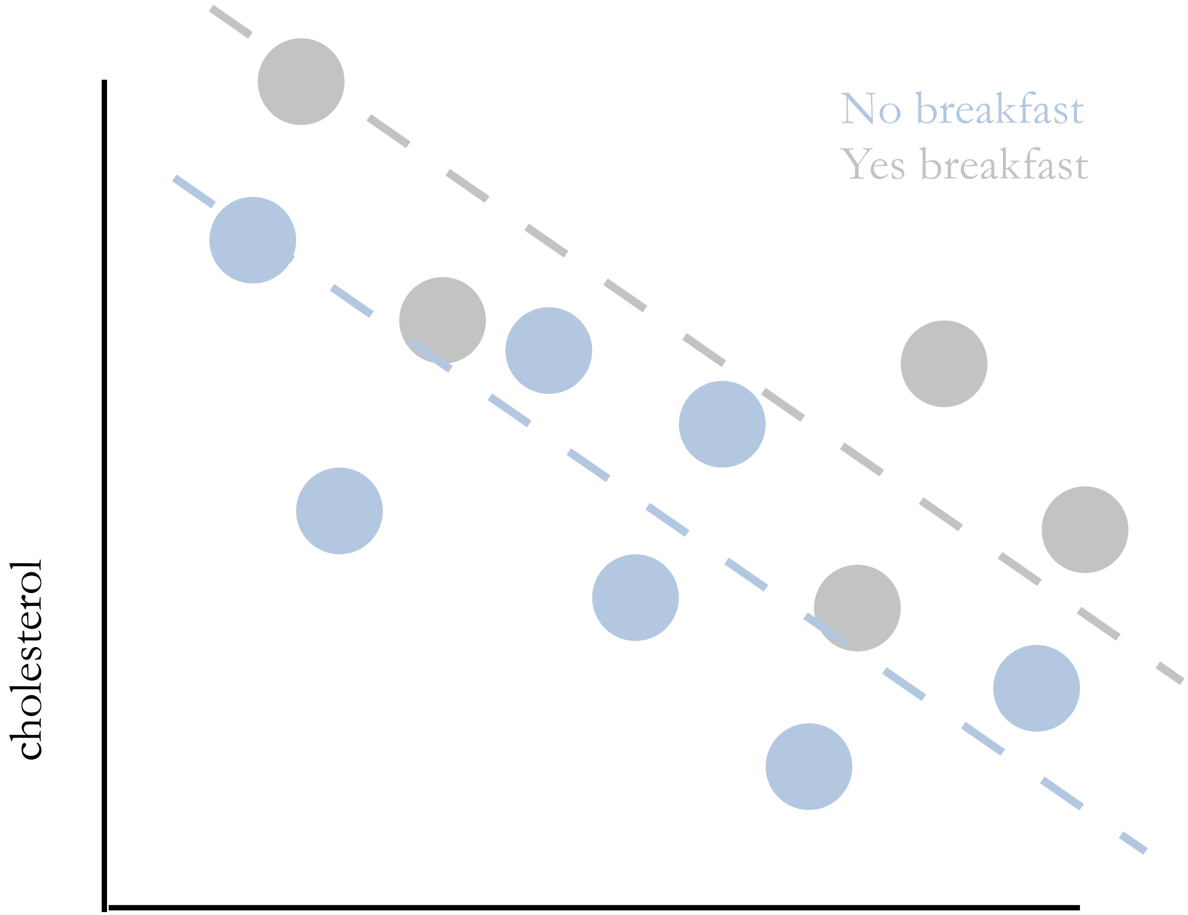
- Indicator variables, Boolean variables, one-hot variables, sparse variables
- E.g., Binary classes, race, political affiliation, etc.

Dummy Variables

- Technically, dummy variables are dichotomous, quantitative variables.
  - Their range of values is small
    - they can take on only two quantitative values.
  - As a practical matter, regression results are easiest to interpret when dummy variables are limited to two specific values, 1 or 0.
    - Typically, 1 represents the presence of a qualitative attribute, and 0 represents the absence.

Dummy Variables

Interpretable as shift in intercept for different groups



How Many Dummy Variables?

- The number of dummy variables required to represent a particular categorical variable depends on the number of values that the categorical variable can assume.
- To represent a categorical variable that can assume *k* different values we would need to define *k - 1* dummy variables.

How Many Dummy Variables?

- Example: we are interested in political affiliation, a categorical variable that might assume three values: {Republican, Democrat, Independent}
- How many dummy variables do we need?

Dummy Variables

cholesterol meds

yes breakfast

constant

$$X = \begin{bmatrix} 20 & 31 & 0 & 1 & 1 \\ 20 & 5 & 0 & 1 & 1 \\ 20 & 40 & 0 & 1 & 1 \\ 25 & 18 & 1 & 0 & 1 \end{bmatrix}$$

eucalyptus

no breakfast

Any problem??

The Dummy variable trap

When defining dummy variables, a common mistake is to define too many variables:

- If a categorical variable can take on *k* values, you only need *k - 1* dummy variables.
- A k^{th} dummy variable is redundant; it carries no new information.
- It creates a severe multicollinearity problem for the analysis.

Question Time!

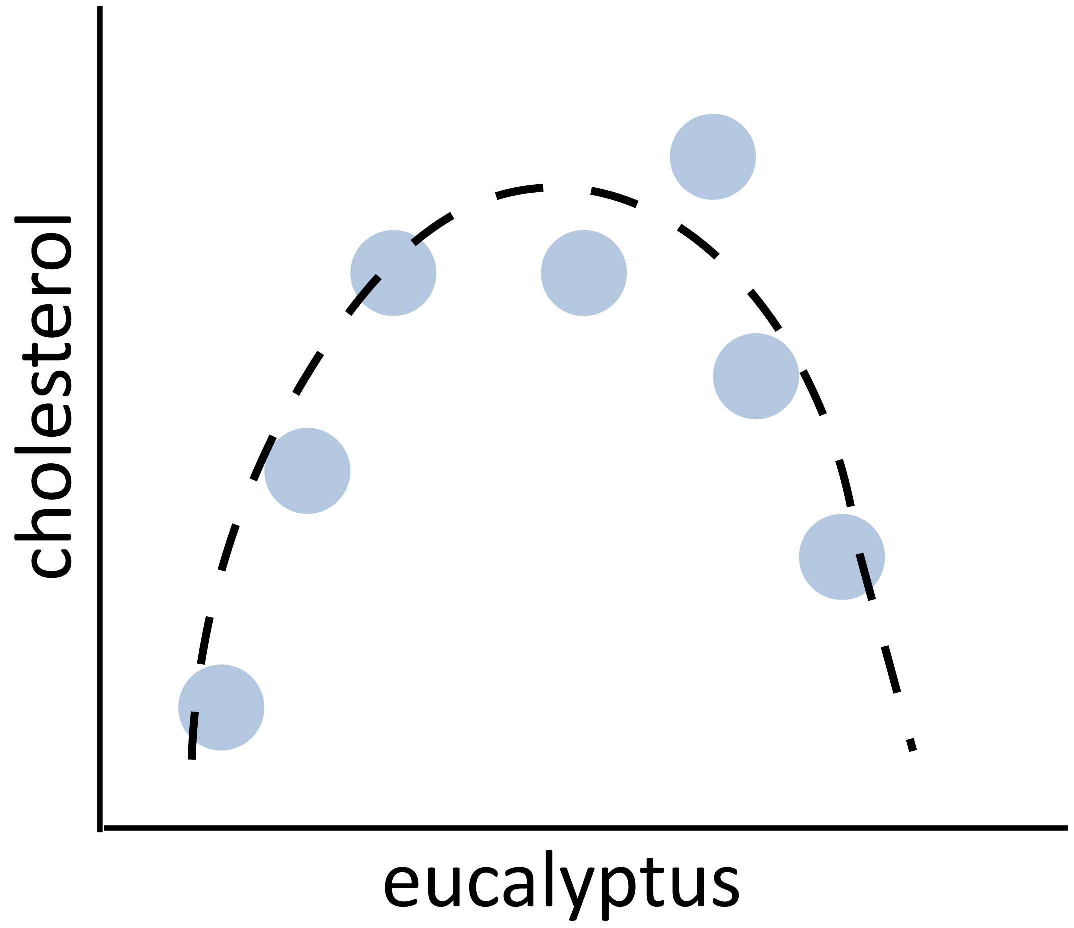
For the below model, how many parameters (coefficients) do we need to estimate?

```
Y = m1X1 + m2X2 + m3X3 + m4X4 +m5X5
```

```
Y: happiness
X1: day of week (dummies M,T,W,Th,F,S,Su)
X2: bank account balance (real value)
X3: breakfast (dummies yes,no)
X4: whether you have found your inner peace
(dummies yes,no,unclear)
```

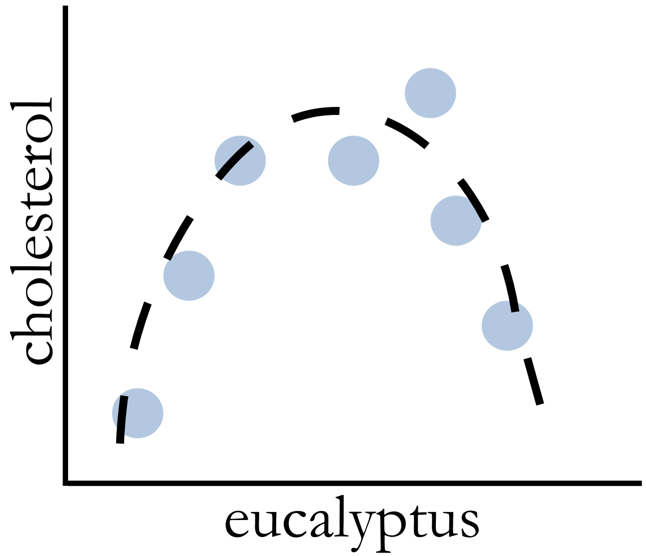
(a)5
(b)10
(c)11
(d)infinite

Nonlinear Relationships



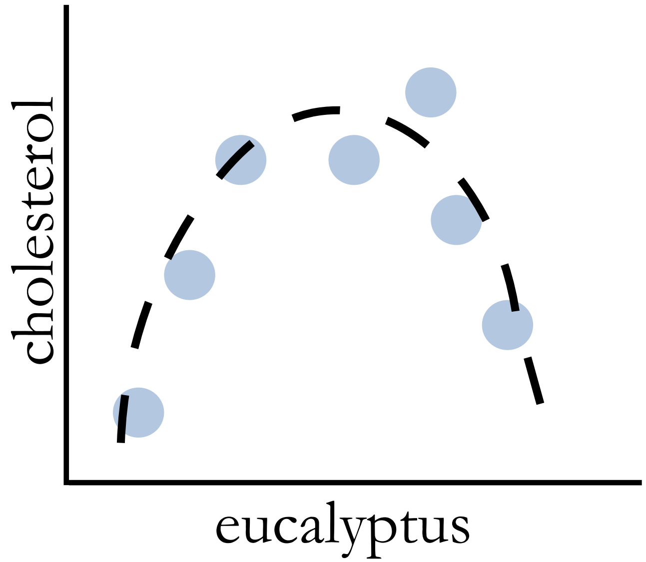
Nonlinear Relationships

Can we model this with linear regression?




```
Y = m1X1 + m2X2 + m3X3
    Y: cholesterol
    X1: eucalyptus
    X2: eucalyptus²
```

Nonlinear Relationships - Interactions




```
Y = m1X1 + m2X2 + m3X3 + m4X4
  Y: cholesterol
  X1: eucalyptus
  X2: cholesterol meds
  X3: X1 x X2
```

variable that models the interaction (composition) of observables

statsmodels

```
import statsmodels.api as sm

y, X = read_data()
X = sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit()
print(results.summary())
```

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html
https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

statsmodels

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
# M has column headers w/ names
M = read_data()
X = sm.add_constant(X)
eq = “chol ~ eucalyptus + meds + breakfast”
model = smf.ols(formula=eq, data=M)
results = model.fit()
print(results.summary())
```

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html
https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

statsmodels

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
# M has column headers w/ names
M = read_data()
X = sm.add_constant(X)
eq = "chol ~ eucalyptus + meds + breakfast
+ eucalyptus:meds"
model = smf.ols(formula=eq, data=M)
results = model.fit()
print(results.summary())
```

interaction term

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html
https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

statsmodels

```
import statsmodels.api as sm
import statsmodels.formula.api as smf
# M has column headers w/ names
M = read_data()
X = sm.add_constant(X)
eq = "chol ~ eucalyptus + meds + breakfast
+ eucalyptus^2"
model = smf.ols(formula=eq, data=M)
results = model.fit()
print(results.summary())
```

squared terms

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html
https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

statsmodels

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 4.020e+06
Date:                Tue, 26 Feb 2019   Prob (F-statistic):          2.83e-239
Time:                        04:42:47   Log-Likelihood:                -146.51
No. Observations:                 100   AIC:                             299.0
Df Residuals:                      97   BIC:                             306.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3423      0.313      4.292      0.000       0.722       1.963
x1            -0.0402      0.145     -0.278      0.781      -0.327       0.247
x2            10.0103      0.014    715.745      0.000       9.982      10.038
==============================================================================
Omnibus:                        2.042   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.360   Jarque-Bera (JB):                1.875
Skew:                           0.234   Prob(JB):                        0.392
Kurtosis:                       2.519   Cond. No.                         144.
==============================================================================
```

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html
https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

statsmodels

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                                                              4.020e+06
Date:                                                                2.83e-239
Time:                                                                  -146.51
No. Observat                                                             299.0
Df Residuals                                                             306.8
Df Model:
Covariance T
============                                                 ================
                                                             025       0.975]
------------                                                 ----------------
const            1.3423      0.313      4.292      0.000      0.722      1.963
x1              -0.0402      0.145     -0.278      0.781     -0.327      0.247
x2              10.0103      0.014    715.745      0.000      9.982     10.038
==============================================================================
Omnibus:                        2.042   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.360   Jarque-Bera (JB):                1.875
Skew:                           0.234   Prob(JB):                        0.392
Kurtosis:                       2.519   Cond. No.                         144.
==============================================================================
```

overall fit of model (SSE)

it signifies the “percentage variation in dependent that is explained by independent variables”. Here, 73.2% variation in y is explained by X1, X2, X3, X4 and X5.

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html
https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

statsmodels

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 4.020e+06
Date:                Tue, 26 Feb 2019   Prob (F-statistic):          2.83e-239
Time:                        04:42:47   Log-Likelihood:                -146.51
No. Observations:                 100   AIC:                             299.0
                                   97   BIC:                             306.8
                                    2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3423      0.313      4.292      0.000       0.722       1.963
x1            -0.0402      0.145     -0.278      0.781      -0.327       0.247
x2            10.0103      0.014    715.745      0.000       9.982      10.038
==============================================================================
Omnibus:                        2.042   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.360   Jarque-Bera (JB):                1.875
Skew:                           0.234   Prob(JB):                        0.392
Kurtosis:                       2.519   Cond. No.                         144.
==============================================================================
```

slopes/coefficients

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html

https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

statsmodels

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                                                              4.020e+06
Date:                                                                2.83e-239
Time:                                                                  -146.51
No. Observ                                                               299.0
Df Residua                                                               306.8
Df Model:
Covariance
==========                                                         ===========
                                                                        0.975]
----------                                                         -----------
const                                                                    1.963
x1                                                                       0.247
x2                                                                      10.038
==============================================================================
Omnibus:                        2.042   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.360   Jarque-Bera (JB):                1.875
Skew:                           0.234   Prob(JB):                        0.392
Kurtosis:                       2.519   Cond. No.                         144.
==============================================================================
```

This tells the overall significance of the regression. This is to assess the significance level of all the variables together unlike the t-statistic that measures it for individual variables. The null hypothesis under this is “all the regression coefficients are equal to zeros”

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html
https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

statsmodels

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 4.020e+06
                                                 tic):               2.83e-239
                                                 :                     -146.51
                                                                         299.0
                                                                         306.8

                                                 =========================
                                                    [0.025      0.975]
------------------------------------------------------------------------------
const          1.3423      0.313      4.292      0.000       0.722       1.963
x1            -0.0402      0.145     -0.278      0.781      -0.327       0.247
x2            10.0103      0.014    715.745      0.000       9.982      10.038
==============================================================================
Omnibus:                        2.042   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.360   Jarque-Bera (JB):                1.875
Skew:                           0.234   Prob(JB):                        0.392
Kurtosis:                       2.519   Cond. No.                         144.
==============================================================================
```

t-statistic and p-values
each computed for the null hypothesis
“the value of the i-th coefficient is equal to zero”

https://www.statsmodels.org/dev/examples/notebooks/generated/ols.html
https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html