More Linear Regression

March 7, 2019
Data Science CSCI 1951A
Brown University
Instructor: Ellie Pavlick
HTAs: Wennie Zhang, Maulik Dang, Gurnaaz Kaur

Announcements

- Blog post and MR due soon (3/15)
- Stats assignment coming out today
- Anything? Questions? Comments? Concerns?

Today

- Finish last lecture—derivation of slope/intercept

- Linear Regression (Part 2) — interpretation, controls, dummy variables

Background

Sample Average: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Sample Variance: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Sample Co-Variance: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$

Linear Regression

$$Q = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

Linear Regression

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

Linear Regression

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

intercept at minimum

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

Linear Regression

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

intercept at minimum

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

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

slope at minimum

Linear Regression

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

Linear Regression

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

$$2\left(nb + m\sum X_i - \sum Y_i\right) = 0$$

$$i=1 \qquad\qquad i=1$$

Linear Regression

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

$$2\left(nb + m\sum^{n} X_i - \sum^{n} Y_i\right) = 0$$

$$i=1 \qquad\qquad i=1$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - m\sum_{i=1}^{n} X_i\right)$$

Linear Regression

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

$$2\left(nb + m\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - m\sum_{i=1}^{n} X_i\right)$$

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

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

Linear Regression

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

$$2\left(nb + m\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0$$

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} Y_i - m\sum_{i=1}^{n} X_i\right)$$

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

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

$$b = \bar{Y} - m\bar{X}$$

Linear Regression

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

intercept at minimum

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

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

slope at minimum

Linear Regression

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

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

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

$$\sum_{i=1}^{n} -2(Y_iX_i - bX_i - mX_i^2) = 0$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

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

$$\sum_{i=1}^{n} -2(Y_i X_i - bX_i - mX_i^2) = 0$$

$$b = \bar{Y} - m\bar{X}$$

Linear Regression

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

$$\sum_{i=1}^{n} -2(Y_iX_i - bX_i - mX_i^2) = 0$$

$$b = \bar{Y} - m\bar{X}$$

$$\sum_{i=1}^{n} -2(Y_iX_i - \bar{Y}X_i + m\bar{X}X_i - mX_i^2) = 0$$

Linear Regression

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

$$\sum_{i=1}^{n} -2(Y_iX_i - bX_i - mX_i^2) = 0$$

$$b = \bar{Y} - m\bar{X}$$

$$\sum_{i=1}^{n} -2(Y_iX_i - \bar{Y}X_i + m\bar{X}X_i - mX_i^2) = 0$$

$$\sum_{i=1}^{n}(Y_iX_i - \bar{Y}X_i) - m\sum_{i=1}^{n} X_i^2 - \bar{X}X_i = 0$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i) - m \sum_{i=1}^{n} X_i^2 - \bar{X} X_i = 0$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i) - m \sum_{i=1}^{n} X_i^2 - \bar{X} X_i = 0$$

$$m = \frac{\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i)}{\sum_{i=1}^{n} X_i^2 - \bar{X} X_i}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i) - m \sum_{i=1}^{n} X_i^2 - \bar{X} X_i = 0$$

$$m = \frac{\sum_{i=1}^{n}(Y_i X_i - \bar{Y} X_i)}{\sum_{i=1}^{n} X_i^2 - \bar{X} X_i}$$

$$m = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$m = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i \bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i \bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_i Y_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$m = \frac{Cov(X,Y)}{Var(X)}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$m = \frac{Cov(X,Y)}{Var(X)}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Sums/Averages of Random Variables

$$m = \frac{Cov(X,Y)}{Var(X)}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

Linear Regression

$$\sum_{i=1}^{n} \bar{X}^2 - X_i\bar{X} = 0$$

$$\sum_{i=1}^{n} \bar{X}\bar{Y} - Y_i\bar{X} = 0$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i) - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}(X_i^2) - n\bar{X}^2}$$

$$m = \frac{\sum_{i=1}^{n}(X_iY_i - X_i\bar{Y}) + \sum_{i=1}^{n}(\bar{X}\bar{Y} - Y_i\bar{X})}{\sum_{i=1}^{n}(X_i^2 - X_i\bar{X}) + \sum_{i=1}^{n}(\bar{X}^2 - X_i\bar{X})}$$

CTL applies, can compute expected value, p-values

$$m = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

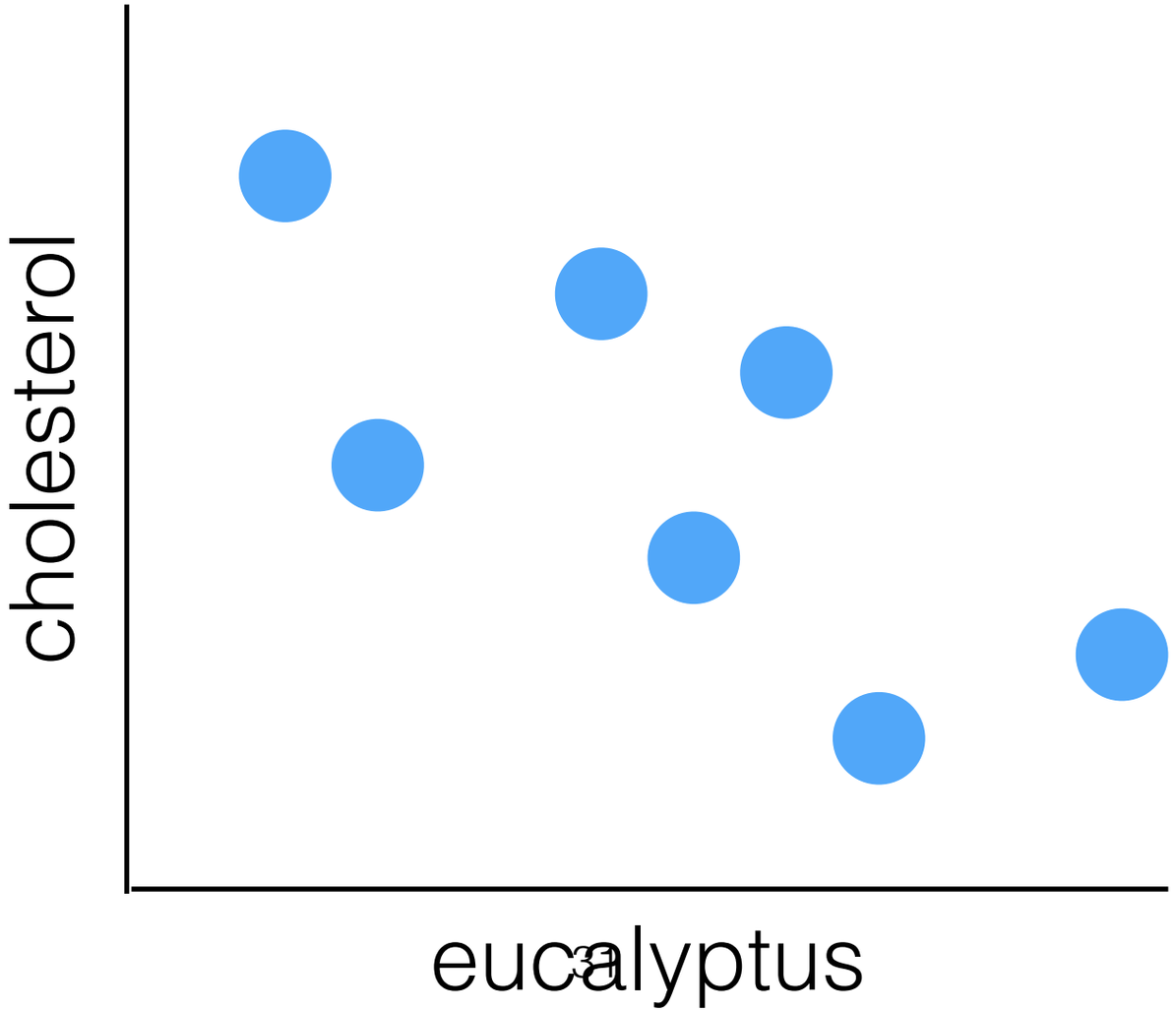
$$m = \frac{Cov(X,Y)}{Var(X)}$$

Data Analysis Toolkit #10: Simple linear regression

http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_10.pdf

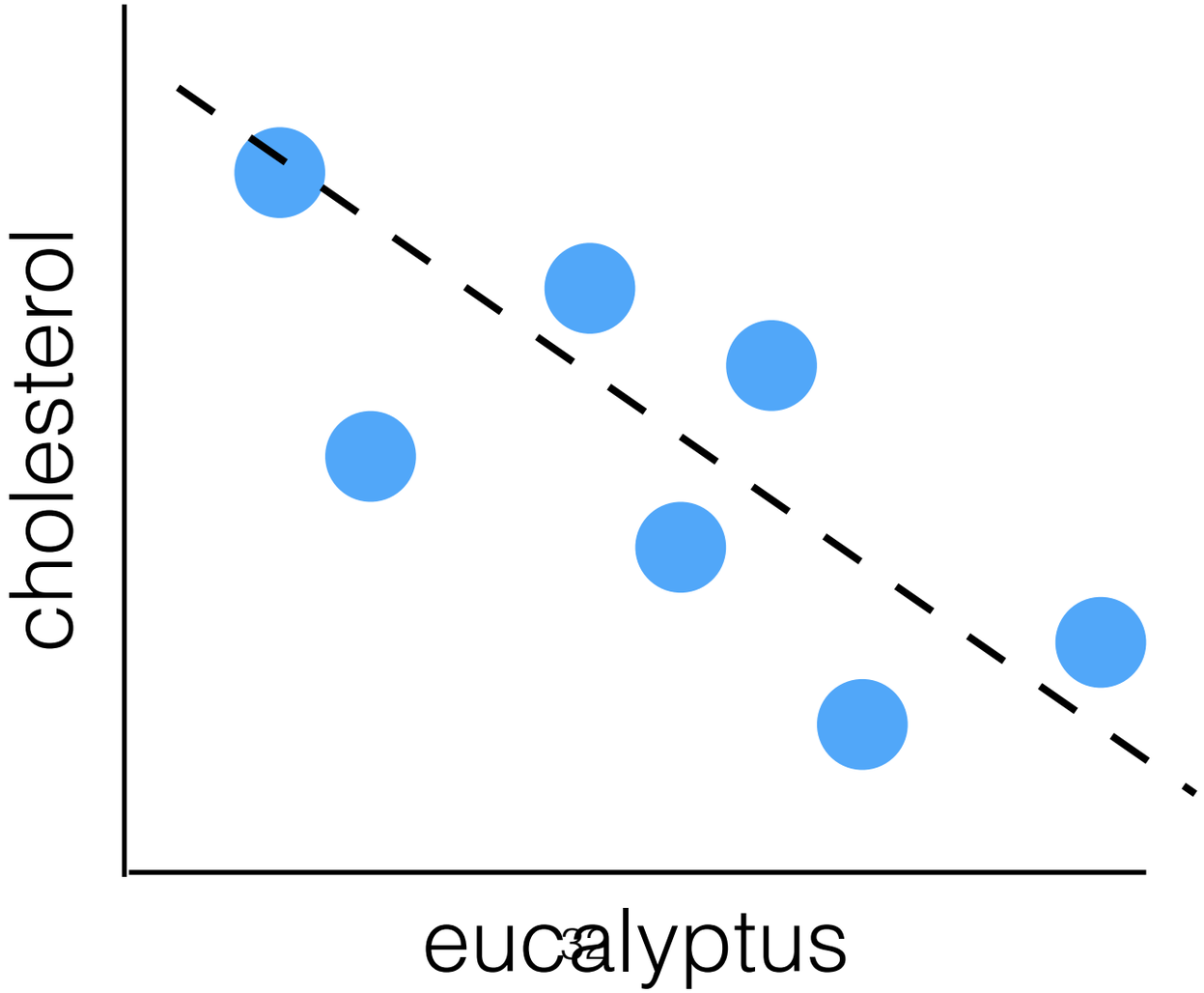
Linear Regression

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



Linear Regression

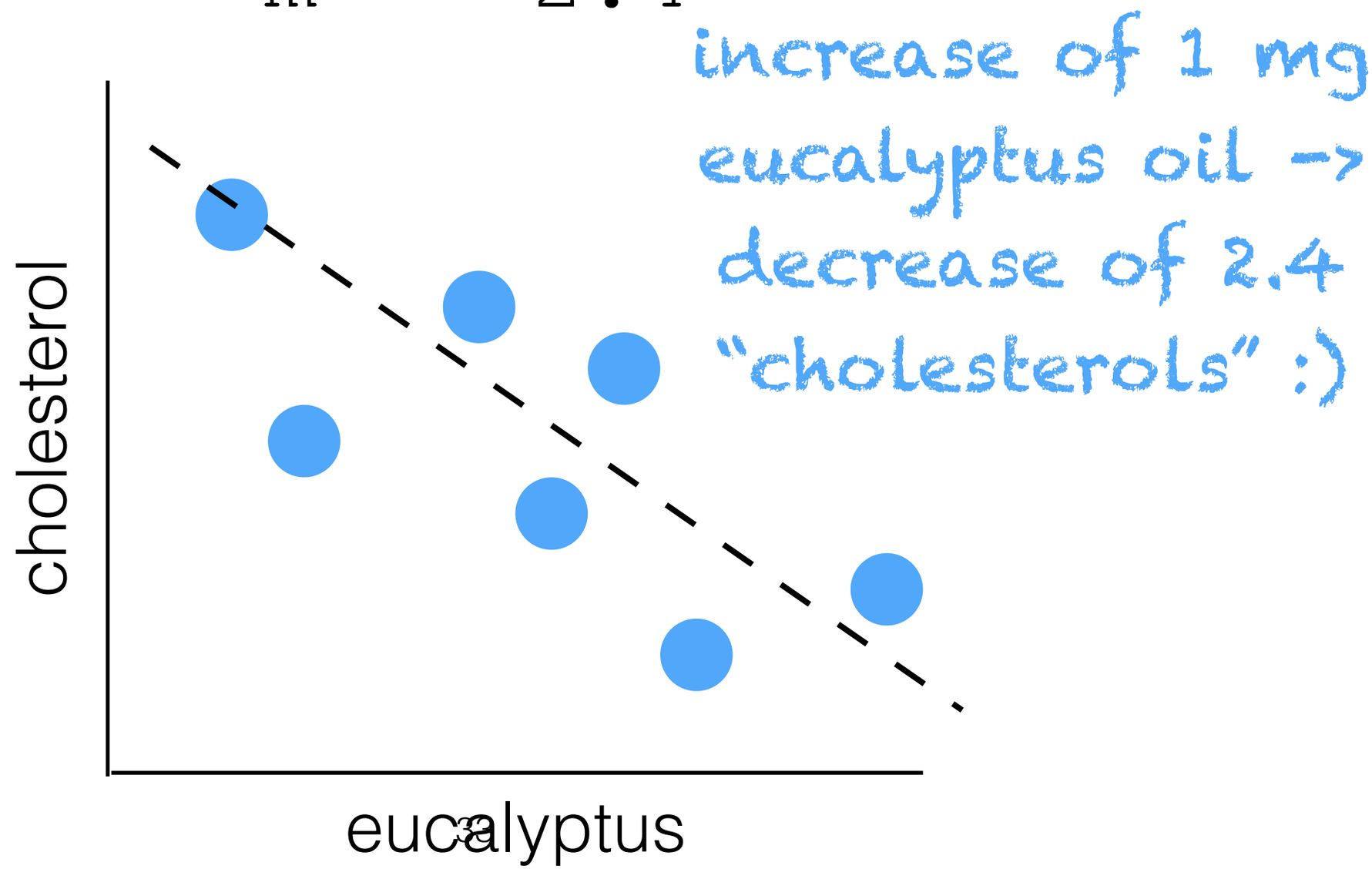
```
cholesterol = m(eucalyptus) + b
          m = -2.4
```



Linear Regression

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

increase of 1 mg eucalyptus oil -> decrease of 2.4 "cholesterols" :)

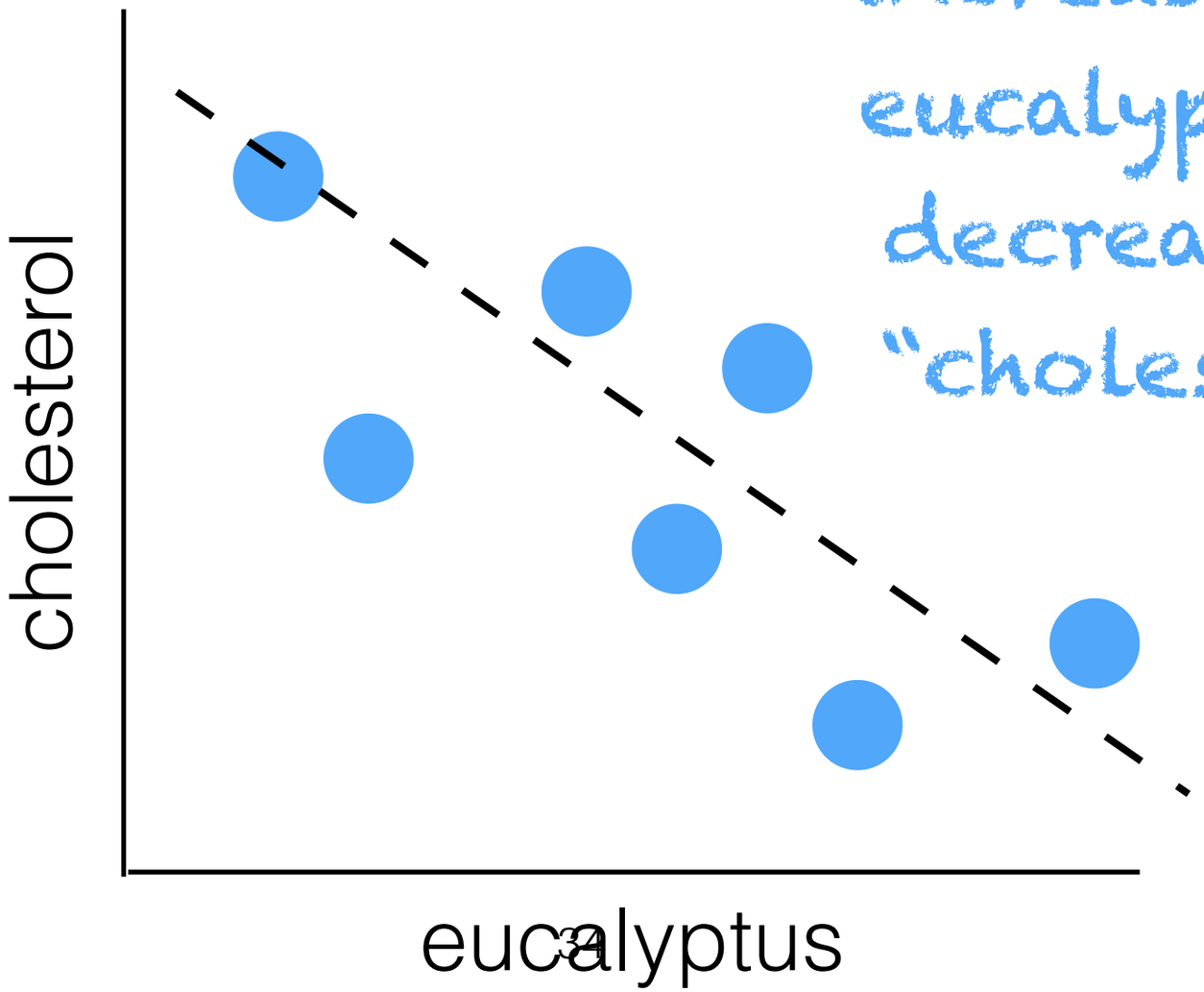


Linear Regression

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

¿que pasa?

increase of 1 mg eucalyptus oil -> decrease of 2.4 "cholesterols" :)



Discussion-question-thinly-veiled-as-a-clicker Question!

Discussion-question-thinly-veiled-as-a-clicker Question!

What should we make 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.
(e) We should click on the obviously snarky answer and see if Ellie gets mad.

Discussion-question-thinly-veiled-as-a-clicker Question!

What should we make 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 correlatio

(c) There is pr easuring eucalyptu s correlated.

(d) There is pr iling to capture o

(e) We should click on the obviously snarky answer and see if Ellie gets mad.

could be the case, but want to do due diligence before concluding this...

Discussion question thinly veiled as a

rel d

Yes and no. We *are* confusing correlation with causation, but linear regression does this by construction (even when we are looking at a real relationship).

(a) There is probably no actual relationship. We are confusing correlation with causation, but 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.

(e) We should click on the obviously snarky answer and see if Ellie gets mad.

Discussion-question-thinly-veiled-as-a-clicker Question!

What should we make of the observed relationship between use of eucalyptus oil and

Units should not matter, since differences in units are usually equivalent up to linear transformation

(a) There proba… …on is a legitimate r…
(b) There is pro… …g correlation …
(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.
(e) We should click on the obviously snarky answer and see if Ellie gets mad.

Discussion-question-thinly-veiled-as-a-clicker Question!

What should we make 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, [illegible]

(b) There is probably no [illegible] correlation with ca[illegible]

(c) There is probably no [illegible] 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.

(e) We should click on the obviously snarky answer and see if Ellie gets mad.

This is a good answer!
Let's spend multiple more slides on it.

Discussion-question-thinly-veiled-as-a-clicker Question!

What should we make 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
correlation with ca
(c) There is probably no
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.
(e) We should click on the obviously snarky answer and see if Ellie gets mad.

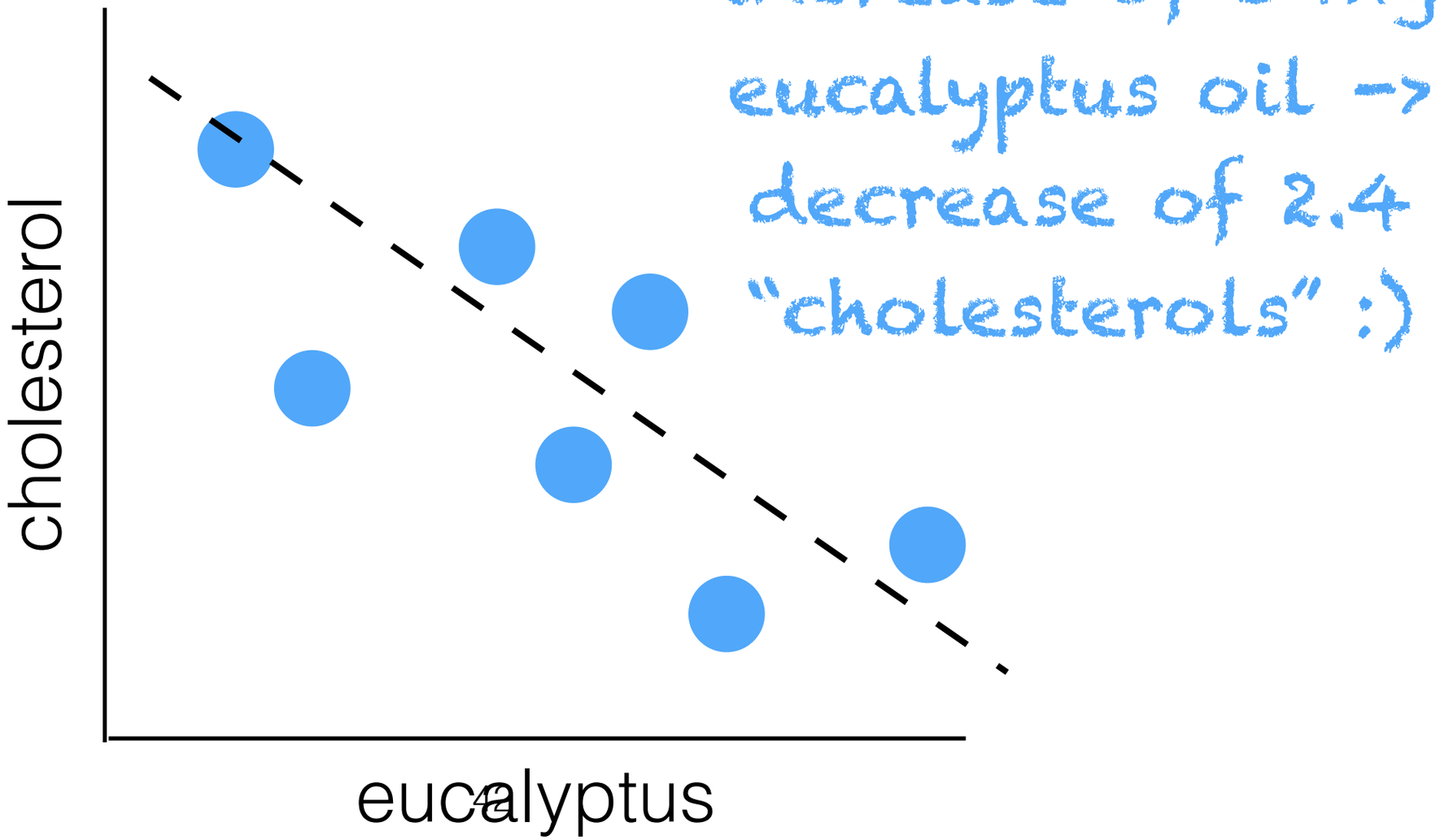
Not mad, just disappointed

Linear Regression

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

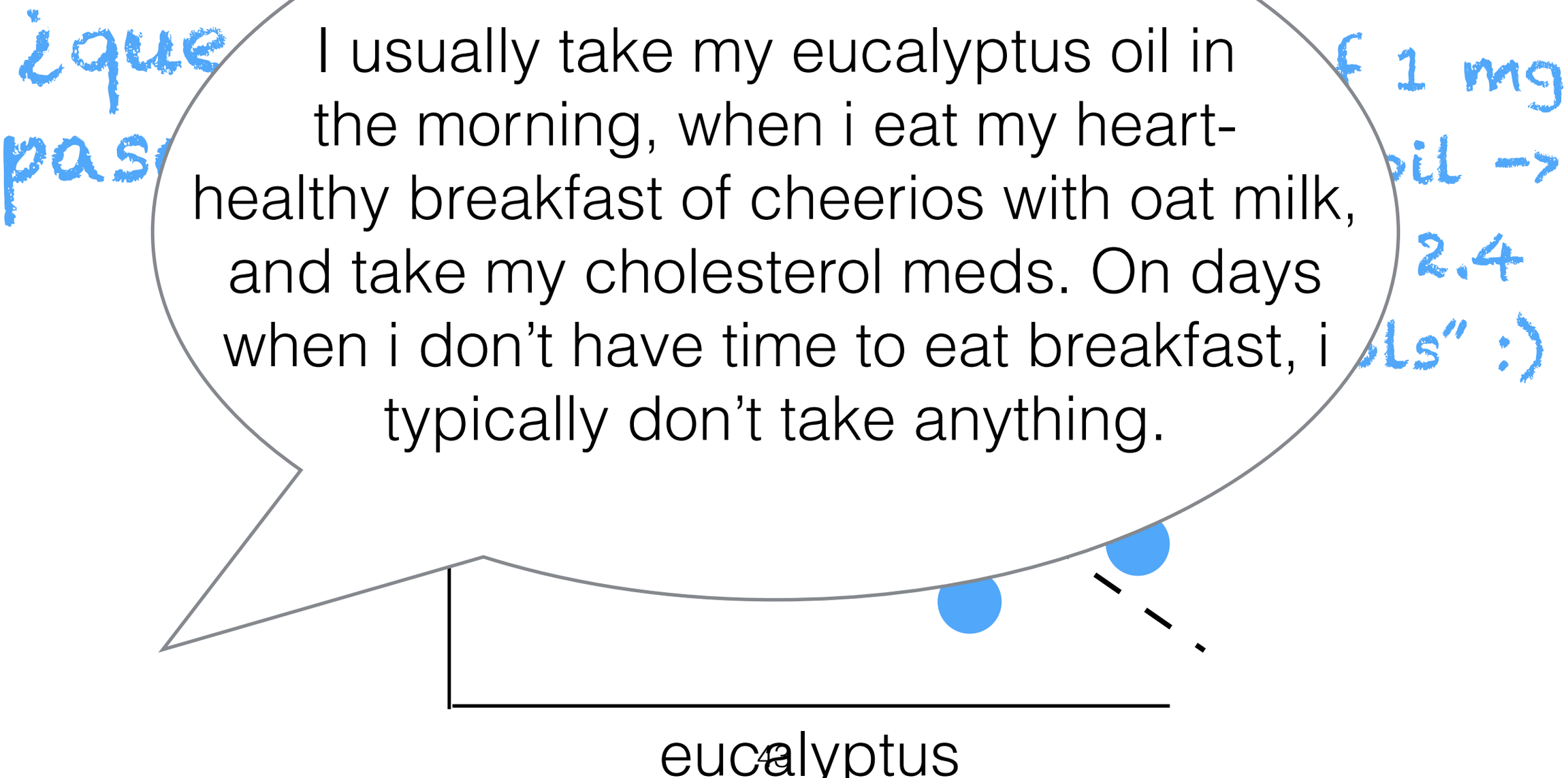
¿que pasa?

increase of 1 mg eucalyptus oil -> decrease of 2.4 "cholesterols" :)



Linear Regression

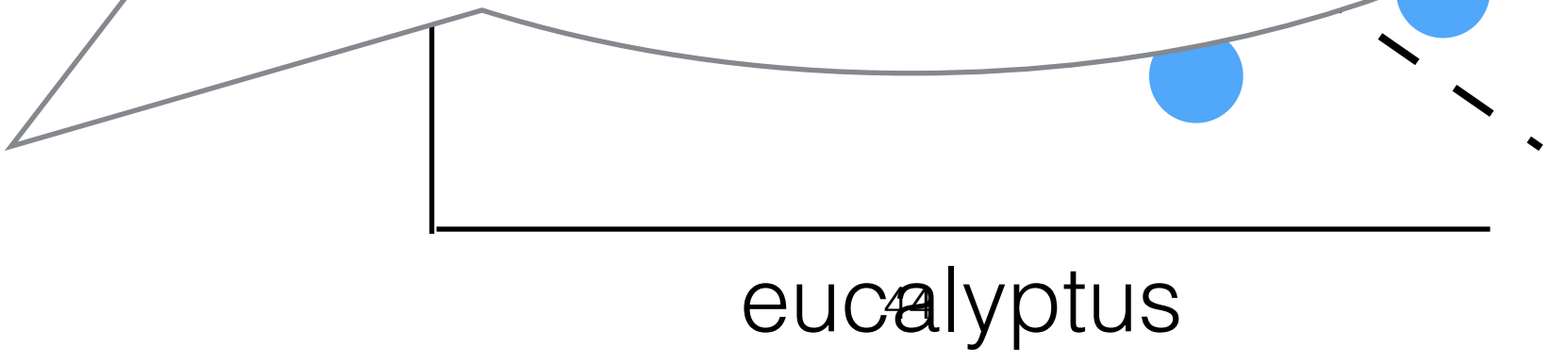
cholester ptus) + b



Linear Regression

cholester ptus) + b

I usually take my eucalyptus oil in the morning, when i eat my heart-healthy **breakfast** of cheerios with oat milk, and take my **cholesterol meds**. On days when i don't have time to eat breakfast, i typically don't take anything.



Omitted Variable Bias

Omitted Variable Bias

- By construction, we assume that the dependent variable can be predicted from the explanatory variables only

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

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
X2: cholesterol meds
X3: breakfast
X4: constant term

Multiple Linear Regression

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

intercept

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

Multiple Linear Regression

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

slopes/
coefficients/
effects

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

Multiple Linear Regression

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

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

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

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

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

depends on other explanatory variables

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

Multiple Linear Regression

$Y = m1X1$ change in cholesterol associated with a increase of 1 mg eucalyptus oil, holding other variables constant

$$Q = \sum_{i=1}^{n} (Y_i -$$

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

Multiple Linear Regression

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

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

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

LinAlg Detour

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

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

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

LinAlg Detour

Y = m Matrices of observations m4X4

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

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

LinAlg Detour

Y = m1X1

Vector of coefficients

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

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

LinAlg Detour

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

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

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

X Transpose

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

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

LinAlg Detour

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

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

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

Inverse

LinAlg Detour

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

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

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

Inverse

doesnt always exist...

LinAlg Detour

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

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

linearly dependent/ co-linear

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

Inverse

LinAlg Detour

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

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

linearly dependent/ co-linear

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

"Pseudo-Inverse"

Dummy Variables

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

Y: cholesterol level

X1: eucalyptus

X2: cholesterol meds

X3: breakfast

X4: constant term

Dummy Variables

- Used to encode qualitative features

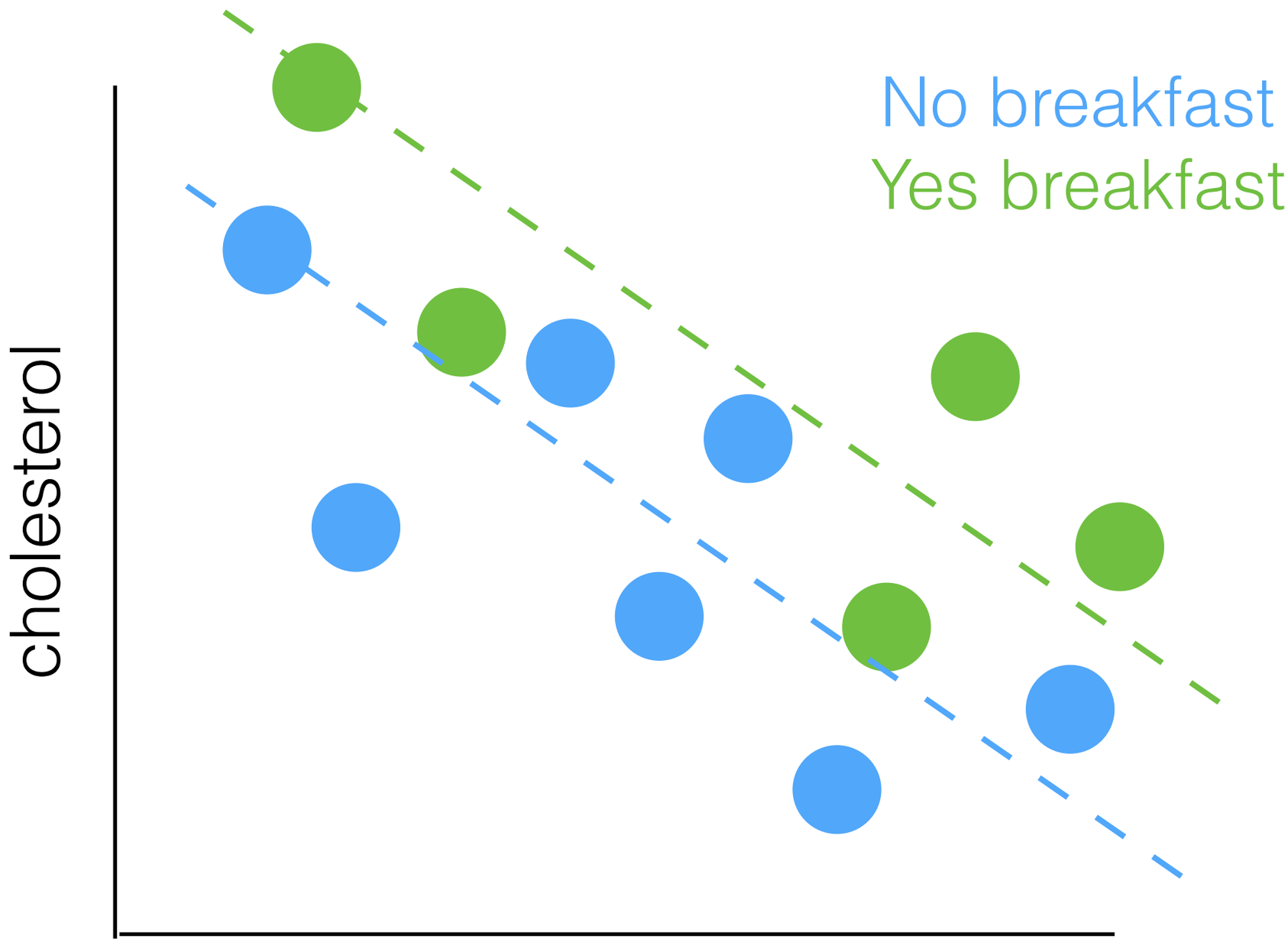
Dummy Variables

- Used to encode qualitative features
- AKA indicator variables, Boolean variables, one-hot variables, sparse variables…

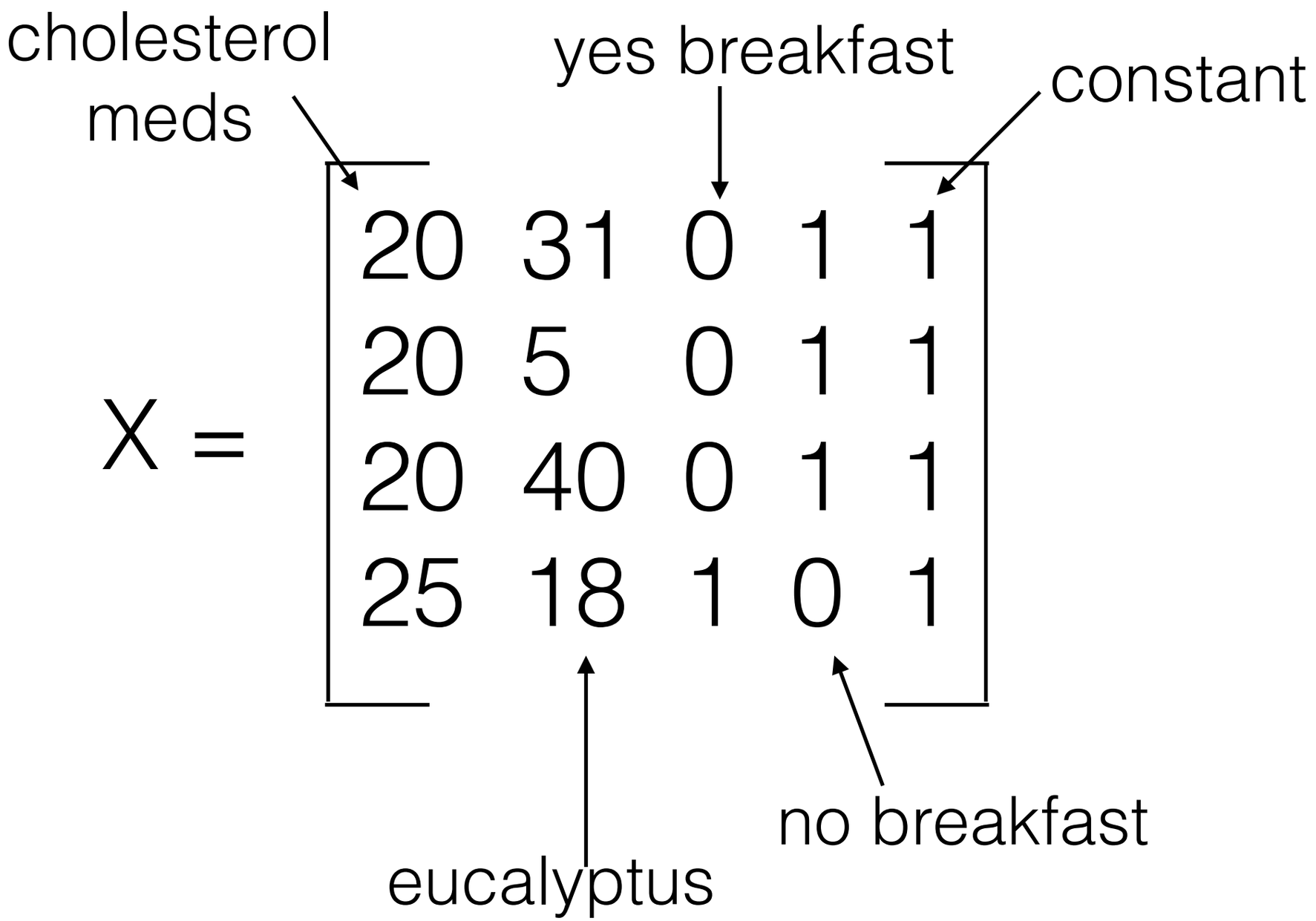
Dummy Variables

- Used to encode qualitative features
- AKA indicator variables, Boolean variables, one-hot variables, sparse variables…
- Interpretable as shift in intercept for different groups

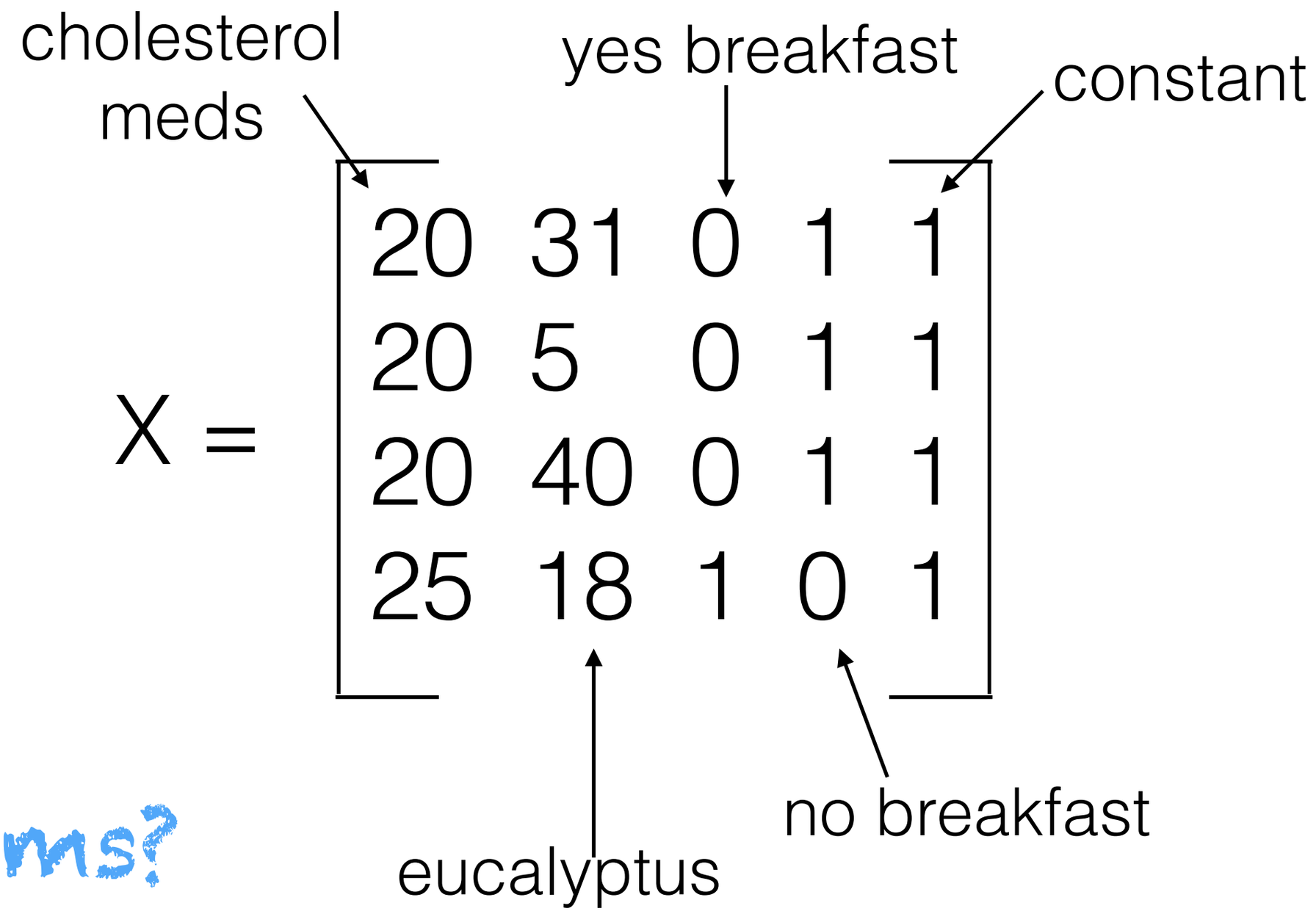
Dummy Variables



Dummy Variables

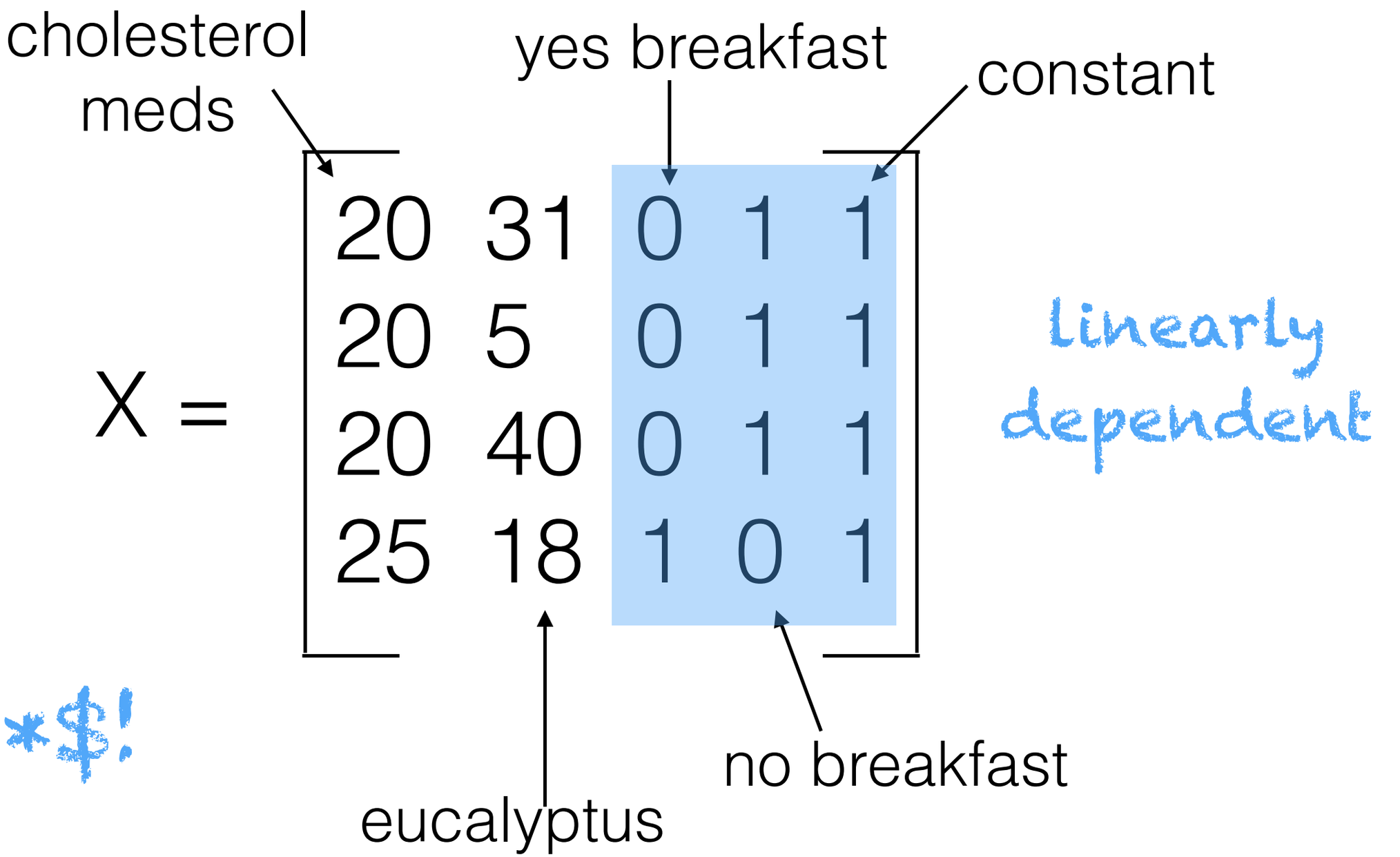


Dummy Variables

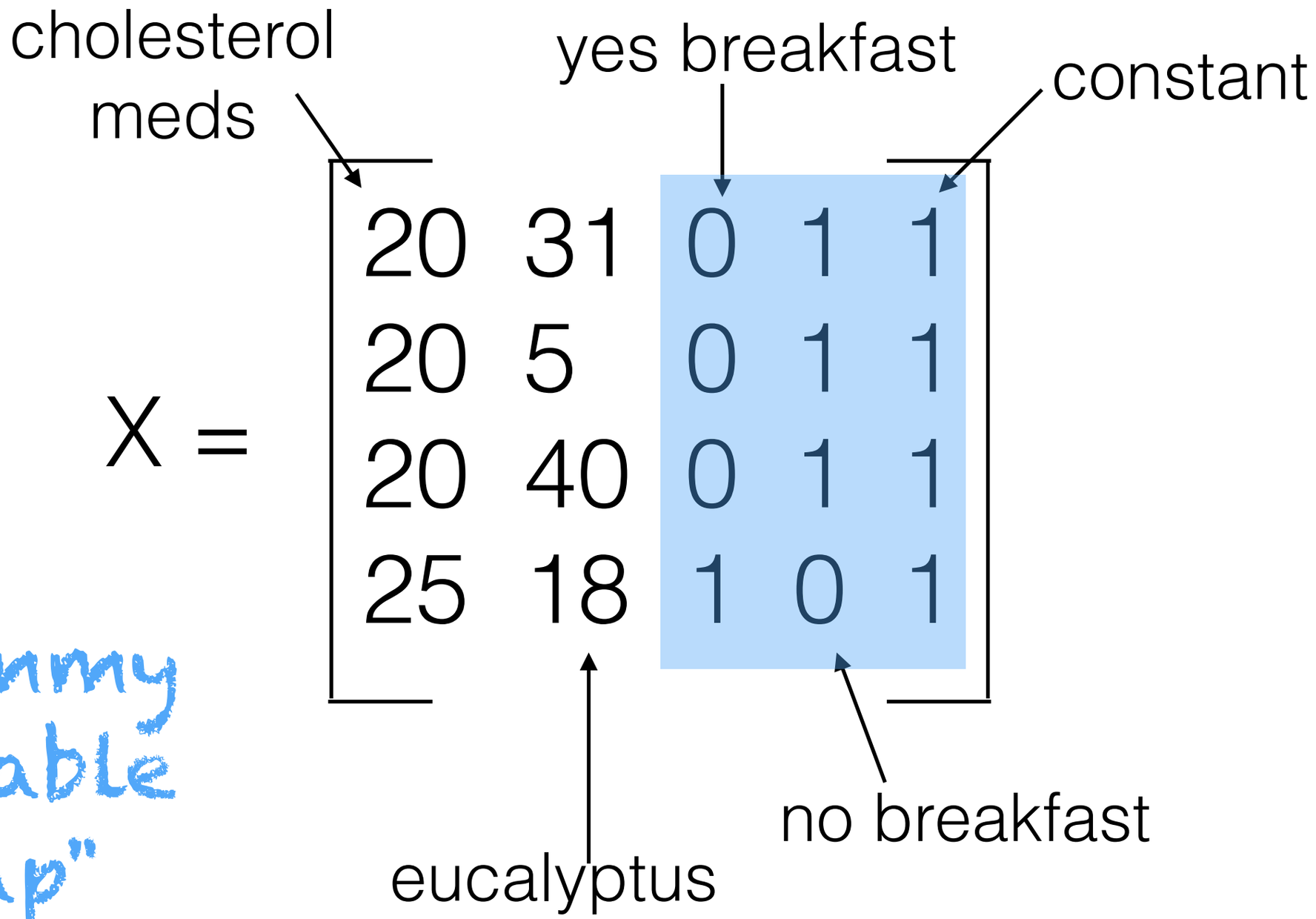


Qualms?

Dummy Variables

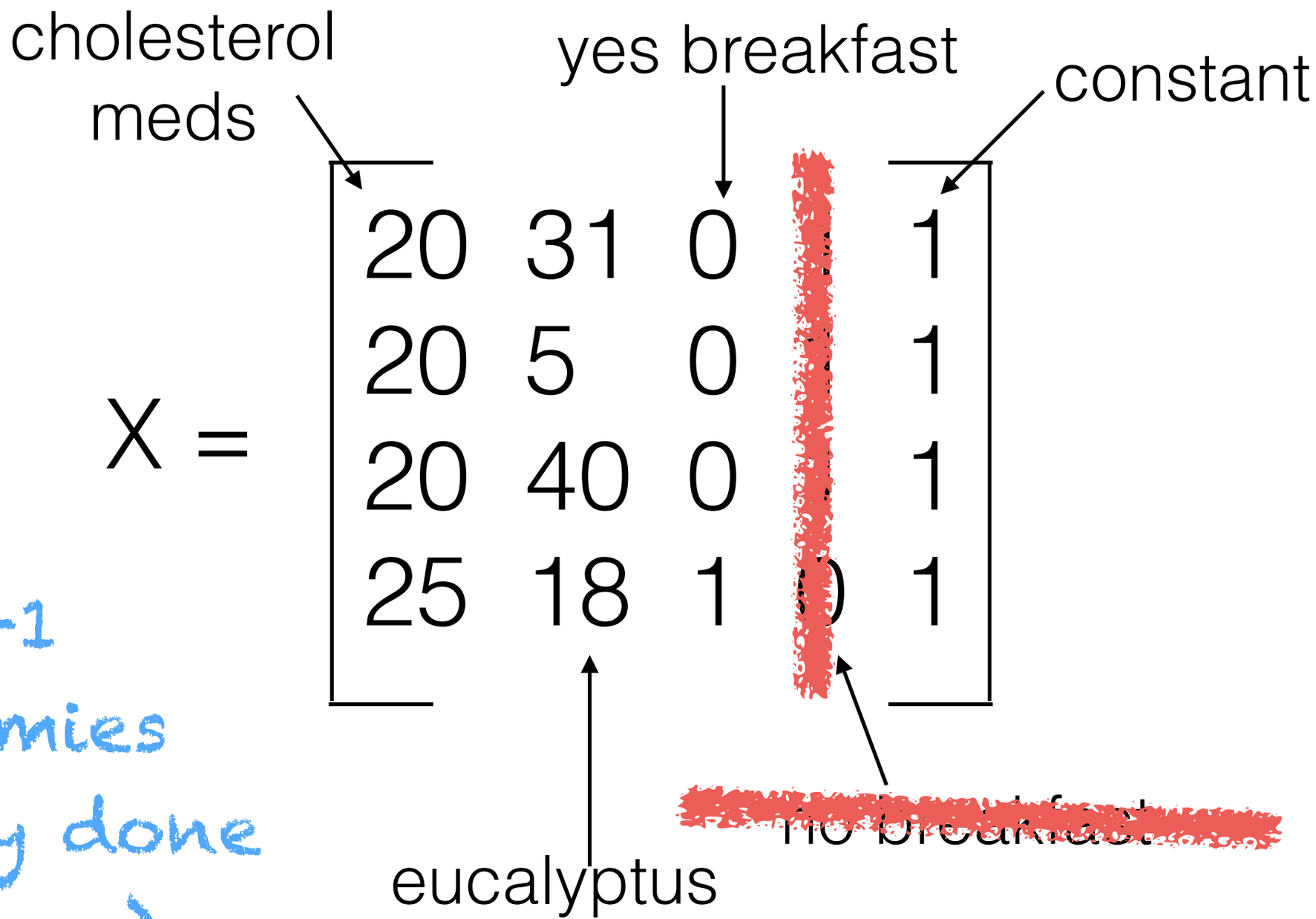


Dummy Variables



"dummy variable trap"

Dummy Variables



n-1 dummies (usually done for you)

Clicker Question!

Clicker Question!

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

Clicker Question!

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

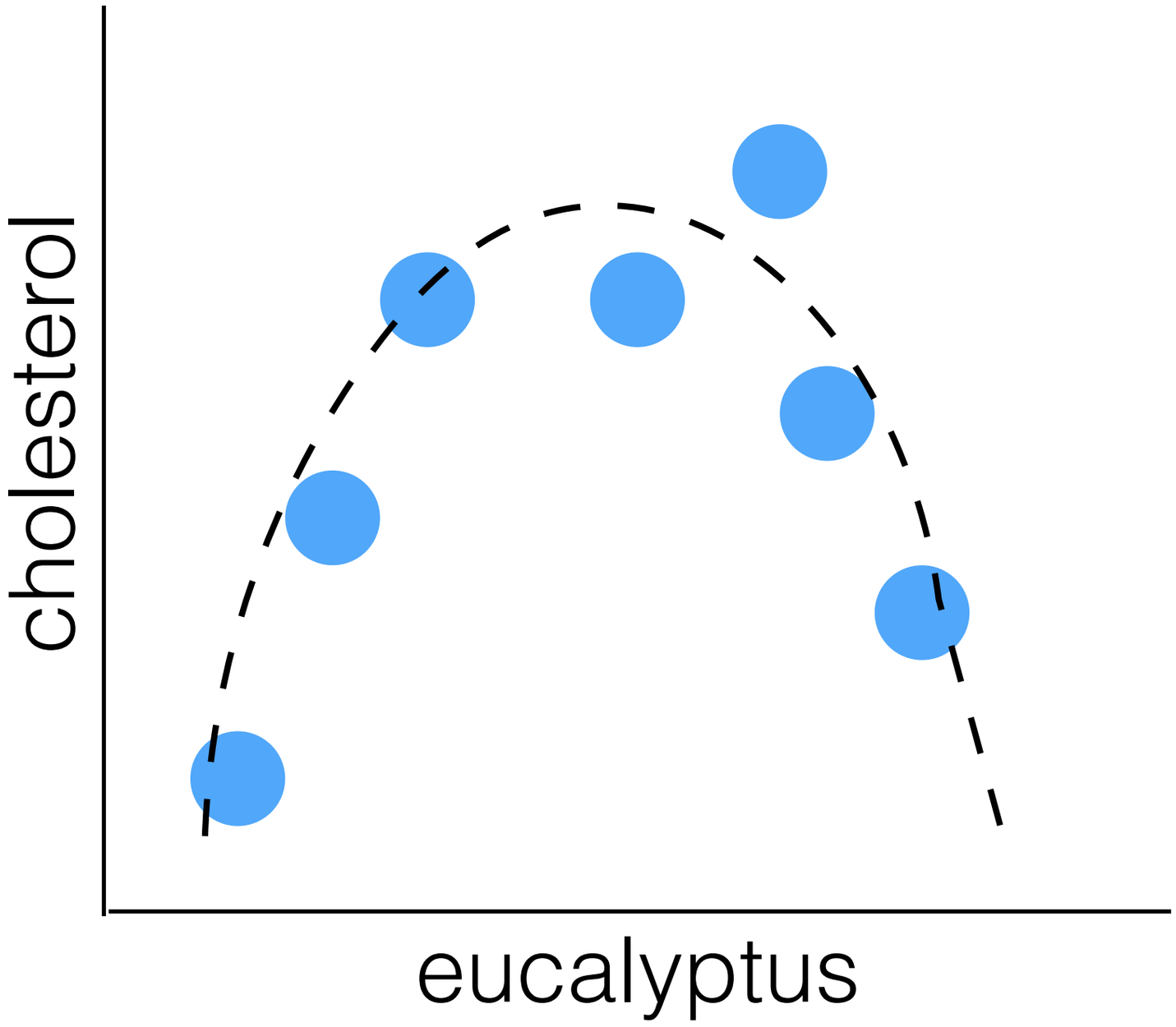
$$Y = m1X1 + m2X2 + m3X3 + m4X4 + m5X5$$

Y: happiness
X1: day of week (dummies M,T,W,Th,F,S,Su) 6
X2: bank account balance (real value) 1
X3: breakfast (dummies yes,no) 1
X4: whether you have found your inner peace (dummies yes,no,unclear) 2

constant = 1

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

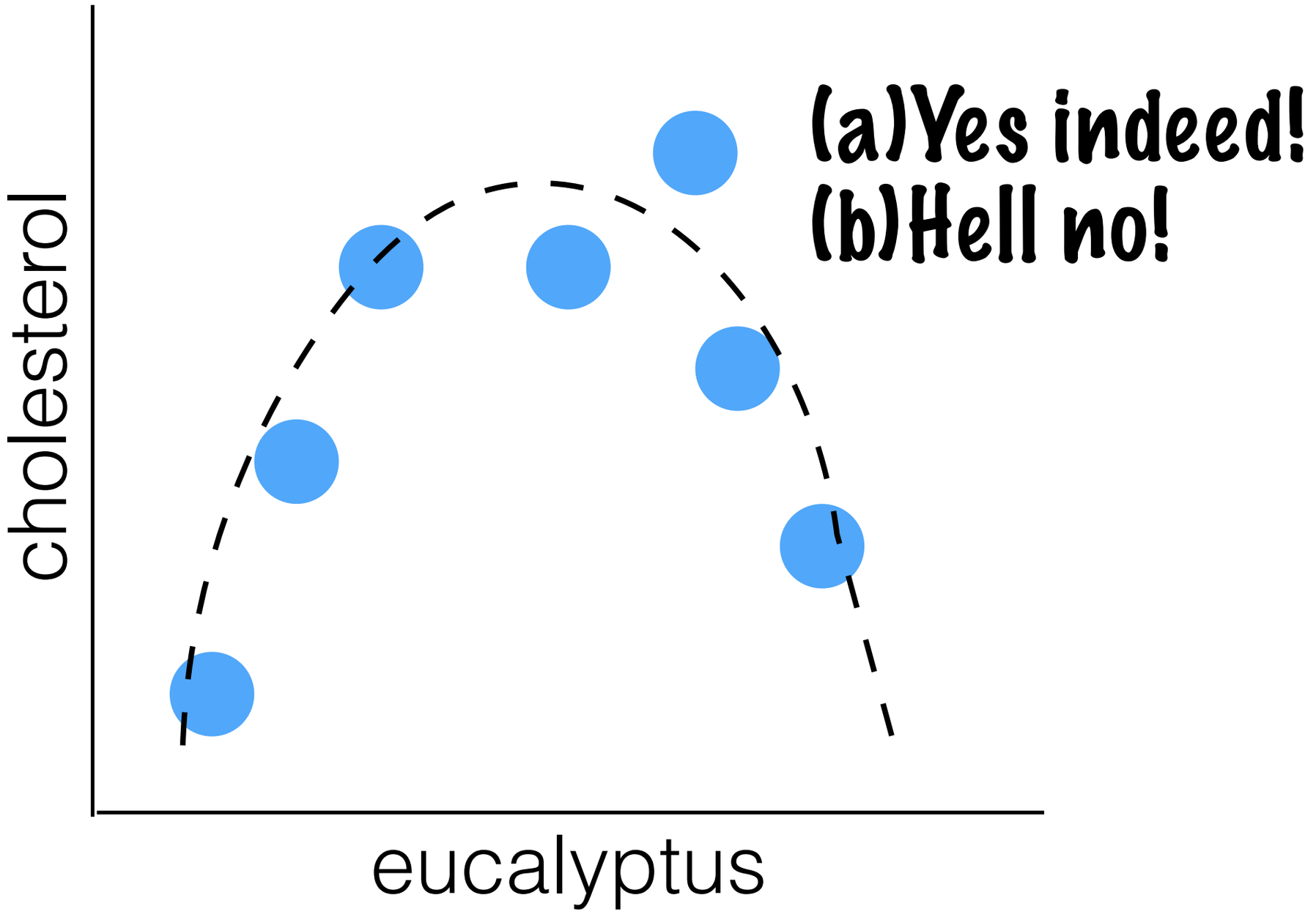
Nonlinear Relationships



Clicker Question!

Clicker Question!

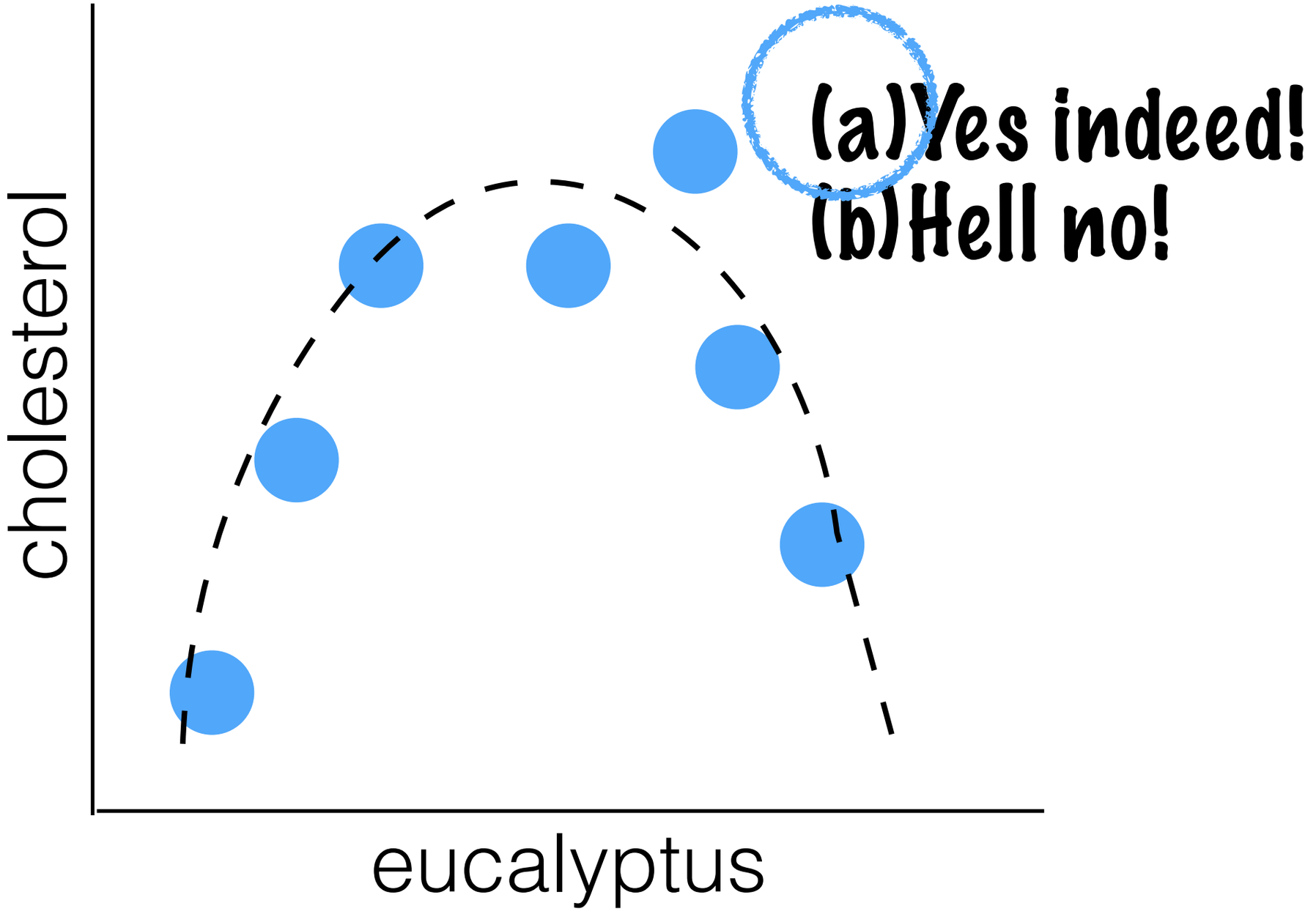
Can we model this with linear regression?



(a)Yes indeed!
(b)Hell no!

Clicker Question!

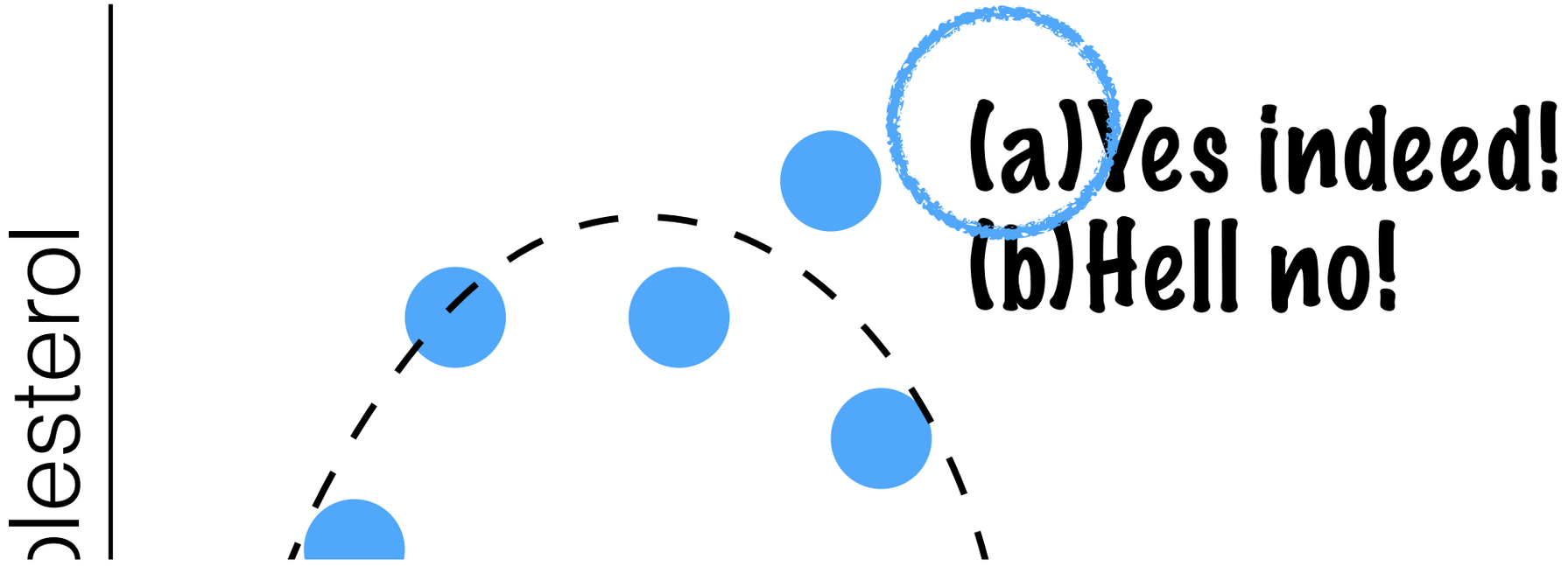
Can we model this with linear regression?



(a) Yes indeed!
(b) Hell no!

Clicker Question!

Can we model this with linear regression?



$$Y = m1X1 + m2X2 + m3X3$$

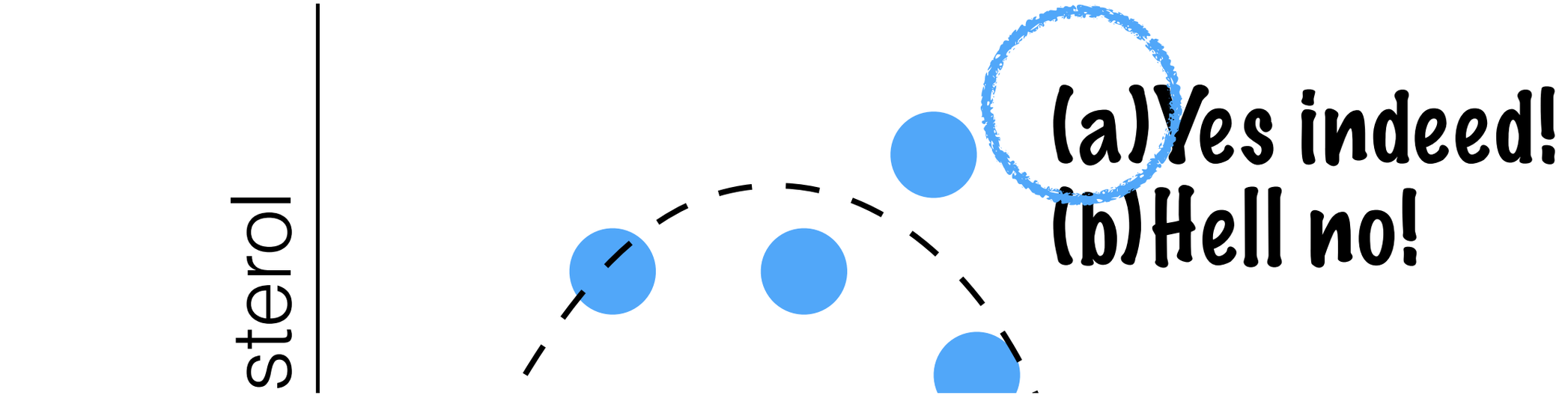
Y: cholesterol

X1: eucalyptus

X2: $eucalyptus^2$

Clicker Question!

Can we model this with linear regression?



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

Y: cholesterol

X1: eucalyptus

X2: cholesterol meds

X3: X1 x X2

"interaction term"

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:	Least Squares	F-sta	.020e+06
Date:	Tue, 26 Feb 2019	Prob	.83e-239
Time:	04:42:47	Log-l	-146.51
No. Observations:	100	AIC:	299.0
Df Residuals:	97	BIC:	306.8
Df Model:	2		
Covariance Type:	nonrobust		

overall fit of model (SSE)

	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:	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
	100	AIC:	299.0
	97	BIC:	306.8
	2		
	bust		

coefficients (i.e. effect sizes)

	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:	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:	
Df Residuals:	97	BIC:	
Df Model:	2		
Covariance Type:	nonrobust		

p-values

	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

Discussion Question!

Discussion Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income:
salary($)

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

* I completely made these numbers up!!

Discussion Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income:
salary($)

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

How to we interpret this?

*** I completely made these numbers up!!**

Discussion Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

```
income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income:
salary($)
```

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

How to we interpret this?

Going to college corresponds to a increase of $20K in salary, assuming other variables are fixed.

*** I completely made these numbers up!!**

Clicker Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income: salary($)

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

Did you get it right

(a) Yes, obv.
(b) No, I cannot tell a lie.

*** I completely made these numbers up!!**

Discussion Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income:
salary($)

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

How to we interpret this?

*** I completely made these numbers up!!**

Discussion Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

```
income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income:
salary($)
```

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

How to we interpret this?

Being female corresponds to a decrease of 12K in salary, holding all other things fixed.

* I completely made these numbers up!!

Clicker Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income: salary($)

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

Did you get it right

(a) Yes, obv.
(b) No, I cannot tell a lie.

*** I completely made these numbers up!!**

Discussion Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

income: salary($)	var	const	P>|t|
edu: 1=college	edu	20000	0.03
gender: 1=F	gender	-12000	0.06
parent_edu: 1=col	parent_edu	15000	0.07
parent_income:	parent_income	1.8	0.01
salary($)	edu:parent_income	2.3	0.02

How to we interpret this?

* I completely made these numbers up!!

Discussion Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

```
income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income:
salary($)
```

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

How to we interpret this?

Conditioned on your having gone to college, an increase of $1 in parents' salary corresponds to an increase of $2.3 in your salary.

Clicker Question!*

```
income ~ education + gender + parent_edu +
  parent_income + education:parent_income
```

```
income: salary($)
edu: 1=college
gender: 1=F
parent_edu: 1=col
parent_income:
salary($)
```

var	const	P>|t|
edu	20000	0.03
gender	-12000	0.06
parent_edu	15000	0.07
parent_income	1.8	0.01
edu:parent_income	2.3	0.02

Did you get it right

(a) Yes, obv.
(b) No, I cannot tell a lie.

*** I completely made these numbers up!!**

run along