Summer Research Symposium
July 23, 2021
California State University, Long Beach

Gaussian Process Regression for Analysis of Educational Data

Kierra Manuel

Dr. Kagba Suaray, Department of Mathematics

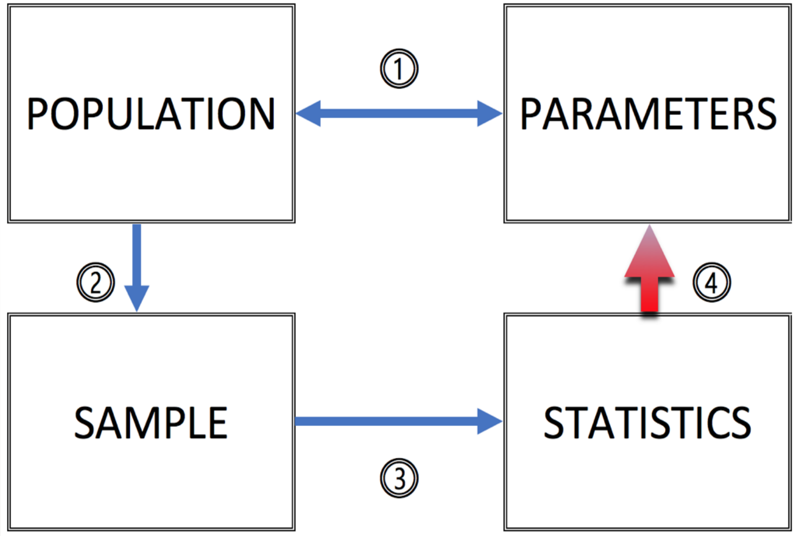
College of Natural Sciences and Mathematics

Overview

- Introduction/Purpose
- Background
  - Univariate Normal Properties and Derivation
- Methods of Theory and Code
  - The Bivariate Normal
  - Multiple Regression
  - R Coding Language
- Future Work
  - Gaussian Process Regression for Educational Data
- References
- Acknowledgements

Introduction/Purpose

- My research focus: investigating patterns and trends in the educational system, and how they correlate with geography, race, and socioeconomic status. I believe once patterns and trends are identified this information can be used to promote equity in the educational system.
- In order to identify these patterns, we use statistics to model these trends, in addition other real-world data.
- At this point in my research, I am still learning methods and will discuss how I plan to apply them in future work.



Background

- The two approaches to statistical analysis are univariate and multivariate.
  - Univariate examines one variable (ex. weight of women)
  - Multivariate examines two or more variables simultaneously (ex. height and weight of women)

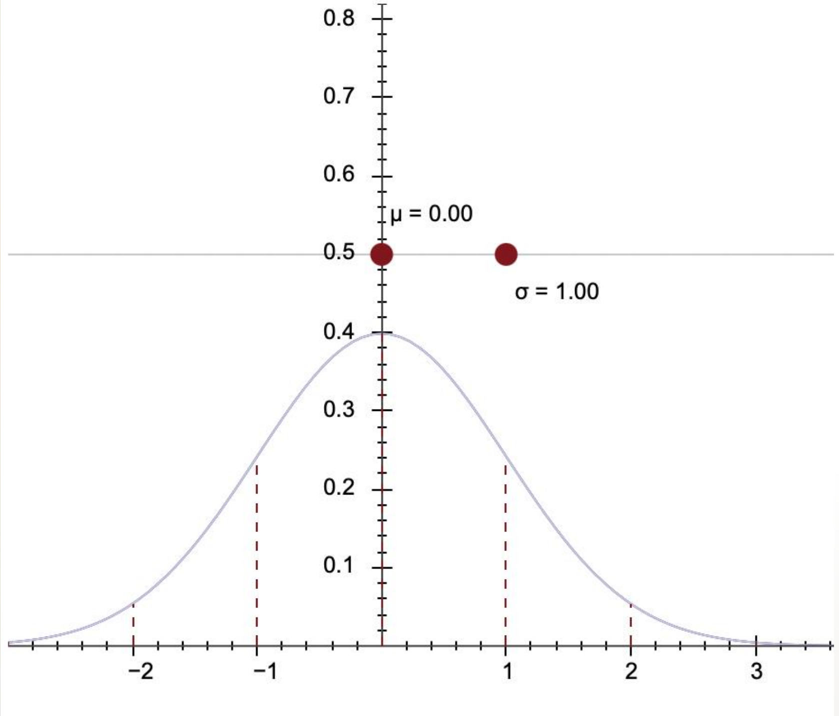
- The Univariate Gaussian Distribution, also known as Normal Distribution, is a continuous probability distribution for one random variable, x
  - The probability distribution function (PDF) is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, -\infty < x < \infty$

- There are two parameters for the distribution
  - $\mu = mean = E(x) = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx$
  - $\sigma^2 = variance = E(x-\mu)^2$

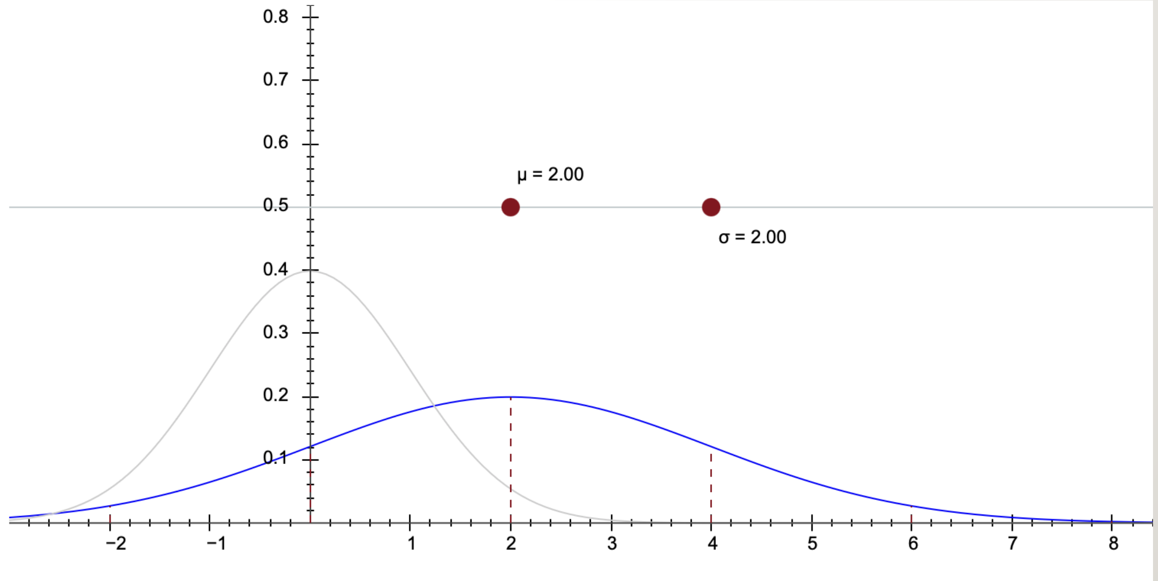
Background cont.

- The parameters μ and σ^2 affect the location and spread of the Gaussian probability density function

- A Standard Gaussian distribution function, where $\mu = 0$ and $\sigma = 1$.

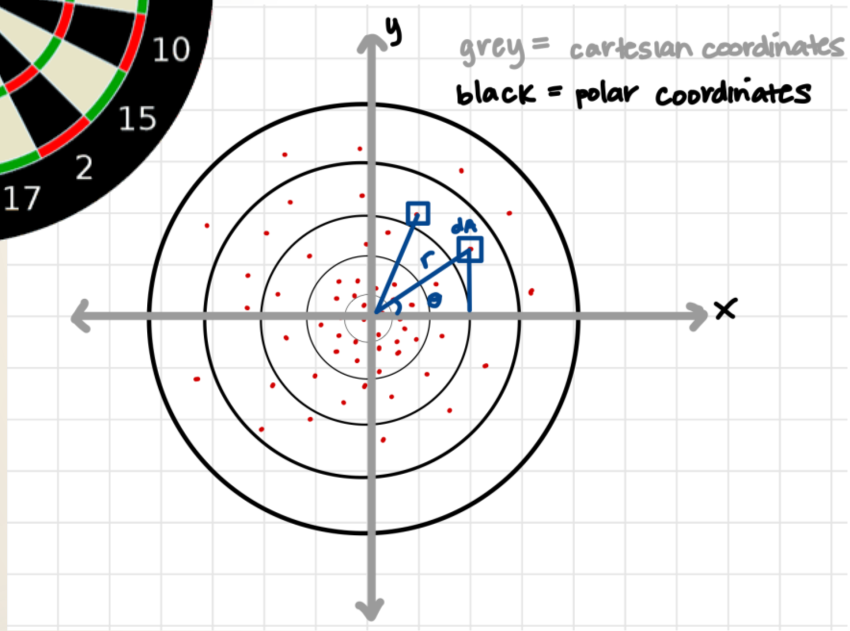


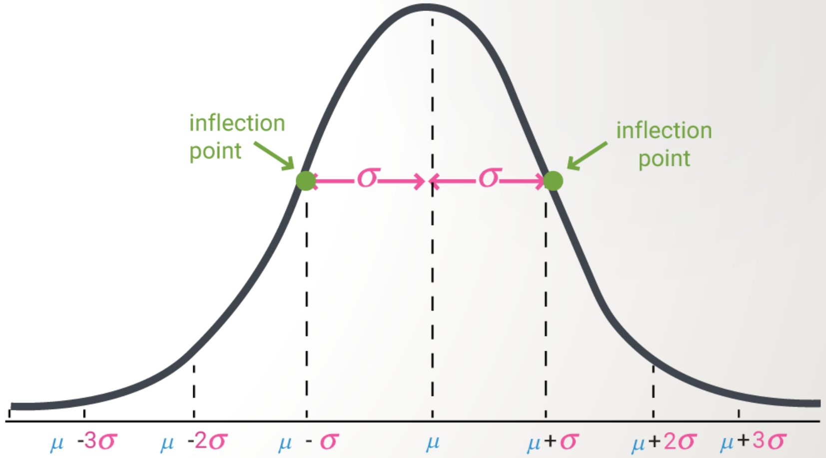
- A Gaussian distribution function where $\mu = 2$ and $\sigma = 2$. Notice how the μ $parameter$ changes the center of the curve and the increase in the σ $parameter$ spreads the curve out.



Background cont.

- How was the univariate probability distribution function derived?

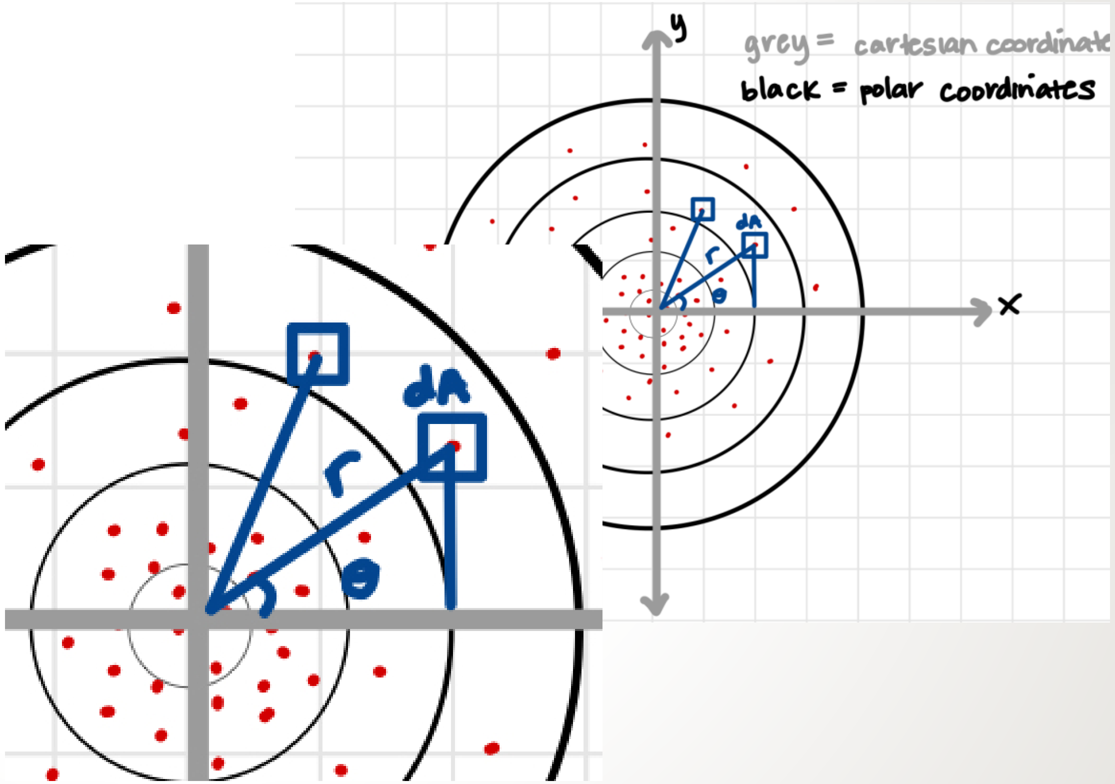




Background cont.

- φ *is* a function that describes the relative probability of finding darts at different locations

Let $\varphi = Probability\ Density\ Function$



$$\varphi dA = \varphi(r)dA$$

$$\varphi(r)dA = \varphi(x) * \varphi(y)dA$$

$$\varphi\left(\sqrt{x^2+y^2}\right) = \varphi(x) * \varphi(y)$$

$$\varphi\left(\sqrt{x^2+y^2}\right) dA = \frac{\varphi(x)\varphi(y)}{\lambda^2}$$

$$The\ most\ general\ case\ \ \varphi(x) = ae^{bx^2}$$

$$\int_{-\infty}^{\infty} \varphi(x)dx = 1 \ \therefore a\int_{-\infty}^{\infty} e^{bx^2} = 1$$

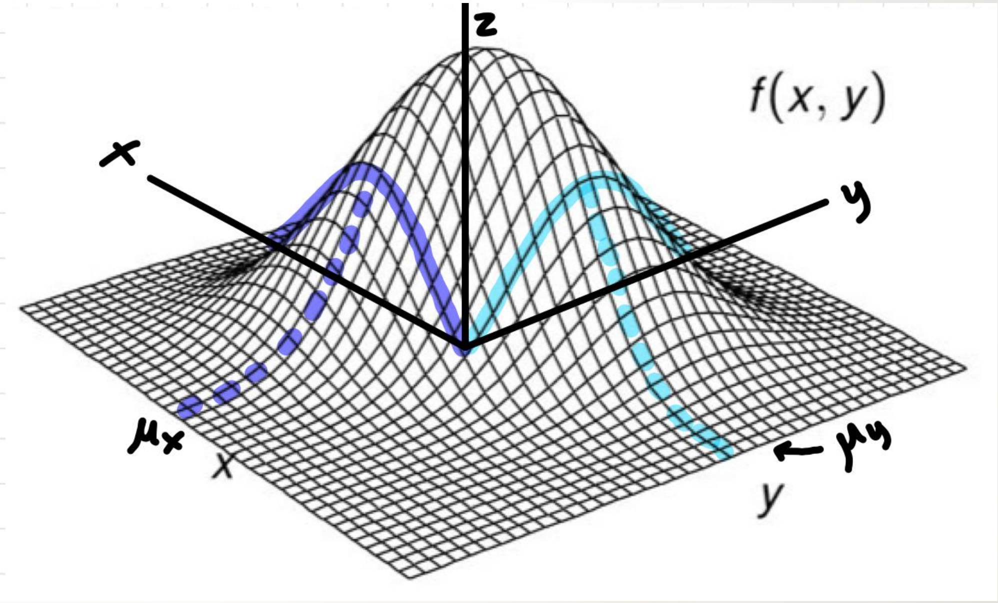
$$\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\ with\ mean = \ \mu\ and\ variance = \ \sigma^2$$

Methods - Theory

- The Bivariate Gaussian Distribution examines two independent random variables.

- The Marginals for random variables X and Y are both normal.
  - $X \sim N(\mu_x, \sigma_x)$
  - $Y \sim N(\mu_y, \sigma_y)$

- The Multivariate Gaussian Distribution probability Function
  - $P(X,Y) = (2\pi)^{-\frac{\kappa}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$



Methods – Theory cont.

Bivariate Case

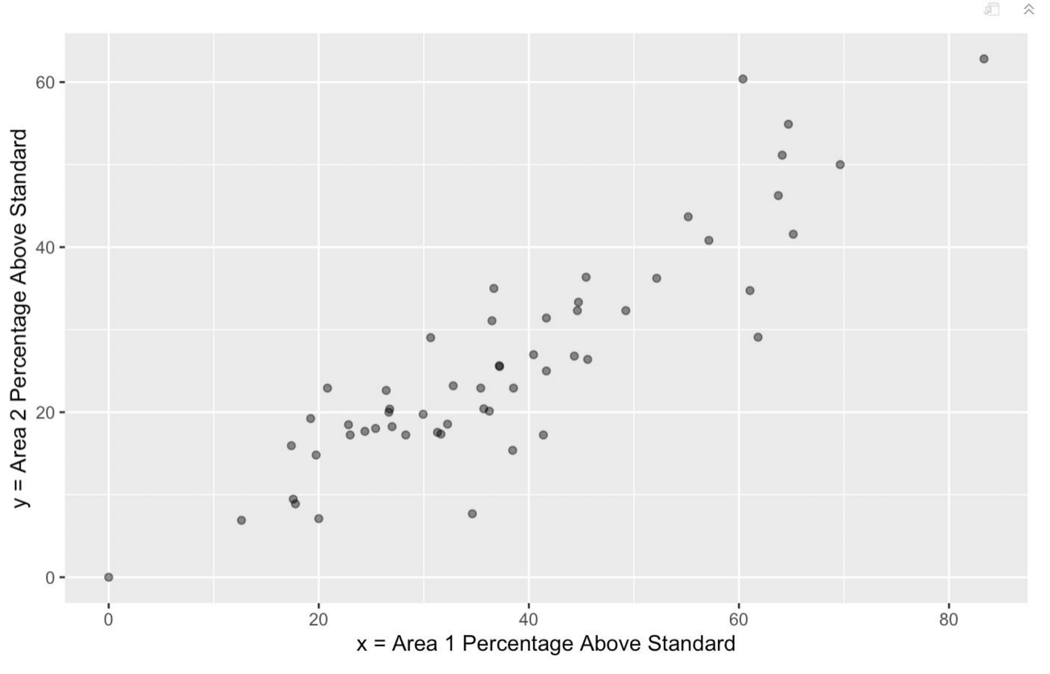
$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

Multivariate Care

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}$$

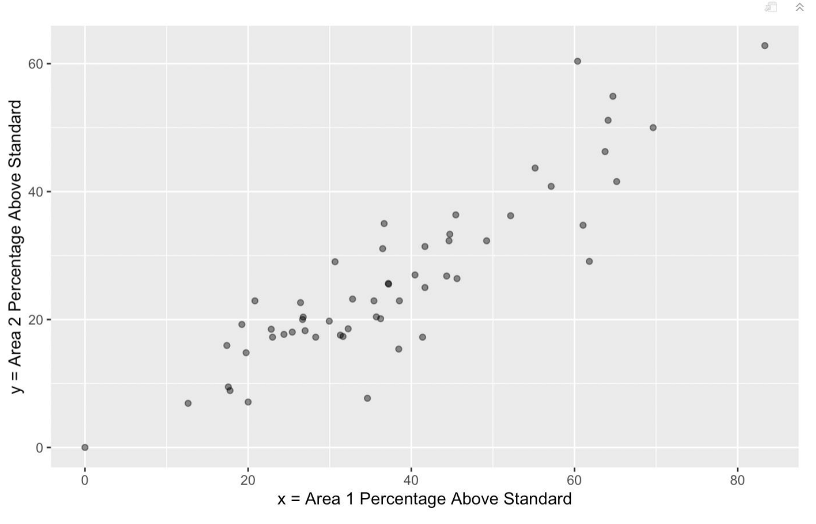


- σ_{21} *and* σ_{12} are always equal = equal to covariance.
- Covariance is proportional is correlation.
- Correlation is the scale at which two variables are positively, negatively, or not correlated.
  - $\rho = corr(x, y) = cov(x, y)/\sigma_x \sigma_y$, $-1 < \rho < 1$

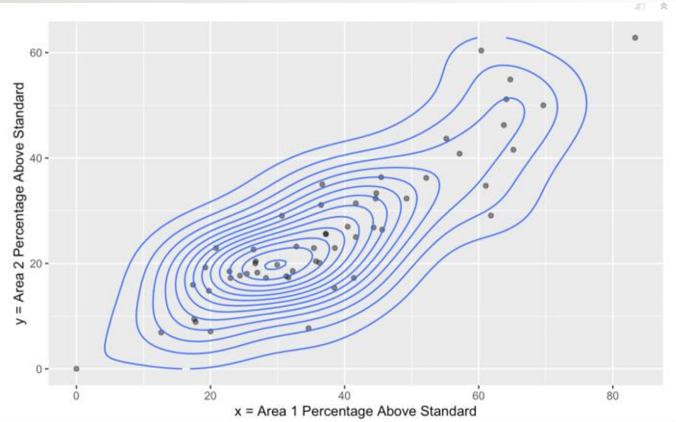
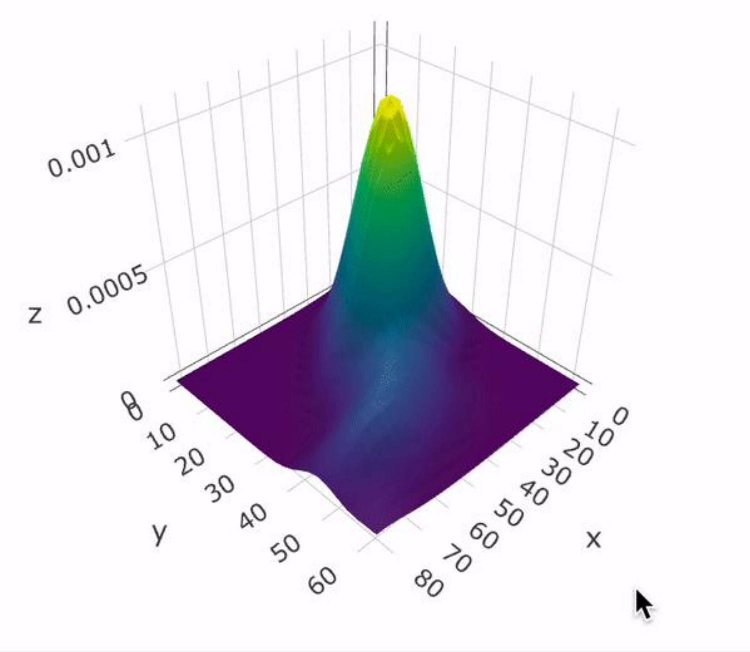
Methods - Code

- R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques.

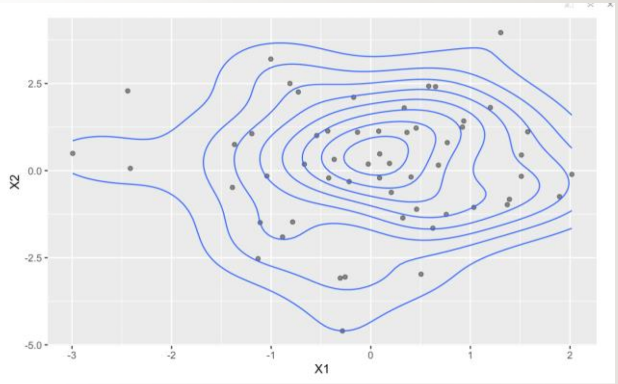
- This scatter plot from the previous slide was produced in R.



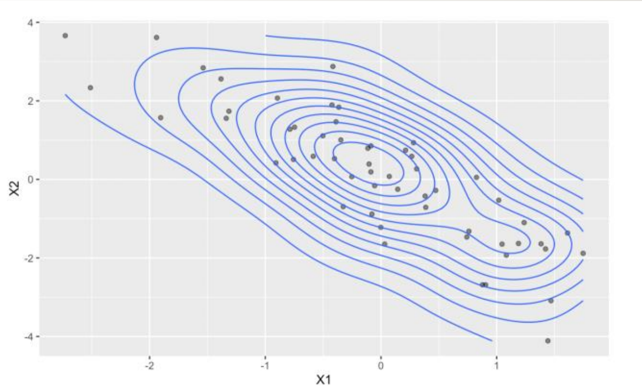
- R code to produce this scatter plot

```
```{r}
xx2<-(as.numeric(LBT2$`Area 1 Percentage Above Standard`))
xx3<-(as.numeric(LBT2$`Area 2 Percentage Above Standard`))


p2 <- ggplot(LBT2, aes(x = xx2, y = xx3)) +
  geom_point(alpha = .5) +
  #geom_density_2d() +
  labs(x="x = Area 1 Percentage Above Standard",y="y = Area 2 Percentage Above Standard")

p2
```
```

Methods - Code




````
```{r}
p2 <- ggplot(LBT2, aes(x = xx2, y = xx3)) +
  geom_point(alpha = .5) +
  geom_density_2d() +
  labs(x="x = Area 1 Percentage Above Standard",y="y = Area 2 Percentage Above Standard")

p2
```
````




````
```{r}
dens <- kde2d(xx2, xx3, h=c(28,28))

plot_ly(x = dens$x,
        y = dens$y,
        z = dens$z) %>% add_surface()
```
````



- Covariance is related to the shape of the bivariate function.



Methods – Theory and Code

- Linear Regression is a model that assumes a linear relationship between two random variables. The analysis is used to predict the value of one variable based on another variable.
  - Variables X1 and X2 in previous scatter diagram
- The linear relationship between the two random variables are expressed in the form
  - $y = \beta_0 + \beta_1 x + \varepsilon$
- Use the Least Squares Method to predict β_0 *and* β_1
- The linear regression formula predicts β_0 *and* β_1 in R and provides other valuable statistical information.

```
Call:
lm(formula = LBT2$`Mean Scale Score` ~ xx2 + xx3, data = LBT2)

Residuals:
    Min      1Q  Median      3Q     Max 
-62.399  -3.832   0.409   7.982  17.777 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2414.5995     4.0251 599.885  < 2e-16 ***
xx2            1.4112     0.2184   6.461 3.32e-08 ***
xx3            0.7910     0.2691   2.940  0.00486 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.88 on 53 degrees of freedom
Multiple R-squared:  0.8903,	Adjusted R-squared:  0.8862 
F-statistic: 215.1 on 2 and 53 DF,  p-value: < 2.2e-16
```

Future Work

Gaussian Process Regression in Educational Data

- Gaussian process regression is the combination of looking at the Gaussian distribution as a flexible tool for modeling complex relationships that are challenging for linear regression to model.

- Use this process to study and analyze actual data sets.
  - $\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix})$
  - $\kappa(X,X') = \sigma_f^2 exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\frac{1}{\ell_d^2}(x_d - x_d{}')^2\right\}$
  - (X,y) data used to build the model
  - (X^*, f_*) data used to validate the model
  - K is covariance based on our choice of covariance function κ

References

- "Multivariate Normal Distribution." *Wikipedia*, Wikimedia Foundation, 21 July 2021, en.wikipedia.org/wiki/Multivariate_normal_distribution.
- "The R Project for Statistical Computing." *R*, www.r-project.org/.
- Wackerly, Dennis D., et al. *Mathematical Statistics with Applications*. Brooks/Cole, 2012.
- Yin, Min, et al. "Key Course Selection for Academic Early Warning Based on Gaussian Processes." *Lecture Notes in Computer Science*, 2016, pp. 240–247., doi:10.1007/978-3-319-46257-8_26.
- *YouTube*, YouTube, 11 Apr. 2021, www.youtube.com/watch?v=N-bI-Dsm-rw&t=1675s.

Acknowledgements

Dr. Kagba Suaray

Dr. Janette Mariscal

McNair and Trio Staff

RONALD E. McNAIR POSTBACCALAUREATE ACHIEVEMENT PROGRAM
McNAIR
SCHOLARS PROGRAM
DIVISION OF STUDENT AFFAIRS

TRIO
STUDENT SUPPORT SERVICES
DIVISION OF STUDENT AFFAIRS