Homework_14

December 9, 2019

In [2]:
```
# HIDDEN
from datascience import *
from prob140 import *
import numpy as np
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
%matplotlib inline
from scipy import stats
```

1 Homework 14

1.0.1 1. 2017 SAT Scores, Part 1

The College Board presents an annual report about participation and performance in the SAT. The 2017 SAT had two parts: Evidence-Based Reading and Writing (ERW) and Math. It is generally believed that the joint distribution of ERW and Math scores of test-takers in any given year is roughly bivariate normal. This can only be a rough approximation because SAT scores aren't continuous variables. But let's make a lazy choice for this exercise: pretend that SAT scores are continuous and that the approximation is exact.

So assume that for the 2017 test-takers in the United States, the joint distribution of ERW and Math scores is bivariate normal. The College Board provides the following summary statistics.

- ERW: Mean 533, SD 100
- Math: Mean 527, SD 107
- Total: Mean 1060, SD 195

Let R be the ERW score and M the math score of a test-taker picked at random.
a) Find the correlation $r(R, M)$.

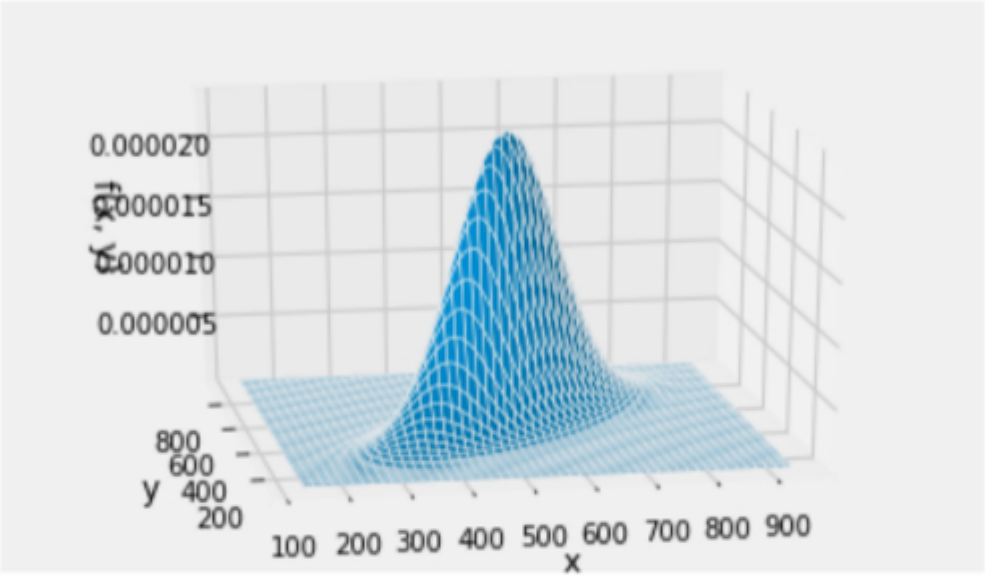
In [4]:
```
#1a
var_R = 100**2
var_M = 107**2
var_T = 195**2
r = (var_T - var_R - var_M) / (2 * 100 * 107)
r
```

Out[4]:
```
0.7745794392523364
```

b) Plot the joint density surface of R and M. The call is `Plot_bivariate_normal(mu, cov)` where the mean vector `mu` is a list and the covariance matrix `cov` is a list of lists specifying the rows.

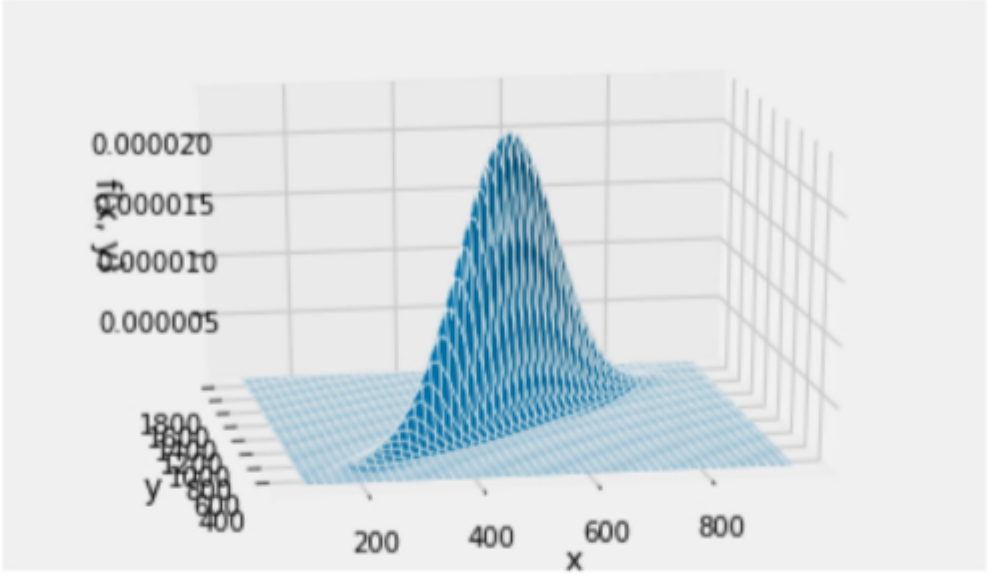
In [6]:
```
#1b

mu = [533, 527]
cov_RM = r * 100 * 107
cov = [[var_R, cov_RM], [cov_RM, var_M]]
Plot_bivariate_normal(mu, cov, figsize=(6,4))
```



c) [WRITTEN & CODE] Let $T = R + M$. In the code cell below, plot the joint density surface of M and T. Use as many lines of code as you need. **On paper, explain** your choice of density, and say what you notice about the plot.

In [8]:
```
#1c
mu2 = [527, 1060]
cov_MT = cov_RM + var_M
cov2 = [[var_M, cov_MT], [cov_MT, var_T]]
Plot_bivariate_normal(mu2, cov2, figsize=(6, 4))
```



d) Write a code expression that evaluates to $r(M, T)$. In the comment, explain why it's high.

```
In [10]: # 1d

         """r(M, T) is high because the math score is a component of the total score."""
         cov_MT / ((var_M ** 0.5) * (var_T ** 0.5))
```

```
Out[10]: 0.9459381739755571
```

1.0.2 2. 2017 SAT Scores, Part 2

This exercise continues the previous one. Please make the same assumptions as before. In your math and code you can use any quantity that has already been defined.

As in the previous exercise, the parts below refer to the scores of a test-taker picked at random.

a) Write an expression that evaluates to the chance that the total score is greater than 1500.

```
In [12]: 1 - stats.norm.cdf((1500 - 1060) / (var_T ** 0.5))
```

```
Out[12]: 0.012022474952684048
```

b) [WRITTEN & CODE] In terms of the standard normal cdf Φ, what is the chance that the student scored higher on ERW than on Math? Answer this on paper (show your work) and then use the code cell below to find the numerical value. Use as many lines of code as you need. The last expression should evaluate to the proportion.

1.0.3 Solution

$P(R > M) = P(D > 0)$ where $D = R - M$.

$\mu_D = E(D) = 533 - 527$

$\sigma_D^2 = Var(D) = Var(R) + Var(M) - 2Cov(R, M)$, all of which we have already calculated.

So we can compute $P(D > 0) = 1 - \Phi\left(\frac{0-\mu_D}{\sigma_D}\right)$

```
In [14]: #2b

         # D = R - M

         mu_D = 533 - 527
         var_D = var_R + var_M - 2*cov_RM
         sigma_D = var_D ** 0.5
         1 - stats.norm.cdf((0 - mu_D) / sigma_D)
```

```
Out[14]: 0.5342474813742879
```

c) Write a code expression that evaluates to the chance that the ERW and Math scores were more than 100 points apart. As before, your expression can involve quantities defined in earlier parts.

```
In [16]: #2c

         stats.norm.cdf((-100 - mu_D) / sigma_D) + (1 - stats.norm.cdf((100 - mu_D) / sigma_D)
```

```
Out[16]: 0.1535063671419105
```

1.0.4 3. Heights of Mothers and Daughters

The heights of a population of mother-daughter pairs have a bivariate normal distribution with correlation 0.5.

a) Of the mothers on the 90th percentile of mothers' heights, what proportion have daughters who are taller than the 90th percentile of daughters' heights?

b) In what proportion of mother-daughter pairs are both women taller than average? (This means the mothers are taller than the average mother and the daughters are taller than the average daughter.)

[Hint: Remember that you can express standard bivariate normal variables in terms of two independent standard normal variables.]

1.0.5 Solution

For a random pair, let M^* be the mother's height in standard units and let D^* be the daughter's height in standard units.

3a) The 90th percentile of the standard normal curve is $\Phi^{-1}(0.9)$ which is about 1.28.

Given $M^* = 1.28$, the conditional distribution of D^* is normal with mean 0.5×1.28 and variance $1 - 0.5^2$. That's normal $(0.64, 0.75)$.

$P(D^* > 1.28 \mid M^* = 1.28) = 1 - \Phi((1.28 - 0.64)/\sqrt{0.75}) \approx 0.23$

3b) We want $P(M^* > 0, D^* > 0)$. Since M^* and D^* are standard bivariate normal with $\rho = 0.5$, we can write $D^* = \frac{1}{2}M^* + \frac{\sqrt{3}}{2}Z$ where Z is standard normal independent of M^*.

$P(M^* > 0, D^* > 0) = P(M^* > 0, \frac{1}{2}M^* + \frac{\sqrt{3}}{2}Z > 0) = P(M^* > 0, Z > -\frac{1}{\sqrt{3}}M^*)$

M^* and Z are i.i.d. standard normal so their joint density has circular symmetry. So we have a "slices of a normal cake" problem. Draw the region and do the trig to see that the ratio of angles is $(90 + 30)/360 = 1/3$.

1.0.6 4. Least Squares Linear Predictor

Suppose that X is normal (μ_X, σ_X^2), Y is normal (μ_Y, σ_Y^2), and the two random variables are independent. Let $S = X + Y$.

a) Find the conditional distribution of X given $S = s$.

b) Find the least squares predictor of X based on S and provide its mean squared error.

c) Find the least squares linear predictor of X based on S and provide its mean squared error.

1.0.7 Solution

4a) Because X and Y are independent normal variables, the joint distribution of X and S is bivariate normal with means μ_X and $\mu_S = \mu_X + \mu_Y$, variances σ_X^2 and $\sigma_S^2 = \sigma_X^2 + \sigma_Y^2$, and correlation ρ calculated using $Cov(X, S) = Var(X)$ to get

$$\rho = \frac{\sigma_X^2}{\sigma_X \sqrt{\sigma_X^2 + \sigma_Y^2}} = \sqrt{\frac{\sigma_X^2}{\sigma_X^2 + \sigma_Y^2}}$$

The conditional distribution of X given $S = s$ is normal with mean $\rho \frac{\sigma_X}{\sigma_S}(s - \mu_S) + \mu_X$ and variance $(1 - \rho^2)\sigma_X^2$.

4b) The best predictor is $E(X \mid S)$ which is $\rho \frac{\sigma_X}{\sigma_S}(S - \mu_S) + \mu_X$ by (a). Also by (a), the MSE is $E(Var(X \mid S)) = (1 - \rho^2)\sigma_X^2$.

4c) Same as answer to (b) because the best predictor is linear in this case.

1.0.8 5. Properties of Multiple Regression Estimates

This exercise assumes the multiple regression model of Section 25.4 of the textbook and uses the same notation as in that section.

a) What are the dimensions of $E(\hat{\boldsymbol{\beta}})$? Show that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

b) Find the covariance matrix of $\hat{\boldsymbol{\beta}}$. The diagonal entries of this matrix tell you how variable the estimates of the coefficients are.

c) What is the distribution of $\hat{\boldsymbol{\beta}}$?

d) Find the distribution of the fitted values $\hat{\mathbf{Y}}$.

e) Find the distribution of the residuals $\mathbf{e}$.

The answers are based on $\mathbf{Y}$ being multivariate normal with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\sigma^2\mathbf{I}$.

a) $p \times 1$. $E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ so $\hat{\boldsymbol{\beta}}$ is unbiased.

b) The covariance matrix of $\hat{\boldsymbol{\beta}}$ is $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

c) Multivariate normal with the parameters in Parts **a** and **b**, because it is a linear combination of independent normal variables.

d) Multivariate normal because $\mathbf{Y}$ is a linear transformation of $\boldsymbol{\beta}$. The mean is $\mathbf{X}E(\hat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}$. The covariance matrix is $\mathbf{X}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

e) $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y} = (\mathbf{I} - \mathbf{H})Y$ is a linear transformation of $\mathbf{Y}$ and hence is multivariate normal. The mean is $E(\mathbf{Y}) - E(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. The covariance matrix is $\sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T$.

This is not required for the solution, but: note that $\mathbf{H}$ is symmetric, and hence so is $(\mathbf{I} - \mathbf{H})$. So $\sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})$, and if you do the algebra you will see that $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H}$.

1.1 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

1.1.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using an application. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.

1.1.2 Code Portion

- Save your notebook using File > Save and Checkpoint.
- Generate a PDF file using File > Download as > PDF via LaTeX. This might take a few seconds and will automatically download a PDF version of this notebook.
  - If you have issues, please make a follow-up post on the general HW 14 Piazza thread.

1.1.3 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Homework 14 on Gradescope.
- **Make sure to assign each page of your pdf to the correct question.**
- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

1.1.4 We will not grade assignments which do not have pages selected for each question.