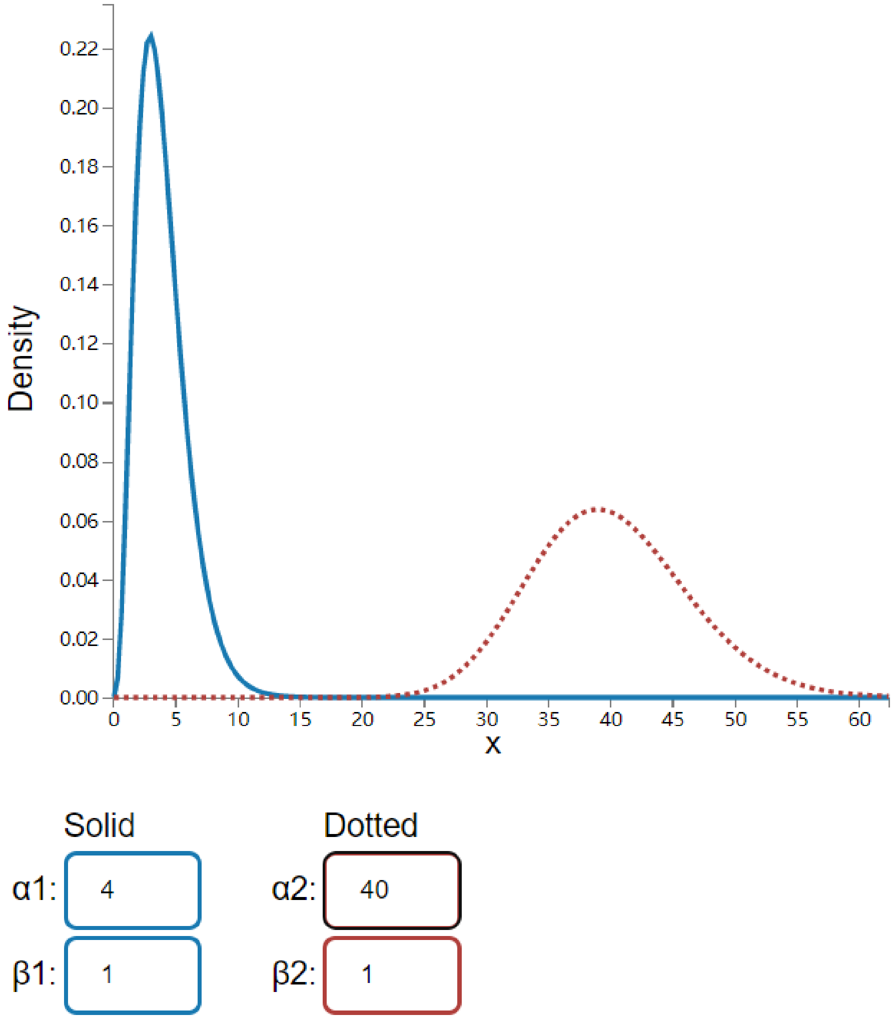
Computer assignment

Wuhao Wang(wuhwa469)

10/23/2021

1 book assignment

4.84



Solid

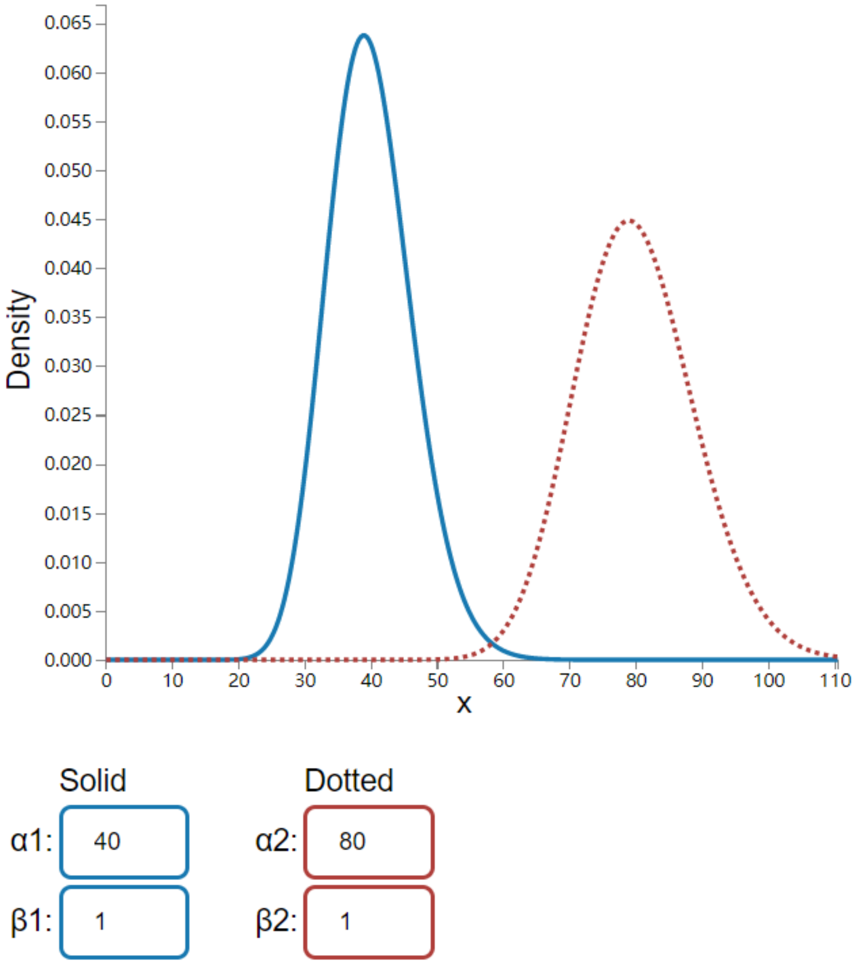
α1: 4

β1: 1

Dotted

α2: 40

β2: 1



Solid

α1: 40

β1: 1

Dotted

α2: 80

β2: 1

a)

As alpha increases so does the abscissa of the highest point, and the width of the bulge decreases. When alpha=4, the shape is more skewed.When alpha = 80, the shape is more symmetric.

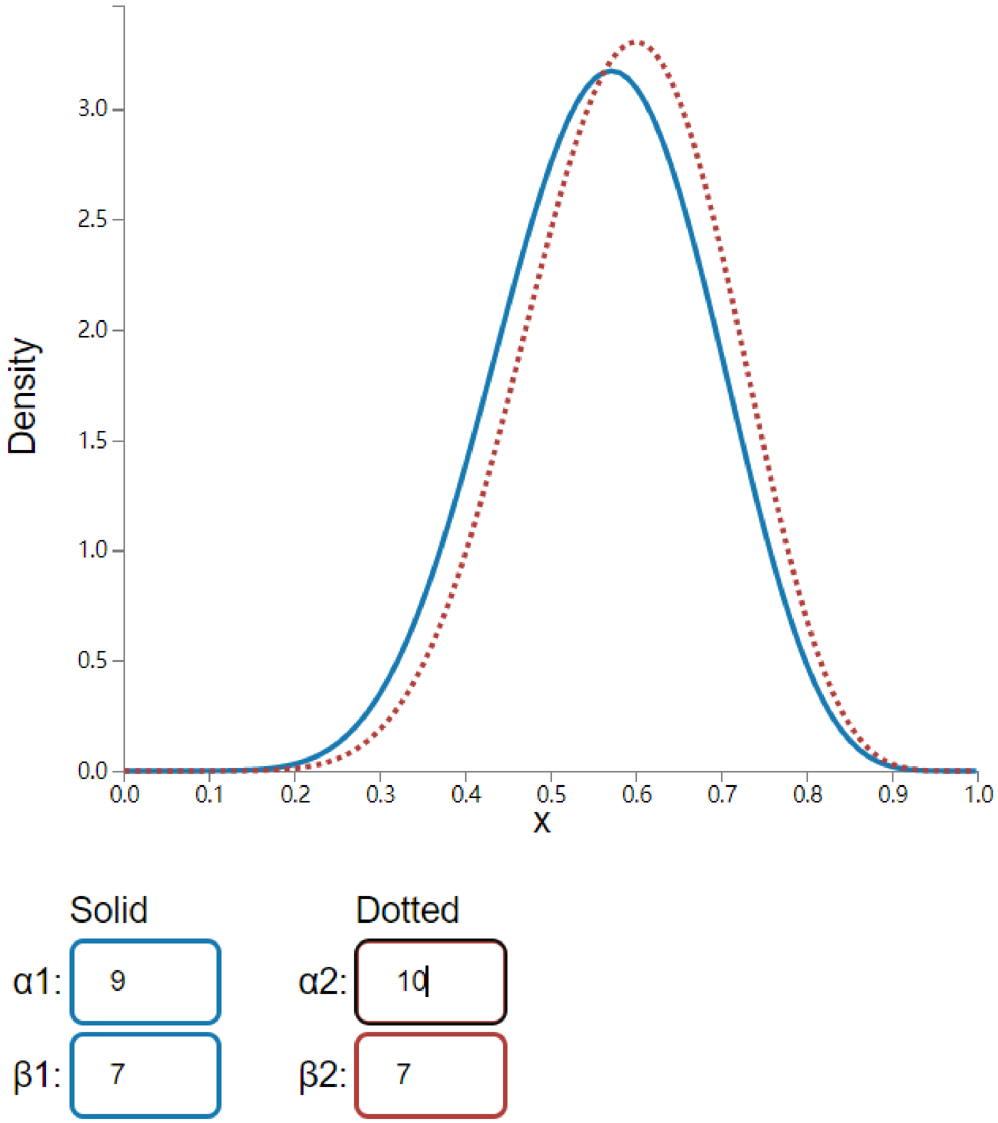
b)

The center of each picture moves right as alpha increases.

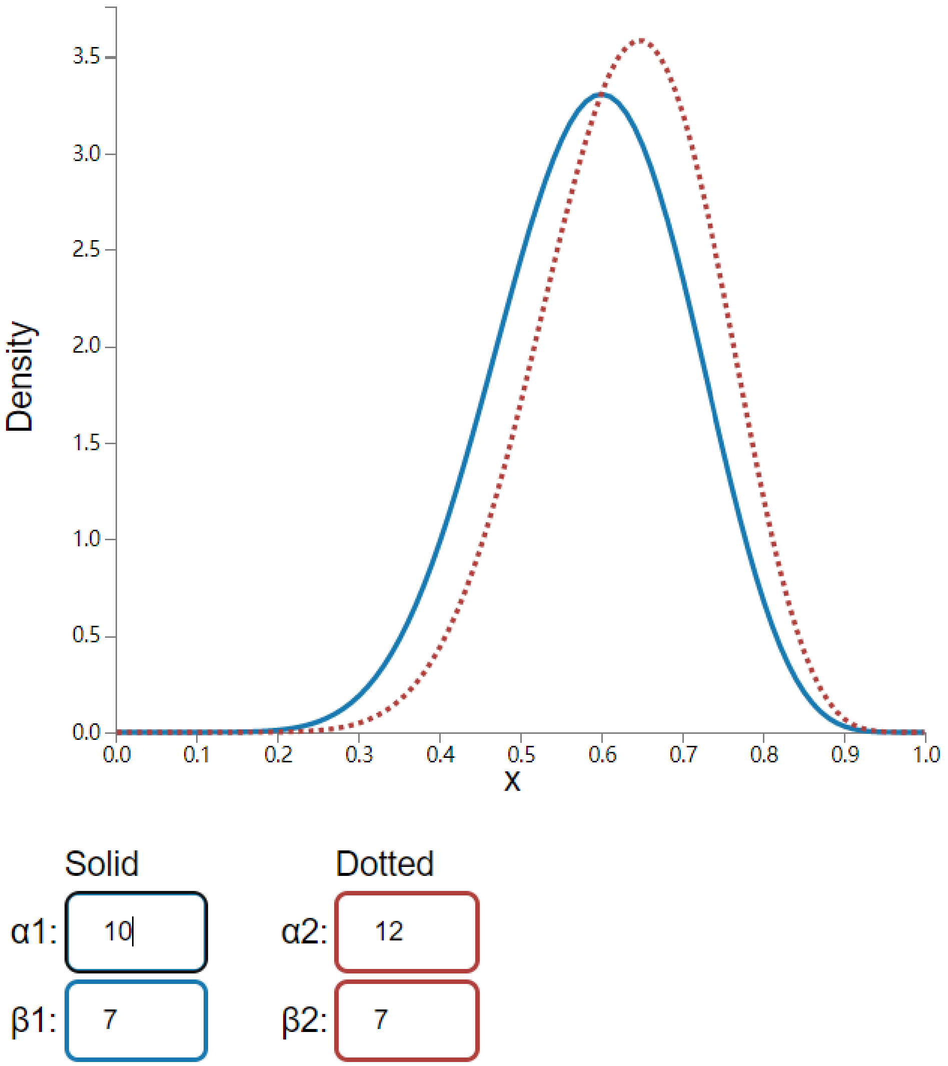
c)

The parameter alpha is used to control the shape of gamma distribution, so change the value of alpha will change the shape.

4.117



Solid		Dotted	
α1:	9	α2:	10
β1:	7	β2:	7



Solid

α1: 10

β1: 7

Dotted

α2: 12

β2: 7

a)

These densities are skewed right.

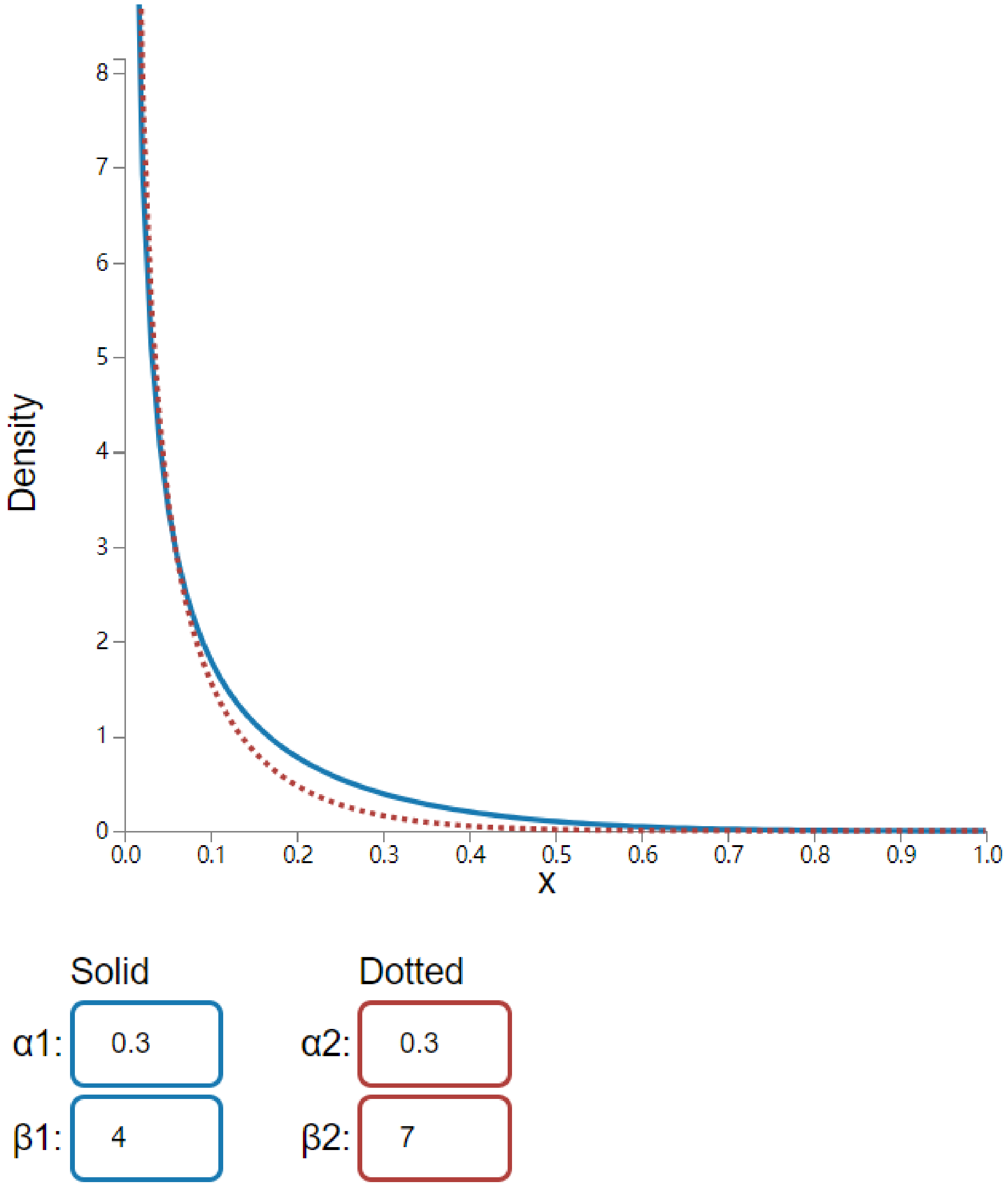
b)

the center of densities moves to right, and the peak value increases.

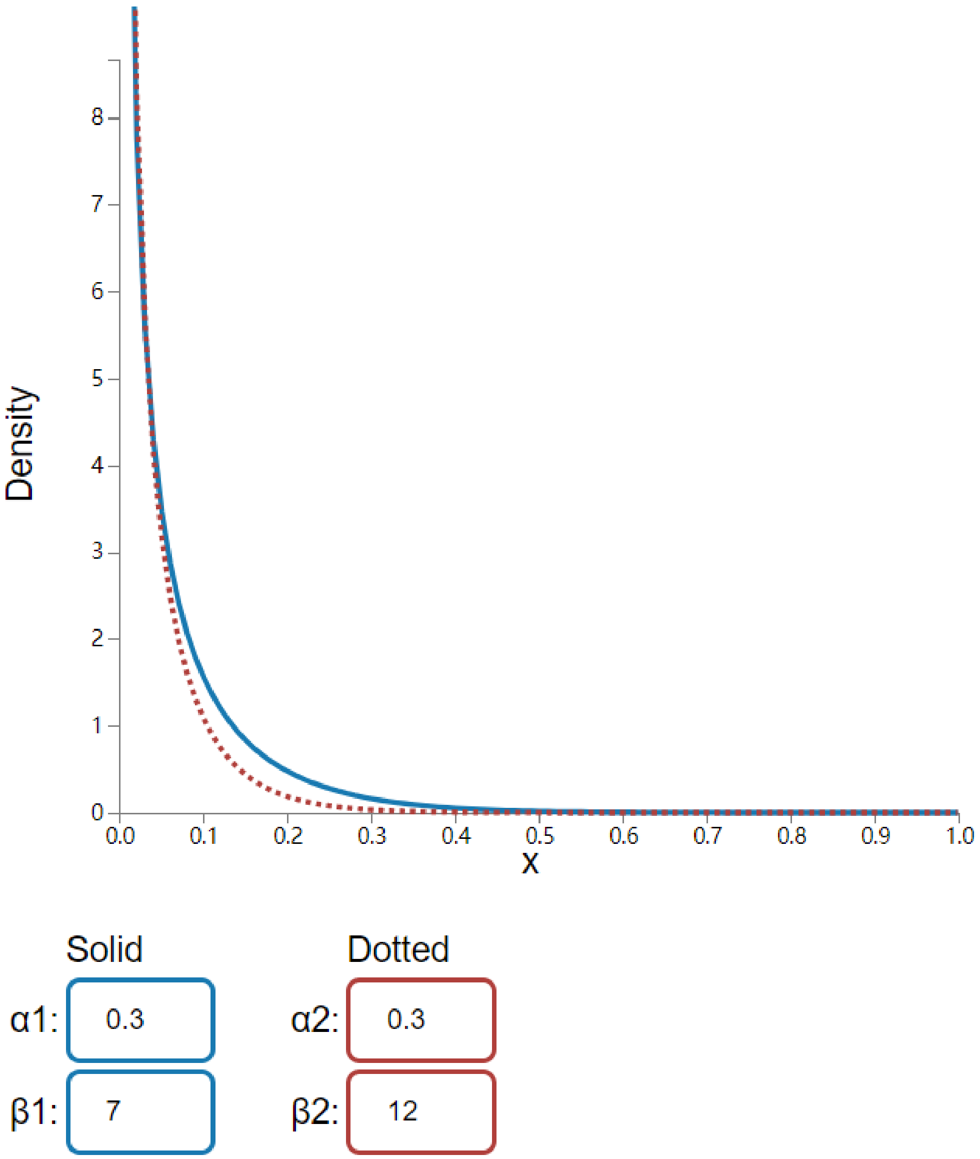
c)

if alpha > beta, then densities will be skewed right. As alpha/beta increases,the center point on the x-coordinate gradually close to 1. Further, I think the center point on the x-coordinate is equal to alpha/ (alpha + beta).

4.118



Solid		Dotted	
α1:	0.3	α2:	0.3
β1:	4	β2:	7



Solid

α1: 0.3

β1: 7

Dotted

α2: 0.3

β2: 12

a)

These densities are skewed left.

b)

the densities become more steep.

c)

the densities where beta = 4.

d)

the bigger beta/alpha is the steeper density function curve will be.

10.19

The output voltage for an electric circuit is specified to be 130. A sample of 40 independent readings on the voltage for this circuit gave a sample mean 128.6 and standard deviation 2.1. Test the hypothesis that the average output voltage is 130 against the alternative that it is less than 130. Use a test with level .05.

Solution: For this test, the deviation is already known, so we can use the T = (128.6-130)/(2.1/squt(40)), the reject region is less than -z0.5 = -1.645

H0: the average voltage is not less than 130; Ha: the average voltage is less than 130

```
t <- (128.6-130)/(2.1/(sqrt(40)))
t

## [1] -4.21637
```

t < -1.645, so reject H0

10.21

Shear strength measurements derived from unconfined compression tests for two types of soils gave the results shown in the following table (measurements in tons per square foot). Do the soils appear to differ with respect to average shear strength, at the 1% significance level?

Solution: H0: appear same ; Ha: appear to differ with respect to average shear strength.

```
t <- (1.65-1.43)/sqrt((0.26**2)/30+(0.22**2)/35)
t

## [1] 3.648374
```

t > 2.576, so we reject H0.

11.31

According to the problem, we set H0: beta1 = 0 against Ha: beta1 != 0, since the p-value is quite small, we reject H0.

Solution:

x is 19.1, 38.2, 57.3, 76.2, 95, 114, 131, 150, 170 y is .095, .174, .256, .348, .429, .500, .580, .651, .722

```
x_1131 <- c(19.1, 38.2, 57.3, 76.2, 95, 114, 131, 150, 170)
y_1131 <- c(.095, .174, .256, .348, .429, .500, .580, .651, .722)
summary(lm(y_1131~x_1131))
```

```
## 
## Call:
## lm(formula = y_1131 ~ x_1131)
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -0.0133264 -0.0042777 -0.0000231  0.0080557  0.0098107 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.875e-02  6.129e-03   3.059   0.0183 *  
## x_1131      4.215e-03  5.771e-05  73.040 2.37e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.008376 on 7 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9985 
## F-statistic:  5335 on 1 and 7 DF,  p-value: 2.372e-11
```

the fitted liner model is y^ = .01875 + .004215x,p-value is quite small(2.372e-11), so we reject H0, thus peak current increases as nickel concentrations increase

11.69

Solution: x is -7, -5, -3, -1, 1, 3, 5, 7 y is 18.5, 22.6, 27.2, 31.2, 33.0, 44.9, 49.4, 35.0 x^2 is 49, 25, 9, 1, 1, 9, 25, 49

```
x_1169 <- c(-7, -5, -3, -1, 1, 3, 5, 7)
y_1169 <- c(18.5,22.6,27.2,31.2,33.0,44.9,49.4,35.0)
x2_1169 <- x_1169**2
lm(y_1169~x_1169)
```

```
##
## Call:
## lm(formula = y_1169 ~ x_1169)
##
## Coefficients:
## (Intercept)       x_1169
##      32.725        1.812
```

```
lm(y_1169~x_1169+x2_1169)
```

```
##
## Call:
## lm(formula = y_1169 ~ x_1169 + x2_1169)
##
## Coefficients:
## (Intercept)       x_1169      x2_1169
##     35.5625       1.8119      -0.1351
```

a) the fitted liner model is y = 32.725 +1.812x

b) the fitted liner model is y = 35.5625 +1.8119x-0.1351x^2

2 imputation techs

a) Which type of missing mechanism do you prefer to get a good imputation?

The simplest one: Missingness completely at random.

b) Say something about simple random imputation and regression imputation of a single variable.

SIMPLE RANDOM IMPUTATION

1) impute missing values of earnings based on the observed data for this variable.

Pick data from observed data. This method will be seriously biased if this variable has several types or regions. For example, in the iris dataset, the variable Petal.Length has obviously two groups, the one is about 1~2, the other one is above 5. If the missing values happen to all belong to the same category and there are a lot of them, then the randomly selected numbers are likely to change the distribution of the category itself.

2) Zero coding and topcoding

It is like classifying data into two class. This method will destroy the distribution of continuous variable. Like what just mentioned in method 1.

REGRESSION IMPUTATION

Using those observation without missing data to do liner regression, target variable is the one contains missing data, other variable act as independent variable. This method is more reasonable than simple those in simple random imputation, it takes other variable in consideration. In other words, this method uses statistic method to find relationship between the variable (with missing data) and others.

c) Explain shortly what Multiple Imputation is.

Multiple imputation is about filling dataset with several variables containing missing data. Under such circumstance, we must think of the dataset as a multivariate outcome, any components of which can be missing. It is like multiple variable regression.