Computer Assignment - Statistical methods

YiHung Chen

2022-10-25

1.Computer exercises from Course's book

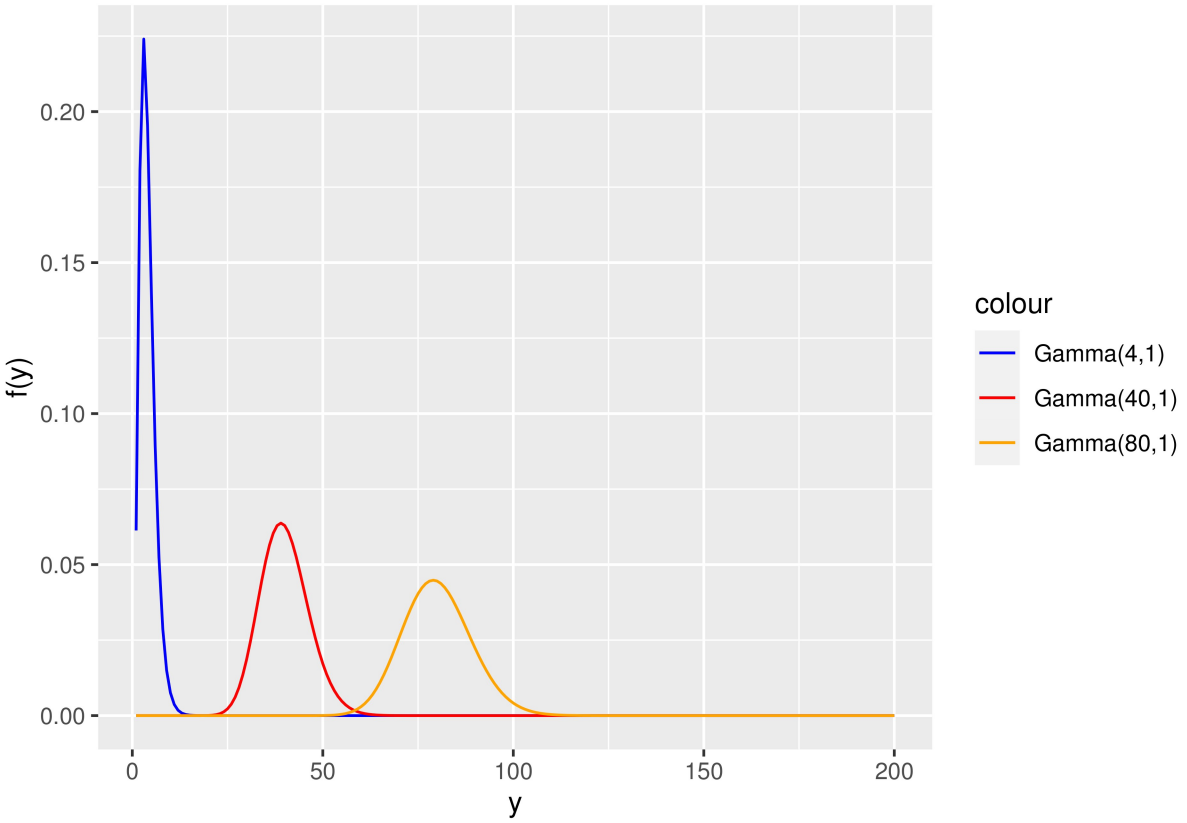
Ex.4.84

Applet Exercise Refer to Exercise 4.83. Use the applet Comparison of Gamma Density Functions to compare gamma density functions with ($\alpha = 4$, $\beta = 1$), ($\alpha = 40$, $\beta = 1$), and ($\alpha = 80$, $\beta = 1$).

```
Gamma <- function(alpha, beta){
  y <- 1:200
  (1/gamma(alpha)*beta^alpha)*(y^(alpha-1))*(exp(1)^(-y/beta))
}

Ga1 <- Gamma(alpha = 4, beta = 1)
Ga2 <- Gamma(alpha = 40, beta = 1)
Ga3 <- Gamma(alpha = 80, beta = 1)
y <- 1:200
df <- data.frame(Ga1,Ga2,Ga3)

colors <- c("Gamma(4,1)" = "blue", "Gamma(40,1)" = "red", "Gamma(80,1)" = "orange")
ggplot(df) + geom_line(aes(x=y, y=Ga1, colour = "Gamma(4,1)"))+
             geom_line(aes(x=y, y=Ga2, colour = "Gamma(40,1)"))+
             geom_line(aes(x=y, y=Ga3, colour = "Gamma(80,1)"))+
             ylab("f(y)") +
             scale_color_manual(values = colors)
```



Answer 4.84

(a) The line with larger α will looks more symmetric.
(b) The larger α is, the center of density function will toward right (larger).
(c) According to the observation from (b), the mean is increase when α increase

Ex.4.117

Applet Exercise Use the applet Comparison of Beta Density Functions to compare beta density functions with ($\alpha = 9$, $\beta = 7$), ($\alpha = 10$, $\beta = 7$), and ($\alpha = 12$, $\beta = 7$).

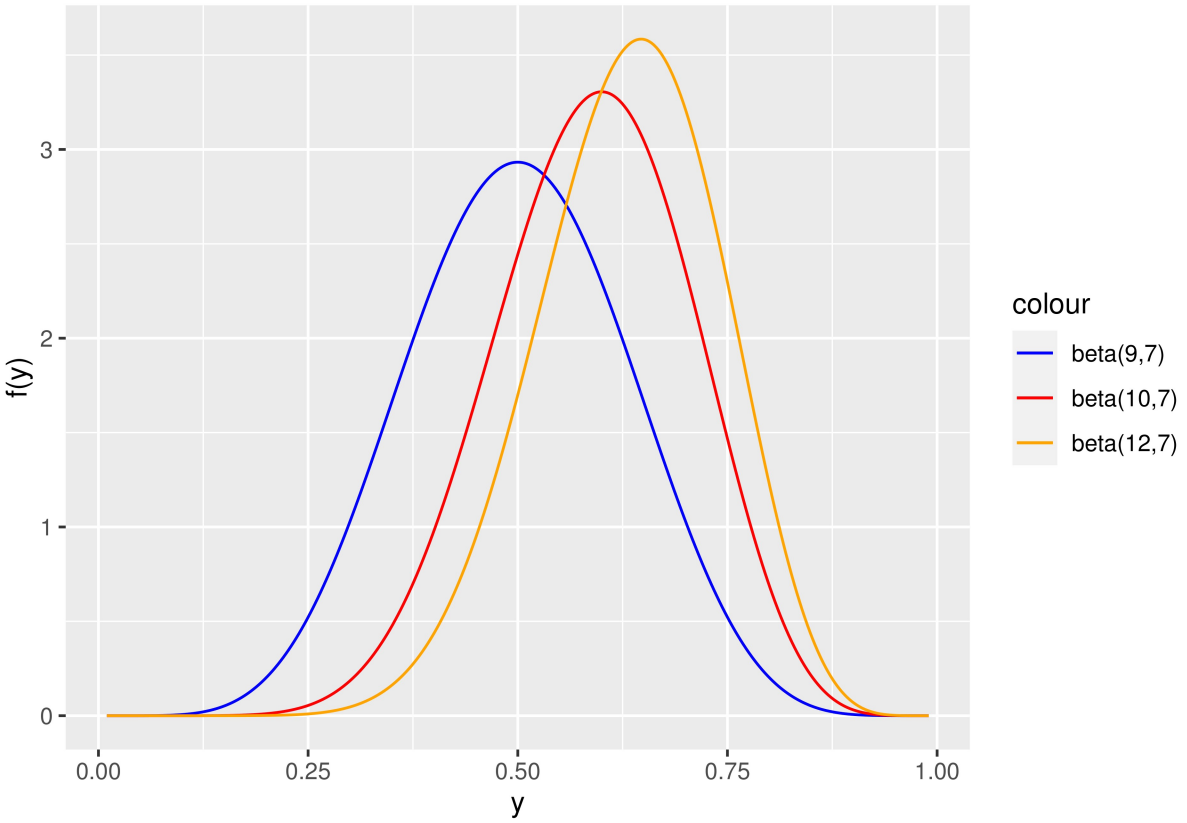
```
Beta <- function(alpha, beta,y){

  (1/beta(alpha,beta)*y^(alpha-1)*(1-y)^(beta-1))

}
y <- seq(0.01,0.99, by=0.001)
beta1 <- Beta(alpha = 7, beta = 7,y)
beta2 <- Beta(alpha = 10, beta = 7,y)
beta3 <- Beta(alpha = 12, beta = 7,y)

df <- data.frame(beta1,beta2,beta3)
colors <- c("beta(9,7)" = "blue", "beta(10,7)" = "red", "beta(12,7)" = "orange")
ggplot(df) + geom_line(aes(x=y, y=beta1, colour = "beta(9,7)"))+
```

```
            geom_line(aes(x=y, y=beta2, colour = "beta(10,7)"))+
            geom_line(aes(x=y, y=beta3, colour = "beta(12,7)"))+
            ylab("f(y)") +
            scale_color_manual(values = colors)
```

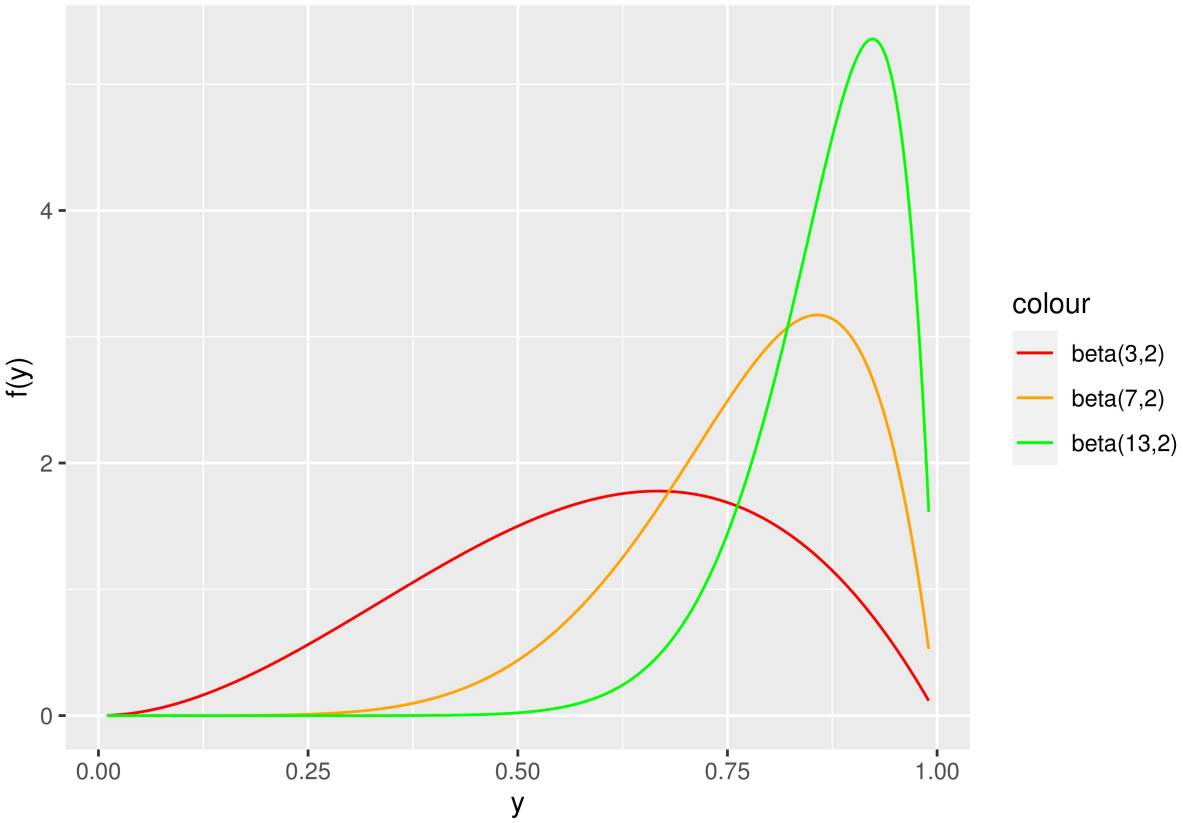


Answer 4.117

(a) Every densities are skewed left
(b) The density function become more un-symmetric when α become closer to 12(away from β).
(c) According to the graph below, the line skewed left when $\alpha > \beta$ and $\alpha > 1$ and $\beta > 1$.

```
y <- seq(0.01,0.99, by=0.001)
beta1 <- Beta(alpha = 3, beta = 2,y)
beta2 <- Beta(alpha = 7, beta = 2,y)
beta3 <- Beta(alpha = 13, beta = 2,y)


df <- data.frame(beta1,beta2,beta3)
colors <- c( "beta(3,2)" = "red", "beta(7,2)" = "orange","beta(13,2)" = "green")
ggplot(df) + geom_line(aes(x=y, y=beta1, colour = "beta(3,2)"))+
            geom_line(aes(x=y, y=beta2, colour = "beta(7,2)"))+
            geom_line(aes(x=y, y=beta3, colour = "beta(13,2)"))+
            ylab("f(y)") +
            scale_color_manual(values = colors)
```

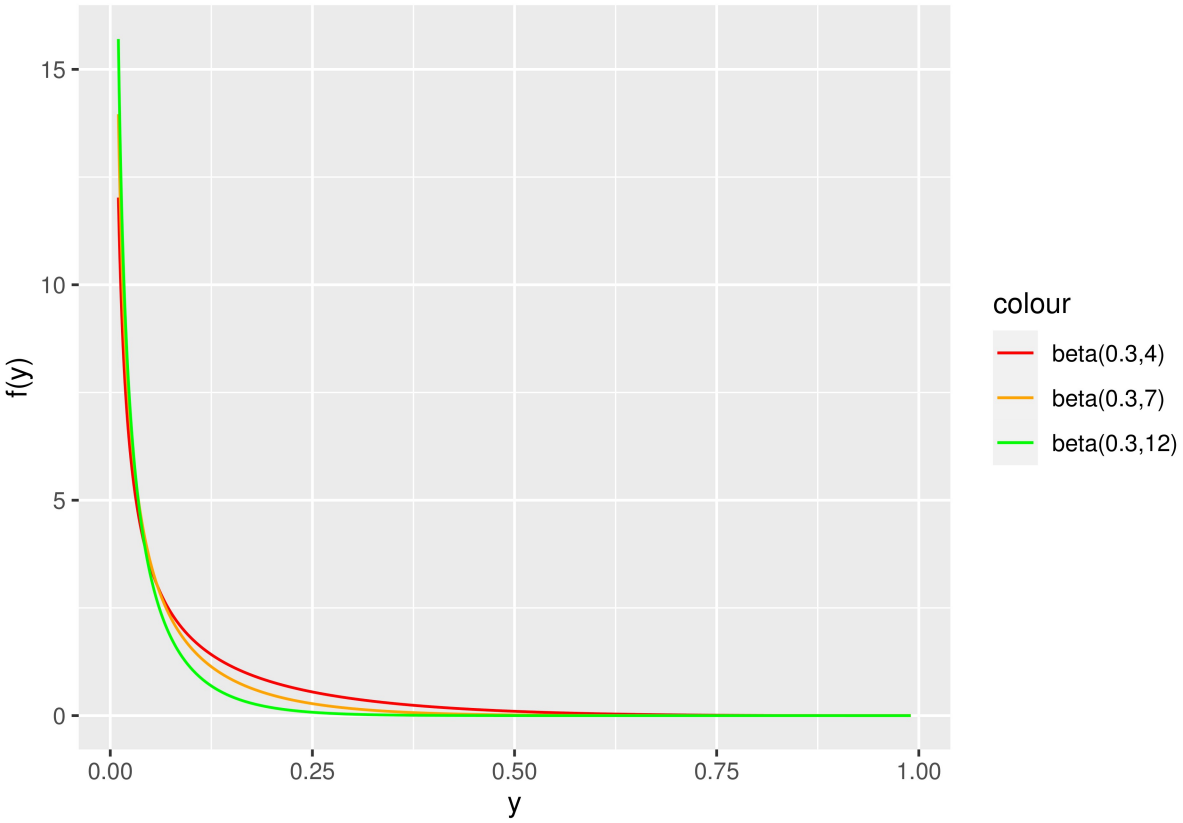


Ex.4.118

Applet Exercise Use the applet Comparison of Beta Density Functions to compare beta density functions with ($\alpha = .3$, $\beta = 4$), ($\alpha = .3$, $\beta = 7$), and ($\alpha = .3$, $\beta = 12$).

```
y <- seq(0.01,0.99, by=0.001)
beta1 <- Beta(alpha = 0.3, beta = 4,y)
beta2 <- Beta(alpha = 0.3, beta = 7,y)
beta3 <- Beta(alpha = 0.3, beta = 12,y)


df <- data.frame(beta1,beta2,beta3)
colors <- c( "beta(0.3,4)" = "red", "beta(0.3,7)" = "orange","beta(0.3,12)" = "green")
ggplot(df) + geom_line(aes(x=y, y=beta1, colour = "beta(0.3,4)"))+
             geom_line(aes(x=y, y=beta2, colour = "beta(0.3,7)"))+
             geom_line(aes(x=y, y=beta3, colour = "beta(0.3,12)"))+
             ylab("f(y)") +
             scale_color_manual(values = colors)
```

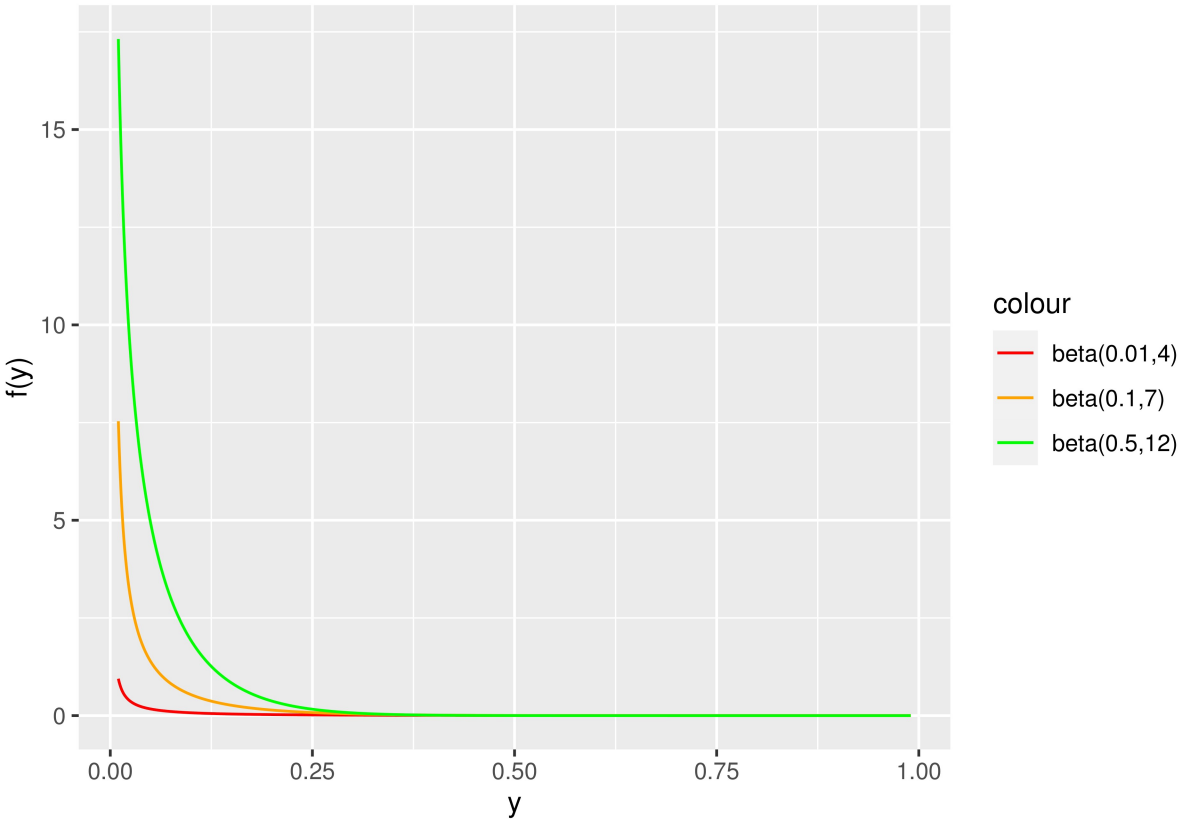


Answer 4.118

(a) Every densities are skewed right
(b) When β become closer to 12, the spread decrease.
(c) With $\alpha = 0.3(\beta{=}4)$ the distributions gives the highest probability.
(d) According to the graph below, the line skewed right when $\alpha < 1$ and $\beta > 1$.

```
y <- seq(0.01,0.99, by=0.001)
beta1 <- Beta(alpha = 0.01, beta = 4,y)
beta2 <- Beta(alpha = 0.1, beta = 7,y)
beta3 <- Beta(alpha = 0.5, beta = 12,y)


df <- data.frame(beta1,beta2,beta3)
colors <- c( "beta(0.01,4)" = "red", "beta(0.1,7)" = "orange","beta(0.5,12)" = "green")
ggplot(df) + geom_line(aes(x=y, y=beta1, colour = "beta(0.01,4)"))+
             geom_line(aes(x=y, y=beta2, colour = "beta(0.1,7)"))+
             geom_line(aes(x=y, y=beta3, colour = "beta(0.5,12)"))+
             ylab("f(y)") +
             scale_color_manual(values = colors)
```



Ex10.19

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.

Answer 10.19

Calculate Z using (sample_mean-population_mean)/standard_deviation/sqrt(sample_size)

```
sample_size <- 40
sample_mean <- 128.6
population_mean <- 130
sd <- 2.1
z <- (sample_mean-population_mean)/(sd/sqrt(sample_size))
z_005 <- qnorm(0.05,mean=0,sd=1, lower.tail = T)

cat("the value of  Z =",z,"\nthe value of Z0.05 =",z_005,"\n")
```

```
## the value of  Z = -4.21637
## the value of Z0.05 = -1.644854
```

Since 0 > $Z0.05$ > Z, H0 is rejected.

Ex10.21

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?

Answer 10.21

```
sample_size1 <- 30
sample_mean1 <- 1.65
sd1 <- 0.26
sample_size2 <- 35
sample_mean2 <- 1.43
sd2 <- 0.22

z <- (sample_mean1-sample_mean2)/(sqrt((sd1^2/sample_size1)+(sd1^2/sample_size2)))

z_001 <- qnorm(0.01/2,mean=0,sd=1, lower.tail = F)
cat("the value of  Z =",z,"\nthe value of Z0.05 Z =",z_001)
```

```
## the value of  Z = 3.400849
## the value of Z0.05 Z = 2.575829
```

Since 0<$Z0.05$<Z, H0 is rejected Which means at 1% significance level, the soils differ with respect to average shear strength

Ex11.31

```
x <- c(19.1, 38.2, 57.3, 76.2, 95, 114, 131, 150, 170)
y <- c(0.095, 0.174, 0.256, 0.348, 0.429, 0.500, 0.580, 0.651, 0.722)
model<- lm(y~x)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## 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           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
```

Answer 11.31

As the p-value of x is very small(even if divided by 2), H_0 is rejected. Thus, the peak current increases as nickel concentrations increase

Ex 11.69

The manufacturer of Lexus automobiles has steadily increased sales since the 1989 launch of that brand in the United States. However, the rate of increase changed in 1996 when Lexus introduced a line of trucks. The sales of Lexus vehicles from 1996 to 2003 are shown in the accompanying table.

```
x <- c(-7,-5,-3,-1,1,3,5,7)
y <- c(18.5,22.6,27.2,31.2,33.0,44.9,49.4,35.0)
linear_model<- lm(y~x)
quadratic_model <-  lm(y~x+I(x^2))

linear_model
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      32.725        1.812
```

```
quadratic_model
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Coefficients:
## (Intercept)            x       I(x^2)
##     35.5625       1.8119      -0.1351
```

Answer 11.69

(a) According to the calculation, the model $\hat{y} = \beta_0 + \beta_1 \text{x} + \epsilon = 32.725 + 1.812\text{x}$

(b) According to the calculation, the model $\hat{y} = \beta_0 + \beta_1 \text{x} + \beta_2 x^2 + \epsilon = 35.5625 + 1.8119\text{x} - 0.1351x^2$

2. Imputations techniques

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

There are four general missing mechanisms, from the simplest to the most general:

(1) Missingness completely at random.
(2) Missingness at random.

(3) Missingness that depends on unobserved predictors.
(4) Missingness that depends on the missing value itself.

For good imputation, I think the "Missigness completely at random" is the best missing mechanism. Since the missing data is completely at random with the same probability distribution, one can remove the case with missing data and this will not bias the inferences.

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

Simple random imputation is the simplest way of imputation. It uses others data from the complete case and put the data randomly in the missing value, This is a convenient way to impute data. However, it is also NOT very useful since it ignores other useful information(data).

Regression imputation can be used to improve the downside of simple random imputation. It fit a regression model on the observed data then predict the missing values according to the model. However, it should be noted although using this method can gives out the most likely value of missing data but it does not supply the uncertainty of the value.

3. Explain shortly what Multiple Imputation is.

Multiple Imputation is a technique that is different from single variable imputation. This method replace missing data with several imputed values to reflect the uncertainty about the imputation model.

The multiple imputation takes prediction from different models to create imputed values, then put them into a "complete data set". After the complete data set being create, the analyze should be done to combined the inerences.