HW #4

Bilal Gilani

9/27/2020

1.

a.

```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100) / 10
y <- 2 + 2 * x2 + rnorm(100)
```

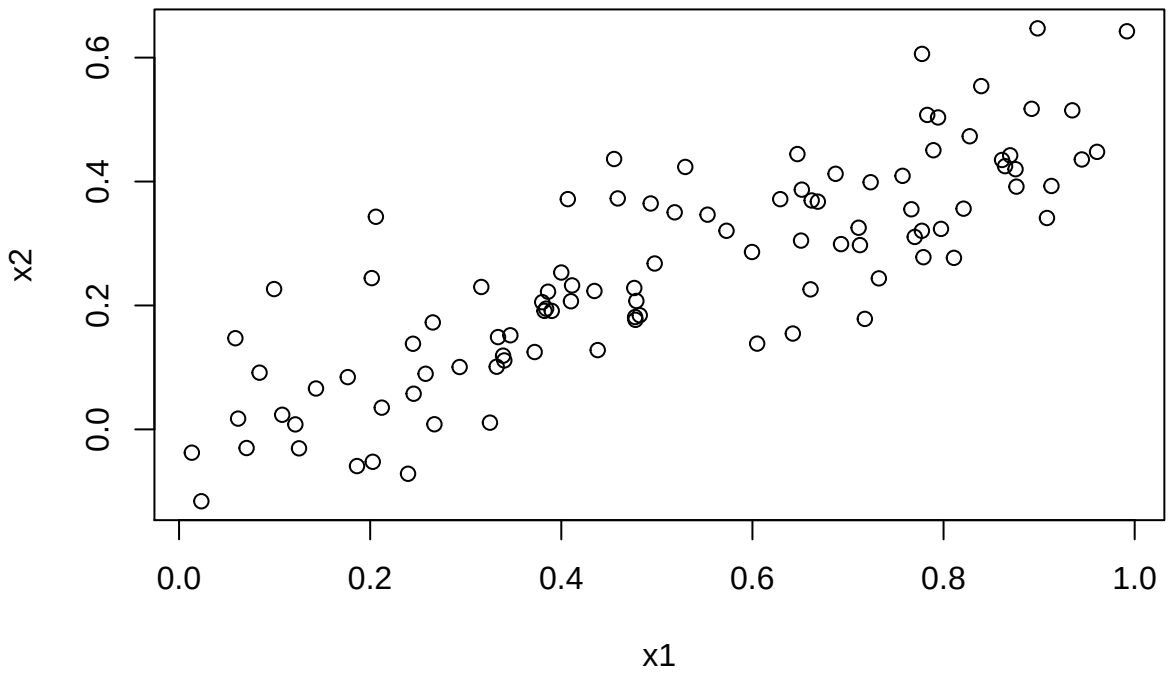
$Y = 2 + 2X_1 + 0.3X_2 + \epsilon$

b.

```
cor(x1, x2)
```

```
## [1] 0.8351212
```

```
plot(x1, x2)
```



c.

```
lm1 <- lm(y ~ x1 + x2)
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1           -0.5604     0.7212  -0.777   0.4390
## x2            2.7097     1.1337   2.390   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.0987, Adjusted R-squared:  0.08012
## F-statistic: 5.311 on 2 and 97 DF,  p-value: 0.006473
```

We can reject the null hypothesis that $\beta_2 = 0$ because of the p-value of 0.0188. In the case of β_1, however, we would fail to reject the null hypothesis that $\beta_1 = 0$ because of the p-value of 0.4390

d.

```
lm.x1<- lm(y ~ x1)
summary(lm.x1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00241 -0.66755 -0.09282  0.71984  2.78124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0819     0.2365   8.804 4.75e-14 ***
## x1            0.8790     0.4061   2.164   0.0329 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.081 on 98 degrees of freedom
## Multiple R-squared:  0.04562,    Adjusted R-squared:  0.03588
## F-statistic: 4.685 on 1 and 98 DF,  p-value: 0.03286
```

In a model where there is no $X2$ We can reject the null hypothesis that $\beta_1 = 0$ because of a p-value of 0.0329.

e.

```
lm.x2 <- lm(y ~ x2)
summary(lm.x2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.91065 -0.65771 -0.06083  0.65167  2.47408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0295     0.1916  10.590  < 2e-16 ***
## x2            1.9739     0.6224   3.172  0.00202 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.054 on 98 degrees of freedom
## Multiple R-squared:  0.09309,    Adjusted R-squared:  0.08384
## F-statistic: 10.06 on 1 and 98 DF,  p-value: 0.002024
```

In this model, we can reject the null hypothesis that $\beta_1 = 0$ because the p-value is 0.00202.

f.

No the results do not contradict each other because multicollinearity exists between `x1` and `x2`, making it difficult to distinguish their effects on `y`.

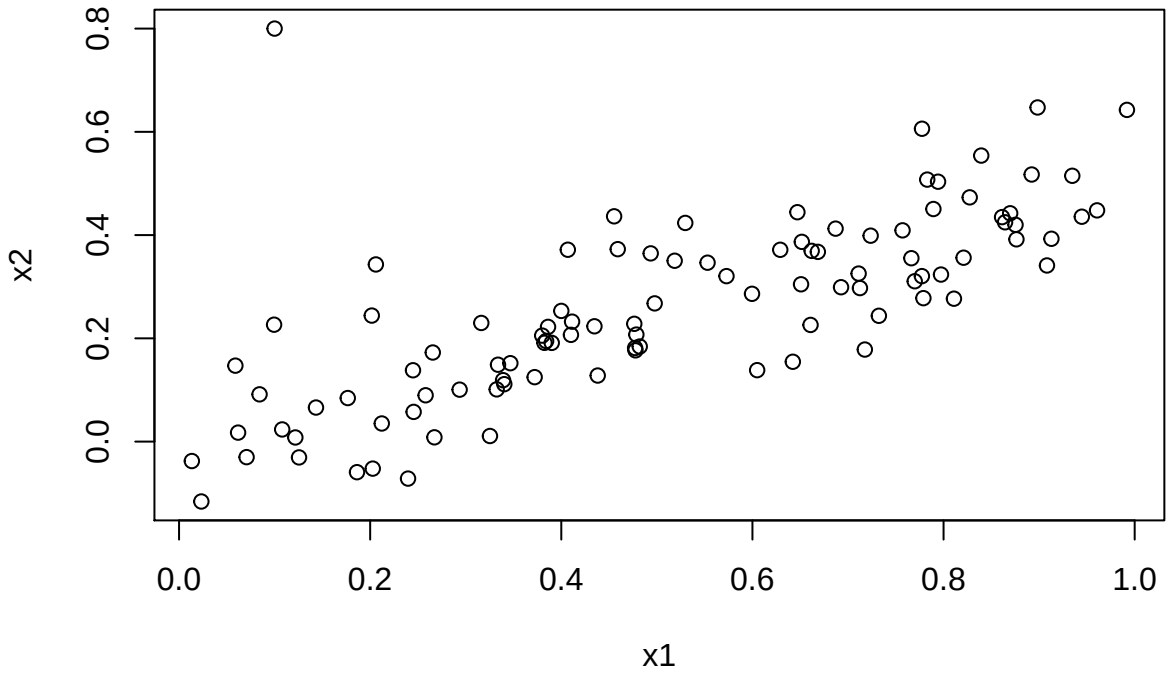
g.

```
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)
```

```
cor(x1, x2)
```

```
## [1] 0.7392279
```

```
plot(x1, x2)
```

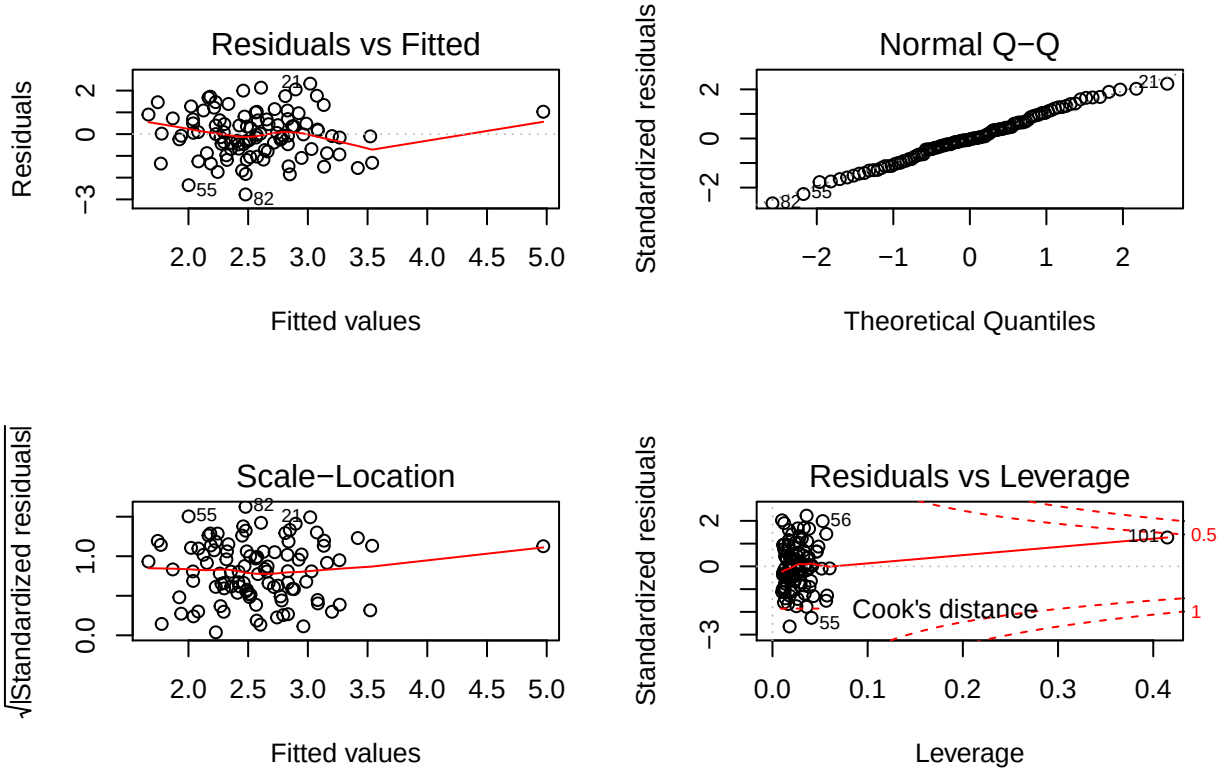


If we include the new observations, we can see that there is a lower correlation $X1$ and $X2$. In addition, there is a clar output in the top-left corner of the plot, which corresponds to out new value.

```r
lm2 <- lm(y ~ x1 + x2)
summary(lm2)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.77230 -0.68497 -0.03604  0.67478  2.31801
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1884     0.2281   9.595 9.18e-16 ***
## x1           -1.1027     0.5838  -1.889   0.0619 .
## x2            3.6163     0.8850   4.086 8.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 98 degrees of freedom
## Multiple R-squared:  0.1661, Adjusted R-squared:  0.1491
## F-statistic: 9.761 on 2 and 98 DF,  p-value: 0.0001363
```

```r
par(mfrow = c(2, 2))
plot(lm2)
```

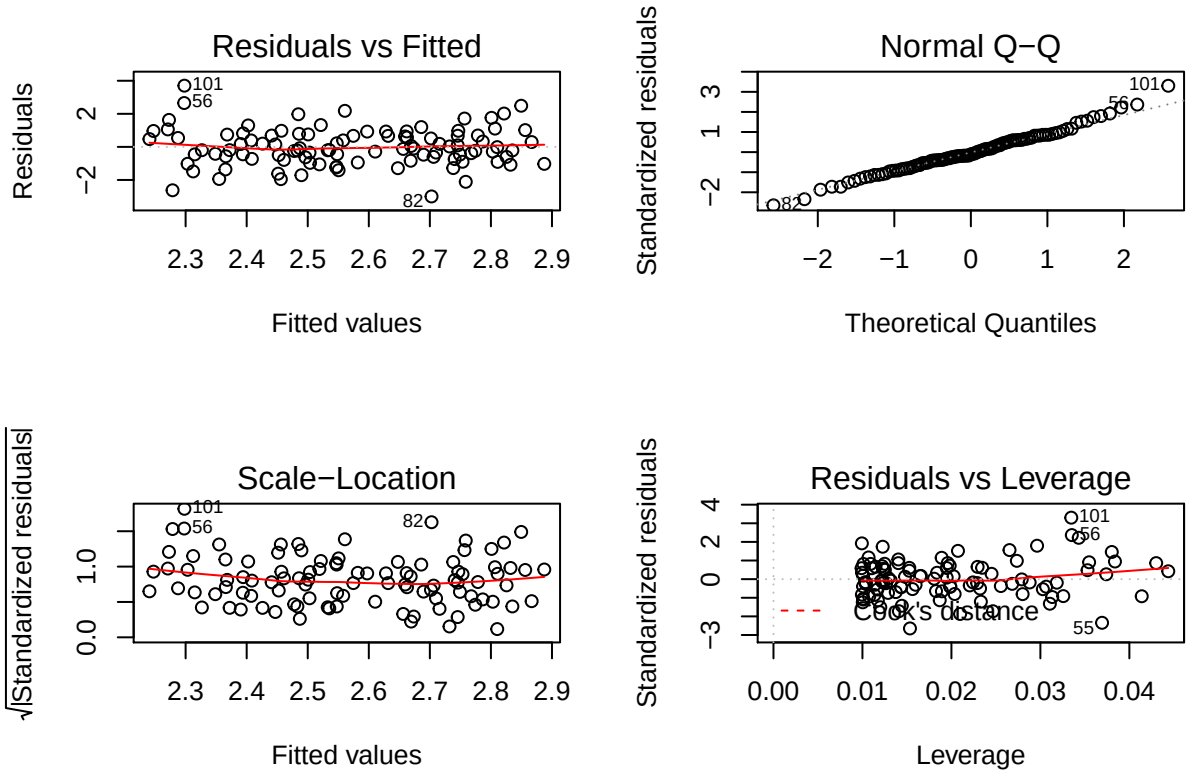


The full model shows that we fail to reject that $\beta_1 = 0$ becayse of the p-value of 0.0619. However, we can still conclude that $\beta_2 \neq 0$. The leverage plot suggests that the 101st oberservation acts as a high leverage point but it does not exceed the 2 outlier threshold in regards to studentized residuals.

```
lm3 <- lm(y ~ x1)
summary(lm3)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9970 -0.7260 -0.1236  0.6885  3.7020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2319     0.2452   9.101 9.99e-15 ***
## x1            0.6608     0.4232   1.561    0.122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.14 on 99 degrees of freedom
## Multiple R-squared:  0.02403,    Adjusted R-squared:  0.01418
## F-statistic: 2.438 on 1 and 99 DF,  p-value: 0.1216
```

```
par(mfrow = c(2, 2))
plot(lm3)
```

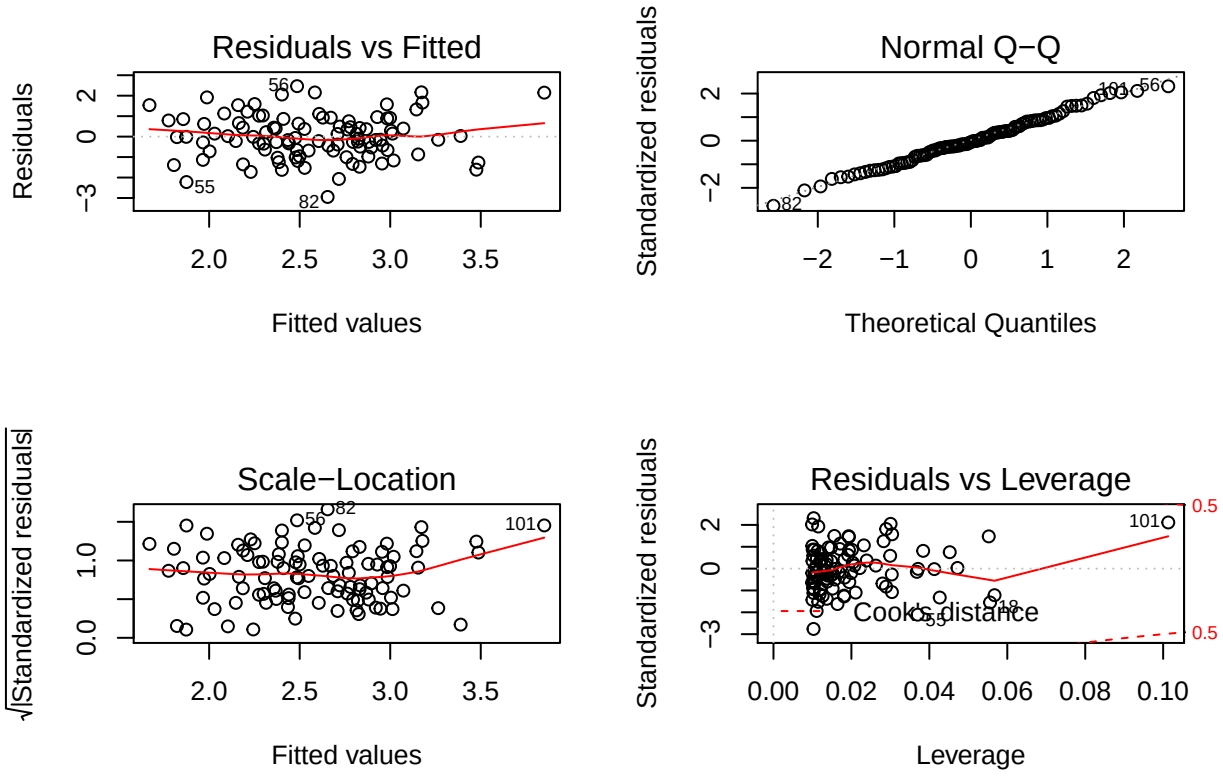


$X1$ on its own does not have a relationship with Y, we come to this conclusion because of the p-value of 0.122. In this model, observation 101 is both a studentized residual (above the value of 2) as well as a high leverage point.

```
lm4 <- lm(y ~ x2)
summary(lm4)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.94849 -0.68322 -0.06569  0.75209  2.46508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9464     0.1911  10.185  < 2e-16 ***
## x2            2.3806     0.6037   3.943  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.073 on 99 degrees of freedom
## Multiple R-squared:  0.1358, Adjusted R-squared:  0.127
## F-statistic: 15.55 on 1 and 99 DF,  p-value: 0.00015
```

```
par(mfrow = c(2, 2))
plot(lm4)
```



$X2$ has a relationship with Y, we come to this conclusion because of the p-value of 0.00015. Similar to the **lm3** model, oberservation 101 again acts as a high leverage point and outlier.

For **lm3**, the slope of $X1$ is reduced compared to the previous iteration. In **lm4** $beta_1$ estimate shows an increase of slope against Y.

h.

Based on the outputs above:

- `lm2`: 1.06
- `lm3`: 1.14
- `lm4`: 1.073

The full model, or `lm2`, has the lowest standard error, meaning that it produces the most reliable estimates despite the lack of significance of `x1`.

i.

```r
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

```r
vif(lm1)
```

```
##       x1       x2
## 3.304993 3.304993
```

```r
vif(lm2)
```

```
##       x1       x2
## 2.204867 2.204867
```

Based off of the output, we can see that the model WITH the outlier has less multicollineatity than the model with the outlier. The reason that the model with the outlier performed better was because the multicollinearity among the predictors allowed us to properly identify $X1$ and $X2$'s effects on Y.

An example with the problem with multicollinearity would be if we were looking to predict how fast someone can run based off of leg length and something else like height. I cannot claim that I know they are correlated but let us assume that they are. It would be pointless to use these two variables because we know that they are correlated. Replacing height with something else like weight would help in being more accurate with a prediction.

3.

a.

$p(X) = 0.37/1.37$

```r
0.37/1.37
```

```
## [1] 0.270073
```

b.

$p(X) = 0.16/0.84$

```r
0.16/0.84
```

```
## [1] 0.1904762
```

4.

β_0 = -6 $\beta_1 = 0.05$ $\beta_2 = 1$

a.

$e^-0.05/(1 + e^-0.05)$

```r
exp(-0.5)/(1 + exp(-0.5))
```

```
## [1] 0.3775407
```

b.

$(log(1) + 2.5)/0.05$

```r
(log(1) + 2.5)/0.05
```

```
## [1] 50
```

5.

5.

a.

Distance formula:
$d(p, q) = \sqrt{(q_1 - p_1)^2 + \cdots + (q_n - p_n)^2}$
Observation distances:

1. 3
2. 2
3. 3.16
4. 2.23
5. 1.41
6. 1.73

b.

When $K = 1$, our prediction should be green because the Euclidean distance to observation 5 is the lowest at 1.41.

c.

When $K = 3$, our prediction should be red because we have 2 red observations (distance of 1.73 and 2) and 1 green observation (distance of 1.41) within our boundary. Because of the majority rule, we must go with red for our test observation.

d.

If the decision boundary is highly non-linerar, then we can can expect that a lower K would work higher than a K of a higher value because it a higher value of K would make our model less flexible. Conversely, the lower valued K would allow the model to be more flexible, which would be better if we find that we find that we have a non-linear decision boundary.

6.

a.

The QDA would perform better on the training set because of its flexibility, meanwhile LDA would perform better on a test set.

b.

When it comes to a training set, we would expect QDA to do better, again because of its high flexibility. In the case of a testing set, it depends on the nonlinearity. We should expect the QDA to perform better but there will be some non-linear relationships that can work with LDA.

c.

We would expect that the "test prediction accuracy" of QDA to be improve compared to LDA, again because of its high flexibility. The larger the sample size, the less of a concern there is for inaccuracies or variances.

d.

False, because if there are not many predictors, than the highly flexibile QDA can overfit. If there are a lower number of observations, then LDA should be preferred, as it will likely perform better.

7.

We can use the formula to obtain the posterior probability of x.

$p_1(4) = \frac{0.8e^{-(1/72)(4-10)^2}}{0.8e^{-(1/72)(4-10)^2}+0.2e^{-(1/72)(4-0)^2}} = 0.752$

If we are given a value of 4, then the probability that a company issues a dividend is 0.752.