Problem 6

```
set.seed(1)
x1 = runif(100)
x2 = (0.5 * x1) + rnorm(100)/10

# Create a linear model in which y is a function of x1 and x2.
y = 2 + (2 * x1) + (0.3 * x2) + rnorm(100)
```

Part A

The form of the linear model is

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\
&= 2 + 2 \cdot x_1 + 0.3 \cdot x_2 + rnorm(100)
\end{aligned}
$$

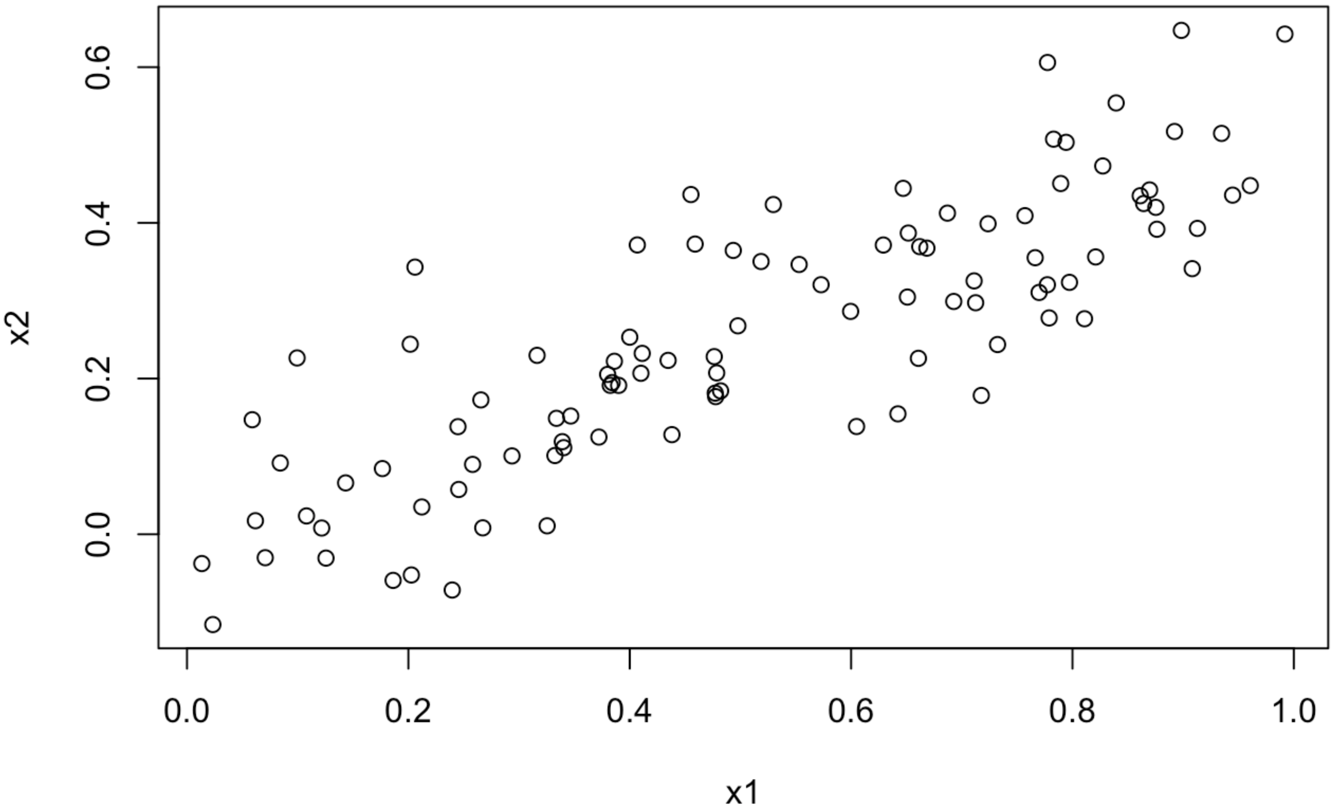
$$\boxed{\beta_0 = 2, \beta_1 = 2, \beta_2 = 0.3, \epsilon = rnorm(100)}$$

Part B

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

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

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



Part C

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

```
##
## 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            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## 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:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

The estimated coefficients are $\beta_0 = 2.1305$, $\beta_1 = 1.4396$, and $\beta_2 = 1.0097$, which are at least in the ballpark of the true coefficients (`2, 2, 0.3`). β_2 is smaller than both β_0 and β_1 in both the true and estimated coefficients.

We can reject the null hypothesis $H_0 : \beta_1 = 0$, because the `p = 0.0487 < 0.05`. However, we cannot reject $H_0 : \beta_2 = 0$, because `p = 0.3754 > 0.05`.

Part D

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

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89495 -0.66874 -0.07785  0.59221  2.45560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
## x1            1.9759     0.3963   4.986 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 98 degrees of freedom
## Multiple R-squared:  0.2024, Adjusted R-squared:  0.1942
## F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

We can reject the null hypothesis $H_0 : \beta_1 = 0$, because the `p = 2.661e-06 < 0.05`. When we throw out x_2, we get a much more impressive `p` value than when we included both x_1 and x_2 in the linear regression.

Part E

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

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62687 -0.75156 -0.03598  0.72383  2.44890
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
## x2            2.8996     0.6330    4.58 1.37e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.072 on 98 degrees of freedom
## Multiple R-squared:  0.1763, Adjusted R-squared:  0.1679
## F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

We can reject the null hypothesis $H_0 : \beta_2 = 0$, because the `p = 1.366e-06 < 0.05`. When we throw out x_1 , we get a much more impressive `p` value than when we included both x_1 and x_2 in the linear regression.

Part F

In a way, yes, I expected Part E to show a non-significant `p`-value, but instead it was even lower than in Part D!

Part G

```
x1 = c(x1, 0.1)
x2 = c(x2, 0.8)
y  = c(y,  6)
```

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

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73348 -0.69318 -0.05263  0.66385  2.30619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
## x1            0.5394     0.5922   0.911  0.36458
## x2            2.5146     0.8977   2.801  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.075 on 98 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2029
## F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06
```

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

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8897 -0.6556 -0.0909  0.5682  3.5665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
## x1            1.7657     0.4124   4.282 4.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.111 on 99 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1477
## F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

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

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64729 -0.71021 -0.06899  0.72699  2.38074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
## x2            3.1190     0.6040   5.164 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.074 on 99 degrees of freedom
## Multiple R-squared:  0.2122, Adjusted R-squared:  0.2042
## F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

This new observation flips the relationship we saw before. Previously, we saw a more significant p value for x_1 when we fit the lm to both x_1 and x_2, while now we have a more significant p value for x_2. However, they still both indicate a low p value in the lms where we fit the two variables independently.