Stat 230: Linear Models
Homework 3
Professor Ding
Lev Golod

Question 5, part A

The regression equation is: $Y = 2 + 2X_1 + 0.3X_2 + \epsilon,\ \epsilon \sim N(0, I)$.
Equivalently we may write $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$.

```
set.seed(240)
x1 <- matrix(runif(100), ncol = 1)
x2 <- 0.5*x1 + rnorm(100)/10
y  <- 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

Question 5, part B

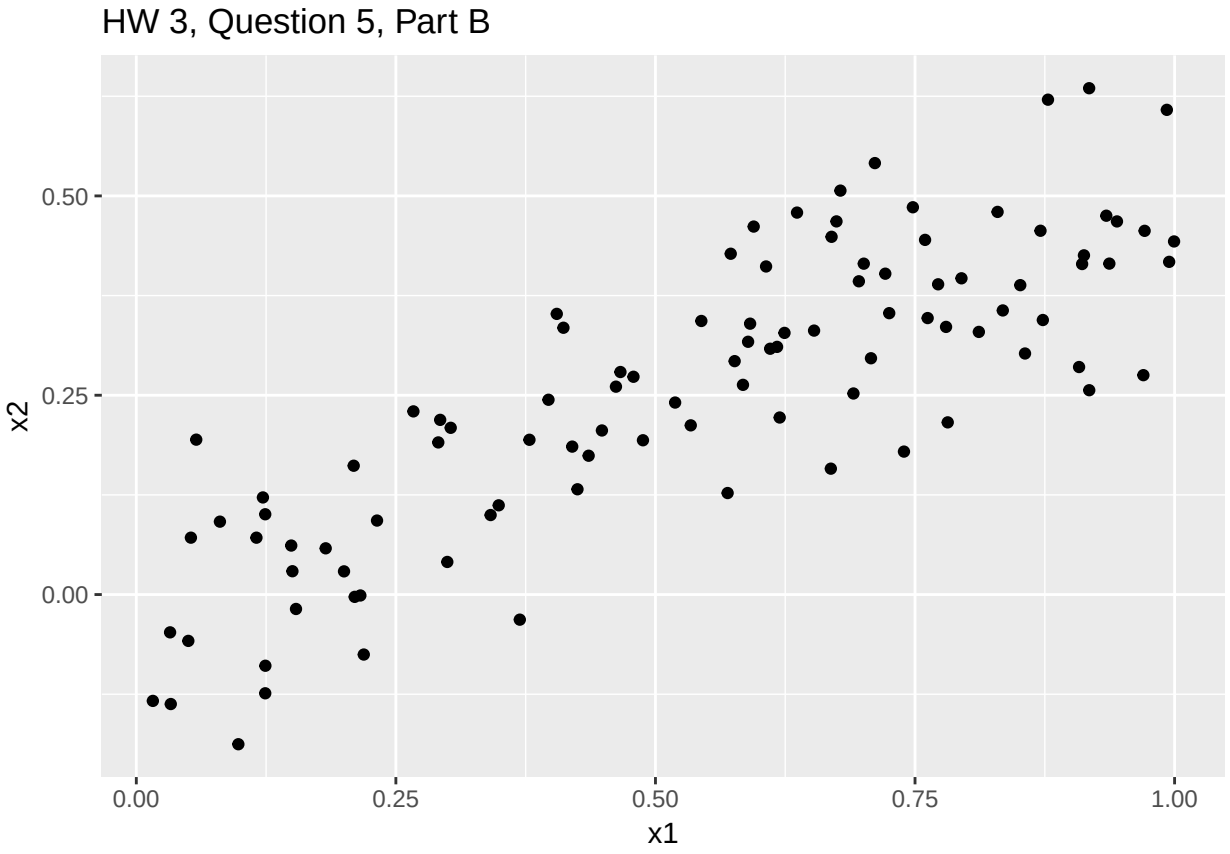
I will calculate the empirical, or observed, correlation.

```
# First I need the covariance. To find this I will center the two vectors
# X1 and X2 and take their dot product.
mycov <- function(x,y) (nrow(x)-1)**-1 * t(x - mean(x)) %*% (y - mean(y))
covx1x2 = mycov(x1,x2)
# I will check this result against R's built-in function
# abs(cov - cov(x1,x2)) < 1e-6
# I will also need the variances of x1 and x2.
varx1 <- mycov(x1,x1)
varx2 <- mycov(x2,x2)

# Correlation(x,y) is defined as Cov(x,y) / sqrt{Varx Vary}
corrx1x2 <- covx1x2 / (varx1 * varx2)**0.5
corrx1x2
```

```
##           [,1]
## [1,] 0.835556
```

```
# Create a scatter plot
dat <- data.frame(cbind(y,x1,x2))
names(dat) <- c('y', 'x1','x2')
myplot <- ggplot(dat, aes(x1,x2)) +
  geom_point() +
  ggtitle("HW 3, Question 5, Part B")
myplot
```



Question 5, part C

```
fit1 <- lm(y ~ x1 + x2, dat)
# summary(fit1)

# The estimated coefficients are as follows
round(summary(fit1)$coefficients[,1], 3)
```

```
## (Intercept)          x1          x2
##       1.970       2.036       0.006
```

```
# Here are the associated p-values
# round(summary(fit1)$coefficients[,4], 4)
format(summary(fit1)$coefficients[,4], digits = 4)
```

```
## (Intercept)          x1          x2
## "1.809e-14" "2.196e-03" "9.955e-01"
```

Describe the results: We see that the model has an adjusted R-squared of around 0.21, which is only modest. This makes sense because when we generated Y we added a large amount of random noise, compared to the 'signals' of x1 and x2. However the p-value for the hypothesis $\beta_0 = ... = \beta_2 = 0$ is small, so the model as a whole is able to explain some of the variation in Y.

How do the estimated regression parameters compare to the true values? The model did a good job estimating β_0 and β_1, both of which have a true value of 2. However it did a poor job estimating β_2 which has a true value of 0.3.

Hypothesis testing: Based on the p-values associated with the coefficients we can reject the hypothesis $\beta_1 = 0$ at the level $\alpha = 0.05$. We cannot reject the hypothesis $\beta_2 = 0$.

Question 5, part D

```
fit2 <- lm(y ~ x1, dat)
# summary(fit2)

# The estimated coefficients are as follows
round(summary(fit2)$coefficients[,1], 3)
```

```
## (Intercept)          x1
##       1.970       2.039
```

```
# Here are the associated p-values
format(summary(fit2)$coefficients[,4], digits = 4)
```

```
## (Intercept)          x1
## "8.279e-15" "9.492e-08"
```

Again, the model did a good job estimating β_0 and β_1. Again, we reject the hypothesis: $\beta_1 = 0$. We observe that the p-value for β_1 has become much more significant. This makes sense, since X_1 and X_2 are highly colinear, and we have removed X_2.

Question 5, part E

```
fit3 <- lm(y ~ x2, dat)
# summary(fit2)

# The estimated coefficients are as follows
round(summary(fit3)$coefficients[,1], 3)
```

```
## (Intercept)          x2
##       2.382       2.680
```

```
# Here are the associated p-values
format(summary(fit3)$coefficients[,4], digits = 4)
```

```
## (Intercept)          x2
## "3.863e-23" "1.308e-05"
```

Interestingly the model now thinks that the coefficient β_2 is around 2.6, whereas we know that the true value is 0.3. In the absense of X_1 β_2 becomes highly significant and we reject the hypothesis $\beta_2 = 0$.

Question 5, part F

These results do not contradict each other.
In (c) we got an underestimate of β_2 and failed to reject the hypothesis $\beta_2 = 0$. That's because the model included both X_1 and X_2, and most of the information in X_2 is in X_1 as well.
In (d) we got a very good estimate of β_1 and found that it was highly significant, as we would expect.
in (e) we got an overestimate of β_2: around 2.6, as opposed to the true

value of 0.3. However if we ignore 'noise' terms we can informally write:

$$X_2 \approx 0.5\ X_1 Y \approx 2 + 2X_1 + 0.3X_2 Y \approx 2 + 2.15X_1$$

When we remove X_2 from the model,we are trying to predict Y based solely on X_2, even though in reality we know Y was constructed as a combination of both of them (plus random noise). As we see in the above approximate equalities, 2.6 is not an unreasonable estimate of the coefficient of X_2 when we remove X_1. In fact, if we divide 2.15 by 0.83 which is the correlation, we get a number very close to 2.6.

Question 5, part G

```
# # Summarize X1,X2, and Y to determine whether the new point is an outlier
# summary(x1)
# summary(x2)
# summary(y)

# update the data with the new unfortunate point
x1b <- c(x1, 0.1)
x2b <- c(x2, 0.8)
yb  <- c(y, 6)
dat2 <- data.frame(cbind(yb,x1b,x2b))
names(dat2) <- c('y', 'x1','x2')
```

Model from Part C

```
### Model from part C
fit1b <- lm(y ~ x1 + x2, dat2)
# The estimated coefficients are as follows
round(summary(fit1b)$coefficients[,1], 3)
```

```
## (Intercept)          x1          x2
##       2.115       0.952       1.812
```

```
# Here are the associated p-values
format(summary(fit1b)$coefficients[,4], digits = 4)
```

```
## (Intercept)          x1          x2
## "1.001e-15" "8.711e-02" "3.337e-02"
```

The new observation has significantly messed up the model from Part (c). We now have a substantial under-estimate for β_1 and an over-estimate for β_1, relative to the true values. The new data point exceeds the previous maxima for X_2 and Y. It is an outlier. It is a leverage point since it changed the estimated regression coefficients bigly.

Model from Part D

```
### Model from part D
fit2b <- lm(y ~ x1, dat2)
# The estimated coefficients are as follows
round(summary(fit2b)$coefficients[,1], 3)
```

```
## (Intercept)          x1
##       2.112       1.843
```

```r
# Here are the associated p-values
format(summary(fit2b)$coefficients[,4], digits = 4)
```

```
## (Intercept)          x1
## "2.358e-15" "2.917e-06"
```

We observe that the estimated regression coefficients did not change that much after adding the new point. In this model is an outlier but not a leverage point.

Model from Part E

```r
fit3b <- lm(y ~ x2, dat2)
# The estimated coefficients are as follows
round(summary(fit3b)$coefficients[,1], 3)
```

```
## (Intercept)          x2
##       2.340       2.899
```

```r
# Here are the associated p-values
format(summary(fit3b)$coefficients[,4], digits = 4)
```

```
## (Intercept)          x2
## "4.651e-23" "1.258e-06"
```

We observe that the estimated regression coefficients did not change that much after adding the new point. In this model is an outlier but not a leverage point.