DSE6011 - Module 2 - Ch 3 HW

Nathan Monges

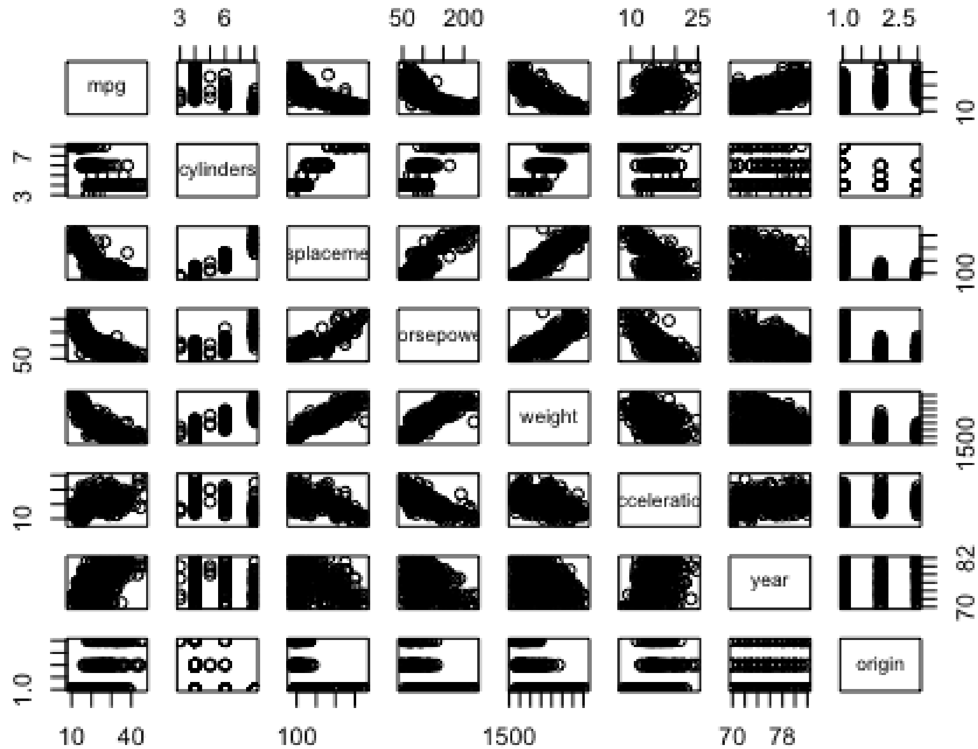
2024-07-14

```
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.0     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2     
## ── Conflicts ────────────────────────────────────────── 
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all 
conflicts to become errors
```

Exercise 9

a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
auto_data <- read.table("Auto.data", header = TRUE, sep = "", na.strings = 
"?")
auto_data <- na.omit(auto_data)

pairs(auto_data[1:8])
```



b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.

```
cor(auto_data[1:8])

##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

i. Is there a relationship between the predictors and the re- sponse?

ii. Which predictors appear to have a statistically significant relationship to the response?

iii. What does the coefficient for the year variable suggest?

```
mpg_regression <- lm(mpg ~ cylinders + displacement + horsepower + weight +
acceleration + year + origin, data = auto_data)

summary(mpg_regression)

##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + year + origin, data = auto_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

From this linear model of mpg regressed on all other variables in the auto data, we can see that the predictors of displacment, weight, year and origin are all statistically significant with p-values less than 0.05. All varuables other than cylinders, horsepower andf weight all show a negative relationship with mpg, which makes sense as these variables can be known to be trade-offs to higher mpg. The 'year' variables stands out showing a strong postive relationship with fuel efficiency which shows the growth in automobile engineering over the years.

d) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```r
mpg_interactive <- lm(mpg ~ (cylinders + displacement + horsepower + weight +
acceleration + year + origin)^2, data = auto_data)

summary(mpg_interactive)

##
## Call:
## lm(formula = mpg ~ (cylinders + displacement + horsepower + weight +
##     acceleration + year + origin)^2, data = auto_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.548e+01  5.314e+01   0.668  0.50475
## cylinders                  6.989e+00  8.248e+00   0.847  0.39738
## displacement              -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                 5.034e-01  3.470e-01   1.451  0.14769
## weight                     4.133e-03  1.759e-02   0.235  0.81442
## acceleration              -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                       6.974e-01  6.097e-01   1.144  0.25340
## origin                    -2.090e+01  7.097e+00  -2.944  0.00345 **
## cylinders:displacement    -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders:horsepower       1.161e-02  2.420e-02   0.480  0.63157
## cylinders:weight           3.575e-04  8.955e-04   0.399  0.69000
## cylinders:acceleration     2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders:year            -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders:origin           4.022e-01  4.926e-01   0.816  0.41482
## displacement:horsepower   -8.491e-05  2.885e-04  -0.294  0.76867
## displacement:weight        2.472e-05  1.470e-05   1.682  0.09342 .
## displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853
## displacement:year          5.934e-03  2.391e-03   2.482  0.01352 *
## displacement:origin        2.398e-02  1.947e-02   1.232  0.21875
## horsepower:weight         -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower:acceleration   -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower:year           -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower:origin          2.233e-03  2.930e-02   0.076  0.93931
## weight:acceleration        2.346e-04  2.289e-04   1.025  0.30596
## weight:year               -2.245e-04  2.127e-04  -1.056  0.29182
## weight:origin             -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration:year          5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration:origin        4.583e-01  1.567e-01   2.926  0.00365 **
## year:origin                1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.695 on 363 degrees of freedom
```

```
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

From this linear model summary the following terms of 'displacement:year', 'acceleration:year', and 'acceleration:origin' appear to be statistically significant with p-values less than 0.05. In the case for displacement, the positive coefficient of 0.005934 indicates that as year increases, the effect that displacement has on mpg becomes more significant.

f) Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings.

```
log_model <- lm(mpg ~ log(cylinders) + log(displacement) + log(horsepower) +
log(weight) +
                  log(acceleration) + log(year) + log(origin), data =
auto_data)

summary(log_model)

##
## Call:
## lm(formula = mpg ~ log(cylinders) + log(displacement) + log(horsepower) +
##     log(weight) + log(acceleration) + log(year) + log(origin),
##     data = auto_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5987 -1.8172 -0.0181  1.5906 12.8132
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -66.5643    17.5053  -3.803 0.000167 ***
## log(cylinders)       1.4818     1.6589   0.893 0.372273
## log(displacement)   -1.0551     1.5385  -0.686 0.493230
## log(horsepower)     -6.9657     1.5569  -4.474 1.01e-05 ***
## log(weight)        -12.5728     2.2251  -5.650 3.12e-08 ***
## log(acceleration)   -4.9831     1.6078  -3.099 0.002082 **
## log(year)           54.9857     3.5555  15.465  < 2e-16 ***
## log(origin)          1.5822     0.5083   3.113 0.001991 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.069 on 384 degrees of freedom
## Multiple R-squared:  0.8482, Adjusted R-squared:  0.8454
## F-statistic: 306.5 on 7 and 384 DF,  p-value: < 2.2e-16
```

In the case of log(x) of the variables in the auto dataset, an R-squared of 0.8482 indicates that ~84.82 of the variance in mgg can be explained by the predictors in the model. Intercept of -66.5643 indicates the estimated mpg when all predictors are zero. A residual standard error of 3.069 indicates the average deviation of the observed values from the values of the fitted line.

```
sqrt_model <- lm(mpg ~ sqrt(cylinders) + sqrt(displacement) +
sqrt(horsepower) + sqrt(weight) + sqrt(acceleration) + sqrt(year) +
sqrt(origin), data = auto_data)

summary(sqrt_model)

##
## Call:
## lm(formula = mpg ~ sqrt(cylinders) + sqrt(displacement) + sqrt(horsepower)
+
##     sqrt(weight) + sqrt(acceleration) + sqrt(year) + sqrt(origin),
##     data = auto_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5250 -1.9822 -0.1111  1.7347 13.0681
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -49.79814    9.17832  -5.426 1.02e-07 ***
## sqrt(cylinders)     -0.23699    1.53753  -0.154   0.8776
## sqrt(displacement)   0.22580    0.22940   0.984   0.3256
## sqrt(horsepower)    -0.77976    0.30788  -2.533   0.0117 *
## sqrt(weight)        -0.62172    0.07898  -7.872 3.59e-14 ***
## sqrt(acceleration)  -0.82529    0.83443  -0.989   0.3233
## sqrt(year)          12.79030    0.85891  14.891  < 2e-16 ***
## sqrt(origin)         3.26036    0.76767   4.247 2.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.21 on 384 degrees of freedom
## Multiple R-squared:  0.8338, Adjusted R-squared:  0.8308
## F-statistic: 275.3 on 7 and 384 DF,  p-value: < 2.2e-16
```

In the case of sqrt(x) of the variables in the auto dataset, an R-squared of 0.8308 indicates that ~83.08 of the variance in mpg can be explained by the predictors in the model. Intercept of -49.79814 indicates the estimated mpg when all predictors are zero. A residual standard error of 3.21 indicates the average deviation of the observed values from the values of the fitted line.

Exercise 10

a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

```
carseats <- read.csv("Carseats.csv")

sales_model <- lm(Sales ~ Price + Urban + US, data = carseats)
```

```
summary(sales_model)

##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

Intercept: When all predictors variables (Price, Urban, and US) are zero, the estimated Sales is 13.043469 units. The interept is statistically significant with a p-value < 0.001 which indicats that if Price, Urban and US are zero, there is still a positive baseline Sales value.

Price: For each 1 unit increase in Price, Sales decrease by ~ -0.054459 units and a p-value < 0.001 shows statistical significance indicating that Price has a strong negative linear relationship with Sales.

Urban: This coefficient refers to the effect of being in an Urban (Urban = "Yes") area compared to rural (Urban = "No"). A p-value of 0.936 shows non-significance between the variables and suggests that whether a store is located in an Urban area or not does significantly impact Sales in this model.

US: This coefficent refers to the effet of being located in the US (US = "Yes) compared to being located not in the US (US ="No"). A p-value < 0.001 shows statistical significance indicating that being in the US is asspciated with an increase of ~ 1.200573 units in Sales.

c) Write out the model in equation form, being careful to handle the qualitative variables properly.

Sales = 13.043 + Price(-0.0544) + Urban(-0.0219) + US(1.201) + error term

d) For which of the predictors can you reject the null hypothesis H0 :βj =0?

We can reject the null hypothesis H_0: Beta(Price) = 0 and H_0: Beta(US) = 0 since their p-values are less than the significance level of 0.05 or 5%. We would fail to reject the null hypothesis H_0: Beta(Urban) because its p-value is much larger than 0.05 or 5%, which indicates that the variable does not have any statistically significant relationship with Sales.

e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

```
small_carseats_model <- lm(Sales ~ Price + US, data = carseats)

summary(small_carseats_model)

##
## Call:
## lm(formula = Sales ~ Price + US, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

f) How well do the models in (a) and (e) fit the data?

When determining how well the models fit the data, it useful to look at the R^2 and residual standard error values. The R^2 of 0.2393 in both models explain the ~23.93% of the variance in Sales suggesting that ~23.93 of the variability in Sales can be explained by the predictors in botg models, since the R^2 value is identical. The residual standard error indicate the average deviation of the observed Sales values from the predicted values by the model and a value of 2.469 in model “e” and 2.472 in model “a” show that these values are very similar in both models. Model “e” having a slightly smaller standard residual error is evidence that model “e” fit the data better than model “a”.

g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

```
confint(small_carseats_model, level = 0.95)

##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
```

```
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

h) Is there evidence of outliers or high leverage observations in the model from (e)?

The minimum residual of -6.9269 and maximum residal of 7.0515 are large values for residuals and could indicate that there may be outliers of observations that can heavily influence the fit of the line.

Exercise 14

a) The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

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

y_model <- lm(y ~ x1 + x2)

summary(y_model)

##
## 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
```

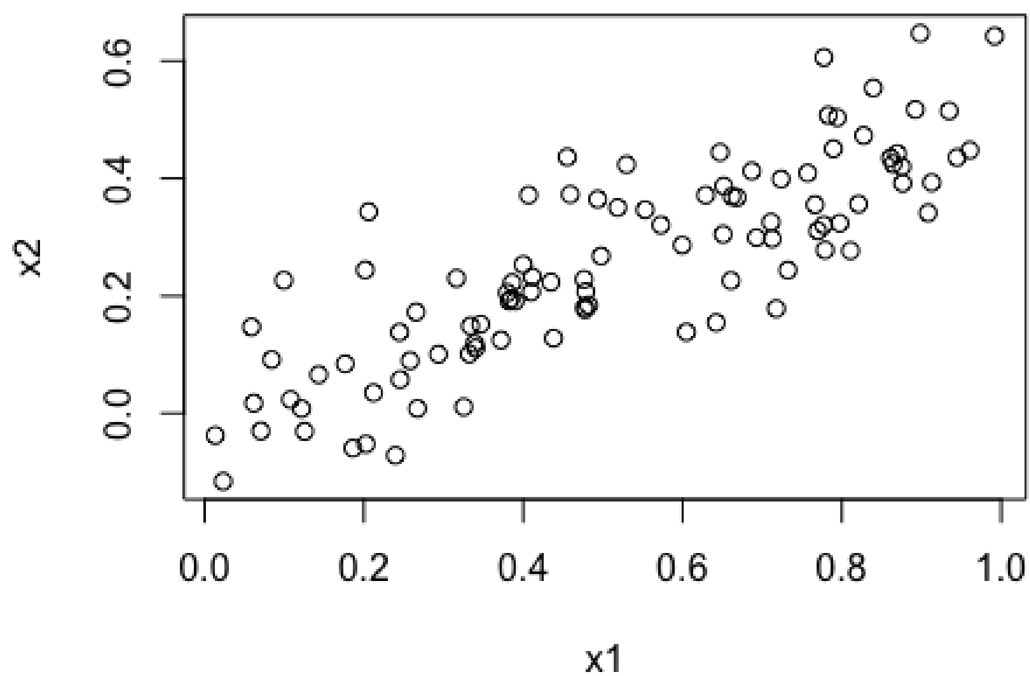
The form of the linear model is: y = Beta_0 + Beta_1(x1) + Beta_2(x2) + error term

x1 coefficient = 1.1436 x2 coefficient = 1.0097

b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

```
cor(x1, x2)

## [1] 0.8351212
```

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



The correlation between x1 and x2 is 0.835 which is a strong positive correaltion.

c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are β̂0, β̂1, and β̂2? How do these relate to the true β0, β1, and β2? Can you reject the null hypothesis H0 : β1 = 0? How about the null hypothesis H0 : β2 = 0?

```r
summary(y_model)

##
## 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
```

Predicted B_0 coefficient is ~2.13 compared to true term of 2. Predicted B_1 coefficient is ~1.43 compared to true term of 2. Predicted B_2 coefficient is ~1.0097 compared to true term of 0.3.

For H_0: B_1 = 0 (Null hypothesis for x1), the p-value is 0.0487 is less than the signifcance level of 0.05 and so we reject the null hypothesis that x1 is signficantly related to Y.

For H_0: B_2 = 0 (Null hypothesis for x2), the p-value associatde with the coefficient of 0.3754 is greater than the significance level of 0.05 and so we would fail to reject the null hypothesis that there is no signifcnat evidence that x2 and y are related.

d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis H0 : $\beta 1 = 0$?

```
x1_model <- lm(y ~ x1)

summary(x1_model)

##
## 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
```

Both coefficents of 2.11 (intercept) and 1.98 (x1) are very close to their true coefficents of 2. A R^2 of 0.2024 indicates that 20.24% of the variance in y is explained by x1. The p-value associated with the coefficient for x1 is 2.66e-06 which is much less than the signifcance level of 0.05 and so we would reject the null hypothesis and conclude that x1 is significantly related to y.

e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis H0 :β1 =0?

```
x2_model <- lm(y ~ x2)

summary(x2_model)

##
## 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
```

Both coefficents of 2.2899 (intercept) and 2.8896 (x2) are slightly higher to their true coefficents of 0.3. A R^2 of 0.1763 indicates that 17.63% of the variance in y is explained by x2. The p-value associated with the coefficient for x2 is 1.37e-05 which is much less than the signifcance level of 0.05 and so we would reject the null hypothesis and conclude that x2 is significantly related to y.

f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

Model c includes both x1 and x2 and suggests that x1 is significant while x2 is not. This could be explained by the multicollinearity between x1 and x2 as they are highly correlated (cor = 0.835).

Model d includes only x1 and shows that x1 is highly significant providing a closer estimate to its true coefficent. Model 3 includes only x2 and shows that x2 is significant but the estimated coefficent of 2.8996 is much higher than the true value of 0.3. This indicates that x2 captures more varaince when acting as the sole variable.

g) Now suppose we obtain one additional observation, which was unfortunately mismeasured. Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

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

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

```
##
## 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
```

```r
d_model <- lm(y ~ x1)
summary(d_model)
```

```
##
## 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
```

```r
e_model <- lm(y ~ x2)
summary(e_model)
```

```
##
## 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
```

Model c: The coefficent for x1 changed dramatically from 1.4396 to -0.1648, and the variable is no longer statistically signigcant with p-value = 0.716. The coefficient for xw increased from 1.0097 to 3.6920 and its remains significant with p-value < 0.05. The new observation has a high leverage since the combination of x1 = 0.1 and x2 = 0.8 is unique and not as common in the dataset and the residuals max and mins indicate the points are not exteree outliers.

Model d: The coeffeficient for x1 decreased from 1.9759 to 1.2117 but is still statistically significant with p-value = 0.00756 < 0.05. The residual standatd error increasing can indicate that the model fit worsened with the new observation and that this observation may be considered an outlier.

Model e: The coefficient for x2 increased fro 2.8996 to 3.5728 and remains significant with p-value = 1.37e-09 is < 0.05. The residual standard error did not change as much as model d which can mean that the new observation did not drastically affect the overall fit model. The residual max and mins indicate that the observation is not an extree outlier but the increease in coefficient can suggest the new observation has an impact on the model.