Assignment16

John Bute

2024-10-15

R Markdown

1. Perform the following commands in R:

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

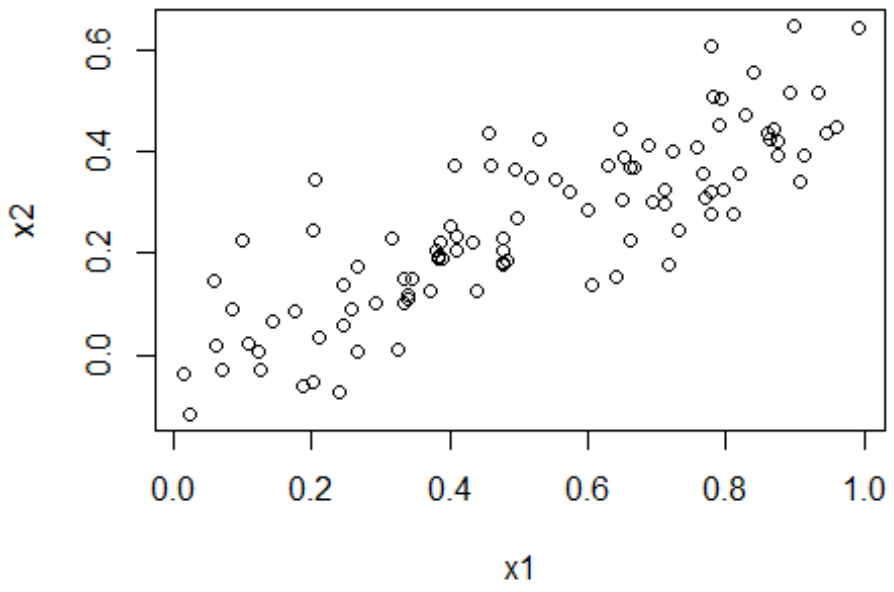
The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

0 = 2 1 = 2 _2 = 0.3

```
cor(x1, x2)

## [1] 0.8351212

plot(x1, x2)
```



The correlation between x1 and x2 is 0.835 indicates a strong positive linear relationship between the two variables, which we can see with the upward linear trend in the scatterplot. The high correlation suggests collinearity, or that two predictor variables in our regression model are highly correlated.

Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are ˆβ0, ˆβ1,, and ˆβ2,? 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?

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

_0 = 2.13

_1 = 1.43

_2 = 1

We determined that the true coefficients are 2, 2, and 0.3 respectively. Thus, the coefficients are not as close to their true values, especially _1 and _2. This may be due to the collinearity between _1 and _2. However, for the null hypothesis of _1 = 0, we check that the p - value is 0.0487 which is less than 0.05, and therefore we can reject the null hypothesis, thus concluding that x1 is a significant predictor for y.

Meanwhile, for x2, the p-value is 0.3754, which is much greater than 0.05, therefore we cannot reject the null hypothesis _2 = 0, which suggests that x2 is not a statistically significant predictor for y

4. Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis H0 : β1 = 0?

```
model1 <- lm(y ~ x1)
summary(model1)
```

```
##
## 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
```

_0 = 2.1124

_1 = 1.9759

In terms of rejecting the null hypothesis, we see that its p-value is significantly low (2.66e-06) (close to 0). As a result, we can safely reject the null hypothesis.

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?

```
model2 <- lm(y ~ x2)
summary(model2)
```

```
##
## 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
```

_0 = 2.3899

_1 = 2.8996

By looking at the p-value of x2 (1.37e-05), we can see that the null hypothesis can be safely rejected as it is very close to 0.

Do the results obtained in 3. to 5. contradict each other? Explain your answer

The results from question 3 show that x2 is not significant. However, the model for x2 shows that it is significant due to its p-value. This contradition happens due to the collinearity between x1 and x2, as the variance caused by x2 is largely shared with x1, thus reducing the significance of x1. However, the significance of x1 is maintained across both models its included in, although its strength is reduced in the model that it shares with x2

Now suppose we obtain one additional observation, which was unfortunately mismeasure

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

refitting model y onto x1 and x2

```
model_y_x1_x2_new <- lm(y ~ x1 + x2)
summary(model_y_x1_x2_new)
```

```
##
## 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
```

refitting model y onto x1

```r
model_y_x1_new <- lm(y ~ x1)
summary(model_y_x1_new)
```

```
##
## 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
```

refitting model y onto x2

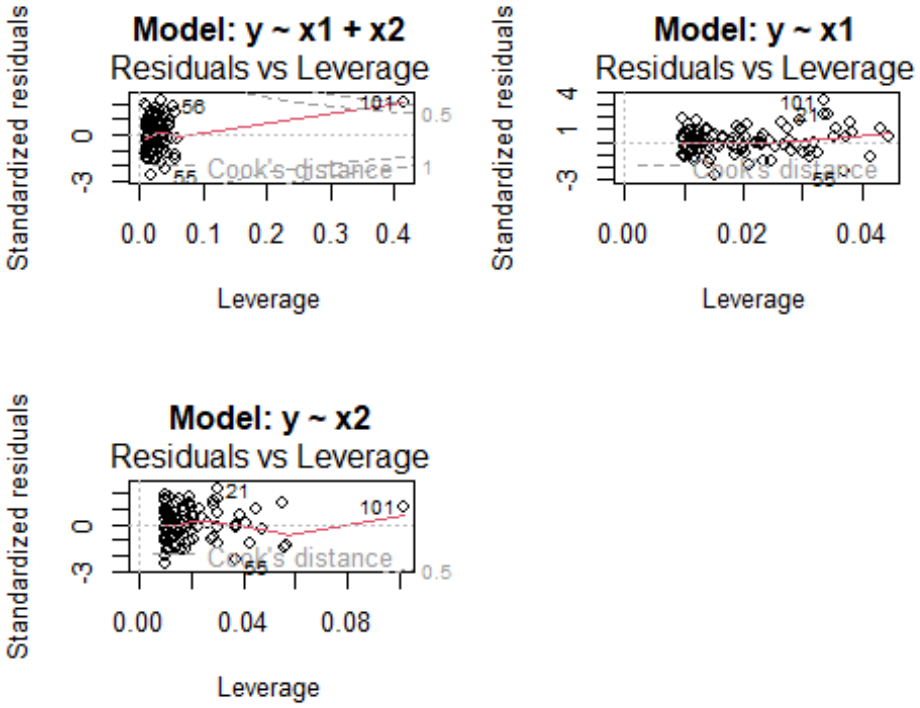
```r
model_y_x2_new <- lm(y ~ x2)
summary(model_y_x2_new)
```

```
##
## 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
```

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

plot(model_y_x1_x2_new, which = 5)
title("Model: y ~ x1 + x2")
plot(model_y_x1_new, which = 5)
```

```
title("Model: y ~ x1")
plot(model_y_x2_new, which = 5)
title("Model: y ~ x2")
```



Re-fit the linear models from 3. to 5. 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

The leverage vs standardized residual plots demonstrate that our added point (labeled 101) is a both a high-leverage point and outlier for our model that uses both x1 and x2. Thus it is a highly influential point as they can distort our model. They pull the regression line significantly away from the true model's fit, leading to bias estimates. Meanwhile, for our model that uses x1, it is not a high-leverage point, but it is an outlier as it is 3 standard deviations up from 0. This means that although this point will be poorly predicted by the model, and increase residual variance (higher standard errors for coefficients), but they would not impact the fit of the model if there are few outliers with low-leverage. Finally, for our model that only uses x2, it is a high-leverage point, but it is not an outlier. Thus, it is significant, but since it is not an outlier, it will not pull the regression line significantly towards itself thus not change the fit of the model significantly.