Q16

Rahul Atre

2023-10-16

Q16 [Regression Modeling]

What's the deal with collinearity?

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 x_1 and x_2. Write out the form of the linear model. What are the regression coefficients?

The form of the linear model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon = 2 + 2x_1 + 0.3x_2 + \epsilon$. The regression coefficients for the given linear model is $\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$, $\epsilon \sim N(0, 1)$.

2. What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

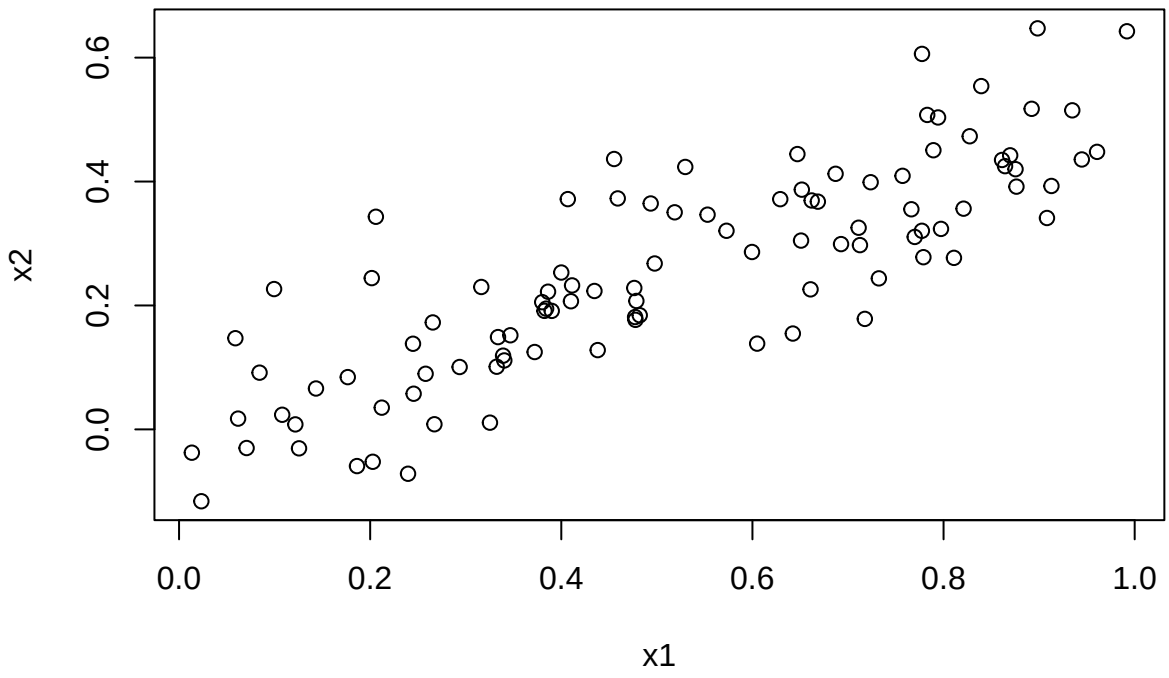
We can run the following commands to obtain the correlation and scatterplot:

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

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

x_1 and x_2 have a 0.8351212 correlation, which is strongly positive.

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



3. Using this data, fit a least squares regression to predict y using x_1 and x_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$? How do these relate to the true β_0, β_1 and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

```
lin_model = lm(y ~ x1 + x2)
summary(lin_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
```

From the summary() function call, we have obtained the estimated beta parameters using the least squares regression model: $\hat{\beta}_0 = 2.1305$, $\hat{\beta}_1 = 1.4396$ and $\hat{\beta}_2 = 1.0097$. It appears that only $\hat{\beta}_0$ is close to β_0. The rest of the estimated parameters are not that good.

As for the hypothesis testing of β_0, since the p-value is 0.0487, which is less than 0.05, we can say that it is statistically significant, and **reject the null hypothesis** $H_0 : \beta_1 = 0$. For β_2 however, the p-value is 0.3754, which is not less than 0.05. Thus, we **cannot reject** $H_0 : \beta_2 = 0$.

4. Now fit a least squares regression to predict y using only x_1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
lin_model2 = lm(y ~ x1)
summary(lin_model2)
```

```
##
## 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
```

For the least squares regression using only x_1, we can immediately see that both the estimated β parameters are very close to 2. As for the null hypothesis, we can **reject** H_0 since the p-value for β_1 is very small. Thus, it is statistically significant.

5. Now fit a least squares regression to predict y using only x_2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

```
lin_model3 = lm(y ~ x2)
summary(lin_model3)
```

```
##
## 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
```

From the above linear regression using only x_2, we can see that the estimated parameter for it is 2.8996, which is quite different from the previous two models. As for the null hypothesis, we can once again **reject** H_0 since the p-value for β_2 is very small. Thus, it is statistically significant.

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

No, the results obtained from 3 to 5 do not contradict each other. We know from part a that there is a strong positive correlation between x_1 and x_2. Due to strong collinearity, even if say, x_1 is significant, but x_2 is not, the response may not be affected. This evidently causes the error in the beta estimates to increase, leading to higher p-values. The results that follow from 4, 5, are not necessarily contradictions, they simply show that collinearity can conceal the errors of one beta estimate, assuming that the two parameters are strongly, positively correlated.

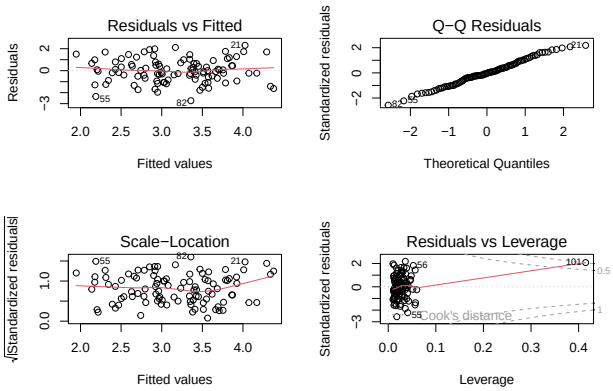
7. Now suppose we obtain one additional observation, which was unfortunately mismeasured.

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

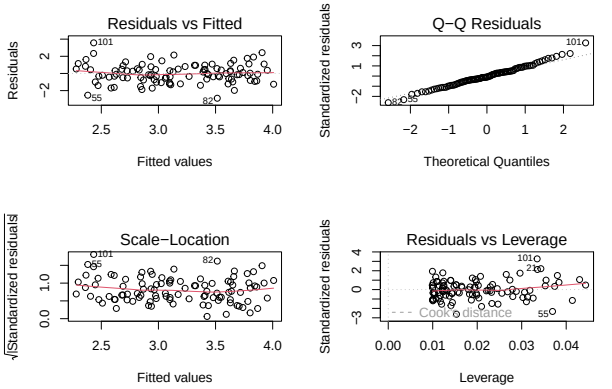
Re-fit the linear models from 3. to 5. using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

```
lin_model4 = lm(y ~ x1 + x2) #First model with both variables + changed
lin_model5 = lm(y ~ x1) #Second model with both variables + changed
lin_model6 = lm(y ~ x2) #Third model with both variables + changed
```

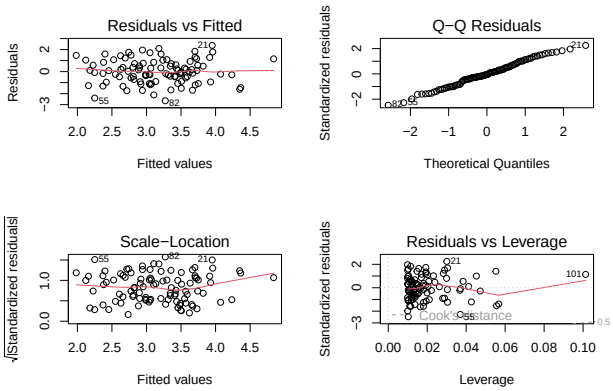
```
par(mfrow=c(2,2))
plot(lin_model4)
```




```
par(mfrow=c(2,2))
plot(lin_model5)
```




```
par(mfrow=c(2,2))
plot(lin_model6)
```



From the summary data of the first model that was changed (hidden due to size constraints), we can see that the β_1 parameter in this case has a high p-value, which contrasts the original model without adding the new observation. However, in the second model, β_1 has a very low p-value like its original, thus making it statistically significant.

As for β_2, it has a low p-value in both the second and third model, making it statistically significant. Notice how in the original, unchanged first model, β_2 had a high p-value, but not when the new observation is added. This new observation is an outlier for only the second changed model.

We can see that in the first and third changed models, the new observation is a very high-leverage point, however, it is also one of the two high-leverage points in the second changed model.