Homework 4

Daniel Tshiani

2025-06-01

1

a

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

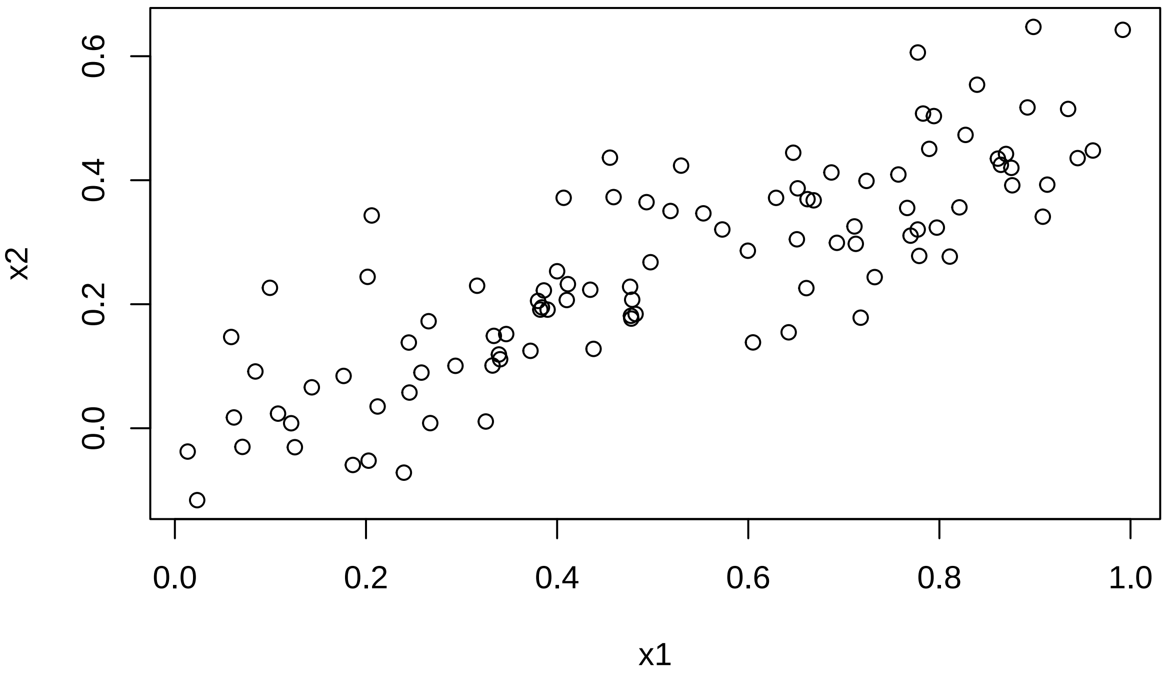
the form of the linear model is: y = 2 + 2*x1 + 0.3*x2 + error

b

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

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

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



c

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

```
##
## 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
```

beta0 is 2.13 beta1 is 1.44 beta2 is 1.01 the p value for beta1 is 0.0487. this is less than our standard p value of 0.05. therefore we reject the null of beta1 = 0. on the other hand, the p value for beta2 is 0.3754. this is greater than 0.05 so we fail to reject the null of beta 2 = 0.

d

```r
lm2 <- lm(y~x1)
summary(lm2)
```

```
##
## 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
```

according to this model x1 is significant to the 1% level meanwhile the previous model was significant to the 5% level. we would still fail to reject the null of beta1 = 0

e

```r
lm3 <- lm(y~x2)
summary(lm3)
```

```
##
## 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
```

in this model x2 is significant to the 1% level meanwhile in the other model x2 was not significant. in this model we would fail to reject the null of x2=0.

f

yes they do contradict eachother, especially for x2 because in one model its not significant at all and the next model it is significant to the 1% level. x1 also increased in significance in the model without x2.

g

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

```r
lm4 <- lm(y~x1+x2)
summary(lm4)
```

```
##
## 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
```

this point seems to have made x2 statistically significant instead of x1. the slope for x1 decreased and the slop for x2 increases.

```
lm5 <- lm(y~x1)
summary(lm5)
```

```
##
## 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
```

the slope for x1 slightly decreased and x1 remained statistically significant.

```
lm6 <- lm(y~x2)
summary(lm6)
```

```
##
## 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
```

the slope increased and x2 remained statistically significant compared to the x2 model above.

```
outliers <- rstudent(lm4)
outliers <- abs(outliers)
outliers <- outliers[outliers > 2]
outliers
```

```
##       16       21       55       82      101 
## 2.006788 2.229257 2.272110 2.644037 2.113479
```

if i am using 2 as the threshold, then there are 5 outliers but if i am using 3 as the threshold there are no outliers.

```
influence <- cooks.distance(lm4)
cook_threshold <- 4/101
influence[influence > cook_threshold]
```

```
##          5         18         21         47         55         56         82 
## 0.04269723 0.04831531 0.05920418 0.04157469 0.07061955 0.06219450 0.04033670 
##        101 
## 1.01902104
```

there appears to be 8 points with a big influence but we can also see that the 101st point the was added last has a massive leverage at 1.019. compared to the others, the next highest is 0.6

h

```
summary(lm)$coefficients
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 2.130500  0.2318817 9.1878742 7.606713e-15
## x1          1.439555  0.7211795 1.9961126 4.872517e-02
## x2          1.009674  1.1337225 0.8905831 3.753565e-01
```

```
summary(lm2)$coefficients
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 2.112394  0.2307448 9.154676 8.269388e-15
## x1          1.975929  0.3962774 4.986227 2.660579e-06
```

```
summary(lm3)$coefficients
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 2.389949  0.1949307 12.260508 1.682395e-21
## x2          2.899585  0.6330467  4.580365 1.366430e-05
```

the intercept in lm3 is the most reliable x1 in lm2 is the most reliable x2 in lm3 is the most reliable these were where the standard error were the lowest.

i

```
library(car )
```

```
## Loading required package: carData
```

```
vif(lm)
```

```
##       x1       x2 
## 3.304993 3.304993
```

I would've expect there to be more multicolinearity. usually 5 and above is problematic and based on the analysis above it sounded like there was alot of multicolinearity but a score of 3 suggest moderate colinieartiy.

2

beta hat1 is an alternative way to estimate the slope in a linear regression model, and it is unbiased. However, I would prefer the least squares estimator because it is more commonly used and generally performs better in practice.

3

a

```
odds <- 0.37
probability <- odds / (1 + odds)
probability
```

```
## [1] 0.270073
```

about 27% of people with an odds of 0.37 will default

b

```
probability <- 0.16
odds <- probability / (1 - probability)
odds
```

```
## [1] 0.1904762
```

The odds of default are 0.1905

4

a

```
beta0 <- 6
beta1 <- 0.05
beta2 <- 1

hours_studied <- 40
gpa <- 3.5

model <- beta0 + beta1 * hours_studied + beta2 * gpa

probability <- 1 / (1 + exp(-model))
probability
```

```
## [1] 0.9999899
```

The estimated probability of getting an A is 99%

b

```
linear_sum <- -(beta0 + beta2 * gpa)
hours_needed <- linear_sum / beta1
hours_needed
```

```
## [1] -190
```

I am getting -190 which is not possible. i think the gpa is already high enough so the model is predicting negative hours.

5

a

```
library(tibble)

df <- data.frame(
  X1 = c(0, 2, 0, 0, 1, 1),
  X2 = c(3, 0, 1, 1, 0, 1),
  X3 = c(0, 0, 3, 2, 1, 1)
)

Y = c("Red", "Red", "Red", "Green", "Green", "Red")

euclidean_distances <- apply(df, 1, function(row) {
  sqrt(sum(row^2))
})

df$Y <- Y

df$DistanceFromOrigin <- euclidean_distances

print(df)
```

```
##   X1 X2 X3     Y DistanceFromOrigin
## 1  0  3  0   Red           3.000000
## 2  2  0  0   Red           2.000000
## 3  0  1  3   Red           3.162278
## 4  0  1  2 Green           2.236068
## 5  1  0  1 Green           1.414214
## 6  1  1  1   Red           1.732051
```

b

when K=1 the 5th observation has the smallest distance so we would predict green.

c

when K=3 the 5th, 6th, and 2nd observations have the smallest distances. there responses are green, red, and red. So we would predict red.

d

I would expect the best value for k to be small because a smalll K would accomadate more flexibility.