Homework 1

Ben Chu

February 5, 2018

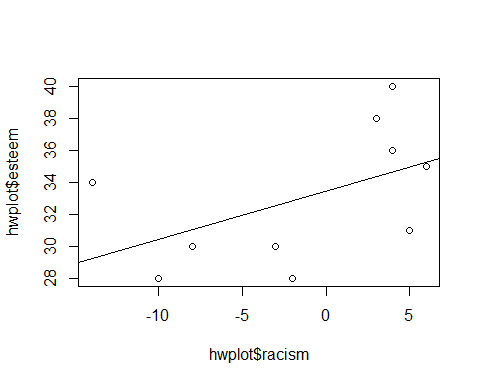
# Question 1

## Hand writing a correlation and double checking with R.

esteem <- c(28,30,34,35,31,36,28,30,38,40)  
racism <- c(-10,-8,-14,6,5,4,-2,-3,3,4)  
mean.x <- mean(esteem)  
mean.y <- mean(racism)   
sd.x <- sd(esteem)  
sd.y <- sd(racism)  
sx.sy <- sd.x\*sd.y  
xprime <- esteem - mean.x  
yprime <- racism - mean.y  
xyprime <- sum(yprime\*xprime)  
covxy <- xyprime/9  
correlation.coefficient <- covxy/sx.sy

## Graph with abline

hwplot <- data.frame(esteem,racism)  
plot(hwplot$racism,hwplot$esteem)  
abline(lm(hwplot$esteem~hwplot$racism))



Feelings of esteem are not positive correlated with feelings of racism *r* = .50, *p* = 0.13

## Question 2

### Generating a regression equation by hand and double checking with R.

sx2 <- (sd.x)^2  
by <- covxy/sx2  
a <- mean.y-(by\*mean.x)  
ycarrot<-a+(by\*esteem)

## Question 3

### Generating sums of squares with hand calculations and double checking with R.

sstotal <- sum(yprime^2)  
ssregression <- sum((ycarrot-mean.y)^2)  
ssresidual <- sum((racism-(a+(by\*esteem)))^2)  
r2 <-ssregression/sstotal  
r2

## [1] 0.2554696

model<-lm(esteem~racism)  
summary(model)

##   
## Call:  
## lm(formula = esteem ~ racism)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8497 -2.5232 -0.6503 3.0724 5.3470   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.4508 1.2502 26.756 4.1e-09 \*\*\*  
## racism 0.3006 0.1814 1.657 0.136   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.859 on 8 degrees of freedom  
## Multiple R-squared: 0.2555, Adjusted R-squared: 0.1624   
## F-statistic: 2.745 on 1 and 8 DF, p-value: 0.1361

## 4

These calculations were calculated with r previously in the aforementioned questions.

correlation.coefficient

## [1] 0.50544

ycarrot

## [1] -5.75 -4.05 -0.65 0.20 -3.20 1.05 -5.75 -4.05 2.75 4.45

r2

## [1] 0.2554696

## 5

cor.test(esteem,racism)

##   
## Pearson's product-moment correlation  
##   
## data: esteem and racism  
## t = 1.6568, df = 8, p-value = 0.1361  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.1821550 0.8610479  
## sample estimates:  
## cor   
## 0.50544

95%[-0.18,0.86]. There is no relationship between esteem and racism. The null hypothesis cannot be rejected because there may be a zero relationship difference. Essentially, we are less than 5% sure that is zero is not potential correlation.

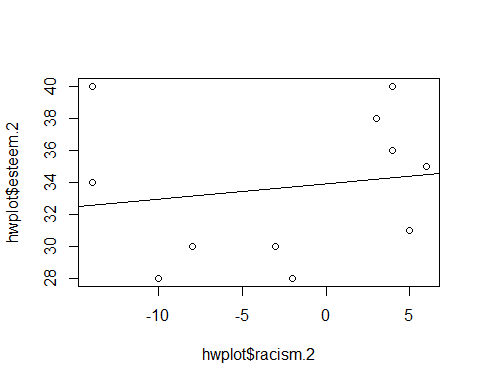
## 6

### Adding in an 11th data point.

esteem.2 <- c(28,30,34,35,31,36,28,30,38,40,40)  
racism.2 <- c(-10,-8,-14,6,5,4,-2,-3,3,4,-14)  
cor.test(esteem.2,racism.2)

##   
## Pearson's product-moment correlation  
##   
## data: esteem.2 and racism.2  
## t = 0.49215, df = 9, p-value = 0.6344  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.4850969 0.6943338  
## sample estimates:  
## cor   
## 0.1618868

hwplot <- data.frame(esteem.2,racism.2)  
plot(hwplot$racism.2,hwplot$esteem.2)  
abline(lm(hwplot$esteem.2~hwplot$racism.2))



The addition of the 11th data point increases the correlation coefficient from .51 to .16. This is dramatically pulling down the correlation. These scores are fairly distanced from the regression line which decreases the explanation power of variance.

## q7

### loading in the data

library(tidyverse)  
load("C:/Users/Branly Mclanbry/Downloads/hw1\_7.RData")  
q7 <- hw1\_7 %>% janitor::clean\_names()

Now looking at the different types of predictions.

examscores <- lm(score~satv, dat = q7)  
summary(examscores)

##   
## Call:  
## lm(formula = score ~ satv, data = q7)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.7161 -2.9347 -0.0879 4.1323 12.2839   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.42295 10.92180 1.046 0.30524   
## satv 0.05812 0.01815 3.202 0.00359 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.808 on 26 degrees of freedom  
## Multiple R-squared: 0.2828, Adjusted R-squared: 0.2552   
## F-statistic: 10.25 on 1 and 26 DF, p-value: 0.003588

lm.beta::lm.beta(examscores)

##   
## Call:  
## lm(formula = score ~ satv, data = q7)  
##   
## Standardized Coefficients::  
## (Intercept) satv   
## 0.000000 0.531767

lmSupport::modelEffectSizes(examscores)

## lm(formula = score ~ satv, data = q7)  
##   
## Coefficients  
## SSR df pEta-sqr dR-sqr  
## (Intercept) 36.8955 1 0.0404 NA  
## satv 345.7545 1 0.2828 0.2828  
##   
## Sum of squared errors (SSE): 877.0  
## Sum of squared total (SST): 1222.7

Higher scores are predicted by previous exam scores *R*2 = .28, *F*(1,26) = 10.25, *p* < .01. *b*\* = .53

## Q7

### loading the data up

load("C:/Users/Branly Mclanbry/Downloads/employee (5).RData")  
q8 <- employee %>% janitor::clean\_names()

regression analysis

emp.mod <- lm(salbegin~jobtime+prevexp, dat = q8)  
emp.mod.2 <- lm(salbegin~jobtime+prevexp + educ, dat = q8)  
lmSupport::modelCompare(emp.mod,emp.mod.2)

## SSE (Compact) = 29229622992   
## SSE (Augmented) = 16156369619   
## Delta R-Squared = 0.4461723   
## Partial Eta-Squared (PRE) = 0.4472604   
## F(1,470) = 380.31, p = 1.704446e-62

summary(emp.mod)

##   
## Call:  
## lm(formula = salbegin ~ jobtime + prevexp, data = q8)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8604 -4506 -2141 520 62845   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17951.985 2960.187 6.064 2.72e-09 \*\*\*  
## jobtime -15.558 36.003 -0.432 0.666   
## prevexp 3.401 3.463 0.982 0.327   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7878 on 471 degrees of freedom  
## Multiple R-squared: 0.002433, Adjusted R-squared: -0.001803   
## F-statistic: 0.5743 on 2 and 471 DF, p-value: 0.5635

lmSupport::modelSummary(emp.mod)

## lm(formula = salbegin ~ jobtime + prevexp, data = q8)  
## Observations: 474  
##   
## Linear model fit by least squares  
##   
## Coefficients:  
## Estimate SE t Pr(>|t|)   
## (Intercept) 17951.985 2960.187 6.064 2.72e-09 \*\*\*  
## jobtime -15.558 36.003 -0.432 0.666   
## prevexp 3.401 3.463 0.982 0.327   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Sum of squared errors (SSE): 29229622991.8, Error df: 471  
## R-squared: 0.0024

lm.beta::lm.beta(emp.mod)

##   
## Call:  
## lm(formula = salbegin ~ jobtime + prevexp, data = q8)  
##   
## Standardized Coefficients::  
## (Intercept) jobtime prevexp   
## 0.00000000 -0.01988807 0.04519486

lmSupport::modelEffectSizes(emp.mod.2)

## lm(formula = salbegin ~ jobtime + prevexp + educ, data = q8)  
##   
## Coefficients  
## SSR df pEta-sqr dR-sqr  
## (Intercept) 234980004 1 0.0143 NA  
## jobtime 82579793 1 0.0051 0.0028  
## prevexp 1324031446 1 0.0757 0.0452  
## educ 13073253372 1 0.4473 0.4462  
##   
## Sum of squared errors (SSE): 16156369619.5  
## Sum of squared total (SST): 29300904965.5

lmSupport::varDescribe(q8)

## vars n mean sd median min  
## id 1 474 2.375000e+02 136.98 2.375000e+02 1.0000e+00  
## gender\* 2 474 1.540000e+00 0.50 2.000000e+00 1.0000e+00  
## bdate 3 473 1.180185e+10 371721297.37 1.196882e+10 1.0929e+10  
## educ 4 474 1.349000e+01 2.88 1.200000e+01 8.0000e+00  
## jobcat\* 5 474 1.410000e+00 0.77 1.000000e+00 1.0000e+00  
## salary 6 474 3.441957e+04 17075.66 2.887500e+04 1.5750e+04  
## salbegin 7 474 1.701609e+04 7870.64 1.500000e+04 9.0000e+03  
## jobtime 8 474 8.111000e+01 10.06 8.100000e+01 6.3000e+01  
## prevexp 9 474 9.586000e+01 104.59 5.500000e+01 0.0000e+00  
## minority\* 10 474 1.220000e+00 0.41 1.000000e+00 1.0000e+00  
## lgsalary 11 474 4.500000e+00 0.17 4.460000e+00 4.2000e+00  
## lgexp 12 474 1.660000e+00 0.64 1.750000e+00 0.0000e+00  
## sqsal 13 474 1.811700e+02 40.01 1.699300e+02 1.2550e+02  
## invsal 14 474 0.000000e+00 0.00 0.000000e+00 0.0000e+00  
## invexp 15 474 9.000000e-02 0.22 2.000000e-02 0.0000e+00  
## sqrtexp 16 474 8.460000e+00 5.04 7.480000e+00 1.0000e+00  
## mah\_2 17 474 2.000000e+00 1.88 1.370000e+00 4.0000e-02  
## mah\_1 18 474 2.000000e+00 2.36 1.080000e+00 7.0000e-02  
## salb\_sq 19 474 1.279300e+02 25.54 1.224700e+02 9.4870e+01  
## salb\_lg 20 474 4.200000e+00 0.15 4.180000e+00 3.9500e+00  
## salb\_in 21 474 0.000000e+00 0.00 0.000000e+00 0.0000e+00  
## current\_salary 22 474 3.442000e+01 17.08 2.888000e+01 1.5750e+01  
## newvar 23 474 1.000000e+00 0.00 1.000000e+00 1.0000e+00  
## max skew kurtosis  
## id 4.740000e+02 0.00 -1.21  
## gender\* 2.000000e+00 -0.18 -1.97  
## bdate 1.225437e+10 -0.85 -0.59  
## educ 2.100000e+01 -0.11 -0.29  
## jobcat\* 3.000000e+00 1.45 0.24  
## salary 1.350000e+05 2.11 5.27  
## salbegin 7.998000e+04 2.83 12.18  
## jobtime 9.800000e+01 -0.05 -1.16  
## prevexp 4.760000e+02 1.50 1.65  
## minority\* 2.000000e+00 1.35 -0.17  
## lgsalary 5.130000e+00 0.99 0.65  
## lgexp 2.680000e+00 -0.81 0.32  
## sqsal 3.674200e+02 1.51 2.26  
## invsal 0.000000e+00 -0.05 -0.27  
## invexp 1.000000e+00 3.73 12.69  
## sqrtexp 2.184000e+01 0.61 -0.41  
## mah\_2 1.109000e+01 1.45 2.02  
## mah\_1 1.425000e+01 2.29 5.93  
## salb\_sq 2.828100e+02 1.91 4.85  
## salb\_lg 4.900000e+00 1.27 1.74  
## salb\_in 0.000000e+00 -0.31 -0.24  
## current\_salary 1.350000e+02 2.11 5.27  
## newvar 1.000000e+00 NaN NaN