Final Midterm Project Simulation

Dr. Phil

2023-02-27

## Introduction

The Purpose of this report is to simulate data for the in-class question where the fitted regression model is:

where the Midterm exam is scored from 0 to 100 and the project is scored from 0 to 30. The question is which predictor variable (Midterm or Project) has a stronger association with the response (Final).

We saw in class that when Midterm=100 and project=30 (full points for both), the predicted Final score is 100. We also noted that when Midterm = 100, it contributes points to the Final score, and when Project = 30, it contributes points to the Final score. So it seems obvious that Midterm has a stronger association with Final.

We made that conclusion, in part, because we were aware of the different ranges of possible values for Midterm and Project. If we did not have that information, we might think that Project has a stronger relationship, because is has a larger coefficient (i.e., slope): 1.2 vs. 0.53. As we begin doing multiple regression with many predictors, we need to keep this in mind.

One way to deal with this issue is to standardize all of the predictor variables, that is, put them all on the same scale. We explore some different options for this in the analysis below. Note that regardless of how the predictor variables are standardized, the t-value and p-value for each predictor remain the same.

### Set up 3D surface plotting

This code just sets up 3D scatter plot and surface plotting.

lm\_surface <- function(lmfit, n=10, alpha=0.2, ...) {   
 coeffs = lmfit$coefficients  
 f = function (x1, x2) coeffs[1] + coeffs[2]\*x1 + coeffs[3]\*x2  
 ranges <- rgl:::.getRanges()  
 x <- seq(ranges$xlim[1], ranges$xlim[2], length=n)  
 y <- seq(ranges$ylim[1], ranges$ylim[2], length=n)  
 z <- outer(x,y,f)  
 surface3d(x, y, z, alpha=alpha, ...)  
}  
  
library(rgl)

## Data Simulation

Next, we simulate some data using the equation above. We take the population parameter values to be:

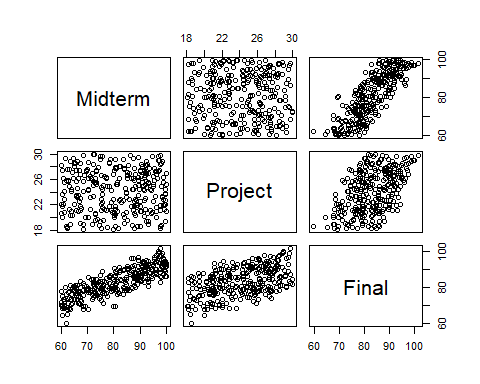
There are different ways to vary this simulation. The “levers” are:

1. The range of values for the Midterm
2. The range of values for the Project
3. The residual standard error
4. The sample size

This first data uses what might be considered a typical range of values for the Midterm (60-100) and the Project (18-30), that is, the possible scores are between 60% and 100% of the total available.

We’ll start with a larger sample size and a lower residual standard error to demonstrate that the estimated coefficients are close to the population values.

# Set the simulation seed value, so that we get the same simulated residuals  
# each time  
  
set.seed (12345)  
  
n = 300  
sig.e = 2.5  
Midterm = runif (n, 60, 100)  
Project = runif (n, 18, 30)  
Final = 11 + 0.53 \* Midterm + 1.2 \* Project + rnorm (n, 0, sig.e)  
  
sim.df = data.frame (Midterm, Project, Final)  
plot (sim.df)



Fit the model and predict Final when Midterm=100 and Project=30.

m1 = lm (Final ~ Midterm + Project)  
summary (m1)

##   
## Call:  
## lm(formula = Final ~ Midterm + Project)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8069 -1.4294 0.0691 1.5833 8.1662   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.46967 1.43076 8.016 2.52e-14 \*\*\*  
## Midterm 0.53883 0.01201 44.858 < 2e-16 \*\*\*  
## Project 1.15889 0.04382 26.450 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.473 on 297 degrees of freedom  
## Multiple R-squared: 0.9032, Adjusted R-squared: 0.9025   
## F-statistic: 1385 on 2 and 297 DF, p-value: < 2.2e-16

predict (m1, list (Midterm=100, Project=30), interval = 'confidence')

## fit lwr upr  
## 1 100.1195 99.40172 100.8374

Here’s a 3D scatterplot with the fitted regression surface (must be run interactively to view):

plot3d (Midterm, Project, Final, type="p", col="red",   
 xlab="Midterm", ylab="Project", zlab="Final")  
  
lm\_surface (m1)

Note in the output above that Midterm has a larger t-value and a smaller p-value, which means it has a stronger statistical relationship with Final.

### Rescale Project to match Midterm

Using the same data set, let’s rescale the Project score to be 0 to 100 (from 0 to 30).

Project100 = Project \* 10/3  
  
m1a = lm (Final ~ Midterm + Project100)  
summary (m1a)

##   
## Call:  
## lm(formula = Final ~ Midterm + Project100)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8069 -1.4294 0.0691 1.5833 8.1662   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.46967 1.43076 8.016 2.52e-14 \*\*\*  
## Midterm 0.53883 0.01201 44.858 < 2e-16 \*\*\*  
## Project100 0.34767 0.01314 26.450 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.473 on 297 degrees of freedom  
## Multiple R-squared: 0.9032, Adjusted R-squared: 0.9025   
## F-statistic: 1385 on 2 and 297 DF, p-value: < 2.2e-16

predict (m1a, list (Midterm=100, Project100=100), interval = 'confidence')

## fit lwr upr  
## 1 100.1195 99.40172 100.8374

Note that the y-intercept and the Midterm slope are exactly the same as before (estimate, standard error, t-value and p-value).

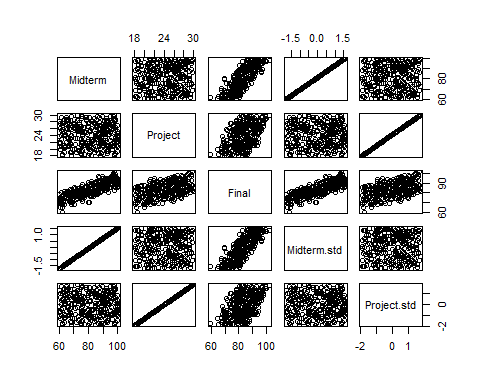
Also note that the coefficient (slope) for Project is exactly the previous coefficient times 3/10: 0.34767 = 1.15887\*3/10. And the t-value and p-value are exactly the same as before.

Also note that the predicted Final score with Midterm=100 and Project100=100 (Project=30) is exactly the same as before, including the confidence interval for the prediction.

## Standardize both predictor variables to have mean=0 and sd=1

Next we standardize both predictor variables to have a mean of 0 and a standard deviation of 1.

sim.df$Midterm.std = scale (sim.df$Midterm)[,1]  
sim.df$Project.std = scale (sim.df$Project)[,1]  
  
plot (sim.df)



m2 = lm (Final ~ Midterm.std + Project.std, data=sim.df)  
summary (m2)

##   
## Call:  
## lm(formula = Final ~ Midterm.std + Project.std, data = sim.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8069 -1.4294 0.0691 1.5833 8.1662   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 83.3025 0.1428 583.50 <2e-16 \*\*\*  
## Midterm.std 6.4166 0.1430 44.86 <2e-16 \*\*\*  
## Project.std 3.7834 0.1430 26.45 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.473 on 297 degrees of freedom  
## Multiple R-squared: 0.9032, Adjusted R-squared: 0.9025   
## F-statistic: 1385 on 2 and 297 DF, p-value: < 2.2e-16

predict (m2, list (Midterm.std=(100 - mean (Midterm))/sd (Midterm),   
 Project.std=(30 - mean(Project))/sd (Project)), interval = 'confidence')

## fit lwr upr  
## 1 100.1195 99.40172 100.8374

In this case, all of the coefficients are different, but the t-values and p-values for the slopes are still the same. And Midterm still has the stronger association with Final (larger t-value).

### 3D Plot with the standardized data

Must be run interactively to view.

plot3d (sim.df$Midterm.std, sim.df$Project.std, Final, type="p", col="red",   
 xlab="Midterm std", ylab="Project std", zlab="Final")  
lm\_surface(m1)

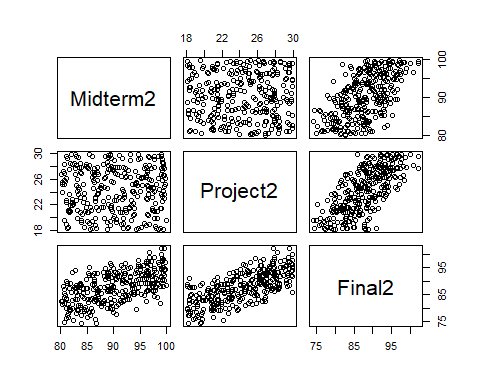
### Conclusions from the analyses above:

* Beware of interpreting the relative sizes of regression coefficients when the predictor variables are on different scales.
* The t-value and p-value for each predictor don’t change when you rescale that predictor, so t-values and/or p-values can be used to judge the relative statistical importance across predictor variables.

# The range of each predictor matters

Next, we will demonstrate what happens when the Midterm has a narrower range of values: 80 to 100 instead of 60 to 100.

Midterm2 = runif (n, 80, 100)  
Project2 = runif (n, 18, 30)  
Final2 = 11 + 0.53 \* Midterm2 + 1.2 \* Project2 + rnorm (n, 0, sig.e)  
  
sim.df2 = data.frame (Midterm2, Project2, Final2)  
plot (sim.df2)



with Midterm having a narrower range, even with the same population coefficient values, it now appears that Project has a stronger association with Final.

3D scatterplot and regression fit:

plot3d (Midterm2, Project2, Final2, type="p", col="red",   
 xlab="Midterm", ylab="Project", zlab="Final")  
  
m3 = lm (Final2 ~ Midterm2 + Project2, data=sim.df2)  
summary (m3)

##   
## Call:  
## lm(formula = Final2 ~ Midterm2 + Project2, data = sim.df2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.2337 -1.5880 -0.0975 1.8123 5.8513   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.26125 2.37795 3.474 0.000589 \*\*\*  
## Midterm2 0.56805 0.02405 23.617 < 2e-16 \*\*\*  
## Project2 1.16161 0.04086 28.428 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.444 on 297 degrees of freedom  
## Multiple R-squared: 0.8239, Adjusted R-squared: 0.8227   
## F-statistic: 694.6 on 2 and 297 DF, p-value: < 2.2e-16

predict (m3, list (Midterm2=100, Project2=30), interval = 'confidence')

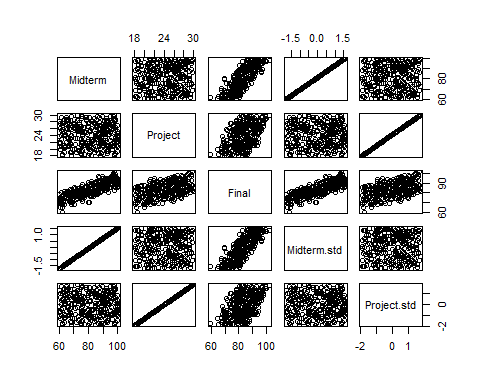
## fit lwr upr  
## 1 99.9148 99.21601 100.6136

lm\_surface (m3)

Now it appears that Project has a stronger relationship with Final. The parameter estimates are sill about the same, so we can still make the argument that Midterm has the strong effect. And it does, if we’re talking about the actual (practical) effect size. Remember that statistical significance and practical importance are two different things. Midterm still has the larger practical importance, but it’s effect is less statistically significant, because we have a narrower range of values (80-100).

### Standardize the predictor variables

sim.df2$Midterm.std = scale (sim.df2$Midterm2)[,1]  
sim.df2$Project.std = scale (sim.df2$Project2)[,1]  
  
plot (sim.df)



m4 = lm (Final2 ~ Midterm.std + Project.std, data=sim.df2)  
summary (m4)

##   
## Call:  
## lm(formula = Final2 ~ Midterm.std + Project.std, data = sim.df2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.2337 -1.5880 -0.0975 1.8123 5.8513   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 87.8199 0.1411 622.41 <2e-16 \*\*\*  
## Midterm.std 3.3384 0.1414 23.62 <2e-16 \*\*\*  
## Project.std 4.0183 0.1414 28.43 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.444 on 297 degrees of freedom  
## Multiple R-squared: 0.8239, Adjusted R-squared: 0.8227   
## F-statistic: 694.6 on 2 and 297 DF, p-value: < 2.2e-16

predict (m4, list (Midterm.std=(100 - mean (Midterm2)) / sd (Midterm2),   
 Project.std=(30 - mean (Project2)) / sd (Project2)), interval = 'confidence')

## fit lwr upr  
## 1 99.9148 99.21601 100.6136

### Standardizing the predictor variables does not affect any of these results:

* Predicted response and its confidence (or prediction) interval
* t-values and p-values on the coefficients
* R^2, R\_adj^2, residual standard error, F-test and its p-value