Stat 415 Regression: Homework 1

Frankie Tyndall

2022-07-03

## Grade point average Dataset

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student’s grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X).The results of the study follow. Assume that first-order regression model is appropriate.

x is the ACT test score Y is grade point average (GPA) at the end of the freshman year

gpaact

## # A tibble: 120 × 2  
## gpa actscore  
## <dbl> <dbl>  
## 1 3.90 21  
## 2 3.88 14  
## 3 3.78 28  
## 4 2.54 22  
## 5 3.03 21  
## 6 3.86 31  
## 7 2.96 32  
## 8 3.96 27  
## 9 0.5 29  
## 10 3.18 26  
## # … with 110 more rows

## Obtain the least squares estimates of Bo and B1 and state the estimated regression function.

lm(gpaact$gpa ~ gpaact$actscore)

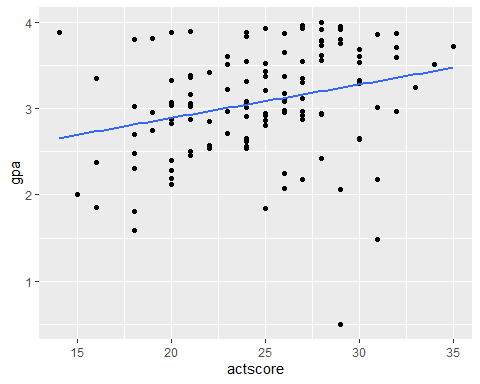
##   
## Call:  
## lm(formula = gpaact$gpa ~ gpaact$actscore)  
##   
## Coefficients:  
## (Intercept) gpaact$actscore   
## 2.11405 0.03883

Bo = y intercept = 2.11405 B1 = slope = 0.03883 y(hat) = Bo + B1x y(hat) = 2.11405 + 0.03883x

## Plot the estimated regression function and the scatter plot. Does the estimated regression function appear to fit the data well?

qplot(x = actscore, y = gpa, data = gpaact, geom = "point") +  
 geom\_smooth(method = lm, se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'



gpaact.out <- lm(gpaact$gpa ~ gpaact$actscore)  
summary(gpaact.out)

##   
## Call:  
## lm(formula = gpaact$gpa ~ gpaact$actscore)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.74004 -0.33827 0.04062 0.44064 1.22737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.11405 0.32089 6.588 1.3e-09 \*\*\*  
## gpaact$actscore 0.03883 0.01277 3.040 0.00292 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6231 on 118 degrees of freedom  
## Multiple R-squared: 0.07262, Adjusted R-squared: 0.06476   
## F-statistic: 9.24 on 1 and 118 DF, p-value: 0.002917

cor(gpaact$gpa, gpaact$actscore)

## [1] 0.2694818

Based on the scatter plot above, it looks like there is a moderately positive linear relationship between the two quantitative variables. The points seem somewhat clustered about the fitted regression line. The estimated regression function is y(hat) = 2.11405 + 0.03883x and the p value of the predictor variable (actscore) is 0.00292 which is below the 0.05 threshold rejecting the null hypothesis that the theoritical slope is equal to 0. The multiple R squared is 0.07262. The correlation coefficient is 0.2694818 which shows a weak relationship between the gpa and act scores. This regression model/ function does not seem to fit the data well. There seems to be extreme outliers which could be having a neagtive impact on the ability for the model to predict.

## Give a full interpretation of the slope for the regression function that you generated in part a.

Bo = y intercept = 2.11405 B1 = slope = 0.03883 y(hat) = Bo + B1x y(hat) = 2.11405 + 0.03883x

The slope indicates that for every increase by 1 point in act score, the grade point average (GPA) at the end of the freshman year increases by 0.03883 point on average.

## Obtain a point estimate of the mean freshman GPA for students with ACT test score X = 28

y(hat) = Bo + B1x y(hat) = 2.11405 + 0.03883x y(hat) = 2.11405 + 0.03883(28) y(hat) = 3.20129

## As demonstrated in class, use R coding to produce a regression output summary table and indicate the typical distance between the sample slope estimate and the true population slope estimate. Also, fine the proportion of variability in the response variable that is explained by your regression model.

gpaact.out <- lm(gpaact$gpa ~ gpaact$actscore)  
summary(gpaact.out)

##   
## Call:  
## lm(formula = gpaact$gpa ~ gpaact$actscore)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.74004 -0.33827 0.04062 0.44064 1.22737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.11405 0.32089 6.588 1.3e-09 \*\*\*  
## gpaact$actscore 0.03883 0.01277 3.040 0.00292 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6231 on 118 degrees of freedom  
## Multiple R-squared: 0.07262, Adjusted R-squared: 0.06476   
## F-statistic: 9.24 on 1 and 118 DF, p-value: 0.002917

The typical distance between a sampled slope estimate and the true slope from the population is called the standard error of slope which is 0.01277. The proportion of variability in the response variable is the multiple r-squared which is 0.07262. This means that 7.262% of the variation in the response variable (GPA) can be explained by the regression model with the predictor variable (actscore).

## Find the residual for X = 28, and determine if the observed value is above or below average.

observed\_for28 <- gpaact%>% filter(actscore == 28)  
observed\_for28

## # A tibble: 10 × 2  
## gpa actscore  
## <dbl> <dbl>  
## 1 3.78 28  
## 2 3.73 28  
## 3 3.56 28  
## 4 2.42 28  
## 5 2.93 28  
## 6 4 28  
## 7 3.62 28  
## 8 3.79 28  
## 9 3.91 28  
## 10 2.95 28

y(hat) = Bo + B1x y(hat) = 2.11405 + 0.03883x y(hat) = 2.11405 + 0.03883(28) y(hat) = expected value = 3.20129

residual = observed - expected residual = 3.778 - 3.20129 = 0.57671. Being that the residual is positive this means that the observed value of 3.778 is above average or above the fitted line.

## As demonstrated in class, use R coding to find a 95% confidence interval for B1

confidence\_B1 <- lm(gpaact$gpa ~ gpaact$actscore)  
confidence\_B1

##   
## Call:  
## lm(formula = gpaact$gpa ~ gpaact$actscore)  
##   
## Coefficients:  
## (Intercept) gpaact$actscore   
## 2.11405 0.03883

confidence\_B1\_1 <- tidy(confidence\_B1, conf.int = TRUE)  
select(confidence\_B1\_1, term, estimate, p.value, conf.low, conf.high)

## # A tibble: 2 × 5  
## term estimate p.value conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 2.11 0.00000000130 1.48 2.75   
## 2 gpaact$actscore 0.0388 0.00292 0.0135 0.0641

We are 95% confident that the true population slope falls between 0.01353307 and 0.06412118

## As demonstrated in class, use R coding to find a 95% confidence interval for an input value of X = 28.

confidence\_y <- lm(gpaact$gpa ~ gpaact$actscore)  
confidence\_y

##   
## Call:  
## lm(formula = gpaact$gpa ~ gpaact$actscore)  
##   
## Coefficients:  
## (Intercept) gpaact$actscore   
## 2.11405 0.03883

new\_df <- data.frame(actscore = 28)  
new\_df

## actscore  
## 1 28

predict(object = confidence\_y, newdata = new\_df, interval = "confidence") %>%  
cbind(new\_df)

## Warning: 'newdata' had 1 row but variables found have 120 rows

## fit lwr upr actscore  
## 1 2.929419 2.782565 3.076273 28  
## 2 2.657629 2.363893 2.951365 28  
## 3 3.201209 3.061384 3.341033 28  
## 4 2.968246 2.836187 3.100305 28  
## 5 2.929419 2.782565 3.076273 28  
## 6 3.317690 3.123060 3.512320 28  
## 7 3.356517 3.140763 3.572272 28  
## 8 3.162382 3.035890 3.288873 28  
## 9 3.240036 3.083891 3.396181 28  
## 10 3.123555 3.006385 3.240725 28  
## 11 3.045900 2.931773 3.160028 28  
## 12 3.278863 3.104246 3.453481 28  
## 13 3.045900 2.931773 3.160028 28  
## 14 3.045900 2.931773 3.160028 28  
## 15 3.395344 3.157650 3.633039 28  
## 16 3.162382 3.035890 3.288873 28  
## 17 3.084727 2.971869 3.197586 28  
## 18 3.317690 3.123060 3.512320 28  
## 19 3.084727 2.971869 3.197586 28  
## 20 2.890592 2.726359 3.054825 28  
## 21 3.045900 2.931773 3.160028 28  
## 22 2.929419 2.782565 3.076273 28  
## 23 3.201209 3.061384 3.341033 28  
## 24 3.162382 3.035890 3.288873 28  
## 25 3.201209 3.061384 3.341033 28  
## 26 3.123555 3.006385 3.240725 28  
## 27 3.201209 3.061384 3.341033 28  
## 28 2.968246 2.836187 3.100305 28  
## 29 3.123555 3.006385 3.240725 28  
## 30 2.929419 2.782565 3.076273 28  
## 31 3.084727 2.971869 3.197586 28  
## 32 2.735283 2.487507 2.983060 28  
## 33 3.201209 3.061384 3.341033 28  
## 34 3.123555 3.006385 3.240725 28  
## 35 2.968246 2.836187 3.100305 28  
## 36 3.045900 2.931773 3.160028 28  
## 37 2.929419 2.782565 3.076273 28  
## 38 3.278863 3.104246 3.453481 28  
## 39 3.162382 3.035890 3.288873 28  
## 40 3.123555 3.006385 3.240725 28  
## 41 3.123555 3.006385 3.240725 28  
## 42 3.278863 3.104246 3.453481 28  
## 43 3.045900 2.931773 3.160028 28  
## 44 3.123555 3.006385 3.240725 28  
## 45 3.240036 3.083891 3.396181 28  
## 46 3.045900 2.931773 3.160028 28  
## 47 3.317690 3.123060 3.512320 28  
## 48 2.696456 2.425906 2.967006 28  
## 49 2.851765 2.668303 3.035227 28  
## 50 2.812938 2.608919 3.016956 28  
## 51 3.162382 3.035890 3.288873 28  
## 52 2.735283 2.487507 2.983060 28  
## 53 3.162382 3.035890 3.288873 28  
## 54 3.123555 3.006385 3.240725 28  
## 55 3.045900 2.931773 3.160028 28  
## 56 3.278863 3.104246 3.453481 28  
## 57 2.929419 2.782565 3.076273 28  
## 58 2.890592 2.726359 3.054825 28  
## 59 3.278863 3.104246 3.453481 28  
## 60 3.240036 3.083891 3.396181 28  
## 61 3.084727 2.971869 3.197586 28  
## 62 3.007073 2.886274 3.127873 28  
## 63 3.084727 2.971869 3.197586 28  
## 64 3.007073 2.886274 3.127873 28  
## 65 3.278863 3.104246 3.453481 28  
## 66 2.929419 2.782565 3.076273 28  
## 67 3.045900 2.931773 3.160028 28  
## 68 3.356517 3.140763 3.572272 28  
## 69 2.812938 2.608919 3.016956 28  
## 70 3.007073 2.886274 3.127873 28  
## 71 2.890592 2.726359 3.054825 28  
## 72 3.007073 2.886274 3.127873 28  
## 73 2.812938 2.608919 3.016956 28  
## 74 2.812938 2.608919 3.016956 28  
## 75 3.240036 3.083891 3.396181 28  
## 76 2.890592 2.726359 3.054825 28  
## 77 3.007073 2.886274 3.127873 28  
## 78 3.123555 3.006385 3.240725 28  
## 79 3.201209 3.061384 3.341033 28  
## 80 3.434172 3.173927 3.694416 28  
## 81 2.890592 2.726359 3.054825 28  
## 82 2.890592 2.726359 3.054825 28  
## 83 3.123555 3.006385 3.240725 28  
## 84 3.356517 3.140763 3.572272 28  
## 85 3.084727 2.971869 3.197586 28  
## 86 3.162382 3.035890 3.288873 28  
## 87 3.162382 3.035890 3.288873 28  
## 88 3.240036 3.083891 3.396181 28  
## 89 2.851765 2.668303 3.035227 28  
## 90 2.929419 2.782565 3.076273 28  
## 91 3.045900 2.931773 3.160028 28  
## 92 3.162382 3.035890 3.288873 28  
## 93 3.084727 2.971869 3.197586 28  
## 94 2.812938 2.608919 3.016956 28  
## 95 3.240036 3.083891 3.396181 28  
## 96 3.045900 2.931773 3.160028 28  
## 97 3.162382 3.035890 3.288873 28  
## 98 2.929419 2.782565 3.076273 28  
## 99 2.851765 2.668303 3.035227 28  
## 100 2.812938 2.608919 3.016956 28  
## 101 3.084727 2.971869 3.197586 28  
## 102 2.812938 2.608919 3.016956 28  
## 103 2.890592 2.726359 3.054825 28  
## 104 3.356517 3.140763 3.572272 28  
## 105 3.045900 2.931773 3.160028 28  
## 106 3.472999 3.189741 3.756256 28  
## 107 3.084727 2.971869 3.197586 28  
## 108 3.201209 3.061384 3.341033 28  
## 109 3.201209 3.061384 3.341033 28  
## 110 3.084727 2.971869 3.197586 28  
## 111 2.968246 2.836187 3.100305 28  
## 112 3.278863 3.104246 3.453481 28  
## 113 2.890592 2.726359 3.054825 28  
## 114 2.890592 2.726359 3.054825 28  
## 115 3.317690 3.123060 3.512320 28  
## 116 2.890592 2.726359 3.054825 28  
## 117 3.240036 3.083891 3.396181 28  
## 118 3.201209 3.061384 3.341033 28  
## 119 2.735283 2.487507 2.983060 28  
## 120 3.201209 3.061384 3.341033 28

## Find an estimate for the variance of the population regression function

# n = 120  
# degrees of freedom = n - 2   
# = 120 - 2 = 118  
  
gpplt <- gpaact  
anova(lm(gpplt$gpa ~ gpplt$actscore))

## Analysis of Variance Table  
##   
## Response: gpplt$gpa  
## Df Sum Sq Mean Sq F value Pr(>F)   
## gpplt$actscore 1 3.588 3.5878 9.2402 0.002917 \*\*  
## Residuals 118 45.818 0.3883   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The estimated variance of the population regression function is 0.3883

## Refer to regression model above, what is the implication for the regression function if B1 = 0 so that the model is Yi = B0 + ei? How would the regression function plot on a graph? Provide a detailed answer of 5 or 6 sentences.

B1 equaling 0 would mean that the slope is 0 indicating that the predictor and response variable do not have a linear relationship and that the predictor variable (X) does not fit as predictor of the distribution of the response variable (Y). A A zero slope is just the slope of a horizontal line! This indicates that y coordinate never changes no matter what the x coordinate. A slope of zero would mean that your regression line would be perfectly horizontal. If the slope is zero, then no x will be put into the regression model so you are left with B0 which is the y intercept and ei which is the residual.

## Again using the regression model above, Is it true or false that the model always produces a data point that is exactly on the regression line regardless of the value of the error term? Justify your answer.

This statement is false because the equation, Yi = B(0) + b(1)X(i) without the inclusion of the error value is what produces data points that are exactly on the regression line for an X input. The equation, Yi = B(0) + B(1)X(i) + E(i), produces data points that may or may not be on the regression line for an X input value and a given error value. So saying that the points will alwat be on the regression line, regardless of the error value is false, making the statement not true.