Lab 1 STA 5303

Salman Imtiaz

Table of Contents

Part 1: Data analysis continuous predictor: Grade point average (GPA.Rdata).

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). Use this data set to answer the questions below

Question 1

Produce summary statistics for each variable separately (mean, median, standard deviation etc) as well as a graph (histogram or boxplot) for each. Are there any features of the data of concern or interest?

TEXT ANSWER HERE

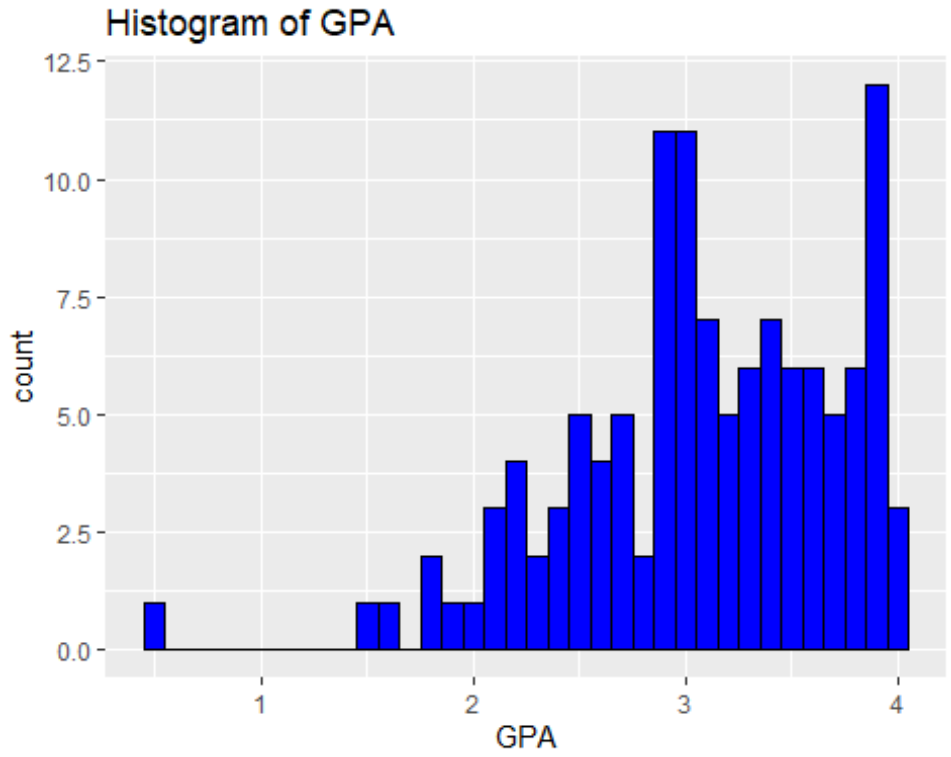
```
## Summary statistics
summary(GPA)
```

```
##       GPA             ACT       
##  Min.   :0.500   Min.   :14.00  
##  1st Qu.:2.689   1st Qu.:21.00  
##  Median :3.078   Median :25.00  
##  Mean   :3.074   Mean   :24.73  
##  3rd Qu.:3.593   3rd Qu.:28.00  
##  Max.   :4.000   Max.   :35.00
```

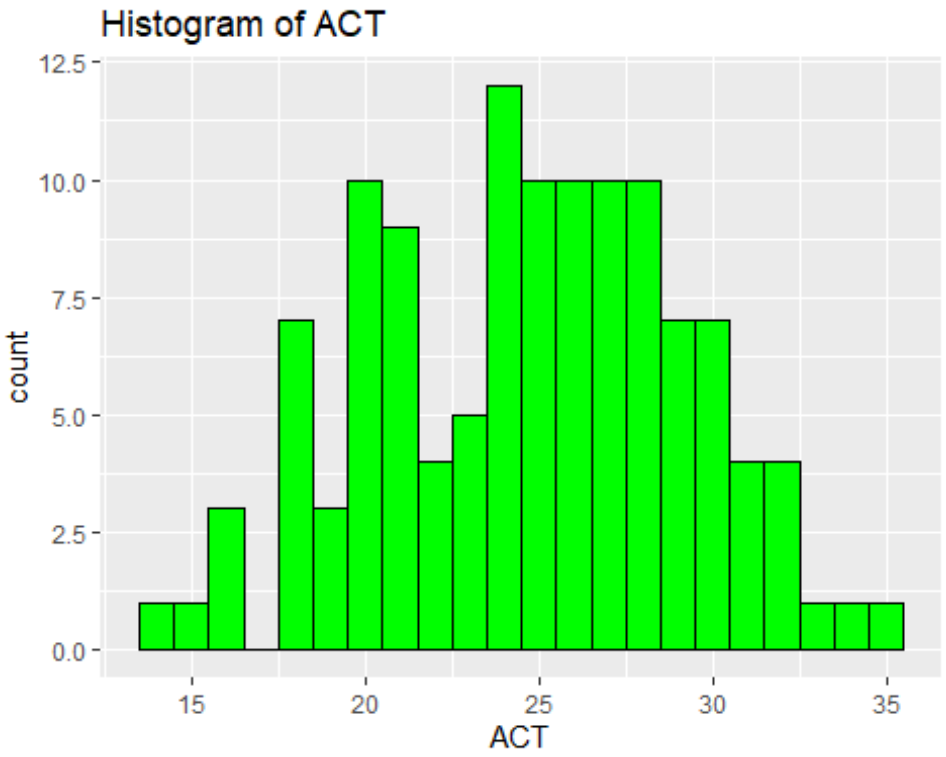
```r
GPA_summary <- GPA %>%
  summarise(across(everything(), list(
    mean = ~mean(.),
    median = ~median(.),
    sd = ~sd(.)
  )))
print(GPA_summary)
```

```
##   GPA_mean GPA_median    GPA_sd ACT_mean ACT_median   ACT_sd
## 1  3.07405     3.0775 0.6443383   24.725         25 4.472065
```

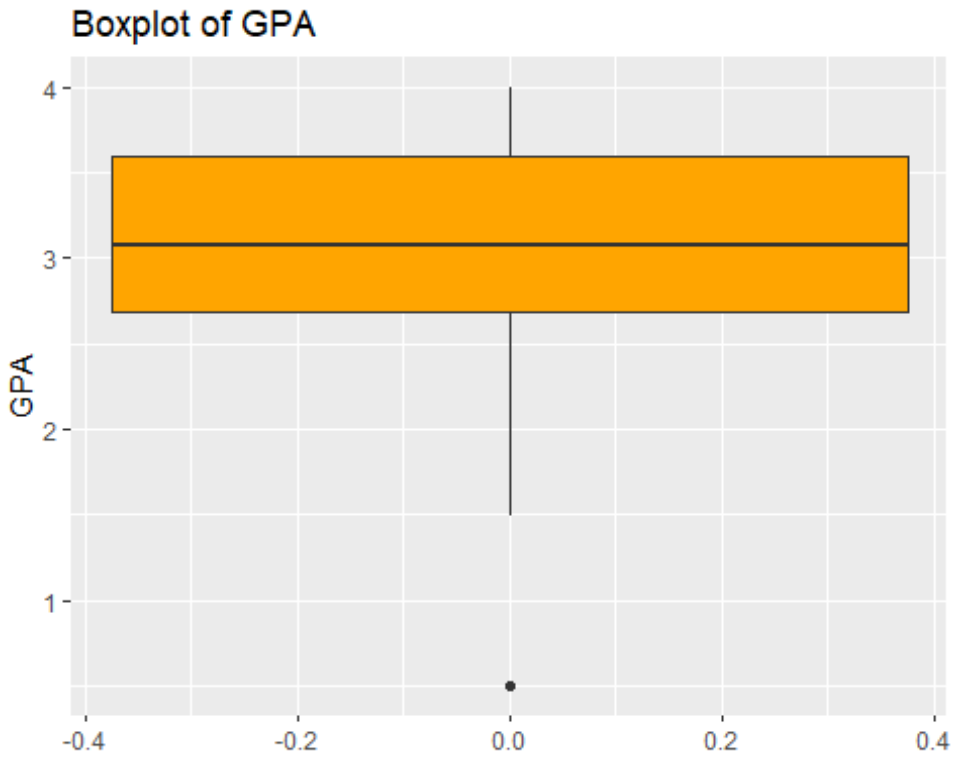
```r
## Histograms
ggplot(GPA, aes(x = GPA)) +
  geom_histogram(binwidth = 0.1, fill = 'blue', color = 'black') +
  labs(title = 'Histogram of GPA')
```



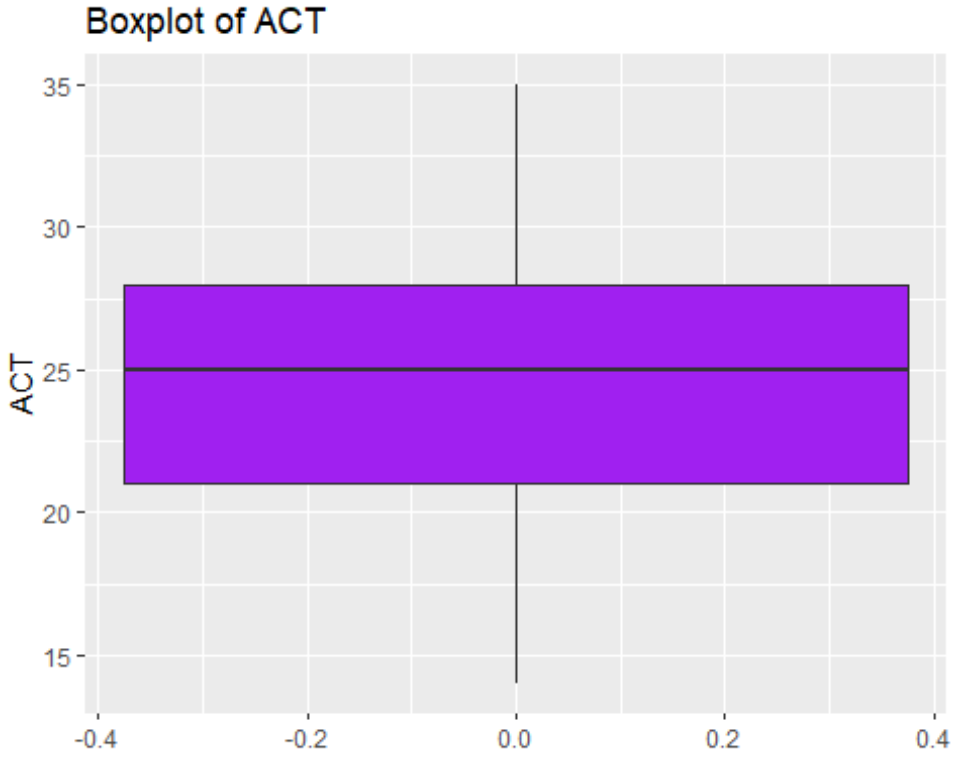

```r
ggplot(GPA, aes(x = ACT)) +
  geom_histogram(binwidth = 1, fill = 'green', color = 'black') +
  labs(title = 'Histogram of ACT')
```




```
## Boxplots
ggplot(GPA, aes(y = GPA)) +
  geom_boxplot(fill = 'orange') +
  labs(title = 'Boxplot of GPA')
```




```
ggplot(GPA, aes(y = ACT)) +
  geom_boxplot(fill = 'purple') +
  labs(title = 'Boxplot of ACT')
```



Question 2

Obtain the least squares estimates of the parameters for a simple linear regression model.

```
## Linear model
lm_GPA_ACT <- lm(GPA ~ ACT, data = GPA)
summary(lm_GPA_ACT)

##
## Call:
## lm(formula = GPA ~ ACT, data = GPA)
##
## 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 ***
## ACT          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

## Estimated regression function
coefficients(lm_GPA_ACT)

## (Intercept)         ACT
##  2.11404929  0.03882713
```

a. State the estimated regression function.

The estimated regression function is GPA = 2.114 + 0.039 * ACT. This equation predicts GPA based on the ACT score.

b. Interpret the parameter estimates in context. Are both relevant/meaningful? What is the point estimate of the change in the mean response when the ACT test score increases by one point?

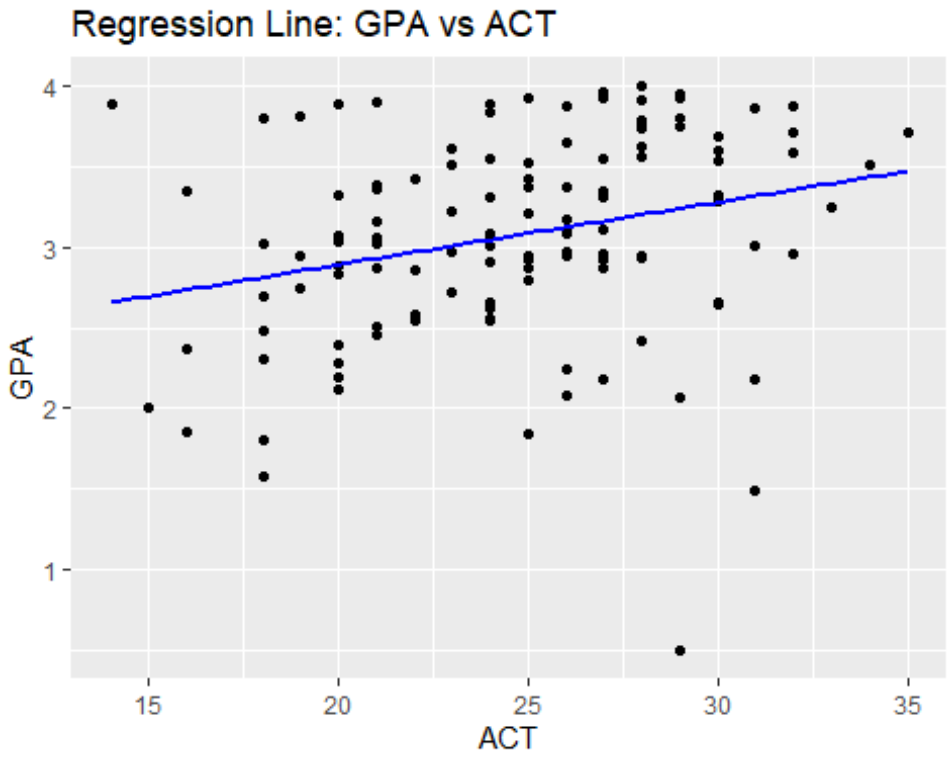
The intercept (2.114) represents the estimated GPA when the ACT score is 0, which, while not practically meaningful since ACT scores can't be zero, serves as a part of the linear model. The slope (0.039) indicates that for each one-point increase in ACT score, the GPA is expected to increase by approximately 0.039 points, suggesting a positive relationship between ACT scores and GPA.

c. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

The plot of the estimated regression line over the scatterplot of the data shows how well the line fits. In this case, if the data points are somewhat scattered around the line, it suggests a modest linear fit.

```
## Plot the estimated regression function
ggplot(GPA, aes(x = ACT, y = GPA)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = 'Regression Line: GPA vs ACT')

## `geom_smooth()` using formula = 'y ~ x'
```



d. Obtain a point estimate of the mean freshman GPA for students with ACT test score X = 30. How does this compare to the overall average GPA for students in the data set? Explain.

The point estimate for GPA when ACT = 30 is approximately 3.28. This is slightly higher than the overall mean GPA, suggesting that students with an ACT score of 30 tend to have a higher GPA.

```
## Point estimate of the mean GPA for ACT = 30
predict(lm_GPA_ACT, newdata = data.frame(ACT = 30))

##        1
## 3.278863
```

Question 3

Residuals and error. Obtain the residuals.

```
## Residuals
residuals <- residuals(lm_GPA_ACT)
residuals
```

```
##           1           2           3           4           5           6
##  0.96758105  1.22737094  0.57679116 -0.42824608  0.09858105  0.54730978
##           7           8           9          10          11          12
## -0.39451735  0.79861829 -2.74003597  0.05444541  0.26409967  0.25913691
##          13          14          15          16          17          18
##  0.03709967 -0.03290033 -0.15034448 -0.19938171  0.43727254 -0.30469022
##          19          20          21          22          23          24
## -0.13772746 -0.77259183 -0.48290033  0.42758105  0.52979116  0.76261829
##          25          26          27          28          29          30
##  0.35479116 -0.02255459 -0.78120884 -0.38924608  0.74744541  0.13058105
##          31          32          33          34          35          36
##  0.84227254 -0.36028332 -0.27220884  0.25144541 -0.11124608  0.02609967
##          37          38          39          40          41          42
##  0.45158105  0.01113691  0.38661829  0.52244541 -0.14555459 -0.62486309
##          43          44          45          46          47          48
## -0.50590033 -0.87355459 -1.17103597 -0.42890033 -1.13469022 -0.69645619
##          49          50          51          52          53          54
##  0.10023530  0.99306243 -0.29138171  0.61671668  0.14261829 -0.17155459
##          55          56          57          58          59          60
##  0.50109967  0.41213691  0.23058105 -0.69659183  0.04413691  0.69596403
##          61          62          63          64          65          66
## -0.16272746 -0.29107321  0.28527254  0.59892679 -0.63686309 -0.47741895
##          67          68          69          70          71          72
## -0.39090033  0.35748265 -1.00693757  0.50892679  0.14840817 -0.04107321
##          73          74          75          76          77          78
## -0.33093757 -0.11293757  0.67996403 -0.05659183  0.21492679 -0.03955459
##          79          80          81          82          83          84
##  0.79879116  0.07682840  0.43240817  0.18140817 -1.04455459  0.51848265
##          85          86          87          88          89          90
##  0.12327254 -0.24238171  0.18261829  0.71596403  0.95623530 -0.42341895
##          91          92          93          94          95          96
##  0.84009967 -0.97938171  0.34427254  0.21106243  0.50996403  0.78709967
##          97          98          99         100         101         102
## -0.04938171 -0.05441895 -0.10476470 -0.50193757 -1.24372746 -1.22993757
##         103         104         105         106         107         108
## -0.01159183  0.23448265 -0.13190033  0.24300127 -0.28472746  0.41979116
##         109         110         111         112         113         114
##  0.59079116 -0.21772746  0.45075392  0.32113691 -0.49659183 -0.60459183
##         115         116         117         118         119         120
## -1.83169022  0.99440817  0.55996403  0.71279116 -0.87528332 -0.25320884
```

a. Confirm that they sum to zero (hint: sum function in R). Are there any particularly large (or small – more negative) values?

The residuals sum to approximately zero (-6.793177e-15), which is expected in a properly fitted linear regression model. Some residuals appear larger in magnitude, indicating some points deviate notably from the fitted line.

```
## Sum of residuals
sum(residuals)

## [1] -6.793177e-15
```

b. What is the largest (in absolute value) residual and what does the residual mean in context.

The largest residual is approximately 2.74. This means that there is a student whose GPA deviates from the predicted GPA by 2.74 points, suggesting this student either overperformed or underperformed compared to the model's expectations.

```
## Largest residual
max_residual <- max(abs(residuals))
max_residual

## [1] 2.740036
```

c. Obtain the estimate of the standard deviation of the errors and interpret this quantity.

The standard deviation of the residuals is approximately 0.62, indicating the average distance of the data points from the regression line. A lower value would suggest a tighter fit of the model.

```
## Estimate of the standard deviation of errors
sd_residuals <- sd(residuals)
sd_residuals

## [1] 0.6205013
```

Question 4

Obtain a 95 percent confidence interval for the true slope parameter. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

The 95% confidence interval for the slope is (0.0135, 0.0641). Since this interval does not include zero, it suggests a statistically significant positive relationship between ACT scores and GPA. The director of admissions might be interested in this interval to understand the strength and direction of the association.

```
## 95% Confidence interval for the true slope parameter
confint(lm_GPA_ACT, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 1.47859015 2.74950842
## ACT         0.01353307 0.06412118
```

Question 5

Conduct a hypothesis test based on the regression model of whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Clearly state the alternatives, decision rule, and your conclusion (include p-value). How does the test compare to the analysis of the confidence interval in the previous question?

The hypothesis test results indicate that the p-value for the ACT coefficient is 0.00292, which is less than 0.05. Therefore, we reject the null hypothesis and conclude that there is a statistically significant linear association between ACT scores and GPA. This is consistent with the confidence interval analysis, which also suggested a significant relationship.

```
## Summary includes p-value for hypothesis test
summary(lm_GPA_ACT)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = GPA)
##
## 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 ***
## ACT          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
```

Question 6

Confidence and prediction intervals. a. Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval.

The 95% confidence interval for the mean GPA for an ACT score of 28 is (3.061, 3.341). This interval estimates the range where the true mean GPA for students with an ACT score of 28 likely falls.

```
## Confidence interval for mean GPA when ACT = 28
predict(lm_GPA_ACT, newdata = data.frame(ACT = 28), interval = "confidence",
level = 0.95)
```

```
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
```

b. Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA using a 95 percent prediction interval. Interpret your prediction interval.

The 95% prediction interval for Mary Jones's GPA is (1.959, 4.443). This wider interval accounts for individual variability, indicating where Mary Jones's GPA is likely to fall.

```
## Prediction interval for Mary Jones with ACT = 28
predict(lm_GPA_ACT, newdata = data.frame(ACT = 28), interval = "prediction",
level = 0.95)

##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
```

c. Is the prediction interval in part b wider than the confidence interval in the part a? Should it be?

Yes, the prediction interval is wider than the confidence interval because it includes the variability in individual GPAs around the regression line, not just the mean estimate.

```
## Compare the width of the intervals
conf_interval <- predict(lm_GPA_ACT, newdata = data.frame(ACT = 28), interval
= "confidence", level = 0.95)
pred_interval <- predict(lm_GPA_ACT, newdata = data.frame(ACT = 28), interval
= "prediction", level = 0.95)

# Width of the confidence interval
conf_width <- conf_interval[3] - conf_interval[2]
conf_width

## [1] 0.279649

# Width of the prediction interval
pred_width <- pred_interval[3] - pred_interval[2]
pred_width

## [1] 2.483708

# Check if the prediction interval is wider
is_wider <- pred_width > conf_width
is_wider

## [1] TRUE
```

Question 7

What is the estimated R2 for this SLR model? Interpret this value in context. Obtain the estimated correlation, r, between GPA and ACT test score. What is the interpretation of this value. Which measure, R2 or r, has the more clear-cut operational interpretation? Explain.

The R2 value is 0.0726, indicating that approximately 7.26% of the variability in GPA can be explained by the ACT score. The correlation coefficient is 0.269, suggesting a weak positive linear relationship. R^2 provides a clear-cut interpretation of how much variance is explained by the model, whereas the correlation coefficient indicates the strength and direction of the relationship.

```
## R-squared
summary(lm_GPA_ACT)$r.squared

## [1] 0.07262044

## Correlation
cor(GPA$GPA, GPA$ACT)

## [1] 0.2694818
```

Part 2: Data analysis categorical predictor: Market Share Data (Market.Rdata)

Executives from a large packaged foods manufacturer are interested in which factors influenced the market share (Y = share, average monthly market share for the product in percent) for one of their products. The data collected is for 36 consecutive months from September 1999 to August 2002.

Question 8

Executives are interested in whether their promotions have been effective. Fit a model to predict the market share (share) based on the predictor variable promo (presence of a promotion: 1 if promotion, 0 otherwise).

```
## Linear model for Market dataset
lm_Market_share <- lm(share ~ promo, data = Market)
summary(lm_Market_share)

##
## Call:
## lm(formula = share ~ promo, data = Market)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.465 -0.205 -0.035  0.160  0.425
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.57500    0.06386  40.321   <2e-16 ***
## promo        0.16000    0.08568   1.867   0.0705 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2555 on 34 degrees of freedom
## Multiple R-squared:  0.09302,    Adjusted R-squared:  0.06635
## F-statistic: 3.487 on 1 and 34 DF,  p-value: 0.07049
```

a. Interpret the estimated coefficient of promo in context.

The coefficient for promo is 0.16, suggesting that when there is a promotion, the market share is expected to increase by 0.16 percentage points, on average.

b. Is there a significant relationship between promotions and market share (given supporting evidence from the model)?

The p-value for the promo coefficient is 0.0705, which is slightly higher than the 0.05 significance level. Therefore, we fail to reject the null hypothesis, suggesting that there is not enough evidence to conclude a statistically significant relationship between promotions and market share at the 5% level.

Question 9

Is there a significant relationship of market share to the year (1999 – 2002). Fit the model to answer this question (treating year as a factor variable!).

```
## Convert year to factor
Market$year <- as.factor(Market$year)

## Linear model with year as a factor
lm_year <- lm(share ~ year, data = Market)
summary(lm_year)

##
## Call:
## lm(formula = share ~ year, data = Market)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3900 -0.2165 -0.0525  0.2106  0.4742
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.71500    0.13654  19.884   <2e-16 ***
## year2000    -0.02917    0.15767  -0.185    0.854
## year2001    -0.10750    0.15767  -0.682    0.500
## year2002    -0.02500    0.16723  -0.149    0.882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2731 on 32 degrees of freedom
## Multiple R-squared:  0.02446,    Adjusted R-squared:  -0.067
## F-statistic: 0.2675 on 3 and 32 DF,  p-value: 0.8483
```

```r
## Overall F-test
anova(lm_year)
```

```
## Analysis of Variance Table
##
## Response: share
##           Df  Sum Sq  Mean Sq F value Pr(>F)
## year       3 0.05984 0.019946  0.2675 0.8483
## Residuals 32 2.38642 0.074576
```

a. Conduct an overall "chunk" test (F test) to determine if year is significant.

The overall F-test for the variable year has a p-value of 0.8483, which is much higher than 0.05. This suggests that the year is not a significant predictor of market share, implying no significant differences in market share across the years 1999 to 2002.


```r
## Overall F-test
anova(lm_year)
```

```
## Analysis of Variance Table
##
## Response: share
##           Df  Sum Sq  Mean Sq F value Pr(>F)
## year       3 0.05984 0.019946  0.2675 0.8483
## Residuals 32 2.38642 0.074576
```


b. What is the reference category in the model you fit?

The reference category is the first level of the year factor, which is 1999. All other year coefficients are interpreted relative to this reference year.

c. Use one of the estimates for year to illustrate how to properly interpret the coefficients from the model.

For year 2000, the coefficient is -0.02917. This means that the market share in the year 2000 is estimated to be 0.029 percentage points lower than in the reference year (1999), though this difference is not statistically significant.