CSCI E-106:Assignment 3

Shreya Bajpai

Problem 1

Refer to the Grade point average Data in Homework2. 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). (30 points, each part is 5 points)

```
library(ggplot2)
library(lattice)
library(caret)
gpa_data <- read.csv('/Users/shreyabajpai/CSCI E-106 - Data Modeling/Assignments/CSCI E-106 Assignment 3
set.seed(1023) # Make code reproducible
DataSplit<-createDataPartition(y = gpa_data$Y, p = 0.7, list = FALSE)
training_data <- gpa_data[DataSplit, ]
testing_data <- gpa_data[-DataSplit, ]
```

a-) By using the regression model in Homework2, Obtain a 99 percent confidence interval for β_1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

```
fit_gpa_act_data <- lm(Y ~ X, data=training_data)
summary(fit_gpa_act_data)
```

```
##
## Call:
## lm(formula = Y ~ X, data = training_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71521 -0.35426  0.04034  0.44288  1.18770
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.21392    0.39896   5.549 3.43e-07 ***
## X            0.03453    0.01584   2.180   0.0321 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6545 on 82 degrees of freedom
## Multiple R-squared:  0.05479,    Adjusted R-squared:  0.04327
## F-statistic: 4.754 on 1 and 82 DF,  p-value: 0.0321
```

```
confint(fit_gpa_act_data, "X", level = 0.99)
```

```
##          0.5 %     99.5 %
## X -0.007234821 0.07628952
```

The 99% confidence interval for the slope (β_1) of the regression line, [-0.0072, 0.07629], quantifies the relationship between ACT scores (X) and GPA (Y). This interval indicates:

- We are 99% confident that for each additional point increase in ACT score, the predicted GPA will increase by between -0.0072 and 0.07629.
- The inclusion of zero within this interval signifies that the slope is not statistically significant at the 99% confidence level, suggesting a weak likelihood that ACT scores positively influence GPA.
- In addition, the modest slope (0.03453) implies that while ACT scores have a positive correlation with GPA, the effect size is relatively small, indicating that an increase of one point in ACT score results in a minor change in predicted GPA.

This suggests that ACT score is a not a significant predictor of GPA.

The director of admissions may be particularly interested in whether this confidence interval includes zero because it suggests that the true slope (β_1) could potentially be zero, meaning ACT scores do not significantly predict GPA. This result reinforces the lack of utility of ACT scores as a significant predictor of student performance, invalidating their relevance in admissions decisions.

b-) Test, using the test statistic t^*, whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of $\alpha = 0.01$. State the alternatives, decision rule, and conclusion.

First, we state the hypotheses:

H_o : $\beta_1 = 0$ There is no linear association between ACT score and GPA.

H_a : $\beta_1 \neq 0$ There is a linear association between ACT score and GPA.

The decision rule is as follows:

1. If $|t^*| \leq t_{\alpha/2,n-2}$, then we fail to reject H_0, concluding that there is no evidence of a linear association between ACT score and GPA.
2. If $|t^*| > t_{\alpha/2,n-2}$, then we reject the null hypothesis and conclude H_a, which indicates that there is evidence of a linear association between ACT score and GPA.

```
# Extract slope coefficient and its standard error
beta1_hat <- coef(summary(fit_gpa_act_data))["X", "Estimate"]  # Slope coefficient
se_beta1 <- coef(summary(fit_gpa_act_data))["X", "Std. Error"] # Standard error of the slope

# Compute the t-statistic for the slope
t_stat_gpa <- beta1_hat / se_beta1

# Determine df for the t-test
df_gpa <- nrow(training_data) - 2

# Find the critical t-value at alpha = 0.01
alpha <- 0.01
crit_val_gpa <- qt(1 - alpha/2, df_gpa)  # Two-tailed critical value

# Calculate p-value
p_val_gpa <- 2 * pt(-abs(t_stat_gpa), df_gpa)

# Output
cat("Test Statistic (t*):", t_stat_gpa, "\n")
```

```
## Test Statistic (t*): 2.180272
```

```
cat("Critical t-value (at alpha = 0.01):", crit_val_gpa, "\n")
```

```
## Critical t-value (at alpha = 0.01): 2.637123
```

```
cat("p-value:", p_val_gpa, "\n")
```

```
## p-value: 0.03210377
```

Given $|t^*| \leq t_{\alpha/2,n-2}$ where $2.1802717 < 2.6371234$ with a non-significant p-value $0.0321038 > \alpha = 0.01$, we cannot reject the null hypothesis and conclude that there is no linear association between ACT Score and GPA.

c-) What is the P-value of your test in part (b)? How does it support the conclusion reached in part (b)?

The p-value from the test in part (b) was 0.0321038, which is not less than our chosen significance level of 0.01. Therefore, we cannot reject the null hypothesis H_0, which states that there is no linear relationship between ACT scores and GPA. A large p-value suggests that observing such an extreme test statistic t^* by chance, assuming H_0 is true, is likely. This evidence indicates that a significant linear relationship does not between ACT scores and GPA. In other words, the slope of the regression line is zero, meaning that ACT scores do not significantly help predict a student's GPA.

d-) Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval.

```
frshmn_act <- data.frame(X = 28) # Set the value we will perform a confidence interval form
predicted_GPA_CI <- predict(fit_gpa_act_data, newdata = frshmn_act, interval = "confidence", level = 0.9
cat("95% Confidence Interval for mean GPA with ACT Score of 28: \n")
```

```
## 95% Confidence Interval for mean GPA with ACT Score of 28:
```

```
print(predicted_GPA_CI)
```

```
##        fit      lwr      upr
## 1 3.180683 3.006225 3.355141
```

```
# Extract values for markdown usage
fit_conf_value_gpa <- predicted_GPA_CI[, "fit"]
lwr_conf_value_gpa <- predicted_GPA_CI[, "lwr"]
upr_conf_value_gpa <- predicted_GPA_CI[, "upr"]
```

This 95% confidence interval suggests that for students with an ACT score of 28, the mean freshman GPA is estimated to be 3.1806831, with a 95% confidence interval ranging from 3.0062254 to 3.3551409. This means we are 95% confident that the true mean GPA of all students with an ACT score of 28 lies within this interval. Essentially, this interval provides insight into the expected academic performance of these students, highlighting the relationship between ACT scores and GPA.

e-) 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.

```
predicted_GPA_PI <- predict(fit_gpa_act_data, newdata = frshmn_act, interval = "prediction", level = 0.9
cat("95% Prediction Interval for Mary Jones' GPA with ACT Score of 28: \n")
```

```
## 95% Prediction Interval for Mary Jones' GPA with ACT Score of 28:
```

```
print(predicted_GPA_PI)
```

```
##        fit      lwr      upr
## 1 3.180683 1.867023 4.494343
```

```
# Extract values for markdown usage
fit_pred_value_gpa <- predicted_GPA_PI[, "fit"]
```

```r
lwr_pred_value_gpa <- predicted_GPA_PI[, "lwr"]
upr_pred_value_gpa <- predicted_GPA_PI[, "upr"]
```

We are 95% confident that Mary Jones' actual freshman GPA will fall between 1.8670233 and 4.494343. Although her predicted GPA is approximately 3.1806831, there's a chance for considerable variation due to factors we can't fully account for in the model.

When comparing this prediction interval with the confidence interval for an ACT score of 28, both point to the same estimated mean GPA. However, the prediction interval is much wider—ranging from 1.8670233 to 4.494343. The reason for this is that the prediction interval reflects both the uncertainty in estimating the mean and the natural variation in individual GPAs. In contrast, the confidence interval only addresses the accuracy of the mean estimate for all students with that ACT score. Ultimately, the broader prediction interval gives a more realistic range for an individual outcome, as it accounts for both the variability in the model and the natural fluctuations in GPA among similar students.

f-) Is the prediction interval in part (e) wider than the confidence interval in part (d)? Should it be?

The prediction interval in part (e) is wider than the confidence interval in part (d) because they serve different purposes. The confidence interval estimates the range for the mean GPA for all students with a specific ACT score, like 28. It tells us how uncertain we are about the true average GPA for students with that score.

On the other hand, the prediction interval is for a single observation—Mary Jones' GPA, for example. Since predicting one student's GPA is trickier than estimating the average for a group, this interval needs to account for both the overall model's error and the natural variability in individual outcomes.

In short, because predicting an individual is less certain than estimating a group average, the prediction interval is wider to reflect that extra uncertainty.

g-) Determine the boundary values of the 95 percent confidence band for the regression line when $X_h = 28$. Is your-confidence band wider at this point than the confidence interval in part (d)? Should it be?

```r
#Set the specific value of ACT score for which we will calculate the confidence band
X_h <- 28

# Use the predict() to get the predicted GPA for X_h
pred_val <- predict(fit_gpa_act_data, newdata = data.frame(X = X_h))

# Extract residuals and calculate their standard deviation
res_gpaact <- fit_gpa_act_data$residuals
n <- length(res_gpaact)
sd_res <- sd(res_gpaact)

# Calculate mean of X and the sum of squared differences from the mean
mean_X <- mean(training_data$X)
ss_X <- sum((training_data$X - mean_X)^2)

# Compute standard error for the prediction at X_h
se_pred <- sd_res * sqrt(1 + (1/n) + ((X_h - mean_X)^2 / ss_X))

# Define confidence level and degrees of freedom
alpha <- 0.05
df_gpa_act <- n - 2

# Get critical t-value for the confidence band, which represents the uncertainty in the regression line
t_crit <- qt(1 - alpha/2, df_gpa_act)
```

```r
# Calculate confidence band boundaries
CI_lower <- pred_val - t_crit * se_pred
CI_upper <- pred_val + t_crit * se_pred

# Print results
cat("95% Confidence Band for the regression line at X =", X_h, ":\n")
```

```
## 95% Confidence Band for the regression line at X = 28 :
```

```r
cat("Lower Bound:", CI_lower, "\n")
```

```
## Lower Bound: 1.874961
```

```r
cat("Upper Bound:", CI_upper, "\n")
```

```
## Upper Bound: 4.486405
```

```r
# Check confidence interval vs confidence band
if (CI_upper - CI_lower > (upr_conf_value_gpa - lwr_conf_value_gpa)) {
  cat("The confidence band is wider than the confidence interval in part (d), as expected.\n")
} else {
  cat("The confidence band is narrower than expected compared to the confidence interval in part (d), wh
}
```

```
## The confidence band is wider than the confidence interval in part (d), as expected.
```

First, we utilize the fitted regression model to derive the estimated coefficients (slope, intercept) and standard errors of the residuals. We then employ the regression equation to calculate the predicted value ($\hat{Y}$), which will be the center of the confidence band at this at $\{X_h = 28\}$ Next, we compute the standard error of the prediction at $\{X_h = 28\}$ using the formula: $SE_{pred} = SE(residuals) \times \sqrt{(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{SS_x})}$ where:

- $SE(residuals)$ is the standard error of the residuals
- n is the number of observations
- X_h is the value of ACT score (28 in this case)
- $\bar{X}$ is the mean of the ACT scores
- SS_x is the sum of squares of the deviations of X from the mean, i.e., $SS_x = \sum(X_i - \bar{X})^2$

Following this, we determine the critical value t^* for a 95% confidence level with *n-2* degrees of freedom. Using the predicted value and standard error, we calculate the upper and lower limits of the 95% confidence band for the regression line at $X_h = 28$:

$CI_{lower} = \hat{Y} - t^* \times SE_{pred}$

$CI_{upper} = \hat{Y} + t^* \times SE_{pred}$.

The resulting 95% confidence band for the regression line at X = 28 indicates that we can be 95% confident that the true mean GPA for students with an ACT score of 28 lies between approximately 1.875 and 4.486. This range suggests that while the predicted mean GPA for this ACT score may vary, we expect that the average GPA of all students achieving this score will fall within these bounds.

It is essential to note that the confidence band for the regression line is typically wider than the confidence interval for the mean response at the same X_h. This is because the confidence band takes into account the variability in the estimated regression line across a range of X values, whereas the confidence interval at a specific X_h (such as 28) only reflects the uncertainty in the mean GPA at that single point. For example, the *confidence interval* for mean response is narrower and focuses on the mean GPA for students with $X_h = 28$, whereas the *confidence band* is broader, accounting for both the uncertainty in the mean and the variability of individual student GPAs around that mean.

Problem 2

Refer to the Crime rate data. A criminologist studying the relationship between level of education-and crime rate in medium-sized U.S. counties collected the following data for a random sample of 84 counties; X is the percentage of individuals in the county having at least a high-school diploma, and Y is the crime rate (crimes reported per 100,000 residents) last year. (45 points, each part is 5 points)

```
crime_data <- read.csv('/Users/shreyabajpai/CSCI E-106 - Data Modeling/Assignments/CSCI E-106 Assignmen
summary(crime_data)
```

```
##        Y               X
##  Min.   : 2105   Min.   :61.00
##  1st Qu.: 5020   1st Qu.:76.00
##  Median : 6930   Median :79.00
##  Mean   : 7111   Mean   :78.60
##  3rd Qu.: 8840   3rd Qu.:82.25
##  Max.   :14016   Max.   :91.00
```

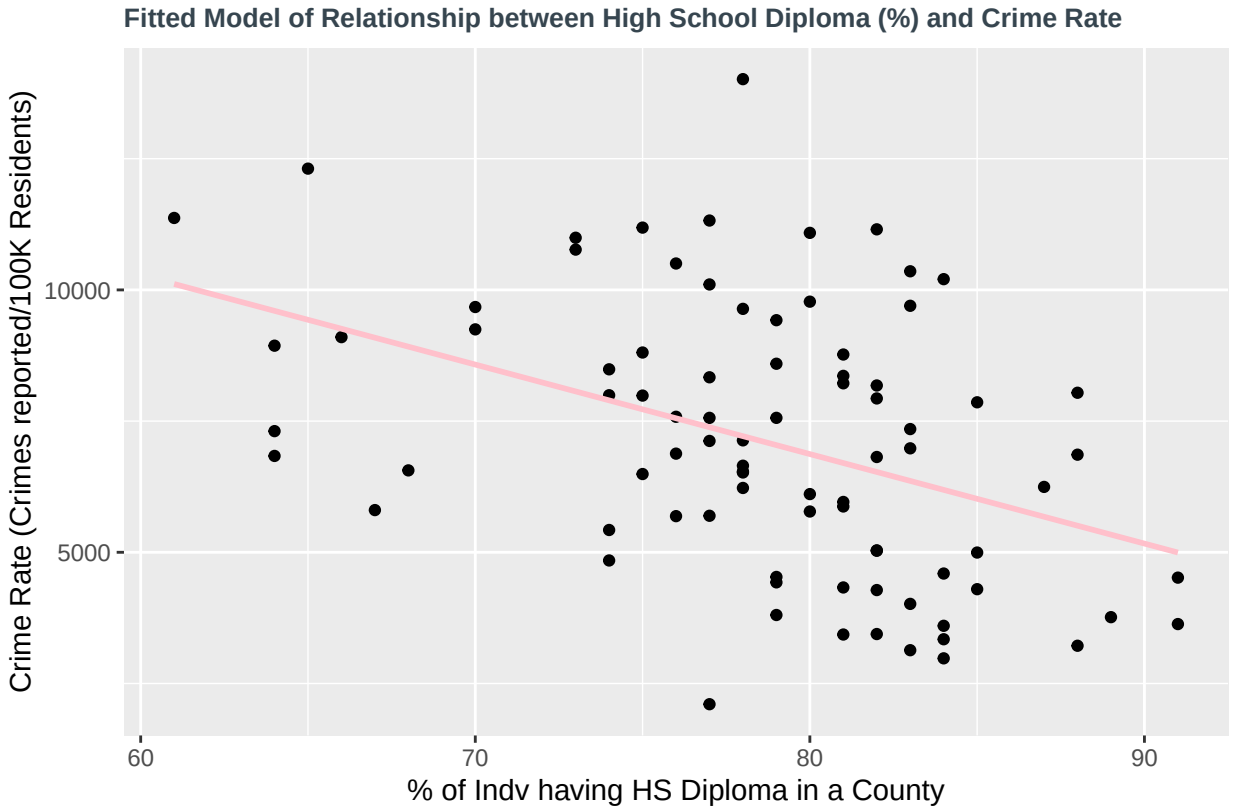
a-) Obtain the estimated regression function. Plot the estimated regression function and the data. Does the linear regression function appear to give a good fit here? Discuss.

```
fit_crmdt <- lm(Y ~ X, data=crime_data) #Y - Crime Rate, X - % of individuals in county having at least
summary(fit_crmdt)
```

```
##
## Call:
## lm(formula = Y ~ X, data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5278.3 -1757.5  -210.5  1575.3  6803.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20517.60    3277.64   6.260 1.67e-08 ***
## X            -170.58      41.57  -4.103 9.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2356 on 82 degrees of freedom
## Multiple R-squared:  0.1703, Adjusted R-squared:  0.1602
## F-statistic: 16.83 on 1 and 82 DF,  p-value: 9.571e-05
```

```
library(ggplot2)
ggplot(data=crime_data, aes(x = X, y = Y)) +
    geom_point(color = "black") +
    labs(title = "Fitted Model of Relationship between High School Diploma (%) and Crime Rate",
    caption = "Source: Crime Data.csv",
    x = "% of Indv having HS Diploma in a County",
    y = "Crime Rate (Crimes reported/100K Residents)") +
    theme(plot.title = element_text(color = "#36454F", size = 10, face = "bold")) + geom_smooth(color='
```

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



Source: Crime Data.csv

To assess the adequacy of the linear regression model for predicting crime rates, we analyze several key outputs:

1. **Residuals** The residuals range from -5278.3 to 6803.3, with a median of -210.5. This wide spread and the presence of large residuals suggest that the model struggles to capture the data's variance effectively.
2. R^2 The multiple R-squared value is 0.1703, indicating that only 17.03% of the variance in crime rates is explained by the model. A good fit would typically show an R^2 closer to 1, signaling that the model performs poorly.
3. **Significance of Coefficients** Both the intercept and the coefficient for the percentage of individuals with a high school diploma are statistically significant (p-values of 1.67e-08 and 9.57e-05, respectively). Although this indicates a relationship between the predictor and the response variable, it does not guarantee a strong model fit.
4. **Residual Standard Error** With a residual standard error of 2356, the average distance of observed values from the regression line indicates that predictions are not closely aligned with actual values, particularly given the crime rate's scale.
5. **F-Statistic** The F-statistic of 16.83, with a p-value (9.57e-05) less than 0.01, confirms the overall significance of the regression model. However, significance alone does not imply a strong fit.

Despite a statistically significant relationship between the percentage of high school graduates and crime rates, the low R-squared, high residual standard error, and extensive residual range suggest that linear regression is not an adequate model for this data.

b-) Test whether or not there is a linear association between crime rate and percentage of high school graduates, using a t test with $\alpha = 0.01$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

First, we state the hypotheses:

H_o : $\beta_1 = 0$ There is no linear association between % of Individuals with High School Diplomas and Crime Rate.

H_a : $\beta_1 \neq 0$ There is a linear association between % of Individuals with High School Diplomas and Crime Rate.

The decision rule is as follows:

1. If $|t^*| \leq t_{\alpha/2,n-2}$, then we fail to reject H_0, concluding there is *no* significant linear association between % of Individuals with High School Diplomas and Crime Rate.
2. If $|t^*| > t_{\alpha/2,n-2}$, then we reject H_0 and conclude H_a, concluding there *is* a significant linear association between % of Individuals with High School Diplomas and Crime Rate.

```
# Get coefficients and standard error for slope
beta1_hat_cd <- coef(summary(fit_crmdt))["X", "Estimate"]
se_beta1_cd <- coef(summary(fit_crmdt))["X", "Std. Error"]

# Calculate test statistic t*
t_stat_cd <- beta1_hat_cd / se_beta1_cd

# Degrees of freedom: n - 2
df_cd <- nrow(crime_data) - 2

# Decision Rule details
alpha <- 0.01
critical_value_cd <- qt(1 - alpha/2, df_cd)

# Calculate the p-value
p_value_cd <- 2 * pt(-abs(t_stat_cd), df_cd)

# Output the results
cat("Test Statistic (t*):", t_stat_cd, "\n")
```

```
## Test Statistic (t*): -4.102897
```

```
cat("Critical t-value (at alpha = 0.01):", critical_value_cd, "\n")
```

```
## Critical t-value (at alpha = 0.01): 2.637123
```

```
cat("p-value:", p_value_cd, "\n")
```

```
## p-value: 9.571396e-05
```

Since ($|t^*| = |$-4.1028971$|$) is greater than the critical value (2.6371234), and the p-value (9.5713958×10^{-5}) is less than the significance level $\alpha = 0.01$, we reject the null hypothesis. This provides significant evidence to suggest that there is a linear association between the crime rate and the percentage of high school graduates. In other words, changes in the percentage of high school graduates are associated with corresponding changes in the crime rate at the 0.01 significance level.

c-) Estimate β_1, with a 99 percent confidence interval. Interpret your interval estimate.

```
beta1 <- coef(fit_crmdt)["X"]

# Calculate the 99% confidence interval
conf_interval <- confint(fit_crmdt, level = 0.99)

# Extract the confidence interval for beta1
beta1_conf_interval <- conf_interval["X", ]
```

```r
# Output the estimate and confidence interval
cat("Estimate of Beta 1:", beta1, "\n")
```

```
## Estimate of Beta 1: -170.5752
```

```r
cat("99% Confidence Interval for Beta 1:", "\n")
```

```
## 99% Confidence Interval for Beta 1:
```

```r
print(beta1_conf_interval)
```

```
##      0.5 %     99.5 %
## -280.21182  -60.93856
```

The estimate of β_1 = -170.5751886 suggests that for each one-unit increase in the percentage of high school graduates, the crime rate is expected to decrease by approximately 170.58 units (crimes reported per 100,000 residents). The *99% confidence interval* for β_1 is $[-280.2118, -60.93856]$, meaning that we are 99% confident the true slope of the regression line lies within this range. Since the entire confidence interval is negative, this indicates a statistically significant negative relationship between the percentage of high school graduates and the crime rate. In other words, as the percentage of high school graduates increases, the crime rate tends to decrease, underscoring the importance of education in influencing societal outcomes.

d-) Set up the ANOVA table.

```r
anova_crm_dta <- anova(fit_crmdt)
print(anova_crm_dta)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df    Sum Sq  Mean Sq F value    Pr(>F)
## X          1  93462942 93462942  16.834 9.571e-05 ***
## Residuals 82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-value of 16.834 and a very low p-value of 9.571e-05 indicate that the predictor variable X, which represents the percentage of high school graduates in the county, significantly impacts the crime rate, Y. This allows us to reject the null hypothesis that the coefficient for X is zero. The ANOVA results show that a large part of the variation in crime rates can be explained by changes in the percentage of graduates. Specifically, X accounts for a sum of squares of 93,462,942, while the remaining variability (the residual sum of squares) is 455,273,165. This comparison clearly shows that X plays a substantial role in explaining the variability in Y. Given the strong F-statistic and low p-value, we can conclude that our model effectively captures the relationship between X and Y, emphasizing the importance of including percentage of High School graduates in the county to improve our predictions of crime rates.

e-) Carry out the test in part a by means of the F test. Show the numerical equivalence of the two test statistics and decision rules. Is the P-value for the F test the same as that for the t test?

```r
# Model is already built, fit_crmdt, so now we extract the required values
n_crmdt <- length(fit_crmdt$fitted.values)  # Number of observations
p_crmdt <- length(coef(fit_crmdt)) - 1      # Number of predictors (subtract 1 for intercept)
SSR_crmdt <- sum((fit_crmdt$fitted.values - mean(fit_crmdt$fitted.values))^2)  # Regression sum of squa
SSE_crmdt <- sum(residuals(fit_crmdt)^2)  # Error sum of squares
alpha_crmdt <- 0.01

# Calculate MSR and MSE
MSR_crmdt <- SSR_crmdt / p_crmdt
```

```r
MSE_crmdt <- SSE_crmdt / (n_crmdt - p_crmdt - 1)

# Calculate the F-statistic
f_stat_crmdt <- MSR_crmdt / MSE_crmdt

# Calculate the p-value
df1 <- p_crmdt
df2 <- n_crmdt - p_crmdt - 1
p_val_crmdt <- pf(f_stat_crmdt, df1, df2, lower.tail = FALSE)

# Calculate the critical F value
F_crit_crmdt <- qf(1 - alpha_crmdt, df1, df2)

# Output the results
cat("F-statistic of Crime Data:", f_stat_crmdt, "\n")
```

```
## F-statistic of Crime Data: 16.83376
```

```r
cat("p-value:", p_val_crmdt, "\n")
```

```
## p-value: 9.571396e-05
```

```r
cat("Critical value of F at alpha = 0.01:", F_crit_crmdt, "\n")
```

```
## Critical value of F at alpha = 0.01: 6.95442
```

Test	Statistic Value	Critical Value	p-value	Decision Rule
t-Test	-4.102897	2.637123	9.571396e-05	If $\|t^*\| \leq t_{\alpha/2,n-2}$, then we fail to reject H_0. If $\|t^*\| > t_{\alpha/2,n-2}$, then we reject H_0, and accept H_a
F-Test	16.83376	6.95442	9.571396e-05	If $\|F^*\| \leq F(1-\alpha; n-2)$, then we fail to reject H_0. If $\|F^*\| > F(1-\alpha; n-2)$, then we reject H_0, and accept H_a

The findings from both the t-test (-4.1028971 with p-value 9.5713958×10^{-5}, $\alpha = 0.01$) and F-test (16.8337645 with p-value 9.5713958×10^{-5}, $\alpha = 0.01$) are consistent, revealing a strong linear association between the percentage of high school graduates in a county and the crime rate. Since the p-values for both tests are below the alpha level, we can confidently reject the null hypothesis in both cases. Notably, the p-values for both tests are intrinsically linked; specifically, the p-value for the F-test arises from squaring the t-statistic, reinforcing the robustness of our findings. The significant t-statistic indicates that the individual predictor is important, while the substantial F-statistic confirms that the overall model explains a significant portion of the variance in the crime rate. Together, this analysis underscores the vital role that educational attainment plays in shaping community safety.

f-) By how much is the total variation in crime rate reduced when percentage of high school graduates is introduced into the analysis? Is this a relatively large or small reduction?

```r
# Calculate the mean of the response variable
mean_crime_rate <- mean(crime_data$Y)

# Calculate SST (Total Sum of Squares)
SST <- sum((crime_data$Y - mean_crime_rate)^2)

# Calculate SSE (Error Sum of Squares)
SSE <- sum(residuals(fit_crmdt)^2)

# Calculate the reduction in variation (SSR)
SSR <- SST - SSE

# Calculate R-squared for reference
R_squared <- 1 - (SSE / SST)

# Output results
cat("Total Variation (SST):", SST, "\n")
```

```
## Total Variation (SST): 548736108
```

```r
cat("Residual Variation (SSE):", SSE, "\n")
```

```
## Residual Variation (SSE): 455273165
```

```r
cat("Reduction in Variation:", SSR, "\n")
```

```
## Reduction in Variation: 93462942
```

```r
cat("R-Squares:", R_squared, "\n")
```

```
## R-Squares: 0.170324
```

To understand how the introduction of the percentage of high school graduates affects the variation in crime rate, we start by calculating the mean crime rate, which serves as a reference point. Next, we calculate the Total Sum of Squares (SST), which measures the total variation in the response variable Y (crime rate) before introducing any predictors. Following this, we compute the Error Sum of Squares (SSE), which represents the variation that remains unexplained after accounting for the percentage of high school graduates. The Sum of Squares for Regression (SSR), or the explained variation by our regression model, *fit_crmdt*, is then calculated as $SSR = SST - SSE$.

To assess the significance of this reduction, we calculate the proportion of the reduction relative to the total variation, commonly expressed as R^2:

$Proportion of Reduction = \frac{Reduction}{SST} \times 100 = \frac{93462942}{548736108} \times 100 \approx 17.03.$

This reduction of approximately 17.03% indicates that the percentage of high school graduates accounts for a meaningful portion of the total variability in crime rates. However, this is a moderate reduction, suggesting that other factors may also play a critical role in explaining crime rates. Thus, while the introduction of this predictor improves our understanding of the variance in crime rate, it also highlights the complexity of the underlying dynamics.

g-) State the full and reduced models (Hint reduced model is a an intercept model without X).

```r
fit_crmdt <- lm(Y ~ X, data=crime_data)

fit_red_crmdt <- lm(Y ~ 1, data = crime_data)
```

For the crime data, the *full model* includes all predictor variables and the response variable $Y = \beta_0 + \beta_1 X + \epsilon$, where:

- Y corresponds to the response variable, *crime rate*.
- X is the predictor variable, *percentage of individuals with high-school diplomas in the county*.
- β_0 is the intercept and β_1 is the slope coefficient for the predictor variable.
- ϵ is the error term.

For the crime data, the *reduced model* includes only the intercept, $Y = \beta_0 + \epsilon$. This model suggests that the response variable is constant at its mean value $\bar{Y}$ without considering any effect from the predictor variable X.

The full model evaluates the relationship between the crime rate and the percentage of high school graduates, providing insights into how this predictor affects crime rates. In contrast, the reduced model serves as a baseline for comparison, allowing us to assess the contribution of the predictor variable through methods such as the F-test or by comparing the explained variation (SST vs. SSE). This comparison enhances our understanding of the significance of the predictor in explaining variability in the response variable.

h-) Obtain (1) SSE(F), (2) SSE(R), (3) dfF. (4) dfR, (5) test statistic F* for the general linear test, (6) decision rule.

```
# Calculate SSE for the full model (SSE(F)). This is calculated above as variable 'SSE', but to track t
SSE_F <- sum(residuals(fit_crmdt)^2)

# Calculate SSE for the reduced model (SSE(R))
SSE_R <- sum(residuals(fit_red_crmdt)^2)

# Calculate df for Full Model
df_F <- n_crmdt - length(coef(fit_crmdt))

# Calculate df for Reduced Model
df_R <- n_crmdt - 1

# Now compute the F-statistic
f_stat_glt <- ((SSE_R - SSE_F)/(df_R - df_F))/(SSE_F/df_F)

# Calculate the critical F-value
alpha_glt <- 0.01
crit_f_val_glt <- qf(1 - alpha_glt, df_F, df_R)
p_value_glt <- 1 - pf(f_stat_glt, df_F, df_R)

# Print the results
cat("SSE(F):", SSE_F, "\n")
```

```
## SSE(F): 455273165
```

```
cat("SSE(R):", SSE_R, "\n")
```

```
## SSE(R): 548736108
```

```
cat("Degrees of Freedom for Full Model (df_F):", df_F, "\n")
```

```
## Degrees of Freedom for Full Model (df_F): 82
```

```
cat("Degrees of Freedom for Reduced Model (df_R):", df_R, "\n")
```

```
## Degrees of Freedom for Reduced Model (df_R): 83
```

```
cat("F-statistic (F*):", f_stat_glt, "\n")
```

```
## F-statistic (F*): 16.83376
```

```
cat("Critical F-value at alpha =", alpha_glt, ":", crit_f_val_glt, "\n")
```

```
## Critical F-value at alpha = 0.01 : 1.675753
```

```
cat("p-value:", p_value_glt, "\n")
```

```
## p-value: 0
```

```
anova(fit_red_crmdt, fit_crmdt)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ 1
## Model 2: Y ~ X
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1     83 548736108
## 2     82 455273165  1  93462942 16.834 9.571e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calculations for General Linear Test

1. The Sum of Squares for the Full Model *(SSE(F))* (residual variation with the predictor) is calculated by: $SSE_F = \sum(Y_i - \hat{Y})^2$ and for our scenario totals 4.5527317×10^8. The full model typically has a lower SSE, indicating a better fit by accounting for the variation explained by predictors.
2. The Sum of Squares for the Reduced Model *(SSE(R))* (residual variation without the predictor) is calculated by: $SSE_R = \sum(\hat{Y} - \bar{Y})^2$ and for our scenario totals 5.4873611×10^8.
3. The df_F for the Full Model uses more degrees of freedom due to the inclusion of the predictor(s), which reduces the number of degrees of freedom available for error (residuals) and is calculated by: $df_F = n - k - 1$ where k is the number of predictors (1), so effectively this is n - *2*.
4. The df_R for the Reduced Model uses fewer degrees of freedom as you're estimating only one parameter (the intercept) and is calculated by: $df_R = n - 1$.
5. Critical F value at $\alpha = 0.01$ is 1.6757533 and the p-value is 0.

The decision rule is as follows:

1. If $F^* \leq F(1 - \alpha; df_R - df_F, df_F)$, we fail to reject H_0: $y_i = \beta_0 + \epsilon_i$ or $\beta_1 = 0$. The full model does not provide a significantly better fit than the reduced model. In other words, the percentage of high school graduates does not have a significant effect on the crime rate.
2. If $F^* > F(1 - \alpha; df_R - df_F, df_F)$, we reject H_0 and conclude H_a : $y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ or $\beta_1 \neq 0$. The full model does provide a significantly better fit, indicating that the percentage of high school graduates has a significant effect on the crime rate.

The F^* Statistic for the general linear test is 16.8337645, which is greater than the critical value (1.6757533), allowing us to reject the null hypothesis. This indicates that the percentage of high school graduates significantly affects the crime rate, highlighting a meaningful linear association between education and crime rates.

i-) Are the test statistic F^* and the decision rule for the general linear test numerically equivalent to those in part a?

Table 1.1: Examining F* Statistic.

Variable	Full Model F*	General Linear Test F*
F*	16.83376	16.83376
Critical Value	6.95442	1.675753
p-value	9.571396e-05	0

Variable	Full Model F*	General Linear Test F*

The F^* statistic in a full linear regression model evaluates the overall effectiveness of the model by testing whether at least one predictor variable has a significant relationship with the response variable. This is achieved by comparing the variation explained by the model (Mean Square Regression, MSR) to the unexplained variation (Mean Square Error, MSE), as shown in the formula:

$F = \frac{MeanSquareRegression}{MeanSquareError} = \frac{SSR/p}{SSE/n-p-1}$.

where MSR captures the variability accounted for by the predictors, while MSE captures the residual, unexplained variability.

Similarly, the F^*-statistic in a general linear test compares a full model (with all predictors) to a reduced model (e.g., a model with just an intercept). Both tests evaluate the same hypothesis: whether adding predictors significantly improves the model's ability to explain the response variable's variation.

In simple linear regression, where there is only one predictor, both the full model test and the general linear regression F^*-test reduce to testing whether the slope ($beta_1$) is significantly different from zero. This is why the numerical values of the F^*-statistic and decision rule for both tests are equivalent. Ultimately, both approaches assess the significance of the predictor(s) in explaining the response variable, offering a flexible interpretation of the model's performance based on the context.

Problem 3

Five observations on Y are to be taken when X = 4, 8, 12, 16, and 20, respectively. The true regression function is E(Y = 20 + 4X, and the ϵ_i are independent N(O, 25). (25 points)}

a-) Generate five normal random numbers, with mean 0 and variance 25. Consider these random numbers as the error terms for the five Y observations at X = 4,8, 12, 16, and 20 and calculate Y_1, Y_2, Y_3, Y_4 , and Y_5. Obtain the least squares estimates b_0 and b_1, when fitting a straight line to the five cases. Also calculate Y_h when X_h = 10 and obtain a 95 percent confidence interval for E(Y_h) when X_h = 10. (10 points)

```
# Generate Random Error terms
set.seed(123)
n <- 5
errors <- rnorm(n, mean = 0, sd = sqrt(25))  # N(0, 25)

# Calculate Y values using true regression function
X <- c(4, 8, 12, 16, 20)
tru_reg_fun <- function(X) {20 + 4 * X}
Y_tru <- tru_reg_fun(X)
Y <- Y_tru + errors

# Display Y values
data.frame(X, Y)
```

```
##    X         Y
## 1  4  33.19762
## 2  8  50.84911
## 3 12  75.79354
## 4 16  84.35254
## 5 20 100.64644
```

```r
# Fit the linear model
fit <- lm(Y ~ X)

# Obtain least squares estimates
b_0 <- coef(fit)[1]
b_1 <- coef(fit)[2]

# Display the estimates
b_0
```

```
## (Intercept)
##    18.44753
```

```r
b_1
```

```
##        X
## 4.210027
```

```r
# Predict Y_h for X_h = 10
X_h <- 10
Y_h <- predict(fit, newdata = data.frame(X = X_h))

# Calculate the confidence interval
conf_interval <- predict(fit, newdata = data.frame(X = X_h), interval = "confidence", level = 0.95)
conf_int_fit_rndm <- conf_interval[, "fit"]
conf_int_lwr_rndm <- conf_interval[, "lwr"]
conf_int_upr_rndm <- conf_interval[, "upr"]

# Display results
Y_h
```

```
##       1
## 60.5478
```

```r
conf_interval
```

```
##       fit      lwr      upr
## 1 60.5478 53.86933 67.22627
```

First, we generate five normally random numbers with a mean of 0 and a variance of 25. Next, using the true regression function, $\mathrm{E}\{\mathrm{Y}\} = 20 + 4\mathrm{X}$, we can compute the corresponding Y values for the given X values, and add the generated error terms. Once the data is ready, we can fit a linear regression model to the data. The estimated intercept b_0 is 18.4475323 and the slope b_1 is 4.2100266. For $X_h = 10$, the predicted value of Y_h is 60.5477981, with a 95% confidence interval, which reflects the uncertainty around the predicted mean response, ranging from 53.8693306 and 67.2262657.

b-) Repeat part (a) 200 times, generating new random numbers each time. (5 points)

```r
set.seed(123)
output <- data.frame(Fit = numeric(200), Lwr = numeric(200), Upr = numeric(200)) # stores the predicted

for (i in 1:200) {
  # Generate five normal random numbers with mean 0 and variance 25
  epsilon <- rnorm(5, mean = 0, sd = sqrt(25))

  # Calculate Y values for X = 4, 8, 12, 16, 20
  X <- c(4, 8, 12, 16, 20)
  Y <- 20 + 4 * X + epsilon
```

```r
  # Fit the linear model
  model <- lm(Y ~ X)

  # Predict for X_h = 10
  prediction <- predict(model, newdata = data.frame(X = 10), interval = "confidence", level = 0.95)

  # Store results
  output[i, ] <- c(prediction)
}

# View the results
cat("The output data frame contains the predicted values and their corresponding 95% confidence interval
```

```
## The output data frame contains the predicted values and their corresponding 95% confidence intervals
```

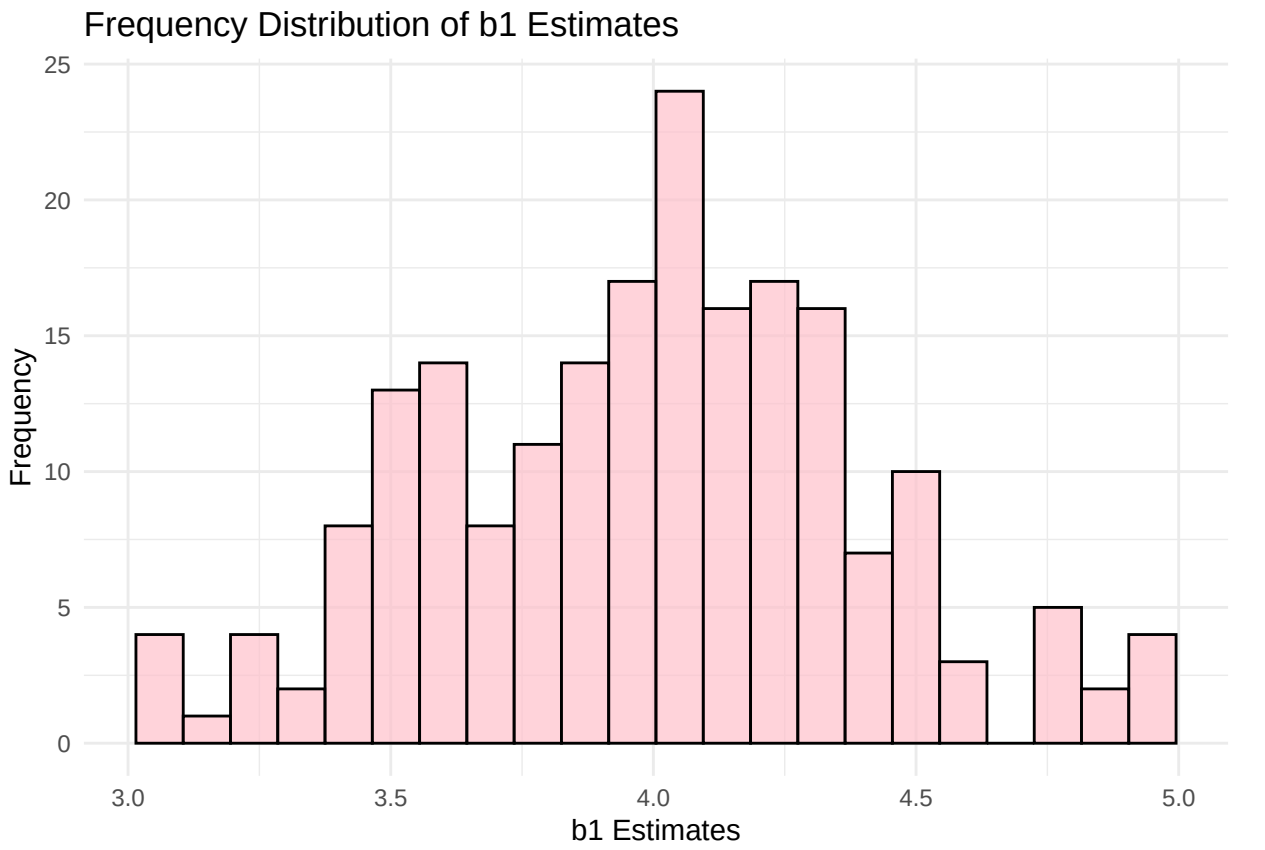
```r
head(output)
```

```
##        Fit      Lwr      Upr
## 1 60.54780 53.86933 67.22627
## 2 61.14571 54.36162 67.92980
## 3 62.49175 60.61832 64.36519
## 4 61.62569 50.86158 72.38980
## 5 56.24060 53.28232 59.19888
## 6 58.44387 48.85136 68.03639
```

c-) Make a frequency distribution of the 200 estimates b_1. Calculate the mean and standard deviation of the 200 estimates b_1. Are the results consistent with theoretical expectations? (5 points)

```r
set.seed(123)
b1_est <- numeric(200) # stores the slope estimates from each iteration

for (i in 1:200) {
  # Generate five normal random numbers with mean 0 and variance 25
  eps <- rnorm(5, mean = 0, sd = sqrt(25))

  # Calculate Y values for X = 4, 8, 12, 16, 20
  X <- c(4, 8, 12, 16, 20)
  Y <- 20 + 4 * X + eps

  # Fit the linear model
  fit_mod <- lm(Y ~ X)

  # Extract the slope (b1)
  b1_est[i] <- coef(fit_mod)[2]
}

# Calculate mean and standard deviation of b1 estimates
mu_b1 <- mean(b1_est)
sd_b1 <- sd(b1_est)

# Create a frequency distribution of b1 estimates using ggplot2
b1_data <- data.frame(b1 = b1_est)

# Create a frequency distribution of b1 estimates
```

```r
ggplot(b1_data, aes(x = b1)) +
  geom_histogram(binwidth = 0.09, fill = "pink", color = "black", alpha = 0.7) +
  labs(title = "Frequency Distribution of b1 Estimates",
       x = "b1 Estimates",
       y = "Frequency") +
  theme_minimal()
```




```r
# Output mean and standard deviation
cat("Mean of b1 estimates:", mu_b1, "\n")
```

```
## Mean of b1 estimates: 3.988583
```

```r
cat("Standard deviation of b1 estimates:", sd_b1, "\n")
```

```
## Standard deviation of b1 estimates: 0.4036461
```

The mean of the b_1 estimates across all iterations is approximately $\mu \approx 3.9885828$, indicating that for each unit increase in X, the predicted value of Y increases by about μ. The standard deviation of the b_1 estimates is approximately $\sigma \approx 0.4036461$, which reflects the variability of the slope estimates across the 200 simulations. A smaller standard deviation suggests that the estimates are relatively close to the mean, while a larger standard deviation would indicate more variability.

Given that the true regression function is $E(Y) = 20 + 4X$, the expected slope (b_1) is 4. The fact that the mean of the estimates (b_1) is close to this value suggests the simulation accurately reflects the real relationship between X and Y. Although there's some variation, the standard deviation shows that the estimates are fairly consistent. This matches what we'd expect theoretically—repeated sampling from a normal distribution should give estimates that hover around the true value. Overall, the results confirm that the regression model is a good fit and that the underlying assumptions hold up well.

d-) What proportion of the 200 confidence intervals for $\mathbf{E}(Y_h)$ when $X_h = 10$ include $\mathbf{E}(Y_h)$? Is

this result consistent with theoretical expectations? (5 points)

First, we calculate $E(Y_h)$ when $X_h = 10$ by substituting the value into the true regression function: $E(Y) = 20 + 4 * 10 = 60$.

Next, we check how many of the confidence intervals from our previous simulation include this value.

```
E_Yh <- 60 # True expected value when X_h = 10

# Grab the output dataframe that stores the lower and upper confidence intervals as the 2nd and 3rd col
lwr <- output[, 2]
upr <- output[, 3]

# Count how many intervals include 60
int_incl_EYh <- sum(lwr <= E_Yh & upr >= E_Yh)

# Calculate the proportion
prop <- int_incl_EYh / 200

cat("Proportion of intervals that include E(Y_h):", prop, "\n")
```

```
## Proportion of intervals that include E(Y_h): 0.945
```

The result is approximately 0.945, which is nearly consistent with the theoretical expectation of 95%. This aligns with the confidence level of the intervals, indicating that our simulation effectively reflects the expected coverage probability of confidence intervals.