STA 471 Homework #3

Richard McCormick

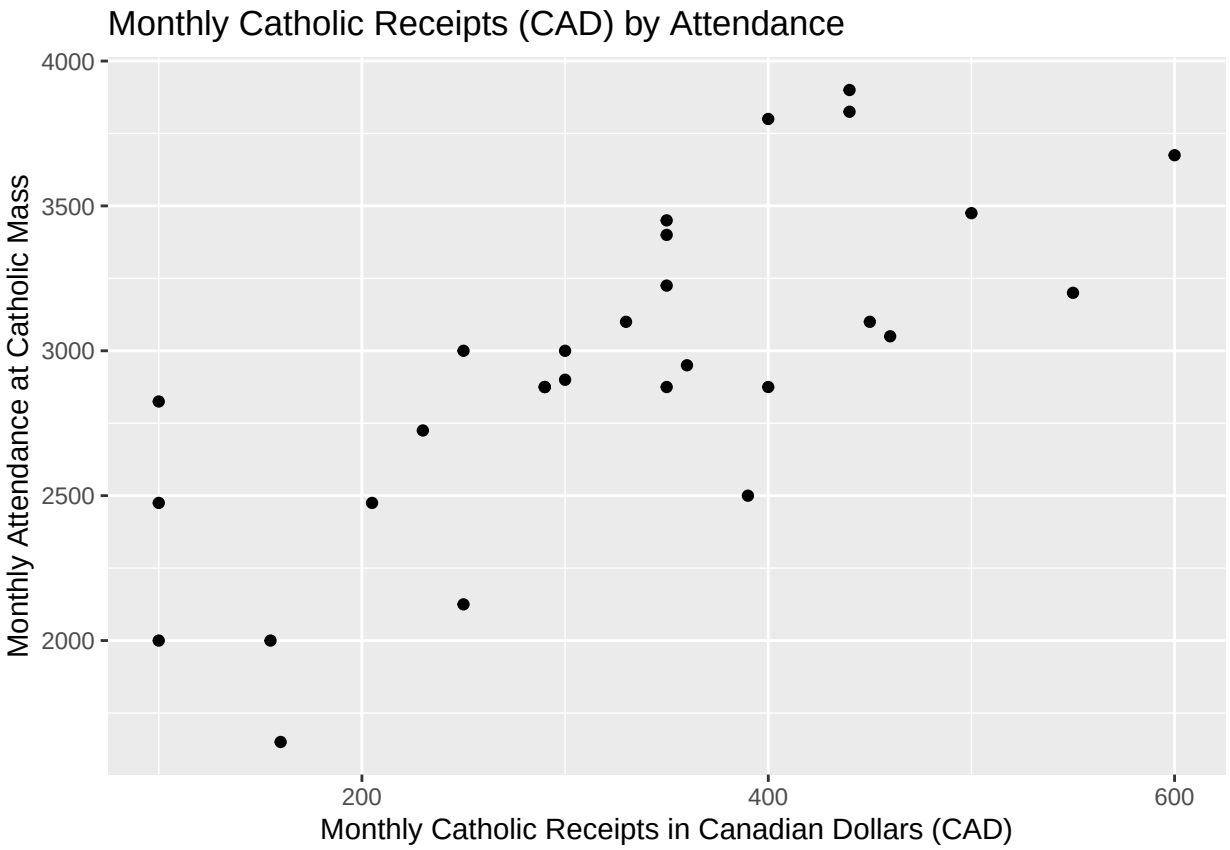
2023-09-29

Refer to the data given in Exercise HH on page 111 to answer the following questions.

```
X <- c( 100, 100, 100, 155, 160, 205, 230, 250, 250, 290, 290, 300, 300, 330,
        350, 350, 350, 350, 360, 390, 400, 400, 440, 440, 450, 460, 500, 550,
        600 )

Y <- c( 2000, 2475, 2825, 2000, 1650, 2475, 2725, 2125, 3000, 2875, 2875, 2900,
        3000, 3100, 2875, 3225, 3400, 3450, 2950, 2500, 2875, 3800, 3825, 3900,
        3100, 3050, 3475, 3200, 3675 )
```

a. Is it adequate to fit the data by a straight line? Use the scatter plot of Y versus X, the Pearson's sample correlation coefficient r_{XY}, and a lack-of-fit test to support your answer.

```
scatter_plot <- ggplot() +
  geom_point( aes( x=X, y=Y ) ) +
  labs( x="Monthly Catholic Receipts in Canadian Dollars (CAD)",
        y="Monthly Attendance at Catholic Mass",
        title="Monthly Catholic Receipts (CAD) by Attendance" )

scatter_plot
```



From the scatter plot, there appears to be a positive linear relationship between X and Y.

```
pearson_coef <- cor( X, Y )
print( paste( "Pearson's Correlation Coefficient: ", pearson_coef ) )
```

```
## [1] "Pearson's Correlation Coefficient:  0.736505444226309"
```

Analyzing Pearson's Correlation Coefficient for this data shows us that there is a moderate, positive linear relationship between X and Y.

```
lack_of_fit <- anovaPE( lm( Y~X ) )
lack_of_fit
```

```
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## X             1 4752525 4752525   34.83 0.0001507 ***
## Lack of Fit  17 2644375  155551    1.14 0.4295404
## Pure Error   10 1364479  136448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lack-of-Fit Test:

i. H_0: $F_{obs} \leq F^1_{test}$; H_A: $F_{obs} > F^1_{test}$

ii. H_0: There is no lack of fit ; H_A: There is a lack of fit.

$F^1_{test} = 4.45$ (p-value $= 0.05$) ; $F_{obs} = 1.14$

iii. Conclusion: $F_{obs} < F^1_{test}$. We can conclude that there is no lack of fit for this model and accept the null hypothesis.

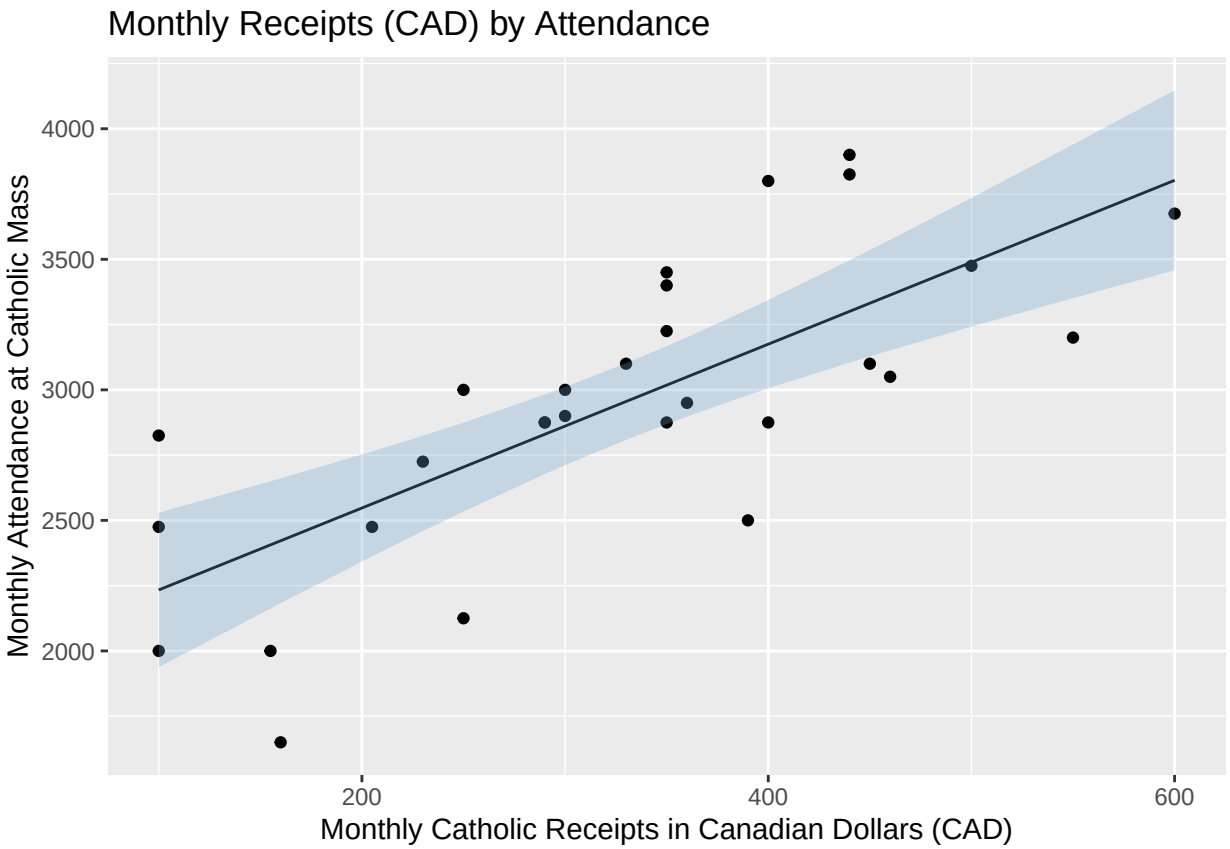
According to the scatter plot, Pearson Correlation Coefficient, and Lack-of-Fit Test for the given data, it appears there is sufficient data to conclude there is a moderate, positive correlation between X and Y. It is thus appropriate to use attempt to use a straight line to fit the data.

b. Fit the simple linear regression model for Y on X. Estimate all parameters in the model and interpret the estimates of the parameters in the context of the problem. How much of the variation in Y is explained by the fitted line? Draw the fitted line on the scatter plot.

```
SLR <- lm( Y ~ X )

summary( SLR )
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -771.8 -233.6   38.9  241.4  625.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1919.8120   194.3671   9.877 1.85e-10 ***
## X              3.1376     0.5546   5.658 5.24e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 385.3 on 27 degrees of freedom
## Multiple R-squared:  0.5424, Adjusted R-squared:  0.5255
## F-statistic: 32.01 on 1 and 27 DF,  p-value: 5.239e-06
```

```
dataframe <- data.frame( X, Y )

dataframe <- dataframe %>%
  dplyr::select( -matches('fit'), -matches('lwr'), -matches('upr') ) %>%
  cbind( predict(SLR, newdata=., interval='confidence') )

scatter_plot <- ggplot( data=dataframe, aes( x=X, y=Y ) ) +
  geom_point() +
  labs( x="Monthly Catholic Receipts in Canadian Dollars (CAD)",
        y="Monthly Attendance at Catholic Mass",
        title="Monthly Receipts (CAD) by Attendance" ) +
  geom_line( aes( y=fit ) ) +
  geom_ribbon( aes( ymin=lwr, ymax=upr ), fill='skyblue3', alpha=0.3 )


scatter_plot
```




```
explained_variance = ( pearson_coef )^2 * 100

print( paste( "The percentage of variance explained by the model is",
              explained_variance, "%." ) )
```

```
## [1] "The percentage of variance explained by the model is 54.2440269374992 %."
```

```
coef( SLR )
```

```
## (Intercept)           X
## 1919.811998    3.137614
```

The intercept parameter is equal to b_0 and the X parameter is equal to b_1. We can see that there is a moderate, positive slope to the model, along with a large Y intercept. This model assumes that each new person at mass donates 3 dollars, and that the Catholic Receipts will be 1,900 dollars even if no one attends mass.

c. Construct the ANOVA table with lack of fit and pure error included. Carry out the test for linear relationship. List all the assumptions you made in order to conduct the test.

```
print( anovaPE( SLR ) )
```

```
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## X             1 4752525 4752525   34.83 0.0001507 ***
## Lack of Fit  17 2644375  155551    1.14 0.4295404
## Pure Error   10 1364479  136448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test for Linear Relationship:

i. $H_0 : y = \beta_0 + \beta_1 X + \epsilon$; $H_A : y \neq \beta_0 + \beta_1 X + \epsilon$

ii. $F_{obs} = \frac{MS(lof)}{MS(pe)} = \frac{155{,}551}{136{,}448} = 1.14$
$F_{m-2,n-m}(1-\alpha) = F_{17,10}(1-0.05) = 2.45$

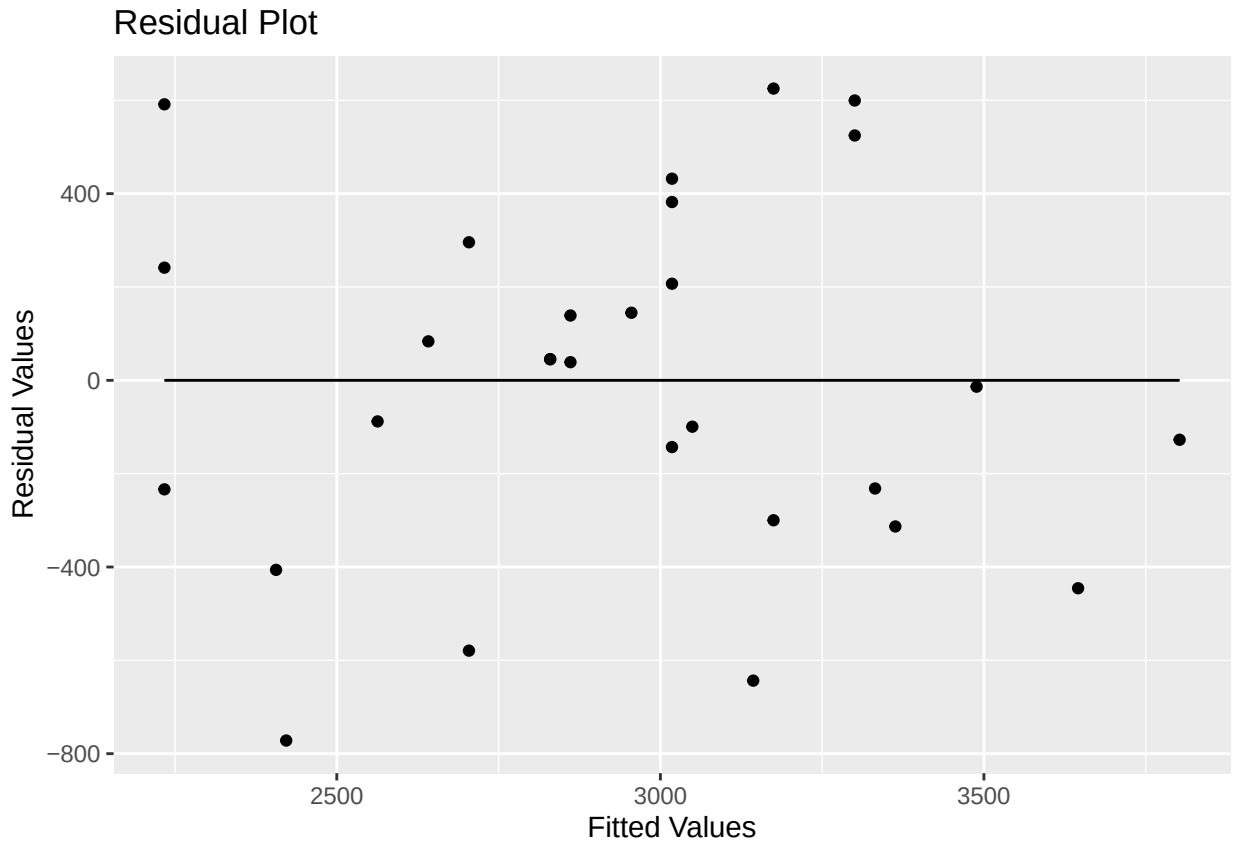
iii. Conclusion: $F_{obs} < F_{m-2,n-m}$, so we can accept the null hypothesis and conclude that there must be a linear relationship in the model.

The assumptions for our model are as follows:

1. $\mathrm{E}(\ \epsilon_i\ ) = 0$
2. $\mathrm{Var}(\epsilon_i) = \sigma^2$
3. $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are indepdendent.
4. ϵ_i is normally distributed.

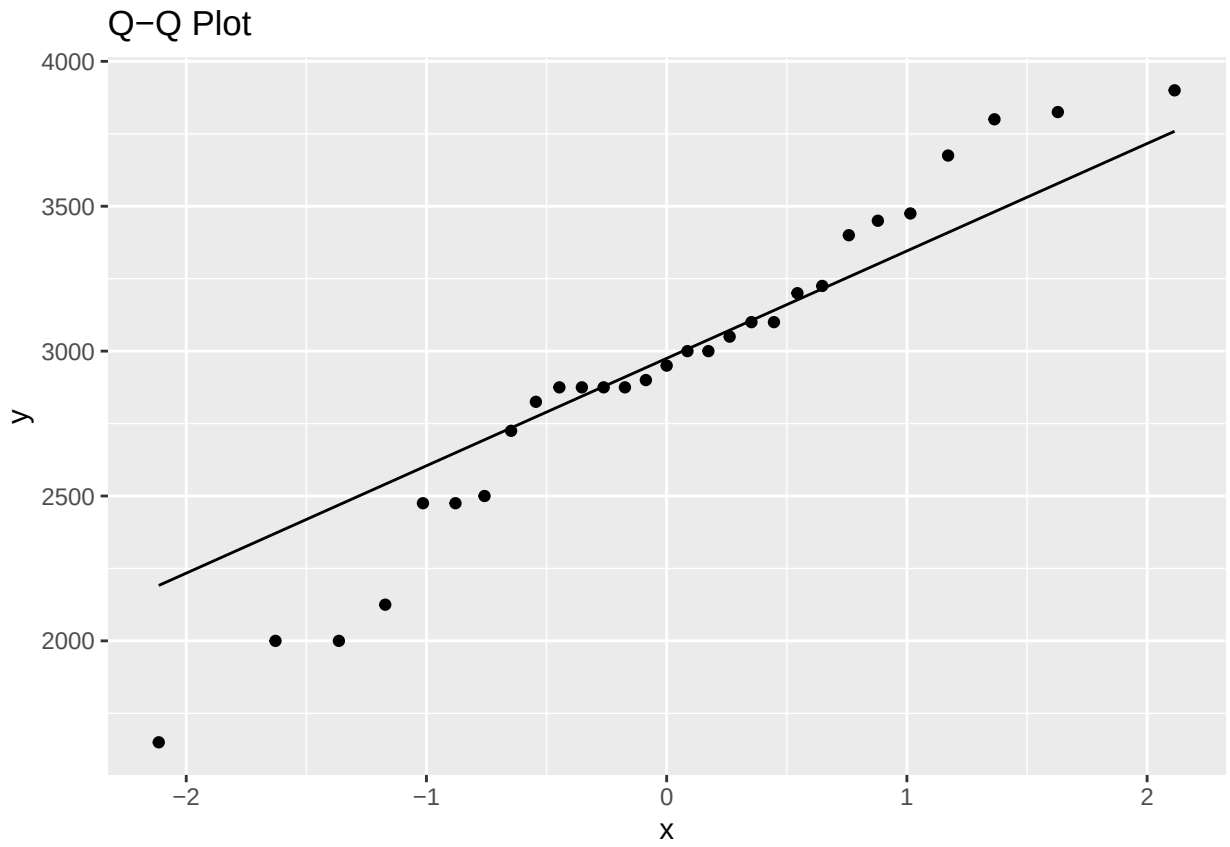
d. Check the assumptions by a residual plot and a $Q-Q$ plot of the residuals. In addition, conduct Shapiro and Wilk test for normality based on residuals.

```
dataframe['resid'] <- resid( SLR )

residual_plot <- ggplot( data=dataframe, aes( x=fit, y=resid ) ) +
  geom_point() +
  geom_line( aes( y=0 ) ) +
  labs( title="Residual Plot", x="Fitted Values",
        y="Residual Values" )

residual_plot
```



The residual plot seems to show a random distribution of residuals across the data, with $e = 0$. This indicates that the linear relationship established by the model is correct. Residuals are evenly distributed.

```
qq_plot <- ggplot( dataframe, aes( sample = Y ) ) +
  stat_qq() +
  stat_qq_line() +
  labs( title='Q-Q Plot' )

qq_plot
```



The $Q-Qplot$ seems to indicate that our data is roughly normal. There may be some more extreme values at either end of the spectrum, but overall it appears as though our data is normally distributed. Variance is static.

```
shapiro_test_result <- shapiro.test( dataframe$X )
print( shapiro_test_result )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataframe$X
## W = 0.97472, p-value = 0.6928
```

Shapiro-Wilk Normality Test:

i. $H_0 : p_{obs} > \alpha$; $H_A : p_{obs} \leq \alpha$

H_0 : The data is normally distributed. H_A : The data is not normally distributed.

ii. $p_{obs} = 0.6928$; $\alpha = 0.05$

iii. Conclusion: We can observe that $p_{obs} > \alpha$. Thus, we fail to reject our null hypothesis, and conclude that the data is normally distributed.

It appears that all assumptions made in the creation of the model correctly hold.

e. Predict the monthly Catholic mass attendance when $X = 50$ and provide a 99% prediction interval. What is the 99% confidence interval for the true mean value of Y at $X = 50$?

Calculating the predicted value for $X = 50$, using the fitted Simple Linear Regression Model.

```
prediction <- predict( SLR, newdata=data.frame( X=50 ), level=0.99 )
print( paste( "Predicted value for X = 50 is: ", prediction ) )
```

```
## [1] "Predicted value for X = 50 is:  2076.69269670746"
```

```
pred_interval = predict( SLR,
                         interval="prediction",
                         level=0.99,
                         newdata=data.frame( X=50 ) )

pred_interval
```

Calculating the 99% prediction interval at $X = 50$.

```
##        fit      lwr      upr
## 1 2076.693 911.0242 3242.361
```

```
conf_interval = predict( SLR,
                         interval="confidence",
                         level=0.99,
                         newdata=data.frame( X=50 ) )

conf_interval
```

Calculating the 99% confidence interval at $X = 50$.

```
##        fit      lwr      upr
## 1 2076.693 1608.741 2544.645
```