Assignment 8 Analysis Key

2023-06-07

Logistics:

Please read Chapters 5.4 and 5.5 in BC&P while doing the "Simple Linear Regression Tutorial Key.r" as you read the chapters. Then do this assignment.

Introduction - The Data.

In this assignment we are going to analyze the relationship between where people live, based on latitude, and the mortality rate due to malignant melanoma (a serious cancer with a high death rate). The following is from the CDC in July, 2019:

"Based on data from 2012 to 2016, about 77,698 new cases of melanoma occurred in the United States each year, including 45,854 among men and 31,845 among women. The overall incidence rate of melanoma was 21.8 per 100,000. The highest incidence rate was among non-Hispanic white males (34.9 per 100,000), and the lowest rate was among black females (0.9 per 100,000)."

Q1. Before we begin, state what you think may come from our analysis. Do you think that there is some relationship between melanoma mortality rates and latitude? What to you think that relationship might be?

I think as latitude approaches the equator, melanoma rates will increase.

Description of the Dataset

The data set consists of mortality data due to malignant
melanoma of the skin (a skin cancer) of white males during the period 1950 - 1969 for each state in the United States as well as the District of Columbia. No data was available for Alaska, Hawaii, Puerto Rico, or the Virgin Islands. It would have been interesting to have those four locations included for our analysis.

The data set has four columns:

1. "state":

The by USPS its abbreviation;

2. "mortality":

the annual number of deaths due to melanoma per 10,000,000;

3. "lat":

the latitude in degrees (estimated at the center of the state); In the northern hemisphere, as latitude increases as one moves further north (closer to the North Pole)

4. "population":

the size of the population in millions as of 1965.

Source of the data:

U.S. Department of Health, Education, and Welfare (1974).

We are going to investigate the relationship between melanoma mortality and latitude for these states. An example of a similar analysis is:

```
 I. K. Crombie. 1979. Variation of Melanoma Incidence with
Latitude in North America and Europe. Br. J. Cnacer 40: 774 - 781.

 Reference link:

 https://www.cdc.gov/cancer/skin/statistics/index.htm
```

This time the five steps wil not be explicitly identified. See if you can find the five step process being used in what we do below.

Q2. In the space below, please clear the R memory; load the libraries "ggplot2", "dplyr", "ggfortify", "gridExtra", and "readr", then load the dataset, "melanoma.csv" and assign it to anobject named "Melanoma".

```
# Clear the memory
rm(list = ls())



# Load libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggfortify)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(readr)

# Read or upload dataset
melanoma <- read.csv('melanoma.csv')
```

Now that we have the data in R, let's take a quick look at it. We will want to find out such things as:

1. What does the dataset look like?
2. what type of data is it?
3. What are the column names in the dataset?
4. How many rows and columns are in the dataset?

Feel free to look at anything else you would like to know about the dataset.

```
View(melanoma)
str(melanoma)
```

```
## 'data.frame':    49 obs. of  4 variables:
##  $ state     : chr  "AL" "AZ" "AR" "CA" ...
##  $ mortality : int  219 160 170 182 149 159 200 177 197 214 ...
##  $ lat       : num  33 34.5 35 37.5 39 41.8 39 39 28 33 ...
##  $ population: num  3.46 1.61 1.96 18.6 1.97 2.83 0.5 0.76 5.8 4.36 ...
```

```
colnames(melanoma)
```

```
## [1] "state"      "mortality"  "lat"        "population"
```

```
nrow(melanoma)
```

```
## [1] 49
```

```
ncol(melanoma)
```

```
## [1] 4
```

```
# Select alpha = 0.05
Alpha_value  = 0.05
```

Part 1: Data Exploration

As always, the first step in the analysis will be to explore the data looking for the structure of the dataset (what is in it), outliers, and if it appears to meet the assumptions of the hypothesis test we will apply in analyzing it. Plotting the data is critical in the exploration as are summary statistics.

Q1. In the space below, create an R script that produces at least a summary of the number of locations for which data was collected and the mean, median, and IRQ of the mortality rates. Assign these to objects named "total_Locations", "meanMortality", "medianMortality", and "irqMoratlity", respectively. You might assign the entire script to an object named "sumData".

Feel free to calculate any other sample statistics that you would like. You may wish to review some prior assignments for ideas for what to request and for how to write the script.

Print a title for your table of summary statistics. You might use "Summary Statistics for Melanoma Mortality Data".

```
total_Locations = nrow(melanoma)
meanMortality = mean(melanoma$mortality)
medianMortality = median(melanoma$mortality)
irqMortality = IQR(melanoma$mortality)
# Standard method we have used before:


library(glue)

glue('# locations = {total_Locations}, Mean mort = {meanMortality}, Median mort = {medianMortality}, IQ
```

```
## # locations = 49, Mean mort = 152.877551020408, Median mort = 147, IQR = 50
```

Q2. What is the total number of locations, and the mean, median, IRQ of the mortality rates. What do you notice about the mean and the median?

locations = 49, Mean mort = 152.877551020408, Median mort = 147, IQR = 50

The mean is higher than the median by about 6

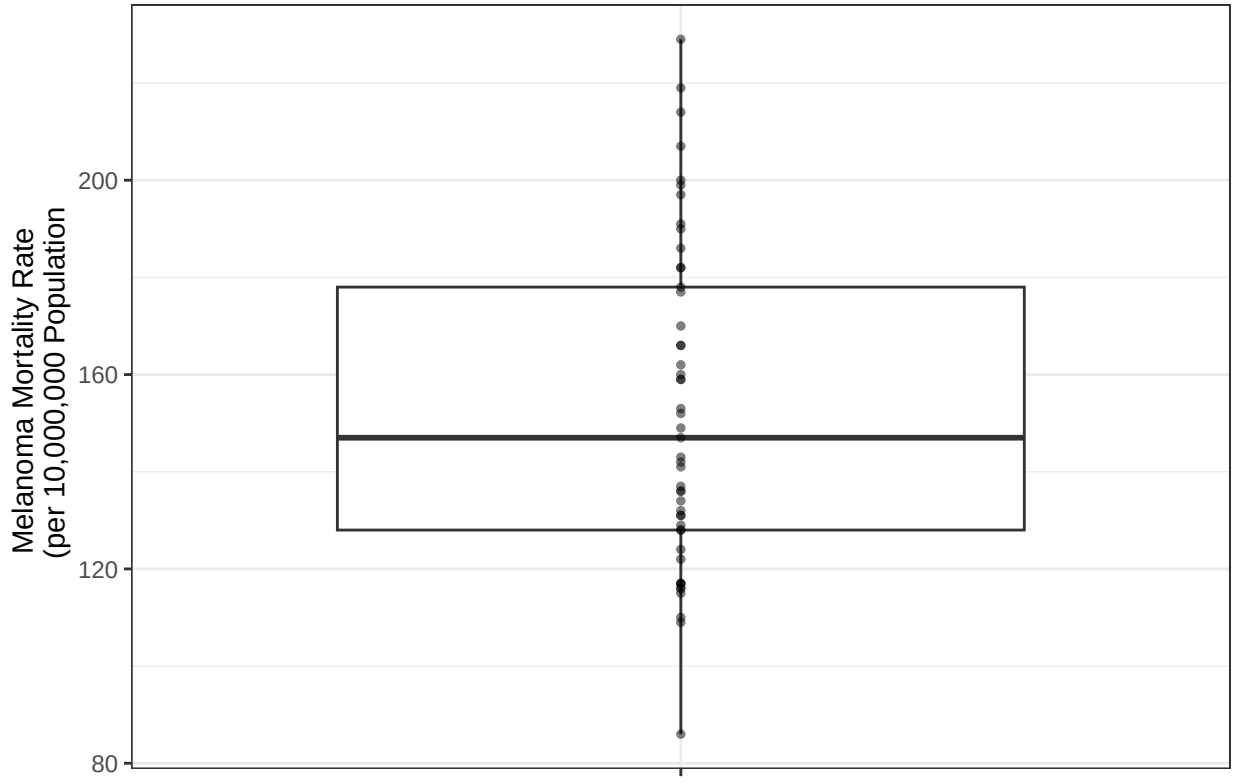
Q3. What is the mean latitude? Place your script and the result in the space below.For ease of later use, assign both the mean latitude and the entire script to an object named "meanLatitude".

```
meanLatitude <- mean(melanoma$lat)
meanLatitude
```

```
## [1] 39.53265
```

###. Create a boxplot and check the data for outliers. Incorporate an identification of any outliers in the boxplot.

```
is_outlier  = function(x) {
  return(x < quantile(x,0.25) - 1.5 * IQR(x) |
           x > quantile(x, 0.75) + 1.5 * IQR(x))
}
melanoma %>%
  mutate(outlier = ifelse(is_outlier(mortality), mortality, as.numeric(NA))) %>%
  ggplot(., aes(x = "", y = mortality)) +
  geom_boxplot() +
  geom_text(aes(label = outlier), na.rm = TRUE, hjust = -0.3)+
  geom_point(size = 1, colour = 'black', alpha = 0.5) +
  xlab("") +
  ylab("Melanoma Mortality Rate \n (per 10,000,000 Population") +
  theme_bw()
```

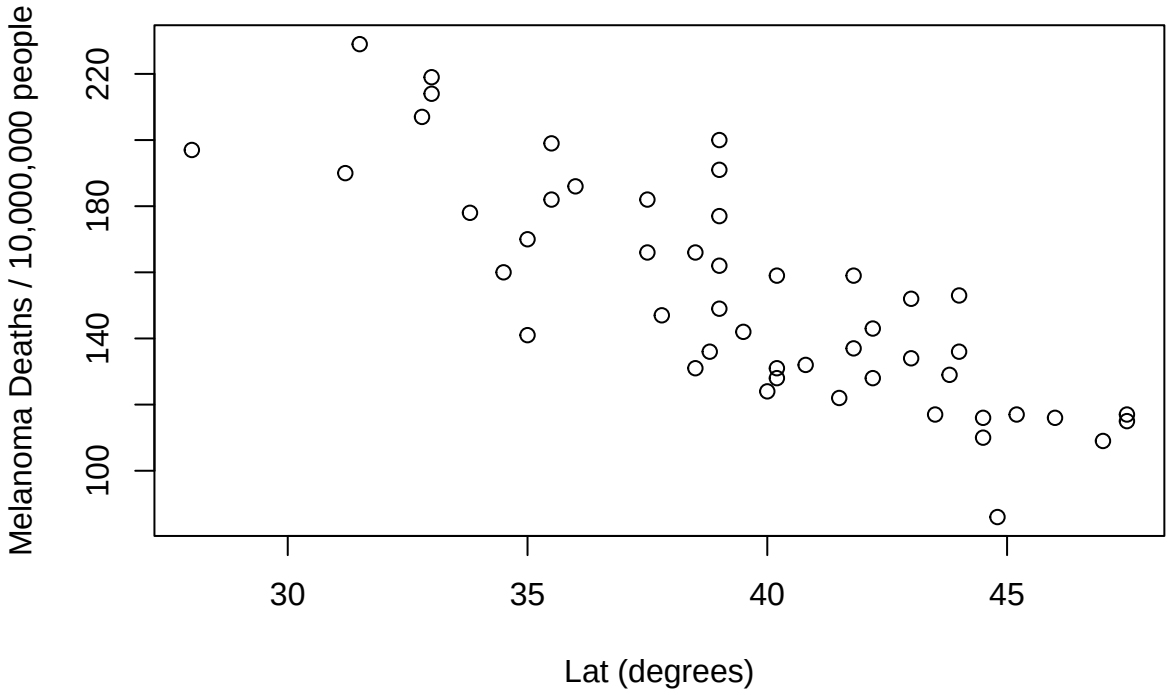


Q4. Please summarize what you conclude about the data. Think about things like symmetry and outliers.

First there appears to be a skew to the right, the mean is less than the median and we can see a demonstration of the tail from the boplot. There doesn't appear to be any outliers; the endpoints are contained within our acceptable quantiles.

Plot the mortality data versus latitude in a scatterplot. Use informative axis labels and remember to include the appropriate units on each axis.

```
x = melanoma$lat
y = melanoma$mortality
xrange <- c(min(x), max(x))
yrange <- c(min(y), max(y))
plot(x,y, xlab = 'Lat (degrees)', ylab = 'Melanoma Deaths / 10,000,000 people')
```




```
(unique(melanoma$mortality))
```

```
##  [1] 219 160 170 182 149 159 200 177 197 214 116 124 128 166 147 190 117 162 143
## [20] 207 131 109 122 191 129 141 152 199 115 136 132 137 178  86 186 229 142 153
## [39] 110 134
```

```
# nrow(unique(melanoma$mortality))
```

Q5. From the scatterplot, what observations do you make about the relationship between melanoma mortality rates and latitude?

Melanoma deaths decrease as latitude increases.

Q6. Is the relationship one-to-one (recall from algebra what that means)?

No, the relationship is not one to one. For one thing, there are more values of mortality than latitude, so we will have to have at least one latitude mapping to multiple mortality rates.

Q7. Is there much scatter?

Some, not much. There is more scatter in the lower latitudes than higher latitudes, but the relationship looks pretty linear.

Linear Models

We want to decide if the scatter seen in the data could be due to some relationship between mortality rate and latitude or chance. We may also be interested in estimating the melanoma rate for various latitudes. In addition, we may also like to estimate the variability that is seen in the scatter plot. How might we do this? That is where linear models (linear in the parameters estimated) play a pivotal role.In the case of our data, the overall relationship is roughly linear (a straight line), so we will apply simple linear regression.

We will hypothesize that the melanoma mortality rate is a function of latitude, using the variables from the Melanoma data frame. That is, we are going model this relationship with a simple linear regression model and estimate its parameters.

Q8. Create a script that does the estimation, save the results as an object, named "model_melanoma_mortality" for later use, and print the results.

```
mort = melanoma$mortality
lat = melanoma$lat
model_melanoma_mortality <- lm(mort ~ lat, data = melanoma)

# Print the estimates of the regression coefficients,i.e., the estimated intercept and slope.

model_melanoma_mortality
```

```
##
## Call:
## lm(formula = mort ~ lat, data = melanoma)
##
## Coefficients:
## (Intercept)          lat
##     389.189       -5.978
```

Q9. Write the estimated linear model relationship between melanoma mortality rate and and latitude.

The model is:

$$\text{MMR} = -5.98 * \text{Latitude} + 389.19$$

Assumptions of a Linear Regression Model

1. The predicted (response) variable come at random from the sampled population and are independent of one another;

2. The explanatory variables are constants known without error;

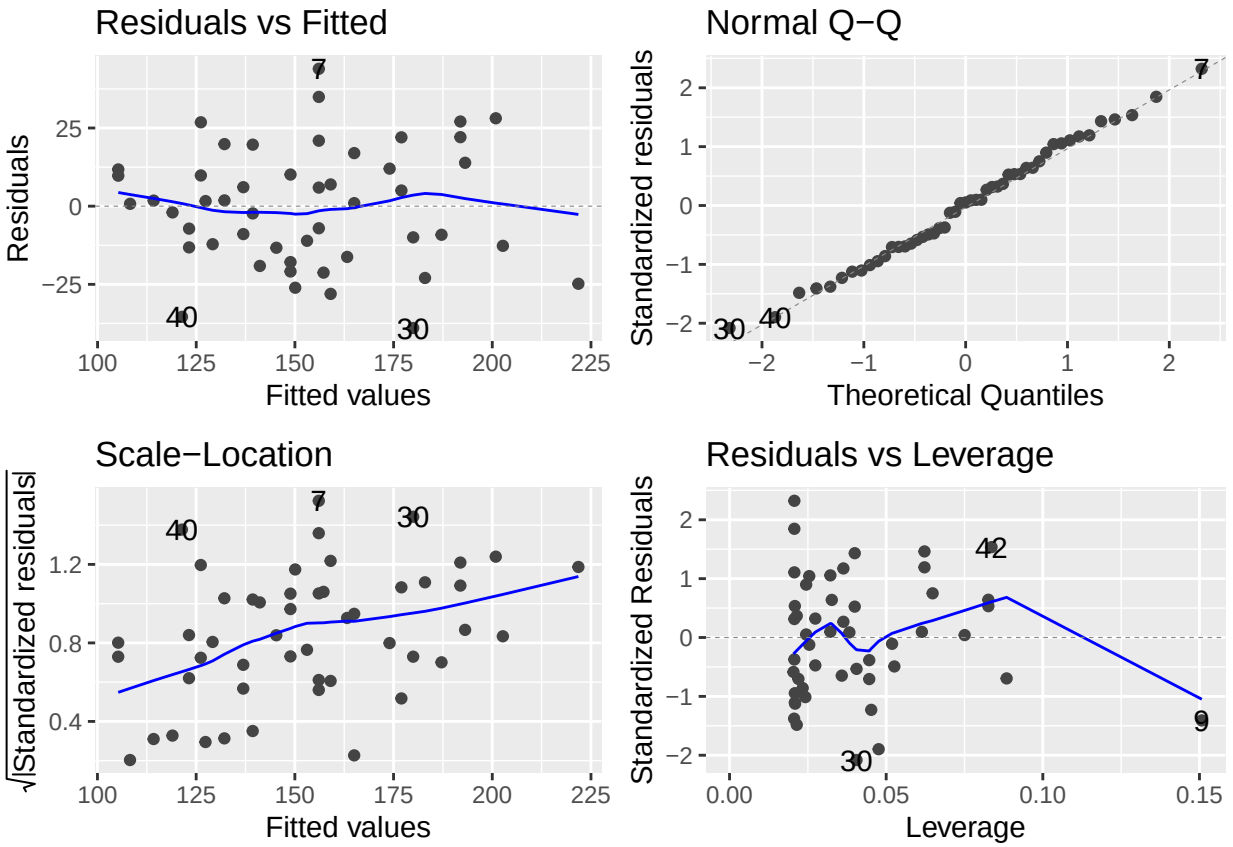
3. The errors are independent and identically distributed as a normal distributed with a mean of 0 and constant variance. This is the same as saying that the predicted variable has a normal distribution; and

4. The variance of the errors is constant (equal to each other; homogeneity).

Check the assumptions of the linear regression model. Do this by visually examining them using the "autoplot()" function in the package "ggfortify".

Q1. Run the script and review the plots, then state wht you conclude form this plot and the rest of our exploration of the data that we did above.

```
autoplot(model_melanoma_mortality)
```



Our data and residuals look good. There are no patterns in the top left plot, our residual QQ plot looks good, there is a sliht pattern in the bottom-right plot, but BC&P says that's okay if variance increases with mean, which is possible here.

Let's check another feature of a simple linear regression equation. In the space below, enter the mean of the latitudes into the linear regression model with the estimated values of the slope and intercept that were

determined above. See if you can do this easily by using some of the objects that we created above.

Q2. What value do you get for the predicted value? Please show your script and cut and paste the results in the space below.

```
mm_int = model_melanoma_mortality$coefficients[[1]]
mm_var = model_melanoma_mortality$coefficients[[2]]

pred_mort = meanLatitude * mm_var + mm_int
glue('Predicted mortality using the mean latitude: {pred_mort}')
```

```
## Predicted mortality using the mean latitude: 152.877551020408
```

Predicted mortality using the mean latitude: 152.877551020408

Q3. What value does this predicted value approximately equal (Hint: You have already calculated this value)? It is only approximately equal because we have used estimates of the slope and intercept that do not include many decimal places.

the mean mortality across all locations.

```
# Recall the mean latitude "meanlatitude" define an object for the mean mortality rate:

meanLatitude
```

```
## [1] 39.53265
```

```
# Print the result:

print(meanLatitude)
```

```
## [1] 39.53265
```

```
# Calculate the predicted mortality rate for the mean latitude:
pred_mort = meanLatitude * mm_var + mm_int



# Print the result:

print(pred_mort)
```

```
## [1] 152.8776
```

```
# Subtract this value from the mean mortality rate estimated eralier:

print(meanMortality - pred_mort)
```

```
## [1] 0
```

You get a result that is close to the mean mortality rate calculated from the data. We know that the regression line goes through the point (mean of the response variable, mean of the predicted value) = (39.5,152.878). How about that!!

Part 2: Analysis

We want to decide if the scatter seen in the data could be due to some relationship between mortality rate and latitude or chance. In a simple linear model the relationship is determined by two parameters, the slope of the equation and the vertical axis intercept. We know that we can, if the assumptions of linear regression analysis are met, conduct statistical tests of two hypotheses; on each parameter.

We will only conduct an hypothesis test for the slope.

State the null and alternative hypotheses

Q1. State the null and alternative hypotheses for that test.

H_o: There is no statistically significant relationship between latitude and melanoma mortality and the slope is equal to 0 $\beta = 0$

H_a: There is a statistically significant relationship between latitude and melanoma mortality and the slope is not equal to zero $\beta \neq 0$.

OLS and ANOVA

Recall that linear regression uses ordinary least squares to minimize the sum of squares (SS) to estimate the coefficients (parameters) in the linear model. If we rearrange the SS, apply some algebra, and recognize that the cross-product SS term equals zero, we arrive at a rearranged equation that states:

$$\text{Total SS} = \text{Regression SS} + \text{Error SS}$$

We use the ANOVA to analyze the sources of the Total SS. Hence, regression and ANOVA are intimately linked. ANOVA will provide a frame- work for testing hypotheses in linear models.

The ANOVA table is a classic assessment of the null hypothesis, that there is no relationship between latitude content and melanoma mortality rate. It presents the F-value, degrees of freedom, and the p-value associated with the explanatory variables in the model (in this case, the model has one explanatory variable, i.e., latitude).

Q2. Create the script that produces the ANOVA table using the object we created from estimating the coefficients ofthe linear regression model.

```
anova(model_melanoma_mortality)
```

```
## Analysis of Variance Table
##
## Response: mort
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lat        1  36464   36464  99.797 3.309e-13 ***
## Residuals 47  17173     365
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q3. Calculate and display the summary(table) that provides the regression analysis.

```
summary(model_melanoma_mortality)
```

```
##
## Call:
## lm(formula = mort ~ lat, data = melanoma)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.972 -13.185   0.972  12.006  43.938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
## lat          -5.9776     0.5984   -9.99 3.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 47 degrees of freedom
## Multiple R-squared:  0.6798, Adjusted R-squared:  0.673
## F-statistic:  99.8 on 1 and 47 DF,  p-value: 3.309e-13
```

Q4. How much of the variability in the data is explained my the relationship estimated by the liner model? Clearly indicate what you looked at to draw this conclusion.

68% This is from the R^2 value output by the linear model.

A Digression.

Let's stop for a moment and consider something mentioned in the class notes. It was stated that the square of the t-distribution is exactly equal to the F-distribution; with the appropriate degrees of freedom taken into account.

In the ANOVA table we see the expected results of an F-test for an F-distributed random variable with (1, n-2) = (1,47) degrees of freedom. The values of that test statistic = 99.8.

In the summary model table we see the results of a t-test for a t-distributed random variable with

$$n - 2 = 47$$

degrees of freedom. The value of that test statistic is -9.99.

Notice that

$$-9.99^2 = 99.9$$

. This means that the F-test in the ANOVA is EXACTLY equal to the square of the t-test of the null hypothesis that the slope parameter = 0. Therefore, the t-test and the F-test will provide exactly the same result in terms of the p-value and the decision when testing the null hypothesis that the population slope parameter = 0. The value of the t-test is in one-tailed tests of hypotheses.

Now we will complete the test of the hypotheses:

Null hypothesis: the population slope parameter equals 0 Alternative hypothesis: the population slope parameter does not equal 0.

Q1. Calculate the number of observations, n, directly from the data.

```
Num_obs  = length(melanoma$lat)
Num_obs
```

```
## [1] 49
```

Calculate the critical value for the F-test.

Q2. Run the scripts below, then in the space provided please cut and paste the alpha value, the numerator and denominator degrees of freedom and the valueof the calculated critical value.

```
Alpha_value
```

```
## [1] 0.05
```

```
Num_DOF  = 1
Num_DOF
```

```
## [1] 1
```

```
Den_DOF  = Num_obs-2
Den_DOF
```

```
## [1] 47
```

```
Critical_F_Value = qf(1 - Alpha_value, Num_DOF, Den_DOF)
print(Critical_F_Value)
```

```
## [1] 4.0471
```

[1] 0.05 [1] 1 [1] 47 [1] 4.0471

Q3. Report the F test statistic and compare it to critical value. State the decision rule and your conclusion.

F-statistic: 99.8 on 1 and 47 DF, p-value: 3.309e-13

This is larger than our critical value and the p value is smaller than alpha.

Our decision rule is to reject the null hypothesis when the test statistic is larger than the critical value.

The decision is to reject the null hypothesis that the slope parameter equals zero.

Calculate the critical value for the t-test.

Q4. Run the scripts below, then in the space provided please cut and paste the alpha value, the degrees of freedom and the value of the calculated critical value.

```
Alpha_value
```

```
## [1] 0.05
```

```
DOF  = Num_obs-2
DOF
```

```
## [1] 47
```

```
print(Criticial_t_Value <- qt(1-(Alpha_value/2), DOF))
```

```
## [1] 2.011741
```

[1] 0.05 [1] 47 [1] 2.011741

Q5. Report the t test statistic and compare it to critical value. State the decision rule and your conclusion.

t statistic: -9.99; p-value: 3.31e-13

The t statistic is less than the critical value, but the p-value is less than our alpha. The decision rule says to reject the null when p is less than alpha, so the conclusion is to reject the null hypothesis that the slope is equal to zero.

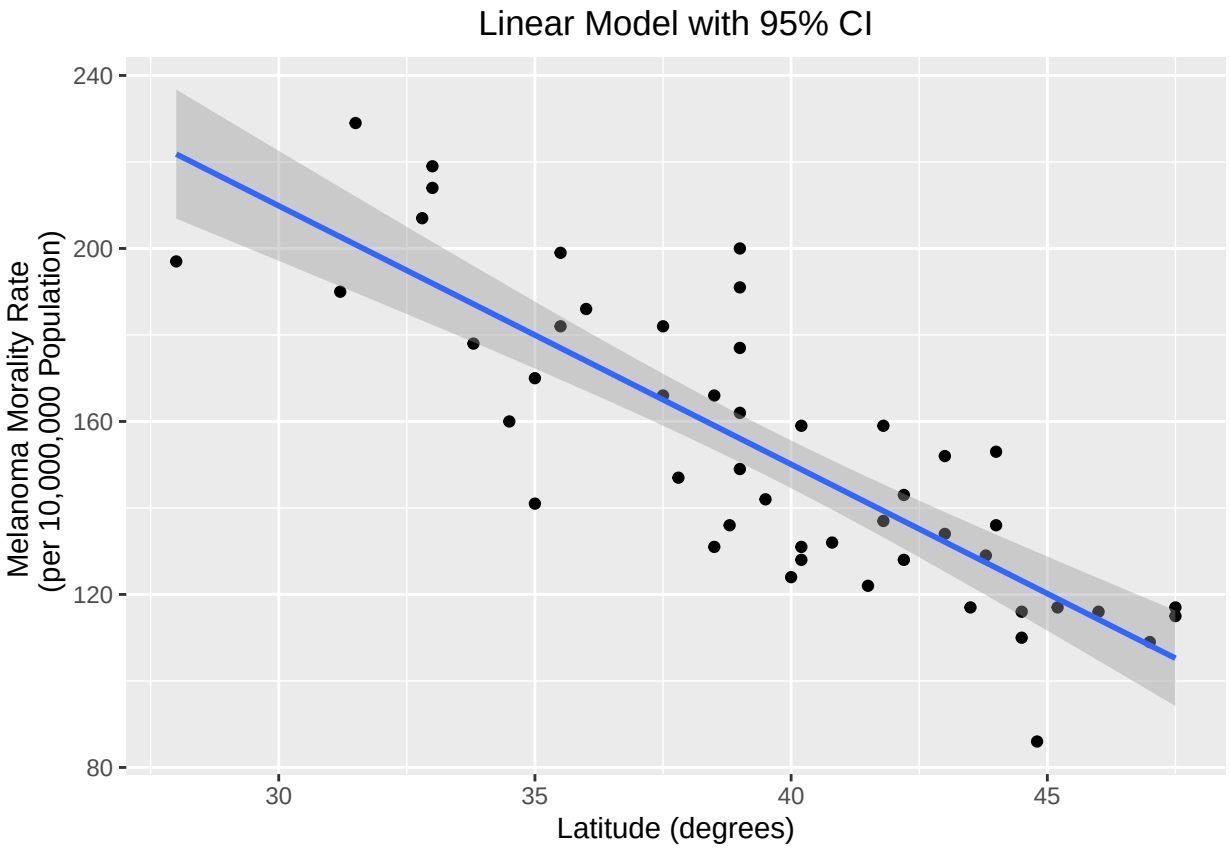
Please take the time to notice that the decisions of each test resulted in the same conclusion.Note also that the square of the critical value of the t-test equals that of the F-test, and that the square of the test statistic for the t-test equals that of the F-test.

Part 3: Predictions and Confidence Intervals

Consider the confidence intervals for both the model and predictions made by using it.Show our estimated model together with the original data and 95% confidence intervals.

```
ggplot(melanoma, aes(x = lat,
                     y = mortality)) +
  geom_point() +
  geom_smooth(method = 'lm', se=TRUE) +
  ggtitle ("Linear Model with 95% CI") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylab("Melanoma Morality Rate \n (per 10,000,000 Population)") +
  xlab("Latitude (degrees)")
```

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



Calculate 95% confidence and prediction intervals. Plot the results.

Compute fitted model output and upper (upr) and lower (lwr) prediction intervals.

```
Mortality_Prediction  = predict(model_melanoma_mortality, interval="prediction")
```

```
## Warning in predict.lm(model_melanoma_mortality, interval = "prediction"): predictions on current data
```

```r
# Combine the upper and lower prediction intervals with the original data.
combined_function  = cbind(melanoma, Mortality_Prediction)


Mortality_Prediction   # Print the results for checking
```

```
##         fit       lwr      upr
## 1  191.9274 152.29449 231.5602
## 2  182.9609 143.64646 222.2754
## 3  179.9721 140.74588 219.1983
## 4  165.0280 126.10613 203.9499
## 5  156.0616 117.21137 194.9117
## 6  139.3242 100.38352 178.2648
## 7  156.0616 117.21137 194.9117
## 8  156.0616 117.21137 194.9117
## 9  221.8156 180.56445 263.0667
## 10 191.9274 152.29449 231.5602
## 11 123.1846  83.88214 162.4870
## 12 150.0839 111.23496 188.9329
## 13 148.8884 110.03520 187.7416
## 14 136.9331  97.95575 175.9105
## 15 159.0504 120.18560 197.9152
## 16 163.2347 124.33388 202.1356
## 17 202.6871 162.56807 242.8062
## 18 119.0002  79.56080 158.4396
## 19 156.0616 117.21137 194.9117
## 20 136.9331  97.95575 175.9105
## 21 129.1622  90.02483 168.2996
## 22 114.2181  74.60075 153.8355
## 23 193.1229 153.44155 232.8042
## 24 159.0504 120.18560 197.9152
## 25 108.2405  68.36909 148.1119
## 26 141.1175 102.20045 180.0345
## 27 156.0616 117.21137 194.9117
## 28 127.3689  88.18583 166.5520
## 29 148.8884 110.03520 187.7416
## 30 179.9721 140.74588 219.1983
## 31 132.1510  93.08252 171.2195
## 32 176.9833 137.83624 216.1303
## 33 105.2517  65.24028 145.2630
## 34 148.8884 110.03520 187.7416
## 35 176.9833 137.83624 216.1303
## 36 126.1734  86.95802 165.3887
## 37 145.3018 106.42698 184.1766
## 38 139.3242 100.38352 178.2648
## 39 187.1453 147.69217 226.5984
## 40 121.3913  82.03229 160.7503
## 41 173.9945 134.91750 213.0714
## 42 200.8938 160.86354 240.9241
## 43 153.0727 114.22783 191.9176
## 44 126.1734  86.95802 165.3887
## 45 165.0280 126.10613 203.9499
## 46 105.2517  65.24028 145.2630
## 47 157.2571 118.40218 196.1120
```

```
## 48 123.1846  83.88214 162.4870
## 49 132.1510  93.08252 171.2195
```

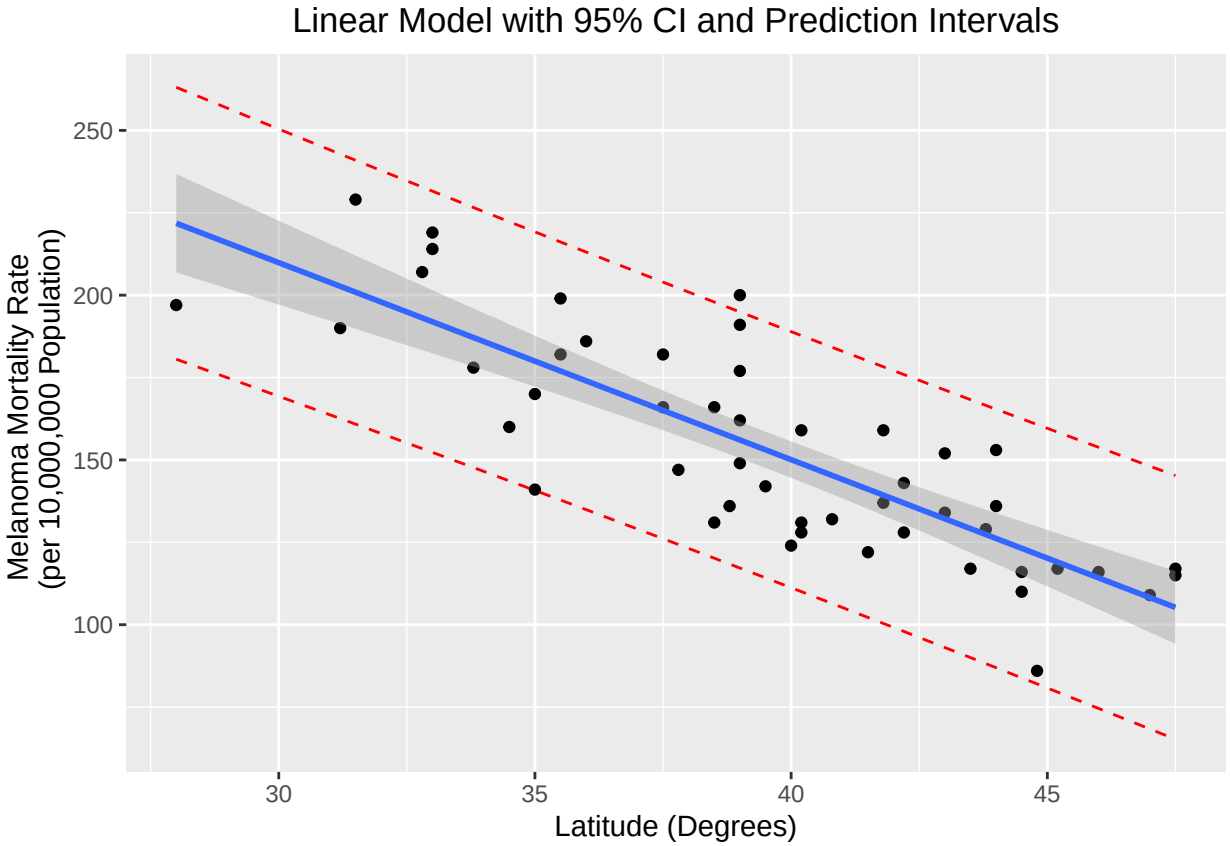
```
combined_function      # Print the results for checking
```

```
##    state mortality  lat population      fit       lwr      upr
## 1     AL       219 33.0       3.46 191.9274 152.29449 231.5602
## 2     AZ       160 34.5       1.61 182.9609 143.64646 222.2754
## 3     AR       170 35.0       1.96 179.9721 140.74588 219.1983
## 4     CA       182 37.5      18.60 165.0280 126.10613 203.9499
## 5     CO       149 39.0       1.97 156.0616 117.21137 194.9117
## 6     CT       159 41.8       2.83 139.3242 100.38352 178.2648
## 7     DE       200 39.0       0.50 156.0616 117.21137 194.9117
## 8     DC       177 39.0       0.76 156.0616 117.21137 194.9117
## 9     FL       197 28.0       5.80 221.8156 180.56445 263.0667
## 10    GA       214 33.0       4.36 191.9274 152.29449 231.5602
## 11    ID       116 44.5       0.69 123.1846  83.88214 162.4870
## 12    IL       124 40.0      10.64 150.0839 111.23496 188.9329
## 13    IN       128 40.2       4.88 148.8884 110.03520 187.7416
## 14    IA       128 42.2       2.76 136.9331  97.95575 175.9105
## 15    KS       166 38.5       2.23 159.0504 120.18560 197.9152
## 16    KY       147 37.8       3.18 163.2347 124.33388 202.1356
## 17    LA       190 31.2       3.53 202.6871 162.56807 242.8062
## 18    ME       117 45.2       0.99 119.0002  79.56080 158.4396
## 19    MD       162 39.0       3.52 156.0616 117.21137 194.9117
## 20    MA       143 42.2       5.35 136.9331  97.95575 175.9105
## 21    MI       117 43.5       8.22 129.1622  90.02483 168.2996
## 22    MN       116 46.0       3.55 114.2181  74.60075 153.8355
## 23    MS       207 32.8       2.32 193.1229 153.44155 232.8042
## 24    MO       131 38.5       4.50 159.0504 120.18560 197.9152
## 25    MT       109 47.0       0.71 108.2405  68.36909 148.1119
## 26    NE       122 41.5       1.48 141.1175 102.20045 180.0345
## 27    NV       191 39.0       0.44 156.0616 117.21137 194.9117
## 28    NH       129 43.8       0.67 127.3689  88.18583 166.5520
## 29    NJ       159 40.2       6.77 148.8884 110.03520 187.7416
## 30    NM       141 35.0       1.03 179.9721 140.74588 219.1983
## 31    NY       152 43.0      18.07 132.1510  93.08252 171.2195
## 32    NC       199 35.5       4.91 176.9833 137.83624 216.1303
## 33    ND       115 47.5       0.65 105.2517  65.24028 145.2630
## 34    OH       131 40.2      10.24 148.8884 110.03520 187.7416
## 35    OK       182 35.5       2.48 176.9833 137.83624 216.1303
## 36    OR       136 44.0       1.90 126.1734  86.95802 165.3887
## 37    PA       132 40.8      11.52 145.3018 106.42698 184.1766
## 38    RI       137 41.8       0.92 139.3242 100.38352 178.2648
## 39    SC       178 33.8       2.54 187.1453 147.69217 226.5984
## 40    SD        86 44.8       0.70 121.3913  82.03229 160.7503
## 41    TN       186 36.0       3.84 173.9945 134.91750 213.0714
## 42    TX       229 31.5      10.55 200.8938 160.86354 240.9241
## 43    UT       142 39.5       0.99 153.0727 114.22783 191.9176
## 44    VT       153 44.0       0.40 126.1734  86.95802 165.3887
## 45    VA       166 37.5       4.46 165.0280 126.10613 203.9499
## 46    WA       117 47.5       2.99 105.2517  65.24028 145.2630
## 47    WV       136 38.8       1.81 157.2571 118.40218 196.1120
## 48    WI       110 44.5       4.14 123.1846  83.88214 162.4870
## 49    WY       134 43.0       0.34 132.1510  93.08252 171.2195
```

Create a scatterplot of the data, the 95% CI for the model, and the prediction intervals. The red dashed lines are the upper and lower prediction intervals.

```
ggplot(combined_function, aes(x = lat,
                              y = mortality)) +
  geom_point() +
  geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y=upr), color = "red", linetype = "dashed")+
  geom_smooth(method = 'lm', se=TRUE) +
  ggtitle ("Linear Model with 95% CI and Prediction Intervals") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ylab("Melanoma Mortality Rate \n (per 10,000,000 Population)") +
  xlab("Latitude (Degrees)")
```

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



Q1. What do you notice about the prediction intervals? What warning would you give to someone who blindly decides to use a linear regression model to make predictions; especially for ranges far outside of the range of the original data used to fit the model parameters?

The prediction intervals are quite a bit wider than the confidence intervals. I would warn someone blindly predicting using this model to be careful, as the uncertainty represented by the prediciton intervals is much

wider than the regression line. There is plenty of uncertainty to take into account.

Q2. Please comment on the following statement:

Living in higher latitudes prevents a person from getting and dying from melanoma.

This statement is basically disinformation. There is a correlation with latitude and melanoma mortality, sure, but latitude is only a factor in what will cause someone to get melanoma. It certainly won't prevent it, but it may reduce the chance.

Part 4: What information can we obtain from the model and predicted values?

Knowing how to access to this information is useful if you wish to do more than test hypotheses with the linear model, e.g., to this information in R.

Check for what information is available for use from the linear model fitting.

```
names(model_melanoma_mortality)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
# Display the estimated coefficients in the linear model.

model_melanoma_mortality$coefficients
```

```
## (Intercept)         lat
##  389.189351   -5.977636
```

```
# Check for what information is available for use in the object "combined_function".
names(combined_function)
```

```
## [1] "state"      "mortality"  "lat"        "population" "fit"
## [6] "lwr"        "upr"
```

```
# Display the mortality rates estimated by the linear regression model.
combined_function$fit
```

```
##  [1] 191.9274 182.9609 179.9721 165.0280 156.0616 139.3242 156.0616 156.0616
##  [9] 221.8156 191.9274 123.1846 150.0839 148.8884 136.9331 159.0504 163.2347
## [17] 202.6871 119.0002 156.0616 136.9331 129.1622 114.2181 193.1229 159.0504
## [25] 108.2405 141.1175 156.0616 127.3689 148.8884 179.9721 132.1510 176.9833
## [33] 105.2517 148.8884 176.9833 126.1734 145.3018 139.3242 187.1453 121.3913
## [41] 173.9945 200.8938 153.0727 126.1734 165.0280 105.2517 157.2571 123.1846
## [49] 132.1510
```

```
# Display the residuals between the linear regression model and the data.
model_melanoma_mortality$residuals
```

```
##           1           2           3           4           5           6
##  27.0726285 -22.9609178  -9.9721000  16.9719894  -7.0615570  19.6758231
##           7           8           9          10          11          12
##  43.9384430  20.9384430 -24.8155502  22.0726285  -7.1845604 -26.0839213
##          13          14          15          16          17          18
## -20.8883941  -8.9331226   6.9496251 -16.2347199 -12.6871158  -2.0002154
##          19          20          21          22          23          24
##   5.9384430   6.0668774 -12.1621961   1.7818932  13.8771014 -28.0503749
##          25          26          27          28          29          30
##   0.7595290 -19.1174676  34.9384430   1.6310946  10.1116059 -38.9721000
##          31          32          33          34          35          36
##  19.8489860  22.0167179   9.7483468 -17.8883941   5.0167179   9.8266217
##          37          38          39          40          41          42
## -13.3018127  -2.3241769  -9.1452629 -35.3912697  12.0055358  28.1061749
##          43          44          45          46          47          48
## -11.0727391  26.8266217   0.9719894  11.7483468 -21.2570841 -13.1845604
```

```
##          49
##   1.8489860
```

CONGRATULATIONS!!!

In the words of Mel Blanc, who was the voice of all the Warner Brothers and Looney Tunes cartoon characters who populated the television of my youth,

```
"That's all folks!"
```

https://en.wikipedia.org/wiki/Mel_Blanc