# MXN500

# PST-2



Submitted By:

Name: Ajay Moktan

ID: n10349561

Submitted To:

Lecturer: Kate Saunders

Tutor: David Rovere

**Q 1.1 Answer:**

Downloading the soi\_data

**soi\_data <- read\_csv(“soi\_data\_wide.csv")**

Converting the Dataset from wide to long format.

**soi\_data <- soi\_data %>% pivot\_longer( !Year, names\_to = "Month",**

**values\_to = "Value")**

**head( soi\_data, 3 )**

**Output:**

*> head(soi\_data,3)*

*# A tibble: 3 x 4*

*Year Month Value Season*

*<dbl> <fct> <dbl> <chr>*

*1 2021 Jan 16.5 Summer*

*2 2021 Feb 11.5 Summer*

*3 2021 Mar -0.3 Autumn*

**Q 1.2 Answer:**

# we are converting the Month into factor variable and also using the abbreviate form of Months of a year in an order.

**library( ggplot2 )**

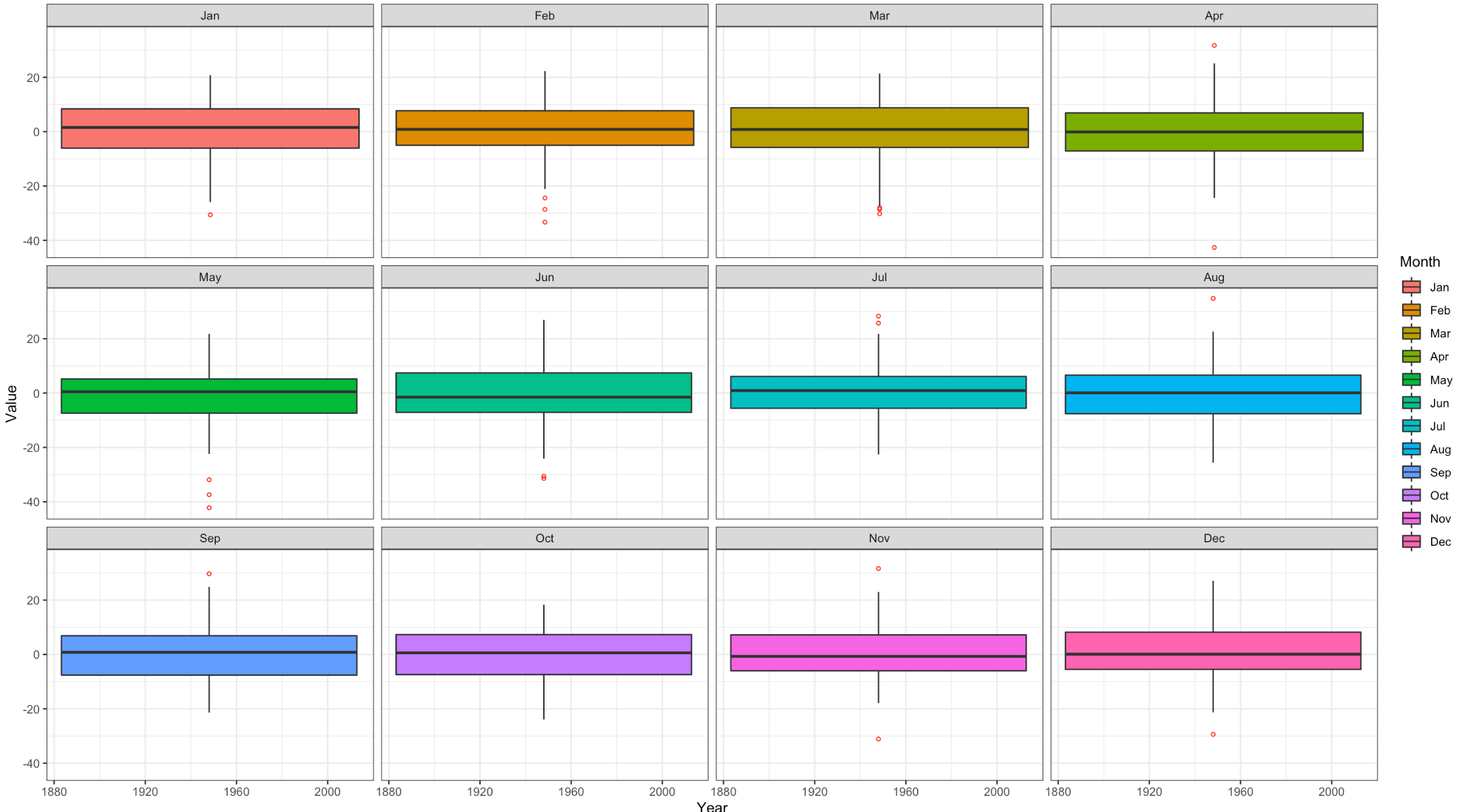
**soi\_data$Month = factor( soi\_data$Month, levels = month.abb )**

**ggplot( soi\_data, aes(x=Year, y=Value, fill = Month ))+**

**geom\_boxplot(outlier.colour = "red", outlier.shape = 1, outlier.size=1)+**

**facet\_wrap(~Month)+**

**theme\_bw()**



Seasonal SOI Data

**Q 1.3 Answer:**

# Categorising Seasons

**summer\_months = c("Dec", "Jan", "Feb")**

**spring\_months = c("Sep", "Oct", "Nov")**

**autumn\_months = c("Mar", "Apr", "May")**

**winter\_months = c("Jun", "Jul", "Aug")**

#Creating "soi\_data" with Season column.

**soi\_data <- soi\_data %>%**

**mutate(Season = NA\_character\_) %>%**

**mutate(Season = if\_else(Month %in% summer\_months, "Summer", Season)) %>%**

**mutate(Season = if\_else(Month %in% spring\_months, "Spring", Season)) %>%**

**mutate(Season = if\_else(Month %in% autum\_months, "Autumn", Season)) %>%**

**mutate(Season = if\_else(Month %in% winter\_months, "Winter", Season))**

# Code to show the rows corresponding to Year 2020.

**soi\_data %>% filter(Year == 2020 )**

**Output:**

# A tibble: 12 x 4

Year Month Value Season

<dbl> <chr> <dbl> <chr>

1 2020 Jan 1.3 Summer

2 2020 Feb -2.2 Summer

3 2020 Mar -5.2 Autumn

4 2020 Apr -0.5 Autumn

5 2020 May 2.8 Autumn

6 2020 Jun -9.6 Winter

7 2020 Jul 4.2 Winter

8 2020 Aug 9.8 Winter

9 2020 Sep 10.5 Spring

10 2020 Oct 4.2 Spring

11 2020 Nov 9.2 Spring

12 2020 Dec 16.9 Summer

**Q 1.4 Answer:**

#Creating data set "seasonal\_soi\_data"

**seasonal\_soi\_data <- soi\_data %>% group\_by(Year, Season) %>%**

**summarise(Mean = mean(Value, na.rm =TRUE)) %>% ungroup() %>%**

**transmute(Year = Year, Season = Season, SeasonalSOI = Mean)**

# Adding columns SeasonalSOI and assigning mean of each season of the each year and

# Keeping Year and Season column as well.

# Printing out the rows corresponding to Year 2020

**seasonal\_soi\_data %>% filter(Year == 2020)**

**Output:**

*# A tibble: 4 x 3*

*Year Season SeasonalSOI*

*<dbl> <chr> <dbl>*

*1 2020 Autumn -0.967*

*2 2020 Spring 7.97*

*3 2020 Summer 5.33*

*4 2020 Winter 1.47*

**Q 1.5 Answer:**

# Adding New column “Phase ” in a existing dataset “seasonal\_soi\_data”.

**seasonal\_soi\_data <- seasonal\_soi\_data %>%**

**mutate(Phase = "Neutral") %>%**

**mutate(Phase = if\_else(SeasonalSOI > 8, "LaNina", Phase)) %>%**

**mutate(Phase = if\_else(SeasonalSOI < -8, "ElNino", Phase))**

# Displaying rows corresponding to Year 2011

**seasonal\_soi\_data %>% filter(Year == 2011)**

**Output:**

*# A tibble: 4 x 4*

*Year Season SeasonalSOI Phase*

*<dbl> <chr> <dbl> <chr>*

*1 2011 Autumn 16.2 LaNina*

*2 2011 Spring 10.9 LaNina*

*3 2011 Summer 21.7 LaNina*

*4 2011 Winter 4.33 Neutral*

**Q 1.6 Answer:**

#Checking the data type of the variables in “seasonal\_soi\_data"

**str(seasonal\_soi\_data)**

**Output:**

*$ Year : num [1:584] 1876 1876 1876 1876 1877 ...*

*$ Season : chr [1:584] "Autum" "Spring" "Summer" "Winter" ...*

*$ SeasonalSOI: num [1:584] 5.4667 -0.0667 6.4333 7.9667 -3.5667 ...*

*$ Phase : chr [1:584] "Neutral" "Neutral" "Neutral" "Neutral" …*

# The variables are not a factor datatype. Hence, converting them into factor.

**seasonal\_soi\_data$Season <- as.factor(seasonal\_soi\_data$Season)**

**seasonal\_soi\_data$Phase <-as.factor(seasonal\_soi\_data$Phase)**

**str(seasonal\_soi\_data)**

**Output:**

*> str(seasonal\_soi\_data)*

*tibble [584 × 4] (S3: tbl\_df/tbl/data.frame)*

*$ Year : num [1:584] 1876 1876 1876 1876 1877 ...*

*$ Season : Factor w/ 4 levels "Autumn","Spring",..: 1 2 3 4 1 2 3 4 1 2 ...*

*$ SeasonalSOI: num [1:584] 5.4667 -0.0667 6.4333 7.9667 -3.5667 ...*

*$ Phase : Factor w/ 3 levels "ElNino","LaNina",..: 3 3 3 3 3 1 1 1 3 2 ...*

# Hence, Variable Season and Phase are converted into a factor datatype.

# Viewing the levels of the factor variables.

**levels(seasonal\_soi\_data$Season)**

**Output:**

*[1] "Autumn" "Spring" "Summer" “Winter"*

**levels(seasonal\_soi\_data$Phase)**

**Output:**

*[1] "ElNino" "LaNina" “Neutral"*

**Section 2 Linear Regression:**

# Creating dataset “total\_seasonal\_rainfall” by adapting the provided code.

**total\_seasonal\_rainfall <- read\_csv("total\_seasonal\_rainfall.csv") %>%**

**mutate(total\_seas\_prcp = total\_seas\_prcp/10) %>%**

**left\_join(seasonal\_soi\_data)**

**Q 2.1 Answer:**

*For the Brisbane Station in Spring, a linear model was specified to model*

*how the total seasonal precipitation, ..****yi.****, is related to the mean seasonal*

*SOI value, ..****xi****.. The parameter .****β1****.. describes the rate of change in the*

*total seasonal precipitation with an increase in mean seasonal SOI value.*

*The parameter .* ***β0****. . represents the total seasonal precipitation when the mean*

*seasonal SOI value is 0.*

**Q 2.2 Answer:**

# Creating a dataset "BRO\_Spring\_rainfall" with the information related to

# BRISBANE REGIONAL OFFICE Station and Spring season only.

**BRO\_Spring\_rainfall <- total\_seasonal\_rainfall %>%**

**filter( Season == "Spring" & name == "BRISBANE REGIONAL OFFICE")**

# Creating Linear Model “BRO\_Spring\_rainfall\_lm”

**BRO\_Spring\_rainfall\_lm <- lm(data = BRO\_Spring\_rainfall,**

**total\_seas\_prcp ~ SeasonalSOI)**

# Viewing the parameters of linear model “BRO\_Spring\_rainfall\_lm”

t**idy(BRO\_Spring\_rainfall\_lm)**

**Output:**

*A tibble: 2 x 5*

*term estimate std.error statistic p.value*

*<chr> <dbl> <dbl> <dbl> <dbl>*

*1 (Intercept) 222. 12.4 17.9 5.00e-30*

*2 SeasonalSOI 3.68 1.43 2.58 1.17e- 2*

Hence, Based on the parameter estimate, the equation is:

Total Seasonal Precipitation = 222.199 + 3.68 \* SeasonalSOI + ε

**Q 2.3 Answer:**

**round( glance ( BRO\_Spring\_rainfall\_lm )$r.squared, 4 )**

*> round(glance(BRO\_Spring\_rainfall\_model)$r.squared,4)*

*[1] 0.0749*

*The variability in data explained by the model is 7.49 %*

**Q 2.4 Answer:**

The code for parameter estimate for 95% confidence.

**tidy(BRO\_Spring\_rainfall\_lm, conf.int = T, conf.level = 0.95) %>%**

**select(term, estimate, conf.low, conf.high, p.value)**

**Output:**

*A tibble: 2 x 5*

*term estimate conf.low conf.high p.value*

*<chr> <dbl> <dbl> <dbl> <dbl>*

*1 (Intercept) 222. 197. 247. 5.00e-30*

*2 SeasonalSOI 3.68 0.840 6.52 1.17e- 2*

**Q 2.5 Answer:**

Let's create a data frame using fortify() function from ggplot2 that has everything to analyse the residuals.

**BRO\_Spring\_rainfall\_lm.fort <- fortify(BRO\_Spring\_rainfall\_lm)**

**head(BRO\_Spring\_rainfall\_lm.fort)**

**Output:**

*total\_seas\_prcp SeasonalSOI .hat . sigma .cooksd . fitted .resid*

*1 215.3 15.100000 0.04818984 114.3712 0.0079992750 277.7602 -62.46024*

*2 168.3 -9.266667 0.02511424 114.5706 0.0003994288 188.1025 -19.80246*

*3 253.5 -3.133333 0.01335298 114.4920 0.0009699179 210.6702 42.82978*

*4 126.3 -10.166667 0.02783457 114.4025 0.0038839100 184.7909 -58.49089*

*5 99.1 -10.966667 0.03046631 114.2112 0.0085544080 181.8473 -82.74726*

*6 108.9 -1.566667 0.01224400 113.9598 0.0055938485 216.4348 -107.53481*

*.stdresid*

*1 -0.5621313*

*2 -0.1760968*

*3 0.3785945*

*4 -0.5208677*

*5 -0.7378728*

*6 -0.9500217*

Now, Visualising the fitted value (y[i] hat), .fitted compared with the residuals (epsilon[i]), .resid

**ggplot(data = BRO\_Spring\_rainfall\_lm.fort, aes(x = .fitted, y = .resid))+**

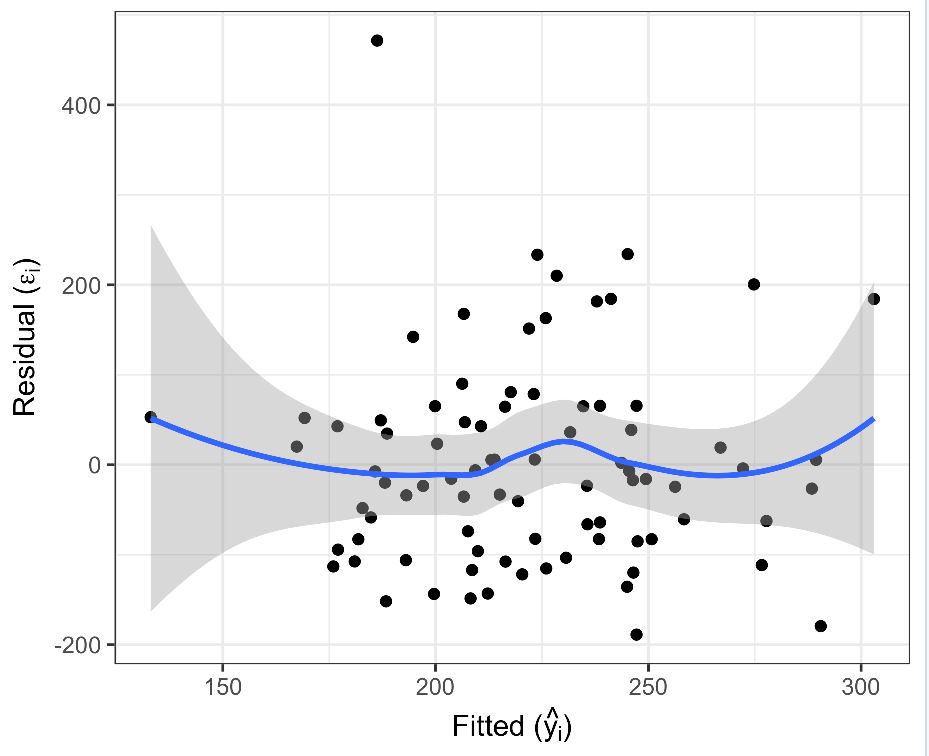
**geom\_point()+**

**theme\_bw()+**

**geom\_smooth()+**

**labs(x = expression(paste("Fitted (",hat(y[i]), ")")),**

**y = expression(paste("Residual (“,epsilon[i],")")))**



**Fig.1 Residual vs. Fitted graph**

Now Visualising the standardised quantiles of the residuals compared with the theoretical quantiles.

**ggplot(data=BRO\_Spring\_rainfall\_lm.fort, aes(sample=.stdresid)) +**

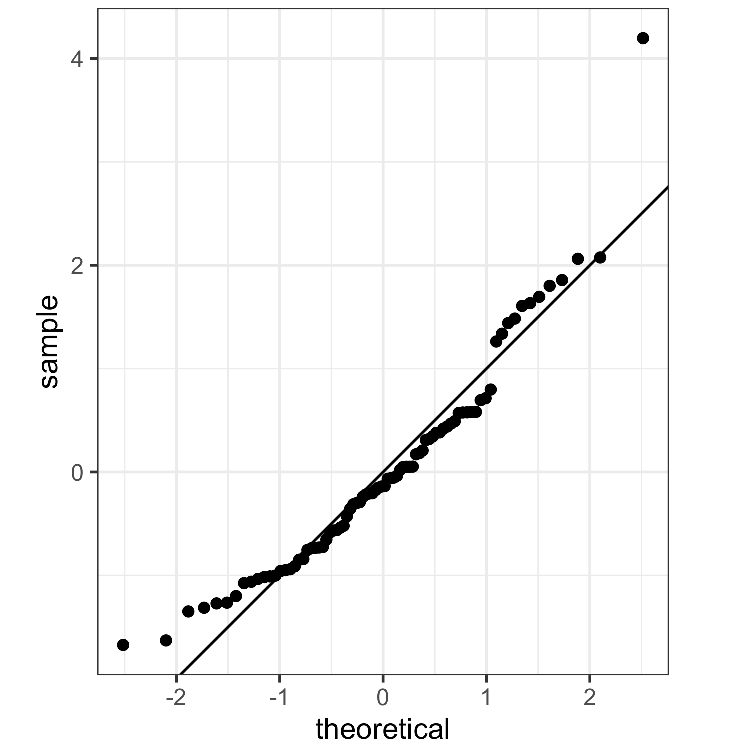
**stat\_qq() +**

**geom\_abline(intercept=0, slope=1) +**

**coord\_equal()+**

**theme\_bw()**

Validity of the underlying assumptions of linear regression:



***Fig.2 Standardised residuals quantiles vs.theoretical quantiles.***

**From the Fig.1 Residual vs. Fitted graph:**

The fitted vs. residual is essential in determining the assumption of Linearity and homoscedasticity of the model.

Having seen the residuals are not too far away from 0, it suggests that the model complies with linearity.

Since the residuals are equally spread around the line y=0 and also, there is no any clear pattern seen, hence, homoscedasticity holds true.

**From Fig. 2 standardised residual quantiles vs theoretical quantiles:**

It is used to test the assumption of Normality. Since the sample observation lies well along the line of 45-degree, we can say that Normality holds.

Hence, it validates the underlying assumption of linear regression.

**Q 2.6 Answer:**

**summary (BRO\_Spring\_rainfall\_lm)**

**Output:**

*Call:*

*lm(formula = total\_seas\_prcp ~ SeasonalSOI, data = BRO\_Spring\_rainfall)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-188.72 -82.69 -15.65 52.31 471.54*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 222.199 12.427 17.880 <2e-16 \*\*\**

*SeasonalSOI 3.680 1.428 2.578 0.0117 \**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 113.9 on 82 degrees of freedom*

*(26 observations deleted due to missingness)*

*Multiple R-squared: 0.07495, Adjusted R-squared: 0.06367*

*F-statistic: 6.644 on 1 and 82 DF, p-value: 0.01174*

**Interpretation of the Model based on above summary:**

For one unit increase in the SeasonalSOI, there is an increase of 3.68 unit in a total\_seaosnal\_rainfall. SeasonalSOI explains 7.49% variability in total\_seaosnal\_rainfall. When testing the null hypothesis that there is no linear association between SeasonalSOI and total seasonal rainfall, we reject the null hypothesis in favour of alternate hypothesis, given by the fact that p-value = 0.0117 < 0.05.

**Q 2.7 Answer:**

A visualisation for each city and for each season that compares the proposed linear model to the null model.

**ggplot(total\_seasonal\_rainfall, aes(x = SeasonalSOI, y = total\_seas\_prcp))+**

**geom\_point()+**

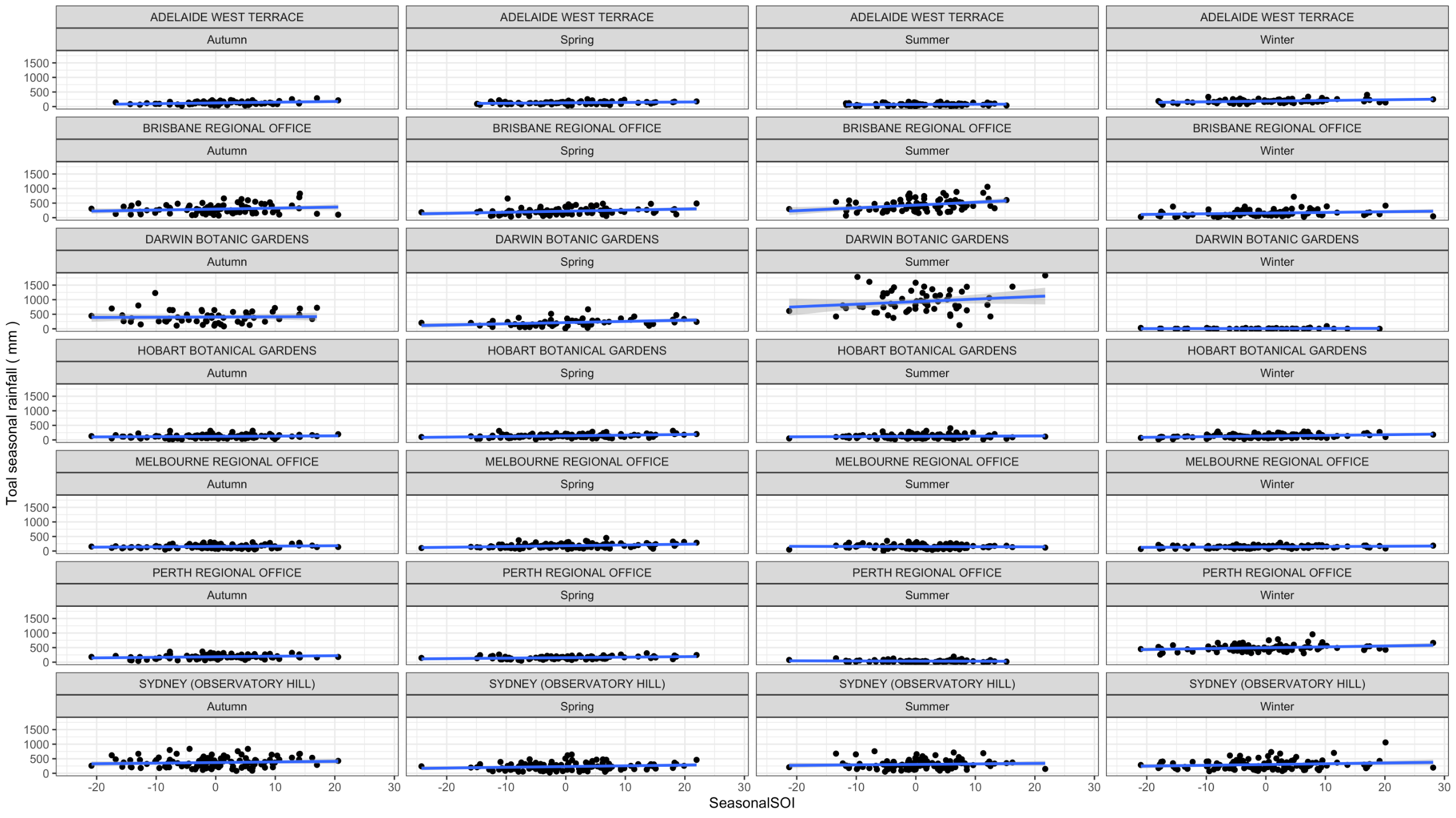
**facet\_wrap(~name + Season, nrow = 7)+**

**theme\_bw()+**

**geom\_smooth(method = 'lm', formula = y~x) +**

**labs(y = "Toal seasonal rainfall ( mm )")**

**Output Graph:**



***Total Rainfall vs. SeasonalSOI for each city and each Season.***

Apart from BRISBANE REGIONAL OFFICE, Summer and DARWIN BOTANIC GARDENS, Summer where there is a significant relationship between total rainfall and seasoanlSOI and a slight response in the locations MELBOURNE REGIONAL OFFICE(winter), PERTH REGIONAL OFFICE(winter) and SYDNEY (OBSERVATORY HILL) (winter), rest of the location shows no response. Those locations with no response favours the null hypothesis and Hence have Beta1=0 which is also called the null model. Hence, we can see the comparison between model that favour the null hypothesis and the model that reject the null hypothesis.

**Q 2.8 Answer:**

In terms of significant linear relationship based on visualisation of total seasonal precipitation and the mean soasonal SOI value, the cities and seasons are:

1. BRISBANE REGIONAL OFFICE, Summer. and

2. DARWIN BOTANIC GARDENS, Summer.

**Q 2.9 Answer:**

All the locations have almost same amount of rainfall recorded However, BRISBANE REGIONAL OFFICE and HOBART BOTANICAL GARDENS seems to have recorded higher rainfall.

The total seasonal rainfall is significantly responsive to ENSO as we move from El Niño phase to La Niña in BRISBANE REGIONAL OFFICE, Summer and DARWIN BOTANIC GARDENS, Summer.

There is a slight response to total seasonal rainfall due to ENSO in the locations like

MELBOURNE REGIONAL OFFICE(winter), PERTH REGIONAL OFFICE(winter) and

SYDNEY (OBSERVATORY HILL) (winter).

However, the rest of the locations with seasons seems to have no effect on total

seasonal rainfall due to ENSO.

**Section 3: Polynomial Lines of Best Fit**

**Q 3.1 Answer:**

For BRISBANE REGIONAL OFFICE, Spring Season, and to fit into polynomial model, a dataset “**Brisbane\_spring\_poly**” with “Log\_prcp“ column is created using  **“BRO\_Spring\_rainfall”** dataset.

**Brisbane\_spring\_poly <- mutate(BRO\_Spring\_rainfall,**

**Log\_prcp = log(total\_seas\_prcp))**

Creating a linear model “Bris\_spring\_poly\_lm” of order 2 using function poly( )

**Bris\_spring\_poly\_lm <- lm( data = Brisbane\_spring\_poly,**

**Log\_prcp ~ poly(SeasonalSOI, 2, raw=T ))**

**tidy(Bris\_spring\_poly\_lm) %>% select(term, estimate)**

**Output:**

*A tibble: 3 x 2*

*term estimate*

*<chr> <dbl>*

*1 (Intercept) 5.26*

*2 poly(SeasonalSOI, 2, raw = T)1 0.0195*

*3 poly(SeasonalSOI, 2, raw = T)2 0.0000342*

*Hence, the equation is :*

log (total\_seas\_prcp) = 5.26 + 0.0195 \* SeasonalSOI + 0.0000342 \* SeasonalSOI^2

**Q 3.2 Answer:**

**summary(Bris\_spring\_poly\_lm)**

**Output:**

*Call:*

*lm(formula = Log\_prcp ~ poly(SeasonalSOI, 2, raw = T), data = Brisbane\_spring\_poly)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-1.47857 -0.30333 0.05821 0.37638 1.42097*

*Coefficients:*

*Estimate Std. Error t value*

*(Intercept) 5.255e+00 7.383e-02 71.176*

*poly(SeasonalSOI, 2, raw = T)1 1.950e-02 6.950e-03 2.806*

*poly(SeasonalSOI, 2, raw = T)2 3.424e-05 5.666e-04 0.060*

*Pr(>|t|)*

*(Intercept) < 2e-16 \*\*\**

*poly(SeasonalSOI, 2, raw = T)1 0.00627 \*\**

*poly(SeasonalSOI, 2, raw = T)2 0.95196*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 0.5499 on 81 degrees of freedom*

*(26 observations deleted due to missingness)*

*Multiple R-squared: 0.09044, Adjusted R-squared: 0.06798*

*F-statistic: 4.027 on 2 and 81 DF, p-value: 0.02151*

**Interpretation of Result of the summary:**

* When SeasonalSOI is 0, The log of total rainfall = 5.255e+00.
* log of total rainfall(Log\_prcp) increases by 1.950e-02 with one unit increase in SeasonalSOI.
* log of total rainfall(Log\_prcp) increases by 3.424e-05 with one unit increase in SeasonalSOI square.
* There is 9% variability on log of total rainfall(Log\_prcp) predicted by SeasonalSOI.
* P-value: 0.02151 which is less than 0.05 indicates that we can reject the null hypothesis which suggests that there is no effect of log of total rainfall (Log\_prcp) due to SeasonalSOI ( that is beta\_one and beta\_one are 0)

**Q 3.3 Answer:**

Predicting for mean SeasonalSOI = 25 using model “Bris\_spring\_poly\_lm”

**predict ( Bris\_spring\_poly\_lm,data.frame ( SeasonalSOI = 25 ),**

**interval = "conf", level = 0.95)**

**Output:**

*fit lwr upr*

*1 5.764111 5.083633 6.44459*

**Comment:** For mean SeaonalSOI of 25 with confidence interval of 95% , Log\_prcp = 5.76411 which is within the range (5.083633 6.44459) of true

population parameter of 95% confidence interval.

Again, Predicting for mean SeasonalSOI = -25 using model “Bris\_spring\_poly\_lm”

**predict(Bris\_spring\_poly\_lm,data.frame(SeasonalSOI = -25),**

**interval = "conf", level = 0.95)**

**Output:**

*fit lwr upr*

*1 4.788972 4.033551 5.544393*

**Comment:** For mean SeaonalSOI of -25 with confidence interval of 95% , Log\_prcp = 4.788972 which is within the range (4.033551 5.544393) of true population parameter of 95% confidence interval.

**Q 3.4 Answer:**

Comparing the linear model and polynomial model

**anova(BRO\_Spring\_rainfall\_lm , Bris\_spring\_poly\_lm)**

**Output:**

*Analysis of Variance Table*

*Response: total\_seas\_prcp*

*Df Sum Sq Mean Sq F value Pr(>F)*

*SeasonalSOI 1 86176 86176 6.6436 0.01174 \**

*Residuals 82 1063642 12971*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

Since p-value: 0.01174 < 0.05, we can reject the null hypothesis that models explain the same amount of variability in data in favour of alternate hypothesis that the quadratic model explains the greater deal of variability in data. Which is true when we see the variability going up from 7% to 9%.

**Section 4: Linear Regression with Categorical Explanatory Variables**

**Q 4.1 Answer:**

Making Phase “ElNino" as a reference Phase.

**BRO\_Spring\_rainfall$Phase <- relevel(BRO\_Spring\_rainfall$Phase, ref = "ElNino")**

Creating Linear model “BRO\_Spring\_rainfall\_phase\_lm”

**BRO\_Spring\_rainfall\_phase\_lm <- lm(data = BRO\_Spring\_rainfall,**

**total\_seas\_prcp ~ Phase)**

Let’s check the estimate parameters:

**tidy(BRO\_Spring\_rainfall\_phase\_lm)**

**Output:**

*# A tibble: 3 x 5*

*term estimate std.error statistic p.value*

*<chr> <dbl> <dbl> <dbl> <dbl>*

*1 (Intercept) 181. 29.1 6.23 0.0000000201*

*2 PhaseLaNina 91.4 45.5 2.01 0.0482*

*3 PhaseNeutral 42.6 32.9 1.30 0.199*

Our Equation is of the form :

total\_seas\_prcp (i) = Beta (0) + Beta (1) I (Phase(i) ==LaNina )+Beta (2) I (Phase(i) == Neutral ) + erroe

Since Phase(i) ==LaNina is true as it is significantly different from 0 given by 0.0482 < 0.05 and the Phase(i) == Neutral is false as its not significantly different from 0 given by 0.1985 > 0.05

Hence, The equation based on the model is :

*mean of total\_seas\_prcp = 180.94 + 91.36*

**Q 4.2 Answer:**

**summary(BRO\_Spring\_rainfall\_phase\_lm)**

**Output:**

*Call:*

*lm(formula = total\_seas\_prcp ~ Phase, data = BRO\_Spring\_rainfall)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-167.48 -81.98 -7.64 47.05 476.86*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 180.94 29.07 6.225 2.01e-08 \*\*\**

*PhaseLaNina 91.36 45.54 2.006 0.0482 \**

*PhaseNeutral 42.64 32.89 1.296 0.1985*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 116.3 on 81 degrees of freedom*

*(26 observations deleted due to missingness)*

*Multiple R-squared: 0.04778, Adjusted R-squared: 0.02427*

*F-statistic: 2.032 on 2 and 81 DF, p-value: 0.1377*

**Interpetation of Result:**

* Intercept = 180.94 is the mean total\_seas\_prcp of ElNino Phase.
* PhaseLaNina = 91.36 is the sum of mean of total\_seas\_prcp of ElNino Phase and LaNina phase.
* PhaseNeutral = 42.64 is the sum of mean of total\_seas\_prcp of ElNino Phase and Neutral phase.
* P-value of baseline (2.01e-08) and the LaNina phase (0.0482) are less than 0.05 which indicates that the beta 0 and beta 1 are not zero.
* However, the P-value of Neutral pahse 0.1985 > 0.05 indicating beta 2 equals to zero.
* On ElNino Phase is significantly different from 0.
* On LaNina phase is significantly different from ElNino.
* On Neutral phase is not significantly different from ElNino.

**Q 4.3 Answer:**