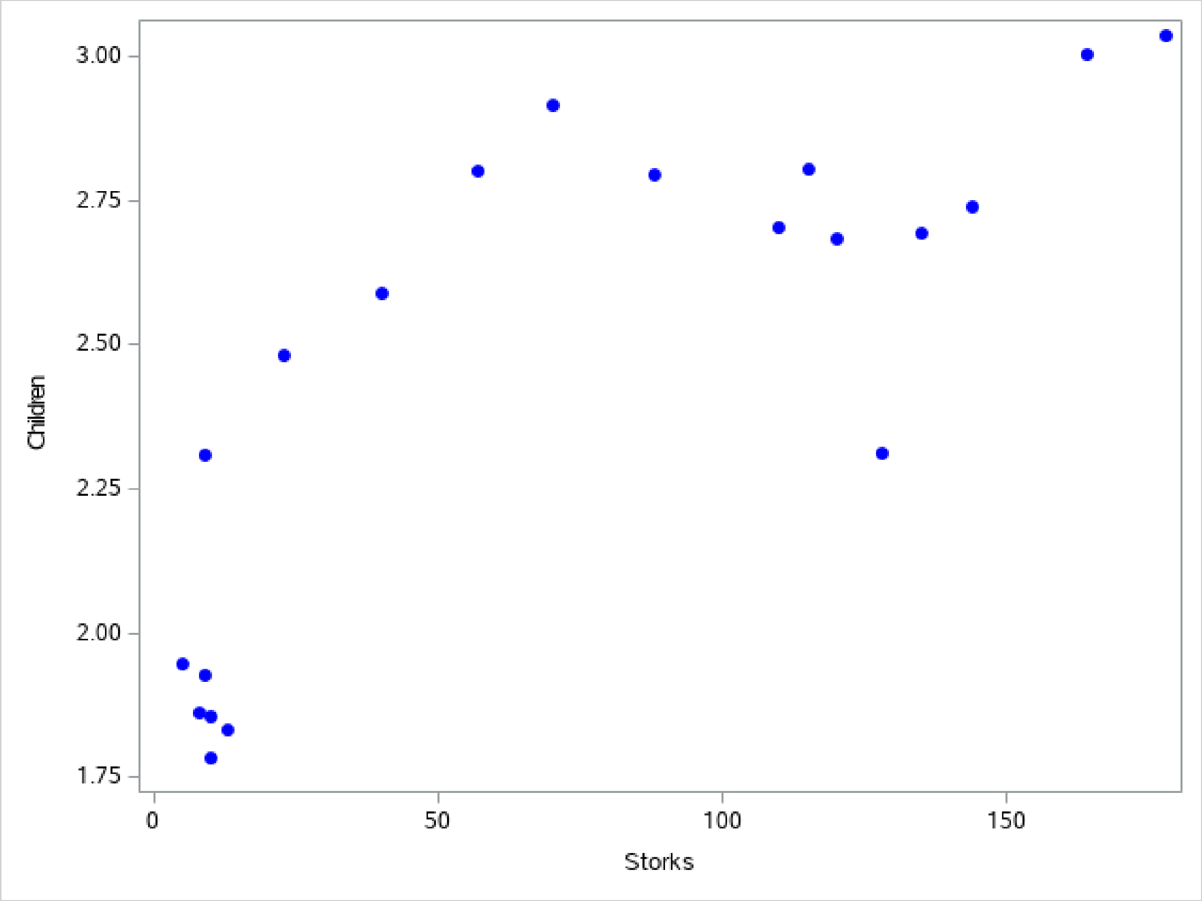
Nikhil Jagatia
18055146

Assignment 7

1.

a) A scatterplot showing Storks Vs Births:



To comment on the scatter plot shown we can visually see 3 distinct zones where we have plots.
The first zone shown as a cluster in the bottom left of the plot shows that as a lower of number of storks are seen so is the number of children that are born. A reason for this could be due to coincidence or that there were not as many storks during those years whether it is due to climate or seasonal changes.

Secondly there is a trend that is noticed whereby as the number of storks increases so does the number of children. This relationship does fit a regression style scheme however more investigation into analysis can be performed (later).
Taking a further look at this section at around 9 storks the value of children broke the 2.25 threshold and there are 5 consistently progressive values that are part of the intial trend. Following this there are a drop in 6 values of children between 100-150 storks as a dip/curve before rising again to 2 of the most maximum points.

Finally there seems to be at least one obvious outlier where at 128 storks there were only 2.3106 children where as part of the trend it would be expected to be slightly more.

Nikhil Jagatia
18055146

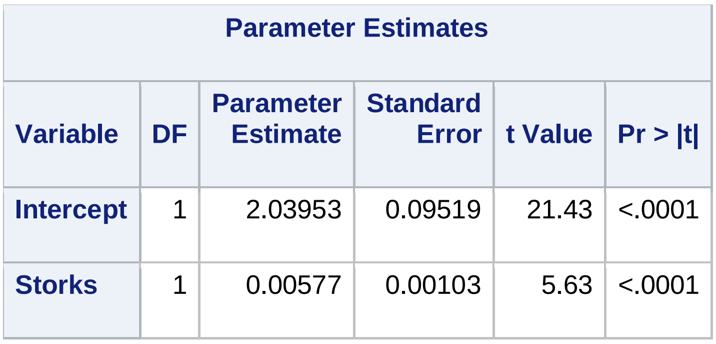
```
FILENAME STORKS '/folders/myfolders/sasuser.v94/storks.csv';

PROC IMPORT DATAFILE=STORKS
        DBMS=CSV
        OUT=STORKS;
        GETNAMES=YES;
RUN;

PROC PRINT DATA=STORKS;
RUN;

PROC SGPLOT DATA=STORKS;
        SCATTER X= STORKS Y= CHILDREN /
        MARKERATTRS= (SYMBOL= CIRCLEFILLED COLOR= BLUE);
        RUN;
```

b) Using linear regression model the parameter estimates are:



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.03953	0.09519	21.43	<.0001
Storks	1	0.00577	0.00103	5.63	<.0001

From these values for the parameter estimates the estimated regression equation is:
Children = 2.155 + 0.00577 Storks .

This means if we have an incidence of 0 storks the value we can approximate for children is 2.155. The coefficient which is not very large is the gradient of the slope fitted to this distribution. The reason of this could be due to the first cluster having an effect of the fit of the line. If the data was capped from above 2.25 we would obviously see a much greater coefficient even with the outlier.

```
PROC REG DATA=STORKS PLOTS=(FITPLOT DIAGNOSTICS);
 MODEL CHILDREN=STORKS;
RUN;
```

Nikhil Jagatia
18055146

c) The ANOVA results as part of the storks diagnostics are:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	2.30485	2.30485	31.67	<.0001
Error	18	1.30987	0.07277		
Corrected Total	19	3.61471			

In terms of linear association if we set the $H_0 = 0$ which would mean there is not a relationship however because the P value is found as <0.0001 which is less than 0.05 at the significance level the P is extremely small and we can thus conclude that we can reject the null hypothesis of there not being a significant relationship between the number of storks and number of children.

d) A 95% confidence interval as calculated where x=1000 storks:

The REG Procedure
Model: MODEL1
Dependent Variable: Children

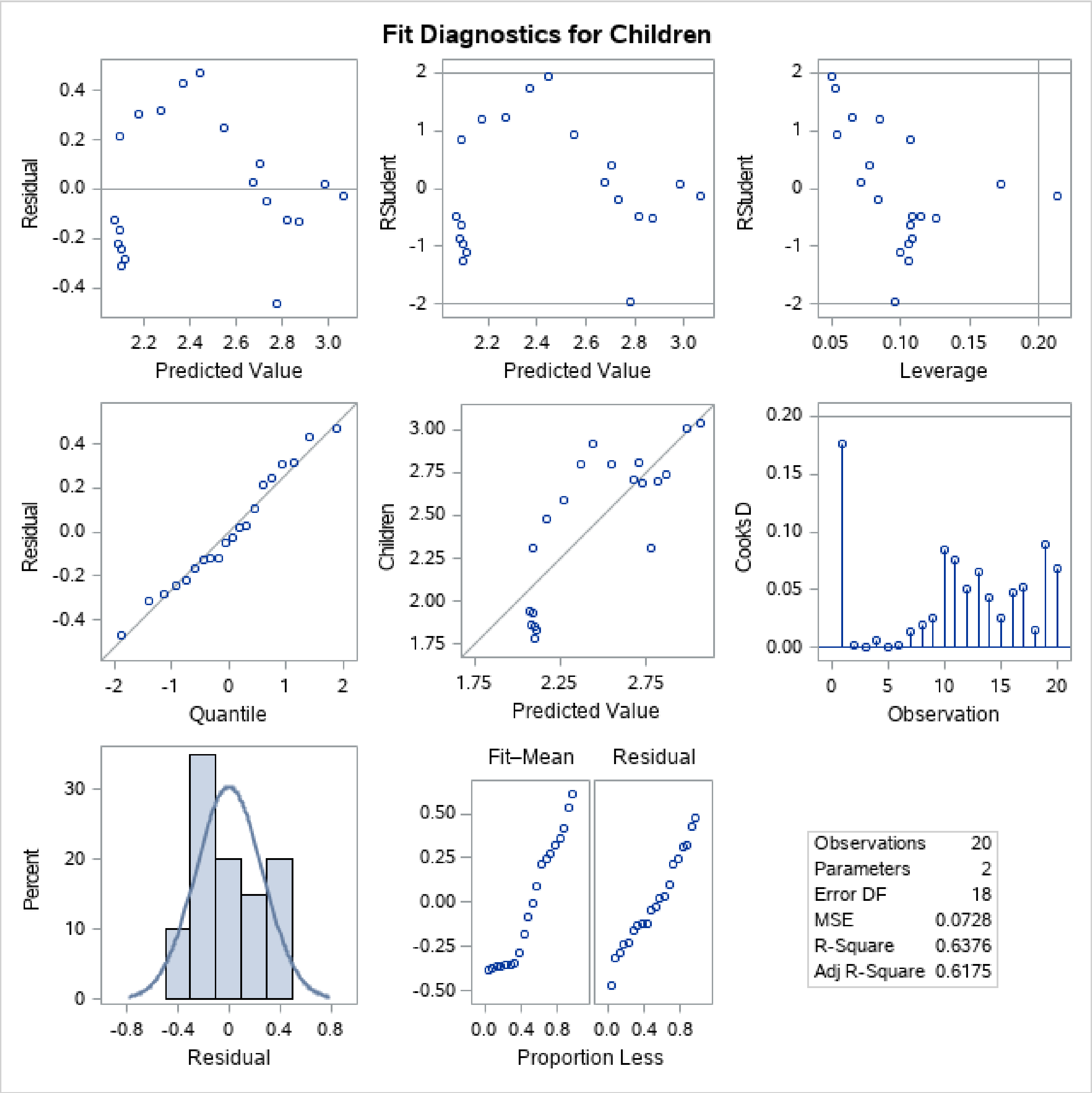
Output Statistics											
Obs	**Dependent Variable**	**Predicted Value**	**Std Error Mean Predict**	**95% CL Mean**		**95% CL Predict**		**Residual**	**Std Error Residual**	**Student Residual**	**Cook's D**
20	1.83	2.1146	0.0853	1.9354	2.2938	1.5202	2.7090	-0.2832	0.256	-1.106	0.068
21	.	7.8113	0.9538	5.8073	9.8152	5.7287	9.8938	.	.	.	.

From the input of data where the number of storks was a large value of 1000 the predicted value is 7.8113 number of children and the 95% prediction value is between 5.7287 and 9.8938. These values have a greater variation due to the added value of 1000 not being part of the initial set data therefore the relationship might slightly different therefore it is wider to ensure correctness. Although these prediction values may fit the equation and model whether it is accurate or not is questionable. The reason this might not be as accurate is due to generating an estimation on the basis of extrapolation. Extrapolating the value of 1000 storks is so extreme the relationship might be totally different (or no relationship) when this value is reached however we do not have enough information to know for sure that this is true.

Nikhil Jagatia
18055146

```
DATA STORKS2; /* Create the new observation */
STORKS=1000;
RUN;
DATA STORKS2; /* Append it to the dataset */
 SET STORKS STORKS2;
RUN;
PROC REG DATA=STORKS2;
 MODEL CHILDREN=STORKS / R CLI CLM;
RUN;
```

e) Charts showing analysis of Storks – Children data:



Observations	20
Parameters	2
Error DF	18
MSE	0.0728
R-Square	0.6376
Adj R-Square	0.6175

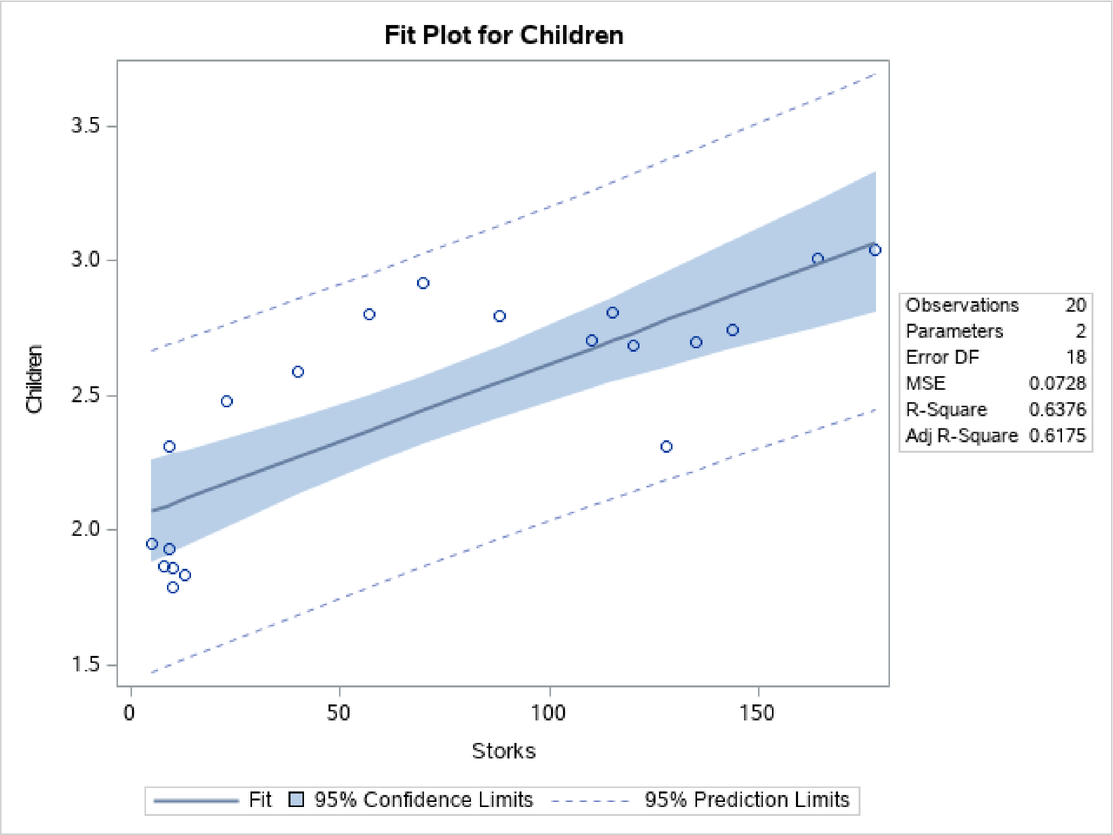
Nikhil Jagatia
18055146

The Residual Vs Predicted values scatter aims to shows the dispersion and along with the histogram which includes the kde curve shows the comparison and how normally distributed the data set and predictions are. A lot of the observations are centrally distributed yet it is still not ideally symmetrical highlighting the lack of the data being normally distributed. Largely there are a good number of points on either side of the line indicating a good set of normally distributed data however due to the clustering of smaller values it changes the shape showing slight skewness and therefore a lack of variability.

The RStudent for Predicted value and Leverage aims to show if there are outliers that cause an effect on the data. There are numerous data points which lie very close to the borders of becoming an outlier which has a trigger on the data. Additionally on the Leverage plot we can clearly see 1 point which is significantly classed as an outlier in this dataset

The QQ plots shown gives a visual representation of the strength of the correlation between the dataset variables via determination of their own respective quantiles. In addition it gives better understanding if the populations have commonalities in their distributions which is given by the strength and closeness to the line with the gradient of 1 whereby the residual shows good strength in its correlation.

The CooksD vs Observations is another interesting plot showing strength of observations and its classification in terms of being an outlier. Here we have many values close to 0 yet there is one significant value which is very high and close to 0.2 . This indicates that although its not classified as an outlier it is an observation that is part of the data which is somewhat unusual as part of the model. Additionally strength of the quantile plots of fitted mean and residual shows that only slightly does the residual have a smaller spread of data than the centred fit


Fit Plot for Children

Observations	20
Parameters	2
Error DF	18
MSE	0.0728
R-Square	0.6376
Adj R-Square	0.6175

Interestingly the fit plot showing the line of best fit along with 95% confidence limits shows that more of the plotted proportion of data fits outside of the 95% confidence limits which is not ideal. Although the majority fit within the prediction limits there is a lot of variation. However the Adjusted R-Squared value shows that 61.75% of observations are explained due to randomness.

f)

Although there is some correlated data the model is weak and not the best predictor given the distribution of data. However there is a statistically significant relationship between storks delivering babies. This is due to the P value found indicating a significant relationship and lack of significant outliers. Furthermore as the number of storks increases as 1000 being used as an example there are more children, yet because this is extrapolated caution needs to be considered when making these assumptions because of a lot of the values not being within the 95% confidence intervals and over half the observations being a cause of randomness. Therefore the two assumptions that can be made are that there is a relationship so that storks deliver babies, but also that there is a coincidence between number of storks and babies. This is because from correlated data and linear regression models for prediction even though there seems to be a relationship it does not mean causation.