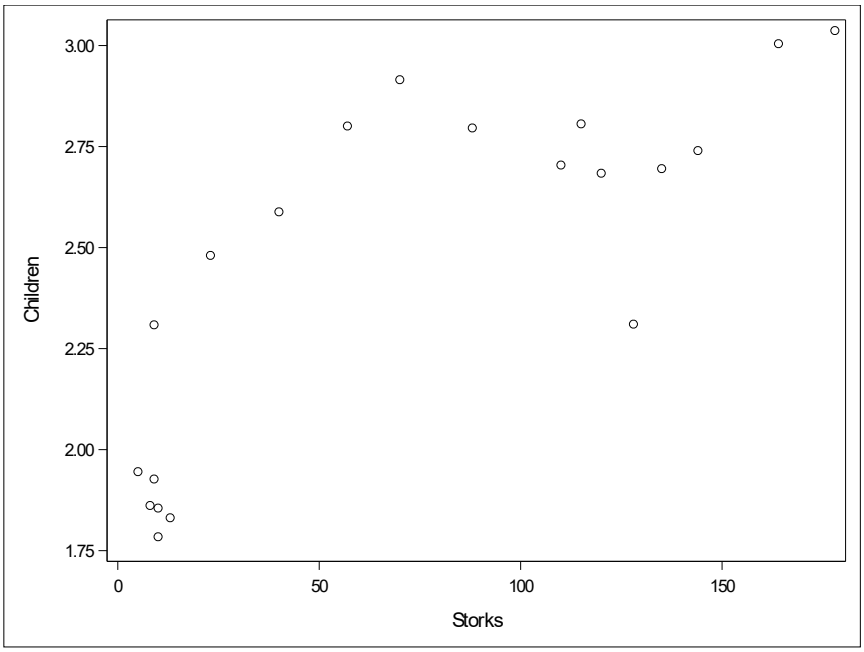
STORKS

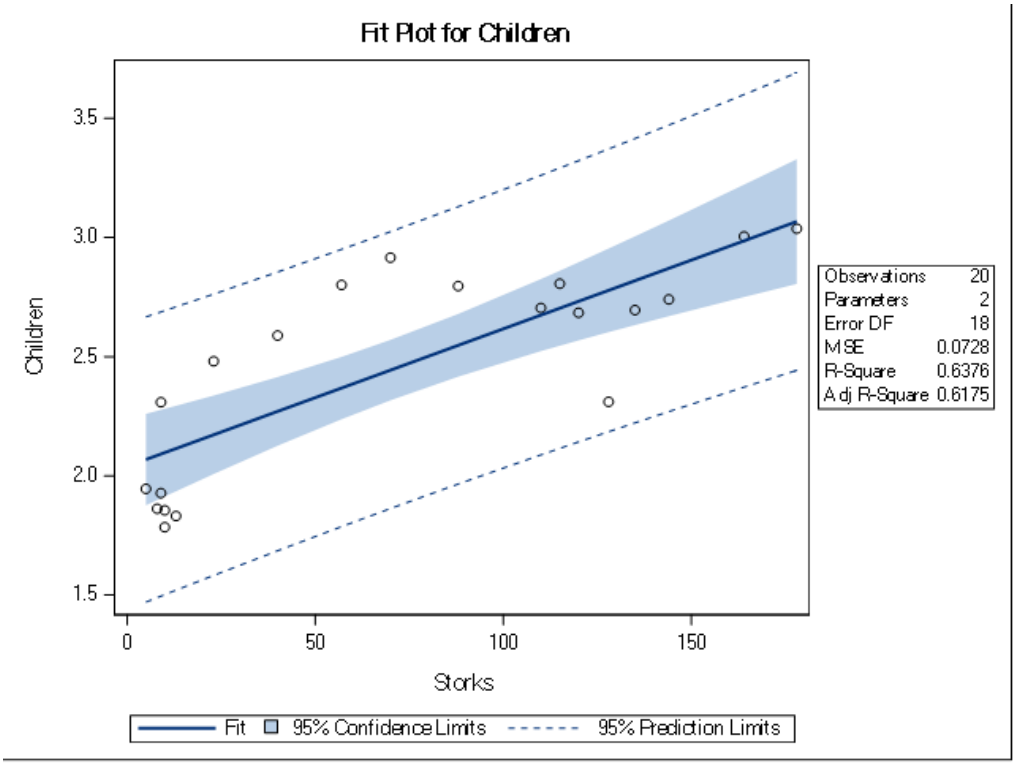
In Assignment 7 we have been given a dataset containing the number of storks observed in the Alsace region of France along with the average number of children per woman in France. Researchers have suggested that storks deliver babies. We have been tasked with proving if this is correct or not.

a) In part A, we have been asked to draw a scatter plot of the two variables. Using this visual statistical technique, we can see if there is any relationship between the two:



Looking at the graph we can see that generally as Storks increase, the average number of children per woman increases. We can also see that the relationship between the two doesn't follow a linear pattern as such. We can see that below 50 storks, the average number of babies per woman ranges from 1.75 – 2.60 and above 50 storks, the average number of babies per woman ranges from 2.75 - 3.

b) In part b we have been tasked with fitting a linear regression model to the data so that we can investigate whether there is a significant linear association between the two variables:



As we can see, SAS has produced a linear regression model with a visual representation. Looking at the regression line we can theorise that when there have been 50 storks observed the average number of children per woman is roughly 2.25 and when there have been 100 storks observed the average number of children per woman is roughly 2.5. Although statistically descriptive, we can produce an equation to find the exact value of the average number of children related to the number of storks by using the formula y = mx + c :

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

Here we can see that the average number of children per woman = 2.04 + 0.0058*storks. This means that when there are 50 storks, we estimate there will be 2.33 average children per woman (2.04 + (0.0058*50)) and when there are 100 storks, we estimate there will be 2.62 average number of children per woman (2.04 + (0.0058*100)).

c) In part C, we want to investigate whether there is a significant linear association between Storks and Births. Looking at the results of the Parameter Estimates and the ANOVA for Regression Coefficients:

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 order to interpret the p values, we must first state the hypothesis, $H_0 = \beta_1 = 0$. If we were to accept this hypothesis, we would be accepting that the two variables are not related. Looking at the p-values for the ANOVA and Parameter Estimations we see that they are <0.0001. This means that we reject the null hypothesis, therefore concluding based on the results, with the p-value being extremely small that there is a relationship between the number of storks and the average number of children born per woman.

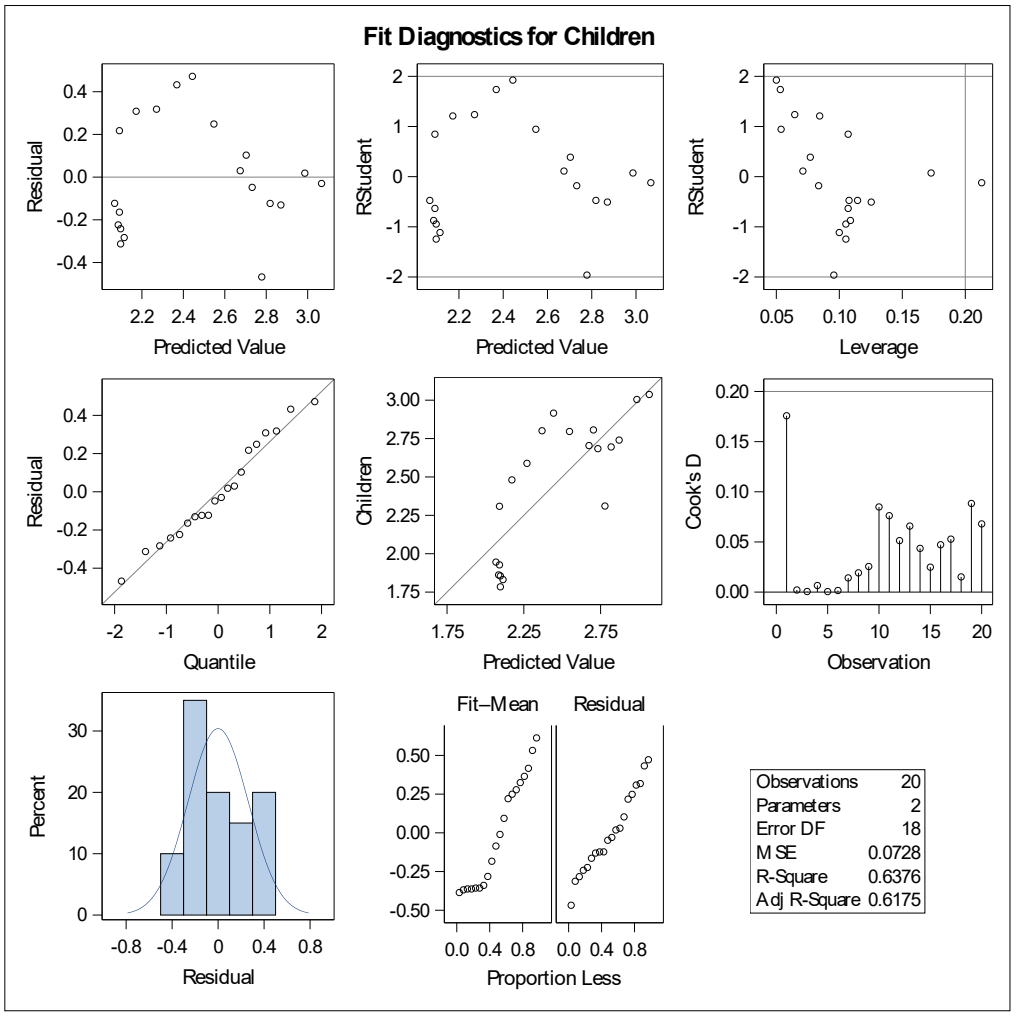
d) Now that we have fitted a model to the data, we have been tasked with predicting how many children are born if 1000 storks have been observed. Using SAS we can produce a 95% prediction interval:

(In order to preserve space I have removed some rows from the output)

Output Statistics											
Obs	**Dependent Variable**	**Predicted Value**	**Std Error Mean Predict**	**95% CL Mean**		**95% CL Predict**		**Residual**	**Std Error Residual**	**Student Residual**	-2-1 0 1 2
1	2.31	2.7783	0.0834	2.6030	2.9536	2.1851	3.3715	-0.4677	0.257	-1.823	| ***| |
2	3.04	3.0669	0.1245	2.8053	3.3285	2.4427	3.6911	-0.0298	0.239	-0.125	| | |
11	2.80	2.3685	0.0622	2.2378	2.4992	1.7869	2.9501	0.4324	0.262	1.647	| |*** |
12	2.59	2.2704	0.0686	2.1263	2.4145	1.6856	2.8552	0.3181	0.261	1.219	| |** |
13	2.48	2.1723	0.0784	2.0076	2.3369	1.5821	2.7625	0.3084	0.258	1.195	| |** |
14	2.31	2.0915	0.0882	1.9061	2.2769	1.4952	2.6878	0.2174	0.255	0.853	| |* |
19	1.78	2.0972	0.0875	1.9134	2.2811	1.5014	2.6931	-0.3128	0.255	-1.226	| **| |
20	1.83	2.1146	0.0853	1.9354	2.2938	1.5202	2.7090	-0.2832	0.256	-1.106	| **| |
21	.	7.8113	0.9538	5.8073	9.8152	5.7287	9.8938	.	.	.	

The row of interest is the last one, here we can see that the predicted value for 1000 storks is 7.8113 children per woman with a 95% CL predict of 5.729 – 9.894. Looking at these results we can conclude that they aren't accurate as it is unlikely that a woman can give birth to on average 5-9 children in a year. It is more likely that the results produced are an example of the dangers of extrapolation as the predicted value is far beyond the range of values in our dataset.

e) In part E, we want to asses the model assumptions using the appropriate plots:



- Residual vs Predicted value – we expect a random scatter plot with no evidence of patterns however looking at the plot we can see it follows a pattern similar to that of a $y=-x^2$ graph.

- Externally studentized residuals (RStudent) vs predicted values and RStudent vs leverage – again follows the same pattern however all points within the expected range

- Normal QQ Plot (Residual vs Normal Quantiles) and Observed vs Fitted – the normal QQ plot follows the straight line with gradient 1 whilst the observed vs fitted doesn't follow the straight line

- *Cook's Distance – we can see there are no outliers as there are no points which exceed the 0.20 cooks d mark.*

- Histogram – the histogram has a skew to the left with just one mode of -0.2, is not symmetric.

- Quantile plots of Fitted-mean and Residuals – the spreads of both graphs are similar however the spread of the right is slightly bigger than the left.

f) Looking at all of the statistical analysis, we can conclude that there is a relationship between the number of storks and the average number of children per women. However, the relationship is spurious as using my domain knowledge I can say with 100% confidence that the number of storks isn't related to the average number of children per woman in France.

CODE

```
PROC IMPORT DATAFILE='/folders/myfolders/7/data/storks.csv'
     DBMS=CSV
     OUT=storks;
     GETNAMES=YES;
RUN;
PROC SGPLOT DATA=storks;
     scatter x=storks y=children;
RUN;
PROC REG DATA=storks;
     Model children = storks;
RUN;
data storks2;
storks = 1000;
Run;
DATA storks2;
     set storks storks2;
run;
PROC REG DATA=storks2 PLOTS(ONLY)=(FITPLOT DIAGNOSTICS);
     Model children = storks / R CLM CLI;
RUN;
```