Simple linear regression (SLR) is used to check if there is a linear relationship between two numerical variables. Based on SLR, we can use one numerical variable (X) to predict the value of the other numerical variable (Y). The Child Health and Development Studies (CHDS) conducted a study in 1960 using data from women in the Kaiser Foundation Health Plan in the San Francisco East Bay area to find birthweight influence factors of infants. We are interested in exploring the relationship between birthweight and length of gestation and we hope to predict birthweights using length of gestation.

Using SAS OnDemand, the correction coefficient value is 0.41 (**Table 1**.) which indicates a moderate positive linear correlation between the birthweight and length of gestation in babies. As length of gestation increases, the birthweight increases.

**Table 1.** Correlation coefficient between gestation length and birthweight.

The CORR Procedure

| 2 Variables: | gestation bwt |
|---|---|

Simple Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| gestation | 1223 | 279.33851 | 16.02769 | 341631 | 148.00000 | 353.00000 |
| bwt | 1236 | 119.57686 | 18.23645 | 147797 | 55.00000 | 176.00000 |

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

| | gestation | bwt |
|---|---|---|
| gestation | 1.00000<br>1223 | 0.40785<br><.0001<br>1223 |
| bwt | 0.40785<br><.0001<br>1223 | 1.00000<br>1236 |

```
proc import
datafile = "/home/u59373588/My SAS/week13/babies.csv"
out = babies
DBMS = csv
REPLACE  ;
GETNAMES = yes;
GUESSINGROWS=3;
run;

proc corr data= babies;
 VAR gestation bwt;
run;
```
(Returns **Table 1**.)

```
proc reg data= babies;
MODEL bwt = gestation / clb alpha=0.05;
run;
```
(Returns **Figure 1.** and **Table 2.**)

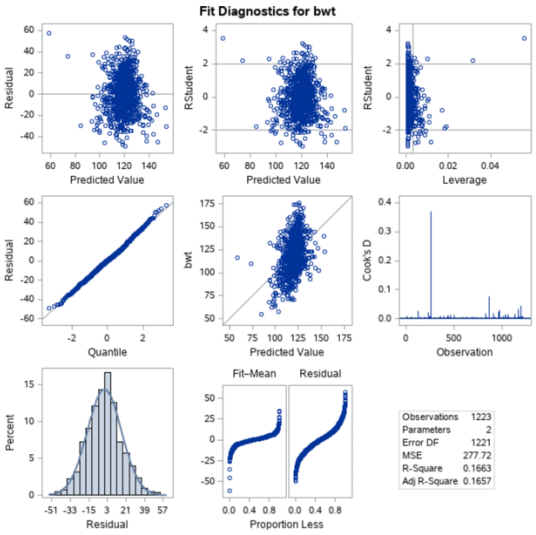


**Figure 1.** Fit diagnostics graphs**.**

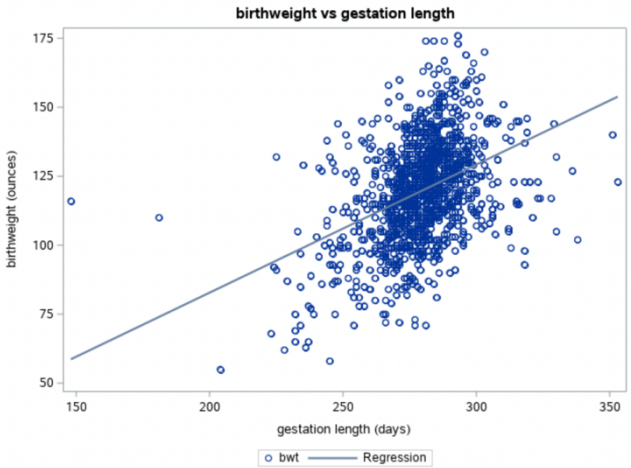


**Figure 2.** Fit plot of birthweight versus gestation length.

**Table 2.** Simple linear regression results between gestation length and birthweight.

The REG Procedure
Model: MODEL1
Dependent Variable: bwt

| | |
|---|---|
| Number of Observations Read | 1236 |
| Number of Observations Used | 1223 |
| Number of Observations with Missing Values | 13 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 67661 | 67661 | 243.63 | <.0001 |
| Error | 1221 | 339092 | 277.71673 | | |
| Corrected Total | 1222 | 406753 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 16.66484 | R-Square | 0.1663 |
| Dependent Mean | 119.62224 | Adj R-Sq | 0.1657 |
| Coeff Var | 13.93122 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | -10.06418 | 8.32220 | -1.21 | 0.2268 | -26.39159 | 6.26322 |
| gestation | 1 | 0.46426 | 0.02974 | 15.61 | <.0001 | 0.40591 | 0.52262 |

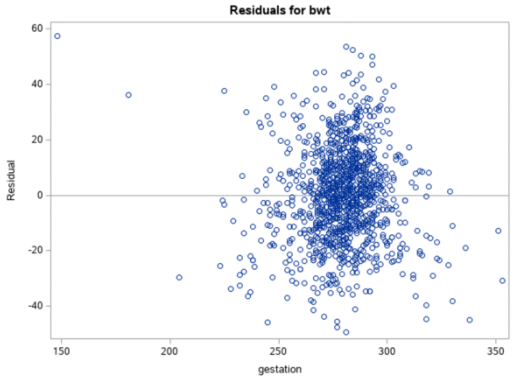


**Figure 3.** Residuals plot of birthweight versus gestation length.

```
proc sgplot data = babies;
SCATTER x = gestation y = bwt;
REG x = gestation y = bwt;
XAXIS label = "gestation length (days)";
YAXIS label = "birthweight (ounces)";
TITLE "birthweight vs gestation length";
run;
```
(Returns **Figure 2**.)

According to **Table 2**., birthweight = 0.46*gestation length -10.06, which -10.06 is the intercept, and 0.46 is the slope. The positive slope further confirms the positive linear relationship between gestation length and birthweight. For every increase of gestation length (a day), our model predicts an increase of birthweight of 0.46 ounce in the gestation length between 148 to 353 days. From **Figure 2.,** a linear, positive (upward trend) with moderate strength relationship can be observed as the points appears to be concentrated around the fitted line and scattered with outliers. Intercept interpretation is meaningless here since it is impossible to give birth babies when gestation length is zero. Since we don't observe any birthweight values close to 0, it's not meaningful. The R square is 0.1663, which indicates that 16.63% of the variation in birthweight is explained by the linear relationship with the gestation length. Other factors may contribute to the other 83.7% of the cases. So, since the model can only be used in a few cases, we should not use it to predict birthweight using gestation length. The results do not provide enough evidence that gestation length causes higher birthweight since the correlation relationship is weak.

From **Figure 1**., we can prove the linear assumption. The scatter plot of the data shows weak positive linear relationship between gestation length and birthweight. When we look at the residual plot, the residual data points are randomly distributed with no obvious trends and models other than linear model can better describe the relationship. Independence can be assumed since participants are not related, and measurements are not repeated. The histogram appears approximately normal and thus the normality assumption is met. The equal variance assumption is met since residuals have constant variability above and below the horizontal line at 0 as we move across the x axis. Based on the four assumptions, it is reasonable to use the inferential results from the model.

The hypothesis H0 is the slope between the birthweight and gestation length is zero (beta 1= 0) and H1 is the slope between the birthweight and gestation length is not zero. From **Table 2**., the t value is 15.61 and the p value is less than 0.0001, which is less than the standard significance level when alpha equals to 0.05. Then, we can reject the null hypothesis that the slope is zero. There is evidence that there is a non-zero slope between the birthweight and gestation length. According to **Table 2.**, we found the 95% confidence interval for the slope is between 0.41 and 0.52. We are 95% confident that with increase of each day of gestation, the birthweight increases on average between 0.41 and 0.52 ounce. Thus, birthweight increases as gestation length increases since the slope is positive.

To conclude, the birthweight (in ounces) and length of gestation (in days) in babies are directly proportional. As birthweight increases as gestation length increases. We can use the equation birthweight = 0.46*gestation length -10.06 to describe the relationship. 16.63% of the variation in birthweight is explained by the linear relationship with the gestation length. Other factors may contribute to the other 83.7% of the cases. So, the model can only explain and predict 16.63% of the cases, which is not through enough to make precise prediction. We may need pregnant women's previous health conditions, diet, daily routine, and other information that influence the birthweight to predict the birthweight of babies.