```yaml
Course: Regression_Models
Lesson: Residual_Variation


- Class: text
  Output: "Residual Variation. (Slides for this and other Data Science courses may be found at github https://github.com/DataScienceSpecialization/courses. If you care to use them, they must be downloaded as a zip file and viewed locally. This lesson corresponds to Regression_Models/01_06_residualVariation. Galton data is from John Verzani's website, http://wiener.math.csi.cuny.edu/UsingR/)"

- Class: text
  Output: As shown in the slides, residuals are useful for indicating how well data points fit a statistical model. They "can be thought of as the outcome (Y) with the linear association of the predictor (X) removed. One differentiates residual variation (variation after removing the predictor) from systematic variation (variation explained by the regression model)."

- Class: text
  Output: It can also be shown that, given a model, the maximum likelihood estimate of the variance of the random error is the average squared residual. However,  since our linear model with one predictor requires two parameters we have only (n-2) degrees of freedom. Therefore, to calculate an "average" squared residual to estimate the variance we use the formula 1/(n-2) * (the sum of the squared residuals). If we divided the sum of the squared residuals by n, instead of n-2, the result would give a biased estimate.

- Class: cmd_question
  Output: To see this we'll use our favorite Galton height data. First regenerate the regression line and call it fit. Use the R function lm and recall that by default its first argument is a formula such as "child ~ parent" and its second is the dataset, in this case galton.
  CorrectAnswer: fit <- lm(child ~ parent, galton)
  AnswerTests: omnitest(correctExpr='fit <- lm(child ~ parent, galton)')
  Hint: Type "fit <- lm(child ~ parent, galton)" at the R prompt.

- Class: text
  Output: First, we'll use the residuals (fit$residuals) of our model to estimate the standard deviation (sigma) of the error. We've already defined n for you as the number of points in Galton's dataset (928).

- Class: cmd_question
  Output:  Calculate the sum of the squared residuals divided by the quantity (n-2). Then take the square root.
  CorrectAnswer: sqrt(sum(fit$residuals^2) / (n - 2))
  AnswerTests: omnitest(correctExpr='sqrt(sum(fit$residuals^2) / (n - 2))')
  Hint: Type "sqrt(sum(fit$residuals^2) / (n - 2))" at the R prompt.


- Class: cmd_question
  Output: Now look at the "sigma" portion of the summary of fit, "summary(fit)$sigma".
  CorrectAnswer: summary(fit)$sigma
  AnswerTests: omnitest(correctExpr='summary(fit)$sigma')
  Hint: Type "summary(fit)$sigma" at the R prompt.

- Class: text
  Output: Pretty cool,  huh?

- Class: cmd_question
  Output: Another cool thing - take the sqrt of  "deviance(fit)/(n-2)" at the R prompt.
  CorrectAnswer: sqrt(deviance(fit)/(n-2))
  AnswerTests: omnitest(correctExpr='sqrt(deviance(fit)/(n-2))')
  Hint: Type "sqrt(deviance(fit)/(n-2))" at the R prompt.

- Class: text
  Output: Another useful fact shown in the slides was

- Class: text
  Output: Total Variation = Residual Variation + Regression Variation
```

```yaml

- Class: mult_question
  Output: Recall the beauty of the  slide full of algebra which proved this fact. It had a bunch of Y's, some with hats and some with bars and several summations of squared values. The Y's with hats were the estimates provided by the model. (They were on the regression line.) The Y with the bar was the mean or average of the data. Which sum of squared term represented Total Variation?
  AnswerChoices: Yi-mean(Yi); Yi-Yi_hat; Yi_hat-mean(Yi)
  CorrectAnswer: Yi-mean(Yi)
  AnswerTests: omnitest(correctVal='Yi-mean(Yi)')
  Hint: Pick the choice which is independent of the estimated or predicted values, the (hat terms).

- Class: mult_question
  Output: Which sum of squared term represents Residual Variation?
  AnswerChoices:  Yi-Yi_hat; Yi-mean(Yi); Yi_hat-mean(Yi)
  CorrectAnswer: Yi-Yi_hat
  AnswerTests: omnitest(correctVal='Yi-Yi_hat')
  Hint: Residuals represent the vertical distance between actual values and estimated (hat) values.

- Class: text
  Output: The term R^2 represents the percent of total variation described by the model, the regression variation (the term we didn't ask about in the preceding multiple choice questions). Also, since it is a percent we need a ratio or fraction of sums of squares. Let's do this now for our Galton data.

- Class: cmd_question
  Output: We'll start with easy steps. Calculate the mean of the children's heights and store it in a variable called mu. Recall that we reference the childrens' heights with the expression 'galton$child' and the parents' heights with the expression 'galton$parent'.
  CorrectAnswer: mu <- mean(galton$child)
  AnswerTests: omnitest(correctExpr='mu <- mean(galton$child)')
  Hint: Type "mu <- mean(galton$child)" at the R prompt.

- Class: cmd_question
  Output: Recall that centering data means subtracting the mean from each data point. Now calculate the sum of the squares of the centered children's heights  and store the result in a variable called sTot. This represents the Total Variation of the data.
  CorrectAnswer: sTot <- sum((galton$child-mu)^2)
  AnswerTests: ANY_of_exprs('sTot <- sum((galton$child-mu)^2)','sTot <- sum((galton$child-mu)*(galton$child-mu))')
  Hint: Type "sTot <- sum((galton$child-mu)^2)" at the R prompt.

- Class: cmd_question
  Output: Now create the variable sRes. Use the R function deviance to calculate the sum of the squares of the residuals. These are the distances between the children's heights and the regression line. This represents the Residual Variation.
  CorrectAnswer: sRes <- deviance(fit)
  AnswerTests: omnitest(correctExpr='sRes <- deviance(fit)')
  Hint: Type "sRes <- deviance(fit)" at the R prompt.

- Class: cmd_question
  Output: Finally, the ratio sRes/sTot represents the percent of total variation contributed by the residuals. To find the percent contributed by the model, i.e., the regression variation,  subtract the fraction sRes/sTot from 1.  This is the value R^2.
  CorrectAnswer: 1-sRes/sTot
  AnswerTests: omnitest(correctExpr='1-sRes/sTot')
  Hint: Type "1-sRes/sTot" at the R prompt.

- Class: cmd_question
  Output: For fun you can compare your result to the values shown in summary(fit)$r.squared to see if it looks familiar. Do this now.
  CorrectAnswer: summary(fit)$r.squared
  AnswerTests: omnitest(correctExpr='summary(fit)$r.squared')
  Hint: Type "summary(fit)$r.squared" at the R prompt.

- Class: cmd_question
```

```yaml
  Output: To see some real magic, compute the square of the correlation of the galton data, the children and parents. Use the R function cor.
  CorrectAnswer: cor(galton$parent,galton$child)^2
  AnswerTests: ANY_of_exprs('cor(galton$parent,galton$child)^2','cor(galton$child,galton$parent)^2')
  Hint: Type "cor(galton$parent,galton$child)^2" at the R prompt.


- Class: text
  Output: We'll now summarize useful facts about R^2. It is the percentage of variation explained by the regression model. As a percentage it is between 0 and 1. It also equals the sample correlation squared. However, R^2 doesn't tell the whole story.

- Class: text
  Output: Congrats! You've finished this lesson on Residual Variation.

- Class: mult_question
  Output: "Would you like to receive credit for completing this course on
    Coursera.org?"
  CorrectAnswer: NULL
  AnswerChoices: Yes;No
  AnswerTests: coursera_on_demand()
  Hint: ""
```