Exploration librairie swirl : Residual Variation

> swirl()

Welcome to swirl! Please sign in. If you've been here before, use the same name as you did then. If you are new, call yourself something unique.

What shall I call you? ***jlbellier***

Would you like to continue with one of these lessons?

1: Regression Models Least Squares Estimation
2: No. Let me start something new.

Selection: ***2***

Please choose a course, or type 0 to exit swirl.

1: Regression Models
2: Statistical Inference
3: Take me to the swirl course repository!

Selection: ***1***

Please choose a lesson, or type 0 to return to course menu.

1: Introduction
2: Residuals
3: Least Squares Estimation
4: Residual Variation
5: Introduction to Multivariable Regression
6: MultiVar Examples
7: MultiVar Examples2
8: MultiVar Examples3
9: Residuals Diagnostics and Variation
10: Variance Inflation Factors
11: Overfitting and Underfitting
12: Binary Outcomes
13: Count Outcomes

Selection: ***4***

0%

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/)

==== ***5%***

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)."

========= **9%**

*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.*

========== **14%**

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.

*> **fit <- lm(child ~parent, data=galton)***

You are doing so well!

========== **18%**

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).

========== **23%**

Calculate the sum of the squared residuals divided by the quantity (n-2). Then take the square root.

*> **sqrt(sum((fit$residuals)^2)/(n-2))***
[1] 2.238547

Not quite right, but keep trying. Or, type info() for more options.
Type "sqrt(sum(fit$residuals^2) / (n - 2))" at the R prompt.

*> **sqrt(sum(fit$residuals^2)/(n-2))***
[1] 2.238547

That's a job well done!

========== **27%**

Now look at the "sigma" portion of the summary of fit, "summary(fit)$sigma".

*> **summary(fit)$sigma***
[1] 2.238547

You are really on a roll!

========== **32%**

Pretty cool, huh?

========== **36%**

Another cool thing - take the sqrt of "deviance(fit)/(n-2)" at the R prompt.

*> **sqrt(deviance(fit)/(n-2))***
[1] 2.238547

Excellent work!

========== **41%**

Another useful fact shown in the slides was

========== **45%**

Total Variation = Residual Variation + Regression Variation

========== **50%**

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?

1: Yi_hat-mean(Yi)
2: Yi-mean(Yi)
3: Yi-Yi_hat

*Selection: **3***

That's not exactly what I'm looking for. Try again.
Pick the choice which is independent of the estimated or predicted values, the (hat terms).

1: Yi-Yi_hat
2: Yi-mean(Yi)
3: Yi_hat-mean(Yi)

*Selection: **2***

Keep working like that and you'll get there!

========== **55%**

Which sum of squared term represents Residual Variation?

1: Yi_hat-mean(Yi)
2: Yi-Yi_hat
3: Yi-mean(Yi)

*Selection: **1***

Try again. Getting it right on the first try is boring anyway!
Residuals represent the vertical distance between actual values and estimated (hat) values.

1: Yi-mean(Yi)
2: Yi-Yi_hat
3: Yi_hat-mean(Yi)

*Selection: **2***

That's correct!
========== ***59%***
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.
========== ***64%***
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'.

*> **mu <- mean(galton$child)***

That's a job well done!
========== ***68%***
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.

*> **sTot <- sum((galton$child-mu)^2)***

You got it!
========== ***73%***
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.

*> **sRes <- deviance(fit)***

All that hard work is paying off!
========== ***77%***
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.

*> **1-sRes/sTot***
[1] 0.2104629

You are really on a roll!
========== ***82%***

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.

*> **summary(fit)$r.squared***
[1] 0.2104629

That's correct!
========== 86%
To see some real magic, compute the square of the correlation of the galton data, the children and parents. Use the R function cor.

*> **cor(galton$child,galton$parent)^2***
[1] 0.2104629

That's the answer I was looking for.
========== 91%
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.
==========95%
Congrats! You've finished this lesson on Residual Variation.
========== 100%