Regression Models (Part 5: Residuals)

A. Estimating residual variation

We've talked at length about how to estimate $\beta_0 and \beta_1$. However, there's another parameter in our model, σ. Recall that our model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

It seems natural to use our residual variation to estimate population error variation. In fact, the maximum likelihood estimate of σ^2 is $\frac{1}{n}\sum_{i=1}^n e_i^2$, the average squared residual. Since the residuals have a zero mean (if an intercept is included), this is close to the the calculating the variance of the residuals. However, to obtain unbiasedness, most people use

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n e_i^2.$$

The $n-2$ instead of n is so that $E[\hat{\sigma}^2] = \sigma^2$. This is exactly analogous to dividing by $n-1$ in the ordinary variance calculation. In fact, the ordinary variance (using `var` in R on a vector) is exactly the same as the residual variance estimate from a model that has an intercept and no slope. The $n-2$ instead of $n-1$ when we include a slope can be thought of as losing a degree of freedom from having to estimate an extra parameter (the slope).

Most of this is typically opaque to the user, since we just grab the correct residual variance output from `lm`. But, to solidify the concepts, let's go through the diamond example to make sure that we can hard code the estiamtes on our own. (And from then on we'll just use lm.)

B. Diamond example

```
library(UsingR)
y <- diamond$price; x <- diamond$carat; n <- length(y)
fit <- lm(y ~ x)
## the estimate from lm
summary(fit)$sigma
## [1] 31.84052
## directly calculating from the residuals
sqrt(sum(resid(fit)^2) / (n - 2))
## [1] 31.84052
```

C. Summarizing variation

A way to think about regression is in the decomposition of variability of our response. The total variability in our response is the variability around an intercept. This is also the variance estimate from a model with only an intercept:

$$\text{Total variability} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

The regression variability is the variability that is explained by adding the predictor. Mathematically, this is:

$$\text{Regression variabilty} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2.$$

The residual variability is what's leftover around the regression line

$$\text{Residual variability} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

It's a nice fact that the error and regression variability add up to the total variability:

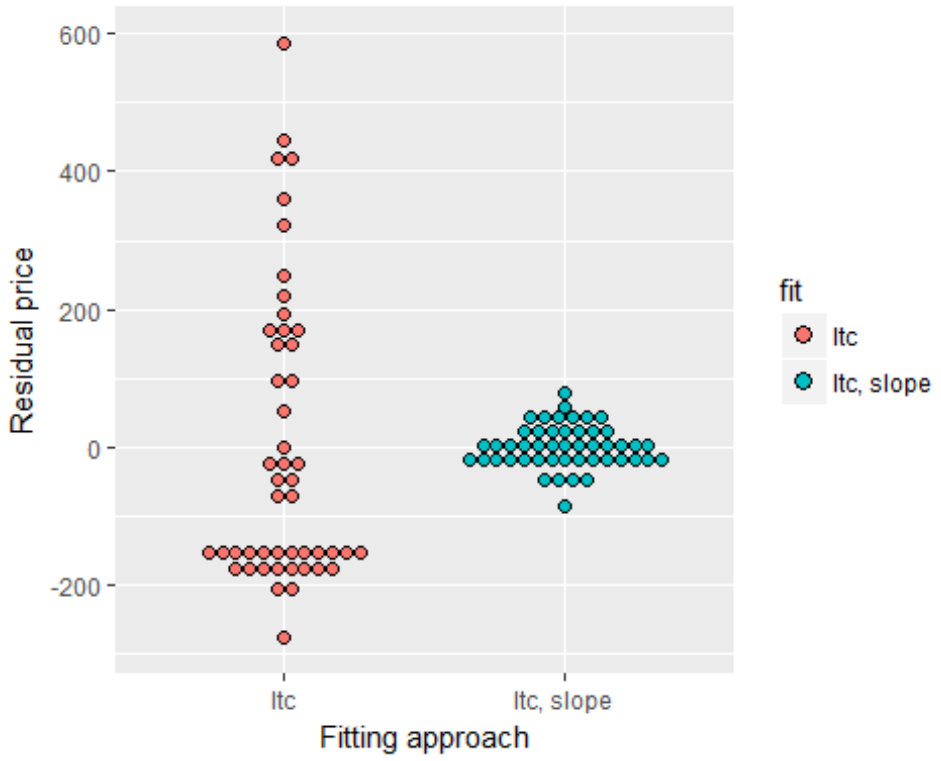
$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

Thus, we can think of regression as explaining away variability. The fact that all of the quantities are positive and that they add up this way allows us to define the proportion of the total variability explained by the model.

Consider our diamond example again. The plot below shows the variation explained by a model with an intercept only (representing total variation) and that when the mass is included as a linear predictor. Notice how much the variation decreases when including the diamond mass.

Here's the code:

```
e = c(resid(lm(price ~ 1, data = diamond)),
      resid(lm(price ~ carat, data = diamond)))
fit = factor(c(rep("Itc", nrow(diamond)),
               rep("Itc, slope", nrow(diamond))))
g = ggplot(data.frame(e = e, fit = fit), aes(y = e, x = fit, fill = fit))
g = g + geom_dotplot(binaxis = "y", size = 2, stackdir = "center", binwidth =
20)
## Warning: Ignoring unknown parameters: size
g = g + xlab("Fitting approach")
g = g + ylab("Residual price")
g
```



D. R squared

R squared is the percentage of the total variability that is explained by the linear relationship with the predictor

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Here are some summary notes about R squared.

R^2 is the percentage of variation explained by the regression model.

$$0 \leq R^2 \leq 1$$

R^2 is the sample correlation squared R^2 can be a misleading summary of model fit. Deleting data can inflate it. (For later.) Adding terms to a regression model always increases R^2.

Anscombe's residual (named after their inventor) are a famous example of our R squared doesn't tell the whole story about model fit. In this example, four data sets have equivalent R squared values and beta values, but dramatically different model fits. The result is to suggest that reducing data to a single number, be it R squared, a test statistic or a P-value, often masks important aspects of the data. The code is quite simple: `data(anscombe);example(anscombe)`.