2 · *Simple statistical models*

Due Monday, February 2, 2015

(1) Brain weight versus body weight

Revisit the data set `mammalsleep` from the R package `faraway` (which you'll have to install the first time you use it). We looked at sleeping patterns for all these critters in the examples from the course packet. We also looked at the relationship between brain weight and body weight in class. These are the first two columns of the data set, labeled "body" and "brain," and measured in kilograms and grams, respectively.

Fit a linear model to these data on an appropriate transformed scale, and answer the following two questions.

1. Extract the coefficients from this linear model, and re-express the estimated model on the original scale. How should one interpret the parameters of your model in terms of the original body weight and brain weight variables?

2. Which mammals have the largest and smallest brains, in an absolute sense? Which ones have the largest and smallest brains, adjusting for body size?

(2) The dangers of engine emissions and naïve polynomial regression

Imagine a combustion engine that burns ethanol. A perfectly balanced fuel-air mixture involves exactly as much oxygen as is required to burn a given volume of fuel: $C_2H_5OH + 3O_2 \longrightarrow 2CO_2 + 3H_2O$. But in practice this is rarely achieved, or even desirable from the standpoint of engine design—and the gas in the mixture is not pure oxygen anyway, but air. As a result, some of the ethanol reacts with nitrogen, yielding nitrogen oxides (NO and NO_2) as emission byproducts.

Load the data set in "ethanol.csv." This summarizes an experiment where an ethanol-based fuel was burned in a one-cylinder combustion engine. The experimenter varied the engine compression (C) and the equivalence ratio (E), which measures the richness of the fuel-air mixture in the combustion chamber. For each setting of C and E, the emissions of nitrogen oxides (NOx) were recorded.

1. Fit a polynomial regression model for NOx emissions versus the equivalence ratio. Choose a sensible order of the polynomial (quadratic, cubic, etc) by eye, superimposing plots of fitted values onto the original data to guide your decision. Write a short

summary of your process and findings. Discuss any shortcomings you see with your final model.

2. There are 88 observations in the ethanol data set. An 88th degree polynomial fit to the data could pass through every single one—perfect fit! Yet presumably you didn't choose an 88th-degree polynomial just now. Why not?

3. Plot C vs E (the two variables under experimental control). What do you notice? Briefly share any insight you may have on why the experimenter chose this pattern of C–E combinations at which to measure the NOx emissions.

(3) Two different inferential principles for linear regression

From the notes, you will recall reading that the least-squares estimate for a simple regression line is $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, where

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \,. \end{aligned}$$

We didn't dwell on this in the notes, because nowadays computers do all the calculations for us. But it's worth remembering how Legendre arrived at these formulas in the first place: he defined the sum-of-squares *objective function*,

$$g(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \{\beta_0 + \beta_1 x_i\})^2 \,,$$

which is a function of two variables. (Here "objective" means something like "target" or "goal," and not the opposite of "subjective.") The objective function is summing up the squared residuals from the regression line with parameters (β_0, β_1). The worse the fit, the bigger the residuals, and the bigger the value of the objective function.

Legendre's inferential principle—least squares—was to choose the parameters so as to make this objective function as small as possible. Remember that a function of more than one variable obtains a minimum when all of its partial derivatives are equal to zero.[1] Applying this fact to get a solution would involve three steps:

1. First, we'd take the partial derivative of $g(\beta_0, \beta_1)$ with respect to β_0, and set the resulting expression equal to zero.

2. Second, we'd take the partial derivative of $g(\beta_0, \beta_1)$ with respect to β_1, and set this second expression equal to zero.

[1] This could mean a maximum or other form of stationary point, too—so yes, technically you'd also need to check the second derivative.

3. Steps 1 and 2 give a system of two equations in two variables. Solve the system for β_0 and β_1 to give the least-squares solution.

You are encouraged, but not required, to try this process on your own to verify the expressions above.

But that's not the main point of this exercise. Instead, you will derive an optimal linear fit to the data using a different inferential principle: *the method of moments*. Rather than forcing the partial derivatives of the above objective function to be zero, you will enforce the following two constraints:

(1) That the sample mean of the residuals $(y_i - \hat{y}_i)$ is zero, so that the line passes "on average" through the middle of the point cloud. To see the intuition here: imagine if the average residual weren't zero. Then you could systematically move the line up or down to get a better fit!

(2) That the sample correlation between the residuals $(y_i - \hat{y}_i)$ and the original predictor variable (x_i) is zero, so that the line "takes the X-ness out of Y."

This is completely different, but entirely sensible, inferential principle for fitting straight lines to data.[2] Notice there's nothing about squared errors here at all, and no calculus. But the same general idea applies: each of these two constraints above implies an equation that the method-of-moments estimator for β_0 and β_1 must satisfy.

[2] Quantities like sample means, sample variances, and sample correlations are called sample moments, in the sense that they are quantitative measures of the shape of a set of points—hence the name "method of moments."

Write these constraints out. You will be left with two equations and two unknowns . . . your job is to derive the solution implied by this system of equations. Make sure to show your work. Once you have derived the solution, compare and contrast the least-squares estimator and the method-of-moments estimator. How similar are they?

Hint: If you find this tough going, resort to the mathemetician's favorite trick of trying a special case of the general problem: assume that $\bar{y}$ and $\bar{x}$ are both zero. (This would be the case, for example, if you defined new predictor and response variables by subtracting the sample means from the original ones. This is called "centering," and is a common thing to do.) At this point, one of the constraints will be trivially satisfied, and the other will involve simpler algebra than you had before. Once you see how this works, try the more general case where $\bar{x}$ and $\bar{y}$ are nonzero.