Homework Solution

Problem 1 and 2

Most of you have no problem getting the scatter plots, the regression equations and residual plots . If you use "Length" to predict "Weight", then your first regression should have

$$\text{Weight} = -483.7 + 35.8 \text{ Length}$$

Your second regression should have

$$log(\text{Weight}) = 5.11 + 0.04 \text{ Length}$$

Pay attentions to the following when you analyze the residuals:

1. Some pointed out that the residuals clustered at large values of the Length. This **has nothing to do** with the regression itself. It is something that happened in data collecting: we caught more fish of that length, so we would surely have points clustering at those lengths. This is NOT a problem.

2. The problem with the first residual plot is that the variance seems to increase with Length. We assume the residual is $N(0, \sigma^2)$ distributed, its variance σ^2 being a constant. But in this case we have a non-constant variance.

3. Some argued that the 2nd model is better **because its residuals are smaller than in the 1st**. In fact, you can't compare residuals from two different models by looking at how big they are. The scales are different, say in this problem the 2nd model is the 1st one taken log, its residuals would definitely be smaller even if it's worse than the 1st. A reasonable thing to look at would be: R-squared, which is$1-$(the proportion of Residual Sum of Squares in Total Sum of Squares). We look for a model with less proportion of residuals, that is, a larger R-squared.

Problem 3

In order to compare the DDT level in Catfish and Non-Catfish, we have to include both kinds in one model. Doing a seperate regression for each kind does NOT help in comparing. So we use a Qualitative variable CF in our regression to see the effect of species in DDT level: CF=1 for Catfish and CF=0 for others. So your model should be something like:

$$\text{DDT} = \beta_0 + \beta_1 \text{CF} + \beta_2 \text{ Weight } + \beta_3 \text{ Length}$$

(you can choose not to include Weight or Length if you like, but you must include CF)

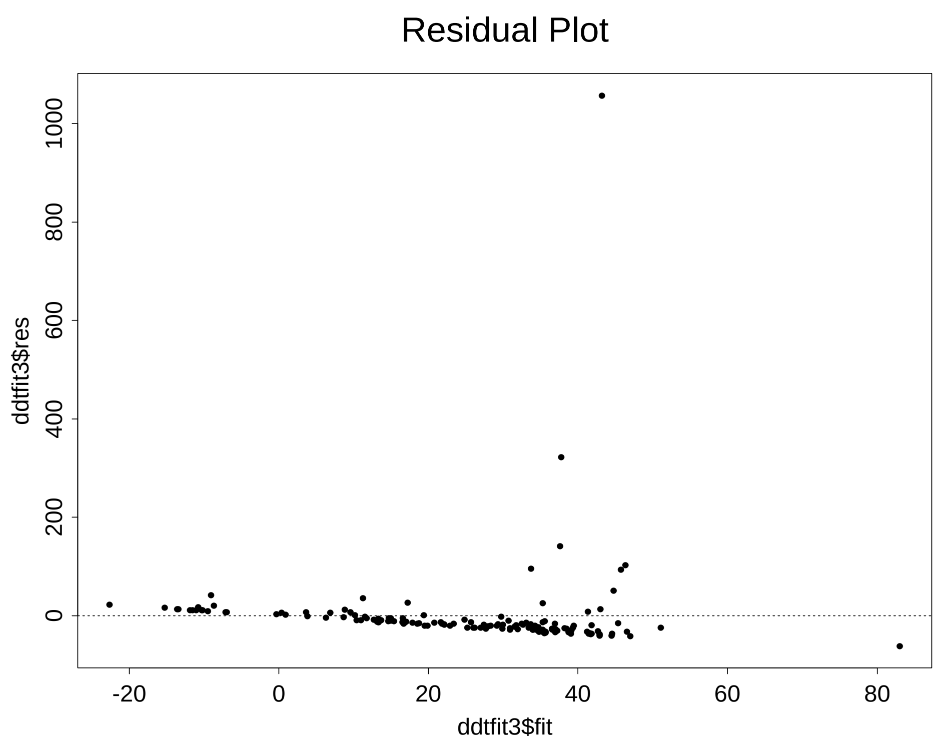
The interpretation: if $\beta_1$turns out to be 5, then for a catfish and a non-catfish **of the same length and weight**, the catfish's DDT will on average beat that of the non-catfish by 5.

Note: I bold-font "of the same length and weight" because many of you said Catfish has a higher DDT level. That is an illusion: Catfish have a bigger size on average, so before adjusted for Weight and Length, they have a higher DDT level.

And in fact β_1 is negative here, but is insignificant, so your conclusion should be: Species doesn't make a difference. (If you need an explanation, refer to the quiz solution at the end of Homework 9 solutions.)

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-74.9632	60.0001	-1.2494	0.2136
CF	-1.2289	26.7927	-0.0459	0.9635
Weight	-0.0445	0.0409	-1.0884	0.2783
Length	3.4305	2.3678	1.4488	0.1496

Few have mentioned the unusually low R-squared value 3.153% for this regression. The reason is: most of the fish has a DDT < 50 regardless of size, while there are a bunch of big outliers. The regression line is mostly brought up by these outliers, so the majority of fish are off-line(see the plot on next page). That gave us a big Residual Sum of Squares, thus a small R-squared.


Residual Plot
ddtfit3$res
ddtfit3$fit
0
200
400
600
800
1000
-20
0
20
40
60
80