

Assignment 5

Q1. A very old problem in regression is that of predicting the height of the son from the height of the father. The available [data](#) summarizes the raw information. The father's height has been rounded to the nearest inch and the average height of the son for fathers of that height is given. The number of fathers in each category is given.

- i. Construct a linear regression model for predicting the height of the son from the height of the father in the best manner given information available.
- ii. Can the model be simplified to

$$\text{height of son} = \text{height of father} + \text{error?}$$

Carry out the appropriate test.

Q2. Researchers at NIST collected [data](#) on ultrasonic measurements of the depths of defects in the Alaska pipeline in the field. The depth of the defects were then re-measured in the laboratory. These measurements were performed in six different batches. It turns out that this batch effect is not significant and so can be ignored in the analysis that follows. The laboratory measurements are more accurate than the in field measurements, but more time consuming and expensive. We want to develop an regression equation for correcting the in field measurements.

- i. Fit a regression model $\text{Lab} \sim \text{Field}$. Check for non-constant variance.
- ii. We wish to use weights to account for the non-constant variance. Here we split the range of `Field` into 12 groups of size 9 (except for the last group which has only 8 values). Within each group, we compute the variance of `Lab` as `varlab` and the mean of `Field` as `meanfield`. Suppose `pipe` is the name of your dataframe, the following R code will make the need computations:

```
> i <- order(pipe$Field)
> npipe <- pipe[i,]
> ff <- gl(12,9)[-108]
> meanfield <- unlist(lapply(split(npipe$Field,ff),mean))
> varlab <- unlist(lapply(split(npipe$Lab,ff),var))
```

Suppose we guess that the the variance in the response is linked to the predictor in the following way:

$$\text{var}(\text{Lab}) = a_0 \text{Field}^{a_1}$$

Regress $\log(\text{varlab})$ on $\log(\text{meanfield})$ to estimate a_0 and a_1 . (You might choose to remove the last point). Use this to determine appropriate weights in a WLS fit of `Lab` on `Field`. Show the regression summary.

Q3. [Data](#) on the outside diameter of crankpins produced by an industrial process over several days are given. All of the crankpins should be between 0.7425 and 0.7430 inches. The numbers given in the table are in units of 0.00001 inches deviation from 0.742 inches.

When the manufacturing process is "under control", the average size of the crankpins produced should (1) fall near the middle of the specified range and (2) should not depend on time. Fit an appropriate model to see if the process is under control and test for lack of fit in the model.