DATA621 LMR Ex 8.1

Chun Yip

2022/4/10

R Markdown

Researchers at National Institutes of Standards and Technology (NIST) collected pipeline data on ultrasonic measurements of the depth of defects in the Alaska pipeline in the field. The depth of the defects were then remeasured 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 a regression equation for correcting the in-field measurements.

(a) Fit a regression model Lab ~ Field. Check for non-constant variance.

```
data(pipeline, package="faraway")
head(pipeline)
```

```
##   Field  Lab Batch
## 1    18 20.2     1
## 2    38 56.0     1
## 3    15 12.5     1
## 4    20 21.2     1
## 5    18 15.5     1
## 6    36 39.0     1
```

```
summary(pipeline)
```

```
##      Field            Lab         Batch
##  Min.   : 5.00   Min.   : 4.30   1:19
##  1st Qu.:18.00   1st Qu.:18.35   2:20
##  Median :35.00   Median :38.00   3:20
##  Mean   :33.58   Mean   :39.10   4:20
##  3rd Qu.:46.50   3rd Qu.:55.55   5:21
##  Max.   :85.00   Max.   :81.90   6: 7
```

```
lmod <- lm(Lab ~ Field, pipeline)
summary(lmod)
```

```
##
## Call:
## lm(formula = Lab ~ Field, data = pipeline)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.985  -4.072  -1.431   2.504  24.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96750    1.57479  -1.249    0.214
## Field        1.22297    0.04107  29.778   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.865 on 105 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8931
## F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```

```r
glmod <- gls(Lab ~ Field, correlation=corAR1(), pipeline)
summary(glmod)
```

```
## Generalized least squares fit by REML
##   Model: Lab ~ Field
##   Data: pipeline
##       AIC      BIC    logLik
##   753.319 763.9349 -372.6595
##
## Correlation Structure: AR(1)
##  Formula: ~1
##  Parameter estimate(s):
##        Phi
## 0.09443647
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept) -2.023771 1.5794450 -1.281318  0.2029
## Field        1.224530 0.0399197 30.674863  0.0000
##
##  Correlation:
##       (Intr)
## Field -0.849
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.8032658 -0.5161639 -0.1772741  0.3175650  3.0902616
##
## Residual standard error: 7.869855
## Degrees of freedom: 107 total; 105 residual
```

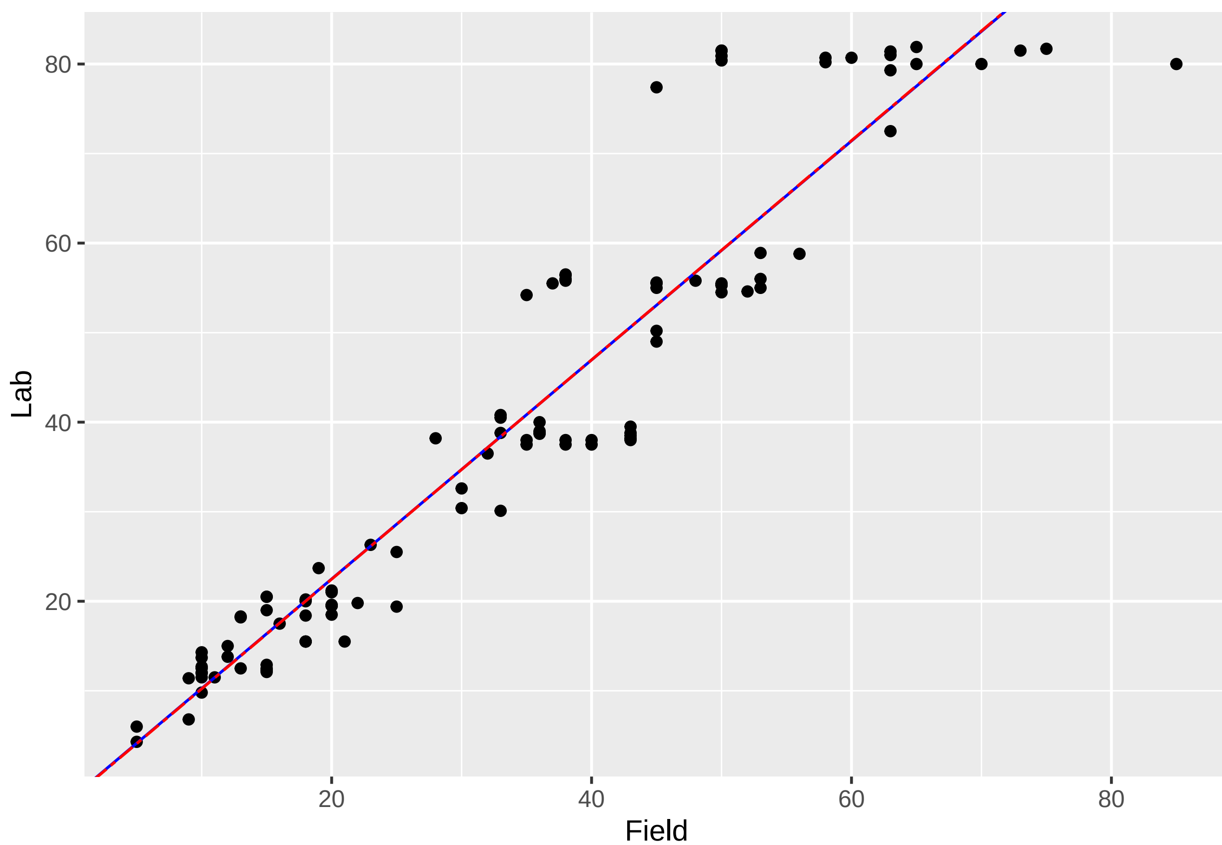
```r
intervals(glmod, which="var-cov")
```

```
## Approximate 95% confidence intervals
##
##  Correlation structure:
##           lower       est.     upper
## Phi -0.09689466 0.09443647 0.2790366
```

```
## attr(,"label")
## [1] "Correlation structure:"
##
##  Residual standard error:
##    lower     est.    upper
## 6.864801 7.869855 9.022057
```

The Phi is 0.0944 and the interval is across 0. It's possible that there is no non-constant variance.

```
ggplot(pipeline, aes(x=Field, y=Lab))+
  geom_point()+
  geom_abline(intercept=lmod$coefficients[1], slope = lmod$coefficients[2], col="Blue")+
  geom_abline(intercept=glmod$coefficients[1], slope = glmod$coefficients[2], col="Red", linetype="twod
```



However, the scatterplot shows the variance increases with Field. There are non-constant variance. Both linear regression and the general linear regression provide the same line.

We wish to use weights to account for the non-constant variance. Here we split the range of Field into 12 groups of size nine (except for the last group which has only eight values). Within each group, we compute the variance of Lab as varlab and the mean of Field as meanfield. Supposing pipeline is the name of your data frame, the following R code will make the needed computations

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

```r
lmodweights <- lm(log(varlab)~log(meanfield))
summary(lmodweights)
```

```
##
## Call:
## lm(formula = log(varlab) ~ log(meanfield))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2038 -0.6729  0.1656  0.7205  1.1891
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.3538     1.5715  -0.225   0.8264
## log(meanfield)   1.1244     0.4617   2.435   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.018 on 10 degrees of freedom
## Multiple R-squared:  0.3723, Adjusted R-squared:  0.3095
## F-statistic: 5.931 on 1 and 10 DF,  p-value: 0.03513
```

```r
a1<-exp(lmodweights$coefficients[1])
a0<-exp(lmodweights$coefficients[2])
```

a0 is 3.08 whilst a1 is 0.702.