Homework 03 Key. PLS206 Fall 2013

Emilio A. Laca

November 28, 2014

1 Model development

1.1 Write a model to describe the lnwt of clover plants as a function of plant age (days) and temperature as a continuous variable.

Assume that temperature treatments were 10, 15 and 20 C for treatments 1-3. Make sure you incorporate terms that will allow you to estimate RGR for different temperatures, even if they were not in the treatments. Define all symbols used in the model except for the usual mathematical operators. [20]

$$lnwt = \beta_0 + \beta_1 * days + \beta_2 * temperature + \beta_{12} * days * temperature + \epsilon \quad (1)$$

This model is good, but it imposes the restriction that the temperature affects RGR in a linear manner. For any given temperature T, the RGR is:

$$\begin{aligned} lnwt &= \beta_0 + \beta_1 * days + \beta_2 * T + \beta_{12} * days * T + \epsilon && (2) \\ &= (\beta_0 + \beta_2.T) + (\beta_1 + \beta_{12} * T) * days + \epsilon && (3) \end{aligned}$$

Note that the slope is now a linear function of *temperature*.

Considering that it is likely that RGR first increases and then decreases with increasing temperature, we can improve the model by including a quadratic temperature term that affects only the coefficient of *days*. Later, we could remove the possibility that *temperature* affects the intercept, as temperature should not affect the weight of the plants before the treatments are applied.

$$lnwt = \beta_0 + \beta_2 * T + (\beta_1 + \beta_{12} * T + \beta_{13} * T^2) * days + \epsilon \quad (4)$$

$$= \beta_0 + \beta_1 * days + \beta_2 * T + \beta_{12} * T * days + \beta_{13} * T^2 * days + \epsilon \qquad (5)$$

1.2 Run a linear model in R to get estimated parameters for your model.

Use the principle of the extra sum of squares to determine if it is necessary to have the quadratic effect of temperature in the model. [20]

The principle essential states that a term (or set of terms) is significant if there is a significant increase of sum of squares of residuals when the term(s) is(are) removed to obtain a **reduced** model. The F-test for the effect is constructed as follows:

$$F_{(\Delta df, dfe)} = \frac{(SSE_{reducedModel} - SSE_{fullModel})/\Delta df}{MSE_{fullModel}} \qquad (6)$$

The test can be performed ”by hand” or using the Type III sum of squares, or using the `anova()` method.

Performing the calculations ”by hand.” First get the data and run the model.

```
> clover<-read.csv("~/Google Drive/PLS206F13/Examples/clover.csv", header=TRUE)
> options(contrasts =c("contr.sum", "contr.poly"))
> names(clover) <- c("group", "days", "lnwt")
> clover$group<-factor(clover$group)
> # Create a continuous variable for temperature
> clover$t <- NA
> clover$t[clover$group==1] <- 10
> clover$t[clover$group==2] <- 15
> clover$t[clover$group==3] <- 20
> FullModel<-lm(lnwt ~ days + t + days:t + days:I(t^2), clover)
```

Now, calculate the components of the test and then obtain the F-value.

```
> anova(FullModel)
```

```
Analysis of Variance Table

Response: lnwt
            Df Sum Sq Mean Sq  F value    Pr(>F)
days         1 36.553  36.553 1279.728 < 2.2e-16 ***
t            1  9.944   9.944  348.143 < 2.2e-16 ***
days:t       1  1.358   1.358   47.537 2.629e-08 ***
```

```
days:I(t^2)  1  0.007   0.007    0.241    0.6262
Residuals   40  1.143   0.029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The type III sum of squares tests each term as if it were entered last, which is equivalent to compare a full and reduced model that is nested in the full model. The result indicates that it is not necessary to include a quadratic term for temperature affecting the slope with respect to days (i.e., the RGR).

```
> ReducedModel <- lm(lnwt ~ days + t + days:t, clover)
> anova(FullModel, ReducedModel) # same as the test above
```

```
Analysis of Variance Table

Model 1: lnwt ~ days + t + days:t + days:I(t^2)
Model 2: lnwt ~ days + t + days:t
  Res.Df    RSS Df  Sum of Sq     F Pr(>F)
1     40 1.1425
2     41 1.1494 -1 -0.0068833 0.241 0.6262
```

```
> SSE.ReducedModel <- deviance(ReducedModel)
> SSE.FullModel <- deviance(FullModel)
> dfe.ReducedModel <- df.residual(ReducedModel)
> dfe.FullModel <- df.residual(FullModel)
> MSEfull <- SSE.FullModel/dfe.FullModel
> (fstat <- ((SSE.ReducedModel-SSE.FullModel)/(dfe.ReducedModel-dfe.FullModel))/MSEfull)
```

```
[1] 0.2409857
```

```
> print("The p value is")
```

```
[1] "The p value is"
```

```
> 1-pf(fstat,dfe.ReducedModel-dfe.FullModel,dfe.FullModel)
```

```
[1] 0.6261777
```

1.3 Final model

Create a model with only those terms that you think are necessary and obtain estimated parameters. Report the formula, summary, R^2, PRESS and diagnostic plots. [20] The results above indicate that although the conceptual model required a quadratic effect for temperature, the plants were grown in temperatures that did not result in a reduction of the rate at which RGR increased with increasing temperature. We use the ReducedModel.

$$lnwt = \beta_0 + \beta_1 * days + \beta_2 * temperature + \beta_{12} * days * temperature \qquad (7)$$

```
> summary(ReducedModel)


Call:
lm(formula = lnwt ~ days + t + days:t, data = clover)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30364 -0.10129 -0.04016  0.15779  0.27292 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.2415753  0.2303887   5.389 3.18e-06 ***
days        0.0286541  0.0093927   3.051  0.00399 ** 
t           0.0229345  0.0146070   1.570  0.12408    
days:t      0.0041750  0.0005999   6.959 1.88e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.1674 on 41 degrees of freedom
Multiple R-squared:  0.9765,	Adjusted R-squared:  0.9748 
F-statistic:   569 on 3 and 41 DF,  p-value: < 2.2e-16 
```

```
> (PRESS.statistic <- sum( (resid(ReducedModel)/(1-hatvalues(ReducedModel)))^2 ))

[1] 1.395954
```

```
> library(car) # awesome diagnostic plots
> outlierTest(ReducedModel) # Bonferonni p-value for most extreme obs


No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
43 -2.057651           0.046182           NA
```

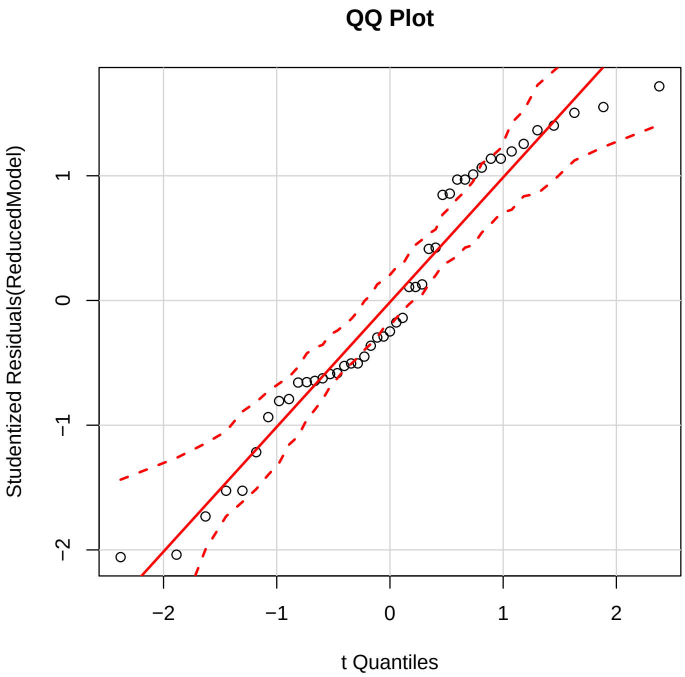

Figure 1: Quantile plot of residuals to test for normality.

2 Model diagnostics

2.1 Interpret the diagnostic plots

Report any problems or potential problems. [20]

```
> qqPlot(ReducedModel, main="QQ Plot") #qq plot for studentized resid
```

The quantile plot shows an s-shaped pattern of points, although few points are aoutside the confidence band. This pattern indicates that there are problems with the normality of the data, if the model is correct. It possible the the pattern results from a misspeified model. For that we look at the residual plots.

```
> residualPlots(ReducedModel)
```

```
           Test stat Pr(>|t|)
days          -5.444    0.000
```

```
           Test stat Pr(>|t|)
days          -5.444    0.000
t             -0.235    0.816
Tukey test    -4.682    0.000
```

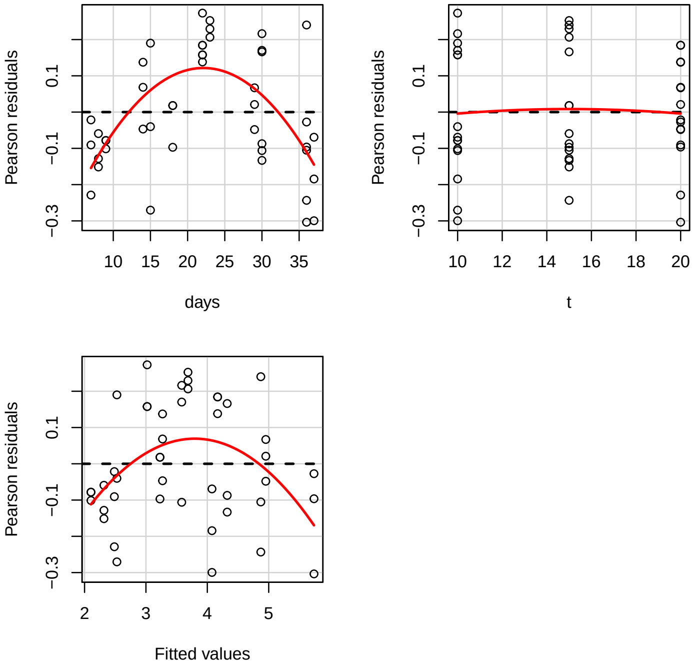

Figure 2: Diagnostic residual plots.

```
t             -0.235    0.816
Tukey test    -4.682    0.000
```

Residual plots obtained with the `residualPlots()` method of the car package indicate that there is a quadratic response to *days* that has to be added to the model, as it is now in the residuals. The quadratic response is significant as indicated by the tests in the output table. The t-tests are obtained by adding the squared predictors to the model last. A significant test indicates that there is a significant quadratic or curvilinear tendency.

```
> influenceIndexPlot(ReducedModel)
```

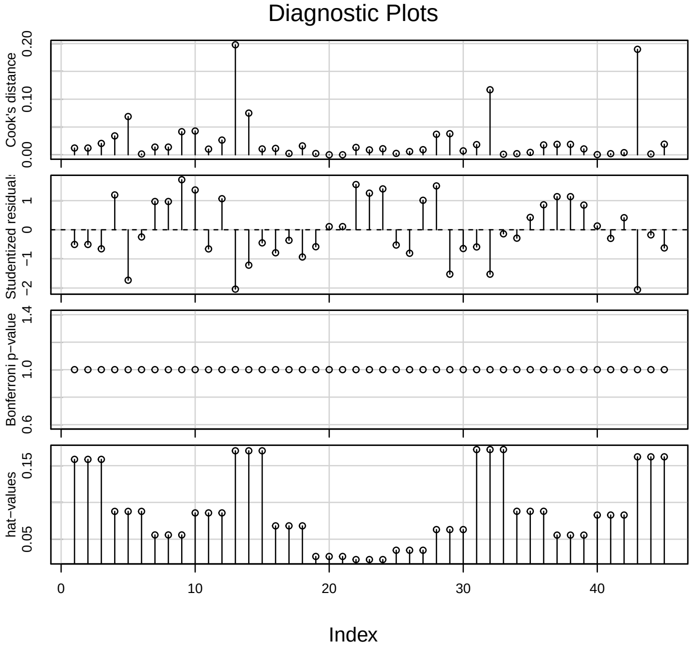

Figure 3: Index plots to assess influential cases.

2.2 Perform a test of Lack of Fit

Perform the test and interpret the results. [20]

For a test of lack-of-fit it is necessary to have ”true” or ”near replicates.” True (near) replicates are observations that have equal (similar) values for all predictors. In this data set we have true replicates in temperature and near replicates in days. The test consists in comparing a fullmodel where each combination of day and temperature gets a parameter, as if both predictors were continuous, vs. the model with continuous predictors. A ”cell” is the set of observations that the same or similar values of all predictors.

First, create the grouping variable for days.

```
> clover$dayg <- NA
> clover$dayg[clover$days<10] <- "A"
> clover$dayg[clover$days>10&clover$days<20] <- "B"
> clover$dayg[clover$days>20&clover$days<25] <- "C"
> clover$dayg[clover$days>25&clover$days<35] <- "D"
> clover$dayg[clover$days>35] <- "E"
```

Next, make the ”fullest” model and compare to the reduced model.

```
> FullestModel <- lm(lnwt ~ group*dayg, clover)
> anova(ReducedModel, FullestModel)
```

```
Analysis of Variance Table

Model 1: lnwt ~ days + t + days:t
Model 2: lnwt ~ group * dayg
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     41 1.14940
2     30 0.48212 11   0.66728 3.7747 0.001851 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is significant lack of fit, meaning that the reduced model is rejected because the deviations from the mean of each ”cell” to the prediction by the reduced model are significantly larger than expected on the basis of the MSE from the fullest model. This is just another results showing that in reality the RGR is not constant as plants age. A quadratic term should be added for days or a better mechanistic model should be constructed.