228.371 Computer Lab: Polynomial Regression; One and Two-way ANOVA

Semester One 2015 - Week 3

Instructions: Read each section and try the commands. Then try the Stream worksheet questions suggested to test your knowledge. The worksheet is "adaptive" which means if you get an answer wrong, you can try again. This quiz is to help you monitor your progress, it does not count toward your mark.

Note that because of fonts, especially for symbols like quotation marks, cutting and pasting commands from this document occasionally will not work - you may have to retype.

1 Polynomial Regression

Consider the diagnostic plots for the model of cherry tree wood volume developed in the lab last week.

```
data (trees)
treemod <- lm(Volume ~ Girth, data=trees)
par  (mfrow=c(2,2))
plot (treemod)
```

A trend is somewhat evident in the fitted values vs residuals plot. Combined with our expectation that the volume be related to the radius squared, ($radius = Girth/(2\pi)$), this suggests we should try a second degree polynomial.

```
treepoly <- lm(Volume ~ poly(Girth,2,raw=TRUE), data=trees)
```

```
summary(treepoly)
....
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                10.78627   11.22282   0.961 0.344728
poly(Girth, 2, raw = TRUE)1 -2.09214    1.64734  -1.270 0.214534
poly(Girth, 2, raw = TRUE)2  0.25454    0.05817   4.376 0.000152 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.335 on 28 degrees of freedom
Multiple R-squared:  0.9616,       Adjusted R-squared:  0.9588
F-statistic: 350.5 on 2 and 28 DF,  p-value: < 2.2e-16
```

Because we used `raw=TRUE` the coefficients are the coefficients of the linear and quadratic terms. Note that although only the quadratic term is significantly different from zero, we should not keep the quadratic term without the lower order term as well. If we want to assess the significance of the terms as a group within a larger model, we can use the `anova` function:

```
treepoly2  <- lm(Volume ~ Height+poly(Girth,2,raw=TRUE), data=trees)
treesimple <- lm(Volume ~ Height, data=trees)
anova(treesimple, treepoly2)
```

```
Analysis of Variance Table

Model 1: Volume ~ Height
Model 2: Volume ~ Height + poly(Girth, 2, raw = TRUE)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     29 5204.9
2     27  186.0  2    5018.9 364.25 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis here is that the larger model provides no benefit. The small p-value means it is clearly rejected.

We can also consider models that use the product of Height and Girth. These are called response surface models.

```
interaction.mod <- lm(Volume ~ Girth * Height, data=trees)
summary (interaction.mod)
```

```
Call:
lm(formula = Volume ~ Girth * Height, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5821 -1.0673  0.3026  1.5641  4.6649

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.39632   23.83575   2.911  0.00713 **
Girth        -5.85585    1.92134  -3.048  0.00511 **
Height       -1.29708    0.30984  -4.186  0.00027 ***
Girth:Height  0.13465    0.02438   5.524 7.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.709 on 27 degrees of freedom
Multiple R-squared:  0.9756,	Adjusted R-squared:  0.9728
F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16
```

Now try questions 1-3 on the lab worksheet.

2 One-way Anova

Load and attach the `chickwts` data set and examine it.

```
data (chickwts)
attach (chickwts)
class(chickwts)
class(chickwts$feed)
class(chickwts$weight)
```

Here our response variable (`weight`) is numeric, but our explanatory variable (`feed`) is a factor (values are discrete levels). We cannot feed a linear regression model, but we can fit an one-way ANOVA (analysis of variance) model. (One-way refers to having only one explanatory variable.) An anova model assumes that observations are normally distributed with common variance with mean level dependant on the level of the factor.

Try the following `R` commands.

```
boxplot (weight ~ feed)
chickmod <- lm(weight ~ feed)
summary (chickmod)
```

```
Call:
lm(formula = weight ~ feed)

Residuals:
     Min       1Q   Median       3Q      Max
-123.909  -34.413    1.571   38.170  103.091

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    323.583     15.834  20.436  < 2e-16 ***
feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
feedsoybean    -77.155     21.578  -3.576 0.000665 ***
feedsunflower    5.333     22.393   0.238 0.812495
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.85 on 65 degrees of freedom
Multiple R-squared:  0.5417,       Adjusted R-squared:  0.5064
F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

The intercept (323.58) gives the group mean for the first group (`casein`).

The following coefficients give the difference between the mean for that group and `casein`. So the mean for the `horsebean` group is 323.58-163.38=160.20. We are primarily interested in the p-value for the overall test of *any* differences between groups. In this case it is 5.936×10^{-10}.

Now try questions 4 and 5 on the lab worksheet.

3 Two-way Anova

We will use the `ToothGrowth` data set in `R`. (Access this data as you did the `chickwts` data in section 2). The response is the tooth length for guinea pigs raised under different conditions. The first factor is a dose level of Vitamin C (0.5, 1 and 2mg); the second factor is the mode of delivery (orange juice or ascorbic acid). There are 10 observations for each treatment combination. Try the following:

```
data (ToothGrowth)
attach (ToothGrowth)
boxplot (len ~ supp + dose)
interaction.plot (dose, supp, len)
#plots the mean value for each factor combination.
```

Does it appear the assumptions for ANOVA are satisfied (roughly)? (*Hint:* Are the boxplots roughly the same width?)

In two-way ANOVA we try to model the response variable with two factors. We would like to know the average **effect** that a change in the level of a factor has on the response varible. If that effect is constant for different levels of the **other** factor, then we say that those factors do not interact. If this is not the case then we need to model the interactions.

Use the following commands to fit the model with interactions.

```
ToothMod <- lm(len ~ supp * factor(dose))
#the factor() notation indicates we are treating dose as three
#different levels, not as a continuous variable
anova(ToothMod)
summary(ToothMod)
```

Now try questions 6-10 on the lab worksheet.