

ENV 710 – Applied Data Analysis

Fall 2014

Lab 10: Multiple Linear Regression

In this lab, we are going to extend linear regression to the case where we have multiple independent variables. In multiple regression we have to deal with some familiar concepts, such as normally distributed errors, and a few new concepts, including multicollinearity and model comparison.

The learning goals of the lab are to:

- Gain an improved understanding of the assumptions of regression
- Learn to implement multiple regression in **R**
- Understand how to interpret the output from multiple linear regression
- Learn to make predictions from regression models.

At the end of the lab, there are two analyses to conduct. Please type your answers to each problem, including any requested graphs, in a Word document. *Submit your answers and your R-code to the class Sakai site under the folder Assignments before 5 pm on either Mon., Nov. 10 (Section 07) or Wed., Nov., 12 (Sections 01 and 02).*

More functions in R

In this lab, we introduce a few new **R** commands.

`update()` – Updates and refits an existing model so that you do not have to rewrite the entire model each time.

`stepAIC()` - Performs stepwise model selection using AIC (Akaike Information Criterion), using either forward or backward selection procedures.

`curve()` – Draws a curve corresponding to a function over a defined interval or using a variable, x .

`vif()` – Function from the **HH** or **car** packages that calculates the variance inflation factor for each of the regression coefficients for the inputted model.

`I()` – Changes the class of an object to indicate that it should be treated “as is”; most often used to allow terms in models to include mathematical symbols.

Table depicting syntax for defining models in **R**.

Syntax	Model	Comments
$Y \sim A$	$Y = \beta_0 + \beta_1 A$	Straight-line with an implicit y-intercept
$Y \sim -1 + A$	$Y = \beta_1 A$	Straight-line with no y-intercept; that is, a fit forced through (0,0)
$Y \sim A + I(A^2)$	$Y = \beta_0 + \beta_1 A + \beta_2 A^2$	Polynomial model; note that the identity function I() allows terms in the model to include normal mathematical symbols.
$Y \sim A + B$	$Y = \beta_0 + \beta_1 A + \beta_2 B$	A first-order model in A and B without interaction terms.
$Y \sim A:B$	$Y = \beta_0 + \beta_1 AB$	A model containing only first-order interactions between A and B.
$Y \sim A*B$	$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB$	A full first-order model with a term; an equivalent code is $Y \sim A + B + A:B$.
$Y \sim (A + B + C)^2$	$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 AB + \beta_5 AC + \beta_6 BC$	A model including all first-order effects and interactions up to the n^{th} order, where n is given by $()^n$. An equivalent code in this case is $Y \sim A*B*C - A:B:C$.

Strategy for Data Analysis

1. *Define your research question.* Write down your question(s) of interest. Why are you analyzing the data? What hypotheses are you testing?
2. *Explore the data.* Summarize the data and use graphical tools such as histograms, boxplots, pairwise plots to evaluate distribution of the variables and the initial relationships between variables.
 - a. *Assess normality of variables.* Transform variables that are not normally distributed (square-root, log, or inverse transformations, or create dummy variables). It is most common to transform the dependent variable to be normally distributed. The goal in transforming independent variables is not to make them prettier and more symmetrical, but to make the relationship between the dependent and independent variable linear.
 - b. *Assess the relationship of each independent variable with the dependent variable.* Are the two variables linearly related? Do you need to consider other relationships between them?
 - c. *Assess the relationship of the independent variables with each other.* Are they too highly correlated? Are some variables redundant or better combined?
3. *Fit a tentative model.* Fit a first model. Plot the residual plots, q-q plots, and leverage plots to check for outliers, non-constant variance, and problems with model fit.
4. *Refine the model.* Test whether extra terms in the full model can be dropped.

Test and possibly drop interactions and quadratic terms or explore other types of non-linearity. Drop non-significant control variables.

- a. *Use stepwise methods to reduce model.* Do not just let **R** choose the best model for you! It is preferable to use the `update()` function and to evaluate each model as you remove variables.
5. *Check the final model.* Run the diagnostics on your final model to check for issues that could influence the model, but that are not exactly assumptions of multiple regression: multicollinearity, outliers and influential points, and missing data.
6. *Infer the answers to the questions of interest.* Calculate confident intervals, prediction intervals, test your hypotheses.
 - a. *Evaluate the relative impact of variables using standardized coefficients.*
 - b. *Make conclusions as to the effect of each variable on the dependent variable.*

Multiple Linear Regression

For this lab, we are going to examine a couple of datasets from the *Statistical Sleuth*. They are conveniently located in an **R** package, `Sleuth3`. Download this package and install it in your workspace. We will also need to download either the `HH` or `car` package to use the `vif()` function and install the `pair.fun` function for plotting pairwise plots of variables.

```
require(Sleuth3)
require(car)
source("pair.fun.R")
```

Load the first dataset from `Sleuth3`:

```
dat1 <- case0901
```

The data is from a study on variables that influence flower production in seedlings. The design consists of 12 treatment groups: 6 light intensities at two timing levels (0 days and 24 days) prior to photoperiodic flower induction (PFI). PFI is the production of a flowering stimulus in the leaves and its translocation to the stem apex under certain day lengths. Ten seedlings were randomly assigned to each treatment group. The number of flowers per plant is the primary measure of production, averaging the number of flowers produced by 10 seedlings in each group.

Our questions of interest are:

1. What are the effects of differing light intensity levels on flower production?
2. What is the effect of timing?
3. Does the effect of intensity depend on timing?

Let's look at the data and run an initial model including all the variables and interactions. Note that we are including `Time` as a factor (categorical variable).

```

pairs(dat1, lower.panel=panel.smooth,
      upper.panel=panel.cor, diag.panel=panel.hist)
hist(dat1$Flowers)
hist(log(dat1$Flowers))
qqnorm(log(dat1$Flowers))
qqline(log(dat1$Flowers))

lm.flw <- with(dat1, lm(log(Flowers) ~ Intensity +
                      factor(Time) + Intensity*factor(Time)))
summary(lm.flw)

```

```

Call:
lm(formula = log(Flowers) ~ Intensity + factor(Time) + Intensity *
    factor(Time))

Residuals:
    Min       1Q   Median       3Q      Max
-0.25748 -0.07343 -0.01337  0.09870  0.18137

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.3087179   0.0810330   53.172 < 2e-16 ***
Intensity      -0.0008102   0.0001387   -5.840 1.03e-05 ***
factor(Time)2    0.1462404   0.1145980    1.276  0.217
Intensity:factor(Time)2 0.0001580   0.0001962    0.805  0.430
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1231 on 20 degrees of freedom
Multiple R-squared:  0.7938, Adjusted R-squared:  0.7629
F-statistic: 25.67 on 3 and 20 DF, p-value: 4.625e-07

```

In the above code we figured out that the dependent variable, `Flowers`, is not at all normally distributed. As a start, we log-transform the variable, which makes it look a little better (the q-q plot looks better than the histogram). The variable `Intensity` is also not normally distributed, but we don't expect it to be since each experiment was conducted the same number of times at a few light intensities. We build our initial model with the log-transformed dependent variable, and will decide if we need to do more transformation or use different probability models if the model fit looks bad.

```

lm.flw <- with(dat1, lm(log(Flowers) ~ Intensity +
                      factor(Time)+ Intensity*factor(Time)))

```

In our initial model, only light intensity is statistically significant. Because the interaction is not significant we need to drop it and evaluate each of the main factors.

We are going to use the `update()` function to run our model, taking out variables one at a time. Look at the `summary()` of each of the models.

```

lm.flw1 <- update(lm.flw, ~.-Intensity:factor(Time))
lm.flw2 <- update(lm.flw1, ~.-factor(Time))
lm.flw3 <- update(lm.flw2, ~.-Intensity)

```

So, what have we learned?

First, the adjusted R^2 did not change much with the removal of the interaction terms (comparison of `lm.flw` and `lm.flw1`). We can safely remove the interaction as it did not contribute much to the model. In the `lm.flw1`, both `Time` and `Intensity` are statistically significant main factors. When we remove `Time` in model `lm.flw2`, our adjusted R^2 falls from 0.7669 to 0.5334, and the residual standard error increases from 0.122 to 0.169. There is very good evidence that both `Time` and `Intensity` should be left in the model as main effects. Incidentally, if `Intensity` is also removed from the model (`lm.flw3`) the residual error goes up to 0.253; there is no R^2 because there are no variables left in the model to explain variation in flower production.

We will take `lm.flw1`, which includes both `Intensity` and `Time`, as our best model. Let's take a look at the diagnostic plots to see if there are any problems.

```
plot(lm.flw1)
mean(residuals(lm.flw1))
vif(lm.flw1)
```

The mean error (mean of the residuals) is very small or close to 0 as expected.

The variance inflation factor (VIF) is an index of how much the variance of an estimated regression coefficient increases because of collinearity. A common rule of thumb is that if $VIF(\hat{\beta}_i) > 5$, then multicollinearity is high; others have proposed a cut off value of 10. If the VIF of a predictor variable were 4.3, this can be interpreted as meaning that the standard error for the coefficient of that variable is $\sqrt{4.3} = 2.07$ times as large as it would be if that predictor were uncorrelated with the other predictor variables.

The residual plots show a nice scattered cloud of points. The q-q plot doesn't look too bad. However, in all the plots, including the q-q plot, observation #9 sticks out as a potential outlier with lower flower production than the rest of the data. Let's redo the analysis, taking that observation out and see if it changes our results.

```
lm.flw4 <- with(dat1[-9,], lm(log(Flowers) ~ Intensity +
                             factor(Time)))
summary(lm.flw4)
```

Taking out the 9th observation improves the model fit and reduces the error leftover in the model; even so, it doesn't change our conclusions about the importance of `Time` and `Intensity`. We don't have any reason to think it is an outlier (rather than just an extreme data point); therefore, we will continue using `lm.flw1` with all the data.

Now let's plot the data for each level of `Time`. Here I back transform the data in the calls to `curve()` to look at the number of flowers produced over different levels of light intensity at the two different PFI times.

```

with(dat1, plot(Intensity[Time == 1], Flowers[Time==1],
  pch=21, bg="red", ylab="No. of flowers",
  xlab = "Light intensity", las=1, cex=1.2))
with(dat1, points(Intensity[Time == 2], Flowers[Time==2],
  pch=21, bg="grey", cex=1.2))
x <- with(dat1, seq(min(Intensity), max(Intensity),
  length=30))
coefs <- coefficients(lm.flw1)

curve(exp(coefs[1]+coefs[2]*x+coefs[3]), add=T, lty=2,
  col = 1)
curve(exp(coefs[1]+coefs[2]*x), add=T, lty=2, col = 2)

```

Because we log-transformed the data, our results are on the geometric scale. The relationship follows the equation:

$$\log(Y_i) = \beta_0 + \beta_1 \text{Intensity}_i + \beta_2 \text{Time}_i + \varepsilon_i.$$

To make the coefficients easier to manipulate, let's first save the regression coefficients to a variable.

```
coefs <- coefficients(lm.flw1)
```

The intercept, e^{β_0} or $\exp(\text{coefs}[1])$ is the mean number of flowers at Time 0 when Intensity is equal to 0. e^{β_1} , $\exp(\text{coefs}[2])$, is the percent change in number of flowers with a 1-unit change in Intensity. This becomes more intuitive if we think about a 100 unit change in light intensity: $\exp(\text{coefs}[2]*100) = 0.929$. Therefore, we conclude that a 100-unit increase in light intensity results in 92.9% of the flower production, or a 7.1% decrease in flower production ($t = -7.57$, $df = 21$, $p < 0.001$). (Note the way the previous sentence was written – this is how you should report statistics, in this class and outside of it.)

We can check the above results by inserting two levels of light intensity, 800 and 700 $\mu\text{mol}/\text{m}^2/\text{sec}$, and see the outcome:

```
(exp(coefs[1]+coefs[2]*800)-exp(coefs[1]+coefs[2]*700)) /
(exp(coefs[1]+coefs[2]*700))
```

The effect of the categorical variable Time is estimated by taking the ratio between the level 24 days and the *reference level* 0 days: $\exp(\text{coefs}[3]) = 1.2575$. There is a ~ 26% increase in flower production by beginning the light treatment 24 days prior to PFI, after taking into account light Intensity ($t = 4.599$, $df = 21$, $p < 0.001$). At the average light intensity, beginning the light treatment 24 days prior to PFI results in 12.5 more flowers per plant compared to beginning at PFI.

```

x1<-mean(dat1$Intensity)
flwrs24 <- exp(coefs[1]+coefs[2]*x1+coefs[3])
flwrs0 <- exp(coefs[1]+coefs[2]*x1+coefs[3]*0)
flwrs24 - flwrs0

```

Or we could have calculated this difference in flowers as:

```
0.26*flwrs0
```

Your assignment

Your assignment is to conduct two different analyses (see descriptions below).

For each regression, write a 1-page description of your analysis, results, and inference. Each write-up should include the following information:

- Null and alternative hypotheses of your tests
- Results of your statistical test, interpreting your test in 2-3 sentences that include the appropriate reporting of the statistics
- A description of how you checked the assumptions of your statistical test, including an interpretation of diagnostic figures
- Create plots of your regression equations to illustrate the results

Problem #1: In this problem, analyze the `TreePlot` data on Sakai. The database contains information on approximately 70 tree plots, including: (1) plot biomass (`AGBH.Mg.ha`), (2) mean tree diameter (`mDBH.cm`), (3) mean height (`mH.m`), (4) mean wood density (`mWD.g.m3`), mean basal area (`mBA.cm2`), and (5) presence of tree falls in the plot (`Tree.Fall`). Note that `Tree.Fall` is a factor with three levels: `majeur` (major tree fall), `mineur` (minor tree fall), and `rien` (no tree fall).

The goal of the analysis is to determine the variables that influence the plot biomass and the direction and magnitude of their effect on biomass.

Problem #2: The dataset `ozone.data.csv` consists of data on air quality (in terms of ozone concentration) and several weather variables, including temperature, solar radiation, and wind speed. Model the relationship of the weather variables to ozone concentration, making sure to consider interactions between the independent variables. (Do not worry about units in this example.) Once you have determined the best model to predict ozone concentration, write a one-page report that includes the following:

- Null and alternative hypotheses of your model
- Results of your statistical test, interpreting the fit of the model in 2-3 sentences that include the appropriate reporting of the statistics
- An interpretation of the regression model coefficients (i.e., what do each of the main effect(s) and interaction effect(s) mean)
- A description of how you checked the assumptions of your statistical test
- Graphs that depict (i) the relationship between the independent variable and dependent variable where a variable is only included as a main effect, (ii) the relationship between of the interacting independent variables with the dependent variable (i.e., plug in 2-3 interesting values for one of the independent and plot the other independent variable against the dependent variable).