Chapter 15: Correlation and regression - Relationships between measured values

Multiple Regression Analysis

TXCL7565/PHSC7565

What This Lecture Covers

- Multiple regression
- Model selection
- Model fit
- Multiple testing
- Incorporating nominal factors into multiple regression models using indicator variables

MULTIPLE REGRESSION

Multiple regression

- Multiple regression is a linear regression model with one response variable and multiple explanatory variables.

- Our model is similar to before, but we have extra regression coefficients and explanatory variables.

- Multiple regression is often used when we want to ‘control’ for a variable.

Multiple regression (cont.)

The relationship between the response for the ith observation (Y_i) and the values for the explanatory variables for the ith observation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

$$Y_i = \mu_{Y|X_1,\ldots,X_{p-1}} + \epsilon_i$$

where X_{i1} is the value of the first explanatory variable for observation i, X_{i2} is the value of the second explanatory variable for the observation i, etc., and β_j is the regression coefficient for the jth explanatory variable.

Interpretation of regression coefficients

- β_0 - the mean value of Y when ALL the explanatory variables are equal to zero.

- β_j - the mean change in Y when the jth explanatory variable increase by 1 unit and the other explanatory variables are held constant.

- $\hat{\beta}_0$ - the estimated mean value of Y when all the explanatory variables are equal to 0.

- $\hat{\beta}_j$ - the estimated mean change in Y when the jth explanatory variable increases by 1 unit and the other explanatory variables are held constant.

Multiple regression (cont.)

We estimate the regression coefficients for multiple regression using the method of least squares, i.e., estimating the βs with $\hat{\beta} s$ that minimize

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} e_i^2$$

where

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta_{p-1}} X_{p-1}$$

- GraphPad cannot do multiple regression (can trick it to do two independent variables)

- The following will be examples in R. You do not need to use R for homework/test. This is just for illustration.

- R can be downloaded at https://www.r-project.org/

Example

Rain ($cm.week^{-1}$)	Noon temp. (°C)	Sunshine ($h.day^{-1}$)	Wind speed ($km.h^{-1}$)	Toxin (µg per 100 g)
1.30	20.9	6.23	13.3	18.1
2.28	25.4	8.13	10.8	28.6
1.11	28.2	10.21	10.9	15.9
0.74	23.7	6.96	8.2	19.2
1.32	26.5	9.04	9.8	19.3
0.51	23.9	7.84	12.3	14.8
1.56	26.7	6.69	10.0	21.7
1.32	30.0	8.30	12.2	16.5
2.05	24.9	9.22	10.7	23.8
1.37	22.0	8.37	15.0	19.0

Example

- We have 4 potential predictors of concentration of fungal toxin in nuts (µg per 100g).

- Fit a multiple regression model using toxin as the response variable and noon_temp and rain as explanatory variables.

1. Find the estimated regression coefficients for this model.

2. Find the estimated regression model.

3. Interpret each of the coefficients in the model.

R code

```
rm(list=ls())

inputFile = “/Volumes/sabal/Teaching/TXCL7565/lectures/
Chapter15.Regression/toxicFungus.multipleRegression.txt"

orig = read.table(file=inputFile,header=TRUE,sep="\t")

full = lm(toxin ~ rain + noon_temp,data=orig)
summary(full)
```

```
Call:
lm(formula = toxin ~ rain + noon_temp, data = orig)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3429 -1.5539 -0.2791  0.8934  3.3513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.9990     6.6870   2.542  0.03855 *
rain          6.8863     1.3641   5.048  0.00148 **
noon_temp    -0.2635     0.2621  -1.006  0.34809
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.159 on 7 degrees of freedom
Multiple R-squared:  0.7857,    Adjusted R-squared:  0.7244
F-statistic: 12.83 on 2 and 7 DF,  p-value: 0.004559
```

1. Find the estimated regression coefficients for this model.

2. Find the estimated regression model.

```
Call:
lm(formula = toxin ~ rain + noon_temp, data = orig)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3429 -1.5539 -0.2791  0.8934  3.3513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.9990     6.6870   2.542  0.03855 *
rain          6.8863     1.3641   5.048  0.00148 **
noon_temp    -0.2635     0.2621  -1.006  0.34809
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.159 on 7 degrees of freedom
Multiple R-squared:  0.7857,    Adjusted R-squared:  0.7244
F-statistic: 12.83 on 2 and 7 DF,  p-value: 0.004559
```

3. Interpret each of the coefficients in the model.

Hypothesis tests for individual regression coefficients

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

Null Hypothesis: There is no linear association between the explanatory variable i and the outcome when the other explanatory variables are held constant

Alternative Hypothesis: There is an association between the explanatory variable i and the outcome when the the other explanatory variables are held constant.

Fit a multiple regression model using Toxin as the response variable and rain and sunshine as explanatory variables.

Test whether Sunshine has an effect on Toxin levels when amount of Rain is held constant.

R Code:

```
x = lm(toxin ~ rain + sunshine, data=orig)
summary(x)
```

```
Call:
lm(formula = toxin ~ rain + sunshine, data = orig)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8011 -0.6223 -0.1775  0.7621  3.0392

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.7390     4.8023   3.277  0.01354 *
rain          6.9916     1.3440   5.202  0.00125 **
sunshine     -0.6828     0.5816  -1.174  0.27881
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.111 on 7 degrees of freedom
Multiple R-squared:  0.795,     Adjusted R-squared:  0.7365
F-statistic: 13.58 on 2 and 7 DF,  p-value: 0.003897
```

Linear regression assumptions

1. **Linearity assumption**: The relationship between the mean response and the explanatory variables is linear in the explanatory variables.

Linear regression assumptions

2. **Assumptions about the errors:**

   a. The errors are assumed to follow a normal distribution.

   b. The errors have a mean of zero.

   c. The errors have constant variance (homoscedasticity or homogeneity of variance).

   d. The errors are independent of one another.

Linear regression assumptions

3. **Assumptions about explanatory variables:**

   a. The explanatory variables are nonrandom (fixed).

   b. The explanatory variables are not highly correlated with one another.

Linear regression assumptions

4. **Assumption about observations:** All observations are equally reliable and have approximately equal role in determining the regression results in influencing conclusions.

MODEL SELECTION

Model selection

- If we have several explanatory variables, there are many models we may potentially fit. How do we choose which one is best?

- There is no best model. In general, we want the simplest model that adequately explains the data, e.g., the most parsimonious model.

Backward selection

- Start with all variables and drop one variable at a time.
- The variable with the largest p-value is considered for removal. The variable is removed if it is not significant and a new model (excluding that variable) is fit.
- Variables are removed in this way until all variables in the model are significant.

Forward selection

- Start with only an intercept in the model.

- Look for the variable most associated with the response variable. If this regression coefficient is significantly different than zero a search for a second variable is made.

- The next variable selected is the most highly correlated with the residuals from the model with the first explanatory variable. If this regression coefficient is significantly different from zero, a search for a third regression variable is made.

- This process continues until there are no additional variables that are significantly different from zero.

Choosing a model

- "When there is sound theoretical literature available, you base your model upon what past research tells you."
  - ➡ e.g., in human genetics studies age, gender, and ethnicity are often always included in the regression model regardless of whether or not they significantly contribute.
- Use caution when using an automated selection method (i.e., forward or backward selection)
  - ➡ These methods rely on a mathematical criterion instead of human intelligence.
  - ➡ Variable selection may depend upon only slight differences in significance. Slight numerical differences can lead to major differences in the model.
  - ➡ Backward selection is preferred (not applicable if the number of predictors is larger than the number of observations.

Find the most parsimonious model for Toxin using all four possible explanatory variables and the backward selection procedure.

R Code:

```
y = lm(toxin ~ rain + noon_temp + sunshine + wind_speed,
data=orig)
summary(y)
```

```
Call:
lm(formula = toxin ~ rain + noon_temp + sunshine + wind_speed,
    data = orig)

Residuals:
      1       2       3       4       5       6       7       8
-1.8818  2.0498 -0.6314  0.4787 -0.5805  1.2508 -0.1921 -0.1813
      9      10
-1.1552  0.8429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.6084     7.1051   4.449  0.00671 **
rain          7.0676     1.0031   7.046  0.00089 ***
noon_temp    -0.4201     0.2413  -1.741  0.14215
sunshine     -0.2375     0.5086  -0.467  0.66018
wind_speed   -0.7936     0.2977  -2.666  0.04458 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.574 on 5 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.8535
F-statistic: 14.11 on 4 and 5 DF,  p-value: 0.006232
```

```
y = lm(toxin ~ rain + noon_temp + wind_speed, data=orig)
summary(y)
```

```
Call:
lm(formula = toxin ~ rain + noon_temp + wind_speed, data = orig)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6394 -0.9308  0.1394  0.6545  2.0909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.5651     6.6253   4.764  0.00311 **
rain          7.0108     0.9285   7.551  0.00028 ***
noon_temp    -0.4790     0.1919  -2.495  0.04682 *
wind_speed   -0.8218     0.2718  -3.023  0.02331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.468 on 6 degrees of freedom
Multiple R-squared:  0.915,     Adjusted R-squared:  0.8726
F-statistic: 21.54 on 3 and 6 DF,  p-value: 0.001298
```

Final regression model:

Interpretation of regression coefficients:

Predict the toxin levels at a site that has 2.05 $cm.week^{-1}$, 26.1 °C noon temperature, and 11.0 $km.h^{-1}$ wind speed.

MODEL FIT

Coefficient of determination

Coefficient of determination, R^2 - measures the proportion of total response variation explained by the linear regression model with p-1 explanatory variables compared to the regression model with only an intercept.

- R^2 measures the relative reduction in the residual sum of squares for our regression model in comparison to the model having only an intercept.

- When R^2 is close to 1, the regression model explains a high proportion of the variation in the observed data compared to the simplest model.

Adjusted Coefficient of Determination

- Since the coefficient of determination, R^2, measures the proportion of additional variance explained by the regression model in comparison to the simple model (just an intercept), it may seem like a reasonable statistic to choose between models.

- However, R^2 will never decrease, and almost always will increase, as you add explanatory variables to the model.

  - If R^2 is used to choose between models, we will always choose the model with the most explanatory variables.

  - This is NOT a good thing. We may make our model too complex. This is called overfitting.

Adjusted Coefficient of Determination (cont.)

- The adjusted coefficient of determination, R^2_a, is similar to the coefficient of determination but penalizes a model having more parameters

```
Residual standard error: 1.468 on 6 degrees of freedom
Multiple R-squared:  0.915,     Adjusted R-squared:  0.8726
F-statistic: 21.54 on 3 and 6 DF,  p-value: 0.001298
```

MULTIPLE TESTING

Multiple testing

- Just as multiple t-tests will increase the risk of false positives, considering lots of explanatory variables will also increase the risk of false positives in our regression model.

- When faced with any claim that a significant regression (or multiple regression) equation has been discovered, it is always worth asking how many potential predictors were initially considered.

INCORPORATING NOMINAL FACTORS INTO MULTIPLE REGRESSION MODELS USING INDICATOR VARIABLES

Nominal variables in linear regression

In the simple and multiple regression models we have looked at previously all variables were continuous. What if we wanted to add a categorical variable, such as gender or race, to our multiple regression model?

With a dichotomous predictor variable, we simply code one category as 1 and the other category as 0.

But with three or more groups we cannot use a single variable that has only zeros and ones to distinguish these groups and we do not want to assign values such as 0,1, and 2 to the categories because most softwares would assume there was an ‘order’ to the categories.

Indicator (dummy) variables

1. Count the number of groups you want to recode and subtract 1.
2. Create as many new variables as the value that you calculated in step 1. These are your dummy variables.
3. Choose one of your groups as a baseline (i.e., a group against which all other groups should be compared). This should usually be a control group, or, if you don't have a specific hypothesis, it should be the group that represents the majority of the people.
4. Having chosen a baseline group, assign that group values of 0 for all of your dummy variables.
5. For your first dummy variable, assign the value 1 to the first group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
6. For your second dummy variable, assign the value 1 to the second group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
7. Repeat until you run out of dummy variables.
8. Place all your dummy variables in the regression analysis.

Example - phosphodiesterase inhibitors

Activity	pK_a	Substituent
5.07	7.9	hydrogen
2.42	8.5	hydrogen
3.04	9.0	hydrogen
0.78	9.6	fluorine
8.07	7.8	fluorine
5.01	8.0	fluorine
5.48	8.4	chlorine
4.63	8.9	chlorine
4.43	9.4	chlorine

```
Call:
lm(formula = Activity ~ pKa + fluorine + chlorine, data = orig)

Residuals:
      1       2       3       4       5       6       7       8       9
-0.02308 -0.99688  1.01996 -0.67384  1.58755 -0.91371 -0.76350 -0.21667  0.98017

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.1631     5.9501   4.565  0.00603 **
pKa          -2.7937     0.6978  -4.003  0.01029 *
fluorine      1.1100     0.9956   1.115  0.31560
chlorine      2.5473     1.0405   2.448  0.05808 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.219 on 5 degrees of freedom
Multiple R-squared:  0.7835,    Adjusted R-squared:  0.6536
F-statistic:  6.03 on 3 and 5 DF,  p-value: 0.04085
```

State the regression model:

Interpret the fluorine and chlorine coefficients:

What did we learn?

- Regression can be extended to multiple regression, allowing several factors to participate in the prediction of the dependent variable.

- Regression coefficients are interpreted in the context that the other explanatory variables are held constant.

- The most parsimonious model can be found using either backward (preferred when the number of explanatory variables is smaller than the number of observations) or forward selection.

- Nominal variables can be incorporated into multiple regression equations by the use of ‘Indicator’ (or Dummy) variables.