Wednesday, February 19, 2014

Scribing 2/19/14 – Multiple Regression

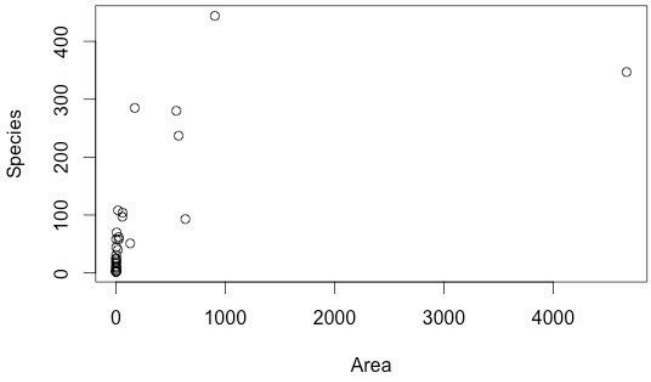
Varun Bhatnagar and Catherine King, 11:00 AM

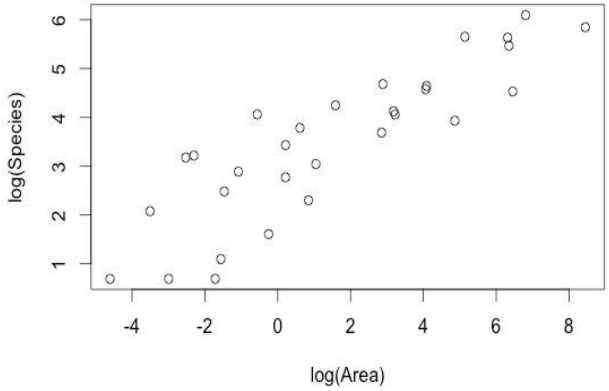
Overall question: How do I interpret multiple regression models?

Example 1: Galapagos Islands

- Variables: number of species, area, elevation
- Broad initial overview
  - High elevation, fewer species? Higher elevation allows for more species?

Step 1: plot the data.





The data are squished in lower left corner, implying that we need to do a log-log transformation. Equation of log-log transformation linear model: species = e^2.9 * Area ^0.39

Do the same process for elevation regarding the log-log scaled plot. This results in a similar plot to the model above.

The issue with this approach to the problem is…

Collinearity!

Collinearity occurs when the predictors in a model are related to one another (in other words, elevation and area are correlated).

- The estimate above is not clean (has dirty variation), thus we cannot determine area's effect on species independently because elevation is a confounding variable. The opposite is also true for elevation.
- Is it really that as area increases, species count increases? OR, is it just because large areas have higher elevation?

How do we deal with dirty variation?

Method 1: Two Stage Regressions

Step 1: Do a regression of `log(Species) ~ log(Area)`

Step 2: Take the residual of this model. The residual quantifies the part of the species variable that cannot be predicted by area.

Step 3: Create a linear model with the residual as your response variable and the elevation as the predictor. The resulting coefficient is -0.05821.

Step 4: Repeat this process with elevation, to get a coefficient of 0.08738.

When adjusting for area, elevation is also implicitly adjusted since the two predictor variables are correlated (they are collinear).

If you repeat this complete this process the starting with elevation rather than area, we get a coefficient of 1.08. Then, when taking the part of species that cannot be predicted by elevation (residuals) regress against area, and the coefficient is close to 0 (0.08).

THEREFORE, THIS STRATEGY DOES NOT WORK, because the order matters and you get a different answer depending on which model you do first. Since the variables are correlated, this method produces the wrong answer.

Method Two: An Order-Independent Method

1) Take residual of a linear model, `elevadj`, comparing `elevation~area` (thus adjusting for elevation). Then create a linear model comparing `species~elevadj`. The resulting slope is -0.317.

2) Take residual of a linear model, `areaadj` comparing `area~elevation` (adjusting for area). Then create a linear model comparing `species~areaadj`.
The resulting slope is 0.47.

We are now able to see the effect of each variable when the other variable is held constant. Unlike the first method, this method is order independent.

Method Three: Multiple Regression

Take the multiple regression equation in R using the following format:

```
lm(log(Species)~log(Area) + log(Elevation), data=gala)
```

This outputs the effect of each predictor variable, the same slope numbers as we determined in Method Two. The difference is that method three retains the values in one equation.

We can also take a 3-dimensional representation of the multiple regression equation. The data points form a three-dimensional "football like" point cloud within which we fit a plane of best fit. This plane describes the systematic variation of area and elevation.

Note that we can rotate this plot to "cancel" the effect of a predictor and see how the other predictor affects the response variable, holding all else equal.

The equation of this model (with only two predictors) is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

We can keep adding predictors to this equation.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_n x_{in} + e_i$$

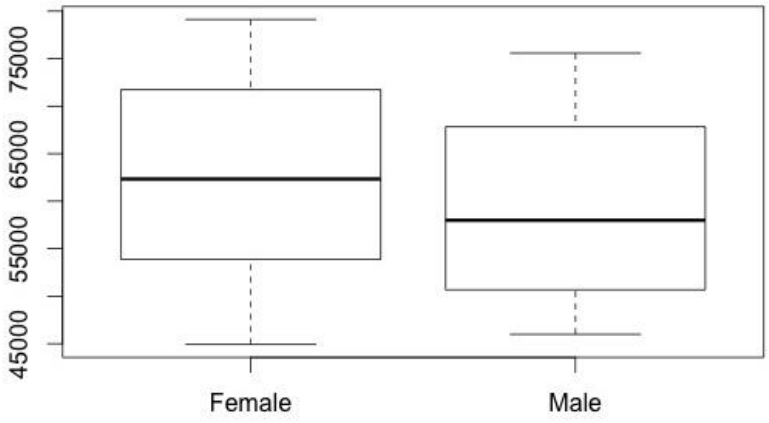
Example 2: Salary

- Variables: years of education, years of experience in the sector, months working with in the company, a dummy sex variable (1=man, 0=woman)

The average salary for each sex by `mean(Salary ~ Sex, data=salary)`:

Men: $62,610

Women: $59,381

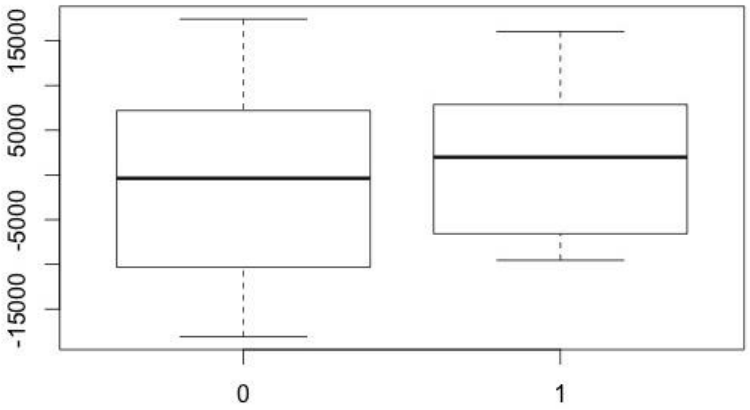


HOWEVER, the numbers above are not the end of the story.

There are confounding variables (qualifications, years of education, experience) that we must take into account when comparing the salaries of men versus women.

If we adjust for experience and then plot the residuals, the opposite case is true.

```
lm1 = lm(Salary~Experience, data=salary)
boxplot(resid(lm1)~salary$Sex)
```



In fact, we can adjust for all of the remaining variables.

```
lm2 = lm(Salary~Experience+Months+Education, data=salary)
```

We can now take the means to compare the average salary of men versus women after taking these variables into account. The slopes for the three variables are as follows:

experience = $65.79

months in department = $265.26

education = $699.73

Finally, we explicitly include a dummy variable for whether the worker is male or female (THIS IS THE CORRECT AND COMPLETE LINEAR MODEL).

```
lm3= lm(Salary~Experience+Months+Education+Sex, data=salary)
```

When the dummy variable is activated (thus the dummy variable indicates a male employee), the modified slopes are as follows:

experience = $122.25

months in department = $263.58

education = $591.08

sex = $2,320.54

When the dummy variable is active, men are getting paid ~$2,300 more than the women at the same experience, months in department, and education.

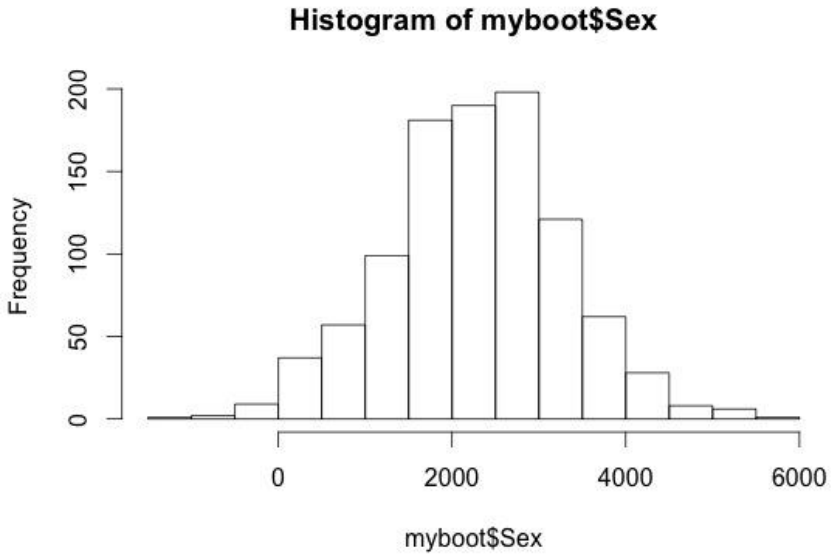
Why is there a gap of ~$1,500 when we compare the slopes with and without dummy variable?

Collinearity! Sex is correlated with the other predictors. Thus, when we include the dummy, it adjusts for the other predictors, changing the values of the variables while fixing the problem of collinearity.

Finally, we can use the `confint()` function or bootstrapping to quantify the uncertainty.

```
myboot = do(1000)*lm(Salary~Experience+Months+Education+Sex,
data=resample(salary))

hist(myboot$Sex)
```



Example 3: Car Auction

-Variables: acquisition price, vehicle age, make, color, transmission, if it is an online sale, auction, mileage (in 000s)

As we have done before, we can perform a multiple regression with all variables included in order to determine the effect of the predictors on price of the car (after we adjust for these predictor variables in relation to every other predictor variable).

```
lm(MMRAcquisitonRetailCleanPrice ~ VehicleAge + Make + Color +
Transmission + IsOnlineSale + Auction + VNST + I(VehOdo/1000),
data=carauction)
```

Conclusion: Use a multiple regression equation to hold some of the confounding predictor variables constant in order to evaluate a MARGINAL effect of a single predictor variable.