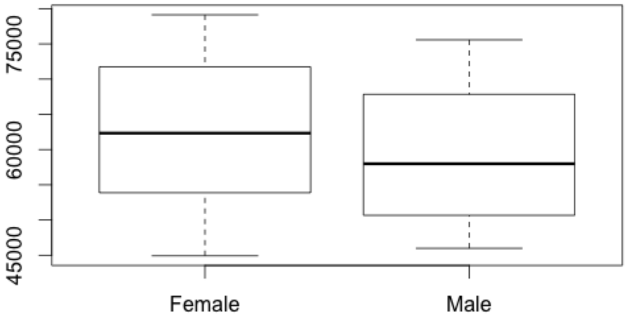
Zach Weissgarber STA371H
Dylan Compo 9:30-11:00

Multiple Regression - 2/19/14

Data files: salary.R & salary.csv

Dummy Variables and Interaction Terms will be used a lot more in the future. As the class progresses, we'll build up the capacity for judging which to use over time.

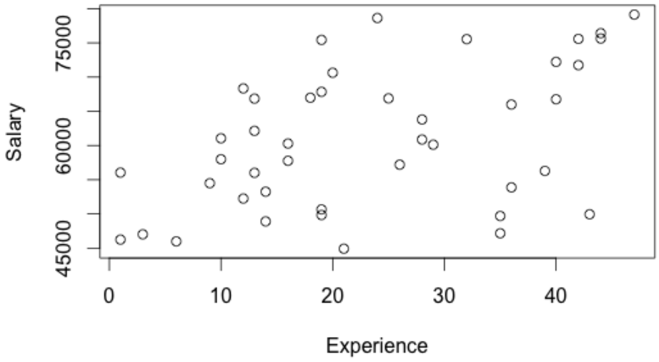
Regression analysis - are women being discriminated against with their salaries?



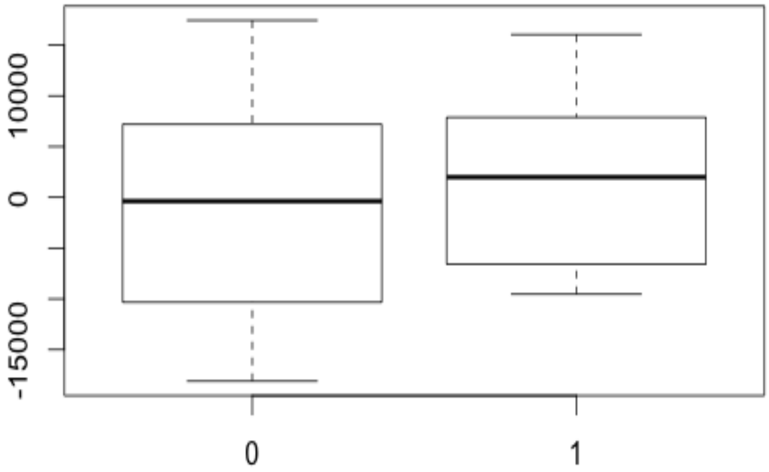
One interpretation of this plot: men are being discriminated against in the company, etc.
But consider: Maybe women, on average have more relevant education, experience, etc. (aka confounding variables)

note: indicator variable and dummy variable are the same thing

Salaries adjusted for experience:



Looking at the residuals: we notice that once adjusted for experience, men are slightly higher than women on average (reverses the sign of the original conclusion).



```
> mean(resid(lm1)~salary$Sex)
       0                1
-1228.776     1287.289
```

This command tells us men have positive residuals, and women have negative. Does this mean that there is discrimination? We have to look at other variables too.

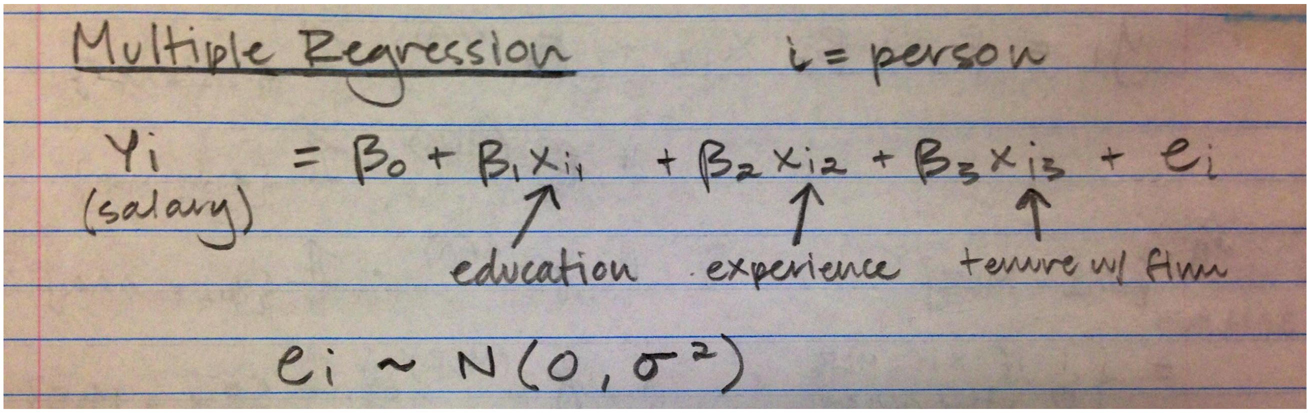
We can take all objections/variables and put them into one regression model.

Today in class, we focused on understanding how multiple regressions work, their geometry, and ultimately how to interpret their coefficients.

For the data set about salary, a multiple regression with 3 predictors uses the following R script:

```
lm2 = lm(Salary~Experience+Months+Education, data=salary)
summary(lm2)
boxplot(resid(lm2)~salary$Sex)
```

Deriving a formula for multiple regression:



Multiple Regression   i = person

$$\underset{\text{(salary)}}{Y_i} = \beta_0 + \underset{\overset{\uparrow}{\text{education}}}{\beta_1 x_{i1}} + \underset{\overset{\uparrow}{\text{experience}}}{\beta_2 x_{i2}} + \underset{\overset{\uparrow}{\text{tenure w/ firm}}}{\beta_3 x_{i3}} + e_i$$

$$e_i \sim N(0, \sigma^2)$$

- The individual predictors are said to be <u>multilinear</u> - each of them is linear.

- For a multiple regression, the ways in which we quantify uncertainty and confidence intervals all stay the same.

- You can assume linearity or bootstrap to perform this linear multiple regression model.

- Multiple regression is different from dummy variables where you have two numerical axes (one x, one y) and the dummy variables shift you up or down.

→Coefficients are very different for each of them.

→This has to do with weighting, not in sense of importance but in numerical values.

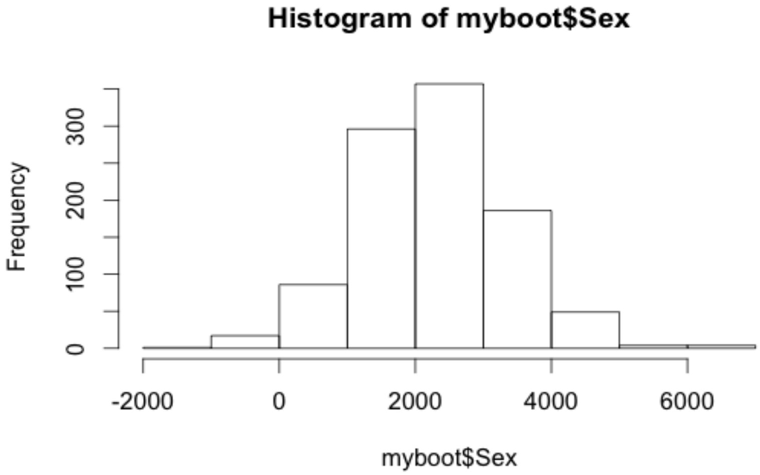
→Some variables are measured in months, some in years.

We can insert a dummy variable for sex. It gives us the following results:

```
              Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)   39305.71    2028.95    19.372   < 2e-16 ***
Experience      122.25     42.63      2.868    0.00671 **
Months          263.58     12.96     20.338   < 2e-16 ***
Education       591.08    424.44      1.393    0.17184
Sex            2320.54   1037.69      2.236    0.03128 *
```

Colinearity briefly introduced - will be discussed in more detail during a later in class.

For uncertainty in a multiple regression, we can bootstrap. Here are the results:

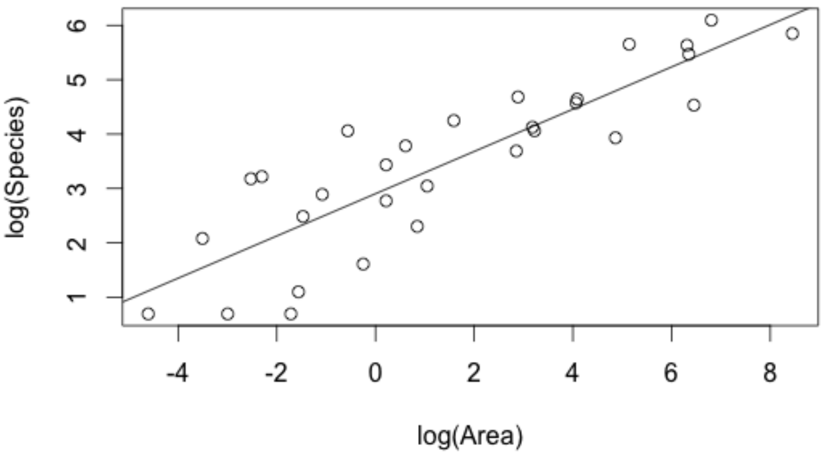


Example #2 - The Galapagos

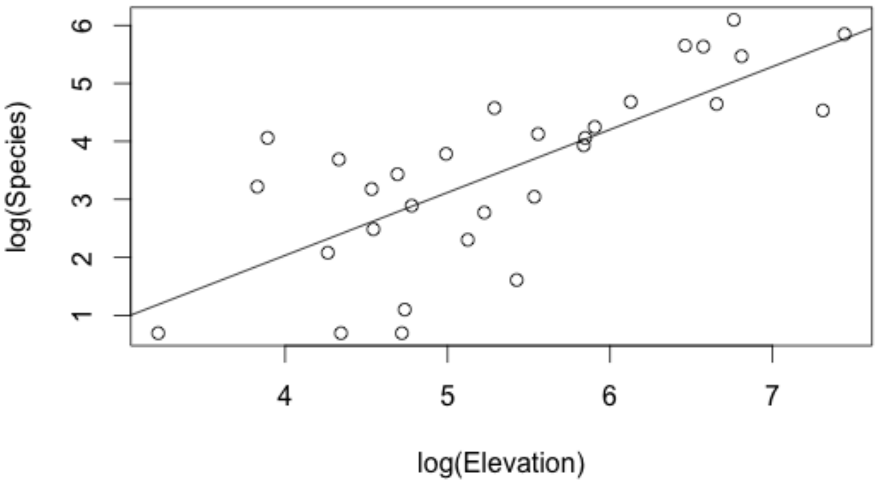
Data files: gala.R & gala.csv

Original graph does not give us an accurate, informative representation of the data, so we need to use…

Power log relationship (log-log scale) for Area vs. Species:



Power log relationship (log-log scale) for Elevation vs. Species:



We want to get a pure species account that's not affected by area and then see how the elevation affects those particular species. So how would we do this?

Steps:

1. Plot Species vs. Area
2. Take residuals of this data
3. Regress residuals against elevation
4. Now perform the same operation the other way around
5. Plot Species vs. Elevation
6. Take residuals of this data
7. Regress residuals against area

this is ultimately not the best method to go about regressing this data

Multiple regression is **not** like a two-stage regression model. The reason that a two-stage regression is not the right thing to do in this situation is because taking the log adjusts for area directly and for elevation indirectly since the two are related.

see R script (gala.r) for the following references

- Taking two predictors and regressing them against each other in lm6, taking out the residuals (the part of elevation that cannot be predicted by area)

  - Take your adjusted elevation and regress against species and you get:

    ```
    (Intercept)     elevadj
      3.5076185  -0.3175439
    ```

- Taking two predictors and regressing them against each other in lm8, taking out the residuals (the part of area that cannot be predicted by elevation)

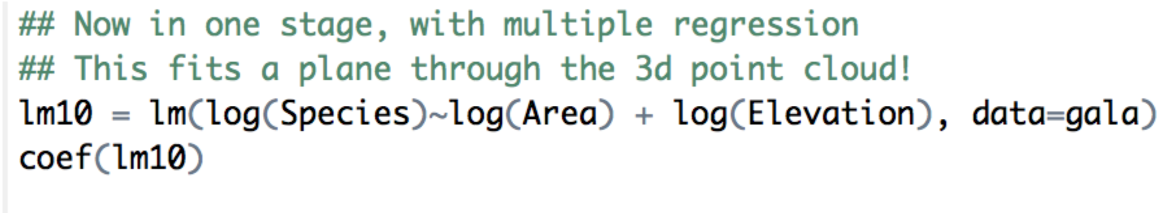
    - Take your adjusted area and regress against species and you get:

```
(Intercept)    areaadj
3.5076185   0.4767297
```

- If you run the multiple regression of both adjusted area and adjusted elevation against species:

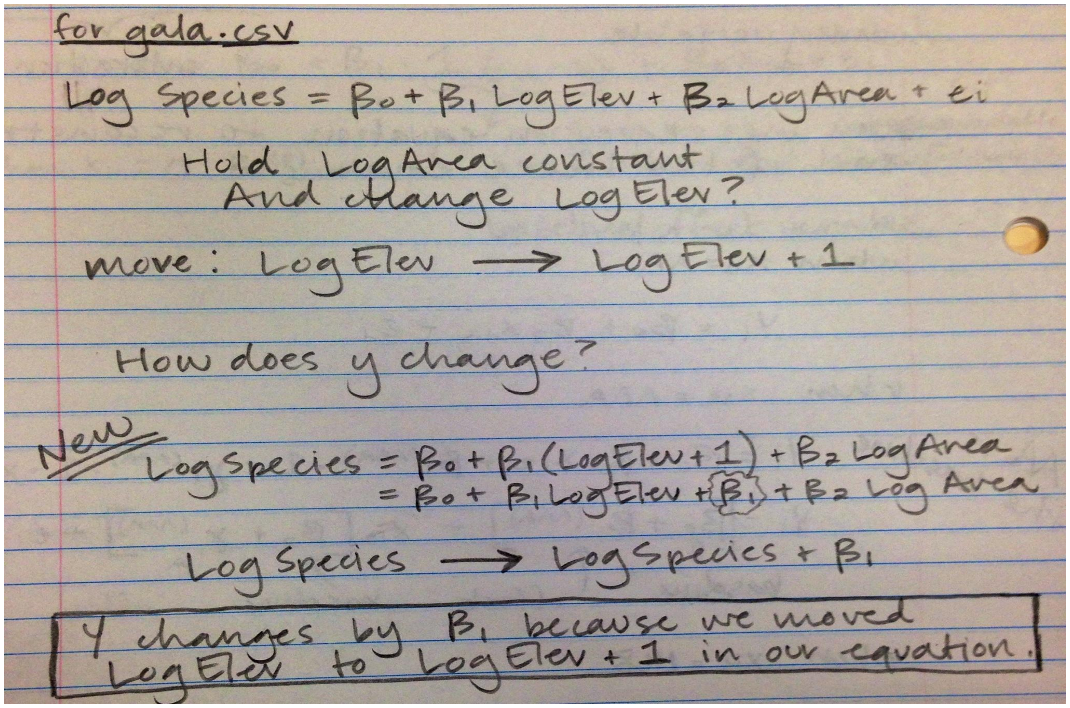
```
(Intercept)      log(Area)      log(Elevation)
4.4682481       0.4767297      -0.3175439
```

BUT, we would NEVER do a multiple regression through a 2-step method. You would go straight to step #10 and do the regression in one step using the following R script command:



```
## Now in one stage, with multiple regression
## This fits a plane through the 3d point cloud!
lm10 = lm(log(Species)~log(Area) + log(Elevation), data=gala)
coef(lm10)
```

A look at how changing LogElevation can alter your whole Y value and LogSpecies equation:



for gala.csv

$\text{Log Species} = \beta_0 + \beta_1 \text{LogElev} + \beta_2 \text{LogArea} + e_i$

Hold LogArea constant
And change LogElev?

move: $\text{LogElev} \longrightarrow \text{LogElev} + 1$

How does y change?

New

$\text{LogSpecies} = \beta_0 + \beta_1(\text{LogElev} + 1) + \beta_2 \text{LogArea}$
$= \beta_0 + \beta_1 \text{LogElev} + \boxed{\beta_1} + \beta_2 \text{Log Area}$

$\text{Log Species} \longrightarrow \text{LogSpecies} + \beta_1$

Y changes by β_1 because we moved LogElev to LogElev + 1 in our equation.