Lecture 19: Qualitative variables as predictors and Interactions

Pratheepa Jeganathan

11/04/2019

Recap

- What is a regression model?
- Descriptive statistics – graphical
- Descriptive statistics – numerical
- Inference about a population mean
- Difference between two population means
- Some tips on R
- Simple linear regression (covariance, correlation, estimation, geometry of least squares)
  - Inference on simple linear regression model
  - Goodness of fit of regression: analysis of variance.
  - F-statistics.
  - Residuals.
  - Diagnostic plots for simple linear regression (graphical methods).

Recap

- Multiple linear regression
  - Specifying the model.
  - Fitting the model: least squares.
  - Interpretation of the coefficients.
  - Matrix formulation of multiple linear regression
  - Inference for multiple linear regression
    - T-statistics revisited.
    - More F statistics.
    - Tests involving more than one β.
- Diagnostics – more on graphical methods and numerical methods
  - Different types of residuals
  - Influence
  - Outlier detection
  - Multiple comparison (Bonferroni correction)
  - Residual plots:
    - partial regression (added variable) plot,
    - partial residual (residual plus component) plot.

Outline

- Qualitative variables as predictors to the regression model (**CH**: Chapter 5)
- Adding interactions to the linear regression model.

Qualitative variables and Interactions

Introduction (Qualitative variables)

- Most predictor variables we have looked at so far were continuous: `height`, `rating`, etc.
- In many situations, we record a categorical variable: `gender`, `state`, `country`, etc.
- We call these variables *categorical* or *qualitative* variables.
  - In `R`, these are referred to as `factors`.
- For our purposes, we want to answer: **How do we include this in our model?**
- This will eventually lead us to the notion of *interactions* and some special regression models called *ANOVA* (analysis of variance) models.

Two-sample problem

- In some sense, we have already seen a regression model with categorical variables: the two-sample model.
- Two sample problem with equal variances: suppose $Z_j \sim N(\mu_1, \sigma^2), 1 \leq j \leq m$ and $W_l \sim N(\mu_2, \sigma^2), 1 \leq l \leq n$.
- For $1 \leq i \leq (m+n)$, let

$$X_i = \begin{cases} 1 & \text{if } \ i \ \text{ is one of } \ j \\ 0 & \text{otherwise.} \end{cases}$$

Two-sample problem

- The design matrix and response look like

$$\boldsymbol{Y}_{(n+m)\times 1} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \\ W_1 \\ \vdots \\ W_n \end{pmatrix}, \qquad \mathbf{X}_{(n+m)\times 2} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}$$

- The regression model is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

Salary example (**CH** Page 130)

- In this example, we have data on salaries of employees in IT (several years ago?) based on their years of experience, their education level and whether or not they are management.
- Outcome: `S`, salaries for IT staff in a corporation.
- Predictors:
  - `X`, experience (years)
  - `E`, education (1=High school diploma, 2= B.S., 3= Advanced degree)
  - `M`, management (1=management responsibility, 0=not management)
- Goal: Measure the effects of experience, education, and management on salary using regression analysis.

Salary example

```
url = 'http://stats191.stanford.edu/data/salary.table'
salary.table = read.table(url, header=T)
salary.table$E = factor(salary.table$E)
salary.table$M = factor(salary.table$M)
```

Salary example

- Let's take a quick look at how `R` treats a `factor`

```
str(salary.table$E)
```

```
##  Factor w/ 3 levels "1","2","3": 1 3 3 2 3 2 2 1 3 2 ...
```

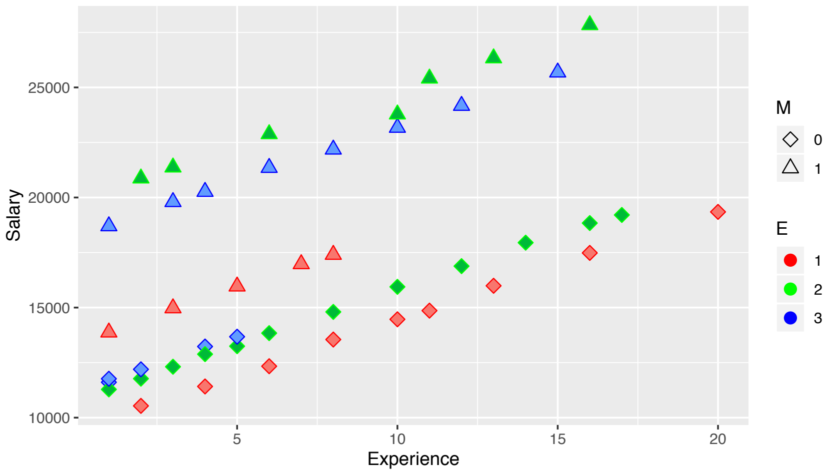
Salary example

- Let's take a look at the data.
  - We will use triangles for management, diamonds for non-management
  - red for education=1, green for education = 2 and blue for education=3.

Salary example

```
p = ggplot(data = salary.table, aes(x = X, y = S,
  shape = M, col = E, fill = E)) +
  geom_point(size = 3) +
  scale_shape_manual(values = c(23,24))+
  scale_color_manual(values = c("red",
    "green", "blue")) +
  xlab("Experience") +
  ylab("Salary")
```

Salary example



Salary example

- If we
  - assume a linear relationship between salary and experience (each additional year of experience is worth a fixed salary increment)
  - add raw education (1,2,3) to the model (each step-up in education is worth a fixed increment in salary)
    - this interpretation is too restrictive.
    - will consider education as a categorical variable with three levels (or categories)
- Effect of experience on salary
  - In these pictures, the slope of each line seems to be about the same.
  - How might we estimate it?

Salary example

- One solution is *stratification*.
  - Make six separate models (one for each combination of `E` and `M`) and estimate the slope.
    - We have few degrees of freedom in each group.

Salary example

- Or, use *qualitative* variables
  - IF it is reasonable to assume that σ^2 is constant for each observation.
  - THEN, we can incorporate all observations into 1 model.

$$S_i = \beta_0 + \beta_1 X_i + \beta_2 E_{i2} + \beta_3 E_{i3} + \beta_4 M_i + \varepsilon_i$$

- Above, the variables are:

$$E_{i2} = \begin{cases} 1 & \text{if } E_i{=}2 \\ 0 & \text{otherwise.} \end{cases}$$

$$E_{i3} = \begin{cases} 1 & \text{if } E_i{=}3 \\ 0 & \text{otherwise.} \end{cases}$$

Use *qualitative* variables

- Notes
  - Although E has 3 levels, we only added 2 variables to the model.
    - In a sense, this is because `(Intercept)` (i.e. β_0) absorbs one level.
  - If we added three variables then the columns of design matrix would be linearly dependent so we would NOT have a unique least squares solution.
  - Assumes β_1 – effect of experience is the same in all groups, unlike when we fit the model separately.
    - This may or may not be reasonable.

Use *qualitative* variables

- According to the model

$$S_i = \beta_0 + \beta_1 X_i + \beta_2 E_{i2} + \beta_3 E_{i3} + \beta_4 M_i + \varepsilon_i$$

  - the indicator variables determine the base salary level as a function of education and management status after adjustment for years of experience.
  - β_2 measures the salary differential for the B.S. relative to the H.S. (every fixed level of experience and management)
  - β_3 measures the salary differential for the A.D. relative to the H.S. (every fixed level of experience and management)
  - $\beta_3 - \beta_2$ measures the salary differential for the A.D. relative to the B.S. (every fixed level of experience and management)
  - β_4 measures the average incremental value in salary associated with a management position (every fixed level of experience and education)

Salary example

```r
salary.lm = lm(S ~ E + M + X, salary.table)
#summary(salary.lm)
```

```
Call:
lm(formula = S ~ E + M + X, data = salary.table)

Residuals:
     Min       1Q   Median       3Q      Max 
-1884.60  -653.60    22.23   844.85  1716.47 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8035.60     386.69  20.781  < 2e-16 ***
E2           3144.04     361.97   8.686 7.73e-11 ***
E3           2996.21     411.75   7.277 6.72e-09 ***
M1           6883.53     313.92  21.928  < 2e-16 ***
X             546.18      30.52  17.896  < 2e-16 ***
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1027 on 41 degrees of freedom
Multiple R-squared:  0.9568,	Adjusted R-squared:  0.9525 
F-statistic: 226.8 on 4 and 41 DF,  p-value: < 2.2e-16
```

Salary example

- Now, let's take a look at our design matrix

```r
head(model.matrix(salary.lm))
```

```
##   (Intercept) E2 E3 M1 X
## 1           1  0  0  1 1
## 2           1  0  1  0 1
## 3           1  0  1  1 1
## 4           1  1  0  0 1
## 5           1  0  1  0 1
## 6           1  1  0  1 2
```

- Comparing to our actual data, we can understand how the columns above were formed.
  - They were formed just as we had defined them above.

Salary example

```r
head(model.frame(salary.lm))
```

```
##       S E M X
## 1 13876 1 1 1
## 2 11608 3 0 1
## 3 18701 3 1 1
## 4 11283 2 0 1
## 5 11767 3 0 1
## 6 20872 2 1 2
```

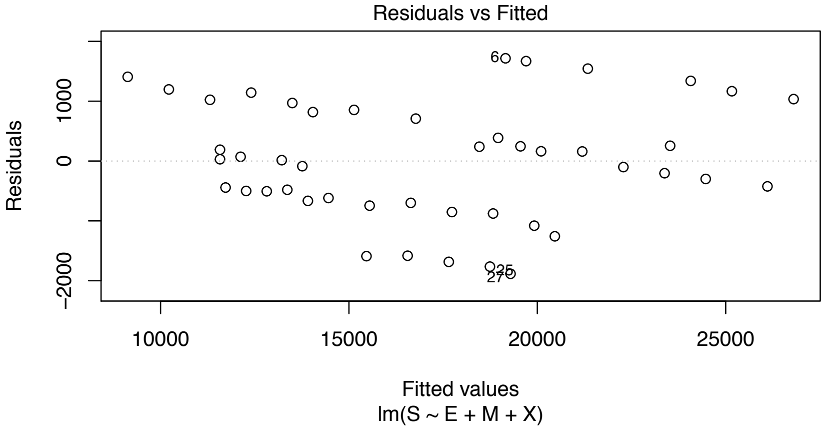
```r
head(data.frame(model.frame(salary.lm),
  model.matrix(salary.lm)), 4)
```

```
##       S E M X X.Intercept. E2 E3 M1 X.1
## 1 13876 1 1 1            1  0  0  1   1
## 2 11608 3 0 1            1  0  1  0   1
## 3 18701 3 1 1            1  0  1  1   1
## 4 11283 2 0 1            1  1  0  0   1
```

Diagnostics

- Assumed that σ^2 is constant for each observation.
- Let us check the diagnostics plot.

```
plot(salary.lm, add.smooth = FALSE, which = 1)
```



Interactions

- Our model has enforced the constraint the β_1 (Effect of experience) is the same within each group.
- We could fit a model with different slopes in each group, but keeping as many degrees of freedom as we can.
- This model has *interactions* in it: the effect of experience depends on what level of education you have.

Interaction between experience and education

- Model:

$$S_i = \beta_0 + \beta_1 X_i + \beta_2 E_{i2} + \beta_3 E_{i3} + \beta_4 M_i$$
$$+ \beta_5 E_{i2} X_i + \beta_6 E_{i3} X_i + \varepsilon_i.$$

- What is the regression function within each group?
- Note that we took each column corresponding to education and multiplied it by the column for experience to get two new predictors.
- To test whether the slope is the same in each group we would just test $H_0 : \beta_5 = \beta_6 = 0$.
- Based on figure, we expect not to reject H_0.

Interaction between experience and education

```r
model_XE = lm(S~ E + M + X + X:E, salary.table)
#summary(model_XE)
```

```
Call:
lm(formula = S ~ E + M + X + X:E, data = salary.table)

Residuals:
     Min       1Q   Median       3Q      Max 
-2013.04  -634.68   -16.71   615.66  2014.14 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7256.28     549.49  13.205 5.65e-16 ***
E2           4172.50     674.97   6.182 2.90e-07 ***
E3           3946.36     686.69   5.747 1.16e-06 ***
M1           7102.45     333.44  21.300  < 2e-16 ***
X             632.29      53.19  11.888 1.53e-14 ***
E2:X         -125.51      69.86  -1.797   0.0801 .  
E3:X         -141.27      89.28  -1.582   0.1216    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1005 on 39 degrees of freedom
Multiple R-squared:  0.9606,    Adjusted R-squared:  0.9546 
F-statistic: 158.6 on 6 and 39 DF,  p-value: < 2.2e-16
```

Testing $H_0 : \beta_5 = \beta_6 = 0$

```r
anova(salary.lm, model_XE)
```

```
## Analysis of Variance Table
##
## Model 1: S ~ E + M + X
## Model 2: S ~ E + M + X + X:E
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     41 43280719
## 2     39 39410680  2   3870040 1.9149  0.161
```

- The notation `X:E` denotes an *interaction*.
  - Generally, `R` will take the columns added for `E` and the columns added for `X` and add their element wise product (Hadamard product) to the design martrix.

Interaction in the model

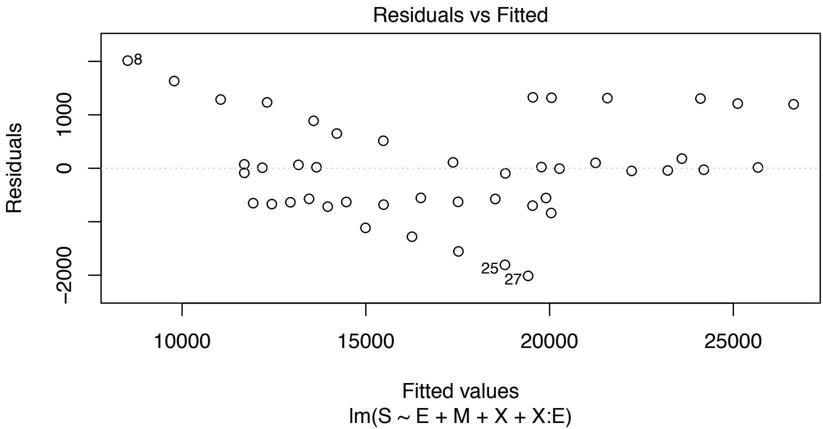
- Let's look at our design matrix again to be sure we understand what model was fit.

```r
model.matrix(model_XE)[10:20,]
```

```
##    (Intercept) E2 E3 M1 X E2:X E3:X
## 10           1  1  0  0 3    3    0
## 11           1  0  0  1 3    0    0
## 12           1  1  0  1 3    3    0
## 13           1  0  1  1 3    0    3
## 14           1  0  0  0 4    0    0
## 15           1  0  1  1 4    0    4
## 16           1  0  1  0 4    0    4
## 17           1  1  0  0 4    4    0
## 18           1  1  0  0 5    5    0
## 19           1  0  1  0 5    0    5
## 20           1  0  0  1 5    0    0
```

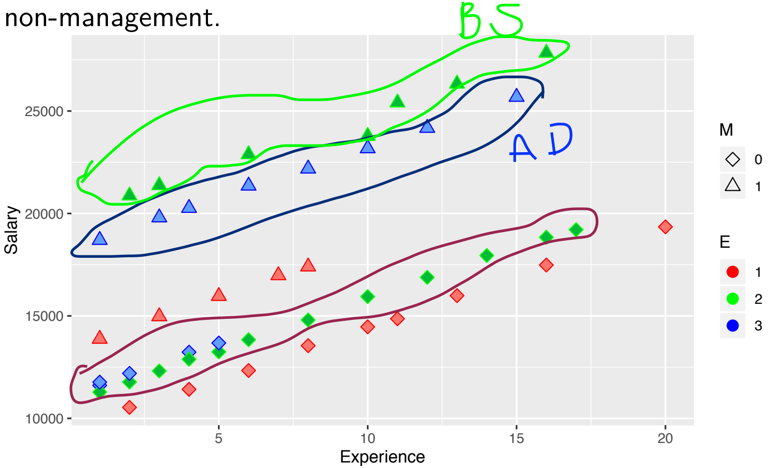
Diagnostics (Interaction between experience and education)

```
plot(model_XE, add.smooth = FALSE, which = 1)
```



Interaction between management and education

- We can also test for interactions between qualitative variables.
- In our plot, note that B.S in management make more than A.D. in management, but this difference disappears in non-management.



Interaction between management and education

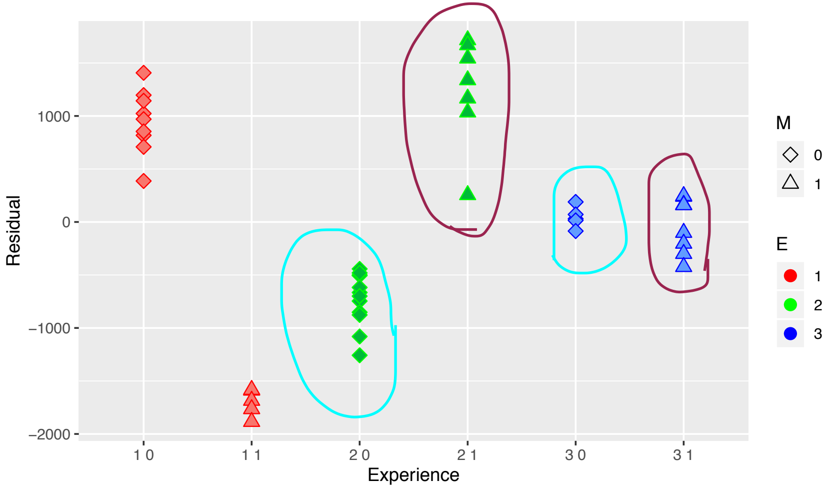
- This means the effect of education is different in the two management levels. This is evidence of an *interaction*.
- To see this, we plot the residuals within groups separately.

```
salary.lm = lm(S ~ E + M + X, salary.table)
df = data.frame(salary.table, res = resid(salary.lm))
df$group = paste(df$E, df$M)
```

Interaction between management and education

```r
p1 = ggplot(data = df, aes(x = group, y = res,
  shape = M, col = E, fill = E, group = group)) +
  geom_point(size = 3) +
  scale_shape_manual(values = c(23,24))+
  scale_color_manual(values = c("red",
    "green", "blue")) +
  xlab("Experience") +
  ylab("Residual")
```

Interaction between management and education



Interaction plot in R

- R has a special plot that can help visualize this effect, called an interaction.plot.

```
interaction.plot(salary.table$E,
  salary.table$M, resid(salary.lm), type='b',
  col=c('red','blue'), lwd=2, pch=c(23,24))
```

Interaction between management and education

- Based on figure, we expect an interaction effect.
- Fit model

$$\begin{aligned} S_i = \beta_0 + \beta_1 X_i + \beta_2 E_{i2} + \beta_3 E_{i3} + \ & \beta_4 M_i \\ & + \beta_5 E_{i2} M_i + \beta_6 E_{i3} M_i + \varepsilon_i. \end{aligned}$$

- Again, testing for interaction is testing $H_0 : \beta_5 = \beta_6 = 0$.
- What is the regression function within each group?

Interaction between management and education

```r
model_EM = lm(S ~ X + E:M + E + M,
  salary.table)
##summary(model_EM)
```

```
Call:
lm(formula = S ~ X + E * M + E + M, data = salary.table)

Residuals:
    Min      1Q  Median      3Q     Max 
-928.13  -46.21   24.33   65.88  204.89 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 9472.685     80.344  117.90   <2e-16 ***
X            496.987      5.566   89.28   <2e-16 ***
E2          1381.671     77.319   17.87   <2e-16 ***
E3          1730.748    105.334   16.43   <2e-16 ***
M1          3981.377    101.175   39.35   <2e-16 ***
E2:M1       4902.523    131.359   37.32   <2e-16 ***
E3:M1       3066.035    149.330   20.53   <2e-16 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 173.8 on 39 degrees of freedom
Multiple R-squared:  0.9988,	Adjusted R-squared:  0.9986 
F-statistic:  5517 on 6 and 39 DF,  p-value: < 2.2e-16
```

Interaction between management and education

- Testing for interaction is testing $H_0 : \beta_5 = \beta_6 = 0$.

```
anova(salary.lm, model_EM)
```

```
## Analysis of Variance Table
##
## Model 1: S ~ E + M + X
## Model 2: S ~ X + E:M + E + M
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     41 43280719
## 2     39  1178168  2  42102552 696.84 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

- We reject the null hypothesis.

Interaction between management and education

Let's look at our design matrix again to be sure we understand what model was fit.

```r
head(model.matrix(model_EM))
```

```
##   (Intercept) X E2 E3 M1 E2:M1 E3:M1
## 1           1 1  0  0  1     0     0
## 2           1 1  0  1  0     0     0
## 3           1 1  0  1  1     0     1
## 4           1 1  1  0  0     0     0
## 5           1 1  0  1  0     0     0
## 6           1 2  1  0  1     1     0
```

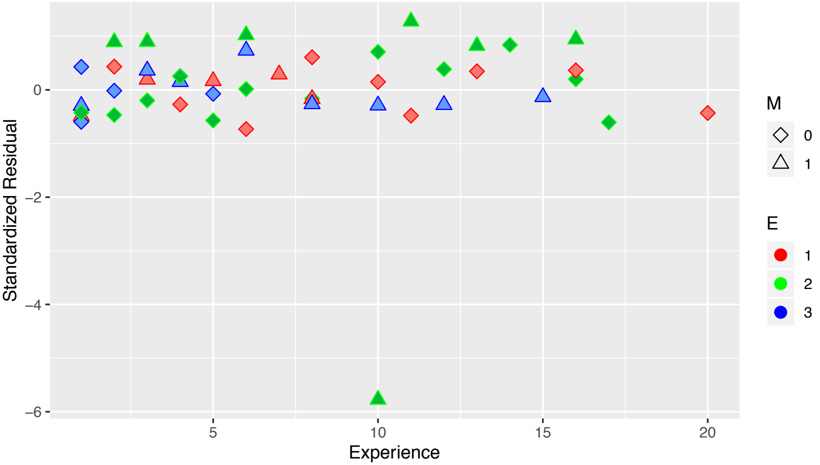
Diagnostics

- We will plot the residuals as functions of experience with each *experience* and *management* having a different symbol/color.

```
df2 = data.frame(salary.table, rs = rstandard(model_EM))
df2$group = paste(df2$E, df2$M)

p2 = ggplot(data = df2, aes(x = X, y = rs,
  shape = M, col = E, fill = E, group = group)) +
  geom_point(size = 3) +
  scale_shape_manual(values = c(23,24))+
  scale_color_manual(values = c("red",
    "green", "blue")) +
  xlab("Experience") +
  ylab("Standardized Residual")
```

Diagnostics

- One observation seems to be an outlier.



Outlier detection

```
library(car)
outlierTest(model_EM)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 33 -14.95083         1.6769e-17    7.714e-16
```

Refit the model

- Let's refit our model to see that our conclusions are not vastly different.

```
subs33 = c(1:length(salary.table$S))[-33]
salary.lm33 = lm(S ~ E + X + M,
  data=salary.table, subset=subs33)
model_EM33 = lm(S ~ E + X + E:M + M,
  data=salary.table, subset=subs33)
anova(salary.lm33, model_EM33)
```

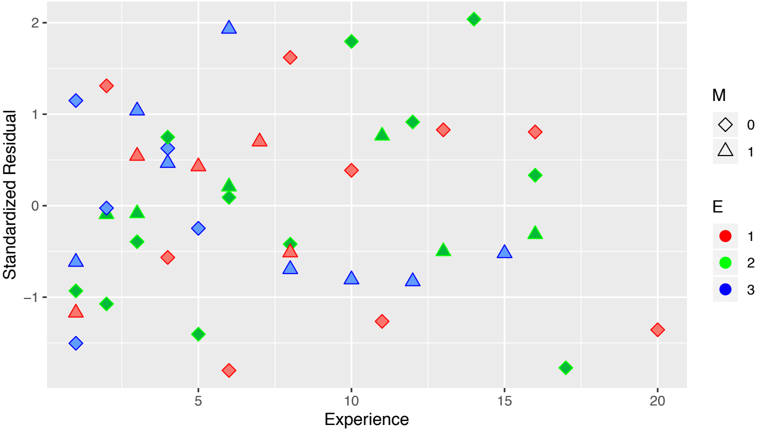
```
## Analysis of Variance Table
##
## Model 1: S ~ E + X + M
## Model 2: S ~ E + X + E:M + M
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     40 43209096
## 2     38   171188  2  43037908 4776.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
```

Diagnostics (refitted model)

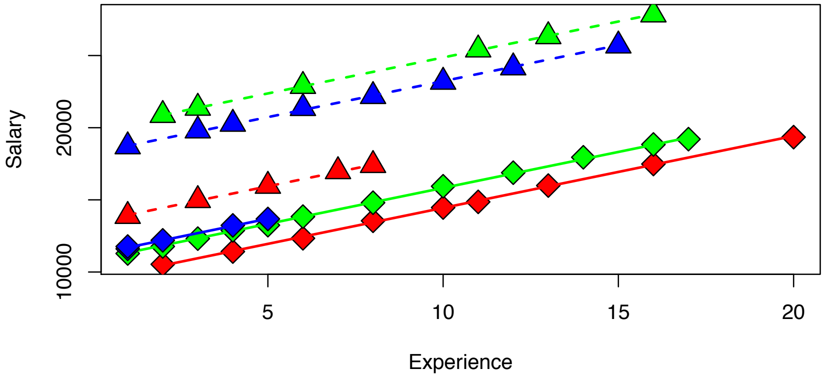
```
df3 = data.frame(salary.table[-33,], rs = rstandard(model_E
df3$group = paste(df3$E, df3$M)

p3 = ggplot(data = df3, aes(x = X, y = rs,
  shape = M, col = E, fill = E, group = group)) +
  geom_point(size = 3) +
  scale_shape_manual(values = c(23,24))+
  scale_color_manual(values = c("red",
    "green", "blue")) +
  xlab("Experience") +
  ylab("Standardized Residual")
```

Diagnostics (refitted model)

- Looks good!



Plot the fitted regression



Visualizing an interaction

- From our first look at the data, the difference between B.S. and A.D in the management group is different than in the non-management group.
  - This is an interaction between the two qualitative variables *management*,M and *education*,E.
  - We can visualize this by first removing the effect of experience, then plotting the means within each of the 6 groups using *interaction.plot*.

Visualizing an interaction

```
U = salary.table$S - salary.table$X * model_EM$coef['X']
interaction.plot(salary.table$E, salary.table$M, U,
  type='b', col=c('red', 'blue'),
  lwd=2, pch=c(23,24))
```

Reference

- **CH**: Chapter 5.
- Lecture notes of Jonathan Taylor .