STAT 224 Lecture 6
Interactions of Categorical & Numerical Predictors

Yibi Huang
Department of Statistics
University of Chicago

Example: Salary Survey Data (p.130, Textbook)

S	X	E	M
13876	1	1	1
11608	1	3	0
18701	1	3	1
11283	1	2	0
11767	1	3	0
20872	2	2	1
11772	2	2	0
10535	2	1	0
⋮	⋮	⋮	⋮
19346	20	1	0

S = Salary

X = Experience, in years

E = Education
(1 if H.S. only,
2 if Bachelor's only,
3 if Advanced degree)

M = Management Status
(1 if manager, 0 if non-manager)

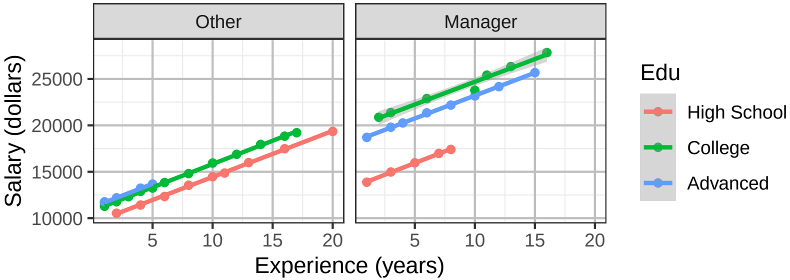
You can download the data at

http://www.stat.uchicago.edu/~yibi/s224/data/P130.txt

change the working directory and load the data using the command

```
p130 = read.table("P130.txt", header=TRUE)
```

```r
p130$Edu = factor(p130$E, labels=c("High School","College","Advanced"))
p130$Mgr = factor(p130$M, labels=c("Other","Manager"))
library(ggplot2)
ggplot(p130, aes(x = X, y = S, color=Edu)) +
  geom_point() + facet_grid(~Mgr) +
  geom_smooth(method="lm", formula='y~x') +
  xlab("Experience (years)") + ylab("Salary (dollars)")
```



Indicator Variables (aka. Dummy Variables)

- Salary (S): response
- Experience (X): numerical
- Education (E): categorical
  - 3 categories, needs 3 indicator variables

$$E_{i1} = \begin{cases} 1 & \text{if } i^{th} \text{ subject has a high school diploma only} \\ 0 & \text{otherwise} \end{cases}$$

$$E_{i2} = \begin{cases} 1 & \text{if } i^{th} \text{ subject has a B.A. or B.S. only} \\ 0 & \text{otherwise} \end{cases}$$

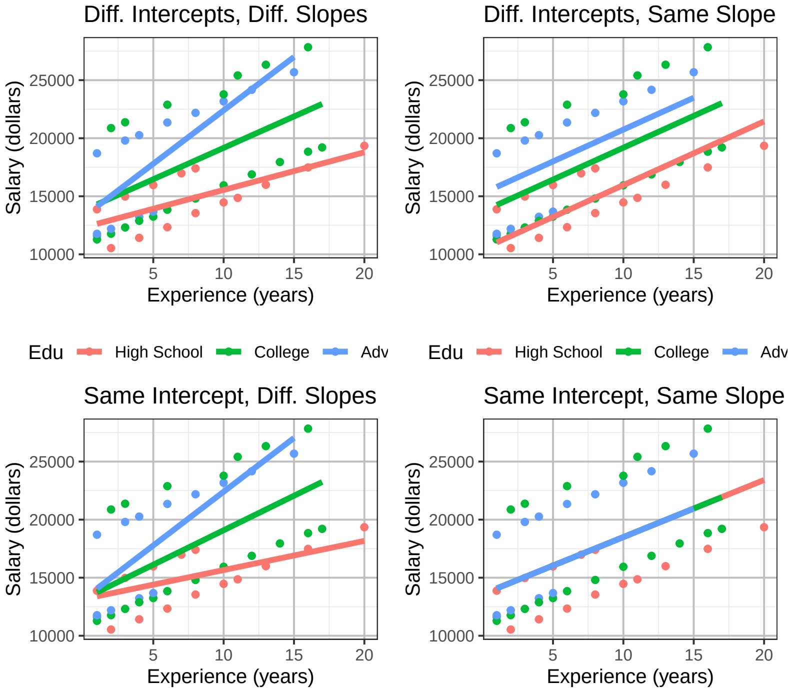
$$E_{i3} = \begin{cases} 1 & \text{if } i^{th} \text{ subject has an advanced degree} \\ 0 & \text{otherwise.} \end{cases}$$

- Cannot include all of E_1, E_2, and E_3 in the model since $E_1 + E_2 + E_3 = 1$. Must drop one of them.
- In general, **a categorical predictor with c categories needs only $c - 1$ indicator variables**

Models w/ Same or Different Intercept/Slopes

If we ignore M and consider models w/ X and E as predictors only, there are 4 possible models

- $S = \beta_{0E} + \beta_{1E}X + \varepsilon$ .... *different intercepts, different slopes*
  - both the intercept β_{0E} and the slope β_{1E} change with E (Edu)
- $S = \beta_{0E} + \beta_1 X + \varepsilon$ .... *different intercepts, same slope*
  - only the intercept β_{0E} changes with E but the slope β_1 doesn't
- $S = \beta_0 + \beta_{1E}X + \varepsilon$ .... *same intercept, different slopes*
  - only the slope β_{1E} changes with E but the intercept β_0 doesn't
- $S = \beta_0 + \beta_1 X + \varepsilon$ .... *same intercept, same slope*
  - neither the intercept β_0 nor the slope β_1 changes with E. Education (E) has no effect


Diff. Intercepts, Diff. Slopes
Diff. Intercepts, Same Slope
Same Intercept, Diff. Slopes
Same Intercept, Same Slope
Salary (dollars)
Experience (years)
25000
20000
15000
10000
5
10
15
20
Edu
High School
College
Adv

Models w/ Different Intercepts but Same Slope

$$S = \beta_0 + \delta_1 E_1 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \varepsilon$$
$$= \begin{cases} \beta_0 + \delta_1 + \beta X + \varepsilon & \text{if HS only} \\ \beta_0 + \delta_2 + \beta X + \varepsilon & \text{if B.A. or B.S. only} \\ \beta_0 + \delta_3 + \beta X + \varepsilon & \text{if advanced deg.} \end{cases}$$

Regardless of which indicator E_1, E_2, E_3 is dropped,

- Same slope β of X across all education levels.
- For all Education levels, people are paid β more on average if having 1 more years of experience.
  - The effect of X on S doesn't change w/ E
- Likewise, the effect of E on S doesn't change on X
  - People w/ a B.A. or B.S. earn $\delta_2 - \delta_1$ more on average than HS graduates w/ same years of experience (X).
    The change $\delta_2 - \delta_1$ doesn't depend on X
  - Ditto for (Advanced - Bachelor's) $= \delta_3 - \delta_2$
    and (Advanced - HS) $= \delta_3 - \delta_1$

Interactions & Additive Models

- If the effect of a predictor on response changes with the level of another predictor, we say there exists *interaction(s)* between the 2 predictors
  Otherwise, we say their effects are *additive*.
- e.g., the model below assumes the effects of education (E) and experience (X) on salary are additive

$$
\begin{aligned}
S &= \beta_0 + \delta_1 E_1 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \varepsilon \\
&= \begin{cases} \beta_0 + \delta_1 + \beta X + \varepsilon & \text{if HS only} \\ \beta_0 + \delta_2 + \beta X + \varepsilon & \text{if B.A. or B.S. only} \\ \beta_0 + \delta_3 + \beta X + \varepsilon & \text{if advanced deg.} \end{cases}
\end{aligned}
$$

- in R:

```
lm1 = lm(S ~ as.factor(E) + X, data=p130)
```

Model w/ Different Intercepts & Different Slopes

Consider the model

$$
\begin{aligned}
S &= \beta_0 + \delta_2 E_2 + \delta_3 E_3 \\
&\quad + \beta X + \gamma_2 (E_2 \cdot X) + \gamma_3 (E_3 \cdot X) + \varepsilon
\end{aligned}
$$

Here $(E_2 \cdot X)$ means the **product** of the indicator E_2 and X. Then

$$
S = \begin{cases}
\beta_0 + (\beta)X + \varepsilon & \text{if HS only} \\
\beta_0 + \delta_2 + (\beta + \gamma_2)X + \varepsilon & \text{if BA or BS only} \\
\beta_0 + \delta_3 + (\beta + \gamma_3)X + \varepsilon & \text{if advanced}
\end{cases}
$$

Here $(E_1 \cdot X)$ is not included since E_1 is dropped

- The model has the same property if a different indicator E_i is dropped

This model has *different intercepts* and *different slopes*!

Fitting Models with Interactions (Different Slopes) In R

In R, the term `E:X` and `E*X` both means interactions of E and X.

```
p130$E = as.factor(p130$E)
lm2 = lm(S ~ E+X+E*X, data = p130)
lm2$coef
(Intercept)          E2          E3           X        E2:X        E3:X
    12299.0      1461.2       898.2       324.5       216.3       595.5
```

Again, R drops the indicator E1 for the lowest level.

```
lm2$coef
(Intercept)          E2          E3           X        E2:X        E3:X
    12299.0      1461.2       898.2       324.5       216.3       595.5
```

$$\widehat{S} = 12299 + 1461.2E_2 + 898.2E_3 + 324.5X + 216.3(E_2 \cdot X) + 595.5(E_3 \cdot X)$$

$$= \begin{cases} 12299 \qquad\qquad + 324.5X & \text{if HS only} \\ 12299 + 1461.2 + (324.5 + 216.3)X & \text{if BA or BS only} \\ 12299 + 898.2 + (324.5 + 595.5)X & \text{if advanced} \end{cases}$$

On average, every extra year of experience worth

- \$324.5 if HS only
- \$324.5+\$216.3 if BA or BS only
- \$324.5+\$595.5 if Adv. deg.

The effect of X on S changes w/ $E \Rightarrow$ *Interactions*!

$$\widehat{S} = 12299 + 1461.2E_2 + 898.2E_3 + 324.5X + 216.3(E_2 \cdot X) + 595.5(E_3 \cdot X)$$

$$= \begin{cases} 12299 \qquad\qquad + 324.5X & \text{if HS only} \\ 12299 + 1461.2 + (324.5 + 216.3)X & \text{if BA or BS only} \\ 12299 + 898.2 + (324.5 + 595.5)X & \text{if advanced} \end{cases}$$

The effect of E on S also changes w/ X.

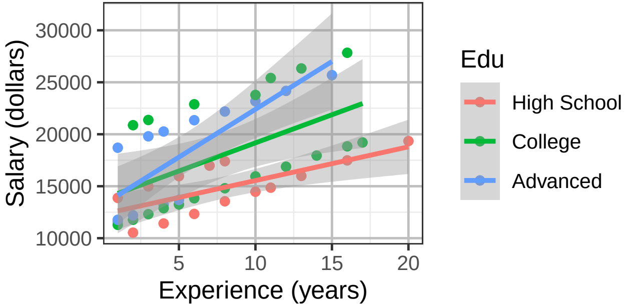
e.g., people with a Bachelor's deg and X years of experience earn on average

$$\underbrace{12299 + 1461.2 + (324.5 + 216.3)X}_{\text{Bachelor's deg}} - \underbrace{(12299 + 324.5X)}_{\text{HS}} = 1461.2 + 216.3X$$

more than people w/ HS diploma only and same years of experience

The difference $1461.2 + 216.3X$ change w/ X

```
ggplot(p130, aes(x = X, y = S, color=Edu)) + geom_point() +
  geom_smooth(method="lm", formula='y~x') +
  xlab("Experience (years)") +
  ylab("Salary (dollars)")
```



Are the slopes of the 3 lines significantly different?

Test Whether the Slopes Are Different

$$S = \beta_0 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \gamma_2 (E_2 \cdot X) + \gamma_3 (E_3 \cdot X) + \varepsilon$$

```
summary(lm2)$coef
            Estimate Std. Error t value       Pr(>|t|)
(Intercept)  12299.0     1740.4  7.0669 0.00000001514
E2            1461.2     2326.4  0.6281 0.53351638090
E3             898.2     2357.1  0.3811 0.70516764730
X              324.5      179.6  1.8065 0.07837469825
E2:X           216.3      238.6  0.9066 0.37004974108
E3:X           595.5      288.9  2.0615 0.04579092275
```

- `X:E2` (γ_2) is not significant (P-value 0.37)
  - No significant diff btw the slopes of the lines for HS & College
- `X:E3` (γ_3) is slightly significant (P-value 0.045).
  - slightly significant diff btw the slopes of the lines for HS v.s. advanced deg.

Test of Interactions

To know whether the effect of experience X on salary S changes with education level, one can test

$$H_0 : \gamma_2 = \gamma_3 = 0$$

by comparing the full model and the reduced model below

$$S = \beta_0 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \gamma_2 (E_2 \cdot X) + \gamma_3 (E_3 \cdot X) + \varepsilon \qquad \text{(full)}$$

$$S = \beta_0 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \varepsilon \qquad \text{(reduced)}$$

```
lm1 = lm(S ~ X+E, data = p130)
lm2 = lm(S ~ X+E+X*E, data = p130)
anova(lm1,lm2)
Analysis of Variance Table

Model 1: S ~ X + E
Model 2: S ~ X + E + X * E
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     42 550853135
2     40 497897342  2  52955792 2.13   0.13
```

Models w/ Same Intercept but Different Slopes — Less Common

$$
\begin{aligned}
S &= \beta_0 + \beta X + \gamma_2(E_2 \cdot X) + \gamma_3(E_3 \cdot X) + \varepsilon \\
&= \begin{cases} \beta_0 \qquad\quad + \beta X + \varepsilon & \text{if HS diploma only} \\ \beta_0 + (\beta + \gamma_2)X + \varepsilon & \text{if college only} \\ \beta_0 + (\beta + \gamma_3)X + \varepsilon & \text{if advanced degree} \end{cases}
\end{aligned}
$$

- Need to include X and $E * X$ but not E in the model
- R will automatically include `E` and `X` if `E*X` is included in the model. R would fit identical models for the 3 commands below.
  - `lm(S ~ X + E*X, data=p130)`
  - `lm(S ~ E + X + E*X, data=p130)`
  - `lm(S ~ E*X, data=p130)`
- Use `lm(S ~ X + E:X, data=p130)` to include only the product but not the `E`. Unlike `E*X`, `E:M` would not automatically include `E` and `M`.
- Does the effect of X on S depend on E?
  Does the effect of E on S depend on S?

```
summary(lm(S ~ X + E*X, data=p130))$coef
            Estimate Std. Error t value       Pr(>|t|)
(Intercept)  12299.0     1740.4  7.0669 0.00000001514
X              324.5      179.6  1.8065 0.07837469825
E2            1461.2     2326.4  0.6281 0.53351638090
E3             898.2     2357.1  0.3811 0.70516764730
X:E2           216.3      238.6  0.9066 0.37004974108
X:E3           595.5      288.9  2.0615 0.04579092275
summary(lm(S ~ X + E + E*X, data=p130))$coef
            Estimate Std. Error t value       Pr(>|t|)
(Intercept)  12299.0     1740.4  7.0669 0.00000001514
X              324.5      179.6  1.8065 0.07837469825
E2            1461.2     2326.4  0.6281 0.53351638090
E3             898.2     2357.1  0.3811 0.70516764730
X:E2           216.3      238.6  0.9066 0.37004974108
X:E3           595.5      288.9  2.0615 0.04579092275
```

```
summary(lm(S ~ E*X, data=p130))$coef
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  12299.0     1740.4  7.0669 0.00000001514
E2            1461.2     2326.4  0.6281 0.53351638090
E3             898.2     2357.1  0.3811 0.70516764730
X              324.5      179.6  1.8065 0.07837469825
E2:X           216.3      238.6  0.9066 0.37004974108
E3:X           595.5      288.9  2.0615 0.04579092275
summary(lm(S ~ X + E:X, data=p130))$coef
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  13144.6      916.3  14.345 8.699e-18
X              251.2      124.2   2.022 4.960e-02
X:E2           343.0      125.0   2.743 8.901e-03
X:E3           674.8      168.2   4.011 2.431e-04
```

Fitting a Model w/ Same Intercept & Diff Slopes in R

```
summary(lm(S ~ X+E:X, data=p130))$coef
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  13144.6      916.3  14.345 8.699e-18
X              251.2      124.2   2.022 4.960e-02
X:E2           343.0      125.0   2.743 8.901e-03
X:E3           674.8      168.2   4.011 2.431e-04
```

$$\widehat{S} = \begin{cases} 13144.6 + 251.2X & \text{if HS diploma only} \\ 13144.6 + (251.2 + 343)X & \text{if college only} \\ 13144.6 + (251.2 + 674.8)X & \text{if advanced degree} \end{cases}$$

Answer Questions w/ Appropriate Hypothesis Tests

Q1. Does salary grow faster w/ experience if one has higher education?

Q2. If equally educated, do those w/ more experience get paid more on average?

Q3. If equally experienced, do people w/ higher education get paid more on average?

Need to *translate* questions in context into tests of models or model parameters.

Q1. Does salary grow faster w/ experience if one has higher education?

Q1. Does salary grow faster w/ experience if one has higher education?

Ans: This asks whether the effect of experience (X) on salary (S) changes w/ Education (E), i.e., whether there are `E*X` interactions.

```
lm1 = lm(S ~ E + X + E*X, data=p130)
lm2 = lm(S ~ E + X, data=p130)
anova(lm2, lm1)
Analysis of Variance Table

Model 1: S ~ E + X
Model 2: S ~ E + X + E * X
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     42 550853135
2     40 497897342  2  52955792 2.13   0.13
```

As the P-value 0.13 is not small, the value of an extra year of experience does not change with significantly w/ education levels.

```
anova(lm2, lm1)
Analysis of Variance Table

Model 1: S ~ E + X
Model 2: S ~ E + X + E * X
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     42 550853135
2     40 497897342  2  52955792 2.13   0.13
```

How is the F-statistic 2.13 computed from the SSE's (`RSS`)?

```
anova(lm2, lm1)
Analysis of Variance Table

Model 1: S ~ E + X
Model 2: S ~ E + X + E * X
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     42 550853135
2     40 497897342  2  52955792 2.13   0.13
```

How is the F-statistic 2.13 computed from the SSE's (RSS)?

$$F = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(\text{dfE}_{reduced} - \text{dfE}_{full})}{\text{MSE}_{full}}$$

$$= \frac{(550853134.6991 - 497897342.452)/(42 - 40)}{497897342.452/40} = 2.1272$$

```
anova(lm2, lm1)
Analysis of Variance Table

Model 1: S ~ E + X
Model 2: S ~ E + X + E * X
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     42 550853135
2     40 497897342  2  52955792 2.13   0.13
```

How is the F-statistic 2.13 computed from the SSE's (RSS)?

$$F = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(\text{dfE}_{reduced} - \text{dfE}_{full})}{\text{MSE}_{full}}$$

$$= \frac{(550853134.6991 - 497897342.452)/(42 - 40)}{497897342.452/40} = 2.1272$$

```
anova(lm2, lm1)
Analysis of Variance Table

Model 1: S ~ E + X
Model 2: S ~ E + X + E * X
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     42 550853135
2     40 497897342  2  52955792 2.13   0.13
```

What are the degrees of freedom of the F statistic?

a. 42 and 40
b. 40 and 42
c. 2 and 40
d. 2 and 42

```
anova(lm2, lm1)
Analysis of Variance Table

Model 1: S ~ E + X
Model 2: S ~ E + X + E * X
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     42 550853135
2     40 497897342  2  52955792 2.13   0.13
```

What are the degrees of freedom of the F statistic?

a. 42 and 40
b. 40 and 42
c. 2 and 40 .... Ans
d. 2 and 42

```
pf(2.13, 2, 40, lower.tail=FALSE)
[1] 0.1321
```

Q2. If equally educated, do those w/ more experience earn more on average?

Ans: This means whether experience X has any effect on salary after accounting for education E.

```
lm3 = lm(S ~ E, data=p130)
anova(lm3, lm2) # if one believes no E*X interactions
```

```
Model 1: S ~ E
Model 2: S ~ E + X
  Res.Df       RSS Df Sum of Sq  F    Pr(>F)
1     43 891962932
2     42 550853135  1 341109797 26 0.0000077
```

or

```
anova(lm3, lm1) # if there might be E*X interactions
```

```
Model 1: S ~ E
Model 2: S ~ E + X + E * X
  Res.Df       RSS Df Sum of Sq    F  Pr(>F)
1     43 891962932
2     40 497897342  3 394065589 10.6 0.00003
```

Q3. If equally experienced, do people w/ higher education get paid more on average?

Ans: This means whether education E has any effect on salary after accounting for experience X.

```
lm4 = lm(S ~ X, data=p130)
anova(lm4, lm2) # if one believes no E*X interactions
```

```
Model 1: S ~ X
Model 2: S ~ E + X
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     44 710380856
2     42 550853135  2 159527722 6.08 0.0048
```

or

```
anova(lm4, lm1) # if there might be E*X interactions
```

```
Model 1: S ~ X
Model 2: S ~ E + X + E * X
  Res.Df       RSS Df Sum of Sq    F Pr(>F)
1     44 710380856
2     40 497897342  4 212483514 4.27 0.0057
```