STAT 224 Lecture 5
Qualitatitive Variables as Predictors (Ch5)

Yibi Huang
Department of Statistics
University of Chicago

MLR Model w/ Qualitative/Categorical Predictors

- In a linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$ what if some predictor X_j is **categorical**?
  - e.g., $X_1 =$ blood type (O, A, B, AB)?
    It makes NO sense to write a model like
    $$Y = \beta_0 + \beta_1(\text{blood type}) + \varepsilon.$$
    since blood type is not a number

MLR Model w/ Qualitative/Categorical Predictors

- In a linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$ what if some predictor X_j is **categorical**?
  - e.g., $X_1 =$ blood type (O, A, B, AB)?
    It makes NO sense to write a model like

    $$Y = \beta_0 + \beta_1(\text{blood type}) + \varepsilon.$$

    since blood type is not a number
- However, many demographics (Gender, marital status, etc) are categorical and can provide useful info for predicting/understanding the response variable Y.

MLR Model w/ Qualitative/Categorical Predictors

- In a linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$ what if some predictor X_j is **categorical**?
  - e.g., $X_1 =$ blood type (O, A, B, AB)?
    It makes NO sense to write a model like

    $$Y = \beta_0 + \beta_1(\text{blood type}) + \varepsilon.$$

    since blood type is not a number
- However, many demographics (Gender, marital status, etc) are categorical and can provide useful info for predicting/understanding the response variable Y.
- How to represent categorical variables “numerically” in a model?

MLR Model w/ Qualitative/Categorical Predictors

- In a linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$ what if some predictor X_j is **categorical**?
  - e.g., $X_1 =$ blood type (O, A, B, AB)? It makes NO sense to write a model like

$$Y = \beta_0 + \beta_1(\text{blood type}) + \varepsilon.$$

    since blood type is not a number
- However, many demographics (Gender, marital status, etc) are categorical and can provide useful info for predicting/understanding the response variable Y.
- How to represent categorical variables "numerically" in a model?
  - Solution: Create an **indicator** or **dummy variable** for each category of the categorical variable

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)

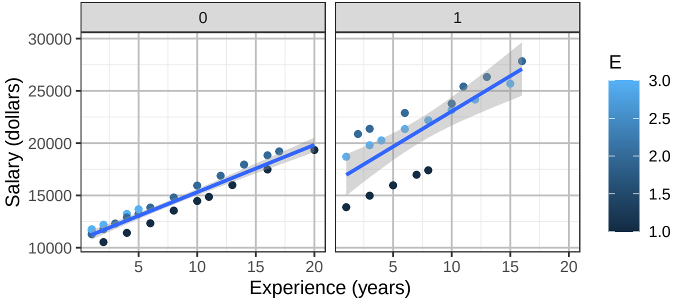
You can download the data at

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

change the working directory and load the data using the command

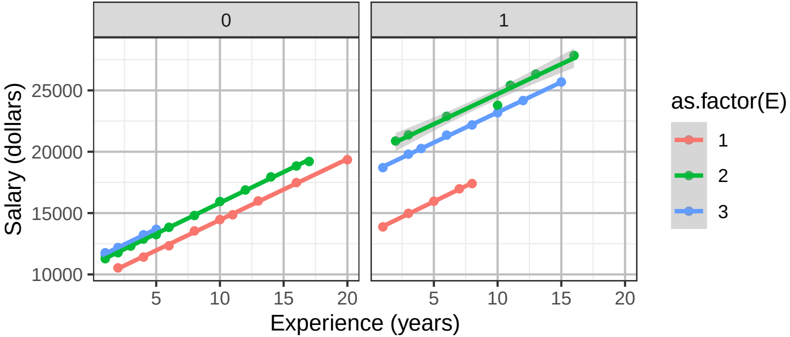
```
p130 = read.table("P130.txt", header=TRUE)
```

```r
library(ggplot2)
ggplot(p130, aes(x = X, y = S, color=E)) +
  geom_point() + facet_grid(~M) +
  geom_smooth(method="lm", formula='y~x') +
  xlab("Experience (years)") +
  ylab("Salary (dollars)")
```



Oops! R regards $E = 1, 2, 3$ as numerical rather than categorical!

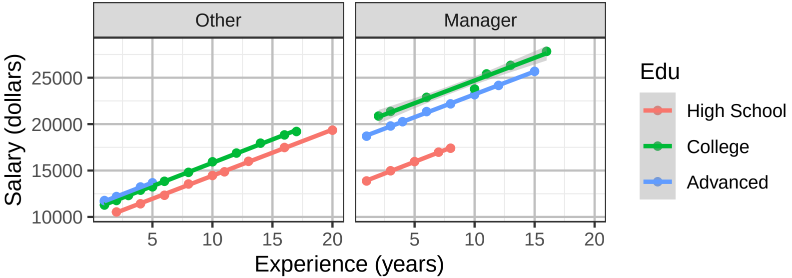
```r
ggplot(p130, aes(x = X, y = S, color=as.factor(E))) +
  geom_point() + facet_grid(~M) +
  geom_smooth(method="lm", formula='y~x') +
  xlab("Experience (years)") +
  ylab("Salary (dollars)")
```



The command `as.factor(E)` let R know that E is categorical.

It'd be better changing the labels of E and M

```r
p130$Edu = factor(p130$E, labels=c("High School","College","Advanced"))
p130$Mgr = factor(p130$M, labels=c("Other","Manager"))
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)")
```



Observe that Salary (S) and Experience (X) are linearly related for each level of Education (E) and Management Status (M).

How to express this as a MLR model?

Indicator Variables (aka. Dummy Variables)

Let's first ignore M and focus on S, X, and E.

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

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

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

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

Can one fit the model?

$$S = \beta_0 + \beta_1 X + \delta_1 E_1 + \delta_2 E_2 + \delta_3 E_3 + \varepsilon?$$

One of the Indicator Variables is Redudant

- Education (E) only has 3 categories

One of the Indicator Variables is Redudant

- Education (E) only has 3 categories
- Each subject must fall in exactly one of the 3 categories. For each subject only one of E_1, E_2, and E_3 can be 1 and the other 2 must be 0.

One of the Indicator Variables is Redudant

- Education (E) only has 3 categories
- Each subject must fall in exactly one of the 3 categories. For each subject only one of E_1, E_2, and E_3 can be 1 and the other 2 must be 0.
- So, the following identity always holds

$$E_1 + E_2 + E_3 = 1$$

One of the Indicator Variables is Redudant

- Education (E) only has 3 categories
- Each subject must fall in exactly one of the 3 categories. For each subject only one of E_1, E_2, and E_3 can be 1 and the other 2 must be 0.
- So, the following identity always holds

$$E_1 + E_2 + E_3 = 1$$

- One of E_1, E_2, and E_3 is redundant. The last one is known once the remaining are known

One of the Indicator Variables is Redudant

- Education (E) only has 3 categories
- Each subject must fall in exactly one of the 3 categories. For each subject only one of E_1, E_2, and E_3 can be 1 and the other 2 must be 0.
- So, the following identity always holds

$$E_1 + E_2 + E_3 = 1$$

- One of E_1, E_2, and E_3 is redundant. The last one is known once the remaining are known
- In general, **a categorical predictor with c categories needs only $c - 1$ indicator variables**

One of the Indicator Variables Must Be Removed

If we keep all indicator variables in the model

$$S = \beta_0 + \beta_1 X + \delta_1 E_1 + \delta_2 E_2 + \delta_3 E_3 + \varepsilon$$

the least square estimate for β_j and δ_j's **cannot be uniquely determined** since

$$\begin{aligned} S &= \beta_0 - c + \beta_1 X + (\delta_1 + c)E_1 + (\delta_2 + c)E_2 + (\delta_3 + c)E_3 + \varepsilon \\ &= \beta_0 + \beta_1 X + \delta_1 E_1 + \delta_2 E_2 + \delta_3 E_3 + c(\underbrace{E_1 + E_2 + E_3 - 1}_{=0}) + \varepsilon \end{aligned}$$

Regardless of the value of c, the coefficients

$$(\beta_0, \beta_1, \delta_1, \delta_2, \delta_3) \quad \text{and} \quad (\beta_0 - c, \beta_1, \delta_1 + c, \delta_2 + c, \delta_3 + c)$$

give identical means for the response.

We thus cannot keep all of E_1, E_2, and E_3 in the model

When E_1 is Removed From the Model...

When E_1 is removed from the model ..., the model becomes

$$S = \beta_0 + \beta_1 X + \delta_2 E_2 + \delta_3 E_3 + \varepsilon,$$

and the mean response $\mathrm{E}[S]$ for the 3 education levels are

Education (E)	Indicator	$\mathrm{E}(S)$
1 (HS diploma)	$E_2 = E_3 = 0$	$\beta_0 \quad\quad + \beta_1 X$
2 (Bachelor's degree)	$E_2 = 1,\ E_3 = 0$	$\beta_0 + \delta_2 + \beta_1 X$
3 (Advanced degree)	$E_2 = 0,\ E_3 = 1$	$\beta_0 + \delta_3 + \beta_1 X$

Based on the model above, for people w/ the same years of experience (X), the diff in their mean salary are

$$\text{(Bachelor's – HS)} = \delta_2$$

$$\text{(advanced – HS)} = \delta_3$$

$$\text{(advanced – Bachelor's)} = \delta_3 - \delta_2$$

Interpretation of Parameters

- To test whether those w/ a Bachelor's degree had a higher mean salary than those w/ only a HS diploma, after accounting for experience, which parameter should we test?

- To test whether a Bachelor's + an advanced degree increases mean salary one should test ...

- To test whether an advanced degree increases mean salary than a Bachelor's degree after accounting for experience, one should test ...

Interpretation of Parameters

- To test whether those w/ a Bachelor's degree had a higher mean salary than those w/ only a HS diploma, after accounting for experience, which parameter should we test?

H$_0$: $\delta_2 = 0$ v.s. H$_1$: $\delta_2 > 0$

- To test whether a Bachelor's + an advanced degree increases mean salary one should test . . .

- To test whether an advanced degree increases mean salary than a Bachelor's degree after accounting for experience, one should test . . .

Interpretation of Parameters

- To test whether those w/ a Bachelor's degree had a higher mean salary than those w/ only a HS diploma, after accounting for experience, which parameter should we test?

H_0: $\delta_2 = 0$ v.s. H_1: $\delta_2 > 0$

- To test whether a Bachelor's + an advanced degree increases mean salary one should test ...

H_0: $\delta_3 = 0$ v.s. H_1: $\delta_3 > 0$

- To test whether an advanced degree increases mean salary than a Bachelor's degree after accounting for experience, one should test ...

Interpretation of Parameters

- To test whether those w/ a Bachelor's degree had a higher mean salary than those w/ only a HS diploma, after accounting for experience, which parameter should we test?

H_0: $\delta_2 = 0$ v.s. H_1: $\delta_2 > 0$

- To test whether a Bachelor's + an advanced degree increases mean salary one should test ...

H_0: $\delta_3 = 0$ v.s. H_1: $\delta_3 > 0$

- To test whether an advanced degree increases mean salary than a Bachelor's degree after accounting for experience, one should test ...

H_0: $\delta_3 = \delta_2$ v.s. H_1: $\delta_3 > \delta_2$

```r
lm0 = lm(S ~ X + E, data=p130)
summary(lm0)$coef
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)   8279.9     1814.6   4.563 0.000041758
X              560.8      105.8   5.299 0.000003781
E             2418.4      706.9   3.421 0.001377546
```

- Something wrong?

```
lm0 = lm(S ~ X + E, data=p130)
summary(lm0)$coef
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)   8279.9     1814.6   4.563 0.000041758
X              560.8      105.8   5.299 0.000003781
E             2418.4      706.9   3.421 0.001377546
```

- Something wrong?
- R treats E (education) as numerical taking values 1, 2, and 3, not a categorical one

Numerical or Categorical?

If one treats E (education) as **numerical** taking values 1, 2, and 3, the model then becomes

$$S = \beta_0 + \beta X + \delta E + \varepsilon.$$

The mean response $\mathrm{E}[S]$ for the 3 education levels would be

Education (E)	Value of E	$\mathrm{E}(S)$
1 (HS diploma)	1	$\beta_0 + \beta_1 X + \delta$
2 (Bachelor's degree)	2	$\beta_0 + \beta_1 X + 2\delta$
3 (Advanced degree)	3	$\beta_0 + \beta_1 X + 3\delta$

The diff in mean salary controlling for experience X would be

$$(\text{Bachelor's} - \text{HS}) = \delta$$

$$(\text{advanced} - \text{Bachelor's}) = \delta$$

That is, the salary bonus for completing college is as much as the bonus for completing an advanced degree . . . . unrealistic and too restrictive.

```
lm1 = lm(S ~ X + as.factor(E), data=p130)
summary(lm1)$coef
               Estimate Std. Error t value        Pr(>|t|)
(Intercept)     10474.3     1305.4   8.024 0.0000000005186
X                 548.6      107.6   5.100 0.0000076946014
as.factor(E)2    3221.1     1275.8   2.525 0.0154427258510
as.factor(E)3    4780.1     1422.7   3.360 0.0016690499444
```

- The command `as.factor()` tells R that E is categorical and the indicator variables E1, E2, E3 are created automatically

```
lm1 = lm(S ~ X + as.factor(E), data=p130)
summary(lm1)$coef
                Estimate Std. Error t value        Pr(>|t|)
(Intercept)      10474.3     1305.4   8.024 0.0000000005186
X                  548.6      107.6   5.100 0.0000076946014
as.factor(E)2     3221.1     1275.8   2.525 0.0154427258510
as.factor(E)3     4780.1     1422.7   3.360 0.0016690499444
```

- The command `as.factor()` tells R that E is categorical and the indicator variables E1, E2, E3 are created automatically
- By default, R drops the indicator E1 for the lowest level

```r
lm1 = lm(S ~ X + as.factor(E), data=p130)
summary(lm1)$coef
                Estimate Std. Error t value         Pr(>|t|)
(Intercept)      10474.3     1305.4   8.024 0.0000000005186
X                  548.6      107.6   5.100 0.0000076946014
as.factor(E)2     3221.1     1275.8   2.525 0.0154427258510
as.factor(E)3     4780.1     1422.7   3.360 0.0016690499444
```

- The command `as.factor()` tells R that E is categorical and the indicator variables E1, E2, E3 are created automatically
- By default, R drops the indicator E1 for the lowest level
- 95% Confidence interval for coefficients:

```r
confint(lm1, level=0.95)
                 2.5 %  97.5 %
(Intercept)     7839.8 13108.7
X                331.5   765.7
as.factor(E)2    646.4  5795.8
as.factor(E)3   1908.9  7651.4
```

From the output of on th previous slide, the predicted salary is

$$\widehat{S} = 10474 + 548X + 3221E_2 + 4780E_3.$$

- This model implies that **on average**:
  - each extra year of experience worths $\widehat{\beta_1} \approx \$548$, with a 95% CI of \$331.5 to \$765.1.
  - completing college increases salary by $\widehat{\delta_2} = \$3221$, with a 95% CI of \$646.4 to \$5795.8.
  - completing college + advanced degree increases salary by $\widehat{\delta_3} = \$4780$, with a 95% CI of \$1908.9 to \$7651.4.
- All the 3 coefficients above are significantly different from 0 (P-value $< 5\%$)
- To compare college graduates with those with an advanced degree, need to test whether $\delta_2 < \delta_3$. What to do?

What if We Drop a Different Indicator Variable?

If we drop E2 (Bachelor's degree) instead if E1, the model becomes

$$S = \beta_0' + \beta_1' X + \delta_1' E_1 + \delta_3' E_3 + \varepsilon,$$

and the mean response $\mathrm{E}[S]$ for the 3 education levels are

Education (E)	Indicator	$\mathrm{E}(S)$
1 (HS diploma)	$E_1 = 1, E_3 = 0$	$\beta_0' + \delta_1' + \beta_1' X$
2 (Bachelor's degree)	$E_1 = E_3 = 0$	$\beta_0' + \beta_1' X$
3 (Advanced degree)	$E_2 = 0,\ E_3 = 1$	$\beta_0' + \delta_3' + \beta_1' X$

The model above means for people w/ the same years of experience (X), the diff in their mean salary are

$$\text{(HS – Bachelor's)} = \delta_1'$$

$$\text{(advanced – Bachelor's)} = \delta_3'$$

$$\text{(advanced – HS)} = \delta_3' - \delta_1'$$

Hence one can compare a advanced degree with a Bachelor's degree by testing whether $\delta_3' = 0$

How to Drop a Different Indicator Variable in R?

If not happy with R's choice of which indicator to drop, one can manually create the indicator variables E1 and E3

```
p130$E1 = ifelse(p130$E==1, 1, 0)
p130$E3 = ifelse(p130$E==3, 1, 0)
```

and fit the model

```
lm1b = lm(S ~ X + E1 + E3, data = p130)
summary(lm1b)$coef
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  13695.4     1225.0  11.180 3.626e-14
X              548.6      107.6   5.100 7.695e-06
E1           -3221.1     1275.8  -2.525 1.544e-02
E3            1559.0     1338.6   1.165 2.507e-01
```

The large P-value 0.251 for E3 (δ_3') indicate an advanced degree did **not** increase salary significantly

It Doesn't Matter Which Indicator is Dropped

If E_1 is dropped,

Education (E)	$\mathrm{E}(S)$
1 (HS)	$\beta_0 \qquad + \beta_1 X$
2 (Bachelor's)	$\beta_0 + \delta_2 + \beta_1 X$
3 (Advanced)	$\beta_0 + \delta_3 + \beta_1 X$

If E_2 is dropped,

Education (E)	$\mathrm{E}(S)$
1 (HS)	$\beta_0' + \delta_1' + \beta_1' X$
2 (Bachelor's)	$\beta_0' \qquad + \beta_1' X$
3 (Advanced)	$\beta_0' + \delta_3' + \beta_1' X$

The 2 models are equivalent in the sense that they give identical mean responses $\mathrm{E}(S)$:

$$\begin{aligned} \beta_0 &= \beta_0' + \delta_1' \\ \beta_0 + \delta_2 &= \beta_0' \\ \beta_0 + \delta_3 &= \beta_0' + \delta_3' \\ \beta_1 &= \beta_1' \end{aligned}$$

The 2 models have identical fitted values $\widehat{y}_i$, residuals e_i, SSE, SSR and hence $\widehat{\sigma}^2 = \mathrm{MSE}$, multiple and adjust R^2.

Observe the 2 models have identical fitted values $\widehat{y_i}$, residuals e_i, SSE, SSR and hence $\widehat{\sigma}^2$ = MSE, multiple and adjust R^2, and many others, despite they drop different indicators

```
> summary(lm1)
...(some output omitted)...
Residual standard error: 3620 on 42 degrees of freedom
Multiple R-squared:  0.45,  Adjusted R-squared:  0.41
F-statistic: 11.4 on 3 and 42 DF,  p-value: 0.0000129

> summary(lm1b)
...(some output omitted)...
Residual standard error: 3620 on 42 degrees of freedom
Multiple R-squared:  0.45,  Adjusted R-squared:  0.41
F-statistic: 11.4 on 3 and 42 DF,  p-value: 0.0000129
```

Model w/ Two Categorical Predictors & Their Interactions

Model w/ 2 Categorical Predictors

Now let's take another categorical predictor, management status (M), into account.

$$M = \begin{cases} 1 & \text{if manager,} \\ 0 & \text{if other} \end{cases}$$

Since M is categorical, just like E, we should create indicator variables M_0 and M_1 for the two categories, and consider the model

$$S = \beta_0 + \alpha_0 M_0 + \alpha_1 M_1 + \delta_1 E_1 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \varepsilon.$$

However, we need to drop one of M_0 and M_1 and one of E_1, E_2 and E_3.

Say we drop M_0 and E_1, and consider the model

$$S = \beta_0 + \alpha_1 M_1 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \varepsilon.$$

Model w/ No Interactions

$$S = \beta_0 + \alpha_1 M_1 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \varepsilon.$$

	E(S)	
Education (E)	Other ($M_1 = 0$)	Manager ($M_1 = 1$)
1 (HS, $E_2 = E_3 = 0$)	$\beta_0 + \beta X$	$\beta_0 + \alpha_1 + \beta X$
2 (Bachelor's, $E_2 = 1, E_3 = 0$)	$\beta_0 + \delta_2 + \beta X$	$\beta_0 + \alpha_1 + \delta_2 + \beta X$
3 (Advanced, $E_2 = 0, E_3 = 1$)	$\beta_0 + \delta_3 + \beta X$	$\beta_0 + \alpha_1 + \delta_3 + \beta X$

This model says, on average

- managers earn α_1 more than non-managers, regardless of E and X;
- completing college increases salary by δ_2, regardless of M and X;
- advanced degree earn δ_3 more than HS, regardless of M and X

The model $S = \beta_0 + \alpha_1 M_1 + \delta_2 E_2 + \delta_3 E_3 + \beta X + \varepsilon$ assumes the effect of management status (M) on salary (S) does not change with education levels E. However, from the plot below ...

```
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)")
```

Interpretation of Interactions (1)

We may consider the model below with $M * E$ interactions.

$$S = \beta_0 + \alpha_1 M_1 + \delta_2 E_2 + \delta_3 E_3 + \theta_2 (M_1 \cdot E_2) + \theta_3 (M_1 \cdot E_3) + \beta X + \varepsilon.$$

Here $(M_1 \cdot E_2)$ means the **product** of the variables M_1 and E_2.

	E(S)	
Education (E)	Other ($M_1 = 0$)	Manager ($M_1 = 1$)
1 (HS, $E_2 = E_3 = 0$)	$\beta_0 \quad + \beta X$	$\beta_0 + \alpha_1 \quad + \beta X$
2 (Bachelor's, $E_2 = 1, E_3 = 0$)	$\beta_0 + \delta_2 + \beta X$	$\beta_0 + \alpha_1 + \delta_2 + \theta_2 + \beta X$
3 (Advanced, $E_2 = 0, E_3 = 1$)	$\beta_0 + \delta_3 + \beta X$	$\beta_0 + \alpha_1 + \delta_3 + \theta_3 + \beta X$

- For HS, managers earns α_1 more than others with the same X
- For B.A. or B.S, managers earns $\alpha_1 + \theta_2$ more than others with the same X
- For advance degree, managers earns $\alpha_1 + \theta_3$ more than others with the same X

Interpretation of Interactions (2)

$$S = \beta_0 + \alpha_1 M_1 + \delta_2 E_2 + \delta_3 E_3 + \theta_2(M_1 \cdot E_2) + \theta_3(M_1 \cdot E_3) + \beta X + \varepsilon.$$

	E(S)	
Education (E)	Other ($M_1 = 0$)	Manager ($M_1 = 1$)
1 (HS, $E_2 = E_3 = 0$)	$\beta_0 + \beta X$	$\beta_0 + \alpha_1 + \beta X$
2 (Bachelor's, $E_2 = 1, E_3 = 0$)	$\beta_0 + \delta_2 + \beta X$	$\beta_0 + \alpha_1 + \delta_2 + \theta_2 + \beta X$
3 (Advanced, $E_2 = 0, E_3 = 1$)	$\beta_0 + \delta_3 + \beta X$	$\beta_0 + \alpha_1 + \delta_3 + \theta_3 + \beta X$

- Non-managers with a B.A. or B.S. earns δ_2 more than non-managers with H.S. diploma with the same X
- Managers with a B.A. or B.S. earns $\delta_2 + \theta_2$ more than managers with H.S. diploma with the same X

Effects of E on S changes with M as well.

Model Without E*M Interactions

```
lm3 = lm(S ~ as.factor(E)+as.factor(M)+X, data = p130)
summary(lm3)$coef
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)      8035.6     386.69  20.781 2.199e-23
as.factor(E)2    3144.0     361.97   8.686 7.733e-11
as.factor(E)3    2996.2     411.75   7.277 6.722e-09
as.factor(M)1    6883.5     313.92  21.928 2.901e-24
X                 546.2      30.52  17.896 5.546e-21
summary(lm3)$sigma
[1] 1027
summary(lm3)$r.squared
[1] 0.9568
```

Model With E*M Interactions

```
p130$E = as.factor(p130$E)
p130$M = as.factor(p130$M)
lm4 = lm(S ~ E+M+E*M+X, data = p130)
summary(lm4)$coef
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)     9473     80.344  117.90 2.074e-51
E2              1382     77.319   17.87 2.211e-20
E3              1731    105.334   16.43 4.013e-19
M1              3981    101.175   39.35 5.253e-33
X                497      5.566   89.28 1.021e-46
E2:M1           4903    131.359   37.32 3.934e-32
E3:M1           3066    149.330   20.53 1.635e-22
summary(lm4)$sigma
[1] 173.8
summary(lm4)$r.squared
[1] 0.9988
```

F-Test of E*M Interactions

```
anova(lm3, lm4)
Analysis of Variance Table

Model 1: S ~ as.factor(E) + as.factor(M) + X
Model 2: S ~ E + M + E * M + X
  Res.Df      RSS Df Sum of Sq   F Pr(>F)
1     41 43280719
2     39  1178168  2  42102552 697 <2e-16
```

There are significant E*M interactions!