Section 3.3: Dummies and Interactions

Jared S. Murray
The University of Texas at Austin
McCombs School of Business

Example: Detecting Sex Discrimination

Imagine you are a trial lawyer and you want to file a suit against a company for salary discrimination... you gather the following data...

	Gender	Salary
1	Male	32.0
2	Female	39.1
3	Female	33.2
4	Female	30.6
5	Male	29.0
...	...	...
208	Female	30.0

Detecting Sex Discrimination

You want to relate salary(Y) to gender(X)... how can we do that?

Gender is an example of a categorical variable. The variable gender separates our data into 2 groups or categories. The question we want to answer is: *"how is your salary related to which group you belong to..."*

Could we think about additional examples of categories potentially associated with salary?

- Level of education
- Length of experience
- What else?

Detecting Sex Discrimination

We can use regression to answer these question but we need to recode the categorical variable into a dummy variable

	Gender	Salary	Male
1	Male	32.00	1
2	Female	39.10	0
3	Female	33.20	0
4	Female	30.60	0
5	Male	29.00	1
...	...	...	
208	Female	30.00	0

Note: In R, categorical variables are known as **factors**. R will turn factor variables into dummies for you.

Detecting Sex Discrimination

```
head(salary)
```

```
## # A tibble: 6 x 10
##   Employee EducLev JobGrade YrHired YrBorn Gender YrsPrior PCJob Salary
##      <int>   <int>    <int>   <int>  <int>  <chr>    <int> <chr>  <dbl>
## 1        1       3        1      92     69   Male        1    No   32.0
## 2        2       1        1      81     57 Female        1    No   39.1
## 3        3       1        1      83     60 Female        0    No   33.2
## 4        4       2        1      87     55 Female        7    No   30.6
## 5        5       3        1      92     67   Male        0    No   29.0
## 6        6       3        1      92     71 Female        0    No   30.5
## # ... with 1 more variables: Exp <dbl>
```

`read_csv` has made Gender into a factor already, but you can also do it yourself:

```
salary$Gender = factor(salary$Gender)
```

Detecting Sex Discrimination

Now you can present the following model in court:

$$Salary_i = \beta_0 + \beta_1 Male_i + \epsilon_i$$

How do you interpret β_1?

$$\begin{aligned} E[Salary|Male = 0] &= \beta_0 \\ E[Salary|Male = 1] &= \beta_0 + \beta_1 \end{aligned}$$

β_1 is the male/female difference

Detecting Sex Discrimination

$$Salary_i = \beta_0 + \beta_1 Male_i + \epsilon_i$$

```
salaryfit = lm(Salary~Gender, data=salary)
coef(salaryfit)

## (Intercept)  GenderMale
##   37.209929    8.295513

confint(salaryfit)

##                 2.5 %   97.5 %
## (Intercept) 35.446314 38.97354
## GenderMale   5.211041 11.37998
```

$\hat{\beta}_1 = b_1 = 8.29...$ on average, a male makes approximately \$8,300 more than a female in this firm.

How should the plaintiff's lawyer use the confidence interval in his presentation?

Detecting Sex Discrimination

How can the defense attorney try to counteract the plaintiff's argument?

Perhaps, the observed difference in salaries is related to other variables in the background and NOT to policy discrimination...

Obviously, there are many other factors which we can legitimately use in determining salaries:

- education
- job productivity
- experience

How can we use regression to incorporate additional information?

Detecting Sex Discrimination

Let's add a measure of experience...

$$Salary_i = \beta_0 + \beta_1 Male_i + \beta_2 Exp_i + \epsilon_i$$

What does that mean?

$$\begin{aligned} E[Salary|Male = 0, Exp] &= \beta_0 + \beta_2 Exp \\ E[Salary|Male = 1, Exp] &= (\beta_0 + \beta_1) + \beta_2 Exp \end{aligned}$$

Detecting Sex Discrimination

	Exp	Gender	Salary	Male
1	3	Male	32.00	1
2	14	Female	39.10	0
3	12	Female	33.20	0
4	8	Female	30.60	0
5	3	Male	29.00	1
...	...	...		
208	33	Female	30.00	0

Detecting Sex Discrimination

$$Salary_i = \beta_0 + \beta_1 Male_i + \beta_2 Exp_i + \epsilon_i$$

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.83075    1.08926  24.632  < 2e-16 ***
## GenderMale   8.01189    1.19309   6.715 1.81e-10 ***
## Exp          0.98115    0.08028  12.221  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.07 on 205 degrees of freedom
## Multiple R-squared:  0.491,Adjusted R-squared:  0.486
## F-statistic: 98.86 on 2 and 205 DF,  p-value: < 2.2e-16
```

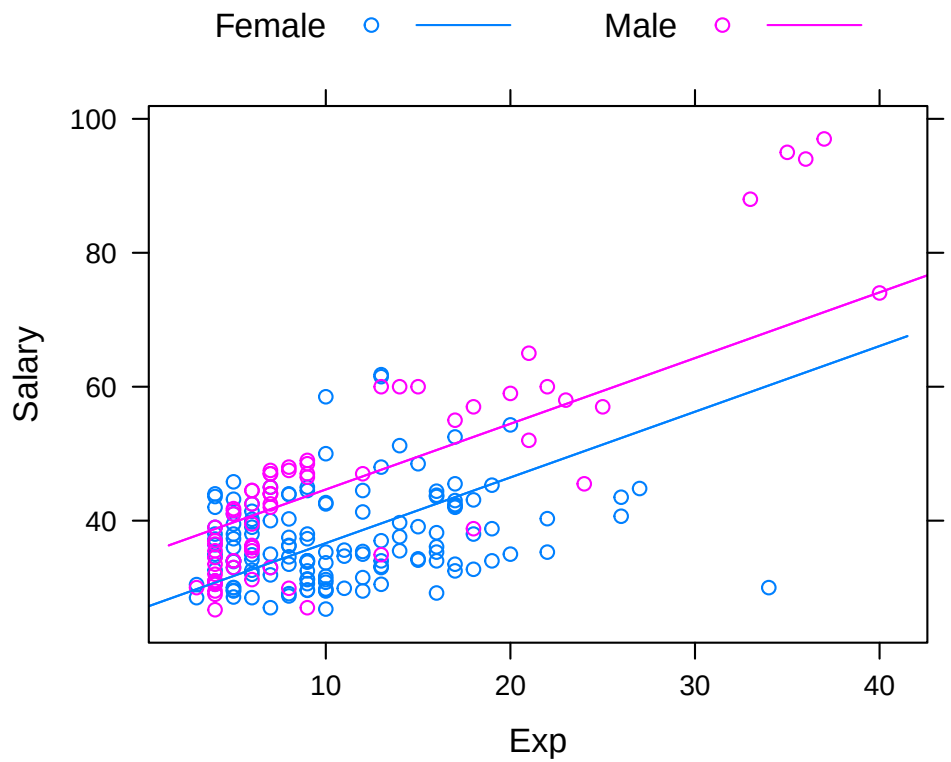
$$Salary_i = 27 + 8 Male_i + 0.98 Exp_i + \epsilon_i$$

Is this good or bad news for the defense?

Detecting Sex Discrimination

$$Salary_i = \begin{cases} 27 + 0.98 Exp_i + \epsilon_i & \text{females} \\ 35 + 0.98 Exp_i + \epsilon_i & \text{males} \end{cases}$$

```
plotModel(salaryfit_exp, Salary~Exp)
```



More than Two Categories

We can use dummy variables in situations in which there are more than two categories. Dummy variables are needed for each category except one, designated as the "base" category.

Why? *Remember that the numerical value of each category has no quantitative meaning!*

Example: House Prices

We want to evaluate the difference in house prices in different neighborhoods.

	Nbhd	SqFt	Price
1	2	1.79	114.3
2	2	2.03	114.2
3	2	1.74	114.8
4	2	1.98	94.7
5	2	2.13	119.8
6	1	1.78	114.6
7	3	1.83	151.6
8	3	2.16	150.7
...	...	...	...

Example: House Prices

Let's create the dummy variables $dn1$, $dn2$ and $dn3$...

	Nbhd	SqFt	Price	dn1	dn2	dn3
1	2	1.79	114.3	0	1	0
2	2	2.03	114.2	0	1	0
3	2	1.74	114.8	0	1	0
4	2	1.98	94.7	0	1	0
5	2	2.13	119.8	0	1	0
6	1	1.78	114.6	1	0	0
7	3	1.83	151.6	0	0	1
8	3	2.16	150.7	0	0	1
...	...	...				

(Again, R will do this for you if you make Nbhd a factor)

Example: House Prices

$$Price_i = \beta_0 + \beta_1 dn2_i + \beta_2 dn3_i + \beta_3 Size_i + \epsilon_i$$

$$\begin{aligned}
E[Price|dn2 = 0, dn3 = 0, Size] &= \beta_0 + \beta_3 Size && \text{(Nbhd 1)} \\
E[Price|dn2 = 1, dn3 = 0, Size] &= \beta_0 + \beta_1 + \beta_3 Size && \text{(Nbhd 2)} \\
E[Price|dn2 = 0, dn3 = 1, Size] &= \beta_0 + \beta_2 + \beta_3 Size && \text{(Nbhd 3)}
\end{aligned}$$

Example: House Prices

$$Price = \beta_0 + \beta_1 dn2 + \beta_2 dn3 + \beta_3 Size + \epsilon$$

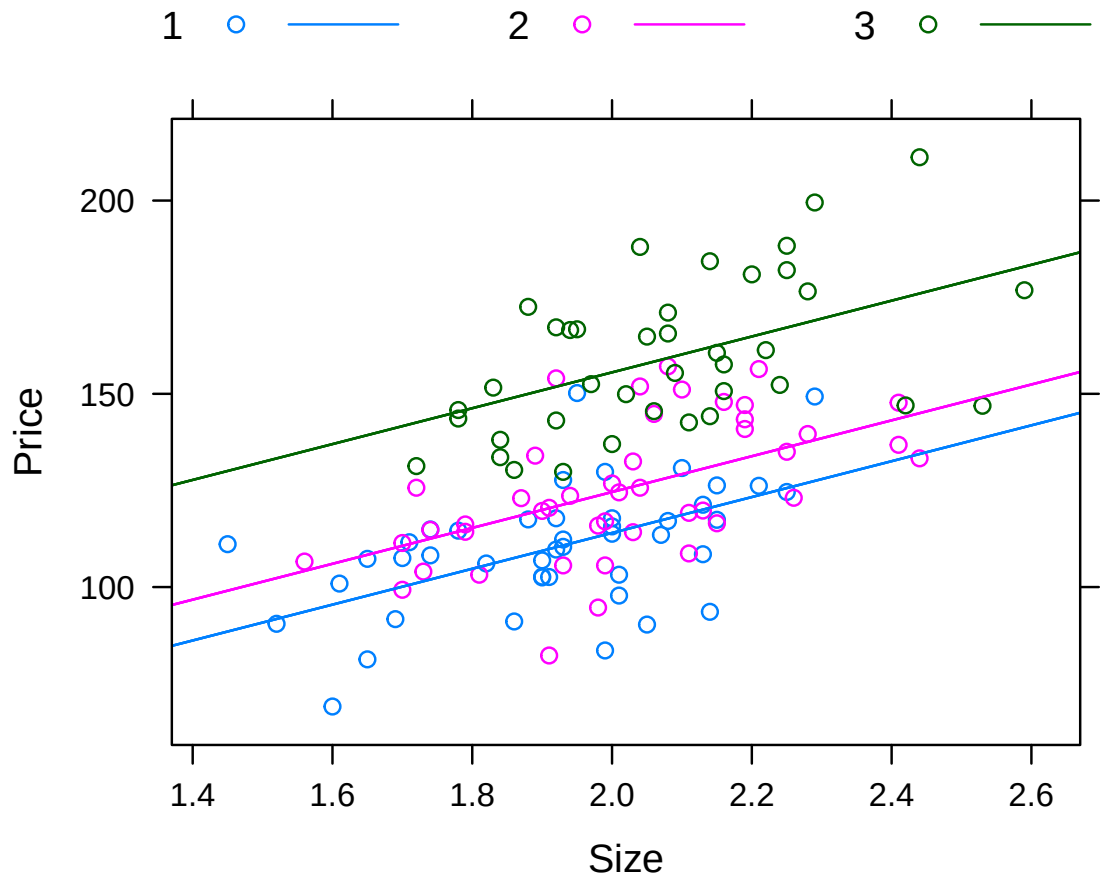
```
housing_fit = lm(Price~factor(Nbhd) + Size, data=housing)
coef(housing_fit)
```

```
##   (Intercept) factor(Nbhd)2 factor(Nbhd)3          Size
##         21.24         10.57         41.54         46.39
```

$$Price = 21.24 + 10.57 dn2 + 41.54 dn3 + 46.39 Size + \epsilon$$

Example: House Prices

```
plotModel(housing_fit, Price~Size)
```



Example: House Prices

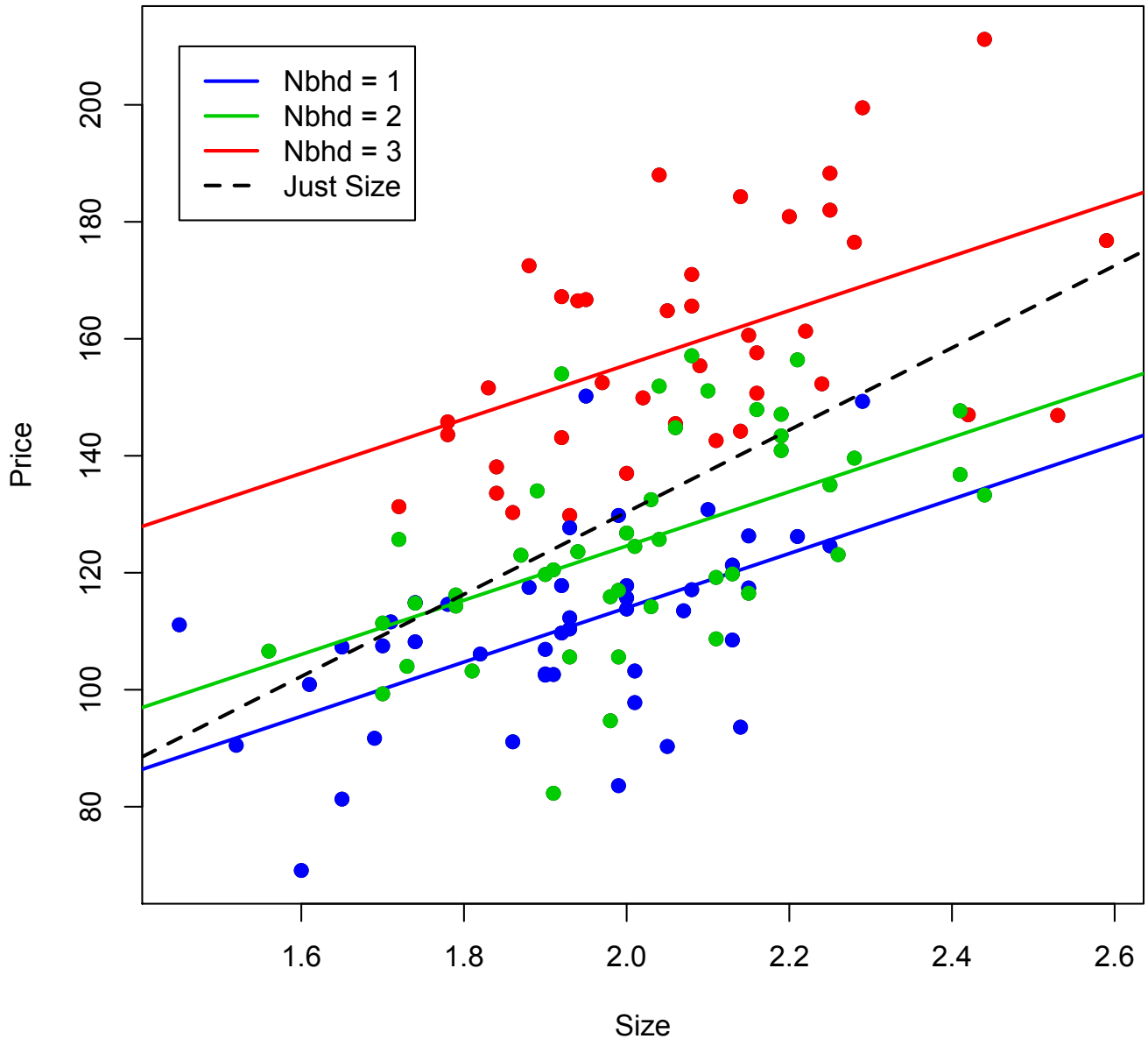
$$Price = \beta_0 + \beta_1 Size + \epsilon$$

```
lm(Price~Size, data=housing)


##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## Coefficients:
## (Intercept)         Size
##      -10.09        70.23
```

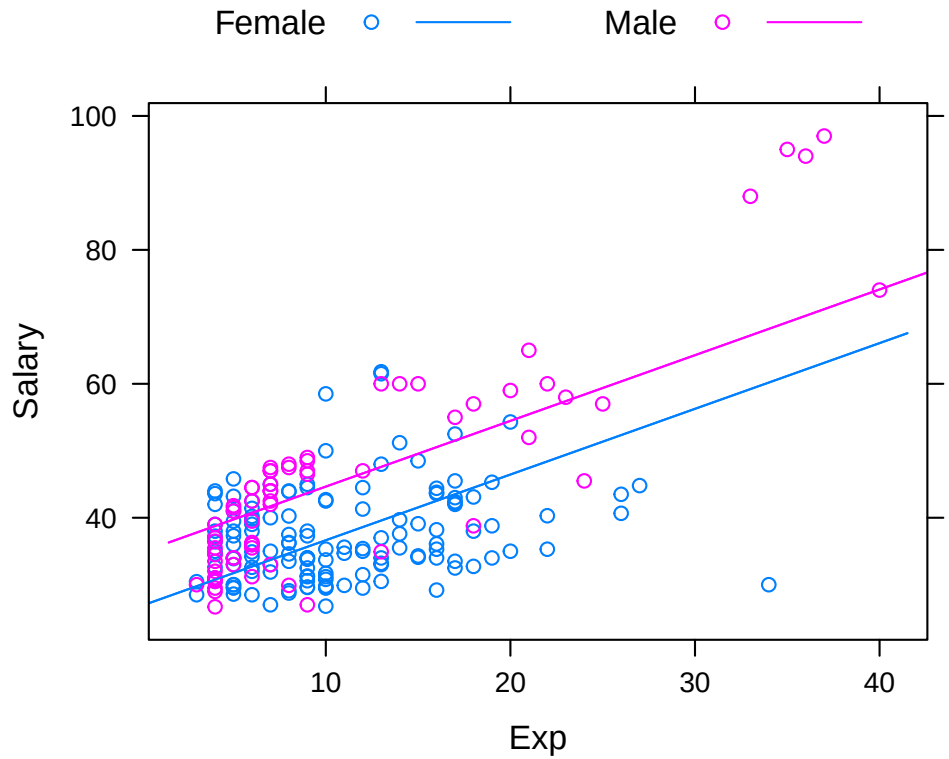
$$Price = -10.09 + 70.23 Size + \epsilon$$

Example: House Prices



Back to the Sex Discrimination Case

```
plotModel(salaryfit_exp, Salary~Exp)
```



Does it look like the effect of experience on salary is the same for males and females?

Back to the Sex Discrimination Case

Could we try to expand our analysis by allowing a different slope for each group?

Yes... Consider the following model:

$$Salary_i = \beta_0 + \beta_1 Exp_i + \beta_2 Male_i + \beta_3 Exp_i \times Male_i + \epsilon_i$$

For Females:

$$Salary_i = \beta_0 + \beta_1 Exp_i + \epsilon_i$$

For Males:

$$Salary_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Exp_i + \epsilon_i$$

Sex Discrimination Case

What do the data look like?

	Exp	Gender	Salary	Male	Exp*Male
1	3	Male	32.00	1	3
2	14	Female	39.10	0	0
3	12	Female	33.20	0	0
4	8	Female	30.60	0	0
5	3	Male	29.00	1	3
...	...	...			
208	33	Female	30.00	0	0

Sex Discrimination Case

```
salaryfit_int = lm(Salary~Gender*Exp, data=salary)
```

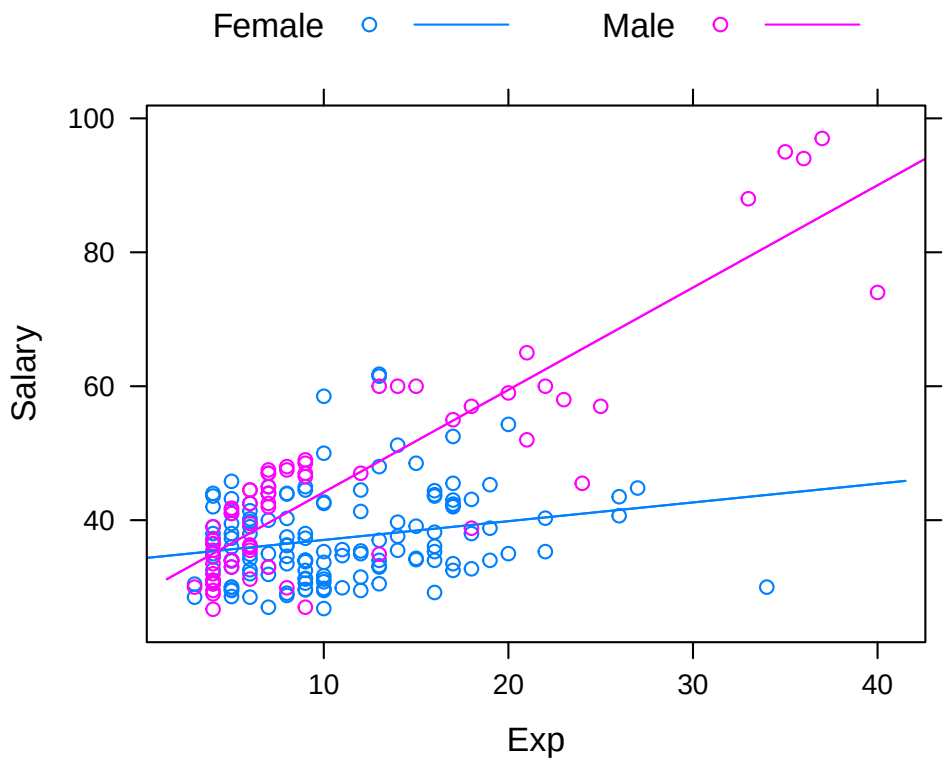
```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     34.2483     1.2274  27.903  < 2e-16 ***
## GenderMale      -5.3461     1.7766  -3.009  0.00295 **
## Exp              0.2800     0.1025   2.733  0.00684 **
## GenderMale:Exp   1.2478     0.1367   9.130  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.816 on 204 degrees of freedom
## Multiple R-squared:  0.6386,Adjusted R-squared:  0.6333
## F-statistic: 120.2 on 3 and 204 DF,  p-value: < 2.2e-16
```

Is this good or bad news for the plaintiff?

Sex Discrimination Case

$$Salary = \beta_0 + \beta_1 Sex + \beta_2 Exp + \beta_3 Exp * Male + \epsilon$$

```
plotModel(salaryfit_int, Salary~Exp)
```



$$Salary = 34 - 4 Sex + 0.28 Exp + 1.24 Exp * Male + \epsilon$$

Variable Interaction

So, the effect of experience on salary is different for males and females... in general, when the effect of the variable X_1 on Y depends on another variable X_2 we say that X_1 and X_2 interact with each other.

We can extend this notion by the inclusion of multiplicative effects by constructing interaction terms.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2) + \varepsilon$$

$$\frac{\partial \mathbb{E}[Y|X_1, X_2]}{\partial X_1} = \beta_1 + \beta_3 X_2$$

Example: College GPA and Age

Consider the relationship between undergrad and MBA grades:

A model to predict McCombs GPA from undergrad GPA could be

$$GPA^{MBA} = \beta_0 + \beta_1 GPA^{Bach} + \varepsilon$$

	Estimate	Std.Error	t value	Pr(>|t|)
BachGPA	0.26269	0.09244	2.842	0.00607 **

For every 1 point increase in college GPA, your expected GPA at McCombs increases by about .26 points.

College GPA and Age

However, this model assumes that the marginal effect of College GPA is the same for any age.

It seems that how you did in college should have less effect on your MBA GPA as you get older (farther from college).

We can account for this intuition with an interaction term:

$$GPA^{MBA} = \beta_0 + \beta_1 GPA^{Bach} + \beta_2 Age + \beta_3 (Age \times GPA^{Bach}) + \varepsilon$$

Now, the college effect is $\frac{\partial \mathbb{E}[GPA^{MBA} | GPA^{Bach} \ Age]}{\partial GPA^{Bach}} = \beta_1 + \beta_3 Age$.

Depends on Age!

College GPA and Age

```
lm(MBAGPA ~ BachGPA*Age, data=gpa)

##
## Call:
## lm(formula = MBAGPA ~ BachGPA * Age, data = gpa)
##
## Coefficients:
## (Intercept)      BachGPA          Age  BachGPA:Age
##    -0.27964      1.36936      0.10974     -0.04181
```

College GPA and Age

Without the interaction term

- Marginal effect of College GPA is $b_1 = 0.26$.

With the interaction term:

- Marginal effect is $b_1 + b_3 Age = 1.37 - 0.042 Age$.

Age	Marginal Effect
24	0.36
27	0.24
30	0.11

Interactions: Things to remember

Never try to interpret/test the main effect of a variable involved in an interaction. (You can't hold the interaction constant and vary the main effect!)

While it can occasionally make sense to omit main effects, usually if an interaction between two variables is present you should include both main effects .