Section 5: Dummy Variables and Interactions

Carlos M. Carvalho
The University of Texas at Austin
McCombs School of Business

http://faculty.mccombs.utexas.edu/carlos.carvalho/teaching/

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?

- MBA education vs. not
- legal vs. illegal immigrant
- quarterback vs wide receiver

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	Sex
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 Excel you can create the dummy variable using the formula:

=IF(Gender="Male",1,0)

Detecting Sex Discrimination

Now you can present the following model in court:

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

How do you interpret β_1?

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

β_1 is the male/female difference

Detecting Sex Discrimination

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

Regression Statistics	
Multiple R	0.346541
R Square	0.120091
Adjusted R Square	0.115819
Standard Error	10.58426
Observations	208

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3149.634	3149.6	28.1151	2.93545E-07
Residual	206	23077.47	112.03		
Total	207	26227.11			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	37.20993	0.894533	41.597	3E-102	35.44631451	38.9735426
Gender	8.295513	1.564493	5.3024	2.9E-07	5.211041089	11.3799841

$\hat{\beta}_1 = b_1 = 8.29...$ 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...

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

What does that mean?

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

Detecting Sex Discrimination

	Exp	Gender	Salary	Sex
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

$$Salary_i = \beta_0 + \beta_1 Sex_i + \beta_2 Exp + \epsilon_i$$

Regression Statistics	
Multiple R	0.701
R Square	0.491
Adjusted R Square	0.486
Standard Error	8.070
Observations	208

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.000	12876.269	6438.134	98.857	0.000
Residual	205.000	13350.839	65.126		
Total	207.000	26227.107			

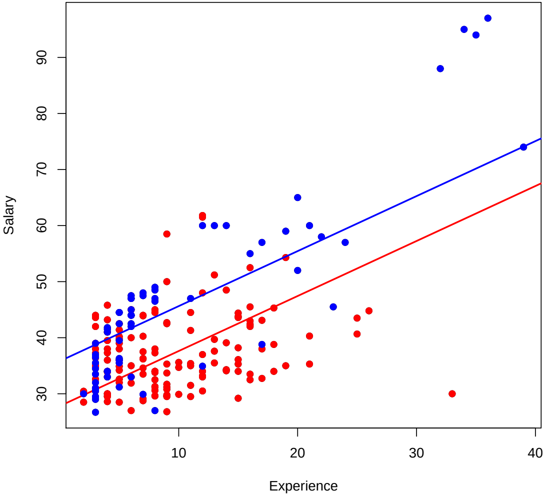
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	27.812	1.028	27.057	0.000	25.785	29.839
Sex	8.012	1.193	6.715	0.000	5.660	10.364
Exp	0.981	0.080	12.221	0.000	0.823	1.139

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

Is this good or bad news for the defense?

Detecting Sex Discrimination

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



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 a couple of 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
...	...	...				

Example: House Prices

$$Price_i = \beta_0 + \beta_1 dn1_i + \beta_2 dn2_i + \beta_3 Size_i + \epsilon_i$$

$$\begin{aligned} E[Price|dn1 = 1, Size] &= \beta_0 + \beta_1 + \beta_3 Size \quad \text{(Nbhd 1)} \\ E[Price|dn2 = 1, Size] &= \beta_0 + \beta_2 + \beta_3 Size \quad \text{(Nbhd 2)} \\ E[Price|dn1 = 0, dn2 = 0, Size] &= \beta_0 + \beta_3 Size \quad \text{(Nbhd 3)} \end{aligned}$$

Example: House Prices

$$Price = \beta_0 + \beta_1 dn1 + \beta_2 dn2 + \beta_3 Size + \epsilon$$

Regression Statistics	
Multiple R	0.828
R Square	0.685
Adjusted R Square	0.677
Standard Error	15.260
Observations	128

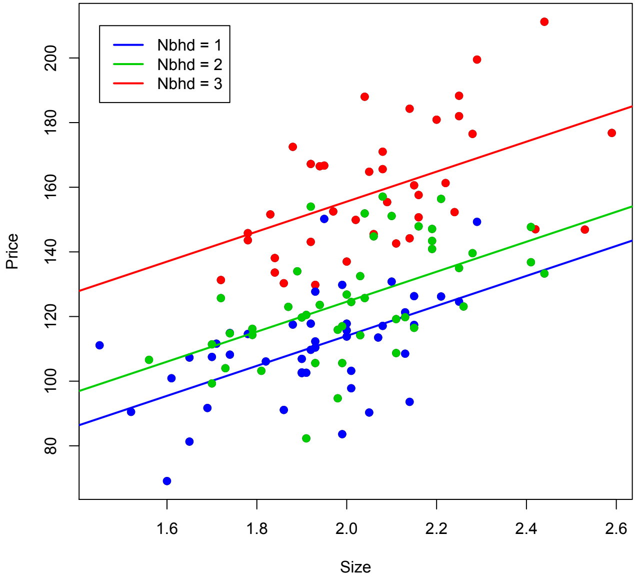
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	3	62809.1504	20936	89.9053	5.8E-31
Residual	124	28876.0639	232.87		
Total	127	91685.2143			

	Coefficients	*Standard Erroi*	*t Stat*	*P-value*	*.ower 95%*	*Jpper 95%*
Intercept	62.78	14.25	4.41	0.00	34.58	90.98
dn1	-41.54	3.53	-11.75	0.00	-48.53	-34.54
dn2	-30.97	3.37	-9.19	0.00	-37.63	-24.30
size	46.39	6.75	6.88	0.00	33.03	59.74

$$Price = 62.78 - 41.54 dn1 - 30.97 dn2 + 46.39 Size + \epsilon$$

Example: House Prices



Example: House Prices

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

Regression Statistics	
Multiple R	0.553
R Square	0.306
Adjusted R Square	0.300
Standard Error	22.476
Observations	128

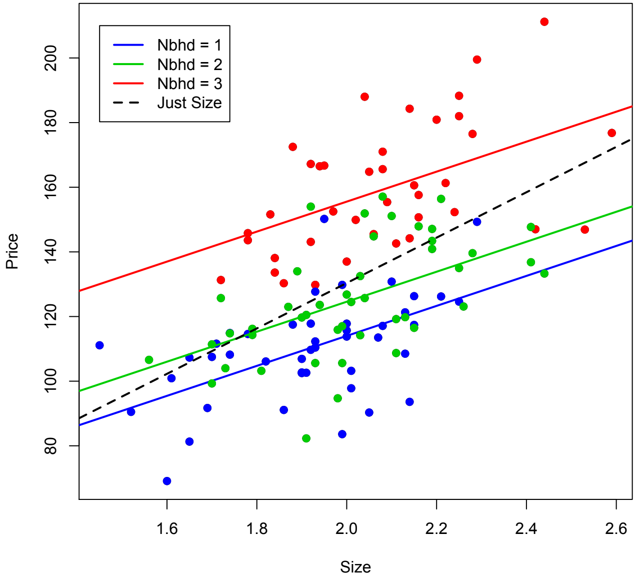
ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	28036.4	28036.36	55.501	1E-11
Residual	126	63648.9	505.1496		
Total	127	91685.2			

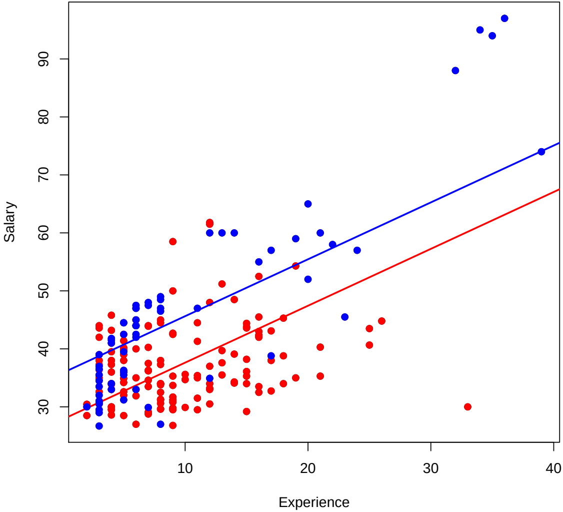
	Coefficients	*andard Err*	*t Stat*	*P-value*	*ower 95%*	*pper 95%*
Intercept	-10.09	18.97	-0.53	0.60	-47.62	27.44
size	70.23	9.43	7.45	0.00	51.57	88.88

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

Example: House Prices



Back to the Sex Discrimination Case



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:

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

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

How does the data look like?

	Exp	Gender	Salary	Sex	Exp*Sex
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

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

Regression Statistics	
Multiple R	0.7991
R Square	0.6386
Adjusted R Square	0.6333
Standard Error	6.8163
Observations	208

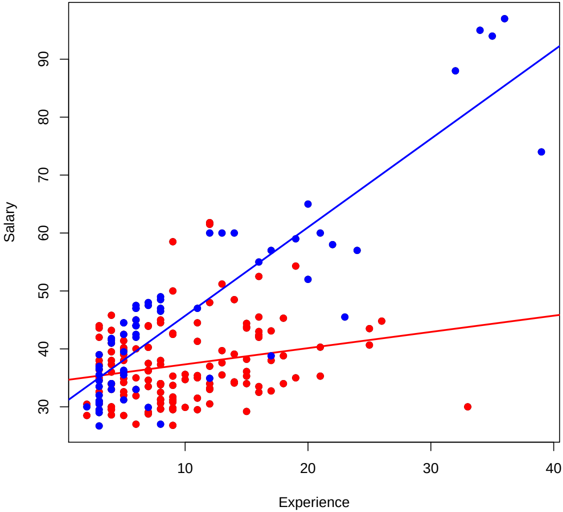
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	16748.875	5582.958	120.162	7.513E-45
Residual	204	9478.2322	46.46192		
Total	207	26227.107			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	34.528	1.138	30.342	0.000	32.285	36.772
Sex	-4.098	1.666	-2.460	0.015	-7.383	-0.814
Exp	0.280	0.102	2.733	0.007	0.078	0.482
Sex*Exp	1.248	0.137	9.130	0.000	0.978	1.517

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

Sex Discrimination Case



Is this good or bad news for the plaintiff?

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 onto 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 through interaction terms.

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

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

We will pick this up in our next section...

Example: College GPA and Age

Consider the connection between college and MBA grades:

A model to predict McCombs GPA from college 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:

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

Now, the college effect is $\frac{\partial E[GPA^{MBA} | GPA^{Bach} \;\; Age]}{\partial GPA^{Bach}} = \beta_1 + \beta_2 Age$.

Depends on Age!

College GPA and Age

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

Here, we have the interaction term but do not the main effect of age... what are we assuming?

	Estimate	Std.Error	t value	Pr(>|t|)	
BachGPA	0.455750	0.103026	4.424	4.07e-05	***
BachGPA:Age	-0.009377	0.002786	-3.366	0.00132	**

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 $b_1 + b_2 Age = 0.46 - 0.0094 Age$.

Age	Marginal Effect
25	0.22
30	0.17
35	0.13
40	0.08