Section 5: Dummy Variables and Interactions

Tengyuan Liang, Chicago Booth

https://tyliang.github.io/BUS41000/

Suggested Reading:
Statistics for Business, Part IV

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 questions 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

The data gives us the "year hired" as a measure of experience...

```
    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

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

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.70068016
R Square	0.49095268
Adjusted R Sq	0.48598637
Standard Erro	8.07007076
Observations	208

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	12876.2686	6438.13431	98.8565267	8.7642E-31
Residual	205	13350.8386	65.126042		
Total	207	26227.1072			

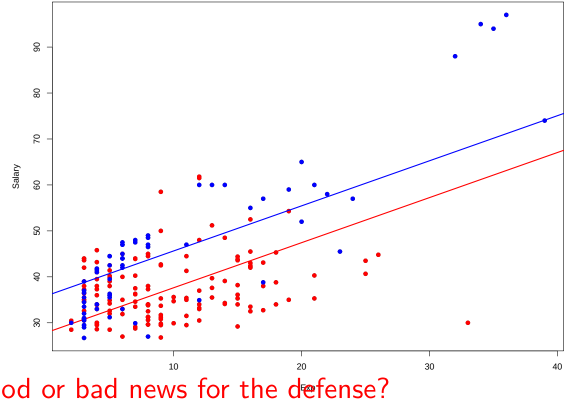
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	27.8119041	1.02789303	27.0571969	1.3985E-69	25.7853066	29.8385016
Exp	0.98115095	0.08028453	12.2209217	3.6995E-26	0.82286169	1.13944021
Sex	8.01188578	1.19308866	6.71524761	1.8094E-10	5.659588	10.3641836

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

Is this good or bad news for the defense?

Detecting Sex Discrimination

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



Is this good or bad news for the defense?

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 *dn*1, *dn*2 and *dn*3...

	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

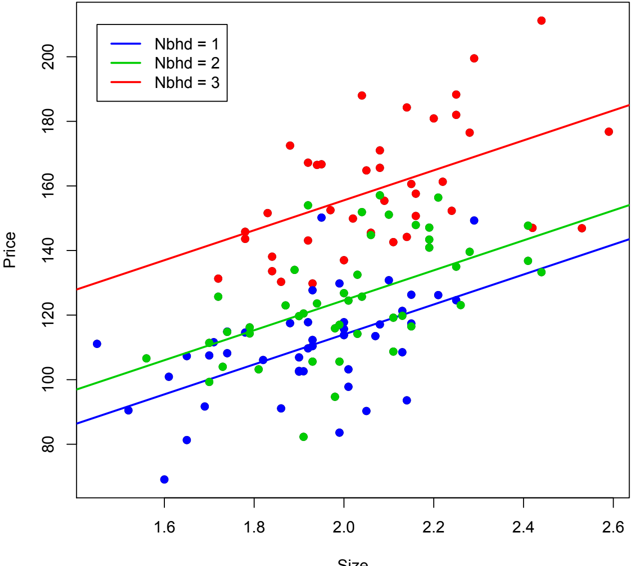
ANOVA

	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 Erro*	*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

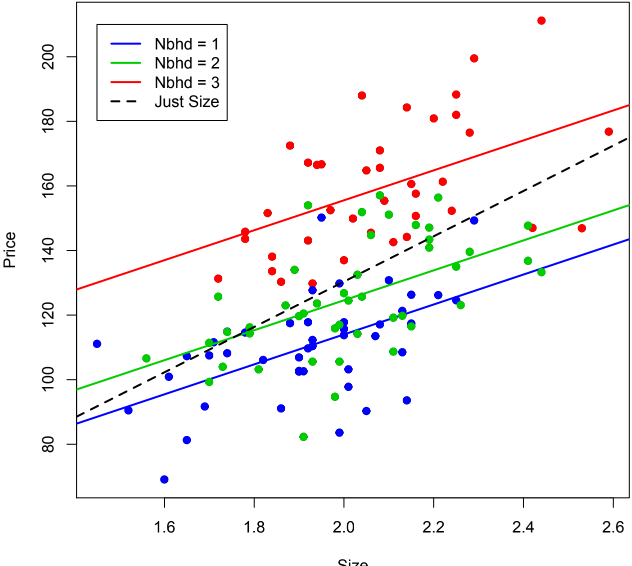
ANOVA

	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			

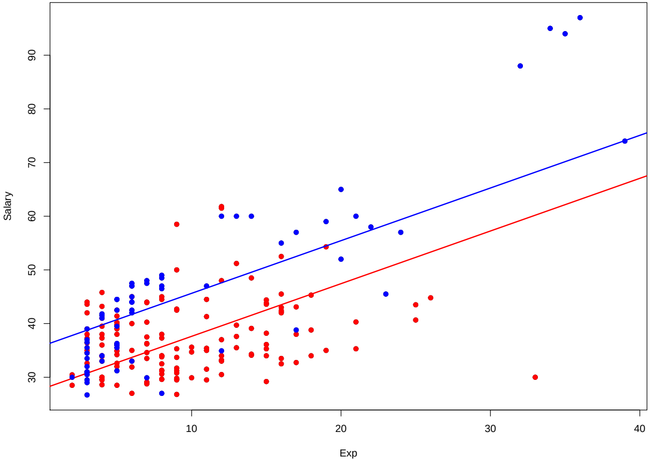
	Coefficients	*andard Err*	*t Stat*	*P-value*	*ower 95%*	*pper 95%*
Intercept	-10.09	18.97	-0.53	0.60	-47.62	27.44
size	70.23	9.43	7.45	0.00	51.57	88.88

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

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

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

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.79913035
R Square	0.63860932
Adjusted R Sq	0.63329475
Standard Err	6.81629829
Observations	208

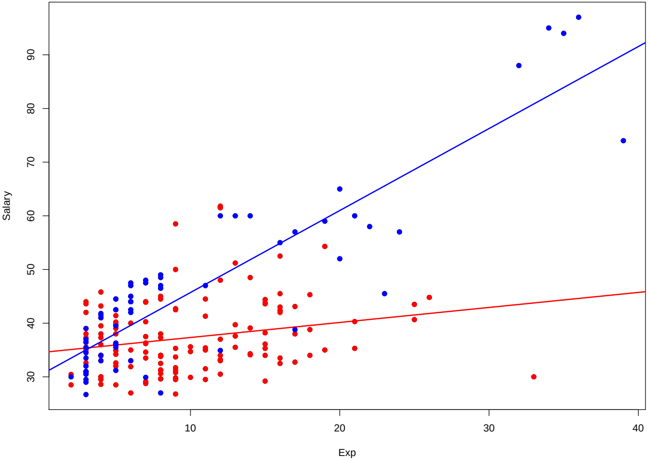
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	16748.8751	5582.95836	120.162018	7.5128E-45
Residual	204	9478.23216	46.4619224		
Total	207	26227.1072			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	34.5282796	1.13797036	30.3419852	1.4745E-77	32.284588	36.7719713
Exp	0.27996335	0.10245572	2.73253013	0.00683654	0.07795541	0.48197129
Sex	-4.0982519	1.66584202	-2.4601684	0.01471882	-7.3827274	-0.8137763
ExpSex	1.24779837	0.1366757	9.12962828	6.8335E-17	0.97832023	1.51727651

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

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.

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

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

Example: College GPA and Age

Consider the connection between college and MBA grades:

A model to predict MBA 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 MBA GPA 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 (further 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

MidCity Data

```r
# install.packages("readr")
library(readr)
MidCity <- read_csv("MidCity.csv",
                    col_types = cols(
                      Nbhd = col_factor(levels = c("1", "2", "3"))
                      )
                    )
head(MidCity)
```

```
## # A tibble: 6 x 8
##    Home Nbhd  Offers  SqFt Brick Bedrooms Bathrooms  Price
##   <dbl> <fct>  <dbl> <dbl> <chr>    <dbl>     <dbl>  <dbl>
## 1     1 2          2  1790 No           2         2 114300
## 2     2 2          3  2030 No           4         2 114200
## 3     3 2          1  1740 No           3         2 114800
## 4     4 2          3  1980 No           3         2  94700
## 5     5 2          3  2130 No           3         3 119800
## 6     6 1          2  1780 No           3         2 114600
```

Dummies for Neighbourhood

```r
reg1 = lm(Price~Nbhd+SqFt, data=MidCity)
summary(reg1)
```

```
##
## Call:
## lm(formula = Price ~ Nbhd + SqFt, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38107 -10924   -305   9643  38506
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21241.174  13133.642   1.617  0.10835
## Nbhd2       10568.698   3301.096   3.202  0.00174 **
## Nbhd3       41535.306   3533.668  11.754  < 2e-16 ***
## SqFt           46.386      6.746   6.876 2.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15260 on 124 degrees of freedom
## Multiple R-squared:  0.6851, Adjusted R-squared:  0.6774
## F-statistic: 89.91 on 3 and 124 DF,  p-value: < 2.2e-16
```

```r
MidCity = cbind(MidCity, pred1 = predict(reg1))
```

Dummies for Neighbourhood

Dummies with Interaction

```r
reg2 = lm(Price~Nbhd+SqFt+Nbhd*SqFt, data=MidCity)
summary(reg2)
```

```
##
## Call:
## lm(formula = Price ~ Nbhd + SqFt + Nbhd * SqFt, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37791 -10287    217   8989  38708
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32906.423  22784.778   1.444 0.151238
## Nbhd2       -7224.312  32569.556  -0.222 0.824831
## Nbhd3       23752.725  33848.749   0.702 0.484183
## SqFt           40.300     11.825   3.408 0.000887 ***
## Nbhd2:SqFt      9.128     16.495   0.553 0.580996
## Nbhd3:SqFt      9.026     16.827   0.536 0.592681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15360 on 122 degrees of freedom
## Multiple R-squared:  0.6861, Adjusted R-squared:  0.6732
## F-statistic: 53.32 on 5 and 122 DF,  p-value: < 2.2e-16
```

```r
MidCity = cbind(MidCity, pred2 = predict(reg2))
```

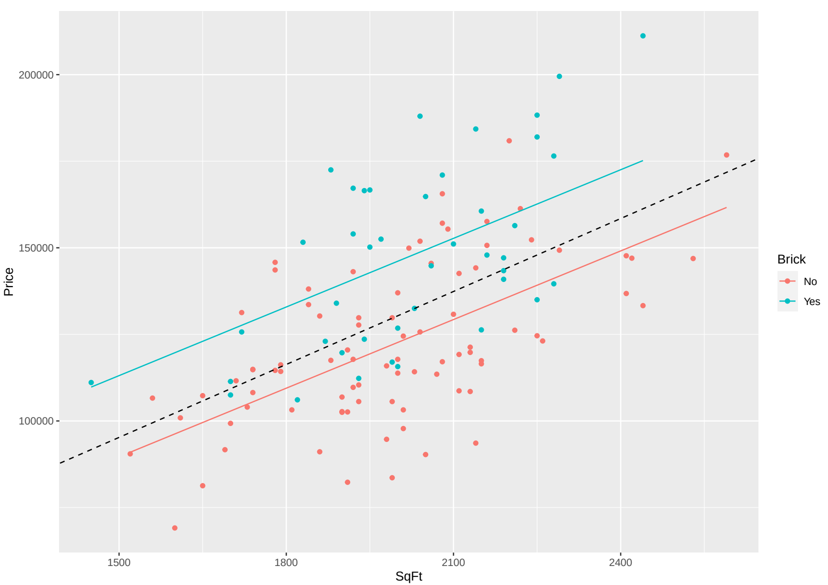
Dummies with Interaction

Dummies for Brick

```r
reg4 = lm(Price~SqFt + Brick, data=MidCity)
summary(reg4)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38412 -14665  -1772  13912  45016
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9444.289  16577.134  -0.570     0.57
## SqFt           66.058      8.265   7.992 7.54e-13 ***
## BrickYes    23445.096   3709.805   6.320 4.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19640 on 125 degrees of freedom
## Multiple R-squared:  0.4739, Adjusted R-squared:  0.4655
## F-statistic:  56.3 on 2 and 125 DF,  p-value: < 2.2e-16
```

Dummies for Brick



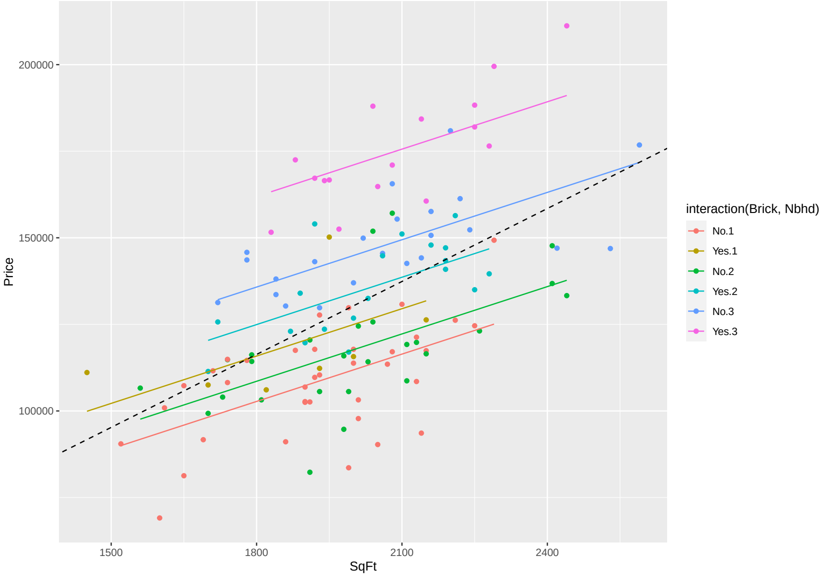
Crazy interaction

Now let's look at a crazy interaction *Brick* $*$ *Nbhd*. How many categories? Answer $2 * 3 = 6$.

```r
reg5 = lm(Price~SqFt+Brick*Nbhd, data=MidCity)
summary(reg5)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick * Nbhd, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31279  -7405   -847   6889  35775
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    20735.558  10766.923   1.926   0.0565 .
## SqFt              45.562      5.484   8.308 1.64e-13 ***
## BrickYes       13106.669   5106.897   2.566   0.0115 *
## Nbhd2           5820.591   3187.082   1.826   0.0703 .
## Nbhd3          33023.314   3375.878   9.782  < 2e-16 ***
## BrickYes:Nbhd2  3267.031   6335.286   0.516   0.6070
## BrickYes:Nbhd3 13053.182   6506.989   2.006   0.0471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12350 on 121 degrees of freedom
## Multiple R-squared:  0.7986, Adjusted R-squared:  0.7886
## F-statistic: 79.95 on 6 and 121 DF,  p-value: < 2.2e-16
```

Crazy interaction

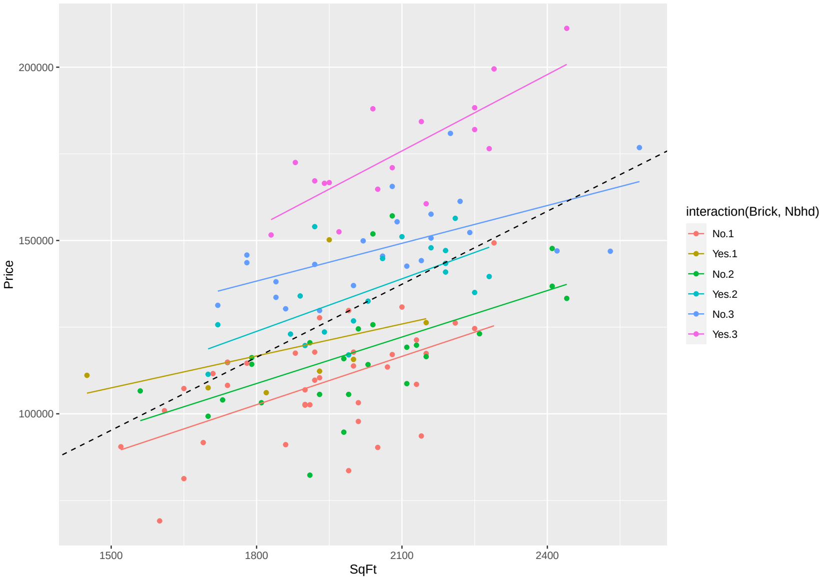


Crazy interaction

```r
reg6 = lm(Price~SqFt+ Brick*Nbhd + SqFt*Brick*Nbhd, data=MidCity)
summary(reg6)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick * Nbhd + SqFt * Brick * Nbhd,
##     data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31359  -7173   -781   6906  35843
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            18969.783  20764.749   0.914   0.3628
## SqFt                      46.478     10.717   4.337  3.1e-05 ***
## BrickYes               42464.160  46497.577   0.913   0.3630
## Nbhd2                   9323.775  30588.472   0.305   0.7611
## Nbhd3                  53901.413  31594.024   1.706   0.0907 .
## BrickYes:Nbhd2        -38015.319  62223.215  -0.611   0.5424
## BrickYes:Nbhd3        -93694.197  64939.291  -1.443   0.1518
## SqFt:BrickYes            -15.773     24.704  -0.638   0.5244
## SqFt:Nbhd2                -1.784     15.469  -0.115   0.9084
## SqFt:Nbhd3               -10.133     15.658  -0.647   0.5188
## SqFt:BrickYes:Nbhd2       21.657     32.015   0.676   0.5001
## SqFt:BrickYes:Nbhd3       52.858     32.843   1.609   0.1102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12430 on 116 degrees of freedom
## Multiple R-squared:  0.8045, Adjusted R-squared:  0.7859
## F-statistic: 43.39 on 11 and 116 DF,  p-value: < 2.2e-16
```

Crazy interaction



Merging Neibhorhood 1 and 2

```r
MidCity <- read_csv("MidCity.csv", col_types = cols(Nbhd = col_factor(levels = c("1", "2", "3"))))
# Merge Nbhd 1&2
MidCity = cbind(MidCity, NbhdNew = MidCity$Nbhd)
levels(MidCity$NbhdNew) <- c("1&2", "1&2", "3")
summary(lm(Price~SqFt+NbhdNew+Brick+Bedrooms+Bathrooms, data = MidCity))
```

```
##
## Call:
## lm(formula = Price ~ SqFt + NbhdNew + Brick + Bedrooms + Bathrooms,
##     data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34382  -7364    -53   7789  35778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16374.106  10531.829   1.555  0.12260
## SqFt           37.111      6.427   5.774 6.03e-08 ***
## NbhdNew3    31046.000   2698.846  11.503  < 2e-16 ***
## BrickYes    19486.156   2353.868   8.278 1.84e-13 ***
## Bedrooms     2280.483   1907.399   1.196  0.23417
## Bathrooms    6972.212   2584.471   2.698  0.00797 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12260 on 122 degrees of freedom
## Multiple R-squared:  0.7999, Adjusted R-squared:  0.7917
## F-statistic: 97.53 on 5 and 122 DF,  p-value: < 2.2e-16
```

Merging Neibhorhood 1 and 2

```
coeff = coefficients(lm(Price~SqFt, data=MidCity))
reg2 = lm(Price~NbhdNew+SqFt, data=MidCity)
summary(reg2)
```

```
##
## Call:
## lm(formula = Price ~ NbhdNew + SqFt, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35396  -9610  -1762   8778  38551
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18152.749  13574.154   1.337    0.184
## NbhdNew3    35699.135   3137.188  11.379  < 2e-16 ***
## SqFt           50.675      6.852   7.396 1.78e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15810 on 125 degrees of freedom
## Multiple R-squared:  0.659,  Adjusted R-squared:  0.6536
## F-statistic: 120.8 on 2 and 125 DF,  p-value: < 2.2e-16
```

```
MidCity = cbind(MidCity, pred2 = predict(reg2))
```

Merging Neibhorhood 1 and 2

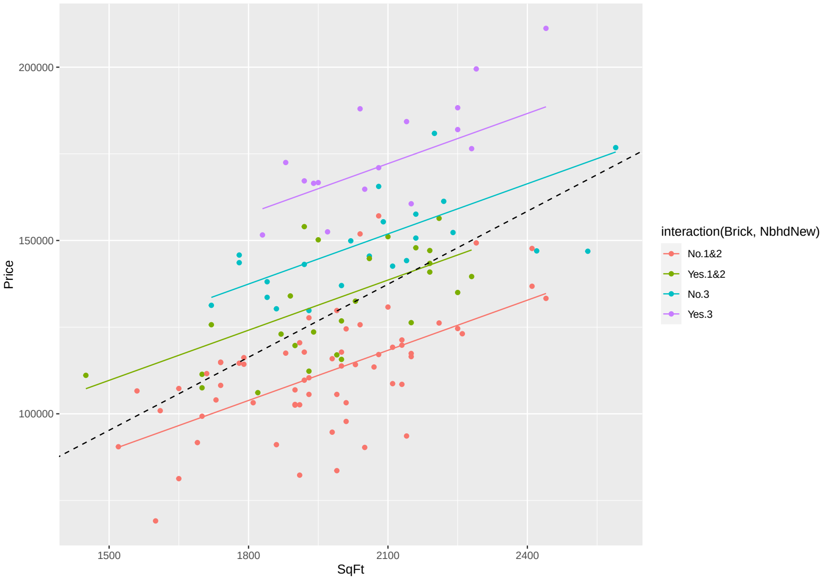
Merging Neibhorhood 1 and 2

```r
reg5 = lm(Price~SqFt+Brick+NbhdNew, data=MidCity)
summary(reg5)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick + NbhdNew, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29415  -7450     47   8343  39744
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17039.80   10861.84   1.569    0.119
## SqFt           48.23       5.49   8.785 1.07e-14 ***
## BrickYes    20271.33    2401.53   8.441 6.96e-14 ***
## NbhdNew3    33585.50    2522.60  13.314  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12650 on 124 degrees of freedom
## Multiple R-squared:  0.7834, Adjusted R-squared:  0.7782
## F-statistic: 149.5 on 3 and 124 DF,  p-value: < 2.2e-16
```

```r
MidCity = cbind(MidCity, pred5 = predict(reg5))
```

Merging Neibhorhood 1 and 2



Merging Neibhorhood 1 and 2

```r
reg6 = lm(Price~SqFt+ Brick*NbhdNew + SqFt*Brick*NbhdNew, data=MidCity)
summary(reg6)
```

```
##
## Call:
## lm(formula = Price ~ SqFt + Brick * NbhdNew + SqFt * Brick *
##     NbhdNew, data = MidCity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30285  -6983   -715   8294  38889
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             18237.214  15123.806   1.206   0.2302
## SqFt                       48.064      7.680   6.258 6.25e-09 ***
## BrickYes                10090.665  29245.328   0.345   0.7307
## NbhdNew3                54633.983  28398.755   1.924   0.0567 .
## BrickYes:NbhdNew3      -61320.701  54307.890  -1.129   0.2611
## SqFt:BrickYes               3.624     14.717   0.246   0.8059
## SqFt:NbhdNew3             -11.720     13.848  -0.846   0.3991
## SqFt:BrickYes:NbhdNew3     33.461     26.341   1.270   0.2064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12550 on 120 degrees of freedom
## Multiple R-squared:  0.7939, Adjusted R-squared:  0.7819
## F-statistic: 66.03 on 7 and 120 DF,  p-value: < 2.2e-16
```

```r
MidCity = cbind(MidCity, pred6 = predict(reg6))
```

Merging Neibhorhood 1 and 2