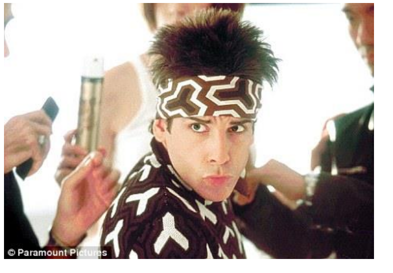
Stat 104: Quantitative Methods for Economists

Class 36: Dummy Variables

Variable selection

- If we have k variables, and assuming a constant term in each model, there are 2^k-1 possible subsets of variables, not counting the null model with no variables. How do we select a subset for our model?
- Two main approaches: Stepwise regressions and all possible regressions.
- A point to note-modelling is hard.



Stepwise Regression

- A full regression course is required to fully understand modeling, but it will be beneficial to begin the thought process of how to work with a lot of variables.
- One easy way to do this is to perform something called "backward stepwise regression".
- Under this scheme, you start with all the variables in the model, and remove them one by one.

Variable Selection: Backward Stepwise Regression

The way hypothesis testing works, you are only allowed to remove *one variable at a time* from the model.

So one way we build models as follows:

- Start with all variables in the model
- at each step, delete the least important variable from the remaining ones based on largest p-value above 0.05.
- stop when you can't delete any more.

Example: Football data; what variables contribute to a winning season ?

Column	Count	Name
C1	28	wins
C2	28	rush
C3	28	pass
C4	28	patt
C5	28	pcomp
C6	28	pint
C7	28	penalty
C8	28	fumble
C9	28	rushopp
C10	28	passopp
C11	28	pattopp
C12	28	pcompopp
C13	28	piopp

The Full Model

```
> fit=lm(wins~rush+pass+patt+pcomp+pint+penalty+fumbles+rushopp+passopp+pattopp+pcompopp+piopp)
> summary(fit)

Call:
lm(formula = wins ~ rush + pass + patt + pcomp + pint + penalty + 
    fumbles + rushopp + passopp + pattopp + pcompopp + piopp)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88194 -0.96564  0.09151  0.75299  2.61849 

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)   
(Intercept) -2.05583056  9.90592844  -0.208  0.83838   
rush         0.00152605  0.00143588   1.063  0.30469   
pass         0.00289591  0.00142320   2.035  0.05994 . 
patt        -0.01161437  0.01950896  -0.595  0.56049   
pcomp        0.00911988  0.02916689   0.313  0.75883   
pint        -0.06647342  0.10589890  -0.628  0.53964   
penalty     -0.00097561  0.00466386  -0.209  0.83712   
fumbles     -0.01763890  0.09579048  -0.184  0.85637   
rushopp      0.00004805  0.00177364   0.027  0.97874   
passopp     -0.00590162  0.00151141  -3.905  0.00141 **
pattopp      0.06102059  0.02325585   2.624  0.01917 * 
pcompopp    -0.02233248  0.01909735  -1.169  0.26049   
piopp       -0.07731961  0.11164524  -0.693  0.49918   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 1.578 on 15 degrees of freedom
Multiple R-squared:  0.8738,    Adjusted R-squared:  0.7729 
F-statistic: 8.659 on 12 and 15 DF,  p-value: 0.0001019
```

Remove RUSHOPP (why?)

```
> fit1=lm(wins~rush+pass+patt+pcomp+pint+penalty+fumbles+passopp+pattopp+pcompopp+piopp)
> summary(fit1)

Call:
lm(formula = wins ~ rush + pass + patt + pcomp + pint + penalty +
    fumbles + passopp + pattopp + pcompopp + piopp)

Residuals:
     Min       1Q   Median       3Q      Max
-1.85945 -0.97045  0.09558  0.74859  2.62532

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.9319164  8.5078391  -0.227 0.823241
rush         0.0015089  0.0012475   1.209 0.244040
pass         0.0028924  0.0013725   2.107 0.051205 .
patt        -0.0114803  0.0182716  -0.628 0.538666
pcomp        0.0089597  0.0276548   0.324 0.750148
pint        -0.0672078  0.0991226  -0.678 0.507442
penalty     -0.0009653  0.0045009  -0.214 0.832886
fumbles     -0.0176719  0.0927435  -0.191 0.851278
passopp     -0.0058881  0.0013816  -4.262 0.000596 ***
pattopp      0.0608653  0.0218231   2.789 0.013135 *
pcompopp    -0.0222535  0.0182747  -1.218 0.240984
piopp       -0.0773400  0.1081002  -0.715 0.484643
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.528 on 16 degrees of freedom
Multiple R-squared:  0.8738,    Adjusted R-squared:  0.7871
F-statistic: 10.07 on 11 and 16 DF,  p-value: 0.0000305
```

Remove the variable with the highest p-value above .05 then refit

Easier way to do this removal

- There is a model update command in R

```
> fit=lm(wins~rush+pass+patt+pcomp+pint+penalty+fumbles+rushopp+passopp+pattopp+pcompopp+piopp)
> fit1=update(fit,.~.-rushopp)
> summary(fit1)

Call:
lm(formula = wins ~ rush + pass + patt + pcomp + pint + penalty +
    fumbles + passopp + pattopp + pcompopp + piopp)

Residuals:
     Min       1Q   Median       3Q      Max
-1.85945 -0.97045  0.09558  0.74859  2.62532

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.9319164  8.5078391  -0.227 0.823241
rush         0.0015089  0.0012475   1.209 0.244040
pass         0.0028924  0.0013725   2.107 0.051205 .
patt        -0.0114803  0.0182716  -0.628 0.538666
pcomp        0.0089597  0.0276548   0.324 0.750148
pint        -0.0672078  0.0991226  -0.678 0.507442
penalty     -0.0009653  0.0045009  -0.214 0.832886
fumbles     -0.0176719  0.0927435  -0.191 0.851278
passopp     -0.0058881  0.0013816  -4.262 0.000596 ***
pattopp      0.0608653  0.0218231   2.789 0.013135 *
pcompopp    -0.0222535  0.0182747  -1.218 0.240984
piopp       -0.0773400  0.1081002  -0.715 0.484643
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.528 on 16 degrees of freedom
Multiple R-squared:  0.8738,    Adjusted R-squared:  0.7871
F-statistic: 10.07 on 11 and 16 DF,  p-value: 0.0000305
```

Now remove FUMBLES

```
> fit2=update(fit1,.~.-fumbles)
> summary(fit2)

Call:
lm(formula = wins ~ rush + pass + patt + pcomp + pint + penalty +
    passopp + pattopp + pcompopp + piopp)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9524 -0.9999  0.1174  0.7394  2.5911 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.491490   7.952383  -0.188 0.853448    
rush         0.001499   0.001211   1.239 0.232350    
pass         0.002954   0.001296   2.279 0.035874 *  
patt        -0.011864   0.017638  -0.673 0.510242    
pcomp        0.008403   0.026709   0.315 0.756890    
pint        -0.064290   0.095117  -0.676 0.508189    
penalty     -0.001362   0.003876  -0.351 0.729567    
passopp     -0.005902   0.001340  -4.405 0.000387 ***
pattopp      0.060776   0.021191   2.868 0.010660 *  
pcompopp    -0.023211   0.017065  -1.360 0.191542    
piopp       -0.072794   0.102403  -0.711 0.486809    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.484 on 17 degrees of freedom
Multiple R-squared:  0.8736,    Adjusted R-squared:  0.7992
F-statistic: 11.74 on 10 and 17 DF,  p-value: 0.000008598
```

After a while get to this

```
> fit10=lm(wins~rush+pass+pint+passopp+pattopp)
> summary(fit10)

Call:
lm(formula = wins ~ rush + pass + pint + passopp + pattopp)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0626 -1.0763  0.0480  0.6624  3.2261 

Coefficients:
              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) -5.7428336  5.2888387  -1.086   0.28930    
rush         0.0020463  0.0010463   1.956   0.06330 .  
pass         0.0029797  0.0005326   5.595 0.0000126 ***
pint        -0.1106437  0.0620384  -1.783   0.08831 .  
passopp     -0.0053287  0.0009882  -5.392 0.0000205 ***
pattopp      0.0401539  0.0120817   3.324   0.00308 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.393 on 22 degrees of freedom
Multiple R-squared:  0.8558,    Adjusted R-squared:  0.823 
F-statistic: 26.11 on 5 and 22 DF,  p-value: 0.00000001461
```

Drop PINT

```
> fit10=lm(wins~rush+pass+passopp+pattopp)
> summary(fit10)

Call:
lm(formula = wins ~ rush + pass + passopp + pattopp)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6100 -1.1772  0.2459  0.8287  2.3167 

Coefficients:
               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept) -10.2492232   4.8615000  -2.108   0.04611 *  
rush          0.0025855   0.0010481   2.467   0.02150 *  
pass          0.0029576   0.0005571   5.309 0.0000217 ***
passopp      -0.0055535   0.0010255  -5.415 0.0000168 ***
pattopp       0.0448382   0.0123391   3.634   0.00139 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 1.457 on 23 degrees of freedom
Multiple R-squared:  0.8349,    Adjusted R-squared:  0.8062 
F-statistic: 29.09 on 4 and 23 DF,  p-value: 0.00000001067
```

Why do we stop here ?

How does this compare to previous model?

R can (sort of) do it automatically

- There are better methods than this but someone wrote a function called model.select to do this.
- Load the function into R as follows

```
source("http://people.fas.harvard.edu/~mparzen/stat100/model_select.txt")
```

Running model.select()

```
> model.select(fit,verbose=FALSE)

Call:
lm(formula = wins ~ rush + pass + passopp + pattopp)

Coefficients:
(Intercept)         rush         pass      passopp      pattopp
  -10.24922      0.00259      0.00296     -0.00555      0.04484

 model.select(fit,verbose=TRUE)
-------------STEP  9 -------------
 The drop statistics :
Single term deletions

Model:
wins ~ rush + pass + passopp + pattopp
        Df Sum of Sq   RSS  AIC F value   Pr(>F)
<none>                48.9 25.6
rush     1      12.9  61.8 30.2    6.09   0.0215 *
pass     1      59.9 108.7 46.0   28.19 0.000022 ***
passopp  1      62.3 111.1 46.6   29.32 0.000017 ***
pattopp  1      28.0  76.9 36.3   13.20   0.0014 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
lm(formula = wins ~ rush + pass + passopp + pattopp)

Coefficients:
(Intercept)         rush         pass      passopp      pattopp
  -10.24922      0.00259      0.00296     -0.00555      0.04484
```

The built in method in R

```
step(fit,model="backward")

Step:  AIC=23.81
wins ~ rush + pass + pint + passopp + pattopp

          Df Sum of Sq     RSS    AIC
<none>                  42.685 23.806
- pint     1     6.171  48.856 25.587
- rush     1     7.422  50.106 26.294
- pattopp  1    21.431  64.116 33.198
- passopp  1    56.419  99.104 45.391
- pass     1    60.736 103.421 46.585

Call:
lm(formula = wins ~ rush + pass + pint + passopp + pattopp)

Coefficients:
(Intercept)         rush         pass         pint      passopp      pattopp
  -5.742834     0.002046     0.002980    -0.110644    -0.005329     0.040154
```

The `step` function in R minimizes AIC – details in more advanced courses.

All Subsets Regression

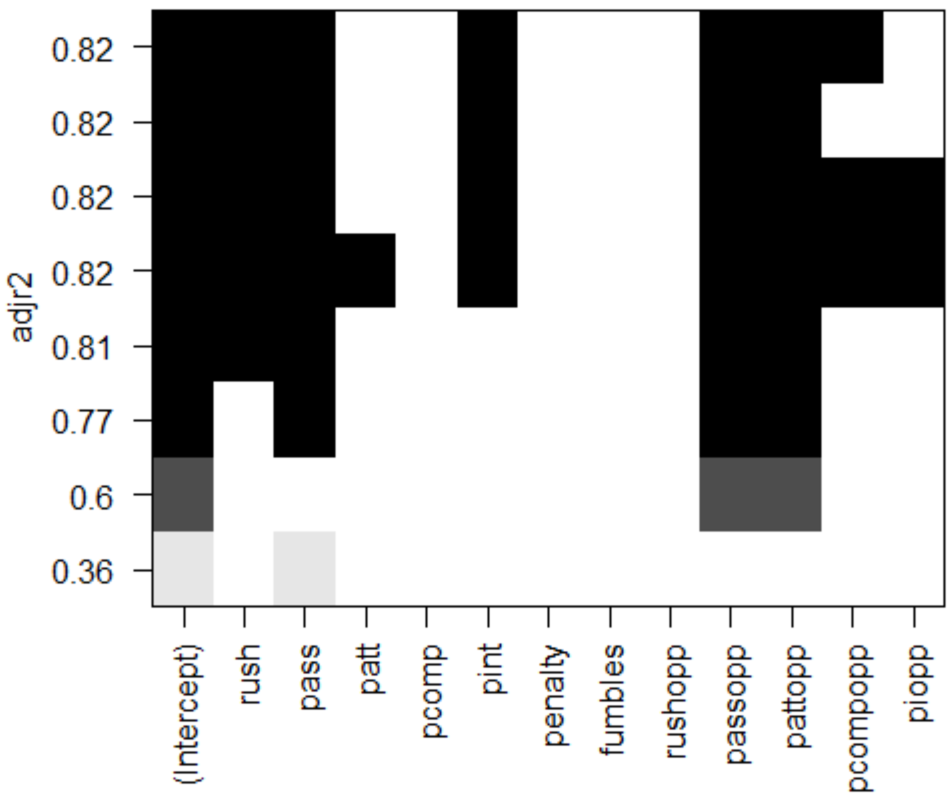
- This procedure runs all 1 variable models, all 2 variable models, all 3 variable models and so on.
- The idea is to pick the model that has the adjusted R-2 [or some other measure].
- The output looks cool at least.

All subsets regression

- The function is call `regsubsets` and is in the `leaps` package:

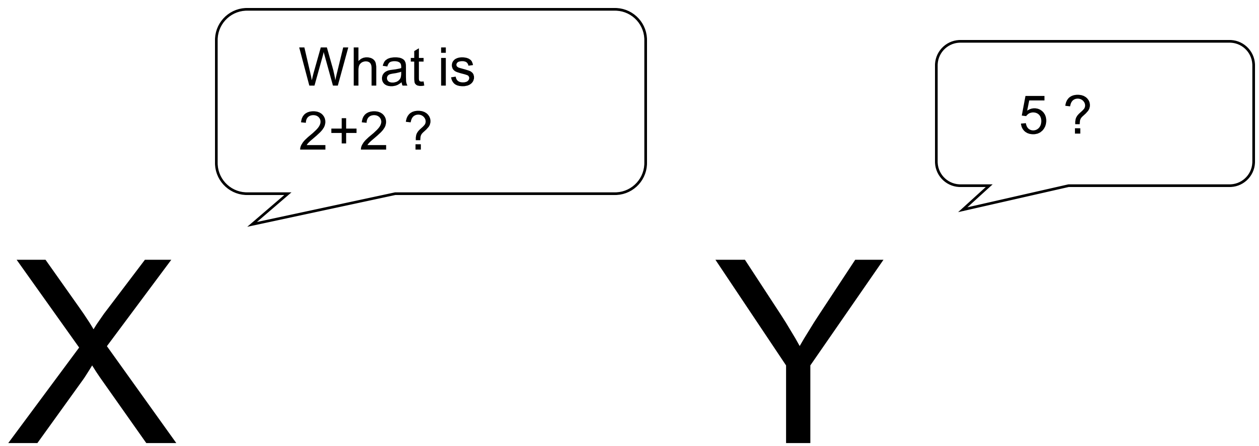
```
library(leaps)

fit=regsubsets(wins~rush+pass+patt+pcomp+pint+penalty+fumb
les+rushopp+passopp+pattopp+pcompopp+piopp,data=mydata)

plot(fit,scale="adjr2")
```

The Output



adjr2
0.82
0.82
0.82
0.82
0.81
0.77
0.6
0.36
(Intercept)
rush
pass
patt
pcomp
pint
penalty
fumbles
rushopp
passopp
pattopp
pcompopp
piopp

Dummy variables



Dummy Variables

- Many variables in the world are fundamentally *discrete* -- - e.g., you are male or female, have your tonsils or not, eat/drink Four Loko or not, etc.

- A **dummy variable** (or **indicator variable**) is a variable that takes on a value of zero or one.

Note: This section considers X variables that are indicator variables. When Y is an indicator variable, least-squares is not the appropriate method (logistic regression is, which is covered in Stat 139).

Some Examples

- categorical variable

  (e.g., 1 if female, 0 if not)

- temporal variable

  (e.g., 1 if Monday, 0 if not)

- spatial variable

  (e.g., 1 if Midwest, 0 if not)

- qualitative variable

  (e.g., 1 if “good at beer pong,” 0 if not)

Example

- A certain drug (Vitamin L) is suspected of having the unfortunate side effect of raising blood pressure.
- To test this, 10 women were randomly sampled, 6 of whom took the drug once per day and 4 who didn't take the drug at all.
- Define the dummy (indicator) variable
- D=1 if took drug, 0 otherwise

See https://www.youtube.com/watch?v=7W8M687hz-c

Example

- Our data looks as follows

BP	D	Age
85	0	30
95	1	40
90	1	40
75	0	20
100	1	60
90	0	40
90	0	50
90	1	30
100	1	60
85	1	30

Here is the regression output

At the same age level, those on the drug had a BP 4.651 times higher than those not on the drug

- How do we interpret this ?

```
> fit=lm(bp~d+age)
> summary(fit)

Call:
lm(formula = bp ~ d + age)

Residuals:
   Min     1Q Median     3Q    Max
-3.372 -1.802 -0.698  2.471  3.140

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept)   69.535      2.905   23.93 0.000000057 ***
d              4.651      1.885    2.47     0.04301 *
age            0.442      0.073    6.05     0.00052 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.76 on 7 degrees of freedom
Multiple R-squared:  0.893,     Adjusted R-squared:  0.862
F-statistic: 29.2 on 2 and 7 DF,  p-value: 0.0004
```

The fitted line

- (Rounded) the fitted line is

$$\hat{y} = 70 + 5D + 0.44Age$$

- This takes on two forms for D=0,1
- For D=0 $\quad \hat{y} = 70 + 0.44Age$
- For D=1 $\quad \hat{y} = 75 + 0.44Age$

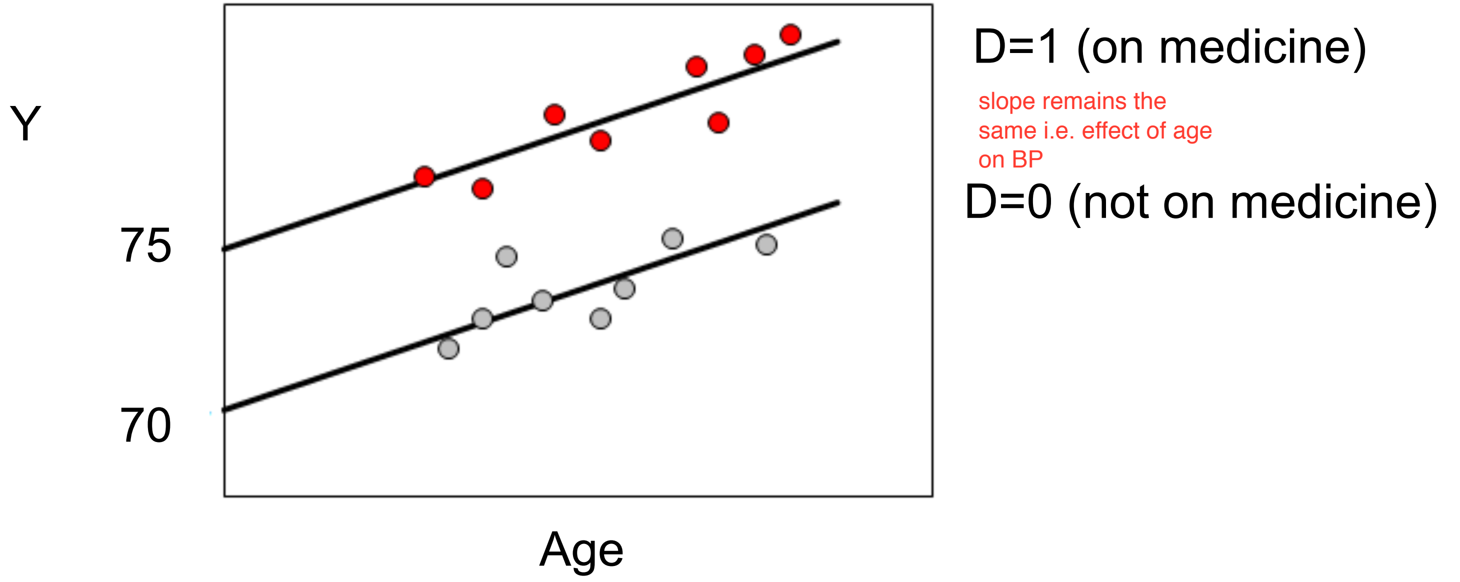
How to Interpret?

- For a given age, those on the medicine (D=1), have on average blood pressure readings 5 points higher than those not on the medicine (D=0).
- This model allowed the effect of Age on Blood Pressure to be the same for both groups-we will show in a little bit how to relax that.

Visually

- The model we are fitting looks as follows

$$\hat{y} = 70 + 5D + 0.44Age$$



Example: Employment Discrimination

- If two groups have apparently different salary structures, you first need to account for differences in education, training and experience before any claim of discrimination can be made.

- Regression analysis with an indicator variable for the group is a way to investigate this.

Bank Teller Salaries

- We have data on salaries of bank tellers, along with their years of experience and gender.
- The bank was sued for discrimination.
- Here we examine how salary depends on experience, also accounting for gender.

First compare salaries by gender

- What does this output imply?

```
> t.test(salary[male==0],salary[male==1],var.equal=FALSE)

        Welch Two Sample t-test

data:  salary[male == 0] and salary[male == 1]
t = -4.14, df = 78.9, p-value = 0.000086
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.2829  -4.3081
sample estimates:
mean of x mean of y
   37.210    45.505
```

Ho: $\mu f = \mu m$
Ha: $\mu f \neq \mu m$
p-value : reject null hypothesis
confidence interval: all negative - male make more

Compare with this regression output

- How does this compare with the previous 2 sample t-test output?

```
> fit=lm(salary~male)
> summary(fit)

Call:
lm(formula = salary ~ male)

Residuals:
   Min     1Q Median     3Q    Max
-18.81  -6.43  -1.86   4.12  51.49

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)   37.210      0.895    41.6 < 0.0000000000000002 ***
male           8.296      1.564     5.3           0.00000029 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.6 on 206 degrees of freedom
Multiple R-squared:  0.12,     Adjusted R-squared:  0.116
F-statistic: 28.1 on 1 and 206 DF,  p-value: 0.000000294
```

Regression Analysis with Experience

```
> fit=lm(salary~male+exper)
> summary(fit)

Call:
lm(formula = salary ~ male + exper)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.190  -5.748  -0.605   4.813  25.855 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  27.8119     1.0279   27.06 < 0.0000000000000002 ***
male          8.0119     1.1931    6.72         0.0000000018 ***
exper         0.9812     0.0803   12.22 < 0.0000000000000002 ***
---
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: 98.9 on 2 and 205 DF,  p-value: <0.0000000000000002
```

You need male and experience in the model

How do we interpret this equation?

We say, after controlling for experience we find….. the average male salary is 8 more than the average female salary

An Intercept Adjuster

For an indicator variable, the b_j is not really a slope. To see this, evaluate the equation for the two groups.

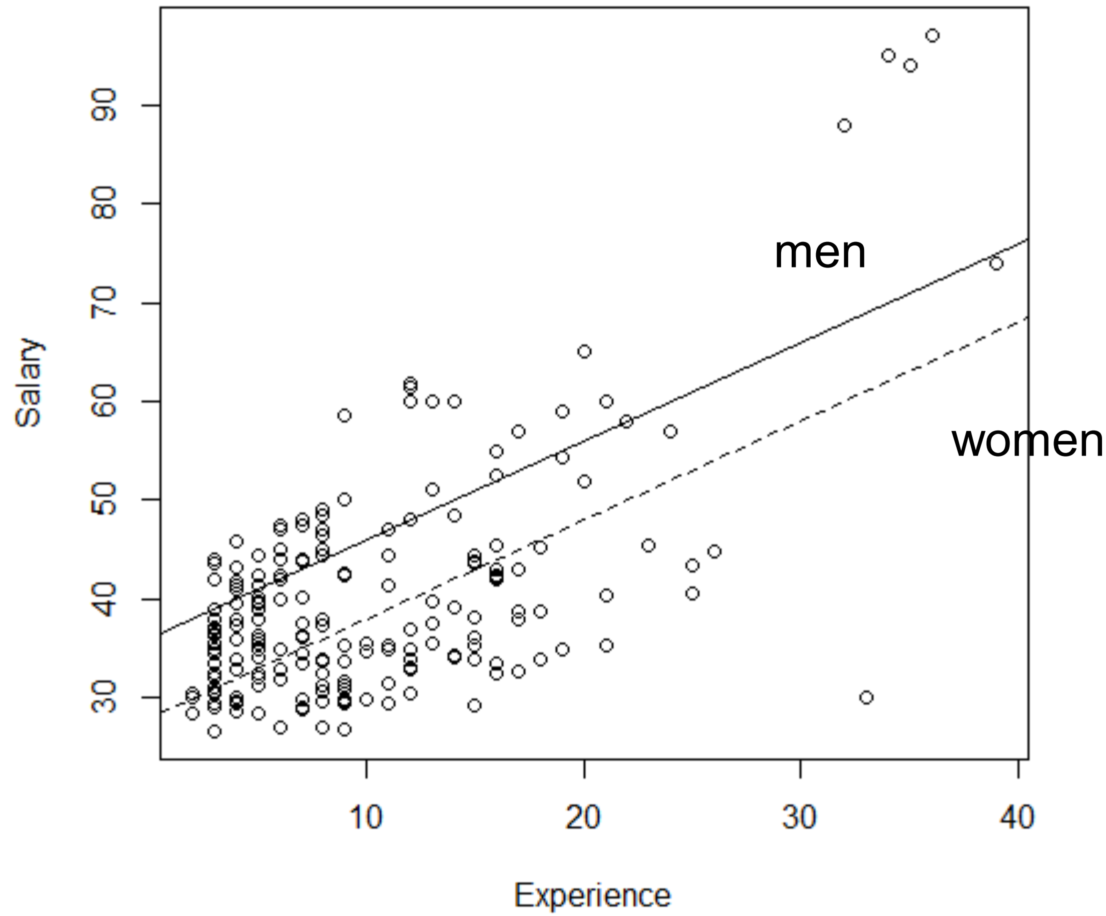
FEMALES (MALES = 0)

SALARY = 28 + 1 EXPER + 8 MALES
= 28 + EXPER + 8 (0)
= 28 + EXPER

MALES (MALES = 1)

SALARY = 28 + 1 EXPER + 8 MALES
= 28 + EXPER + 8 (1)
= 28 + EXPER + 8
= 36 + EXPER

Parallel Salary Equations



This basic model forces the two lines to be parallel (same slope)

Is The Difference Significant?

H_0: $\beta_{MALES} = 0$

(After accounting for years of education, there is no salary difference)

H_a: $\beta_{MALES} \neq 0$

(After accounting for education, there IS a salary difference)

Use $t = b/SE_b$ as usual

$t = 6.72$ is significant (p-value also $< .05$)

What if the Coding Was Different?

- If we had an indicator for females and used it, the equation would be:

```
> fit=lm(salary~female+exper)
> summary(fit)

Call:
lm(formula = salary ~ female + exper)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.190  -5.748  -0.605   4.813  25.855 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  35.8238     1.2591   28.45 < 0.0000000000000002 ***
female       -8.0119     1.1931   -6.72         0.0000000018 ***
exper         0.9812     0.0803   12.22 < 0.0000000000000002 ***
---
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: 98.9 on 2 and 205 DF,  p-value: <0.0000000000000002
```

- Note how this is related to the previous output with males.

Multiple Categories

- Pick one category as the "base category".
- Create one indicator variable for each other category.
- In general, if there are *m* categories, use *m – 1* indicator variables.

Example: Meddicorp Sales

Y = Sales in one of 25 territories

X_1 = advertising in territory

X_2 = bonuses paid in territory

Also Region: 1 = South

2 = West

3 = Midwest

SALES	ADV	BONUS	REGION
963.50	374.27	230.98	1
893.00	408.50	236.28	1
1057.25	414.31	271.57	1
1183.25	448.42	291.20	2
1419.50	517.88	282.17	3
1547.75	637.60	321.16	3
1580.00	635.72	294.32	3
1071.50	446.86	305.69	1
1078.25	489.59	238.41	1
1122.50	500.56	271.38	2
1304.75	484.18	332.64	3
1552.25	618.07	261.80	3
1040.00	453.39	235.63	1
1045.25	440.86	249.68	2

How do you use region?

What happens if you just put it in the model?

```
> fit=lm(sales~adv+bonus+region)
> summary(fit)

Call:
lm(formula = sales ~ adv + bonus + region)

Residuals:
   Min     1Q Median     3Q    Max 
-105.2  -63.2    7.8   43.7  112.2 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -84.219    177.907   -0.47  0.64082    
adv            1.546      0.306    5.05 0.000053 ***
bonus          1.106      0.573    1.93  0.06699 .  
region       118.899     28.687    4.14  0.00046 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 68.9 on 21 degrees of freedom
Multiple R-squared:  0.92,      Adjusted R-squared:  0.909 
F-statistic: 80.7 on 3 and 21 DF,  p-value: 0.0000000000108
```

Region as an X

This implies the difference between Region 3 (MW) and Region 2 (W) = b_3 = 119

And the difference between Region 2 (W) and Region 1 (S) is also 119

The sales differences may not be equal but this **<u>forces</u>** them to be estimated that way

A more flexible approach

- Use two indicator variables to tell the three regions apart

- Can use any one of the three as the “base” category.

- Here is what it looks like if Midwest is selected as the base.

Coding scheme

Region	D_1 South	D_2 West
SOUTH	1	0
WEST	0	1
MIDWEST	0	0

Creating Indicators inR

```
> south=1.0*(region==1)
> west=1.0*(region==2)
> fit=lm(sales~adv+bonus+south+west)
```

Results

```
> fit=lm(sales~adv+bonus+south+west)
> summary(fit)

Call:
lm(formula = sales ~ adv + bonus + south + west)

Residuals:
   Min     1Q Median     3Q    Max
-117.0  -24.5   -1.1   35.9  102.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  435.099    206.234    2.11    0.048 *
adv            1.368      0.262    5.22 0.000042 ***
bonus          0.975      0.481    2.03    0.056 .
south       -257.892     48.413   -5.33 0.000033 ***
west        -209.746     37.420   -5.61 0.000017 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.6 on 20 degrees of freedom
Multiple R-squared:  0.947,     Adjusted R-squared:  0.936
F-statistic:   89 on 4 and 20 DF,  p-value: 0.00000000000189
```

Both indicators are significant

This Defines Three Equations

SALES = 435 + 1.37ADV + .975 BONUS
- 258 South – 210 West

S: SALES = 177 + 1.37ADV + .975 BONUS

W: SALES = 225 + 1.37ADV + .975 BONUS

MW: SALES = 435 + 1.37ADV + .975 BONUS

More on Nominal Variables

Dummy Variables are especially nice because they allow us to use *nominal variables* in regression.

A nominal variable has no rank or order, rendering the numerical coding scheme useless for regression.

More Than 2 Categories

- There may be more than two categories that apply to a variable of interest:

- Region: West, Midwest, South, Northeast
- Season: Winter, Spring, Summer, Fall
- Quality: Poor, Fair, Good, Excellent

- If C is the number of categories, create (C–1) dummy variables for describing the variable.
- **One category is always the "baseline", which is included in the intercept.**

Nominal Variables

- The way you use nominal variables in regression is by converting them to a series of dummy variables.

Nomimal Variable

Race

1 = White

2 = Black

3 = Other

Recode into different

Dummy Variables

1. White
   0 = Not White; 1 = White
2. Black
   0 = Not Black; 1 = Black
3. Other
   0 = Not Other; 1 = Other

Multiple Regression

- When you need to use a nominal variable in regression (like race), just convert it to a series of dummy variables.
- When you enter the variables into your model, you MUST LEAVE OUT ONE OF THE DUMMIES.

Leave Out One	Enter Rest into Regression
White	Black
	Other

Example

- Y = measure of self-esteem
- White = 1 if white, 0 otherwise
- Black = 1 if black,0 otherwise
- Other = 1 if not white or black, 0 otherwise

$$\hat{Y} = b_0 + b_1 Black + b_2 Other$$

b_0 = the y-intercept, which in this case is the predicted value of self-esteem for the excluded group, **white**.

b_1 = the slope for variable **black**

b_2 = the slope for variable **other**

Example

- If our equation were:

$$\hat{Y} = 28 + 5Black - 2Other$$

Plugging in values for the dummies tells you each group's self-esteem average:

White = 28

Black = 33

Other = 26

Example

- Dummy variables can be entered into multiple regression along with other dichotomous and continuous variables.
- For example, you could regress self-esteem on sex, race, and years of education:
- How would you interpret this?

$$\hat{Y} = 30 - 4Female + 5Black - 2Other + 0.3Edu$$

Interpret

$$\hat{Y} = 30 - 4Female + 5Black - 2Other + 0.3Edu$$

1. Women's self-esteem is 4 points lower than men's.
2. Blacks' self-esteem is 5 points higher than whites'.
3. Others' self-esteem is 2 points lower than whites' and consequently 7 points lower than blacks'.
4. Each year of education improves self-esteem by 0.3 units.

Make sure you get into the habit of saying the slope is the effect of an independent variable ***"while holding everything else constant."***

Example

STOCK RETURNS AND THE WEEKEND EFFECT

Kenneth R FRENCH*

University of Rochester, Rochester, NY 14627, USA

Received October 1979, final version received February 1980

This paper examines two alternative models of the process generating stock returns Under the calendar time hypothesis, the process operates continuously and the expected return for Monday is three times the expected return for other days of the week Under the trading time hypothesis, returns are generated only during active trading and the expected return is the same for each day of the week During most of the period studied, from 1953 through 1977, the daily returns to the Standard and Poor's composite portfolio are inconsistent with both models Although the average return for the other four days of the week was positive, the average for Monday was significantly negative during each of five five-year subperiods

The Model

- Y = daily return on S&P
- X_1 = 1 if Tuesday, 0 otherwise
- X_2 = 1 if Wednesday, 0 otherwise
- X_3 = 1 if Thursday, 0 otherwise
- X_4 = 1 if Friday, 0 otherwise

- $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i$

- β_0 --- expected Monday return
- β_1 --- expected difference between Tuesday return and Monday return

How the paper expressed it

the week The regression,

$$R_t = \alpha + \gamma_2 d_{2t} + \gamma_3 d_{3t} + \gamma_4 d_{4t} + \gamma_5 d_{5t} + \varepsilon_t, \qquad (1)$$

is used to formally test this proposition In this regression, R_t is the return to the Standard and Poor's portfolio and the dummy variables indicate the day of the week on which the return is observed (d_{2t} = Tuesday, d_{3t} = Wednesday, etc) The expected return for Monday is measured by α, while γ_2 through γ_5 represent the difference between the expected return for Monday and the expected return for each of the other days of the week If the expected return is the same for each day of the week, the estimates of γ_2 through γ_5 will be close to zero and an F-statistic measuring the joint significance of the dummy variables should be insignificant

Interpreting the Model

Day of Week	Return
Monday	β_0
Tuesday	$\beta_0 + \beta_2$
Wednesday	$\beta_0 + \beta_3$
Thursday	$\beta_0 + \beta_4$
Friday	$\beta_0 + \beta_5$

Running Model in R

```
> fit=lm(rets~tues+wed+thur+fri)
> summary(fit)

Call:
lm(formula = rets ~ tues + wed + thur + fri)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06520 -0.00379  0.00008  0.00390  0.04917

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -0.001558   0.000218   -7.13      0.0000000000011 ***
tues         0.001673   0.000306    5.47      0.0000000480860 ***
wed          0.002607   0.000307    8.50 < 0.0000000000000002 ***
thur         0.002130   0.000307    6.94      0.0000000000042 ***
fri          0.002640   0.000307    8.59 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00749 on 6016 degrees of freedom
Multiple R-squared:  0.0163,    Adjusted R-squared:  0.0157
F-statistic: 24.9 on 4 and 6016 DF,  p-value: <0.0000000000000002
```

For hardcore R fans

```
getSymbols("^GSPC",from="1953-01-01",to="1977-01-01")
rets=dailyReturn(GSPC)
dadates=time(rets)
wdays=weekdays(as.Date(dadates,'%d-%m-%Y'))
mon=1.0*(wdays=="Monday")
tues=1.0*(wdays=="Tuesday")
wed=1.0*(wdays=="Wednesday")
thur=1.0*(wdays=="Thursday")
fri=1.0*(wdays=="Friday")
fit=lm(rets~tues+wed+thur+fri)
```

What does the model say?

- Suggests Monday is a downer

Day of week	Return
Monday	-.0016
Tuesday	-.0016+.0017
Wednesday	-.0016+.0026
Thursday	-.0016+.0021
Friday	-.0016+.0026

A Trading Scheme

5. Potential profit from the negative returns for Monday

Even if one were to conclude that the negative returns for Monday are evidence of market inefficiency, the profit to any individual from knowledge of the negative returns is more limited than it may appear One simple trading strategy based on this information would be for an individual to purchase the Standard and Poor's composite portfolio every Monday afternoon and to sell these investments on Friday afternoon, holding cash over the weekend Ignoring transactions costs, this trading rule would have generated an average annual return of 13 4 percent from 1953 to 1977, while a buy and hold policy would have yielded a 5 5 percent annual return However, no investor can ignore transactions costs If these costs are only 0 25 percent per transaction, the buy and hold policy would have yielded a higher return in each of the 25 years studied

Does it still work?

- The hedge funds want to find an edge, exploit as much as possible, then move on.
- With R its easy enough to see what happens if we use data from 1990 to 2015.

Data from 1990 to 2015

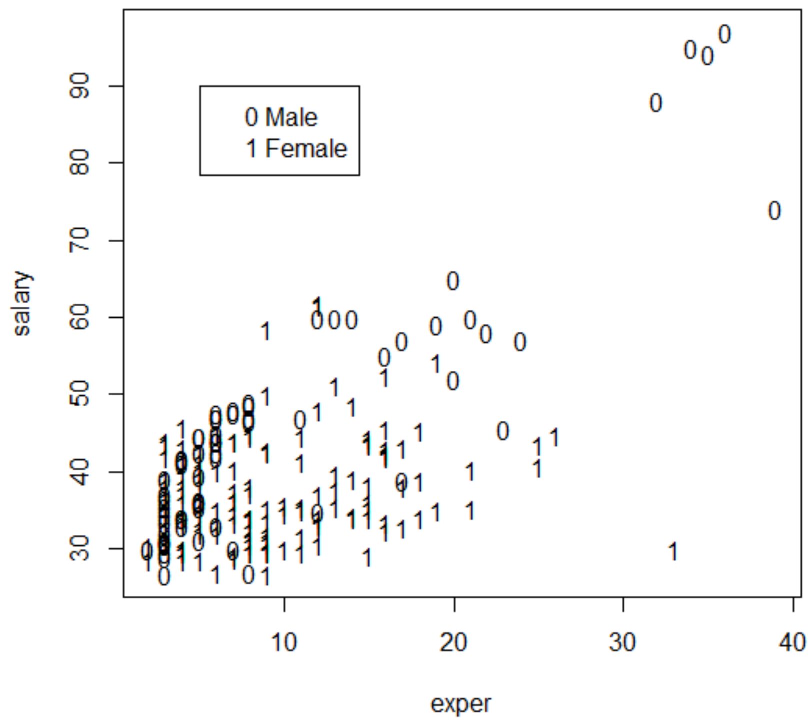
```
Call:
lm(formula = rets ~ tues + wed + thur + fri)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.090683 -0.005029  0.000227  0.005340  0.115374 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.263e-04  3.306e-04   1.289    0.197
tues         2.313e-04  4.585e-04   0.505    0.614
wed         -9.326e-05  4.585e-04  -0.203    0.839
thur        -2.448e-04  4.606e-04  -0.532    0.595
fri         -3.022e-04  4.612e-04  -0.655    0.512

Residual standard error: 0.01141 on 6296 degrees of freedom
Multiple R-squared:  0.0002798, Adjusted R-squared:  -0.0003553 
F-statistic: 0.4406 on 4 and 6296 DF,  p-value: 0.7794
```

Interaction Variables

- Another type of variable used in regression models is an interaction variable.
- This is usually formulated as the product of two variables; for example, $x_3 = x_1 x_2$
- With this variable in the model, it means the level of x_2 changes how x_1 affects Y

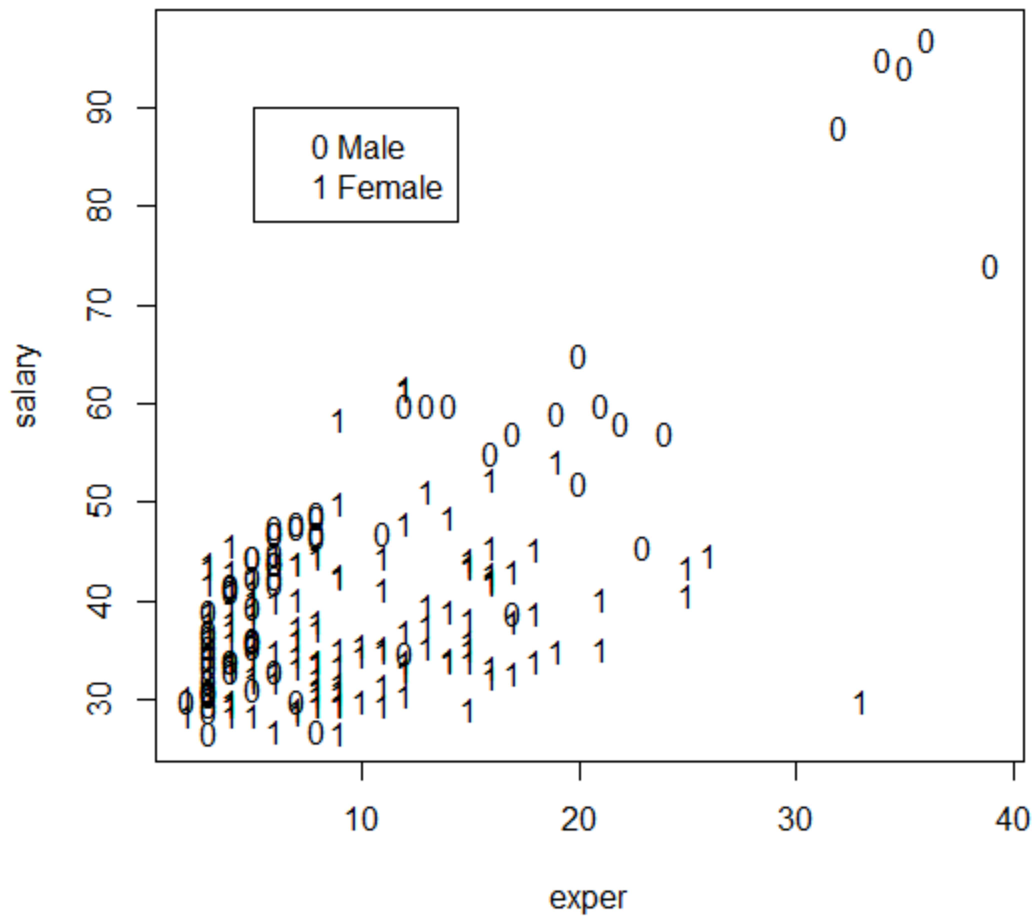
Bank Data Again

- Examine the graph-do you see two lines with different intercepts and slopes?



To model different slopes you need an interaction term.

Salary Versus Years of Experience



1 female

0 male

At all levels of experience, the **male** salaries appear higher.

The Interaction Model

With two *x* variables the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

If we factor out x_1 we get:

$$y = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + e$$

so each value of x_2 yields a different slope in the relationship between *y* and x_1

Interaction Involving an Indicator

If one of the two variables is binary, the interaction produces a model with two different slopes.

When $x_2 = 0$

$$y = \beta_0 + \beta_1 x_1 + e$$

When $x_2 = 1$

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + e$$

Example: Discrimination (again)

- In the Bank Case, suppose we suspected that the salary difference by gender changed with different levels of experience
- To investigate this, we created a new variable MEXP = EXPER*MALES and added it to the model.

Regression Output

```
> mexp=exper*male
> fit=lm(salary~exper+male+mexp)
> summary(fit)

Call:
lm(formula = salary ~ exper + male + mexp)

Residuals:
     Min       1Q   Median       3Q      Max
-20.0685  -4.6506  -0.7679   4.4034  23.9122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.5283     1.1380  30.342  < 2e-16 ***
exper         0.2800     0.1025   2.733  0.00684 **
male         -4.0983     1.6658  -2.460  0.01472 *
mexp          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
```

How do we interpret the equation this time?

A Slope Adjuster

To see the interaction effect, once again evaluate the equation for the two groups.

FEMALES (MALES = 0)

SALARY = 35 + 0.3 EXPER - 4 MALES + 1.25 MEXP

= 35 + 0.3 EXPER - 4 (0) + 1.25 (EXPER*0)

= 35 + 0.3 EXPER

MALES (MALES = 1)

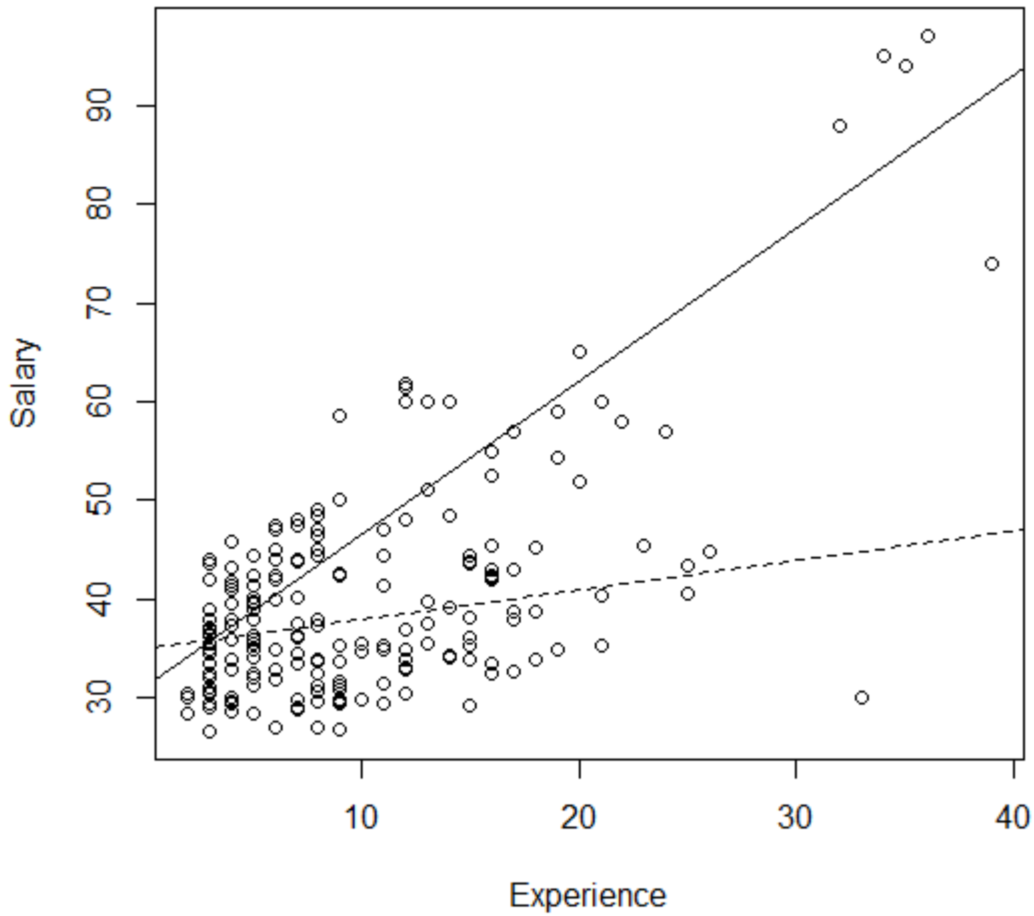
SALARY = 35 + 0.3 EXPER - 4 MALES + 1.25 MEXP

= 35 + 0.3 EXPER - 4 (1) + 1.25 (EDUCAT*1)

= 35 + 0.3 EXPER – 4 + 1.25 EXPER

= 31 + 1.55 EXPER

Lines With Two Different Slopes



Women start out at a higher rate, but men make much more money per year of experience.

Are these results significant? What do we examine in the regression output?

Example: Brick Houses

- We have data on 128 recent sales in Mid City.
- For each sale, the file shows the neighborhood (1, 2, or 3) in which the house is located, the number of offers made on the house, the square footage, whether the house is made primarily of brick, the number of bathrooms, the number of bedrooms, and the selling price.
- Neighborhoods 1 and 2 are more traditional neighborhoods, whereas neighborhood 3 is a newer, more prestigious neighborhood.

Snapshot of Data

Home	Nbhd	Offers	Sq Ft	Brick	Bedrooms	Bathrooms	Price	Nbhd1	Nbhd2	Nbhd3
1	2	2	1790	0	2	2	114300	0	1	0
2	2	3	2030	0	4	2	114200	0	1	0
3	2	1	1740	0	3	2	114800	0	1	0
4	2	3	1980	0	3	2	94700	0	1	0
5	2	3	2130	0	3	3	119800	0	1	0
6	1	2	1780	0	3	2	114600	1	0	0
7	3	3	1830	1	3	3	151600	0	0	1
8	3	2	2160	0	4	2	150700	0	0	1
9	2	3	2110	0	4	2	119200	0	1	0
10	2	3	1730	0	3	3	104000	0	1	0
11	2	3	2030	1	3	2	132500	0	1	0

Is there a brick premium

- All else equal, do buyers pay a premium for a brick house?

```
> fit=lm(Price~Offers+Sq.Ft+Brick+Bedrooms+Bathrooms+Nbhd2+Nbhd3,data=foo)
> summary(fit)

Call:
lm(formula = Price ~ Offers + Sq.Ft + Brick + Bedrooms + Bathrooms +
    Nbhd2 + Nbhd3, data = foo)

Residuals:
     Min       1Q   Median       3Q      Max
-27337.3  -6549.5    -41.7   5803.4  27359.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2159.498   8877.810   0.243  0.80823
Offers      -8267.488   1084.777  -7.621 6.47e-12 ***
Sq.Ft          52.994      5.734   9.242 1.10e-15 ***
Brick       17297.350   1981.616   8.729 1.78e-14 ***
Bedrooms     4246.794   1597.911   2.658  0.00894 **
Bathrooms    7883.278   2117.035   3.724  0.00030 ***
Nbhd2       -1560.579   2396.765  -0.651  0.51621
Nbhd3       20681.037   3148.954   6.568 1.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10020 on 120 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.861
F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```

Is there a Neighborhood 3 Premium?

```
> fit=lm(Price~Offers+Sq.Ft+Brick+Bedrooms+Bathrooms+Nbhd2+Nbhd3,data=foo)
> summary(fit)

Call:
lm(formula = Price ~ Offers + Sq.Ft + Brick + Bedrooms + Bathrooms +
    Nbhd2 + Nbhd3, data = foo)

Residuals:
     Min       1Q   Median       3Q      Max
-27337.3  -6549.5    -41.7   5803.4  27359.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2159.498   8877.810   0.243  0.80823
Offers      -8267.488   1084.777  -7.621 6.47e-12 ***
Sq.Ft          52.994      5.734   9.242 1.10e-15 ***
Brick       17297.350   1981.616   8.729 1.78e-14 ***
Bedrooms     4246.794   1597.911   2.658  0.00894 **
Bathrooms    7883.278   2117.035   3.724  0.00030 ***
Nbhd2       -1560.579   2396.765  -0.651  0.51621
Nbhd3       20681.037   3148.954   6.568 1.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10020 on 120 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.861
F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```

What does the following imply?

```
> inter=Brick*Nbhd3
> fit=lm(Price~Offers+Sq.Ft+Brick+Bedrooms+Bathrooms+Nbhd2+Nbhd3+inter)
> summary(fit)

Call:
lm(formula = Price ~ Offers + Sq.Ft + Brick + Bedrooms + Bathrooms +
    Nbhd2 + Nbhd3 + inter)

Residuals:
     Min       1Q   Median       3Q      Max
-26939.1  -5428.7   -213.9   4519.3  26211.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3009.993   8706.264   0.346  0.73016
Offers      -8401.088   1064.370  -7.893 1.62e-12 ***
Sq.Ft          54.065      5.636   9.593  < 2e-16 ***
Brick       13826.465   2405.556   5.748 7.11e-08 ***
Bedrooms     4718.163   1577.613   2.991  0.00338 **
Bathrooms    6463.365   2154.264   3.000  0.00329 **
Nbhd2        -673.028   2376.477  -0.283  0.77751
Nbhd3       17241.413   3391.347   5.084 1.39e-06 ***
inter       10181.577   4165.274   2.444  0.01598 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9817 on 119 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8665
F-statistic:   104 on 8 and 119 DF,  p-value: < 2.2e-16
```