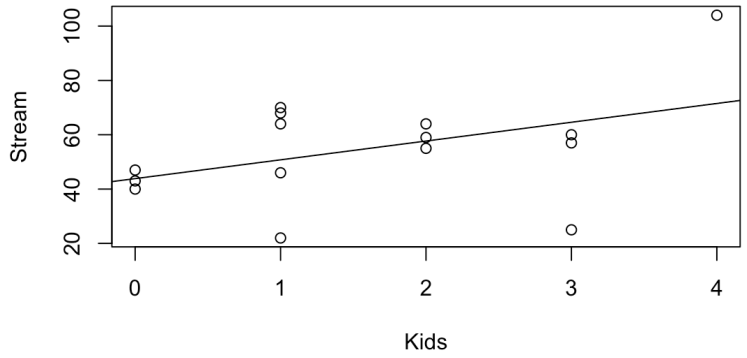
Session 2

Statistical Significance vs. Strength of Relationship: Review (Using Assignment 1 Framework: Sales~Kids)

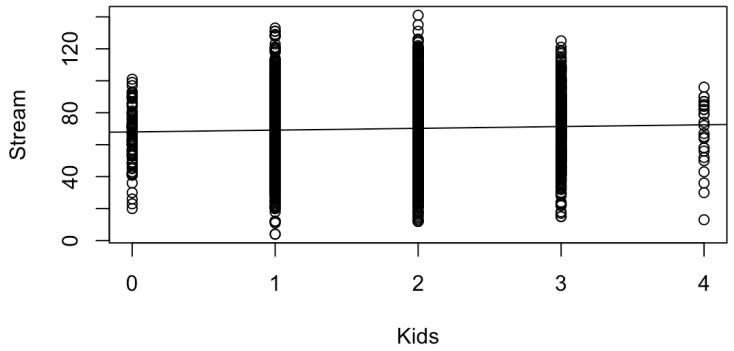
Moderate relationship, not statistically significant




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   43.852      7.979   5.496 0.000103 ***
Kids           6.926      3.989   1.736 0.106165
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.54 on 13 degrees of freedom
Multiple R-squared:  0.1882     Adjusted R-squared:  0.1258
F-statistic: 3.014 on 1 and 13 DF,  p-value: 0.1062
```

```
FlixIt15 <- read.table("FlixIt15.dat",
                       header = TRUE)
FlixIt15.slr <- lm(Stream~Kids,
                   data=FlixIt15)
summary(FlixIt15.slr)
plot(Stream~Kids,data = FlixIt15)
abline(FlixIt15.slr)
```

Weak relationship, statistically significant




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.9937     0.8524  79.769   <2e-16 ***
Kids         1.1248     0.4438   2.534   0.0113 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.99 on 3998 degrees of freedom
Multiple R-squared:  0.001604,  Adjusted R-squared:  0.001354
F-statistic: 6.423 on 1 and 3998 DF,  p-value: 0.0113
```

```
FlixIt4000 <- read.table("FlixIt4000.dat",
                         header = TRUE)
FlixIt4000.slr <- lm(Stream~Kids,
                     data=FlixIt4000)
summary(FlixIt4000.slr)
plot(Stream~Kids,data = FlixIt4000)
abline(FlixIt4000.slr)
rm(FlixIt4000,FlixIt4000.slr)
```

Introduction to Today's Discussion

This week, we will consider whether the *conditional* relationship between a given independent variable and the dependent variable changes once the effects of other variables are considered.

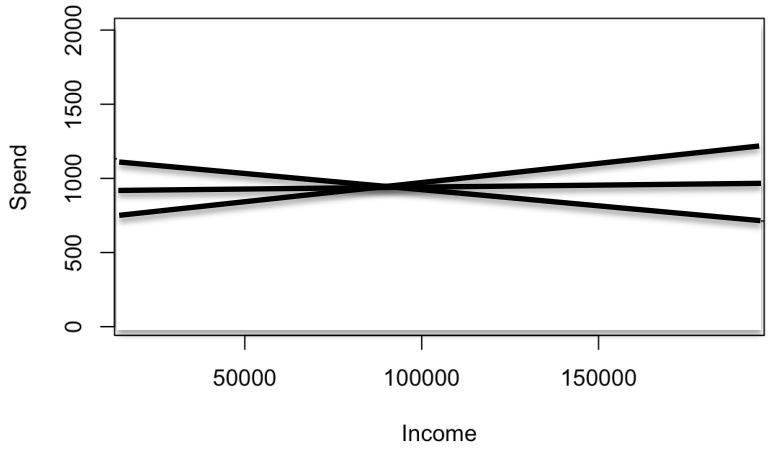
In other words: two variables which appear to be related to each other in a specific way …

… may be related to each other in a different way …

… or may be unrelated to each other …

… once the effects of other variables are considered.

This is important in terms of determining any action that you might (or might not) take in view of the conditional relationship.



Introduction to Multiple Regression

Generic form: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

Rationale:

1. **Gain a better prediction of the dependent variable**

2. **Assess "value added" (if any) by each independent variable in the prediction of the dependent variable (coefficients of partial determination)**

3. **Assess the magnitude and direction of the *conditional* relationship between each independent variable and the dependent variable**

Importantly, multiple regression is NOT a convenient summary of the collection of simple linear regression results.

Important differences can, and often do, emerge in the results of a multiple regression analysis compared to the results of a series of simple linear regression analyses

Market Experts, Inc.

- **Market Experts, Inc.: large point-to-point marketing firm**
- **Objective: Create linear model to predict employee sales (continuous) from tenure as a marketing employee (continuous) and age (continuous)**
- **Data collected from a random sample of Market Experts' marketing employees**
- **Collect: tenure, age, 12-month sales (in thousands of dollars)**
  - **Importantly, "tenure" is defined as the number of years the individual has been employed by the company in the marketing division**

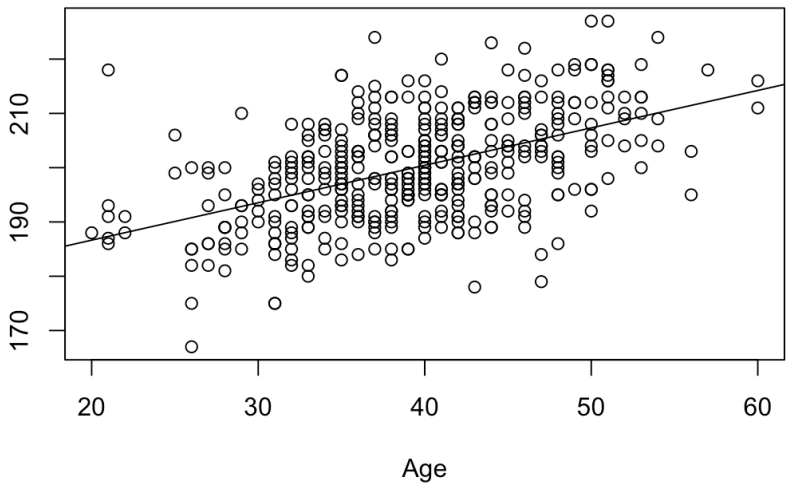
Age	Tenure	Sales
Min. :20.00	Min. : 1.000	Min. :167
1st Qu.:35.00	1st Qu.: 7.000	1st Qu.:193
Median :40.00	Median : 9.000	Median :200
Mean :39.36	Mean : 8.837	Mean :200
3rd Qu.:44.00	3rd Qu.:10.000	3rd Qu.:207
Max. :60.00	Max. :15.000	Max. :227

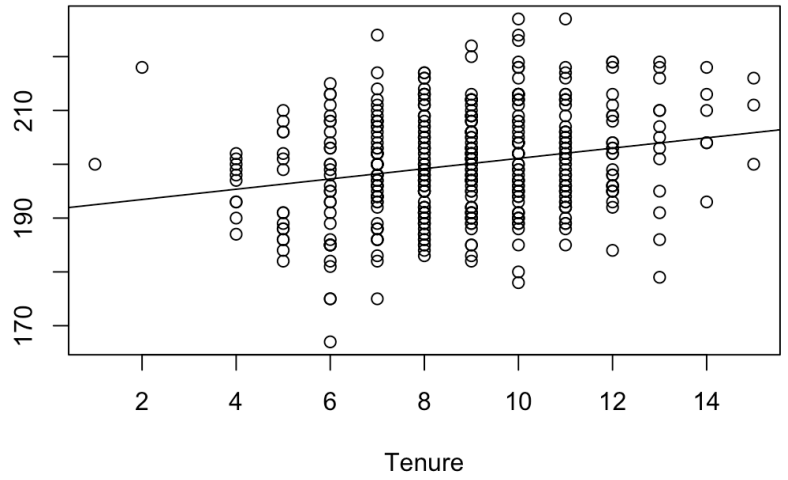
- **Pass the "smell test?"**

Simple Linear Regressions of Sales on Age, Tenure (Separately)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 172.85362    2.33069   74.16   <2e-16 ***
Age           0.68960    0.05824   11.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.585 on 414 degrees of freedom
Multiple R-squared:  0.253,     Adjusted R-squared:  0.2512
F-statistic: 140.2 on 1 and 414 DF,  p-value: < 2.2e-16
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 191.5385     1.9090 100.335  < 2e-16 ***
Tenure        0.9570     0.2092   4.574 6.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.69 on 414 degrees of freedom
Multiple R-squared:  0.0481,    Adjusted R-squared:  0.0458
F-statistic: 20.92 on 1 and 414 DF,  p-value: 6.336e-06
```



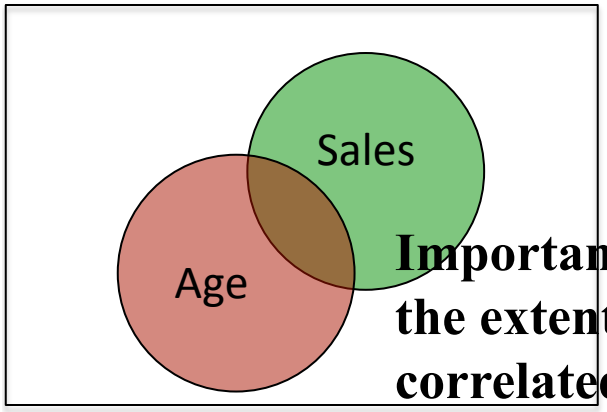


What conclusions do you draw from these bivariate relationships?

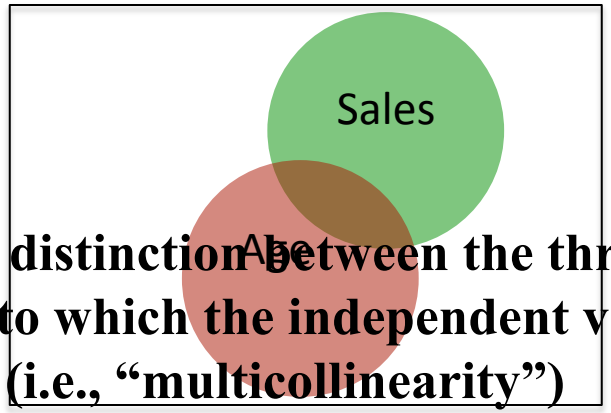
What are the personnel policy implications?

Simple Linear Regression vs. Multiple Regression: Multicollinearity

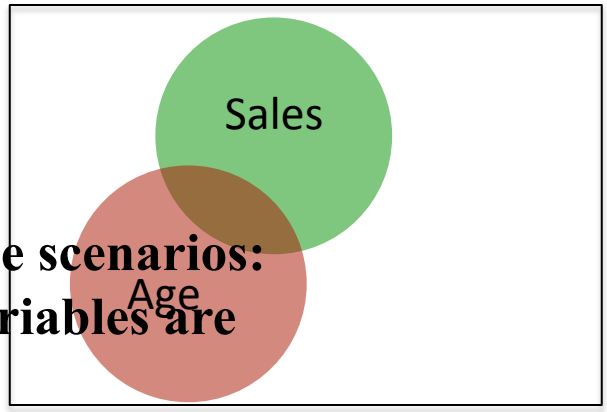
Scenario A



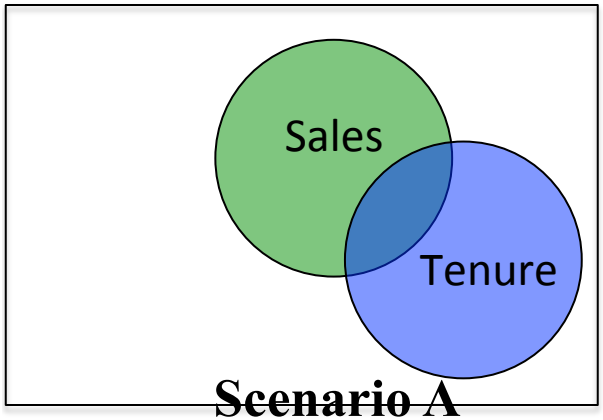
Scenario B

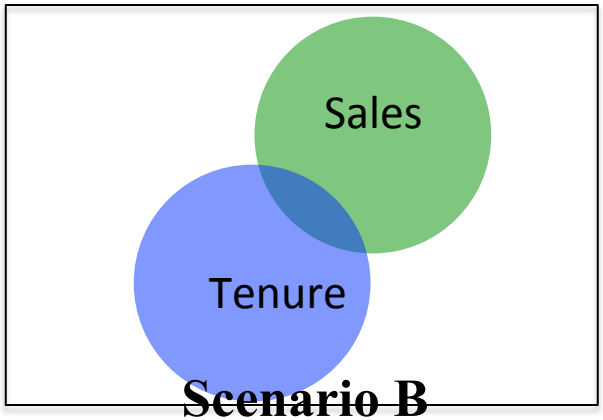


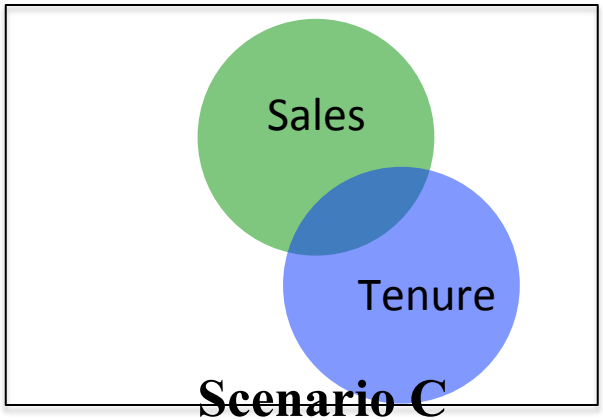
Scenario C



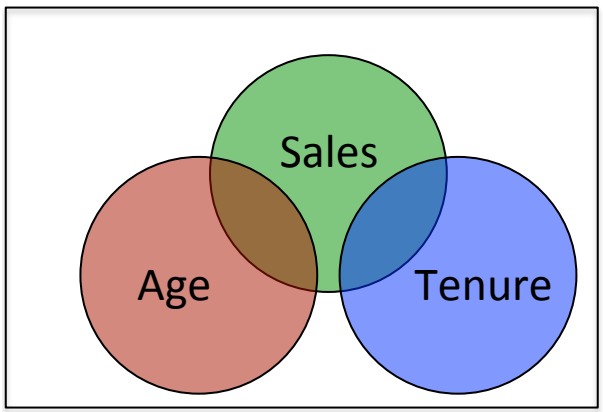
Important distinction between the three scenarios: the extent to which the independent variables are correlated (i.e., "multicollinearity")



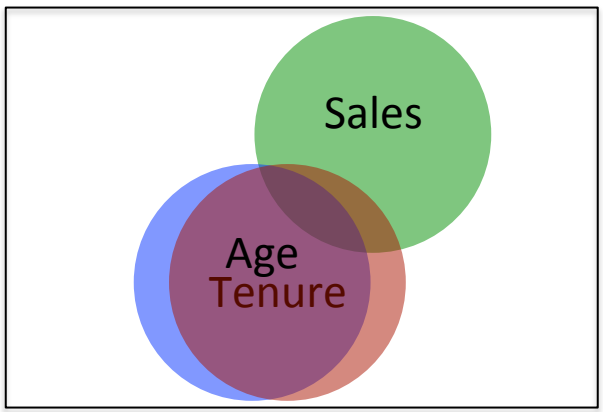




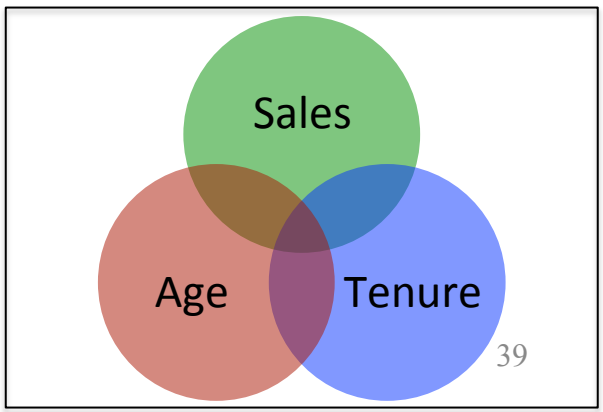
Scenario A



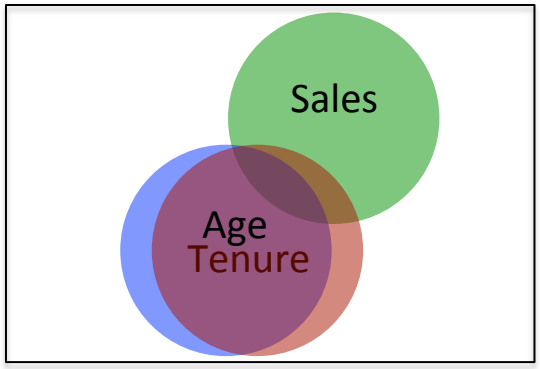
Scenario B

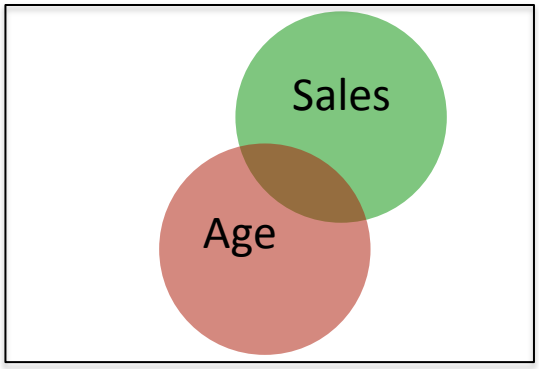


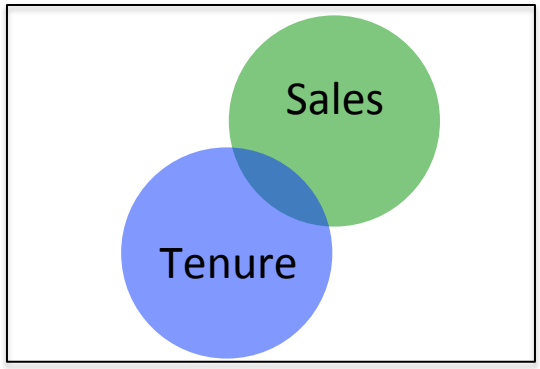
Scenario C



Scenario B








```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 185.8405     5.4966  33.810   <2e-16 ***
Age           0.3794     0.1429   2.655   0.0107 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.423 on 48 degrees of freedom
Multiple R-squared:  0.1281,    Adjusted R-squared:  0.1099
F-statistic: 7.051 on 1 and 48 DF,  p-value: 0.01072
```

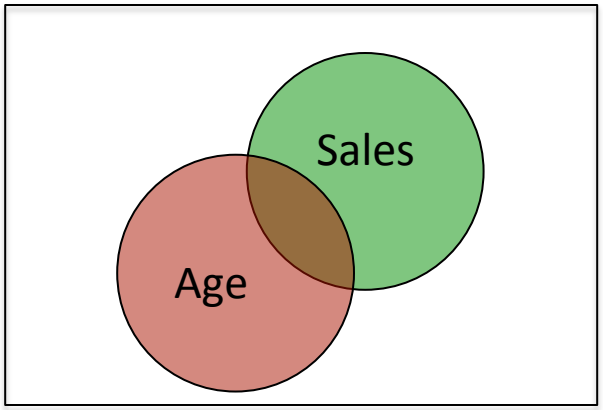
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.1057     3.5805   53.65   <2e-16 ***
Tenure        0.9868     0.4146    2.38   0.0213 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.544 on 48 degrees of freedom
Multiple R-squared:  0.1056,    Adjusted R-squared:  0.08695
F-statistic: 5.666 on 1 and 48 DF,  p-value: 0.02131
```

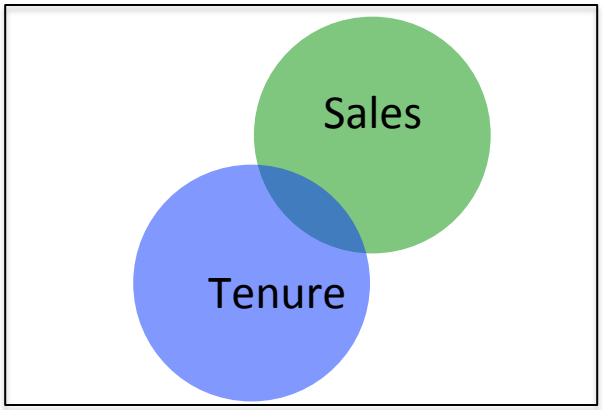
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 186.6221     5.7495  32.459   <2e-16 ***
Age           0.2853     0.2348   1.215    0.230
Tenure        0.3414     0.6725   0.508    0.614
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.497 on 47 degrees of freedom
Multiple R-squared:  0.1328,    Adjusted R-squared:  0.09593
F-statistic:   3.6 on 2 and 47 DF,  p-value: 0.03511
```

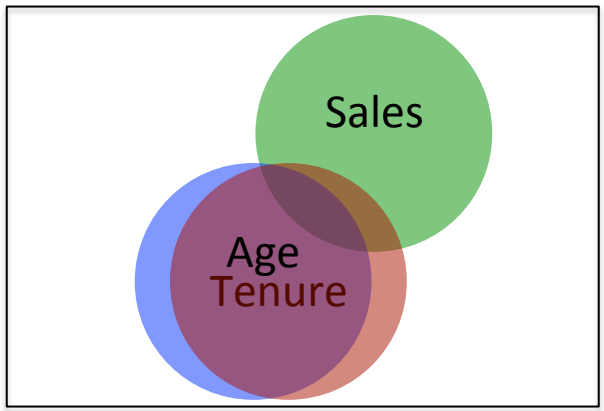
```
ScenarioB <- lm(Sales~Age + Tenure,
                data=MarketExperts)
summary(ScenarioB)
```

E(Sales) = b0 + b1*Age + b2*Tenure
= 186.62 + .2853*Age + .3414*Tenure

Introducing the Coefficient of Multiple Determination, Coefficients of Partial Determination, and the Global F: Scenario B








```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 185.8405     5.4966  33.810   <2e-16 ***
Age           0.3794     0.1429   2.655   0.0107 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.423 on 48 degrees of freedom
Multiple R-squared:  0.1281,    Adjusted R-squared:  0.1099
F-statistic: 7.051 on 1 and 48 DF,  p-value: 0.01072
```

```
Response: Sales
           eta^2 Sum Sq Df F value  Pr(>F)
Age      0.12807    626  1  7.0505 0.01072 *
Residuals          4262 48
```

```
require(heplots)
etasq(ScenarioB,anova=TRUE,partial=FALSE)
```

“Statistical significance” of a particular independent variable in multiple regression implies “reasonable confidence that the variable’s coefficient of partial determination is greater than zero in the population.”
Based on these results, what recommendations would you offer the CEO about Market Experts’ personnel policy?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.1057     3.5805   53.65   <2e-16 ***
Tenure        0.9868     0.4146    2.38   0.0213 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.544 on 48 degrees of freedom
Multiple R-squared:  0.1056,    Adjusted R-squared:  0.08695
F-statistic: 5.666 on 1 and 48 DF,  p-value: 0.02131
```

```
Response: Sales
            eta^2 Sum Sq Df F value  Pr(>F)
Tenure    0.10558  516.1  1  5.6663 0.02131 *
Residuals          4371.9 48
```

```
etasq(ScenarioB,anova=TRUE,partial=FALSE)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 186.6221     5.7495  32.459   <2e-16 ***
Age           0.2853     0.2348   1.215    0.230
Tenure        0.3414     0.6725   0.508    0.614
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.497 on 47 degrees of freedom
Multiple R-squared:  0.1328,    Adjusted R-squared:  0.09593
F-statistic:  3.6 on 2 and 47 DF,  p-value: 0.03511
```

Global F and associated p-value

Coefficient of Multiple Determination

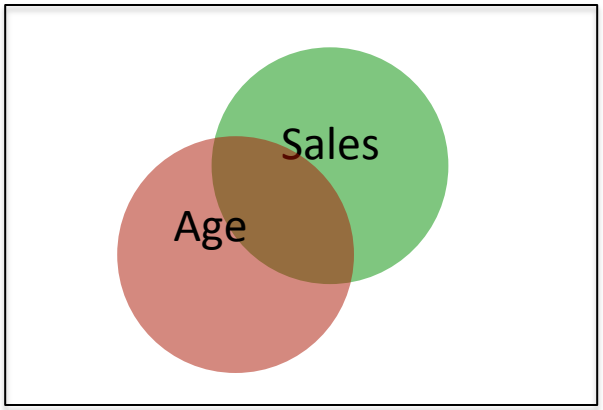
Adjusted Coefficient of Multiple Determination

```
Response: Sales
             eta^2 Sum Sq Df F value Pr(>F)
Age      0.0303011  133.2  1  1.4767 0.2304
Tenure   0.0052893   23.2  1  0.2578 0.6140
Residuals          4238.7 47
```

```
etasq(ScenarioB,anova=TRUE,partial=FALSE)
```

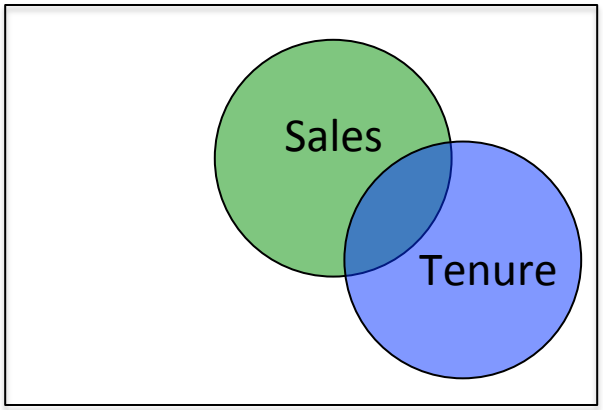
Coefficients of *partial* determination

Multiple Regression Results: Scenario A



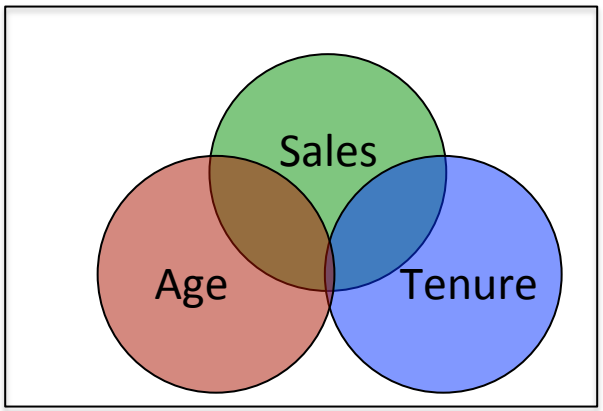

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 185.6636     6.8271  27.195   <2e-16 ***
Age           0.3748     0.1746   2.146   0.0369 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 9.668 on 48 degrees of freedom
Multiple R-squared:  0.08757,   Adjusted R-squared:  0.06856
F-statistic: 4.607 on 1 and 48 DF,  p-value: 0.03693
```

```
Response: Sales
             eta^2 Sum Sq Df F value  Pr(>F)
Age        0.08757  430.6  1  4.6068 0.03693 *
Residuals          4486.4 48
```




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.4834     3.9685  48.503   <2e-16 ***
Tenure        0.7933     0.3919   2.024   0.0485 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.715 on 48 degrees of freedom
Multiple R-squared:  0.07866,   Adjusted R-squared:  0.05946
F-statistic: 4.098 on 1 and 48 DF,  p-value: 0.04852
```

```
Response: Sales
              eta^2 Sum Sq Df F value  Pr(>F)
Tenure     0.078655  386.7  1  4.0978 0.04852 *
Residuals           4530.2 48
```




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 177.6860     7.5514  23.530   <2e-16 ***
Age           0.3821     0.1684   2.269   0.0279 *
Tenure        0.8104     0.3761   2.155   0.0363 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.32 on 47 degrees of freedom
Multiple R-squared:  0.1696,    Adjusted R-squared:  0.1343
F-statistic:   4.8 on 2 and 47 DF,  p-value: 0.01268
```

```
Response: Sales
              eta^2 Sum Sq Df F value  Pr(>F)
Age        0.090657  447.3  1  5.1486 0.02790 *
Tenure     0.081772  403.4  1  4.6441 0.03632 *
Residuals           4083.0 47
```

Based on these results, what recommendations would you offer the CEO about Market Experts' personnel policy?

Summary: Unconditional vs. Conditional Relationships

Last week, we focused on the relationship between two continuous variables, and learned about the distinction between statistical significance and strength of relationship, but did not focus on that relationship in the context of additional variables in the model.

This week, we have looked at the relationship under two situations: unconditional (bivariate), and conditional (multivariable). We learned that…

Two continuous variables which might have a moderate or strong unconditional (i.e., bivariate) relationship…

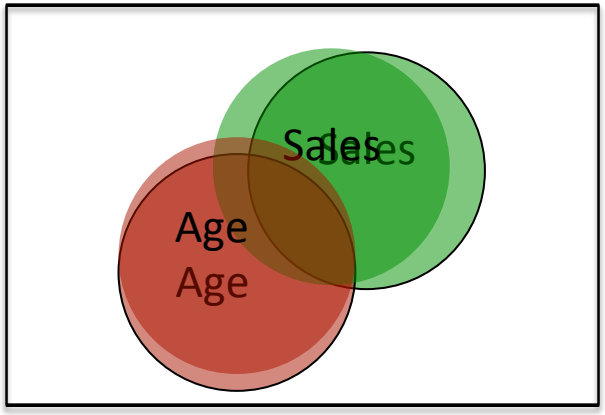
May have only a weak conditional (i.e., multivariable, or "unique") relationship.

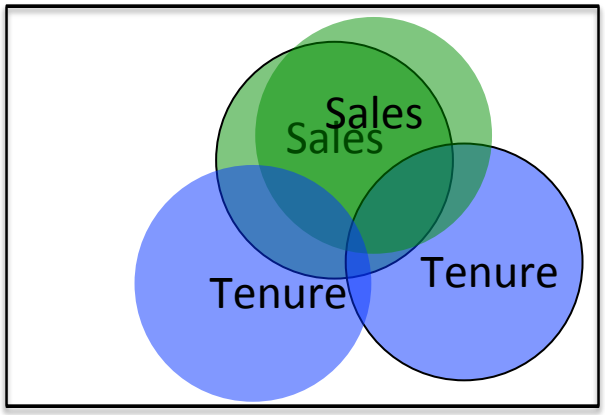
Thus, multiple regression is not the union of a set of simple linear regressions: multivariable results can be quite different from bivariate results.

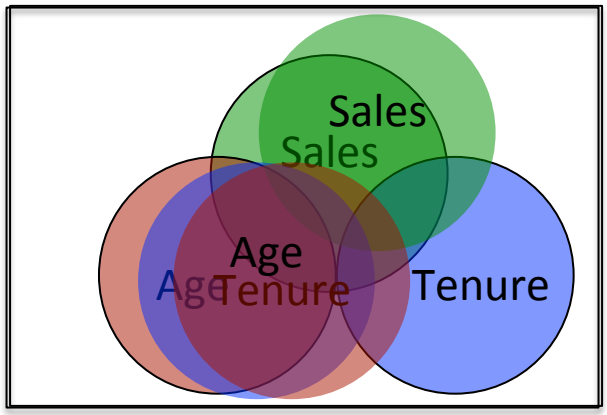
The one case where multiple regression *is* the union of the set of simple linear regressions is when there is no multicollinearity.

When interpreting results, it is important to consider both the unconditional relationships and the conditional relationships: they can provide complementary information.

Thus there are six (next week: eight) measures of importance: unconditional p-values, bivariate coefficients of determination, conditional p-values, coefficients of partial determination, coefficients of multiple determination, and adjusted coefficients of multiple determination. These measures all tell you different things about your data.







Group Project: NewsData, Inc.

- **NewsData, Inc.: provides data analytics to news organizations such as the Washington Post, Fox News, etc.**
- **Objective: Create linear model to predict characteristics of people who prefer getting their news in print vs. via social media, based on Age and Income**
- **Data collected from a random sample of adults**
  - **Available on Blackboard: Outline/Session 2/Datasets Used in Today's Session/Newspaper.dat**
  - **There is a header line at the beginning of the dataset**
- **Variables: "Newspaper" (ranging from -2=prefer social media to +2=prefer print), "Age", "Income"**
- **Class has been divided into 6 groups (announced in class). Please do not switch to another group.**
- **<u>Every</u> member of <u>every</u> group is responsible for being able to answer <u>all</u> questions in <u>every one</u> of the following six question sets at our next class meeting**
- **Please meet with your group this week, and be sure that every member of your group is prepared to answer all questions in all question sets.**
- **The Group Convenor is the first person listed on the next slide. This person should arrange for the first meeting of your group.**
- **If you are unfamiliar with Blackboard's group tools, a brief review is provided at https://www.youtube.com/watch?v=B82oJfxmgXA**

Group Assignments

Group 1	Group 2	Group 3
Kelvin Chang	Adel Hassen	Charlotte Grayson
Seungheon Han	Matthew Arnaut	Hrolfur Sveinsson
Alex Krasner	Yuwen Luo	Griffin Faulkner
Yihang Zhao	Lucas Okwudishu	Ting Huang
Lu Li	Yixuan Yang	Lourdes Siman Ghattas
Calvin Ji	Leqi Yin	Matias Roca-Guiulfo
Vi Pham	Qunzhe Ding	Adetoun Elizabeth Adeyemi
Weike Zhou	Ben Katz	Alexis Yang
Mete Ozmen	Michael Kelly	

Group 4	Group 5	Group 6
Akansha Rathore	Garrett Ramela	Lily Zeng
Kewei Chen	Zach Vila	Brendan Carney
Raquel Kerber	Abenezer Tekle	Peijia Wu
Chengshu Yang	Yu Luo	Jason Liu
Youssef Ragab	Qian Xie	Yiliang Xu
Yachao He	Carlos Machado Rios	Justin Sherman
Dahyun Choi	Tivon Johnson	Elias Issa Issa
Jaime Sarmiento-Monroy	Sun Pil Howang	Runzhe Tang

Question Sets 1 and 2

Question Set 1

- Do the data "smell" right?
- What is the difference between what the mean measures and what the median measures?
- What is meant by the "1st Quartile"?
- What is meant by the 3rd Quartile"?
- What would we EXPECT the relationship to be between Newspaper and Age?
- What would we EXPECT the relationship to be between Newspaper and Income?
- Would we expect these relationships to change from the bivariate analysis to the multivariable analysis?
- What other variables would we like to see in this analysis?

Question Set 2 (bivariate relationship with Age)

- Is there a relationship?
- Is it significant?
- How strong is it?
- What is the interpretation of the intercept?
- What is the interpretation of the slope?
- What is the null hypothesis?
- What is the alternative hypothesis?
- Based on this output, what would you report to the CEO?

Question Sets 3 and 4

Question Set 3 (bivariate relationship with Income)

- Is there a relationship?
- Is it significant?
- How strong is it?
- What is the interpretation of the intercept?
- What is the interpretation of slope?
- What is the null hypothesis?
- What is the alternative hypothesis?
- Based on this output, what would you report to the CEO?

Question Set 4 (multivariable relatioship with Age)

- Is there a relationship?
- Is it significant?
- How strong is it?
- What is the interpretation of the intercept?
- What is the interpretation of slope?
- What is the null hypothesis?
- What is the alternative hypothesis?
- Based on this output, what would you report to the CEO?

Question Sets 5 and 6

Question Set 5 (bivariate relationship with Income)

- Is there a relationship?
- Is it significant?
- How strong is it?
- What is the interpretation of the intercept?
- What is the interpretation of slope?
- What is the null hypothesis?
- What is the alternative hypothesis?
- Based on this output, what would you report to the CEO?

Question Set 6

- How would you summarize the total set of bivariate and multivariable analyses?
- Based collectively on thisall output, what would you report to the CEO?

Administrivia

- **PINs, Master Keys, Marking, and Grading**

- **Assignment 2 is due at 4:25pm next Wednesday (via Blackboard). *Blackboard will prohibit submitting assignments after 4:25pm*. Assignments will NOT be accepted by email;**

***HAVE A GREAT WEEK*!**