Session 3

Today’s Quiz

Administrivia

- ✧ Logistical issues:
  - ✧ TWO class sessions next week: Wednesday (11/18) and Friday (11/20) 4:30pm → 7:00pm;
  - ✧ Assignment 3 will be due (as usual) next Wednesday (11/18);
  - ✧ No class on 11/25;
  - ✧ Assignment 4 will be due on Wednesday 12/2;
  - ✧ Assignment 5 (final assignment) will be due on 12/9 (our last class session);
  - ✧ Final Examination: Wednesday 12/16 4:30pm → 7pm;
- ✧ PINs, master keys, marks, and grades

Picking Up From Our Last Session:
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

✧ Collect: “Newspaper” (ranging from -2=prefer social media to +2=prefer print), “Age”, “Income”

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

<u>Question Set 3 (bivariate relationship with Income)</u>

- 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?

<u>Question Set 4 (multivariable relatioship with Age)</u>

- 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

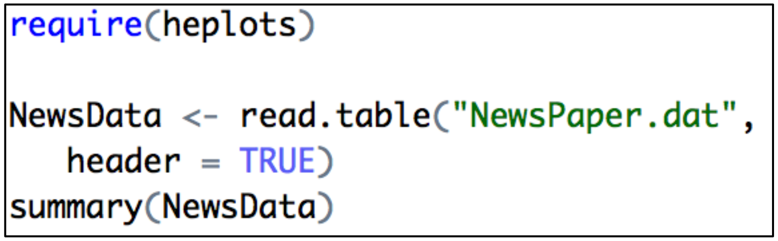
Question Set 5 (multivariable relationship with Income)

- 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?

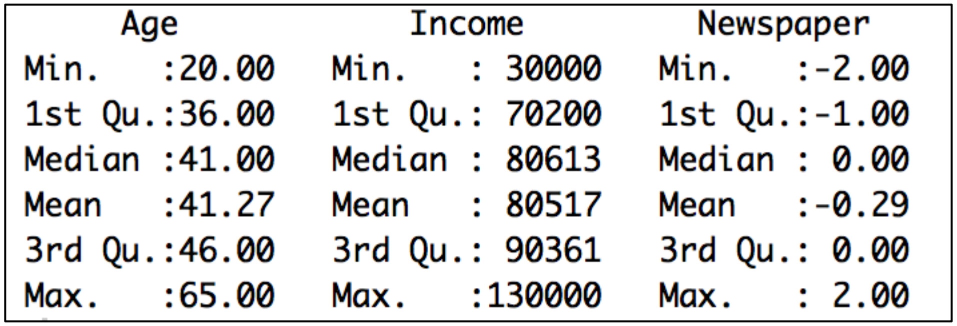
Question Set 6

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

Overview of the NewsData File

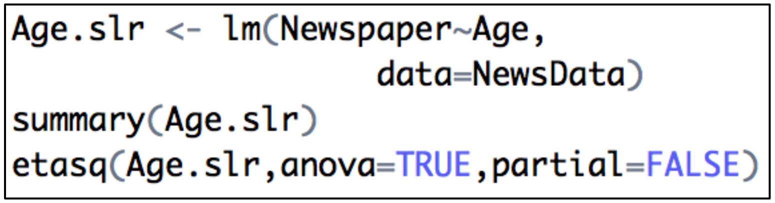


```r
require(heplots)

NewsData <- read.table("NewsPaper.dat",
   header = TRUE)
summary(NewsData)
```

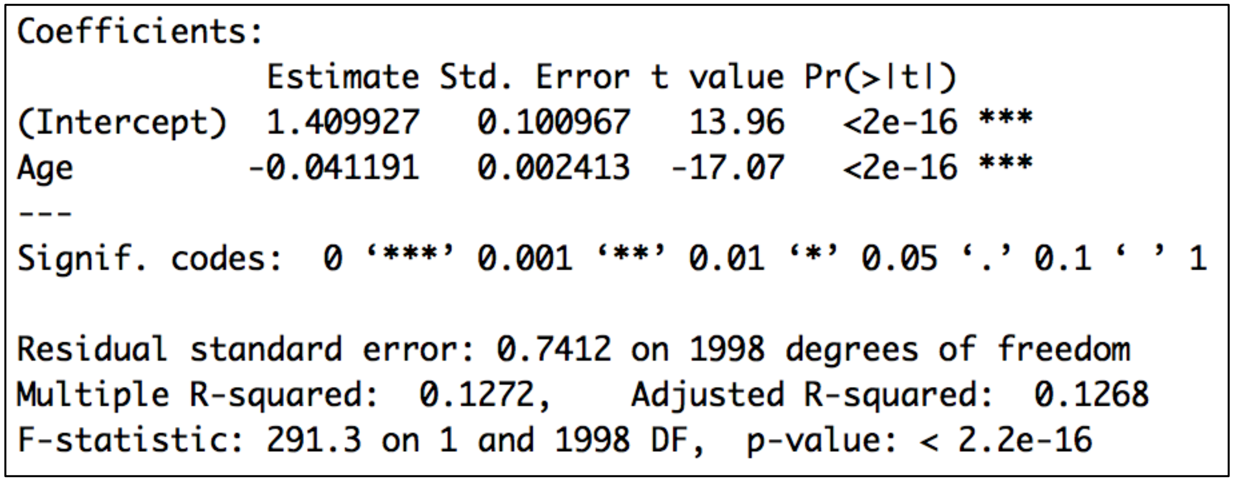


```
      Age             Income          Newspaper   
 Min.   :20.00   Min.   : 30000   Min.   :-2.00  
 1st Qu.:36.00   1st Qu.: 70200   1st Qu.:-1.00  
 Median :41.00   Median : 80613   Median : 0.00  
 Mean   :41.27   Mean   : 80517   Mean   :-0.29  
 3rd Qu.:46.00   3rd Qu.: 90361   3rd Qu.: 0.00  
 Max.   :65.00   Max.   :130000   Max.   : 2.00  
```

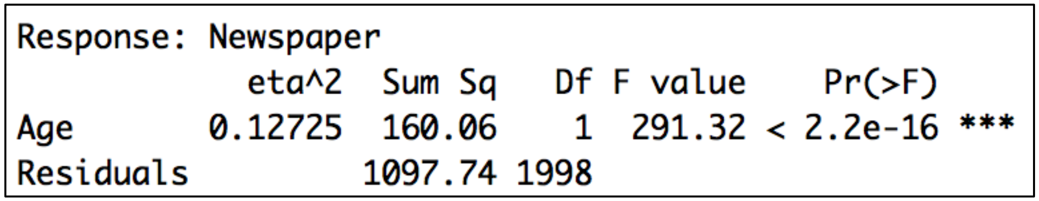
Bivariate: Age



```
Age.slr <- lm(Newspaper~Age,
                   data=NewsData)
summary(Age.slr)
etasq(Age.slr,anova=TRUE,partial=FALSE)
```

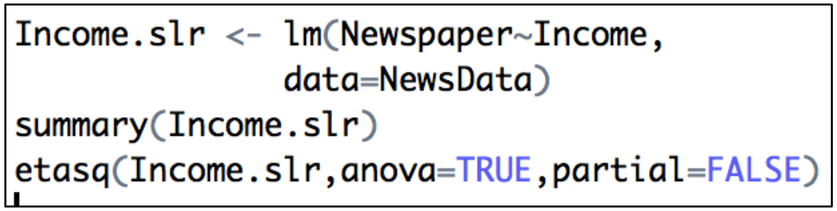


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.409927   0.100967   13.96   <2e-16 ***
Age         -0.041191   0.002413  -17.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7412 on 1998 degrees of freedom
Multiple R-squared:  0.1272,    Adjusted R-squared:  0.1268
F-statistic: 291.3 on 1 and 1998 DF,  p-value: < 2.2e-16
```

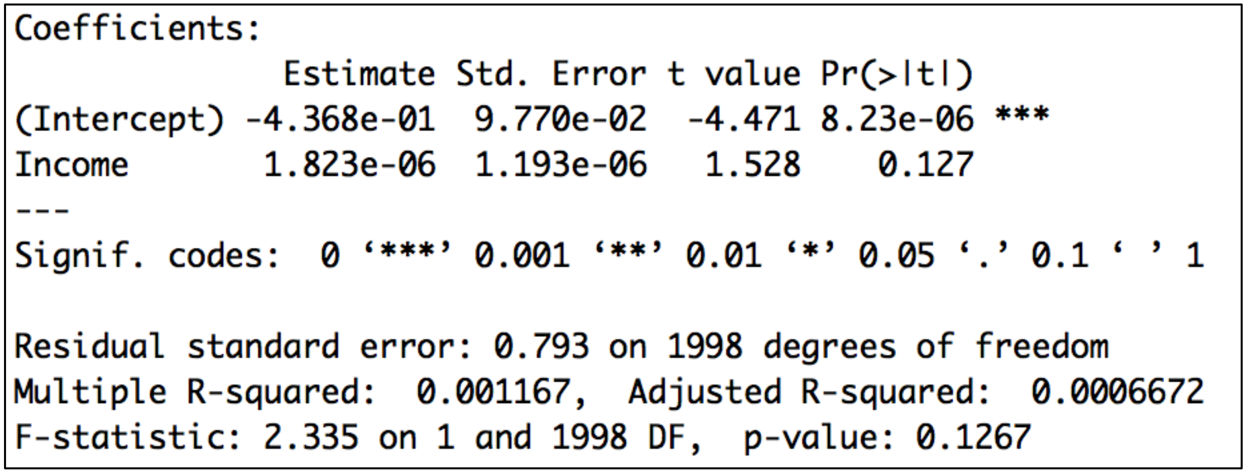


```
Response: Newspaper
            eta^2  Sum Sq   Df F value    Pr(>F)
Age       0.12725  160.06    1  291.32 < 2.2e-16 ***
Residuals         1097.74 1998
```

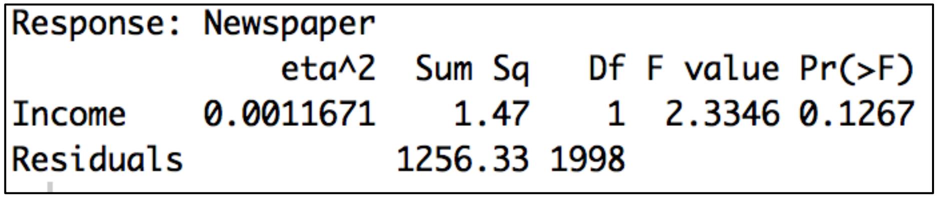
Bivariate: Income



```
Income.slr <- lm(Newspaper~Income,
                 data=NewsData)
summary(Income.slr)
etasq(Income.slr,anova=TRUE,partial=FALSE)
```

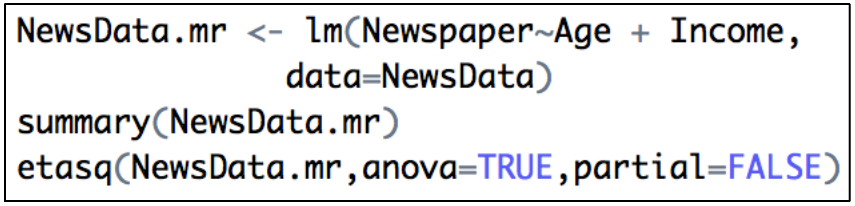


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.368e-01  9.770e-02  -4.471 8.23e-06 ***
Income       1.823e-06  1.193e-06   1.528    0.127    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.793 on 1998 degrees of freedom
Multiple R-squared:  0.001167,  Adjusted R-squared:  0.0006672 
F-statistic: 2.335 on 1 and 1998 DF,  p-value: 0.1267
```

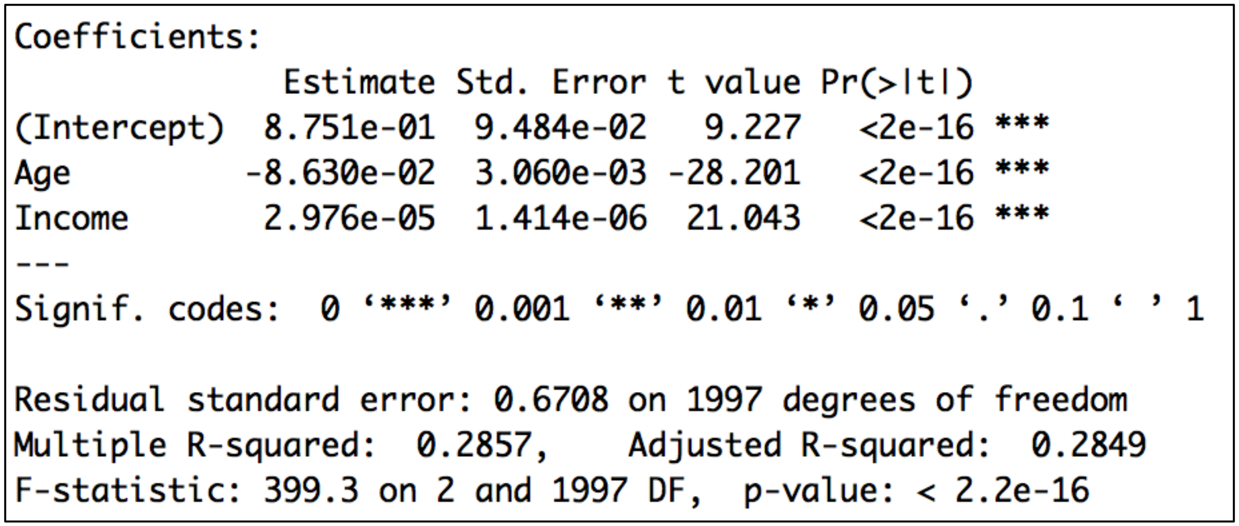


```
Response: Newspaper
              eta^2  Sum Sq   Df F value Pr(>F)
Income    0.0011671    1.47    1  2.3346 0.1267
Residuals           1256.33 1998               
```

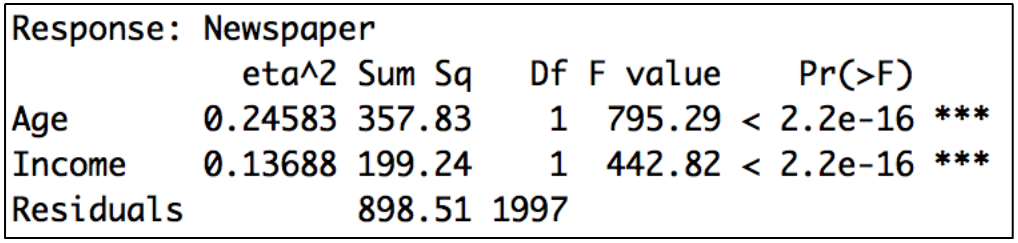
Multivariable: Age + Income



```
NewsData.mr <- lm(Newspaper~Age + Income,
                  data=NewsData)
summary(NewsData.mr)
etasq(NewsData.mr,anova=TRUE,partial=FALSE)
```



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.751e-01  9.484e-02   9.227   <2e-16 ***
Age         -8.630e-02  3.060e-03 -28.201   <2e-16 ***
Income       2.976e-05  1.414e-06  21.043   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6708 on 1997 degrees of freedom
Multiple R-squared:  0.2857,	Adjusted R-squared:  0.2849 
F-statistic: 399.3 on 2 and 1997 DF,  p-value: < 2.2e-16
```



```
Response: Newspaper
            eta^2 Sum Sq   Df F value    Pr(>F)    
Age       0.24583 357.83    1  795.29 < 2.2e-16 ***
Income    0.13688 199.24    1  442.82 < 2.2e-16 ***
Residuals         898.51 1997
```

Returning to 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
- Important note: the data used in today's session differ slightly from the data used in last week's session. Feel free to download today's data from Outline/Session 3 in Blackboard.

Review: Unconditional vs. Conditional Relationships

Two weeks ago, 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.

Last week, we 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 <u>unconditional</u> (i.e., bivariate) relationship…

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

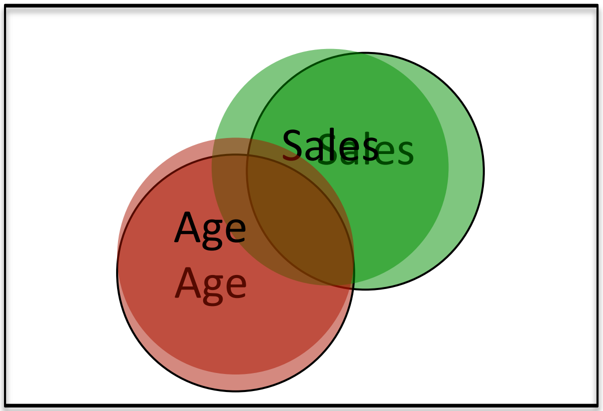
Thus, multiple regression is not the union of a set of simple linear regressions: multivariable results can be quite different from bivariate results.

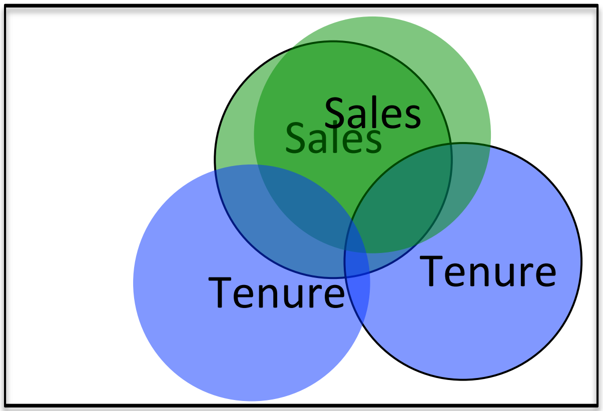
The one case where multiple regression <u>*is*</u> the union of the set of simple linear regressions is when there is no multicollinearity.

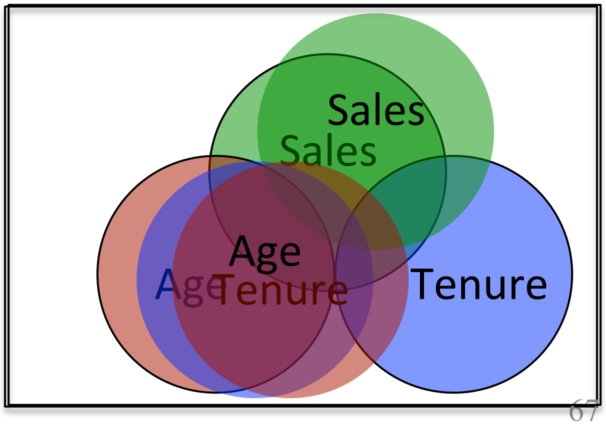
When interpreting results, it is important to consider both the unconditional relationships and the conditional relationships: they can provide complementary information.

Thus there are six (soon to be eight) measures of importance: <u>unconditional</u> p-values, bivariate coefficients of determination, <u>conditional</u> p-values, coefficients of partial determination, the coefficient of multiple determination, and the adjusted coefficient of multiple determination. These measures all tell you different things about your data.

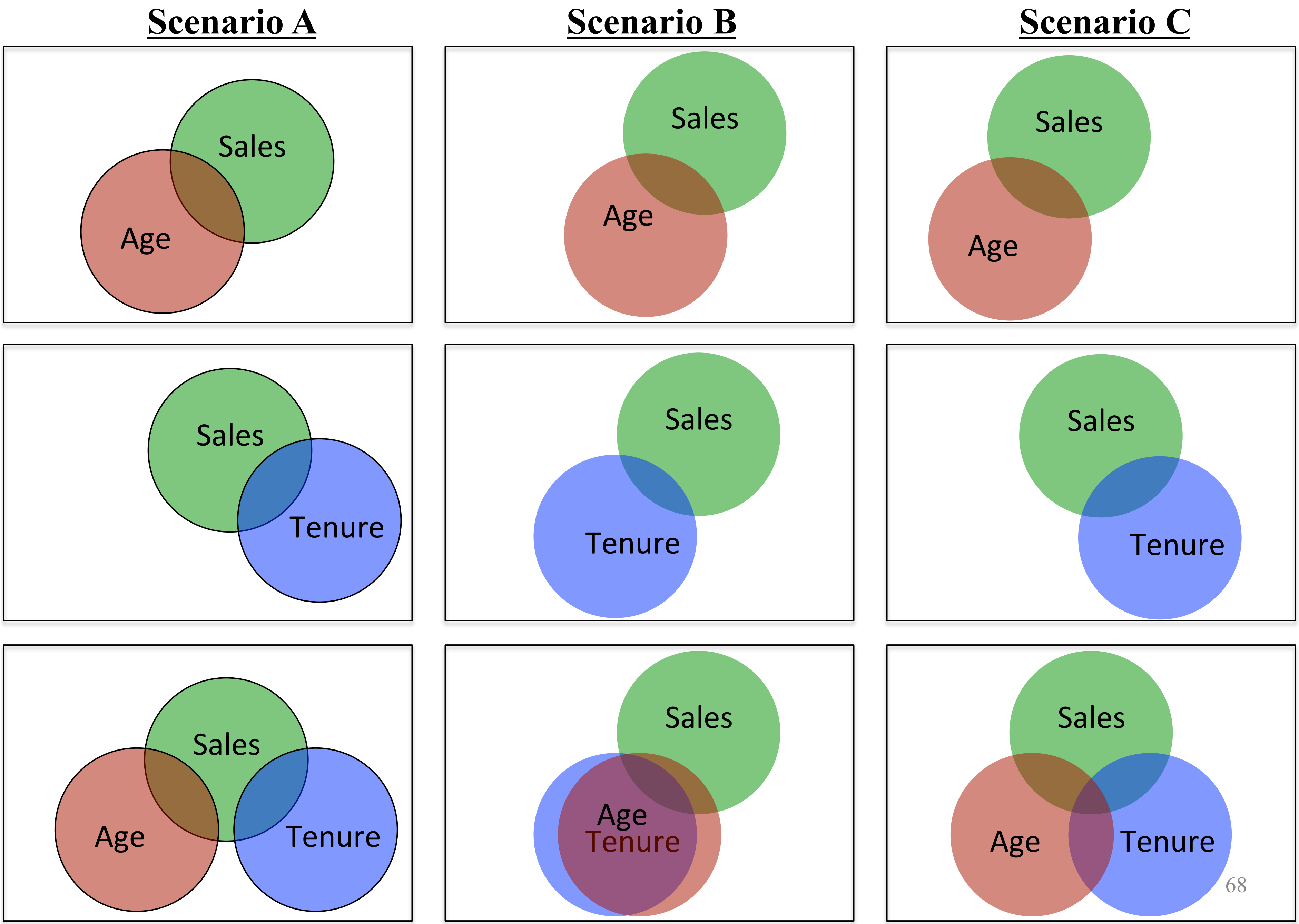
Today we will see how the <u>slope</u> can (and often does) differ between the unconditional and conditional paradigms.







Review: Three Scenarios

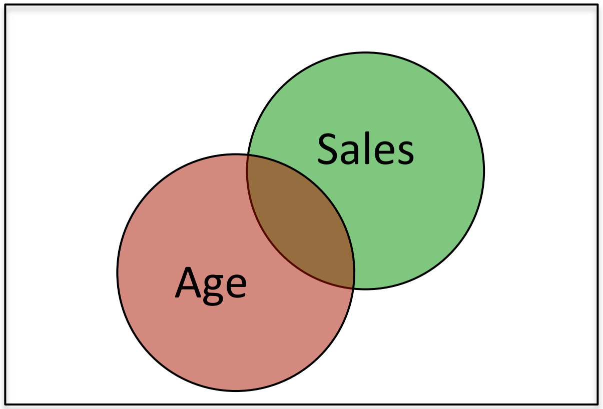


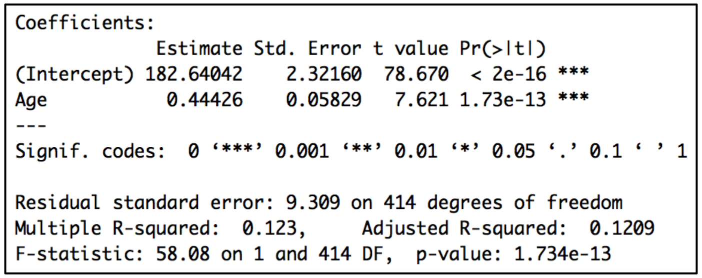
Scenario A

Scenario B

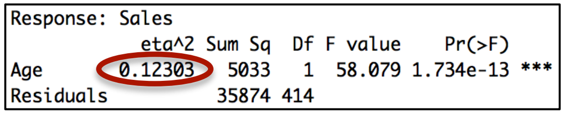
Scenario C

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

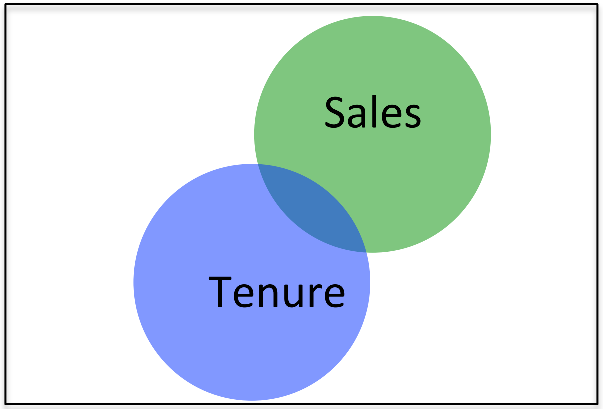


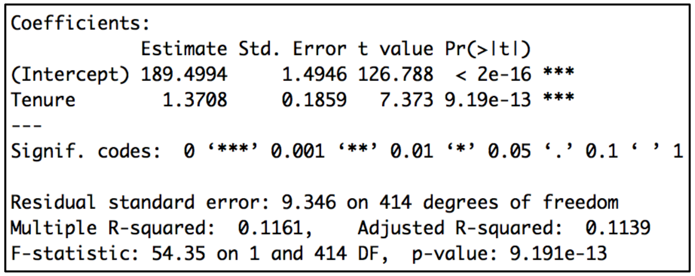


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 182.64042    2.32160  78.670  < 2e-16 ***
Age           0.44426    0.05829   7.621 1.73e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.309 on 414 degrees of freedom
Multiple R-squared:  0.123,	Adjusted R-squared:  0.1209
F-statistic: 58.08 on 1 and 414 DF,  p-value: 1.734e-13
```

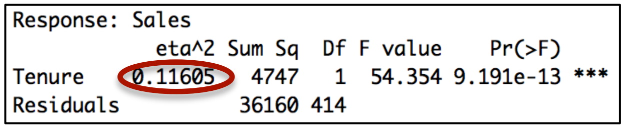


```
Response: Sales
            eta^2 Sum Sq  Df F value    Pr(>F)
Age       0.12303   5033   1  58.079 1.734e-13 ***
Residuals          35874 414
```

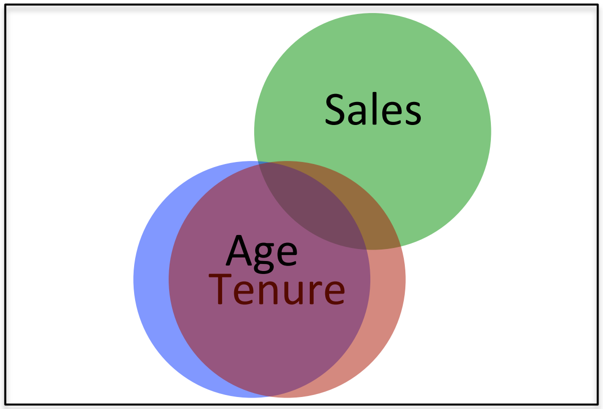


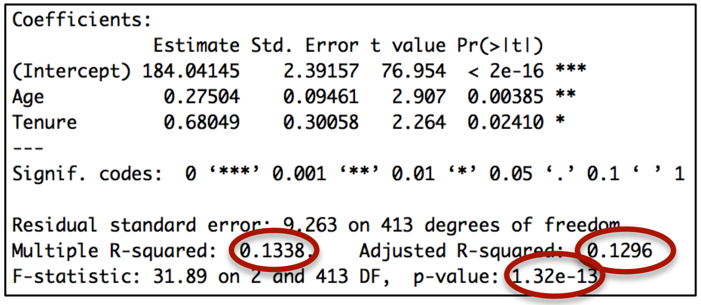


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 189.4994     1.4946 126.788  < 2e-16 ***
Tenure        1.3708     0.1859   7.373 9.19e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.346 on 414 degrees of freedom
Multiple R-squared:  0.1161,	Adjusted R-squared:  0.1139
F-statistic: 54.35 on 1 and 414 DF,  p-value: 9.191e-13
```

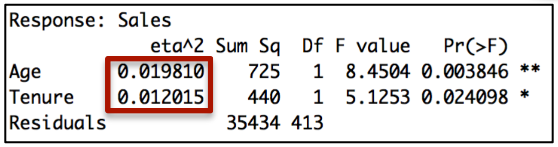


```
Response: Sales
            eta^2 Sum Sq  Df F value    Pr(>F)
Tenure    0.11605   4747   1  54.354 9.191e-13 ***
Residuals          36160 414
```



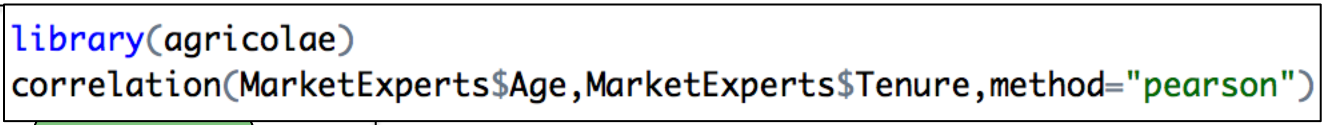


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 184.04145    2.39157  76.954  < 2e-16 ***
Age           0.27504    0.09461   2.907  0.00385 **
Tenure        0.68049    0.30058   2.264  0.02410 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.263 on 413 degrees of freedom
Multiple R-squared:  0.1338,	Adjusted R-squared:  0.1296
F-statistic: 31.89 on 2 and 413 DF,  p-value: 1.32e-13
```

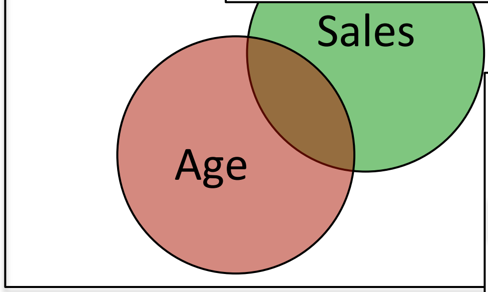


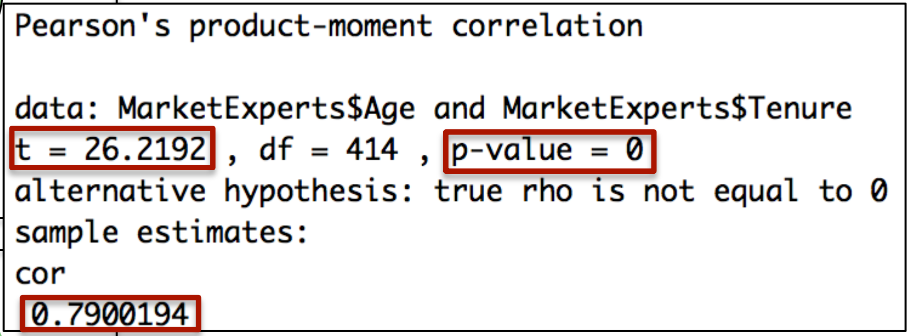
```
Response: Sales
             eta^2 Sum Sq  Df F value   Pr(>F)
Age       0.019810    725   1  8.4504 0.003846 **
Tenure    0.012015    440   1  5.1253 0.024098 *
Residuals           35434 413
```

Measuring the Co-Relationship Between Two Variables: The Coefficient of Correlation (Scenario B)

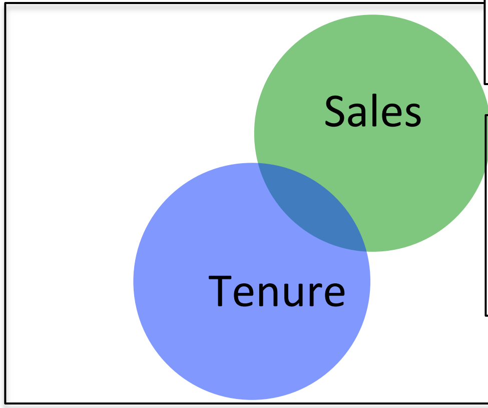


```
library(agricolae)
correlation(MarketExperts$Age,MarketExperts$Tenure,method="pearson")
```

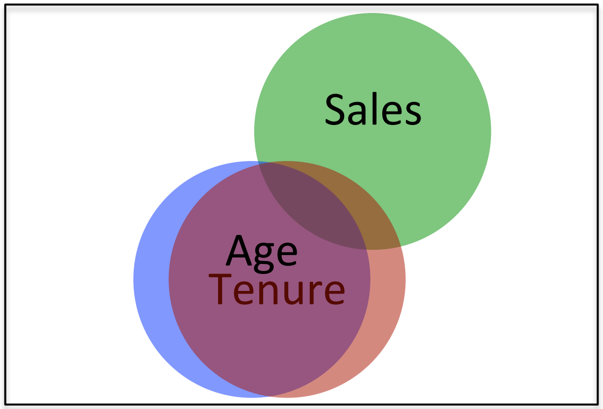


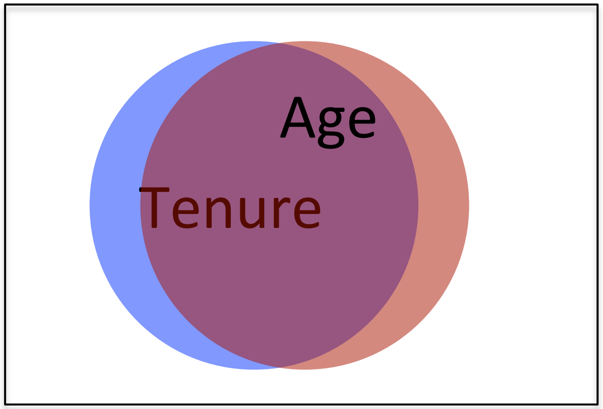


```
Pearson's product-moment correlation

data: MarketExperts$Age and MarketExperts$Tenure
t = 26.2192 , df = 414 , p-value = 0
alternative hypothesis: true rho is not equal to 0
sample estimates:
cor
 0.7900194
```

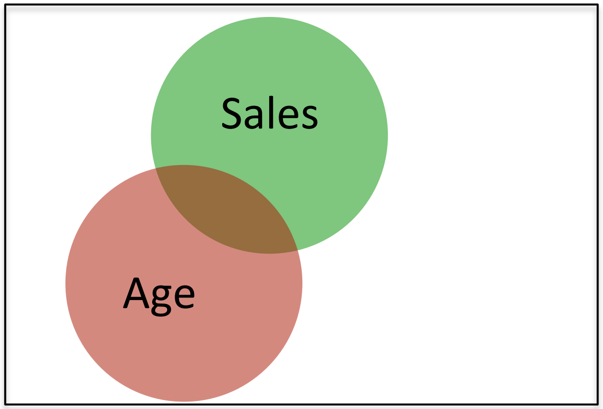


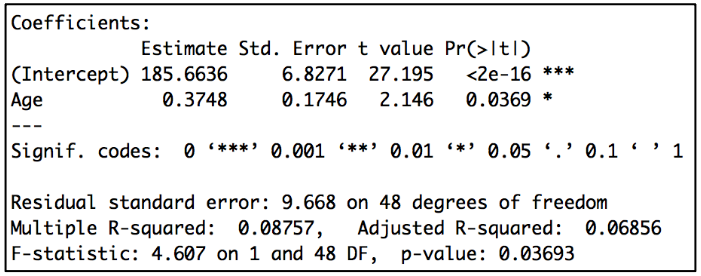
Based on these results, would it be appropriate to write a report to the CEO suggesting that he/she concentrate on hiring older people, because older people tend to have higher sales?



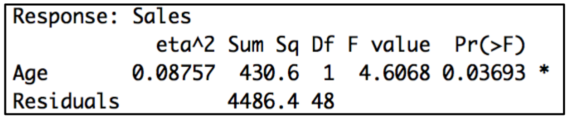


Multiple Regression Results: Scenario C

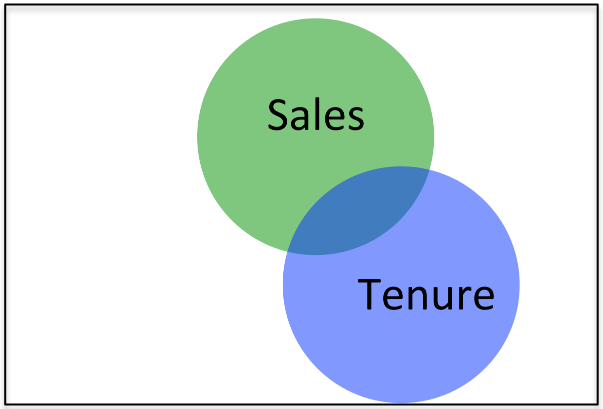


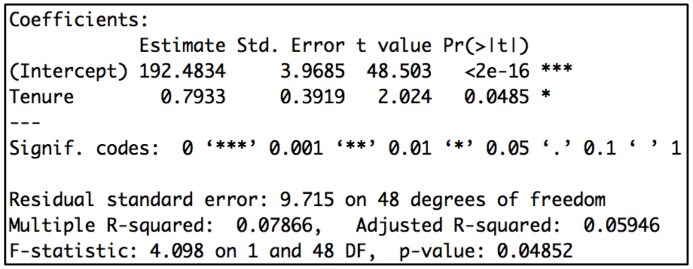


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 185.6636     6.8271  27.195   <2e-16 ***
Age           0.3748     0.1746   2.146   0.0369 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

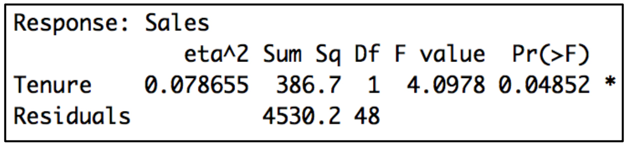


```
Response: Sales
           eta^2 Sum Sq Df F value  Pr(>F)
Age      0.08757  430.6  1  4.6068 0.03693 *
Residuals        4486.4 48
```

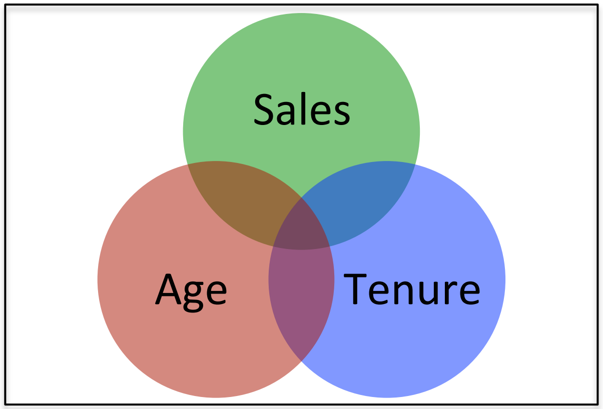


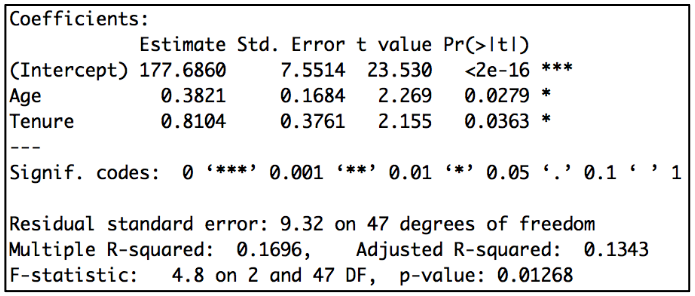


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.4834     3.9685  48.503   <2e-16 ***
Tenure        0.7933     0.3919   2.024   0.0485 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

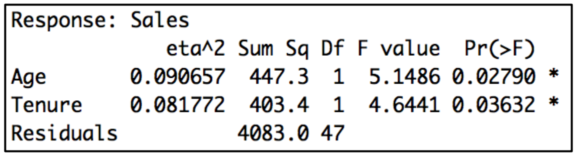


```
Response: Sales
             eta^2 Sum Sq Df F value  Pr(>F)
Tenure    0.078655  386.7  1  4.0978 0.04852 *
Residuals          4530.2 48
```





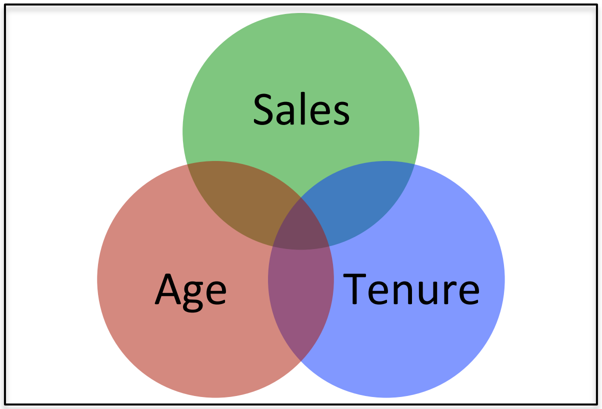
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 177.6860     7.5514  23.530   <2e-16 ***
Age           0.3821     0.1684   2.269   0.0279 *
Tenure        0.8104     0.3761   2.155   0.0363 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```



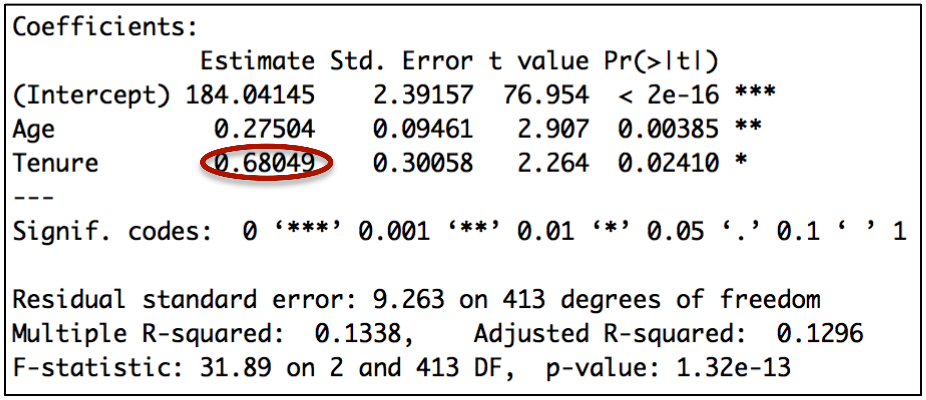
```
Response: Sales
             eta^2 Sum Sq Df F value  Pr(>F)
Age       0.090657  447.3  1  5.1486 0.02790 *
Tenure    0.081772  403.4  1  4.6441 0.03632 *
Residuals          4083.0 47
```

Would your recommendation to the CEO differ from the recommendation for Scenario B?

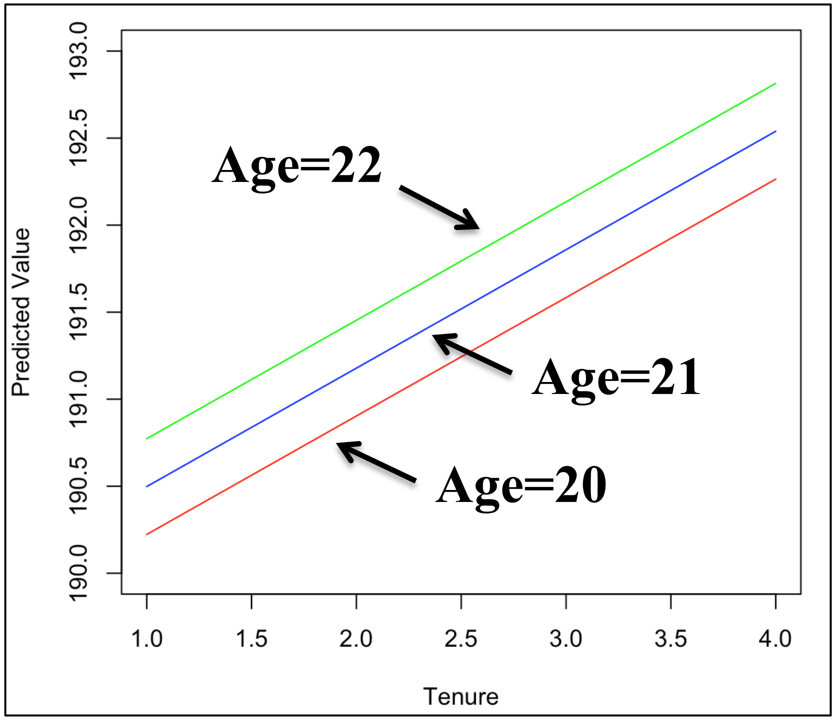
The Meaning of “Slope” in Multiple Regression: “Controlling For” (or “Holding Constant”) Age (Scenario C)

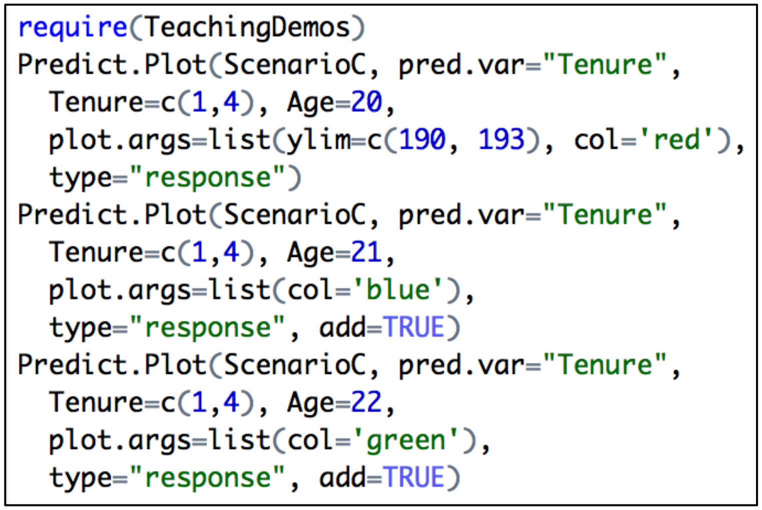


Population regression equation:
$\text{Sales} = \beta_0 + \beta_1 * \text{Age} + \beta_2 * \text{Tenure} + \varepsilon$



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 184.04145    2.39157  76.954  < 2e-16 ***
Age           0.27504    0.09461   2.907  0.00385 ** 
Tenure        0.68049    0.30058   2.264  0.02410 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.263 on 413 degrees of freedom
Multiple R-squared:  0.1338,    Adjusted R-squared:  0.1296 
F-statistic: 31.89 on 2 and 413 DF,  p-value: 1.32e-13
```

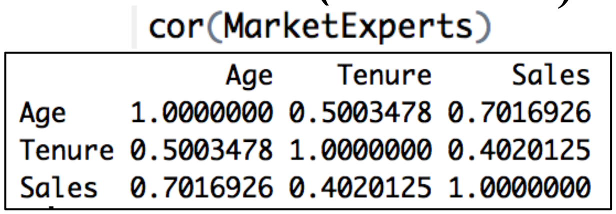




```
require(TeachingDemos)
Predict.Plot(ScenarioC, pred.var="Tenure",
  Tenure=c(1,4), Age=20,
  plot.args=list(ylim=c(190, 193), col='red'),
  type="response")
Predict.Plot(ScenarioC, pred.var="Tenure",
  Tenure=c(1,4), Age=21,
  plot.args=list(col='blue'),
  type="response", add=TRUE)
Predict.Plot(ScenarioC, pred.var="Tenure",
  Tenure=c(1,4), Age=22,
  plot.args=list(col='green'),
  type="response", add=TRUE)
```

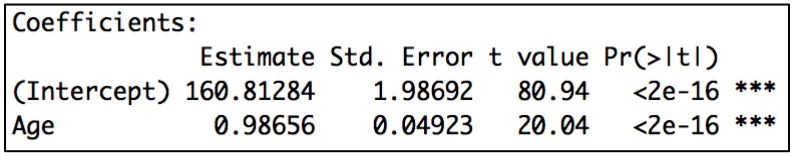
Simple Linear Regression vs. Multiple Regression: An Example of the Effects of Moderate Multicollinearity

Correlations (Scenario D)



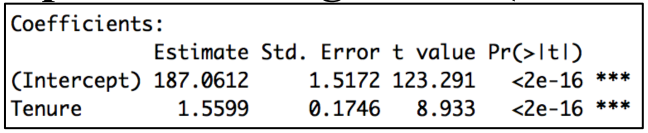
```
cor(MarketExperts)
              Age     Tenure     Sales
Age     1.0000000 0.5003478 0.7016926
Tenure  0.5003478 1.0000000 0.4020125
Sales   0.7016926 0.4020125 1.0000000
```

Simple Linear Regression (Age)



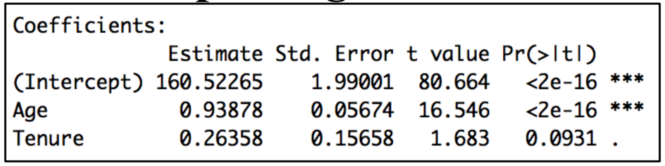
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 160.81284    1.98692   80.94   <2e-16 ***
Age           0.98656    0.04923   20.04   <2e-16 ***
```

Simple Linear Regression (Tenure)

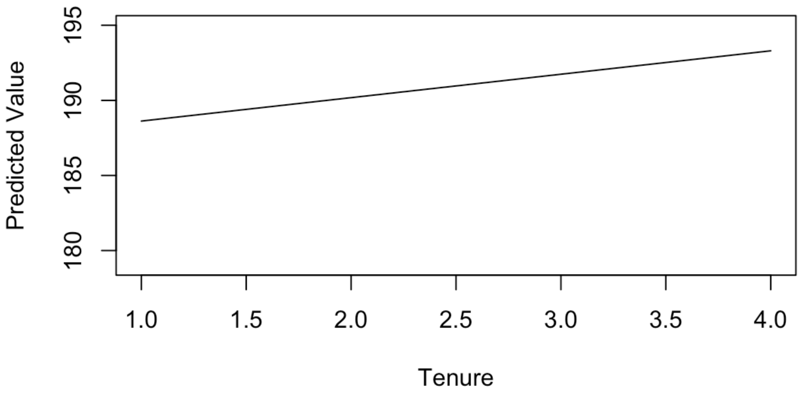


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 187.0612     1.5172 123.291   <2e-16 ***
Tenure        1.5599     0.1746   8.933   <2e-16 ***
```

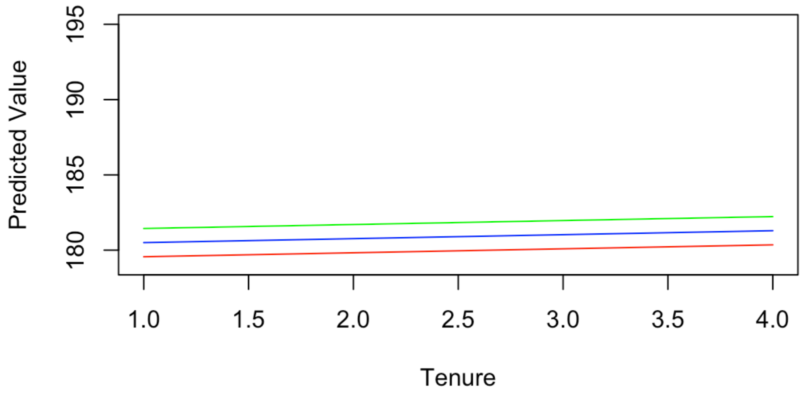
Multiple Regression



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 160.52265    1.99001  80.664   <2e-16 ***
Age           0.93878    0.05674  16.546   <2e-16 ***
Tenure        0.26358    0.15658   1.683   0.0931 .  
```



b_{Tenure}=1.5599

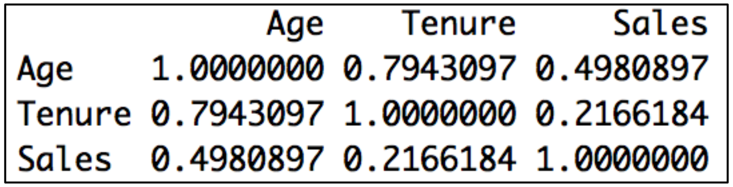


b_{Tenure}=0.2636

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

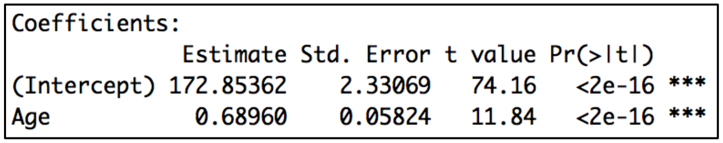
A Continuous Variable Variation of Simpson's Paradox: An Example of Severe Multicollinearity

Correlations (Scenario E)



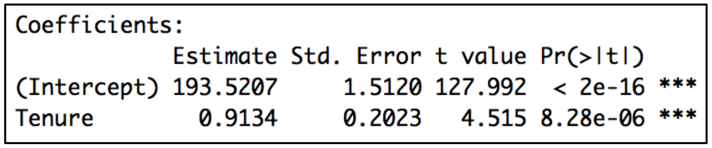
```
            Age    Tenure     Sales
Age    1.0000000 0.7943097 0.4980897
Tenure 0.7943097 1.0000000 0.2166184
Sales  0.4980897 0.2166184 1.0000000
```

Simple Linear Regression (Age)



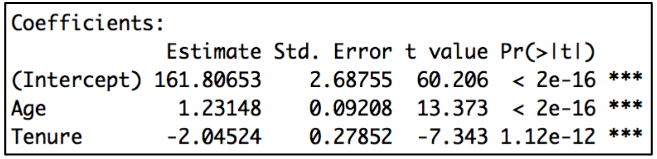
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 172.85362    2.33069   74.16   <2e-16 ***
Age           0.68960    0.05824   11.84   <2e-16 ***
```

Simple Linear Regression (Tenure)

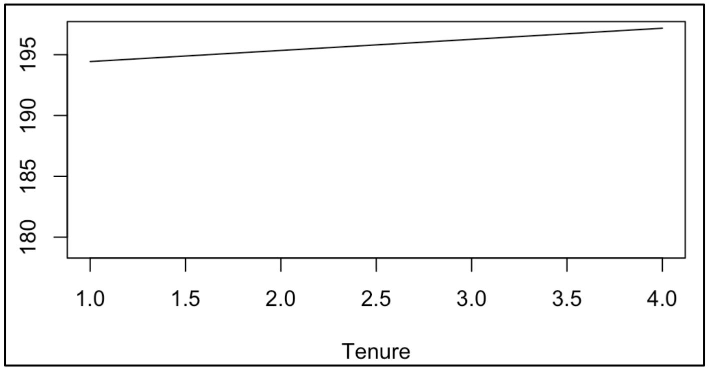


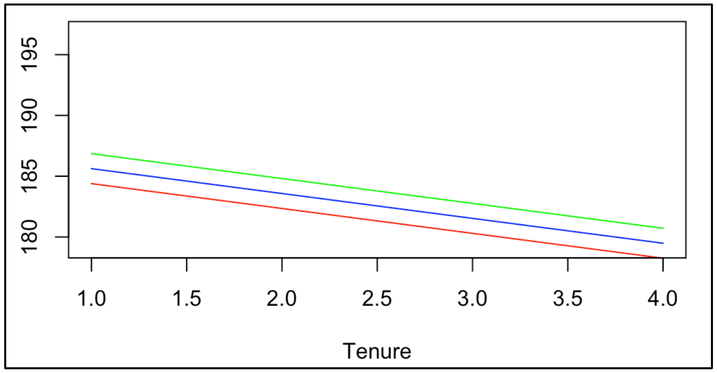
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 193.5207     1.5120 127.992  < 2e-16 ***
Tenure        0.9134     0.2023   4.515 8.28e-06 ***
```

Multiple Regression



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 161.80653    2.68755  60.206  < 2e-16 ***
Age           1.23148    0.09208  13.373  < 2e-16 ***
Tenure       -2.04524    0.27852  -7.343 1.12e-12 ***
```





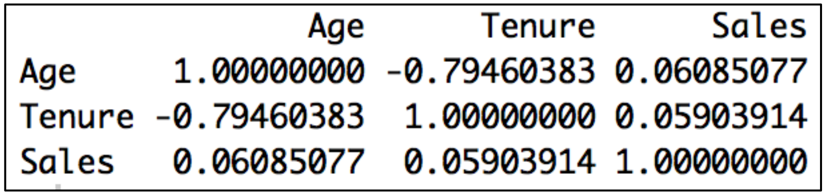
$b_{Tenure} = 0.9134$

$b_{Tenure} = -2.0452$

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

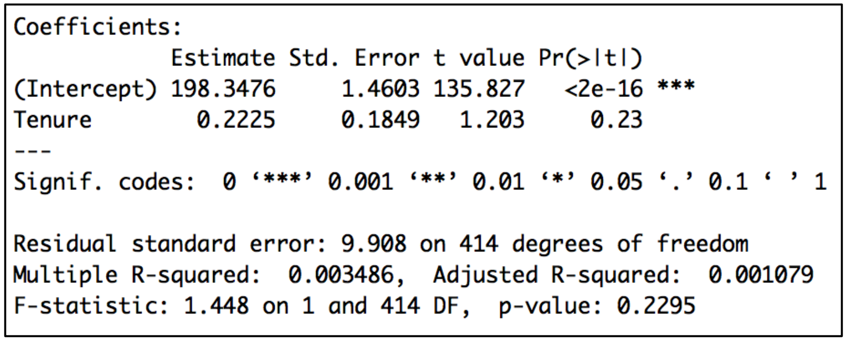
The Suppressor Effect: Introduction

Correlations (Scenario H)



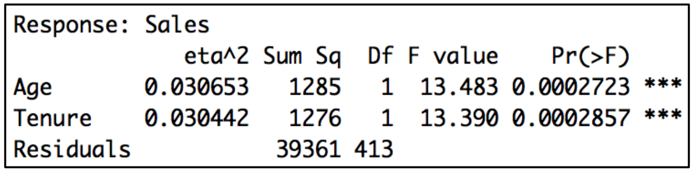
```
                Age      Tenure      Sales
Age      1.00000000 -0.79460383 0.06085077
Tenure  -0.79460383  1.00000000 0.05903914
Sales    0.06085077  0.05903914 1.00000000
```

Simple Linear Regression (Tenure)



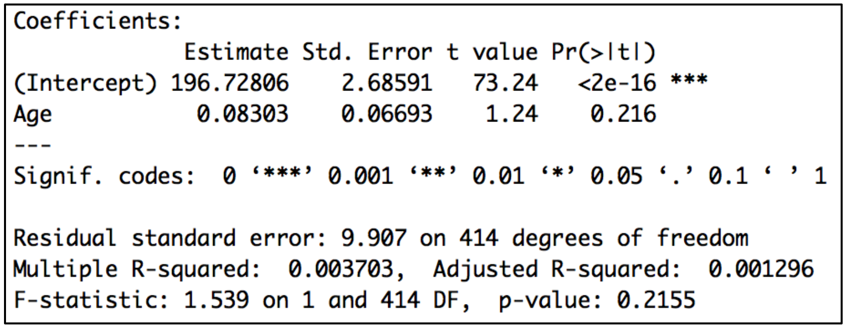
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 198.3476     1.4603 135.827   <2e-16 ***
Tenure        0.2225     0.1849   1.203     0.23    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.908 on 414 degrees of freedom
Multiple R-squared:  0.003486,	Adjusted R-squared:  0.001079 
F-statistic: 1.448 on 1 and 414 DF,  p-value: 0.2295
```

Coefficients of Partial Determination



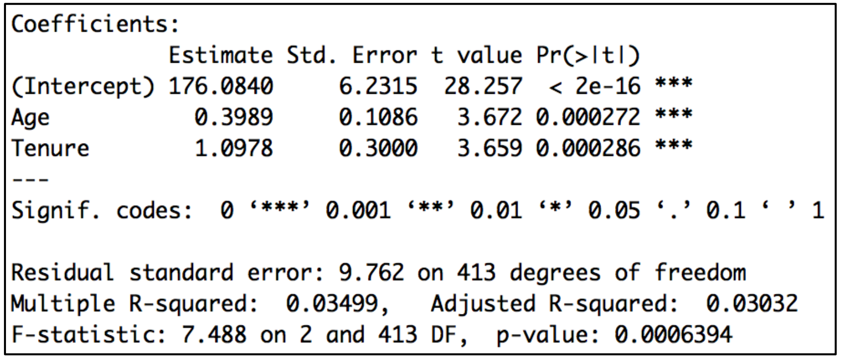
```
Response: Sales
             eta^2 Sum Sq  Df F value    Pr(>F)    
Age       0.030653   1285   1  13.483 0.0002723 ***
Tenure    0.030442   1276   1  13.390 0.0002857 ***
Residuals           39361 413                      
```

Simple Linear Regression (Age)



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 196.72806    2.68591   73.24   <2e-16 ***
Age           0.08303    0.06693    1.24    0.216    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.907 on 414 degrees of freedom
Multiple R-squared:  0.003703,	Adjusted R-squared:  0.001296 
F-statistic: 1.539 on 1 and 414 DF,  p-value: 0.2155
```

Multiple Regression

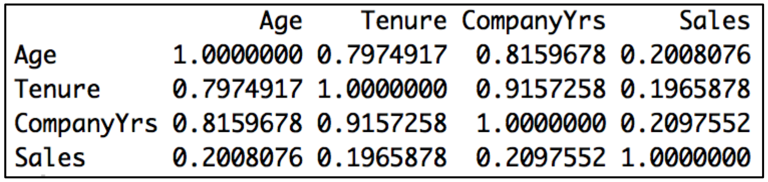


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 176.0840     6.2315  28.257  < 2e-16 ***
Age           0.3989     0.1086   3.672 0.000272 ***
Tenure        1.0978     0.3000   3.659 0.000286 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.762 on 413 degrees of freedom
Multiple R-squared:  0.03499,	Adjusted R-squared:  0.03032 
F-statistic: 7.488 on 2 and 413 DF,  p-value: 0.0006394
```

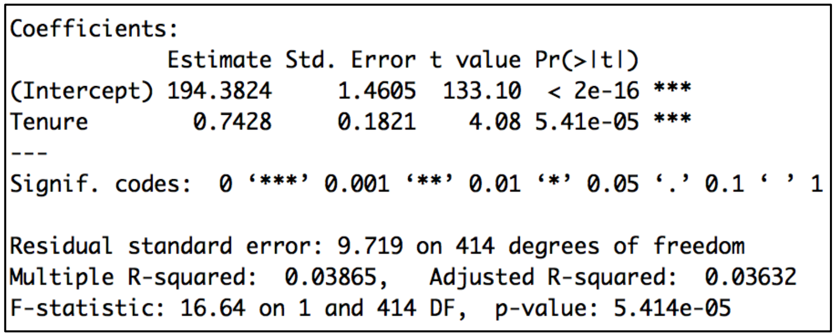
Based on these results, what advice would you offer to the CEO about Market Experts’ personnel policy?

Adding More Independent Variables

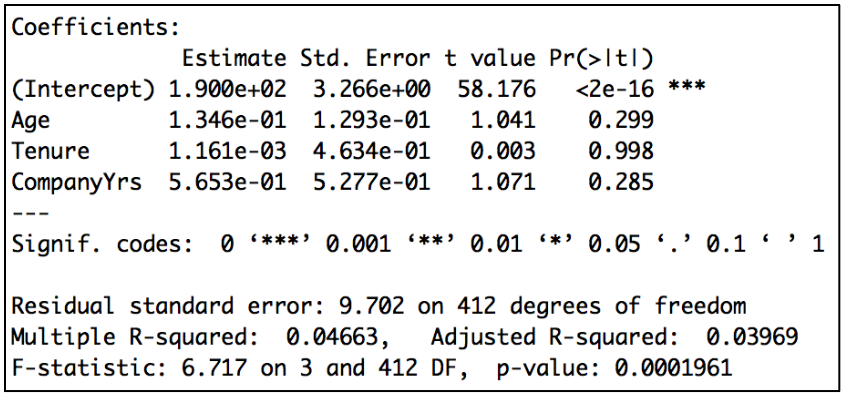
Scenario F: Age, Tenure, *CompanyYrs*, and Sales are measured. "CompanyYrs" is the number of years working for the Company in any capacity (not just as a marketer); the other variables are defined as before. How would you interpret this output, and which independent variable is the best predictor of the dependent variable?



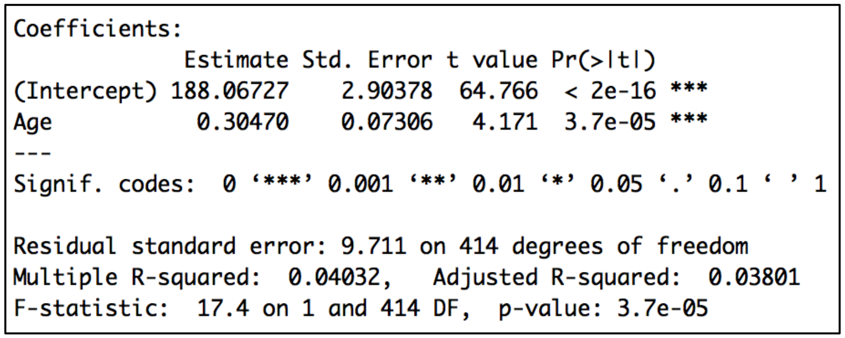
```
                 Age    Tenure CompanyYrs     Sales
Age        1.0000000 0.7974917  0.8159678 0.2008076
Tenure     0.7974917 1.0000000  0.9157258 0.1965878
CompanyYrs 0.8159678 0.9157258  1.0000000 0.2097552
Sales      0.2008076 0.1965878  0.2097552 1.0000000
```



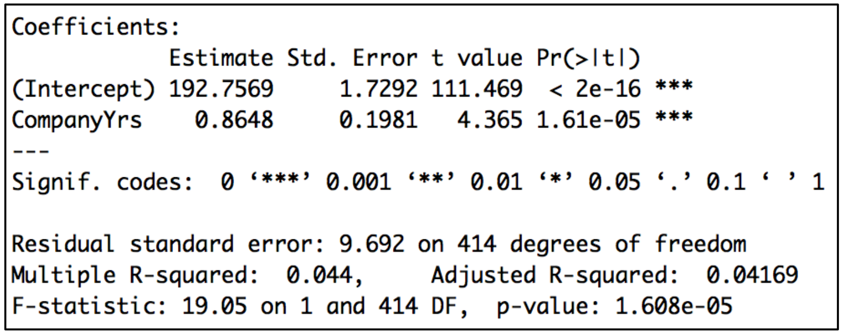
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 194.3824     1.4605  133.10  < 2e-16 ***
Tenure        0.7428     0.1821    4.08 5.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.719 on 414 degrees of freedom
Multiple R-squared:  0.03865,	Adjusted R-squared:  0.03632 
F-statistic: 16.64 on 1 and 414 DF,  p-value: 5.414e-05
```



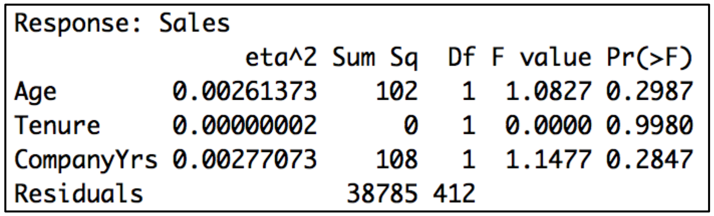
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.900e+02  3.266e+00  58.176   <2e-16 ***
Age         1.346e-01  1.293e-01   1.041    0.299    
Tenure      1.161e-03  4.634e-01   0.003    0.998    
CompanyYrs  5.653e-01  5.277e-01   1.071    0.285    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.702 on 412 degrees of freedom
Multiple R-squared:  0.04663,	Adjusted R-squared:  0.03969 
F-statistic: 6.717 on 3 and 412 DF,  p-value: 0.0001961
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 188.06727    2.90378  64.766  < 2e-16 ***
Age           0.30470    0.07306   4.171  3.7e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.711 on 414 degrees of freedom
Multiple R-squared:  0.04032,	Adjusted R-squared:  0.03801 
F-statistic:  17.4 on 1 and 414 DF,  p-value: 3.7e-05
```

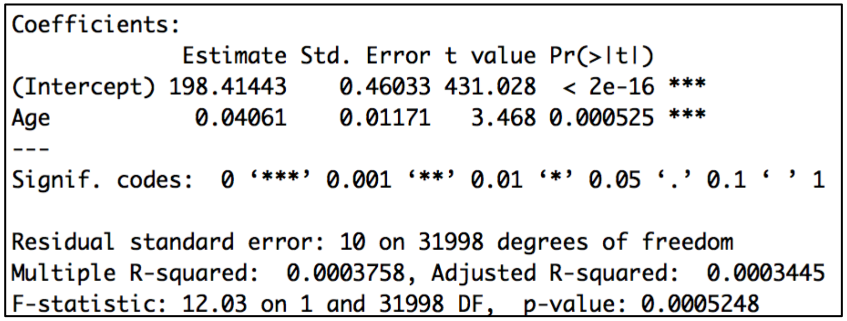


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 192.7569     1.7292 111.469  < 2e-16 ***
CompanyYrs    0.8648     0.1981   4.365 1.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.692 on 414 degrees of freedom
Multiple R-squared:  0.044,	Adjusted R-squared:  0.04169 
F-statistic: 19.05 on 1 and 414 DF,  p-value: 1.608e-05
```

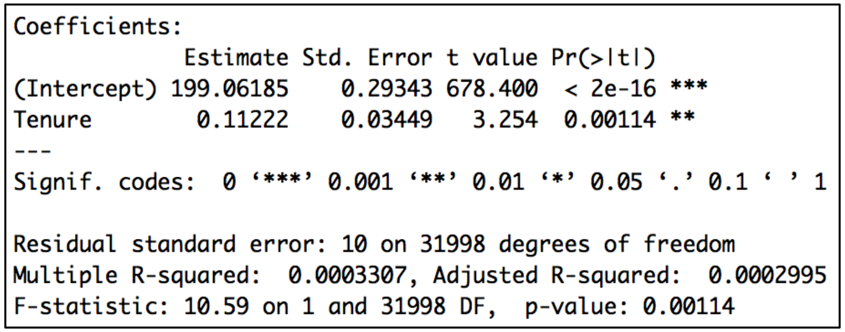


```
Response: Sales
                eta^2 Sum Sq  Df F value Pr(>F)
Age        0.00261373    102   1  1.0827 0.2987
Tenure     0.00000002      0   1  0.0000 0.9980
CompanyYrs 0.00277073    108   1  1.1477 0.2847
Residuals              38785 412               
```

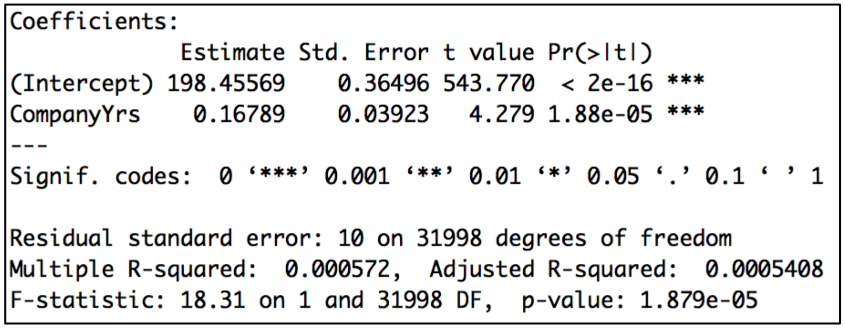
Regression and Big Data (Scenario G: 32000 observations)



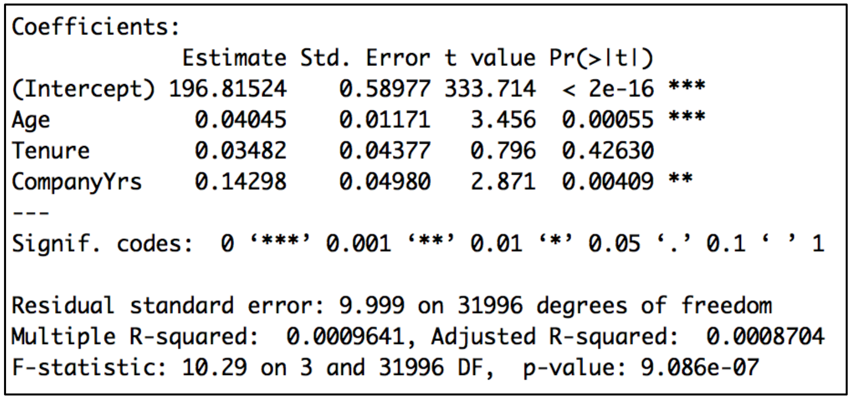
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 198.41443    0.46033 431.028  < 2e-16 ***
Age           0.04061    0.01171   3.468 0.000525 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10 on 31998 degrees of freedom
Multiple R-squared:  0.0003758, Adjusted R-squared:  0.0003445 
F-statistic: 12.03 on 1 and 31998 DF,  p-value: 0.0005248
```



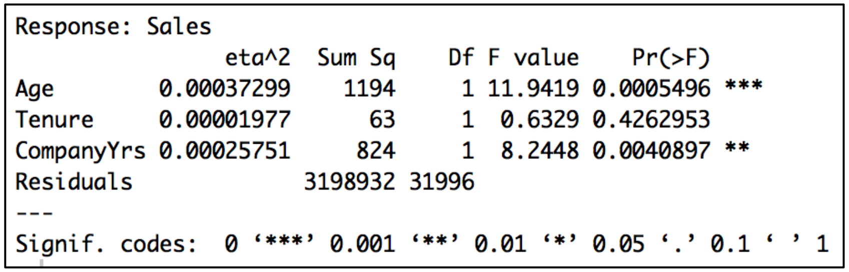
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 199.06185    0.29343 678.400  < 2e-16 ***
Tenure        0.11222    0.03449   3.254  0.00114 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10 on 31998 degrees of freedom
Multiple R-squared:  0.0003307, Adjusted R-squared:  0.0002995 
F-statistic: 10.59 on 1 and 31998 DF,  p-value: 0.00114
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 198.45569    0.36496 543.770  < 2e-16 ***
CompanyYrs    0.16789    0.03923   4.279 1.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10 on 31998 degrees of freedom
Multiple R-squared:  0.000572,  Adjusted R-squared:  0.0005408 
F-statistic: 18.31 on 1 and 31998 DF,  p-value: 1.879e-05
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 196.81524    0.58977 333.714  < 2e-16 ***
Age           0.04045    0.01171   3.456  0.00055 ***
Tenure        0.03482    0.04377   0.796  0.42630    
CompanyYrs    0.14298    0.04980   2.871  0.00409 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.999 on 31996 degrees of freedom
Multiple R-squared:  0.0009641, Adjusted R-squared:  0.0008704 
F-statistic: 10.29 on 3 and 31996 DF,  p-value: 9.086e-07
```



```
Response: Sales
                eta^2  Sum Sq    Df F value    Pr(>F)    
Age        0.00037299    1194     1 11.9419 0.0005496 ***
Tenure     0.00001977      63     1  0.6329 0.4262953    
CompanyYrs 0.00025751     824     1  8.2448 0.0040897 ** 
Residuals             3198932 31996                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

How would you summarize these data in your report to the CEO? What are the most salient features of these results?

Summary: Unconditional vs. Conditional Relationships

Last week, we learned about the distinction between statistical significance and strength of relationship in a multivariable context, but did not focus specifically on the nature of the slope.

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

Two continuous variables which are unconditionally related to each other in a specific way…

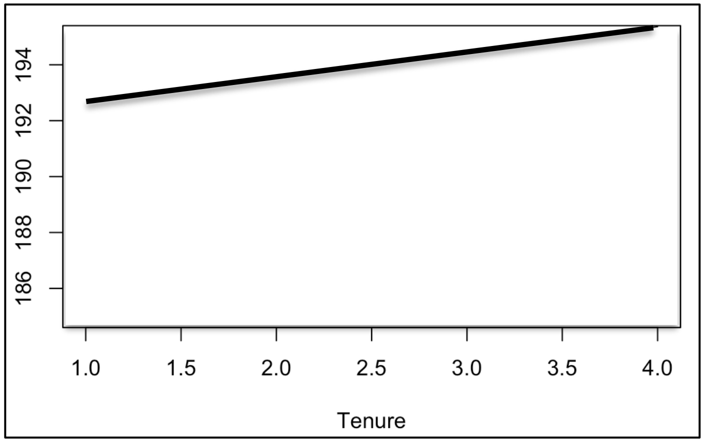
may be conditionally related to each other in a very different way.

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

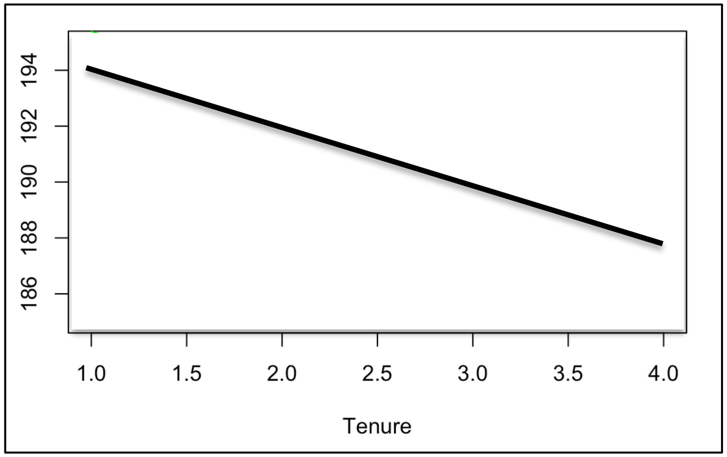
The one case where multiple regression *is* the union of the set of simple linear regressions is when there is no multicollinearity.

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

Thus there are eight measures of importance: unconditional p-values, bivariate coefficients of determination, and slope; conditional p-values, coefficients of partial determination, and slope; the coefficient of multiple determination, and the adjusted coefficient of multiple determination. These measures all tell you different things about your data.



b_{Tenure}=0.9570



b_{Tenure}=-2.09386

Administrivia

- **Assignment 3 is due at 7:00pm next Wednesday (via Blackboard). *Blackboard will prohibit submitting assignments after 4:25pm*. Assignments will NOT be accepted by email to the instructor or the GTA;**

- **See you next Wednesday and next Friday.**

***HAVE A GREAT WEEK*!**