Session 4

Administrivia

- ✧ Logistical issues:
  - ✧ TWO class sessions this week: today (11/18) and Friday (11/20) 4:30pm → 7:00pm;
  - ✧ NO class session next week (Thanksgiving week);
  - ✧ Assignment 4 will be available at 9:40pm <u>this Friday</u> (not today), and will be due at 7:00pm on Wednesday 12/2;
  - ✧ Assignment 5 (final assignment) will be assigned on Wednesday 12/2 and due on 12/9 (our last class session);
  - ✧ Final Examination: Wednesday 12/12 4:30pm → 7:00pm

Finishing 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

<u>Question Set 1</u>

- 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?

<u>Question Set 2 (bivariate relationship 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 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 all this 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                      
```

Question Sets 1 and 2

<u>Question Set 1</u>

- 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?

<u>Question Set 2 (bivariate relationship 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 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

<u>Question Set 5 (multivariable 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 6</u>

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

Overview of the NewsData File

```
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
```

Last Week's Discussion:
Unconditional vs. Conditional Relationships

Two weeks ago, 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.

Last week, we looked at the <u>slope</u> under two situations: unconditional (bivariate), and conditional (multivariable). We learned that…

Two continuous variables which are <u>unconditionally</u> related to each other in a specific way…

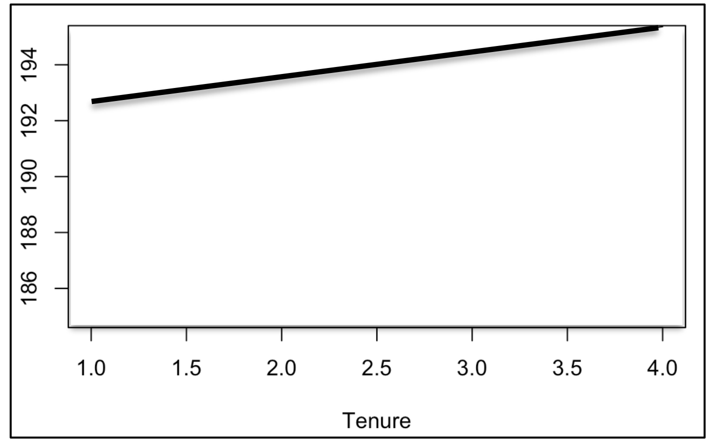
may be <u>conditionally</u> related to each other in a very different way.

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

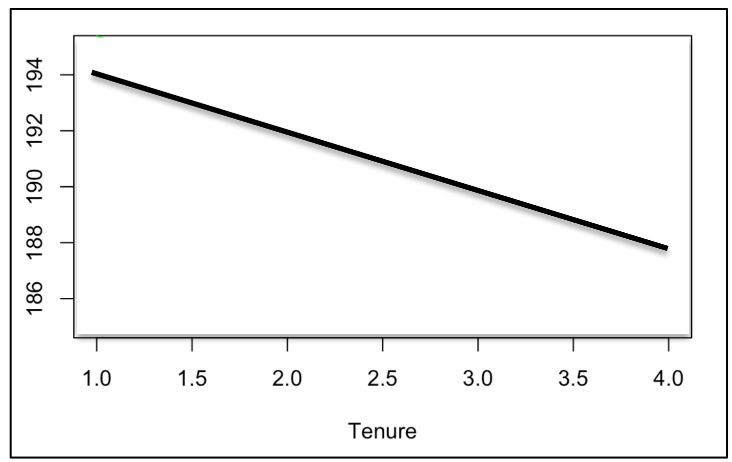
The one case where multiple regression *<u>is</u>* the union of the set of simple linear regressions is when there is no multicollinearity.

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

Thus there are eight measures of importance: <u>unconditional</u> p-values, bivariate coefficients of determination, and slope; <u>conditional</u> 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

Prelude to Today's Discussion

Up to this point in the course, we have been focusing on continuous independent variables.

This week we will introduce discrete independent variables (…the dependent variable will, as always in this course, be continuous…).

When <u>all</u> of the independent variables are discrete, the model is called an "Analysis of Variance" (ANOVA) model, not a regression model;

As we will see, in this context, we are interested in eight measures (similar to multiple regression), but the "slope" is replaced by the "difference in means".

In today's discussion, will be looking only at models where <u>all</u> of the independent variables are discrete; in our next class session, we will be introducing models which include a mixture of discrete and continuous variables ("ANACOVA Models").

Stevens' Levels of Measurement Typology

Nominal

Ordinal

"Discrete"

Interval

Ratio

"Continuous"

Three Primary Forms of the General Linear Model (GLM)

GLM Form	Dependent Variable	Independent Variables
Regression	Continuous	All continuous
Analysis of Variance	Continuous	All discrete
Analysis of Covariance	Continuous	Mixture

Regression sub-forms:

- **Simple linear regression: *single* independent variable which is continuous**
- **Multiple regression: *multiple* independent variables, all of which are continuous**

Analysis of variance sub-forms:

- **One-way ANOVA: *single* independent variable which is discrete**
- **n-way ANOVA: "*n*" independent variables, all of which are discrete**

Case 1 (Analysis of Variance): Worldwide Wholesale, Inc.

- WWI: a chain of membership-only retail warehouse clubs
- Objective: Predict "cost per trip" (CPT: continuous) from weekday (discrete: M T W R F Sa Su) and employment status (discrete: not employed, employed part-time, employed fulltime)
- Data collected from seven random samples of WWI cash register receipts records: one for each day of the week
- Collect: ID, CPT, day of week, employment status.
- Data are available on Blackboard (Outline/Session 4: Scenario4.dat)

```
       ID               Day       EmpStat      CPT
 Min.   : 104123   1-M:34    FT:62    Min.   : 31.00
 1st Qu.:2470562   2-T:33    NE:75    1st Qu.: 80.00
 Median :5353738   3-W:40    PT:88    Median :100.00
 Mean   :5159247   4-R:21             Mean   : 99.99
 3rd Qu.:7621615   5-F:34             3rd Qu.:123.00
 Max.   :9935214   6-S:34             Max.   :163.00
                   7-S:29
```

```
WWI.dat <- read.table("WWI.dat", header=TRUE,
  sep="", na.strings="NA", dec=".", strip.white=TRUE)
summary(WWI.dat)
```

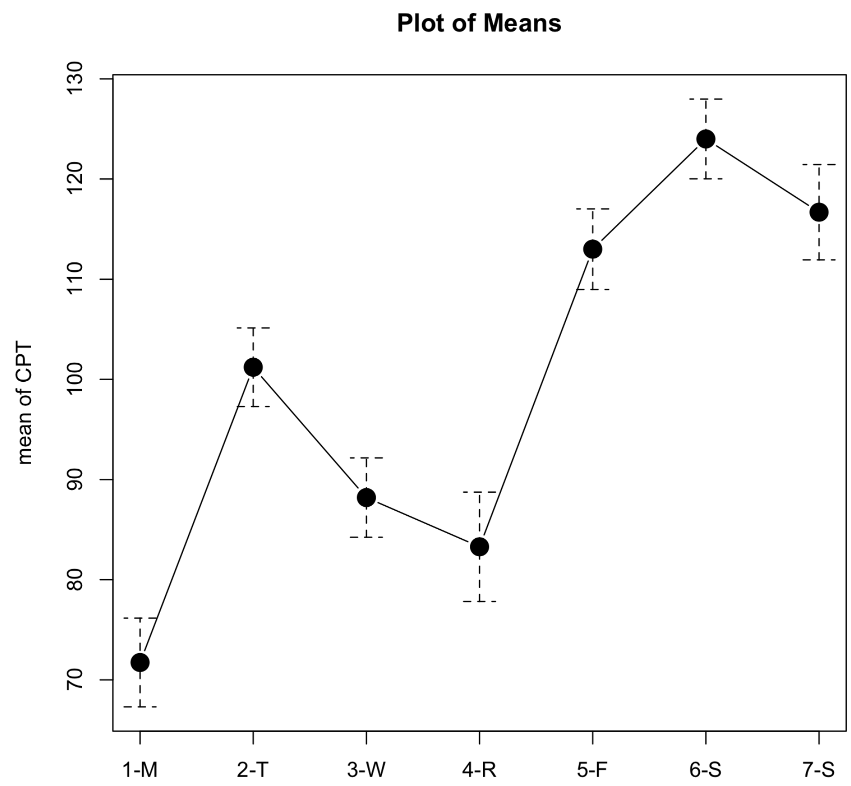
- Pass the "smell test?"

Category Profiles

```
$`1-M`
   vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 34 71.74 25.83     71   71.93 40.03  31 111    80 -0.02     -1.5 4.43

$`2-T`
   vars  n   mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 33 101.21 22.52    102  101.93 28.17  61 135    74 -0.17     -1.2 3.92

$`3-W`
   vars  n mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 40 88.2 25.07     86   88.66 27.43  43 127    84 -0.08    -1.18 3.96

$`4-R`
   vars  n  mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 21 83.29 25.04     88   84.24 29.65  40 117    77 -0.35    -1.39 5.46

$`5-F`
   vars  n mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 34  113 23.44  114.5  113.96 24.46  68 149    81 -0.27    -1.07 4.02

$`6-S`
   vars  n mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 34  124 23.23  132.5  123.89 31.13  86 163    77 -0.01    -1.36 3.98

$`7-S`
   vars  n   mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 29 116.69 25.62    116  116.64 28.17  71 160    89 0.13    -1.15 4.76
```

```r
library(psych)
describeBy(WWI.dat$CPT, WWI.dat$Day)
```

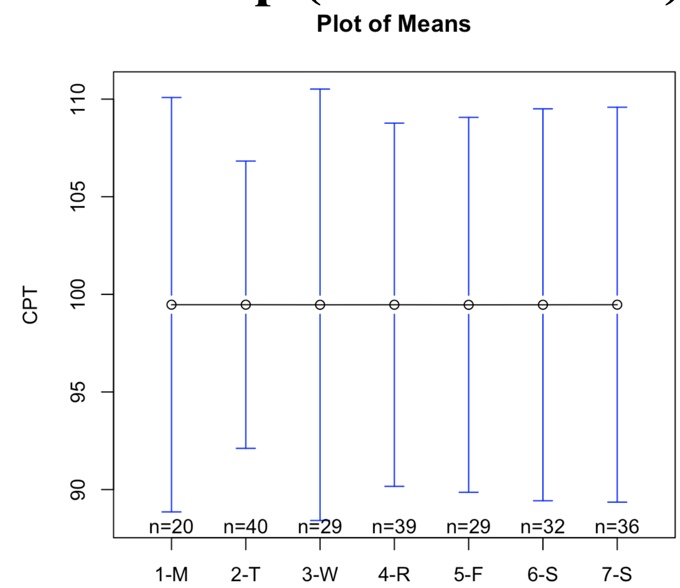
Plotting the Relationship Between a Continuous Variable And a Discrete Variable: "Mean Plots"




```
Boxplot(CPT~Day, data=WWI.dat, id.method="y")
with(WWI.dat, plotMeans(CPT, Day,
  error.bars="se", connect=TRUE))
```

Describing Relationships: No Relationship (Scenario1.dat)

```r
plotmeans(CPT~Day,data=WWI.dat,main="Plot of Means")
```




```r
ANOVA <- lm(CPT~Day, data=WWI.dat)
summary(ANOVA)
```

```
Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept)  9.947e+01  6.029e+00  16.497   <2e-16 ***
Day2-T      -1.349e-14  7.385e+00   0.000    1.000
Day3-W      -4.828e-03  7.837e+00  -0.001    1.000
Day4-R      -2.051e-03  7.416e+00   0.000    1.000
Day5-F      -6.897e-03  7.837e+00  -0.001    0.999
Day6-S      -3.750e-03  7.686e+00   0.000    1.000
Day7-S      -1.189e-14  7.520e+00   0.000    1.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.96 on 218 degrees of freedom
Multiple R-squared:  8.664e-09, Adjusted R-squared:  -0.02752
F-statistic: 3.148e-07 on 6 and 218 DF,  p-value: 1
```

H_0: All population means are identical

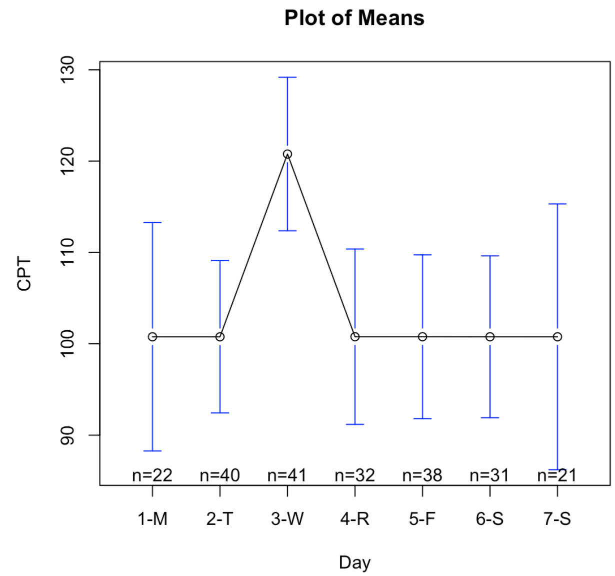
H_A: Not all population means are identical

```r
etasq(ANOVA,anova=TRUE,partial=FALSE)
```

```
Response: CPT
              eta^2 Sum Sq  Df F value Pr(>F)
Day      9.5315e-09      0   6       0      1
Residuals           152672 218
```

A Slightly Stronger Relationship (Scenario2.dat)

```r
plotmeans(CPT~Day,data=WWI.dat,main="Plot of Means")
```




```r
ANOVA <- lm(CPT~Day, data=WWI.dat)
summary(ANOVA)
```

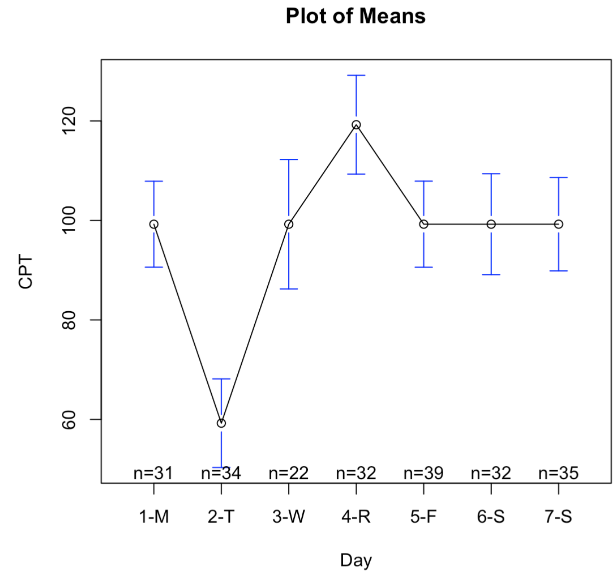
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.008e+02  5.758e+00  17.500  < 2e-16 ***
Day2-T      -1.818e-03  7.169e+00   0.000  0.99980    
Day3-W       2.001e+01  7.138e+00   2.803  0.00553 ** 
Day4-R       5.682e-03  7.480e+00   0.001  0.99939    
Day5-F       6.603e-03  7.236e+00   0.001  0.99927    
Day6-S       7.625e-04  7.529e+00   0.000  0.99992    
Day7-S      -8.658e-04  8.240e+00   0.000  0.99992    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.01 on 218 degrees of freedom
Multiple R-squared:  0.07779,	Adjusted R-squared:  0.05241 
F-statistic: 3.065 on 6 and 218 DF,  p-value: 0.006711
```

```r
etasq(ANOVA,anova=TRUE,partial=FALSE)
```

```
Response: CPT
             eta^2 Sum Sq  Df F value   Pr(>F)   
Day       0.077794  13416   6  3.0649 0.006711 **
Residuals          159036 218                    
```

A Yet Slightly Stronger Relationship (Scenario3.dat)

```
plotmeans(CPT~Day,data=WWI.dat,main="Plot of Means")
```




```
ANOVA <- lm(CPT~Day, data=WWI.dat)
summary(ANOVA)
```

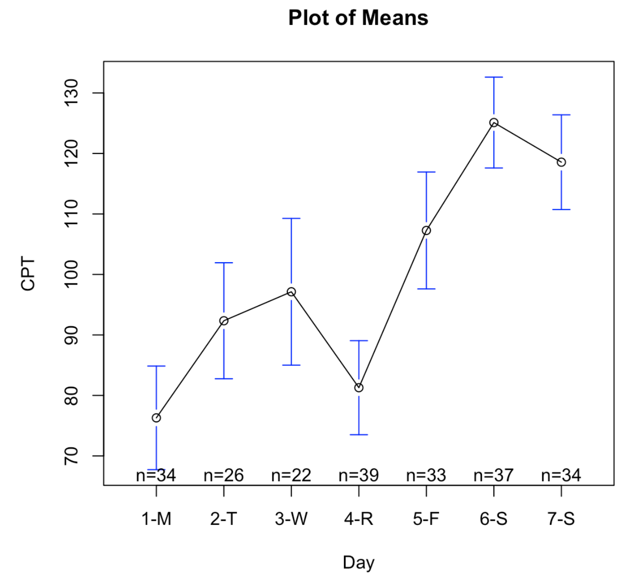
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.924e+01  4.815e+00  20.609  < 2e-16 ***
Day2-T      -4.000e+01  6.658e+00  -6.008 7.81e-09 ***
Day3-W      -1.613e-03  7.474e+00   0.000  0.99983
Day4-R       2.000e+01  6.757e+00   2.961  0.00341 **
Day5-F       3.772e-03  6.451e+00   0.001  0.99953
Day6-S      -3.629e-04  6.757e+00   0.000  0.99996
Day7-S       4.101e-03  6.613e+00   0.001  0.99951
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.81 on 218 degrees of freedom
Multiple R-squared:  0.2929,	Adjusted R-squared:  0.2734
F-statistic: 15.05 on 6 and 218 DF,  p-value: 2.155e-14
```

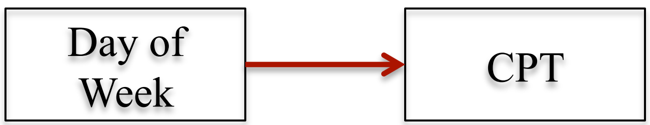
```
etasq(ANOVA,anova=TRUE,partial=FALSE)
```

```
Response: CPT
            eta^2 Sum Sq  Df F value    Pr(>F)
Day       0.29287  64904   6  15.048 2.155e-14 ***
Residuals         156709 218
```

A More Typical Scenario (Scenario4.dat)

```r
plotmeans(CPT~Day,data=WWI.dat,main="Plot of Means")
```






```r
ANOVA <- lm(CPT~Day, data=WWI.dat)
summary(ANOVA)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   76.294      4.191  18.206  < 2e-16 ***
Day2-T        16.052      6.366   2.522  0.01240 *
Day3-W        20.842      6.686   3.117  0.00207 **
Day4-R         4.988      5.733   0.870  0.38526
Day5-F        30.979      5.971   5.188 4.85e-07 ***
Day6-S        48.814      5.805   8.409 5.41e-15 ***
Day7-S        42.265      5.926   7.132 1.44e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.44 on 218 degrees of freedom
Multiple R-squared:  0.3538,	Adjusted R-squared:  0.336
F-statistic: 19.89 on 6 and 218 DF,  p-value: < 2.2e-16
```

Based on these results from this 1-way Analysis of Variance (ANOVA), what conclusions would you draw?

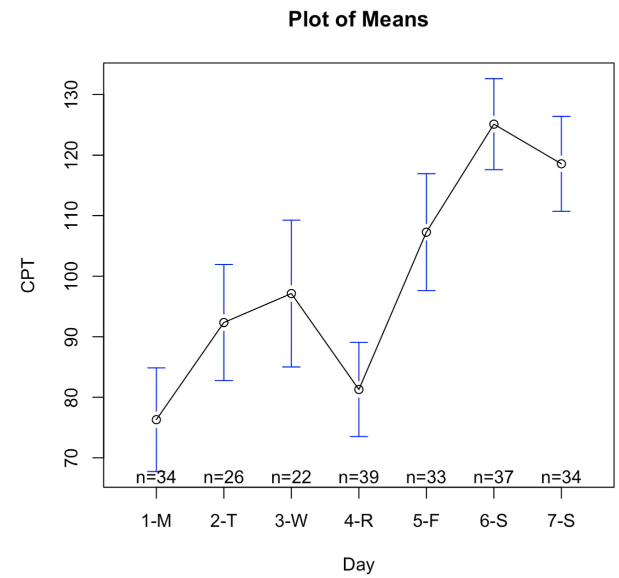
```r
etasq(ANOVA,anova=TRUE,partial=FALSE)
```

```
Response: CPT
            eta^2 Sum Sq  Df F value    Pr(>F)
Day       0.35377  71256   6   19.89 < 2.2e-16 ***
Residuals         130162 218
```

A-Posteriori Test: Scheffé Test

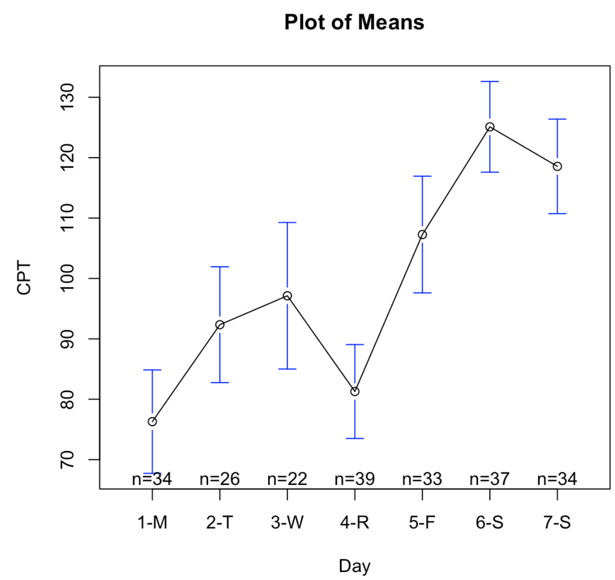
```r
library(agricolae)
scheffe.test(ANOVA,"Day", group=TRUE,console=TRUE,main="Scheffe Test")
```

```
Scheffe Test for CPT

Mean Square Error  : 597.0731

Day,  means

          CPT      std  r Min Max
1-M  76.29412 24.54128 34  39 118
2-T  92.34615 23.73679 26  63 138
3-W  97.13636 27.35184 22  55 132
4-R  81.28205 23.98239 39  45 127
5-F 107.27273 27.25375 33  62 150
6-S 125.10811 22.51763 37  87 167
7-S 118.55882 22.43670 34  73 161

alpha: 0.05 ; Df Error: 218
Critical Value of F: 2.140338

Harmonic Mean of Cell Sizes  31.01315
Minimum Significant Difference: 22.23681

Means with the same letter are not significantly different.

Groups, Treatments and means
a        6-S      125.1
ab       7-S      118.6
abc      5-F      107.3
bcd      3-W      97.14
cd       2-T      92.35
d        4-R      81.28
d        1-M      76.29
```



Scheffé grouping letters

Analysis of Variance: Assumptions

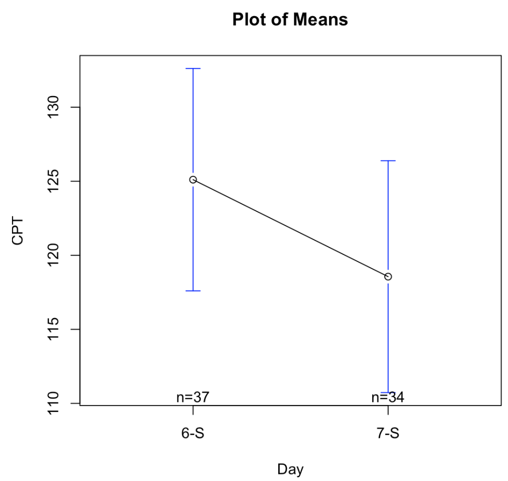


1. Normality
2. Homoskedasticity
3. Uncorrelated error terms

A-Priori Test (Two Groups): Independent Samples *t*-test

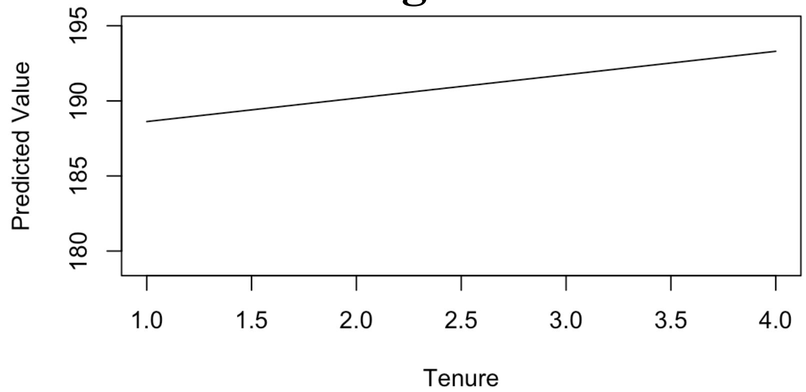
```
SatSun <- subset (WWI.dat, Day=="6-S"| Day=="7-S")
library(lsr)
independentSamplesTTest(CPT~Day, data=SatSun, var.equal=TRUE)
```

```
   Student's independent samples t-test

Outcome variable:   CPT
Grouping variable:  Day

Descriptive statistics:
                6-S     7-S
   mean     125.108 118.559
   std dev.  22.518  22.437

Hypotheses:
   null:        population means equal for both groups
   alternative: different population means in each group

Test results:
   t-statistic:  1.226
   degrees of freedom:  69
   p-value:  0.224

Other information:
   two-sided 95% confidence interval:  [-4.104, 17.203]
   estimated effect size (Cohen's d):  0.291
```

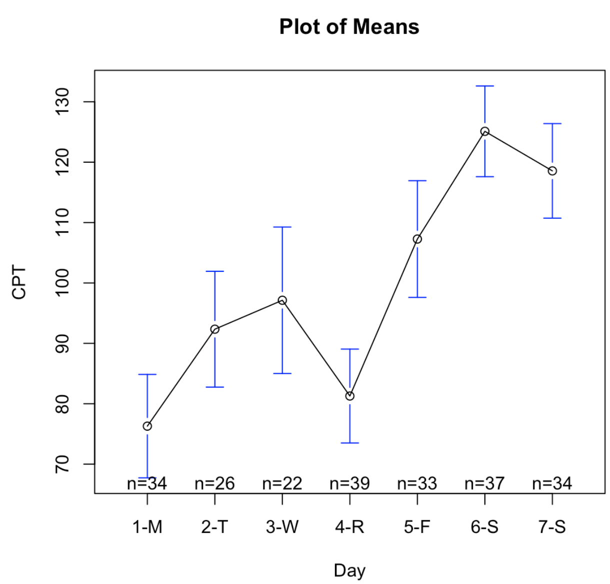


Regression vs. Analysis of Variance: The General Linear Model

Regression



ANOVA



Key Points in Today's Discussion

- **In much the same way that a <u>continuous</u> independent variable can be related to a dependent variable, a <u>discrete</u> independent variable can be related to the dependent variable. We call a model in which all of the independent variables are discrete an "Analysis of Variance" (ANOVA) model;**

- **In this case, the focus is on whether the population means differ rather than whether the slope of the best-fitting straight line has a non-zero slope in the population;**

- **In the same way that we can distinguish between statistical significance and strength of relationship in regression models, we can make the same distinction in ANOVA models;**

- **In an ANOVA model, a "reference group" is arbitrarily selected, and the focus is on the difference between this reference group and each of the other groups;**

- **The primary statistical test in ANOVA tests the null hypothesis that the population means of all groups are identical;**

- **If the null hypothesis is rejected, a followup ("a-posteriori") test, such as the Scheffé Test, can be conducted to identify which pairs of populations differ in their means;**

- **When we have just two populations, a "single sample t-test" is often employed to test the null hypothesis that the two population means are identical,**

Parting Thoughts …

SEE YOU FRIDAY

<u>HAVE A GREAT THURSDAY</u>!