Session 5

Administrivia

Logistical issues:

- ✧ Optional VOH Sessions 2 (Monday) and 3 (Tuesday) **will** be held;
- ✧ There **will** be an optional Quiz, as usual, next Tuesday at 8:30am Eastern Time;
- ✧ No <u>class</u> session next week (Thanksgiving week);
- ✧ Assignment 4 will be available at 7:00pm tonight, and will be due at 4:25pm on Wednesday 12/2;
- ✧ Assignment 5 (final assignment) will be assigned on Wednesday 12/2 and due on 12/9 (our last class session);

Key Points in Our Last Discussion

In much the same way that a continuous independent variable can be related to a dependent variable, a discrete 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.

Prelude to Today's Discussion

In our last session, we focused on a single discrete independent variable.

In the same way that multiple regression can qualify the results of simple linear regression, an n-way ANOVA can qualify the results of one-way ANOVA;

As in regression, the coefficient of *partial* determination in an ANOVA model provides a measure of the extent to which a given independent variable is *uniquely* related to the dependent variable;

Reminder: in all forms of the General Linear Model discussed in this course, the dependent variable is continuous;

In the last half of today's discussion, we will be introducing models which include a mixture of discrete and continuous variables ("ANACOVA Models").

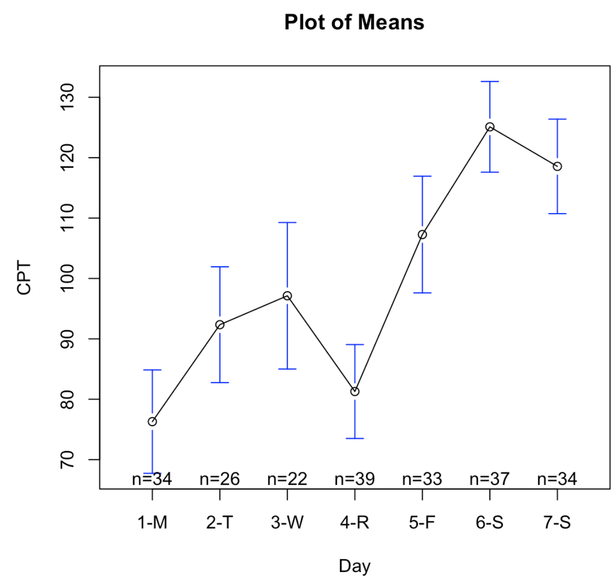
Reminder: Worldwide Wholesale, Inc.

- ✧ **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)
```

Analysis of Variance: Assumptions



1. Normality

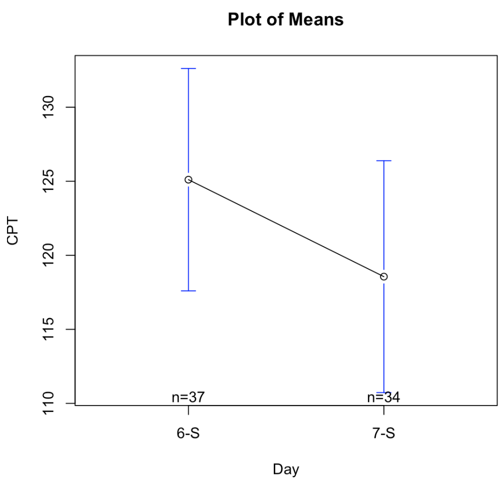
2. Homoskedasticity

3. Uncorrelated error terms

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

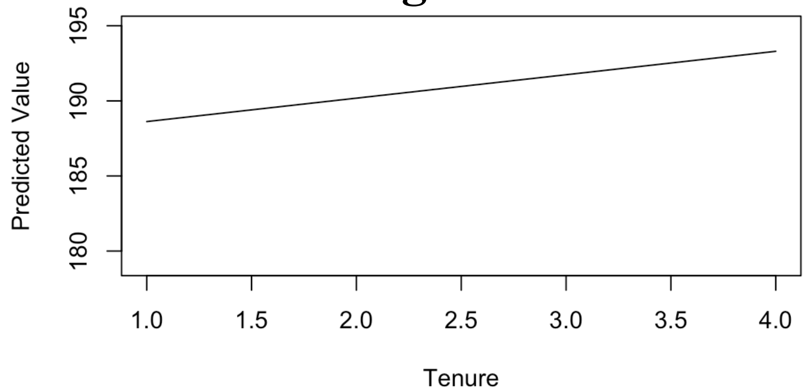
```
SatSun <- subset (WWI.dat, Day=="6-S"| Day=="7-S")
library(lsr)
independentSamplesTTest(CPT~Day, data=SatSun, var.equal=TRUE)
```

```
   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
```

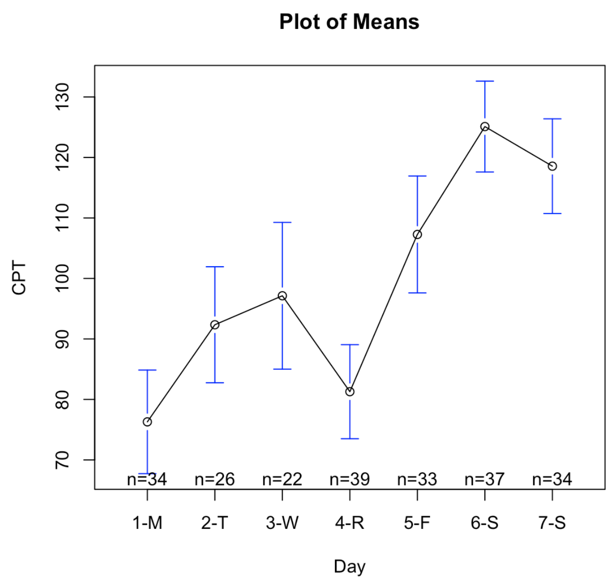


Regression vs. Analysis of Variance: The General Linear Model

Regression

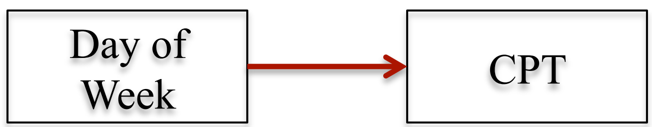


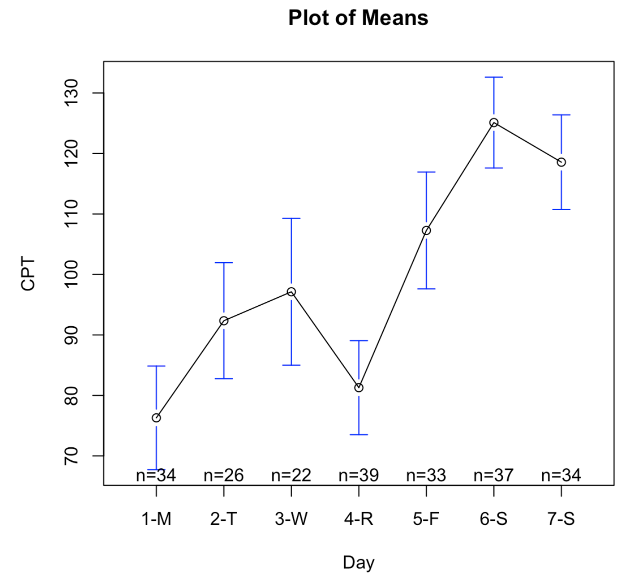
ANOVA



Review of One-Way ANOVA Models (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
```

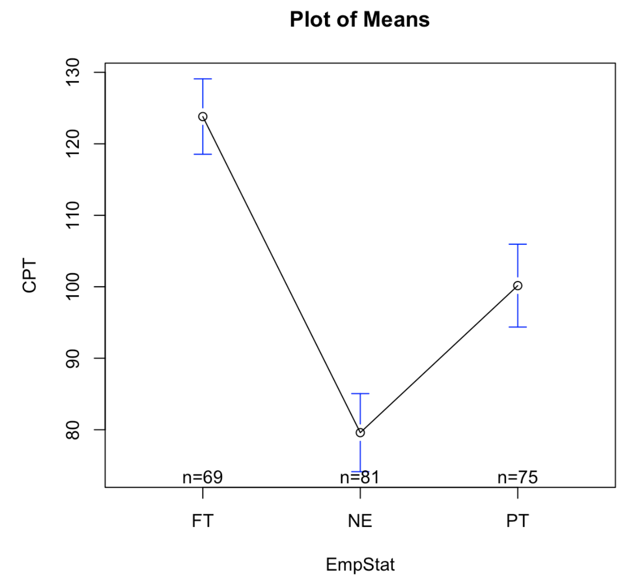
Based on these results from this 1-way Analysis of Variance (ANOVA), what conclusions would you draw?

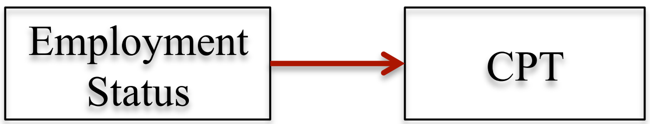
```r
etasq(ANOVA,anova=TRUE,partial=FALSE)
```

```
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 Second Discrete Independent Variable: Employment Status

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






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

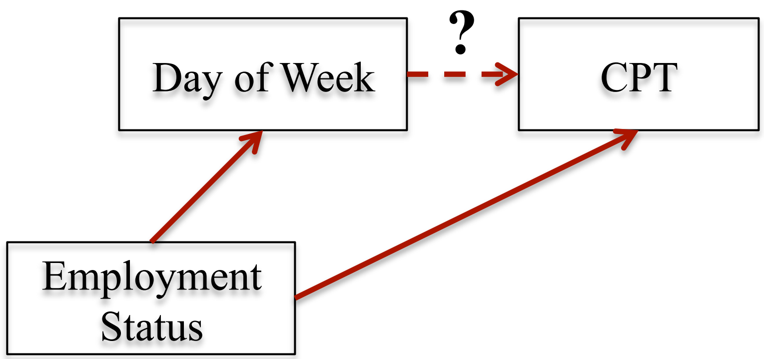
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  123.812      2.897  42.737  < 2e-16 ***
EmpStatNE    -44.219      3.942 -11.216  < 2e-16 ***
EmpStatPT    -23.652      4.014  -5.892  1.4e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.06 on 222 degrees of freedom
Multiple R-squared:  0.3617,    Adjusted R-squared:  0.356
F-statistic: 62.91 on 2 and 222 DF,  p-value: < 2.2e-16
```

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

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

```
Response: CPT
           eta^2  Sum Sq  Df F value    Pr(>F)
EmpStat  0.36172   72858   2  62.906 < 2.2e-16 ***
Residuals         128560 222
```

Does Employment Status *Explain* the Day-of-Week → CPT Relationship? (Part 1)




```
library(stats)
xtabs(~EmpStat+Day, data=WWI.dat)
```

```
       Day
EmpStat 1-M 2-T 3-W 4-R 5-F 6-S 7-S
     FT   0   0   0   0   3  34  25
     NE  34   0  21  20   0   0   0
     PT   0  33  19   1  31   0   4
```

```
chisq.test(xtabs(~EmpStat+Day,
  data=WWI.dat), correct=FALSE)
```

```
	Pearson's Chi-squared test

data:  xtabs(~EmpStat + Day, data = WWI.dat)
X-squared = 351.03, df = 12, p-value < 2.2e-16
```

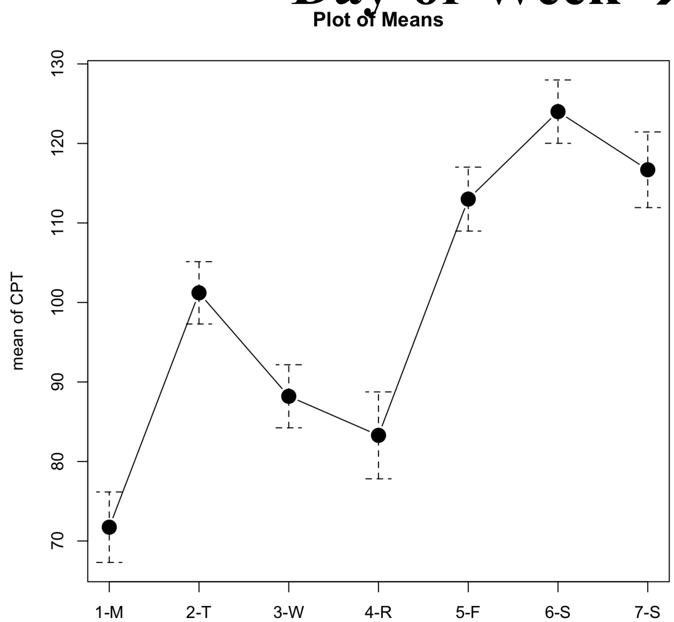
```
ANOVA <- lm(CPT~Day + EmpStat, data=WWI.dat)
summary(ANOVA)
```

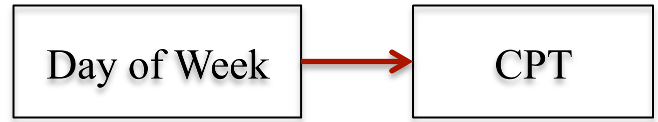
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  110.801     12.709   8.718 7.58e-16 ***
Day2-T        -1.420     11.458  -0.124   0.9015
Day3-W        10.518      8.687   1.211   0.2273
Day4-R         4.540      5.651   0.803   0.4226
Day5-F        10.409     11.327   0.919   0.3591
Day6-S        14.307     13.310   1.075   0.2836
Day7-S        11.766     12.520   0.940   0.3484
EmpStatNE    -34.507     12.021  -2.871   0.0045 **
EmpStatPT    -17.034      7.246  -2.351   0.0196 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.06 on 216 degrees of freedom
Multiple R-squared:  0.3792     Adjusted R-squared:  0.3562
F-statistic: 16.49 on 8 and 216 DF,  p-value: < 2.2e-16
```

```
etasq(ANOVA, data=WWI.dat, anova=TRUE,
  partial=FALSE)
```

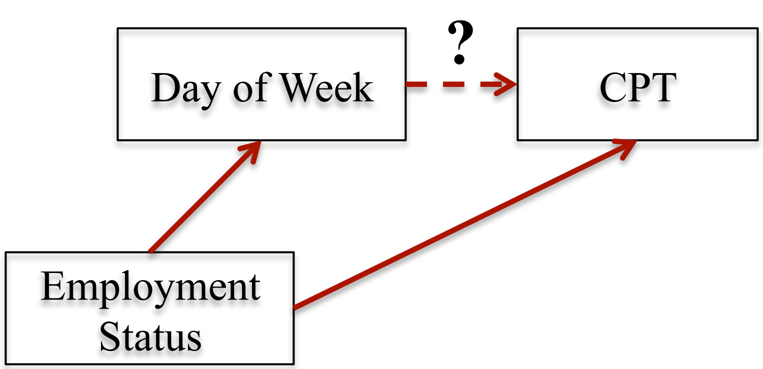
```
Response: CPT
             eta^2 Sum Sq  Df F value   Pr(>F)
Day       0.049433   6741   6  1.9267 0.07772 .
EmpStat   0.026943   3674   2  3.1504 0.04481 *
Residuals          125955 216
```

Does Employment Status *Explain* the Day-of-Week → CPT Relationship? (Part 2)





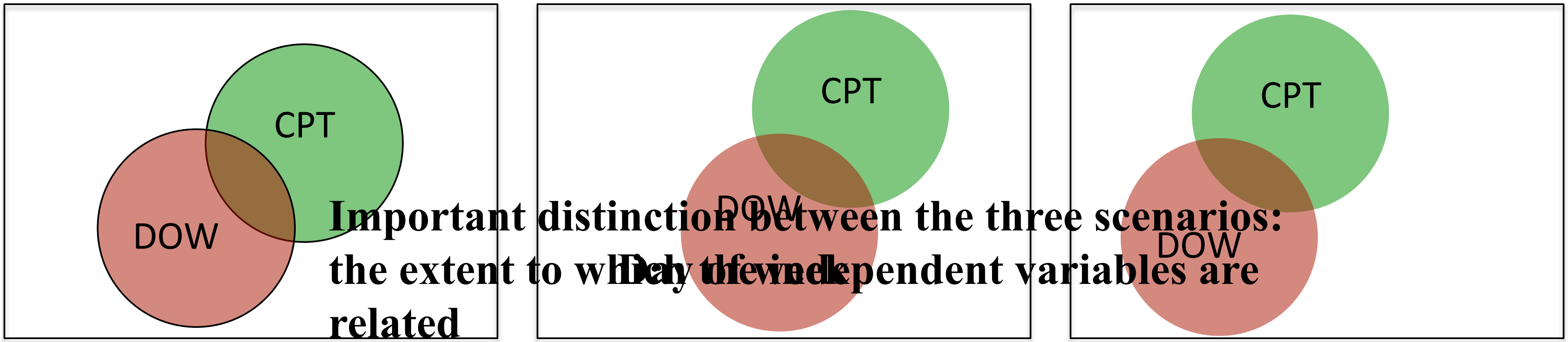

```
Response: CPT
            eta^2 Sum Sq  Df F value    Pr(>F)
Day       0.35727  72055   6  20.196 < 2.2e-16 ***
Residuals         129629 218
```




```
Response: CPT
              eta^2 Sum Sq  Df F value  Pr(>F)
Day        0.049433   6741   6  1.9267 0.07772 .
EmpStat    0.026943   3674   2  3.1504 0.04481 *
Residuals           125955 216
```

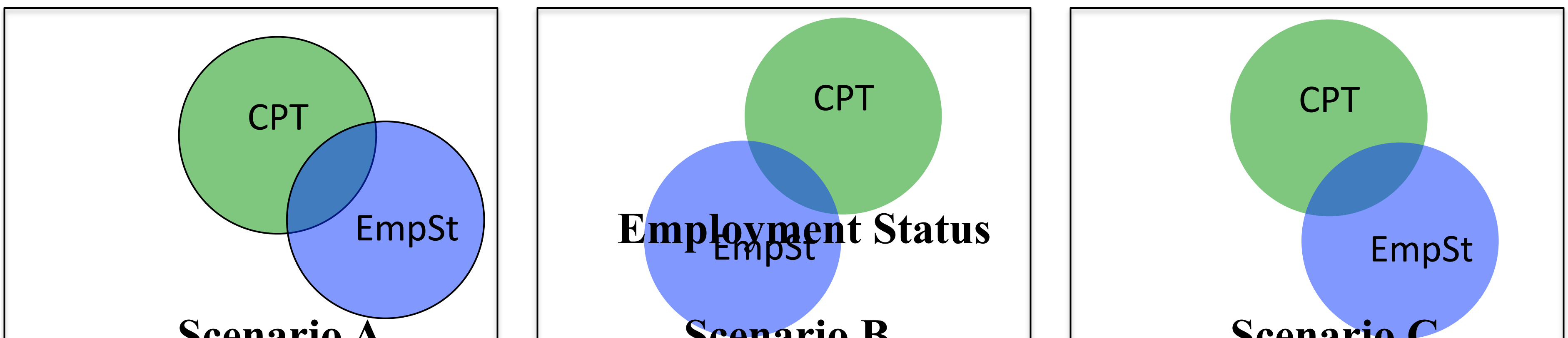
1-Way ANOVA vs. 2-Way ANOVA

Scenario A **Scenario B** **Scenario C**



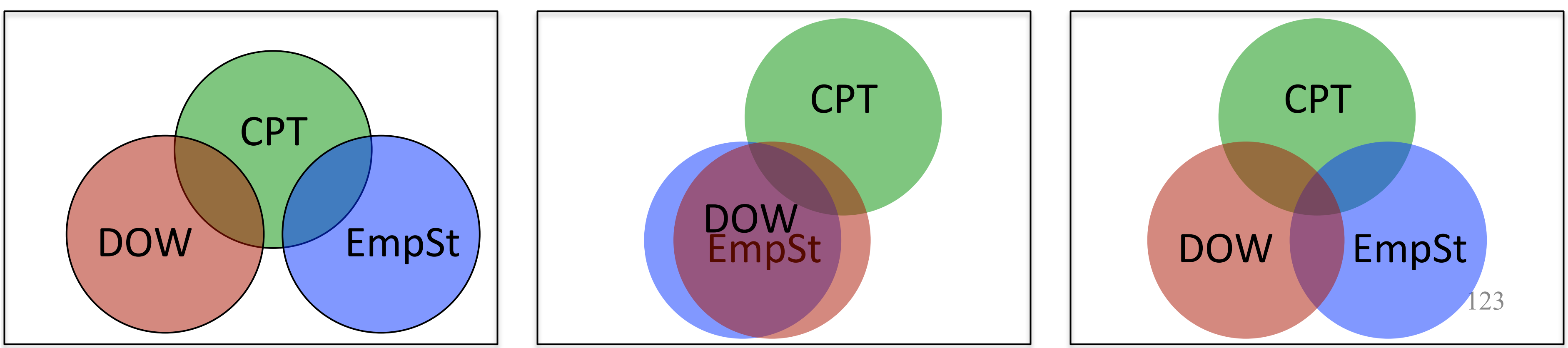
Important distinction between the three scenarios: the extent to which the independent variables are related

Day of Week

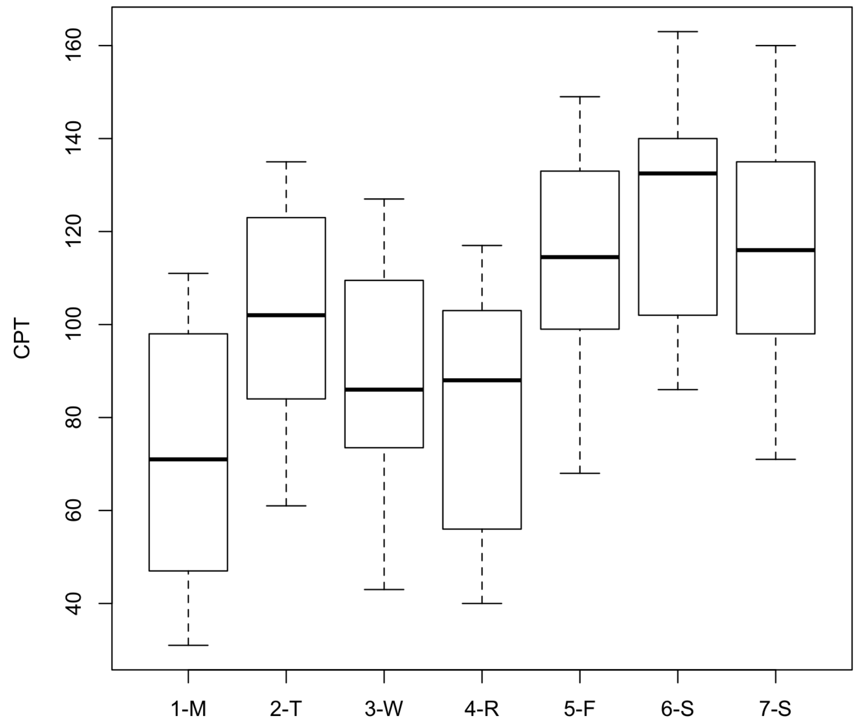


Employment Status

Scenario A **Scenario B** **Scenario C**



Useful Descriptive Tool: Box and Whisker Plot




```
library(car)
Boxplot(CPT~Day, data=WWI.dat)
```

Stevens' Levels of Measurement Typology (revisited)

$$\left.\begin{array}{l}\textbf{Nominal}\\[2ex]\textbf{Ordinal}\end{array}\right\}\ \textbf{“Discrete”}$$

$$\left.\begin{array}{l}\textbf{Interval}\\[2ex]\textbf{Ratio}\end{array}\right\}\ \textbf{“Continuous”}$$

Three Primary Forms of the General Linear Model (revisited)

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**
- **Multiple regression: *multiple* independent variables**

Analysis of variance sub-forms:

- **One-way ANOVA: single independent variable**
- ***n*-way ANOVA: "*n*" independent variables**

Case 2 (Analysis of Covariance): Kitridge Hosts, Inc.

- **KHI: an American multinational diversified hospitality company that manages and franchises a broad portfolio of hotels and related lodging facilities.**
- **Objective: Predict guest satisfaction (<u>continuous</u>) from property type (<u>discrete</u>: 1-Classic, 2-Premium, 3-Luxury) and guest age (<u>continuous</u>).**
- **Data collected from <u>*three*</u> random samples of KHI populations: guests staying at Classic, Premium, and Luxury hotels.**
- **Collect: satisfaction (scale: 0 to 100), property type, age of registered guest.**
- **Data are available on Blackboard (Outline/Session 5: KHI.dat).**

```
       ID             PropType          Age             SAT
 Min.   : 104123   1-Class:70   Min.   :20.00   Min.   :38.00
 1st Qu.:2470562   2-Premi:72   1st Qu.:32.00   1st Qu.:46.00
 Median :5353738   3-Luxur:83   Median :39.00   Median :50.00
 Mean   :5159247                Mean   :39.04   Mean   :50.02
 3rd Qu.:7621615                3rd Qu.:46.00   3rd Qu.:54.00
 Max.   :9935214                Max.   :59.00   Max.   :62.00
```

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

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

Focusing on Property Type: Category Profiles

Property Type (discrete) → Satisfaction (continuous)

Plausible?

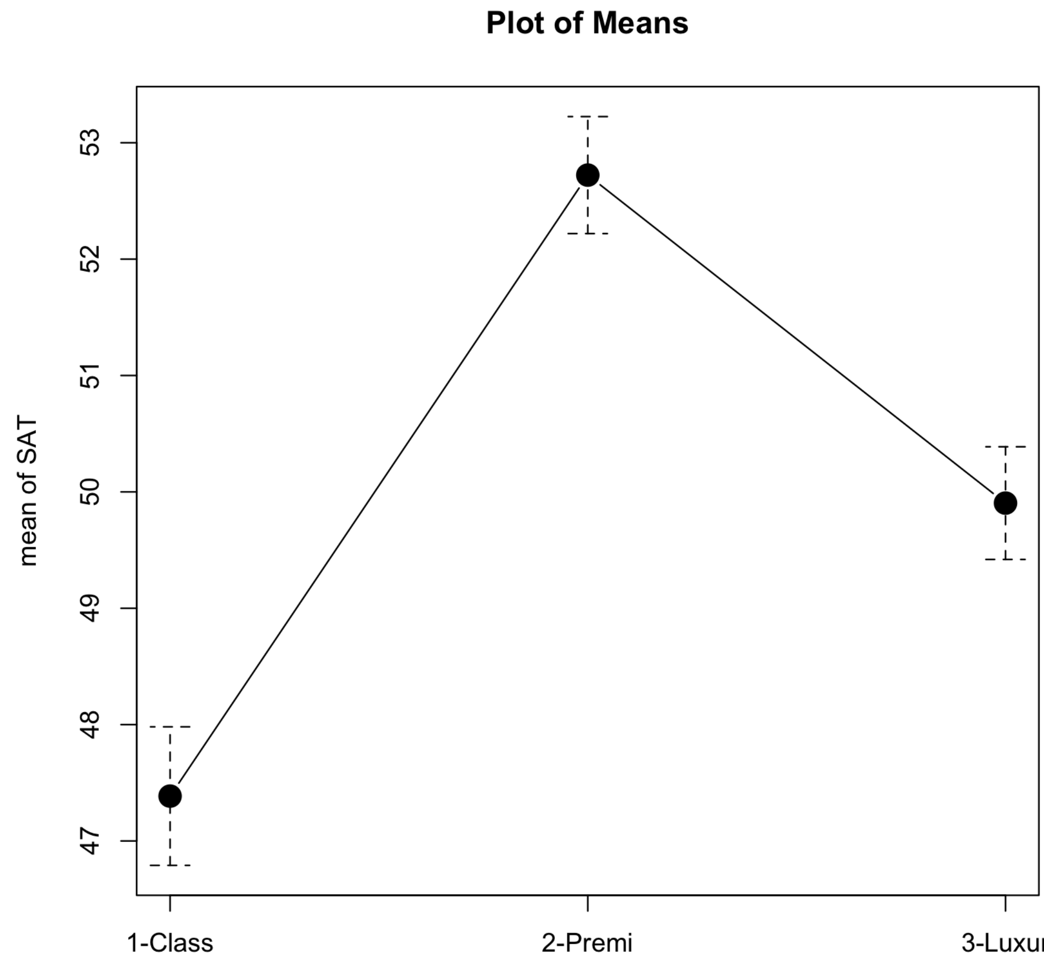
```
$`1-Class`
   vars  n  mean   sd median trimmed  mad min max range  skew kurtosis  se
X1    1 70 47.39 4.98   47.5   47.46 6.67  38  56    18 -0.13     -1.1 0.6

$`2-Premi`
   vars  n  mean   sd median trimmed  mad min max range skew kurtosis  se
X1    1 72 52.72 4.27     53   52.71 4.45  44  62    18 0.05    -0.74 0.5

$`3-Luxur`
   vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 83 49.9 4.41     50   49.91 5.93  42  58    16 0.01    -1.28 0.48
```

```
describeBy(KHI.dat$SAT, KHI.dat$PropType)
```

Conclusions Based on One-Way ANOVA Results?

Property Type (discrete) → Satisfaction (continuous)

```
Boxplot(SAT~PropType, data=KHI.dat, id.method="y")
with(KHI.dat, plotMeans(SAT, PropType,
  error.bars="se", connect=TRUE))
```



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

```
etasq(lm(SAT~PropType,data=KHI.dat),
      anova=TRUE, partial=FALSE)
```

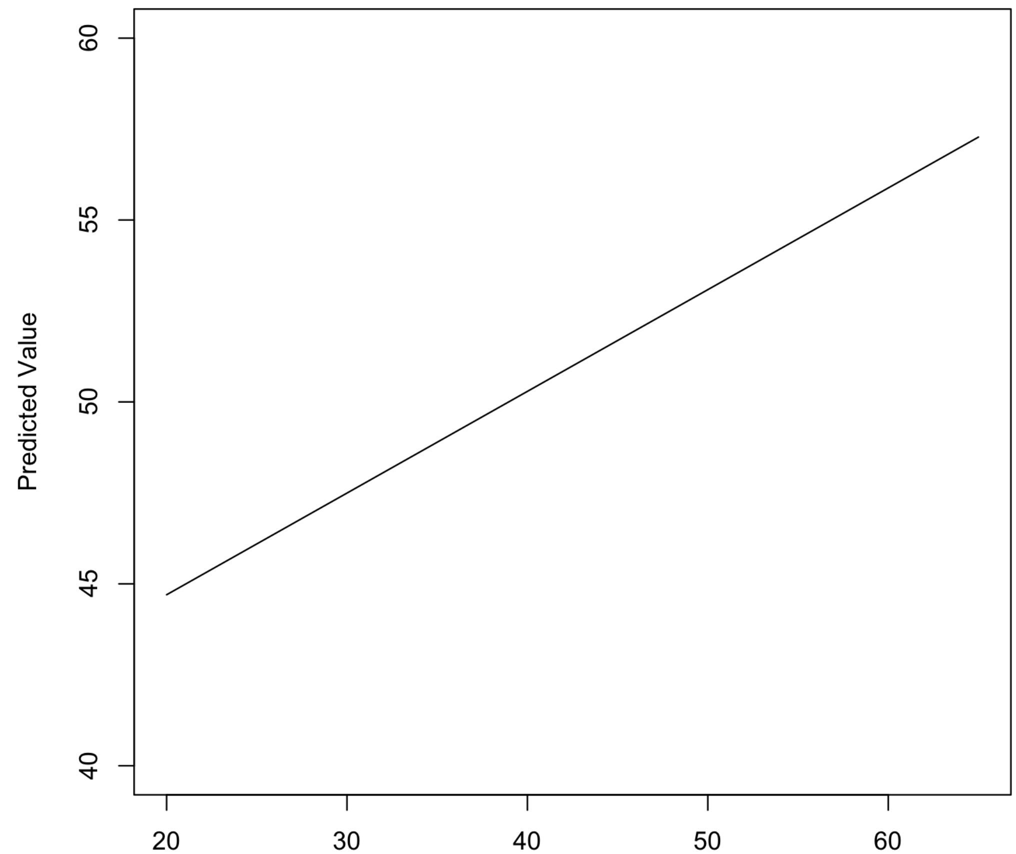
```
Response: SAT
              eta^2 Sum Sq  Df F value    Pr(>F)
PropType    0.18048 1012.6   2  24.444 2.543e-10 ***
Residuals           4598.3 222
```

Conclusions Based on Simple Linear Regression Results?

Age of Guest (continuous) → Satisfaction (continuous)

```
SLR<-lm(SAT~Age,data=KHI.dat)
library(TeachingDemos)
Predict.Plot(SLR, pred.var="Age",Age=c(20,65),
  plot.args=list(ylim=c(40, 60),col='black'),
  type="response")
```

Based on these results from this simple linear regression (SLR), what conclusions would you draw?




```
etasq(lm(SAT~Age,data=KHI.dat),
      anova=TRUE, partial=FALSE)
```

```
Response: SAT
             eta^2 Sum Sq  Df F value    Pr(>F)
Age        0.26619 1493.5   1  80.892 < 2.2e-16 ***
Residuals          4117.3 223
```

Are “Property Type” and “Age of Guest” Related?

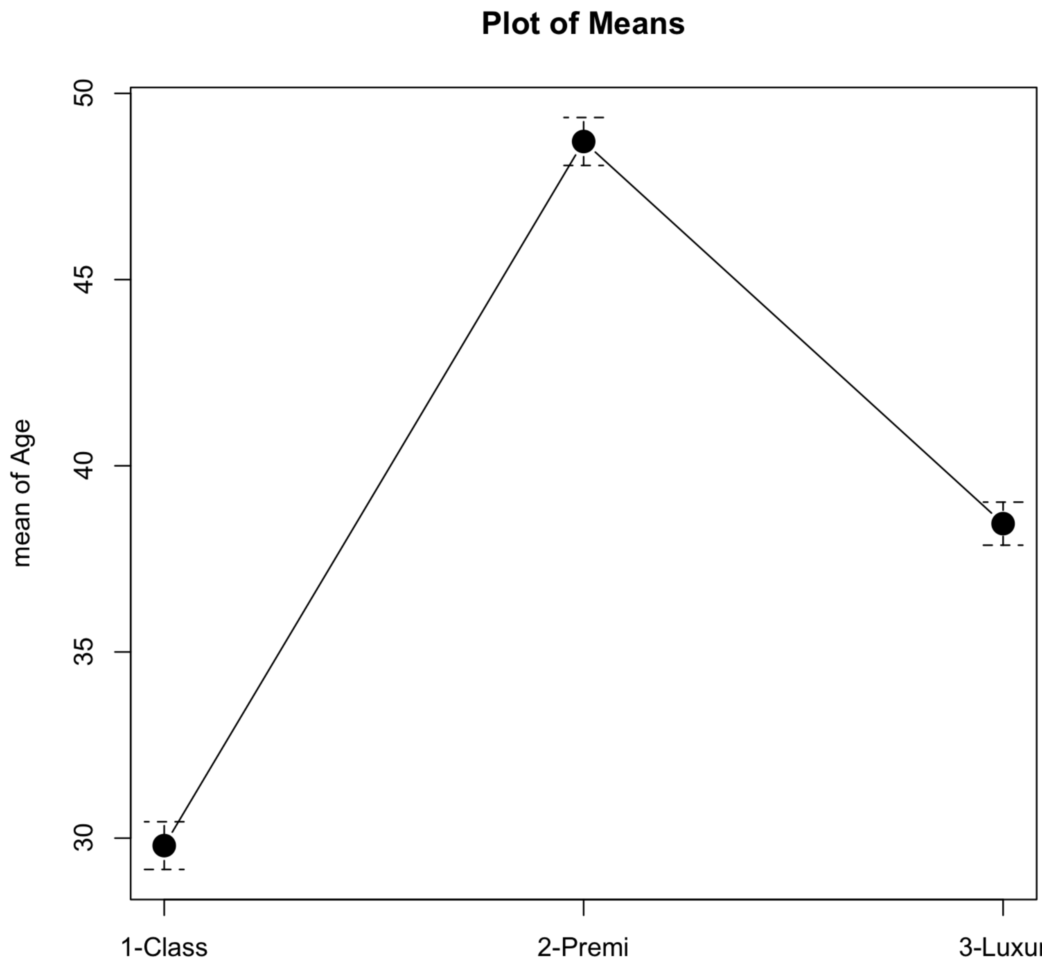
Property Type
(discrete)

↕

Age of Guest
(continuous)

```
Boxplot(Age~PropType, data=KHI.dat, id.method="y")
with(KHI.dat, plotMeans(Age, PropType,
  error.bars="se", connect=TRUE))
```

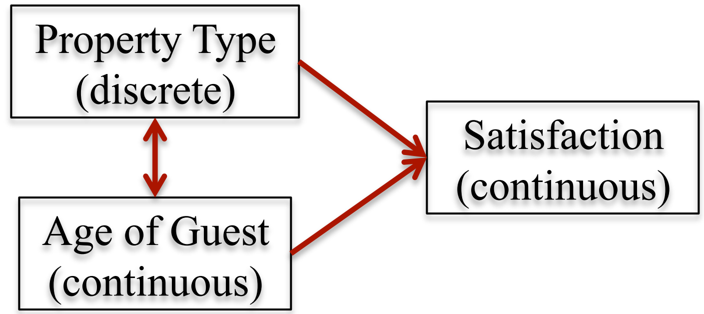
Based on these results from this one-way ANOVA, what conclusions would you draw?




```
etasq(lm(Age~PropType,data=KHI.dat),
      anova=TRUE, partial=FALSE)
```

```
Response: Age
            eta^2  Sum Sq  Df F value    Pr(>F)
PropType  0.66637 12736.1   2   221.7 < 2.2e-16 ***
Residuals          6376.6 222
```

Analysis of Covariance (ANCOVA) Results, Least Squares Means




```
etasq(lm(SAT~PropType + Age,data=KHI.dat),
      anova=TRUE, partial=FALSE)
```

```
Response: SAT
             eta^2 Sum Sq  Df F value    Pr(>F)
PropType  0.000116    0.5   2  0.0143    0.9858
Age       0.104689  481.4   1 25.8450 7.878e-07 ***
Residuals          4116.8 221
```

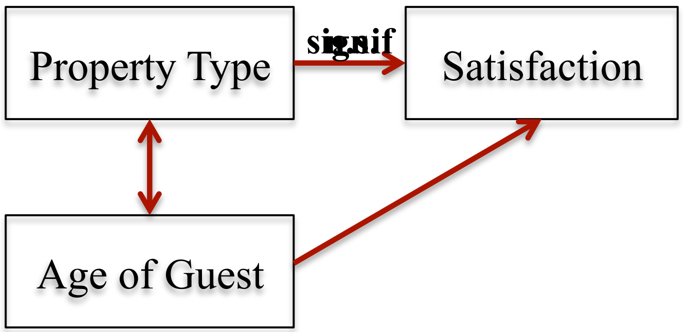
```
ANACOVA <- lm(SAT~PropType + Age,
  data=KHI.dat)
library(lsmeans)
lsmeans(ANACOVA,"PropType")
```

PropType	lsmean	SE	df	lower.CL	upper.CL
1-Class	49.92465	0.7180059	221	48.50963	51.33966
2-Premi	50.06559	0.7292465	221	48.62843	51.50276
3-Luxur	50.06689	0.4748333	221	49.13111	51.00267

Based on these results from this Analysis of Covariance (ANCOVA), what conclusions would you draw?

How do these conclusions reconcile with the one-way ANOVA and SLR conclusions you drew?

Confounding and Decision Making: What Action Should We Take?



This provides a classic example of *<u>confounding</u>* in business analytics: what initially appears to be a causal relationship (due to which intervention might be taken) has a possible alternative causal explanation.

This leads us to a very different intervention: market to older guests, rather than attempt to improve satisfaction at classic and luxury brands.

In our one-way ANOVA, we found that guests were more satisfied (on average) at premium (i.e., mid-scale) brands than at either classic (lower-scale) or luxury (upper-scale) brands.

This suggests as a possible intervention that KHI should work to improve whatever it is at classic and luxury brands that guests find unsatisfactory.

We also found that classic brands drew mostly younger guests, premium brands drew mostly older guests, and luxury brands drew mostly middle-age guests. This raised the question of whether it was BRAND or AGE that was causing the differences in satisfaction levels.

To address this question we controlled for age, and found (holding age constant) very little difference in average satisfaction level across brands.

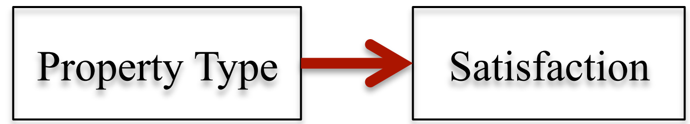
This led us to conclude that age may the dominating causal factor, and that attempting to improve the satisfaction level of certain brands may be futile: older guests may simply be easier to satisfy.

Assessing the Magnitude of Confounding

```
       ID               PropType       Age              SAT       
 Min.   : 104123   1-Class:70   Min.   :20.00   Min.   :38.00  
 1st Qu.:2470562   2-Premi:72   1st Qu.:32.00   1st Qu.:46.00  
 Median :5353738   3-Luxur:83   Median :39.00   Median :50.00  
 Mean   :5159247                Mean   :39.04   Mean   :50.02  
 3rd Qu.:7621615                3rd Qu.:46.00   3rd Qu.:54.00  
 Max.   :9935214                Max.   :59.00   Max.   :62.00  
```

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

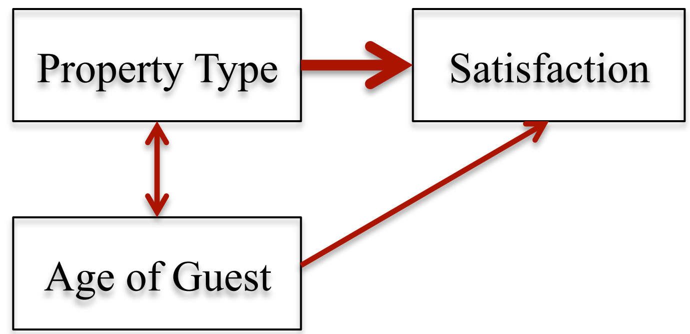
Unconditional effect




```
etasq(lm(SAT~PropType,data=KHI.dat),
     anova=TRUE, partial=FALSE)
```

```
Response: SAT
           eta^2 Sum Sq  Df F value    Pr(>F)    
PropType 0.18048 1012.6   2  24.444 2.543e-10 ***
Residuals         4598.3 222
```

Conditional effect




```
Response: SAT
             eta^2 Sum Sq  Df F value    Pr(>F)    
PropType  0.000116    0.5   2  0.0143    0.9858    
Age       0.104689  481.4   1 25.8450 7.878e-07 ***
Residuals          4116.8 221
```

```
etasq(lm(SAT~PropType + Age,data=KHI.dat),
     anova=TRUE, partial=FALSE)
```

- **By definition, "confounding" occurs only when the unconditional effect is larger than the conditional effect;**

- **The magnitude of the confounding effect is reflected by the difference in the unconditional and conditional coefficients of partial determination.**

Key Points in Today's Discussion

In much the same way that multiple regression results qualify the results of a simple linear regression, n-way ANOVA results can qualify the results of a one-way ANOVA;

When we have just two populations, an "independent samples t-test" is often employed to test the null hypothesis that the two population means are identical.

When we have more than one discrete independent variable, we can still meaningfully talk about the coefficients of partial determination, the Global F, the model coefficient of determination, and the adjusted model coefficient of determination;

Models in which we have at least one continuous independent variable and at least one discrete independent variable are called "Analysis of Covariance" (ANCOVA) models. Here, too, we can still meaningfully talk about the coefficients of partial determination, the Global F, the model coefficient of determination, and the adjusted model coefficient of determination;

Multivariable models allow us to begin to address issues of causality. Although being able to predict the value of the dependent variable is often thought of as the primary rationale for the various general linear model forms (analysis of variance, regression, analysis of covariance), the issue of cause is also frequently of interest;

A confounding model is one of the basic forms of causal models. Confounding models allow for better decisions to be made for resource allocation.

Administrivia

- **No class session next week;**

- **There will be an optional Quiz at the usual time next week;**

- **Assignment 4 will be available tonight at 7:00, and will be due at 4:25pm on 12/2 (via Blackboard).**

***HAVE A GREAT THANKSGIVING BREAK*!**