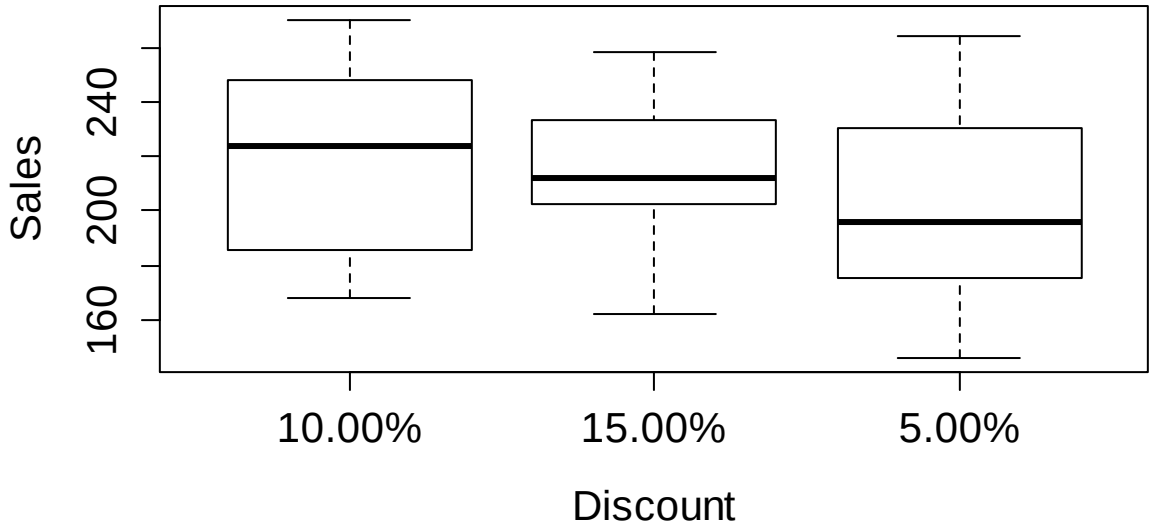
Example of More than 2 Categories, and Analysis of Covariance Example

```
> attach(Grocery)
> boxplot(Sales~Discount, ylab="Sales",xlab="Discount")
```




```
>
tapply(Sales,Discount,mean)
  10.00%   15.00%    5.00%
217.7500 213.5833 203.5000

> tapply(Sales,Discount,sd)
  10.00%   15.00%    5.00%
35.00162 26.66785 37.86939
```

Question: Is there a statistically significant difference in *population* mean sales for the different discount levels?

Two versions in R: The aov command, and the lm command as covered in Friday discussion.

See next page for output.

Using the aov command, followed by “summary”:

```
> AOVModel<-aov(Sales~Discount)
> summary(AOVModel)
             Df Sum Sq Mean Sq F value Pr(>F)
Discount      2   1288   644.2   0.573  0.569
Residuals    33  37074  1123.5
```

Using the lm command, followed by “anova”:

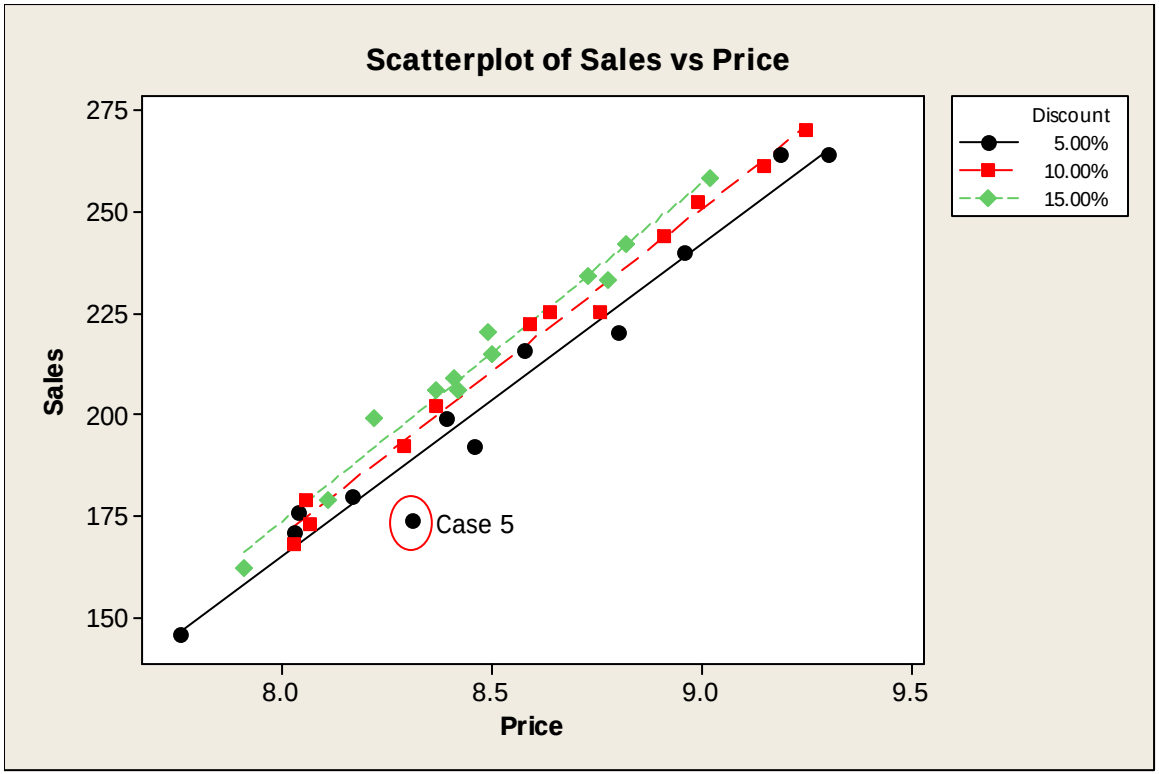
```
> LMVersion<-lm(Sales~as.factor(Discount))
> anova(LMVersion)
Analysis of Variance Table

Response: Sales
                     Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Discount)   2   1288  644.19  0.5734 0.5691
Residuals            33  37074 1123.46
```

Using the lm command, followed by “summary”:

```
> summary(LMVersion)
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 217.750      9.676  22.505   <2e-16 ***
as.factor(Discount)15.00%   -4.167     13.684  -0.304    0.763
as.factor(Discount)5.00%   -14.250     13.684  -1.041    0.305
Residual standard error: 33.52 on 33 degrees of freedom
Multiple R-squared:  0.03358, Adjusted R-squared:  -0.02499
F-statistic: 0.5734 on 2 and 33 DF,  p-value: 0.5691
```

Using any of the versions, do not reject H_0; conclude discount levels don’t have significant effect on sales.

NOW add a “covariate” of X = Price of the item. Explanatory notes on white board.




```
> AOC<-lm(Sales~Price+as.factor(Discount))
> anova(AOC)
Analysis of Variance Table
Response: Sales
                    Df Sum Sq Mean Sq  F value    Pr(>F)
Price                1  36718   36718 1391.366 < 2.2e-16 ***
as.factor(Discount)  2    800     400   15.149 2.348e-05 ***
Residuals           32    844      26
```

NOW we *can* reject H_0 and conclude Discount *does* have an effect on Sales, after accounting for Price.

Sample version of model for each group and tests with conclusions on board. (See next page for adjusted R^2, which is now almost 0.98.)

NOTE: Order matters for anova command but not for summary command:

```
> AOC<-lm(Sales~Price+as.factor(Discount))#Tests Price, then Discount
> anova(AOC)
Analysis of Variance Table
Response: Sales
                    Df Sum Sq Mean Sq  F value    Pr(>F)
Price                1  36718   36718 1391.366 < 2.2e-16 ***
as.factor(Discount)  2    800     400   15.149 2.348e-05 ***
Residuals           32    844      26

> AOCOrder<-lm(Sales~as.factor(Discount)+Price)#Discount, then Price
> anova(AOCOrder)
Analysis of Variance Table
Response: Sales
                    Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(Discount)  2   1288     644   24.41 3.648e-07 ***
Price                1  36230   36230 1372.84 < 2.2e-16 ***
Residuals           32    844      26
```

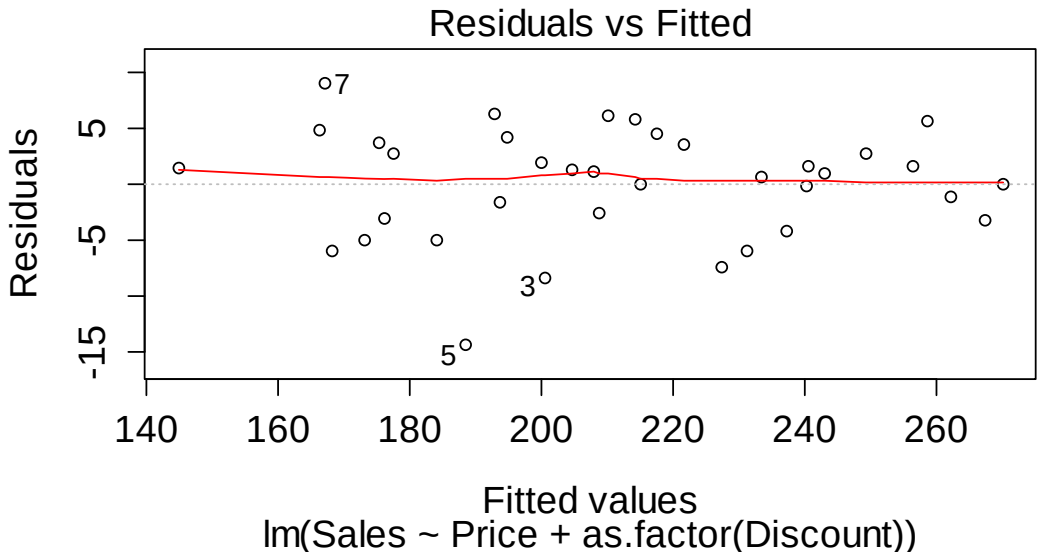
Question: Why is the Factor (Discount) now statistically significant *even before adding Price*, when it wasn't when the model was run without price at all???

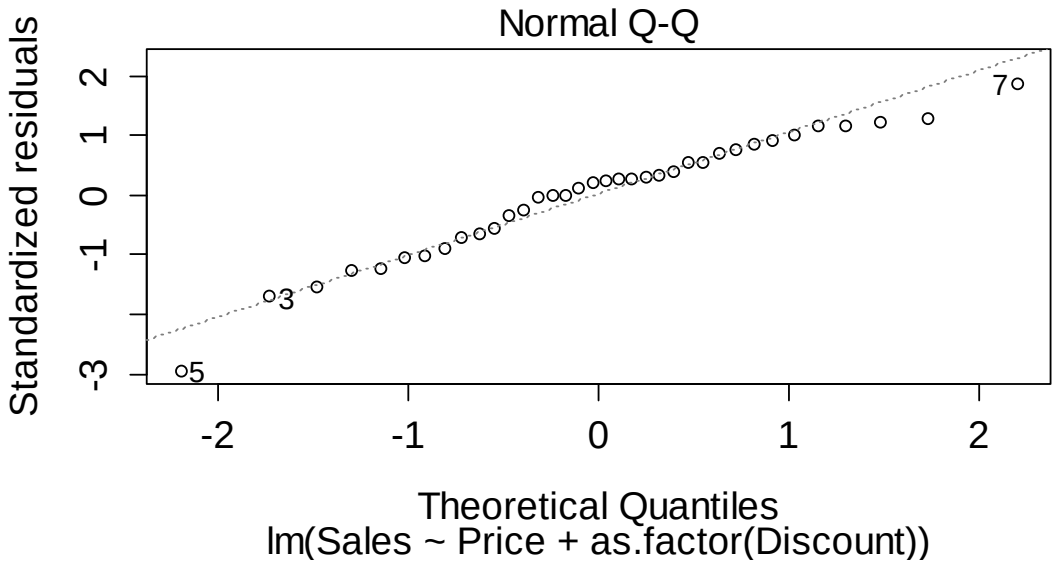
Answer: The MSE is now computed after accounting for Price. It's the MSE for the *full* model. Adding price has explained a very large amount of the previous "unexplained" residual/error!

Question: Does it matter whether you put the covariate or the factor in the model first?
Answer:
Order does not matter for the "Summary" command, but it does matter for the anova table. And the results of "summary" never test the Factor as a whole. Individual added intercept terms are tested:

```
> summary(AOC)
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -466.132     18.517 -25.173   <2e-16 ***
Price                       79.591      2.148  37.052   <2e-16 ***
as.factor(Discount)15.00%    4.655      2.111   2.205   0.0347 *
as.factor(Discount)5.00%    -6.822      2.107  -3.238   0.0028 **
Residual standard error: 5.137 on 32 degrees of freedom
Multiple R-squared:  0.978,   Adjusted R-squared:  0.9759
F-statistic: 473.9 on 3 and 32 DF,  p-value: < 2.2e-16

> summary(AOCOrder)
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -466.132     18.517 -25.173   <2e-16 ***
as.factor(Discount)15.00%    4.655      2.111   2.205   0.0347 *
as.factor(Discount)5.00%    -6.822      2.107  -3.238   0.0028 **
Price                       79.591      2.148  37.052   <2e-16 ***
Residual standard error: 5.137 on 32 degrees of freedom
Multiple R-squared:  0.978,   Adjusted R-squared:  0.9759
F-statistic: 473.9 on 3 and 32 DF,  p-value: < 2.2e-16
```

Assessing fit: Both plots look good.





The only case that may be a problem is the one labeled as “5.” It has a large standardized residual. Its predicted Sales = 188.45, actual Sales = 174 and estimated s.d. = 5.137. No obvious explanation, so *don’t* remove case!