Reminder: The Linear Model

\[ Y_i = \beta_0 + \beta_u U_i + \beta_w W_i + \beta_x X_i + \epsilon_i \]

where the \( \beta \)'s are unknown constants, and the \( U \)'s, \( W \)'s and \( X \)'s are known constants. The linear model includes several special cases better known by other names:
$Y_i = \beta_0 + \beta_u U_i + \beta_w W_i + \beta_x X_i + \epsilon_i$

where the $\beta$’s are unknown constants, and the $U$’s, $W$’s and $X$’s are known constants. The linear model includes several special cases better known by other names:

**Multiple Regression**  All explanatory variables are numeric.
Where the $\beta$’s are unknown constants, and the $U$’s, $W$’s and $X$’s are known constants. The linear model includes several special cases better known by other names:

**Multiple Regression** All explanatory variables are numeric.

**Analysis of Variance** All explanatory variables are categorical.
$Y_i = \beta_0 + \beta_u U_i + \beta_w W_i + \beta_x X_i + \epsilon_i$

where the $\beta$’s are unknown constants, and the $U$’s, $W$’s and $X$’s are known constants. The linear model includes several special cases better known by other names:

**Multiple Regression**  All explanatory variables are numeric.

**Analysis of Variance**  All explanatory variables are categorical.

**Analysis of Covariance**  Some explanatory variables are numeric, and some are categorical.
One special type of explanatory variable is the so-called *dummy variable* or *indicator variable* to mark membership in a category. Closely related are *contrasts*. The terminology you will see in R help files is *treatment coding* for a dummy variable:

<table>
<thead>
<tr>
<th>categorical</th>
<th>logical variable</th>
<th>treatment coding</th>
<th>contrast coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>A vs B</td>
<td>A</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>A</td>
<td>FALSE</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>B</td>
<td>TRUE</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
The Berkeley Longitudinal Study dataset includes a variable called `sex`, with values Male and Female.

In this case, R will automatically code a dummy variable with values 1 for Male and 0 for Female (based on alphabetical order).
Example: categorical explanatory variable

```r
> ht2.lm <- lm(ht2 ~ sex, data=Berkeley)
> summary(ht2.lm)

                Estimate Std. Error t value Pr(>|t|)
(Intercept)      87.4656   0.5842 149.718  <2e-16 
sexMale           0.9344   0.8725  1.071   0.289 

Residual standard error: 3.305 on 56 df
```

```r
> with(Berkeley, tapply(ht2, sex, mean))
   Female    Male
     87.46563  88.40000
```

The `summary()` function reports the treatment coding by combining the name of the variable (sex) with the category coded as 1: `sexMale`. The mean height for females is the intercept, for males it is the sum $87.4656 + 0.9344$. 

Albyn Jones  
Math 141
```r
> with(Berkeley,t.test(ht2~sex,var.equal=TRUE))

Two Sample t-test

data:  ht2 by sex
t = -1.0709, df = 56, p-value = 0.2888
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -2.6822985  0.8135485
sample estimates:
mean in group Female       mean in group Male
   87.46563                 88.40000
```
Anova is just another linear model!

(Assuming it is a fixed effects model. For random effects or a mixed model don’t use least squares!)
Anova Example: Michelson’s 1879 data

Michelson's Data

Albyn Jones  Math 141
Example: Anova

Call: lm(Speed ~ Run, data = Michelson)

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 299909.0 | 16.60      | 18067.73| < 2e-16  |
| Run2       | -53.0    | 23.47      | -2.25   | 0.026251 |
| Run3       | -64.0    | 23.47      | -2.72   | 0.007627 |
| Run4       | -88.5    | 23.47      | -3.77   | 0.000283 |
| Run5       | -77.5    | 23.47      | -3.30   | 0.001356 |

# Coincidence ??
> sqrt(2)*16.6
[1] 23.47595
# compute means by group of test runs
> with(Michelson, tapply(Speed, Run, mean))
1 2 3 4 5
299909.0 299856.0 299845.0 299820.5 299831.5

# compare to lm output
> coef(Michelson.lm)
(Intercept)  Run2  Run3  Run4  Run5
 299909.0  -53.0  -64.0  -88.5  -77.5

> 299909.0 + c(-53.0, -64.0, -88.5, -77.5)
[1] 299856.0 299845.0 299820.5 299831.5
> contrasts(Michelson$Run)
   2 3 4 5
1 0 0 0 0
2 1 0 0 0
3 0 1 0 0
4 0 0 1 0
5 0 0 0 1
> contrasts(Michelson$Run) <-
   contr.treatment(5, base=5)

> contrasts(Michelson$Run)
     1 2 3 4
  1 1 0 0 0
  2 0 1 0 0
  3 0 0 1 0
  4 0 0 0 1
  5 0 0 0 0
# contrasts(Michelson$Run) <-
# contr.treatment(5,base=5)
# make group 5 the baseline

> coef(lm(Speed ~ Run, data=Michelson))
(Intercept)  Run1  Run2  Run3  Run4
     299831.5   77.5   24.5   13.5  -11.0

> with(Michelson,tapply(Speed,Run,mean))
 1     2     3     4     5
299909.0 299856.0 299845.0 299820.5 299831.5
The data again...

Michelson's Data

Albyn Jones  Math 141
Michelson Data: Code Your Own Dummy Var.

> Run1 <- Michelson$Run == 1
# create a dummy variable for first run

> coef(lm(Speed ~ Run1, data=Michelson))
(Intercept) Run1TRUE
 299838.25  70.75

> with(Michelson,tapply(Speed,Run1,mean))
   FALSE TRUE
 299838.2  299909.0

> 299838.25 + 70.75
[1] 299909
> summary(aov(Speed ~ Run, Michelson))

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run</td>
<td>4</td>
<td>94514</td>
<td>23628</td>
<td>4.288</td>
<td>0.00311</td>
</tr>
<tr>
<td>Residuals</td>
<td>95</td>
<td>523510</td>
<td>5511</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

More on this next week...
Analysis of Variance is a Linear Model!

# using lm()
> coef(lm(Speed ~ Run, data=Michelson))
(Intercept)  Run2  Run3  Run4  Run5
 299909.0   -53.0  -64.0  -88.5  -77.5

# using aov(), the Anova version
> coef(aov(Speed ~ Run, data=Michelson))
(Intercept)  Run2  Run3  Run4  Run5
 299909.0   -53.0  -64.0  -88.5  -77.5
Choose other coding

Options: contr.helmert(), contr.sum(), contr.poly(), contr.treatment() — see ?contr.sum

Create your own contrast matrix:

```r
> contrasts(Michelson$Run) <- matrix(scan(), ncol=4)

1:  -4  1  1  1  1
6:   0 -1  1  0  0
11:  0  0  0 -1  1
16:  0  1  1 -1 -1
21:
```

Read 20 items
The Contrast matrix

> contrasts(Michelson$Run)
1  1   -4   0   0   0
2  1    1  -1   0   1
3  1    1   1   0   1
4  1    0  -1  -1 -1
5  1    0   1 -1  -1

What hypotheses will our t-tests test?
lm() with user specified contrasts

> summary(lm(Speed ~ Run, data=Michelson))
Coefficients:

                            Estimate Std. Error t value Pr(>|t|)
(Intercept)       299852.40      7.423 40393.06  < 2e-16
Run1              -14.15       3.712  -3.81  0.00024
Run2               -5.50      11.737  -0.46  0.64043
Run3                5.50      11.737   0.46  0.64043
Run4              12.25       8.300   1.47  0.14325

Warning! Run1 now means the first contrast, not the first run! The intercept is the overall mean.
> colnames(contrasts(Michelson$Run))<-
  c("2345-1","2-3","4-5","23-45")

> summary(lm(Speed ~ Run,data=Michelson))

Coefficients:

                   Estimate Std. Error     t value     Pr(>|t|)
(Intercept) 299852.4     7.423    40393.068 < 2e-16
Run2345-1   -14.1     3.712     -3.812  0.000245
Run2-3      -5.5    11.737     -0.469  0.640437
Run4-5      5.5    11.737      0.469  0.640437
Run23-45    12.2     8.300      1.476  0.143256
Analysis of Covariance is just another linear model!

(Note: Assuming it is a fixed effects model. For random effects or a mixed model don’t use least squares!)
Berkeley Longitudinal Study, Again

ht18.lm0 <- lm(ht18 ~ ht2, data=Berkeley)

Berkeley Residuals

![Graph showing residual values against fitted values]
Berkeley Normal Quantile Plot

Theoretical Quantiles vs. Sample Quantiles

Normal Q–Q Plot

Sample Quantiles

Theoretical Quantiles

Albyn Jones  Math 141
Berkeley Analysis of Covariance

> ht18.lm1 <- lm(ht18 ~ ht2 + sex, data=Berkeley)
> summary(ht18.lm1)

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 49.4174  | 16.4052    | 3.012   | 0.00391  |
| ht2            | 1.3416   | 0.1873     | 7.162   | 2.05e-09 |
| sexMale        | 12.0192  | 1.2356     | 9.727   | 1.49e-13 |

Residual standard error:

4.633 on 55 degrees of freedom
Berkeley Longitudinal Study, ANCOVA

Berkeley Analysis of Covariance

The scatter plot shows the relationship between two variables, likely height measurements over time, with annotations and data points plotted accordingly.

Albyn Jones  Math 141
R code for previous plot

```
coef(ht18.lm1)
  (Intercept)   ht2   sexMale
  49.417437  1.341649  12.019233

plot(ht2, ht18, pch=19,
     col=ifelse(Berkeley$sex=='Male','blue','red'))

abline(49.417, 1.342, lwd=2, col='red')

abline(49.417 + 12.019, 1.342, lwd=2, col='blue')

title('Berkeley Analysis of Covariance')
```
A Dummy Variable Test for Outliers!

```r
> PalmBeach <- rep(0, 67)
> PalmBeach[50] <- 1
> summary(lm(log(Buchanan) ~
    log(Bush) + PalmBeach, data=FL))

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| -2.3179  | 0.3548     | -6.533  | 1.22e-08 |
| log(Bush)  | 0.7297   | 0.0360     | 20.270  | < 2e-16  |
| PalmBeach  | 1.7411   | 0.4308     | 4.041   | 0.000145 |

Residual standard error:

0.4204 on 64 degrees of freedom

Compare the coefficient for PalmBeach to the res. SE!
Linear Models and Dummy Variables

- Dummy Variables are used to build models with categorical explanatory variables.
- Anova (Analysis of Variance) is a linear model with only categorical explanatory variables.
- Ancova (Analysis of Covariance) is a linear model with both numerical and categorical explanatory variables.
- Dummy variables and contrast variables can be constructed to test special hypotheses.