Fun with models (part 1): So far have seen:
To "control" for important confounding variables
To allow lines to curve

What's to come:
Using regression to fit T-test and ANOVA models
Models with both categorical and continuous variables:
Analysis of covariance
Heterogeneous regression lines models

Connecting regression and T-test:
T-test model using means: $Y_{ij} = \mu_i + \varepsilon_{ij},\ \ i =' a', 'b'$
2 groups ('a', and 'b'), 2 μ_i parameters
Indicator variable:

$$X = I(\text{something}) \text{ means } X = \begin{cases} 1 & \text{when something is true} \\ 0 & \text{when something is false} \end{cases}$$

So I(group = 'b') is 1 when the group = "b" and 0 when the group = "a"
Write T-test model as a regression, using $X_{ij} = I(\text{obs } i, j \text{ in group b})$
T-test model using regression: $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$
Relationship between the two T-test models

group	mean	X_{ij}	regression
a	μ_a	0	β_0
b	μ_b	1	$\beta_0 + \beta_1$

Interpretation of the regression coefficients with X = I(group="b") (R only):
JMP and SAS define the indicator variable, X, differently
More a bit later

coefficient	In terms of means
β_0	μ_a
β_1	$\mu_b - \mu_a$

Connecting regression and ANOVA:
T-test ideas, just more groups and a complication
ANOVA model: $Y_{ij} = \mu_i + \varepsilon_{ij}$
Define 3 indicator variables, one for each group:
So I(group = "b") is 1 when the group = "b" and 0 when the group = "a" or "c"
$X_{1i} = I$(i'th obs has group ="a"),
$X_{2i} = I$(i'th obs has group ="b"),
$X_{3i} = I$(i'th obs has group ="c")
Fit the model $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon i$ (Note: no β_0, so no intercept)

group	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1	0	0	$\beta_1 = \mu_a$
b	0	1	0	$\beta_2 = \mu_b$
c	0	0	1	$\beta_3 = \mu_c$

Add an intercept to previous model

Write as a regression using a column of 1's for β_0

Model is $Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon i$

group	X_{0i}	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1	1	0	0	$\beta_0 + \beta_1 = \mu_a$
b	1	0	1	0	$\beta_0 + \beta_2 = \mu_b$
c	1	0	0	1	$\beta_0 + \beta_3 = \mu_c$

Nasty numerical problem: $\boldsymbol{X}$ has 4 columns, but 1 is redundant

Choose any three, fourth can be computed from them. fourth is not new information.

Called a "non-full rank" $\boldsymbol{X}$ matrix

Can not use the matrix equation $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ because $\boldsymbol{X}'\boldsymbol{X}$ has no inverse.

Software "fix" the problem differently

R: Drop the first column (X_1). Remaining three are full rank.

SAS: uses generalized inverse methods for non-full rank matrix,

equivalent to dropping last column

JMP: uses "effects" coding, +1, 0 or −1 and drops the last column

Can request indicator parameterization (drop last column)

The choice changes the estimated regression coefficients

So we have a problem with interpretation if you focus on coefficients:

R, SAS, and JMP give different estimates for β_1 !!

NOT GOOD. Answer depends on arbitrary choice of parameterization

Take home point: use 2 programs to fit same model, will get different $\hat{\beta}$'s

But: Many important quantities do not depend on the parameterization

R:

group	X_{0i}	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1		0	0	$\beta_0 = \mu_a$
b	1		1	0	$\beta_0 + \beta_2 = \mu_b$
c	1		0	1	$\beta_0 + \beta_3 = \mu_c$

SAS:

group	X_{0i}	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1	1	0		$\beta_0 + \beta_1 = \mu_a$
b	1	0	1		$\beta_0 + \beta_2 = \mu_b$
c	1	0	0		$\beta_0 = \mu_c$

JMP:

group	X_{0i}	X_{1i}	X_{2i}	X_{3i}	predicted value
a	1	1	0		$\beta_0 + \beta_1 = \mu_a$
b	1	0	1		$\beta_0 + \beta_2 = \mu_b$
c	1	-1	-1		$\beta_0 - \beta_1 - \beta_2 = \mu_c$

Problem: All β's have different estimates in R, SAS, or JMP !!
Example: 3 groups, means are $\overline{Y}_1 = 5$, $\overline{Y}_2 = 10$, $\overline{Y}_3 = 9$

Parameter	JMP	R	SAS
β_0	8	5	9
β_a	-3	–	-4
β_b	2	5	1
β_c	–	4	–

NOT GOOD. Estimates of β's depend on arbitrary choice of parameterization
My advice: don't look at estimates of β's in ANOVA models
In R, don't look at summary() output
unless you understand how to interpret the coefficients
SAS and JMP: don't show the estimates unless you specifically request them

Estimable functions:
Good news: some quantities, such as mean for group same for all three param.
Estimable function: an estimate that does not depend on arbitrary choices
Theory (not in this course): defines what is and what is not an estimable function
Some estimable functions: μ_a, $\mu_a - \mu_b$, $\mu_a - (\mu_b + \mu_c)/2$
Some non-estimable functions: β_1, $\mu_a - (\mu_b + \mu_c)$
If software tells you 'non-est', you either
wrote the wrong quantity (bad contrast or estimate statement)
wrote the wrong model
or the data is insufficient to fit the model

Software	β_0	β_1	β_2	β_3	μ_a	$\hat{\mu}_a$	μ_b	$\hat{\mu}_b$	μ_c	$\hat{\mu}_c$
JMP	8	-3	2	–	$\beta_0 + \beta_1$	8 - 3 = 5	$\beta_0 + \beta_2$	8 + 2 = 10	$\beta_0 - \beta_1 - \beta_2$	8 + 3 - 2 = 9
R	5	–	5	4	β_0	5	$\beta_0 + \beta_2$	5 + 5 = 10	$\beta_0 + \beta_3$	5 + 4 = 9
SAS	9	-4	1	–	$\beta_0 + \beta_1$	9 - 4 = 5	$\beta_0 + \beta_2$	9 + 1 = 10	β_0	9

More examples of estimable functions:

Software	β_0	β_1	β_2	β_3	$\mu_a - \mu_b$	$\hat{\mu}_a - \hat{\mu}_a$	$\mu_c - (\mu_a + \mu_b)/2$	$\hat{\mu}_c - (\hat{\mu}_a + \hat{\mu}_a)/2$
JMP	8	-3	2	–	$\beta_1 - \beta_2$	-3 -2 = -5	$-1.5(\beta_1 + \beta_2)$	-1.5(-3 + 2) = 1.5
R	5	–	5	4	$-\beta_2$	-5	$\beta_3 - \beta_2/2$	4 - 2.5 = 1.5
SAS	9	-4	1	–	$\beta_1 - \beta_2$	-4 -1 = -5	$-(\beta_1 + \beta_2)/2$	-(-4 + 1)/2 = 1.5

Fun with models (part 2):
Combining groups and continuous predictor variables

ANCOVA: analysis of covariance

$$Y_{ij} = \mu_i + \beta\, X_{ij} + \varepsilon_{ij}$$

i indicates groups, j observation within group
parallel lines, each group has a different intercept
ANCOVA as a regression model

$$Y_{ij} = \beta_0 + \beta_1 G_{ij} + \beta_2\, X_{ij} + \varepsilon_{ij}$$

Define $G_{ij} = I(group =' b')$

Group	G_{ij}	Equation
a	0	$\beta_0 + \beta_2 X_{ij}$
b	1	$\beta_0 + \beta_1 + \beta_2 X_{ij}$

Heterogeneous regression lines

$$Y_{ij} = \mu_i + \beta_i\, X_{ij} + \varepsilon_{ij}$$

each group (i) has a different intercept and a different slope
As a regression

$$Y_{ij} = \beta_0 + \beta_1 G_{ij} + \beta_2\, X_{ij} + \beta_3 G_{ij}\, X_{ij} + \varepsilon_{ij}$$

Define $G_{ij} = I(group = b)$

Group	G_{ij}	Equation
a	0	$\beta_0 + \beta_2 X_{ij}$
b	1	$(\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_{ij}$

Additive effects:
Most previous regression models have had additive effects
Example: model with 2 continuous predictor variables, X_1 and X_2
$Y_i = \beta_0 + \beta_1\, X_{1i} + \beta_2\, X_{2i} + \varepsilon_i$
Changing X_1 by 1 unit changes $\hat{Y}$ by β_1 units,
no matter what value X_2 has
Analogous consequence for changing X_2
Example: model with sex (indicator for female) and age (continuous)
M and F have same change in $\hat{Y}$ when age increased by 1
difference (female - male) = sex effect same for all ages
plot of Y vs age has two parallel lines (same difference at all ages)

Interaction:
In general, effect of one X variable depends on level of a second
Example: 2 continuous predictor variables

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$
Change in $\hat{Y}$ when X_{1i} increased by one is $\beta_2 + \beta_3 X_2$
Change depends on value of X_2
Example: model with sex (indicator for female) and age (continuous)
Heterogeneous regression lines model has an interaction
$Y_{ij} = \mu_i + \beta_i X_{ij} + \varepsilon_{ij}$
difference (female - male) of same age is $\mu_f - \mu_m + (\beta_f - \beta_m)X$
depends on age, not constant
Interactions can be between:
a grouping variable (e.g. sex) and a continuous one (e.g., age)
so slope relating Y to age is different for M and F
other examples are light/flowering time, bat echolocation
two continuous variables (e.g., litter size and body weight)
so slope relating brain size to litter size depends on body weight
two grouping variables (e.g., sex and ethnicity)
So difference between sexes, M-F, is not constant, depends on ethnicity
We'll talk a lot more about this situation in 2 way ANOVA

Diagnostics: old tools
Residual vs predicted value plot: equal variances, outliers, lack of fit
And 3 new tools: influence, standardized residuals, multicollinearity

Influence:
"Outliers" in X space. Outliers pull the fitted line to the obs.
Cook's distance: How much do fitted values change when delete one obs.
computed for each point

$$D_i = \frac{\Sigma_{allobs.}(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{p\, s^2}$$

$\hat{Y}_{j(i)}$ is predicted value for j'th obs. when i'th obs. is deleted
$D_i \approx 0$: good, deleting that obs. doesn't not change predicted values
$D_i > 1$: deleting that obs. really changes predicted values
an unusually influential obs.

Standardized (Studentized) residuals.
Residuals can have very different variances even if errors have constant variance
Happens in SLR, but often much worse in MLR
Small variance when an outlier pulls the fitted line to that obs.

$$\begin{aligned} r_i &= Y_i - \hat{Y}_i \text{ usual residual} \\ r_i^s &= \frac{r_i}{\sqrt{\text{Var } r_i}} \end{aligned}$$

slight differences (not important) between standardized and studentized versions
If model fits and errors are normal,
standardized residuals are normally distributed with mean 0 and sd 1
95% of standardized residuals between -2 and 2

Multicollinearity

Two (or more) X variables highly correlated in the data set.

"hard" to separate the effects of the two X variables

Consequence: very large se for a regression coefficient

so non-signif. p-value

Should suspect multicollinearity when overall F test has $p < 0.05$

but all coefficient-specific T-tests have $p > 0.05$ or > 0.10

Assess by variance-inflation factor (VIF)

$$\text{VIF}_i = \frac{\text{Var } \hat{\beta}_i \text{ in MLR}}{\text{Var } \hat{\beta}_i \text{ when X's uncorrel.}} = \frac{1}{1 - R_i^2}$$

R_i^2 measures how well (0-1 scale) X_i predicted by other X variables

VIF ≈ 1 is great; > 10 is bad