

Regression Models & ANCOVA

Advanced Biostatistics

Dean Adams

Lecture 4

EEOB 590C

Regression: $Y \sim X$

-X & Y are continuous

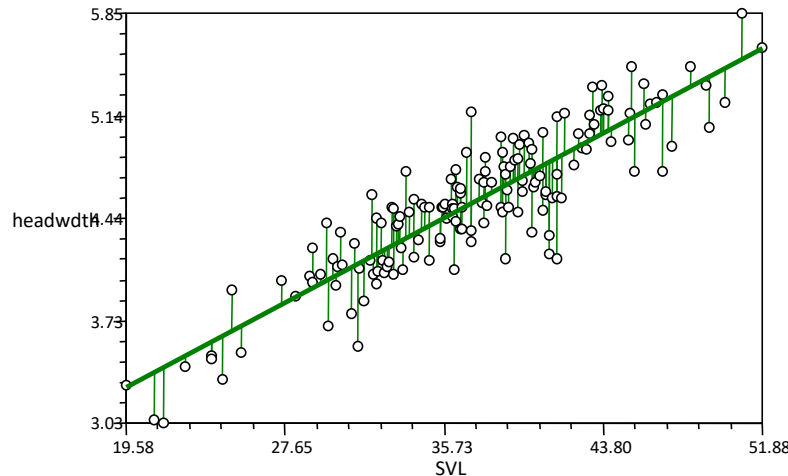
H_0 : no relationship between X & Y

-Compare variation explained by model to residual error variation

Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ (β_0 is grand mean, β_1 is slope, and ε_{ij} is error)

-Standard test statistic: F-ratio (ratio of variances)

$$F \approx \frac{\sigma_{\text{model}}^2}{\sigma_{\text{error}}^2}$$



Name from Galton's work: offspring values 'regress' (tend towards) to that of parents

1: Independence: ε_{ij} of variates must be independent

2: Normality: requires normally distributed ε_{ij}

3: Homoscedasticity: equal variance

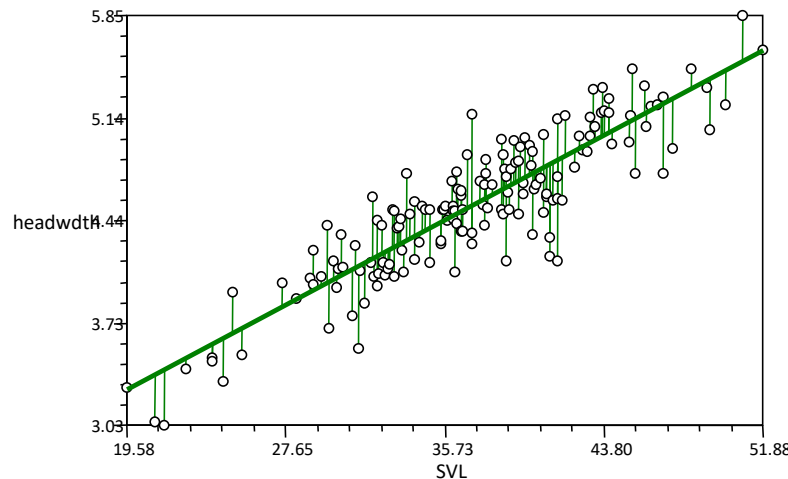
-Samples along regression line are homoscedastic (variance doesn't 'balloon' along regression)

4: X values are independent and measures *without error*

Fit line that minimizes sum of squared deviations (LS fit) from Y-variates to line
(vertical deviations b/c no error in X)

Slope calculated as: $b_{Y \cdot X} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ ($\frac{\sum(\text{cross products})}{SS_x}$)

Regression line always crosses (\bar{X}, \bar{Y}) so intercept is: $\beta_0 = \bar{Y} - b_{Y \cdot X} \bar{X}$



Note: this is Model I regression with 1 Y for each X. Slight alterations exist for multiple Y for each X: see Biometry.

Can obtain SS_{Model} & SS_{Error}

$$Y \sim X$$

SS explained: $SSR = SSM = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ where \hat{Y} are predicted values

SS error: $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

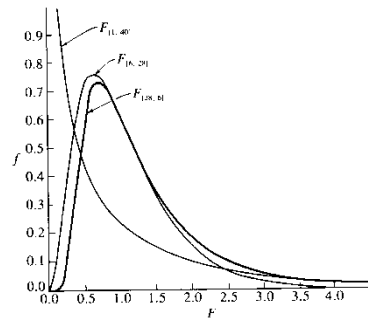
Traditional test: F-ratio (MSR / MSE)

Source	df	SS	MS	F	P
Regression	1	SSM	SSM/df	MSM/MSE	
Error	n-2	SSE	SSE/df		
Total	n-1	SST=SSM+SSE			

Tests whether regression explains a significant portion of variation

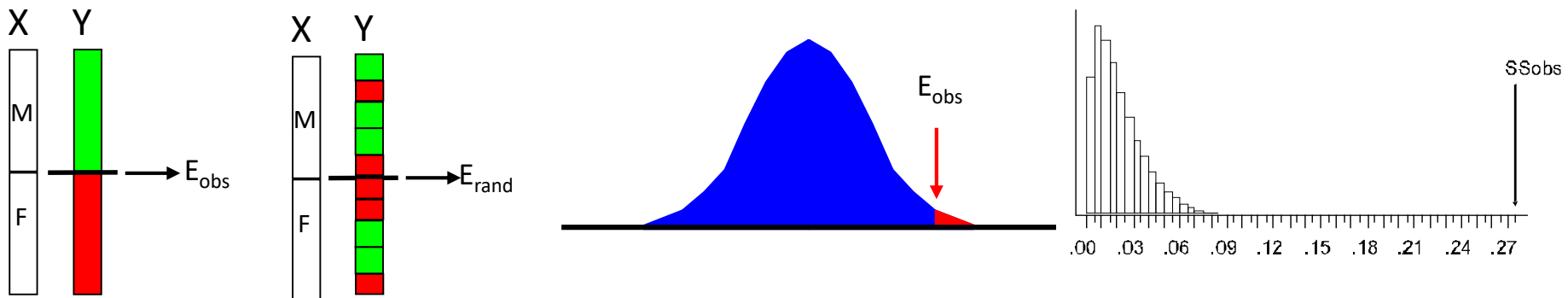
Standard approach:

- Compare F-ratio to F-distribution with appropriate df



Resampling Alternative:

- Shuffle **Y** relative to **X** & generate distribution of possible outcomes



F-test assesses variance explained, but not how much Y changes with X

Can evaluate the model parameters separately

Slope test ($\beta_1 \neq 0$): $t = \frac{\beta_1 - 0}{s_{\beta_1}}$ with $s_{\beta_1} = \sqrt{\frac{MSE}{\sum_{i=1}^n (X_i - \bar{X})^2}}$ (n-2 df)

Intercept test ($\beta_0 \neq 0$): $t = \frac{\beta_0 - 0}{s_{\beta_0}}$ with $s_{\beta_0} = \sqrt{MSE \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$ (n-2 df)

Can also test against particular values (e.g., Isometry: is $\beta_1 = 1$?)

Very useful for certain biological hypotheses

```
> summary(lm(Y~X1))#TotalLength ~ AlarExtent
```

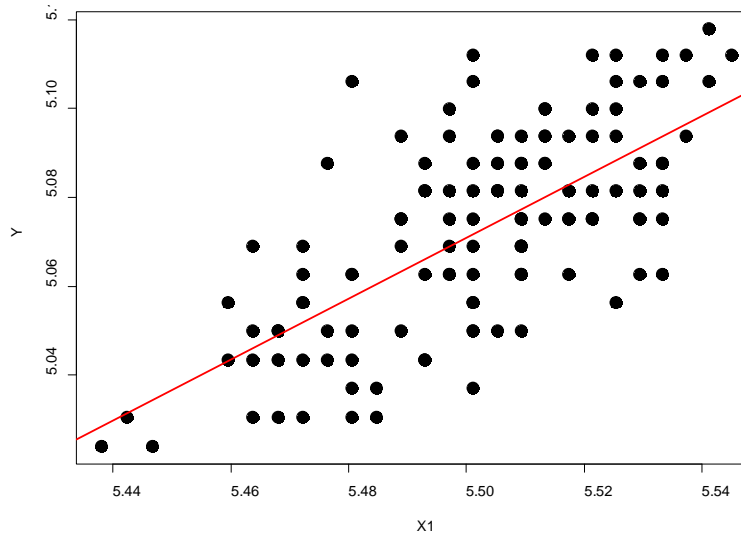
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.3045	0.3383	3.856	0.000179	***
X1	0.6848	0.0615	11.136	< 2e-16	***

Multiple R-squared: 0.4806, Adjusted R-squared: 0.4768

```
> anova(lm(Y~X1))#provides regression coefficients
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	0.032412	0.032412	124.01	< 2.2e-16 ***



$$Y = 1.3045 + 0.6848X$$

Correlation and regression closely linked

Mathematically: $b_{Y \cdot X} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{SS_x}$ vs. $r_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_x s_y}$

Difference in denominator, so related as: $r_{xy} = b_{Y \cdot X} \left(\frac{s_x}{s_y} \right)$

r is a *standardized* regression coefficient (i.e. slope in standard deviation units)

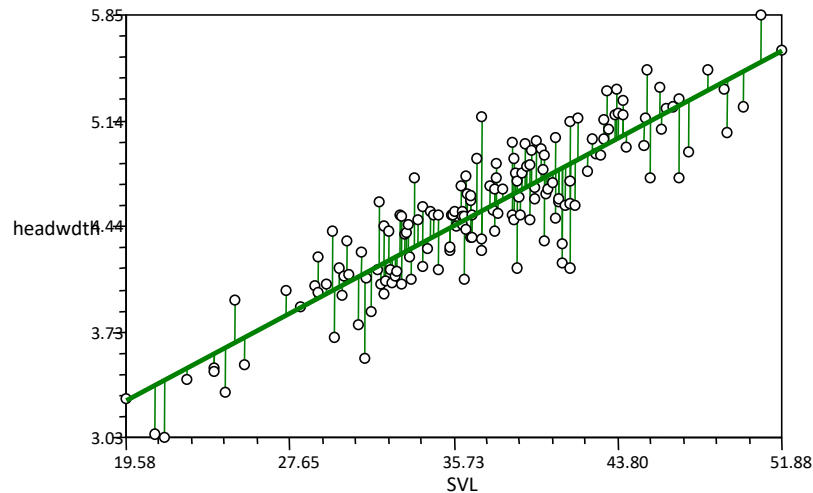
- β : error in Y direction only (direction of scatter)

- r : error in X & Y (dispersion of scatter)

-Thus, regression models *causation* (Y as function of X) and correlation models association (covariation in X & Y)

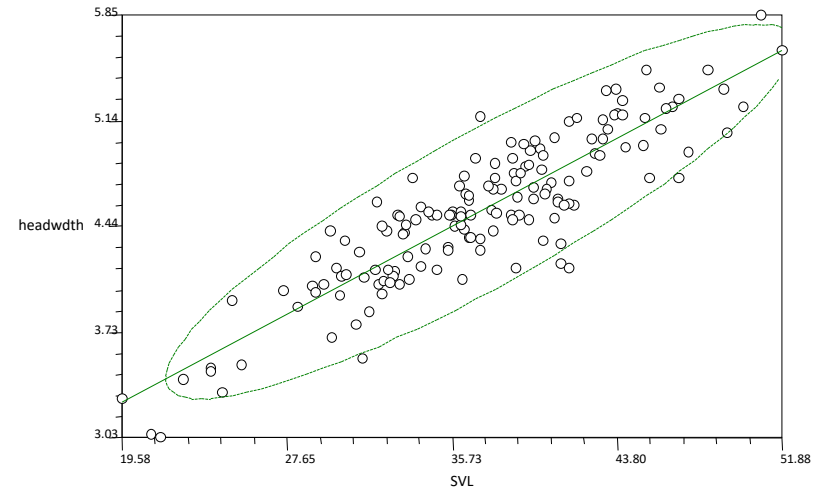
Regression

$$b_{Y.X} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{SS_x}$$



Correlation

$$r_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_x s_y}$$



When both X & Y contain measurement error, model I regression underestimates slope

- Model II regression accounts for this by minimizing deviations perpendicular to regression line (not in Y direction only)

- Different types of model II regression, depending on data (major axis, reduced major axis, etc.)

- X & Y in 'same' units/scale: major axis regression (PCA)

- X & Y not in same units/scale: standard (reduced) major axis regression

When X & Y in 'same' units/scale: major axis regression (PCA)

Compute covariance matrix for X & Y:

$$VCV = \begin{bmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{bmatrix}$$

Obtain principle eigenvalue (λ) and eigenvector

$$\lambda_1 = \frac{s_X^2 + s_Y^2 + \sqrt{(s_X^2 + s_Y^2)^2 - 4(s_X^2 s_Y^2 - s_{XY}^2)}}{2}$$

Slope is:
$$b_{Y \cdot X} = \frac{s_{XY}}{\lambda_1 - s_X^2}$$

When X & Y not in same units/scale: standard (reduced) major axis regression

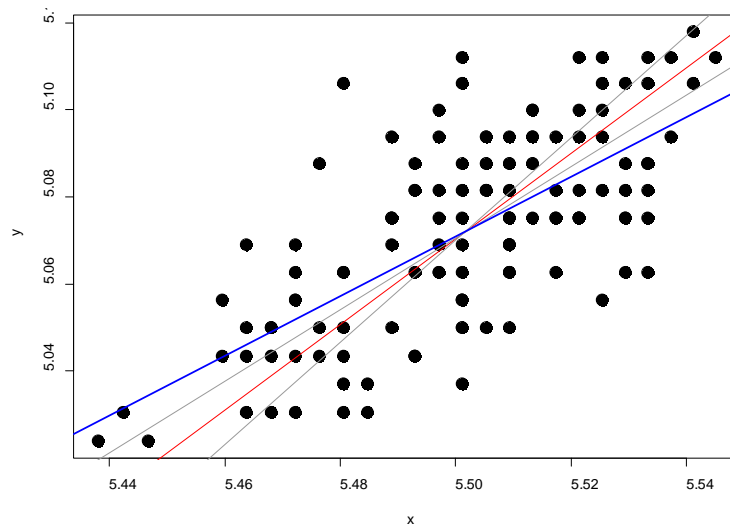
Standardize variables to $N(0,1)$ (standard normal deviates)

Calculate SMA slope as:
$$b_{Y \cdot X rma} = \frac{b_{Y \cdot X}}{r_{XY}}$$

```
> lmodel2(Y~X1,nperm=999)
```

Angle between the two OLS regression lines = 20.53316 degrees

Method		Intercept	Slope	Angle (degrees)	P-perm (1-tailed)
1	OLS	1.3044593	0.6847984	34.40328	0.001
2	MA	-0.3329159	0.9824064	44.49152	0.001
3	SMA	-0.3624212	0.9877692	44.64746	NA



OLS slope (blue)

MA slope (red)

*SMA slope = β_{OLS}/r

***NOTE: Justification for RMA usage has been overplayed in biology.**

- What matters is *NOT* whether X has measurement error: $\text{var}(\varepsilon_x)$
- Instead, *ONLY* when $\text{var}(\varepsilon_x)/\text{var}(X)$ is large, might there be an issue with Model I regression
- But even in this case, RMA is not guaranteed to alleviate the problem
(see Kilmer & Rodriguez, 2017. *J. Evol. Biol.*)

Identify relationship between several X and continuous Y

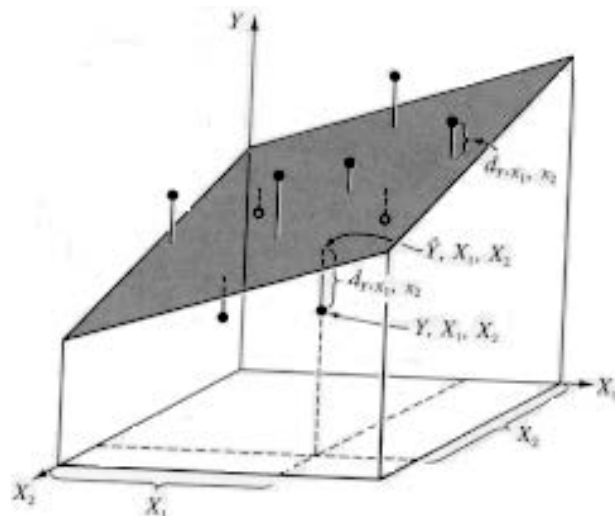
Predict Y using several variables simultaneously

$$Y \sim A + B$$

Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \varepsilon_i$

β_i are partial regression coefficients (effect of X_i while holding effects of other X constant)

For 2X, think of fitting a plane to the data



```
> summary(lm(Y~X1+X2))
```

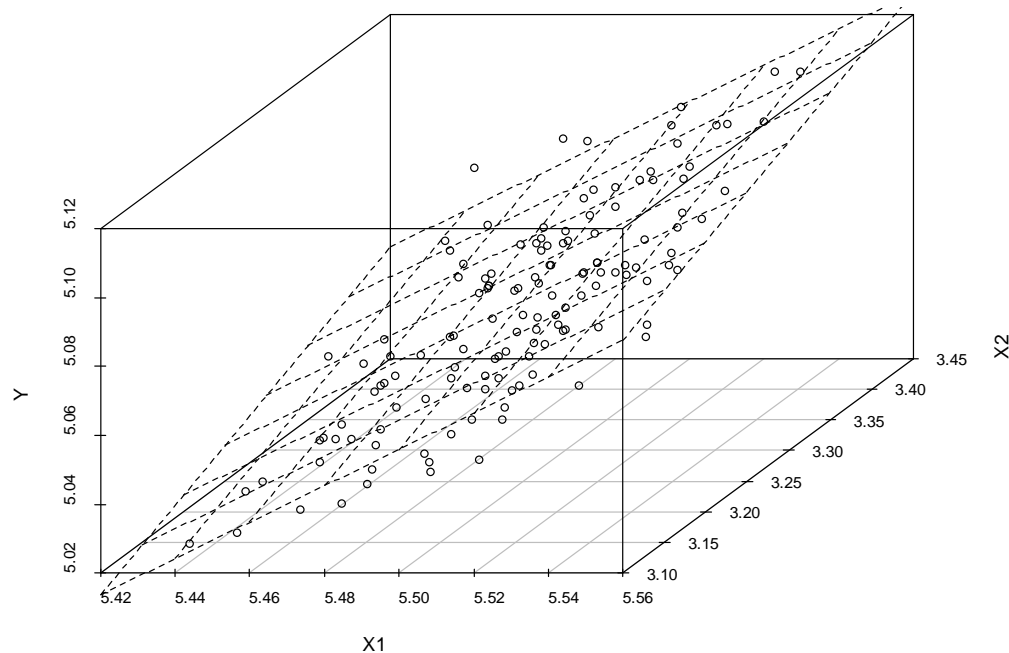
```
Estimate Std. Error t value Pr(>|t|)
```

(Intercept)	1.82521	0.34843	5.238	6.18e-07	***
X1	0.52516	0.07142	7.354	1.77e-11	***
X2	0.11042	0.02835	3.895	0.000155	***

```
> anova(lm(Y~X1+X2))
```

```
Df Sum Sq Mean Sq F value Pr(>F)
```

X1	1	0.032412	0.032412	137.118	< 2.2e-16	***
X2	1	0.003585	0.003585	15.168	0.0001551	***



Standard partial regression coefficients $b'_{YX_i \cdot X_j}$: β of Y & X_i holding X_j constant.

Expresses change in normalized units (standard normal deviates: $\frac{Y - \bar{Y}}{\sigma_Y}$)

ADVANTAGE: can be directly compared for each variable

Found from variable correlations: $b'_{YX_1 \cdot X_2} = \frac{r_{X_1 Y} - r_{X_2 Y} r_{X_1 X_2}}{1 - r_{X_1 X_2}^2}$

Interpret, then calculate back to original units and for conventional partial regression coefficients

$$b_{YX_1 \cdot X_2} = b'_{YX_1 \cdot X_2} \frac{S_Y}{S_{X_1}}$$

For all models, R^2 : proportion of variance explained by model

Should I add another variable? Test difference in R^2

NOTE: R^2 will 'top' out as you add variables (so frequently adjusted R^2 used)

Can also compare models using 'aov' in R, using AIC, etc.

How to add variables: stepwise addition, stepwise deletion, random addition, etc.

NOTE: different methods yield different results (b/c $b'_{YX_i \cdot X_j}$ depend on other variables in model)

CAREFUL WITH BIOLOGICAL INTERPRETATION!!!

To compare 2 regression lines, calculate F-ratio as:

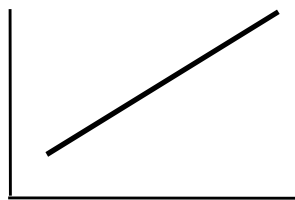
$$F = \frac{(b_1 - b_2)^2}{\frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{\left(\sum (X_1 - \bar{X}_1)^2\right)\left(\sum (X_2 - \bar{X}_2)^2\right)} \bar{s}_{Y \cdot X}^2}$$

Df = 1, (n₁ + n₂ - 4)

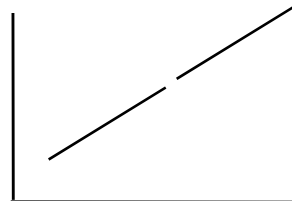
Where $\bar{s}_{Y \cdot X}^2$ is the weighted average of $s_{Y \cdot X}^2$

Procedure can be generalized to compare > 2 regression lines (see *Biometry*)

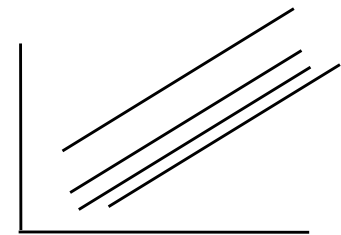
- Often, one wishes to compare regression lines AND the groups
- ANCOVA 'combines' regression and ANOVA
- H_0 : no difference among slopes, no difference among groups
- Must first compare slopes, then compare groups (ANOVA) while holding effects of covariate constant
- Several possible outcomes



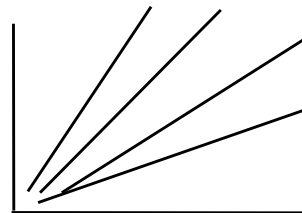
Groups the same,
slopes the same
(overlapping data)



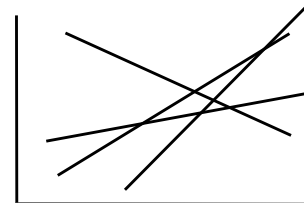
Groups the same, slopes
the same
(non -overlapping data)



Slopes the same,
groups different



Slopes different, 'easily'
identified pattern



Slopes different,
'complicated' pattern

Model:
$$Y_{ij} = \mu + \alpha_i + \beta_{within} (X_{ij} - \bar{X}_i) + \varepsilon_{ij}$$

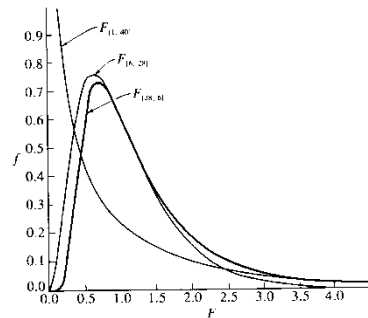
Calculate SS for:

- 1: covariate (common regression)
- 2: group x covariate interaction (slopes test)
- 3: groups

Source	df	SS	MS	F	P
Covariate	1	SSR	SSR/df		
Group	a-1	SSG	SSG/df		
Cov x Group	a-1	SSInt	SSInt/df		
Error	n-2(a-1)+1	SSE	SSE/df		

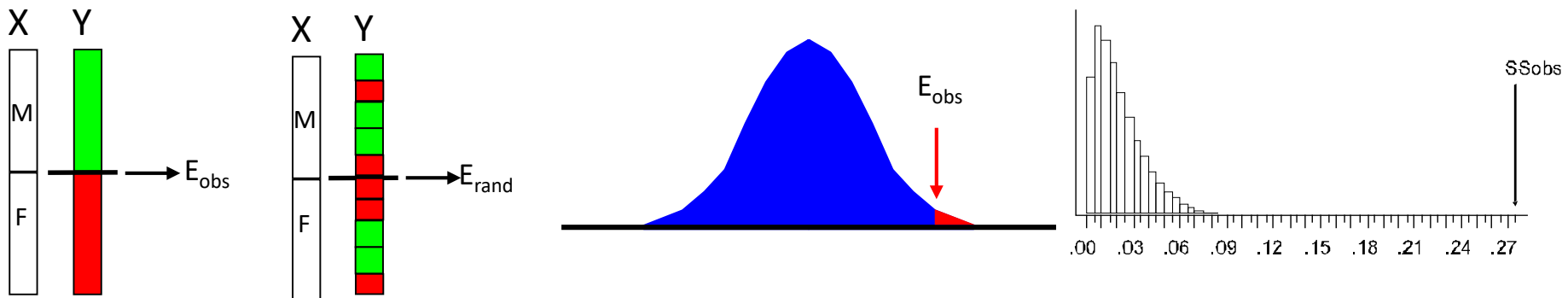
Standard approach:

- Compare F-ratio for each factor to F-distribution with appropriate df



Resampling Alternative:

- Residual randomization: shuffle residuals from reduced model to assess that factor (e.g., remove $A \times cov$ and use randomization to test $A \times cov$ factor)



- 1: Are slopes for groups different ($SS_{\text{cov} \times \text{group}} > 0?$)
- 2: If interaction significant, then *we cannot compare groups*, because slopes differ
Group comparisons not useful, as the variation is in their group-specific slopes
- 3: If interaction term is NOT significant, then test groups while holding covariate constant (ie, refit new model as $Y \sim \text{cov} + \text{group}$)

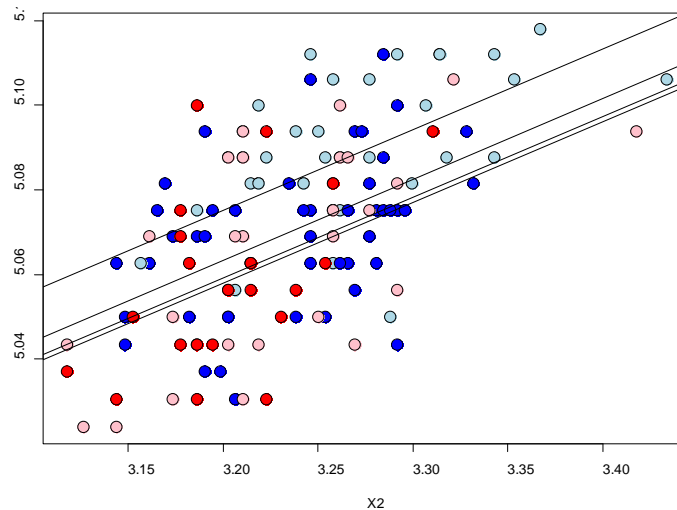
```
> anova(lm(Y~X2*SexBySurv))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2	1	0.023215	0.0232149	77.3890	8.139e-15	***
SexBySurv	3	0.005172	0.0017239	5.7467	0.001014	**
X2:SexBySurv	3	0.000651	0.0002172	0.7239	0.539496	
Residuals	128	0.038397	0.0003000			

Fit common-slope model

```
> anova(lm(Y~X2+SexBySurv))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2	1	0.023215	0.0232149	77.8814	6.015e-15	***
SexBySurv	3	0.005172	0.0017239	5.7833	0.0009591	***



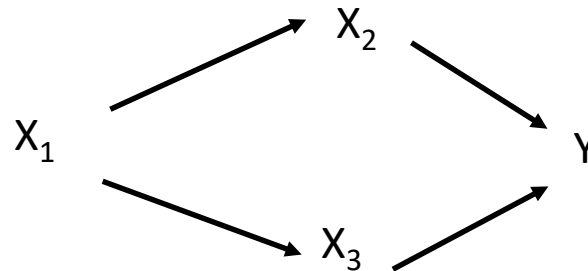
1: Perform ANOVA on regression residuals: NOT the same as ANCOVA (different df, pooled β , etc.). Also, lose test of slopes, which is important (see J. Anim. Ecol. 2001. 70:708-711)

2: Significant cov:gp interaction, but still compare groups: not useful, as answer depends upon where along regression you compare

3: “Size may be a covariate, so I’ll use a small size range to ‘standardize’ for it”: choosing animals of similar sizes will eliminate covariate, but also will eliminate potentially important biological information (e.g., what if male head width grows relatively faster than females (i.e. size:head interaction?)

Many other models possible

-PATH ANALYSIS & SEM (structural equation modeling): looks at partial correlation coefficients and partial regression coefficients to explain variance in data according to hypothesized 'causal' path between variables



Curvilinear regression: Fit polynomials of X : X , X^2 , X^3 , etc.

Regression an incredibly flexible tool for testing hypotheses

Can add factors, based on theory of partitioning SS

Careful in decisions here!

Permutations also incredibly useful for hypothesis testing