

Model Building

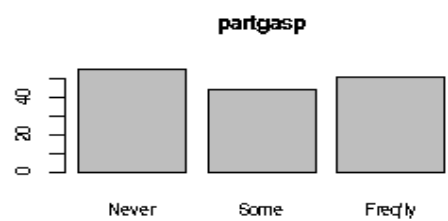
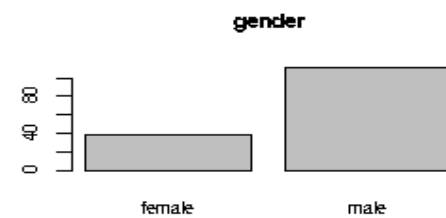
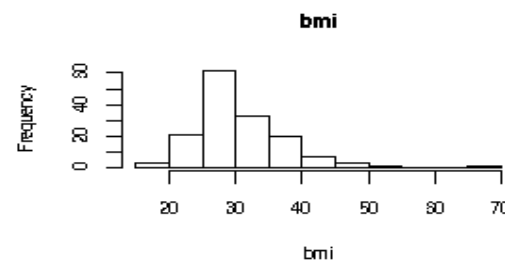
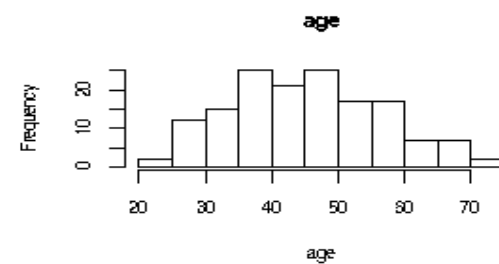
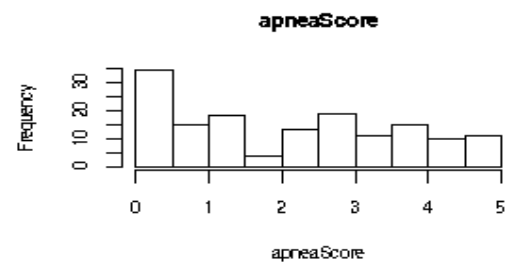
Example: Diagnosing sleep apnea

Response: apneaScore = score on overnight sleep test

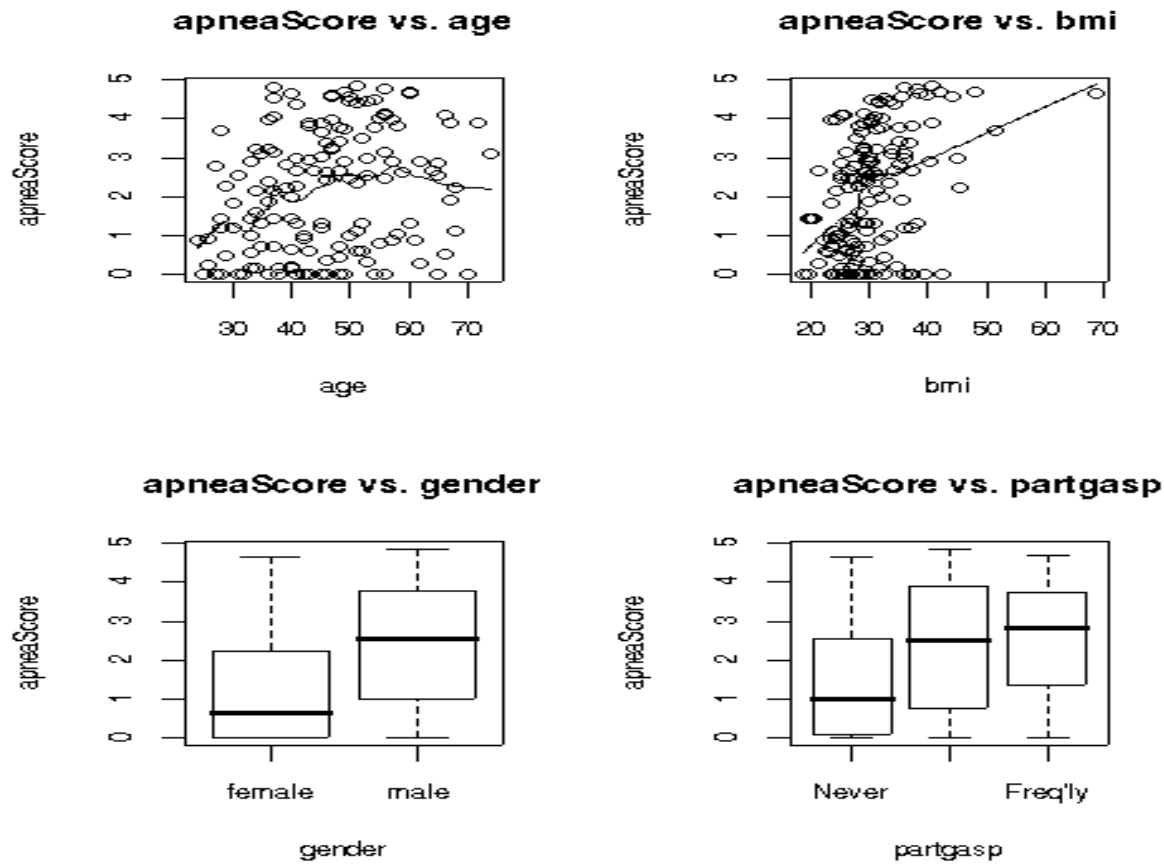
Possible predictor variables:

Demographic characteristics: age, sex, BMI

partgasp = partner report of gasping



Preliminary examination



Additive model:

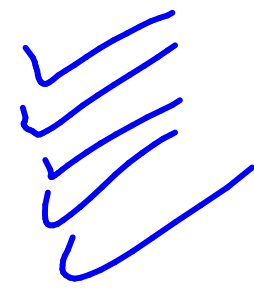
```
lm(formula = apneaScore ~ age + bmi + gender + partgasp)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.728733	0.655531	-4.163	5.39e-05	***
age	0.025295	0.009606	2.633	0.009378	**
bmi	0.086577	0.016566	5.226	5.96e-07	***
gendermale	0.872773	0.252911	3.451	0.000733	***
partgaspSome	0.563673	0.270722	2.082	0.039101	*
partgaspFreq'ly	0.617452	0.266779	2.314	0.022056	*

Signif. codes: 0 0.001 0.01 0.05 0.1 1

Residual standard error: 1.312 on 144 degrees of freedom
Multiple R-Squared: 0.3259, Adjusted R-squared: 0.3025
F-statistic: 13.92 on 5 and 144 DF, p-value: 4.269e-11



Confidence intervals

	2.5 %	97.5 %
(Intercept)	-4.024438667	-1.4330269
age	0.006308605	0.0442807
bmi	0.053832684	0.1193213
gendermale	0.372876253	1.3726699
partgaspSome	0.028571533	1.0987748
partgaspFreq'ly	0.090144115	1.1447603

Rescaling for Interpretability

```
lm(formula = apneaScore ~ I(age/20) + I(bmi/10) + gender + partgasp)
```

	Estimate	Std. Error	t value	Pr(> t)
...				
I(age/20)	0.5059	0.1921	2.633	0.009378 **
I(bmi/10)	0.8658	0.1657	5.226	5.96e-07 ***
...	2.5 %	97.5 %		
I(age/20)	0.12617209	0.885614		
I(bmi/10)	0.53832684	1.193213		

Model with Interactions with Gender

```
lm(formula = apneaScore ~ gender * (age + bmi + partgasp))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.34413	0.99458	-1.351	0.1787
gendermale	-1.57477	1.34082	-1.174	0.2422
age	0.01725	0.01703	1.013	0.3130
bmi	0.05007	0.02773	1.806	0.0731 .
partgaspSome	0.38114	0.54444	0.700	0.4851
partgaspFreq'ly	1.27931	0.68765	1.860	0.0649 .
gendermale:age	0.01586	0.02090	0.759	0.4492
gendermale:bmi	0.05766	0.03626	1.590	0.1141
gendermale:partgaspSome	0.33006	0.63188	0.522	0.6023
gendermale:partgaspFreq'ly	-0.63705	0.75110	-0.848	0.3978

Signif. codes: 0 0.001 0.01 0.05 0.1 1

Residual standard error: 1.31 on 140 degrees of freedom
 Multiple R-Squared: 0.347, Adjusted R-squared: 0.305
 F-statistic: 8.265 on 9 and 140 DF, p-value: 8.263e-10

dummy variables

This weeks office hour:
 Thursday 11-12am
 Office 308D

Review: Estimation of $\beta = \hat{\beta}$

σ^2 - residual variance

$\hat{\sigma}^2$ - residual s.d.

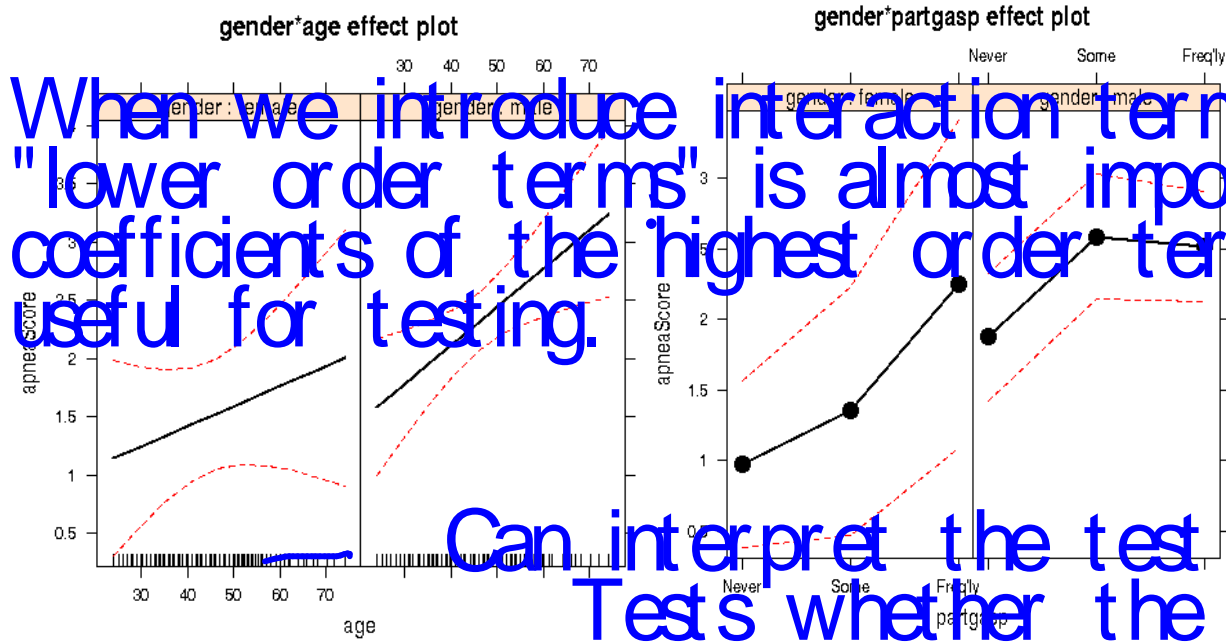
interaction generator

S.E. ($\hat{\beta}$)

Test $H_0: \beta = 0$

$$t = \hat{\beta} / \text{se}(\hat{\beta})$$

Conf. int. $\hat{\beta} \pm t \times \text{se}(\hat{\beta})$



When we introduce interaction terms into a model, the interpretation of "lower order terms" is almost impossible. In some sense, only the coefficients of the highest order terms have transparent interpretations, useful for testing.

Can interpret the test for gender:age parameter - Tests whether the slope is the same for males and females. in this case, P=.44, non-significant, so we have no evidence for a sex - age interaction.

Comparing models: assessing statistical significance

The F-test for nested models:

To test the sex x partner gasping report interaction, we need to test two coefficients simultaneously.

$$F = \frac{(Residual\ SS - Residual\ SS_f) / 2}{\hat{\sigma}_f^2}$$

need a modification of the overall F-test for significance.

Model 1: apneaScore ~ gender + age + bmi + partgasp
 Model 2: apneaScore ~ gender * (age + bmi + partgasp)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	144	247.903				
2	140	240.146	4	7.757	1.1305	0.3447

To implement this test, we compare the fits of two models, Full model = model with all the variables, Reduced model = model, with the variables to test omitted

F statistic is referred to F distribution, numerator df = # variables omitted, denominator = df. for error in full model.

This weeks office hour:
Thursday 11-12am
Office 308D

Re-review:
Estiamtion of

When we introduce interaction terms into a model, the interpretation of "lower order terms" is almost impossible. In some sense, only the coefficients of the highest order terms have transparent interpretations, useful for testing.

Can interpret the test for gender:age parameter -
Tests whether the slope is the same for males and females.
in this case, $P=.44$, non-significant, so we have no evidence for a sex - age interaction.

To test the sex x partner gasping report interaction,
we need to test two coefficients simultaneously.

need a modification of the overall F-test for significance.

Comparing models: assessing predictive utility

Model building - t-tests are very poor tool for model building (i.e. choosing which variables to use for prediction).

$$MSE \text{ for prediction} = E \left[\sum_{i=1}^n (y_i - \hat{y}_i)^2 \right]$$

Proper use of t-test is in confirmatory analysis, with a pre-specified hypothesis of interest.

eg. Is BMI a predictor of apneaScore,

$$\text{Mallow's } C_p \text{ statistic} = \frac{SSE_p}{\hat{\sigma}^2} + n - 2p$$

if we also adjust for age and sex, and partner report of gasping.

eg. common when we are exploring the potential effects of possible risk factors - epidemiological studies.

$$AIC = -2L + 2p = n \log(\hat{\sigma}_s^2) + 2p$$

Exploratory model building - different kettle of fish!!

For example:

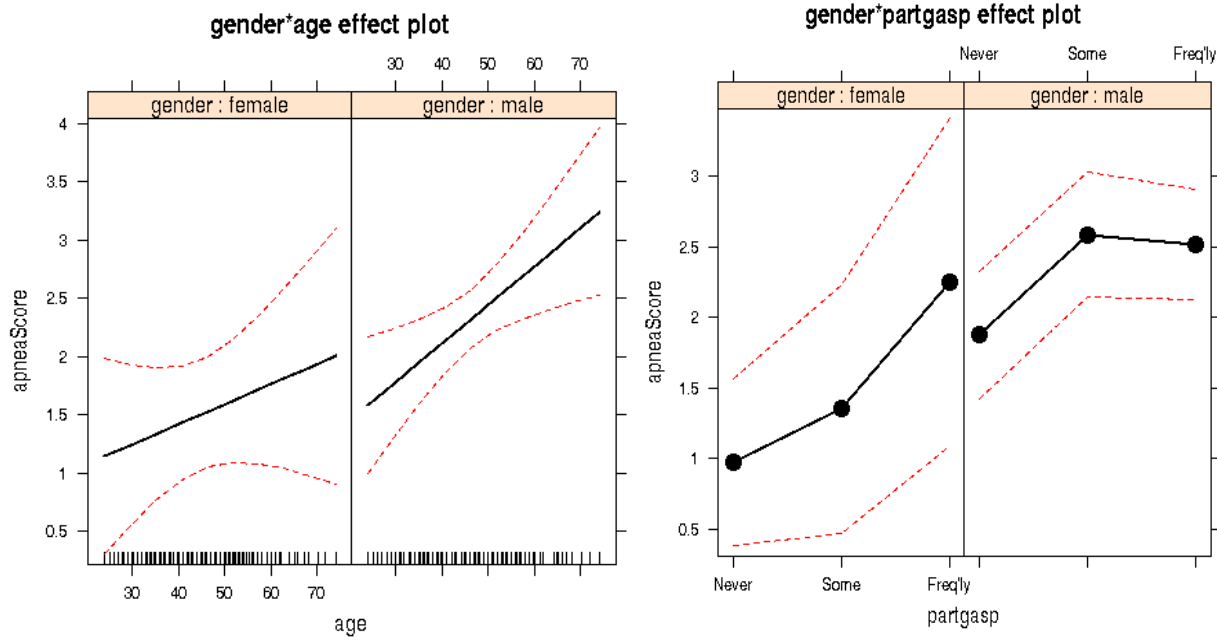
One problem of using t-tests, is the problem of multiple testing. - if you do enough t-tests, something will pop up as statistically significant just by chance.

```
lm(apneaScore ~ gender + age + bmi + partgasp)      df      AIC
lm(apneaScore ~ gender * (age + bmi + partgasp))  17    518.2734
```

One last detail: weighted least squares - the type I error rate or "alpha" level is only valid for a single test - when we we do multiple tests

Homogeneous variance assumption: $Var(y) = \sigma^2 I$ the effective type I error rate is approximately number of t-tests performed x "nominal" type I error rate
 Unequal variances, lack of independence: $Var(y) = \sigma^2 W$ = alpha that we've used for the individual tests

Weighted least squares estimate: $\hat{\beta} = (X^t W^{-1} X)^{-1} X^t W^{-1} y$



~~α~~ $\beta = 0$?

$$t = \hat{\beta} / se(\hat{\beta})$$

Comparing models: assessing statistical significance

The F-test for nested models:

$$F = \frac{(Residual\ SS_s - Residual\ SS_f) / 2}{\hat{\sigma}_f^2}$$

Model 1: apneaScore ~ gender + age + bmi + partgasp
 Model 2: apneaScore ~ gender * (age + bmi + partgasp)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	144	247.903				
2	140	240.146	4	7.757	1.1305	0.3447

Limitations

full model - only tests one β at a time
 - sensitive to effects of including other variables
 - collinearity
 - when strong correlations exist among X's

variables left out

→ anova

- see my commands on the web page

- only tests one β at a time
 - sensitive to effects of including other variables
 - collinearity
 - when strong correlations exist among X's

Comparing models: assessing predictive utility

F-tests are better model building tool than t-tests

1

$$MSE \text{ for prediction} = E\left\{\sum_{i=1}^n (\hat{y}_i - \mu_i)^2\right\}$$

$$\text{Mallow's } C_p \text{ statistic} = \frac{SSE_s}{\hat{\sigma}_f^2} - N + 2p$$

$$AIC = -2L + 2\nu = n \log(\tilde{\sigma}_s^2) + 2\nu$$

- results of stepwise variable selection procedures are not statistically reproducible.
- not a valid statistical procedure.

- to decide on a valid approach, firstly we need to have valid measures of the "performance" of a model.

For example:

→ simplified version 3.22

	df	AIC
<code>lm(apneaScore ~ gender + age + bmi + partgasp)</code>	7	515.0419
<code>lm(apneaScore ~ gender * (age + bmi + partgasp))</code>	11	518.2734

implemented in extractAIC

One last detail: weighted least squares

Homogeneous variance assumption: $Var(\mathbf{y}) = \sigma^2 \mathbf{I}$

Unequal variances, lack of independence: $Var(\mathbf{y}) = \sigma^2 \mathbf{W}$

Weighted least squares estimate: $\hat{\beta} = (X^t W^{-1} X)^{-1} X^t W^{-1} \mathbf{y}$

known AIC - builtin to main R

87.36 } 3.23
90.59 }

Announcements - update to assignment - do Q. 1 only
- correction on data management part - step 1
- please note - killer tornadoes defined by forceScale = 3, 4

Amendment to office hours: Tuesdays 10-11am
Thursdays 11am-noon
Room 308d

Review: Fitting a regression model - Y and some X's
- plot the data
- estimate the parameters
- check the model for fit
- interpret the results - !!!!!!!
- just looking at the beta hats and
P-values is not enough - often need to plot the fit.

- especially logistic regression - odds ratios
are not always really helpful in interpreting.

Model building!!! - have Y specified but may have many X's to
choose from

- choosing X's
- do we need to transform Y or X
- how about interactions?

t

Model building is very challenging, process must be informed by our scientific purpose and by background knowledge.

Model building tools -

- individual tests of coefficients - i.e. t-tests are not robust building tools - collinearity (interpretation of a t-test always depends on what other variables are in the model).
- multiplicity - if we do multiple t-tests, we lose the mathematical properties of type I error rate.
- can be argued that t-tests don't really answer the real question - why $P < .05$??

F-tests for model comparisons, help to alleviate some of the issues.

Today look at a different perspective in creating a tool.

