

Model Building

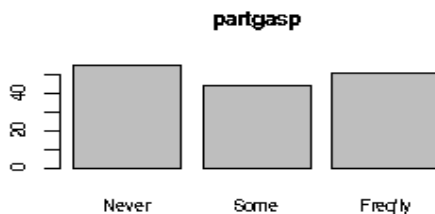
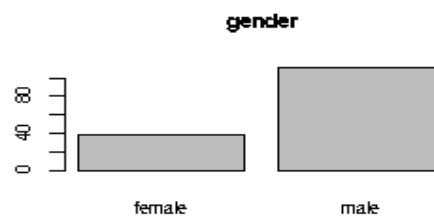
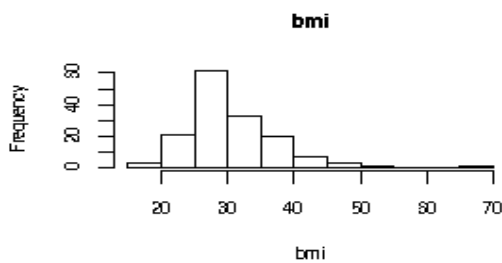
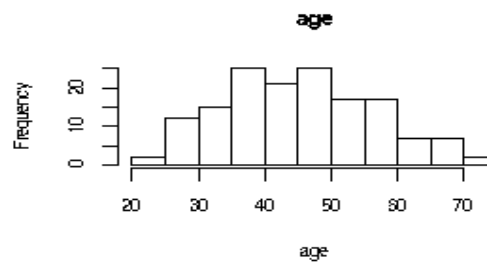
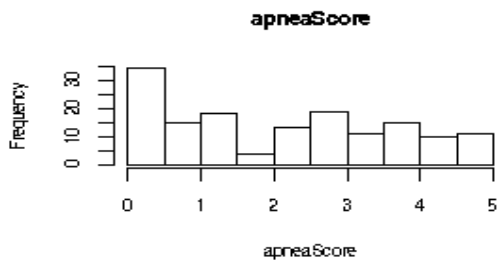
Example: Diagnosing sleep apnea

Response: `apneaScore` = score on overnight sleep test

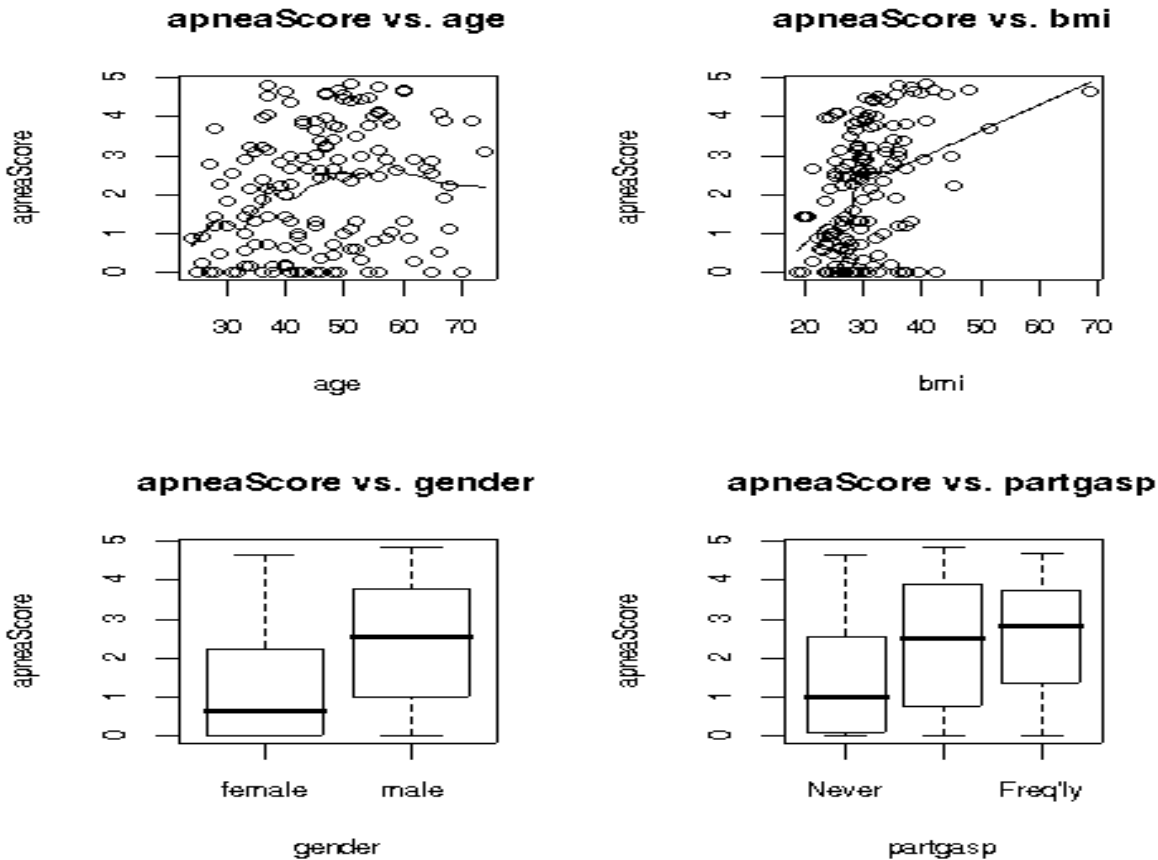
Possible predictor variables:

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		

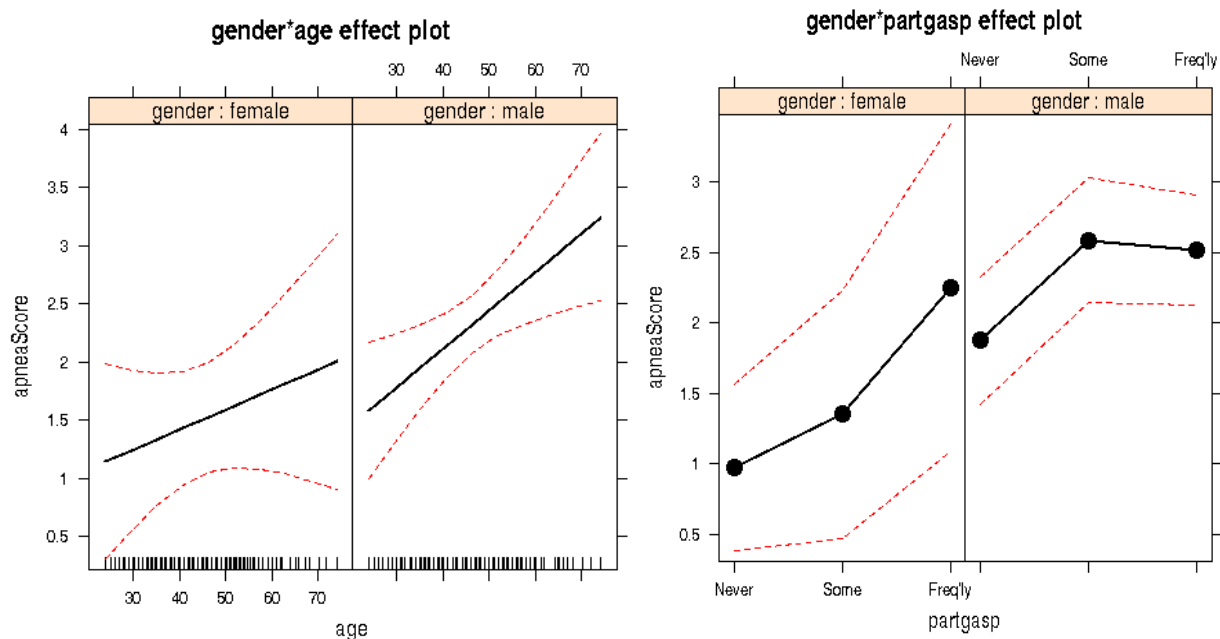
Model with Interactions with Gender

```
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
```



Comparing models: assessing statistical significance

The F-test for nested models:

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

$q =$ difference in number of parameters between full and reduced model

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

Comparing models: assessing predictive utility

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

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

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

For example:

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

One last detail: weighted least squares

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

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

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