Multiple Regression

The General Linear Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Making Interaction and Quadratic Terms in R

Make your own by multiplying (particularly for quadratic or cubic terms)

```
x1sq=x1*x1
lm(y~x1 + x1sq, data=mydata)
```

Let R do the interactions, just enter them as x1*x2 or x1:x2 in the formula call like this

```
lm(y~x1 + x2 + x1:x2, data=mydata)
```

The Hierarchical Approach to Model-Building

- Hierarchical
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_1 x_2 + \epsilon$
- NOT hierarchical
  - $y = \beta_0 + \beta_2 x_1^2 + \epsilon$
  - $y = \beta_0 + \beta_1 x_1 + \beta_k x_1 x_2 + \epsilon$

Example: Model with One Quantitative and one Categorical Predictor

Consider the HealthExam data set with the following definitions and linear model

$$
\begin{aligned}
x_1 &= 1 \text{ if AgeGroup is 36 to 64}, \quad && x_1 = 0 \text{ otherwise} \\
x_2 &= 1 \text{ if AgeGroup is 65+}, && x_2 = 0 \text{ otherwise} \\
x_3 &= \text{SysBP}
\end{aligned}
$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon$$

Example: How R Handles Categorical Predictors

The following R code would be used to fit the model shown previously

```
model<-lm(Cholesterol~AgeGroup+SysBP+AgeGroup:SysBP,data=HealthExam)
summary(model)
```

Example: The Output

```
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             523.980    502.232   1.043   0.3002
## AgeGroup36 to 64      -1399.606    676.160  -2.070   0.0419 *
## AgeGroup65+            -407.333    688.104  -0.592   0.5557
## SysBP                    -2.392      4.616  -0.518   0.6060
## AgeGroup36 to 64:SysBP   13.041      6.047   2.157   0.0343 *
## AgeGroup65+:SysBP         4.213      6.025   0.699   0.4866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Interpreting the Output

The least squares estimated regression line is

$$\widehat{y} = 524 - 1399.6x_1 - 407.3x_2 - 2.39x_3 + 13.04x_1x_3 + 4.21x_2x_3$$

Example: Estimated Models for Each Age Group

For AgeGroup 18 to 35, let $x_1 = 0$ and $x_2 = 0$, which gives

$$\widehat{y} = 524 - 2.39x_3$$

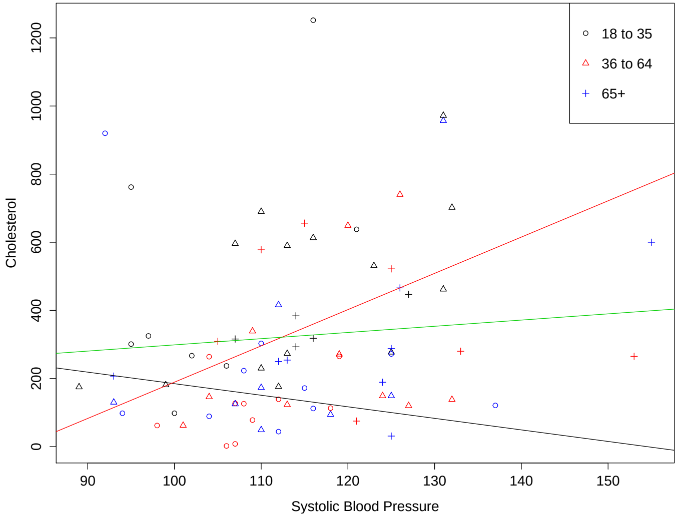
For AgeGroup 36 to 64, let $x_1 = 1$ and $x_2 = 0$, which gives

$$\widehat{y} = -875.6 + 10.65x_3$$

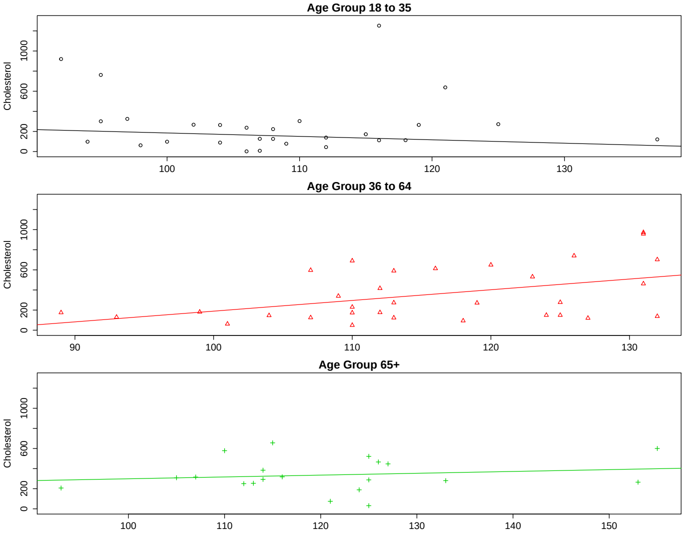
For AgeGroup 65+, let $x_1 = 0$ and $x_2 = 1$, which gives

$$\widehat{y} = 116.7 + 1.82x_3$$

Example: The Scatterplot



Example: Separate Plots and Lines for Each Age Group



Model Selection Criterion

- R^2 - higher is better
- R^2_{adj} - higher is better
- c_p - closer to the number of parameters (p) is better
- PRESS - lower is better
- AIC - lower is better
- BIC - lower is better

What's AIC?

Akaike Information Criterion

$$AIC = n \cdot \ln SSE - n \cdot \ln n + 2p$$

(or some variation on this formula)

AIC is a measure of information loss and is particularly useful for comparing models - lower is better.

Collinearity/Multicollinearity

- When two or more predictor variables are highly correlated in a linear regression model.
- Indicated by large values of the Variance Inflation Factor (VIF), such as 10 or more

The Effects of Collinearity

- Large standard error for estimated regression coefficients
- Higher p-values for tests of individual coefficients
- Wider confidence intervals for coefficients

What Collinearity Does NOT Affect

- F-statistics & p-values for the full model or subsets of coefficients
- R^2
- R^2_{adj}
- AIC
- Predicted values
- Standard errors of predicted values (these can be slightly affected)

Can anything be done to alleviate collinearity?

- Ignore it if it doesn't affect what you are doing with the regression model
- Combine correlated variables in a meaningful way to make a single variable
- Omit the predictor with the highest VIF
- Centering often removes collinearity for quadratic, cubic and interaction terms (centering is to subtract the mean from each data value for a given variable)
- Employ factor analysis to reduce the number of predictors (may be difficult to interpret results)