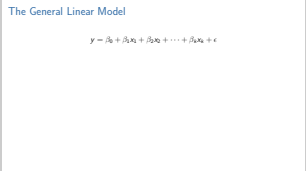
Multiple Regression

The General Linear Model

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



Let me introduce you to the general linear model. This is a very versatile and handy probabilistic model. On the left hand side we have the response variable. On the right we have a linear combination of independent variables or predictors.

The independent variables can be quantitative, categorical, or a mixture of both. For the multiple regression models in sections 12.1 through 12.7, as well chapter 13, we'll assume that the error term epsilon represents independent random variables that are normally distributed with mean 0 and a common variance. The parameters in the model are an intercept term, beta naught, and a coefficient for each independent variable called a partial slope. The partial slope tells us how much the average value of y would change - and in which direction - for each one unit increase in the variable that it is

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

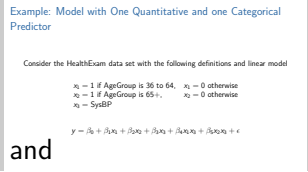
```
lm(y~x1 + x2 + x1:x2, data=mydata)
```

Example: Model with One Quantitative and one Categorical Predictor

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

$x_1 = 1$ if AgeGroup is 36 to 64, $x_1 = 0$ otherwise
$x_2 = 1$ if AgeGroup is 65+, $x_2 = 0$ otherwise
$x_3 =$ SysBP

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



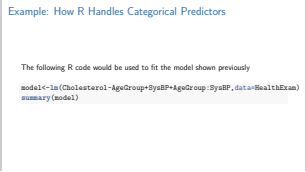
I'm sure you remember the HealthExam data set we've been using from time to time in this course.
While it isn't necessary to write statistical models with notation like x1, x2, x3 and so on, it can make it easier, so we'll relabel the 3 categories for Age Group using 2 dummy variables, x1 and x2 here. Dummy variables are also called indicator variables.

Consider the model with the categorical predictor Age Group, represented by the two indicator variables previously defined, the quantitative predictor variable Systolic Blood Pressure, and the interaction terms associated with these two variables to predict Cholesterol level.

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: 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)
```

One very nice feature in R is that we do NOT have to manually convert the categorical variable AgeGroup to the two dummy variables. We can simply enter the variable names from the data file and R will convert them to dummy variables and use them accordingly.

You'll see in the output on the next slide that R recoded the categorical variable (as we showed it in the previous slide) and there will be two first-order terms - which will correspond to what we called x1 and x2, and by simply specifying the interaction between the categorical predictor AgeGroup and the quantitative predictor Systolic Blood pressure, R automatically fits both second-order interaction terms.

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: The Output

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               523.980    502.252   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   0.0363 *
## AgeGroup65+:SysBP           4.213      6.025   0.699   0.4866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the R output here we see the estimated coefficients for the linear model. The estimated intercept, beta naught hat, is 523.98. What R has automatically labeled as "AgeGroup36 to 64" corresponds to what we called x1, the dummy variable indicating that a subject was in the 36 to 64 age group. The estimated coefficient for that dummy variable, beta 1 hat, is -1399.606.

What R has labeled as "AgeGroup65+" corresponds to what we called x2, the dummy variable indicating that a subject was in the 65 and over age group. The estimated coefficient for that dummy variable, beta 2 hat, is -407.333.

We called SysBP x3, and we see the estimate partial slope here of -2.392.

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: Interpreting the Output

The least squares estimated regression line is

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

Taking the estimated intercept and partial slopes from the R output, we put them in the regression equation to get the model. I took the liberty of rounding some of them.

Recall that x1 was the indicator for the 36 to 64-year-old AgeGroup, x2 was the indicator for the 65 and over Age Group, and x3 represented Systolic Blood Pressure. The regression equation could be written with the actual variable names in it, but it was rather long here and wouldn't have fit on one line.

This regression model is actually 3 simple linear regression models in one! There is a separate regression model for each category of the qualitative predictor variable AgeGroup. To get these regression models for each age group, simply the corresponding combination of

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: Estimated Models for Each Age Group

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

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

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

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

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

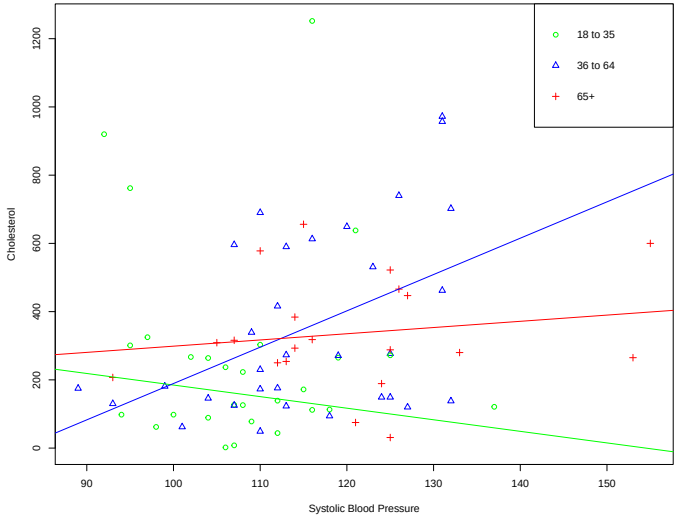
$$\hat{y} = 116.7 + 1.82x_3$$

The model for Age Group 18 to 35 is easy, when x1 and x2 are 0, any term containing either one drops out and we're left with the original intercept and the slope for Systolic Blood Pressure.

With x1=1 and x2=0 for the 36 to 64 Age Group, the intercept is 524 + -1399.6, which gives -875.6 and the slope is -2.39 + 13.04, which is 10.65.

When the AgeGroup is 65+, x1=0 and x2=1, so by substituting these into the equation, it simplifies to have an intercept of 524 + -407.3, which is where the 116.7 comes from, and a slope of -2.39 + 4.21, or 1.82.

Example: The Scatterplot



Example: The Scatterplot

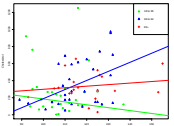
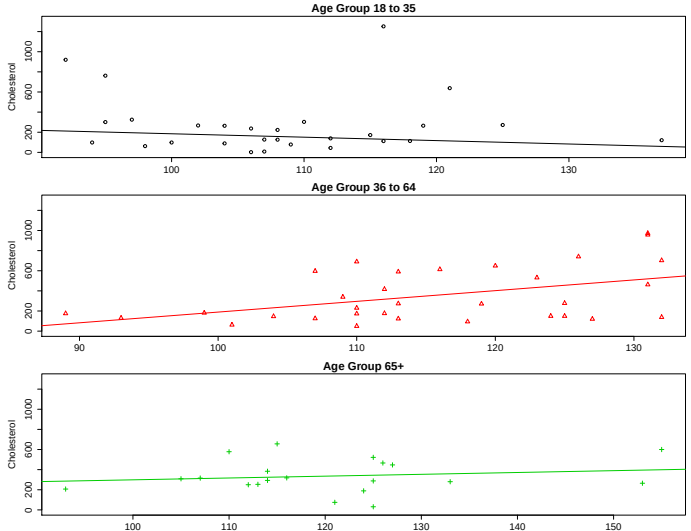


Figure 1:

The scatterplot of Cholesterol by Systolic Blood Pressure, using Age Group as a plotting symbol is shown here with each separate estimated regression model added to the plot. It looks pretty busy, but you can see that the lines have different y-intercepts and different slopes.

A difference in y-intercepts would be indicated by the statistical significance of the main effects (that is , the first order terms) for AgeGroup, and a difference of slopes would be indicated by the the statistical significance of the interaction effects between Age Group and Systolic Blood Pressure.

Example: Separate Plots and Lines for Each Age Group



Example: Separate Plots and Lines for Each Age Group

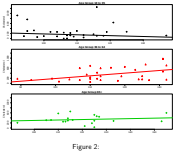


Figure 2:

Separating the plots makes it a bit easier to see how the resulting fitted line for each AgeGroup represents the relationship between Cholesterol and Systolic Blood Pressure for each Age Group.

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

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

Collinearity is likely to be common in the life of a data scientist because the data a data scientist will generally be working with is not from a designed experiment and often contains so many variables that high correlations among predictors are inevitable.

Highly correlated variables basically are providing overlapping explanations of the same variation in the response variable.

The Effects of Collinearity

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

The Effects of Collinearity

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

The effect of collinearity are primarily seen in the form of inflated standard error for the estimated regression coefficients, which in turn produces larger p-values for tests of individual coefficients and wider confidence intervals for coefficients.

A signature clue is when the F test for the full model has a very low p-value, but no individual coefficient has a small enough p-value to indicate that it differs from 0.

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)

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)

Much work can still be done with a regression model even in the presence of collinearity because there are a number of measures that are unaffected by it. Particularly if your aim is to use the model to estimate values the response variable and make predictions.

Bottom panel note: AIC will be covered in a future lesson.

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)