```
  Course: Regression_Models
  Lesson: MultiVar_Examples3

- Class: text
  Output: "MultiVar_Examples3. (Slides for this and other Data Science courses may be
  found at github https://github.com/DataScienceSpecialization/courses. If you care to
  use them, they must be downloaded as a zip file and viewed locally. This lesson
  corresponds to Regression_Models/02_02_multivariateExamples.)"

- Class: text
  Output: This is the third and final lesson in which we'll look at regression models
  with more than one independent variable or predictor. We'll begin with  WHO hunger
  data which we've taken the liberty to load for you. WHO is the World Health
  Organization and this data concerns young children from around the world and rates of
  hunger among them which the organization compiled over a number of years. The
  original csv file was very large and we've subsetted just the rows which identify the
  gender of the child as either male or female. We've read the data into the data frame
  "hunger" for you, so you can easily access it.

- Class: cmd_question
  Output: As we did in the last lesson let's first try to get a better understanding of
  the dataset. Use the R function dim to find the dimensions of hunger.
  CorrectAnswer: dim(hunger)
  AnswerTests: omnitest(correctExpr='dim(hunger)')
  Hint: Type "dim(hunger)" at the R prompt.

- Class: exact_question
  Output: How many samples does hunger have?
  CorrectAnswer: 948
  AnswerTests: omnitest(correctVal=948)
  Hint: The R function dim returns two numbers, the number of rows and the number of
  columns. Each row represents one sample.

- Class: cmd_question
  Output: Now use the R function names to find out what the 13 columns of hunger
  represent.
  CorrectAnswer: names(hunger)
  AnswerTests: omnitest(correctExpr='names(hunger)')
  Hint: Type "names(hunger)" at the R prompt.

- Class: text
  Output: The Numeric column for a particular row  tells us the percentage of children
  under age 5 who were underweight when that sample was taken. This is one of the
  columns we'll be focussing on in this lesson. It will be the outcome (dependent
  variable) for the models we generate.

- Class: cmd_question
  Output: Let's first look at the rate of hunger and see how it's changed over time.
  Use the R function lm to generate the linear model in which the rate of hunger,
  Numeric, depends on the predictor, Year. Put the result in the variable fit.
  CorrectAnswer: fit <- lm(hunger$Numeric ~ hunger$Year)
  AnswerTests: creates_lm_model('fit <- lm(hunger$Numeric ~ hunger$Year)')
  Hint: Remember you need to pass a formula, dependent ~ independent, to the model.
  Also, you may need to specify the data set if it isn't clear from the variables you
  enter in the formula. So type "fit <- lm(Numeric ~ Year, hunger)" at the R prompt or
  more simply fit <- lm(hunger$Numeric ~ hunger$Year)

- Class: cmd_question
  Output: Now look at the coef portion of the summary of fit.
  CorrectAnswer: summary(fit)$coef
  AnswerTests: omnitest(correctExpr='summary(fit)$coef')
  Hint: Type "summary(fit)$coef" at the R prompt.

- Class: mult_question
  Output: What is the coefficient of hunger$Year?
  AnswerChoices: -0.30840; 0.06053; 634.47966;  121.14460
  CorrectAnswer: -0.30840
  AnswerTests: omnitest(correctVal='-0.30840')
  Hint: Look at the hunger$Year row and Estimate column of the summary output.
```

```
- Class: mult_question
  Output: What does the negative Estimate of hunger$Year show?
  AnswerChoices: As time goes on, the rate of hunger decreases; As time goes on, the rate of hunger increases; I haven't a clue
  CorrectAnswer: As time goes on, the rate of hunger decreases
  AnswerTests: omnitest(correctVal='As time goes on, the rate of hunger decreases')
  Hint: Recall the meaning of the slope of a line. For every unit change in the independent variable (Year) there is a -.3084 change (decrease) in the dependent variable (percentage of hungry children).

- Class: mult_question
  Output: What does the intercept of the model represent?
  AnswerChoices: the percentage of hungry children at year 0; the number of hungry children at year 0; the number of children questioned in the survey
  CorrectAnswer: the percentage of hungry children at year 0
  AnswerTests: omnitest(correctVal='the percentage of hungry children at year 0')
  Hint: Numeric gives a percentage of hungry children, and an intercept is the point at which a line intersects the axis. The axis represents a 0 value.

- Class: cmd_question
  Output: Now let's use R's subsetting capability to look at the rates of hunger for the different genders to see how, or even if, they differ.  Once again use the R function lm to generate the linear model in which the rate of hunger (Numeric) for female children depends on Year. Put the result in the variable lmF. You'll have to use the R construct x[hunger$Sex=="Female"] to pick out both the correct Numerics and the correct Years.
  CorrectAnswer: lmF <- lm(hunger$Numeric[hunger$Sex=="Female"] ~ hunger$Year[hunger$Sex=="Female"])
  AnswerTests: creates_lm_model('lmF <- lm(hunger$Numeric[hunger$Sex=="Female"] ~ hunger$Year[hunger$Sex==\"Female\"])')
  Hint: Type lmF <- lm(hunger$Numeric[hunger$Sex=="Female"] ~ hunger$Year[hunger$Sex=="Female"]) at the R prompt or more simply lmF <- lm(Numeric[Sex=="Female"] ~ Year[Sex=="Female"],hunger)

- Class: cmd_question
  Output: Do the same for male children and put the result in lmM.
  CorrectAnswer: lmM <- lm(hunger$Numeric[hunger$Sex=="Male"] ~ hunger$Year[hunger$Sex=="Male"])
  AnswerTests: creates_lm_model('lmM <- lm(hunger$Numeric[hunger$Sex=="Male"] ~ hunger$Year[hunger$Sex=="Male"])')
  Hint: Type lmM <- lm(hunger$Numeric[hunger$Sex=="Male"] ~ hunger$Year[hunger$Sex=="Male"]) at the R prompt or more simply lmM <- lm(Numeric[Sex=="Male"] ~ Year[Sex=="Male"],hunger)

- Class: figure
  Output: Now we'll plot the data points and fitted lines using different colors to distinguish between males (blue) and females (pink).
  Figure: plot1.R
  FigureType: new

- Class: mult_question
  Output: We can see from the plot that the lines are not exactly parallel. On the right side of the graph (around the year 2010) they are closer together than on the left side (around 1970). Since they aren't parallel, their slopes must be different, though both are negative. Of the following R expressions which would confirm that the slope for males is negative?
  AnswerChoices: lmM$coef[2]; lmF$coef[2]; lmM$coef[1]
  CorrectAnswer:  lmM$coef[2]
  AnswerTests: omnitest(correctVal='lmM$coef[2]')
  Hint: First, eliminate the female choice since the question refers to males. Then recall that the first coefficient is the intercept of the line and the second is the slope.

- Class: text
  Output: Now instead of separating the data by subsetting the samples by gender we'll use gender as another predictor to create the linear model lmBoth. Recall that to do this in R we place a plus sign "+" between the independent variables, so the formula looks like dependent ~ independent1 + independent2.
```

```
90
91  - Class: cmd_question
92    Output:  Create  lmBoth now. Numeric is the dependent, Year and Sex are the
      independent variables. The data is "hunger". For lmBoth, make sure Year is first and
      Sex is second.
93    CorrectAnswer: lmBoth <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex)
94    AnswerTests: creates_lm_model('lmBoth <- lm(hunger$Numeric ~ hunger$Year +
      hunger$Sex)')
95    Hint: Type lmBoth <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex) or more simply
      lmBoth <- lm(Numeric ~ Year+Sex,hunger)
96
97  - Class: cmd_question
98    Output: Now look at the summary of lmBoth with the R command summary.
99    CorrectAnswer: summary(lmBoth)
100   AnswerTests: omnitest(correctExpr='summary(lmBoth)')
101   Hint: Type summary(lmBoth) at the R prompt.
102
103 - Class: text
104   Output: Notice that three estimates are given, the intercept, one for Year and one
      for Male. What happened to the estimate for Female? Note that Male and Female are
      categorical variables hence they are factors in this model. Recall from the last
      lesson (and slides) that R treats the first (alphabetical) factor as the reference
      and its estimate is the intercept which represents the percentage of hungry females
      at year 0. The estimate given for the factor Male is a distance from the intercept
      (the estimate of the reference group Female). To calculate the percentage of hungry
      males at year 0 you have to add together the intercept and the male estimate given by
      the model.
105
106 - Class: mult_question
107   Output: What percentage of young Males were hungry at year 0?
108   AnswerChoices: 635.431; 1.9027; 633.2199; I can't tell since the data starts at 1970.
109   CorrectAnswer: 635.431
110   AnswerTests: omnitest(correctVal='635.431')
111   Hint: The intercept is the percentage of females hungry at year 0 and the intercept
      plus hunger$SexMale is the percentage of males hungry at year 0.
112
113 - Class: mult_question
114   Output: What does the estimate for hunger$Year represent?
115   AnswerChoices: the annual decrease in percentage of hungry children of both genders;
      the annual decrease in percentage of hungry females;  the annual decrease in
      percentage of hungry males;
116   CorrectAnswer: the annual decrease in percentage of hungry children of both genders
117   AnswerTests: omnitest(correctVal='the annual decrease in percentage of hungry
      children of both genders')
118   Hint: The model looked at all the data and didn't specify which gender to consider.
119
120 - Class: figure
121   Output: Now we'll replot the data points along with two new lines using different
      colors.  The red line will have the female intercept and the  blue line will have the
      male intercept.
122   Figure: parallelplot.R
123   FigureType: new
124
125 - Class: mult_question
126   Output: The lines appear parallel. This is because
127   AnswerChoices: they have the same slope; they have slopes that are very close; I have
      no idea
128   CorrectAnswer: they have the same slope
129   AnswerTests: omnitest(correctVal='they have the same slope')
130   Hint: By definition parallel lines have the same slope.
131
132 - Class: text
133   Output: Now we'll consider the interaction between year and gender to see how that
      affects changes in rates of hunger. To do this we'll add a third term to the
      predictor portion of our model formula, the product of year and gender.
134
135 - Class: cmd_question
136   Output: Create the model lmInter. Numeric is the outcome and the three predictors are
      Year, Sex, and Sex*Year. The data is "hunger".
```

```
    CorrectAnswer: lmInter <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex + hunger$Year
    * hunger$Sex)
    AnswerTests: creates_lm_model('lmInter <- lm(hunger$Numeric ~ hunger$Year +
    hunger$Sex + hunger$Year * hunger$Sex)')
    Hint: Type lmInter <- lm(hunger$Numeric ~ hunger$Year + hunger$Sex + hunger$Year *
    hunger$Sex) or lmInter <- lm(Numeric ~ Year + Sex + Year*Sex, hunger)

- Class: cmd_question
  Output: Now look at the summary of lmInter with the R command summary.
  CorrectAnswer: summary(lmInter)
  AnswerTests: omnitest(correctExpr='summary(lmInter)')
  Hint: Type summary(lmInter) at the R prompt.

- Class: mult_question
  Output: What is the percentage of hungry females at year 0?
  AnswerChoices: 603.5058; 61.94772; The model doesn't say.
  CorrectAnswer: 603.5058
  AnswerTests: omnitest(correctVal='603.5058')
  Hint: As before, the intercept is the percentage of hunger for the first factor, in
    this case, females.

- Class: mult_question
  Output: What is the percentage of hungry males at year 0?
  AnswerChoices: 665.4535; 603.5058; 61.94772; The model doesn't say.
  CorrectAnswer: 665.4535
  AnswerTests: omnitest(correctVal='665.4535')
  Hint: As before, the estimate associated with SexMale is the distance from the
    intercept, so the intercept of the line associated with males is the intercept plus
    the estimate associated with males.

- Class: mult_question
  Output: What is the annual change in percentage of hungry females?
  AnswerChoices: -0.29340; -0.03000; 0.08547; The model doesn't say.
  CorrectAnswer: -0.29340
  AnswerTests: omnitest(correctVal='-0.29340')
  Hint: The estimate associated with Year represents the annual change in percent of
    hungry females.

- Class: mult_question
  Output: What is the annual change in percentage of hungry males?
  AnswerChoices: -0.32340; -0.03000; 0.12087; The model doesn't say.
  CorrectAnswer: -0.32340
  AnswerTests: omnitest(correctVal='-0.32340')
  Hint: The estimate associated with Year:SexMale represents the distance of the annual
    change in percent of males from that of females.

- Class: figure
  Output: Now we'll replot the data points along with two new lines using different
    colors to distinguish between the genders.
  Figure: interactplot.R
  FigureType: new

- Class: mult_question
  Output: Which line has the steeper slope?
  AnswerChoices: Male; Female; They look about the same
  CorrectAnswer: Male
  AnswerTests: omnitest(correctVal='Male')
  Hint: The lines are not parallel and will eventually intersect. The line that is
    further from horizontal (which has slope 0) has a steeper slope and indicates a
    faster rate of change. Which line has a slope further from 0?

- Class: text
  Output: Finally, we note that things are a little trickier when we're dealing with an
    interaction between predictors which are continuous (and not factors). The slides
    show the underlying algebra, but we can summarize.

- Class: text
  Output: Suppose we have two interacting predictors and one of them is held constant.
    The expected change in the outcome for a unit change in the other predictor is the
```

```
    coefficient of that changing predictor  + the coefficient of the interaction * the
    value of the predictor held constant.
192
193 - Class: text
194   Output: Suppose the linear model is Hi = b0 + (b1*Ii) + (b2*Yi)+ (b3*Ii*Yi) + ei.
      Here the H's represent the outcomes, the I's and Y's the predictors, neither of which
      is a category, and the b's represent the estimated coefficients of the predictors. We
      can ignore the e's which represent the residuals of the model. This equation models a
      continuous interaction since neither I nor Y is a category or factor. Suppose we fix
      I at some value and let Y vary.
195
196 - Class: mult_question
197   Output: Which expression represents the change in H per unit change in Y given that I
      is fixed at 5?
198   AnswerChoices: b2+b3*5; b1+5*b3; b0+b2; b2+b3*Y
199   CorrectAnswer: b2+b3*5
200   AnswerTests: omnitest(correctVal='b2+b3*5')
201   Hint: The expected change in the outcome is the estimate of the changing predictor
      (Y) + the estimate of the interaction  (b3) * the value of the predictor held
      constant (5).
202
203 - Class: text
204   Output: Congratulations! You've finished this final lesson in multivariable
      regression models.
205
206 - Class: mult_question
207   Output: "Would you like to receive credit for completing this course on
208     Coursera.org?"
209   CorrectAnswer: NULL
210   AnswerChoices: Yes;No
211   AnswerTests: coursera_on_demand()
212   Hint: ""
213
```