```
  Course: Regression Models
  Lesson: Variance_Inflation_Factors


- Class: text
  Output: "Variance Inflation Factors. (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_04_residuals_variation_diagnostics.)"

- Class: text
  Output: "In modeling, our interest lies in parsimonious, interpretable
  representations of the data that enhance our understanding of the phenomena under
  study. Omitting variables results in bias in the coefficients of interest - unless
  their regressors are uncorrelated with the omitted ones. On the other hand, including
  any new variables increases (actual, not estimated) standard errors of other
  regressors. So we don't want to idly throw variables into the model. This lesson is
  about the second of these two issues, which is known as variance inflation."

- Class: text
  Output: "We shall use simulations to illustrate variance inflation. The source code
  for these simulations is in a file named vifSims.R which I have copied into your
  working directory and tried to display in your source code editor. If I've failed to
  display it, you should open it manually."

- Class: mult_question
  Output: "Find the function, makelms, at the top of vifSims.R. The final expression in
  makelms creates 3 linear models. The first, lm(y ~ x1), predicts y in terms of x1,
  the second predicts y in terms of x1 and x2, the third in terms of all three
  regressors. The second coefficient of each model, for instance coef(lm(y ~ x1))[2],
  is extracted and returned in a 3-long vector. What does this second coefficient
  represent?"
  AnswerChoices: The coefficient of x1.; The coefficient of the intercept.; The
  coefficient of x2.
  CorrectAnswer: The coefficient of x1.
  AnswerTests: omnitest(correctVal= 'The coefficient of x1.')
  Hint: "The first coefficient is that of the intercept. The rest are in the order
  given by the formula."

- Class: mult_question
  Output: "In makelms, the simulated dependent variable, y, depends on which of the
  regressors?"
  AnswerChoices: x1;x1 and x2;x1, x2, and x3;
  CorrectAnswer: x1
  AnswerTests: omnitest(correctVal= 'x1')
  Hint: The dependent variable, y, is formed by the expression, y <- x1 +
  rnorm(length(x1), sd = .3). Which of the regressors, x1, x2, x3, appears in this
  expression?

- Class: mult_question
  Output: "In vifSims.R, find the functions, rgp1() and rgp2(). Both functions generate
  3 regressors, x1, x2, and x3. Compare the lines following the comment Point A in
  rgp1() with those following Point C in rgp2(). Which of the following statements
  about x1, x2, and x3 is true?"
  AnswerChoices: x1, x2, and x3 are uncorrelated in rgp1(), but not in rgp2().;x1, x2,
  and x3 are correlated in rgp1(), but not in rgp2().;x1, x2, and x3 are uncorrelated
  in both rgp1() and rgp2().;x1, x2, and x3 are correlated in both rgp1() and rgp2().
  CorrectAnswer: x1, x2, and x3 are uncorrelated in rgp1(), but not in rgp2().
  AnswerTests: omnitest(correctVal= 'x1, x2, and x3 are uncorrelated in rgp1(), but not
  in rgp2().')
  Hint: "In rgp2(), in the lines following Point C, x1 appears in the expressions which
  form x2 and x3. In rgp1(), in the lines following Point A, the regressors are formed
  by independent calls to rnorm(), which simulates independent, identically distributed
  samples from a normal distribution."

- Class: mult_question
  Output: "In the line following Point B in rgp1(), the function maklms(x1, x2, x3) is
  applied 1000 times. Each time it is applied, it simulates a new dependent variable,
  y, and returns estimates of the coefficient of x1 for each of the 3 models, y ~ x1, y
```

```
    ~ x1 + x2, and y ~ x1 + x2 + x3. It thus computes 1000 estimates of the  3
    coefficients, collecting the results in 3x1000 array, beta. In the next line, the
    expression, apply(betas, 1, var), does which of the following?"
37    AnswerChoices: Computes the variance of each row.;Computes the variance of each column.
38    CorrectAnswer: Computes the variance of each row.
39    AnswerTests: omnitest(correctVal= 'Computes the variance of each row.')
40    Hint: "We hope to illustrate the effect of extra variables on the variance of x1's
    coefficient. For this purpose we have 3 models, y ~ x1, y ~ x1 + x2, and y ~ x1 + x2
    + x3. The three rows of beta correspond to the three models. The columns correspond
    to the 1000 simulated situations in which we estimate the coefficients of x1 for each
    of the three models. We are interested in the variance of the x1 coefficient for each
    of those three models."
41
42  - Class: cmd_question
43    Output: "The function rgp1() computes the variance in estimates of the coefficient of
    x1 in each of the three models, y ~ x1, y ~ x1 + x2, and y ~ x1 + x2 + x3. (The
    results are rounded to 5 decimal places for convenient viewing.) This simulation
    approximates the variance (i.e., squared standard error) of x1's coefficient in each
    of these three models. Recall that variance inflation is due to correlated regressors
    and that in rgp1() the regressors are uncorrelated. Run the simulation rgp1() now. Be
    patient. It takes a while."
44    CorrectAnswer: rgp1()
45    AnswerTests: omnitest(correctExpr='rgp1()')
46    Hint: Just enter rgp1() at the R prompt.
47
48  - Class: mult_question
49    Output: "The variances in each of the three models are approximately equal, as
    expected, since the other regressors, x2 and x3, are uncorrelated with the regressor
    of interest, x1. However, in rgp2(), x2 and x3 both depend on x1, so we should expect
    an effect. From the expressions assigning x2 and x3 which follow Point C, which is
    more strongly correlated with x1?"
50    AnswerChoices: x3;x2
51    CorrectAnswer: x3
52    AnswerTests: omnitest(correctVal= 'x3')
53    Hint: "In vifSims.R, look at the lines following Point C again, and note that
    1/sqrt(2) in the expression for x2 is much smaller than 0.95 in the expression for x3."
54
55  - Class: cmd_question
56    Output: "Run rgp2() to simulate standard errors in the coefficient of x1 for cases in
    which x1 is correlated with the other regressors"
57    CorrectAnswer: rgp2()
58    AnswerTests: omnitest(correctExpr='rgp2()')
59    Hint: Just enter rgp2() at the R prompt.
60
61  - Class: text
62    Output: "In this case, variance inflation due to correlated regressors is clear, and
    is most pronounced in the third model, y ~ x1 + x2 + x3, since x3 is the regressor
    most strongly correlated with x1."
63
64  - Class: text
65    Output: "In these two simulations we had 1000 samples of estimated coefficients,
    hence could calculate sample variance in order to illustrate the effect. In a real
    case, we have only one set of coefficients and we depend on theoretical estimates.
    However, theoretical estimates contain an unknown constant of proportionality. We
    therefore depend on ratios of theoretical estimates called Variance Inflation
    Factors, or VIFs."
66
67  - Class: text
68    Output: "A variance inflation factor (VIF) is a ratio of estimated variances, the
    variance due to including the ith regressor, divided by that due to including a
    corresponding ideal regressor which is uncorrelated with the others. VIF's can be
    calculated directly, but the car package provides a convenient method for the purpose
    as we will illustrate using the Swiss data from the datasets package."
69
70  - Class: cmd_question
71    Output: "According to its documentation, the Swiss data set consists of a
    standardized fertility measure and socioeconomic indicators for each of 47
    French-speaking provinces of Switzerland in about 1888 when Swiss fertility rates
    began to fall. Type head(swiss) or View(swiss) to examine the data."
```

```
72      CorrectAnswer: head(swiss)
73      AnswerTests: ANY_of_exprs('head(swiss)', 'View(swiss)')
74      Hint: Enter either head(swiss) or View(swiss) at the R prompt.
75
76    - Class: cmd_question
77      Output: "Fertility was thought to depend on five socioeconomic factors: the percent
        of males working in Agriculture, the percent of draftees receiving the highest grade
        on the army's Examination, the percent of draftees with Education beyond primary
        school, the percent of the population which was Roman Catholic, and the rate of
        Infant Mortality in the province. Use linear regression to model Fertility in terms
        of these five regressors and an intercept. Store the model in a variable named mdl."
78      CorrectAnswer: mdl <- lm(Fertility ~ ., swiss)
79      AnswerTests: creates_lm_model('mdl <- lm(Fertility ~ ., swiss)')
80      Hint: "Entering mdl <- lm(Fertility ~ ., swiss) is the easiest way to model Fertility
        as a function of all five regressors. The dot after the ~ means to include all
        (including an intercept.)"
81
82    - Class: cmd_question
83      Output: "Calculate the VIF's for each of the regressors using vif(mdl)."
84      CorrectAnswer: vif(mdl)
85      AnswerTests: omnitest('vif(mdl)')
86      Hint: "Just enter vif(mdl) at the R prompt."
87
88    - Class: text
89      Output: "These VIF's show, for each regression coefficient, the variance inflation
        due to including all the others. For instance, the variance in the estimated
        coefficient of Education is 2.774943 times what it might have been if Education were
        not correlated with the other regressors. Since Education and score on an Examination
        are likely to be correlated, we might guess that most of the variance inflation for
        Education is due to including Examination."
90
91    - Class: cmd_question
92      Output: "Make a second linear model of Fertility in which Examination is omitted, but
        the other four regressors are included. Store the result in a variable named mdl2."
93      CorrectAnswer: mdl2 <- lm(Fertility ~ . -Examination, swiss)
94      AnswerTests: creates_lm_model('mdl2 <- lm(Fertility ~ . -Examination, swiss)')
95      Hint: "Entering mdl2 <- lm(Fertility ~ . -Examination, swiss) is the easiest way to
        model Fertility as a function of all the regressors except Examination. The dot after
        ~ means all, and the minus sign in front of Examination means except."
96
97    - Class: cmd_question
98      Output: "Calculate the VIF's for this model using vif(mdl2)."
99      CorrectAnswer: vif(mdl2)
100     AnswerTests: omnitest(correctExpr='vif(mdl2)')
101     Hint: Just enter vif(mdl2) at the R prompt.
102
103   - Class: text
104     Output: "As expected, omitting Examination has markedly decreased the VIF for
        Education, from 2.774943 to 1.816361. Note that omitting Examination has had almost
        no effect the VIF for Infant Mortality. Chances are Examination and Infant Mortality
        are not strongly correlated. Now, before finishing this lesson, let's review several
        significant points."
105
106   - Class: mult_question
107     Output: "A VIF describes the increase in the variance of a coefficient due to the
        correlation of its regressor with the other regressors. What is the relationship of a
        VIF to the standard error of its coefficient?"
108     AnswerChoices: "VIF is the square of standard error inflation.;They are the
        same.;There is no relationship."
109     CorrectAnswer: VIF is the square of standard error inflation.
110     AnswerTests: omnitest(correctVal= 'VIF is the square of standard error inflation.')
111     Hint: "Variance is the square of standard deviation, and standard error is the
        standard deviation of an estimated coefficient."
112
113   - Class: mult_question
114     Output: "If a regressor is strongly correlated with others, hence will increase their
        VIF's, why shouldn't we just exclude it?"
115     AnswerChoices: "Excluding it might bias coefficient estimates of regressors with
        which it is correlated.;We should always exclude it.;We should never exclude anything."
```

```
    CorrectAnswer: Excluding it might bias coefficient estimates of regressors with which
    it is correlated.
    AnswerTests: omnitest(correctVal= 'Excluding it might bias coefficient estimates of
    regressors with which it is correlated.')
    Hint: "Excluding a regressor can bias estimates of coefficients for correlated
    regressors."

  - Class: mult_question
    Output: "The problems of variance inflation and bias due to excluded regressors both
    involve correlated regressors. However there are methods, such as factor analysis or
    principal componenent analysis, which can convert regressors to an equivalent
    uncorrelated set. Why then, when modeling, should we not just use uncorrelated
    regressors and avoid all the trouble?"
    AnswerChoices: "Using converted regressors may make interpretation difficult.; Factor
    analysis takes too much computation.; We should always use uncorrelated regressors."
    CorrectAnswer: Using converted regressors may make interpretation difficult.
    AnswerTests: omnitest(correctVal= 'Using converted regressors may make interpretation
    difficult.')
    Hint: "In modeling, our interest lies in parsimonious, interpretable representations
    of the data that enhance our understanding of the phenomena under study."

  - Class: text
    Output: That completes the exercise in variance inflation. The issue of omitting
    regressors is discussed in another lesson.

  - Class: mult_question
    Output: "Would you like to receive credit for completing this course on
      Coursera.org?"
    CorrectAnswer: NULL
    AnswerChoices: Yes;No
    AnswerTests: coursera_on_demand()
    Hint: ""
```