```
Course: Regression Models
Lesson: Overfitting and Underfitting


- Class: text
  Output: "Overfitting and Underfitting. (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: "The Variance Inflation Factors lesson demonstrated that including new
  variables will increase standard errors of coefficient estimates of other, correlated
  regressors. Hence, we don't want to idly throw variables into the model. On the other
  hand, omitting variables results in bias in coefficients of regressors which are
  correlated with the omitted ones. In this lesson we demonstrate the effect of omitted
  variables and discuss the use of ANOVA to construct parsimonious, interpretable
  representations of the data."

- Class: text
  Output: "First, I would like to illustrate how omitting a correlated regressor can
  bias estimates of a coefficient. The relevant source code is in a file named
  fitting.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 simbias() at the top of fitting.R. Below the comment
  labeled Point A three regressors, x1, x2, and x3, are defined. Which of these two are
  correlated?"
  AnswerChoices: "x1 and x2;x1 and x3;x2 and x3"
  CorrectAnswer: "x1 and x2"
  AnswerTests: omnitest(correctVal= 'x1 and x2')
  Hint: The variable, temp, is involved in forming both x1 and x2.

- Class: mult_question
  Output: "Within simbias() another function, f(n), is defined. It forms a dependent
  variable, y, and at Point C returns the coefficient of x1 as estimated by two models,
  y ~ x1 + x2, and y ~ x1 + x3. One regressor is missing in each model. In the
  expression for y (Point B,) what is the actual coefficient of x1?"
  AnswerChoices: 1;0.3;1/sqrt(2)
  CorrectAnswer: 1
  AnswerTests: omnitest(correctVal= '1')
  Hint: "What is the coefficient of x1 in the sum, x1 + x2 + x3?"

- Class: cmd_question
  Output: "At Point D in simbias() the internal function, f(), is applied 150 times and
  the results returned as a 2x150 matrix. The first row of this matrix contains
  independent estimates of x1's coefficient in the case that x3, the regressor
  uncorrelated with x1, is omitted. The second row contains estimates of x1's
  coefficient when the correlated regressor, x2, is omitted. Use simbias(), accepting
  the default argument, to form these estimates and store the result in a variable
  called x1c. (The default argument just guarantees a nice histogram, in a figure to
  follow.)"
  CorrectAnswer: "x1c <- simbias()"
  AnswerTests: omnitest(correctExpr='x1c <- simbias()')
  Hint: Just enter x1c <- simbias() at the R prompt.

- Class: cmd_question
  Output: "The actual coefficient of x1 is 1. Having been warned that omitting a
  correlated regressor would bias estimates of x1's coefficient, we would expect the
  mean estimate of x1c's second row to be farther from 1 than the mean of x1c's first
  row. Using apply(x1c, 1, mean), find the means of each row."
  CorrectAnswer: apply(x1c, 1, mean)
  AnswerTests: omnitest(correctExpr='apply(x1c, 1, mean)')
  Hint: Enter apply(x1c, 1, mean) at the R prompt.

- Class: figure
  Output: "Histograms of estimates from x1c's first row (blue) and second row (red) are
  shown. Estimates from the second row are clearly more than two standard deviations
```

```
    from the correct value of 1, and the bias due to omitting the correlated regressor is
    evident. (The code which produced this figure is incidental to the lesson, but is
    available as the function x1hist(), at the bottom of fitting.R.)"
    Figure: histograms.R
    FigureType: new

  - Class: figure
    Output: "Adding even irrelevant regressors can cause a model to tend toward a perfect
    fit. We illustrate this by adding random regressors to the swiss data and regressing
    on progressively more of them. As the number of regressors approaches the number of
    data points (47), the residual sum of squares, also known as the deviance, approaches
    0. (The source code for this figure can be found as function bogus() in fitting.R."
    Figure: bogus.R
    FigureType: new

  - Class: text
    Output: "In the figure, adding random regressors decreased deviance, but we would be
    mistaken to believe that such decreases are significant. To assess significance, we
    should take into account that adding regressors reduces residual degrees of freedom.
    Analysis of variance (ANOVA) is a useful way to quantify the significance of
    additional regressors. To exemplify its use, we will use the swiss data."

  - Class: cmd_question
    Output: "Recall that the Swiss data set consists of a standardized fertility measure
    and socioeconomic indicators for each of 47 French-speaking provinces of Switzerland
    in 1888. Fertility was thought to depend on an intercept and five factors denoted as
    Agriculture, Examination, Education, Catholic, and Infant Mortality. To begin our
    ANOVA example, regress Fertility on Agriculture and store the result in a variable
    named fit1."
    CorrectAnswer: fit1 <- lm(Fertility ~ Agriculture, swiss)
    AnswerTests: creates_lm_model('fit1 <- lm(Fertility ~ Agriculture, swiss)')
    Hint: Enter fit1 <- lm(Fertility ~ Agriculture, swiss) or something equivalent at the
    R prompt.

  - Class: cmd_question
    Output: "Create another model, named fit3, by regressing Fertility on Agriculture and
    two additonal regressors, Examination and Education."
    CorrectAnswer: fit3 <- lm(Fertility ~ Agriculture + Examination + Education, swiss)
    AnswerTests: creates_lm_model('fit3 <- lm(Fertility ~ Agriculture + Examination +
    Education, swiss)')
    Hint: "Enter fit3 <- lm(Fertility ~ Agriculture + Examination + Education, swiss) or
    something equivalent at the R prompt."

  - Class: cmd_question
    Output: "We'll now use anova to assess the significance of the two added regressors.
    The null hypothesis is that the added regressors are not significant. We'll explain
    in detail shortly, but right now just apply the significance test by entering
    anova(fit1, fit3)."
    CorrectAnswer: anova(fit1, fit3)
    AnswerTests: omnitest(correctExpr='anova(fit1, fit3)')
    Hint: Enter anova(fit1, fit3) at the R prompt.

  - Class: mult_question
    Output: "The three asterisks, ***, at the lower right of the printed table indicate
    that the null hypothesis is rejected at the 0.001 level, so at least one of the two
    additional regressors is significant. Rejection is based on a right-tailed F test,
    Pr(>F), applied to an F value. According to the table, what is that F value?"
    AnswerChoices: 20.968;3102.2;45
    CorrectAnswer: 20.968
    AnswerTests: omnitest(correctVal= '20.968')
    Hint: "It's the only number in the column labeled F in the printed table."

  - Class: mult_question
    Output: "An F statistic is a ratio of two sums of squares divided by their respective
    degrees of freedom. If the two scaled sums are independent and centrally chi-squared
    distributed with the same variance, the statistic will have an F distribution with
    parameters given by the two degrees of freedom. In our case, the two sums are
    residual sums of squares which, as we know, have mean zero hence are centrally
    chi-squared provided the residuals themselves are normally distributed. The two
```

```
    relevant sums are given in the RSS (Residual Sum of Squares) column of the table.
    What are they?"
80    AnswerChoices: 6283.1 and 3180.9;2 and 3102.2;45 and 43
81    CorrectAnswer: 6283.1 and 3180.9
82    AnswerTests: omnitest(correctVal= '6283.1 and 3180.9')
83    Hint: "The two numbers are under the RSS label in the table printed by anova(fit1,
      fit3)."
84
85  - Class: cmd_question
86    Output: "R's function, deviance(model), calculates the residual sum of squares, also
      known as the deviance, of the linear model given as its argument. Using
      deviance(fit3), verify that 3180.9 is fit3's residual sum of squares. (Of course,
      fit3 is called Model 2 in the table.)"
87    CorrectAnswer: deviance(fit3)
88    AnswerTests: omnitest(correctExpr='deviance(fit3)')
89    Hint: "Enter deviance(fit3) at the R prompt."
90
91  - Class: cmd_question
92    Output: "In the next several steps, we will show how to calculate the F value,
      20.968, which appears in the table printed by anova(). We'll begin with the
      denominator, which is fit3's residual sum of squares divided by its degrees of
      freedom. Fit3 has 43 residual degrees of freedom. This figure is obtained by
      subtracting 4, the the number of fit3's predictors (the 3 named and the intercept,)
      from 47, the number of samples in swiss. Store the value of deviance(fit3)/43 in a
      variable named d."
93    CorrectAnswer: d <- deviance(fit3)/43
94    AnswerTests: "ANY_of_exprs('d <- deviance(fit3)/43', 'd <-
      deviance(fit3)/df.residual(fit3)', 'd <- deviance(fit3)/fit3$df.residual')"
95    Hint: "Enter d <- deviance(fit3)/43 at the R prompt."
96
97  - Class: cmd_question
98    Output: "The numerator is the difference, deviance(fit1)-deviance(fit3), divided by
      the difference in the residual degrees of freedom of fit1 and fit3, namely 2. This
      calculation requires some theoretical justification which we omit, but the essential
      idea is that fit3 has 2 predictors in addition to those of fit1. Calculate the
      numerator and store it in a variable named n."
99    CorrectAnswer: n <- (deviance(fit1) - deviance(fit3))/2
100   AnswerTests: "ANY_of_exprs('n <- (deviance(fit1) - deviance(fit3))/2', 'n <-
      (deviance(fit1) - deviance(fit3))/(45-43)', 'n <- (deviance(fit1) -
      deviance(fit3))/(df.residual(fit1)-df.residual(fit3))', 'n <- (deviance(fit1) -
      deviance(fit3))/(fit1$df.residual - fit3$df.residual)')"
101   Hint: "Enter n <- (deviance(fit1) - deviance(fit3))/2 at the R prompt."
102
103 - Class: cmd_question
104   Output: "Calculate the ratio, n/d, to show it is essentially equal to the F value,
      20.968, given by anova()."
105   CorrectAnswer: n/d
106   AnswerTests: omnitest(correctExpr='n/d')
107   Hint: Just enter n/d at the R prompt.
108
109 - Class: cmd_question
110   Output: "We'll now calculate the p-value, which is the probability that a value of
      n/d or larger would be drawn from an F distribution which has parameters 2 and 43.
      This value was given as 4.407e-07 in the column labeled Pr(>F) in the table printed
      by anova(), a very unlikely value if the null hypothesis were true. Calculate this
      p-value using pf(n/d, 2, 43, lower.tail=FALSE)."
111   CorrectAnswer: pf(n/d, 2, 43, lower.tail=FALSE)
112   AnswerTests: omnitest(correctExpr='pf(n/d, 2, 43, lower.tail=FALSE)')
113   Hint: Just enter pf(n/d, 2, 43, lower.tail=FALSE) at the R prompt.
114
115 - Class: cmd_question
116   Output: "Based on the calculated p-value, a false rejection of the null hypothesis is
      extremely unlikely. We are confident that fit3 is significantly better than fit1,
      with one caveat: analysis of variance is sensitive to its assumption that model
      residuals are approximately normal. If they are not, we could get a small p-value for
      that reason. It is thus worth testing residuals for normality. The Shapiro-Wilk test
      is quick and easy in R. Normality is its null hypothesis. Use
      shapiro.test(fit3$residuals) to test the residual of fit3."
117   CorrectAnswer: shapiro.test(fit3$residuals)
```

```
    AnswerTests: ANY_of_exprs('shapiro.test(fit3$residuals)',
    'shapiro.test(residuals(fit3))')
    Hint: Enter shapiro.test(fit3$residuals) at the R prompt.

  - Class: cmd_question
    Output: "The Shapiro-Wilk p-value of 0.336 fails to reject normality, supporting
    confidence in our analysis of variance. In order to illustrate the use of anova()
    with more than two models, I have constructed fit5 and fit6 using the first 5 and all
    6 regressors (including the intercept) respectively. Thus fit1, fit3, fit5, and fit6
    form a nested sequence of models; the regressors of one are included in those of the
    next. Enter anova(fit1, fit3, fit5, fit6) at the R prompt now to get the flavor."
    CorrectAnswer: anova(fit1, fit3, fit5, fit6)
    AnswerTests: omnitest(correctExpr='anova(fit1, fit3, fit5, fit6)')
    Hint: Enter anova(fit1, fit3, fit5, fit6) at the R prompt.

  - Class: text
    Output: "It appears that each model is a significant improvement on its predecessor.
    Before ending the lesson, let's review a few salient points."

  - Class: mult_question
    Output: "Omitting a regressor can bias estimation of the coefficient of certain other
    regressors. Which ones?"
    AnswerChoices: Correlated regressors;Uncorrelated regressors
    CorrectAnswer: Correlated regressors
    AnswerTests: omnitest(correctVal= 'Correlated regressors')
    Hint: The other one.

  - Class: mult_question
    Output: "Including more regressors will reduce a model's residual sum of squares,
    even if the new regressors are irrelevant. True or False?"
    AnswerChoices: True;False;It depends on circumstances.
    CorrectAnswer: True
    AnswerTests: omnitest(correctVal= 'True')
    Hint: It doesn't depend on circumstances.

  - Class: mult_question
    Output: "When adding regressors, the reduction in residual sums of squares should be
    tested for significance above and beyond that of reducing residual degrees of
    freedom. R's anova() function uses an F-test for this purpose. What else should be
    done to insure that anova() applies?"
    AnswerChoices: "Model residuals should be tested for normality.;Regressors should be
    tested for normality.;The residuals should be tested for having zero means."
    CorrectAnswer: Model residuals should be tested for normality.
    AnswerTests: omnitest(correctVal= 'Model residuals should be tested for normality.')
    Hint: F-tests are sensitive to the assumption of normality.

  - Class: text
    Output: "That completes the lesson on underfitting and overfitting."

  - 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: ""
```