We have seen that adding useless predictors to a regression model will increase R2. Here, let's examine what our inference methods say if the predictors are in fact useless. Suppose the true/pop fit is \( y = 1 \), (i.e., no \( x \) at all), and so a possible sample from the population could be the following:

```r
cat("set.seed(123)  # Use this line to make sure we all get the same answers.
    n = 20
    y = 1 + rnorm(n,0,1)\n")
a) Write code to make data on 10 useless predictors (and no useful predictors), fit the model \( y = \alpha + \beta_1 x_1 + \ldots + \beta_{10} x_{10} \), perform the test of model utility, and perform t-tests on each of the 10 coefficients to see if they are zero. Show/turn-in your R code.

```r
def x = runif(n,-1,1)
    x2 = runif(n,-1,1)
    x3 = runif(n,-1,1)
    x4 = runif(n,-1,1)
    x5 = runif(n,-1,1)
    x6 = runif(n,-1,1)
    x7 = runif(n,-1,1)
    x8 = runif(n,-1,1)
    x9 = runif(n,-1,1)
    x10 = runif(n,-1,1)
    lm.1 = lm(y~x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10)
syntax(lm.1)
# Coefficients:
# Estimate    Std. Error t value  Pr(>|t|)
# (Intercept)  0.94895    0.30159   3.146   0.0118 *
# x1          -1.12382    0.52108  -2.157   0.0594 .
# x2           0.06083    0.57739   0.105   0.9184
# x3          -0.27627    0.54053  -0.511   0.6216
# x4          -0.91901    0.45966  -1.999   0.0766 .
# x5           0.41941    0.50238   0.835   0.4254
# x6          -0.33002    0.66481  -0.496   0.6315
# x7           1.14796    0.62241  1.844   0.0982 .
# x8          -1.06642    0.71855  -1.484   0.1719
# x9          -0.18101    0.73512  -0.246   0.8110
# x10         0.14813    0.49721  0.298   0.7725
#
# Residual standard error: 0.8696 on 9 degrees of freedom
# Multiple R-squared:  0.6214,  Adjusted R-squared:  0.2006
# F-statistic: 1.477 on 10 and 9 DF,  p-value: 0.2846
```

b) According to the F-test of model utility, are any of the predictors useful at alpha = 0.1?

The p-value (0.2846) is larger than alpha, and so, we cannot reject H0 in favor of H1. In English: There is no evidence that any of the regression coefficients are nonzero. I.e., there is no evidence that any of the predictors are useful.

c) According to the t-tests, are any of the predictors useful at alpha = 0.1? See the solns to make sure you understand the moral of this exercise.

Three of the p-values are less than 0.1, and so there is evidence that three of the predictors are useful. Your count may be different from mine here, but the point is that about alpha% of the predictors will be appear to be useful, when in fact they are not. The F-test of model utility, however, does not have that defect, because the test is done only 1 time.
Consider 5 brands of computers. A code has been run on each of the brands 6 times, and the completion times have been recorded. Here are the data:

<table>
<thead>
<tr>
<th>Brand</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{y}_1 )</td>
<td>13.68</td>
<td>13.97</td>
<td>13.67</td>
<td>14.73</td>
<td>13.88</td>
</tr>
<tr>
<td>( s_1 )</td>
<td>1.194</td>
<td>1.167</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

\[ \overline{y}_1 = 13.68, \overline{y}_2 = 15.97, \overline{y}_3 = 13.67, \overline{y}_4 = 14.73, \overline{y}_5 = 13.88 \]

\[ s_1 = 1.194, s_2 = 1.167, \text{---} = \text{---} \]

a) Write the 1st term in each of the following sums:
(Hint for \( SS_{\text{within}} \): The double-sum is actually a single-sum!)

\[ \overline{Y} = \frac{\sum}{n} \left( \frac{n_i}{n} \overline{y}_i \right) = \frac{6}{30} (13.68) + \cdots = 14.22 \]

\[ SS_{\text{between}} = \sum \frac{n_i}{n} (\overline{Y}_i - \overline{Y})^2 = 6(13.68 - 14.22)^2 + \cdots = 30.88 \]

\[ SS_{\text{within}} = \sum \frac{n_i}{n} \left( \overline{y}_i - \overline{Y} \right)^2 = \frac{5}{6} (n_i - 1) s_i^2 = (6-1)(1.194)^2 + \cdots = 22.83 \]

b) Compute the observed F-ratio.

\[ F = \frac{30.88/(5-1)}{22.83/(30-5)} = 8.45 \]

df = (5-1, 30-5)

\[ p\text{-value} = p(F > F_{0.01 \text{, df}}) = p(F > 8.45) < 0.001 \text{ Table VIII} \]

Conclusion: \( H_0 (\mu_1 = \mu_2 = \cdots) \) in favor of \( H_1 (4 \text{ at least 2 } \mu_1 \text{'s are diff}) \)

In English: Brand has an effect on speed.

which 2?

Section 9.3

Skipped!