More Multiple Linear Regression - Inference in MLR

Revisiting hypothesis testing in MLR, we discussed

- **1. Testing whether a single coefficient is 0 (or some other value of interest)** - **t-test with n-p-1 df**, standard error is a function of MSE for model:

$$\text{var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}$$

where

$$\hat{\sigma}^2 = MSE = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - p - 1}.$$

We then compute the test as

$$t = \frac{\hat{\beta}_1 - value}{\sqrt{\text{var}(\hat{\beta}_1)}}$$

and compare to a critical value from a t dist'n with $n - p - 1$ df

More Multiple Linear Regression - Inference in MLR

– **2. Testing whether all coefficients are simultaneously 0 (whole model is 'null') - F-test**

H_0: all $\beta_s = 0$

The Global F-test in Regression (conceptually):

* $SSR\ (SS_{between}) = \sum(\hat{Y}_i - \bar{Y})^2$ - variation in a predicted 'group mean' $\hat{Y}$ around the overall mean of Y. MSR = SSR/p
* $SSE\ (SS_{within}) = \sum(Y_i - \hat{Y}_i)^2$ - variation between predicted and observed Y_is. Quantifies how group-specific values vary around their group specific mean ($\hat{Y}$ is a group-specific mean, with the group being those with specific value of covariate X) . The MSE = SSE/(n-p-1)

If the ratio MSR/MSE (**F Statistic**) is **large** (far away from 1.0), then ' some X's identify group-specific means', or the population means likely differ across X (as a linear combination of X's weighted by β coefficients)

p-value is small.

Testing in Multiple Regression -
-subsets of coefficients-

(SPRM Chapter 4)

H_0: some $\beta_s = 0$

– **Testing whether several coefficients are simultaneously 0 (or some other value)**

These tests are useful for refining a multivariable model down to those predictors that contributing meaningfully to explaining/predicting Y. (as defined by statistical significance criteria we have specified)

Pairwise

This approach basically compares two models – the model with all predictors in it, or "the full model(FM)" and a model with fewer predictors, "the restricted model (RM)". There are m parameters $1 \leq m \leq p$ that could be omitted. The text (SPRM) also calls the reduced model the 'incomplete' model.

or reduced

Assess if including additional variables improves the model's fit.

Testing in Multiple Regression - subsets of coefficients

SPRM 4.3

The hypothesis setup is:

$H_0 : \beta_j = \beta_k = \ldots = \beta_m = 0$

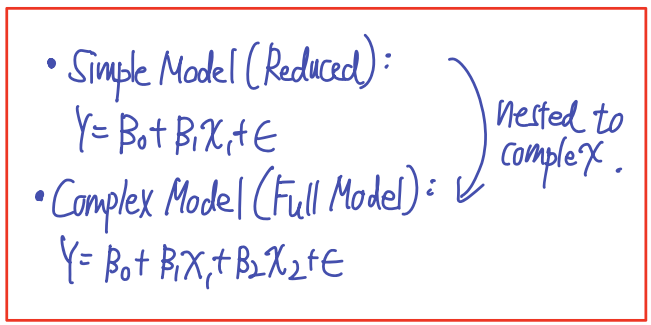
H_1 : at least one of these $\beta \neq 0$

We use properties of the sum of squared errors

$$\sum (y_i - \hat{y}_i)^2$$

from the different models to formulate the test.

Models are said to be *nested* if one is a subset of another with respect to the predictors included. For now we consider these type of tests only



Testing in Multiple Regression - subsets of coefficients

Some facts:

- SST does not change from model to model (for the same fixed n cases included). Recall that SST is simply $\sum(y_i - \bar{y})^2$, without regard to any X

  Full v.s Reduced Model

- SSE $\sum(y_i - \hat{y}_i)^2$ will become smaller in models with a larger number of parameters (i.e., predictors) included, although the gain may be trivial, if the added predictors do not contribute materially.

  Technically adding predictors should ↑ accuracy, if the variable is appropriate.

- Since $R^2 = 1 - SSE/SST$, then R^2 always gets larger as we include more predictors relative to fewer (from a set of p predictors and fixed n), again possibly trivially so

Contrasting Models via F-tests

Tests contrasting nested models are referred to as *partial F-tests*. The key idea is to contrast the SSEs from both models (Full Model and Reduced Model) with respect to the full model's SSE. In this sense we look at the relative loss in information for dropping some predictors. The partial F-test is:

$$F = \frac{(SSE_{RM} - SSE_{FM})/(df_{RM} - df_{FM})}{SSE_{FM}/df_{FM}}$$

Reduced − Full

(Always positive value)

Bigger errors, bigger value.

Notes:

- This statistic takes on a valid F statistic value (> 0) because SSE_{RM} is always greater than SSE_{FM} . Keep in mind that fewer parameters means more variability in prediction, so SSE_{RM} is **always larger** than SSE_{FM}

↑ Variability ↑ Errors

Contrasting Models via F-tests

$$F = \frac{(SSE_{RM} - SSE_{FM})/(df_{RM} - df_{FM})}{SSE_{FM}/df_{FM}}$$

Notes (cont):

- This statistic comes from a distribution that has two defining parameters (called degrees of freedom), $df_{RM} - df_{FM}$ and df_{FM}.
- Note that numerator degree of freedom is just the difference in the numbers of predictors between the two models – **that is, the number of βs being tested**. ex) Page 62 : then $df_{RM} - df_{FM} = 2$.
- The FM degrees of freedom is always $n - p - 1$ (full model minus intercept and all predictors).

Partial → "Partial of the Full Model": Nested.

Contrasting Models via F-tests

- As an example, let's assume that we are testing two parameters (from an initial model with 4 predictors)

  RM FM

  $H_0 : \beta_3 = \beta_4 = 0$ (in a model that with $\beta_1, \beta_2, \beta_3, \beta_4$)

  H_1 : at least one of β_3 or $\beta_4 \neq 0$

- A way to test this hypothesis is by contrasting models using the partial F-test. In Stata, we use the "test" command. After fitting the full model (all predictors), we test whether two predictors (inc and road) can be omitted:

If $\beta_3 = \beta_4 = 0$, then that means not significant and can be omitted.

Contrasting Models via F-tests - Example

Full Model

```
. reg fuel tax dlic inc road
```

```
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  4,    43) =   22.71
       Model |  399316.478     4  99829.1195            Prob > F      =  0.0000
    Residual |  189050.001    43  4396.51165            R-squared     =  0.6787
-------------+------------------------------           Adj R-squared =  0.6488
       Total |  588366.479    47  12518.4357            Root MSE      =  66.306

------------------------------------------------------------------------------
        fuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |  -34.79016    12.9702    -2.68   0.010    -60.94706   -8.633249
        dlic |   13.36449   1.922982     6.95   0.000     9.486431    17.24255
         inc |  -66.58875   17.22175    -3.87   0.000    -101.3197   -31.85778
        road |   -2.42589   3.389175    -0.72   0.478    -9.260812    4.409032
       _cons |   377.2913   185.5412     2.03   0.048     3.111754    751.4708
------------------------------------------------------------------------------
```

SSE FM

Restricted Model

```
. test road inc
```

"Nested"

```
 ( 1)  road = 0
 ( 2)  inc = 0

       F(  2,     43) =    8.16
            Prob > F =    0.0010
```

You can also calculate all the above F statistics by hand as shown in SPRM 4.6: fit two regressions – one using all predictors (the full model) and the other excluding the predictors you wish to test (the restricted model). With the SSEs from the two 'nested' models, you can calculate the F-statistic.The two-variable reduced model:

```
. reg fuel tax dlic
```

"Nested"

```
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  2,    45) =   28.25
       Model |  327532.469     2  163766.234            Prob > F      =  0.0000
    Residual |  260834.01     45  5796.31134            R-squared     =  0.5567
-------------+------------------------------           Adj R-squared =  0.5370
       Total |  588366.479    47  12518.4357            Root MSE      =  76.134
```

SSE_RM

Contrasting Models via F-tests

- **Reduced model** has SSE = 260834.01; **Full model** has SSE = 189050.001; both models have same SST = 588366.479

  And the F-statistic is calculated as

$$\begin{aligned} F &= \frac{(SSE_{RM} - SSE_{FM})/(df_{RM} - df_{FM})}{SSE_{FM}/df_{FM}} \\ &= \frac{(260834.01 - 189050.001/2}{189050.001/43} \\ &= 8.164 \end{aligned}$$

  with dfs 2 and 43. The corresponding p-value is obtained by "display Ftail (2,43,8.164)" is .00098739

  same as given above by *test* command

- ✓ **Conclusion:** Of these two variables, at least one contributes significantly to the model., making $H_A: \beta_3 \text{ or } \beta_4 \neq 0$ correct.

Reiterating other Uses of F-tests

- The partial F-test can also be used (and in fact, already has) to test the scenarios 1 and 2 discussed earlier. These are just special cases of the partial F-test approach:

  1. Testing whether one single coefficient is 0. t-test o.k.
  This scenario is equivalent to the comparison of two models: one with all covariates (the full model), and the other without X_i (the restricted model). The test statistic has the form

$$F = \frac{(SSE_{RM} - SSE_{FM})/(1)}{SSE_{FM}/(n-p-1)}$$

  It is equivalent to the t-test for that parameter and comes directly out of the analysis in the table of coefficients (so we don't usually need this test).

2. Testing whether all coefficients are simultaneously 0. F-test o.k.

This scenario is equivalent to the comparison of two models: the full model and the model with only the intercept. Note that for the latter, SSE is just the total sum of squared errors (SST), ignoring X.

This test is obtained by default as the global test provided in the ANOVA table and equals

$$F = \frac{(SST - SSE_{FM})/(p)}{SSE_{FM}/(n-p-1)}$$

which equals

$$F = \frac{(SSR)/(p)}{SSE/(n-p-1)}$$

Again, this is the typical F-test for the whole model and is always generated from the model run.

Contrasting Models via F-tests in R

```
> library(foreign)
> fuel = read.dta("fuel.dta")
> fuel
   state   pop   tax  nlic   inc   road fuelc dlic fuel
1     ME  1029  9.00   540 3.571  1.976   557 52.5  541
2     NH   771  9.00   441 4.092  1.250   404 57.2  524
.   .   .
>
> bigmodel = lm(fuel$fuel ~ fuel$tax +fuel$dlic +fuel$inc +fuel$road)
> summary(bigmodel)
. . .
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  377.291    185.541   2.033 0.048207 *
fuel$tax     -34.790     12.970  -2.682 0.010332 *
```

```
fuel$dlic     13.364      1.923   6.950 1.52e-08 ***
fuel$inc     -66.589     17.222  -3.867 0.000368 ***
fuel$road     -2.426      3.389  -0.716 0.477999
---
Res std err: 66.31 on 43 df; R-squared: 0.6787, Adj R-squared: 0.6488
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10

> smallmodel = lm(fuel$fuel ~ fuel$tax + fuel$dlic)
> anova(smallmodel, bigmodel)
Analysis of Variance Table
 Model 1: fuel$fuel ~ fuel$tax + fuel$dlic
 Model 2: fuel$fuel ~ fuel$tax + fuel$dlic + fuel$inc + fuel$road
   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
 1     45 260834
 2     43 189050  2     71784 8.1637 0.0009876 ***
```

The Multiple Correlation Coefficient and R^2

- In SLR, recall that

$$R^2 = \frac{SSR}{SST}$$

  equals the the correlation coefficient estimate r squared. Note that $\text{corr}(y, x) = \text{corr}(Y, \hat{Y})$

- In multiple linear regression, each X may be correlated with Y to a different degree. The set of fitted values $\hat{Y}$ reflect the linear correlation of Y with all Xs via the linear model

- Thus, the quantity $\text{corr}(Y, \hat{Y})$ or *multiple correlation coefficient*, equals $\sqrt{(R^2)}$ from the model

- As we've discussed, the R^2 from the MLR model has the same interpretation with respect to proportion of variability in Y that is explained by the model

The Multiple Correlation Coefficient and R^2

Illustrating the correlation between $\hat{Y}$ and Y

```
. reg fuel tax dlic inc road
      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(4, 43)        =     22.71
       Model |  399316.478         4  99829.1195   Prob > F        =    0.0000
    Residual |  189050.001        43  4396.51165   R-squared       =    0.6787
-------------+----------------------------------   Adj R-squared   =    0.6488
       Total |  588366.479        47  12518.4357   Root MSE        =    66.306
. . . . .
. predict yhat
(option xb assumed; fitted values)
. corr fuel yhat
(obs=48)
             |     fuel     yhat
-------------+------------------
        fuel |   1.0000
        yhat |   0.8238   1.0000

. display 0.8238^2
.67864644
```

Limitations of R^2 dealing with multiple predictors

Contrasting Models R^2 values

We sometimes may want to consider an alternative to the usual R^2. Two reasons:

1. In models with many parameters relative the sample size n, we may want to apply a 'penalty' for the model complexity, for including unecessary variables. (Overfitting)
2. We may wish to contrast non-nested models with respect to fit (via R^2). For example, we might have two models that have partially overlapping predictor sets. Both are subsets of the full model but not of each other. We may want evaluate the models in relation to the numbers of parameters, as models with more parameters may be better but less favorable for other reasons. Avoiding inflation with additional parameters.

The *adjusted* R^2 (Discussed in SPRM Section 5.4.2)

$$\text{Adjusted-}R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Contrasting Models R^2 values

it can also be written as

$$\text{Adjusted-}R^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

So, the adjusted R^2 is always smaller than R^2, in effect 'handicapped' by the number of parameters relative to n.

Example: for n=100 and unadjusted $R^2 = .90$ in a model with 20 predictors, the adjusted $R^2 = 0.87$. If one had a 10-parameter model with unadjusted $R^2 = 0.88$, it might be preferred (because in that case the adjusted R^2 is also about 0.87)

Adjusted and unadjusted preferred to be similar

We will discuss uses of quantities such as this when we review modeling strategy

Multiple Linear Regression - Trade-offs Between R^2 and Precision

Some questions:

- What are the consequences of included non-significant variables on the model? If R^2 always improves, why would I omit variables (aside from penalty just described, which can be modest)? too many variables
- What are the consequences of omitted important variables? If fewer parameters is better for MSE estimate, what is the trade-off?

β_1 β_2 β_3 β_4

```
. regress fuel tax dlic inc road
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  4,    43) =   22.71
       Model |  399316.478     4  99829.1195            Prob > F      =  0.0000
    Residual |  189050.001    43  4396.51165            *R-squared    =  0.6787*
-------------+------------------------------           Adj R-squared =  0.6488
       Total |  588366.479    47  12518.4357           * Root MSE      =  66.306*

------------------------------------------------------------------------------
        fuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |  -34.79016    *12.9702*    -2.68   0.010    -60.94706   -8.633249
        dlic |   13.36449   1.922982     6.95   0.000     9.486431    17.24255
         inc |  -66.58875   17.22175    -3.87   0.000    -101.3197   -31.85778
        road |   -2.42589   3.389175    -0.72   0.478    -9.260812    4.409032
       _cons |   377.2913   185.5412     2.03   0.048     3.111754    751.4708
------------------------------------------------------------------------------
. * now drop one variable
```

Omitting road, β_4

```
. regress fuel tax dlic inc

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  3,    44) =   30.44
       Model |   397063.99     3  132354.663            Prob > F      =  0.0000
```

```
    Residual |  191302.489    44  4347.78385           *R-squared     =  0.6749*
-------------+----------------------------           Adj R-squared =  0.6527
       Total |  588366.479    47  12518.4357           *Root MSE      =  65.938*
------------------------------------------------------------------------------
        fuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |  -29.48381   *10.58358*    -2.79   0.008    -50.81361   -8.154015
        dlic |   13.74768   1.836696      7.49   0.000     10.04607    17.4493
         inc |  -68.02286   17.00975     -4.00   0.000    -102.3038   -33.74196
       _cons |    307.328   156.8307      1.96   0.056    -8.743502    623.3994
------------------------------------------------------------------------------
```

Better (Adj R-squared)

Better (Std. Err.)

Note: In smaller model:

- R^2 is only slightly reduced- 0.6749 (small model) vs. 0.6787 (full model) Reason to adjust R^2.
- root MSE is a little bit smaller (this is good) - 65.938 vs. 66.306 - (SSE is larger as expected, but now we divide it by 44 instead of 43 to get MSE) Account for extra parameter +1
- standard error on $\hat{\beta}_1$ smaller 10.58358 vs. 12.9702 - more precision

Thus, there is 'cost' to including unnecessary variable

Go further, omit a potentially important variable:

```
. regress fuel tax dlic
```

Omitting income, β_3

```
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  2,    45) =   28.25
       Model |  327532.469     2  163766.234           Prob > F      =  0.0000
    Residual |   260834.01    45  5796.31134           *R-squared     =  0.5567*
-------------+------------------------------           Adj R-squared =  0.5370
       Total |  588366.479    47  12518.4357           *Root MSE      =  76.134*

------------------------------------------------------------------------------
        fuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |  -32.07532   *12.19716*  -2.63   0.012    -56.64166   -7.508984
        dlic |   12.51486   2.090614     5.99   0.000     8.304147    16.72557
       _cons |   108.9709   171.7859     0.63   0.529    -237.0236    454.9655
------------------------------------------------------------------------------
```

Worse

Worse

Now:

- R^2 is substantially smaller in this model - 0.5567 vs. 0.6749
- root MSE is a larger - 76.134 vs. 65.938 (because we have given up an important predictor, model fit is worse)

Not appropriate to remove income.

– s.e($\hat{\beta}_1$, tax) is larger - 12.19716 vs. 10.58358 - affects significance level a bit here

Some decisions in modeling will not be strictly based on statistical tests, but balancing different aspects of the model depending on context and goals. For example, explaining the relationship vs making best prediction

Misc Topics - Centering and Scaling of Variables

We discussed earlier that linear regression models relate predictors to outcomes in the units associated with the variables. We may want to make some transformations of the data for a number of reasons:

- To make βs unitless, put all predictors on same relative scale
- To render the intercept term (when $X = 0$) meaningful in the model
  ex): Age = 0, unrealistic: Instead, center around mean.
- To deal with collinearity among Xs, we will discuss this later

Misc Topics - Centering and Scaling of Variables

SPRM Section 3.18

- **Centering** refers to subtracting the mean from each variable. For example, $y - \bar{Y}$ or $X_j - \bar{X}_j$. Note that these variables will have mean zero. Also, if one plugs in 1 for X in the model, the prediction is at one unit above the mean. ex) 80-X : To test when someone is 70, plug X=10.
- **Unit length scaling** refers to dividing the centered value by the 'length' of the data, defined as the square root of the sum of squared deviations around the mean. After unit-length scaling, variables have mean 0 and length 1. For example

$$\tilde{Z}_i = \frac{(y_i - \bar{Y})}{\sqrt{\sum_i^n (y_i - \bar{Y})^2}}$$

Ex- Centering a covariate at the mean
Fish Mercury Data

```
. reg mercury weight

      Source |       SS           df       MS      Number of obs   =       171
-------------+----------------------------------   F(1, 169)       =     74.77
       Model |  30.2510497         1  30.2510497   Prob > F        =    0.0000
    Residual |  68.3712259       169  .404563467   R-squared       =    0.3067
-------------+----------------------------------   Adj R-squared   =    0.3026
       Total |  98.6222756       170  .580131033   Root MSE        =    .63605
------------------------------------------------------------------------------
     mercury |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0004818   .0000557     8.65   0.000     .0003718    .0005918
       _cons |   .6386813   .0803536     7.95   0.000     .4800552    .7973073
------------------------------------------------------------------------------
```

Center "weight"

```
. egen meanweight = mean(weight)
. gen weight_c = weight - meanweight
. reg mercury weight_c
```

```
      Source |       SS           df       MS      Number of obs   =       171
-------------+----------------------------------   F(1, 169)       =     74.77
       Model |  30.2510494         1  30.2510494   Prob > F        =    0.0000
    Residual |  68.3712262       169  .404563469   R-squared       =    0.3067
-------------+----------------------------------   Adj R-squared   =    0.3026
       Total |  98.6222756       170  .580131033   Root MSE        =    .63605
------------------------------------------------------------------------------
     mercury |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    weight_c |   .0004818   .0000557     8.65   0.000     .0003718    .0005918
       _cons |   1.191754   .0486402    24.50   0.000     1.095734    1.287775
------------------------------------------------------------------------------
```

Ratio always the same

Note: Inference is identical to earlier. Slope is identical (y was transformed linearly), so R^2, F, etc, same. The intercept now equals the mercury level at the mean fish weight of the sample

Closer to the meaning rather than only statistical #.

ex) Econ/Business Statistics

Misc Topics - Centering and Scaling of Variables

- **Standardizing** refers to centering and dividing by the sample standard deviation. Can be applied to predictors or response

$$\tilde{Z}_j = \frac{(y_j - \bar{Y})}{\hat{\sigma}}$$

where $\hat{\sigma} = \sqrt{\sum_i^n (y_j - \bar{Y})^2 / n - 1}$, the sample standard deviation

- This representation reflects how extreme a value is from the mean or center, taking into account variability
  - Compare relative importance of predictors
  - Transform variables to mean of 0 and standard deviation for easier comparison.

Regression Analysis - with Centering and Scaling

Ex: Clinical features used to characterize prostate cancer include Gleason score - a quantitative measure that rates tumor aggressiveness, Prostate Specific Antigen (PSA), which may reflect extent of disease burden, including sub-clinical disseminated tumor cells. New tumor molecular markers are intended to better characterize and refine prospective risk at diagnosis, to make decisions about extent of treatment and subsequent surveillance

- The main purpose of our recent study was to determine what these biomarkers offer in terms of future recurrence risk*. Before that, there were some relevant questions that could be addressed with linear regression

*Pollack, Dignam, Diaz DA, et al. A tissue biomarker based model that identifies patients at high risk of distant metastases . *Clin Canc Res* 2014 - online Oct 7

Regression Analysis of Prostate Cancer Biomarkers

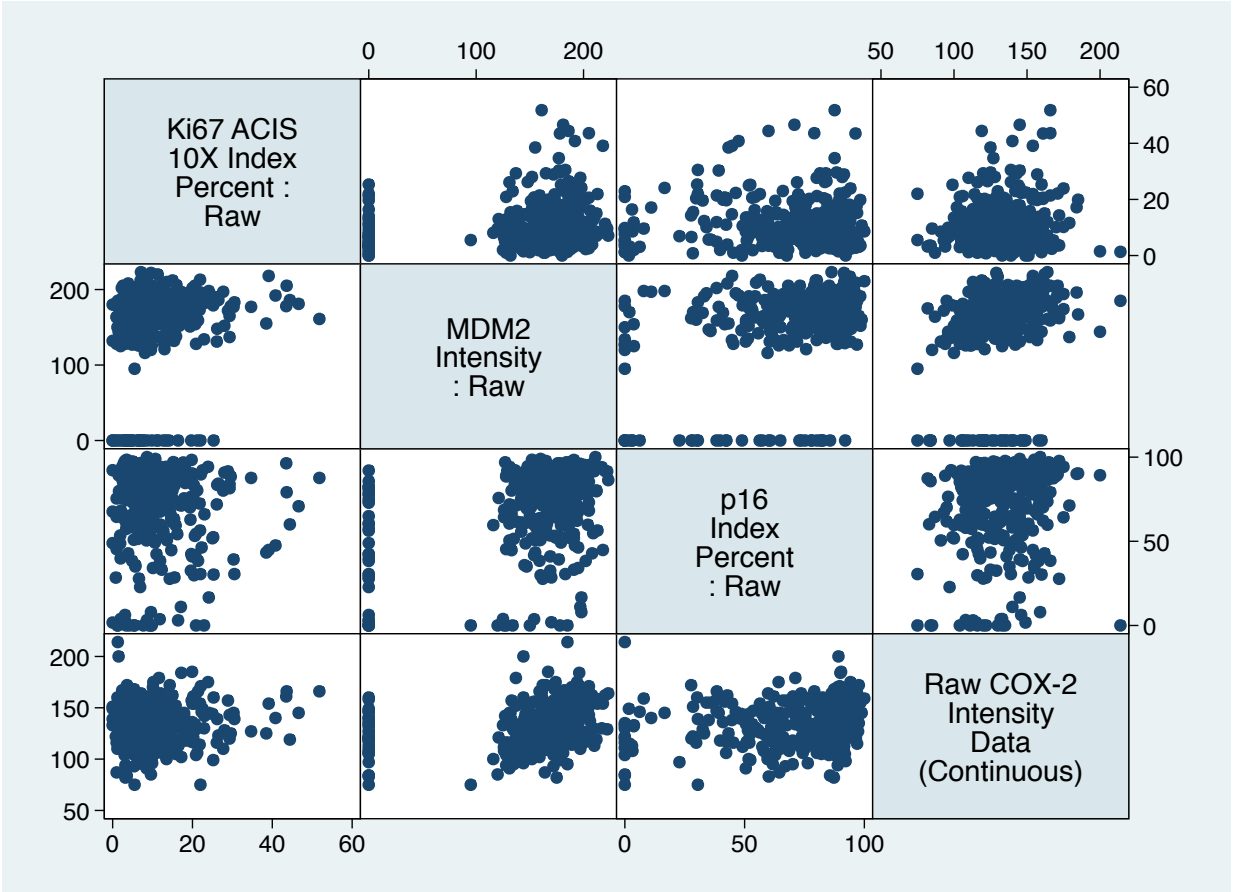
In this analysis, Ki-67 emerged as an important molecular factor.
We might ask:

- How is Ki-67 related to standard prognostic variables, in particular Gleason and PSA, since there are already measured as part of standard diagnostic work-up?
- How is Ki-67 related to other markers that were measured? Could we impute missing Ki-67 from other molecular measurements?

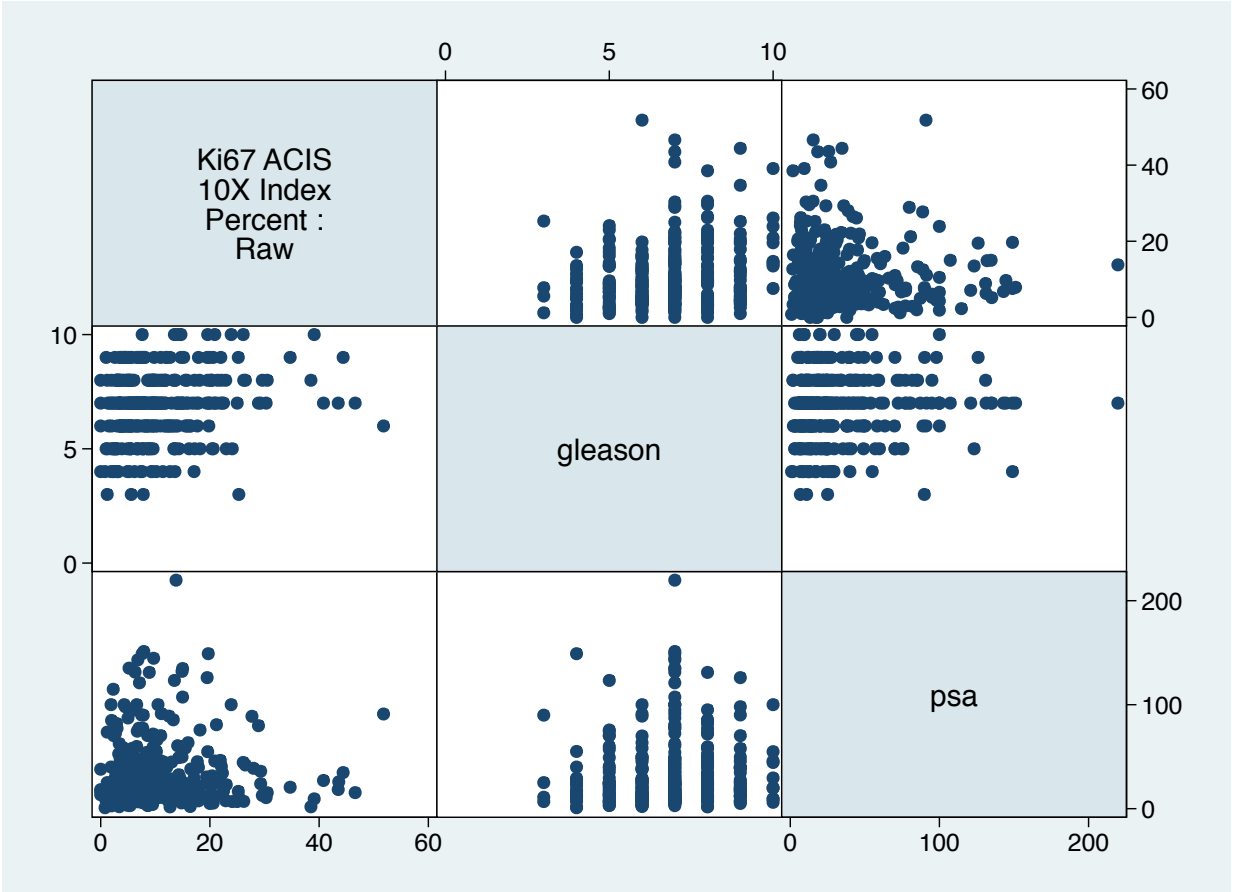
Regression Analysis of Prostate Cancer Biomarkers

We can use correlation and regression to address these questions. Because molecular markers are measured on different platforms, they may be determined as a binding capacity, percent staining positive, staining intensity, etc. We may wish to standardize marker data to assess relationships on the same unitless scale. This may also be useful for identifying extreme values that may be lab error

Scatterplot of Biomarkers $n > 400$



Scatterplot of Response and Clinical Markers



The Model with Covariates as Measured

```
. regress ki67_acis10_index_percent mdm2_intensity p16_index_percent cox2_intensity psa gleason

      Source |       SS       df       MS              Number of obs =     343
-------------+------------------------------           F(  5,   337) =    6.39
       Model |  1931.90354     5  386.380709            Prob > F      =  0.0000
    Residual |  20389.4445   337  60.5028027            R-squared     =  0.0865
-------------+------------------------------           Adj R-squared =  0.0730
       Total |   22321.348   342  65.2670996           Root MSE      =  7.7784

------------------------------------------------------------------------------
ki67_acis10_ind~t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------
   mdm2_intensity |   .0247216   .0085994     2.87   0.004     .0078062    .0416369
p16_index_percent |  -.0496094   .0184626    -2.69   0.008    -.0859259   -.0132929
   cox2_intensity |   .0328408   .0218002     1.51   0.133    -.0100408    .0757224
              psa |  -.0022764     .01326    -0.17   0.864    -.0283593    .0238064
          gleason |   1.136884   .2929631     3.88   0.000     .5606176    1.713151
            _cons |  -1.824927   3.648162    -0.50   0.617    -9.000965    5.351111
------------------------------------------------------------------------------
```

Analysis of Prostate Cancer Biomarkers

Comments

- the global F-test indicates presence of statistically significant predictors of Ki-67. However, the R^2 is only about 8%
- Other molecular features are associated with Ki-67.
- Gleason score is positively associated. This is biologically plausible. Given the very low R^2, we would not conclude that we can capture Ki-67 information with Gleason score.

Many variables are in disparate units. The effect of Gleason (a discrete ordinal variable) appears huge compared to the others. it might be helpful to put the response on a standardized scale, as well as the other molecular predictors. We will leave PSA as is, as the range and the meaning of values for this variable is well known.

The Model with Transformed Covariates

The same model with standardized molecular marker variables

```
. regress stan_ki67 stan_mdm2 stan_p16 stan_cox2 psa gleason

      Source |       SS       df       MS              Number of obs =     343
-------------+------------------------------           F(  5,   337) =    6.39
       Model |  32.4817786     5  6.49635572            Prob > F      =  0.0000
    Residual |   342.81495   337  1.01725504            R-squared     =  0.0865
-------------+------------------------------           Adj R-squared =  0.0730
       Total |  375.296728   342  1.09735886            Root MSE      =  1.0086

------------------------------------------------------------------------------
   stan_ki67 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   stan_mdm2 |   .1947285   .0677366     2.87   0.004     .0614886    .3279683
    stan_p16 |  -.1638238   .0609686    -2.69   0.008    -.2837507   -.0438969
   stan_cox2 |   .0921248   .0611538     1.51   0.133    -.0281665    .2124161
         psa |  -.0002952   .0017194    -0.17   0.864    -.0036772    .0030869
     gleason |   .1474156   .0379875     3.88   0.000     .0726932     .222138
       _cons |  -.8964865   .2660011    -3.37   0.001    -1.419718   -.3732549
------------------------------------------------------------------------------
```

Analysis of Prostate Cancer Biomarkers

Comments

- Only the effect estimates (coefficient values and associated standard errors) change. Inference on these coefficients is identical
- In the ANOVA table, SSR, SSE, SST and associated mean quantities change, all other quantities do not change
- When using the model to, say, predict, we must remember to use transformed versions of predictors

Conclusion: Ki-67 associated with, but not redundant with known prognostic factors or other markers. It is useful to consider jointly with these other factors in survival modeling and risk quantification.

Multiple Linear Regression
Summary so Far

- Multiple regression is a natural extension of simple linear regression, with the principles of SLR still applying. Some additional important aspects
  - Graphical exploration of relationship of X's to Y is more complex, but still important - we need to rely on residual $(y_i - \hat{y}_i)$ examination more to critique model - we will discuss soon
  - Testing can involve multiple parameters, hypotheses. The ANOVA table and partial F-tests/nested models (Reduced Model / Full Model) provide a means to conduct additional tests
  - Must balance model complexity and fit, predictive ability