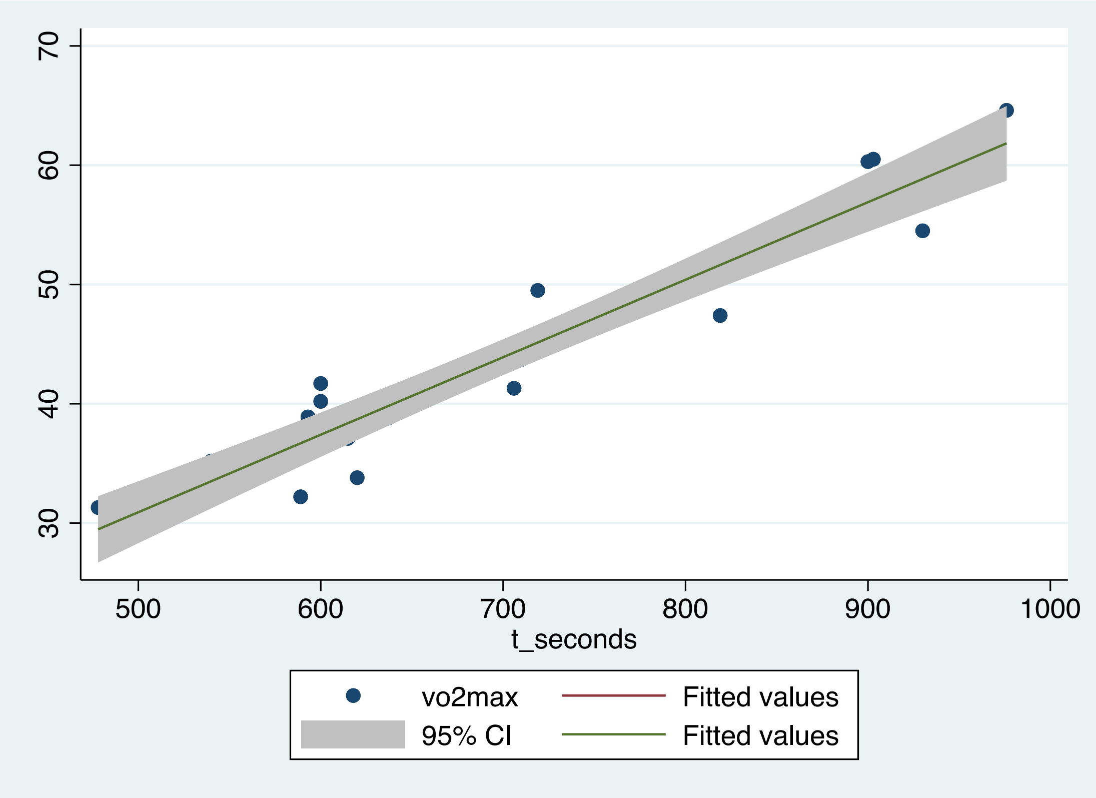
PBHS 32410 / STAT 22401

Linear Regression

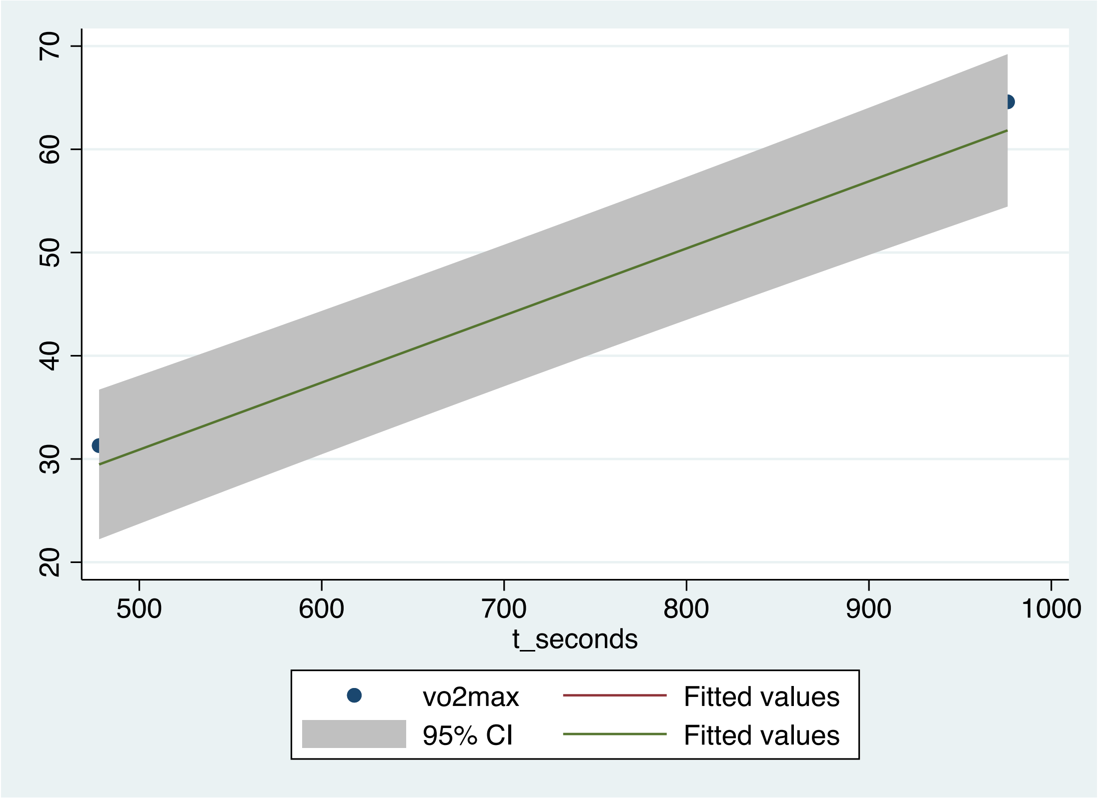
A bit more on prediction: The two relevant prediction intervals for the VO_2MAX data.

CI on the mean VO_2MAX:



Linear Regression

Forecast interval for any given running time:



Linear Regression

The standard error expressions:

Mean prediction

$$s.e.(\hat{\mu_0}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Forecast

$$s.e.(\hat{y}_j|x_j) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

What happens when

- new X value far from $\bar{X}$? Equal $\bar{X}$?
- n very large?
- Fit very good?

Multiple Linear Regression

- We started by the simplest statistical model that makes some physical sense and fits the assumptions we impose. From there, we might naturally build a more elaborate models.
- Specifically, if there are other variables that may help us predict Y, we could extend our simple model to include them.
- Furthermore, we might *need* to incorporate other variables, because the one we have chosen may be a substitute or proxy for another variable, and without it our interpretation, especially any causal inference, would be flawed.

  Thus we extend the approach to **multiple linear regression**. We call it **multiple** vs. **multivariate**, as it is sometimes misnamed, because there is only one Y variable and multiple X variables.

Multiple Linear Regression

In multiple linear regression (MLR), several predictors are available to help explain the behavior of the response variable. We will retain our usual notation of denoting the response variable as Y and the p predictors (or covariates) as $X_1, X_2, \ldots, X_p$. For simple linear regression (SLR), $p = 1$ (one predictor).

Our model is an equation that expresses the response as a linear function of the predictors:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_p X_{p,i} + \epsilon,$$

where subscript i denotes the observation number – i ranging from 1 (first observation in sample) to n (last observation).

Multiple Linear Regression - coefficients

As earlier, the regression coefficients (βs) can be estimated via the least squares estimation procedure, by simultaneous minimization of the sum of squared errors. Find set of βs such that :

$$\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1,i} - \ldots - \beta_p X_{p,i})^2$$

is minimized.

- This involves solving a system of linear equations, and whether there is a unique solution depends on some aspects of the data. We will discuss this later.
- For the problems we will work with, we assume that we have a properly specified model and data to support estimating it. There are some other conditions under which the solution (the βs) is not unique or not obtainable, we will also discuss in relation to diagnostics.

Entering the Matrix

Matrix Representation of Regression Model

Recall that the (least squares) estimation equations for SLR:

$$\hat{\beta}_1 = \frac{\sum X_i(Y_i - \bar{Y})}{\sum X_i(X_i - \bar{X})}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

These equations grow for more predictor (X) variables; we need $p + 1$ equations for p predictors. Also, other important information, such as covariance among βs is needed.

To facilitate and unify SLR and MLR, we use matrix algebra methods to express the models and solve.

For SLR, specify

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \mathbf{and\ e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Then by matrix multiplication rules

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{pmatrix}$$

and the model can be written as: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$

Estimation of Regression Model via Matrix Form

Recall that for SLR, we obtain the $\hat{\beta}$s by minimizing the sum of squared differences between y and $\hat{y}$ (the residuals).

$$\sum (Y_i - (\beta_0 + \beta_1 X_i))^2 = \sum_{i=1}^{n} \mathbf{e_i^2}$$

Using matrix algebra notation, we can define this quantity as

$$S = \sum_{i=1}^{n} \mathbf{e_i^2} = \mathbf{e^T e}$$

and we want to find the β values that minimize

$$S = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathbf{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

where $(X)^T$ means "X transpose", or swap rows and the columns of X (this is the sum of squared differences in matrix notation)

Matrix Representation of Regression Model

By calculus and matrix algebra operations, we can generate and solve the equations that yield the β values (eqn 3.11)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T Y}$$

This is the means by which the computer fits the model (i.e., estimates the βs).

This approach naturally expands to accommodate multiple linear regression by expanding the $\mathbf{X}$ and $\boldsymbol{\beta}$ components.

Matrix Representation of Regression Model

In MLR, the model is $Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_1 X_{p,i} + \epsilon_i$,

Define the following vectors and matrices as before

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The model can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$.

Recall that for a MLR $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \ldots + \hat{\beta}_1 X_{p,i}$, we obtain the $\hat{\beta}$s by

$$\min \sum (Y_i - \beta_0 - \beta_1 X_{1,i} - \ldots - \beta_1 X_{p,i})^2$$

Matrix Representation of Regression Model

Again, by doing the math, it is found that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y}$$

This is a compact way of writing (and programming) least squares solutions for multiple regression with any number of predictors.

We will not use this information further, except that we will extract specific elements from these computations for model diagnostics later - computer programs provide these

Multiple Linear Regression - coefficients

Recall, in SLR, the interpretation of β_1 was that it represented the slope of the regression function. The slope is in turn interpreted as the change in expected Y induced by a one-unit increase of X. To see this, observe that if:

$$\mathrm{E}(Y^{old}) = \beta_0 + \beta_1 X^{old} \text{ and}$$
$$\mathrm{E}(Y^{new}) = \beta_0 + \beta_1 (X^{old} + 1),$$

then

$$\mathrm{E}(Y^{new}) - \mathrm{E}(Y^{old}) = \beta_0 + \beta_1 (X^{old} + 1) - (\beta_0 + \beta_1 X^{old}) = \beta_1$$

So, when X changes by one unit, Y changes by β_1

Multiple Linear Regression - coefficients

SPRM 3.14

In multiple regression, here for just two predictors, note that

$$\mathrm{E}(Y^{old}) = \beta_0 + \beta_1 X_1^{old} + \beta_2 X_2 \text{ and}$$
$$\mathrm{E}(Y^{new}) = \beta_0 + \beta_1 (X_1^{old} + 1) + \beta_2 X_2,$$

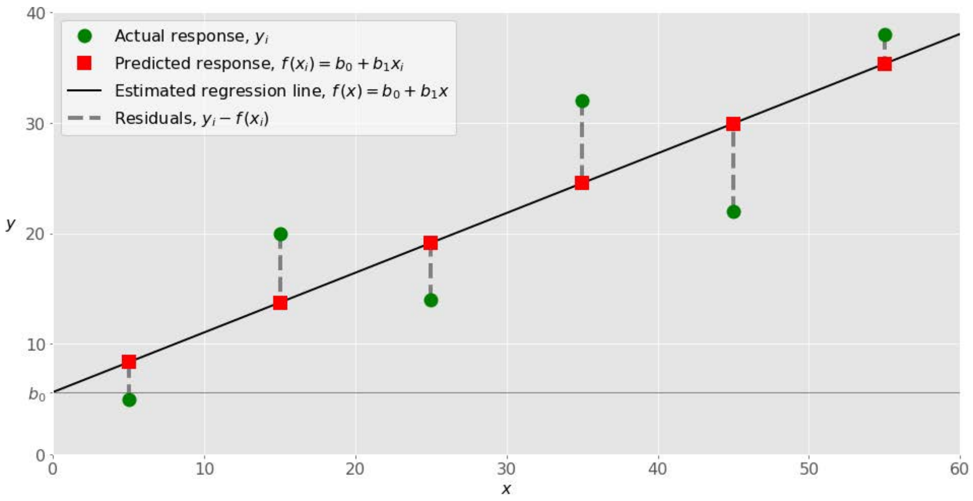
then

$$\mathrm{E}(Y^{new}) - \mathrm{E}(Y^{old}) = \beta_0 + \beta_1 (X_1^{old} + 1) + \beta_2 X_2 - (\beta_0 + \beta_1 X_1^{old} + \beta_2 X_2) = \beta_1.$$

Interpretation: β_1 is the expected change in Y when X_1 increase by 1-unit, with all other covariates (X_2 in this example) unchanged (but present in the model). This is also described as the expected change in Y when X_1 increases by 1 unit, adjusting for all other covariates. More on what this means shortly.

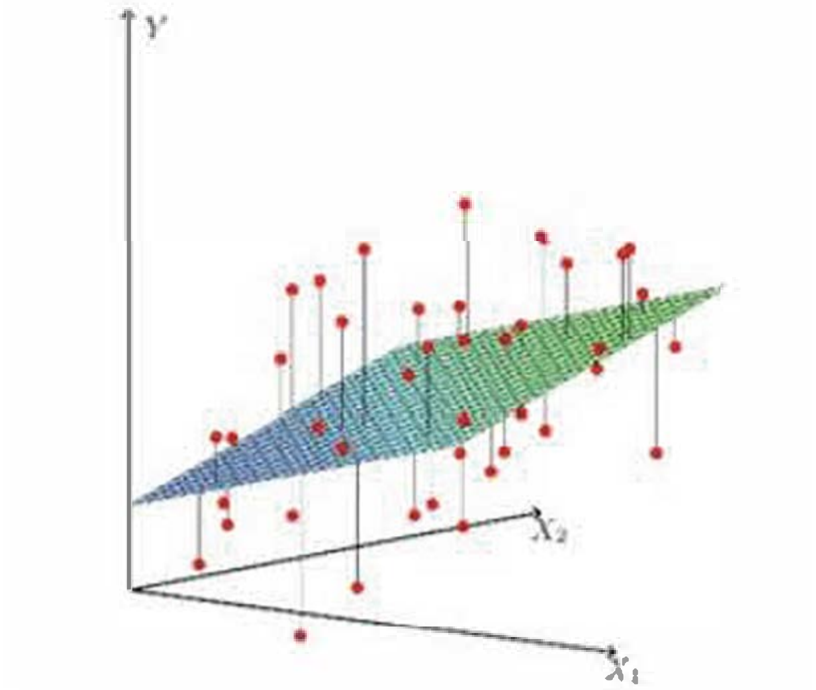
Graphical interpretation of MLR model and its coefficients

- **In SLR** the fitted $\hat{Y}$ all lie on the estimated regression line. β_1 is the slope of the line, β_0 is the intercept. Observed (real) Y values can lie on, above, or below the regression line in the XY plane



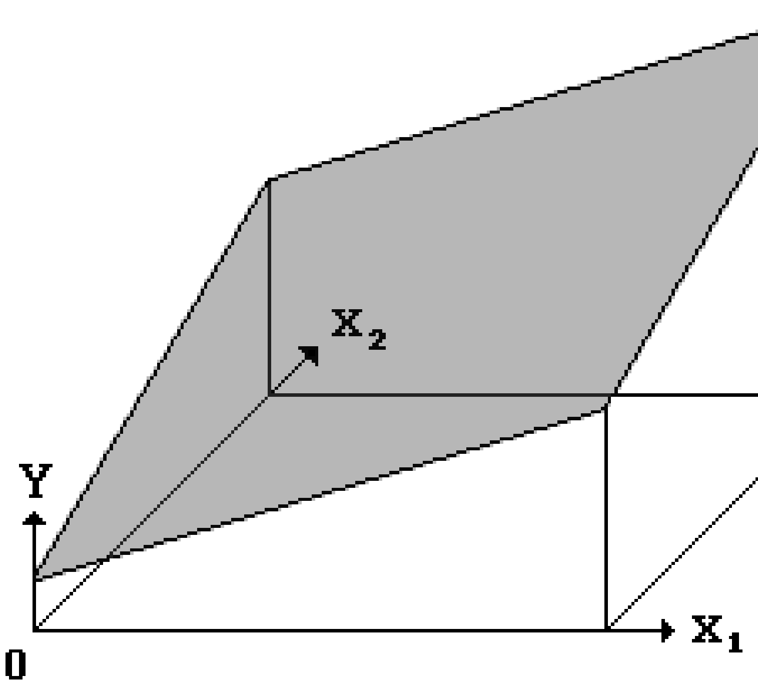
Graphical interpretation of MLR model and its coefficients

- **In MLR**, the fitted $\hat{Y}$ all lie on the same regression *surface.* In MLR with 2 predictors this surface is a plane, with β_1 being the slope of the plane along the X_1 direction, and β_2 the slope of the plane along the X_2 direction. The observed Y lie on, above, or below the regression surface.



Graphical interpretation of MLR model and its coefficients

The interpretation of β_1 is the effect of X_1 on the expected (fitted) Y, but only when we hold the other Xs in the model fixed. In other words, holding all else constant, when we increase X_1 by one unit, the expected Y would change by β_1 units. All other βs in the model are interpreted analogously.



Graphical interpretation of MLR model and its coefficients

How do βs and predicted Y change in MLR?

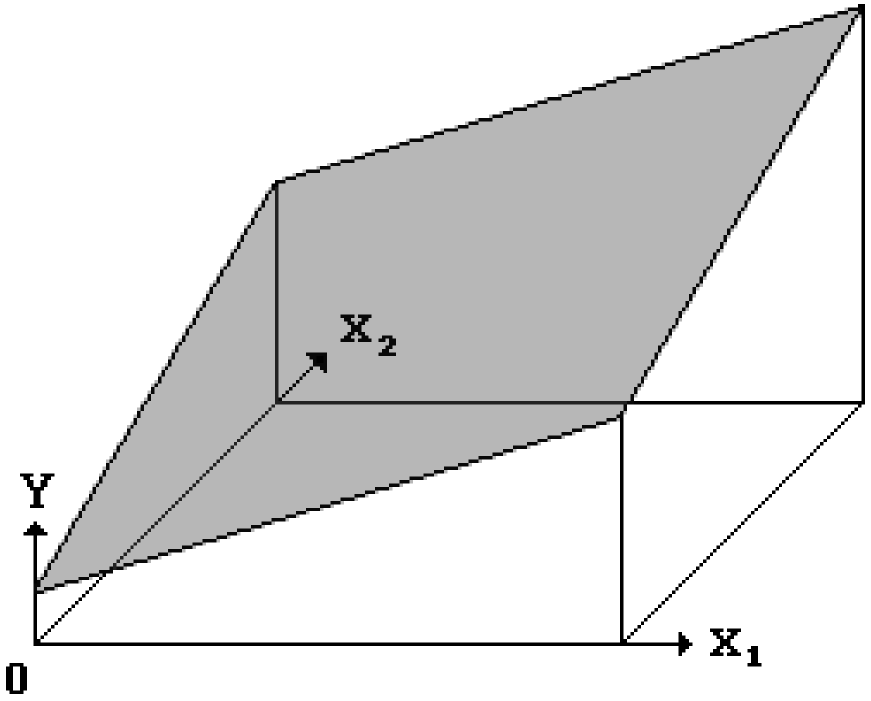
–

– Additional regression assumption is that predictors (Xs) are not highly correlated (called collinearity). If completely correlated, model cannot be estimated.

– In real life, there is some relationship between Xs, as the predictors may come from the same conceptual domain

– Whether correlated or not, effect on Y will be different with multiple predictors vs. one predictor, as the estimate $\hat{Y}$ is the **sum of the effects from each predictor**

To get a sense of what is meant by 'adjusted for', note that the plane is not parallel to either the X_1 or X_1 axis, but changes in both directions (with two meaningful predictors of Y). Note also though that at *any given* X_2, the <u>incremental</u> change due to X_1 is the same, but the resultant predicted Y is not the same, due to X_2s contribution.



Example: Fuel Consumption Data

As an illustration of MLR modeling, we look at a dataset on U.S. fuel consumption (C. Bingham, American Almanac). We will investigate fuel consumption as a linear function of demographic, policy, and other factors. Variables in this dataset are:

```
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
state           str2   %9s                    State
pop             int    %8.0g                  Population in each state
tax             float  %8.0g                  1972 motor fuel tax rate, cents/gal
nlic            int    %8.0g                  1971 thousands of drivers
inc             float  %8.0g                  1972 per capita income, thousands of dollars
road            float  %8.0g                  1971 federal-aid highways, thousands of miles
fuelc           int    %8.0g                  fuel consumption, millions of gallons
dlic            float  %8.0g                  Percent of population with driver's license
fuel            int    %8.0g                  Motor fuel consumption, gal/person
------------------------------------------------------------------------------
```

Example: Fuel Consumption Data

In particular, we will look at the relationship of fuel consumption (variable *fuel*) and fuel *tax* (measured in cents per gallon).

Note that some of the variables are direct transformations of others. There are totals, and per-capita variants of the same variables. Because fuel consumption and number of licensed drivers (variables *fuelc* and *nlic*) are measured for entire states, they will vary with the state population size.

To assure that we work with comparable quantities, we only look at per-capita versions of variables *fuel* and *dlic* (which are *fuelc* and *nlic* divided by population, *pop*).

Example: Fuel Consumption Data

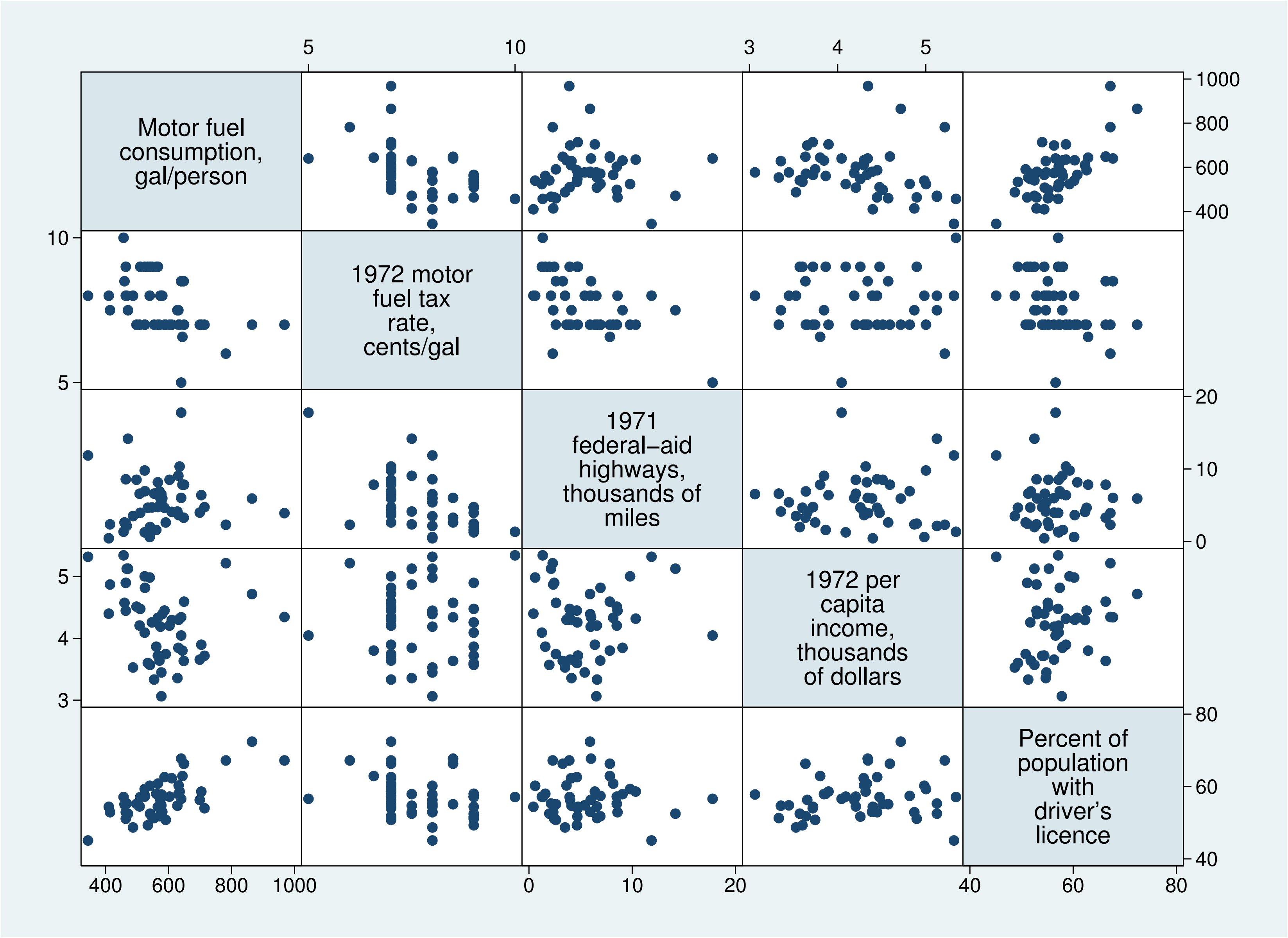
- So for the initial analysis, we will look at the graphical relationship of the following variables which we think might be important in explaining fuel consumption:

  tax, *inc*, *road*, *dlic*.

- The variable *pop* is not considered because its effects will be accounted for (partially) by modeling the per-capita fuel consumption as a function of other per-capita variables (except for the roads). A more thorough analysis can be done with *pop* included.

- The graphical relationships can be best examined by scatterplot matrix, producing all pairwise scatterplots we can make from the data. To produce this matrix of scatterplots in Stata use the following command:

```
. graph matrix fuel tax inc dlic road
```



Scatterplot Matrix: Fuel Consumption Data

We can examine the relationship between per-capita fuel consumption and predictor variables, as well as the relationships amongst the predictor variables themselves. This inter-predictor relationship, known as **collinearity**, will be of greater interest later in this course. For now, just think of it as a nuisance we have to live with.

From the scatterplots in the top row (those describing relationships between fuel and taxes, road length, per-capita income, and the proportion licensed), we can see that the per-capita fuel consumption seems to be linearly related only to two variables, *tax* and *dlic*, while the other variables (*road*, *inc*) seem to have a rather weak, possibly curvi-linear relationship with fuel consumption.

Modeling Fuel Consumption

Hence for now we choose to examine only *tax* and *dlic* and their effect on *fuel*. So our goal is to estimate the model:

$$\widehat{\mathrm{FUEL}} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{DLIC} + \hat{\beta}_2 \mathrm{TAX}$$

Before we go directly to this model, let us consider what adding the single new predictor to a simple linear regression model would do. Specifically, let us consider what would happen if we added *tax* to a SLR model relating *fuel* to *dlic*.

In fact, the main point of adding *tax* variable to the model
$F\widehat{UE}L = \hat{\beta}_0 + \hat{\beta}_1 DLIC$
is to explain the part of *fuel* that hasn't already been explained by *dlic* variable, to assess the influence of taxing fuel. Assessing one predictor after others are accounted for is often the central goal in multiple regression.

Modeling Fuel Consumption

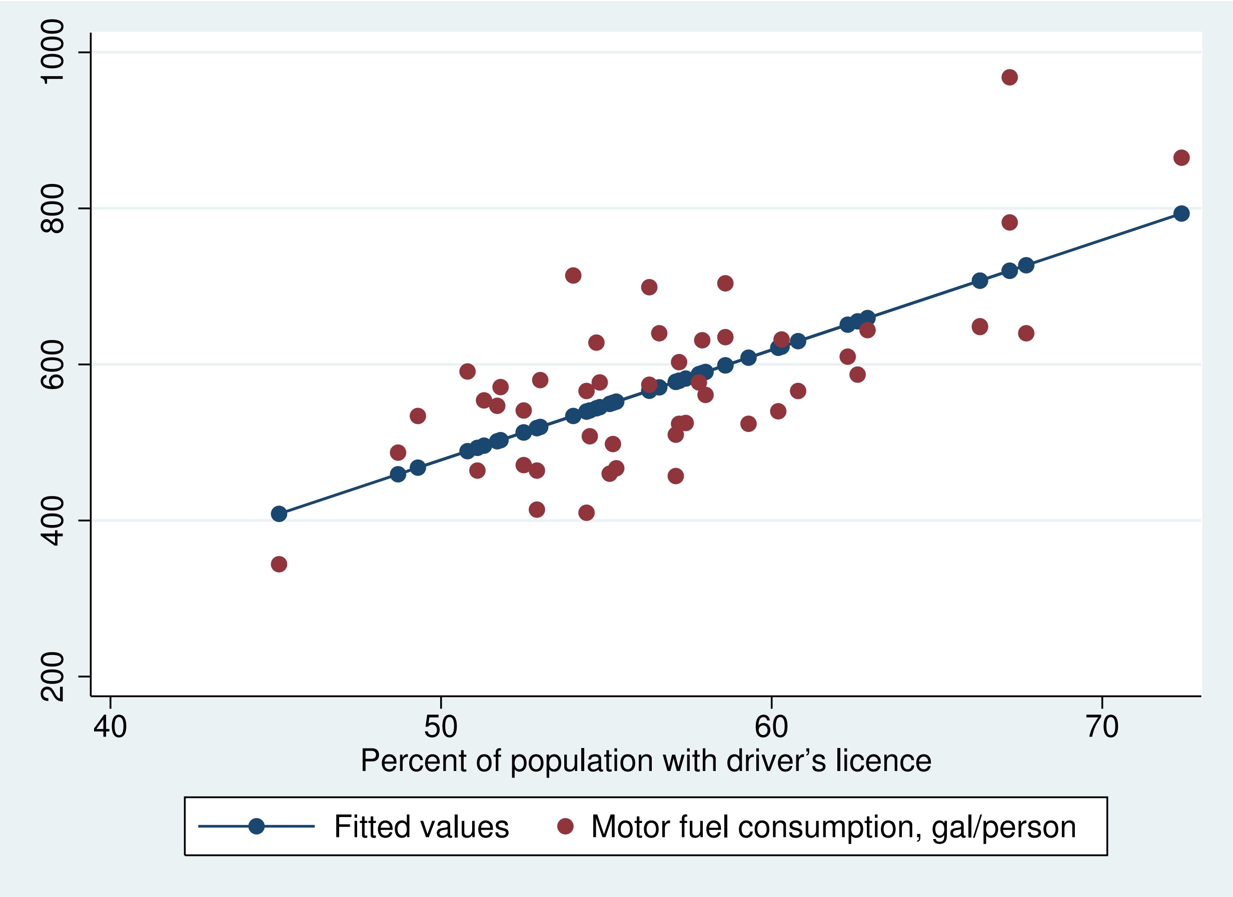
So, we first run the regression predicting fuel use *fuel* using *dlic*.

```
. reg fuel dlic

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  1,    46) =   43.94
       Model |  287447.975     1  287447.975           Prob > F      =  0.0000
    Residual |  300918.504    46   6541.7066           R-squared     =  0.4886
-------------+------------------------------           Adj R-squared =  0.4774
       Total |  588366.479    47  12518.4357           Root MSE      =  80.881

------------------------------------------------------------------------------
        fuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dlic |   14.09842   2.126848     6.63   0.000     9.817298    18.37954
       _cons |  -227.3091   121.8617    -1.87   0.069    -472.6039    17.98576
------------------------------------------------------------------------------
```

We obtain the following fitted model: $\widehat{\text{FUEL}} = -227.3 + 14.1\text{DLIC}$. The $R^2 = 0.49$, indicating that about 50% of variability in *fuel* is explained by *dlic*.

The graph:

```
. predict fhatd
. twoway (scatter fuel dlic) (lfit fuel dlic)
```

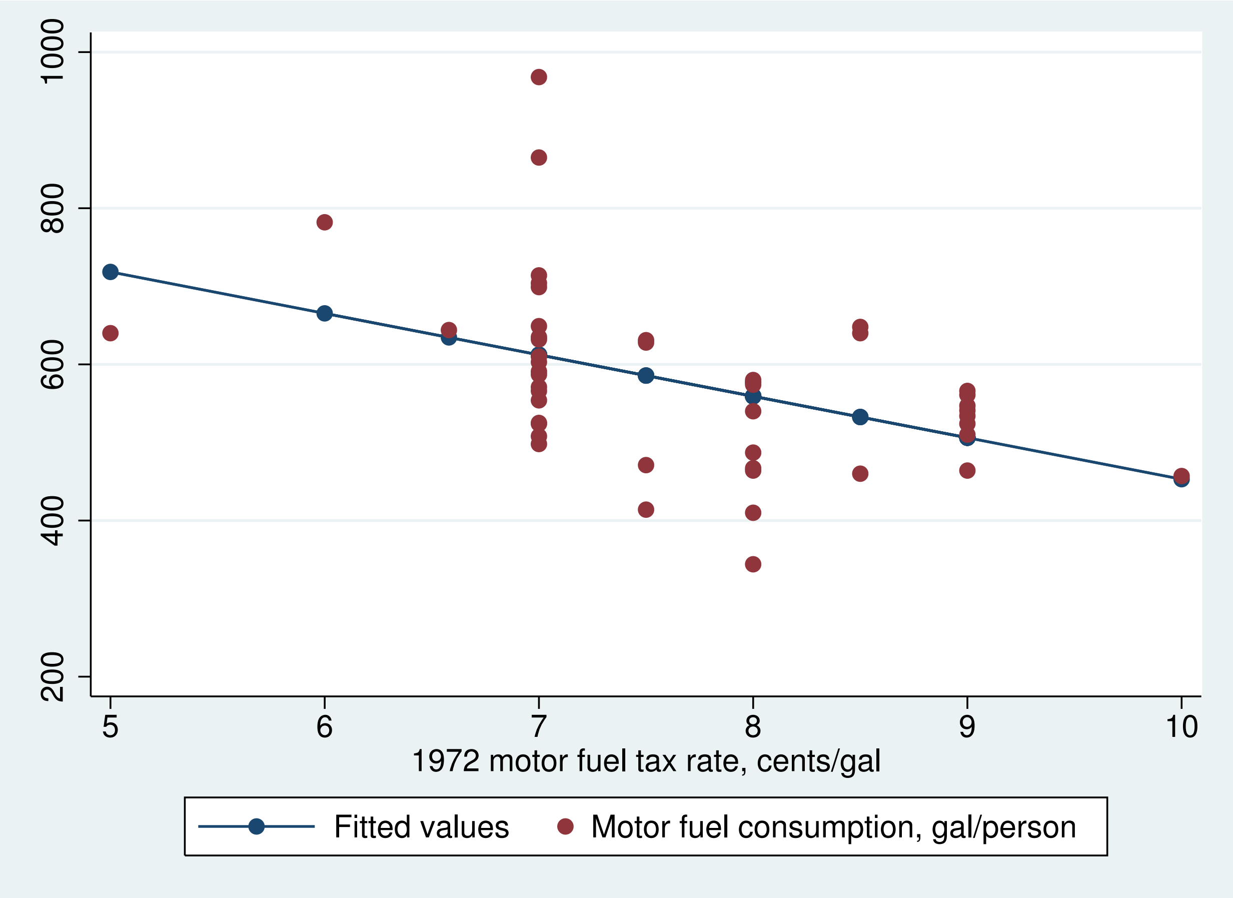


Per 1% increase in licensed driver proportion, about 14 more gallons of fuel is used per year

Modeling Fuel Consumption

The other predictor alone - tax

```
. reg fuel tax

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  1,    46) =   11.76
       Model |   119823.12     1   119823.12           Prob > F      =  0.0013
    Residual |  468543.359    46  10185.7252           R-squared     =  0.2037
-------------+------------------------------           Adj R-squared =  0.1863
       Total |  588366.479    47  12518.4357           Root MSE      =  100.92

------------------------------------------------------------------------------
        fuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |   -53.1063   15.48359    -3.43   0.001    -84.27315   -21.93945
       _cons |   984.0076   119.6236     8.23   0.000     743.2178    1224.797
------------------------------------------------------------------------------
. predict fhatt
(option xb assumed; fitted values)

. twoway (scatter fuel tax) (lfit fuel tax)
```



This regression model predicting *fuel* with *tax* yields:

$$\widehat{\mathrm{FUEL}} = 984.0 - 53.1\mathrm{TAX},$$

Modeling Fuel Consumption

Tax is in cents, so this estimate means that for every penny increase in taxes, the average per-capita fuel consumption goes down by 53 gallons per year.

That amounts to about 1 gallon less per week per person, which would appear to be an effective policy if the objective of the tax is in part to reduce fuel use. On the other hand, there may be a disincentive to making the tax too high

The $R^2 = 0.20$, indicating that about 20% of variability in *fuel* is explained by *tax* alone, **ignoring** *dlic*(leaving it out of the model)

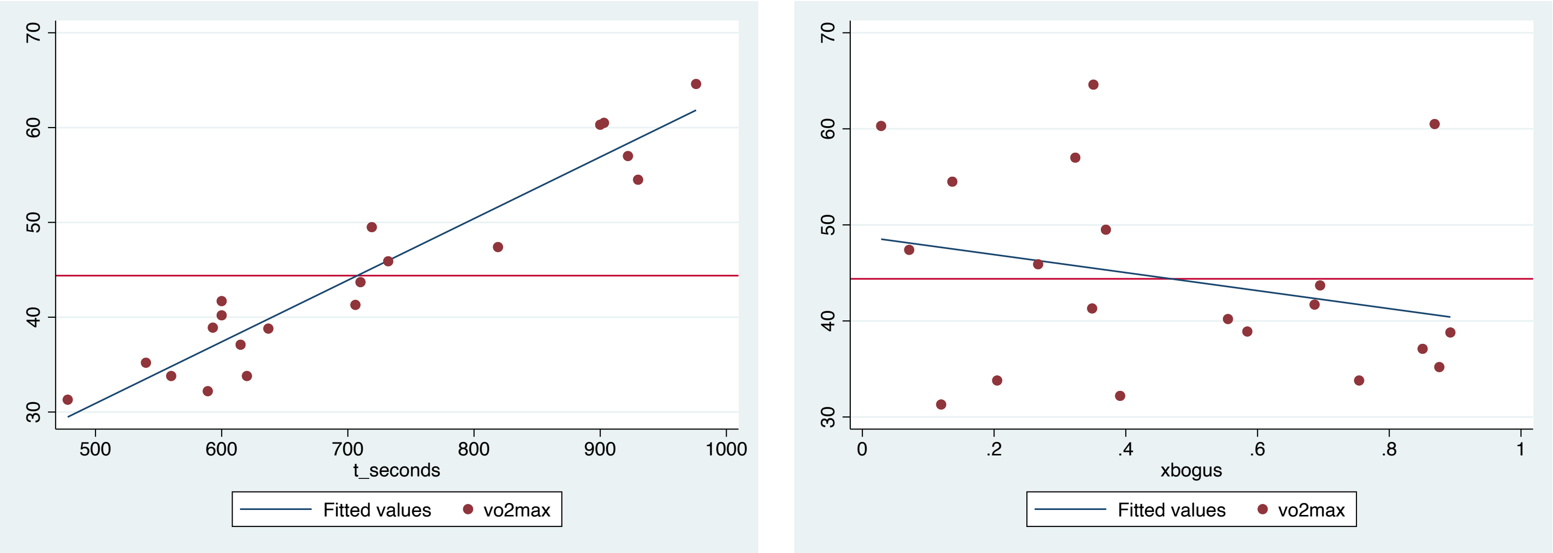
Explained Variation via R^2

Recall in linear regression, we have 3 summaries of the variation in response variable Y:

1. $SSR = \sum(\hat{Y}_i - \bar{Y})^2$ - variation in a predicted Y around the mean of Y. This is how group-specific means vary around the overall mean, since at a given X, $\hat{Y}$ can be thought of as a 'group-specific mean' (recall that model predicts $\mu_y|X$, or mean of Y at X)
2. $SSE = \sum(\hat{Y}_i - Y_i)^2$ - variation between predicted and observed Y_is. This is how group-specific values vary around their group specific mean (again, $\hat{Y}$ is a group-specific mean, with the group being those with specific value of covariate X)
3. $SST = \sum(Y_i - \bar{Y})^2$ - variation of individual Y_is around the overall mean of Y. Defined the same as numerator of $\mathrm{var}(Y)$. **Does not depend on predictor(s) X at all.**

Then $SST = SSR + SSE$ and we derive R^2 as $\frac{SSR}{SST}$ or $1 - \frac{SSE}{SST}$

Explained Variation via R^2 - $VO_2$2MAX data



(a) good fitting model

(b) weak predictor

-**For good model:** SSR $=$ 1815.55, SST $=$ 1997.82 so $R^2 =$ 0.9088, 9-% of variability in VO_2MAX explained by running time

-**For weak model:** SSR $=$ 141.74, SST $=$ 1997.82 (same, this is variance of response variable, and does not change when X changes), so $R^2 = 0.0709$ - 7% explained by null predictor

Explained Variation in Multiple Regression

- **In the fuel consumption study, what can we say about two-variable model based on these two single regressions?**

- Heuristically, we would expect to be able to explain more variability in *fuel* by having both variables in the model, than by having any single variable. **This is a mathematical fact:** R^2 will increase with the number of predictors

- So, the R^2 for the combined model must be larger than 48.9%, the larger of the two individual R^2s. However, the total variation explained will generally NOT be additive.
  $R^2_{comb} \leq R^2_1 + R^2_2 = 48.9\% + 20.4\% = 69.3\%$

- The equality holds only if the two variables *tax* and *dlic* are completely unrelated, and tell us completely separate information about fuel consumption. This is unlikely, and the total will be less than 69.3% if *tax* and *dlic* are related to each other

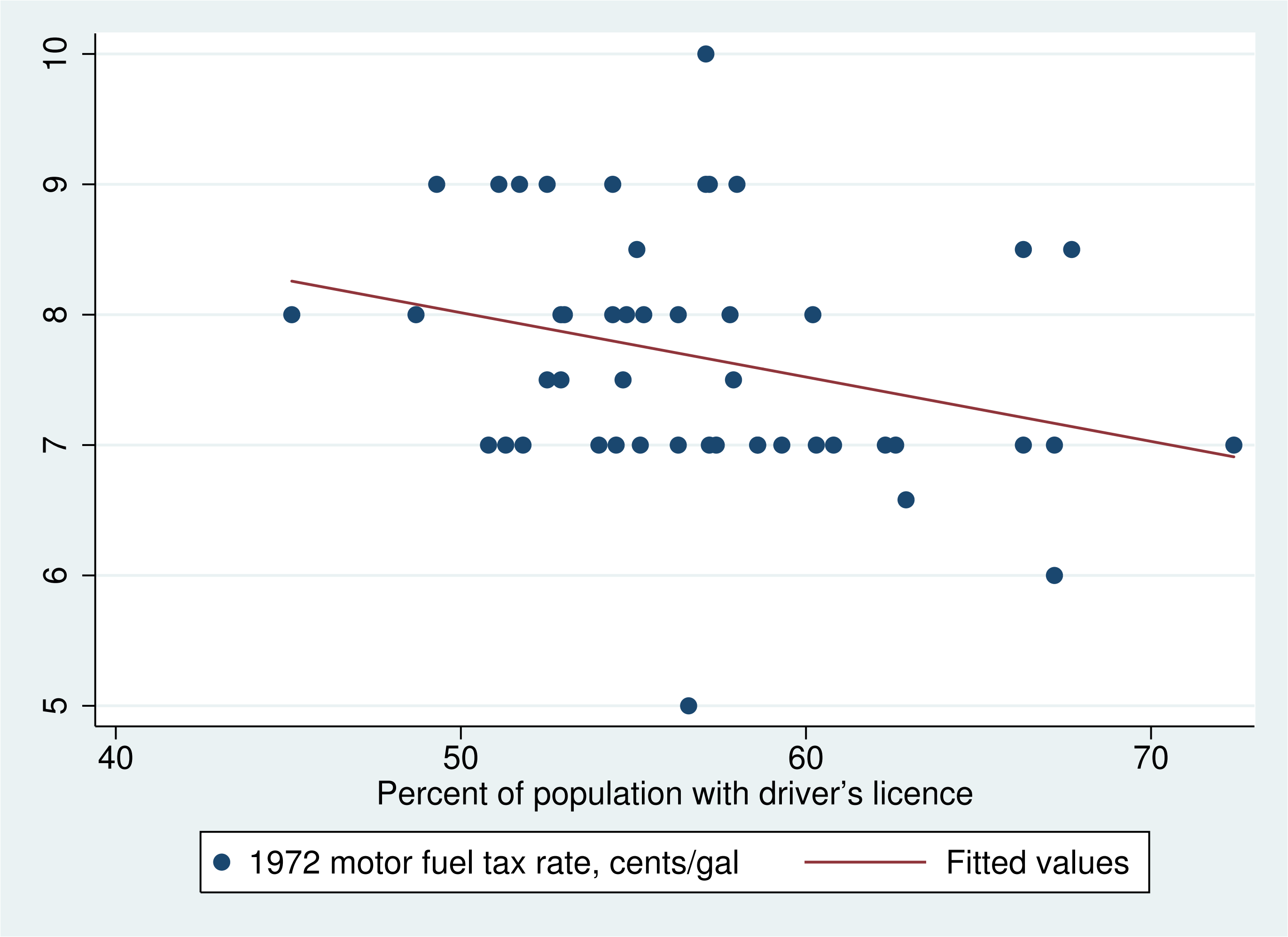
Explained Variation

To understand what the new variable contributes in addition to the one already in the model, we can examine how these two predictor variables relate to each other. We can do this by running a regression of one predictor on the other (this is done for pedagogical purposes here, not something we always do in analyses):

```
. reg tax dlic
      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  1,    46) =    4.16
       Model |  3.52489012     1  3.52489012           Prob > F      =  0.0471
    Residual |  38.9613767    46   .84698645           R-squared     =  0.0830
-------------+------------------------------           Adj R-squared =  0.0630
       Total |  42.4862668    47  .903963124           Root MSE      =  .92032

------------------------------------------------------------------------------
         tax |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dlic |  -.0493701   .0242008    -2.04   0.047    -.0980837   -.0006564
       _cons |   10.48407   1.386628     7.56   0.000     7.692936    13.27521
------------------------------------------------------------------------------
```

And examining the fit and scatterplot:

```
. twoway (scatter tax dlic) (lfit tax dlic)
```



Explained Variation

We see that there is a relationship here, so these two variables, which are both associated with fuel consumption, are associated with each other (weakly so).

This is heuristically why there is a lack of additivity of the R^2s. Yet, as both variables are importantly related to fuel consumption, the model should be improved by using both.

Let's go to the two-variable model

The Two-variable MLR Model

```
. reg fuel dlic tax

      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]
-------------+----------------------------------------------------------------
        dlic |   12.51486   2.090614     5.99   0.000     8.304147    16.72557
         tax |  -32.07532   12.19716    -2.63   0.012    -56.64166   -7.508984
       _cons |   108.9709   171.7859     0.63   0.529    -237.0236    454.9655
------------------------------------------------------------------------------
```

Both predictors remain important: $R^2 = 0.56$

Interpreting coefficients in SLR and MLR: Marginal vs. Partial coefficients

- When we model FUEL via TAX only using a SLR, we ignore all other variables. The model and coefficient depicts the "marginal" relationship between FUEL and TAX. ignoring all other variables.

- The interpretation of β_2 from the MLR is that it describes the effect of TAX FUEL adjusted for DLIC. Sometimes called a partial coefficient.

- What this means is that for any given level of DLIC, one-penny increase in taxes will result in β_2 change in expected fuel consumption. If you take any two states with the same percentage of drivers, and if they differ only by one penny in their fuel taxes, then you would expect their fuel consumption to differ by β_2 gallons per capita.

- **How much different is the effect of tax after accounting for driver's license proportion?**
  - SLR model: $\beta_{tax} = -53.1063.$ So, 53 gallon decrease per cent tax increase
  - MLR model: $\beta_{tax} = \beta_2 = -32.07532.$ 32 gallon decrease per cent tax increase
- **The effect of tax is attenuated** by accounting for the fact that fuel consumption increases simply as a function of the proportion of the state with driver's licenses.
- **The 2-variable model is 'better'** in that it may give a more accurate impression of the effect of tax.

Inference in MLR

In multiple linear regression, inference is expanded to reflect the greater number of parameters estimated and hypotheses of interest.

There are three types of tests one will use in MLR inference:

1. Testing whether one single coefficient is 0 (or some value of interest).
2. Testing whether all coefficients (other than the intercept) are simultaneously 0 (whole model is 'null')
3. Testing whether several coefficients are simultaneously 0 or equal some value(s)

To carry out the testing, we need to state the assumed distribution of regression errors. Most often, this is a $N(0, \sigma^2)$ distribution.

Inference in MLR

Recall that in SLR

- We assume normally distributed residuals, and we estimate the variance of these via MSE.
- Our least-squares estimators of the coefficients, $\hat{\beta}$s, have defined standard errors that are functions of the MSE and X and follow a t distribution with $n - 2$ degrees of freedom. The standard errors of the $\hat{\beta}$ are part of the regression model output

Inference in MLR

- **Same holds in the MLR case**, except that we have a more general expression for degrees of freedom for the t distributions applied to the β's,

  This general expression accounts for the number of parameters (i.e., βs) to be estimated with the data. For n independent sets of observations $(Y_i, X_{i1}, X_{i2}, \ldots, X_{ip})$, the degrees of freedom parameter the t critical values associated with testing individual βs is

  $$n - p - 1$$

  where (p is the number of predictors in the model).

  Let's now examine each of the three testing scenarios separately, assuming that $\epsilon \sim N(0, \sigma^2)$, with $\sigma^2)$ estimated from the data

Inference in MLR - individual coefficients

1. **Testing whether a single coefficient is 0.**

   This basically tests whether one particular predictor matters in our model. The hypotheses are:

   $H_0 : \beta_i = 0, \text{ other } \beta\text{s arbitrary}$
   $H_1 : \beta_i \neq 0, \text{ other } \beta\text{s arbitrary}$

   Ways of testing this hypothesis:

   (1) `STATA` and R output gives t-statistic and the p-value for each coefficient. So, look at the t-statistic for β_i, and compare its p-value to the significance level of the test, α, or report the p-value directly. Or if looking up significance level the old-fashioned way, the reference t dist'n is one with $n - p - 1$, or $48 - 4 - 1 = 43$ for the 4-parameter model

Inference in MLR - individual coefficients

For a model with a few more predictors (adding road miles and per capita income)

```
. 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
------------------------------------------------------------------------------
```

Inference in MLR - individual coefficients

Equivalently, examine the confidence interval associated with that particular β_i. If the confidence interval contains 0, do not reject the null. You can produce any level of confidence interval in Stata by adding ", `level(conflevel)`" to your regression command. For example: `reg Y X1 X2, level(90)` will produce 90% confidence intervals for all coefficients.

```
. reg fuel tax dlic inc road, level (90)

--omitted
------------------------------------------------------------------------------
        fuel |      Coef.   Std. Err.      t    P>|t|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
         tax |  -34.79016    12.9702    -2.68   0.010    -56.59398   -12.98633
        dlic |   13.36449   1.922982     6.95   0.000     10.13182    16.59716
         inc |  -66.58875   17.22175    -3.87   0.000    -95.53973   -37.63777
        road |   -2.42589   3.389175    -0.72   0.478    -8.123332    3.271552
       _cons |   377.2913   185.5412     2.03   0.048     65.38337    689.1991
------------------------------------------------------------------------------
```

Testing in Multiple Regression - all coefficients

2. **Testing whether all coefficients are simultaneously 0.**

   This is basically testing whether your entire model matters. The hypothesis setup is:
   $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$
   H_1 : AT LEAST ONE $\beta \neq 0$

   The null hypothesis can be interpreted as the model with intercept only, and the alternative as the model with intercept and at least one of the predictors in it. This is an overall test of model worth.

   We use the overall F-test, given in STATA on the top of the ANOVA table (the top table in the output of the regress command). The F-statistic and associated p-value are provided.

Testing in Multiple Regression - all coefficients

```
. 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
```

This is the omnibus test of whether the model has any significant predictors

What? The F-test? And why do we need it?

Recall (from introductory Statistics), if we have several groups to compare means among, we can

(a) **Make all pairwise comparisons** - number of comparisons for k groups is

$$\frac{k!}{2!(k-2)!} \equiv {}_kC_2 \equiv \text{“k choose 2”} \equiv \frac{k!}{2!(k-2)!}$$

(b) **Perform an omnibus test** of

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

versus

$$H_A : \mu_i \neq \mu_j$$

The latter approach is preferred to control $multiplicity$, that is, multiple significance tests, which increases the chance that one or more will be 'significant' by chance.

The F-test

- Such a test can be formulated by partitioning variability in the data into *sources of variation*. This is called Analysis of Variance (ANOVA) for this reason (even though we are comparing means)
  - ∗ Specifically, the variance of group-specific means around the mean overall (MSR) and variance of observations within groups around their group-specific mean (MSE)
  - ∗ If the variability in the former is large relative to the latter, then 'groups matter', or the population means are likely different among the groups
- the ratio of these two quantities (which are both estimates of the variance overall)

$$F = \frac{MSR}{MSE}$$

Follows an F-distribution,

The F-test

- F - a continuous probability density with two parameters, corresponding to degrees of freedom for the numerator and denominator for the variance estimates, or $k-1$ and $n-k$ It is typically asymmetric, with a long right ‘tail’
- The F-statistic will be close to 1.0 under the null hypothesis and will be large under the alternative hypothesis (that some means differ from others)
- **If we have only two groups, then the F-test reduces to the two-sample t-test**, and thus serves as a generalization of the latter for more than two groups

The F-test in Linear Regression

- For testing multiple coefficients simultaneously, the F-test is more useful, as it provides a natural way to control for *multiple testing*. Alternatively, we can use methods such as adjusting the α or Type I error level to account for the fact that the error level increases above the desired level when more than one hypothesis test is conducted.
- The degrees of freedom parameter values for the given F test depend on the number of coefficients being testing, ranging from 1 to p
- **See SPRM Section 3.15** - the ANOVA table associated with the regression model

Testing in Multiple Regression - all coefficients

```
. 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
```

$F = \frac{99829.12}{434396.51} = 22.71$

In regression models

– numerator DF parameter is always the number of parameters, so $p = 4$ here

– denominator DF parameter is $n - p - 1 = 43$

Testing in Multiple Regression - all coefficients

Alternatively, use the “`test`” command in Stata. After your regression command (and after you see the regression table come up on the screen), type:

`test X1 X2 ...  Xp` (to test all, make sure you list all your predictor variables here)

This command gives you a p-value, which you should compare to the significance level α.

```
. test tax dlic inc road

 ( 1)  tax = 0
 ( 2)  dlic = 0
 ( 3)  inc = 0
 ( 4)  road = 0
       F(  4,    43) =   22.71
            Prob > F =    0.0000
```

MLR continues next time . . .

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:

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

where

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

We then compute the test as

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

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**

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)

Testing in Multiple Regression -
-subsets of coefficients-

(SPRM Chapter 4)

– **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)

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.

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

– 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.

– 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:

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

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}

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**.
- The FM degrees of freedom is always $n - p - 1$ (full model minus intercept and all predictors).

Contrasting Models via F-tests

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

  $H_0 : \beta_3 = \beta_4 = 0$ (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:

Contrasting Models via F-tests - Example

```
. 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
------------------------------------------------------------------------------
```

```
. test road inc

 ( 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
      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
```

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.

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.
  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.

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 $\mathrm{corr}(y, x) = \mathrm{corr}(Y, \hat{Y})$

- 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 $\mathrm{corr}(Y, \hat{Y})$ 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
```

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
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.

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)

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)?
- What are the consequences of omitted important variables? If fewer parameters is better for MSE estimate, what is the trade-off?

```
. 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

. 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
------------------------------------------------------------------------------
```

Note: In smaller model:

- R^2 is only slightly reduced- 0.6749 (small model) vs. 0.6787 (full model)
- 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)
- 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

      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
------------------------------------------------------------------------------
```

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)

– 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
- 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.
- **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
------------------------------------------------------------------------------
```

```
. 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
------------------------------------------------------------------------------
```

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

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

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

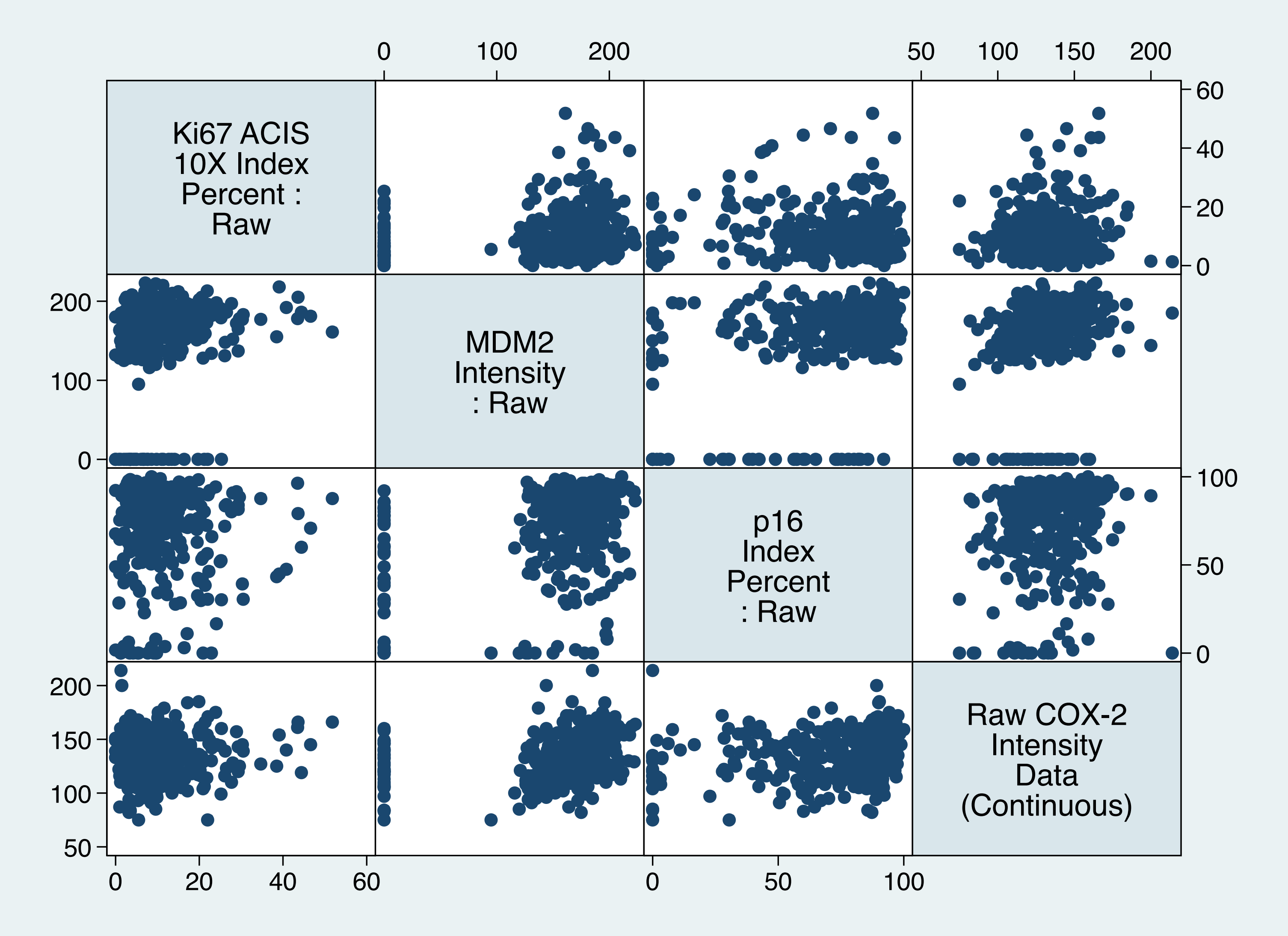
In this analysis, Ki-67 emerged as an important molecular factor.
We might ask:

- 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?

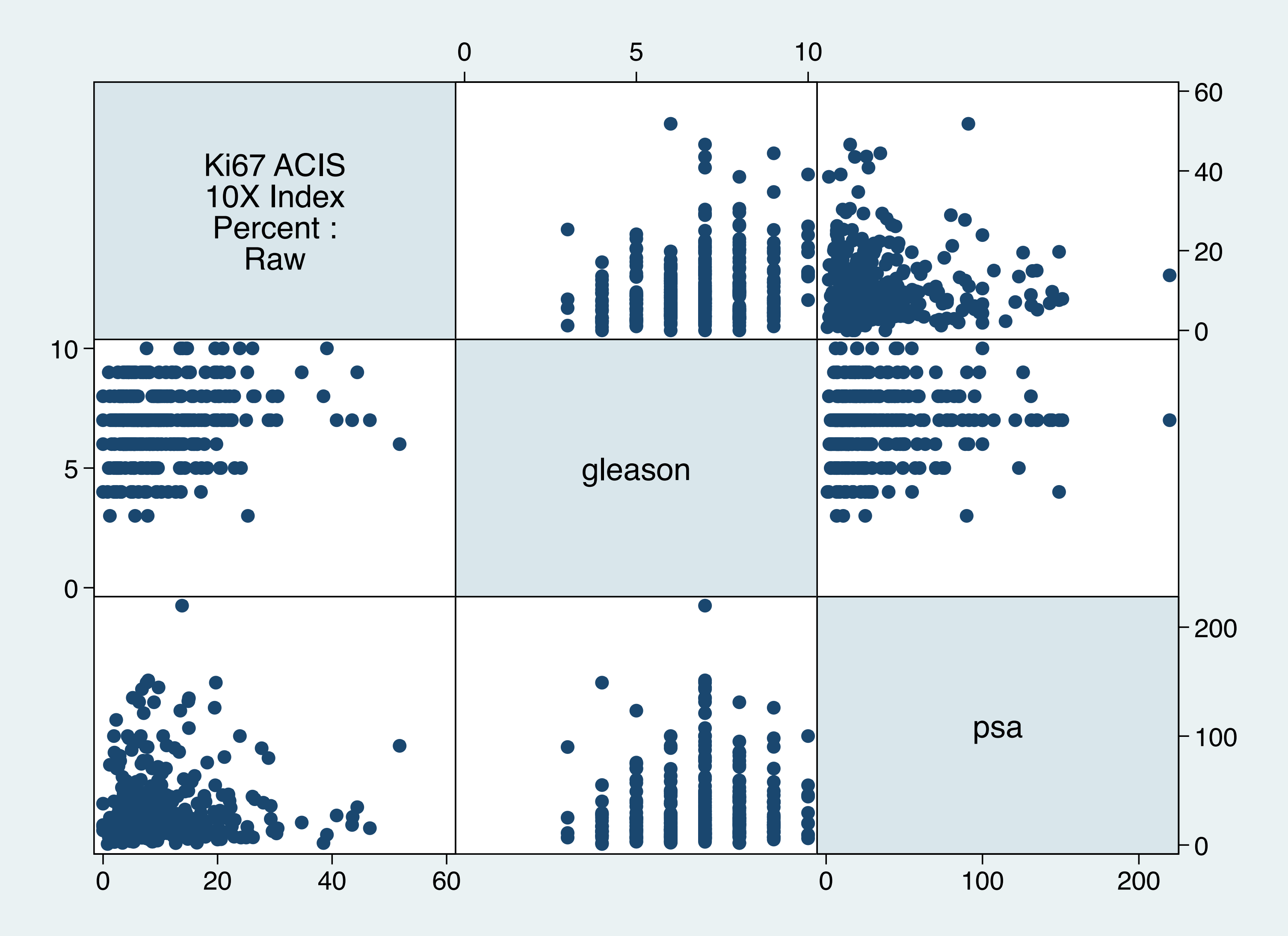
Regression Analysis of Prostate Cancer Biomarkers

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



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 provide a means to conduct additional tests
  - Must balance model complexity and fit, predictive ability