PBHS 32410 / STAT 22401

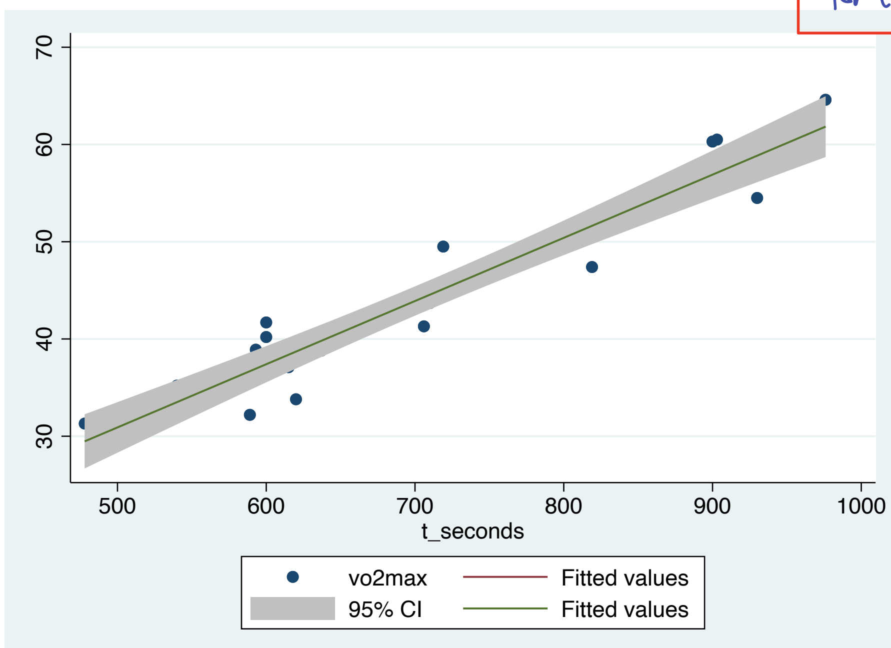
Linear Regression

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

CI on the mean VO_2MAX:

More accurate!

95% probability that the true best-fit line for the population lies within the confidence interval.

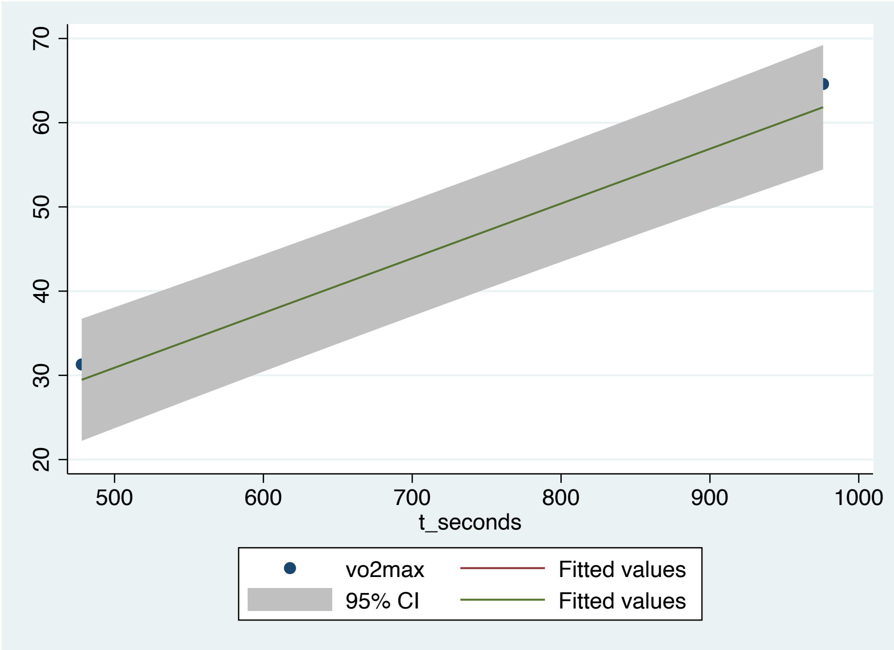


- Predicts mean value.
- Focuses on past and current events.

95% of Y-values to be found for a certain X-value will be within the interval range around the linear regression line.

Linear Regression

Forecast interval for any given running time:



- Less Certain than confident interval
- Predicts individual number.
- Focuses on future events.

Linear Regression

The standard error expressions:

Mean prediction CI

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

Forecast always bigger than CI.

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

What happens when

- new X value far from $\bar{X}$? Equal $\bar{X}$? away from mean ↓

  X value far from $\bar{X}$: ↑ variability, ↑ s.e.
  X value = $\bar{X}$: ↓ variability, ↓ s.e.

- n very large? Large n: ↓ s.e. (↑ precise), more narrow CI.

- Fit very good? Model is accurately capturing patterns of data: s.e ↓, prediction is closer to data.

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. → ex) Confounding, interaction

Multiple Variables of X to predict one Y.

- 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).

"predictor"

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

$$Y_i = \beta_0 + \underbrace{\boxed{\beta_1 X_{1,i}}}^{1^{st}} + \overbrace{\boxed{\beta_2 X_{2,i}}}^{2^{nd}} + \ldots + \beta_p X_{p,i} + \epsilon,$$

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

i: the #th observation out n observations

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 :

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

Sum of squared differences between observed values and the predicted values (values of beta coefficients) to minimize.

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:

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

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

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

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

Formula for vector Y. Matrix X Vector β

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

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: $\mathbf{Y} = \boxed{\mathbf{X}\boldsymbol{\beta}} + \epsilon$

Single Independent Variable

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 of X the columns of X (this is the sum of squared differences in matrix notation)

(Do not need to know Matrix calculation)

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)

Inverse of the product of transpose X and X

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

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

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

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

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

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:

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

$$\mathrm{E}(Y^{new}) = \beta_0 + \beta_1 (X^{old} + 1),$$ Increment X by 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 hence, the "slope"

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

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

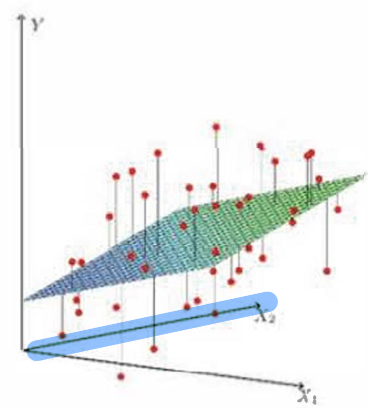
Interpretation: β_1 is the expected change in Y when X_1 increase by 1-unit, with all other covariates (X_2 in this example) unchanged (but present in the model). This is also described as the expected change in Y when X_1 increases by 1 unit, adjusting for all other covariates. More on what this means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

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

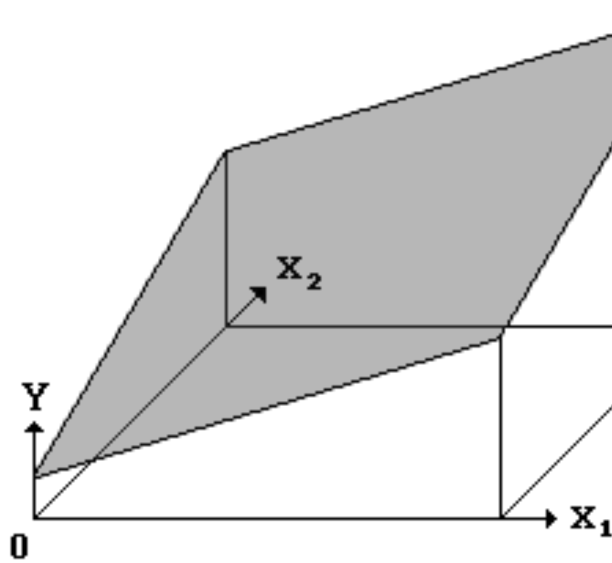
- **In MLR**, the fitted $\hat{Y}$ all lie on the same regression *surface.* In MLR with 2 predictors this surface is a plane, with β_1 being the slope of the plane along the X_1 direction, and β_2 the slope of the plane along the X_2 direction. The observed Y lie on, above, or below the regression surface.

Multiple dimensions



Graphical interpretation of MLR model and its coefficients

The interpretation of β_1 is the effect of X_1 on the expected
All other βs unchanged
(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.

  Representing the same information

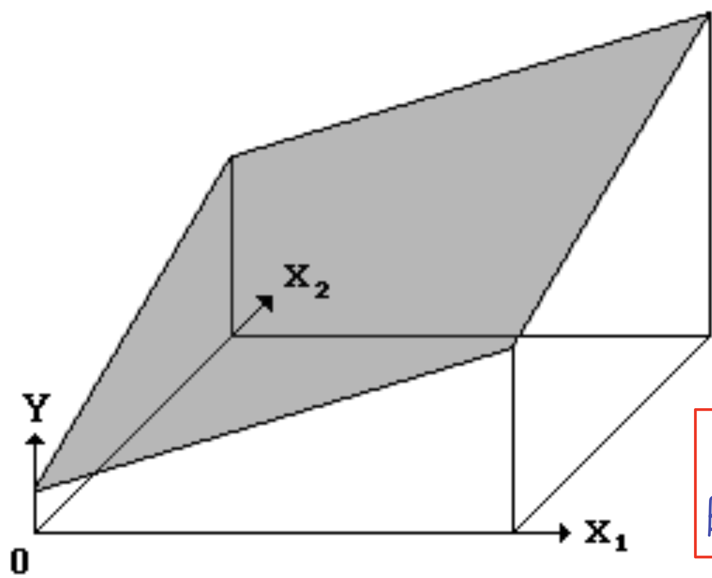
  Essentially just "linear" of eachother

- In real life, there is some relationship between Xs, as the predictors may come from the same conceptual domain Xs Related, but not functions of eachother.

- 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**

Y depends on all the variables within the function.
But variables should NOT be highly correlated with eachother.

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 incremental change due to X_1 is the same, but the resultant predicted Y is not the same, due to X_2s contribution.



X^1 and X^2 are differerent but not parallel to e.o.
Both affect Y : Sum of Y.

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:

variable name	storage type	display format	value 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 Y (variable *fuel*) and fuel *tax* (measured in cents per gallon). X

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*).

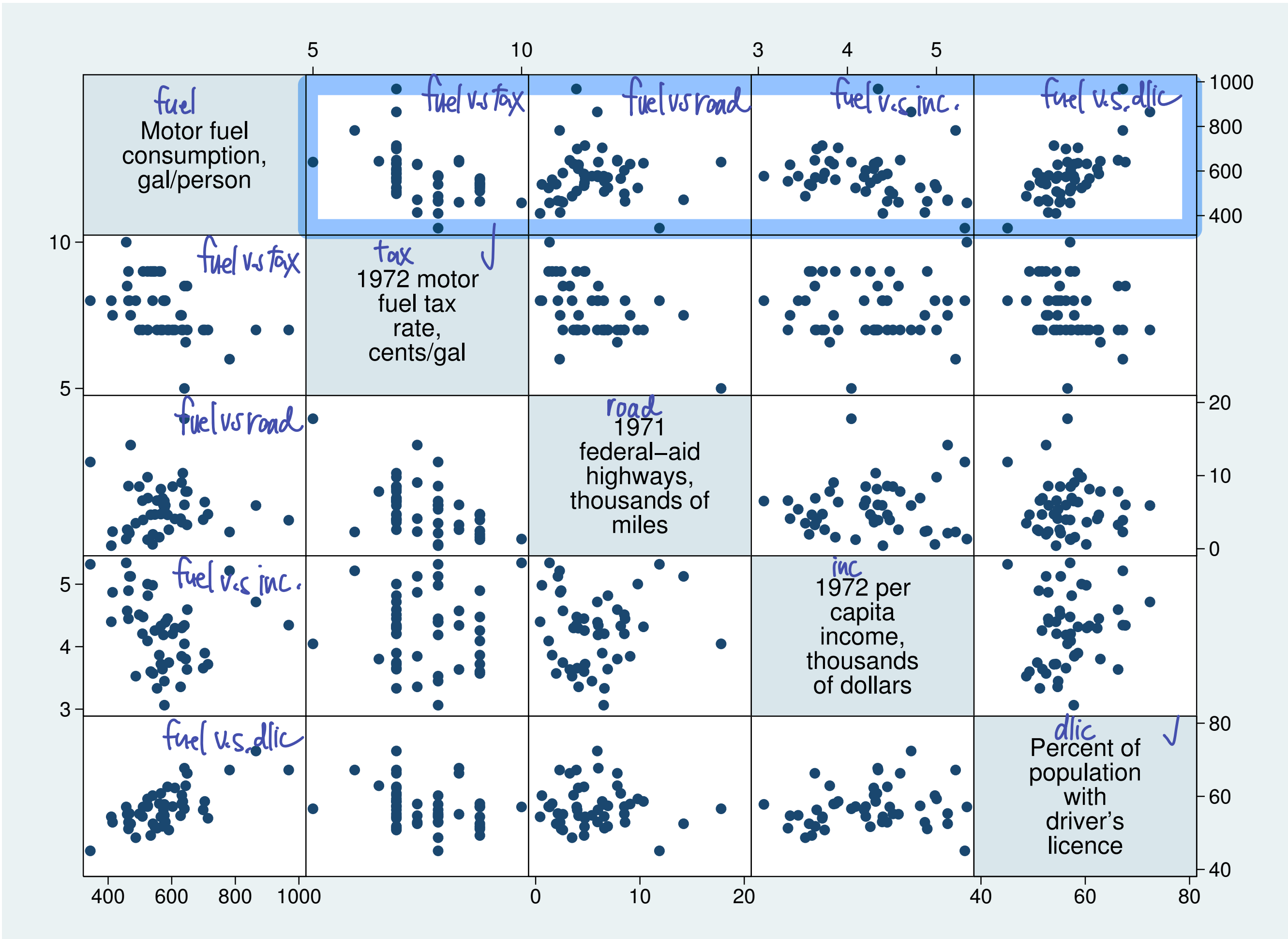
Example: Fuel Consumption Data

- 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:
  β_1 β_2 β_3 β_4
  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:

$$\text{. graph matrix } \overbrace{\text{fuel}}^{Y}\ \overbrace{\text{tax inc dlic road}}^{X\text{s}}$$



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

Underlying influence that's not captured

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.

Driver's License only

Modeling Fuel Consumption

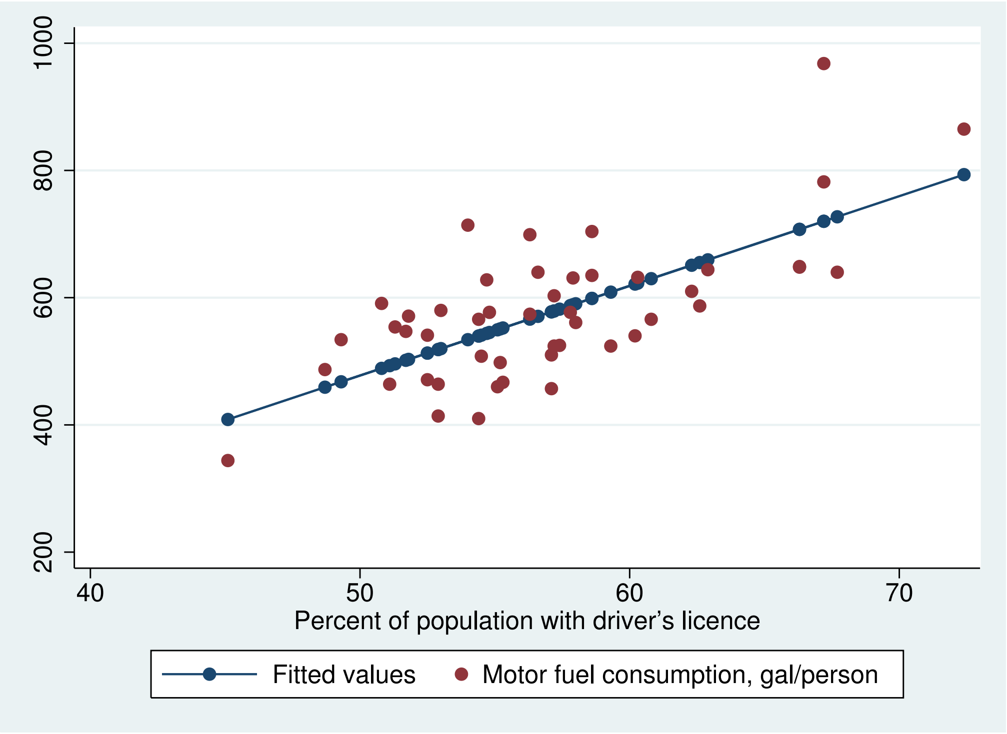
So, we first run the regression predicting fuel use *fuel* using *dlic*.

```
. 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]
-------------+----------------------------------------------------------------
  β1    dlic |   14.09842   2.126848     6.63   0.000     9.817298    18.37954
  β0   _cons |  -227.3091   121.8617    -1.87   0.069    -472.6039    17.98576
------------------------------------------------------------------------------
```

We obtain the following fitted model:
$\widehat{\mathrm{FUEL}} = -227.3 + 14.1\mathrm{DLIC}$. 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)
```



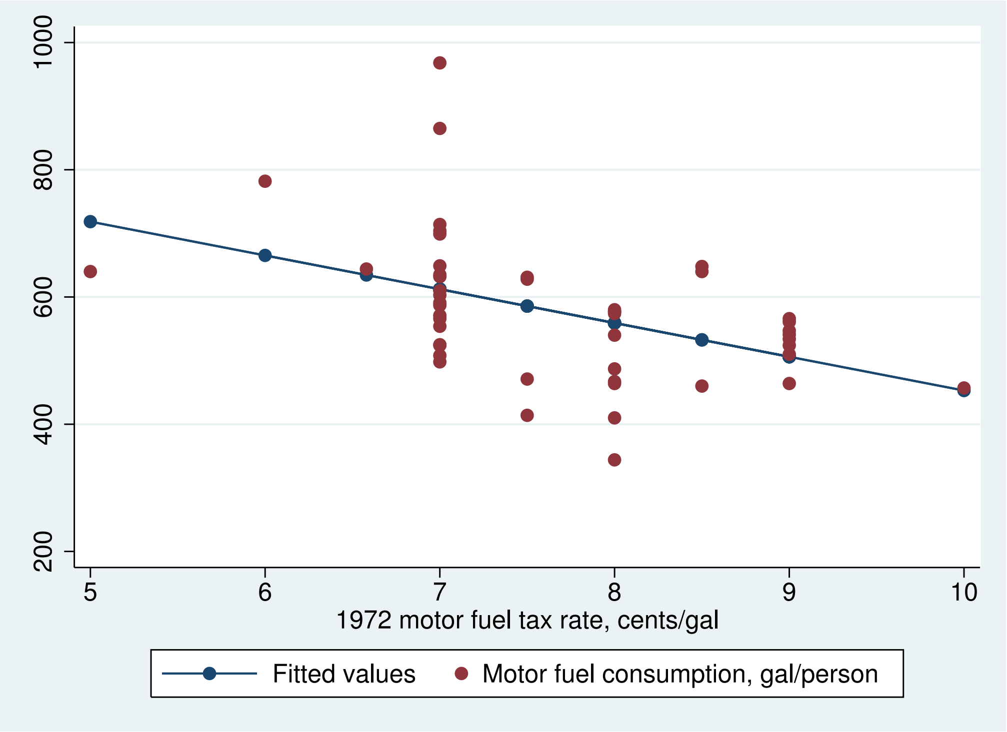
Per 1% increase in licensed driver proportion, about 14 more gallons of fuel is used per year

Tax only

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]
-------------+----------------------------------------------------------------
    β1   tax |  -53.1063   15.48359    -3.43   0.001    -84.27315   -21.93945
    β0 _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:

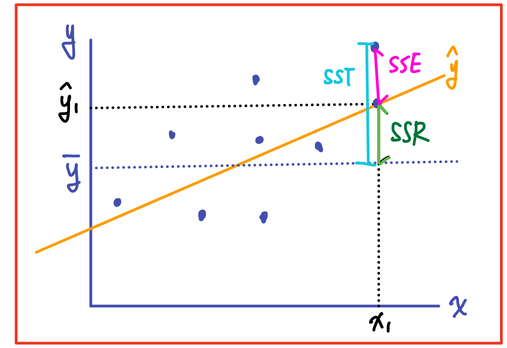
$$\widehat{\text{FUEL}} = 984.0 - 53.1\text{TAX},$$

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:

"Line to bar"

Sum of Squares Regression 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) ↑ Better

"Line to point"

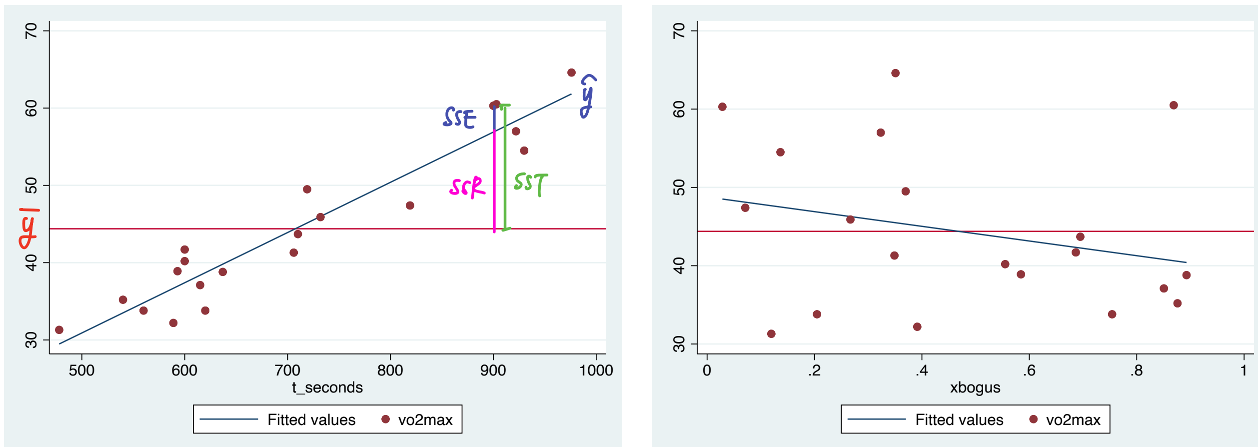
Sum of Squares Error 2. $SSE = \sum(\hat{Y}_i - Y_i)^2$ - variation between predicted and observed Y_is. This is how group-specific values vary around their group specific mean (again, $\hat{Y}$ is a group-specific mean, with the group being those with specific value of covariate X) ↓ Better

"Point to bar"

Sum of Squares Total 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 $\boxed{SST = SSR + SSE}$ and we derive $\boxed{R^2 \text{ as } \frac{SSR}{SST} \text{ or } 1 - \frac{SSE}{SST}}$

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.

$$R^2_{comb} \leq R^2_1 + R^2_2 = \underset{0.4886}{48.9\%} + \underset{0.2037}{20.4\%} = 69.3\%$$

- 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):

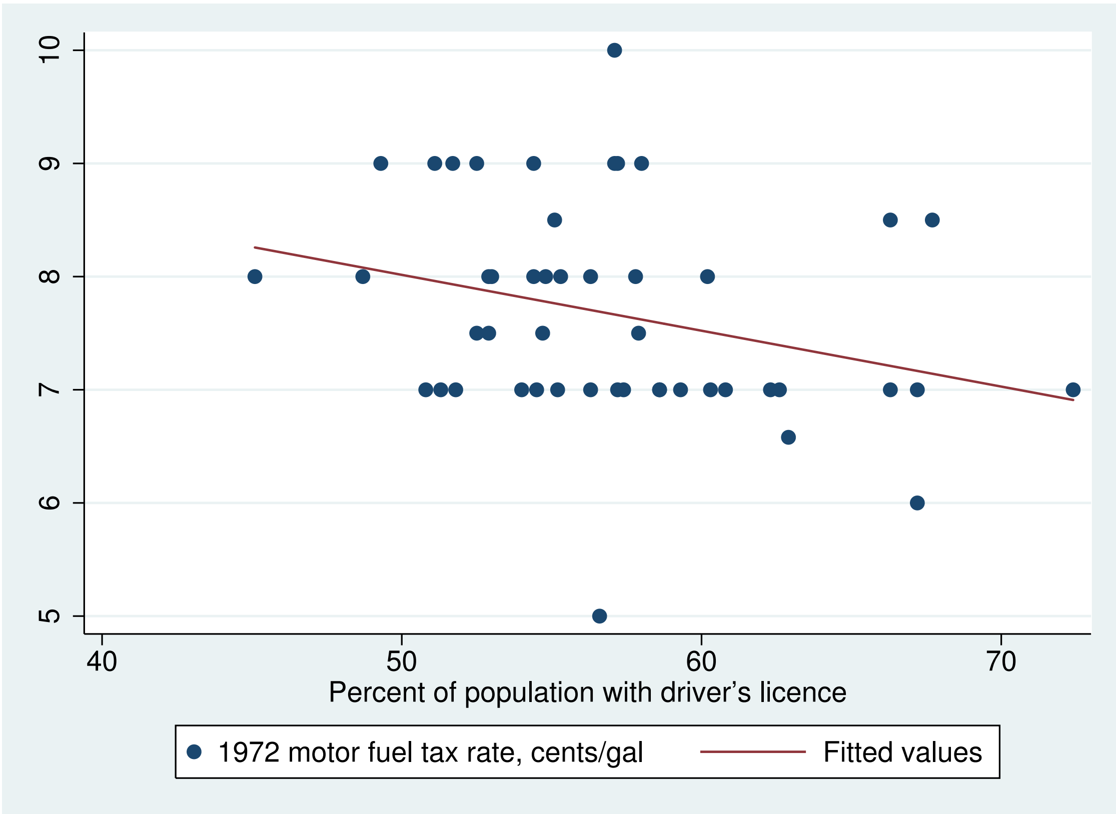
no fuel

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

Two X variables are affecting eachother: NOT completely unrelated.

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: $R^2 = 0.56$

Less than the addition of two Rs from the individual variables, but shows improvement.

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" "Lack of" 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 "Not the main slope, β_1" 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 dof: n-2

- 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 d.f: n-p-1

- **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).
  βs

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

   $H_0 : \beta_i = 0, \text{ other } \beta\text{s arbitrary}$ (other βs remain unchanged)
   $H_1 : \beta_i \neq 0, \text{ other } \beta\text{s arbitrary}$

   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)

$48-4-1=43$

β_1 β_2 β_3 β_4 $\}4$

```
. 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 $\beta \neq 0$

   Multiply by 0=0 so only β_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

Multiple Comparison Test

H_0: reduced model sufficient

H_A: full model provides better fit

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

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

Overall Test

H_0: All βs $= 0$

H_A: At least one $\beta \neq 0$

(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 ↑MSR relative to the latter, then 'groups matter', or the population means are likely different among the groups ↑F, ↓p-value, significant.

– the ratio of these two quantities (which are both estimates of the variance overall)

$$\boxed{F = \frac{MSR}{MSE}} \begin{matrix}\uparrow \\ \downarrow\end{matrix}$$

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** SLR **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

```
         Y    β1   β2   β3   β4
. reg fuel tax dlic inc road

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

$$F = \frac{\overset{MMR}{99829.12}}{\underset{MME}{434396.51}} = 22.71$$

In regression models

– numerator DF parameter is always the number of parameters, so

β1 β2 β3 β4

$p = 4$ here

48 - 4 - 1

– denominator DF parameter is $\underbrace{\boxed{n - p - 1}}_{\#\text{ of obs.}} = 43$

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
          p     n-p-1
       F(  4,    43) =   22.71
            Prob > F =    0.0000
```

MLR continues next time . . .