SINGLE EQUATION LINEAR MODEL WITH CROSS SECTIONAL DATA: OLS

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Asymptotic Results for the Linear Model
2. Practical Regression Hints
3. Linear Regression as the best Mean Squared Error Approximation

1. ASYMPTOTIC RESULTS FOR THE LINEAR MODEL

- The workhorse in empirical research of all kinds is still a model linear in parameters. The model stated in terms of a (well-defined) population is

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u \\ &= \beta_0 + \mathbf{x}\boldsymbol{\beta} + u, \end{aligned} \tag{1}$$

where **x** is 1 × K and observed, and **β** is the $K \times 1$ vector of unknown "slope" parameters.

- Equation (1) is fairly general, as **x** can include nonlinear functions of underlying variables, such as logarithms, squares, reciprocals, log-odds, and interactions.

- Example:

$$\begin{aligned}\log(wage) = {} & \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper \\ & + \beta_4 exper^2 + \beta_5 female + u.\end{aligned}$$

Violations: (i) A model nonlinear in parameters *may* be more appropriate (for example, when the range of y is restricted, such as binary, fractional, or nonnegative). (ii) Perhaps the coefficients on the independent variables should also be viewed as random variables (although this does not prevent us from writing (1) with a complicated error term).

- In what follows, we assume that we can collect a random sample – that is, independent and identically distributed outcomes – from the underlying population.Given randomly sampled observations $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$ satisfying the population model (1), we can write for the random draws

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK} + u_i, i = 1, \ldots, N, \tag{2}$$

where N is the sample size.

- Later we will discuss violations of random sampling, including (i) stratified sampling; (ii) missing data (sample- or self-selection?); (iii) cluster sampling.
- For panel data applications we will explicitly allow a time dimension and discuss how random sampling applies.

- For notational convenience, it is often useful to absorb the intercept into the vector **x** and write

$$y = \mathbf{x}\boldsymbol{\beta} + u \tag{3}$$

where **x** is $1 \times K$, with the convention that the first element x_1 is (almost always) unity.

- None of the main large-sample results rely on $x_1 \equiv 1$, but it is almost always true in practice.

- For a random draw i we write $y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i$.

- With random sampling, the remaining assumptions can (and should) be stated in terms of the population.

Assumption OLS.1 (Zero Correlation): The error has a zero mean and is uncorrelated with each explanatory variables:

$$E(\mathbf{x}'u) = \mathbf{0}. \tag{4}$$

with the last equality a normalization (with an intercept in the model).

- (4) is sometimes called “orthogonality conditions.”

- Because **x** almost always has an intercept, (4) is practically the same as

$$E(u) = 0, Cov(x_j, u) = 0, j = 2, \ldots, K. \tag{5}$$

Violations: This is subtle, because one can use (3) and (4) to simply define **β**, as we will see. Nevertheless, when we begin with an underlying “structural” model, (4) is often violated by (i) omitted variables; (ii) measurement error; (iii) simultaneity.

- Sufficient for (4) [or (5)] is the stronger zero conditional mean assumption,

$$E(u|\mathbf{x}) = E(u) = 0. \tag{6}$$

- Under (3) and (6), we have specified $E(y|\mathbf{x})$:

$$E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} = \beta_1 + \beta_2 x_2 + \ldots \beta_K x_K. \tag{7}$$

- The difference between (4) and (6) is substantive: under (4), discussions of “functional form misspecification” is meaningless, while (6) means that all functions of the covariates affecting the population regression $E(y|\mathbf{x})$ have been accounted for in our choices of $x_2, \ldots, x_K$.

- In most cases we hope to have (6) when the explanatory variables are "exogenous" – typically, if a nonlinear function of a regressor is statistically and practically significant, we leave it in the model – but in reality we should probably settle for (4). For example, when y has discreteness, or its range is limited in some important way – say $y \in \{0,1\}$, $0 \leq y \leq 1$, $y > 0$ – the linear model for $E(y|\mathbf{x})$ cannot hold over a wide range of x_j, but it might provide a useful approximation.

- How might a linear model “approximate” a general regression function? Let $\mu(\mathbf{x}) = E(y|\mathbf{x})$ denote the true regression function. For emphasis we separate out the intercept. The linear projection of y on $(1, \mathbf{x})$, denoted

$$L(y|1, \mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta},$$

is such that the parameters solve

$$\min_{a,\mathbf{b}} E[(y - a - \mathbf{x}\mathbf{b})^2].$$

It is easily shown (later) that α and $\boldsymbol{\beta}$ also solve

$$\min_{a,\mathbf{b}} E[(\mu(\mathbf{x}) - a - \mathbf{x}\mathbf{b})^2].$$

- In other words, the parameters in the linear projection $L(y|1, \mathbf{x})$ provide the best (population) mean square error approximation to the true regression function.
- If we make some strong assumptions we can say more about the interpretation of the slope parameters, β_j.
- First, for a continuous variable x_j, its **partial effect** is

$$\frac{\partial \mu(\mathbf{x})}{\partial x_j},$$

which is a function of **x**.

- Its **average partial effect** (averaging across the distribution of **x** is)

$$APE_j = E_{\mathbf{x}}\left[\frac{\partial\mu(\mathbf{x})}{\partial x_j}\right] \equiv \gamma_j$$

- For a discrete change in, say, x_K, we must specify two values (such as zero and one for a binary variable), take the difference, and the average out the other explanatory variables:

$$APE_K(x_K^{(0)},x_K^{(1)}) = E_{\mathbf{x}_{(K)}}\left[\mu(x_1,\ldots,x_{K-1},x_K^{(1)}) - \mu(x_1,\ldots,x_{K-1},x_K^{(0)})\right]$$

where $\mathbf{x}_{(K)} \equiv (x_1,\ldots,x_{K-1})$.

- In any case, APE_j is a constant (parameter).
- Is it possible that a linear regression consistently estimates the APEs?

Let $\boldsymbol{\beta}$ be the vector of slope parameters in $L(y|1,\mathbf{x})$. Stoker (1986, *Econometrica*) showed that if $\mathbf{x}$ has a multivariate normal distribution then

$$\beta_j = APE_j$$

for all j.

- Of course, multivariate normality is very strong and usually unrealistic, but it suggests that linear regression more generally approximates quantities of interest: APEs.
- We can allow nonconstant partial effects in regression by using flexible functions of the covariates, so we need not settle for approximating only estimate partial effects averaged across the distribution of the covariates.
- Back to the population representation

$$y = \mathbf{x}\boldsymbol{\beta} + u, \; E(\mathbf{x}'u) = \mathbf{0}$$

Assumption OLS.2 (No Perfect Collinearity): In the population, there are no exact linear relationships among the covariates:

$$rank\, E(\mathbf{x}'\mathbf{x}) = K. \tag{8}$$

Violations: None in interesting applications. High correlation among regressors often cannot be avoided, but not a violation of assumptions. Sometimes high correlation among regressors (multicollinearity) *is* the researcher's fault because parameterization has not been carefully chosen. (Example later.)

- When an intercept is included, (8) says the population variance-covariance matrix of the regressors is invertible.

- Under OLS.1 and OLS.2, $\boldsymbol{\beta}$ is *identified*, that is, we can write it as a function of population moments in observable variables:

$$\begin{aligned}
\mathbf{x}'y &= (\mathbf{x}'\mathbf{x})\boldsymbol{\beta} + \mathbf{x}'u \\
E(\mathbf{x}'y) &= E(\mathbf{x}'\mathbf{x})\boldsymbol{\beta} + E(\mathbf{x}'u) \\
E(\mathbf{x}'y) &= E(\mathbf{x}'\mathbf{x})\boldsymbol{\beta} \text{ by OLS.1} \qquad (9) \\
\boldsymbol{\beta} &= [E(\mathbf{x}'\mathbf{x})]^{-1}E(\mathbf{x}'y) \text{ by OLS.2} \qquad (10)
\end{aligned}$$

- This has nothing to do with data! $\mathbf{A} \equiv E(\mathbf{x}'\mathbf{x})$ is a $K \times K$ matrix of variances and covariances in the population; $E(\mathbf{x}'y)$ is essentially a $K \times 1$ vector of population covariances.

• Now apply the logic of method of moments given the random sample: replace population means with sample means:

$$\hat{\boldsymbol{\beta}} = \left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'y_i\right) \tag{11}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{12}$$

where **X** is $N \times K$ with i^{th} row $\mathbf{x}_i$ and **Y** is $N \times 1$ with i^{th} entry y_i.

• (12) is fine for computations and finite-sample analysis, but large-sample properties are derived from (11).

Key Result 1: Under Assumptions OLS.1 and OLS.2, OLS on a random sample is consistent (as $N \to \infty$) for $\boldsymbol{\beta}$: $\text{plim}_{N\to\infty}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

- Can work directly off of (11):

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= plim\left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'y_i\right)\right] \\
&= plim\left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\right]plim\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'y_i\right) \\
&= \left(plim\, N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}plim\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'y_i\right) \\
&= [E(\mathbf{x}'\mathbf{x})]^{-1}E(\mathbf{x}'y) = \boldsymbol{\beta}
\end{aligned}$$

- Or, work off of the error term:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right)$$

and use a similar argument.

- Note: Under $E(u|\mathbf{x}) = 0$, the OLS estimator is unbiased conditional on $\mathbf{X}$: $E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$. This is because, under random sampling, $E(u_i|\mathbf{X}) = 0$ for all i under OLS.1.

- One way to discover our treatment of OLS so far. Define the linear projection of y on $\mathbf{x}$ (where typically $x_1 = 1$) as $L(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = [E(\mathbf{x}'\mathbf{x})]^{-1}E(\mathbf{x}'y)$ is the solution to

$$\min_{\mathbf{b}} E[(y - \mathbf{x}\mathbf{b})^2].$$

(Note where OLS.2 is used, and OLS.1 holds defintionally from this view.) Then OLS consistently estimates $\boldsymbol{\beta}$.

Assumption OLS.3 (Homoskedasticity): With $\sigma^2 = E(u^2)$,

$$E(u^2\mathbf{x}'\mathbf{x}) = \sigma^2 E(\mathbf{x}'\mathbf{x}). \qquad (13)$$

- Says that u^2 is uncorrelated with each x_j and all functions $x_j x_h$ for all j and h (including $j = h$).
- Sufficient is

$$E(u^2|\mathbf{x}) = \sigma^2. \qquad (14)$$

- If we start with $E(u|\mathbf{x}) = 0$, then (14) is the same as

$$Var(u|\mathbf{x}) = Var(u) \equiv \sigma^2, \qquad (15)$$

which essentially gets us to the Gauss-Markov assumptions for cross section data.

Violations: Whether OLS.3 is satisfied is always an empirical issue. Homoskedasticity is often violated, especially if the range of y is limited in some way (especially discrete).

- If $\mathbf{x}\boldsymbol{\beta}$ represents a linear projection and $E(y|\mathbf{x}) \neq \mathbf{x}\boldsymbol{\beta}$, heteroskedasticity (violation of Assumption OLS.3) is almost certain: even if $Var(y|\mathbf{x})$ is constant, $E(u^2|\mathbf{x}) = Var(y|\mathbf{x}) + [\mu(\mathbf{x}) - \mathbf{x}\boldsymbol{\beta}]^2$, and the second term is a function of $\mathbf{x}$ if $\mu(\mathbf{x}) \neq \mathbf{x}\boldsymbol{\beta}$. (Write $y = \mu(\mathbf{x}) + e$ and $u = y - \mathbf{x}\boldsymbol{\beta} = e + [\mu(\mathbf{x}) - \mathbf{x}\boldsymbol{\beta}]$, square both sides, and then condition on $\mathbf{x}$. Note that $E(e|\mathbf{x}) = 0$ and so e is uncorrelated with any function of $\mathbf{x}$.)

- To get the limiting distribution of OLS:

$$\sqrt{N}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i \right). \tag{16}$$

- Now, by the central limit theorem for i.i.d. random vectors,

$$N^{-1/2} \sum_{i=1}^{N} \mathbf{x}_i' u_i \overset{d}{\to} Normal(\mathbf{0}, \mathbf{B}) \tag{17}$$

$$\mathbf{B} = Var(\mathbf{x}_i' u_i) = E(u_i^2 \mathbf{x}_i' \mathbf{x}_i). \tag{18}$$

- An implication of (17) is $N^{-1/2}\sum_{i=1}^{N}\mathbf{x}_i' u_i = O_p(1)$. After a little algebra, and using $O_p(1)\cdot o_p(1) = o_p(1)$,

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{A}^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{x}_i' u_i\right) + o_p(1) \tag{19}$$

where

$$\mathbf{A} = E(\mathbf{x}_i'\mathbf{x}_i) \tag{20}$$

is $K \times K$ and nonsingular by OLS.2. Therefore, from (17) and (19),

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{\rightarrow} Normal(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}).$$

- This variance matrix, $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, is a “robust sandwich” form: it does not assume homoskedasticity.
- Roughly, we act as if

$$\text{“}Var(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/N\text{”}$$

which shrinks to zero at rate $1/N$, just like the variance of a sample average.

- If we add OLS.3, then

$$\mathbf{B} = \sigma^2\mathbf{A}, \tag{21}$$

and the usual OLS inference is asymptotically valid.

Key Result 2: Under Assumptions OLS.1, OLS.2, and OLS.3,

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{\to} Normal(\mathbf{0}, \sigma^2 \mathbf{A}^{-1}). \tag{22}$$

$$\hat{\sigma}^2 = (N - K)^{-1} \sum_{i=1}^{N} \hat{u}_i^2 \overset{p}{\to} \sigma^2 \tag{23}$$

$$\hat{\mathbf{A}} = N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \overset{p}{\to} \mathbf{A}, \tag{24}$$

where $\hat{u}_i \equiv y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$ are the OLS residuals.

- Our estimate of $\sigma^2\mathbf{A}^{-1}/N$ is

$$\hat{\sigma}^2(\mathbf{X}'\mathbf{X}/N)^{-1}/N = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1},$$

the usual formula for estimating $Var(\hat{\boldsymbol{\beta}})$ under the Gauss-Markov assumptions.

- The Gauss-Markov assumptions imply $Var(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ and $E(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$. Under the assumptions we have made, $\hat{\boldsymbol{\beta}}$ is not even unbiased in general. But the same formula works for estimating its asymptotic variance.

- Where is normality used? Not for large-sample analysis. The assumption

$$u|x_1,\ldots,x_K \sim Normal(0,\sigma^2), \tag{25}$$

which implies $E(u|\mathbf{x}) = 0$ and $Var(u|\mathbf{x}) = \sigma^2$, does imply that the MLE is the best linear unbiased estimator (conditonal on $\mathbf{X}$), but this is very strong. Normality underlies exact inference: with random sampling, (25) gives us the classical linear model assumptions. But normality is needed for large-sample inference.

- Note: The CLT does *not* say anything about the population distribution of u. The distribution of u in the population is fixed and has nothing to do with the size of the sample we draw. The CLT implies that standardized sample averages, such as $\sqrt{N}\bar{u} = N^{-1/2}\sum_{i=1}^{N} u_i$, has an approximate $Normal(0, \sigma^2)$ distribution for large N.

- To make inference robust to arbitrary heteroskedasticity, just drop OLS.3. Then

$$\hat{\mathbf{B}} = (N-K)^{-1}\sum_{i=1}^{N}\hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i \xrightarrow{p} \mathbf{B} \tag{26}$$

whether or not OLS.3 holds.

- Avar($\hat{\boldsymbol{\beta}}$) is estimated with a "sandwich" form:

$$\begin{aligned}\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) &= \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N \qquad (27)\\ &= \frac{N}{(N-K)}\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i\right)\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\end{aligned}$$

- The factor $N/(N - K)$ is a finite-sample correction. There are others.
- Remember, $\hat{\boldsymbol{\beta}}$ is still the OLS estimator. We are adjusting the inference for OLS.
- Note: R-squared is perfectly valid as a goodness-of-fit measure under heteroskedasticity: R^2 is a consistent estimate of $\rho^2 = 1 - \sigma_u^2/\sigma_y^2$, which is a function of the unconditional variances. Whether $Var(u|\mathbf{x})$ is constant is irrelevant for estimating ρ^2.

- **Question**: Suppose in the equation $y = \mathbf{x}\boldsymbol{\beta} + u$, OLS.1 holds, that is, $E(\mathbf{x}'u) = \mathbf{0}$. Suppose we think $Var(u|\mathbf{x}) = \omega(\mathbf{x})$, where $\omega(\mathbf{x})$ is a known function (and this may or may not be the correct variance function). Is the weighted least squares estimator, that is, the solution to

$$\min_{\mathbf{b}} \sum_{i=1}^{N}(y_i - \mathbf{x}_i\mathbf{b})^2/\omega(\mathbf{x}_i)$$

generally consistent for $\boldsymbol{\beta}$? What if $E(u|\mathbf{x}) = 0$?

2. PRACTICAL REGRESSION HINTS

● Do not always attempt to maximize *R*-squared, adjusted *R*-squared, or some other goodness-of-fit measure. Might include in **x** factors that should not be held fixed.

EXAMPLE: y is individual or family demand for a product, $x_1, \ldots, x_{k-1}$ include various product prices, income, and demographics. Should we include the demand for a competing product as x_k? Usually does not make sense to hold a quantity demanded fixed and change the price of any good.

- It is possible to obtain a convincing estimate of a causal effect with a low *R*-squared. For example, under random assignment, a simple regression estimate consistently estimates the causal effect, but the "treatment" may not explain much of the variation in y.
- More precisely, in the equation

$$y = \alpha + \beta w + u, \tag{28}$$

the question of whether u is correlated with w is very different from the relative sizes of $Var(y)$ and $Var(u)$.

- Include covariates that help predict the outcome if they are uncorrelated (in the population) with the covariate(s) of interest. So, if w is the explanatory variable of interest, and it has been randomized with respect to the response and controls, say $\mathbf{z}$, then estimate

$$y = \alpha + \beta w + \mathbf{z}\boldsymbol{\gamma} + u. \tag{29}$$

Because $Cov(\mathbf{z}, w) = 0$, adding $\mathbf{z}$ will not cause collinearity (except slightly in any sample), but it will generally reduce the error variance.

- In large samples,

$$Var(\hat{\beta}) \approx \frac{\sigma_u^2}{n\sigma_w^2}. \tag{30}$$

As more (relevant) covariates are added to **z**, σ_u^2 gets smaller. (And, of course, maximizing the variance of w in a designed experiment helps, too.)

- A control that can substantially reduce the error variance is a lagged value of y.

- Be careful in using models nonlinear in explanatory variables, especially with interactions.

Coefficients on level terms may become essentially meaningless.

EXAMPLE: 401(k) pension plan contributions:

$$\begin{aligned} contribs = {} & \beta_0 + \beta_1 match + \beta_2 income + \beta_3 female \\ & + \beta_4 match \cdot income + \beta_5 match \cdot female + u \end{aligned} \tag{31}$$

The coefficient on *match*, β_1, measures the sensitivity of contributions to the match rate for a male worker with zero income!

- For prediction purposes, unimportant. But centering can make coefficients more interesting:

$$\begin{aligned} contribs_i &= \beta_0 + \alpha_1 match_i + \beta_2 income_i + \beta_3 female_i \\ &\quad + \beta_4 match_i \cdot (income_i - \overline{income}) + \beta_5 match_i \cdot female_i + u \end{aligned}$$

Now, α_1 is the effect of the match rate for women at the average income level.

- β_2 is still the effect of *income* on *contribs* when $mrate = 0$. Because many 401(k) plans offer a zero match rate, this is not a crazy parameter. If we center *match* also, we can write

$$\begin{aligned} contribs_i = {} & \beta_0 + \alpha_1 match_i + \alpha_2 income_i + \beta_3 female_i \\ & + \beta_4 (match_i - \overline{match}) \cdot (income_i - \overline{income}) \\ & + \beta_5 match_i \cdot female_i + u, \end{aligned}$$

and α_2 measures the partial effect of *income* on *contribs* at the average match rate. (Note that we could do a similar centering before interacting *match* with *female*.)

● Without centering: the variables $match_i$ and $match_i \cdot income_i$ are probably highly collinear because the partial effect of *match* at $income = 0$ is poorly identified and uninteresting.

EXAMPLE: Firm participation rates in 401(k) plans and the firm match rate.

```
. sum prate mrate age ltotemp sole

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       prate |      4075     .840607    .1874841   .0036364          1
       mrate |      4075     .463519    .4187388          0          2
         age |      4075    8.186503    9.257011          1         71
     ltotemp |      4075     6.97439    1.539165    4.65396   13.00142
        sole |      4075    .3693252    .4826813          0          1
```

```
. reg prate mrate age ltotemp sole, robust

Linear regression                                      Number of obs =    4075
                                                       F(  4,  4070) =  202.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1755
                                                       Root MSE      =  .17033

------------------------------------------------------------------------------
             |               Robust
       prate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       mrate |   .1072729   .0060035    17.87   0.000     .0955027    .1190432
         age |      .0037   .0002493    14.84   0.000     .0032113    .0041887
     ltotemp |  -.0281719   .0021148   -13.32   0.000    -.0323181   -.0240257
        sole |   .0177024   .0059192     2.99   0.003     .0060977    .0293072
       _cons |   .9505378   .0149728    63.48   0.000     .9211829    .9798927
------------------------------------------------------------------------------
```

```
. gen mrateage = mrate*age

. gen mrateltotemp = mrate*ltotemp

. reg prate mrate age mrateage ltotemp mrateltotemp sole, robust

Linear regression                                      Number of obs =    4075
                                                       F(  6,  4068) =  156.51
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1940
                                                       Root MSE      =  .16845

------------------------------------------------------------------------------
             |               Robust
       prate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       mrate |  -.0014222   .0275289    -0.05   0.959     -.055394    .0525496
         age |   .0066224   .0004247    15.59   0.000     .0057898     .007455
    mrateage |  -.0054106   .0005122   -10.56   0.000    -.0064148   -.0044065
     ltotemp |  -.0390588   .0032932   -11.86   0.000    -.0455153   -.0326023
mrateltotemp |   .0240843   .0044453     5.42   0.000     .0153691    .0327995
        sole |   .0170137   .0058649     2.90   0.004     .0055153    .0285121
       _cons |   1.001494   .0219434    45.64   0.000     .9584733    1.044515
------------------------------------------------------------------------------
```

```
. gen mrateage0 = mrate*(age - 8.19)

. gen mrateltotemp0 = mrate*(ltotemp - 6.974)

. reg prate mrate age mrateage0 ltotemp mrateltotemp0 sole, robust

Linear regression                                      Number of obs =    4075
                                                       F(  6,  4068) =  156.51
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1940
                                                       Root MSE      =  .16845

------------------------------------------------------------------------------
             |               Robust
       prate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       mrate |   .1222283   .0066737    18.32   0.000     .1091443    .1353124
         age |   .0066224   .0004247    15.59   0.000     .0057898     .007455
   mrateage0 |  -.0054106   .0005122   -10.56   0.000    -.0064148   -.0044065
     ltotemp |  -.0390588   .0032932   -11.86   0.000    -.0455153   -.0326023
mrateltot~p0 |   .0240843   .0044453     5.42   0.000     .0153691    .0327995
        sole |   .0170137   .0058649     2.90   0.004     .0055153    .0285121
       _cons |   1.001494   .0219434    45.64   0.000     .9584733    1.044515
------------------------------------------------------------------------------
```

```
. corr mrate mrateltotemp
(obs=4075)

             |    mrate mratel~p
-------------+------------------
       mrate |   1.0000
mrateltotemp |   0.9479   1.0000


. corr mrate mrateltotemp0
(obs=4075)

             |    mrate mrate~p0
-------------+------------------
       mrate |   1.0000
mrateltot~p0 |  -0.1918   1.0000
```

3. LINEAR REGRESSION AS THE BEST MEAN SQUARED ERROR APPROXIMATION

- If we write

$$\begin{aligned} y &= \mathbf{x}\boldsymbol{\beta} + u \\ E(\mathbf{x}'u) &= \mathbf{0}, \end{aligned}$$

then $\mathbf{x}\boldsymbol{\beta} = L(y|\mathbf{x})$ is the best linear approximation to the true conditional mean function, $\mu(\mathbf{x}) = E(y|\mathbf{x})$, in mean squared error. Why? By definition, $\boldsymbol{\beta}$ solves

$$\min_{\mathbf{b} \in \mathbb{R}^K} E[(y - \mathbf{x}\mathbf{b})^2].$$

But $y = \mu(\mathbf{x}) + e$ where $E(e|\mathbf{x}) = 0$. Write

$$\begin{aligned}(y - \mathbf{xb})^2 &= [\mu(\mathbf{x}) + e - \mathbf{xb}]^2 \\ &= [\mu(\mathbf{x}) - \mathbf{xb}]^2 + 2[\mu(\mathbf{x}) - \mathbf{xb}] \cdot e + e^2.\end{aligned}$$

Now any function of **x** is uncorrelated with e, so $E\{[\mu(\mathbf{x}) - \mathbf{xb}] \cdot e\} = 0$. It follows that, for any $\mathbf{b} \in \mathbb{R}^K$,

$$\begin{aligned}E[(y - \mathbf{xb})^2] &= E\{[\mu(\mathbf{x}) - \mathbf{xb}]^2\} + E(e^2) \\ &= E\{[\mu(\mathbf{x}) - \mathbf{xb}]^2\} + \sigma_e^2.\end{aligned}$$

It follows immediately that because **β** minimizes the left hand size, it also solves

$$\min_{\mathbf{b}\in\mathbb{R}^K} E\{[\mu(\mathbf{x}) - \mathbf{x}\mathbf{b}]^2\}$$

(because σ_e^2 does not depend on **b**).

- Exercise: Show that **β** is the vector of coefficients from the population regression of $\mu(\mathbf{x})$ on $\mathbf{x}$. In other words, the β_j also measure the partial effects of the true conditional mean with respect to the x_j.