MIT 14.32/14.320 **Spring 2023**
J. Angrist

Lecture Note 8

Residuals, Fitted Values, and Goodness of Fit

1 Regression Recap

Conceptual Recap

Regression is a many-splendored thing: If $E[Y_i|X_i] = a + bX_i$, then $b = \beta = C(X_i, Y_i)/V(X_i)$ and $a = \alpha = E[Y_i] - \beta E[X_i]$. If the CEF is nonlinear, the regression slope and intercept provide the best linear approximation to it (and the best linear predictor for Y_i). Regression estimates of treatment effects approximate what we'd get by matching on the values of regressors included as controls and then averaging these conditional effects. This note discusses further regression features – properties that hold for all regressions, regardless of your reason for running 'em.

Estimation Recap

We estimate the population slope and intercept using their sample analogs. With a single regressor, these OLS estimators are:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$
$$\hat{\beta} = s_{XY}/s_X^2.$$

Multivariate regression models replace each regressor in the bivariate formula with the residuals remaining after partialing out all others. For example, the first regressor, X_{1i}, in a model with $k-1$ additional controls has sample slope:

$$\hat{\beta}_1 = s_{\tilde{x}_1 Y}/s_{\tilde{x}_1}^2$$

where $\tilde{x}_{1i}$ is the residual from a regression of X_{1i} on $X_{2i}, ..., X_{ki}$. This *OLS estimator* minimizes the *residual sum of squares* (RSS),

$$RSS = \sum (Y_i - a - b_1 X_{1i} - ... - b_k X_{ki})^2$$

in your data.

Regression Inference Recap

As we've seen, under classical assumptions, OLS estimates are unbiased, Normally distributed, and BLUE. More generally, assuming only random sampling, OLS estimates are consistent and asymptotically Normally distributed. We use this fact to test hypotheses and construct confidence intervals for regression parameters under very weak assumptions.

- We've focused so far on interpreting regression coefficients and on the sampling distributions of OLS estimates.
- Regression models generate additional information as well. The nature of this information is the same whether there's one regressor or many, so we'll do the math for bivariate regression only.

2 Regression Fission: Two Pieces of Y

2.1 Residuals

- *Population residuals* are defined as:

$$\varepsilon_i \equiv Y_i - \alpha - \beta X_i.$$

By definition, these regression satisfy:

$$E(\varepsilon_i) = 0 \tag{1}$$
$$E(X_i \varepsilon_i) = 0 \tag{2}$$

- *Estimated regression residuals are*

$$e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i,$$

with properties analogous to (1) and (2) in the sample (assumed to be of size n):

$$\frac{1}{n} \sum_{i=1}^{n} e_i = 0 \tag{3}$$

$$\frac{1}{n} \sum_{i=1}^{n} X_i e_i = 0 \tag{4}$$

Proof : Substitute sample resids in sums,

$$\sum e_i = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) = \sum (Y_i - \bar{Y}) - \hat{\beta} \sum (X_i - \bar{X}) = 0$$

$$\sum X_i e_i = \sum X_i (Y_i - \bar{Y}) - \hat{\beta} \sum X_i (X_i - \bar{X}) = 0$$

These are the first-order conditions for the OLS minimization problem. Leaving the $\frac{1}{n}$ out front reminds is that (3) and (4) are sample moments analogous to (1) and (2).

- Note again that we can't use (3) and (4) to "check" whether $E(\varepsilon_i) = 0$ and $E(X_i \varepsilon_i) = 0$: these properties hold in both population and sample

2.2 Fitted values

- Population *fitted values* are defined as:

$$\hat{Y}_i^* = \alpha + \beta X_i$$

So,

$$Y_i = \hat{Y}_i^* + \varepsilon_i$$

This decomposes Y_i into a piece "explained by X_i" and the piece that's left over, ε_i.

- *Population fits and resids are uncorrelated*:

$$E\left[\hat{Y}_i^* \varepsilon_i\right] = 0$$

Proof. Substitute for $\hat{Y}_i$:

$$E\left[\hat{Y}_i^* \varepsilon_i\right] = E\left[(\alpha + \beta X_i)\varepsilon_i\right] = \alpha E[\varepsilon_i] + \beta E[X_i \varepsilon_i]$$

Now use (1) and (2).

- *Estimated* fitted values are defined as:
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i,$$
so,
$$Y_i = \hat{Y}_i + e_i$$

- By virtue of (3) and (4), estimated resids and fits are uncorrelated in the sample that made them:
$$\sum \hat{Y}_i e_i = 0.$$
  - MM Chpt 2 garbles the distinction between population fitted values and estimated fitted values (sorry!)

3 R-squared

How much of the variance in Y_i can be attributed to variation in X_i? We have seen that resids and fits are uncorrelated. So,
$$s_Y^2 = s_{\hat{Y}}^2 + s_e^2 = \hat{\beta}^2 s_X^2 + s_e^2, \tag{5}$$
in your data. Likewise, in the population from which you're sampling,
$$V(Y_i) = V(\hat{Y}_i^*) + V(\varepsilon_i).$$
Be sure you can show this.

- In both sample and pop, regression ANOVA holds:
$$\textit{Total variance} = \textit{explained variance} + \textit{residual variance}$$

- It's customary to report the following as a measure of regression "goodness of fit":
$$R^2 \equiv \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{\hat{\beta}^2 s_X^2}{s_Y^2} = 1 - \frac{s_e^2}{s_Y^2}$$

- This value, called R^2, is necessarily between 0 and 1
  - R-squared is the fraction of variation in Y_i that is accounted for (in a correlational sense) by the regressor, X_i
  - R^2 is also equal to $\left[CORR(\hat{Y}_i, Y_i)\right]^2$, the square of the coefficient of multiple correlation, defined as $CORR(\hat{Y}_i, Y_i)$ (show this)

- Down the road, we'll use R^2 to construct hypothesis tests that *contrast* alternative multivariate regression models, such as long vs. short models (the latter omits some regressors included in the former)

- The R^2 from any *single* model is hard to interpret without a standard of comparison
  - Life is random; stuff happens. How much randomness should we expect? Hard to say!

- See the regression output below, which shows that schooling is better than sex ... in an R^2 sense

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     10054    31.99881    1.419198         30         34
     incwage |     10054    34664.13    39664.05          0     496759
         uwe |      8468    783.3483    530.3319   .1041667       4000
      loguwe |      8468     6.42724    .7603143  -2.261763   8.294049
     yearsEd |     10054    13.43973    2.521863          0         16
-------------+--------------------------------------------------------
       white |     10054           1           0          1          1
     working |     10054    .8490153    .3580518          0          1
      female |     10054    .5277501    .4992542          0          1

. // regress log weekly wage on education
. reg loguwe yearsEd, robust

Linear regression                                      Number of obs =    8468
                                                       F(  1,  8466) = 1022.00
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1135
                                                       Root MSE      =  .71591

------------------------------------------------------------------------------
             |               Robust
      loguwe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsEd |   .1075207   .0033633    31.97   0.000     .1009278    .1141136
       _cons |   4.965565   .0466539   106.43   0.000     4.874112    5.057018
------------------------------------------------------------------------------

. // regress log weekly wage on female
. reg loguwe female, robust

Linear regression                                      Number of obs =    8468
                                                       F(  1,  8466) =  494.69
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0561
                                                       Root MSE      =  .73873

------------------------------------------------------------------------------
             |               Robust
      loguwe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3605673   .0162113   -22.24   0.000    -.3923455   -.3287891
       _cons |   6.597901   .0102161   645.83   0.000     6.577875    6.617927
------------------------------------------------------------------------------
```

- Our discussion of resids, fits, and R^2 carries over to multivariate models (e.g., R^2 is the proportion of dependent variable variance explained by *all* regressors in a multivariate model)
- Sex and schooling together explain more than either alone
- R^2 increases with additional regressors unless the additional regressors have coefficients exactly equal to zero; this non-decreasing property is implied by least-squares logic

```
. // regress log weekly wage on female
. reg loguwe female yearsEd, robust

Linear regression                               Number of obs     =      8,468
                                                F(2, 8465)        =     963.42
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1973
                                                Root MSE          =     .68128

------------------------------------------------------------------------------
             |               Robust
      loguwe | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |  -.4456252   .0151169   -29.48   0.000    -.4752581   -.4159924
     yearsEd |   .1212531   .0033215    36.51   0.000     .1147422    .1277641
       _cons |   4.989801   .0454427   109.80   0.000     4.900722     5.07888
------------------------------------------------------------------------------

. log close
```

3.1 Regression F-Statistics

In the classical bivariate regression model (Normal errors, etc), the statistic:

$$\frac{s_X^2 \hat{\beta}^2}{s_e^2/(n-2)} = (n-2)\frac{R^2}{1-R^2}$$

is distributed $F_{1,n-2}$ under H_0: $\beta = 0$

- An F-stat with one numerator d.f. is the square of the t-statistic for the single regressor in a bivariate model, so there's no new information here (check this above).
- F tests shine when it's time to *compare* multivariate regression models