SO5032 Lecture 4

Brendan Halpin

March 12, 2024

Outline

SO5032 Lecture 4

SO5032 Lecture 4

Formula

Formula for multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k + e$$

$$e \sim N(0, \sigma)$$

- Interpretation of β_j
  - How much $\hat{Y}$ changes for a 1-unit in X_j holding all other values constant
  - The estimated effect on Y of a 1-unit change in X_j, "controlling for" or "taking account" of all the other Xs

Predictions

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k$$

- Enter values for all X variables to get a prediction for those values
- If we increase X_i by 1, holding all others the same, $\hat{Y}$ changes by β_i

Simplest example

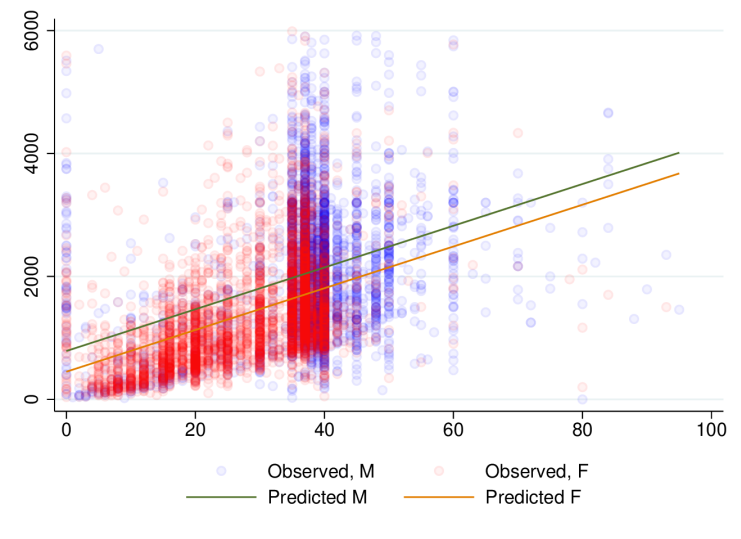
- Simplest multiple regression model adds a binary variable to a model with a continuous X

```
. reg income hours i.sex

      Source |       SS           df       MS      Number of obs   =     7,945
-------------+----------------------------------   F(2, 7942)      =    794.96
       Model |  1.8935e+09         2  946761687   Prob > F        =    0.0000
    Residual |  9.4586e+09     7,942  1190962.07   R-squared       =    0.1668
-------------+----------------------------------   Adj R-squared   =    0.1666
       Total |  1.1352e+10     7,944  1429021.17   Root MSE        =    1091.3

------------------------------------------------------------------------------
      income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       hours |   33.96065   1.123629    30.22   0.000     31.75804    36.16326
             |
         sex |
     female  |  -337.0889   26.44232   -12.75   0.000    -388.9228    -285.255
       _cons |   787.1759   45.73595    17.21   0.000     697.5214    876.8304
------------------------------------------------------------------------------
```

Predicted lines: one for each value of sex



More general 2 X-variable example

```
. reg wage  ttl_exp grade

      Source |       SS           df       MS      Number of obs   =     2,244
-------------+----------------------------------   F(2, 2241)      =    194.77
       Model |    11010.6         2     5505.3   Prob > F        =    0.0000
    Residual |  63343.7305     2,241  28.2658325   R-squared       =    0.1481
-------------+----------------------------------   Adj R-squared   =    0.1473
       Total |  74354.3305     2,243  33.1495009   Root MSE        =    5.3166

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     ttl_exp |   .2616056   .0248373    10.53   0.000     .2128992     .310312
       grade |   .6483343    .045426    14.27   0.000     .5592528    .7374158
       _cons |  -4.002059   .6245962    -6.41   0.000    -5.226906   -2.777211
------------------------------------------------------------------------------
```

Effect of experience on wage, controlling for grade

Effect of grade on wage, controlling for experience

See https://teaching.sociology.ul.ie/so5032/ttlgrade.html

Residuals

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k$$

$$Y = \hat{Y} + e$$

$$e \sim N(0, \sigma)$$

- Mean of zero
- Standard deviation of σ (RMSE)
- Normally distributed
- Should have no structured relationship to X variables

SO5032 Lecture 4

R^2

R^2

- R^2: coefficient of multiple determination
- TSS = sum of squared deviation from the mean = $\sum(Y_i - \bar{Y})^2$
- RSS = sum of squared deviation from the regression prediction = $\sum(Y_i - \hat{Y})^2$
- $R^2 = \frac{TSS - RSS}{TSS}$
- Range: 0 (no relationship) to 1 (perfect linear relationship)
- PRE: Proportional Reduction in Error

R^2 and correlation

- In bivariate regression, R^2 is the square of the correlation coefficient between Y and X
- In multiple regression, it is the square of the correlation between Y and $\hat{Y}$
- (In bivariate regression the correlation between X and $\hat{Y}$ is 1)

SO5032 Lecture 4

Hypothesis testing

Hypothesis testing: one parameter at a time

- t-test: *abs*$(\hat{\beta}_j/\mathrm{se}_j) > t$
- Interpretation:
  - Null: population value of β is 0; this variable has no influence once the other variables are taken account of

Example

```
. reg income age i.sex

      Source |       SS           df       MS      Number of obs   =       959
-------------+----------------------------------   F(2, 956)       =     45.72
       Model |  33922983.9         2   16961492   Prob > F        =    0.0000
    Residual |   354670636       956  370994.389   R-squared       =    0.0873
-------------+----------------------------------   Adj R-squared   =    0.0854
       Total |   388593620       958  405630.083   Root MSE        =    609.09

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -3.144945   1.083398    -2.90   0.004    -5.271057   -1.018833
             |
         sex |
     female  |   -352.678   39.51326    -8.93   0.000    -430.2208   -275.1353
       _cons |   1035.878   54.58935    18.98   0.000     928.7494    1143.007
------------------------------------------------------------------------------
```

Hypothesis testing: all parameters together

- F-test:
  - $\beta_1 = \beta_2 \ldots = \beta_k = 0$
- Null hypothesis: no X variable has an effect once the others are taken care of.
- A "global" test: the null is that there is no relevant variable in the model
- Calculation based on TSS and RSS, but also number of cases and number of parameters estimated
- Uses F distribution (two df parameters: k and n-k-1, k is number of parameters, n the number of cases)

Hypothesis testing: additional parameters

- Delta F-test compares "nested" models
  - Model 1: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_g X_g$
  - Model 1: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_g X_g + \beta_h X_h ... + \beta_k X_k$
- Null hypothesis: $\beta_{\mathrm{h}} = \ldots = \beta_{\mathrm{k}} = 0$
- That is, given the variables already in the model, the additional variables contribute no explanatory power.
- Useful when adding multi-category variables, or related groups of variables

Dummy variables

In regression models we often use "indicator coding" or "dummy coding"

With a two-category variable, we set one category to 0 and the other to 1 and interpret it as the effect of being in the second category (e.g., female) compared with the first.

```
. reg income age i.sex

      Source |       SS           df       MS      Number of obs   =       959
-------------+----------------------------------   F(2, 956)       =     45.72
       Model |  33922983.9         2  16961492    Prob > F        =    0.0000
    Residual |   354670636       956  370994.389   R-squared       =    0.0873
-------------+----------------------------------   Adj R-squared   =    0.0854
       Total |   388593620       958  405630.083   Root MSE        =    609.09

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -3.144945   1.083398    -2.90   0.004    -5.271057   -1.018833
             |
         sex |
    female   |   -352.678   39.51326    -8.93   0.000    -430.2208   -275.1353
       _cons |   1035.878   54.58935    18.98   0.000     928.7494    1143.007
------------------------------------------------------------------------------
```

More than two categories

With more that two categories we create a set of binary variables, "indicator variables" or "dummy variables":

	d1	d2	d3	d4
a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1

For m categories, m-1 dummy variables are sufficient.

We interpret the parameter as the estimated effect of being in that category relative to the omitted or "reference" category.

Stata handles this automatically with the `i.` prefix.

Example

```
. reg income age i.sex i.qual

      Source |       SS           df       MS      Number of obs   =       959
-------------+----------------------------------   F(5, 953)       =     54.14
       Model |  85960604.5         5  17192120.9   Prob > F        =    0.0000
    Residual |   302633015       953  317558.253   R-squared       =    0.2212
-------------+----------------------------------   Adj R-squared   =    0.2171
       Total |   388593620       958  405630.083   Root MSE        =    563.52

------------------------------------------------------------------------------
                 income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------+----------------------------------------------------
                    age |  -.3897295    1.04777    -0.37   0.710    -2.445933    1.666474
                        |
                    sex |
                 female |  -336.9623   36.75947    -9.17   0.000    -409.1011   -264.8234
                        |
                   qual |
A-level, other sub-d.. |  -459.9208   78.54165    -5.86   0.000    -614.0554   -305.7862
O-level, commercial,.. |   -701.695   77.16016    -9.09   0.000    -853.1185   -550.2716
  Sub-O-level, no qual |  -864.9695   76.41768   -11.32   0.000    -1014.936   -715.0032
                       |
                 _cons |   1563.508   81.83797    19.10   0.000     1402.904    1724.111
------------------------------------------------------------------------------
```