SO5032 Lecture 4

Brendan Halpin

April 1, 2025

Outline

SO5032 Lecture 4

SO5032 Lecture 4

Outline

Outline

- Multiple regression
- Formula, Interpretation
- Hypothesis testing
- Goodness of fit: residuals and R^2
- Agresti, Ch 11

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

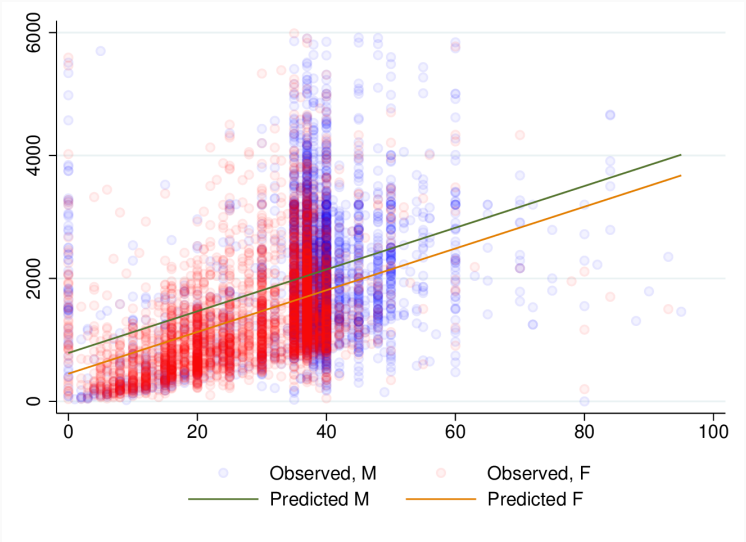
$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k$$

- 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

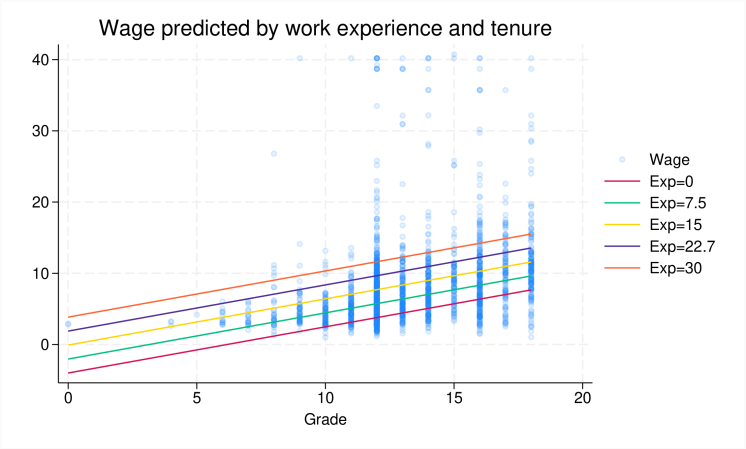
Predicted lines: one for each value of sex



More general 2 X-variable example

Effect of experience on wage, controlling for grade

Effect of grade on wage, controlling for experience



See https://teaching.sociology.ul.ie/so5032/ttlgradelin.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

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: $\beta_h = \ldots = \beta_k = 0$
- 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.

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