Lecture 15: Predicted Values and Qualitative Independent Variables

Pierre Biscaye

Fall 2022

Agenda

1. Predicted values: average and individual
2. Binary independent variables
3. Categorical independent variables (time permitting)

Predicted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} \quad (1)$$

- So for values $x_1 = c_1, x_2 = c_2, ..., x_k = c_k$
- If we want to estimate $\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + ... + \beta_k c_k = E[y_i | x_1 = c_1, x_2 = c_2, ..., x_k = c_k]$
- We would estimate

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + ... + \hat{\beta}_k c_k \quad (2)$$

- If MLR assumptions hold, then $\hat{\beta}_j$s are consistent estimators for β_j and so $\hat{\theta}_0$ is also a consistent estimator.
- This is useful!
  - Think back to our early model where we wanted to predict how CO_2/cap would change across countries as GDP/cap increases.

Variance of predicted values

- We care not just about the predicted value but also its precision.
  - A very imprecise predicted value is not very useful.
- Calculating the variance of the predicted value is not straightforward:

$$\begin{aligned} var(\hat{\theta_0}) &= var(\hat{\beta_1}c_1 + \hat{\beta_2}c_2 + ... + \hat{\beta_k}c_k) \qquad (3) \\ &\neq var(\hat{\beta_1}c_1) + var(\hat{\beta_2}c_2) + ... + var(\hat{\beta_k}c_k) \end{aligned}$$

- We can use a variable substitution trick:

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + ... + \beta_k c_k \qquad (4)$$
$$\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - ... - \beta_k c_k \qquad (5)$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad (6)$$
$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + ... + \beta_k(x_k - c_k) \qquad (7)$$

- So we can regress y on $x_1 - c_1$, $x_2 - c_2$, ..., $x_k - c_k$ and use the constant to estimate θ_0 and its standard error. **Why?**

Example: predicting college GPA

- Suppose we were running college admissions.
- We want to predict success in college (proxied by GPA) using SAT scores, high school class sizes (in 100s), and high school class percentile (lower percentile indicates higher relative rank).
- Using data on our current students, we regress

$$colgpa_i = \beta_0 + \beta_1 SAT_i + \beta_2 hsperc_i + \beta_3 hsize_i + \beta_4 hsize_i^2 + u_i \quad (8)$$

- What is the predicted college gpa for someone with a SAT of 1200, in the 30th percentile of their graduating class, with a high school graduating class of 500?

To Jupyter!

Predicting college GPA

- $E[colgpa|SAT = 1200, hsperc = 30, hsize = 5] = 2.70$
- And $SE(E[colgpa|SAT = 1200, hsperc = 30, hsize = 5]) = 0.02$
- So, our 95% confidence interval for $E[colgpa|SAT = 1200, hsperc = 30, hsize = 5]$ is $[2.66, 2.74]$
  - Pretty small range: 0.08 GPA points
- Note, this does *not* mean we expect 95% of people with these characteristics to have a college GPA in this range.
- Instead, we expect people with these characteristics to have a college GPA in this range *on average* (with 95% probability). There will still be variation!

X values and prediction precision

- Does the choice of X values affect the precision of our prediction?
- Suppose we want to predict the GPA from a student with similar performance as before but at one of the biggest schools, with a graduating class of 900.
- Do you expect the SE for the predicted value to be the same, smaller, or larger?
- To Jupyter!

X values and prediction precision

- Does the choice of X values affect the precision of our prediction?
- Suppose we want to predict the GPA from a student with similar performance as before but at one of the biggest schools, with a graduating class of 900.
- Do you expect the SE for the predicted value to be the same, smaller, or larger?
- To Jupyter!
- $E[colgpa|SAT = 1200, hsperc = 30, hsize = 9] = 2.76$
- And $SE(E[colgpa|SAT = 1200, hsperc = 30, hsize = 5]) = 0.07$: This is much bigger!
- Our 95% confidence interval is $[2.63, 2.90]$, a range of 0.27 GPA points.
- Why is this estimate less precise?
- For what X values would our prediction be most precise?

What if we want a confidence interval for an individual unit?

- What we have done so far is estimate $\hat{E}[y_i|x_1 = c_1, x_2 = c_2, ..., x_k = c_k]$: the *average* predicted value of y for the subpopulation with given characteristics
- Suppose we want a confidence interval for an *individual* observation with those characteristics.
  - For example, predict college GPA for a given new applicant.
  - What is the difference?

What if we want a confidence interval for an individual unit?

- What we have done so far is estimate $\hat{E}[y_i|x_1 = c_1, x_2 = c_2, ..., x_k = c_k]$: the *average* predicted value of y for the subpopulation with given characteristics
- Suppose we want a confidence interval for an *individual* observation with those characteristics.
  - For example, predict college GPA for a given new applicant.
  - What is the difference?
- When we are not talking about averages, we need to account for the unobserved residual.
  - While $E[u|x] = 0$, individual u_i take on a variety of values.
- Variance in the unobserved residual will affect our standard error when estimating predicted values for individual units.

How do we predict the value for a new observation?

- Suppose we label a new observation with 0.

$$y^0 = \beta_0 + \beta_1 x_1^0 + ... + \beta_k x_k^0 + u^0 \tag{9}$$

$$\hat{y^0} = \hat{\beta_0} + \hat{\beta_1} x_1^0 + ... + \hat{\beta}_k x_k^0 \tag{10}$$

- Thus, if we predict y^0 using $\hat{y^0}$ our prediction error $\hat{u}^0$ will be

$$\begin{aligned}
\hat{u}^0 &= y^0 - \hat{y^0} \\
&= \beta_0 + \beta_1 x_1^0 + ... + \beta_k x_k^0 + u^0 - \hat{\beta_0} + \hat{\beta_1} x_1^0 + ... + \hat{\beta}_k x_k^0 \\
&= (\beta_0 - \hat{\beta_0}) + (\beta_1 - \hat{\beta_1}) x_1^0 + ... + (\beta_k - \hat{\beta}_k) x_k^0 + u^0
\end{aligned}$$

What do we know about $\hat{u}^0$?

$$\begin{aligned} var(\hat{u}^0) &= var(y^0 - \hat{y}^0) \\ &= var(\beta_0 + \beta_1 x_1^0 + ... + \beta_k x_k^0 + u^0 - \hat{y^0}) \\ &= var(u^0 - \hat{y^0}) \end{aligned}$$

- This is because we observe the values $x_1^0, \ldots, x_k^0$ but we do not know u^0, while $\hat{y^0}$ is a predicted value with an associated variance.
  - For average predicted values, all we care about is the prediction $\hat{y}$, not the difference between that and any true observed value.
- Note that, *for a new observation*, u^0 is uncorrelated with $\hat{y^0}$
  - This *would not* be true for an observation in the sample used to estimate the $\hat{\beta}_j$ terms.
- This means that $var(\hat{u}^0) = var(u^0 - \hat{y^0}) = var(u^0) + var(\hat{y^0}) = var(\hat{y^0}) + \sigma_u^2$

Confidence Intervals for $\hat{y}^0$

- Two sources of variance for predicted values of new observations: the variance from our predictions, and the underlying variance in the population.
- We don't observe σ_u^2 so we estimate it with $\hat{\sigma_u^2}$.
- $SE(\hat{u}^0) = \sqrt{var(\hat{y^0}) + \hat{\sigma_u^2}}$
- It turns out that

$$\frac{\hat{u}^0}{SE(\hat{u}^0)} \sim t_{n-k-1} \tag{11}$$

- So that a CI for $\hat{y^0}$ is given by

$$(\hat{y^0} - t_{\frac{\alpha}{2}} * SE(\hat{u}^0), \hat{y^0} + t_{\frac{\alpha}{2}} * SE(\hat{u}^0)) \tag{12}$$

- How do we estimate this? To Jupyter!

Estimate and CI for a new observation

- We estimated that $SE(\hat{u}^0) = \sqrt{0.02^2 + 0.56^2} \approx 0.56$
- Note that the prediction error really doesn't matter relative to the unobserved error: we can predict averages accurately but not individual GPA.
- 95% CI = [2.7-1.96*0.56, 2.7 + 1.96*0.56] = [1.60,3.80]
- We conclude that we have little predictive power over individual performance.
- Aside: these predicted values do not always translate to transformations of y.
  - We can predict $\widehat{log(y)}$ by regressing $log(y)$ on x, but $E[y] \neq e^{(\widehat{log(y)})}$.

Binary independent variables

- Many important research and policy questions involve qualitative data.
  - E.g., in focus on gender wage gaps, gender is qualitative data - there is no natural ordering or quantitative valuation.
  - E.g., in the impact of irrigation on crop yields, treatment v. control assignment is qualitative data.
- Qualitative data is often binary ("yes or no"); may also be categorical (multiple categories/options).
- We propose a simple solution for binary variables: define $X = 1$ if "yes", $X = 0$ if no.
  - Any value assignment is permitted, but interpretations are most intuitive with this approach.

Binary variable in simple regression models

$$y_i = \beta_0 + \beta_1 X + u \tag{13}$$

- Let X be binary.
- Interpretation: β_1 is the average difference in Y between observations where $X = 1$ and where $X = 0$.
  - E.g., the mean difference in wages between women and men.
  - Or, the mean difference in the outcome between treatment and control.
- $X = 0$ is the *excluded* group: β_0 gives the average value of the outcome for this group.

Binary variable in MLR models

- Consider

$$wage_i = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + u_i \tag{14}$$

- If we assume MLR4 then $E[u|Female, Educ] = 0$, and

$$\delta_0 = E[wage|Female = 1, Educ] - E[wage|Female = 0, Educ] \tag{15}$$

- This is the mean difference in wages between men and women, holding education constant.

Binary variable in MLR models

- Consider

$$wage_i = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + u_i \tag{14}$$

- If we assume MLR4 then $E[u|Female, Educ] = 0$, and

$$\delta_0 = E[wage|Female = 1, Educ] - E[wage|Female = 0, Educ] \tag{15}$$

- This is the mean difference in wages between men and women, holding education constant.
- Why don't we estimate it this way?

$$wage_i = \beta_0 + \delta_1 Female_i + \delta_2 NotFemale_i + \beta_1 Educ_i + u_i \tag{16}$$

Binary variable in MLR models

- Consider

$$wage_i = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + u_i \tag{14}$$

- If we assume MLR4 then $E[u|Female, Educ] = 0$, and

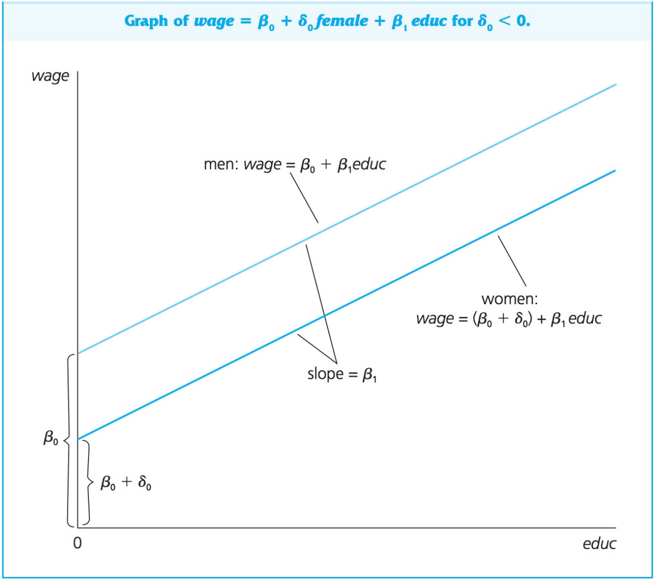
$$\delta_0 = E[wage|Female = 1, Educ] - E[wage|Female = 0, Educ] \tag{15}$$

- This is the mean difference in wages between men and women, holding education constant.
- Why don't we estimate it this way?

$$wage_i = \beta_0 + \delta_1 Female_i + \delta_2 NotFemale_i + \beta_1 Educ_i + u_i \tag{16}$$

- Multicollinearity!
- With qualitative data, always need to have an excluded/left-out category: here, it's $Female = 0$. Values for the excluded category are reflected in β_0.

Binary variable in MLR, graphically



Not much changes when we add more controls

- We could also estimate

$$wage_i = \beta_0 + \delta_0 Female_i + \beta_1 educ_i + \beta_2 tenure_i + \beta_3 exper_i + u_i \quad (17)$$

- How to interpet δ_0?

Not much changes when we add more controls

- We could also estimate

$$wage_i = \beta_0 + \delta_0 Female_i + \beta_1 educ_i + \beta_2 tenure_i + \beta_3 exper_i + u_i \quad (17)$$

- How to interpet δ_0?
- We can test $H_0 : \delta_0 = 0$ against $H_1 : \delta_0 \neq 0$.
- This is a test for a wage penalty against women, which is likely discrimination if we are holding those other factors constant (may also be omitted variable bias).

To Jupyter!

What about non-binary qualitative data?

- Not all qualitative data have two categories. What to do? Construct *categorical* variable.
- For example, could think of categories of educational achievement:
  - $edcat = 1$ if did not complete primary school
  - $edcat = 2$ if completed primary but not secondary school
  - $edcat = 3$ if completed secondary school but not tertiary/vocational school
  - $edcat = 4$ if completed some post-secondary school
- This is a qualitative representation of years of education. The numbers are meaningless.
- Binary variables are just categorical variables with only two categories.

Modeling categorical variables

- How to include a categorical variable in the model?
- Any concerns about this model?

$$wage_i = \beta_0 + \delta_0 Female_i + \beta_1 edcat_i + \beta_2 tenure_i + \beta_3 exper_i + u_i \quad (18)$$

Modeling categorical variables

- How to include a categorical variable in the model?
- Any concerns about this model?

$$wage_i = \beta_0 + \delta_0 Female_i + \beta_1 edcat_i + \beta_2 tenure_i + \beta_3 exper_i + u_i \quad (18)$$

- β_1 would be the effect of moving up by one educational category, all else equal.
- But the values of the categories are arbitrary and the jumps are not equivalent, so this is not very meaningful.
- Instead we will incorporate the categorical variable using individual binary variables for the different categories.

Modeling categorical variables

- How to include a categorical variable in the model?
- Use individual binary variables for the different categories:
  - $noprim = 1$ if did not complete primary school, 0 otherwise
  - $prim = 1$ if completed primary but not secondary school, 0 otherwise
  - $sec = 1$ if completed secondary school but not tertiary/vocational school, 0 otherwise
  - $postsec = 1$ if completed some post-secondary school, 0 otherwise
- Incorporate dummies (binary variables) for each individual category except one:

$$\begin{aligned} wage_i = \gamma_0 + \gamma_1 prim_i + \gamma_2 sec_i + \gamma_3 postsec_i + \\ \beta_1 Female_i + \beta_2 tenure_i + \beta_3 exper_i + u_i \end{aligned} \tag{19}$$

- Excluded/reference category is *noprim* (could choose any). Why do we need to omit one?

Interpreting coefficients with categorical variables

$$wage_i = \gamma_0 + \gamma_1 prim_i + \gamma_2 sec_i + \gamma_3 postsec_i + \beta_1 Female_i + \beta_2 tenure_i + \beta_3 exper_i + u_i \quad (20)$$

- Interpretation of γs (holding all else constant):
  - γ_0: mean wages for a man who did not complete primary school, with 0 experience and tenure
    - Intercept still gives information about the excluded category.
  - γ_1: difference in mean wages for completing primary school relative to no primary.
  - γ_2: difference in mean wages for completing secondary school relative to no primary.
  - γ_3: difference in mean wages for completing post-secondary school relative to no primary.
- All coefficients for categorical variable dummies are interpreted *relative to* the excluded category.
- To Jupyter!

Multiple categorical variables

- Suppose we had data on men and women who were single and married.
- We could compare people who were female or not, and people who are married or not, by estimating

$$
\begin{aligned}
log(wage_i) = \beta_0 + \delta_1 married_i + \delta_2 female_i + \beta_1 educ_i \\
+ \beta_2 exper_i + \beta_3 tenure_i + u_i
\end{aligned} \tag{21}
$$

- But what if marriage impacts wages different for women and men?

Two approaches

1) Can treat the model as interactive:

$$wage_i = \delta_0 + \delta_1 married_i + \delta_2 female_i + \delta_3 female_i * married_i + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \tag{22}$$

- How to interpret the δ coefficients (holding other variables constant)?

Two approaches

1) Can treat the model as interactive:

$$wage_i = \delta_0 + \delta_1 married_i + \delta_2 female_i + \delta_3 female_i * married_i + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \quad (22)$$

- How to interpret the δ coefficients (holding other variables constant)?
  - δ_0: mean wages for a single man with 0 education, experience, and tenure
  - δ_1: difference in mean wages between single and married men
  - δ_2: difference in mean wages between single men and women
  - δ_3: *additional* difference in mean wages for women when married (beyond the difference for single women); *also* the *additional* difference in mean wages with marriage for women (beyond the difference for men)
- What is the mean wage for 1) a single woman and 2) a married woman with 0 education, experience, and tenure?

Two approaches

1) Can treat the model as interactive:

$$wage_i = \delta_0 + \delta_1 married_i + \delta_2 female_i + \delta_3 female_i * married_i + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \quad (22)$$

- How to interpret the δ coefficients (holding other variables constant)?
  - δ_0: mean wages for a single man with 0 education, experience, and tenure
  - δ_1: difference in mean wages between single and married men
  - δ_2: difference in mean wages between single men and women
  - δ_3: *additional* difference in mean wages for women when married (beyond the difference for single women); *also* the *additional* difference in mean wages with marriage for women (beyond the difference for men)
- What is the mean wage for 1) a single woman and 2) a married woman with 0 education, experience, and tenure?
  1. $\delta_0 + \delta_2$
  2. $\delta_0 + \delta_1 + \delta_2 + \delta_3$

Alternate approach

2) Rewrite the model using a new categorical variable

- Female and Married both have 2 categories.
- Can think of the interaction of the two binary variables as a new *categorical* variable with 4 categories.
  - Single female, single male, married female, married male
  - Obviously can't do this for interactions involving a continuous variable.
- How to include this in a regression?

$$\begin{aligned} wage_i = \gamma_0 + \gamma_1 marrmale_i + \gamma_2 singfem_i + \gamma_3 marrfem_i \\ + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \end{aligned} \quad (23)$$

Alternate approach

2) Rewrite the model using a new categorical variable

- Female and Married both have 2 categories.
- Can think of the interaction of the two binary variables as a new *categorical* variable with 4 categories.
  - Single female, single male, married female, married male
  - Obviously can't do this for interactions involving a continuous variable.
- How to include this in a regression?

$$\begin{aligned} wage_i = \gamma_0 + \gamma_1 marrmale_i + \gamma_2 singfem_i + \gamma_3 marrfem_i \\ + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \end{aligned} \tag{23}$$

- γ_0 is the mean wage (when $educ, exper, tenure = 0$) for reference category: single males.
- Other γ coefficients are difference in means for included group compared to the reference group.

How do the two models relate?

$$\begin{aligned} wage_i = \delta_0 + \delta_1 married_i + \delta_2 female_i + \delta_3 female_i * married_i + \\ \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \end{aligned} \tag{24}$$

$$\begin{aligned} wage_i = \gamma_0 + \gamma_1 marrmale_i + \gamma_2 singfem_i + \gamma_3 marrfem_i \\ + \beta_1 educ_i + \beta_2 exper_i + \beta_3 tenure_i + u_i \end{aligned} \tag{25}$$

- $\delta_0 = \gamma_0$
- $\delta_1 = \gamma_1$
- $\delta_2 = \gamma_2$
- $\delta_3 + \delta_2 + \delta_1 = \gamma_3$
- To Jupyter!