Linear Probability Models

2020-04-06

Preparation

In this set of notes, you will learn how about linear probability models, and why they are not typically used to model dichotomous categorical outcome variables (e.g., dummy coded outcome). We will use data from the file *graduation.csv* (see the data codebook) to explore predictors of college graduation.

```r
# Load libraries
library(broom)
library(corrr)
library(educate)
library(patchwork)
library(tidyverse)

# Read in data
grad = read_csv(file = "~/Documents/github/epsy-8252/data/graduation.csv")

# View data
head(grad)
```

```
# A tibble: 6 x 6
  degree   act scholarship    ap firstgen nontrad
   <dbl> <dbl>       <dbl> <dbl>    <dbl>   <dbl>
1      1  21            0      0        0       0
2      1  19.0          0      0        0       0
3      1  27            0      0        1       0
4      1  25            0.5    0        1       0
5      0  28            0     17        1       0
6      1  21            0      0        0       1
```

Note that in these analyses the outcome variable (`degree`) is a dichotomous dummy-coded categorical variable indicating whether or not a student graduated.

Data Exploration

To begin the analysis, we will explore the outcome variable degree. Since this is a categorical variable, we can look at counts and proportions. The analysis suggests that most students in the sample tend to graduate (73%).

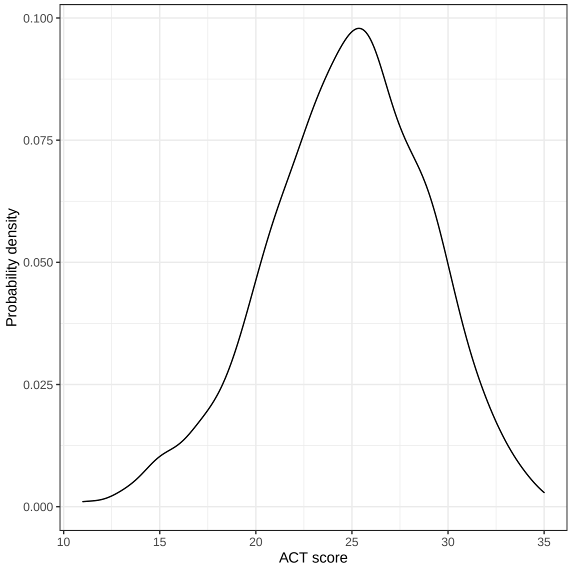
```
grad %>%
  group_by(degree) %>%
  summarize(
    Count = n(),
    Prop = n() / nrow(grad)
    )
```

```
# A tibble: 2 x 3
  degree Count  Prop
   <dbl> <int> <dbl>
1      0   627 0.267
2      1  1717 0.733
```

We will also explore the act variable, which we will use as a predictor in the analysis.

```
# Density plot
ggplot(data = grad, aes(x = act)) +
  geom_density() +
  theme_bw() +
  xlab("ACT score”) +
  ylab("Probability density”)
```

Figure 1
Density plot of the ACT scores.



```
# Summary measures
grad %>%
  summarize(
    M = mean(act),
    SD = sd(act)
    )
```

```
# A tibble: 1 x 2
      M    SD
  <dbl> <dbl>
1  24.8  4.15
```

The distribution is unimodal and symmetric. It indicates that the sample of students have a mean ACT score near 25. While there is a great deal of variation in ACT scores (scores range from 10 to 36), most students have a score between 21 and 29.

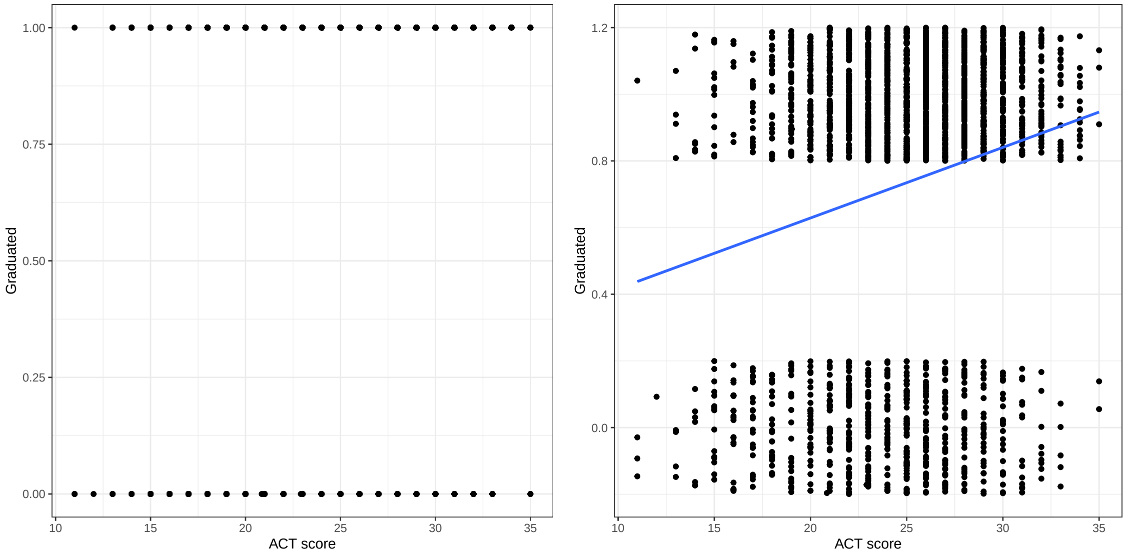
Relationship between ACT Score and Graduation

When the outcome variable was continous, we examined relationships graphically via a scatterplot. If we try that when the outcome is dichotomous, we run into problems. Since there are only two values for the outcome, many of the observations are over-plotted, and it is impossible to tell anything about the relationship between the variables.

```
# Scatterplot
ggplot(data = grad, aes(x = act, y = degree)) +
  geom_point() +
  theme_bw() +
  xlab("ACT score”) +
  ylab("Graduated”)

# Jittered scatterplot
ggplot(data = grad, aes(x = act, y = jitter(degree))) +
  geom_point() +
  geom_smooth(method = "lm”, se = FALSE) +
  theme_bw() +
  xlab("ACT score”) +
  ylab("Graduated”)
```

Figure 2
Scatterplot of whether a student graduated versus ACT score. In the right-hand plot, the points have been jittered to alleviate over-plotting, and the regression smoother has also been added to the plot.



One solution to this problem is to add (or subtract) a tiny amount to each student's `degree` value. This is called "jittering" the observations". To do this, use the `jitter()` function. (Note that the amount of jittering can be adjusted by including an additional argument to the `jitter()` function.) This spreads out the observations vertically, so we no longer have the problem of overplotting. But, the relationship is still difficult to see in this scatterplot. Adding the regression smoother helps us see that there is a positive relationship between ACT scores and graduation.

What does this mean? To understand this, we have to think about what the linear regression is modeling. Remember that the regression model is predicting the average Y at each X. Since our Y is dichotomous, the average represents the proportion of students with a 1. In other words, the regression predicts the proportion of students who graduate for a particular ACT score. The positive relationship between ACT score and graduation suggests that higher ACT scores are associated with higher proportions of students who graduate. We can also see this relationship by examining the correlation matrix between the two variables.

```
grad %>%
  select(degree, act) %>%
  correlate()
```

```
# A tibble: 2 x 3
  rowname degree    act
  <chr>    <dbl>  <dbl>
1 degree  NA      0.195
2 act      0.195 NA
```

Fitting the Linear Probability (Proportion) Model

We can fit the linear model shown in the plot using the `lm()` function as we always have.

```
# Fit the model
lm.1 = lm(degree ~ 1 + act, data = grad)

# Model-level- output
glance(lm.1)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1    0.0381        0.0377 0.434      92.8 1.48e-21     2 -1370. 2746. 2764.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Differences in ACT score account for 3.8% of the variation in graduation. This is more than we would expect by chance if the null hypothesis was true, $F(1,\ 2342) = 92.8$, $p < .001$.

```
# Coefficient-level- output
tidy(lm.1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   0.217    0.0543       3.99 6.83e- 5
2 act           0.0208   0.00216      9.63 1.48e-21
```

The fitted model is

$$\hat{\pi}_i = 0.22 + 0.02(\text{ACT Score}_i)$$

where $\hat{\pi}_i$ is the predicted proportion of students who graduate. Interpreting the coefficients,

- On average, 0.22 of students having an ACT score of 0 are predicted to graduate. (Extrapolation)
- Each one-point difference in ACT score is associated with an additional 0.02 predicted improvement in the proportion of students graduating, on average.

Let's examine the model assumptions.

```
# Augment the model
out = augment(lm.1)
```

```
# View augmented data
head(out)
```

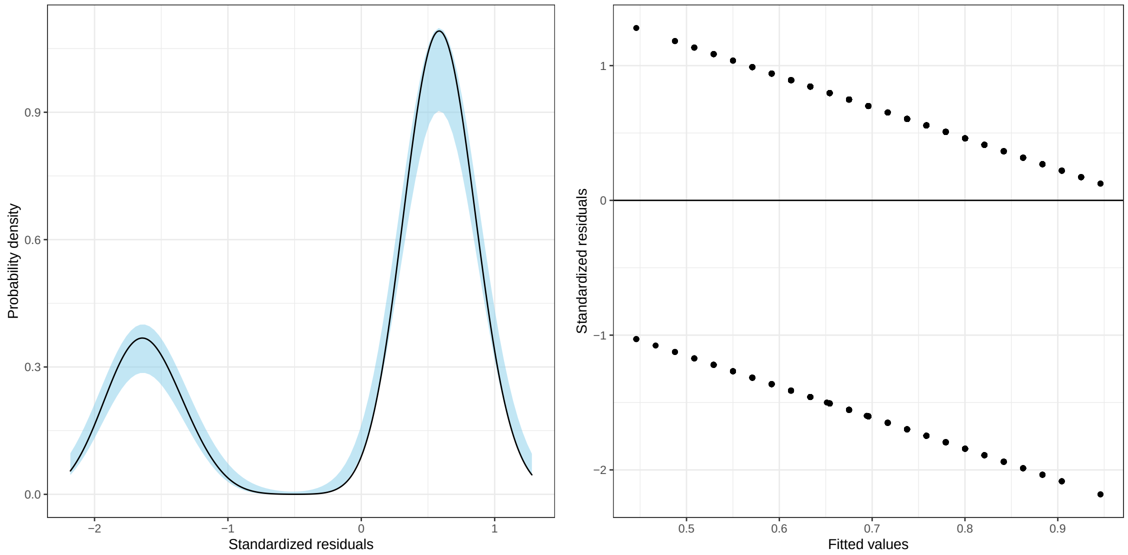
```
# A tibble: 6 x 9
  degree   act .fitted .se.fit .resid     .hat .sigma   .cooksd .std.resid
   <dbl> <dbl>   <dbl>   <dbl>  <dbl>    <dbl>  <dbl>     <dbl>      <dbl>
1      1  21     0.654 0.0121   0.346 0.000777  0.434 0.000246       0.796
2      1  19.0   0.613 0.0153   0.387 0.00124   0.434 0.000493       0.891
3      1  27     0.779 0.0102   0.221 0.000552  0.434 0.0000713      0.508
4      1  25     0.738 0.00899  0.262 0.000428  0.434 0.0000782      0.604
5      0  28     0.800 0.0114  -0.800 0.000688  0.434 0.00117       -1.84
6      1  21     0.654 0.0121   0.346 0.000777  0.434 0.000246       0.796
```

```
# Examine normality assumption
ggplot(data = out, aes(x = .std.resid)) +
  stat_density_confidence() +
  geom_density() +
  theme_bw() +
  xlab("Standardized residuals") +
  ylab("Probability density")

# Examine linearity and homoskedasticity
ggplot(data = out, aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw() +
  xlab("Fitted values") +
  ylab("Standardized residuals")
```

Figure 3

LEFT: Density plot of the standardized residuals from the linear probability model. RIGHT: Scatterplot of the standardized residuals versus the fitted values from the linear probability model.


Probability density
Standardized residuals
Standardized residuals
Fitted values