Linear Probability Models

2019-03-21

In this set of notes, you will learn how about linear probability models, and why they should not be used when the outcome is a dummy coded categorical variable.

Dataset and Research Question

In this set of notes, we will use data from the file *graduation.csv* (see the data codebook here). These data include student-level attributes for $n = 2344$ randomly sampled students who were first-time, full-time freshman from the 2002 cohort at a large, midwestern research university.

```
# Load libraries
library(broom)
library(corrr)
library(dplyr)
library(ggplot2)
library(readr)
library(sm)

# 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
```

We will use these data to explore predictors of college graduation.

Data Exploration

To begin the analysis, we will explore the outcome variable `degree`. Since this is a dichotomous variable, we can look at counts/proportions. The analysis suggests that most freshmen who enroll at this institution 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
sm.density(grad$act, xlab = "ACT score")
```

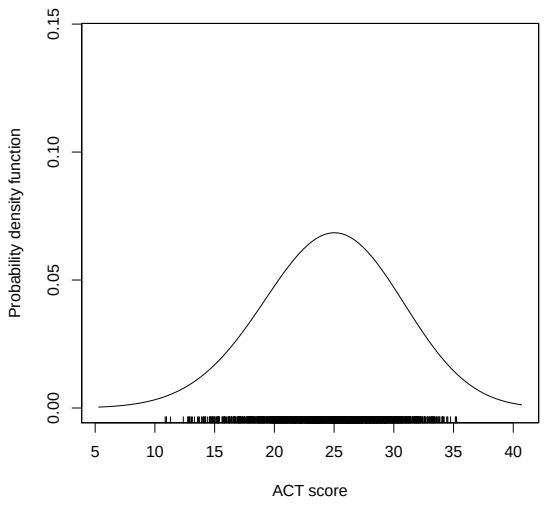

Figure 1: Density plot of the ACT scores.

```
# 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 approximately normal and indicates that freshmen at this institution 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.

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

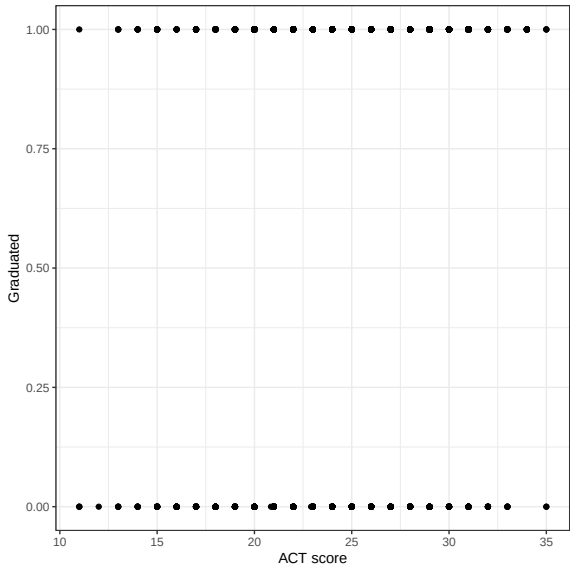

Figure 2: Scatterplot of whether a student graduated versus ACT score.

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. 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.)

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

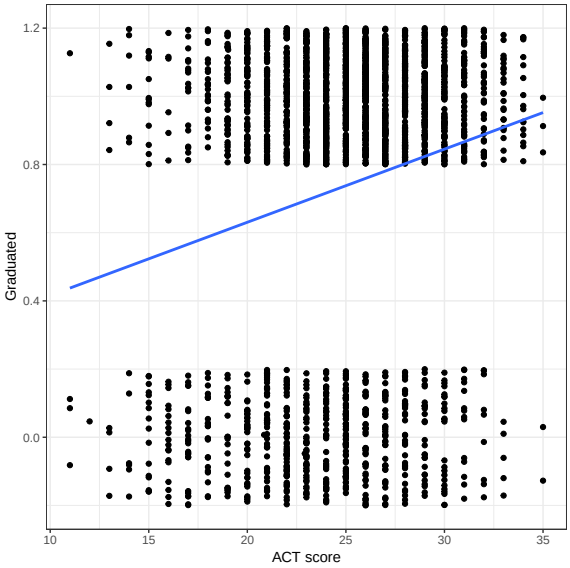

Figure 3: Scatterplot of whether a student graduated versus ACT score. The points have been jittered to alleviate over-plotting, and the regression smoother has also been added to the plot.

This spreads out the observations vertically, so we no longer have the problem of overplotting. But, the relationship is still difficult to discern. 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 statistically significant, $F(1,\ 2342) = 92.8$, $p < .001$.

```
# 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.

```
out = augment(lm.1)
#head(out)

# Examine normality assumption
sm.density(out$.std.resid, xlab = "Standardized residuals")
```

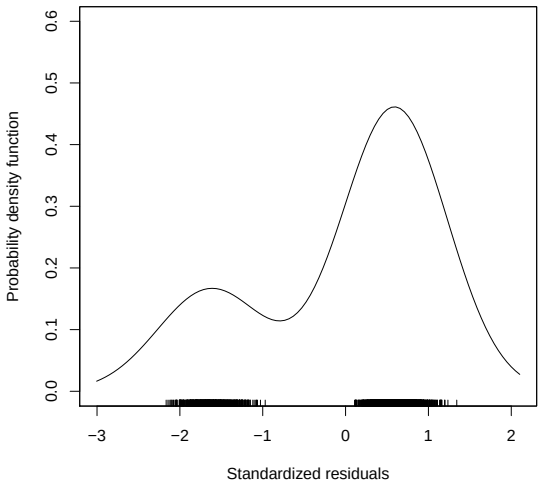

Figure 4: Density plot of the standardized residuals from the linear probability model.

```
# 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")
```

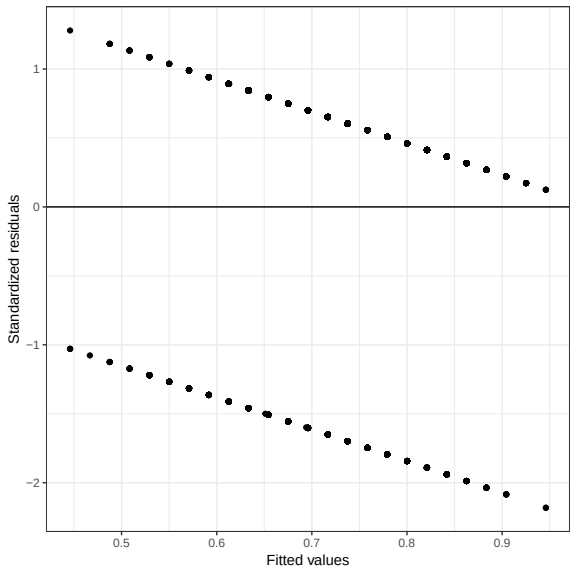

Figure 5: Scatterplot of the standardized residuals versus the fitted values from the linear probability model.

It is clear that the assumptions associated with linear regression are violated. First off, the residuals are not normally distributed. They are in fact, bimodal. The scatterplot of the residuals versus the fitted values also indicates violation of the linearity assumption, as the average residual at each fitted value is not zero.

Understanding the Residuals from the Linear Probability Model

Look closely at the scatterplot of the residuals vs. the fitted values. At each fitted value, there are only two residual values. Why is this? Recall that residuals are computed as $\epsilon_i = Y_i - \hat{Y}_i$. Now, remember that Y_i can only be one of two values, 0 or 1. Also remember that in the linear probability model $\hat{Y}_i = \hat{\pi}_i$. Thus, for $Y = 0$,

$$\begin{aligned} \epsilon_i &= 0 - \hat{Y}_i \\ &= -\hat{\pi}_i \end{aligned}$$

And, if $Y = 1$,

$$\begin{aligned} \epsilon_i &= 1 - \hat{Y}_i \\ &= 1 - \hat{\pi}_i \end{aligned}$$

This means that the residual computed using a particular fitted value (or from a particular ACT value given that the fitted values is a function of X) can only take on one of two values: $-\hat{\pi}_i$ or $1 - \hat{\pi}_i$. (Plotting the residuals for all fitted values, the marginal distribution of the residuals, results in a bimodal distribution for this same reason.)