Chapter 11
Regression with a Binary Dependent Variable

■ Solutions to Exercises

1. (a) The *t*-statistic for the coefficient on *Experience* is 0.031/0.009 = 3.44, which is significant at the 1% level.

   (b) $z_{Matthew} = 0.712 + 0.031 \times 10 = 1.022; \Phi(1.022) = 0.847$

   (c) $z_{Christopher} = 0.712 + 0.031 \times 0 = 0.712; \Phi(0.712) = 0.762$

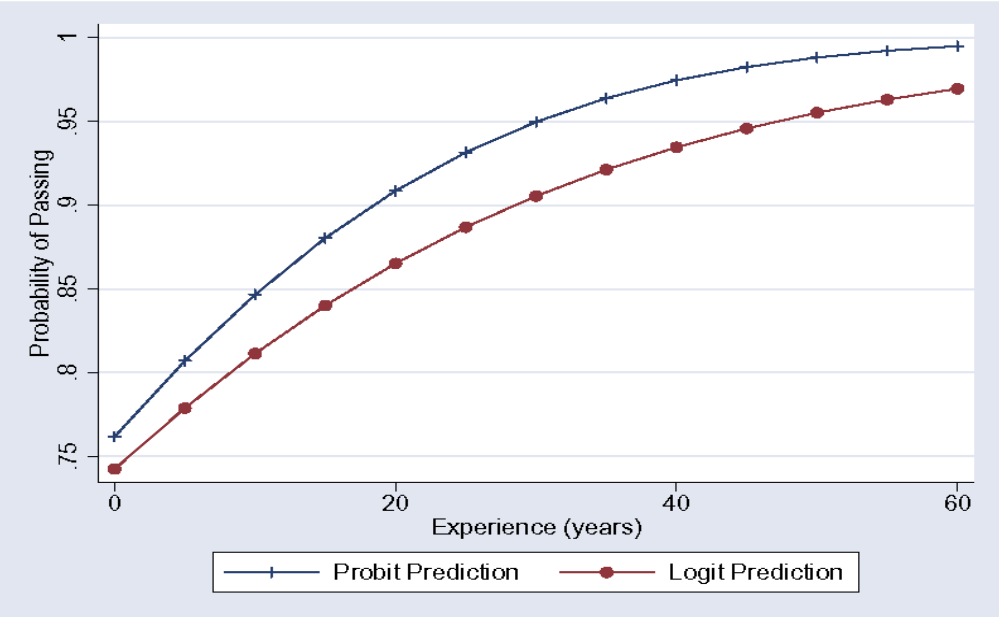
   (d) $z_{Jed} = 0.712 + 0.031 \times 80 = 3.192; \Phi(3.192) = 0.999,$ this is unlikely to be accurate because the sample did not include anyone with more that 40 years of driving experience.

2. (a) The *t*-statistic for the coefficient on *Experience* is $t = 0.040/0.016 = 2.5$, which is significant at the 5% level.

   (b) $\text{Prob}_{Matthew} = \dfrac{1}{1+e^{-(1.059 + 0.040 \times 10)}} = \dfrac{1}{1+e^{-1.459}} = 0.811$

   (c) $\text{Prob}_{Christopher} = \dfrac{1}{1+e^{-(1.059 + 0\ .040 \times 0)}} = \dfrac{1}{1+e^{-1.059}} = 0.742$

   (d)



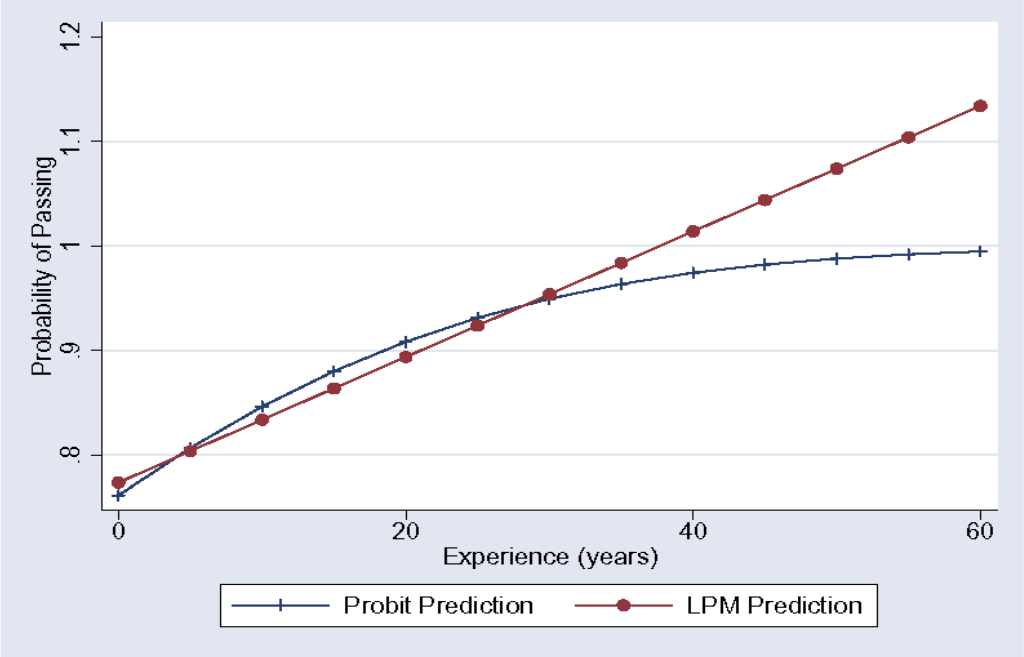
The shape of the regression functions are similar, but the logit regression lies below the probit regession for experience in the range of 0 = 60 years.

3. (a) The t-statistic for the coefficient on *Experience* is $t = 0.006/0.002 = 3$, which is significant a the 1% level.

(b) $Prob_{Matther} = 0.774 + 0.006 \times 10 = 0.836$

(c) $Prob_{Christopher} = 0.774 + 0.006 \times 0 = 0.774$

(d)



The probabilities are similar except when experience in large (>40 years). In this case the LPM model produces nonsensical results (probabilities greater than 1.0).

4. (a)

Group	Probit	Logit	LPM
Men	$\Phi(1.282 - 0.333) = 0.829$	$\frac{1}{1+e^{-(2.197-0.622)}} = 0.829$	0.829
Women	$\Phi(1.282) = 0.900$	$\frac{1}{1+e^{-(2.197)}} = 0.900$	0.900

(b) Because there is only regressor and it is binary (*Male*), estimates for each model show the fraction on males and females passing the test. Thus, the results are identical for all models.

5. (a) $\Phi(0.806 + 0.041 \times 10 \times 0.174 \times 1 - 0.015 \times 1 \times 10) = 0.814$
   (b) $\Phi(0.806 + 0.041 \times 2 - 0.174 \times 0 - 0.015 \times 0 \times 2) = 0.813$
   (c) The *t*-stat on the interaction term is –0.015/0.019 = –0.79, which is insignificant at the 10% level.

6. (a) For a black applicant having a P/I ratio of 0.35, the probability that the application will be denied is $\Phi(-2.26 + 2.74 \times 0.35 + 0.71) = \Phi(-0.59) = 27.76\%$.
   (b) With the P/I ratio reduced to 0.30, the probability of being denied is $\Phi(-2.26 + 2.74 \times 0.30 + 0.71) = \Phi(-0.73) = 23.27\%$. The difference in denial probabilities compared to (a) is 4.4 percentage points lower.
   (c) For a white applicant having a P/I ratio of 0.35, the probability that the application will be denied is $\Phi(-2.26 + 2.74 \times 0.35) = 9.7\%$. If the P/I ratio is reduced to 0.30, the probability of being denied is $\Phi(-2.26 + 2.74 \times 0.30) = 7.5\%$. The difference in denial probabilities is 2.2 percentage points lower.
   (d) From the results in parts (a)–(c), we can see that the marginal effect of the P/I ratio on the probability of mortgage denial depends on race. In the probit regression functional form, the marginal effect depends on the level of probability which in turn depends on the race of the applicant. The coefficient on *black* is statistically significant at the 1% level.

7. (a) For a black applicant having a P/I ratio of 0.35, the probability that the application will be denied is $F(-4.13 + 5.37 \times 0.35 + 1.27) = \frac{1}{1+e^{0.9805}} = 27.28\%$.
   (b) With the P/I ratio reduced to 0.30, the probability of being denied is $F(-4.13 + 5.37 \times 0.30 + 1.27) = \frac{1}{1+e^{1.249}} = 22.29\%$. The difference in denial probabilities compared to (a) is 4.99 percentage points lower.
   (c) For a white applicant having a P/I ratio of 0.35, the probability that the application will be denied is $F(-4.13 + 5.37 \times 0.35) = \frac{1}{1+e^{2.2505}} = 9.53\%$. If the P/I ratio is reduced to 0.30, the probability of being denied is $F(-4.13 + 5.37 \times 0.30) = \frac{1}{1+e^{2.519}} = 7.45\%$. The difference in denial probabilities is 2.08 percentage points lower.
   (d) From the results in parts (a)–(c), we can see that the marginal effect of the P/I ratio on the probability of mortgage denial depends on race. In the logit regression functional form, the marginal effect depends on the level of probability which in turn depends on the race of the applicant. The coefficient on *black* is statistically significant at the 1% level. The logit and probit results are similar.

8. (a) Since Y_i is binary variable, we know $E(Y_i|X_i) = 1 \times \Pr(Y_i = 1|X_i) + 0 \times \Pr(Y_i = 0|X_i) = \Pr(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i$. Thus

$$\begin{aligned} E(u_i|X_i) &= E[Y_i - (\beta_0 + \beta_1 X_i)|X_i] \\ &= E(Y_i|X_i) - (\beta_0 + \beta_1 X_i) = 0 \end{aligned}$$

   (b) Using Equation (2.7), we have

$$\begin{aligned} \text{var}(Y_i|X_i) &= \Pr(Y_i = 1|X_i)[1 - \Pr(Y_i + 1|X_i)] \\ &= (\beta_0 + \beta_1 X_i)[1 - (\beta_0 + \beta_1 X_i)]. \end{aligned}$$

   Thus

$$\begin{aligned} \text{var}(u_i|X_i) &= \text{var}[Y_i - (\beta_0 + \beta_1 X_i)_i|X_i] \\ &= \text{var}(Y_i|X_i) = (\beta_0 + \beta_1 X_i)[1 - (\beta_0 + \beta_1 X_i)]. \end{aligned}$$

(c) $\text{var}(u_i|X_i)$ depends on the value of X_i, so u_i is heteroskedastic.

(d) The probability that $Y_i = 1$ conditional on X_i is $p_i = \beta_0 + \beta_1 X_i$. The conditional probability distribution for the ith observation is $\Pr(Y_i = y_i | X_i) = p_i^{y_i}(1-p_i)^{1-y_i}$. Assuming that (X_i, Y_i) are i.i.d., $i = 1,\ldots, n$, the joint probability distribution of $Y_1,\ldots, Y_n$ conditional on the $X's$ is

$$\begin{aligned}\Pr(Y_1 = y_1,\ldots, Y_n = y_n | X_1,\ldots, X_n) &= \prod_{i=1}^{n} \Pr(Y_i = y_i | X_i) \\ &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \prod_{i=1}^{n} (\beta_0 + \beta_1 X_i)^{y_i}[1-(\beta_0 + \beta_1 X_i)]^{1-y_i}.\end{aligned}$$

The likelihood function is the above joint probability distribution treated as a function of the unknown coefficients (β_0 and β_1).

9. (a) The coefficient on *black* is 0.084, indicating an estimated denial probability that is 8.4 percentage points higher for the black applicant.

(b) The 95% confidence interval is $0.084 \pm 1.96 \times 0.023 = [3.89\%, 12.91\%]$.

(c) The answer in (a) will be biased if there are omitted variables which are race-related and have impacts on mortgage denial. Such variables would have to be related with race and also be related with the probability of default on the mortgage (which in turn would lead to denial of the mortgage application). Standard measures of default probability (past credit history and employment variables) are included in the regressions shown in Table 9.2, so these omitted variables are unlikely to bias the answer in (a). Other variables such as education, marital status, and occupation may also be related the probability of default, and these variables are omitted from the regression in column. Adding these variables (see columns (4)–(6)) have little effect on the estimated effect of *black* on the probability of mortgage denial.

10. (a) Let $n_1 = \#(Y = 1)$, the number of observations on the random variable Y which equals 1; and $n_2 = \#(Y = 2)$. Then $\#(Y = 3) = n - n_1 - n_2$. The joint probability distribution of $Y_1,\ldots, Y_n$ is

$$\Pr(Y_1 = y_1,\ldots, Y_n = y_n) = \prod_{i=1}^{n} \Pr(Y_i = y_i) = p^{n_1} q^{n_2} (1-p-q)^{n-n_1-n_2}.$$

The likelihood function is the above joint probability distribution treated as a function of the unknown coefficients (p and q).

(b) The MLEs of p and q maximize the likelihood function. Let's use the log-likelihood function

$$\begin{aligned}L &= \ln[\Pr(Y_1 = y_1,\ldots, Y_n = y_n)] \\ &= n_1 \ln p + n_2 \ln q + (n - n_1 - n_2)\ln(1-p-q).\end{aligned}$$

Using calculus, the partial derivatives of L are

$$\frac{\partial L}{\partial p} = \frac{n_1}{p} - \frac{n-n_1-n_2}{1-p-q}, \text{ and}$$
$$\frac{\partial L}{\partial q} = \frac{n_2}{q} - \frac{n-n_1-n_2}{1-p-q}.$$

Setting these two equations equal to zero and solving the resulting equations yield the MLE of p and q:

$$\hat{p} = \frac{n_1}{n}, \quad \hat{q} = \frac{n_2}{n}.$$

11. (a) This is a censored or truncated regression model (note the dependent variable might be zero).
    (b) This is an ordered response model.
    (c) This is the discrete choice (or multiple choice) model.
    (d) This is a model with count data.