1. Let *noPC* be a dummy variable equal to one if the student does not own a PC, and zero otherwise.

   i. If *noPC* is used instead of *PC* in the model of:

   \[ \text{colGPA} = \beta_0 + \delta_0 \text{PC} + \beta_1 \text{hsGPA} + \beta_2 \text{ACT} + u, \]

   what happens to the intercept in estimated equation? What will be the coefficient of *noPC*? [Hint: Write *PC* = 1 – *noPC*, and plug this into the estimated equation]

   (Ans)

   Following the hint, \( \text{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 (1 - \text{noPC}) + \hat{\beta}_1 \text{hsGPA} + \hat{\beta}_2 \text{ACT} = \) \((\hat{\beta}_0 + \hat{\delta}_0) - \hat{\delta}_0 \text{noPC} + \hat{\beta}_1 \text{hsGPA} + \hat{\beta}_2 \text{ACT}\). For the specific estimates in equation (7.6) in the textbook, \( \hat{\beta}_0 = 1.26 \) and \( \hat{\delta}_0 = .157 \), so the new intercept is \( 1.26 + .157 = 1.417 \). The coefficient on *noPC* is \(-.157\).

   ii. What will happen to the R-squared if *noPC* is used instead of *PC*?

   (Ans)

   Nothing happens to the R-squared. Using *noPC* in place of *PC* is simply a different way of including the same information on PC ownership.

   iii. Should *PC* and *noPC* both be included as independent variables in the model? Explain.

   (Ans)

   It makes no sense to include both dummy variables in the regression: we cannot hold *noPC* fixed while changing *PC*. We have only two groups based on *PC* ownership so, in addition to the overall intercept, we need only to include one dummy variable. If we try to include both along with an intercept we have perfect multicollinearity (the dummy variable trap).

2. Let *d* be a dummy (binary) variable and let *z* be a quantitative variable. Consider the model:

   \[ y = \beta_0 + \delta_0 d + \beta_z z + \delta_d d \cdot z + u; \]
this is a general version of a model with an interaction between a dummy variable and a quantitative variable.

i. Since it changes nothing important, set the error to zero, \( u = 0 \). Then, when \( d = 0 \) we can write the relationship between \( y \) and \( z \) as the function \( f_0(z) = \beta_0 + \beta_1 z \).

Write the same relationship when \( d = 1 \), where you should use \( f_1(z) \) on the left-hand side to denote the linear function of \( z \).

(Ans)

Plugging in \( u = 0 \) and \( d = 1 \) gives \( f_1(z) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)z \).

ii. Assume that \( \delta_1 \neq 0 \) (which means the two lines are not parallel), show that the value of \( z^* \) such that \( f_0(z^*) = f_1(z^*) \) is \( z^* = -\delta_0 / \delta_1 \). This is the point at which the two lines intersect. Argue that \( z^* \) is positive if and only if \( \delta_0 \) and \( \delta_1 \) have opposite sign.

(Ans)

Setting \( f_0(z^*) = f_1(z^*) \) gives \( \beta_0 + \beta_1 z^* = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)z^* \) or \( 0 = \delta_0 + \delta_1 z^* \).

Therefore, provided \( \delta_1 \neq 0 \), we have \( z^* = -\delta_0 / \delta_1 \). Clearly, \( z^* \) is positive if and only if \( \delta_0 / \delta_1 \) is negative, which means \( \delta_0 \) and \( \delta_1 \) must have opposite signs.

iii. Using the data in TWOYEAR.DTA, the following equation can be estimated:

\[
\log(\text{wage}) = 2.289 - 0.357 \text{ female} + 0.50 \text{ totcoll} + 0.030 \text{ female} \cdot \text{totcoll}
\]

\[
(0.011) \quad (0.015) \quad (0.003) \quad (0.005)
\]

\( n = 6,763 \), \( R^2 = 0.202 \)

where all coefficients and standard errors have rounded to three decimal places. Using this equation, find the value of \( \text{totcoll} \) such that the predicted values of \( \log(\text{wage}) \) are the same for men and women.

(Ans)

Using part (ii) we have \( \text{totcoll}^* = 0.357 / 0.030 = 11.9 \) years.

iv. Based on the equation in part iii, can women realistically get enough years of college so that their earning catch up to those of men? Explain.
The estimated years of college where women catch up to men is much too high to be practically relevant. While the estimated coefficient on female·totcoll shows that the gap is reduced at higher levels of college, it is never closed – not even close. In fact, at four years of college, the difference in predicted log wage is still $-.357 + .030(4) = -.237$, or about 21.1% less for women.

**Computer Exercise**

3. Use the data in SLEEP75.DTA for this exercise. The equation of interest is

\[ \text{sleep} = \beta_0 + \beta_1 \text{totwrk} + \beta_2 \text{educ} + \beta_3 \text{age} + \beta_4 \text{age}^2 + \beta_5 \text{yngkid} + u \]

i. Estimate this equation separately for men and women and report the results in the usual form. Are there notable differences in the two estimated equation?

(Ans)

The estimated equation for men is

\[
\hat{\text{sleep}} = 3,648.2 - .182 \text{totwrk} - 13.05 \text{educ} + 7.16 \text{age} - .0448 \text{age}^2 + 60.38 \text{yngkid}
\]

\[
(310.0) \quad (.024) \quad (7.41) \quad (14.32) \quad (.1684) \quad (59.02)
\]

\[ n = 400, \quad R^2 = .156 \]

and the estimated equation for women is

\[
\hat{\text{sleep}} = 4,238.7 - .140 \text{totwrk} - 10.21 \text{educ} - 30.36 \text{age} - .368 \text{age}^2 - 118.28 \text{yngkid}
\]

\[
(384.9) \quad (.028) \quad (9.59) \quad (18.53) \quad (.223) \quad (93.19)
\]

\[ n = 306, \quad R^2 = .098. \]

There are certainly notable differences in the point estimates. For example, having a young child in the household leads to less sleep for women (about two hours a week) while men are estimated to sleep about an hour more. The quadratic in age is a hump-shape for men but a U-shape for women. The intercepts for men and women are also notably different.

ii. Compute the Chow test for equality of parameters in the sleep equation for men and women. Use the form of the test that adds male and the interaction term male·totwork, ..., male·yngkid and uses the full set of observations. What are the
relevant $df$ for the test? Should you reject the null at the 5% level?

(Ans)
The $F$ statistic (with 6 and 694 $df$) is about 2.12 with $p$-value $\approx .05$, and so we reject the null that the sleep equations are the same at the 5% level.

iii. Now, allow for a different intercept for males and females and determine whether the interaction terms involving $male$ are jointly significant.

(Ans)
If we leave the coefficient on $male$ unspecified under $H_0$, and test only the five interaction terms, $male \cdot totwrk$, $male \cdot educ$, $male \cdot age$, $male \cdot age^2$, and $male \cdot yngkid$, the $F$ statistic (with 5 and 694 $df$) is about 1.26 and $p$-value $\approx .28$.

iv. Given the results from parts ii and iii, what would be your final model?

(Ans)
The outcome of the test in part (iii) shows that, once an intercept difference is allowed, there is not strong evidence of slope differences between men and women. This is one of those cases where the practically important differences in estimates for women and men in part (i) do not translate into statistically significant differences. We need a larger sample size to confidently determine whether there are differences in slopes. For the purposes of studying the sleep-work tradeoff, the original model with $male$ added as an explanatory variable seems sufficient.

4. There has been much interest in whether the presence of 401(k) pension plans, available to many U.S. workers, increases net savings. The data set 401KSUBS.DTA contains information on net financial assets ($nettfa$), family income ($inc$), a binary variable for eligibility in a 401(k) plan ($e401k$), and several other variables.

i. What fraction of the families in the sample are eligible for participation in a 401(k) plan?

(Ans)
About .392, or 39.2%.

ii. Estimate a linear probability model explaining 401(k) eligibility in terms of income, age, and gender. Include income and age in quadratic form, and report the results in
the usual form.

(Ans)

The estimated equation is

\[ \hat{e}_{401k} = -0.506 + 0.0124 \text{inc} - 0.000062 \text{inc}^2 + 0.0265 \text{age} - 0.0031 \text{age}^2 - 0.0035 \text{male} \]

\[ (0.081) (0.0006) (0.000005) (0.0039) (0.00005) (0.0121) \]

\[ n = 9,275, \quad R^2 = 0.094. \]

iii. Would you say that 401(k) eligibility is independent of income and age? What about gender? Explain.

(Ans)

401(k) eligibility clearly depends on income and age in part (ii). Each of the four terms involving inc and age have very significant t statistics. On the other hand, once income and age are controlled for, there seems to be no difference in eligibility by gender. The coefficient on male is very small – at given income and age, males are estimated to have a .0035 lower probability of being 401(k) eligible – and it has a very small t statistic.

iv. Obtain the fitted values from the linear probability model estimated in part ii. Are any fitted values negative or greater than one?

(Ans)

Somewhat surprisingly, out of 9,275 fitted values, none is outside the interval [0,1]. The smallest fitted value is about .030 and the largest is about .697. This means one theoretical problem with the LPM – the possibility of generating silly probability estimates – does not materialize in this application.

v. Using the fitted values \( \hat{e}_{401k_i} \) from part iv, define \( \hat{e}_{401k_i} = 1 \) if \( \hat{e}_{401k_i} \geq 0.5 \) and \( \hat{e}_{401k_i} = 0 \) if \( \hat{e}_{401k_i} < 0.5 \). Out of 9,275 families, how many are predicted to be eligible for a 401(k) plans?

(Ans)

Using the given rule, 2,460 families are predicted to be eligible for a 401(k) plan.

vi. For the 5,638 families not eligible for a 401(k), what percentage of these are
predicted not to have a 401(k), using the predictor $e^{401k}$? For the 3,637 families eligible for a 401(k) plan, what percentage are predicted to have one?

(Ans)

Of the 5,638 families actually ineligible for a 401(k) plan, about 81.7 are correctly predicted not to be eligible. Of the 3,637 families actually eligible, only 39.3 percent are correctly predicted to be eligible.

vii. The overall percent correctly predicted is about 64.9%. Do you think this is a complete description of how well the model does, given your answer in part vi?

(Ans)

The overall percent correctly predicted is a weighted average of the two percentages obtained in part (vi). As we saw there, the model does a good job of predicting when a family is ineligible. Unfortunately, it does less well – predicting correctly less than 40% of the time – in predicting that a family is eligible for a 401(k).

viii. Add the variable $pria$ as an explanatory variable to the linear probability model. Other things equal, if a family has someone with an individual retirement account, how much higher is the estimated probability that the family is eligible for 401(k) plan? Is it statistically different from zero at the 10% level?

(Ans)

The estimated equation is

$$
e^{401k} = -0.502 + 0.0123 \text{inc} - 0.00061 \text{inc}^2 + 0.0265 \text{age} - 0.0031 \text{age}^2$$

$$-0.0038 \text{male} + 0.0198 \text{pira}$$

$$n = 9,275, \quad R^2 = 0.095.$$