**Exercise Solutions Part I: Instructor and Student Access**

**Chapter 2**

**2.6.2 Exercise 2: Predicting Educational Attainment**

**1.**

 R2 is low (0.228), indicating a rather unsatisfactory fit of the regression model. Only 22.8% of the variance in education in this sample is explained by this set of X’s, and there will be substantial differences between actual and predicted Y values for many of the cases in the sample. Among the independent variables, number of siblings and childhood household income had statistically significant effects on education, controlling for other X’s. From this sample, a person’s predicted education decreases by 0.121 years as their number of siblings increases by one, controlling for the other independent variables. The mean of number of siblings is 2.6, and its range is between 0 and 6, so a one-unit increase in the number of siblings is appropriately sized for interpretation. Relative to the mean (14.3) and range (9–21) of education in the sample, a decrease of 0.121 years (approximately a month and a half) in predicted education seems too small to suggest much of a real-world impact.

 Results also show that a person’s predicted education increases 0.002 years as their childhood household income increases $1, controlling for other X’s in the model. It is more informative to consider a larger change in childhood household income, as there is a very large mean ($54,567) and wide range ($24,000 to $78,000) for this variable in the sample. If childhood household income increases by $1,000, predicted education increases by 2 years, holding other X’s constant. This change in predicted education seems large in light of the sample mean (14.3) and range (9–21) of respondent’s education, so this effect appears to have real-life significance.

 The p-value of parents’ education (0.056) indicates that the effect of parent’s education on one’s education is borderline statistically significant, so it is reasonable to proceed with interpretation of the effect. b shows that the estimate from this sample is that a person’s predicted education increases by 1.122 years as their parent’s education increases by one year, holding other X’s constant. Considering the descriptive statistics of parents’ education, a one-year increase seems appropriate to examine. In light of the descriptive statistics for respondent’s education in this sample, an increase of 1.122 years in predicted education seems to be large enough to indicate a real impact.

**2.**

(i) = 1.647

(ii) = –2.086

(iii) = 1.934

**3.**

Ŷ = a + b1 X1 + b2 X2 + b3 X3 + b4 X4

 = –95.5 + 0.028(105) + (–0.121)(3) + (1.122)(12) + (0.002)(46,000)

 = –95.5 + 2.94 – 0.363 + 13.464 + 92 = 12.541 years.

 None of these X-values are unusually large or small relative to the descriptive statistics, so in that respect it is reasonable to calculate the predicted income from them. However, the low R2 indicates that in the sample the predictions were not so successful, leading to some skepticism about the quality of a prediction like this as well.

**4.**

N = df + number of X’s + 1 = 103 + 4 + 1 = 108.

**5.**

 Three of the four X’s have statistically significant effects on respondent’s education. The question did not specify the hypothesized directions of the effects, but the directions found here seem sensible, even though the effect of the number of siblings on a person’s predicted education was not large enough to indicate a real impact. R2 is rather low, especially considering the relatively small sample size, so future analysis should consider additional X’s which might influence one’s education. Researchers should use existing social theories or past research to identify such independent variables. (Note that some likely choices, such as gender and race, are categorical rather than numerical, and so would fall under Chapter 3’s material.)

**2.6.3 Exercise 3**

**1.**

 R2 can indicate whether a set of X’s predicts Y well, so that is an assessment of the overall performance of the regression model. It helps indicate if the set of X’s being used to predict Y is reasonably comprehensive, or if the theory left out some important factors related to Y. Statistical significance helps assess whether each individual X suggested by the theory is actually related to Y.

**2.**

 Multiple regression attempts to separate out the effect of each X on Y while controlling for, or holding constant, other X’s. Just pasting together a series of simple regression results would not include this idea of “control” or “holding other X’s constant”.

**3.**

 Collinearity among X’s (high correlations among X’s) increases standard errors of the b’s in the regression. This leads to smaller t-statistics because the standard error of b forms the denominator of the expression for the t statistic: . Smaller t-statistics lead to larger p-values, and larger p-values make us less apt to reject H0, leading to artificial non-statistical significance of X’s in the presence of collinearity.

**Chapter 3**

Note that for exercises in Chapter 3 and the rest of the book, solutions reflect one choice of reference category for categorical independent variables rather than all possible choices. While, as discussed in the text, much of the regression output will be the same no matter which category is chosen as reference, there will be cosmetic differences in some parts of the output when different categories are chosen.

**3.7.2 Exercise 2: Fire Department Size**

**1*.***

 = = 2.757.

**2.**

 Ŷ = 8.56 – 1.02 X1 + 0.74 X2 + 2.13 X3 + 1.52 X4 – 0.87 X5 (where X1 = % unemployed, X2 = % bachelor’s degree, X3 = NE dummy, X4 = MW dummy, and X5 = W dummy)

 = 8.56 – 1.02(8) + 0.74(22) + 2.13(0) + 1.52(1) – 0.87(0) = 18.2 firefighters per 10,000 population.

 None of these X-values are unusually large or small relative to the descriptive statistics, so in that respect it is reasonable to calculate the predicted income from them. The moderate R2 indicates that the model fit and predictions in the sample were adequate, but we still should be cautious about the quality of a prediction, and there may be additional X’s to consider.

**3.**

 From the output (with “S” as the reference category), the order of Ŷ (predicted number of firefighters per 100,000 population) is “W”, “S”, “MW”, and “NE”. The difference in Ŷ between “W” and “S” is 0.87, between “S” and “MW” is 1.52, and between “MW” and “NE” is 0.61. Even if a different reference category is chosen, the order and differences in Ŷ among categories will stay the same. If the analyst had chosen “W” as reference, “W” would be given zero on the number line. Retaining the order and spacing in Ŷ among categories, so that each region is in the same relative position on the number line as before, the b’s would be 0.87 for “S” (this b indicates that Ŷ is 0.87 greater in the South than in the reference West; the b in the original analysis indicated that Ŷ is 0.87 less in the West than in the reference South), 2.39 (0.87 + 1.52) for “MW”, and 3.00 (2.39 + 0.61) for “NE”.

**4.**

 R2 and results for % unemployed and % holding bachelor’s degree (b’s, standard errors, t-statistics, and p-values) will remain the same as in the analysis with “S” as reference.

**5.**

 R2 is 0.534, indicating that the regression model is performing moderately well, but there may be other X’s that could be added to improve the model’s predictions of Y.53.4% of variance in the number of firefighters per 10,000 population is explained by the set of X’s.

 The effect of percentage unemployed on number of firefighters per 10,000 population is statistically significant, controlling for percentage holding a bachelor’s degree and region. Holding other X’s constant, predicted number of firefighters per 10,000 population decreases 1.02 as percentage unemployed increases 1%. The sample of 106 cities averages 9.3% for unemployment, with a range between 7.1% and 11.6%, so a 1% increase seems like a reasonable change in unemployment to consider. The corresponding increase of 1.02 in predicted number of firefighters does not seem too trivial, given the mean (17.3) and range (11.28 to 24.81) of number of firefighters per 10,000 population. Percentage holding a bachelor’s degree did not have a statistically significant effect on the number of firefighters per 10,000 population, controlling for percentage unemployed and region.

 The F calculation for region is below, and indicates that there is a statistically significant impact of region on number of firefighters per 10,000 population (p < 0.05; more precisely, 0.020 as reported from an online F-calculator or statistical software), controlling for other X’s:

 Holding percentage unemployed and percentage holding a bachelor’s degree constant, the predicted number of firefighters per 10,000 population for the Northeastern region is 2.13 higher than for the Southern region, and 1.52 higher for the Midwestern region than for the Southern region. These differences are statistically significant (from the individual dummies’ p-values) and certainly appear large enough to have real significance, based on the descriptive statistics of number of firefighters per 10,000 population. From the output, the p-value from the t-statistic for the W dummy is greater than 0.05 (p = 0.144), indicating that there is no statistically significant difference in predicted number of firefighters between the Southern and Western regions, holding other X’s constant. In sum, the predicted number of firefighters per 10,000 population is the highest in the Northeastern region, followed by the Midwestern region, controlling for other X’s. Holding other X’s constant, the Southern and Western regions had the lowest number of firefighters per 10,000 population.

**3.7.3 Exercise 3**

**1.**

 The null hypothesis (H0) states that the β’s for the set of dummy variables would all be zero in the true regression model (that we would obtain if our data included the whole population or superpopulation instead of a sample). The alternative hypothesis (H1) states that at least one of these β’s is not zero. To determine whether to reject or not reject the null hypothesis, we need to calculate the F-statistic and find the corresponding p-value. The F-statistic is calculated as SSE1 refers to the sum of squared errors from Model 1 that includes all the X’s (that is, the set of dummy variables we are testing, plus whatever other X’s are in the regression). SSE0 refers to the sum of squared errors from Model 0 that excludes the set of dummy variables (therefore including only the other X’s). s represents the number of dummy variables in the set we are testing, and df1 represents the degrees of freedom for t in Model 1, so that df1 = N – (# of X’s in Model 1) – 1.

 With the calculated F-statistic, s, and df1, we can find a p-value from the F-table (or an online calculator). If the p-value that we determine from the table is smaller than 0.05, we conclude that the categorical independent variable has a statistically significant impact on Y, and that the improvement in the accuracy of the predictions is substantial enough to be worth the complication of the additional variables being used in the regression. If the p-value is larger than 0.05, then we conclude that the categorical independent variable does not have a statistically significant impact on Y, and that the improvement in the accuracy of the predictions is not great enough to justify including the set of dummy variables.

**2.**

 The researcher is making assumptions that 1) Ŷ for the various categories is, holding other X’s constant, in the order or the reverse order of “Republican”, “Democratic”, and “Other”, and 2) the spacing between racial categories in terms of Ŷ is equal, meaning that the difference in Ŷ is the same between Republican and Democratic as it is between Democratic and Other.

**Chapter 4**

**4.6.2 Exercise 3: Public Transportation and Health Benefit**

**1.**

 Ŷ = a + b1 X1 + b2a X2a + b3 X2b + b4 X1X2a + b5 X1X2b.

**2.**

 A city with mild weather will have X2a (dummy for mild weather) = 1 and X2b (dummy for moderate weather) = 0, so

Ŷ = a + b1 X1 + b2 (1) + b3 (0) + b4 X1 (1) + b5 X1 (0)

 = a + b1 X1 + b2 + b4 X1

 = (a + b2) + (b1 + b4) X1.

From the rewritten equation for a city with mild weather, the estimated effect of the public transit score for a city with mild winter weather is b1 + b4.

**4.6.3 Exercise 3: Interaction between Two Numerical Variables**

**(a)**

 If X1 = age, and X2 = education, the regression equation with interaction terms between age and education can be written as Ŷ = a+ b1 X1 + b2 X2 + b3 X1X2. To express the estimated effect of age on happiness as a function of education, the above equation can be rewritten as

Ŷ = a + (b1 + b3 X2) X1 + b2 X2

 = a + (–3.32 + 0.16 X2) X1 + 7.85 X2.

Then (–3.32 + 0.16 X2) is the estimated effect of age on happiness as a function of education.

**(b)**

 To express the estimated effect of education on happiness as a function of age, the above equation can be rewritten as

Ŷ = a + b1 X1 + b2 X2 + b3 X1X2

 = a + b1 X1 + (b2 + b3 X1) X2

 = a + (–3.32) X1 + (7.85 + 0.16 X1) X2.

From this, (7.85 + 0.16 X1) is the estimated effect of education on happiness as a function of age.

**(c)**

 If X2 = 16, the estimated effect of age on happiness is –3.32 + 0.16 (16) = –0.764.

**4.6.4 Exercise 4: Interaction between Two Categorical Variables**

**1.**

 Ŷ = a + b1 X1 + b2 X2 + b3 X3 + b4 X4 + b5 X5 + b6 X1X3 + b7 X1X4 + b8 X1X5 + b9 X2X3 + b10 X2X4 + b11 X2X5,

where X1 = Urban dummy, X2 = Suburban dummy, X3 = Democratic dummy, X4 = Republican dummy, and X5 = minor party dummy. (Here we assume that there are no other independent variables, but the equation could be extended with additional terms involving those as needed.)

**2.**

 There will be 12 possible combinations of types of residence (three categories) and political party identification (4 categories). The predicted values (or relative predicted values, if there are additional independent variables in the equation) of Y for each combination of values of the type of residence and party identification are as follows:

Ŷ = a + b1 + b3 + b6 for urban Democrats (X1 = 1, X2 = 0, X3 = 1, X4 = 0, and X5 = 0),

Ŷ = a + b1 + b4 + b7 for urban Republicans (X1 = 1, X2 = 0, X3 = 0, X4 = 1, and X5 = 0),

Ŷ = a + b1 + b5 + b8 for urban minor identifiers (X1 = 1, X2 = 0, X3 = 0, X4 = 0, and X5 = 1),

Ŷ = a + b1 for urban independents (X1 = 1, X2 = 0, X3 = 0, X4 = 0, and X5 = 0),

Ŷ = a + b2 + b3 + b9 for suburban Democrats (X1 = 0, X2 = 1, X3 = 1, X4 = 0, and X5 = 0),

Ŷ = a + b2 + b4 + b10 for suburban Republicans (X1 = 0, X2 = 1, X3 = 0, X4 = 1, and X5 = 0),

Ŷ = a + b2 + b5 + b11 for suburban minor identifiers (X1 = 0, X2 = 1, X3 = 0, X4 = 0, and X5 = 1),

Ŷ = a + b2 for suburban independents (X1 = 0, X2 = 1, X3 = 0, X4 = 0, and X5 = 0),

Ŷ = a + b3 for rural Democrats (X1 = 0, X2 = 0, X3 = 1, X4 = 0, and X5 = 0),

Ŷ = a + b4 for rural Republicans (X1 = 0, X2 = 0, X3 = 0, X4 = 1, and X5 = 0),

Ŷ = a + b5 for rural minor identifiers (X1 = 0, X2 = 0, X3 = 0, X4 = 0, and X5 = 1),

Ŷ = a for rural independents (X1 = 0, X2 = 0, X3 = 0, X4 = 0, and X5 = 0).

The predicted values (or relative predicted values, if there are additional independent variables) of Y for each combination of type of residence and party identification can be summarized in the chart as follows. If desired, we could remove the intercept a from each expression, because it appears in every cell; if we did, then the lower right corner cell would become 0.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Urban  | Suburban | Rural |
| Democrat  | a + b1 + b3 + b6 | a + b2 + b3 + b9 | a + b3 |
| Republican | a + b1 + b4 + b7 | a + b2 + b4 + b10 | a + b4 |
| Minor party | a + b1 + b5 + b8 | a + b2 + b5 + b11 | a + b5 |
| Independent  | a + b1 | a + b2 | a |

**Chapter 5**

**5.6.1 Exercise 1 – Calculator Work with Logs**

**1.**

 **(a)**

 Controlling for X2, Ŷ is multiplied by e0.08 = 1.083 as X1 increases by 1 unit. In other words, Ŷ increases by 8.3% as X1 increases by 1 unit. Controlling for X1, Ŷ is multiplied by e.-0.15 = 0.861 as X2 increases 1 unit. That is, Ŷ decreases by 13.9% as X2 increases by 1 unit. These numbers make sense because a positive b translates into a multiplicative increase in Ŷ, while a negative b translates into a multiplicative decrease in Ŷ. If b > 0, then eb > 1, so Ŷ increases as X increases (positive relationship between X and Y). If b < 0, then eb < 1, so Ŷ decreases as X increases (negative relationship between X and Y).

 **(b)**

 For X1, Ŷ for the named dummy category is 1.083 times Ŷ for the reference category, holding X2 constant; that is, Ŷ for the named dummy category is 8.3% greater than Ŷ for the reference category. For X2, Ŷ for the dummy category is 0.861 times Ŷ for the reference category, holding X1 constant, which means that Ŷ for the dummy category is 13.9% less than Ŷ for the reference category.

 **(c)**

 = 1.03 + 0.08 X1 – 0.15 X2.

 = 1.03 + 0.08 (8) – 0.15 (12).

 = –0.13.

 Ŷ = = = 0.878.

 **(d)**

 One cannot make a general interpretation involving an additive change in Ŷ when Y has been logged.

**2.**

 **(a)**

 Ŷ increases (additively) by 85.32 as X1 is multiplied by e (or 2.718), while controlling for X2.

 **(b)**

 Ŷ increases by 85.32 × log 1.3 = 85.32 × 0.262 = 22.35 as X1 is increased to 1.3 times its previous value, holding X2 constant. For the second part, 25% smaller is equivalent to multiplying by 0.75, so in that case Ŷ increases by 85.32 × log 0.75 = 85.32 × –0.288 = –24.57—that is, decreases by 24.57—as X1 becomes 25% smaller (is multiplied by 0.75), holding X2 constant.

 **(c)**

 Ŷ = 8.1 + 85.32 log (4) + 1.24 (3)

 = 8.1 + (85.32 × 1.386) + 3.72

 = 8.1 + 118.25 + 3.72

 = 130.07.

**3.**

 **(a)**

 Ŷ is multiplied by e–1.1 = 0.333 as X1 is multiplied by e, holding X2 constant. Ŷ is multiplied by e0.83 = 2.293 as X2 is multiplied by e, holding X1 constant.

 **(b)**

 Ŷ is multiplied by e–1.1\*log3 = e–1.1\*1.099= e–1.209 = 0.298 as X1 is tripled (multiplied by 3), holding X2 constant. This can also be found by 3–1.1 = 0.299; the small difference in these results is simply due to rounding in the various steps of the first approach.

 **(c)**

 = 2.03 – 1.1 log (15) + 0.83 log (46).

 = 2.03 – (1.1 × 2.708) + (0.83 × 3.829).

 = 2.03 – 2.979 + 3.178.

 = 2.229.

 Ŷ = = e2.229 = 9.291.

It also works to start with the model for Ŷ written in multiplicative terms and work through the calculations from that, but most students find it much more straightforward to calculate the predicted log Y and then use the antilog at the end.

**Chapter 6**

**6.4.1 Exercise 1: Interpreting with X and X2**

 The F-test comparing the models with and without X1 and X12 (, giving p = 0.013) is supportive of the statistically significant relationship between X1 and Y, and warrants further investigation of the relationship. The p-value from the individual t-statistic for X12 is small enough to show statistical significance (p = 0.030), and thus the relationship between X1 and Y appears to be nonlinear. The b’s for X1 and X12 have different signs, positive for X1 and negative for X12, indicating that X1 and Ŷ have an upside-down U-shaped relationship, while holding X2 constant. (If X1 is not restricted to positive values, then the upside-down U-shaped relationship is indicated by the negative coefficient on X12, with this conclusion not relying on the sign of the b for X1.) At lower values of X1, there is a positive relationship with Ŷ, but beyond some value of X1 the relationship becomes negative. The estimated value of X1 at which the relationship changes sign is calculated as –b1/2b2 = –0.450 / 2(–0.032) = 7.031. The effect of X2 on Y is statistically significant, net of X1 and X12. For each additional one-unit increase in X2, Ŷ increases by 0.231, holding other X’s constant.

**Chapter 7**

**7.5.2 Exercise 2**

**1.**

 **(a)**

The odds ratio for the “foreign born” dummy is e0.078 = 1.081. The estimated odds of agreeing for a foreign-born person are 1.081 times those for a U.S.-born person, holding other independent variables constant. In other words, the estimated odds of agreement for foreign-born individuals are 8.1% higher than those for U.S.-born individuals.

 **(b)**

The odds ratio for the foreign-born dummy is e-0.025 = 0.975. The estimated odds of agreeing for foreign-born people are 0.975 times those for U.S.-born people, holding other independent variables constant. In other words, the estimated odds of agreement for foreign-born individuals are 2.5% lower than those for U.S.-born individuals.

**2.**

 **(a)**

The odds ratio for years living at current residence is e0.12 = 1.127. As years living at current residence increase by one, the estimated odds of agreeing are multiplied by 1.127, controlling for other independent variables. This can also be expressed as a 12.7% increase in the estimated odds.

 **(b)**

The odds ratio for years living at current residence is e-0.07 = 0.932. As years living at current residence increase by one, the estimated odds of agreeing are multiplied by 0.932, holding other independent variables constant. This can also be expressed as a 6.8% decrease in the estimated odds.

**3.**

As study time increases by seven hours, holding other independent variables constant, the estimated odds of passing are multiplied by e0.096 ×7 = e0.672 = 1.958, equivalent to a 95.8% increase in the estimated odds. This result can also be obtained from the odds ratio e0.096 = 1.101 by calculating 1.1017 = 1.961, with the slight difference in the result due only to rounding in the odds ratio.

**7.5.3 Exercise 3: Voting on Property Tax Referendum**

**1.**

 Holding other independent variables constant, there are statistically significant differences in the estimated odds of supporting the referendum between homeowners and renters, and between city natives and residents who were raised elsewhere. The estimated odds of supporting the referendum for homeowners are e-0.376 = 0.687 times those for renters (or the estimated odds of supporting the referendum are 31.3% lower for homeowners than for renters). The estimated odds of supporting the referendum for city natives are e0.583 = 1.791 times greater, or 79.1% more, than for residents who were raised elsewhere. Both of these differences in the odds are large enough to be meaningful.

 The effect of number of children on supporting the referendum, or more specifically on log odds, odds, or probability of supporting the referendum, is statistically significant, controlling for other independent variables. Holding other independent variables constant, the estimated odds of supporting the referendum are multiplied by e0.474 = 1.606 for each additional child; in other words, estimated odds of supporting the referendum increase by 60.6% for each additional child. This effect also seems quite large in real terms. Education and gender do not have statistically significant relationships with opinions on the referendum, holding other independent variables constant.

**2.**

 Because the current model includes all the variables in the alternative model, plus some more, the current and alternative models are nested, and a chi-square test is appropriate to compare two nested models in a logistic regression setting. Chi-square is calculated as (–2 × log-likelihood0) – (–2 × log-likelihood1) = (–2 × –546.256) – (–2 × –540.637) = 1092.512 – 1081.274 = 11.238. The nested models differ by two independent variables (homeownership and number of children), so df = 2. According to the chi-square table, this corresponds to a p-value < 0.01. (An online calculator gives a more precise p-value of 0.004.) The additional independent variables in the current model significantly improve fit, so that the current model is preferable to the alternative.

**Chapter 8**

**8.4.2 Exercise 2**

**1.**

 **(a)**

If b = 0.19, then the expected number of traffic citations for employed people is estimated to be e0.19 = 1.209 times the expected number of traffic citations for unemployed people. In other words, the expected number of traffic citations is 20.9% more for the employed than for the unemployed.

 **(b)**

If b is –0.15, then the expected number of traffic citations for employed people is estimated to be e-0.15 = 0.861 times the expected number of traffic citations for unemployed people. In other words, the expected number of traffic citations is 13.9% less for the employed than for the unemployed.

**2.**

 **(a)**

For each additional $1,000 increase in annual income, the expected number of traffic citations is estimated to be multiplied by e0.008 = 1.008, or a 0.8% increase.

 **(b)**

For each additional $1,000 increase in annual income, the expected number of traffic citations is estimated to be multiplied by e-0.021 = 0.979, or a 2.1% decrease.

**8.4.3 Exercise 3 Transportation Use Patterns**

**1.**

 First assessing the overall fit of the Poisson regression model, the chi-square value of 91.25, with 64 df, corresponds to a p-value of 0.014. As this p-value is small, it indicates that the model does not adequately fit the sample data. Turning to the various independent variables, both the “Male” dummy and age show p > 0.05, so the gender difference in expected number of train trips and the effect of age are not statistically significant when controlling for the other independent variables. The number of plane rides is statistically significant (p = 0.007). The coefficient shows that the expected of train trips is estimated to be multiplied by e0.163 = 1.177, or increase 17.7%, for each additional plane ride, holding other independent variables constant. This seems large enough to have real-life significance.

 For the car ownership variable, there are two dummies, so statistical significance requires comparison of this model to that in which these dummies have been omitted. Because the information for Model 1 (the full model) and Model 0 (the model without these dummies) reports the deviance, the chi-square statistic will be calculated as the difference in deviance between the two models; df = 2 because there are two dummies in the set being tested. The output shows 108.09 as the deviance for Model 1, so the chi-square statistic is 128.23 – 108.09 = 20.14. From the chi-square table, this corresponds to a p-value < 0.01, so car ownership has a statistically significant impact on the number of train trips. The p-values for the specific dummies show that the difference between those who own an unshared car and those who have no car is statistically significant, while the difference between those who share ownership of a car and those who have none is not statistically significant.

 From the b’s, the expected number of train trips for those who own an unshared car is estimated to be e-0.692 = 0.501 times that for those who own no car. That is, the expected number of train trips is 49.9% less for those who own an unshared car than for those who own none. Also, the expected number of train trips for those who share ownership of a car is estimated to be e0.140 = 1.150 times, or 15.0% greater than, the expected number for those who do not own a car. Note, however, that this difference is not statistically significant (p = 0.496). We can also calculate the estimated difference between those who own an unshared car and those who share ownership. (–0.692 – 0.140) = –0.832, so that the expected number of train trips for those who own an unshared car is estimated to be e-0.832 = 0.435 times, or 56.5% less than, the expected number of train trips for those who share ownership of a car. We could also calculate this as 0.501 / 1.150 = 0.436, with the small difference due to rounding. The estimated differences among the car ownership categories seem large enough to indicate real-world significance of this independent variable.

**2.**

 Because the chi-square statistic and associated p-value indicated that this Poisson regression model did not adequately fit the sample data, it is appropriate to consider additional analyses. One straightforward response would be to identify additional independent variables that are theoretically meaningful and also available in the data being used by the analyst, as the inclusion of these might improve fit. Another possible step would be to analyze the sample data via negative binomial regression instead of Poisson regression. The lack of fit of the Poisson regression model suggests the presence of overdispersion, though also note the practical concern that the sample size is not very large for a negative binomial regression analysis. In principle, zero-inflated models would be another alternative to consider, but this approach may be less attractive here. That is because there does not seem to an obvious process at work that would generate excess zeros in these data; that is, there is not a clear analogue to the fishing story described in the text.