Chapter 17 Regression Models with Dummy Variables

Solutions

1. 1. The coefficient of implies that if the dummy variable , will be ―3.8 units less than if holding the same. In other words, the intercept shifts down by 3.8 units when .



* 1. For



* 1. For (Note that this is 3.8 units greater than when .











1. 1. With , With ,



* 1. The coefficient of ―13.22 for implies that when , is 13.22 units less than when , holding everything else constant. The coefficient of 5.35 for



implies that when is 5.35 units greater than when holding everything else constant.



Since the *p-*value for the coefficient of is 0.40, we cannot conclude that is statistically significant. However, with a *p-*value of zero for , we can conclude at the 5% significance level that is significant.



1. 1. Given that the response variable is transformed into logs, the estimated coefficient of 0.15 implies that when is approximately 15% greater than when holding constant (see Chapter 16 for the interpretation of the coefficients of an exponential model) .



* 1. For



For Note that the actual increase is 16% whereas in part a we had approximated it as 15%. Recall from Chapter16, the coefficient interpretation for the log models is always an approximation.



* 1. Since the *p*-value for is 0.0008, which is less than 0.05, we can conclude that is significant at the 5% level.



1. 1. If there are four dummy variables for four categories, it would cause perfect multicollinearity because . The model would not be able to be estimated. We must use one less dummy variable than the number of categories.



* 1. Since the *p-*values for all three variables are above 0.05, we cannot conclude that any of the dummy variables is individually significant at the 5% level. Therefore, the analyst cannot obtain a higher return in any particular quarter.

1. 1. Make two columns, one for players with nicknames and one for players without; a portion of the data used is shown below.

|  |  |
| --- | --- |
| Years with Nickname | Years with No Nickname |
| 74 | 62 |
| 62 | 56 |
| .  .  . | .  .  . |
| 68 | 64 |

The average lifespan for players with nicknames is 68.05, or approximately 68 years. Without nicknames, the average lifespan is 64.08, or approximately 64 years. The difference in lifespan for players with and without nicknames is 3.97 (= 68.05 ― 64.08)

* 1. Relevant regression results for :



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 64.08 | 1.80 | 35.55 | 0.0000 |
| Nicknames | 3.97 | 2.33 | 1.71 | 0.0989 |

With a nickname, and .



Without a nickname, . Note that with a single regression with a dummy variable, we get the same results as above.



* 1. The hypotheses are:



For this one-tailed test, we must divide the reported *p*-value in half, so the *p*-value = 0.0989/2 Since 0.0494 < 0.05, we reject at the 5% significance level and conclude that players with a nickname do live longer than players without a nickname.



1. 1. Relevant regression results for :



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | | *t Stat* | *P-value* |
| Intercept | 316.41 | 32.88 | 9.62 | | 0.0000 |
| GPA | 81.66 | 10.90 | 7.49 | | 0.0000 |
| Female | 16.59 | 9.17 | 1.81 | | 0.0882 |

* 1. For a male student with a GPA of 3.5, we set .



For a female student with a GPA of 3.5, we set



* 1. Using the hypotheses H0: HA: Since the *p*-value for the dummy variable (Female) is 0.0882, which is greater than α = 0.05, we do not reject and therefore we cannot conclude that there is a statistically significant gender difference in writing scores at a 5% level.







Relevant regression results for :



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 274.12 | 37.85 | 7.24 | 0.0000 |
| GPA | 98.71 | 12.54 | 7.87 | 0.0000 |
| Female | ‒21.10 | 10.55 | ‒2.00 | 0.0618 |

* 1. For a male student with a GPA of 3.5, we set



For a female student with a GPA of 3.5, we set



* 1. Using the hypotheses : *H*A: Since the *p*-value for the dummy variable (Female) is 0.0618, which is greater than α = 0.05, we do not reject and therefore we cannot conclude that there is a statistically significant gender difference in math scores at the 5% level.



1. 1. Relevant regression results for :



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | ‒74.6949 | 75.1623 | ‒0.9938 | 0.3292 |
| Temperature | 6.9624 | 0.9852 | 7.0672 | 0.0000 |
| Weekend | 201.8710 | 15.5131 | 13.0129 | 0.0000 |

* 1. For a Sunday with a temperature of 80 degrees, set The predicted number of customers is , or approximately 684 customers.



* 1. The coefficient of 201.87 for weekend implies that on a weekend day, the predicted number of customers is approximately 202 customers more than on a week day, holding temperature the same. With a *p*-value of approximately zero, we can conclude that this variable is significant at the 5% significance level, and therefore

the manager may want to plan on staffing the store with more employees on the weekends.

1. 1. Relevant regression results for :



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 8.6848 | 4.0926 | 2.1220 | 0.0394 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| EDUC | 1.2327 | 0.3673 | 3.3559 | 0.0016 |
| EXPER | 0.4166 | 0.1425 | 2.9244 | 0.0054 |
| AGE | ‒0.0190 | 0.0828 | ‒0.2294 | 0.8196 |
| Gender | 2.2897 | 1.6762 | 1.3660 | 0.1787 |

* 1. For a 40-year-old male with 10 years of education and 5 years of experience, set and calculate /hour.



For a female with the same qualifications, set and calculate /hour.



* 1. The gender coefficient of 2.29 implies that males earn $2.29 per hour more than females at the firm, holding everything else the same. Given a *p*-value of 0.1787, which is greater than we cannot conclude at the 5% level that the gender



variable is significant and, therefore, we cannot conclude that gender discrimination exists at the firm.

* 1. Relevant regression results for :



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 263995.07 | 37720.51 | 7.00 | 0.0000 |
| GNP (in billions) | 53.13 | 2.92 | 18.23 | 0.0000 |
| d1 | ‒98646.08 | 13631.46 | ‒7.24 | 0.0000 |
| d2 | ‒26350.62 | 13605.46 | ‒1.94 | 0.0609 |
| d3 | ‒38084.93 | 13592.18 | ‒2.80 | 0.0082 |

* 1. For quarter 2 and GNP of $13,000 billion, set . The predicted retail sales for quarter 2 is million.



For quarter 4 and GNP of $13,000 billion, set . The predicted retail sales for quarter 2 is million.



* 1. Since the *p*-values corresponding to and are less than 0.05, we conclude, at the 5% level, that the quarterly sales for quarter 1 and quarter 3 are different from those of the 4th quarter. However, with a *p*-value of 0.06, we cannot conclude that quarter 2 sales are significantly different from quarter 4 sales.



* 1. The competing hypotheses for the partial *F* test are:



The unrestricted model is the complete model above, and the restricted model is

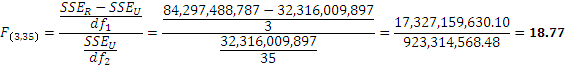
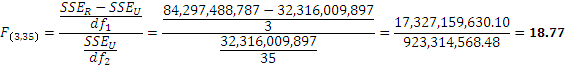


SSER = 84,297,488,787, SSEU = 32,316,009,897 (both of these are reported in the regression results from Excel, in the ANOVA table).

because there are 3 restrictions,



.



With α = 0.05, , . Therefore, reject if > 2.87. Since 18.77 > 2.87, we reject . At the 5% significance level, we can conclude that the three seasonal variables are jointly significant.



1. 1. Relevant regression results for :



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | ‒1.4391 | 0.2865 | ‒5.0239 | 0.0000 |
| ln(Assets) | 0.4042 | 0.0310 | 13.0575 | 0.0000 |
| d1 | ‒0.0822 | 0.1533 | ‒0.5365 | 0.5919 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| d2 | ‒0.2915 | 0.1456 | ‒2.0015 | 0.0459 |
| d3 | ‒0.8253 | 0.1635 | ‒5.0486 | 0.0000 |

* 1. The coefficient 0.40 of ln(Assets) suggests that in any industry, for a 1% change in assets, the predicted increases by about 0.40%.



The coefficient of suggests that that compensation in Manufacturing Technology is predicted to be about 8% (=0.08 lower than that of Nonfinancial Services, holding total assets the same.



The coefficient of suggests that compensation in Manufacturing Other is predicted to be about 29% (= 0.29 lower than that of Nonfinancial Services, holding total assets the same.



The coefficient of suggests that compensation for Financial Services is predicted to be about 83% (= 0.83 lower than that of Nonfinancial Services, holding total assets the same.



* 1. Since the *p*-values corresponding to and are less than 0.05, we conclude at the 5% level that executive compensation in Manufacturing Other firms and Financial Services firms are different from Nonfinancial Services firms. However, with a *p*-value of 0.59, we cannot conclude that executive compensation for Manufacturing Technology firms is different from Nonfinancial Services firms, at the 5% level.



* 1. The competing hypotheses for the partial *F* test are:

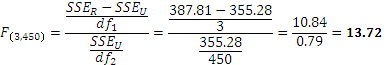
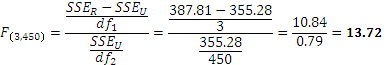


The unrestricted model is the complete model above, and the restricted model is



SSER =387.81, SSEU = 355.28 (both of these are reported in the regression results from Excel, in the ANOVA table).

because there are 3 restrictions,



With α = 0.05, , . Therefore, reject if > 2.62.



Since 13.72 > 2.62, we reject . At the 5% significance level, we can conclude that the three industry dummy variables are jointly significant.







1. 1. For



For



* 1. With a *p*-value of 0.32, the dummy variable is not significant at the 5% level. However, the interaction variable *x* has a *p*-value of 0.02 and therefore is individually significant at the 5% level.



* 1. For a house with ocean views and square footage of 2000, let (thousand), or $380,000.



The corresponding price with 3000 square feet is predicted as (thousand),or $510,000.



* 1. For a house without ocean views and square footage of 2000, let (thousand), or $320,000.



The corresponding price with 3000 square feet is predicted as (thousand),or $440,000.



* 1. Ocean views cause housing prices to start at a higher price with the same square footage (the coefficient of the dummy variable is positive). Moreover, as square footage increases, the price difference between houses with and without ocean views becomes larger as well (the coefficient of the interaction variable is also positive).



Model 1:



Model 2:



Model 3:



Relevant regression results:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| Intercept | 8160.34  (0.1551) | 13007.26\*  (0.0385) | ‒1676.58  (0.8482) |
| Income | 0.55\*  (0.0000) | 0.44\*  (0.0000) | 0.66\*  (0.0000) |
| Urban dummy | NA | 6544.43  (0.0717) | 36361.71\*  (0.0010) |
|  | NA | NA | ‒0.38\*  (0.0274) |
|  | | | |
| Adjusted R2 | 0.5806 | 0.6006 | 0.6332 |
| Notes: The top portion of the table contains parameter estimates with *p*-values in parentheses; \* represents significance at the 5 percent level; Adjusted R2, reported in the last row, is used for model selection. | | | |

1. The linear model, Model 1, is estimated as .



For a family with income of $75,000, the predicted consumption is .



1. With a dummy, Model 2 is estimated as



For a family with income of $75,000 in an urban community, let The predicted consumption is .



For a comparable family in a rural community, let The predicted consumption is .



1. With a dummy and an interaction variable, Model 3 is estimated as . For a family with income of $75,000 in an urban community, let The predicted consumption is . For a comparable family in a rural community, let The predicted consumption is



1. Since each model has a different number of explanatory variables, we use Adjusted *R*2 to compare the models. Model 3 has the highest value of 0.6332 and is therefore the most suitable model.
2. Model 1:



Model 2:



Relevant regression results:

|  |  |  |
| --- | --- | --- |
|  | Model 1 | Model 2 |
| Intercept | 62.34\*  (0.0000) | 56.35\*  (0.0000) |
| BMI | ‒0.96\*  (0.0000) | ‒0.74\*  (0.0002) |
| White dummy | 4.48\*  (0.0036) | 16.69\*  (0.0166) |
|  | NA | ‒0.46  (0.0667) |
|  | | |
| Adjusted R2 | 0.7082 | 0.7344 |
| Notes: The top portion of the table contains parameter estimates with *p*-values in parentheses; \* represents significance at the 5 percent level; Adjusted *R*2, reported in the last row, is used for model selection. | | |

* 1. The estimated model is



For a white college-educated worker with a BMI of 30, set and The predicted salary is



The predicted salary for a corresponding non-white worker (, is .



* 1. The estimated model is



For a white college-educated worker with a BMI of 30, set and The predicted salary is



The predicted salary for a corresponding non-white worker (, is .



Model 1:



Model 2:



Relevant regression results:

|  |  |  |
| --- | --- | --- |
|  | Model 1 | Model 2 |
| Intercept | 8.42\*  (0.0000) | 8.34\*  (0.0000) |
| Price Revision | 0.22\*  (0.0000) | 0.20\*  (0.0000) |
| High-Tech dummy | 3.68\*  (0.0043) | 4.01\*  (0.0033) |
|  | NA | 0.05  (0.4498) |
|  | | |
| Adjusted *R*2 | 0.1585 | 0.1571 |
| Notes: The top portion of the table contains parameter estimates with *p*-values in parentheses; \* represents significance at the 5 percent level; Adjusted R2, reported in the last row, is used for model selection. | | |

* 1. The estimated model is .



* 1. The estimated model is



* 1. Using adjusted *R*2, Model 1 is better with the higher value of 0.1585. Notice also that the interaction variable is insignificant with a *p*-value of 0.45, which further shows that Model 1 is better.

We define *d* = 0 for all observations prior to August, 2008 and d = 1 since August, 2008.

Model 1:



Model 2:



Relevant regression results:

|  |  |  |
| --- | --- | --- |
|  | Model 1 (Restricted) | Model 2  (Unrestricted) |
| Intercept | ‒112.09\*  (0.00) | ‒93.36\*  (0.00) |
| ln(Income) | 12.20\*  (0.00) | 10.17\*  (0.00) |
| Savings dummy | NA | 66.24\*  (0.00) |
|  | NA | ‒7.07\*  (0.0005) |
|  | | |
| Adjusted R2 | 0.7373 | 0.8619 |
| SSE | 2.6916 | 1.3522 |
| Notes: The top portion of the table contains parameter estimates with *p*-values in parentheses; \* represents significance at the 5 percent level; Adjusted R2, reported in the last row, is used for model selection. | | |

* 1. The estimated model is



For this log-log model, for a one percent increase in Income, savings is predicted to increase by 12.2%.

* 1. The estimated model is .



Prior to August 2008 (*d*=0), for a one percent increase in income, savings is predicted to increase by 10.17%. Starting in August 2009 (*d*=1), for a one percent increase in income, savings is predicted to increase by 10.17 ‒ 7.07 = 3.10%.

It seems that the savings rate has decreased since the financial crisis started in August, 2008.

* 1. The competing hypotheses for the test are:

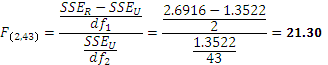
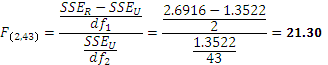


The unrestricted model is Model 2 and the restricted model is Model 1.

because there are 2 restrictions,



.



With α = 0.05, . Therefore, we reject if > 3.21. Since 21.30 > 3.21, we reject . At the 5% significance level, we conclude that and are jointly significant. In other words, the financial crisis has caused a structural shift in the savings model.



1. 1. With , or 42%.



With , or 12%.



* 1. The estimated probability is negative if , which is equivalent to  Since the explanatory variable cannot exceed 50, the final answer is



1. 1. For the LPM model, to test the significance of the intercept coefficient, the competing hypotheses are specified as : *β*0 = 0, *H*A: *β*0 ≠ 0. Since the *p*-value = 0.04 < 0.05 = α, we reject *H0* and conclude that the intercept coefficient is significant at the 5% level. For the slope coefficient, the *p*-value = 0.06 > 0.05 = α. Therefore, we do not reject and cannot conclude that the slope coefficient is significant at the 5% level.



For the logit model, since both *p*-values (0.04 and 0.02) are less than α = 0.05, we can conclude that both the intercept and the slope coefficients are significant at the 5% level.

* 1. The LPM model is estimated as With . With .



* 1. The logit model is estimated as



With .



With



* 1. The LPM model is estimated as .



When



The logit model is estimated as .



When



The predictions for and 5 are made similarly and the results are shown below for both models:



|  |  |  |
| --- | --- | --- |
|  | LPM | Logit |
| 1 | ‒0.08 | 0.05 |
| 2 | 0.24 | 0.19 |
| 3 | 0.56 | 0.53 |
| 4 | 0.88 | 0.84 |
| 5 | 1.2 | 0.96 |

* 1. The LPM model is not always appropriate as it can give negative probabilities and probabilities over 1 for small values and large values of The logit model always gives predicted probabilities between 0 and 1.



1. 1. The LPM model has *p*-values all less than 0.05, so we can conclude that all of the variables are individually significant at the 5% level.

The *p*-value for in the logit model is 0.06, so we cannot conclude that is significant at the 5% level. We can, however, conclude significance of at the 10% level. The variable, with the *p*-value of 0.01, is significant at the 5% level.



* 1. The LPM is estimated as When . When



* 1. The Logit model is estimated as When Similarly, when



1. 1. The LPM is estimated as When .



* 1. The competing hypotheses are specified as : *β*1 = 0, *H*A: *β*1 ≠ 0. Since the *p*-value = 0.0125 is less than α = 0.05, we reject and conclude that is significant at the 5% level.



1. 1. The logit model is estimated as When



* 1. The competing hypotheses are specified as : β1 = 0, *H*A: β1 ≠ 0. Since the *p*-value = 0.030 is less than α = 0.05, we reject and conclude that is significant at the 5% level.



1. 1. The logit model is estimated as



When .



* 1. With a *p*-value of 0.01, is the only variable that is significant at the 5% level. All of the other variables have *p*-values greater than 0.05 and are therefore not significant at the 5% level.



* 1. Relevant regression results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0.8298 | 0.2728 | 3.0423 | 0.0051 |
| Age | ‒0.0192 | 0.0085 | ‒2.2707 | 0.0311 |

The model is estimated as



* 1. For a 20-year-old customer, .



For a 30-year-old customer,



* 1. The competing hypotheses are specified as : β1 ≥ 0, *H*A: β1 < 0. Since the *p*-value = 0.0311/2 = 0.0155 is less than α = 0.05, we reject and conclude that β1 is less than zero at the 5% level. Therefore, Annabel’s belief is supported by the sample data: Age and the probability of purchasing Under Armour are negatively related.



1. Relevant regression results from Minitab:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coef | SE | Z | *P* |
| Constant | 2.3069 | 1.6962 | 1.36 | 0.1740 |
| Age | ‒0.1215 | 0.0609 | ‒1.99 | 0.0460 |

The model is estimated as.



* 1. For a 20-year-old customer,



For a 30-year-old customer,



* 1. The competing hypotheses are specified as : *β*1 ≥ 0, *H*A: *β*1 < 0. Since the *p*-value = 0.0460/2 = 0.0230 is less than α = 0.05, we reject *H*0 and conclude that *β*1  is less than 0 at the 5% level. Therefore, Annabel’s belief that Under Armour attracts a younger clientele is again supported by the data.



1. 1. Relevant regression results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | ‒0.7947 | 0.4421 | ‒1.7976 | 0.0834 |
| Premium% | 0.0092 | 0.0032 | 2.8864 | 0.0076 |
| Income | 0.0172 | 0.0057 | 3.0207 | 0.0055 |

The estimated LPM is .



* 1. With an income of $60,000 , and with an employee contribution of 50 percent of the premium, . The associated predicted probability of having insurance coverage is



* 1. If the employer were to contribute 75 percent of the premium, the estimated probability of coverage would be .



1. Relevant regression results from Minitab:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coef | SE | Z | *P* |
| Constant | ‒9.6504 | 4.1710 | ‒2.31 | 0.021 |
| Premium% | 0.0654 | 0.0287 | 2.28 | 0.023 |
| Income | 0.1291 | 0.0553 | 2.34 | 0.020 |

The model is estimated as.



For an individual with an income of $60,000 and an employee contribution of 50 percent of the premium, let The predicted probability of having insurance coverage is .



If the employer were to contribute 75 percent of the premium, let . The estimated probability of coverage would be .



* 1. Relevant regression results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0.3786 | 0.2167 | 1.7472 | 0.0924 |
| Age | 0.4074 | 0.1622 | 2.5118 | 0.0186 |
| Income | 0.0009 | 0.0030 | 0.2831 | 0.7793 |
| Children | ‒0.1778 | 0.0709 | ‒2.5061 | 0.0188 |

The estimated LPM is . This model suggests that 25-29 year-olds have a 41% higher probability of divorce than other age groups. Also, an increase in income by $1000 increases the probability of a divorce by about 0.1%, holding age and number of children constant. Also, having children has a negative impact on probability of divorce, by a factor of 18% per child, holding income and age constant.



* 1. The competing hypotheses are specified as : *β*1 ≤ 0, *H*A: *β*1 > 0. For this upper-tailed test, the *p*-value equals 0.0186/2 = 0.0093, and is therefore individually significant at the 5% level. We conclude that the divorce rate is higher for this age group.



* 1. For an individual who is 27 years old, has $60,000 in income and has one child, let The probability of divorce for an individual with these characteristics is .



* 1. The corresponding probability with three children is .



1. Relevant regression results from Minitab:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coef | SE | Z | *P* |
| Constant | ‒0.8103 | 1.1594 | ‒0.70 | 0.485 |
| Age | 2.5615 | 1.1783 | 2.17 | 0.030 |
| Income | 0.0094 | 0.0184 | 0.51 | 0.608 |
| Children | ‒1.2436 | 0.5964 | ‒2.09 | 0.037 |

The model is estimated as



* 1. The competing hypotheses are specified as : *β*1 ≤ 0, *H*A: *β*1 > 0. For this upper-tailed test, the *p*-value equals 0.030/2 = 0.015, and is therefore individually significant at the 5% level. We conclude that the divorce rate is higher for this age group.



* 1. For an individual who is 27 years old, has $60,000 in income and has one child, let The probability of divorce for an individual with these characteristics is .



The corresponding probability with three children is .



1. 1. Relevant regression results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 1.606 | 3.17 | 0.51 | 0.6156 |
| *d*1 | ‒3.642 | 4.48 | ‒0.81 | 0.4221 |
| *d*2 | 1.424 | 4.48 | 0.32 | 0.7527 |
| *d*3 | ‒4.454 | 4.48 | ‒0.99 | 0.3273 |

The estimated equation is



* 1. The coefficient ‒3.642 for indicates that returns are lower in quarter 1 by 3.642% compared to quarter 4.



The coefficient 1.434 for indicates that returns are higher in quarter 2 by 1.424% compared to quarter 4.



The coefficient ‒4.454 for indicates that returns are lower in quarter 3 by 4.454% compared to quarter 4.



However, none of the dummy variables is significant at the 5% level.

* 1. For quarter 2, set Calculate



For quarter 4, set Calculate .



* 1. Relevant regression results for :



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 25.1466 | 1.84 | 13.70 | 0.0000 |
| Age | ‒0.3196 | 0.04 | ‒7.31 | 0.0000 |
| Caucasian | 9.4504 | 1.01 | 9.36 | 0.0000 |

* 1. For a 30-year-old applicant with a Caucasian name, set . The predicted call-back rate is %.



The corresponding call-back rate for a non-Caucasian name ( is .



* 1. To test for race discrimination, the hypotheses would be : , *H*A: . With a *p-*value of approximately zero, we reject and conclude that the Caucasian dummy variable is significant at the 5% level. Therefore, the data suggest that there is race discrimination.



1. 1. The intercept dummy has an estimated coefficient of 112605.8, which implies that 4th quarter sales are greater than other quarters by $112,605.80 million, holding everything else the same. However, with a *p*-value of 0.34, the dummy variable is not significant at the 5% level.
   2. The interaction variable has an estimated coefficient of ‒4.7, which implies that the increase in 4th quarter sales is $4.7 million lower than other quarters resulting from a one $billion increase in GNP, holding everything else the same. However, with a *p*-value of 0.62, the interaction variable is not significant at the 5% level.



1. 1. Relevant regression results for :



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 28.26 | 0.29 | 98.26 | 0.0000 |
| Female | ‒3.45 | 0.61 | ‒5.67 | 0.0000 |
| Black | ‒1.31 | 0.43 | ‒3.03 | 0.0030 |
| Female×Black | 6.66 | 0.91 | 7.30 | 0.0000 |

The predicted BMI for white males, set is Calculate



The predicted BMI for white females, set . Calculate



The predicted BMI for black males, set . Calculate



The predicted BMI for black females, set . Calculate



* 1. To test for a difference between white females and white males, we look at With a *p*-value of approximately zero, we can conclude that the female dummy is significant and therefore there is a difference between white females and white males at the 5% level, holding all other variables constant.



* 1. To test for a difference between white males and black males, we look at With a *p*-value of 0.003, we can conclude that the black dummy is significant and therefore there is a difference between white males and black males at the 5% level, holding all other variables constant.



1. 1. Relevant regression results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 17.65 | 1.45 | 12.19 | 0.00 |
| Income | 0.02 | 0.02 | 0.94 | 0.35 |
| Woman | 3.20 | 0.74 | 4.35 | 0.00 |
| Drinks | ‒0.69 | 0.26 | ‒2.62 | 0.01 |

The estimated model is .



* 1. In order to determine if woman live longer than men, we specify the competing hypotheses as:



For a one-tailed test, we must divide the reported *p*-value in half, so the *p*-value = 0.00/2=0.00. Since the *p*-value is less than we reject and conclude that women live longer than men at the 1% significance level.



* 1. For a man with an income of $40,000 and a consumption of 2 drinks per day, The predicted life expectancy is years. The corresponding prediction for a woman is years.



1. 1. Relevant regression results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *p-value* |
| Intercept | ‒2.2461 | 0.6456 | ‒3.4792 | 0.0017 |
| SAT | 0.0010 | 0.0002 | 4.3546 | 0.0002 |
| GPA | 0.3780 | 0.1285 | 2.9428 | 0.0066 |

The estimated model is . Since both SAT and GPA have *p*-values less than 0.05, we can conclude that the variables are individually significant at the 5% level.



* 1. The predicted probability of admission for an individual with a GPA of 3.5 and an SAT score of 1700, .



* 1. With an SAT score of 1800, the predicted probability is .



1. Relevant regression results from Minitab:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coef | SE | Z | *P* |
| Constant | ‒19.4786 | 8.2917 | ‒2.35 | 0.019 |
| SAT | 0.0065 | 0.0025 | 2.58 | 0.010 |
| GPA | 2.7800 | 1.4081 | 1.97 | 0.048 |

The model is estimated as



1. Since SAT and GPA have *p*-values less than 0.05, we can conclude that the variables are individually significant at the 5% level.
2. To find the predicted probability of admission for an individual with a GPA of 3.5 and an SAT score of 1700, let The estimated probability is



1. With an SAT score of 1800, the estimated probability is .



* 1. Relevant regression results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 1.3863 | 0.3691 | 3.7557 | 0.0016 |
| Age | ‒0.0360 | 0.0099 | ‒3.6363 | 0.0020 |
| Gender | 0.2511 | 0.1734 | 1.4479 | 0.1658 |

* 1. Age has a *p*-value of 0.0020, and is therefore significant in explaining the probability of returning to crime. However, Gender has a *p*-value of 0.1658 and is therefore not significant at even the 10% level. Therefore, the claim that women are less likely to re-offend is not supported.
  2. For a 25 year-old male, let . The predicted probability of re-offending iThe corresponding prediction for a female is



1. Relevant regression results for Minitab:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coef | SE | Z | *P* |
| Constant | 9.0293 | 5.6389 | 1.60 | 0.109 |
| Age | ‒0.3454 | 0.1890 | ‒1.83 | 0.068 |
| Gender | 1.2729 | 1.4816 | 0.86 | 0.390 |

The model is estimated as



* 1. Age and Gender have *p*-values of 0.068 and 0.39, respectively. Neither of them is significant in explaining the probability of returning to crime at the 5% level. Therefore, the claims that women and older parolees are less likely to re-offend are not supported.
  2. For a 25 year-old male, let . The predicted probability of re-offending is . The corresponding prediction for a female is .



**Case Study 17.1**

1. Relevant regression results for :



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 4 | 4.2501 | 1.0625 | 3.6436 | 0.0118 |
| Residual | 45 | 13.1228 | 0.2916 |  |  |
| Total | 49 | 17.3730 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 4.2970 | 0.1708 | 25.1627 | 0.0000 |
| Smoking | 0.1710 | 0.2415 | 0.7081 | 0.4826 |
| Tourist | ‒0.5220 | 0.2415 | ‒2.1615 | 0.0360 |
| Elderly | ‒0.5120 | 0.2415 | ‒2.1201 | 0.0395 |
| Obese | ‒0.4590 | 0.2415 | ‒1.9006 | 0.0638 |

1. The variable tourist has a coefficient of ‒0.52. This implies that tourists have a speed of 0.52 feet per second less than non-tourists. The intercept of 4.30 is the estimated speed of a person who is not distracted or exhibiting other traits, or in other words, is not smoking, is not a tourist, and is not elderly or obese.

The predicted speed of an elderly pedestrian is feet per second



The predicted speed of an obese pedestrian is feet per second



1. We conduct an *F* test to determine the joint significance of the explanatory variables. With a *p*-value of 0.0118 (as reported in the ANOVA table above), we conclude that the explanatory variables are jointly significant at the 5% level.

For individual significance, we find that the tourist and elderly variables both have *p*-values less than 0.05 and are therefore individually significant. The smoking and obese variables, however, have *p*-values above 0.05 and are therefore not individually significant at the 5% level. Moreover, since the tourists and the elderly variables have negative effect on speed, a ‘sidewalk rager’ should avoid tourists and the elderly.

**Case Study 17.2**

1. Model 1*:*



Model 2:



Model 3:



Relevant regression results:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| Intercept | 165,888.66\*  (0.0034) | 190,265.22\*  (0.0006) | 240,311.22\*  (0.0001) |
| Sqft | 91.68\*  (0.0080) | 87.27\*  (0.0073) | 79.45\*  (0.0092) |
| Beds | 4372.36  (0.8153) | 4337.70  (0.8045) | 1801.40  (0.9125) |
| Baths | 66,619.61\*  (0.0111) | 53,008.37\*  (0.0348) | 36,897.43  (0.1258) |
| Colonial | 74,557.88\*  (0.0105) | NA | ‒225,683.25\*  (0.0248) |
|  | NA | 45.41\*  (0.0012) | 153.66\*  (0.0029) |
|  | | | |
| Adjusted R2 | 0.7483 | 0.7790 | 0.8075 |
| Notes: The top portion of the table contains parameter estimates with *p*-values in parentheses; NA denotes not applicable; \* represents significance at the 5 percent level; Adjusted *R*2, reported in the last row, is used for model selection. | | | |

1. Model 3 appears to be the most reliable as the Adjusted *R*2 is the highest. Both the dummy and the interaction variables are individually significant at the 5% level since they have *p*-values less than 0.05. We can conclude that there are price differences between colonial homes and other styles at the 5% level that are both fixed and changing.



1. Model 3 is estimated as Using averages for Square feet, bedrooms and bathrooms, we set . The predicted price for a colonial home with these features is . The corresponding estimate for a non-colonial style is



**Case Study 17.3**

1. Relevant regression results for the LPM:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | ‒1.1361 | 0.7101 | ‒1.5999 | 0.1213 |
| GPA | 0.4286 | 0.2130 | 2.0121 | 0.0543 |
| Experience | 0.0339 | 0.0179 | 1.8937 | 0.0690 |

The LPM model is estimated as



The probability of success for someone with a GPA of 3.0 and 2 years of experience if predicted as 75.



Similarly, the estimated probability of success for various values of GPA and experience can be calculated. The following table shows a few combinations and the associated probabilities:

|  |  |  |
| --- | --- | --- |
| GPA | Experience | Predicted Probability |
| 3 | 2 | 0.2175 |
| 3 | 5 | 0.3192 |
| 3 | 10 | 0.4887 |
| 3.5 | 2 | 0.4318 |
| 3.5 | 5 | 0.5335 |
| 3.5 | 10 | 0.7030 |
| 3.9 | 2 | 0.6032 |
| 3.9 | 5 | 0.7049 |
| 3.9 | 10 | 0.8744 |

1. Relevant regression results for the logit model from Minitab:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coef | SE | Z | *P* |
| Constant | ‒8.9244 | 4.4143 | ‒2.02 | 0.043 |
| GPA | 2.3481 | 1.2576 | 1.87 | 0.062 |
| Experience | 0.1826 | 0.0999 | 1.83 | 0.068 |

The logit model is estimated as



The probability of success for someone with a GPA of 3.0 and 2 years of experience if predicted as



Similarly, the estimated probability of success for various values of GPA and experience can be calculated. The following table shows a few combinations and the associated probabilities:

|  |  |  |
| --- | --- | --- |
| GPA | Experience | Predicted Probability |
| 3 | 2 | 0.1802 |
| 3 | 5 | 0.2755 |
| 3 | 10 | 0.4865 |
| 3.5 | 2 | 0.4156 |
| 3.5 | 5 | 0.5516 |
| 3.5 | 10 | 0.7540 |
| 3.9 | 2 | 0.6453 |
| 3.9 | 5 | 0.7588 |
| 3.9 | 10 | 0.8869 |

1. Although both models appear to give similar results in this application, the logit model is probably more reliable. For very low or high levels of GPA or Experience, the LPM may give a probability outside of the 0 to 1 range, whereas the Logit model will stay within those parameters, making it a more reliable model.