MGT6203 - HW2 Solutions

Sowmya Tata

9/9/2019

## Questions 1 to 3 - Weeks 1 & 2 Content

1. C Explanation - “R squared” individually can’t tell whether a variable is significant or not because each time when we add a feature, “R squared” can either increase or stay constant. But, it is not true in case of “Adjusted R squared” (increases when features found to be significant).
2. D Explanation - Correlation coefficient range is between [-1 ,1]. So -1.09 is not possible.
3. B Explanation - Since Linear Regression models and predicts continuous variables, the predicted probability may lie outside the prescribed values for a probability (that is, outside the range of 0 and 1). This is where a Logistic Regression helps.

## Questions 4 and 5 - Week 3 Content

nba <- read.csv("nba2017.csv", header=TRUE)  
model\_1 <- lm(Salary ~ Ht+Exp, data = nba)  
summary(model\_1)

##   
## Call:  
## lm(formula = Salary ~ Ht + Exp, data = nba)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11584274 -2907609 -1371039 1880620 19884672   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12353214 5785552 -2.135 0.0332 \*   
## Ht 2253985 874758 2.577 0.0103 \*   
## Exp 677390 61100 11.087 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5728000 on 496 degrees of freedom  
## Multiple R-squared: 0.2042, Adjusted R-squared: 0.201   
## F-statistic: 63.65 on 2 and 496 DF, p-value: < 2.2e-16

model\_2 <- lm(log(Salary) ~ Ht+Exp, data = nba)  
summary(model\_2)

##   
## Call:  
## lm(formula = log(Salary) ~ Ht + Exp, data = nba)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0100 -0.6698 0.2475 1.0030 2.6381   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.39419 1.39946 6.713 5.24e-11 \*\*\*  
## Ht 0.68885 0.21159 3.256 0.00121 \*\*   
## Exp 0.16738 0.01478 11.325 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.385 on 496 degrees of freedom  
## Multiple R-squared: 0.2151, Adjusted R-squared: 0.2119   
## F-statistic: 67.96 on 2 and 496 DF, p-value: < 2.2e-16

1. A Explanation - From the output summaries above, in model\_1 the coefficient of Ht is 2253985. This means that a one unit increase in height increases salary by 2253985 units.
2. D Explanation - In model\_2, the coefficient of Ht is 0.68885. This means that the a one unit increase in height, increases the ln(salary) by 0.68. This can be interpreted as one unit increase in height roughly increases the salary by 68.85%.

## Questions 6 to 9 - Week 4 Content

1. B (False) Explanation – As dependent variable (whether a person defaults or not) is binary, therefore, a classification model like logistic regression should be used.
2. B (1:5) Explanation – Probability of winning:16.67%. Probability of not winning:83.33%. Odds for = 16.67%/83.33% = 1:5.
3. A (230/(230+54)) Explanation – Specificity measures the proportion of actual negatives that are correctly identified as such. Specificity = true negative / (true negative + false positive)
4. C (Lower the cutoff value p) Explanation – If we decrease the cutoff threshold, false negatives will decrease. False negatives in medical applciations are bad because we don’t want to diagnose a patient as not having a particular disease when they actually do; they wouldn’t get the necessary treatment.

## Questions 10 to 14 - Week 4 Content

#install.packages("epitools")  
library(epitools)  
data(wcgs)  
  
logReg <- glm(chd69 ~ ncigs0 + age0, data=wcgs, family = binomial(link="logit"))  
summary(logReg)

##   
## Call:  
## glm(formula = chd69 ~ ncigs0 + age0, family = binomial(link = "logit"),   
## data = wcgs)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1662 -0.4505 -0.3615 -0.2906 2.6137   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.360059 0.562943 -11.298 < 2e-16 \*\*\*  
## ncigs0 0.023872 0.004085 5.844 5.09e-09 \*\*\*  
## age0 0.076353 0.011448 6.670 2.56e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1781.2 on 3153 degrees of freedom  
## Residual deviance: 1705.9 on 3151 degrees of freedom  
## AIC: 1711.9  
##   
## Number of Fisher Scoring iterations: 5

1. A; 𝑝(chd69) = exp(-6.36 + 0.02*ncigs0 + 0.08*age0)/[1 + exp(-6.36 + 0.02*ncigs0 + 0.08*age0)] Explanation – The output of logistic regression can be seen above. p is given by the logistic function p = e^(z)/ (1 + e^(z)) where z is the linear expression of variables and coefficients (b0 + b1*x1 + b2*x2)
2. D (All of the above statements.) Explanation - One unit increase in age corresponds to 0.076353 (approx 0.08) units increase in natural log odds of getting heart disease. This is the same as saying odds of getting heart disease increases by exp(0.08) and for small changes, exp(0.08) can be approximated to 100\*0.08%. Therefore, all statements are correct.
3. A (If ncigs0 increases by 1 unit, the natural log of the odds of getting heart disease increases by 0.02) Explanation - This is direct interpretation of the coefficient of ncigs0 = 0.0238.
4. B (exp(-6.36 + 0.02*10 + 0.08*35)/[1 + exp(-6.36 + 0.02*10 + 0.08*35)]) Explanation - p = e^(z)/ (1 + e^(z)) where z is the linear expression of variables and coefficients (b0 + b1*x1 + b2*x2). In this case, x1=10 (number of cigarettes a day) and x2=35 (person’s age)
5. C (0.02) Explanation - Logistic regression has the form log(p/(1-p)) = b0 + b1*x1 + b2*x2. The left-hand side is the natural log of predicted odds of getting coronary heart disease.

**Note: Since many students misinterpreted absolute change, we will award credit for options C and D.**

## Questions 15 and 16 - Week 5 Content

1. D (All of the above) Explanation - these examples are directly taken from week 5 notes, slide 15. In a natural experiment, the subjects who might be undergoing treatment are not able to choose if they are in the treatment or control group. This choice is made by an external agent, or a factor like weather, a policy change, etc.
2. B (Regressing on other independent variables and checking for significant coefficients) Explanation - If there is random assignment, there should not be any significant coefficients, i.e., the p-values for all other independent variables should be larger than 0.05.

## Questions 17 to 20 - Week 5 Content

1. B (D-F) Explanation - For the treatment group (County A), the After values of sales is D and the baseline is F. Therefore, the difference D-F is the number of cigarettes sold after the law was passed.
2. D (E-F) Explanation - For the control group (County B), the After values of sales is E. Similar to above, the difference E-F is the number of cigarettes sold after the law was passed.
3. B (b2 + b3) Explanation - Treatment Before: b0 + b1; Treatment After: b0 + b1 + b2 + b3; Difference (Before - After) = b2 + b3
4. B (C-D) Explanation - To find out what is the treatment effect, we need to add parallels to A-D line. The Diff-in-Diff is (D-A) - (E-B) = (D-E) - (A-B) = (D-E) - (C-E) = D-C. Negative of D-C is C-D, since the sales did not go up as much in the treatment group as in the control group. We can say that the County A law requiring that all cigarette packets contain pictorial images of adverse effects of cigarette smoking worked!