**ENV 503: Statistics for Bioinformatics**

**Homework Set #8**

**Due: November 7, 2018**

*Instructions:*

*Use R to complete this assignment. Assignment is to be submitted via Blackboard.*

The R dataset **Pima.tr** (from the MASS package) contains data on 200 women of Pima Indian heritage living near Phoenix, Arizona. All women were tested for diabetes according to World Health Organization criteria. The dataset contains whether or not they met WHO criteria for diabetes (variable type=Yes for diabetes, No otherwise), their plasma glucose concentration (variable glu), as well as variables npreg (number of pregnancies), bp (diastolic blood pressure), skin (triceps skin fold thickness), bmi (body mass index) and age.

1. Fit a series of bivariate linear regression models having plasma glucose(glu) as the dependent variable and each of these independent variables (npreg, bp, skin, bmi, and age). Report the results in terms of regression coefficients and 95% confidence intervals. Which variables are significant predictors of plasma glucose?
2. npreg<-Pima.tr$npreg
3. > glu<-Pima.tr$glu
4. > fit1<-lm(glu~npreg)
5. > fit1
6. Call:
7. lm(formula = glu ~ npreg)
8. Coefficients:
9. (Intercept) npreg
10. 118.243 1.604
11. > coef(fit1)
12. (Intercept) npreg
13. 118.243138 1.604163
14. > confint(fit1)
15. 2.5 % 97.5 %
16. (Intercept) 111.8763839 124.60989
17. npreg 0.3050962 2.90323
18. > confint(fit1,level=0.90)
19. 5 % 95 %
20. (Intercept) 112.9076846 123.578590
21. npreg 0.5155219 2.692804
22. > fit1<-summary(lm(glu~npreg))
23. > fit1
24. Call:
25. lm(formula = glu ~ npreg)
26. Residuals:
27. Min 1Q Median 3Q Max
28. -75.681 -22.451 -4.962 19.601 79.757
29. Coefficients:
30. Estimate Std. Error t value Pr(>|t|)
31. (Intercept) 118.2431 3.2285 36.624 <2e-16 \*\*\*
32. npreg 1.6042 0.6588 2.435 0.0158 \*
33. ---
34. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
35. Residual standard error: 31.28 on 198 degrees of freedom
36. Multiple R-squared: 0.02908, Adjusted R-squared: 0.02418
37. F-statistic: 5.93 on 1 and 198 DF, p-value: 0.01577
38. bp<-Pima.tr$bp
39. > glu<-Pima.tr$glu
40. > fit2<-lm(glu~bp)
41. > fit2
42. Call:
43. lm(formula = glu ~ bp)
44. Coefficients:
45. (Intercept) bp
46. 71.0163 0.7431
47. > coef(fit2)
48. (Intercept) bp
49. 71.0162894 0.7431057
50. > confint(fit2)
51. 2.5 % 97.5 %
52. (Intercept) 44.1452493 97.887329
53. bp 0.3707974 1.115414
54. > confint(fit2,level=0.90)
55. 5 % 95 %
56. (Intercept) 48.4978790 93.534700
57. bp 0.4311047 1.055107
58. > fit2<-summary(lm(glu~bp))
59. > fit2
60. Call:
61. lm(formula = glu ~ bp)
62. Residuals:
63. Min 1Q Median 3Q Max
64. -73.465 -21.534 -3.534 19.716 84.828
65. Coefficients:
66. Estimate Std. Error t value Pr(>|t|)
67. (Intercept) 71.0163 13.6262 5.212 4.68e-07 \*\*\*
68. bp 0.7431 0.1888 3.936 0.000115 \*\*\*
69. ---
70. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
71. Residual standard error: 30.57 on 198 degrees of freedom
72. Multiple R-squared: 0.07257, Adjusted R-squared: 0.06788
73. F-statistic: 15.49 on 1 and 198 DF, p-value: 0.0001146
74. skin<-Pima.tr$skin
75. > glu<-Pima.tr$glu
76. > fit3<-lm(glu~skin)
77. > fit3
78. Call:
79. lm(formula = glu ~ skin)
80. Coefficients:
81. (Intercept) skin
82. 106.8000 0.5877
83. > coef(fit3)
84. (Intercept) skin
85. 106.8000105 0.5877114
86. > confint(fit3)
87. 2.5 % 97.5 %
88. (Intercept) 95.1737582 118.4262627
89. skin 0.2182599 0.9571629
90. > confint(fit3,level=0.90)
91. 5 % 95 %
92. (Intercept) 97.0570039 116.5430170
93. skin 0.2781045 0.8973184
94. > fit3<-summary(lm(glu~skin))
95. > fit3
96. Call:
97. lm(formula = glu ~ skin)
98. Residuals:
99. Min 1Q Median 3Q Max
100. -71.545 -20.087 -3.554 21.479 76.797
101. Coefficients:
102. Estimate Std. Error t value Pr(>|t|)
103. (Intercept) 106.8000 5.8956 18.115 < 2e-16 \*\*\*
104. skin 0.5877 0.1873 3.137 0.00197 \*\*
105. ---
106. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
107. Residual standard error: 30.99 on 198 degrees of freedom
108. Multiple R-squared: 0.04735, Adjusted R-squared: 0.04254
109. F-statistic: 9.841 on 1 and 198 DF, p-value: 0.001967
110. bmi<-Pima.tr$bmi
111. > glu<-Pima.tr$glu
112. > fit4<-lm(glu~bmi)
113. > fit4
114. Call:
115. lm(formula = glu ~ bmi)
116. Coefficients:
117. (Intercept) bmi
118. 87.79 1.12
119. > coef(fit4)
120. (Intercept) bmi
121. 87.786509 1.119885
122. > confint(fit4)
123. 2.5 % 97.5 %
124. (Intercept) 64.5463923 111.026625
125. bmi 0.4131453 1.826625
126. > confint(fit4,level=0.90)
127. 5 % 95 %
128. (Intercept) 68.3108771 107.262140
129. bmi 0.5276246 1.712146
130. > fit4<-summary(lm(glu~bmi))
131. > fit4
132. Call:
133. lm(formula = glu ~ bmi)
134. Residuals:
135. Min 1Q Median 3Q Max
136. -67.519 -20.292 -5.239 19.483 79.104
137. Coefficients:
138. Estimate Std. Error t value Pr(>|t|)
139. (Intercept) 87.7865 11.7849 7.449 2.84e-12 \*\*\*
140. bmi 1.1199 0.3584 3.125 0.00205 \*\*
141. ---
142. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
143. Residual standard error: 30.99 on 198 degrees of freedom
144. Multiple R-squared: 0.047, Adjusted R-squared: 0.04218
145. F-statistic: 9.764 on 1 and 198 DF, p-value: 0.002046

|  |
| --- |
| age<-Pima.tr$age  > glu<-Pima.tr$glu  > fit5<-lm(glu~age)  > fit5  Call:  lm(formula = glu ~ age)  Coefficients:  (Intercept) age  92.1546 0.9908  > coef(fit5)  (Intercept) age  92.1545835 0.9908258  > confint(fit5)  2.5 % 97.5 %  (Intercept) 79.2709309 105.038236  age 0.6110578 1.370594  > confint(fit5,level=0.90)  5 % 95 %  (Intercept) 81.3578531 102.951314  age 0.6725735 1.309078  > fit5<-summary(lm(glu~age))  > fit5  Call:  lm(formula = glu ~ age)  Residuals:  Min 1Q Median 3Q Max  -76.733 -18.093 -4.806 20.320 85.047  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 92.1546 6.5332 14.106 < 2e-16 \*\*\*  age 0.9908 0.1926 5.145 6.41e-07 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 29.82 on 198 degrees of freedom  Multiple R-squared: 0.1179, Adjusted R-squared: 0.1135  F-statistic: 26.47 on 1 and 198 DF, p-value: 6.415e-07 |
|  |
| |  | | --- | | > | |

1. Fit a multiple linear regression model of predictors of plasma glucose:
   1. In which order should IVs enter the model?

Age,bp,skin,bmi,npreg

* 1. Are any IVs removed? If so, why?

No.

* 1. What is the final model? Report the results in terms of regression coefficients and 95% confidence intervals.

1. summary(lm(glu ~ age + bp + skin + bmi+npreg,Pima.tr))
2. Call:
3. lm(formula = glu ~ age + bp + skin + bmi + npreg, data = Pima.tr)
4. Residuals:
5. Min 1Q Median 3Q Max
6. -80.398 -16.558 -2.406 18.508 87.546
7. Coefficients:
8. Estimate Std. Error t value Pr(>|t|)
9. (Intercept) 48.83674 16.19545 3.015 0.002909 \*\*
10. age 0.86912 0.25333 3.431 0.000735 \*\*\*
11. bp 0.34774 0.20179 1.723 0.086425 .
12. skin 0.07238 0.24281 0.298 0.765959
13. bmi 0.68362 0.45451 1.504 0.134183
14. npreg -0.49202 0.77323 -0.636 0.525318
15. ---
16. Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1
17. Residual standard error: 29.34 on 194 degrees of freedom
18. Multiple R-squared: 0.163, Adjusted R-squared: 0.1414
19. F-statistic: 7.556 on 5 and 194 DF, p-value: 1.664e-06

d. Interpret the meaning of each regression coefficient in your final model. What does the model tell you about predictors of blood glucose?

|  |
| --- |
| model$coefficients  Estimate Std. Error t value Pr(>|t|)  (Intercept) 48.83673947 16.1954454 3.0154613 0.0029092054  age 0.86911940 0.2533301 3.4307785 0.0007353392  bp 0.34773957 0.2017857 1.7233117 0.0864254763  skin 0.07237889 0.2428143 0.2980833 0.7659585552  bmi 0.68362368 0.4545085 1.5040942 0.1341833333  npreg -0.49202122 0.7732301 -0.6363192 0.5253183977 |
|  |
| |  | | --- | | > | |

* The result shows that the coefficient for age is 0.869. The coefficient indicates that for every additional increase of age in year we can expect glucose to increase by an average of 0.869.
* The result shows that the coefficient for bp is 0.3477. The coefficient indicates that for every additional increase of bp in mm hg we can expect glucose to increase by an average of 0.3447.
* The result shows that the coefficient for skin is 0.0723. The coefficient indicates that for every additional increase of skin thickness in mm we can expect glucose to increase by an average of 0.869.
* The result shows that the coefficient for bmi is 0.6836. The coefficient indicates that for every additional increase of bmi in weight in kg/(height in m)*\^2*) we can expect glucose to increase by an average of 0.869.
* The result shows that the coefficient for npreg is -0.492. The coefficient indicates that for every additional increase of pregnancy we can expect glucose to increase by an average of-0.492.

3.Now fit a multiple logistic regression model of predictors of diabetes, following the same steps as for question 2. Refer to your HW7 bivariate logistic regression results to determine which IVs should go in the model and in which order. You will also need to create the binary indicator variable diabetes from the type variable as follows using the code provided for HW7. How do the conclusions from this model differ from those you get from the multiple linear regression model predicting plasma glucose?

|  |
| --- |
| > Pima.tr$diabetes[Pima.tr$type == "Yes"] <- 1  > Pima.tr$diabetes[Pima.tr$type == "No"] <- 0  > fit <- glm(diabetes ~ age+bp+skin+bmi+npreg, Pima.tr, family=binomial(logit))  > summary(fit)  Call:  glm(formula = diabetes ~ age + bp + skin + bmi + npreg, family = binomial(logit),  data = Pima.tr)  Deviance Residuals:  Min 1Q Median 3Q Max  -1.7739 -0.8205 -0.4869 0.9682 2.2461  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -6.7736376 1.4809934 -4.574 4.79e-06 \*\*\*  age 0.0573533 0.0192087 2.986 0.00283 \*\*  bp 0.0063662 0.0158073 0.403 0.68714  skin -0.0006967 0.0189430 -0.037 0.97066  bmi 0.1063965 0.0379535 2.803 0.00506 \*\*  npreg 0.0645135 0.0572186 1.127 0.25953  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 256.41 on 199 degrees of freedom  Residual deviance: 214.46 on 194 degrees of freedom  AIC: 226.46  Number of Fisher Scoring iterations: 4  > exp(coef(fit))  (Intercept) age bp skin bmi npreg  0.001143527 1.059029941 1.006386487 0.999303532 1.112262756 1.066639976  > exp(confint(fit))  Waiting for profiling to be done...  2.5 % 97.5 %  (Intercept) 5.428757e-05 0.01848306  age 1.020609e+00 1.10119930  bp 9.754613e-01 1.03830078  skin 9.634293e-01 1.03923655  bmi 1.033317e+00 1.20028184  npreg 9.535291e-01 1.19529731 |
|  |
| |  | | --- | | > | |

step1<-stepAIC(fit,direction="forward")

Start: AIC=226.46

diabetes ~ age + bp + skin + bmi + npreg

> summary(step1)

Call:

glm(formula = diabetes ~ age + bp + skin + bmi + npreg, family = binomial(logit),

data = Pima.tr)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7739 -0.8205 -0.4869 0.9682 2.2461

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.7736376 1.4809934 -4.574 4.79e-06 \*\*\*

age 0.0573533 0.0192087 2.986 0.00283 \*\*

bp 0.0063662 0.0158073 0.403 0.68714

skin -0.0006967 0.0189430 -0.037 0.97066

bmi 0.1063965 0.0379535 2.803 0.00506 \*\*

npreg 0.0645135 0.0572186 1.127 0.25953

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 256.41 on 199 degrees of freedom

Residual deviance: 214.46 on 194 degrees of freedom

AIC: 226.46

Number of Fisher Scoring iterations: 4

step2<-stepAIC(fit,direction="backward")

Start: AIC=226.46

diabetes ~ age + bp + skin + bmi + npreg

Df Deviance AIC

- skin 1 214.46 224.46

- bp 1 214.62 224.62

- npreg 1 215.74 225.74

<none> 214.46 226.46

- bmi 1 222.50 232.50

- age 1 223.76 233.76

Step: AIC=224.46

diabetes ~ age + bp + bmi + npreg

Df Deviance AIC

- bp 1 214.62 222.62

- npreg 1 215.75 223.75

<none> 214.46 224.46

- age 1 224.05 232.05

- bmi 1 227.79 235.79

Step: AIC=222.62

diabetes ~ age + bmi + npreg

Df Deviance AIC

- npreg 1 215.93 221.93

<none> 214.62 222.62

- age 1 225.82 231.82

- bmi 1 229.02 235.02

Step: AIC=221.93

diabetes ~ age + bmi

Df Deviance AIC

<none> 215.93 221.93

- bmi 1 229.94 233.94

- age 1 239.97 243.97

> summary(step2)

Call:

glm(formula = diabetes ~ age + bmi, family = binomial(logit),

data = Pima.tr)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7935 -0.8368 -0.5033 1.0211 2.2531

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.49870 1.17459 -5.533 3.15e-08 \*\*\*

age 0.07104 0.01538 4.620 3.84e-06 \*\*\*

bmi 0.10519 0.02956 3.558 0.000373 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 256.41 on 199 degrees of freedom

Residual deviance: 215.93 on 197 degrees of freedom

AIC: 221.93

Number of Fisher Scoring iterations: 4

step3<-stepAIC(fit,direction="both")

Start: AIC=226.46

diabetes ~ age + bp + skin + bmi + npreg

Df Deviance AIC

- skin 1 214.46 224.46

- bp 1 214.62 224.62

- npreg 1 215.74 225.74

<none> 214.46 226.46

- bmi 1 222.50 232.50

- age 1 223.76 233.76

Step: AIC=224.46

diabetes ~ age + bp + bmi + npreg

Df Deviance AIC

- bp 1 214.62 222.62

- npreg 1 215.75 223.75

<none> 214.46 224.46

+ skin 1 214.46 226.46

- age 1 224.05 232.05

- bmi 1 227.79 235.79

Step: AIC=222.62

diabetes ~ age + bmi + npreg

Df Deviance AIC

- npreg 1 215.93 221.93

<none> 214.62 222.62

+ bp 1 214.46 224.46

+ skin 1 214.62 224.62

- age 1 225.82 231.82

- bmi 1 229.02 235.02

Step: AIC=221.93

diabetes ~ age + bmi

Df Deviance AIC

<none> 215.93 221.93

+ npreg 1 214.62 222.62

+ bp 1 215.75 223.75

+ skin 1 215.92 223.92

- bmi 1 229.94 233.94

- age 1 239.97 243.97

> summary(step3)

Call:

glm(formula = diabetes ~ age + bmi, family = binomial(logit),

data = Pima.tr)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7935 -0.8368 -0.5033 1.0211 2.2531

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.49870 1.17459 -5.533 3.15e-08 \*\*\*

age 0.07104 0.01538 4.620 3.84e-06 \*\*\*

bmi 0.10519 0.02956 3.558 0.000373 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 256.41 on 199 degrees of freedom

Residual deviance: 215.93 on 197 degrees of freedom

AIC: 221.93

Number of Fisher Scoring iterations: 4

**bivariate logistic regression**

|  |  |  |
| --- | --- | --- |
| Variable | Coefficient (p value) | Odds Ratio (95% CI) |
| npreg | 0.17(0.00) | 1.18(1.08,1.30) |
| bp | 0.04(0.00) | 1.04(1.01,1.07) |
| skin | 0.04(0.00) | 1.04(1.01,1.07) |
| bmi | 0.10(0.00) | 1.11(1.05,1.17) |
| age | 0.07(1.58e-06) | 1.07(1.04,1.10) |

Results have indicated that all variables have causal association with diabetes since in the all cases the odd ratio is less than 2 and greater than 0.05.

In both case the result and order of variables are same.

Note: Building multiple regression models is a very subjective process! There is no single “correct” answer. You will get full credit for these problems if you follow the correct procedure and explain what you did at each step and your rationale.