## Exercise 1

TTTTF. A 4-level categorical factor has three dummy variables. If the effect of white cell count were due entirely to smoking, it would have disappeared when smoking was included in the model.

## Exercise 2

a) In order to decipher the output from the regression analysis we need to know how the computer decided to encode the categories. Therefore we need the table below.

**Categorical Variables Codings**

| | | Frequency | Parameter coding (1) |
|---|---|---|---|
| age40 | 39 or younger | 3514 | 1.000 |
| | 40 or older | 1534 | .000 |
| daytime | 6 PM - 6AM | 2601 | 1.000 |
| | 6 AM - 6PM | 2447 | .000 |
| weekday | Monday - Thursday | 2642 | 1.000 |
| | Friday - Sunday | 2406 | .000 |
| sex | Female | 1457 | 1.000 |
| | Male | 3591 | .000 |
| race | non-caucasian | 1758 | 1.000 |
| | caucasian | 3290 | .000 |
| injtype | Unintentional | 3966 | 1.000 |
| | Intentional | 1082 | .000 |

b) From the logistic model it can be seen that age is the only factor that does not significantly affect the probability of high alcohol concentration. All other factors are significant. Especially daytime seems important The odds of high alcohol is 5.3 times higher from 6 PM to 6AM than from 6 AM – 6PM.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95.0% C.I.for EXP(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower | Upper |
| Step 1 | injtype(1) | -.362 | .090 | 16.285 | 1 | .000 | .697 | .584 | .830 |
| | daytime(1) | 1.661 | .080 | 427.809 | 1 | .000 | 5.263 | 4.497 | 6.160 |
| | weekday(1) | -.635 | .071 | 80.979 | 1 | .000 | .530 | .461 | .608 |
| | sex(1) | -.540 | .085 | 40.433 | 1 | .000 | .583 | .494 | .689 |
| | race(1) | -.542 | .082 | 44.078 | 1 | .000 | .582 | .496 | .683 |
| | age40(1) | .069 | .082 | .710 | 1 | .399 | 1.071 | .913 | 1.258 |
| | Constant | -1.266 | .129 | 96.648 | 1 | .000 | .282 | | |

c) There are many ways to describe the goodness of fit. SPSS provides a few:

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 4941.031[a] | .140 | .207 |

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Personally I like the figure telling how many correct predictions the model made:

**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | bac50 | | |
| | | | <50 mg/dL | >50 mg/dL | Percentage Correct |
| Step 1 | bac50 | <50 mg/dL | 3533 | 241 | 93.6 |
| | | >50 mg/dL | 913 | 361 | 28.3 |
| | | Overall Percentage | | | 77.1 |

a. The cut value is .500

d) In general the predictive value is:

z = -1.266 (intercept)

-0.362 (if unintentional injury)

+1.661 (if 6PM – 6AM, that is at night if Im not mistaking)

-0.635 (if weekday)

-0.540 (if female)

-0.542 (if non Caucasian)

+0.069 (if 39 or younger)

For this young man, z = -1.266 – 0.362 + 1.661 +0.069 = 0.102

$$p = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-0.102}} = \frac{1}{1+0.903} = 0.53$$