I. Hints for taking the quiz and interpreting the output

a. Read everything carefully; answer what is asked.

b. Make sure you understand (1) Chi-square test and its use for data in contingency tables, (2) basic principles of logistic regression and output interpretation, (3) similarity and difference between methods covered in the class from the beginning of the semester.

c. Don’t lose sight of the scientific question(s). When possible refer to meaningful outcome, factor, and level names rather than codes, “X”, or “Y”.

d. Show your work to get partial credit.

e. When it is appropriate, refer to the specific area of the output that supports your conclusions. E.g. “there is strong evidence (F=15.6, p<0.0005) that differences in time of contact cause changes in rat biting behavior.”

f. State conclusions so that they will answer the questions that someone interested in the topic would naturally have.
   i. Include the direction of an effect when discussing contrasts or main effects with two levels.
   ii. Include magnitude of effects (either point estimates or confidence intervals) where appropriate.

g. Use the words like “association” for observational studies and reserve “causation” for experiments.

h. Avoid the word “prove”. Use “non-significant” rather than “insignificant” to describe high p-values. Reserve “insignificant” to describe substantively small changes.

i. Round final answers to 3 significant figures. Do not write p=0.000! write “p<0.0005”.

j. Understanding ‘design variables’ in SAS output. Consider the example we discussed in class from TITANIC data:

<table>
<thead>
<tr>
<th>Design Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
</tr>
<tr>
<td>CLASS</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

CLASS is a categorical random variable with 3 levels, that is CLASS takes values 1, 2 and 3. Design Variables are like dummy variables (remember ANCOVA). Since CLASS has 3 levels we create 2 new variables (number of levels of the original variable – 1), Class1 and Class2. Class1 takes values 1, 0, and -1 and Class2 takes values 0, 1, and -1. If a person is in the first class, Class1 variable will be equal to 1 or takes value 1, Class2 variable will equal 0. If a person is in the second class, then variable Class1 will be equal to 0 and Class2 will be equal to 1. If a person is in the third class, Class1 variable will be equal to -1 and Class2 variable will be equal to -1. So if I want to calculate the probability of survival for a person in the second class, in your model you would have $b_0+b_1*$Class1 $+b_2*$Class2 $=b_0+b_1*(0)+b_2*(1)$. For a person in the third class, $b_0+b_1*(-1)+b_2*(-1)$
II. Review of regression, ANCOVA, & logistic regression modeling

a. Converting a list of coefficients to a linear contrast (\( \eta \)) equation.
   i. Format: \( \eta = b_0 + b_1 X_1 + \cdots + b_k X_k \)
   ii. Linear regression or ANCOVA: \( \hat{\mu} (Y|X_1, \cdots, X_k) = \eta \)
   iii. Logistic regression: \( \log \left[ \frac{\Pr(Y=1| X_1, \cdots, X_k)}{\Pr(Y=0| X_1, \cdots, X_k)} \right] = \eta \)

b. Categorical variables (easiest forms):
   i. Two levels: 0/1, name as “1” level
   ii. k levels: k-1 0/1 variables, named as “1” levels

c. Interactions:
   i. \( X_i \) is the Product of two (or more) variables
   ii. If one i/a variable is categorical, there are k-1 i/a variables
   iii. If both i/a variables are categorical, there are (k-1)*(m-1)

d. Interpretation of the intercept (constant)
   i. Linear reg.: predicted mean of Y when all X’s are zero
   ii. Logistic reg: predicted log odds of Y=1 when all X’s are zero.
   iii. Null hypothesis is \( \beta_0 = 0 \). For logistic regression, zero log odds means odds=1 which means \( \Pr(Y=1| X_1=0, \ldots, X_k=0) = 0.5 \).

e. Interpretation of slopes that are not in an interaction
   i. The slope is the (additive) change in mean (or log odds of Y) for a 1-unit increase in the specific explanatory variable when all other explanatory variables in the model are held constant. (Think about the constancy or variability of what is not in the model.) Consider talking of a larger or smaller change if it is more meaningful.
   ii. For logistic regression, the antilog, \( \exp(b_i) \), is the multiplicative change in odds for a one-unit increase in the specific explanatory variable when all other explanatory variables in the model are held constant.
   iii. Null hypothesis is \( \beta_i = 0 \), which says that Y does not change when \( X_i \) changes.
f. Interpretation of slopes that are in an interaction

i. Much trickier to interpret. Most people argue that if the interaction is included in a model, the main effects should also be included, whether or not they are “significant”.

ii. Obviously the above interpretation for “no interaction” explanatory variables does not apply!

iii. Null hypothesis is $\beta_{ij}=0$, which says that $Y$ does not change differently when $X_i$ changes (i.e., follow a different pattern) when comparing the effect at different fixed levels of $X_j$ (or vice versa).

iv. Usually the best way to express the meaning of an interaction is to write separate prediction equations and/or plots for all of some combinations of the interacting explanatory variables.

g. Log transformations (similar for other transformations)

i. Outcome is $\log_{10}(Y)$: $Y$ increases (decreases) 10 fold for a one-unit change in $\eta$, which is a $1/b_i$ unit change in $X_i$, when all other explanatory variables in the model are held constant.

ii. $X_i$ is, e.g. $\log_{10}$(minutes): $b_i$ has the usual interpretation in terms of log(minutes), but it is often better to express the interpretation in terms of minutes, e.g. for every 10 fold increase in minutes (corresponding to a 1 unit increase in log(minutes)), $Y$ increases by $b_i$.

h. Example with outcome equal to milliseconds until target is acquired. Color is one of the three primary colors. Age is years over 18. VAI is one less than the $\log_{10}$ of the second number in “20:20” exam score.

<table>
<thead>
<tr>
<th>Source</th>
<th>Coefficient</th>
<th>SE(coefficient)</th>
<th>t</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>1</td>
<td>0.2</td>
<td>5</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Color.red</td>
<td>2</td>
<td>0.4</td>
<td>5</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Color.blue</td>
<td>3</td>
<td>0.6</td>
<td>5</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Male</td>
<td>4</td>
<td>0.8</td>
<td>5</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Age</td>
<td>5</td>
<td>1.0</td>
<td>5</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Male:Age</td>
<td>6</td>
<td>1.2</td>
<td>5</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Visual Acuity Index</td>
<td>7</td>
<td>1.4</td>
<td>5</td>
<td>&lt;0.0005</td>
</tr>
</tbody>
</table>

i. Prediction equation:

$$\hat{\mu}(Y|\text{color, male, age, VAI})=1+2*\text{red}+3*\text{blue}+4*\text{male}+5*\text{age}+6*\text{male*age}+7*\text{VAI}$$

ii. Interpretation of constant: The mean time in milliseconds until target acquisition for an 18-year-old female with 20:10 vision when the target is green is 1.0. The null hypothesis of $\beta_0=0$ is meaningless because, although “$X_i=0$ for all $i$” is a meaningful group encompassed by the data, it always takes some time to acquire the target.
iii. Interpretation of $b_{\text{VAI}}$: For every 10 fold increase in the second number of the eye exam score, the time till target acquisition increases by 7.0 when gender, age and target color are held constant. If we are unsure whether or not visual acuity affects the outcome, the null hypothesis $\beta_{\text{VAI}}=0$ is meaningful.

iv. Interpretation of $b_{\text{red}}$ and $b_{\text{blue}}$: If $X_{\text{red}}=1$ means “red target” and $X_{\text{blue}}=1$ means blue target, green is the baseline ($X_{\text{red}}=0, X_{\text{blue}}=0$). So target acquisition time increases by 2 for red vs. green targets when gender, age and visual acuity are held constant. This can also be phrased as “For each age, gender and VA combination, the predicted target acquisition time is 2 ms longer for a red than a green target.” It is 3 ms longer for blue vs. green.

v. Interpretations regarding age and gender: The prediction equation for females is:

$$\hat{\mu} (Y|\text{color,female,age})=1+2*\text{red}+3*\text{blue}+5*\text{age}+7*\text{VAI}$$

For males it is:

$$\hat{\mu} (Y|\text{color,male,age})=5+2*\text{red}+3*\text{blue}+11*\text{age}+7*\text{VAI}$$

This implies, e.g. that 18-year-old males are 4 ms slower and 28-year-old males are 114 ms slower than 18-year-old females for each color and VAI. Appropriate plots might include time (y-axis) vs. age with separate lines for each of the three colors for one (or a few) visual acuities, and time vs. visual acuity (perhaps on the original scale) for one or a few ages.