**1. Introduction**

Oropharyngeal squamous cell carcinoma affects 13,000 people in the United States each year. Most cases are attributable to tobacco and alcohol. Human papilloma virus (HPV) has also been confirmed as a significant risk factor in recent years. Men tend to be affected more than women, with a male to female diagnosis ratio of 2.7:1. [1]

Oropharyngeal squamous cell carcinoma presents itself in the tonsil, base and posterior one third of the tongue, soft palate and posterior lateral and pharyngeal walls. [1]

The combination of heavy smoking and drinking exacerbates risk levels, placing an individual at 21 times the risk of a non-smoker or social drinker. The average five-year survival rate is about 60%. However, HPV associated patients have a five-year survival rate of 75% and a three-year survival rate of nearly 90%.

It is the latter point: survival time; we wish to investigate. Typically, patients with oropharyngeal squamous cell carcinoma are treated with radiation therapy or a combination of radiation therapy and a chemotherapeutic agent. Therefore, we intend to investigate if treatment as a factor in the presence of other common factors associated with survival time is a critical component in extending a patient’s time of survival in days.

Formally, we wish to investigate if days of survival are positively affected by ANY treatment in the presence of gender, institution, grade, age in years, physical condition, cancer site, tumor stage and node metastases. An additional investigation if treatment is deemed significant: is the test (combination) treatment more effective than standard radiation?

**2. Experimental Design**

A large clinical trial carried out by the Radiation Therapy and Oncology Group was undertaken from 1968 to 1972 involving 195 subjects. Each subject was diagnosed with oropharyngeal squamous cell carcinoma and was not in remission. The full study included patients with squamous carcinoma in 15 sites of the mouth and throat. However, the available data set only contains three sites.

Patients were randomized into one of two treatment groups: radiation therapy and radiation therapy with chemotherapeutic agent. There is no formal mention of the sample selection techniques employed in the study. Therefore, inferences to larger populations would likely be inaccurate. Further, analysis periods were staggered, with different entry dates over a four-year period from 1968 to 1972. Measures were taken at entry date into the study and days were subsequently tracked until the study’s end date, which was not stated explicitly, nor derivable from the data set itself. Therefore, analysis results should be considered with caution.

**3.** **Exploratory Data Analysis**

Full exposition to explanatory and response variables is required to understand model set up and additional analysis. Data were obtained for the squamous carcinoma study via umass.edu [2] public data. Nine explanatory variables and one response, time, are considered in the study. Variable descriptions are listed below.

**3.1 Explanatory Variables**

Nine explanatory variables are considered. Eight of the nine variables are categorical with multiple levels. An asterisk represents a categorical variable.

|  |  |  |
| --- | --- | --- |
| **Variable 1** | **Levels** | **Description** |
| INST\* | 6 | The medical institution participating in the study |
| Institution contains six levels. The institution names are held anonymous and are coded 1,2,3,4,5,6. | | |
| **Variable 2** | **Levels** | **Description** |
| SEX\* | 2 | The gender of the subject in the study |
| Male is coded as 1, female is coded as 2. | | |

|  |  |  |
| --- | --- | --- |
| **Variable 3** | **Levels** | **Description** |
| TX\* | 2 | The treatment each subject receives in the study. |
| TX is coded to mean treatment. Treatment has two levels, coded 1 for standard treatment, which is radiation and 2 for a combination of chemotherapy and radiation. | | |
| **Variable 4** | **Levels** | **Description** |
| GRADE\* | 3 | Degree of differentiation of the tumor |
| More specifically, grade is the degree to which the tumor cell resembles the host cell. The more abnormal (differentiated), the more risk for cancer growth and spread.    **1:** Well differentiated **2:** Moderately differentiated **3:** Poorly differentiated  Grade is NOT fully reported in the data set. There is one observation missing a grade. Incomplete records will be excluded in this analysis due to missing values and potential confounding influence on further analysis. | | |

|  |  |  |
| --- | --- | --- |
| **Variable 5** | **Levels** | **Description** |
| COND\* | 4 | The physical condition of the subject upon entry into the study |
| Condition is a factor with four levels:  **1:** No disability **2:** Restricted work **3:** Requires assistance with self care **4:** Bed confined  Obviously, the condition of entry is an extremely important covariate in survival time. The expectation, on face value, would be a subject with no disability would survive longer than a subject who is likely bed ridden from cancer. | | |
| **Variable 6** | **Type – Levels** | **Description** |
| T\_STAGE\* | 4 | The clinical stage of the tumor by size in cm |
| The stage of the tumor has four levels:  **1:** < 2cm in largest diameter **2:** Between 2 and 4cm **3:** > 4cm **4:** Massive invasive tumor  The size of the tumor does not automatically result in a worse result, however, medical history shows a strong association between tumor size and a decreased survival time. | | |

|  |  |  |
| --- | --- | --- |
| **Variable 7** | **Type – Levels** | **Description** |
| N\_STAGE\* | 4 | Presence, quantity and status of positive node metastases (by count, size in cm, fixed/non-fixed) |
| N\_STAGE represents spread of carcinoma. Specifically, the presence, quantity and status (fixed or not fixed) of metastases. These statuses are categorical and represented by four levels:    **1:** No metastases **2:** Single node <= 3cm **3:** Single node >3cm **4:** Multiple or fixed node(s)  Fixed, larger metastases are indicators for a degradation in health status and survival days. | | |

|  |  |  |
| --- | --- | --- |
| **Variable 8** | **Type – Levels** | **Description** |
| SITE\* | 3 | The oropharyngeal site of the tumor |
| The site is the physical location the tumor in the oropharynx. The original study contained 16 different sites. The data set explicitly states only five variables were captured in the public data set. However, the public data set itself contains only three sites:  1: Faucial arch (uvula) **2:** Tonsillar fossa (mid-tonsil) **3:** Pharyngeal tongue (rear tongue) | | |

**3.1.1 Quantitative Explanatory Variables**

|  |  |  |
| --- | --- | --- |
| **Variable 9** | **Type – Levels** | **Description** |
| Age | Continuous | The age in years of the subject |
| The age of the subject, represented in years. Descriptive statistics for age with incomplete records removed:  Screen%20Shot%202015-10-08%20at%209.51.48%20PM.png  Visual exploration of age with histogram and QQ plot:  Screen%20Shot%202015-10-08%20at%209.52.39%20PM.pngScreen%20Shot%202015-10-08%20at%209.52.46%20PM.pngScreen%20Shot%202015-10-08%20at%209.58.03%20PM.png  Based on assumptions for multiple linear regression, age as a normal variable would fit into a multiple regression model without much trouble. However, scatterplots do not show a strong linear relationship to survival time in days. Analysis will bear out if age is an important covariate. | | |

**3.2 The Response Variable**

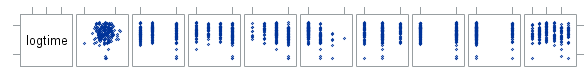
|  |  |  |
| --- | --- | --- |
| **Response** | **Type – Levels** | **Description** |
| Time | Continuous | Survival time in days of each subject |
| The survival time is measured in days for each subject from the entry point into the study until final analysis. Researchers measured previously mentioned explanatory variables one time: at the entry date into the clinical trial. Subjects were then tracked until the clinical trial ended or a final analysis date was determined.  Screen%20Shot%202015-10-09%20at%2011.14.05%20AM.png  Visual exploration of survival time with histogram and QQ plot:  Screen%20Shot%202015-10-09%20at%2011.17.29%20AM.pngScreen%20Shot%202015-10-09%20at%2011.17.54%20AM.png  Using tenants of the central limit theorem and specifics for large sample sizes, analysis can likely move forward using multiple linear regression. However, a conservative approach to meeting assumptions for multiple linear regression is in order, therefore a transformation is attempted. Below we try a log and square root + 0.5 transform:   |  |  | | --- | --- | | **Square Root + 0.5 Transform** | **Log Transform** | | Screen%20Shot%202015-10-09%20at%2012.52.10%20PM.png | Screen%20Shot%202015-10-09%20at%2012.47.35%20PM.png |   Both distributions after transformation look more normal than the original time data. The square transform still maintains a long right tail; however, it is not as explicit. The log, transformation creates a left skew, although relatively slight. For interpretation and assumption, we will move forward with the log transformed time data. | | |

**3.3 Additional Assumption Checking**

Multiple linear regression (MLR) is parametric, uses means and is sensitive to non-constant variance. MLR is relatively robust to departures from normality, however egregious departures warrant transformation. Further, we need to check for linearity in our data set as a baseline to begin a multiple linear regression analysis.

**3.3.1 A Visual Test For Linearity**

Age Site N\_STAGE T\_STAGE COND GRADE TX SEX INST



Linearity assumptions for MLR are met via visual inspection. Eight categorical variables and one continuous variable are viewed as a scatterplot matrix against logtime. Each variable shows weak to moderate linear relationships where categories are assumed ordinal. This relationship can be seen for variables such as N\_STAGE, T\_STAGE, COND and TX. Interesting outcomes for institution, which is not ordinal, show that institution three may be more successful in prolonging patient days than others.

**3.3.2 Model Testing For Overall Normality And Non-Constant Variance**

Univariate normality can be tested at the variable level, as seen previously. However, multiple linear regression normality assumptions pertain to the error associated with the model, and whether those errors are normally distributed. Normally distributed quantitative variables help to ensure model normality, however, the model must still be tested.

While MLR is relatively robust to departures from normality, it is not robust to non-constant variance or heteroscedasticity. This means the variance of errors need to be constant across the regression line. Little non-constant variance can be acceptable, however, moderate to extreme non-constant variance can lead to inaccurate conclusions.

All categorical variables were dummy categorized with level zero or one (where applicable) used as the reference level in order to determine level specific effects via multiple linear regression. An initial model for testing normality and non-constant variance was fit:

|  |  |  |
| --- | --- | --- |
| Assumptions | Graphic | Conclusion |
| Model Normality | Screen%20Shot%202015-10-09%20at%202.31.43%20PM.png | Model normality is acceptable for multiple linear regression techniques |
| Non-Constant Variance | Screen%20Shot%202015-10-09%20at%202.49.26%20PM.png | The model shows some non-constant variance, especially for smaller outcomes. This is expected, as the amount of observations for logtime are fewer on the left tail. The variance violation is not serious enough to warrant further transformation. However, further investigation of leverage points is warranted given regression results. |
| Leverage Plots  And Removal Effects | 1 2  Screen%20Shot%202015-10-09%20at%202.52.08%20PM.pngScreen%20Shot%202015-10-09%20at%203.06.38%20PM.png | Removing six of the seven outlier and potentially influential leverage points eliminates any semblance of non-constant variance (fig. 2). However, there is no reason or evidence to remove the leverage points from the analysis. Further, the violation of non-constant variance is not serious enough to abandon multiple linear regression. Therefore, we move forward with assumptions met. |

**3.3.3 Model Testing For Independence**

Based on the study parameters and experimental design, we assume that all observations were collected independently. Thus, we assume no serial or cluster correlations.

In regard to MLR, the regression of log transformed time on age, condition, treatment, metastases stage, tumor stage, grade, site and institution meets all critical assumptions. Multiple linear regression would give appropriate results.

**4.** **Multiple Linear Regression Analysis**

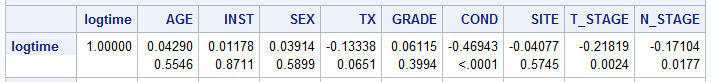
The initial regression equation used to fit a full model is parameterized as follows:

**µ {log(time) | age INST SEX TX GRADE COND SITE T\_STAGE N\_STAGE } =**

β0 + β₁(age) + β₂(inst2) + β₃(inst3) + β5(inst4) + β6(inst5) + β7(inst6) + β8(sexfem) + β9(txcombo) + β10(grademod) + β11(gradepoor) + β12(condres) + β13(condassist)+ β14(condbed) + β15(t\_stage2to4) + β16(t\_stage4cm) + β17(t\_stagemassive) + β18(n\_stages3cm)+ β19(n\_stages3cm+) + β20(n\_stagemultilrg)+ β21(sitetonsil) + β22(sitetongue)

As evidenced by a scatterplot matrix; no curvilinear relationships were present and most categorical variables showed a lack of correlation except for condition, n\_stage and t\_stage.

To confirm visual assumptions of correlations, a pearson correlation matrix was used on the aggregated explanatory variables and logtime:



As expected, the only significant correlations seen are for n\_stage, t\_stage and condition. These variables show weak to moderate negative relationships and no cross-correlation.

Our initial question about treatment type also shows a very weak but insignificant *negative* correlation with logtime. Instead of discarding all other explanatory variables, model selection methods will be employed.

The formal hypothesis to test is:

**Ho: TX(treatment) = 0 in presence of all other covariates**

**Ha: TX (treatment) > 0 in presence of all other covariates**

**4.1 Model Selection and Fit**

After transforming the response variable, all explanatory variables must be back transformed for interpretation. However, first, we must refine the MLR model to ensure we can answer the posed hypothesis.

Under the full model previously mentioned, the R2 value is 0.3752. Such a small R2 for survival analysis is concerning, however, the model does explain a moderate amount of the variance in the response with the predictors in the full model. Leverage points have been considered and kept as they are not influential. All variance inflation values are within acceptable limits, however, as originally suggested by correlation matrices and scatterplots, many predictors are not significant. In order to acquire a significant model for analysis and answer the hypothesis directly, two algorithmic approaches were utilized for variable selection. Traditional selection methods were avoided due to multiple comparison deficiencies:

|  |  |  |
| --- | --- | --- |
| **Selection Method** | **Selection Criteria** | **Resulting Model** |
| LASSO | SBC | µ {log(time) | INST COND T\_STAGE N\_STAGE } |
| LARS | SBC | µ {log(time) | INST COND T\_STAGE N\_STAGE } |

Selection methods employed agree on the same model for predicting the natural logarithm of survival time. In fact, as evidenced by early visual inspection, institution three is a significant explanatory variable, along with condition one and three and later stage metastases presence and tumor size. Treatment is nowhere to be found.

For prediction purposes, which is not our original goal, cross-validation is employed:

|  |  |  |
| --- | --- | --- |
| **Selection Method** | **Selection Criteria** | **Resulting Model** |
| CV Ten Fold | CV | **µ {log(time) | COND T\_STAGE N\_STAGE }** |

Again, treatment falls outside of the model for predicting survival time in days.

**5.** **Conclusion**

Multiple linear regression analysis indicates that treatment has no effect on the log survival time in days in the presence of other covariates associated with oropharyngeal squamous cell carcinoma.

Because we do intend to use this model to predict survival time in days on new data sets, the final parameterized model is:

**µ {log(time) | INST COND T\_STAGE N\_STAGE } =**

**log(time)** = 6.005+ 0.115(inst2) + 0.388(inst3) + 0.389(inst4) + 0.159(inst5) + 0.265(inst6) + -0.828(cond2) + -1.578(cond3) + -0.909(cond4) + 0.369(t\_stage2) + 0.227(t\_stage3) + -0.087(t\_stage4) + 0.095(n\_stage2) + -0.035(n\_stage3) + -0.261(n\_stage4)

Based on exploratory data analysis and multiple model selection methods, we fail to reject the null hypothesis that treatment effects are zero in the presence of other covariates for oropharyngeal squamous cell carcinoma. Our final model results in a factorial equation with institution, condition, metastases stage and tumor stage as significant explanatory variables. Because of the model’s factorial nature, all levels of a factor are included in the model if one level presents as statistically significant.

Interpretations are in median days. For instance, a patient at institution one who has no disability, a small tumor with no evidence of metastases will live for 405 days. Institution three, with all other variables held constant, adds 1.5 days to a patient’s survival time in days. One perplexing outcome is tumor stage three adds net positive days (1.25) to survival time. This is likely due to measurement inconsistencies with the study. Patients were not measured over a fixed period of time. Further, there are very few observations for later tumor stages, which could influence the model when other levels of the factor are included.

There is evidence to show that treatment, in the presence of other covariates, is not significant in extending patient survival time in days. Because of inconsistencies in study design and data set make up, causal conclusions should not be inferred. An association can be stated within the scope of the experiment with caveats taken into account for missing and truncated data.

**REFERENCES**

1. http://www.merckmanuals.com/professional/ear-nose-and-throat-disorders/tumors-of-the-head-and-neck/oropharyngeal-squamous-cell-carcinoma
2. <http://www.umass.edu/statdata/statdata/data/pharynx.txt>

**APPENDIX 1 – SAS CODE**

**proc** **import** out = pharynx

datafile='\\Client\C$\Users\patrickcorynichols\desktop\pharynx.xls'

DBMS =xls REPLACE;

sheet = 'PHARYNX';

getnames=yes;

**RUN**;

/\* remove censored or missing data records (9's and 0's) as these are not complete records and should not be modeled\*/

/\* obs 141 & 136 & 156 & 157 \*/

**data** pharynx2;

set pharynx;

IF \_n\_ = **136** THEN delete;

IF \_n\_ = **141** THEN delete;

IF \_n\_ = **159** THEN delete;

**RUN**;

/\* set up dummy variables for categorical factors \*/

**data** pharynx3;

set pharynx2;

/\* logtransform time to fix non-normality \*/

logtime=log(time);

sqrtime=sqrt(time+**0.5**);

IF inst=**2** THEN inst2=**1**; ELSE inst2=**0**;

IF inst=**3** THEN inst3=**1**; ELSE inst3=**0**;

IF inst=**4** THEN inst4=**1**; ELSE inst4=**0**;

IF inst=**5** THEN inst5=**1**; ELSE inst5=**0**;

IF inst=**6** THEN inst6=**1**; ELSE inst6=**0**;

IF sex=**2** THEN sex2=**1**; ELSE sex2=**0**;

IF tx=**2** THEN tx2=**1**; ELSE tx2=**0**;

IF grade=**2** THEN grade2=**1**; ELSE grade2=**0**;

IF grade=**3** THEN grade3=**1**; ELSE grade3=**0**;

IF cond=**2** THEN cond2=**1**; ELSE cond2=**0**;

IF cond=**3** THEN cond3=**1**; ELSE cond3=**0**;

IF cond=**4** THEN cond4=**1**; ELSE cond4=**0**;

IF site=**2** THEN site2=**1**; ELSE site2=**0**;

IF site=**4** THEN site4=**1**; ELSE site4=**0**;

IF t\_stage=**2** THEN tstage2=**1**; ELSE tstage2=**0**;

IF t\_stage=**3** THEN tstage3=**1**; ELSE tstage3=**0**;

IF t\_stage=**4** THEN tstage4=**1**; ELSE tstage4=**0**;

IF n\_stage=**1** THEN nstage1=**1**; ELSE nstage1=**0**;

IF n\_stage=**2** THEN nstage2=**1**; ELSE nstage2=**0**;

IF n\_stage=**3** THEN nstage3=**1**; ELSE nstage3=**0**;

**run**;

**proc** **print** data = pharynx3;**run**;

/\* basic means investigation \*/

**PROC** **MEANS** data = pharynx3;

VAR time age;

**RUN**;

/\* treatment and condition investigation \*/

**proc** **means** data = pharynx3 n range std var min max median mean;

CLASS tx cond;

VAR time age;

**RUN**;

/\* data investigation on age begins here \*/

**proc** **means** data = pharynx3;

VAR age;

**RUN**;

/\* treatment and condition investigation \*/

**proc** **sgscatter** data = pharynx3;

MATRIX time age;

**RUN**;

/\* age distribution with normal curve and tests (even though we don’t care about the tests – use visuals) \*/

**proc** **univariate** data = pharynx3 NOPRINT;

histogram age /normal(percents=**20** **40** **60** **80** midpercents);

inset n normal(ksdpval)/ pos =ne format = **6.3**;

qqplot age;

**run**;

/\* more EDA, this time on response \*/

**proc** **means** data = pharynx3 mean median std range min max;

var time;

**run**;

**proc** **univariate** data = pharynx3 NOPRINT;

histogram time;

qqplot time;

**run**;

/\* log transform response vs normal curve \*/

**proc** **univariate** data = pharynx3 NOPRINT;

histogram logtime /normal(percents=**20** **40** **60** **80** midpercents);

inset n normal(ksdpval)/ pos =ne format = **6.3**;

**run**;

/\* sqrt transform response vs normal curve \*/

**proc** **univariate** data = pharynx3 NOPRINT;

histogram sqrtime /normal(percents=**20** **40** **60** **80** midpercents);

inset n normal(ksdpval)/ pos =ne format = **6.3**;

qqplot sqrtime;

**run**;

/\* scatter matrix for linearity checks\*/

**PROC** **SGSCATTER** data =pharynx3;

MATRIX logtime age SITE N\_STAGE T\_STAGE COND GRADE TX SEX INST /diagonal =(histogram);

**RUN**;

/\* institution 3 shows deviance from average in the positive (more time) direction, interested to see if it's significant in analysis \*/

**proc** **means** data = pharynx3 n range std var min max median mean;

CLASS inst;

VAR time;

**RUN**;

/\* set up initial model to check for assumptions \*/

**PROC** **REG** data = pharynx3;

MODEL logtime = age

inst2 inst3 inst4 inst5 inst6

sex2

tx2

grade2 grade3

cond2 cond3 cond4

site2 site4

tstage2 tstage3 tstage4

nstage1 nstage2 nstage3 / r VIF influence;

**RUN**;

**QUIT**;

/\* test out removing leverage points to check non-constant variance change \*/

**DATA** pharynxlev;

SET pharynx3;

if \_n\_= **46** then delete;

if \_n\_= **65** then delete;

if \_n\_= **89** then delete;

if \_n\_= **99** then delete;

if \_n\_= **115** then delete;

if \_n\_= **185** then delete;

**RUN**;

/\*re run proc reg without leverage points \*/

**PROC** **REG** data = pharynxlev;

MODEL logtime = age

inst2 inst3 inst4 inst5 inst6

sex2

tx2

grade2 grade3

cond2 cond3 cond4

site2 site4

tstage2 tstage3 tstage4

nstage1 nstage2 nstage3 / r VIF influence;

**RUN**;

**QUIT**;

/\* check initial model, variable significance using PROC GLM \*/

**PROC** **GLM** data=pharynx3;

CLASS inst sex tx grade cond site t\_stage n\_stage;

MODEL logtime = age inst sex tx grade cond site t\_stage n\_stage /solution e;

**RUN**;

/\* patient condition, tumor size and metastases status showing most significance with time \*/

**PROC** **CORR** data = pharynx3;

VAR logtime age inst sex tx grade cond site t\_stage n\_stage;

**RUN**;

/\* apply selection methods to the factorial and quantitative predictors \*/

**PROC** **GLMSELECT** data = pharynx3;

CLASS inst sex tx grade cond site t\_stage n\_stage;

MODEL logtime = age inst sex tx

grade cond site t\_stage n\_stage

/SELECTION = LASSO (choose=SBC);

**RUN**;

**PROC** **GLMSELECT** data = pharynx3;

CLASS inst sex tx grade cond site t\_stage n\_stage;

MODEL logtime = age inst sex tx

grade cond site t\_stage n\_stage

/SELECTION = LARS (choose=SBC);

**RUN**;

**PROC** **PRINT** data = pharynx3;

**run**;

/\*auto selection via lasso based on CV PRESS - cross validate the model \*/

**PROC** **GLMSELECT** data = pharynx3;

CLASS inst sex tx grade cond site t\_stage n\_stage;

MODEL logtime = age inst sex tx grade cond site t\_stage n\_stage

/SELECTION = LASSO (choose=CV) CVMethod=Random(**10**);

**RUN**;

/\* LARS and LASSO agree, we use entire factors in our analysis, including insignificant levels \*/

**PROC** **REG** data = pharynx3;

MODEL logtime = inst2 inst3 inst4 inst5 inst6

cond2 cond3 cond4

tstage2 tstage3 tstage4

nstage1 nstage2 nstage3 / influence CLM CLI PARTIAL;

**RUN**;