## 1. Introduction

### 1.1 Background[1]

> "Paying coaches excessive compensation also makes less revenue available for other sports, causes many athletic departments to operate at a net loss, and may call into question the priorities of educational institutions." Congressmen Bill Thomas quoted in (Upton and Wieberg, 2006)

As the number one sport for revenue in college athletics, football provides a source of income, prestige, and pride for colleges throughout the country. But with more than one-third of the current Division I coaches making in excess of a million dollars a year, some observers are asking if colleges and universities have lost sight of their primary mission - education. And at least one university is asking how they can balance the calling of education with the siren sound of national championships.

At the University of Iowa, Gary Barta is the current athletic director and is responsible for running a successful 22-sports athletic program that represents the University in competitions across the nation (UI, 2007). An always pressing concern for Mr. Barta is hiring and retaining the best coaches for each of the athletic teams. His recent selection of Todd Lickliter to lead the men's basketball program stands out as a success. But for many, the most important coaching position is the head football coach. With the end of the 2007 season only a few weeks away, Mr. Barta is preparing for the annual discussion of compensation for the head football coach.

As Mr. Barta is preparing for the contract negotiations, one of his assistants mentions the practice of "evidence-informed" decision making and hierarchical modeling. Intrigued, Mr. Barta asks for, and receives, the following synopsis:

---

[1] *Author makes no statement supporting or criticizing the schools' decisions. The author found the initial study interesting and wanted to examine the potential relationships between available data and compensation. The following scenario is purely hypothetical and does not represent the official policy or positions of the Gary Barta, the Athletic Department, or the University of Iowa.*

**Specify**: Specify the problem.
**Observe**: Observe the data, or evidence, available related to the problem.
**Analyze**: Analyze the relationships between the evidence and the problem.
**Recommend**: Recommend a solution based on the evidence.

Recognizing the elegant simplicity and yet mathematical complexity involved in "evidence-informed" decision making (Brown, 2006), Mr. Barta asked his assistant to contact some graduate students for assistance with his preparations.

The graduate research team has a history of using computational statistics to provide a better option than guessing. Using a variety of data sources, the research team will conduct the initial analysis of current market salaries in order to assist the Hawkeye athletic department with making better decisions regarding coaches' salaries. The initial focus of the students' efforts will be determining relationships between current school and coaching metrics and salaries. Based on the developed models, the graduate students will provide a recommendation on contract negotiations for the head football coach annual salary.

## 1.2 Problem Specification

*Goals*:

1. Develop a model for explaining the relationship between collected variables and salaries of college football coaches.

2. Predict an appropriate salary for a football coach at the University of Iowa.

3. Recommend a potential range of salary based on current market conditions to be used in future contract negotiations.

*Hypothesis*: Available data suggest that coaches' salaries cannot be predicted.

*Data*

The research team collected a variety of data to assist with the coaches' salary project. The primary source used for the coaches' salaries was a recent report in USA Today on the contracts for college football coaches. This report provided access to current salaries, contracts, and bonus pay for the majority of the 119 NCAA Division I college football coaches. For this study, the research team used the total income figure provided by USA Today. The total income figure includes the base salary and other income such as media contracts. The total income figure does not include performance or other bonuses (Upton and Wieberg, 2006).

The research team collected additional data through a variety of sources as outlined in the references. The following table provides a quick description of the *Coaches* dataset:

| Variable Name | Description | Comments |
|---|---|---|
| School | Name of university | |
| Conference | Name of athletic conference | |
| Wins | Number of wins during 2005 season | |
| Losses | Number of losses during 2005 season | |
| Stadium | Seating capacity of stadium | |
| CoachTenure | Number of years head coaching experience | |
| CoachRecord | Overall winning percentage as head coach | |
| FootballYears | Number of years with football program | |
| WinPercent | Overall winning percentage of school | |
| GradRate | Graduation success rate of football program | As measured by NCAA |
| Enrollment | Size of student body | Enrollment during 2005 |
| Tuition | Cost of tuition ($) | Based on in-state rates when available |
| Rank | Overall team rankings | Lower is better – As observed by Foxsports |
| Salary | Coaches salary for 2005 ($) | Does not include potential bonuses |

**Table 1: *Coaches* data descriptions**

While the original US Today coaches' salary data set is rather comprehensive, several colleges are missing from the study. The total number of

colleges examined for this study is 105. The following table outlines the 12 schools not included in this study.

| Schools not included in study | Remarks |
|---|---|
| Penn State<br>Pittsburgh | Release of Records pending court decision |
| United States Military Academy | Information not released |
| University of Miami – Florida<br>Stanford<br>Vanderbilt<br>Notre Dame<br>Rice<br>Temple<br>Tulane<br>FIU | Private Schools |
| United States Naval Academy | Independent Schools (not included for hierarchical modeling) |

**Table 2: Missing Schools from *Coaches* data**

*Metrics*

As means of comparing a predictive model's performance, the graduate research team divided the data into two sets; a test set and a training set. The test set was used to validate the performance of the model prior to making predictions for the Hawkeye coach. An overall comparison of models will examine the mean absolute error between the actual and predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left| \; y_i - \hat{y}_i \; \right| \qquad (1)$$

## 2 Observation

> "players watch the coach roaming the sidelines in his $1,500 custom made suit. They read about his $500,000 salary and$250,000 per [year] from some sneaker deal. They watch the schools sell jerseys with the players' names … They see the athletic director getting rich and the college president getting rich and NCAA officials getting rich and the coach's dog getting rich." Tony Kornheiser, sportswriter and broadcaster, quoted in (Sperber, 2000)

The graduate research team focused their initial analysis on examining relationships found across the data set. This initial analysis provided additional insight into which variables would be more influential during model development. Of immediate note was the inverse linear relationship between *wins* and *losses*. During the model development portion, the study calculated this variable as a winning percentage.

First observations also indicated that *stadium* capacity, *wins*, and *winpercent* have a positive linear relationship with salary. The variables *losses*, *rank* and *gradpercent* appeared to have a negative linear relationship. However, the *rank* variable is ordinal in the sense that lower is better so we should expect that coaches with a lower number ranking (i.e. in the top 10) are paid more. Figure 1 shows a correlation chart that compares relationships between the variables and depicts these relationships as ellipses. The tighter the ellipse, the stronger the relationship or the more correlation between variables. The orientation of the ellipse indicates whether the correlation is positive or negative.

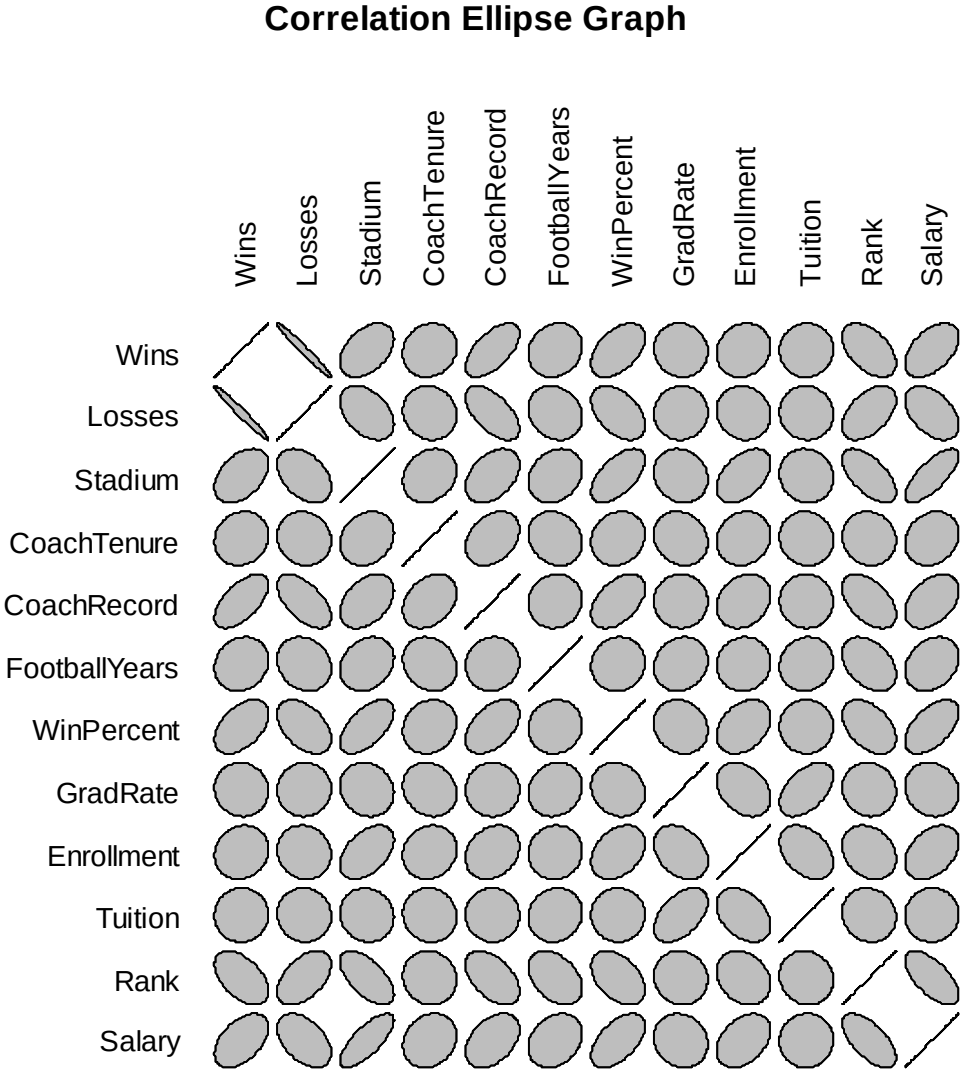



**Figure 1: Relationships between variables**

The histogram provided the research team a quick overview of the coaches' salaries. Using frequency rather than counts we can see that just over one third of the coaches are making more than a $1,000,000 a year.

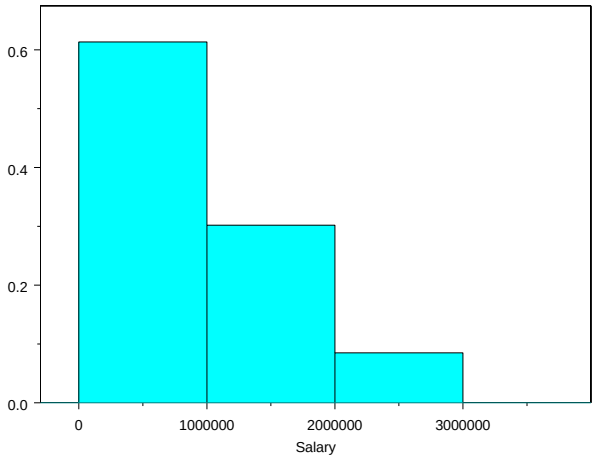



**Figure 2: Coaches' salaries - Histogram**

Since one goal of the research team was to examine hierarchical relationships, the study needed to examine the impact of the conference on a coach's salary. Figure 3 depicts traditional box plots conditioned on the conference. Strong variability across conferences would indicate potential significance of that variable for estimating differences in salary for that conference.

The larger variation across conferences indicates that the qualitative variable *conference* will probably be a significant factor to consider in the model development process. With much mentioned in the media about the relative strengths or weakness of the different conferences, the above plot demonstrates the strength of the SEC when it comes to the coaches paycheck. The boxplots also highlight some of the individual schools with outlier salaries; USC in the PAC 10, Oklahoma in the Big 12, and Iowa in the Big 10. These observations will need to be re-examined in the model development process.

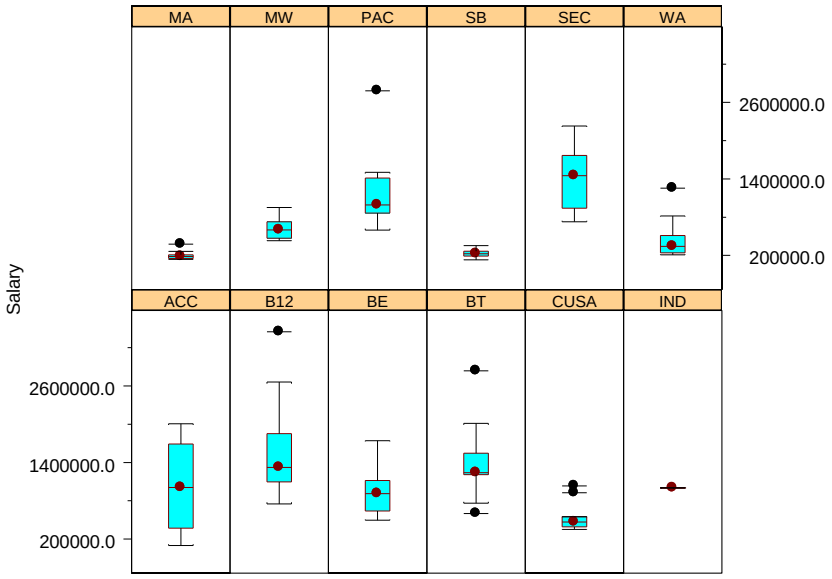
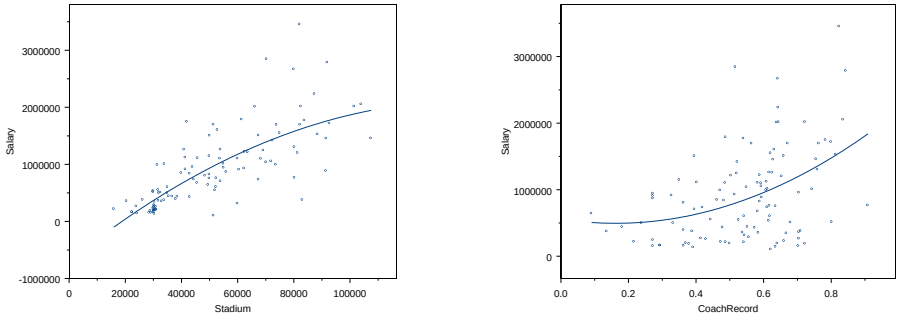



**Figure 3: Coaches' salaries by Conference**

Prior to beginning model development, the research team continued the data analysis with a deeper exploration of the relationships between the different variables. Looking back at the initial correlation plot (Fig 1), we can identify four potential variables that have a more pronounced linear relationship with *Salary*: *Stadium*, *CoachRecord*, *GradRate*, and *Rank*. The following plots depict the relationship between each variable with *Salary*. While some general relationships exist, the stadium size and the coach's record appear to be most significant in the positive direction; however, none of the relationships alone account for all the variance in the dataset.



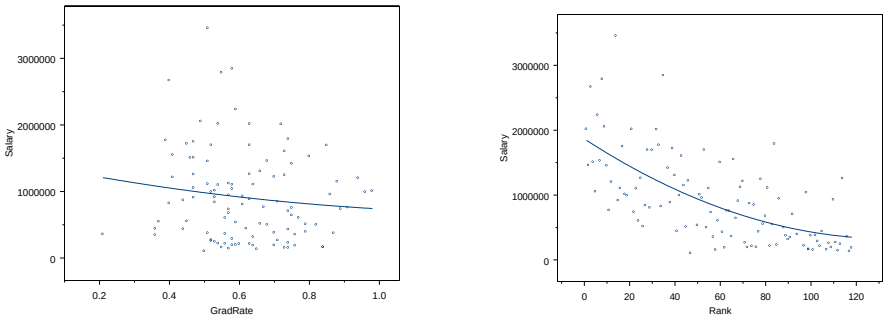



**Figure 4: Smoother Plots with Potentially Significant Variables and Salary**

The principal component analysis allows the research team to display all the variables in one view and examine which components account for the most variance within the data. Figure 4 depicts the scree plot and the biplot indicating that the first two principal components account for a large portion of the variance and that the *stadium* variable is the strongest factor in the first principal component. We can expect that during model development, the *stadium* variable will be significant in most models. On a logical level this makes sense. A large stadium means more people in the seats and more revenue. A successful program (winning) will probably expand their stadium to accommodate more people, creating more revenue. A coach at a school with a large stadium is probably running a successful program and will reap the financial rewards. University of Alabama Birmingham is probably an exception since the coach makes "only" $371,000 with a stadium that holds 83,000 people.

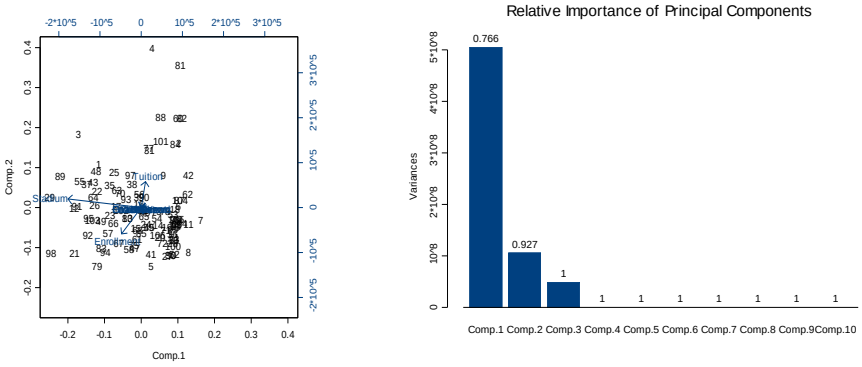



**Figure 5: Principal Component Analysis**

Additional efforts by the research team focused on examining each of the variables in turn and the interactions between the variables. Information gained from this analysis was helpful in developing models that could be used to predict a salary range for a coach at a given school.

## 3 Analysis

> "I believe the high-priced temporary coach is inimical to the development of a permanently high tone in the athletic affairs of a college." Dean Small of the University of Chicago, speaking for a fellow dean at the 13th annual NCAA convention, December 1918 (New York Times, 2007)

*Model Development*

Expanding on some of the observations developed above, the research group proceeded to model development with the focus on creating accurate models for predicting the coach's salary based on a selected group of independent variables. Following the methodology presented in (Gelman and Hill, 2007), the research group developed basic statistical models that summarize how the outcome variable of a coach's salary is a function of several predictor variables. The initial approach was two-fold. First, examine pooled and no-pooled models for the coaches' salaries. Gelman and Hill define a pooled model as one that does not include group level indicators. A no-pooled model includes the group indicators (Gelman and Hill, 2007).

For this study the research group developed basic pooled and no-pooled models and then developed multilevel models that would account for salary differences between the conferences. The equation below provides an overview of the multilevel model used for the final model:

$$\begin{aligned} \log\ Y_i &= \alpha_{j\,i} + \boldsymbol{\beta}\mathbf{X}_i, \text{ for } i = 1,\ldots,n \\ \alpha_j &\sim N\ \ \mu_\alpha + \gamma_j, \sigma^2_{Conference}\ \text{, for } j = 1,\ldots,Conference \end{aligned} \tag{2}$$

The response of $Y_i$ is the coach's salary. The research group takes the log of the salary in order to normalize the response and based on previous salary research work (Corcoran, et al., 2006). $\mathbf{X}$ is a vector of predictor variables and

the intercept is varied based on the conference. While we might expect the non-conference (or pooled) models to be more robust, the research group expected the conference models to be more accurate predictions of how a school must compete within the local market (conference).

*Model Comparison*

Using multiple methods for selecting predictor variables, the research group developed five models for comparison. The least complex model is the pooled model which only included stadium size as a predictor. The more complex multilevel models (MLM) included the conference as the group indicator with multiple predictors. To ensure that the selection of the test set did not overly impact the results, the research group used cross validation with bootstrapping (Faraway, 2006). This important sounding term simply means that the research group split the data into test and training sets a thousand times and took the average MAE for each model across all results. The following table provides an overview of each model with the results for the mean absolute error (MAE) calculation against the test set. A lower MAE means the model provided a closer prediction to the actual salaries in the test set.

| **Model** | **Predictor Variables** | **MAE** |
|---|---|---|
| Pooled | Stadium size | $316,308 |
| No-Pooled | Stadium size & Conference | $290,909 |
| MLM – 1 | Stadium size & Conference | $286,723 |
| MLM – 2 | Stadium size, Conference, & Win Percentage | **$282,651** |
| MLM – 3 | Stadium size, Conference, Win Percentage, & Rank | $382,838 |

**Table 3: Comparison of models – based on Mean Absolute Error**

As can be seen from the table, the multilevel model with stadium size, win percentage, and conference provides the best overall fit. However, the first multilevel model and the no-pooled model are not far off in terms of overall accuracy. A visual depiction of these three models' performance offers some additional insights.

The following graphs depict the model's performance on one test set that includes the Iowa coach. Each graph has been sorted based on the predicted salary. While visually the no-pooled model offers a better overall fit to actual salaries, the impact of the large salaries (UVA, Virginia Tech, and Iowa) raises the MAE. Removing Iowa from the test set reduces the MAE for the no-pooled

model by almost $50,000. Similar improvements would also be seen in the MAE for MLMs. As a general comment the no-pooled model provides a closer fit but fails to account for potential outliers while the multilevel models "split the difference" between pooled and no-pooled models (Gelman and Hill, 2007). The predictions for the MLM were built using simulations that included the uncertainty of the model fit.

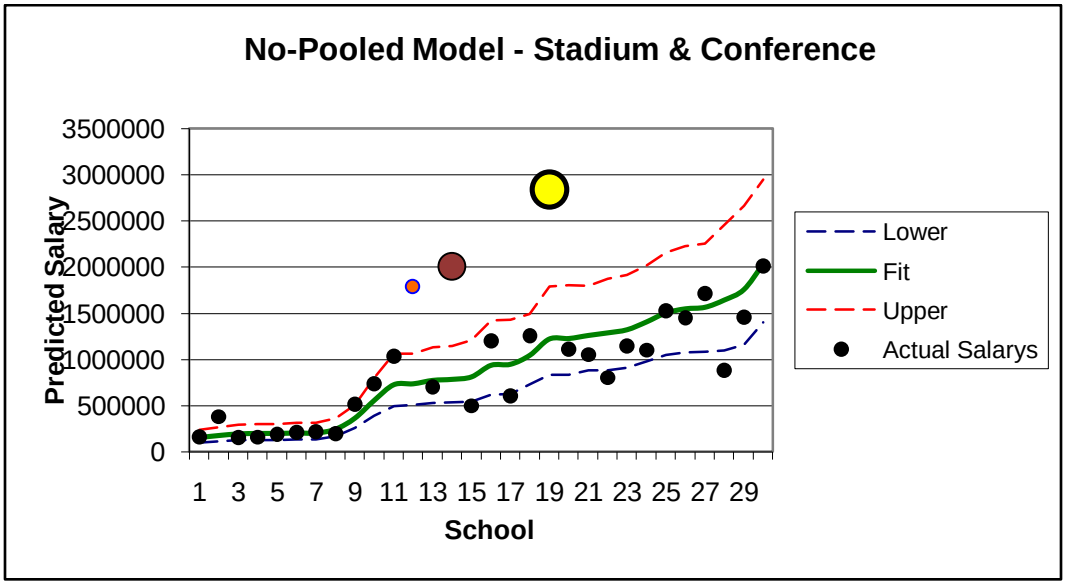



**Figure 6: Predictions for No-Pooled model**

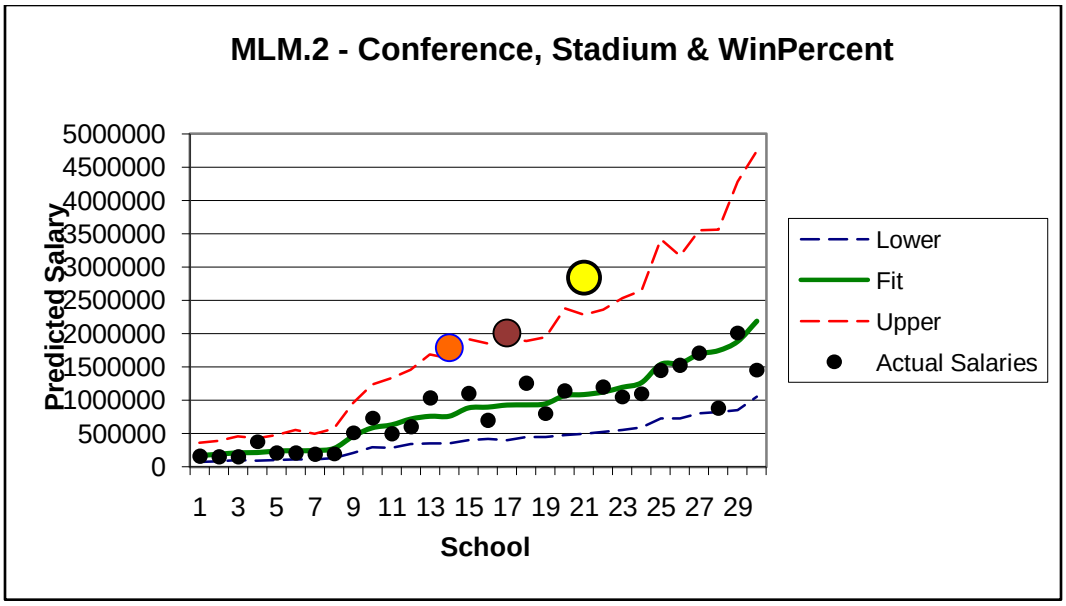



**Figure 7: Predictions for MLM model**

Using the multilevel model with two predictors, the research group predicted the following salary range for the current Hawkeye coach, the other two high salaried coaches in the test set, and the relative bargain of UCLA.

| School | Coach | Current Salary | Predicted Salary | Lower Bound | Upper Bound |
|---|---|---|---|---|---|
| Iowa | Frentz | $2,840 ,002 | $1,084,083 | $496,147 | $2,287,067 |
| **Other schools of interest** | | | | | |
| Virginia Tech | Beamer | $2,008,002 | $926,774 | $400,164 | $1,929,739 |
| Virginia | Groh | $1,785,000 | $758,672 | $358,000 | $1,627,669 |
| UCLA | Dorrell | $881,002 | $1,745,416 | $821,571 | $3,568,393 |

**Table 4: Comparison of selected coaches from test set**

## 4. Recommendation

"I've learned it's silly to leave money on the table … they fire you besides hire you." Kirk Ferentz, University of

Iowa Head Football Coach quoted in (Upton and Wieberg, 2006)

*Prediction*

The graduate student research group provided the following predictions based on the non-conference and conference models:

| School | Coach | Current Salary | Predicted Salary | Lower Bound | Upper Bound |
|---|---|---|---|---|---|
| Iowa | Frentz | $2,840 ,002 | | | |
| **Models** | | | | | |
| Non-Pooled | Stadium Size, Conference, WinPercent | | $1,218,703 | $831,142 | $1,786,983 |
| MLM | Stadium Size, Conference, WinPercent | | $1,084,083 | $496,147 | $2,287,067 |

**Table 5: Recommended models for two recommended models**

Based on the results, the research group recommended that Mr. Barta restructure the Hawkeye coaching contract based on the conference model with the bonus package built upon rankings and other on-field performance.

*Future Coach Selection Strategy*

The results of this study provide two usable models for future coaching selection and salary negotiations. If the Hawkeyes decide to select a new coach in the future, that coach's information can be factored into the model for establishing a salary range based on market conditions and the coach's proven record.

Graduate students will always be available to assist the athletic department in taking the guess work out of their future salary decisions. Additional data collection might refine the ability to account for all unexplained residuals in the models. A Bayesian approach could examine this hierarchical relationship in detail and provide a more concise prediction interval.

**References**:


Brown, D.E., *Data Driven Systems Assessment*, Department of Systems and Information Engineering, University of Virginia, P.O. Box 400747, Charlottesville, VA 22904.

Corcoran, M.E, P.N. Courant, and P.A. Raymond, "University of Michigan Gender Salary Study: Summary of Initial Findings." Available on-line at http://www.provost.umich.edu/reports/U-M_Gender_Salary_Study.pdf. Last accessed 5 Dec 2006.

Faraway, J.J., *Linear Models with R*, Chapman and Hall, CRC, 2005.

Fox, J., *An R and S-Plus Companion to Applied Regression*, Sage Publications, 2002.

Gelman, Andrew, and Jennifer Hill. *Data Analysis using Regression and Multilevel / Hierarchical Models.* Cambridge: Cambridge University Press, 2007.

New York Times, "Evils of College Sports Denounced", 28 Dec, 1918. Available on-line at http://query.nytimes.com/mem/archive-free/pdf?_r=1&res=9902E7D81339E13ABC4051DFB4678383609EDE&oref=slogin. Last accessed on 27 October 2007.

Sperber, M., *Beer and Circus: How Big Time College Sports is Crippling Undergraduate Education*, Henry Holt & Company, 2000.

Upton, J., and S. Wieberg, "Contracts for College Coaches cover more than Salaries", USA Today, 11 November 2006. Available on-line at http://www.usatoday.com/sports/college/football/2006-11-16-coaches-salaries-cover_x.htm, Last accessed on 1 Dec 2006.

UA, http://hawkeyesports.cstv.com/administration/athletic-director.html, last accessed on 28 October 2007.

**Other references consulted:**

The study consulted the following sources to collect the assortment of independent variables. Where data was not available in a central source, the study turned to each university's web site.

**NCAA Records / Stadium Size**

http://www.ncaa.org/library/records/football/football_records_book/2006/2006_d1_football_records_book.pdf

**Graduation Rates**

http://www.ncaa.org/grad_rates/2005/d1_school_data.html

**Enrollment numbers**

http://www.schoolguides.com/

**Tuition**

http://www.usnews.com/usnews/edu/college/rankings/rankindex_brief.php

Meroney, J. "Is the Naval Academy off Course?", *The American Enterprise*, Vol. 10, July 1999.

**Rankings**

http://msn.foxsports.com/cfb/polls

**NOTES:**

Author makes no statement supporting or criticizing the schools' decisions. The author found the initial study interesting and wanted to examine the potential relationships between available data and compensation. The provided scenario is purely hypothetical and does not represent the official policy or positions of the Gary Barta, the Athletic Department, or the University of Iowa. The author selected the Hawkeye's based on his appreciation of the aforementioned and repeated below quote. While the author feels that the "excessive" salaries for all

coaches should be a point of discussion, from the coaches' view, the author agrees with Coach Ferentz of Iowa:

"I've learned it's silly to leave money on the table … they fire you besides hire you"

Just ask Larry Coker.