

WILLIAM L WELBOURN, JR

Candidate, PhD in Statistics

Department of Mathematics and Statistics

Utah State University

3900 Old Main Hill

Logan, UT 84322-3900

Fax: (435) 797-1822

Email: bill.welbourn@aggiemail.usu.edu

Website: <http://osastatistician.com>

RESEARCH INTERESTS

Statistical Methodology. Doctoral Dissertation investigated and corrected problems associated with multiple hypothesis testing in genome-wide association studies (GWAS):

- Permutation procedures, such as maxT and minP, are the widely accepted gold standard for control of the family-wise Type I error rate in GWAS. Implementation of said procedures are extremely computational, and as such are seldom applied in GWAS. For example: One million maxT permutations upon a balanced case-control GWAS data set comprising a 500K biallelic SNP panel and 5000 study subjects could take upwards of 4 CPU years of computational time (based upon executing under the PLINK software package). Have developed a graphical processing unit (GPU) algorithm for maxT permutation which is upwards of 750 times faster than PLINK. That is, 2 CPU years of computing time could be performed upon my desktop computer in about one day.
- Have found that multiplicity adjustment under maxT/minP can lead to inaccurate adjustment. That is, even when one is under the impression s(he) is ‘doing the right thing’ in adapting maxT (or, minP) for multiplicity adjustment, this may not be the case. Have developed a minP method – based on the unconditional distribution of the asymptotic test statistics null distribution – which corrects for this inaccuracy.

Doctoral Dissertation developed a novel tool for detecting genes involved in a gene-environment interaction, called GEM. Presentation to the USU Center for Epidemiologic Studies Seminar Group, February 6, 2012: http://osastatistician.com/Welbourn_EpiSeminarPresentationGxE_2_6_2012.pdf.

Statistical Computing.

- GPU clustering.
- Programming within the SAS, R (including the Bioconductor Project and interfacing with L^AT_EX by way of Sweave), C, and NVIDIA CUDA for C languages.
- Developing C and CUDA code to harness the power of intra-PC parallel (i.e., multi-threaded applications) computing. Exceptional computational gains are realized (see e.g., <http://osastatistician.com/CUDAAbst.pdf>; <http://osastatistician.com/clustermaxt.pdf>). Very active research area in the fields of computer science and electrical engineering, as personal computer central processing units now possess multiple cores.
- Database management within the Structured Query Language (SQL), interfacing with each of the R and C languages. Particularly, management of extravagant genetic databases.
- Graphical user interfacing (GUI) within the R programming environment.

Statistical tools applied to biomedical and genomic studies.

- Generalized linear models (GLM) theory. In particular, linear and logistic regression, as these methods are intimate to GWAS.
- Survival analysis methods: parametric and non-parametric methods; semi-parametric methods (i.e., cox proportional hazards regression models).
- Experimental design: completely randomized designs; block designs; split-plot designs; fully specified fixed effects models; mixed effects models; fully random effects models; unbalanced designs; multi-way ANOVA and MANOVA approaches.
- General least-squares methodology to handle correlated quantitative responses.
- Genetic mapping of complex traits: SNP-based Association Studies in the context of GWAS; microarray-based mapping of mRNA expression profiles.
- Statistical bioinformatics: Analysis of mRNA expression among microarrays (e.g., Affymetrix GeneChips); Permutation based multiple hypothesis testing procedures (e.g., maxT, minP, SAM).
- Semi-conditional statistical tools within GWAS, applied to: Allelic, Cochran-Armitage Trend, Dominant, Recessive, and Additive models, when the phenotypic trait is binary.

EDUCATION

PhD in Statistics, Utah State University, February 2012 (Projected).

Inductee, Golden Key International Honour Society, February 17, 2009.

Nominee, International Scholar Laureate Program, December 2009.

Dissertation title: *Robust Computational Tools for Multiple Testing With Genetic Association Studies.*

For a copy of the abstracts, table of contents, and introduction chapter, see http://osastatistician.com/Welbourn_DissertationChapter1_1_21_2012.pdf.

Dissertation committee: Christopher Corcoran, Chair; Adele Cutler; Kady Schneiter; John Stevens; Ronald Munger.

MSc in Biostatistics, University of Southern California, May 2006.

Inductee, The Honor Society of Phi Kappa Phi, May 10, 2006.

Thesis title: *An Investigation of the Performance for Ordered Subset Analysis to a Gene-Environment Interaction Model.* For a copy of the manuscript, see <http://osastatistician.com/MSThesis.pdf>.

Thesis committee: Kimberly Siegmund, Chair; W. James Gauderman; Bryan Langholz.

BA in Mathematics, Emphasis: Probability and Statistics, California State University, Fullerton, January 2001.

AWARDS AND HONORS

- Technology Expert Award. Department of Mathematics and Statistics, Utah State University, April 19, 2010.
- Nominated to the International Scholar Laureate Program of the Golden Key International Honour Society, December 2009.
- Research Writing Award. Department of Mathematics and Statistics, Utah State University, April 17, 2009.
- Inducted into the Utah State University Chapter of the Golden Key International Honour Society, February 17, 2009.
- Inducted into the University of Southern California Chapter of the Honor Society of Phi Kappa Phi, May 10, 2006.
- Sole recipient of the “Star of the Quarter” award for the fiscal quarter ended June 30, 2001. Co-recipient of the “Star of the Quarter” award for the fiscal quarters ended September 30, 2001 to September 30, 2002, Actuarial Consultants, Inc.
- Dean’s Honor Role List: final three consecutive semesters of California State University, Fullerton Tenure, January 2000 - January 2001.
- High Honors Distinction: May 1990 Golden State Examination, Geometry.

EMPLOYMENT

Graduate Teaching Instructor, August 2007 – May 2012.

Department of Mathematics and Statistics, Utah State University, Logan, UT.

- **Fall 2007: MATH 1050, Section 011** – *College Algebra*.
- **Spring 2008: MATH 1050, Section 009** – *College Algebra*.
- **Fall 2008: STAT 2000, Section 001** – *Statistical Methods*.
- **Spring 2009: STAT 2000, Section 002** – *Statistical Methods*.
- **Summer 2009: STAT 2300, Section 002** – *Business Statistics*.
- **Fall 2009: SAS/R Computer Lab Instructor, STAT 5100** – *Linear Regression and Time Series*.
- **Fall 2009: SAS Computer Lab Instructor, STAT 5200** – *Experimental Designs*.

- **Spring 2010: Recitation Leader (three sections), STAT 3000** – *Statistics for Scientists*.
- **Summer 2010: STAT 3000, Section 001** – *Statistics for Scientists*.
- **Fall 2010: SAS Computer Lab Instructor, STAT 5100** – *Linear Regression and Time Series*.
- **Fall 2010: SAS Computer Lab Instructor, STAT 5200** – *Experimental Designs*.
- **Summer 2011: STAT 3000, Section 001** – *Statistics for Scientists*.
- **Spring 2012: STAT 3000, Section 001** – *Statistics for Scientists*.

Research Assistant, January 2009 – May 2012.

Department of Mathematics and Statistics, Utah State University, Logan, UT.

- **Spring 2009:** Data manager of an extensive genetic database. Transposing data for more than 660,000 SNP loci.
Goal: Obtain the data structure for which GWAS software can analyze.
- **Summer 2009:** Data manager of an extensive genetic database.
Goal: Concordance of genotyping between buccal swabs and blood samples for more than 1.2 million SNP loci.
- **Spring 2010:** Data management and analysis of a GWAS database.
Goal: Reanalyze a previously published GWAS dataset, applying the maxT multiple hypothesis testing procedure, thereby achieving the potential for greater statistical power. Database: 1,001 cases of bipolar disorder and 1,034 controls; 697,266 SNP loci.
- **Fall 2010:** Data management and analysis of a GWAS database.
Goal: Reanalyze a previously published GWAS dataset, applying my developed multiple hypothesis testing procedure methodology, thereby achieving the potential for greater statistical power and accuracy. Database: 1,001 cases of bipolar disorder and 1,034 controls; 875,578 SNP loci.

Homework Grader, January 2009 – May 2009.

Department of Mathematics and Statistics, Utah State University, Logan, UT.

- **Spring 2009: MATH 5720, Section 001** – *Introduction to Mathematical Statistics*.

Data Analyst, June 2006 – November 2006.

Department of Care Management, Corporate Office, Banner Health, Phoenix, AZ.

Graduate Teaching Assistant, May 2005 – December 2005.

Division of Biostatistics, Keck School of Medicine
University of Southern California, Los Angeles, CA.

– **Summer 2005/Fall 2005: PM 510** – *Principles of Biostatistics*.

Consultant, April 2005 – December 2005.

Statistical Consultation and Research Center, Keck School of Medicine
University of Southern California, Los Angeles, CA.

Substitute Teacher, December 2003 – June 2004.

Downey Unified School District, Downey, CA.

Pension Administrator, March 2003 – October 2003.

Actuarial Consultants, Incorporated, Torrance, CA.

Administrator Assistant, February 2001 – March 2003.

Actuarial Consultants, Incorporated, Torrance, CA.

PAPERS

- M. Slattery, A. Lundgreen, **B. Welbourn**, B. Caan, and C. Corcoran, “Oxidative balance and colon and rectal cancer: interaction of lifestyle factors and genes,” *Carcinogenesis*, under review, 2011.

PRESENTATIONS

- Summer 2005 – Final project for Biomedical Informatics Course, University of Southern California, August 3, 2005.
Artificial Neural Networks: Numerical Approximations to Complex Problems.
- Fall 2007 – Final project for STAT 6710: Mathematical Statistics I, Utah State University, December 5, 2007.
The Moment Generating Function for the Multivariate Normal Distribution.
- Fall 2007 – Final project for STAT 6410: Applied Spatial Statistics, Utah State University, December 10, 2007.
A Summary and Implementation of the Generalized Least Squares Algorithm.
<http://osastatistician.com/Stat6410/GLSpres.ppt>.
- Spring 2008 – Presentation to the Epidemiology Seminar Group, Utah State University, January 28, 2008; February 4, 2008.
Introduction to Survival Analysis.
<http://osastatistician.com/Stat6910/Survpres.ppt>.

- Spring 2008 – Final project for STAT 6720: Mathematical Statistics II, Utah State University, April 2, 2008.
More on Hypothesis Testing: Likelihood Ratio Tests.
- Spring 2008 – Final project for STAT 6820: Statistical Bioinformatics, Utah State University, April 25, 2008.
The Power and Integrity of Multiple Hypothesis Testing Procedures in the Realm of Microarray Technology.
<http://osastatistician.com/Stat6820/MHTpres.ppt>.
- Fall 2008 – Final project for STAT 7550: Statistical Computing, Utah State University, November 21, 2008.
Data Management Using R - Interfacing with the Structured Query Language.
<http://osastatistician.com/Stat7550/SQLpres.ppt>.
- Fall 2008 – Final project for STAT 7810: Advanced Regression Analysis, Utah State University, December 1 – 5, 2008.
Survival Analysis.
<http://osastatistician.com/Stat7100/Survival.pdf>.
- Spring 2009 – Project for STAT 6560: Graphical Methods, Utah State University, January 28, 2009.
John Snow's Contribution to Modern Epidemiology.
<http://osastatistician.com/STAT6560/JS.pdf>.
- Spring 2009 – Presentation to the Epidemiology Seminar Group, Utah State University, March 16, 2009.
Data Management of a Large Genetic Database: The Utility of the SQL Database Management System.
<http://osastatistician.com/Stat6910/Dataman.pdf>.
- Fall 2009 – Guest Lecturer, STAT 5100 Class, Utah State University, November 13, 2009.
Ordered Subset Analysis: The Development of a Two-Dimensional Permutation Based Statistical Methodology for Association Testing of Gene-Environment Interaction.
http://osastatistician.com/OSA_11_13_09.pdf.
- Fall 2009 – Presentation to the Epidemiology Seminar Group, Utah State University, November 16, 2009.
Investigating Genetic Markers Associated with Venous Thromboembolism: A Genome Wide Association Study (GWAS).
<http://osastatistician.com/Stat6910/DWsum111609.pdf>.

- Spring 2010 – Presentation to the Epidemiology Seminar Group, Utah State University, February 8, 2010.
Multiple Hypothesis Testing in Gene Expression and Genome Wide Association Studies.
http://osastatistician.com/Stat6910/Welbourn_Epipres2_2010_MHT.pdf.
- Fall 2010 – Presentation to the Epidemiology Seminar Group, Utah State University, December 6, 2010.
Enhancements to Multiple Hypothesis Testing in Genomewide Association Studies.
http://osastatistician.com/Stat6910/Welbourn_EpiSeminar_12_6_2010.pdf.
- Spring 2012 – Presentation to the Epidemiology Seminar Group, Utah State University, February 6, 2012.
A Permutation Approach to Detecting Gene-Environment Interaction in Genetic Association Studies.
http://osastatistician.com/Welbourn_EpiSeminarPresentationGxE_2_6_2012.pdf.

PROFESSIONAL MEMBERSHIPS

- Utah Chapter of the American Statistical Association, Member
- Mathematical Association of America, Member
- Golden Key International Honour Society, Member
- The Honor Society of Phi Kappa Phi, Member

COMPUTER HARDWARE–SOFTWARE EXPERIENCE

- Building personal desktop computer systems. Current system specs:
 - Central Processing Unit: Intel Core i7 920 @ 2.66GHZ, Quad-Core, capable of handling eight (or more) simultaneous computational threads.
 - Random Access Memory: 3GB of third generation double data rate (DDR3) @ 1600MHZ.
 - Graphics Processing Units: 2 x NVIDIA GeForce GTX 470 1.2GB (448 processor cores; 625 MHz) in GPU cluster configuration.
 - Hard Drives: Four 250GB SATA II Seagate Baraccuda drives in RAID 0 configuration; Corsair 120GB solid state drive (SSD).
 - Operating System: Windows XP Home (32-bit).

- Operating Systems: Windows XP (32-bit), Windows 7 (62-bit), Beta tested Windows 7 RC.
- Languages: C, CUDA for C, R, SAS, Mathematica, L^AT_EX(including Beamer class and interfacing with R via Sweave), HTML, SQL (including interfacing with the C and R languages), MS-DOS.