STOR 455
STATISTICAL METHODS I

Jan Hannig

Final Exam

- Held Thursday, December 16 at 4pm – 6pm.
- Facts
  - Cumulative
  - Multiple choice
  - Slightly more emphasis on material after exam 2
- Closed book/notes. No computer.
- You can bring two REGULAR (letter) sheets of paper with formulas, etc.
  - Two sided
  - You must prepare the sheets yourself.

Final Exam

- Bring your calculator, #2 pencil, scantron sheet.
- Office hours
  - Today and tomorrow at usual times
  - Next week:
    - Wednesday and Thursday 10:30-11:30am
  - You can also post question to blackboard

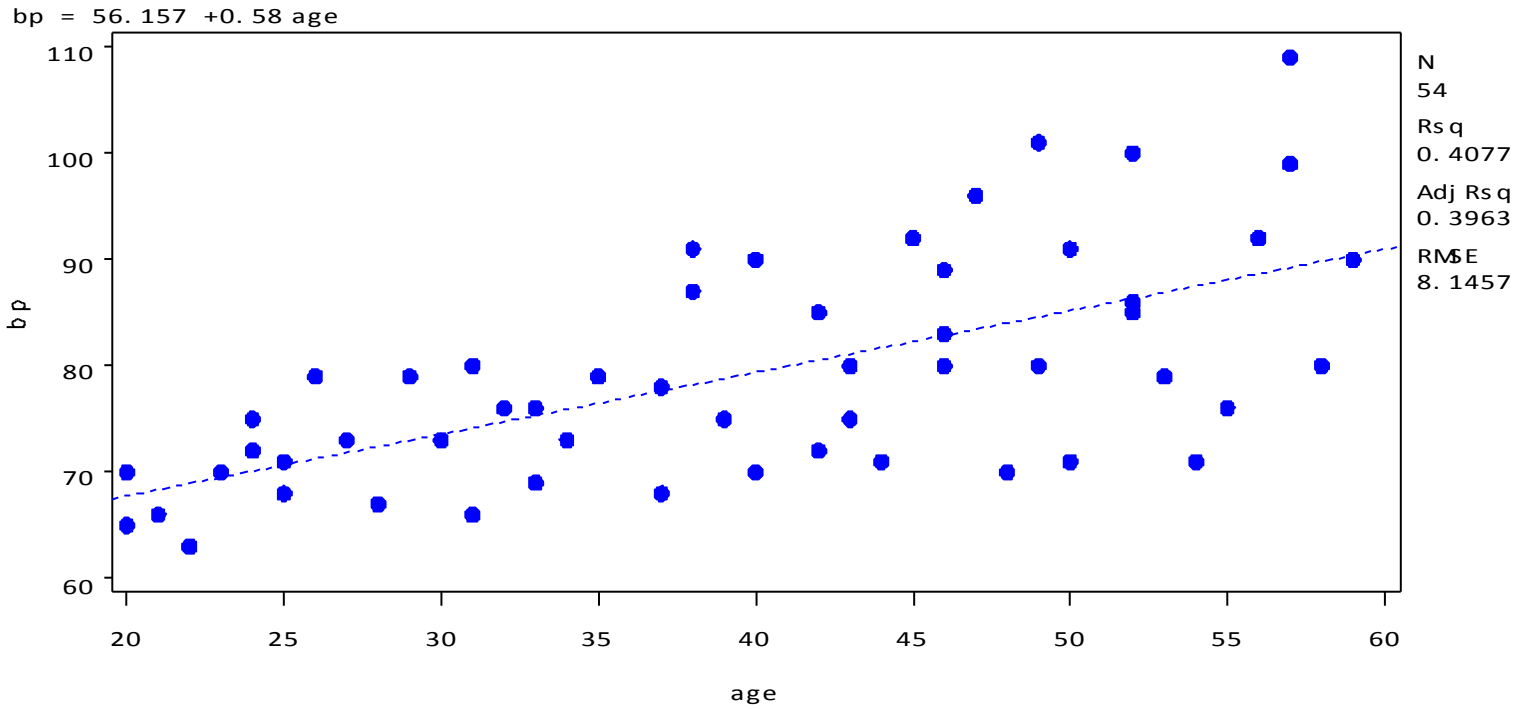
Grading Scheme

- Recall:
  - Homework sets – 20%
  - Midterm Exams (exam 1 – 20%, exam 2 – 20%, diamond project – 10%)
  - Final exam – 30%
- The multiple choice midterms are curved (I use the boxcox variance stabilizing transformation on the percentage)
- The letter grade will be computed from the total score using the usual cutoffs (approximately 90%A, 80%B,70%C,60%D,…)

Blood Pressure Example

- Y is diastolic blood pressure
- X is age
- n = 54 healthy adult women aged 20 to 60 years old
- Scatter plot show non-constant variance

Do it in SAS


bp = 56.157 +0.58 age
110
100
90
80
70
60
bp
20
25
30
35
40
45
50
55
60
age
N
54
Rsq
0.4077
Adj Rsq
0.3963
RMSE
8.1457

Weighted Least Squares (Section 8.2)

- Transformation may create other problems
- Generalize regression model: relax the assumption to allow different variances
- LS estimators still unbiased and consistent, but no longer have minimum variance
- WLS: minimize the the sum of weighted squared residuals

Weighted Least Squares

- OLS minimize

$$SSE = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{1,i} + \cdots + b_{p-1}x_{p-1,i})^2 = (Y - \mathbf{X}b)'(Y - \mathbf{X}b)$$

- WLS minimize

$$WSSE = \sum_{i=1}^{n} w_i(y_i - b_0 - b_1 x_{1,i} + \cdots + b_{p-1}x_{p-1,i})^2 = (Y - \mathbf{X}b)'W(Y - \mathbf{X}b)$$

$$W = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & w_n \end{pmatrix}$$

  - This gives $b_w = (X' WX)^{-1}(X' WY)$
  - Confidence intervals and tests are similar to before (see the book for formulas)
  - W and cW give the same results

Weighted Least Squares

- If the model is
  $Yi=\beta_0+b_1\, x_{1i} + \ldots +b_{p-1}\, x_{p-1i}+g_i\, \xi_i \quad \xi_i$ iid $N(0,\sigma^2)$
- Optimal weights: proportional to inverse variance $w_i=1/g_i^2$
- Often g_i themselves are related to the $\mathbf{x}_i$ and can be estimated from the residuals.

Determine the weights

- Method I: find a relationship between the absolute/ squared residuals and another variable and use this as a model for the standard deviation/variance
- Method II: use grouped data or approximately grouped data to estimate the variance
- Method III: use nonparametric method to estimate variance function
- Weights are proportional to the inverse of the estimated variance

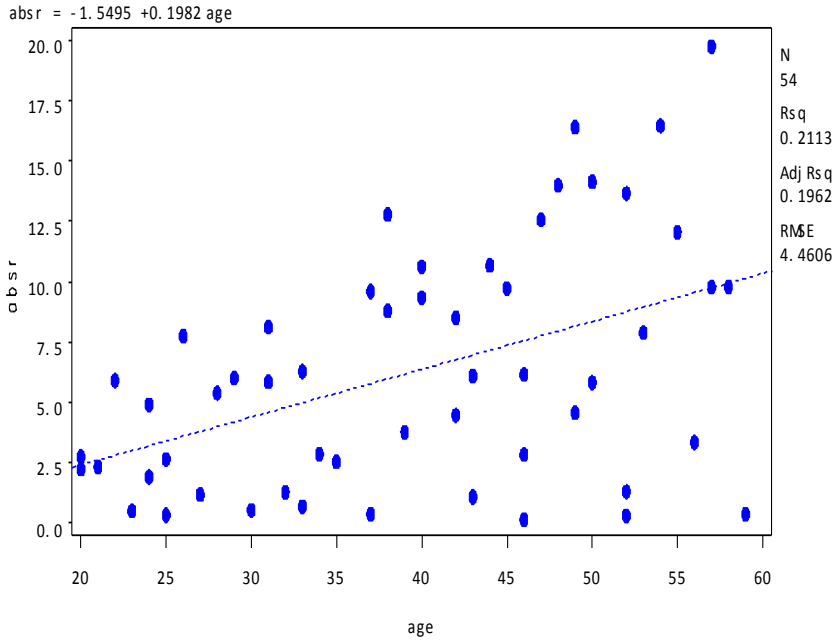
Do it in SAS

```
* Output residuals from
  proc reg;
proc reg data=dias;
  model bp = age /clb;
  output out=d1
  r=residual;
run;
* transform residuals;
data d1; set d1;
  absr = abs(residual);
run;
```

```
* estimate the s.d. using
  LS;
proc reg data = d1;
  model absr = age;
  output out = d2 p = s ;
run;

* Weights correspond to
  inverse variance;
data d2;
  set d2;
  w = 1/(s**2);
run;
```

Do it in SAS




```
*regression with
  weights;
proc reg data =
  d2;
  weight w;
     model bp =
  age / clb;
run;
```

Do it in SAS: OLS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2374.96833	2374.96833	35.79	<.0001
Error	52	3450.36501	66.35317		
Corrected Total	53	5825.33333			

Root MSE	8.14575	R-Square	0.4077
Dependent Mean	79.11111	Adj R-Sq	0.3963
Coeff Var	10.29659		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	56.15693	3.99367	14.06	<.0001	48.14304	64.17082
age	1	0.58003	0.09695	5.98	<.0001	0.38548	0.77458

Do it in SAS: WLS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	83.34082	83.34082	56.64	<.0001
Error	52	76.51351	1.47141		
Corrected Total	53	159.85432			

Root MSE	1.21302	R-Square	0.5214
Dependent Mean	73.55134	Adj R-Sq	0.5122
Coeff Var	1.64921		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	55.56577	2.52092	22.04	<.0001	50.50718	60.62436
age	1	0.59634	0.07924	7.53	<.0001	0.43734	0.75534

Review

- How to study?
  - Go over the lecture notes
  - Go over the homework & old exams
  - Make sure you really understand the project
- Book coverage
  - Chapters 1, 2, 3 ,4 (except 4.10), 5, 7 (except 7.5), 8 (except 8.3).
  - Additional topics not in the book (polynomial regression, interactions, boxcox)
- The book has a number of extra exercises in the last section of each chapter

Topics

- Fundamental concepts (population, sample, model, parameter, statistic, normal distribution, mean, standard deviation, correlation)
- Inferential procedures (point estimate - ubiased, confidence interval, hypothesis testing – p-value, bonferroni adjustment)
- Matrices (basic operations, inverse)

Review

- Regression (basic concepts, population, linear vs non-linear regression)
- Simple linear regression (model, least square estimation, prediction, CIs – parameters & predicted values, hypothesis tests, residual analysis – studentized residual, normal QQ plot, SAS)

Review

- Multiple linear regression (similar to SLR, additionally matrix calculations, ANOVA- sum of squares, extra sum of squares – type 1 and type 2, partial correlations, F tests, multicolinearity, interactions)
- Model Selection (R^2, adjusted R^2, C_p, $Press_p$; all subsets, forward backward stepwise; boxcox transformations)
- Lack of fit (DFFIT, DFBETA, Cook's D, studentized deleted residuals, influential observations, added variables plot)
- Weighted least squares

Evaluations

- UNC is using on-line evaluations for the first time this year!
- https://www.digitalmeasures.com/login/unc/user/authentication/authenticateShibboleth.do
- Happy evaluating!