Estimation of GLMs – Least Squares and Least Other Things

PSYC 943 (930): Fundamentals of Multivariate Modeling

Lecture 5: September 7, 2012

A Note About The Schedge…

- We have had a slight change in the schedule:
  - Today's class:
    - Wrapping up from Wednesday (putting our new found mathematical statistics knowledge to use with GLMs)
    - Estimation of GLMs using Least Squares
    - Various *other* forms of estimation using "least" somethings
  - Next Wednesday's class:
    - Maximum likelihood estimation of GLMs for continuous data (with normally distributed error terms)
  - Next Friday's class:
    - Introduction to generalized models (not normally distributed error terms); more maximum likelihood estimation (just with different distributions)

Today's Class

- An introduction to estimation…#wtftemplin
- Least squares estimation for GLMs
- Other "least" type estimators for GLMs
  - "Quantile"/median regression (GLMs)

Why Estimation is Important

- In "applied" statistics courses, estimation is not discussed very frequently
  - Can be very technical…very intimidating

- Estimation is of critical importance
  - Quality and validity of estimates (and of inferences made from them) depends on how they were obtained
  - New estimation methods appear from time to time and get widespread use without anyone asking whether or not they are any good

- Consider an absurd example:
  - I say the mean for IQ should be 20 – just from what I feel
  - Do you believe me? Do you feel like reporting this result?
    - Estimators need a basis in reality (in statistical theory)

How Estimation Works (More or Less)

- Most estimation routines do one of three things:

1. **Minimize Something:** Typically found with names that have "least" in the title. Forms of least squares include "Generalized", "Ordinary", "Weighted", "Diagonally Weighted", "WLSMV", and "Iteratively Reweighted." Typically the estimator of last resort…

2. **Maximize Something:** Typically found with names that have "maximum" in the title. Forms include "Maximum likelihood", "ML", "Residual Maximum Likelihood" (REML), "Robust ML". Typically the gold standard of estimators (and next week we'll see why).

3. **Use Simulation to Sample from Something:** more recent advances in simulation use resampling techniques. Names include "Bayesian Markov Chain Monte Carlo", "Gibbs Sampling", "Metropolis Hastings", "Metropolis Algorithm", and "Monte Carlo". Used for complex models where ML is not available or for methods where prior values are needed.

ESTIMATION OF GLMS USING LEAST SQUARES

Estimation of General Linear Models

- Recall our GLM (shown here for the prediction of a dependent variable Y_p by two independent variables X_p and Z_p):

$$Y_p = \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p$$

- Traditionally (dating to circa 1840), general linear models can be estimated via a process called least squares

- Least squares attempts to find the GLM parameters (the βs) that minimize the **squared residual** terms:

$$\min_{\{\beta_0, \beta_1, \beta_2, \beta_3\}} \left\{ \sum_{p=1}^{N} e_p^2 \right\}$$

Where We Are Going (and Why We Are Going There)

- Because the basics of estimation are critical to understanding the validity of the numbers you will use to make inferences from, we will detail the process of estimation
  - Today with Least Squares and then ending with Maximum Likelihood

- The LS estimation we will discuss is to get you to visualize functions of statistical parameters (the βs here) and data in order to show which estimates we would choose
  - To be repeated: In practice LS estimation for GLMs does not do this (by the magic of calculus and algebra)

- In the end, we would like for you to understand that not all estimators are created equally and that some can be trusted more than others
  - We would also like for you to see how estimation works so you can fix it when it goes wrong!

How Least Squares Estimation Works

- How Least Squares works is through minimizing the squared error terms…but its what goes into error that drives the process:

$$e_p = Y_p - \hat{Y}_p = Y_p - \left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p\right)$$

- If you were to do this (and you wouldn't), the process called optimization would go like this:

1. Pick values for regression slopes
2. Calculate $\hat{Y}_p$ and then e_p for each person p
3. Calculate $OF = \sum_{p=1}^{N} e_p^2$ (letters OF stand for **objective function**)
4. Repeat 1-3 until you find the values of regression slopes that lead to the smallest value of OF

Today's Example Data #1

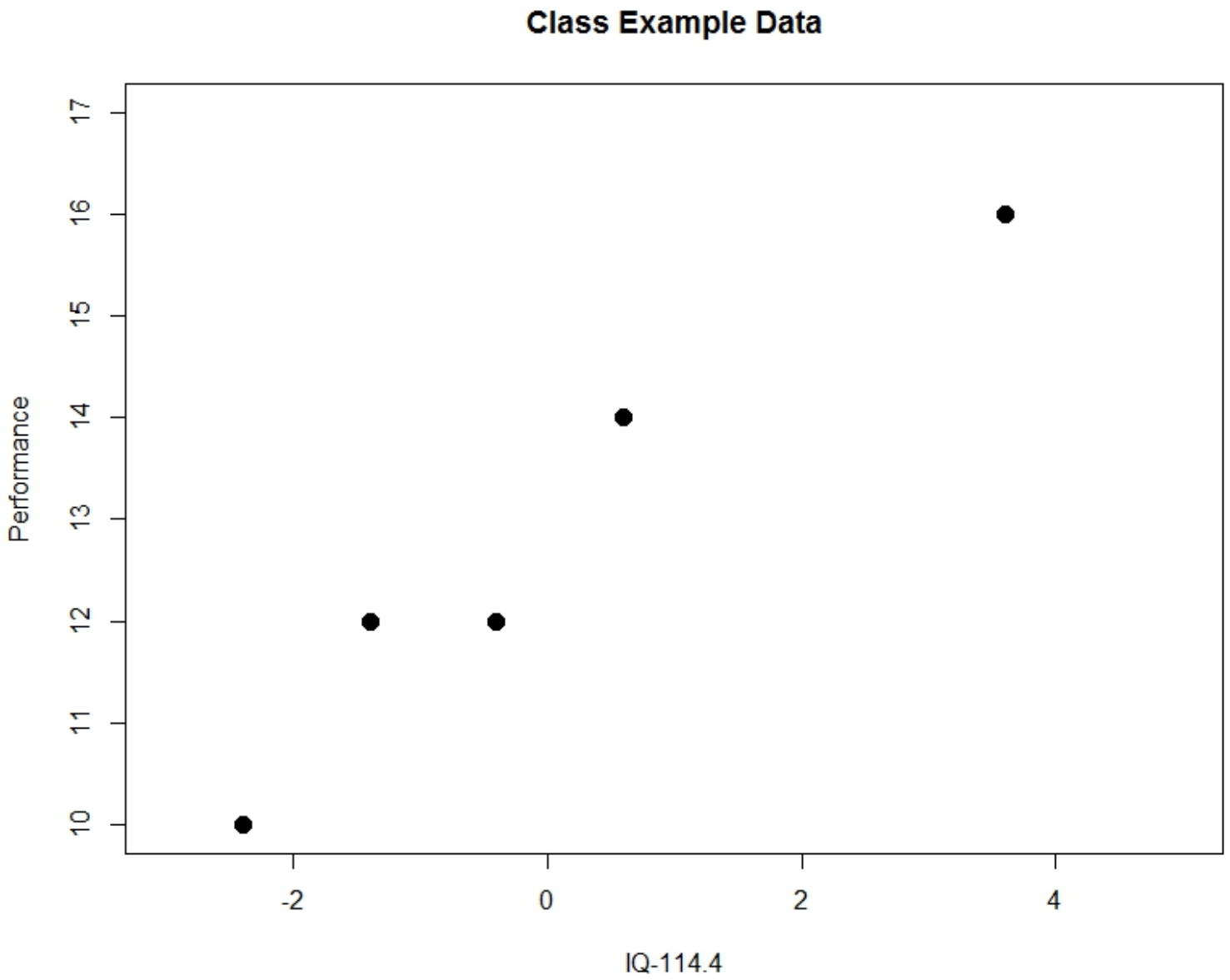
- Imagine an employer is looking to hire employees for a job where IQ is important
  - We will only use 5 observations so as to show the math behind the estimation calculations
- The employer collects two variables:
  - IQ scores
  - Job performance
- Descriptive Statistics:

Variable	Mean	SD
IQ	114.4	2.30
Performance	12.8	2.28

Covariance Matrix		
IQ	5.3	5.1
Performance	5.1	5.2

Observation	IQ	Performance
1	112	10
2	113	12
3	115	14
4	118	16
5	114	12

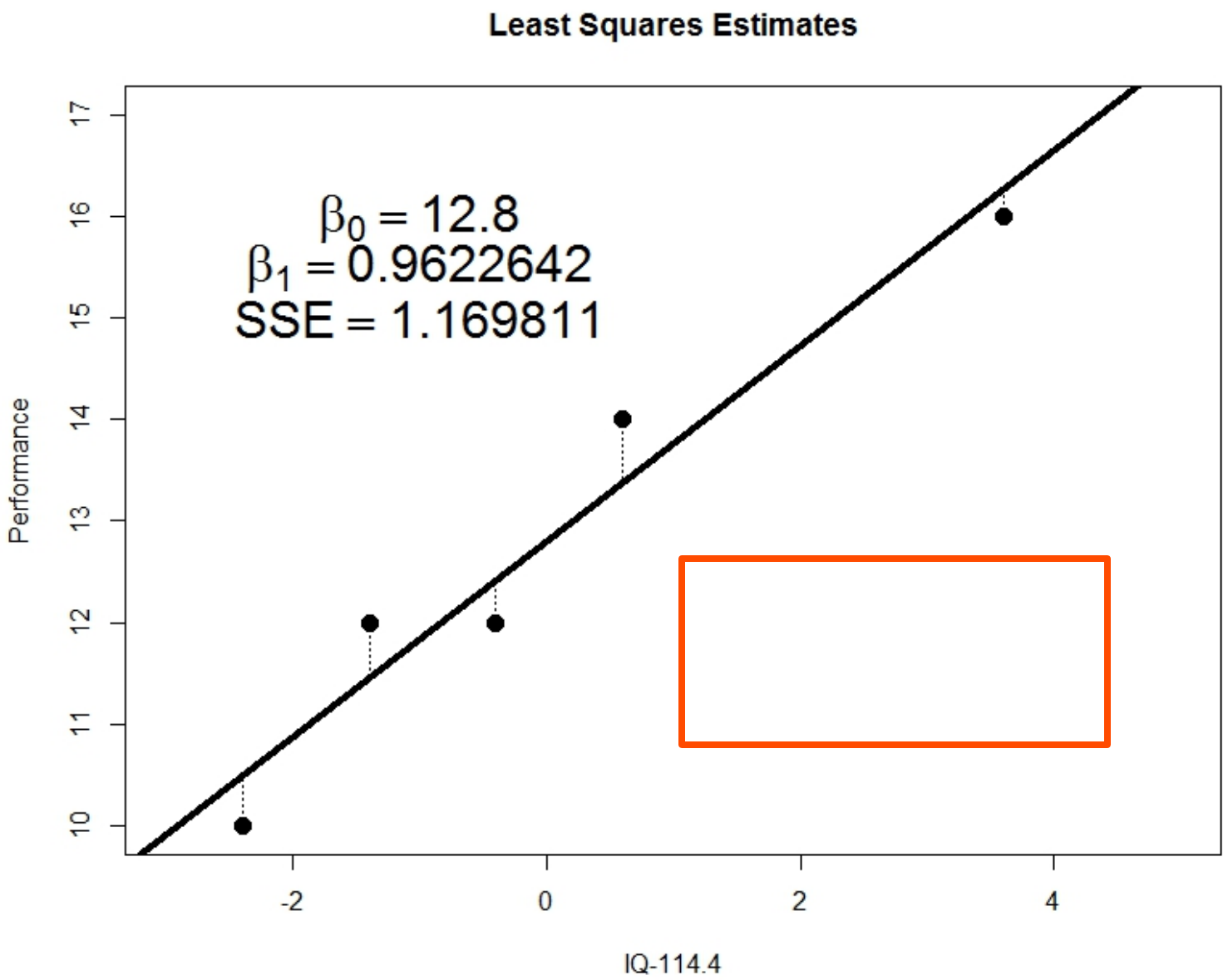
Visualizing the Data



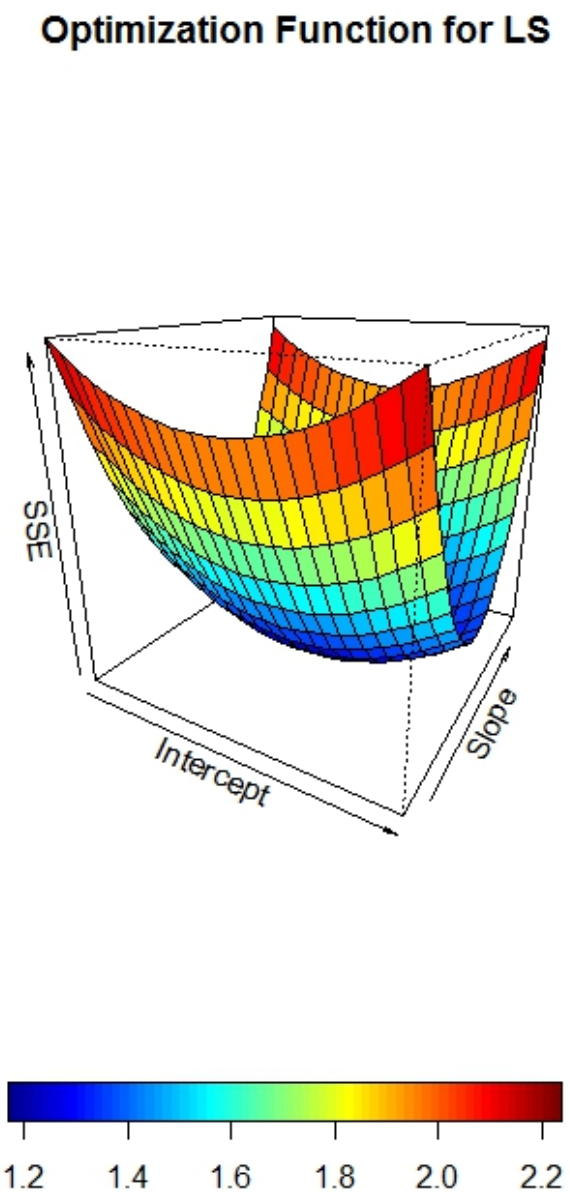
Let's Play…Pick the Parameters…

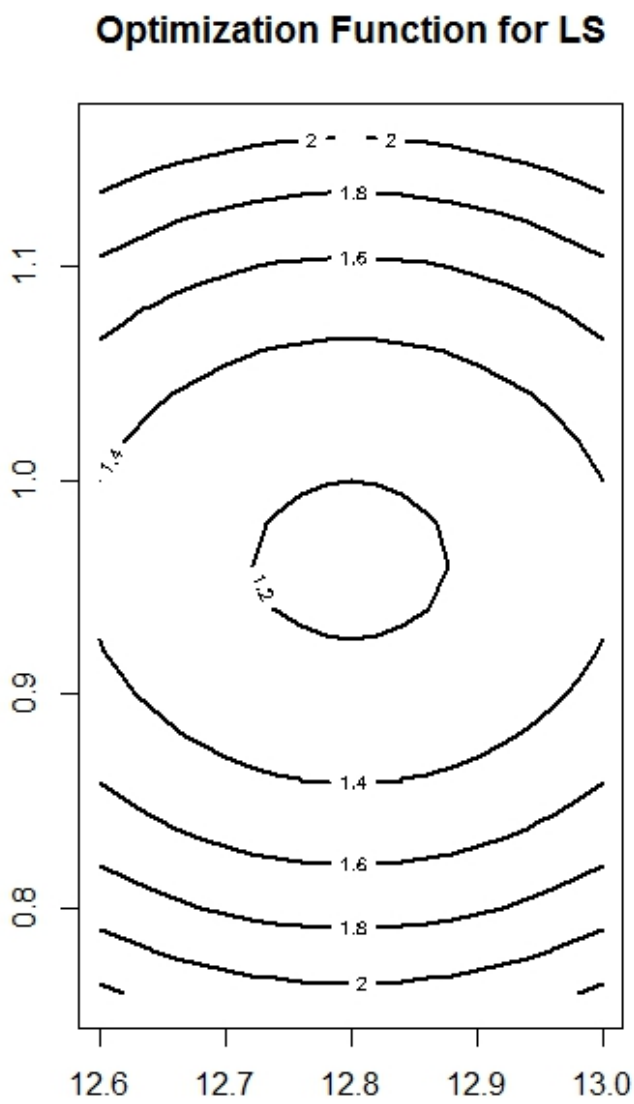
- This slide is left as a placeholder for the Camtasia recording – we will now do a demonstration in R

And…The Winner Is…



Examining the Objective Function Surface





LS Estimates of GLMs

- The process of least squares estimation of GLMs does not need an iterative search

- Using calculus, a minimum of the objective function can be found
  - This involves taking the first derivative of the objective function with respect to each parameter
    - Derivative = slope of the tangent line for a given point
  - The first derivative is then set equal to zero
    - Flat slope = minimum (or maximum or saddle point – neither apply here)
  - The equation is then solved for the parameter
    - Producing the equations you know and love

- For simple linear regression (one predictor):
  - $\beta_X = \frac{\frac{1}{\mathrm{N-k}}\sum_{p=1}^{N}(X_p-\bar{X})(Y_p-\bar{Y})}{\frac{1}{\mathrm{N-k}}\sum_{p=1}^{N}(X_p-\bar{X})(X_p-\bar{X})} = \frac{covariance\ of\ X\ and\ Y}{covariance\ of\ X\ and\ X} = \frac{sum\ of\ cross-products}{sum\ of\ squared\ Xs}$

- When we get to matrix algebra, you will know this as

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Ramifications of Closed-Form LS Equations

- Because least squares estimates have a **closed form** (equations that will provide statistics directly), they will work nearly every time
  - Only fail when collinearity is present (soon you'll know this to mean $\mathbf{X}^T\mathbf{X}$ is singular and cannot be inverted)

- Virtually all other estimators you will encounter in statistics will not have a closed form
  - Even least squares estimators for other types of data (not continuous)

- Without a closed form, least squares estimates are found by search the objective function for its minimum
  - Like finding the drain of a pool….

Why LS is Still in Use

- Least squares estimates still make up the bulk of GLM analyses because:
  - They are easy to compute
  - They pretty much always give you an answer
  - They have been shown to have good statistical properties

- The good statistical properties actually come because LS estimates of GLMs match the **Maximum Likelihood Estimates**
  - We will learn more about maximum likelihood estimation next week
  - For now, know that MLEs are the gold standard when it comes to estimates

Where LS Fails

- For all their flexibility, least squares estimates are somewhat limited
  - Only have good properties for basic univariate GLM for continuous data
    - Normally distributed error terms with homogeneous variance

- When data are not continuous/do not have normally distributed error terms, least squares estimates are not preferred

- For multivariate models with continuous data (repeated measures, longitudinal data, scales of any sort), least squares estimates quickly do not work
  - Cannot handle missing outcomes (deletes entire case)
  - Limited in the types of ways of modeling covariance between observations

OTHER TYPES OF "LEAST" ESTIMATORS: ROBUST GLMS

Other "Least" Estimators

- In order to expand your estimation horizons and to introduce a new statistical technique, let's consider what would happen if we were to make a seemingly small change to our least squares objective function

- Recall our GLM (shown here for the prediction of a dependent variable Y_p by two independent variables X_p and Z_p):

$$Y_p = \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p$$

- Least squares attempts to find the GLM parameters (the βs) that minimize the **squared residual** terms:

$$\min_{\{\beta_0, \beta_1, \beta_2, \beta_3\}} \left\{ \sum_{p=1}^{N} e_p^2 \right\}$$

- Instead of using the squared residual terms, we find the parameters that minimize the **absolute residual** terms:

$$\min_{\{\beta_0, \beta_1, \beta_2, \beta_3\}} \left\{ \sum_{p=1}^{N} |e_p| \right\}$$

Minimizing Absolute Residuals

- The use of absolute residuals changes several properties of the estimates
  - Often this is called **median regression** or **quantile regression (at quantile .5)**
- The estimator itself is called an L_1 or robust estimator
  - The least squares estimator is called an L_2 estimator
- The regression line now includes the median of X and Y
- The regression line is now "robust" to outliers
  - Think median instead of mode
- More generally, any quantile τ (from 0-1, like %) can be specified:

$$\min_{\{\beta_0,\beta_1,\beta_2,\beta_3\}} \left\{ \sum_{i\in\{i:y_i\geq\hat{y}_i\}} \tau|e_p| + \sum_{i\in\{i:y_i<\hat{y}_i\}} (1-\tau)|e_p| \right\}$$

Quantile Regression in SAS: PROC QUANTREG

- Want to do quantile regression? There's a PROC for that…

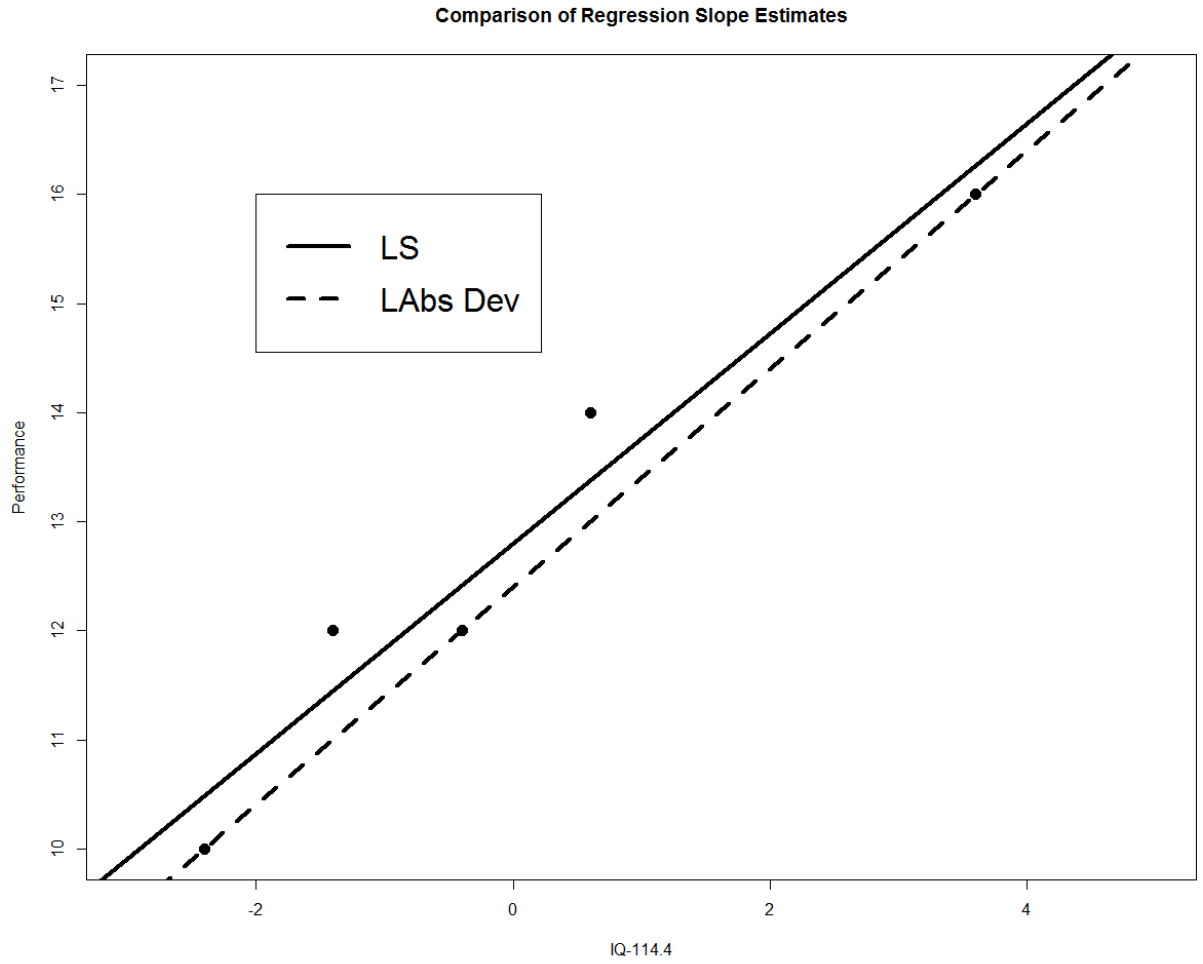
```
DATA work.exampledata;
INPUT perf iq;
iqMC = iq - 114.4;
DATALINES;
10 112
12 113
14 115
16 118
12 114
;
RUN;

PROC QUANTREG DATA=work.exampledata CI=RESAMPLING;
    MODEL perf = iqMC / QUANTILE = .5 SEED=7;
RUN;
```

- Note: MODEL statement is identical to PROC GLM before "/"
  - QUANTILE = .5 represents the median (trend for the median)
  - SEED = 7 represents the random number seed for resampling

Estimated Value from SAS/Comparison with LS Estimates

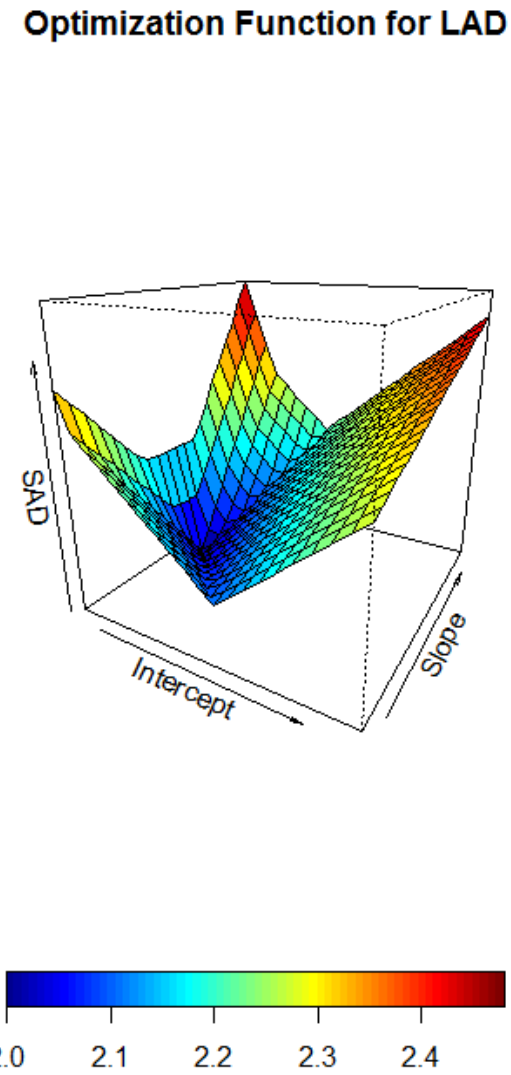
Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**t Value**	**Pr > |t|**
Intercept	1	12.4	1.5208	7.5603	17.2397	8.15	0.0039
iqMC	1	1	0.8863	-1.8207	3.8207	1.13	0.3413

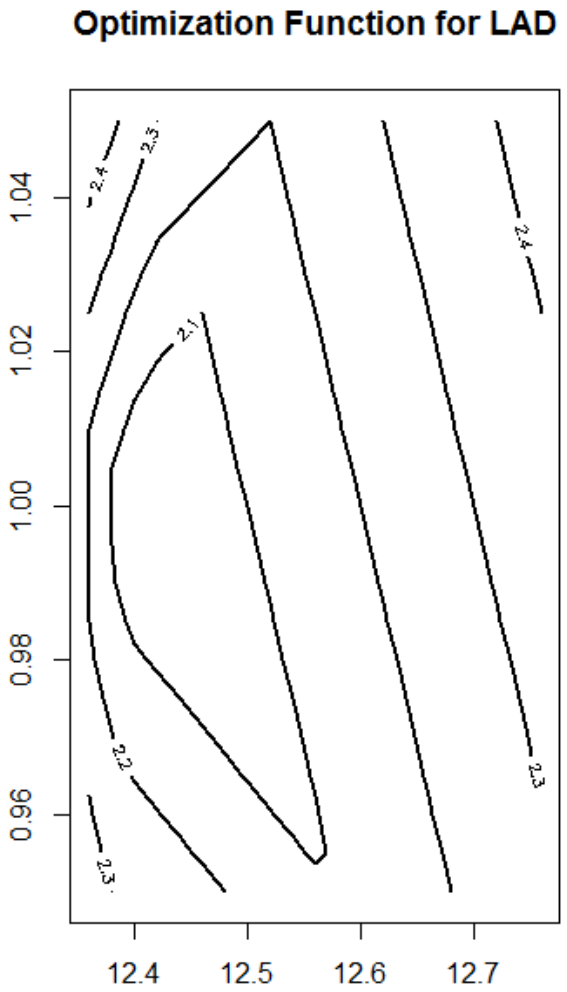


Objective Function Image

- Here, the objective function is:

$$OF = \sum_{i=1}^{N} |e_p|$$



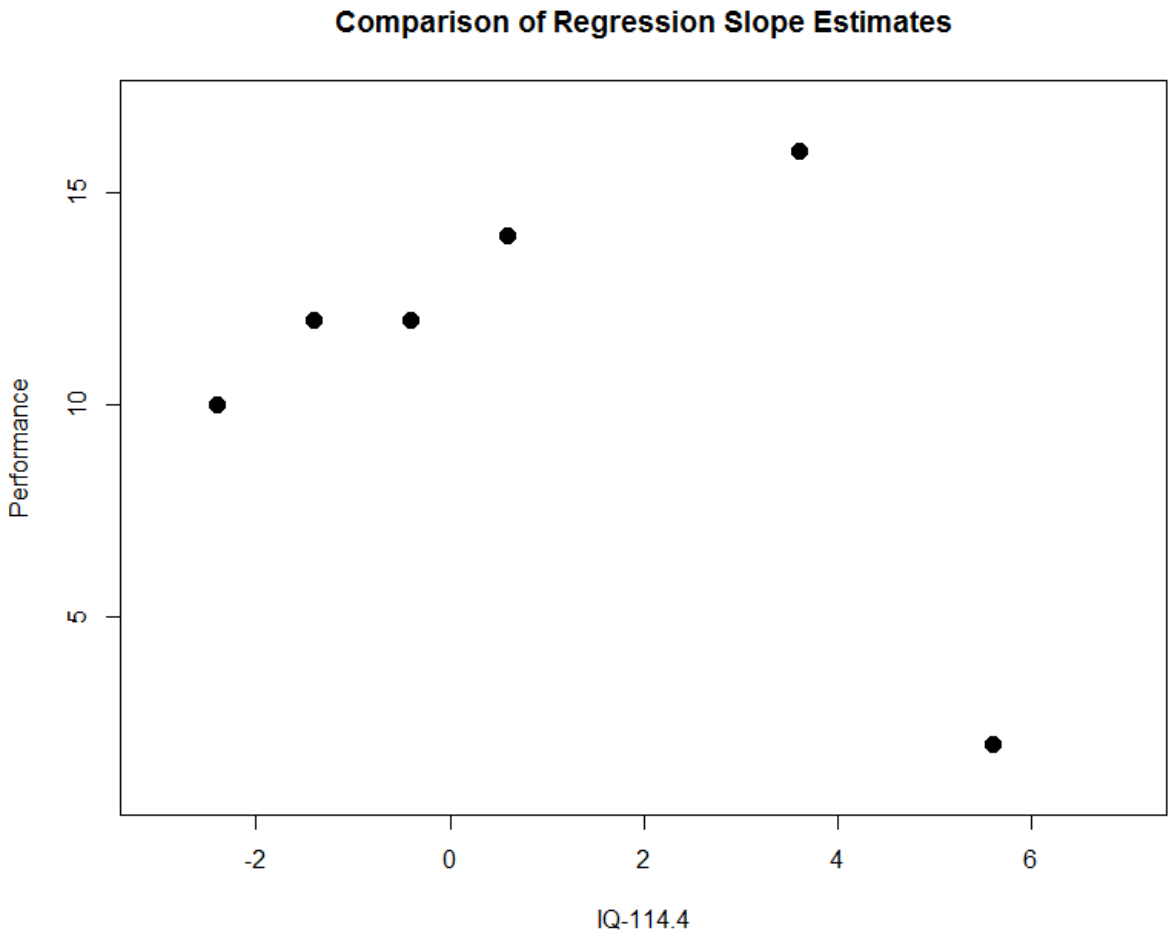


Mathematical Issues with Quantile Regression

- Because the regression function features an absolute value, the process of estimation gets complicated
  - Minimum point is not differentiable – so calculus can't be used
- The L_1 estimator does not have a closed form (no equations)
- SAS (or other software) uses a search algorithm to find the GLM slopes that minimize the objective function
  - This process can take a long time
- The statistical properties of the slopes are not well known
  - Resampling is used to calculate standard errors
    - Resampling: running multiple analyses with a random set of the same data
  - If random number seed differs, results will differ…

Where Quantile Regression Helps…Outliers

- Imagine we had observed a 6th case: a person with an IQ of 120 and a performance of 2



Comparing LS and Quantile Regression Results

- Least Squares WITHOUT Outlier:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	12.8	0.27926228	45.84	<.0001
iqMC	0.96226415	0.13562175	7.1	0.0058

- Least Squares WITH Outlier:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.5915493	2.14005638	5.42	0.0056
iqMC	-0.63380282	0.72305804	-0.88	0.4302

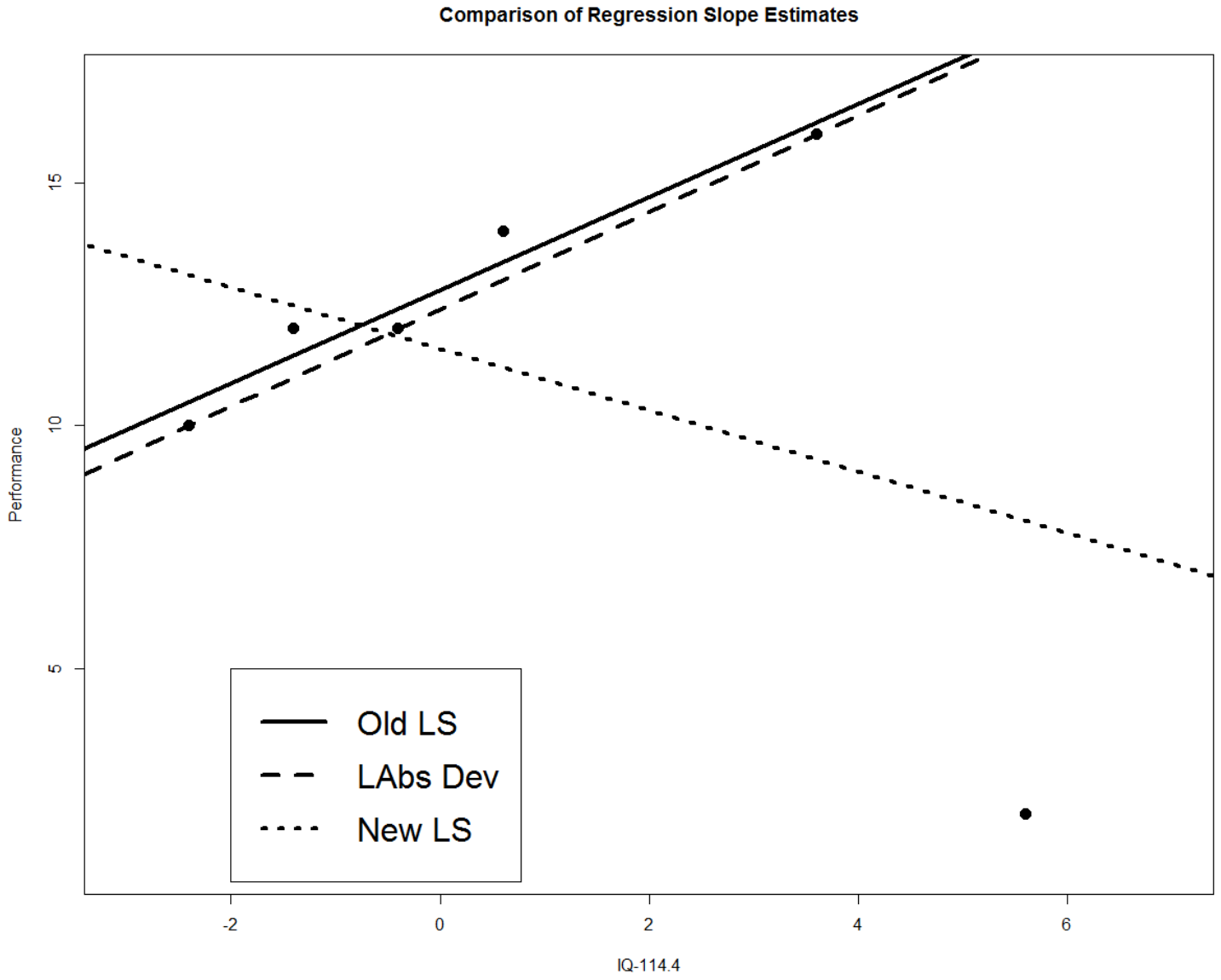
- Quantile Regression WITHOUT Outlier:

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	12.4	1.5208	7.5603	17.2397	8.15	0.0039
iqMC	1	1	0.8863	-1.8207	3.8207	1.13	0.3413

- Quantile Regression WITH Outlier:

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	12.4	22.3349	-49.6116	74.4116	0.56	0.6084
iqMC	1	1	8.1904	-21.7401	23.7401	0.12	0.9087

Graphical Comparison



QUANTILE REGRESSION EXAMPLE

Using Quantile Regression

- Quantile regression is a useful research tool for:
  - When data are skewed
  - Influential (potentially outlying) observations are present
  - Interactions between your IVs and your DV

- Quantile regression (as in PROC QUANTREG) cannot help:
  - Dependency within or between cases
  - Non-constant variance of residual terms

Data Example #3: Change Detection Speed

- To demonstrate quantile regression, we will use data where:
  - Y = Reaction Time (in seconds) to detect the change between two otherwise same pictures (mean across 60ish trials)
  - Age65 = Age in years, centered at 65
  - Nearvis_4 = near vision in logarithmic units (center point is 20/20 vision = .4)
    - Higher scores = worse vision
- Our goal: to predict reaction time across all quantiles to see if age, near vision, or their interaction have an effect
  - In aging literature there is a debate about where slowing in reaction time happens: whether it is a general shift or it is more spread out
- Our process: we will first fit (estimate) a least squares GLM and then use the quantile approach to investigate these data

First: Least Squares Estimates

- We first analyze these data using PROC GLM:

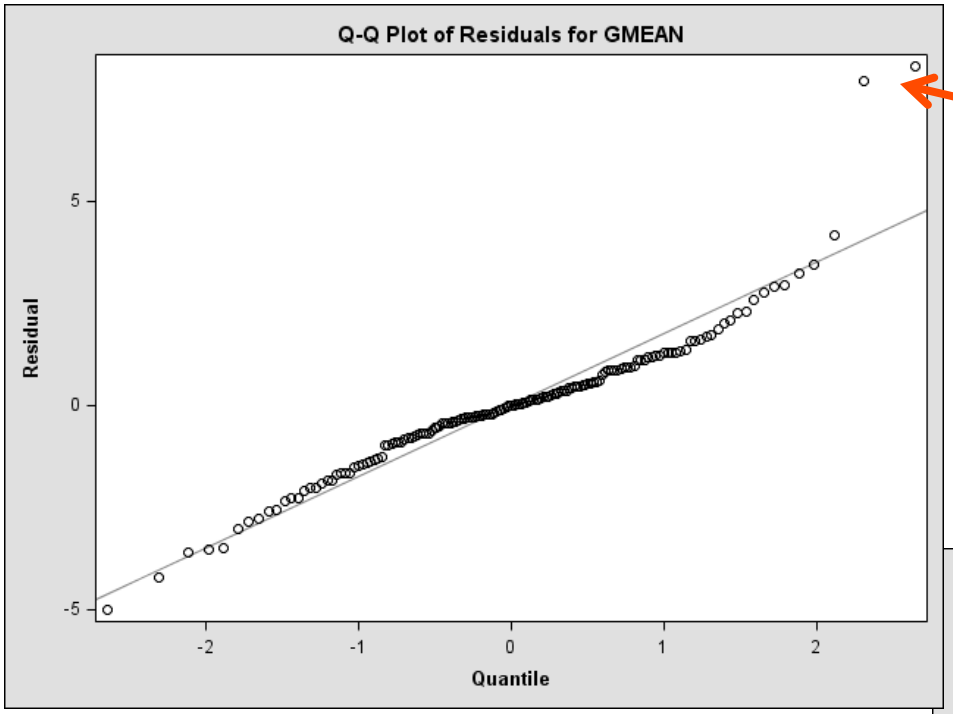
```
PROC GLM DATA=work.lesama PLOTS=(ALL DIAGNOSTICS(UNPACK));
MODEL gmean = age65 nearvis_4 age65*nearvis_4 / SOLUTION;
RUN;
```

- Here were our results:

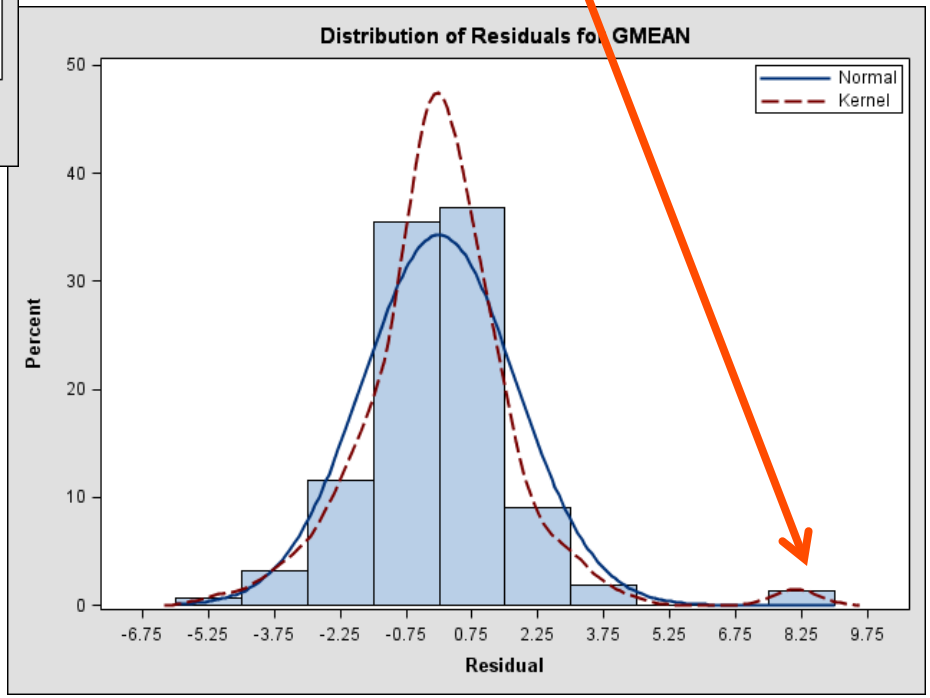
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	12.45918412	0.30396622	40.99	<.0001
age65	0.12193431	0.00745474	16.36	<.0001
nearvis_4	1.35334485	0.89722472	1.51	0.1335
age65*nearvis_4	0.04905639	0.04234832	1.16	0.2485

- From this we *would* conclude that:
  - There was no age by near vision interaction
  - There was no simple **(conditional)** near vision main effect
  - As age increased, the **conditional mean** response time also increased

But…What About *THESE GUYS* (in residuals)



THESE GUYS

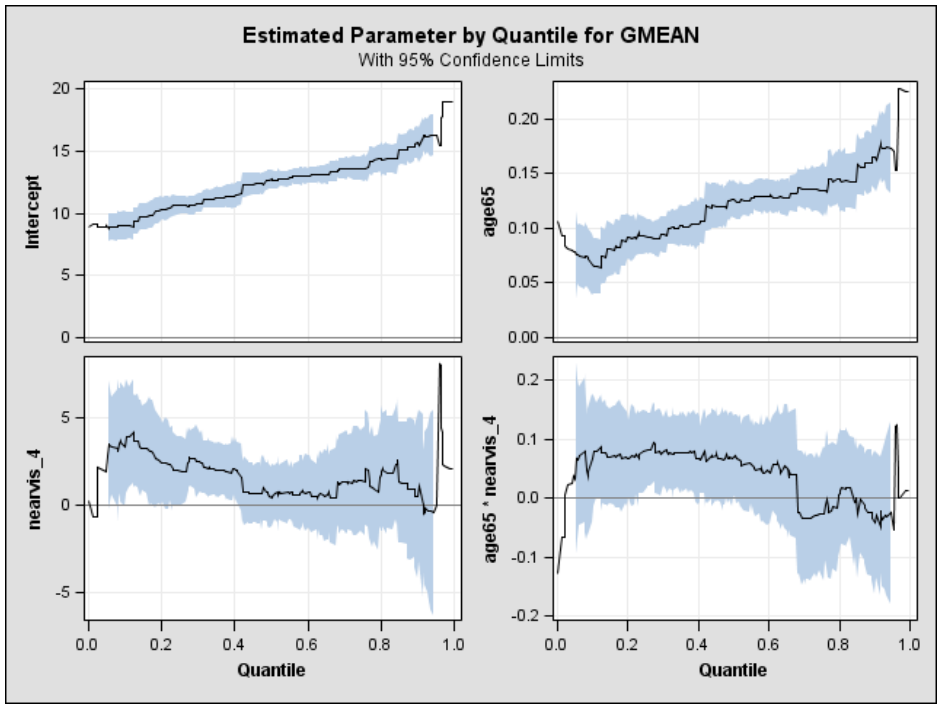


From Least Squares to Quantiles

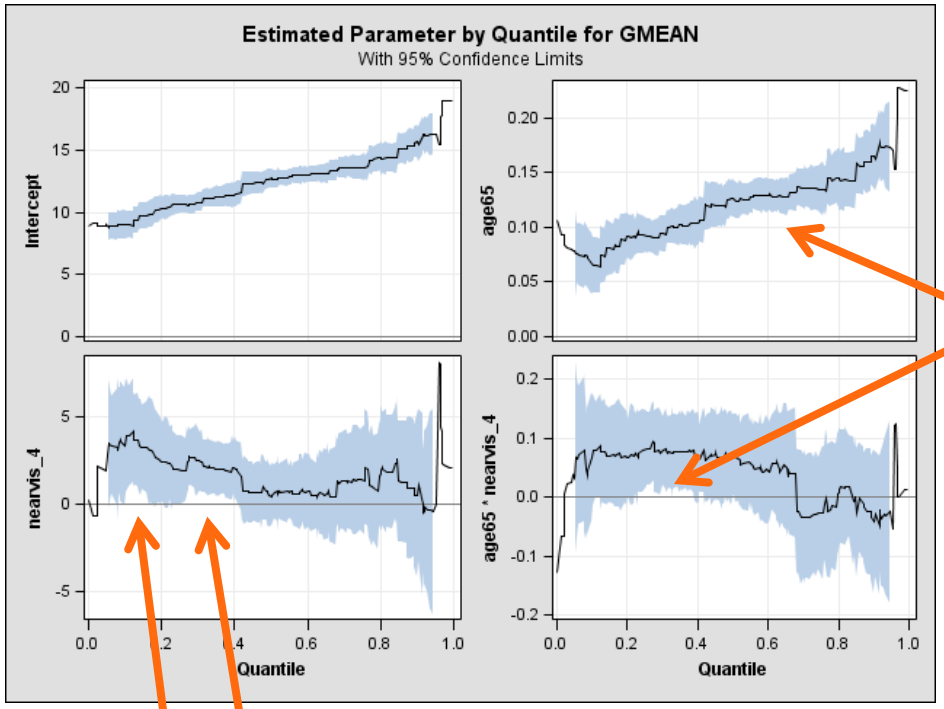
- Next, we will subject these data to a quantile analysis using PROC QUANTREG

```
PROC QUANTREG DATA=work.lesama PLOTS=(ALL) CI=RESAMPLING;
MODEL gmean = age65 nearvis_4 age65*nearvis_4 /QUANTILE = PROCESS SEED=8675309 PLOT=QUANTPLOT;
RUN;
```

  - QUANTILE = PROCESS literally tries every quantile possible for these data
  - SEED = 8675309 http://www.youtube.com/watch?v=FkpGQUflBwU
    - Keeps the CIs the same



Interpreting Quantile Regression Results



Interaction is significant for quantiles between .25ish and .4

Main Effect of Age: Significant across all quantiles

- BUT INTERPRETATION CHANGES DEPENDING ON SIGNIFICANCE OF INTERACTION

Main Effect of Near Vision:
Significant across all some lower quantiles

- BUT INTERPRETATION CHANGES DEPENDING ON SIGNIFICANCE OF INTERACTION

A Simple Description:
http://www.youtube.com/watch?v=Ek0SgwWmF9w

Upon Further Review: For Specific Quantiles

- PROC QUANTREG can give specific parameter estimates for any quantile, allowing for the same linear model interpretation for any part of the **conditional distribution of the response variable (DV)**
  - Interactions and main effects run the world

```
PROC QUANTREG DATA=work.lesama PLOTS=(ALL) CI=RESAMPLING;
MODEL gmean = age65 nearvis_4 age65*nearvis_4 /QUANTILE = 0.25 0.5 0.75 SEED=8675309 PLOT=QUANTPLOT;
RUN;
```

- We will inspect the results from quantiles .25, .50, and .75

Results from Quantile .25 (FAST PEOPLE)

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	10.6233	0.3836	9.8654	11.3812	27.69	<.0001
age65	1	0.0929	0.0092	0.0747	0.1111	10.10	<.0001
nearvis_4	1	1.9850	1.0753	-0.1396	4.1096	1.85	0.0669
age65*nearvis_4	1	0.0789	0.0369	0.0061	0.1518	2.14	0.0340

- Intercept: predicted value for 25th percentile of reaction time for when age is 65 and near vision is perfect
- Age65: increase in predicted value for 25th percentile of reaction time for every year of age **when near vision is perfect**
- Nearvis_4: increase in predicted value of 25th percentile of reaction time for every one-unit decrease in near vision **when age is 65**
- Age65*Nearvis_4: increase in the effect of age on the 25th percentile per unit increase in near vision **–or–** increase in the effect of near vision per year increase in age

Results from Quantile .5 (MIDDLE SPEED PEOPLE)

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	12.6823	0.3620	11.9671	13.3974	35.04	<.0001
age65	1	0.1264	0.0083	0.1100	0.1427	15.25	<.0001
nearvis_4	1	0.4893	0.8563	-1.2025	2.1811	0.57	0.5685
age65*nearvis_4	1	0.0665	0.0380	-0.0086	0.1416	1.75	0.0822

- Intercept: predicted value for 50th percentile of reaction time for when age is 65 and near vision is perfect
- Age65: increase in predicted value for 50th percentile of reaction time for every year of age **when near vision is perfect**
- Nearvis_4: increase in predicted value of 50th percentile of reaction time for every one-unit decrease in near vision **when age is 65**
- Age65*Nearvis_4: increase in the effect of age on the 50th percentile per unit increase in near vision **–or–** increase in the effect of near vision per year increase in age

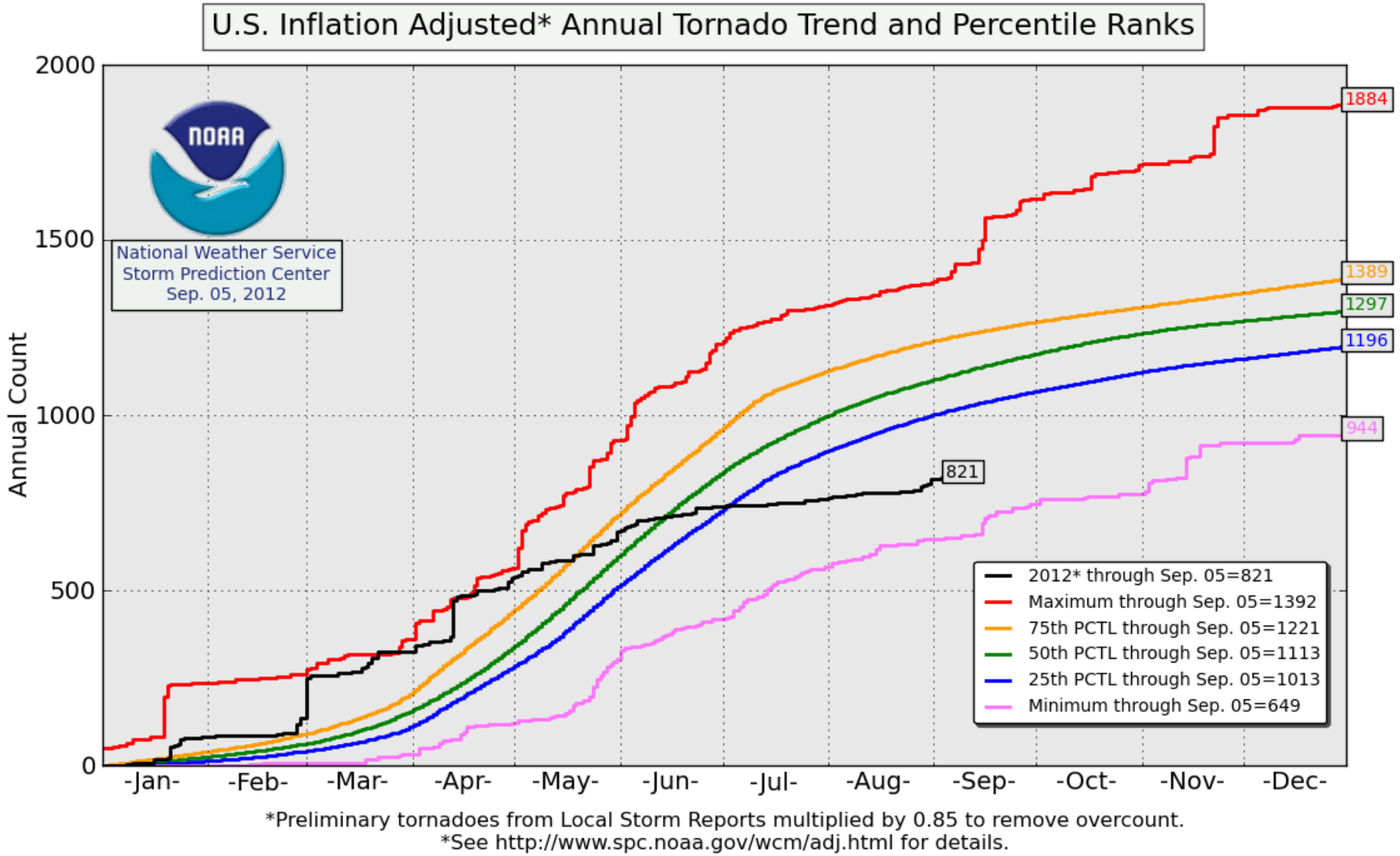
Results from Quantile .75 (SLOW PEOPLE)

Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	13.6076	0.5483	12.5243	14.6909	24.82	<.0001
age65	1	0.1351	0.0114	0.1125	0.1577	11.83	<.0001
nearvis_4	1	1.3750	1.5973	-1.7810	4.5309	0.86	0.3907
age65*nearvis_4	1	-0.0275	0.0560	-0.1382	0.0831	-0.49	0.6239

- Intercept: predicted value for 75th percentile of reaction time for when age is 65 and near vision is perfect
- Age65: increase in predicted value for 75th percentile of reaction time for every year of age **when near vision is perfect**
- Nearvis_4: increase in predicted value of 75th percentile of reaction time for every one-unit decrease in near vision **when age is 65**
- Age65*Nearvis_4: increase in the effect of age on the 75th percentile per unit increase in near vision **–or–** increase in the effect of near vision per year increase in age

QUANTILE REGRESSION IN OTHER FIELDS

Quantile Regression Use in Other Fields: Weather



http://www.spc.noaa.gov/wcm/adj.html

Quantile Regression Use in Other Fields: Salaries

WRAPPING UP

Wrapping Up

- Today discussed estimation, and in the process showed how differing estimators can give you different statistics

- The key today was to shake your statistical view point:
  - There are many more ways to arrive at statistical results than you may know

- The take home point is that **not all estimators are created equal**
  - If ever presented with estimates: ask how the numbers were attained
  - If ever getting estimates: get the best you can with your data

- Next week your world will further be expanded when we introduce maximum likelihood estimators