Prediction and Explanation

November 1, 2007

ERSH 8320

Today's Lecture

- Prediction from regression.

Today's Example Data Set

From Weisberg (1985, p. 240).

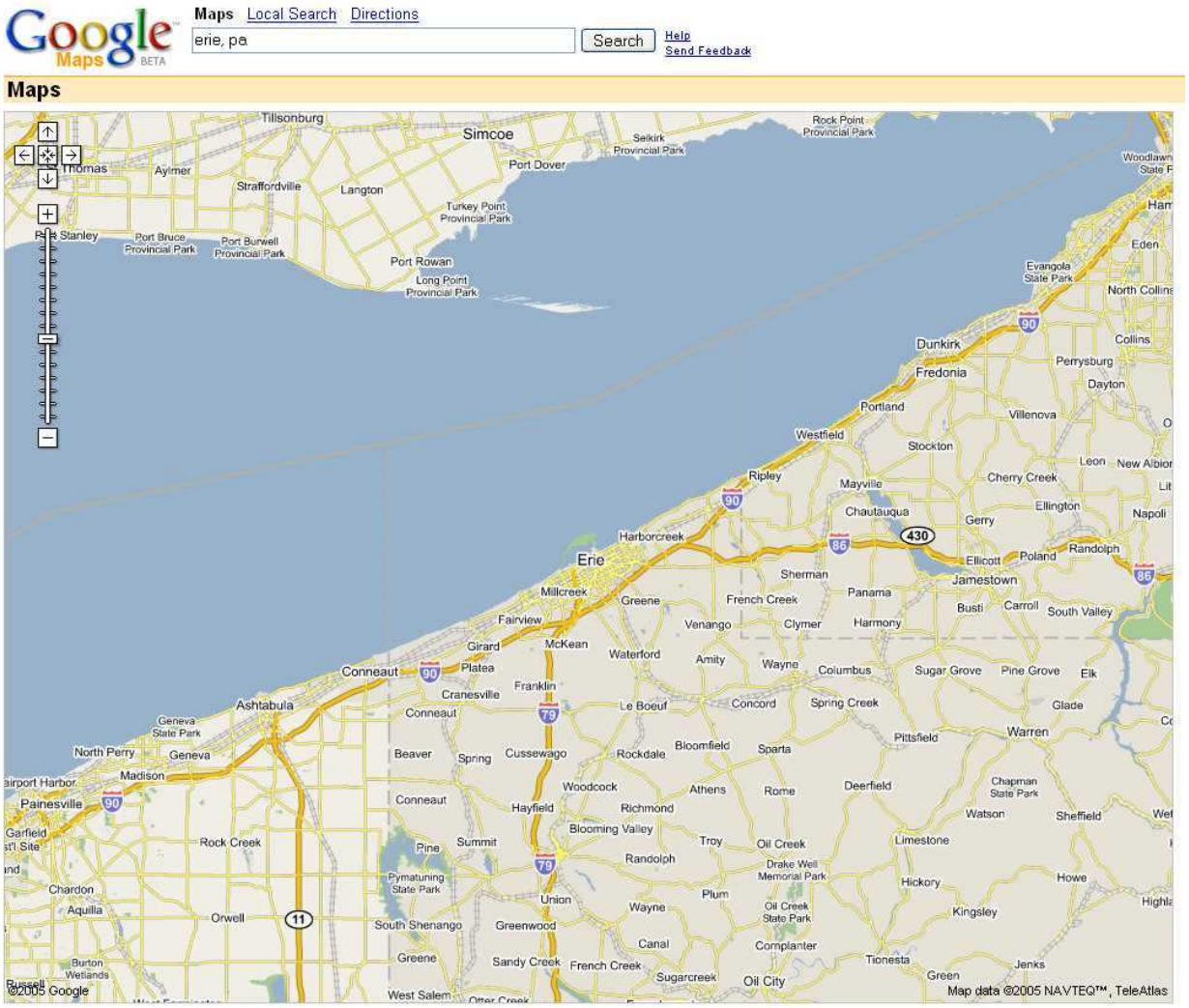
"Property taxes on a house are supposedly dependent on the current market value of the house. Since houses actually sell only rarely, the sale price of each house must be estimated every year when property taxes are set. Regression methods are sometimes used to make up a prediction function."

Erie, Pennsylvania

We have data for 27 houses sold in the mid 1970's in Erie, Pennsylvania:

- X_1: Current taxes (local, school, and county) $\div$ 100 (dollars).
- X_2: Number of bathrooms.
- X_3: Lot size $\div$ 1000 (square feet).
- X_4: Living space $\div$ 1000 (square feet).
- X_5: Number of garage spaces.
- X_6: Number of rooms.
- X_7: Number of bedrooms.
- X_8: Age of house (years).
- X_9: Number of fireplaces.
- Y: Actual sale price $\div$ 1000 (dollars).

Erie, Pennsylvania



Lake Erie

Erie, Pennsylvania

Jack

Erie, Pennsylvania

Paula

Prediction

- To make things easy, let's begin with trying to predict a home's sale price based on a single X variable: living space.

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.929[a]	.863	.857	5.4061

a. Predictors: (Constant), living space

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4591.667	1	4591.667	157.108	.000[a]
	Residual	730.653	25	29.226		
	Total	5322.320	26			

a. Predictors: (Constant), living space

b. Dependent Variable: sale price in thousands

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.506	3.054		.820	.420
	living space	23.804	1.899	.929	12.534	.000

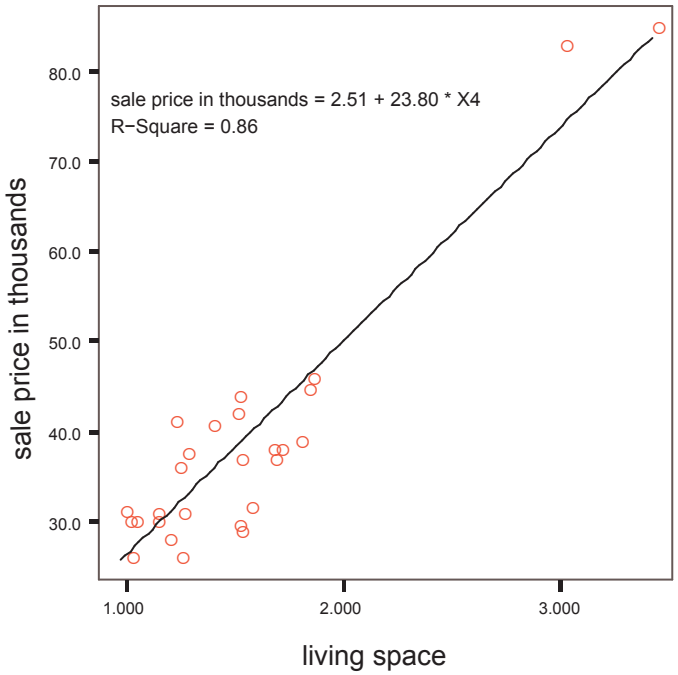
a. Dependent Variable: sale price in thousands

Predicted Value: Y'

- From the results of the analysis, for any size of X (any living space size), we will predict the sale price of the house to be:

$$Y' = 2.506 + 23.804X$$

- But knowing the predicted value is just the beginning...



What Does the Predicted Value Represent?

- The predicted value given by a regression equation is the *expected value* of Y conditional on X.
- This means that for every value of X, the mean value $[E(Y|X)]$ is given by Y'.
- Recall from basic statistics that if a variable Z is distributed with a mean μ_Z and variance σ_Z^2 then:
  - As sample size, N, goes to infinity, the distribution of $\bar{Z}$ is $N(\mu_Z, \frac{\sigma_Z^2}{N})$ (Central Limit Theorem).
  - Using the CLT, we are able to build a confidence bound around our sample mean so that we can be $100(\alpha/2)\%$ certain that the true μ_Z lies in the interval.
- The same principal can be applied to Y', which is the mean of the distribution of $Y|X$.

Confidence Interval for Mean Predicted Value

- A confidence interval for the mean predicted value of a regression is given by:

$$Y' \pm t_{(\alpha/2,df)} s_{\mu'},$$

where:

$$s_{\mu'} = \sqrt{s^2_{y.x} \left[\frac{1}{N} + \frac{(X_i - \bar{X})^2}{\Sigma_{i=1}^{N} (X_i - \bar{X})^2} \right]}$$

- $s^2_{y.x}$ is variance of estimate, or mean squared residual (also known as conditional variance of y given x).

- $t_{(\alpha/2,df)}$ is the value of the t distribution corresponding the the α degree of confidence, with $df = N - k - 1$, the degrees of freedom for the residuals.

Predicted Value Example

- Imagine a set of houses, all with 2,000 square feet of living area.
- Using the estimated regression parameters, we could predict that the average sale price would be:

$$Y' = 2.506 + 23.804X = 2.506 + 23.804(2) = 50.114$$

- Using a 95% degree of confidence, the confidence interval for this value is found from:

$$Y' \pm t_{(\alpha/2,df)} s_{\mu'}$$

$$t_{(95\%/2,25)} = 2.06$$

$$s_{\mu'} = \sqrt{s_{y.x}^2 \left[\frac{1}{N} + \frac{(X_i - \bar{X})^2}{\Sigma_{i=1}^{N} (X_i - \bar{X})^2} \right]}$$

Predicted Value Example

$$\bar{X} = 1.530$$

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = 8.333$$

$$s^2_{y.x} = MS_{res} = 29.226$$

$$s_{\mu'} = \sqrt{29.226\left[\frac{1}{27} + \frac{(2.000 - 1.530)^2}{8.333}\right]} = 1.363$$

$$50.114 \pm 2.06 \times 1.363 \rightarrow (47.245, 52.984)$$

So, we are 95% confident that the mean sale price for a 2,000 square foot home would fall between $47,245 and $52,984.

Mean Predicted Value Interval Properties

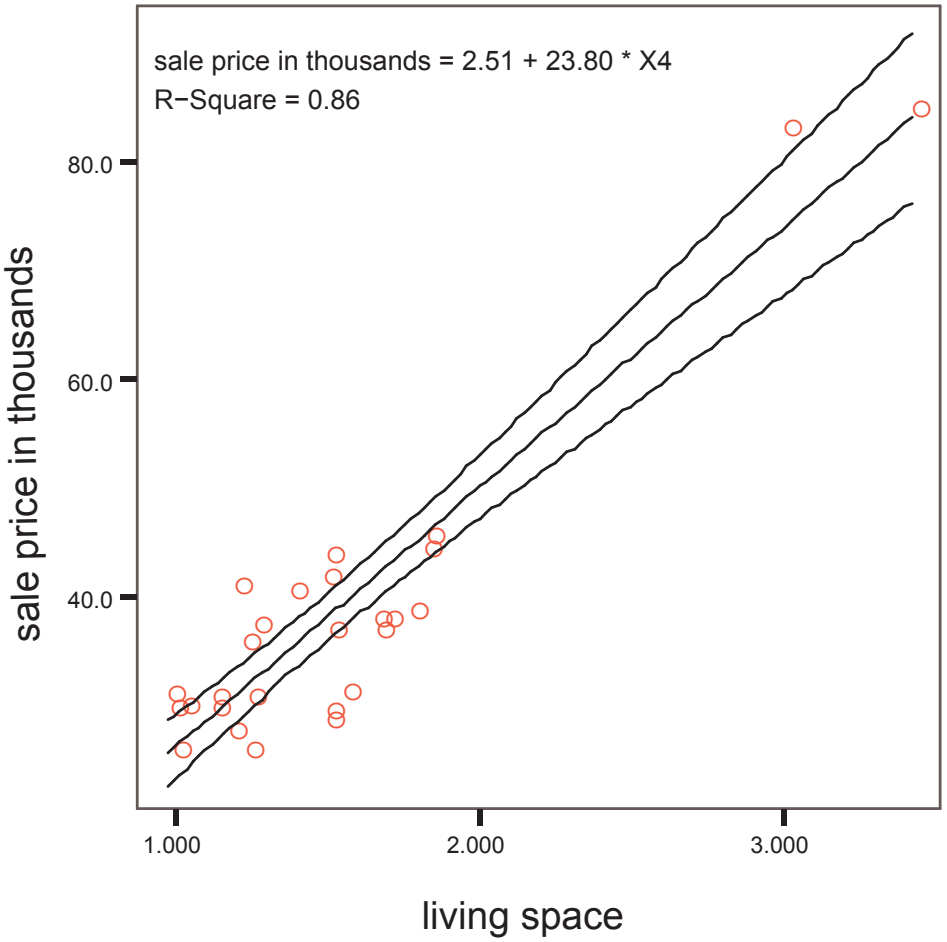
- Notice that $s_{\mu'}$ has $(X_i - \bar{X})^2$ in its equation:

$$s_{\mu'} = \sqrt{s^2_{y.x}\left[\frac{1}{N} + \frac{(X_i - \bar{X})^2}{\Sigma_{i=1}^{N}(X_i - \bar{X})^2}\right]}$$

- This means that as the value of X gets farther from its mean, the prediction interval increases in range.

- Using SPSS you can plot the mean prediction interval from Graphs...Interactive...Scatterplot. Be sure to check Mean under the Prediction Lines of the Fit tab.

Graph of the Mean Predicted Value Interval



Linear Regression with
95.00% Mean Prediction Interval

CI for Single Predicted Values

- Often times, the mean value is not of interest in prediction.
- Imagine you are a homeowner in Erie, PA. Because of all the weirdos fishing by your house, you decide to sell and move to Youngstown, OH. Your house has 2,000 square feet.
- You want to know what you should ask for when putting your house on the market.
- Furthermore, you want to know what range of values you should expect from the final sale price.
- Clearly, in your case, the CI for the mean isn't very helpful. The CI for a single predicted value would be.

$$Y' \pm t_{(\alpha/2,df)} s_{y'}$$

SE of a Single Predicted Value

$$s_{y'} = \sqrt{s^2_{y.x}\left[1 + \frac{1}{N} + \frac{(X_i - \bar{X})^2}{\Sigma_{i=1}^{N}(X_i - \bar{X})^2}\right]}$$

$s^2_{y.x}$ is variance of estimate, or mean squared residual (also known as conditional variance of y given x).

- This formula is very similar to that for the mean.
- Again, because of $(X_i - \bar{X})^2$, as X gets farther from its mean, the CI gets wider.

$$Y' = 2.506 + 23.804X = 2.506 + 23.804(2) = 50.114$$

- Using a 95% degree of confidence, the confidence interval for this value is found from:

$$Y' \pm t_{(\alpha/2,df)} s_{y'}$$

Predicted Value Example

$$t_{(95\%/2,25)} = 2.06$$

$$\bar{X} = 1.530$$

$$\sum_{i=1}^{N}(X_i - \bar{X})^2 = 8.333$$

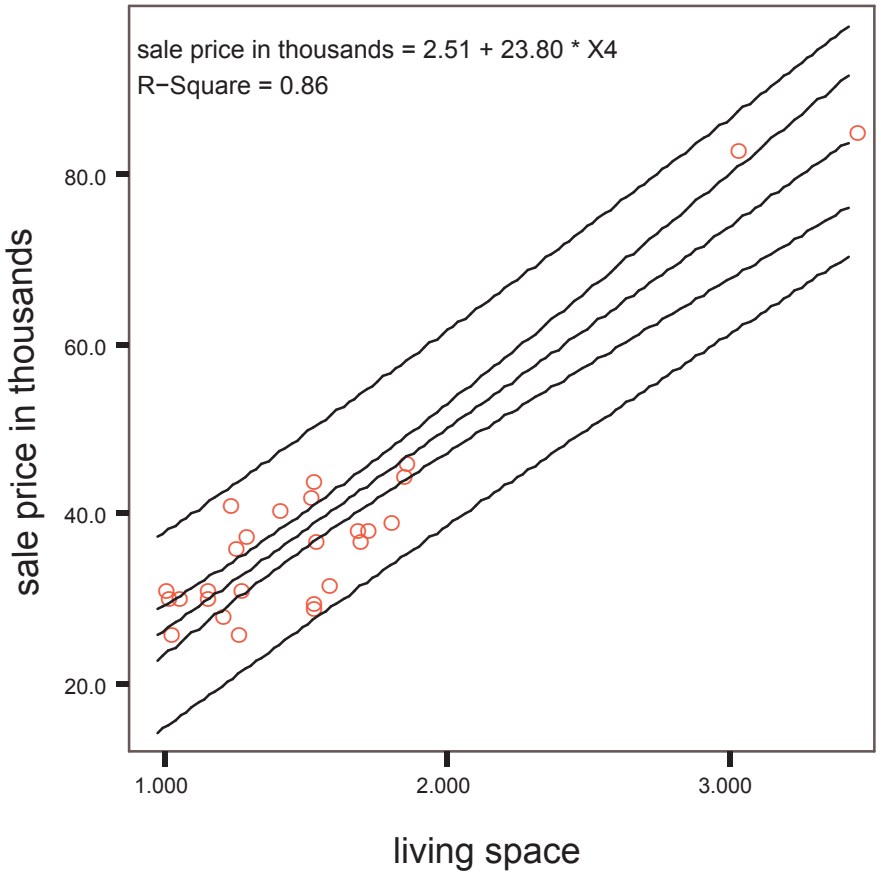
$$s^2_{y.x} = MS_{res} = 29.226$$

$$s_{\mu'} = \sqrt{29.226\left[1 + \frac{1}{27} + \frac{(2.000 - 1.530)^2}{8.333}\right]} = 5.582$$

$$50.114 \pm 2.06 \times 5.582 \rightarrow (38.617, 61.613)$$

So, we are 95% confident that the sale price for *your* 2,000 square foot home would fall between $38,617 and $61,613.

Graph of the Mean Predicted Value Interval



Linear Regression with
95.00% Mean Prediction Interval and
95.00% Individual Prediction Interval

Standard Error Formulas for Multiple X

$$s_{\mu'} = \sqrt{s^2_{y.x}[\mathbf{X}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h]}$$

$$s_{y'} = \sqrt{s^2_{y.x}[1 + \mathbf{X}'_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h]}$$

$s^2_{y.x}$ is variance of estimate, or mean squared residual (also known as conditional variance of y given x).

$\mathbf{X}_h$ is a row vector of values to predict from (including intercept).

X is the data matrix of predictors (including intercept) from which the regression parameter estimates were obtained.

Prediction in Practice

- To augment the topics discussed here, I introduce another example, this one being near to my heart.
- The National Basketball Association (NBA) is the premier professional basketball league in North America.
- A total of 30 NBA teams each play an 82 game regular season.
- For almost all games, you can go to Las Vegas and make wagers as to which team wins (plus or minus the number of points you are given by the casino).
- For example, a casino would say that in a game between the Sacramento Kings and L.A. Lakers, the “point spread” would be -5 for the Kings.
- This means that if you placed a bet on the Kings, if the (Kings’ score -5) was greater than the Lakers’ score, you would win the bet (usually your amount minus about 10%).

NBA Example

NBA Prediction

- I constructed a simple model for the difference between the home team score and away team score using as predictors:
  - An intercept (meaning the effective points a team gets for playing at home).
  - The home team.
  - The away team.
  - An indicator for whether the home team played the previous night.
  - An indicator for whether the away team played the previous night.
- These are all categorical predictors, a topic that will be covered in the upcoming weeks.
- Using the results of the 781 NBA games played up until the All Star break, I estimated the above model (an incorrect model).

NBA Prediction

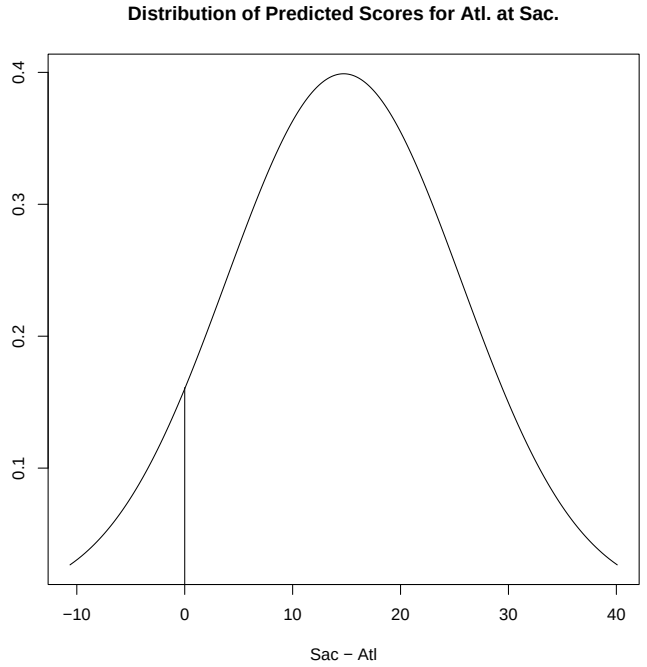
- I sought to predict the scores of the eight games played on Tuesday, 2/22/2005.
- Here are my predicted values (with individual value 95% CI given):

Game	Predicted Home-Away	LCB	UCB	$s_{y'}$
Ind. @ Orl.	3.14	-18.21	24.50	10.90
Mil. @ Cha.	-1.42	-22.79	19.94	10.90
Tor. @ NJN.	1.17	-20.17	22.51	10.89
NYK. @ Det.	8.19	-13.17	29.55	10.90
Mia. @ Chi.	-2.68	-24.03	18.67	10.89
Sea. @ Hou.	-0.06	-21.43	21.32	10.91
Atl. @ Sac.	14.73	-6.63	36.08	10.90
Bos. @ LAL.	0.84	-20.52	22.21	10.90

NBA Prediction

- From the distribution of possible outcomes as predicted by our regression model, we could determine the probability the **home team wins**:

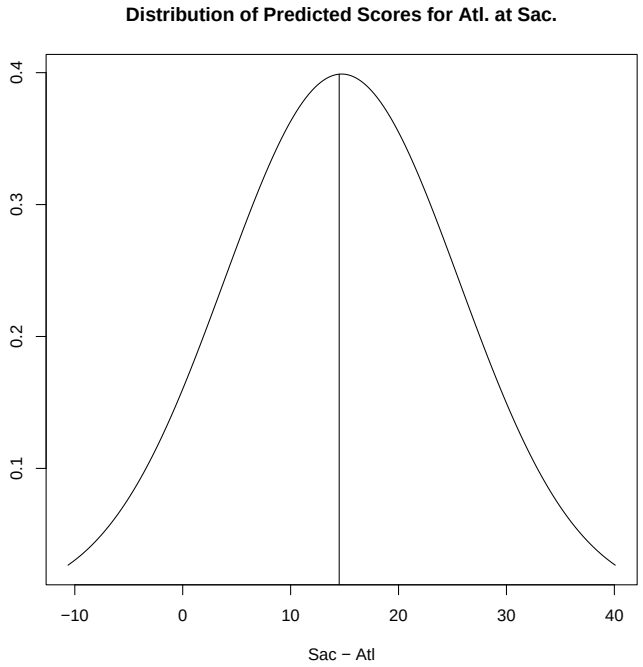
$$P(\text{Home} - \text{Away} > 0) = P\left(\frac{Y' - 0}{s_{y'}} > 0\right)$$



NBA Prediction

- From the distribution of possible outcomes as predicted by our regression model and the point spread at a casino, we could determine the probability we **win a bet placed on the home team**:

$$P(\text{Home} - \text{Away} > \text{Spread}) = P\left(\frac{Y' - \text{Spread}}{s_{y'}} > 0\right)$$



NBA Prediction

- Using this distributional information, and the Vegas spread, I can now estimate the probability I win the bet if I bet on each of the home teams.

Game	Actual Spread	Predicted Home-Away	Estimated P(Home Wins)	Estimated P(I Win Bet)	Actual Home-Away
Ind. @ Orl.	2.0	3.14	0.613	0.542	-24
Mil. @ Cha.	-2.5	-1.42	0.447	0.539	-10
Tor. @ NJN.	6.5	1.17	0.543	0.312	-18
NYK. @ Det.	10.0	8.19	0.774	0.405	12
Mia. @ Chi.	-3.5	-2.68	0.403	0.530	4
Sea. @ Hou.	3.5	-0.06	0.498	0.372	-2
Atl. @ Sac.	14.5	14.73	0.912	0.508	10
Bos. @ LAL.	3.5	0.84	0.531	0.404	9

Final Thought

- Prediction is an important part of regression.
- Be sure to differentiate between differing types of prediction intervals.

- Find a good model and go make some money in Las Vegas (do not take seriously).

Next Time

- Shrinkage adjustments for R^2.
- Cross-validation.
- Variable selection techniques.