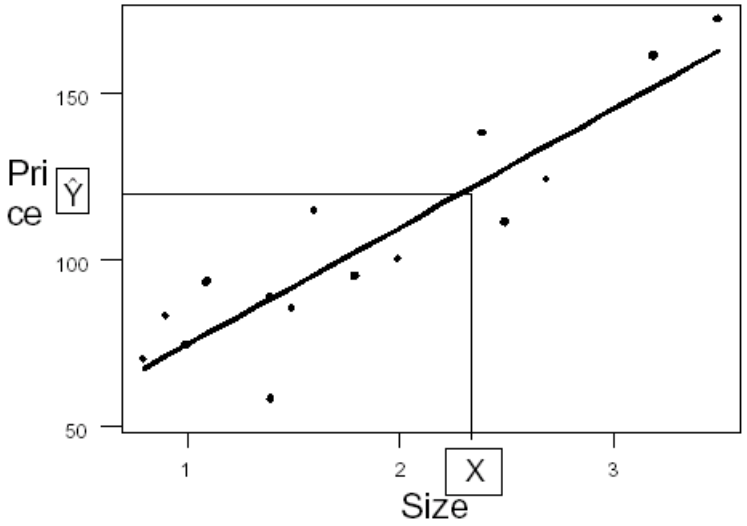
PREDICTIVE MODELING: REGRESSION
REGRESSION: DATA ANALYSIS
REGRESSION: MACHINE LEARNING

What is Regression?

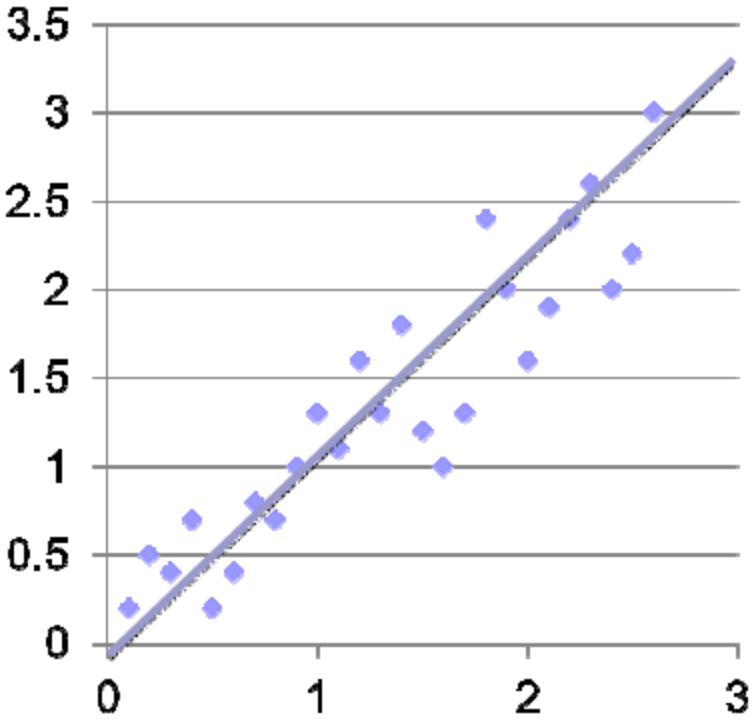
- Type of Data Mining
  - Data Mining is the analysis of large amounts of data in order to discover meaningful patterns
- Regression models analyze the correlation of several variables
- Ex: Linear Regression models a linear graph

Purpose of Modeling

- Prediction: The fitted regression line is a prediction rule!
  - The regression equation is Price = 38.9 + 35.4 Size
- What is the definition of a Prediction Rule?
  - Put in a value of *X for a Y we haven't yet seen*, and out comes a prediction (a function or black box).
  - $f(X) = \beta_0 + \beta_1 X$
  - You give me **any** new value of *X* and I can predict *Y*.

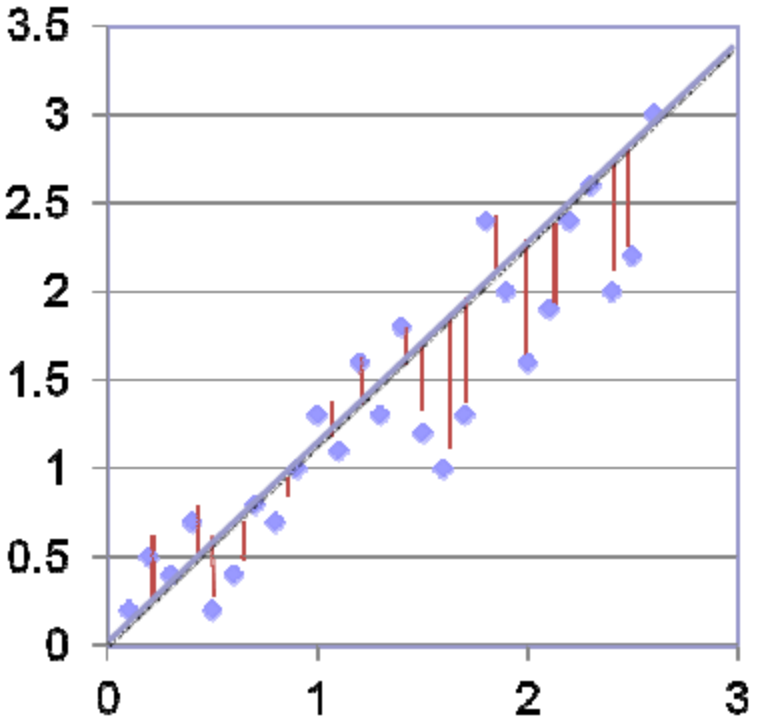


Predictions



- Regression is about making predictions
- Look for a pattern to build a model
- Equation: y = mx + b
- How do you find the best model?

Minimizing Errors



- To make a better model, we minimize the errors
- An error is considered the distance between the actual data and the model data

$$\varepsilon_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Least Squares Linear Regression "Best Fit Line"

- Actual Data: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Error: $\varepsilon_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
- Sum of Errors: $S = \sum \varepsilon_i^2$
- Best Fit Y-Intercept:

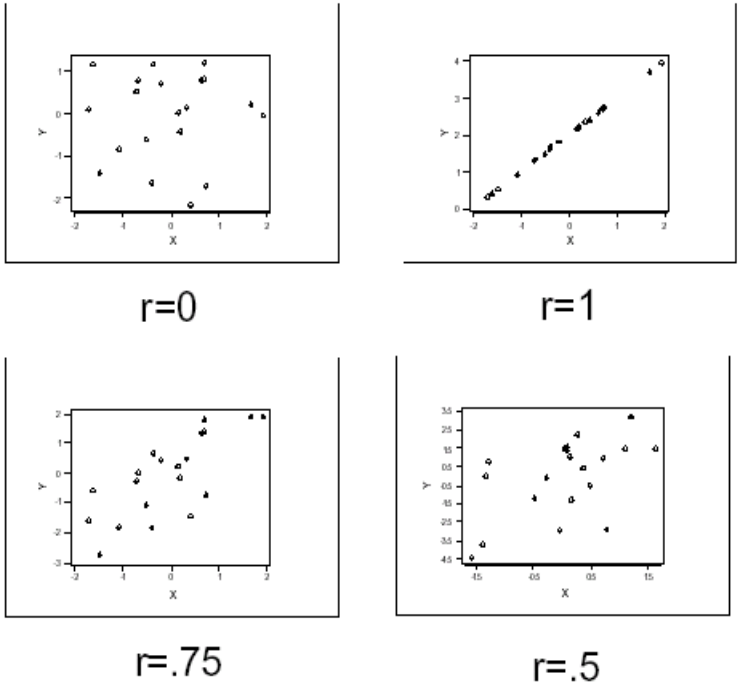
$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

- Best Fit Slope:

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Simple Linear Regression

- Observe the data recorded as the pairs (X_i, Y_i), $i = 1,\ldots,n$.
  - X_i is called independent variables or explanatory variables
  - X is used to explain part of the variability in the dependent variable Y .
- Look at some scatterplots of samples with varying degrees of correlation.



What if your data isn't a straight line…?

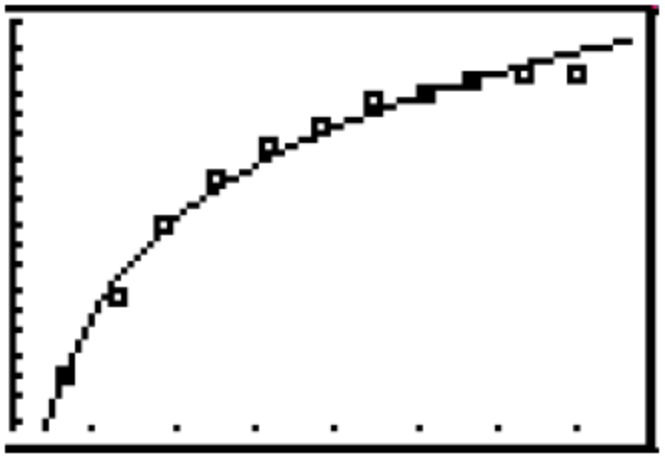
- Logarithmic Regression

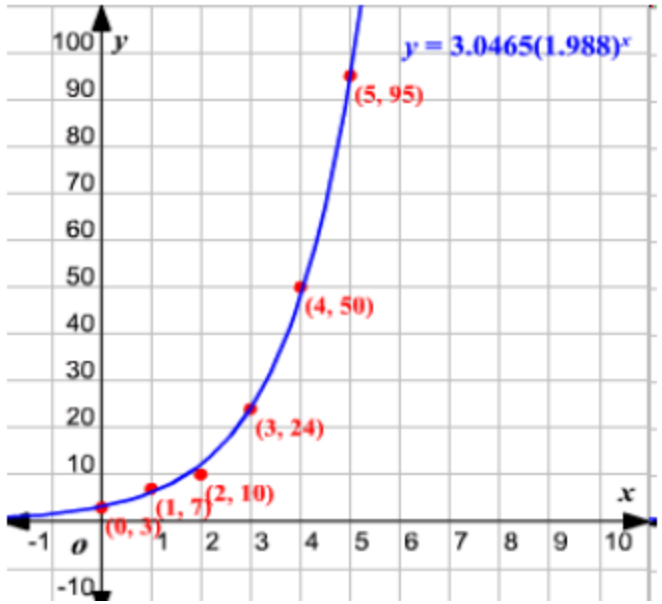
  $Y = a + b\ (\ln x)$
- Quadratic Regression

  $Y = a*x^2+b*x+c$
- Power Regression

  $Y = a*x^b$
- Exponential Regression

  $Y = a*b^x$





Data in Matrix Form

Measurements →

Entities ↓

ID	Income	Age	….	Monthly Debt	Good Risk?
18276	65,000	55	….	2200	Yes
72514	28,000	19	….	1500	No
28163	120,000	62	….	1800	Yes
17265	90,000	35	….	4500	No
…	…	…	….	…	…
…	…	…	….	…	…
61524	35,000	22	….	900	Yes

"Measurements" may be called "variables",
"features", "attributes", "fields", etc

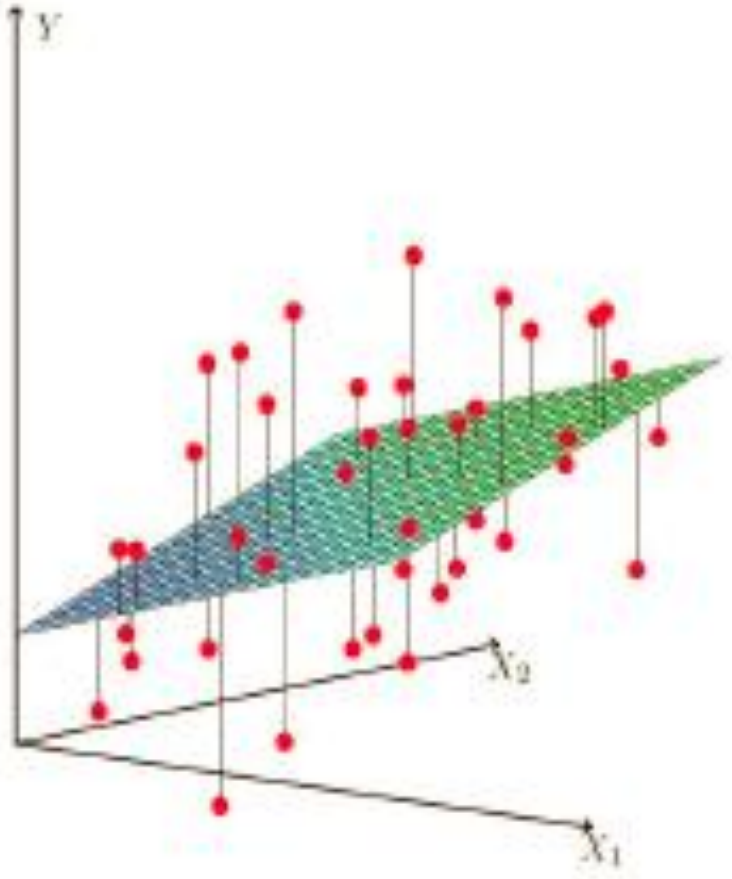
Multiple Regression

Mile Time	Gender	Height (inches)	Weight (lbs)	Age	GPA
10	1	62.2	120	20	3.3
11	0	64.5	166	21	2.8
7	1	70.1	132	18	4.0
8	0	75.0	133	23	1.6
14	0	58.9	121	19	3.7
10	0	68.8	100	25	3.5

Multiple Regression

$$Y = \beta_0 + \beta_1(\text{gender}) + \beta_2(\text{height}) + \beta_3(\text{weight}) + \beta_4(\text{age}) + \beta_5(\text{gpa})$$

- 3+ variables and 1000+ observations
  - Use a computer!
- As before, you can transform these variables if the model does not fit
- By cautious of independence of variables


Y
X2
X1

Growth of the economy

- Consider a simplified version of economic forecasts using regression models.
- Consider the problem of predicting growth of the economy in the next quarter.
  - Some relevant factors might be last quarter's growth, this quarter's growth, the index of leading economic indicators, total factory orders this quarter, aggregate wholesale inventory levels, etc.
  - A linear model for predicting growth would then take the following form:

$$\begin{aligned}\text{next qtr growth} = \beta_0 &+ \beta_1(\text{last qtr growth}) + \beta_2(\text{this qtr growth}) \\ &+ \beta_3(\text{index value}) + \beta_4(\text{factory orders}) \\ &+ \beta_5(\text{inventory levels}) + \text{error}\end{aligned}$$

- Estimate β_0 and the coefficients $\beta_1, \ldots, \beta_5$ from historical data, in order to make predictions.

Levels of advertising

- Determine appropriate levels of advertising and promotion for a particular market segment.
- Consider the problem of managing sales of beer at large college campuses.
  - Sales over, say, one semester might be influenced by ads in the college paper, ads on the campus radio station, sponsorship of sports-related events, sponsorship of contests, etc.
- Use data on advertising and promotional expenditures at many different campuses to tell us the marginal value of dollars spent in each category.
- A marketing strategy is designed accordingly.
- Set up a model of the following type:

$$\text{sales} = \beta_0 + \beta_1(\text{print budget}) + \beta_2(\text{radio budget}) + \beta_3(\text{sports promo budget}) + \beta_4(\text{other promo}) + \text{error}$$

Motivating Examples

- Suppose we have data on sales of houses in some area.
  - For each house, we have complete information about its size, the number of bedrooms, bathrooms, total rooms, the size of the lot, the corresponding property tax, etc., and also the price at which the house was eventually sold.
  - Can we use this data to predict the selling price of a house currently on the market?
  - The first step is to postulate a model of how the various features of a house determine its selling price.
  - A linear model would have the following form:

$$\text{selling price} = \beta_0 + \beta_1(\text{sq.ft.}) + \beta_2\,(\text{no. bedrooms}) + \beta_3\,(\text{no. bath}) + \beta_4\,(\text{no. acres}) + \beta_5\,(\text{taxes}) + \text{error}$$

    - In this expression, β_1 represents the increase in selling price for each additional square foot of area: it is the marginal cost of additional area.
    - β_2 and β_3 are the marginal costs of additional bedrooms and bathrooms, and so on.
    - The intercept β_0 could in theory be thought of as the price of a house for which all the variables specified are zero; of course, no such house could exist, but including β_0 gives us more flexibility in picking a model.

What is Machine learning?

- Part of artificial intelligence
- Creating algorithms allowing computers to evolve behaviors based on empirical data
  - Regression
- Automatically learn and recognize complex patterns and make decisions

Algorithm Types

- Supervised learning
  - Output of training data provided by the programmer
- Unsupervised learning
  - Output of training data not provided (clustering)
- Reinforcement learning
  - learns how to act given an observation of the world
- Transduction
  - uses training input, training output and test output

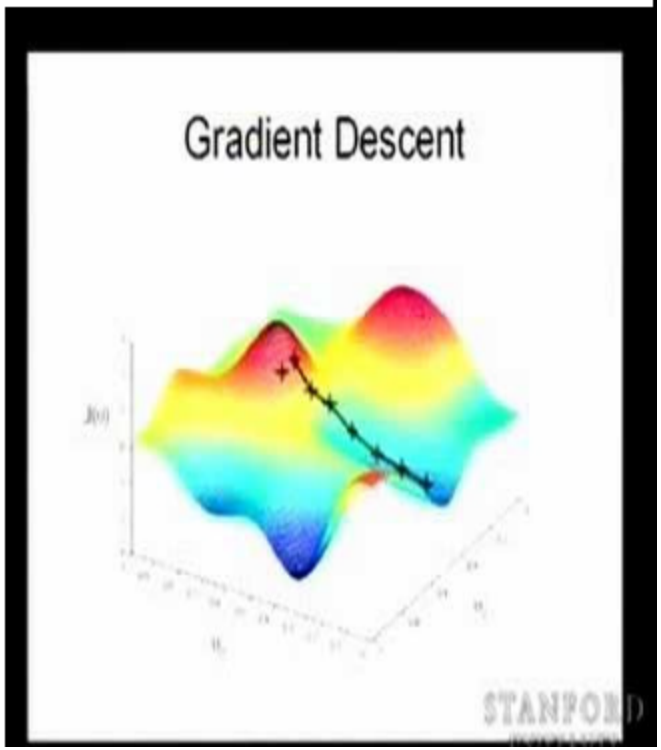
Supervised learning Regression Problem Autonomous Driving

- Learning algorithm-gradient descent
- Digitizes the road ahead and records the person's steering directions
- Once learned…
  - digitizes the road
  - feeds the image to its neural networks.
- Measure each steering direction's confidence

Alvin-system of artificial neural networks

Gradient Descent

- Similar to Directed Random Search
- Find an optimal hypothesis function
- Pick a random point on the graph and go down in the direction that gives most downward descent.
- Repeat until local minimum reached
- Update parameters
- Continue until the hypothesis function converges
- This function has the least overall error



How to make money with machine learning?

Relevance

- Social Media
- Twitter, facebook, youtube, yelp, etc.
- Track ad campaigns

Notation

- Variables X, Y….. with values x, y (lower case)
  - Vectors indicated by $\underline{X}$
- Components of X indicated by X_j with values x_j
- "Matrix" data set D with n rows and p columns
  - jth column contains values for variable X_j
  - ith row contains a vector of measurements on object i, indicated by x(i)
  - The jth measurement value for the ith object is $x_j(i)$
- Unknown parameter for a model = θ
  - Can also use other Greek letters, like α, β, δ, γ
  - Vector of parameters = $\underline{\theta}$

Example: Multivariate Linear Regression

- Task: predict real-valued Y, given real-valued vector $\underline{X}$
- Score function, e.g., least squares is often used
-

$$S(\underline{\theta}) = \Sigma_i \left[y(i) - f\left(\underline{x}(i) ; \underline{\theta}\right) \right]^2$$

target value

predicted value

- Model structure: e.g., linear $f(\underline{x} ; \underline{\theta}) = \alpha_0 + \Sigma\, \alpha_j\, x_j$
- Model parameters = $\underline{\theta} = \{\alpha_0, \alpha_1, \ldots\ldots \alpha_p\}$

Note that we can write

$$S(\theta) = \Sigma_i [y_{(i)} - \Sigma\, \alpha_j\, x_j]^2$$

$$= \Sigma_i\, e_i^2$$

$$= \underline{e}'\, \underline{e}$$

$$= (y - X\,\theta)'\,(y - X\,\theta)$$

where $\underline{e} = y - X\,\theta$

(p+1) x 1 vector of parameter values

y = N x 1 vector of target values

N x (p+1) vector of input values

$$S(\theta) = \Sigma e^2 = e' e = (y - X\theta)' (y - X\theta)$$

$$= y' y - \theta' X' y - y' X \theta + \theta' X' X \theta$$

$$= y' y - 2\theta' X' y + \theta' X' X \theta$$

Taking derivative of $S(\theta)$ with respect to the components of θ gives….

$dS/d\theta = -2 X' y + 2 X' X \theta$

Set this to 0 to find the extremum (minimum) of S as a function of θ …

Set to 0 to find the extremum (minimum) of S as a function of θ …

$\Rightarrow \; -2\, X' y + 2\, X' X \theta = 0$

$\Rightarrow \quad X' X \theta = X' y$ (known in statistics as the Normal Equations)

Letting $X' X = C$, and $X' y = b$,

we have $C \theta = b$, i.e., a set of linear equations

We could solve this directly, e.g., by matrix inversion

$$\theta = C^{-1} b = (X' X)^{-1} X' y$$

Solving for the θ's

- Problem is equivalent to inverting X' X matrix
  - Inverse does not exist if matrix is not of full rank
    - E.g., if 1 column is a linear combination of another (collinearity)
    - Note that X'X is closely related to the covariance of the X data
      - So we are in trouble if 2 or more variables are perfectly correlated
    - Numerical problems can also occur if variables are almost collinear
- Equivalent to solving a system of p linear equations
  - Many good numerical methods for doing this, e.g.,
    - Gaussian elimination, LU decomposition, etc
  - These are numerically more stable than direct inversion
- Alternative: gradient descent
  - Compute gradient and move downhill
    - Will say more later on why this is better than direct solutions for certain types of problems

Comments on Multivariate Linear Regression

- Prediction model is a linear function of the parameters

- Score function: quadratic in predictions and parameters
  - $\Rightarrow$ Derivative of score is linear in the parameters
  - $\Rightarrow$ Leads to a linear algebra optimization problem, i.e., $C\,\theta = b$

- Model structure is simple….
  - p-1 dimensional hyperplane in p-dimensions
  - Linear weights => interpretability

- Often useful as a baseline model
  - e.g., to compare more complex models to

- Note: even if it's the wrong model for the data (e.g., a poor fit) it can still be useful for prediction

Limitations of Linear Regression

- True relationship of X and Y might be non-linear
  - Suggests generalizations to non-linear models
- Complexity:
  - $O(N p^2 + p^3)$ - problematic for large p
- Correlation/Collinearity among the X variables
  - Can cause numerical instability (C may be ill-conditioned)
  - Problems in interpretability (identifiability)
- Includes all variables in the model…
  - But what if p=1000 and only 3 variables are actually related to Y?

Non-linear models, but linear in parameters

- We can add additional polynomial terms in our equations, e.g., all "2nd order" terms

$$f(\underline{x} ; \underline{\theta}) = \alpha_0 + \Sigma \alpha_j x_j + \Sigma \beta_{ij} x_i x_j$$

- Note that it is a non-linear functional form, but it is linear in the parameters (so still referred to as "linear regression")
  - We can just treat the X_i X_j terms as additional fixed inputs
  - In fact we can add in any non-linear input functions!, e.g.

$$f(\underline{x} ; \underline{\theta}) = \alpha_0 + \Sigma \alpha_j f_j(\underline{x})$$

Comments:

- Exact same linear algebra for optimization (same math)
- Number of parameters has now exploded -> greater chance of overfitting
  - Ideally would like to select only the useful quadratic terms
  - Can generalize this idea to higher-order interactions

Non-linear (both model and parameters)

- We can generalize further to models that are nonlinear in all aspects

$$f(\underline{x} ; \underline{\theta}) = \alpha_0 + \Sigma \alpha_k g_k(\beta_{k0} + \Sigma \beta_{kj} x_j)$$

where the g's are non-linear functions with fixed functional forms.

In machine learning this is called a neural network

In statistics this might be referred to as a generalized linear model or projection-pursuit regression

For almost any score function of interest, e.g., squared error, the score function is a non-linear function of the parameters.

Closed form (analytical) solutions are rare.

Thus, we have a multivariate non-linear optimization problem (which may be quite difficult!)

Optimization in the Non-Linear Case

- We seek the minimum of a function in d dimensions, where d is the number of parameters (d could be large!)

- There are a multitude of heuristic search techniques (see chapter 8)
  - Steepest descent (follow the gradient)
  - Newton methods (use 2nd derivative information)
  - Conjugate gradient
  - Line search
  - Stochastic search
  - Genetic algorithms

- Two cases:
  - Convex (nice -> means a single global optimum)
  - Non-convex (multiple local optima => need multiple starts)

Other non-linear models

- Splines
  - "patch" together different low-order polynomials over different parts of the x-space
  - Works well in 1 dimension, less well in higher dimensions

- Memory-based models

  $y' = \Sigma\, w_{(x',x)}\, y,$ where y's are from the training data

  $w_{(x',x)}$ = function of distance of x from x'

- Local linear regression

  $y' = \alpha_0 + \Sigma\, \alpha_j\, x_j$, where the alpha's are fit at prediction time just to the (y,x) pairs that are close to x'

Selecting the k best predictor variables

- Linear regression: find the best subset of k variables to put in model
  - This is a generic problem when p is large
    (arises with all types of models, not just linear regression)

- Now we have models with different complexity..
  - E.g., p models with a single variable
  - p(p-1)/2 models with 2 variables, etc…
  - 2^p possible models in total
    - Can think of space of models as a lattice
  - Note that when we add or delete a variable, the optimal weights on the other variables will change in general
    - k best is not the same as the best k individual variables

- Aside: what does "best" mean here? (will return to this shortly…)

Search Problem

- How can we search over all 2^p possible models?
  - exhaustive search is clearly infeasible
  - Heuristic search is used to search over model space:
    - Forward search (greedy)
    - Backward search (greedy)
    - Generalizations (add or delete)
      - Think of operators in search space
    - Branch and bound techniques
  - This type of variable selection problem is common to many data mining algorithms
    - Outer loop that searches over variable combinations
    - Inner loop that evaluates each combination

Empirical Learning

- Squared Error score (as an example: we could use other scores)

$$S(\underline{\theta}) = \Sigma_i \left[y(i) - f\left(\underline{x}(i) ; \underline{\theta}\right) \right]^2$$

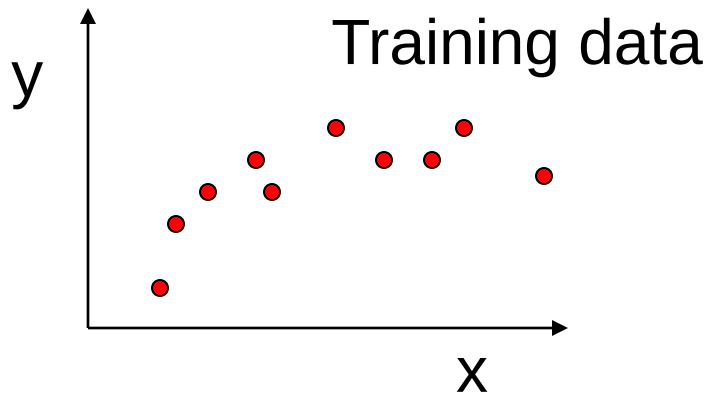
where $S(\underline{\theta})$ is defined on the training data D

- We are really interested in finding the $f(x; \underline{\theta})$ that best predicts y on **future** data, i.e., minimizing

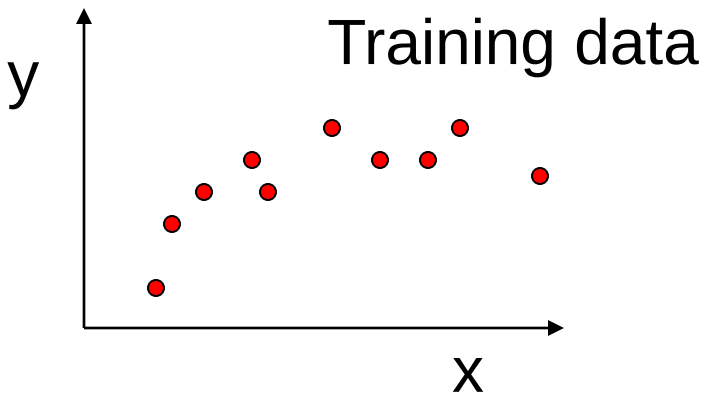
$$E[S] = E\left[y - f\left(\underline{x} ; \underline{\theta}\right) \right]^2$$ (where the expectation is over future data)

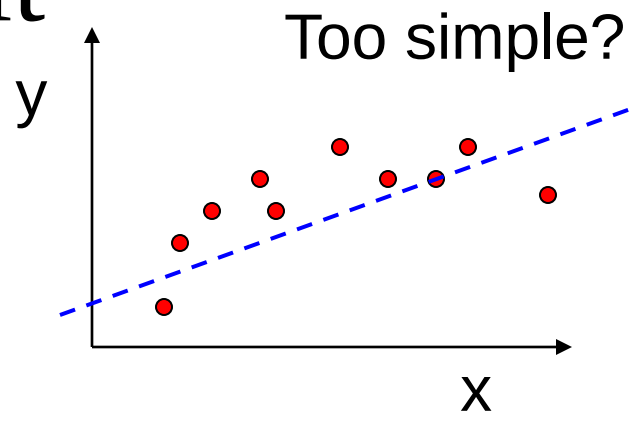
- Empirical learning
  - Minimize $S(\underline{\theta})$ on the training data D_{train}
  - If D_{train} is large and model is simple we are assuming that the best f on training data is also the best predictor f on future test data D_{test}

Complexity versus Goodness of Fit

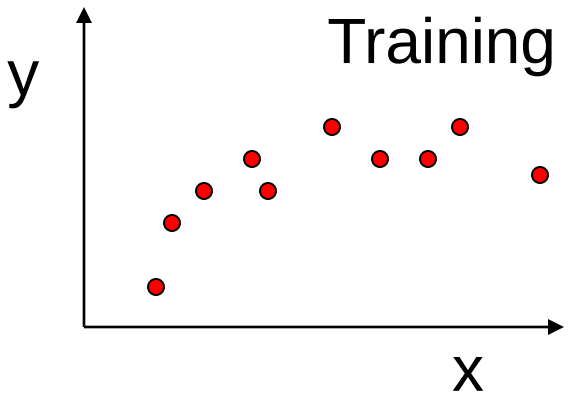


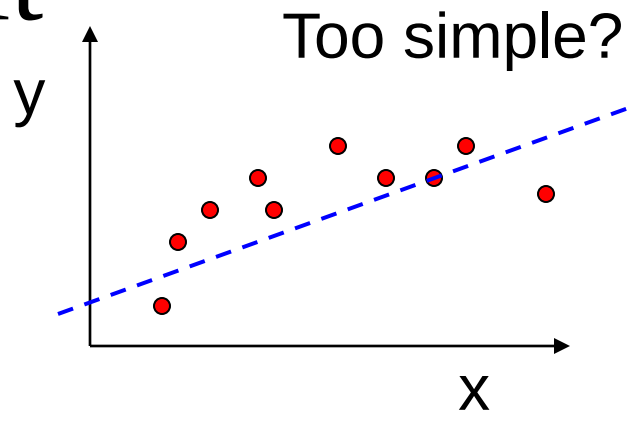
Complexity versus Goodness of Fit

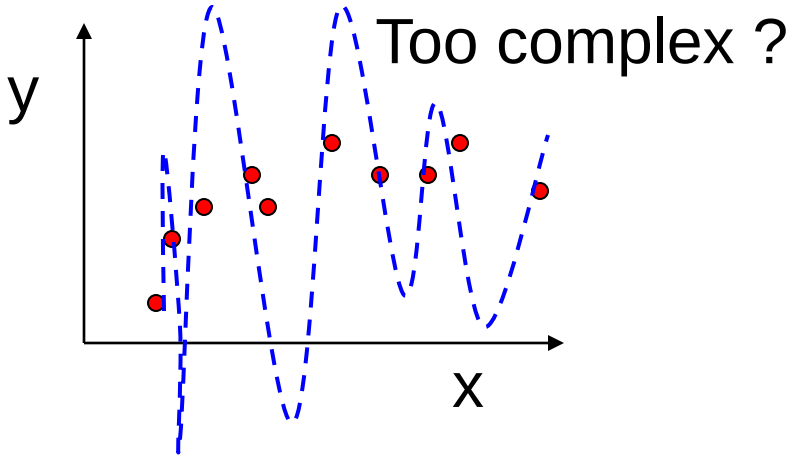




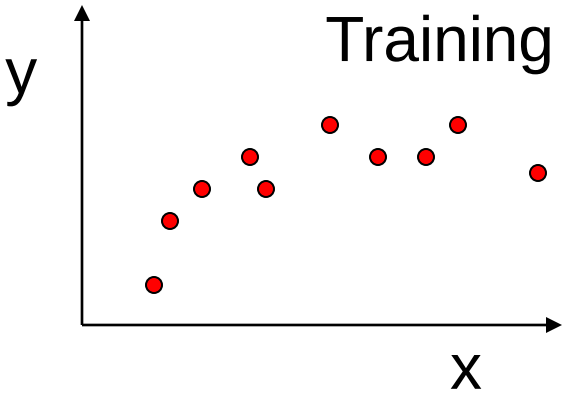
Complexity versus Goodness of Fit

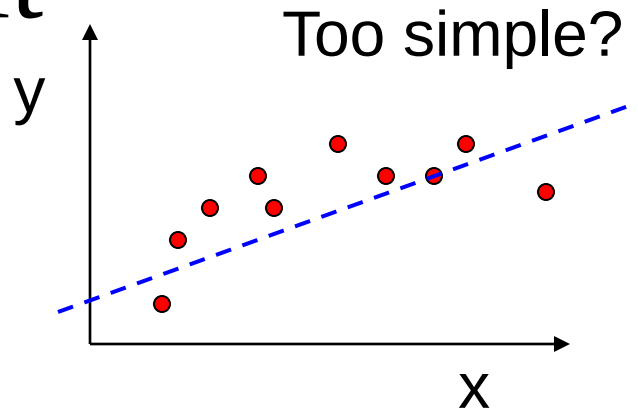


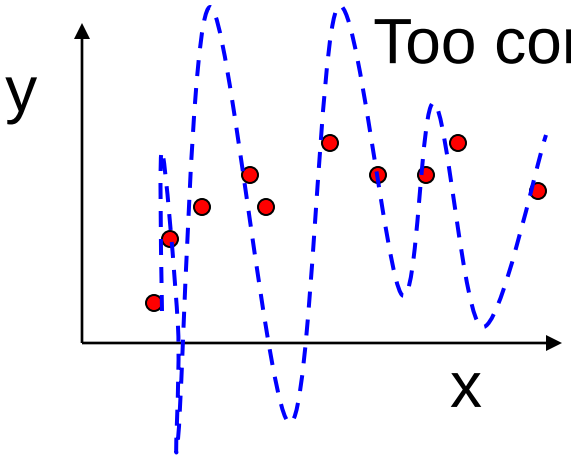


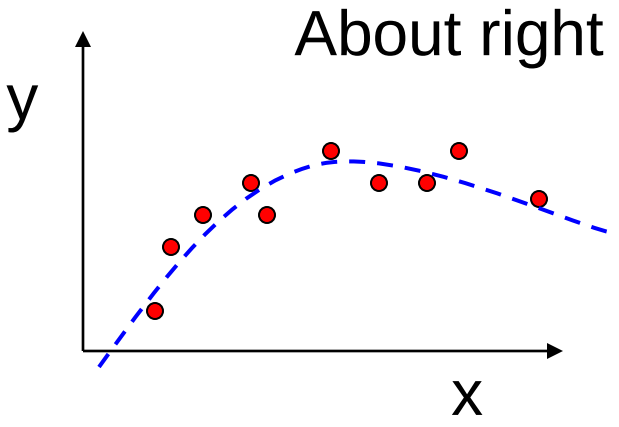


Complexity versus Goodness of Fit

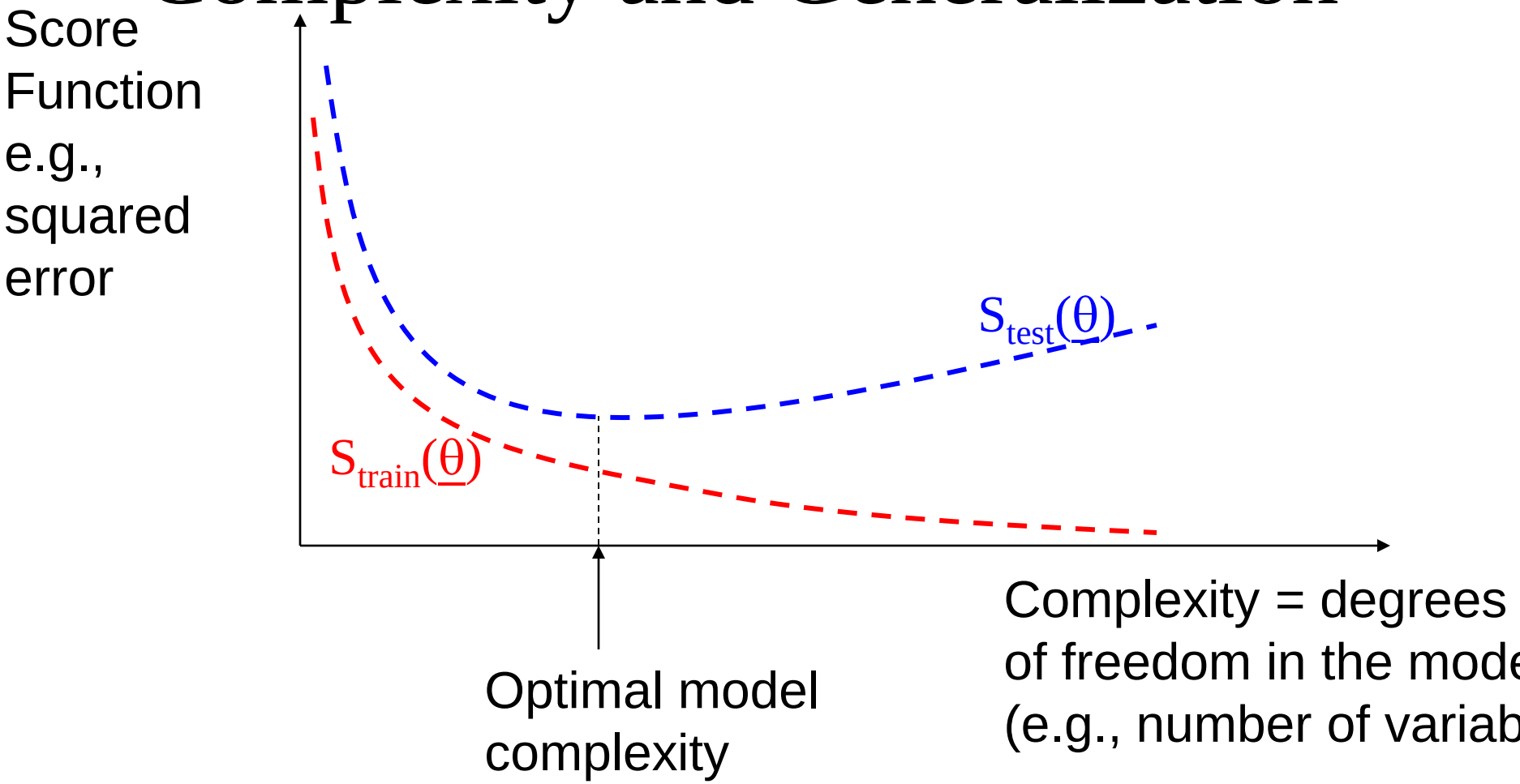




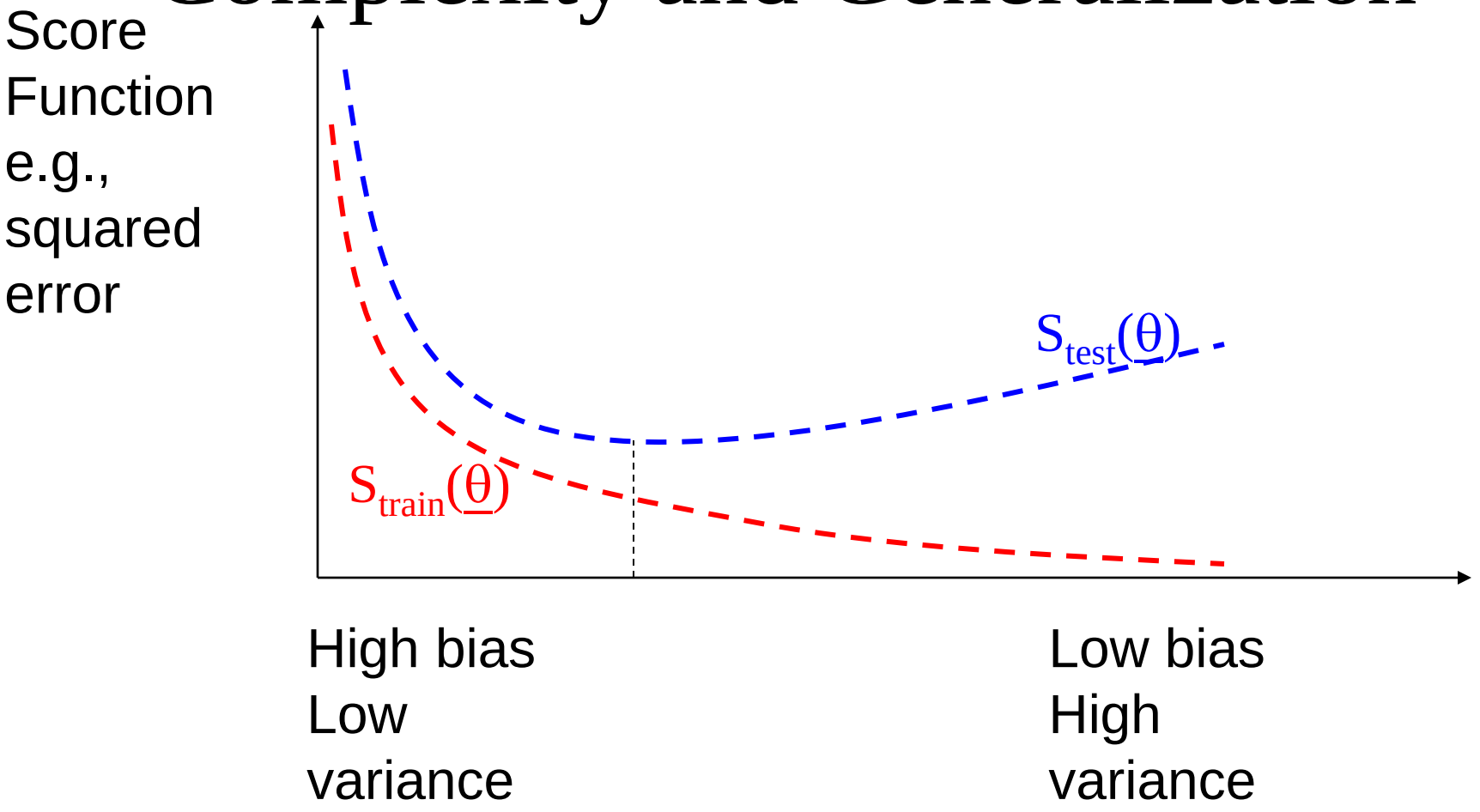




Complexity and Generalization



Complexity and Generalization

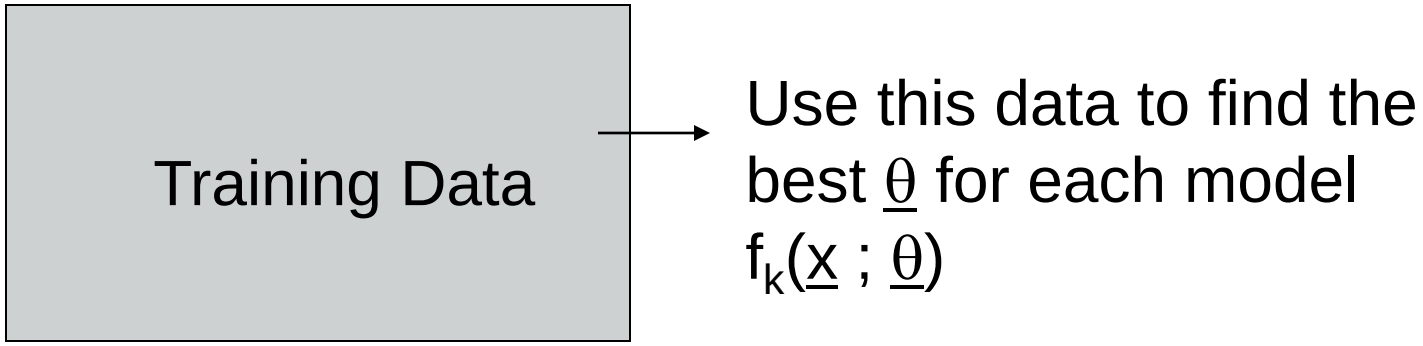


Defining what "best" means

- How do we measure "best"?
  - Best performance on the training data?
    - K = p will be best (i.e., use all variables), e.g., p=10,000
    - So this is not useful in general
  - Performance on the training data will in general be optimistic

- Practical Alternatives:
  - Measure performance on a single validation set
  - Measure performance using multiple validation sets
    - Cross-validation
  - Add a penalty term to the score function that "corrects" for optimism
    - E.g., "regularized" regression: SSE + λ sum of weights squared

Training Data



Validation Data

Use this data to find the best $\underline{\theta}$ for each model $f_k(\underline{x} ; \underline{\theta})$

Use this data to
(1)calculate an estimate of $S_k(\underline{\theta})$ for each $f_k(\underline{x} ; \underline{\theta})$ and
(2)select $k^* = \arg\min_k S_k(\underline{\theta})$

Validation Data

can generalize to **cross-validati**

Use this data to find the best $\underline{\theta}$ for each model $f_k(\underline{x} ; \underline{\theta})$

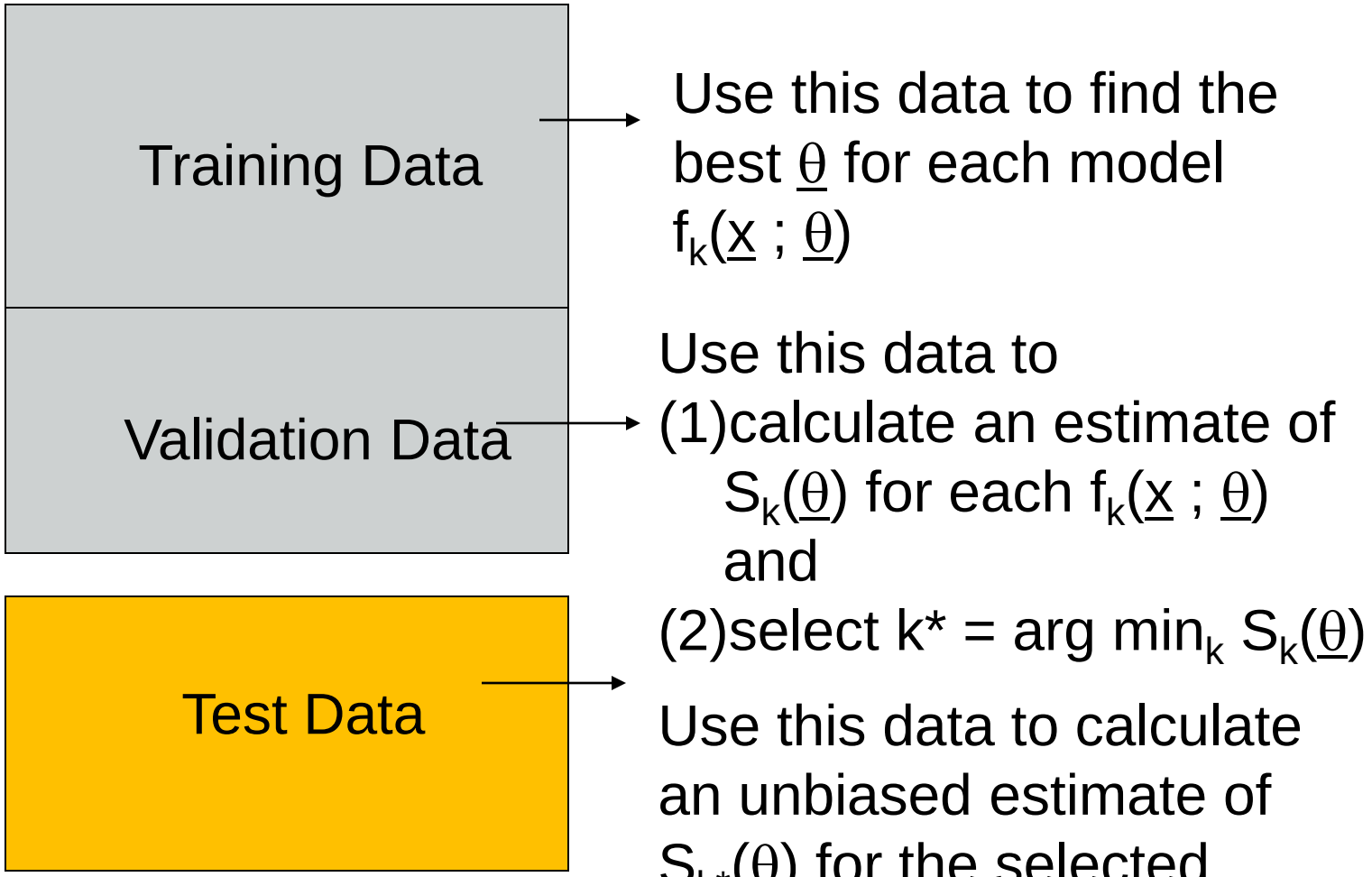
Use this data to
(1)calculate an estimate of $S_k(\underline{\theta})$ for each $f_k(\underline{x} ; \underline{\theta})$ and
(2)select $k^* = \arg\min_k S_k(\underline{\theta})$

2 different (but related) issues here

1. Finding the function f that minimizes $S(\underline{\theta})$ for future data

2. Getting a good estimate of $S(\underline{\theta})$, using the chosen function, on future data,
   - e.g., we might have selected the best function f, but our estimate of its performance will be optimistically biased if our estimate of the score uses any of the same data used to fit and select the model.

Test Data


Training Data
Validation Data
Test Data
Use this data to find the best $\underline{\theta}$ for each model $f_k(\underline{x} ; \underline{\theta})$
Use this data to
(1)calculate an estimate of $S_k(\underline{\theta})$ for each $f_k(\underline{x} ; \underline{\theta})$ and
(2)select $k^* = \arg\min_k S_k(\underline{\theta})$
Use this data to calculate an unbiased estimate of $S_{k^*}(\underline{\theta})$ for the selected model

Another Approach with Many Predictors: Regularization

- Modified score function:

$$S_\lambda(\underline{\theta}) = \Sigma_i [y(i) - f(\underline{x}(i) ; \underline{\theta})]^2 + \lambda \Sigma \theta_j^{\,2}$$

- The second term is for "regularization"
  - When we minimize -> encourages keeping the θ_j's near 0
  - Bayesian interpretation: minimizing - log P(data|q) - log P(q)

- L1 regularization

$$S_\lambda(\underline{\theta}) = \Sigma_i [y(i) - f(\underline{x}(i) ; \underline{\theta})]^2 + \lambda \Sigma | \theta_j |$$

(basis of popular "Lasso" method, e.g., see Rob Tibshirani's page on lasso methods: http://www-stat.stanford.edu/~tibs/lasso.html)

Time-series prediction as regression

- Measurements over time $x_1, \ldots\ldots x_t$
- We want to predict x_{t+1} given $x_1, \ldots\ldots x_t$
- Autoregressive model

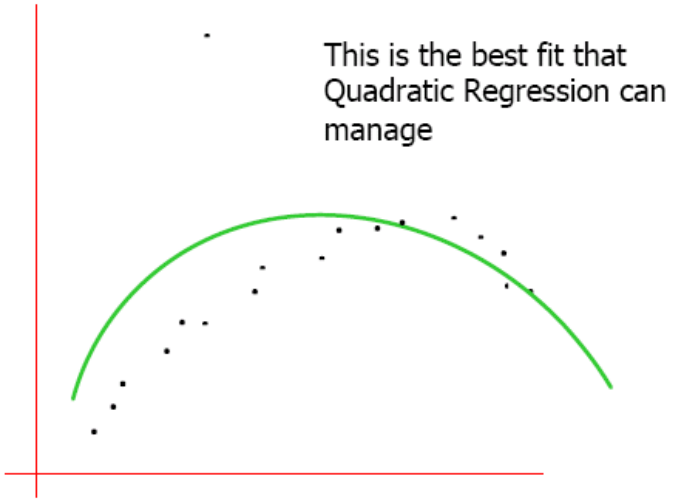
$$x_{t+1} = f(x_1, \ldots\ldots x_t ; \underline{\theta}) = \sum \alpha_k x_{t-k}$$

  - Number of coefficients K = memory of the model
  - Can take advantage of regression techniques in general to solve this problem (e.g., linear in parameters, score function = squared error, etc)

- Generalizations
  - Vector x
  - Non-linear function instead of linear
  - Add in terms for time-trend (linear, seasonal), for “jumps”, etc

Other aspects of regression

- Diagnostics
  - Useful in low dimensions



- Weighted regression
  - Useful when rows have different weights
- Different score functions
  - E.g. absolute error, or additive noise varies as a function of x
- Predicting y values constrained to a certain range, e.g., $y > 0$, or $0 < y < 1$
- Predicting binary y values
  - Regression as a generalization of classification

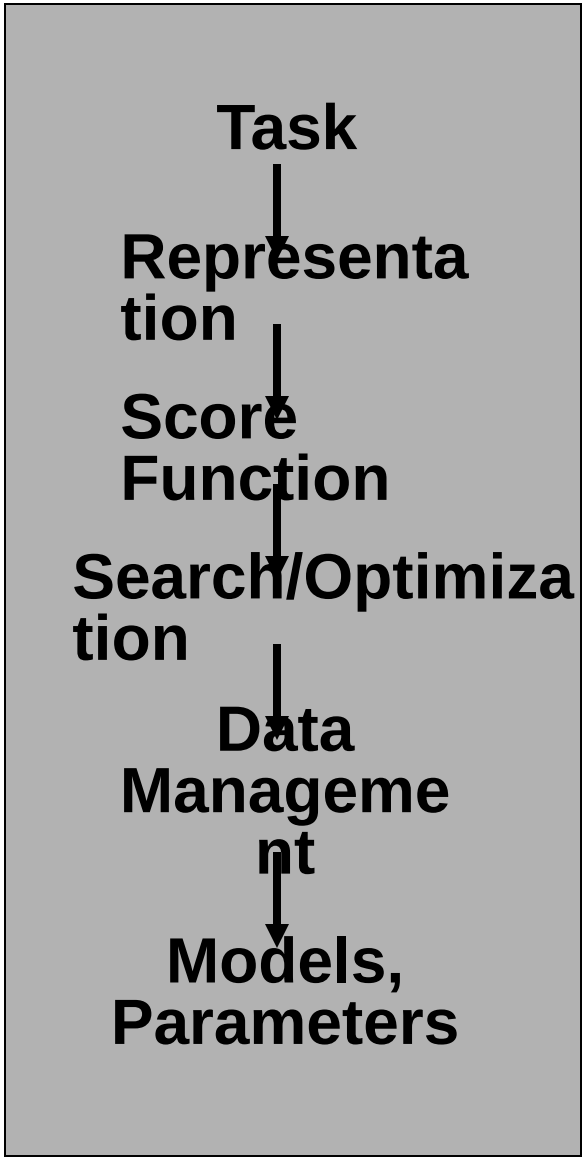
Generalized Linear Models (GLMs)

- g(y) = u(x) $= \alpha_0 + \Sigma \alpha_j x_j$
  - Where g [ ] is a "link" function
  - u(x) is a linear function of the vector x

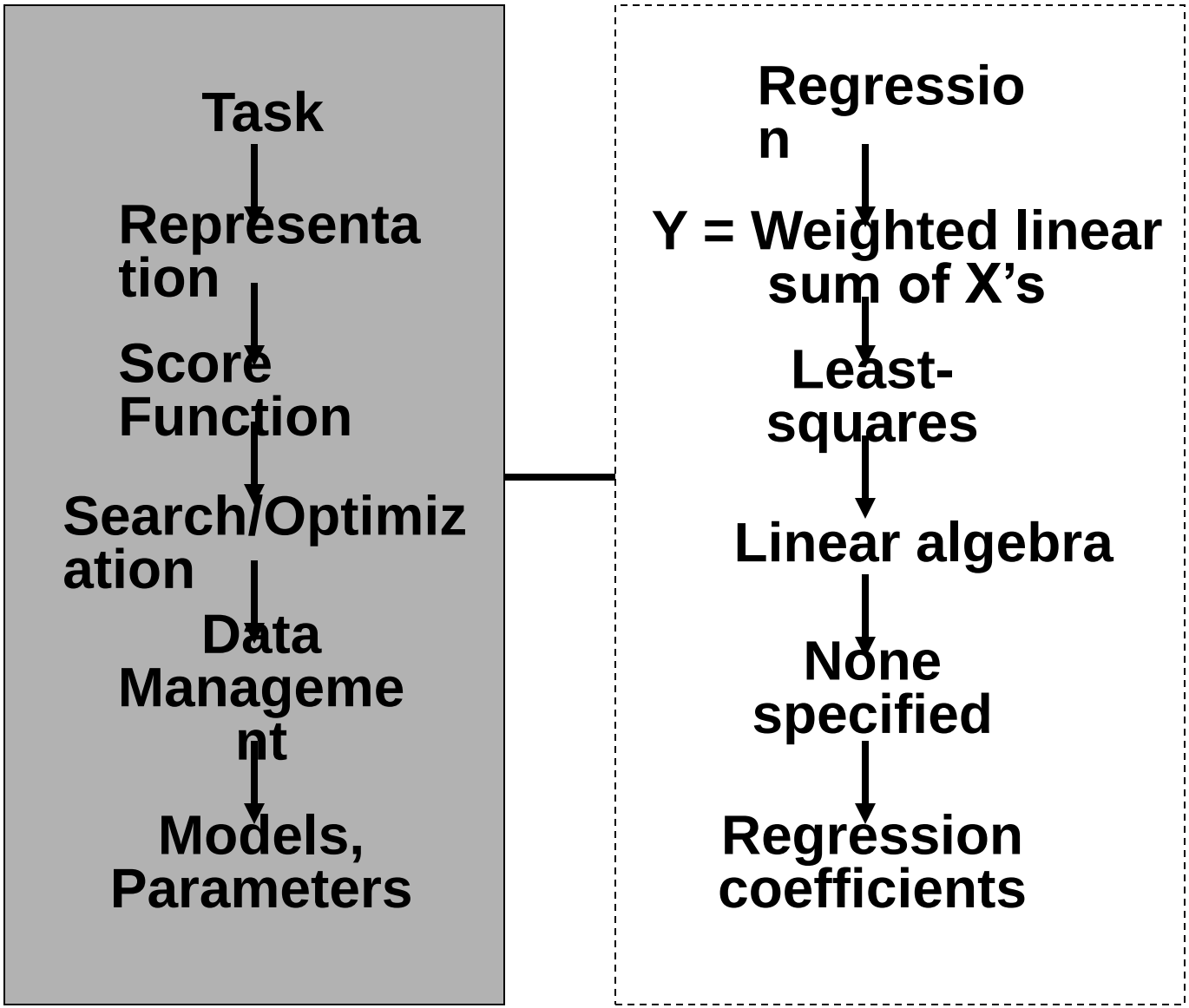
(McCullagh and Nelder, 1989)

- Examples:
  - g = identity function -> linear regression
  - Logistic regression: g(y) = log(y / 1-y) = $\alpha_0 + \Sigma \alpha_j x_j$
  - Logarithmic link: g(y) = log(y) = $\alpha_0 + \Sigma \alpha_j x_j$
  - GLMs are widely used in statistics
  - Details of learning/fitting algorithm depend on the specifics of the link function

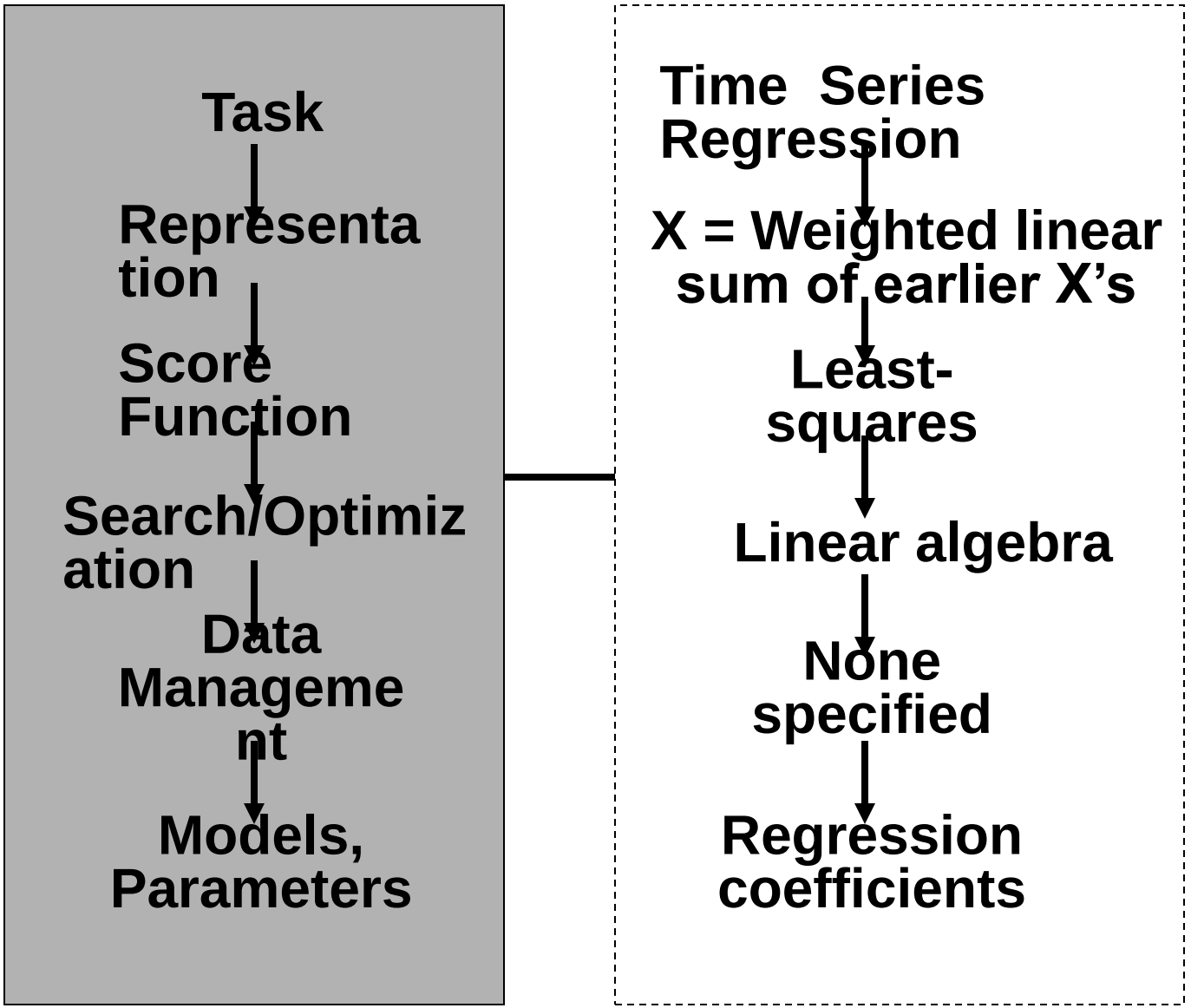
What's in a Data Mining Algorithm?



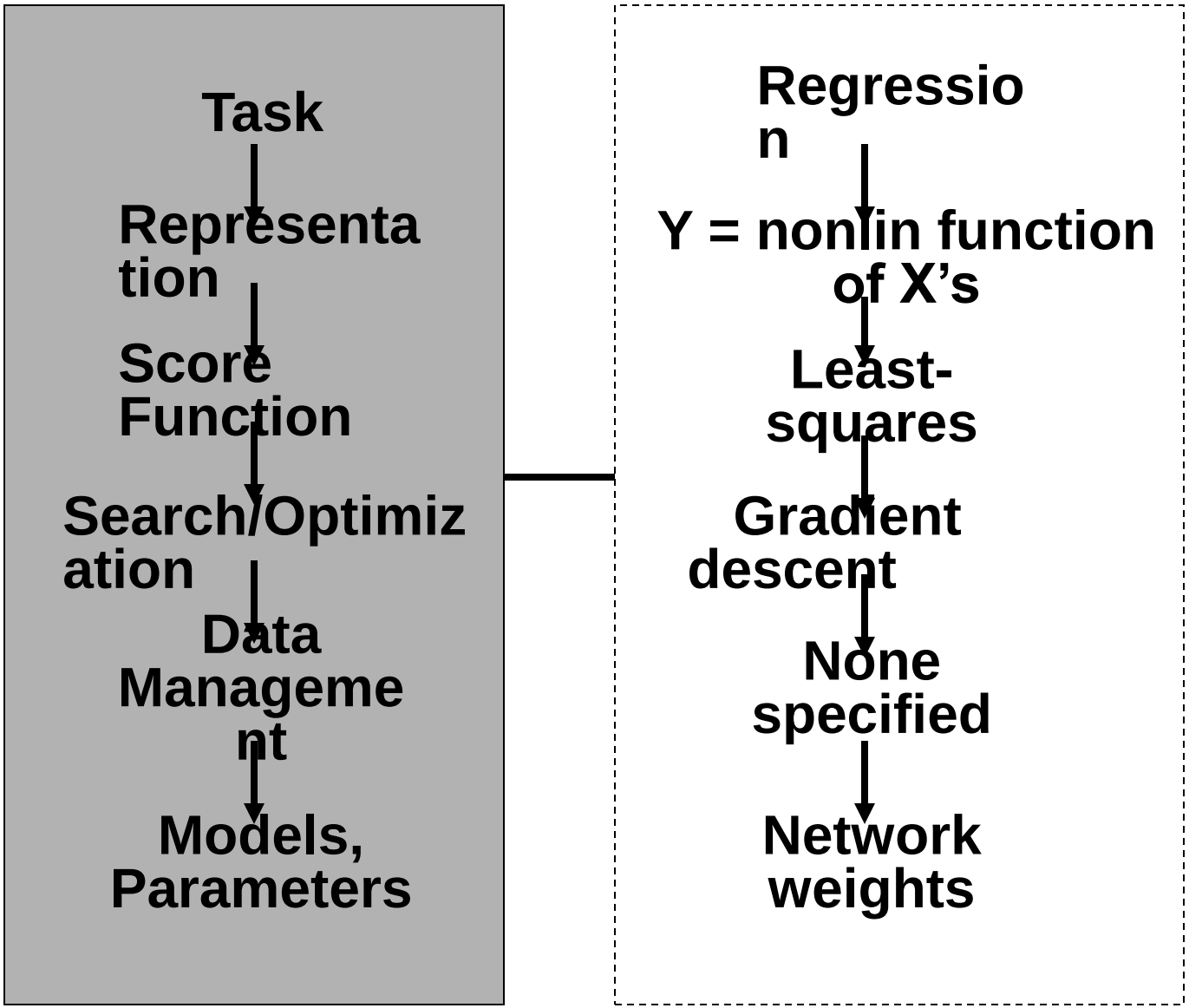
Multivariate Linear Regression



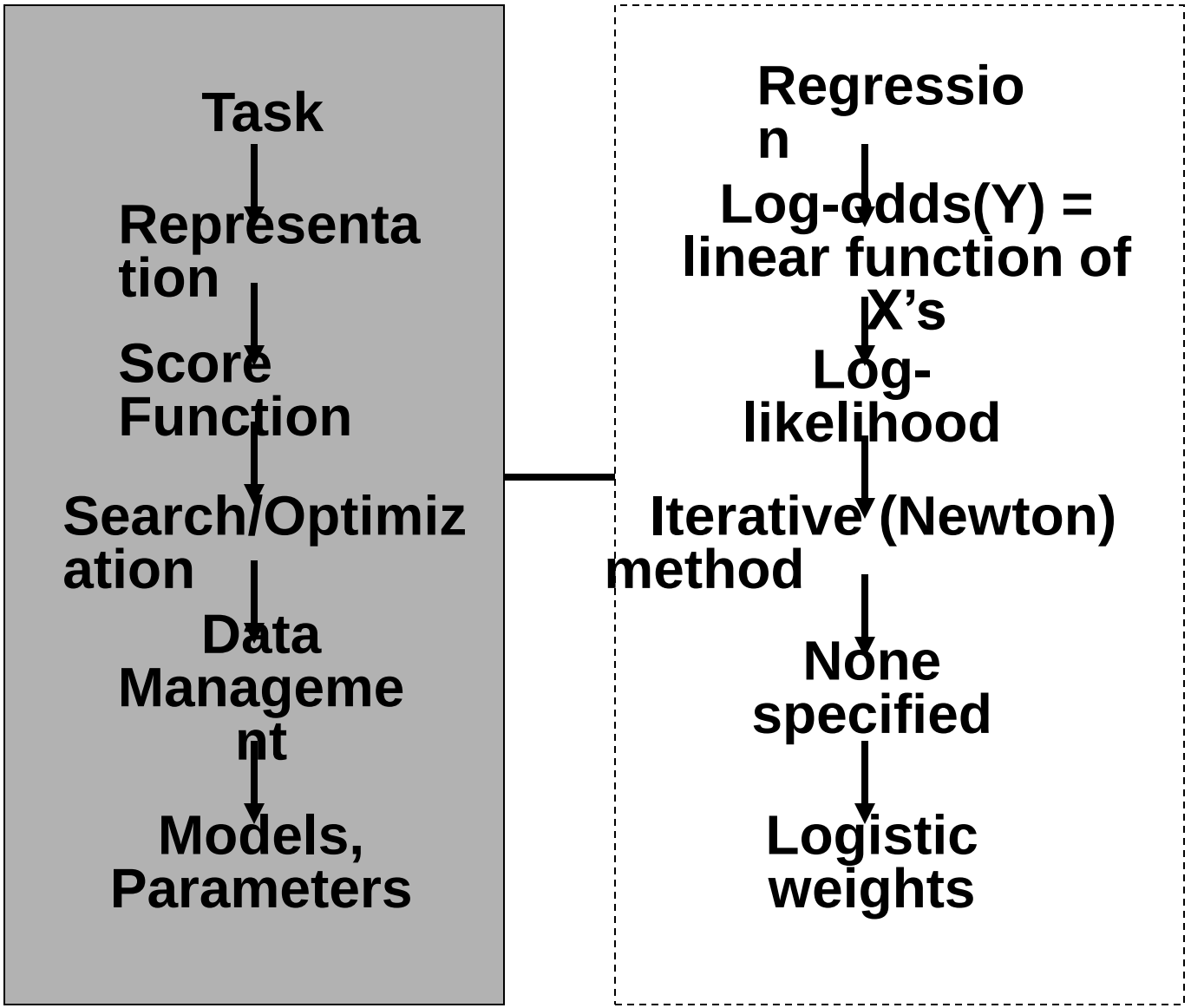
Autoregressive Time Series Models



Neural Networks



Logistic Regression



Sales of Houses

- The error reflects the fact that two houses with exactly the same characteristics need not sell for exactly the same price.
  - There is always some variability left over, even after we specify the value of a large number variables.
  - This variability is captured by an error term, which we will treat as a random variable.

- Regression analysis is a technique for using data to identify relationships among variables and use these relationships to make predictions.

Forecast Accuracy

- Our forecast is not going to be right on the money every time and we need to develop the notion of forecast accuracy.
- Two things we want:
  - What kind of Y can we expect for a given value of X?
  - How sure are we about this forecast?
  - How different could y be from what we expect?
- Goal: Provide a measure of the accuracy of forecasts or equivalently how much **uncertainty** is there in our forecasts.
- Proposal: Provide a range of possible Y values that are likely given this x value.

Prediction Interval

- Prediction Interval: range of possible *Y* values that are likely given *X*
- What influences the length of the prediction interval?
  - Intuitively, the answer must lie in observed variation of the data points about the prediction rule or fitted line.
- **Key Insight:** To construct a prediction interval, we have to assess the the likely range of residual values which will occur for an **as yet unobserved** *Y* value!
- How can we achieve this?
  - Develop a **probability model** for distribution of these residuals values.
  - If the residuals were normally distributed with a given standard deviation, then we could make formal probability statements about the range of likely residuals!!
  - With 95% probability, the residuals will be between -$28,000 and $28,000.

Simple Linear Regression Model

- Once we come to the view that the residuals might come from a probability distribution, we must also acknowledge that the “fitted” line might be fooled by the particular realizations of the residuals.
- The model will enable us to think about uncertainty and which uses a particular distribution for the deviations from the line.
- The power of statistical inference comes from our ability to make very precise probability statements about the accuracy of estimates and forecasts.
  - There is no free lunch, in order to make these statements and to understand the output from the regression procedure, we must *invest in a probability model*.

Simple Linear Regression Model

- $Y = \beta_0 + \beta_1 X + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$
- Part of Y related to X (What we can expect): $\beta_0 + \beta_1 X$
- Part of Y independent of X (How different can Y be): ε
- Note that ε is a random variable which is called the error term.
  - This can be thought of as a sort of "trash can" which contains all of the omitted influences on the Y variable. As an example, it represents the other omitted factors which change the price of the house *other* than the house size.
  - $E(\varepsilon)=0$.
  - standard deviation: The **size** of ε is measured by $[Var(\varepsilon)]^{1/2}$.
- The "systematic" part is given by the term $\beta_0 + \beta_1 X$.
- The conditional expectation of Y given X, E(Y|X), is $\beta_0 + \beta_1 X$.

Linear Regression

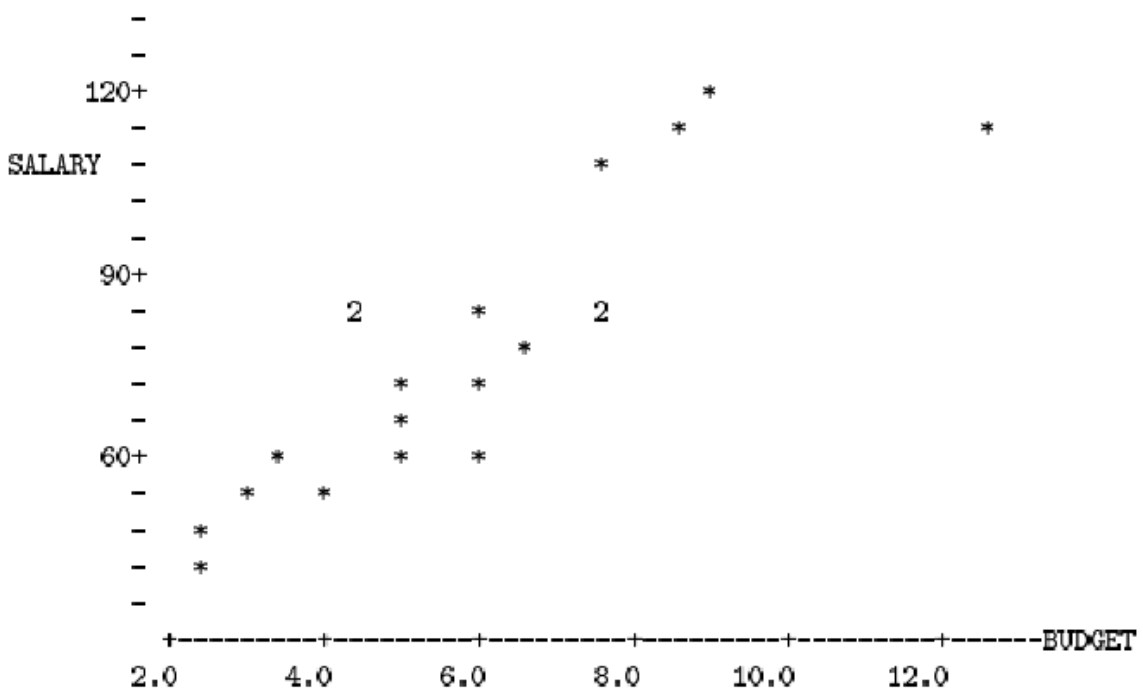
- A regression model species a relation between a dependent variable Y and certain independent variables $X_1,\ldots,X_K$.
- A simple linear regression refers to a model with just one independent variable, K=1.

$$Y = \beta_0 + \beta_1 \mathrm{X} + \varepsilon$$

  - Independent variables are also called explanatory variables; in the equation above, we say that X explains part of the variability in the dependent variable Y.
- Example: A large corporation is concerned about maintaining parity in salary levels across different divisions.
  - As a rough guide, it determines that managers responsible for comparable budgets in different divisions should have comparable compensation.
  - Data Summary

Example

- The following is a list of salary levels ($1000s) for 20 managers and the sizes of the budgets ($100,000s) they manage: (59.0,3.5), (67.4,5.0), (50.4,2.5), (83.2,6.0), (105.6, 7.5), (86.0,4.5), (74.4,6.0), (52.2,4.0), (59.0,3.5), (67.4,5.0), (50.4,2.5), (83.2,6.0), (105.6,7.5), (86.0,4.5), (74.4,6.0), (52.2, 4.0)



Best Line

- Want to fit a straight line to this data.
  - The slope of this line gives the marginal increase in salary with respect to increase in budget responsibility.
- We need to define what we mean by the best line.
  - Regression uses the least squares criterion, which we now explain.
  - Any line we might come up with has a corresponding intercept b_0 and a slope b_1.
  - This line may go through some of the data points, but it typically does not go through all of them.
- The least squares criterion chooses b_0 and b_1 to minimize the sum of squared errors $\Sigma_{1 \le i \le n} (y_i - b_0 - b_1 x_i)^2$ where n is the number of data points.

Least Squares

- For the budget level X_i, the least squares line predicts the salary level

  SALARY = 31.9 + 7.73 BUDGET or $PY_i = 31.9 + 7.73X_i$

  – Unless the line happens to go through the point $(X_i; Y_i)$, the predicted value PY_i will generally be different from the observed value Y_i.

  – Each additional $100,000 of budget responsibility translates to an expected additional salary of $7,730.

  – The average salary corresponding to a budget of 6.0, we get a salary of 31:9 + 7:73(6:0) = 78:28.

  – The difference between the two is the error or residual $e_i = Y_i - PY_i$.

- The least squares criterion chooses b_0 and b_1 to minimize the sum of squared errors $\Sigma_{1 \le i \le n} e_i^2$.

  – A consequence of this criterion is that the estimated regression line always goes through the point $(\overline{X}, \overline{Y})$

Questions

- Q1: Why is the least squares criterion the correct principle to follow?
- Q2: How do we evaluate and use the regression line?
- Assumptions Underlying Least Squares
  - The errors $\varepsilon_1,\ldots,\varepsilon_n$ are independent of the values of $X_1,\ldots,X_n$.
  - The errors have expected value zero; i.e., $E[\varepsilon_i] = 0$.
  - All the errors have the same variance: $Var[\varepsilon_i] = \sigma^2$, for all $i = 1,\ldots,n$.
  - The errors are uncorrelated; i.e., $Corr[\varepsilon_i, \varepsilon_j] = 0$ if $i \neq j$.
- Q1: What are the angle between $(1,\ldots,1)$ and $(\varepsilon_1,\ldots,\varepsilon_n)$ and that of $(\varepsilon_1,\ldots,\varepsilon_n)$ and $(X_1,\ldots,X_n)$?

Discussion on Assumptions

- The first two are very reasonable: if the e_i's are indeed random errors, then there is no reason to expect them to depend on the data or to have a nonzero mean.
- The second two assumptions are less automatic.
  - Do we necessarily believe that the variability in salary levels among managers with large budgets is the same as the variability among managers with small budgets? Is the variability in price really the same among large houses and small houses?
  - These considerations suggest that the third assumption may not be valid if we look at too broad a range of data values.
  - Correlation of errors becomes an issue when we use regression to do forecasting. If we use data from several past periods to forecast future results, we may introduce correlation by overlapping several periods and this would violate the fourth assumption.

Linear Regression

- We assume that the outcome we are predicting depends linearly on the information used to make the prediction.
  - Linear dependence means constant rate of increase of one variable with respect to another (as opposed to, e.g., diminishing returns).
  - E(Y|X) is the population "average" value of *Y* for any given value of *X*. For example, the average house price for a house size = 1,000 sq ft.
- Regression models are really all about modeling the conditional distribution of *Y* given *X*.
  - Distribution of House Price given Size
  - Distribution of Portfolio return given return on market
  - Distribution of wages given IQ or educational attainment
  - Distribution of sales given price

Evaluating the Estimated Regression Line

- Feed data into the computer and get back estimates of the model parameters β_0 and β_1.
  - Is this estimated line any good?
  - Does it accurately reflect the relation between the *X* and *Y* variables?
  - Is it a reliable guide in predicting new *Y* values corresponding to new *X* values?
    - predicting the selling price of a house that just came on the market, or setting the salary for a newly defined position
- Intuitively, the estimated regression line is useful if the points (X_i, Y_i) are pretty well lined up. The more they look like a cloud of dots, the less informative the regression will be.

Reduction of Variability

- Our goal is to determine how much of the variability in Y values is explained by the X values.
  - –We measure variability using sums of squared quantities.
  - –Consider the salary example. The Y_i's (the salary levels) exhibit considerable variability |-not all managers have the same salary.
  - –We conduct the regression analysis to determine to what extent salary is tied to responsibility as measured by budget: the 20 managers have different budgets as well as different salaries.
- What extent the differences in salaries are explained by differences in budgets?

Analysis of Variance Table

- s = 12.14, R^2 = 72.2%, R^2 (adj) = 70.7%

SOURCE	DF	SS	MS
Regression	1	6884.7	6884.7 . . .
Error	18	2651.1	147.3
Total	19	9535.8	

  - DF stands for degrees of freedom, SS for sum of squares, and MS for mean square. The mean squares are just the sum of squares divided by the degrees of freedom: MS = SS/DF.
- A sum of squares measures variability.
  - The Total SS (9535.8) measures the total variability in the salary levels.
  - The Regression SS (6884.7) is the explained variation. It measures how much variability is explained by differences in budgets.
  - The Error SS (2651.1) is the unexplained variation.

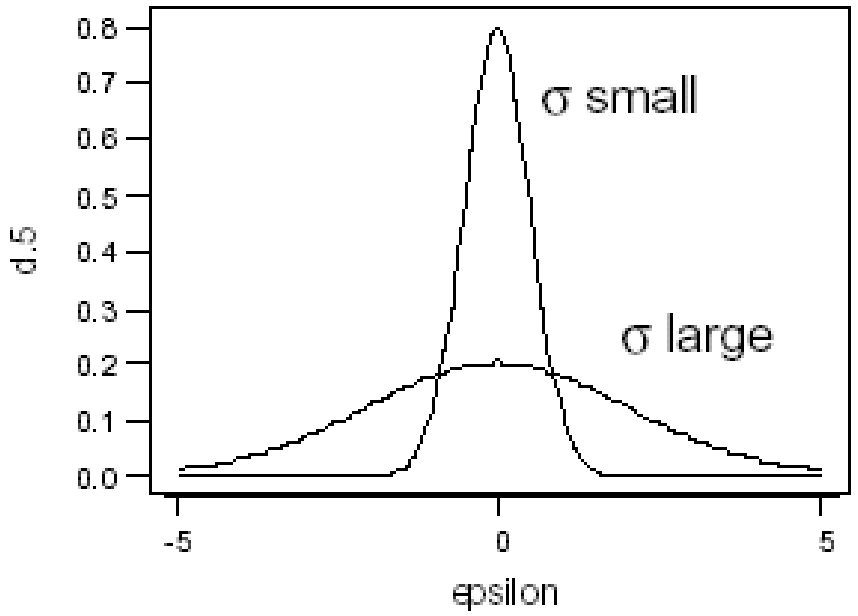
The Error SS

- This reflects differences in salary levels that cannot be attributed to differences in budget responsibilities.
- The explained and unexplained variation sum to the Total SS.
- How much of the original variability has been explained?
  - The answer is given by the ratio of the explained variation to the total variation, which is

    R^2 =SSR/SST=6884.7/9538.8= 72.2%
  - This quantity is the coefficient of determination, though everybody calls it R-squared.

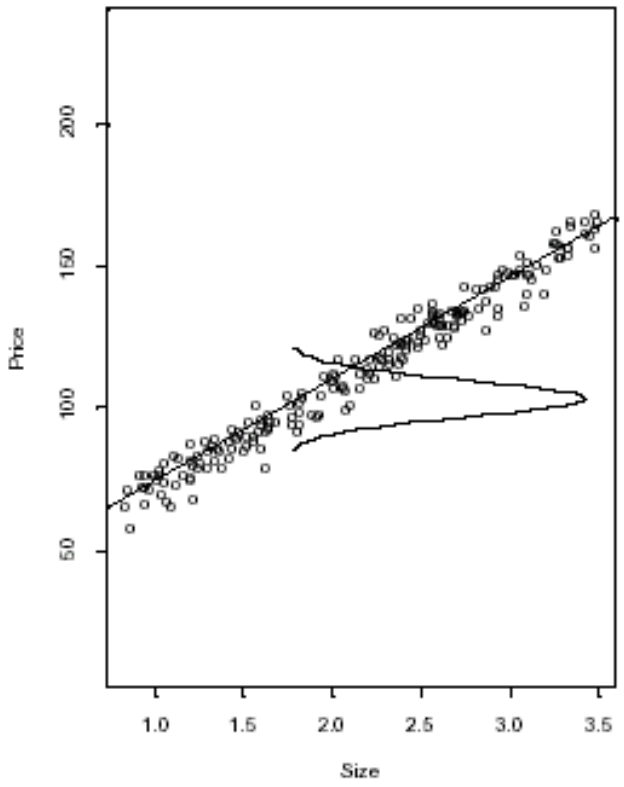
Normal Distribution

- The following figure depicts two different normal distributions both with mean 0 one with σ=.5 one with σ=2.
  - **one σ: 68%, two σ: 95.44%, two σ: 99.7%**

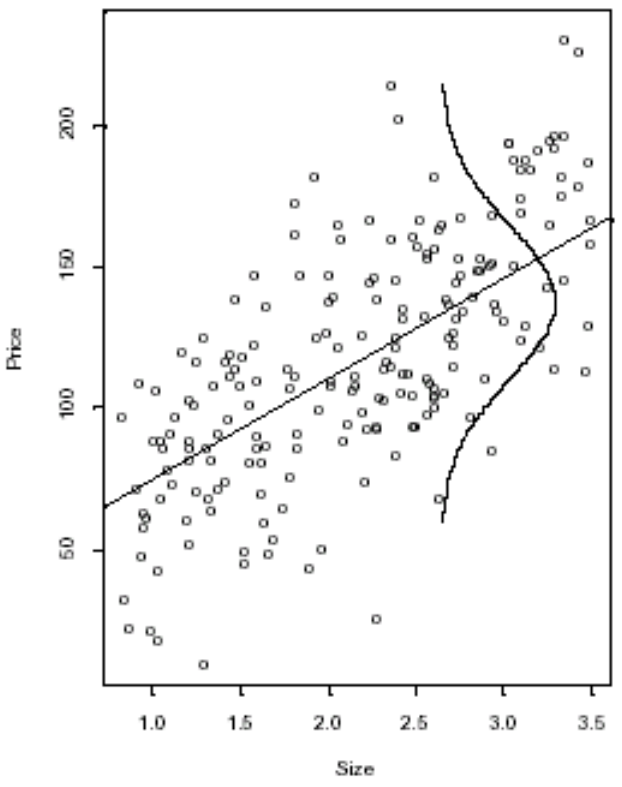


How does σ determine the dispersion of points about the true regression line?

σ small/ε small



σ large/ε large

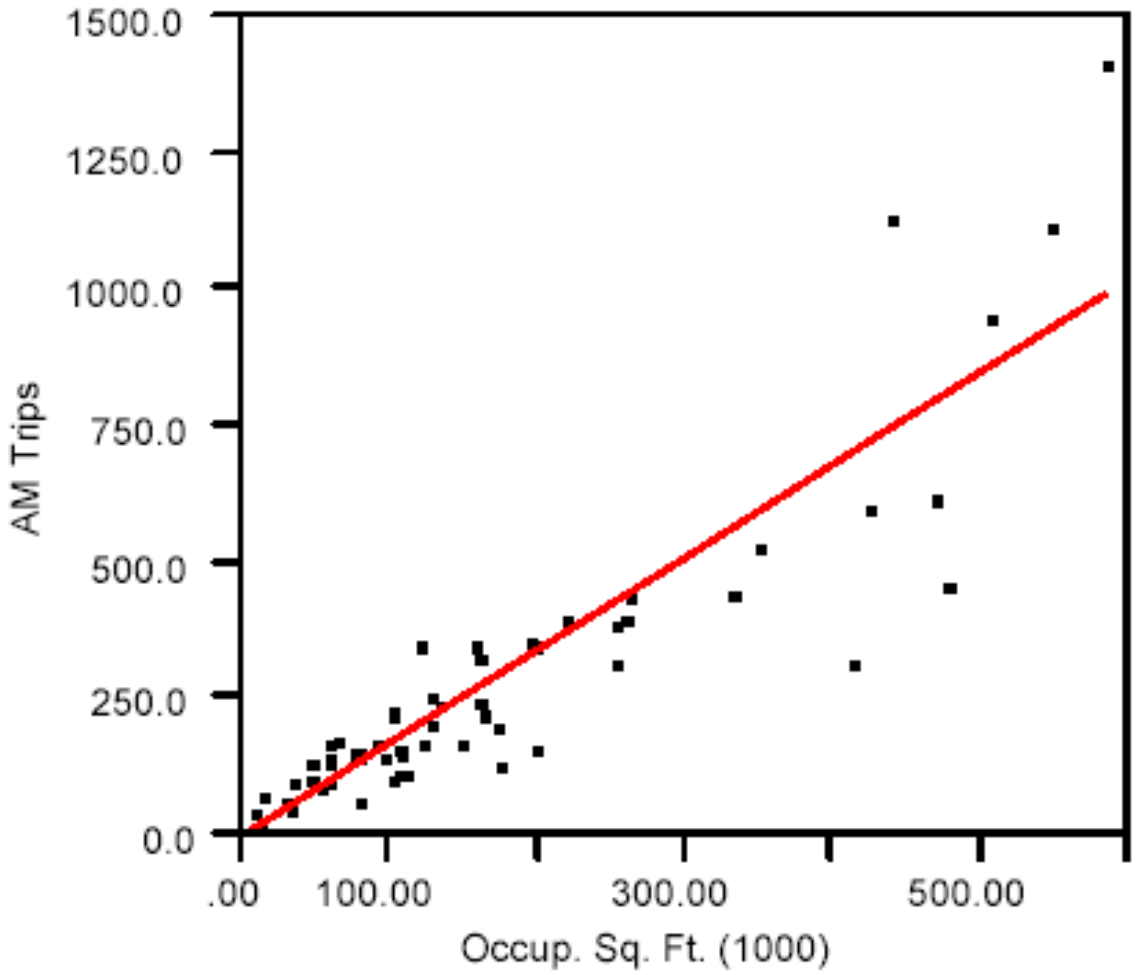


Office Trip Study

- Traffic Planners often refer to results from a classic study done in 1989 for the Maryland Planning Commission by Douglas & Douglas Inc.
  - The study was done in Montgomery County, MD.
  - Goal: Predict the volume of traffic associated with office buildings.
  - Such information is useful for several purposes.
    - For example if a new office building of 350,000 sq. ft. were being planned, planners and zoning administrators, etc., would need to know how much additional traffic to expect after the building was completed and occupied.
- Data AM: traffic counts over a period of time at 61 office building sites within the county.
  - **X-variable**: size of the building measured in occupied gross square feet of floor space (in 1000 sq. ft. units).
  - **Y-variable**: average daily number of vehicle trips originating at or near the building site during the AM hours.
- Data PM: Similar data for PM hours was also measured, and some other information about the size, occupancy, location, etc., of the building was also recorded.

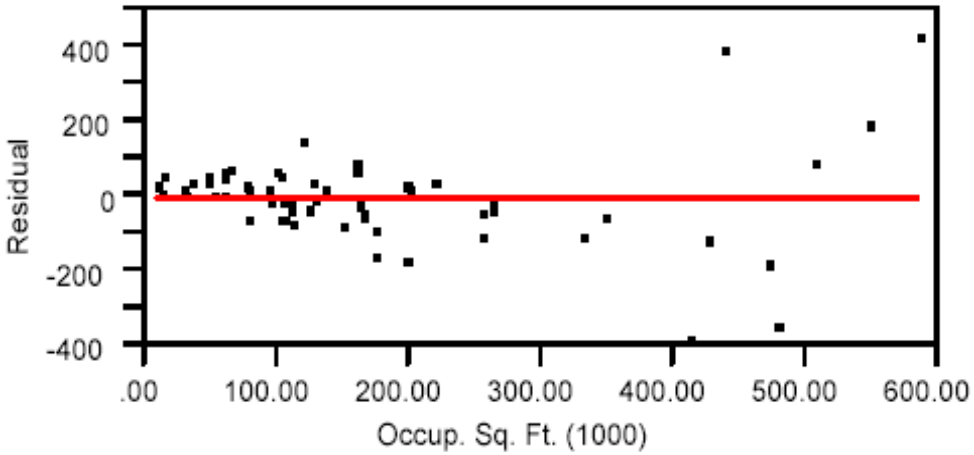
Scatterplot: AM Trips By Occup. Sq. Ft. (1000)

- **Fit**：AM Trips = -7.939 + 1.697 Occup. Sq. Ft. (1000)
- Summary of Fit：$R^2 = 0.800$

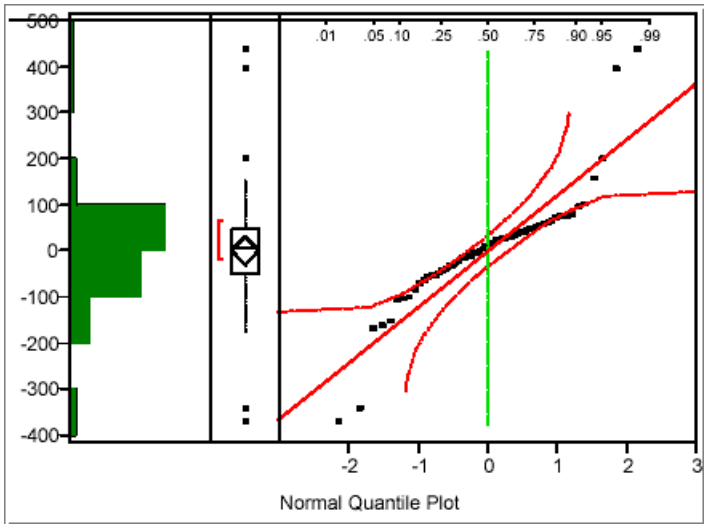


Residual Plot

- How do you know that a correct model is being fitted?
  - Prediction: For a 350,000 sq. ft. bldg, it generates -7.939 + 1.697×350 = 586.0 trips. The 95% confidence interval for this prediction is 535.8 to 636.1.



Noticeable **heteroscedasticity** by looking at scatter plot.

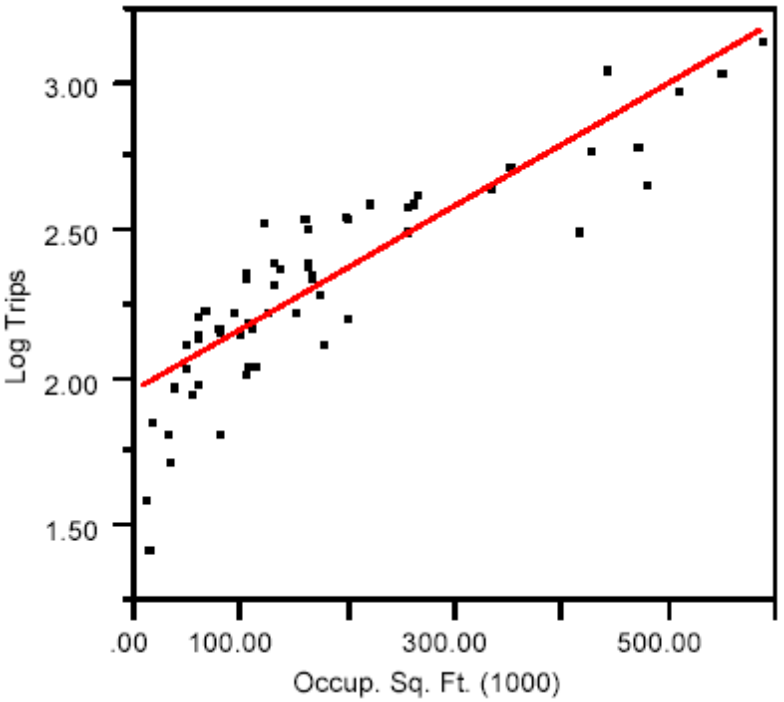


undesirable histogram of residuals

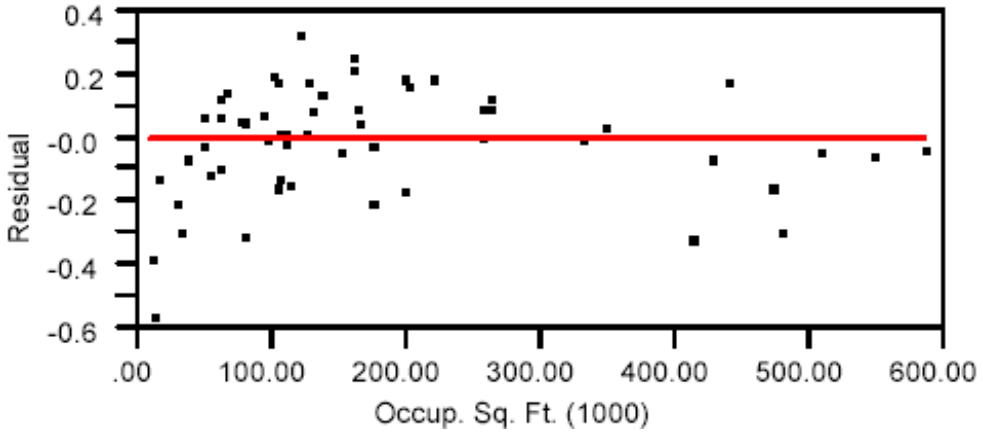
Transformation Attempt #1

- Since the (vertical) spread of the data increases as the *y*-value increases a log transformation may cure this heteroscedasticity.

Scatter plot after transforming *y* to Log(*y*)



Residual Plot



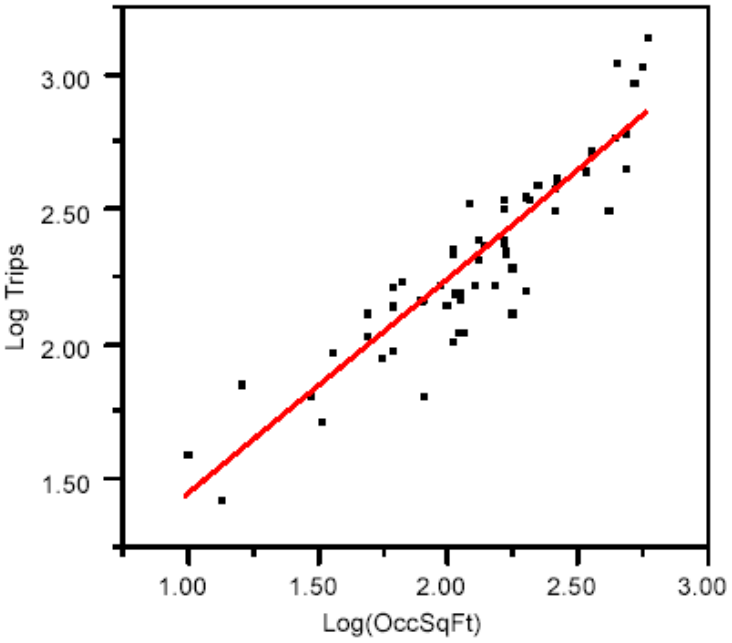
Linear Fit: Log Trips = 1.958 + 0.00208 Occup. Sq. Ft. (1000)
Summary of Fit: $R^2 = 0.761$

New analysis introduces a new problem!

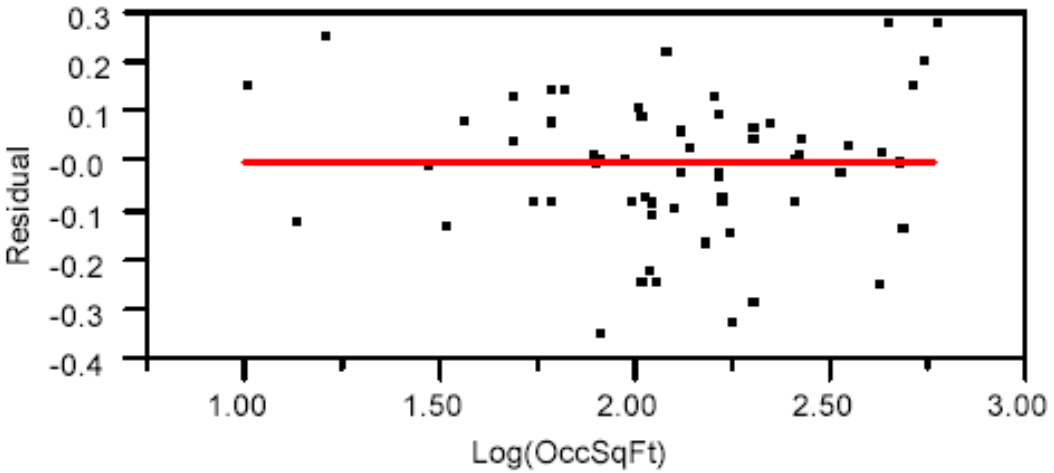
- An undesirable degree of **non-linearity:** It is evident in both the residual plot and the scatterplot.

Transformation Attempt #2

- Try to fix nonlinearity with an additional transformation from *x* to *log*(*x*).



Residual Plot

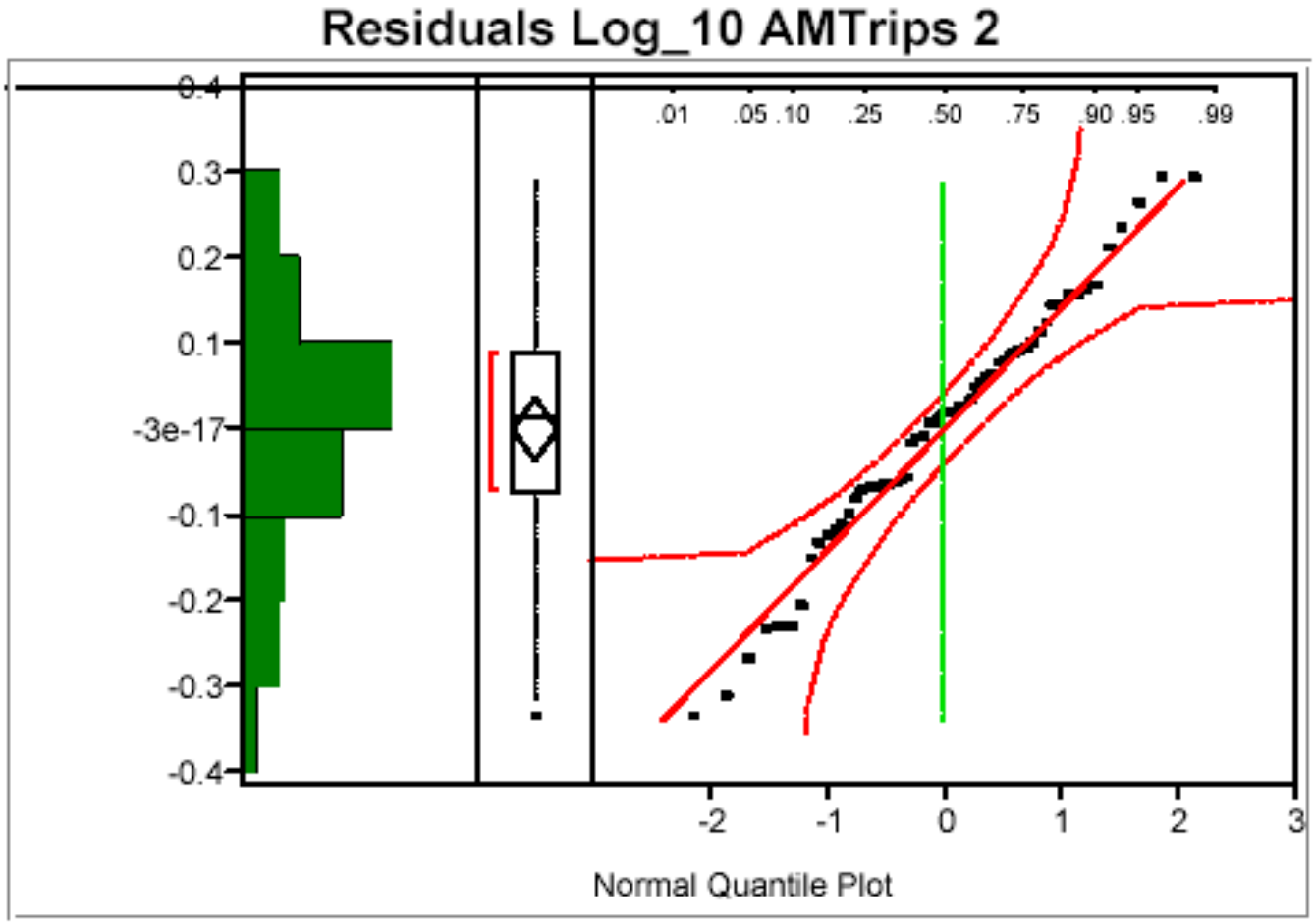


Linear Fit: log Trips = 0.639 + 0.803 log(OccSqFt)

Summary of Fit: $R^2 = 0.827$

Standard Assumptions

- After log-log transformation: Linearity, homoscedasticity, and normality of residuals are all quite OK.



Prediction

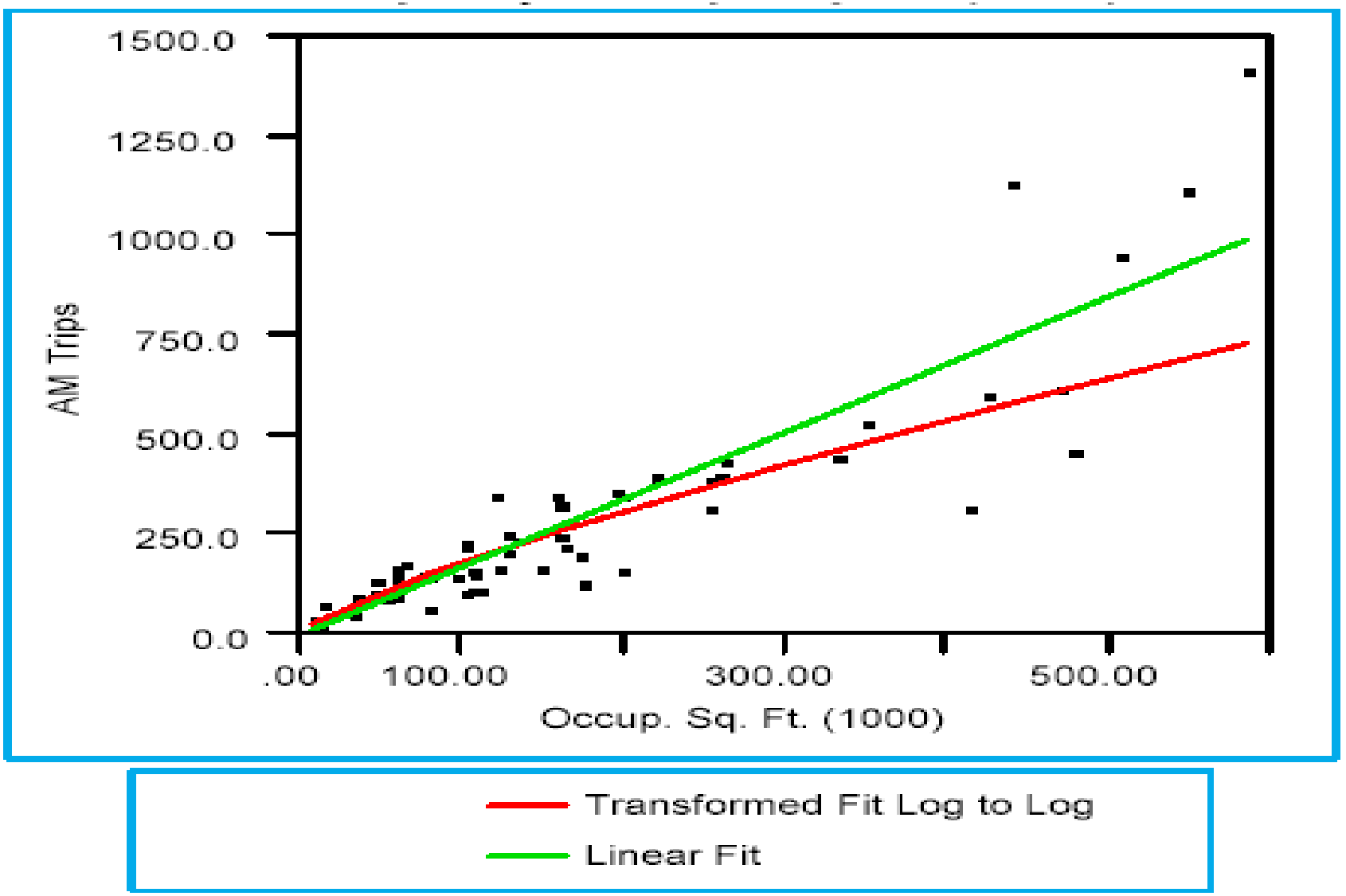
- If a zoning board were in the process of considering the zoning application from a proposed 350,000 sq. ft. office bldg, what is their primary concern?
  - –Proposal I: Find the 95% confidence limits for 350,000 sq. ft. office buildings.

	Lower 95% "Pred"	Upper 95% "Pred"
Log (Trips)	2.6262	2.7399
Number of Trips:	$422.9 = 10^{2.6262}$	549.4

    - •Compare this to the confidence interval of **535.8** to **636.1** from the initial model. These CIs are very different. The latter one, based on the log-log analysis, is the ***valid*** one, since the analysis leading to it is ***valid***.
  - –Proposal II: Consider 95% Individual Prediction intervals - that is, in 95% intervals for the actual traffic that might accompany the particular proposed building. These are

	Lower 95% "Indiv"	Upper 95% "Indiv"
Log (Trips)	2.3901	2.9760
Number of Trips:	245.5	946.2

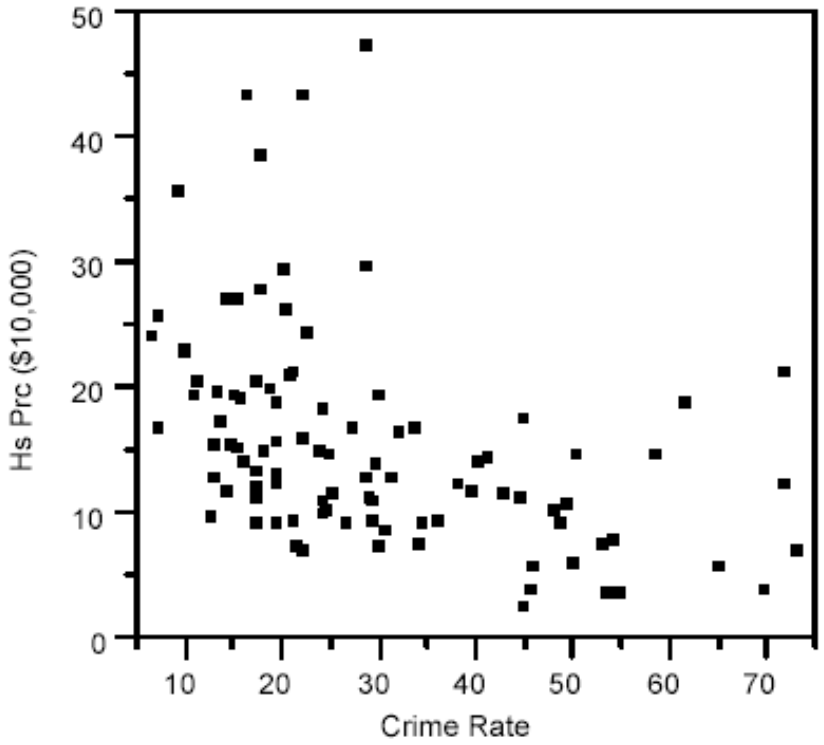
Comparison of the two analyses on a single plot



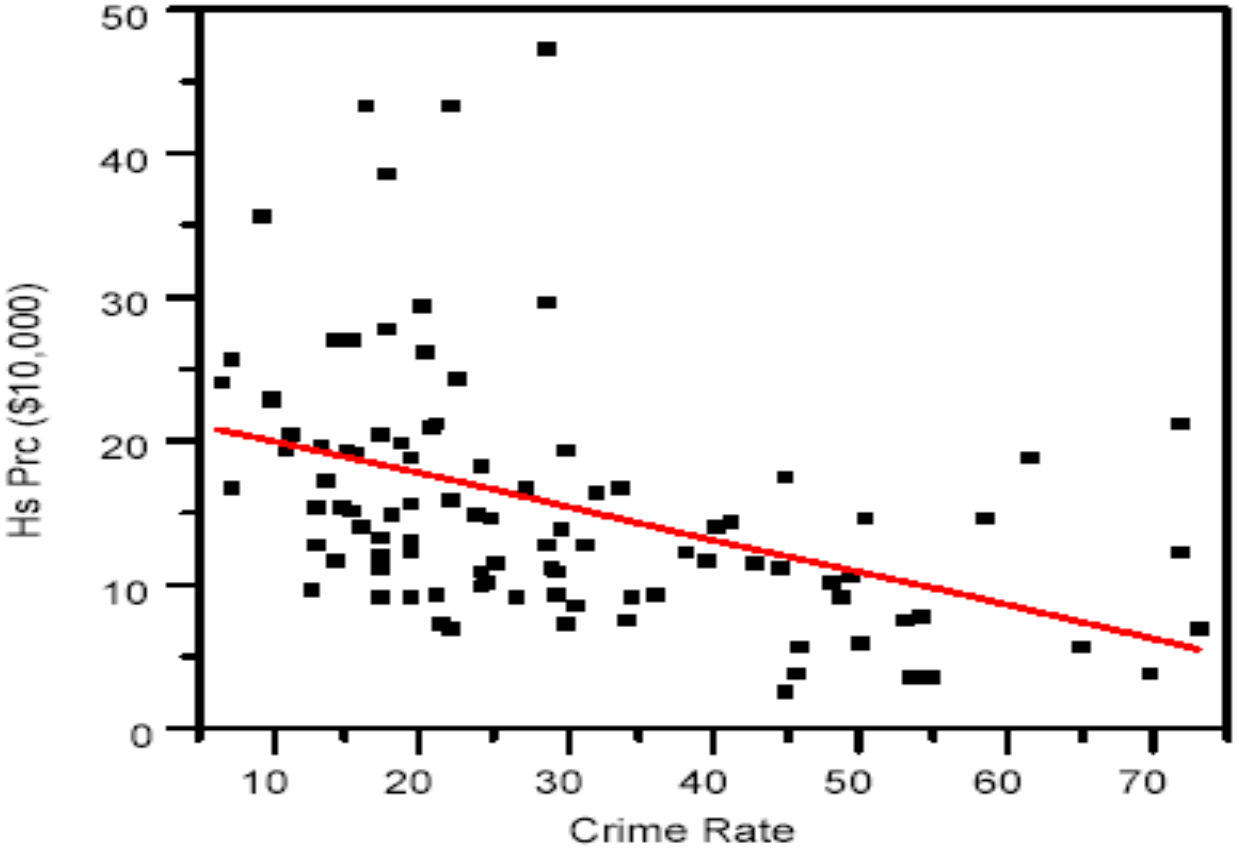
Does an Increasing Crime Rate Decrease House Prices?

- This data was gathered in 1996.
  - –For each community in the Philadelphia area, it gives the crime rate (reported crimes/1000 population/year) and the average sale price of a single family home.
  - –Center City Philadelphia is not included on the following plot and data analyses.

House Price ($10,000) versus Crime Rate



Least squares straight line fit to this data



Summary of Fit

R^2 = 0.184

R^2 Adj = 0.176

Root Mean Square Error = 7.886

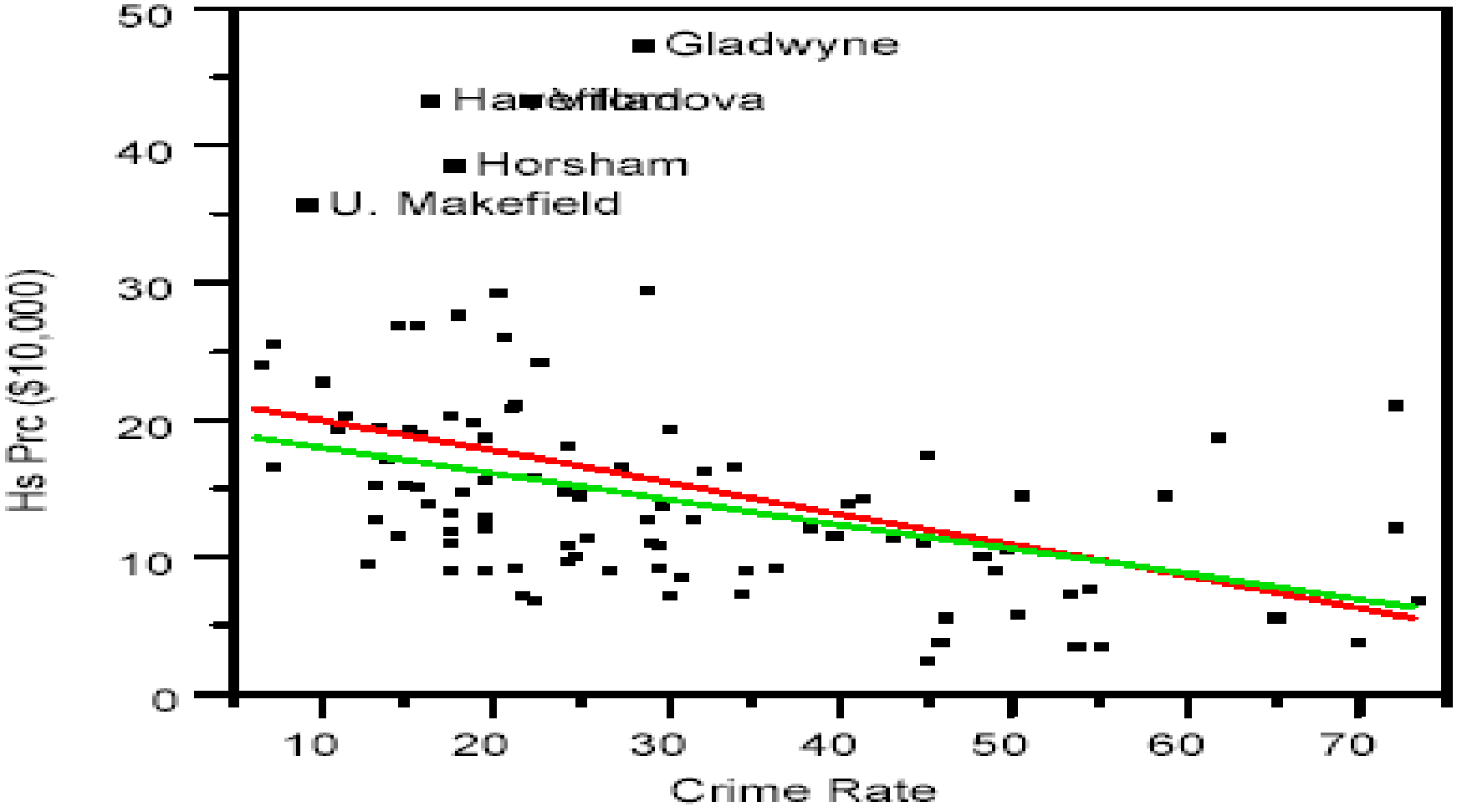
Observations=98

Linear Fit

Hs Prc ($10,000) = 22.53 – 0.229 Crime Rate

(1) Linear fit with all data
(2) Linear fit with five points removed
Linear Fit number **(2)**

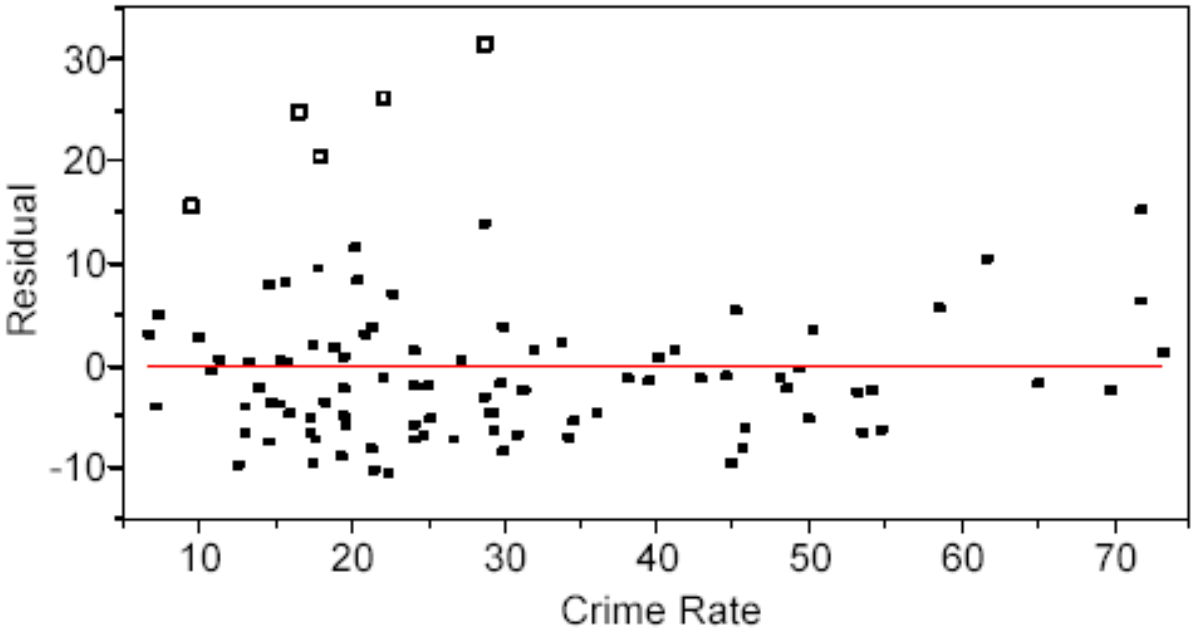
Hs Prc ($10,000) = 19.91 - 0.184 xCrime Rate



Linear Fit Number (1)

- Hs Prc ($10,000) = 22.52 – 0.229×Crime Rate
- Summary of Fit
  - R^2 =0.184, RMSE = 7.886
- Analysis of Variance

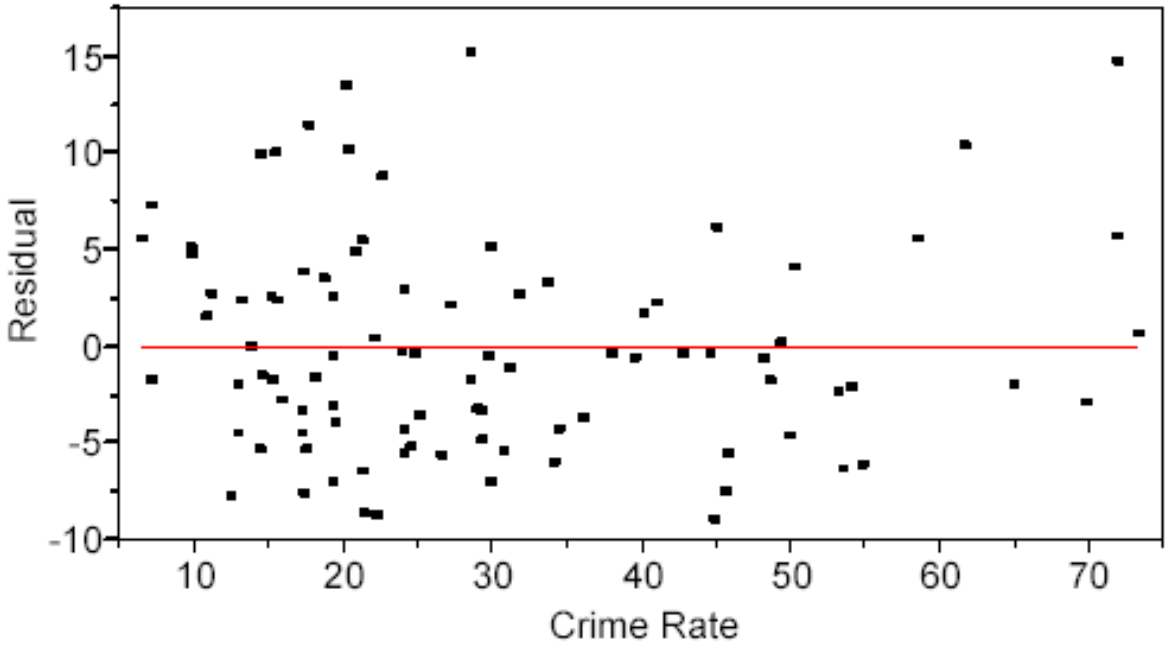
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	22.52	1.649	13.73	<.0001
Crime Rate	-0.229	0.0499	-4.66	<.0001

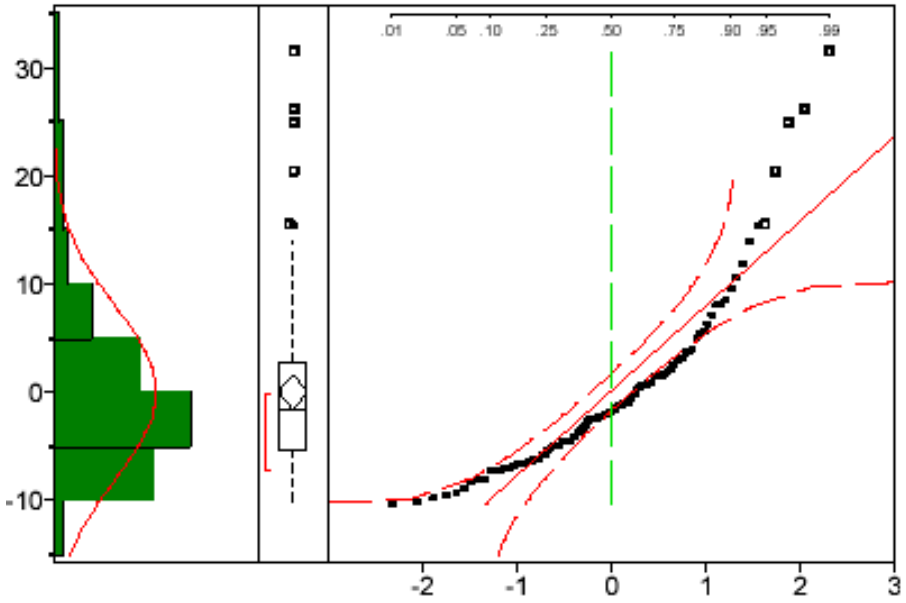


Linear Fit Number (2)

- Hs Prc ($10,000) = 19.91 – 0.184xCrime Rate
- Summary of Fit
  - R^2 =0231, RMSE = 5.556
- Analysis of Variance

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	19.91	1.193	16.69	<.0001
Crime Rate	-0.184	0.0351	-5.23	<.0001

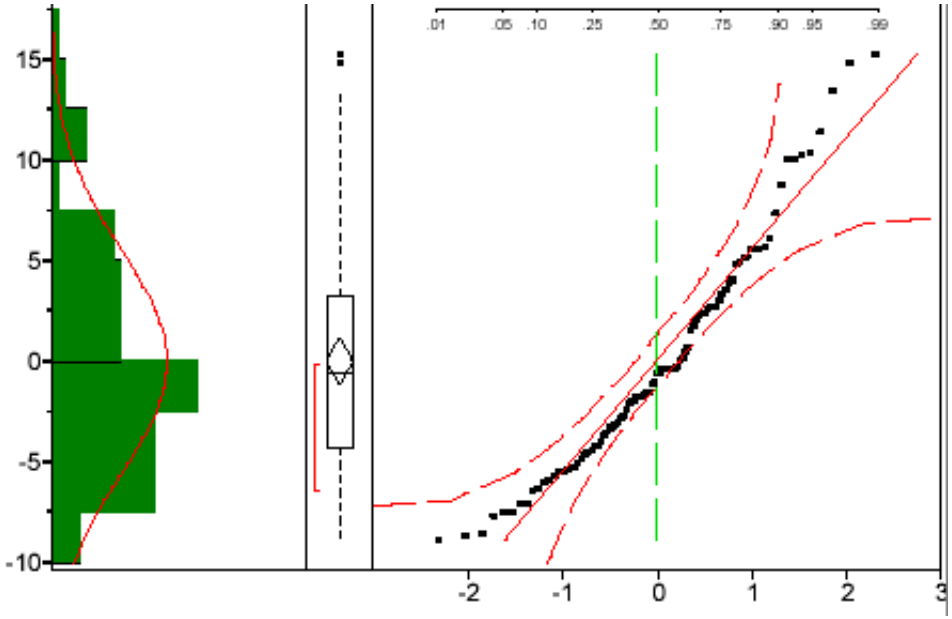




Normal Quantile Plot

Residuals Hs Prc ($10,000); Analysis 1

Residuals Hs Prc ($10,000); Analysis 2



Normal Quantile Plot

Analysis of Car Mileage Data

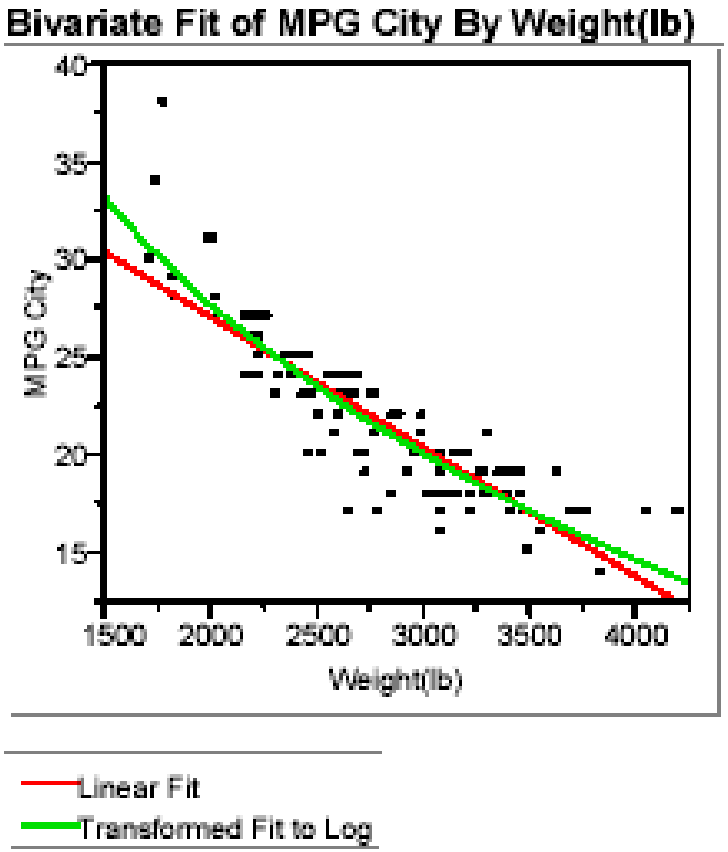
- Data set: It gives mileage figures (in MPG (City)) for various makes of cars, along with various characteristics of the car engine and body as well as the base price for the car model.
- Objective:
  - Create a multiple regression equation that can be used to predict the MPG for a car model having particular characteristics.
  - Get an idea as to which characteristics are most prominent in relation to a car mileage.
- Build a regression model:
- Step 1: Examine the data

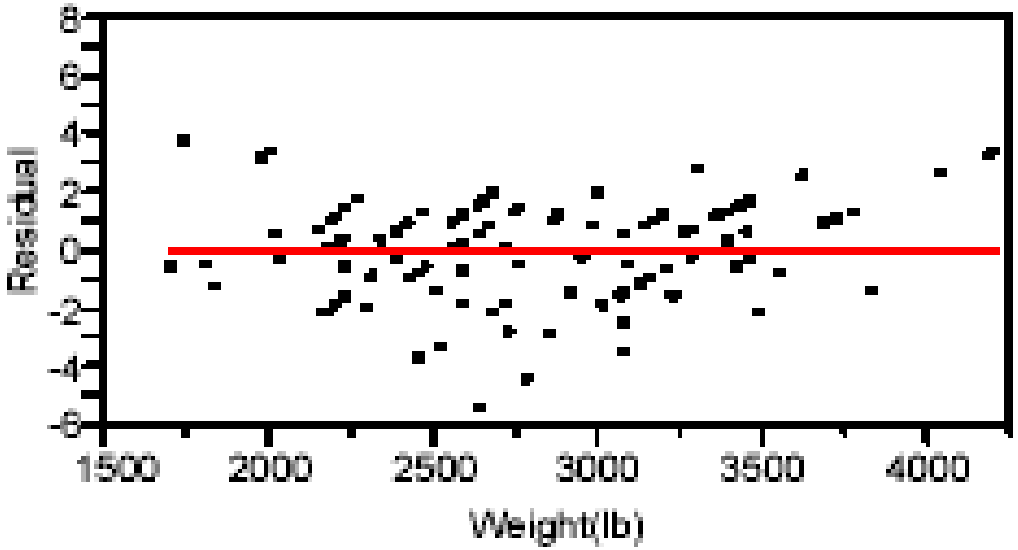
  a. Look for outliers and influential points.

  b. Decide whether to proceed by predicting *Y* or some function of *Y*.

  – Preliminary analysis: Consider *X* to be weight.
    - Linear Fit: MPG City = 40.266964 - 0.006599 Weight(lb)
    - Transformed Fit to Log: MPG City = 171.42143 - 18.898537 Log(Weight(lb))
    - R^2 changes from 0.75093 to 0.803332.
    - RMSE changes from 2.136611 to 1.898592.

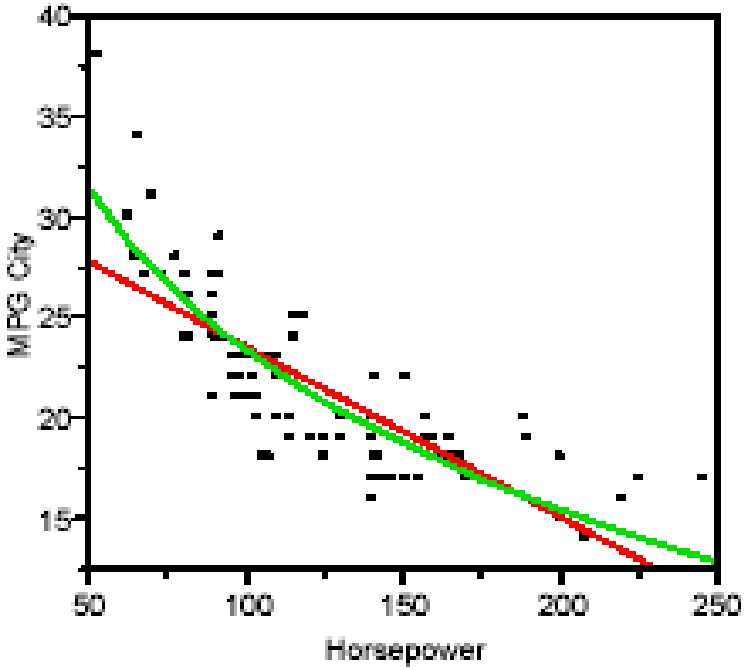
Log(Weight) version provides a somewhat better fit.
A ***slight curve*** is evident in the residual plot (and in the original scatter plot if you look carefully).
No outliers or influential points that seem to warrant special attention.



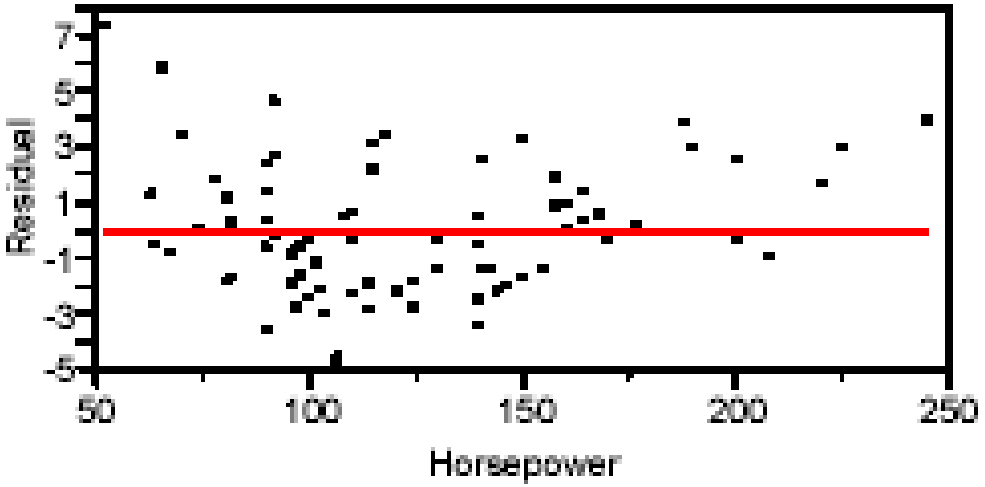


Another Fit

- Fit of MPG City By Horsepower
  - –MPG City = 32.057274 - 0.0843973 Horsepower
  - –MPG City = 76.485987 - 11.511504 Log(Horsepower)



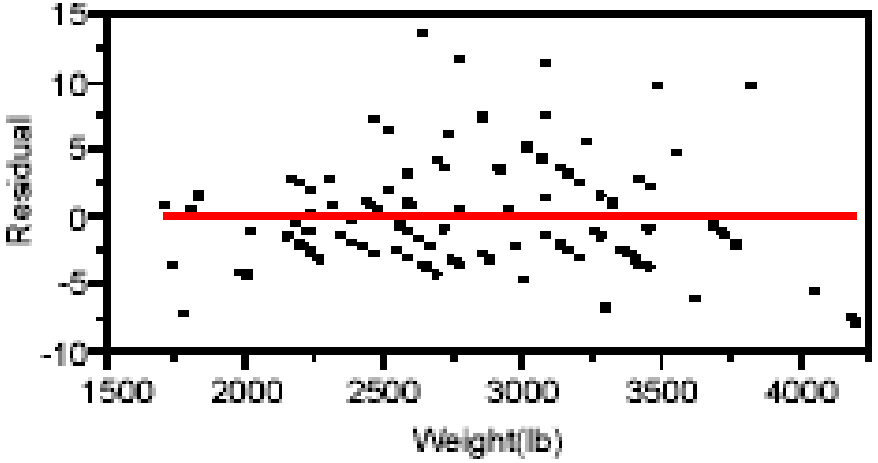
--- Linear Fit
--- Transformed Fit to Log



This suggests that we might want to try also transforming the Y-variable somehow, in order to remove the remaining curved pattern.

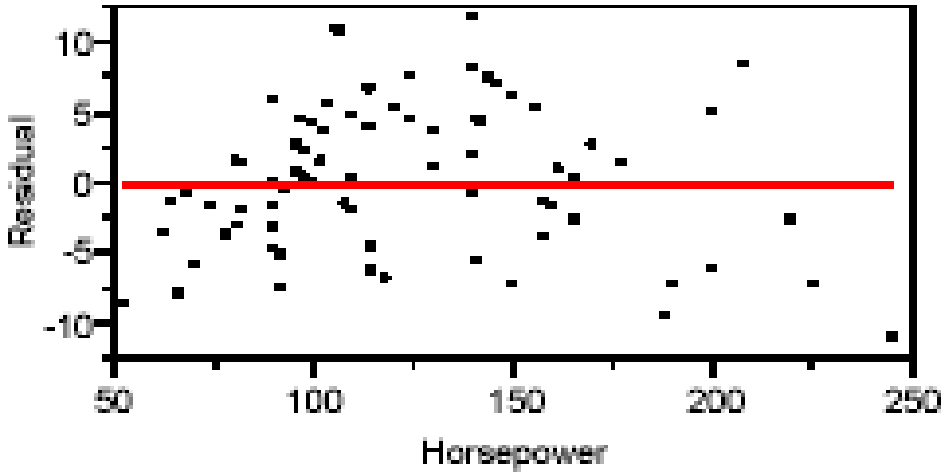
Step 1b: Try transformations of Y

- One transformation that has been suggested is to transform Y to $1/Y$ = **G**allons **P**er **M**ile.
  - –Since this is a rather small decimal, we consider Y* = **G**allons **P**er 1000**M**iles.
  - –Linear Fit: GP1000M City = 9.2452733 + 0.0136792 Weight(lb)
  - –Another Fit: GP1000M City = 25.559124 + 0.1806558 Horsepower



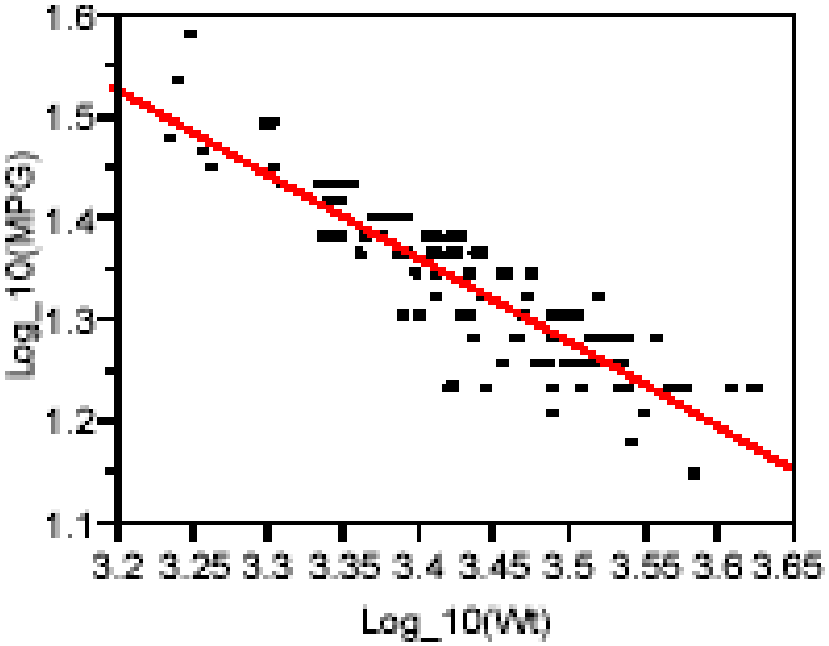
R^2 = 0.774344
RMSE = 4.151475

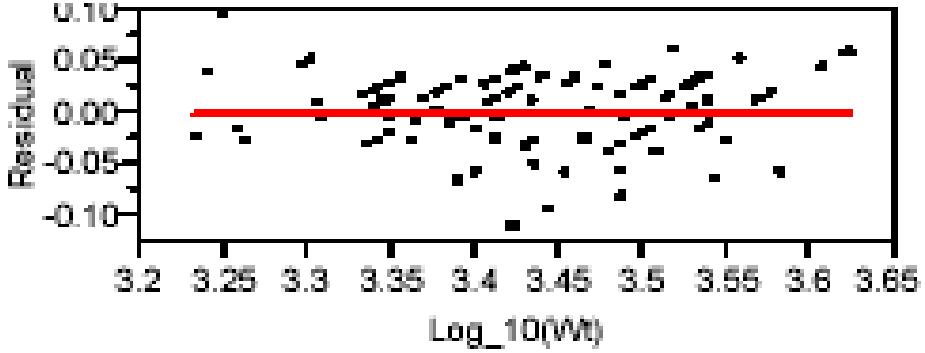
R^2 = 0.705884
RMSE = 4.739567



Transformation

- The first analysis looks nicely linear, but there is some evident heteroscedasticity.
- The second analysis seems to be slightly curved; maybe we could try using log(HP) as a predictor.
- It seems reasonable to also try transforming to Log(Y) = Log_{10} (MPG).
  - Since MPG was *nearly* linear in Wt, it seems more reasonable to try Log_{10} (Wt) as a predictor here, and similarly for Log_{10} (HP).

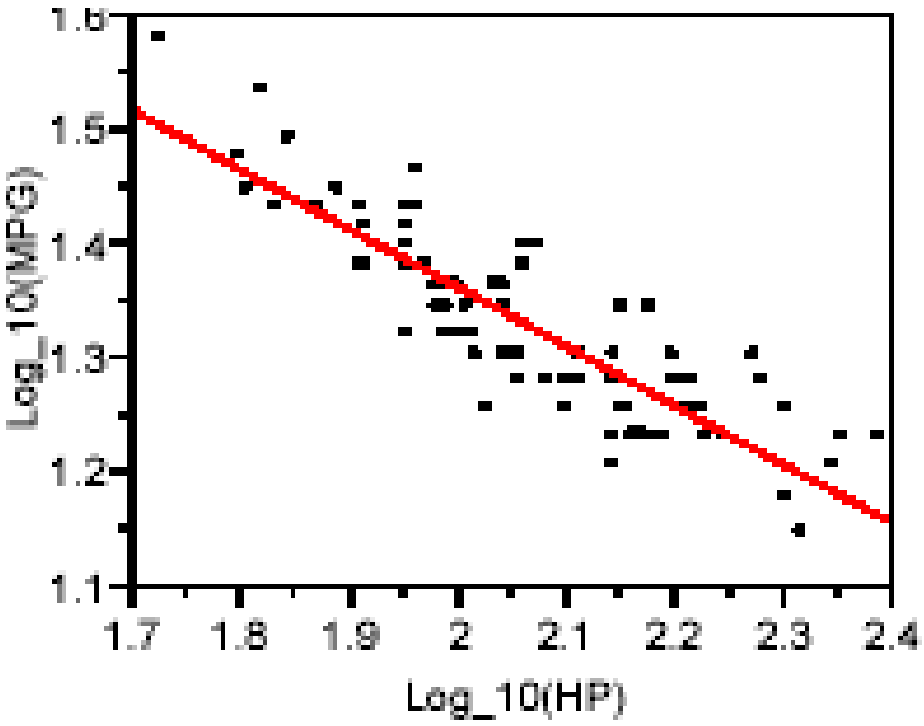


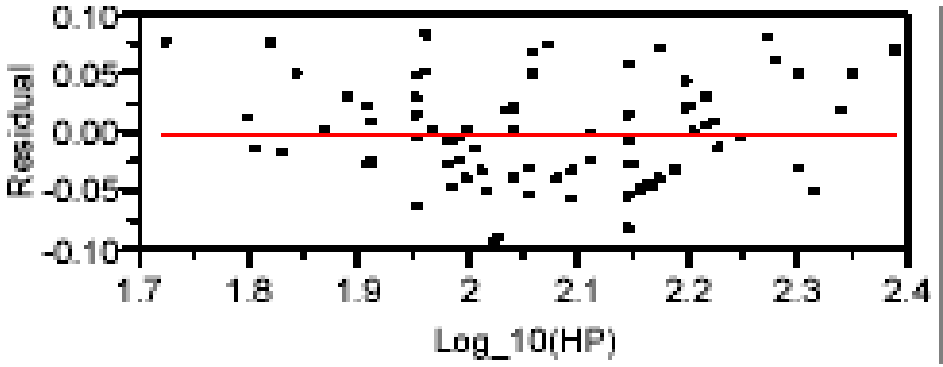


$$Log_{10}(MPG) = 4.2147744 - 0.8388428\ Log_{10}(Wt)$$

Transformation

- Linearity looks fine on these plots.
- There may be a hint of heteroscedasticity - but not close to enough to worry about.
- Again, there are no leverage points or outliers that seem to need special care.
- $\text{Log}_{10}(\text{MPG}) = 2.3941295 - 0.5155074\ \text{Log}_{10}(\text{HP})$





Step 2: Choose predictor variables for the Multiple Regression

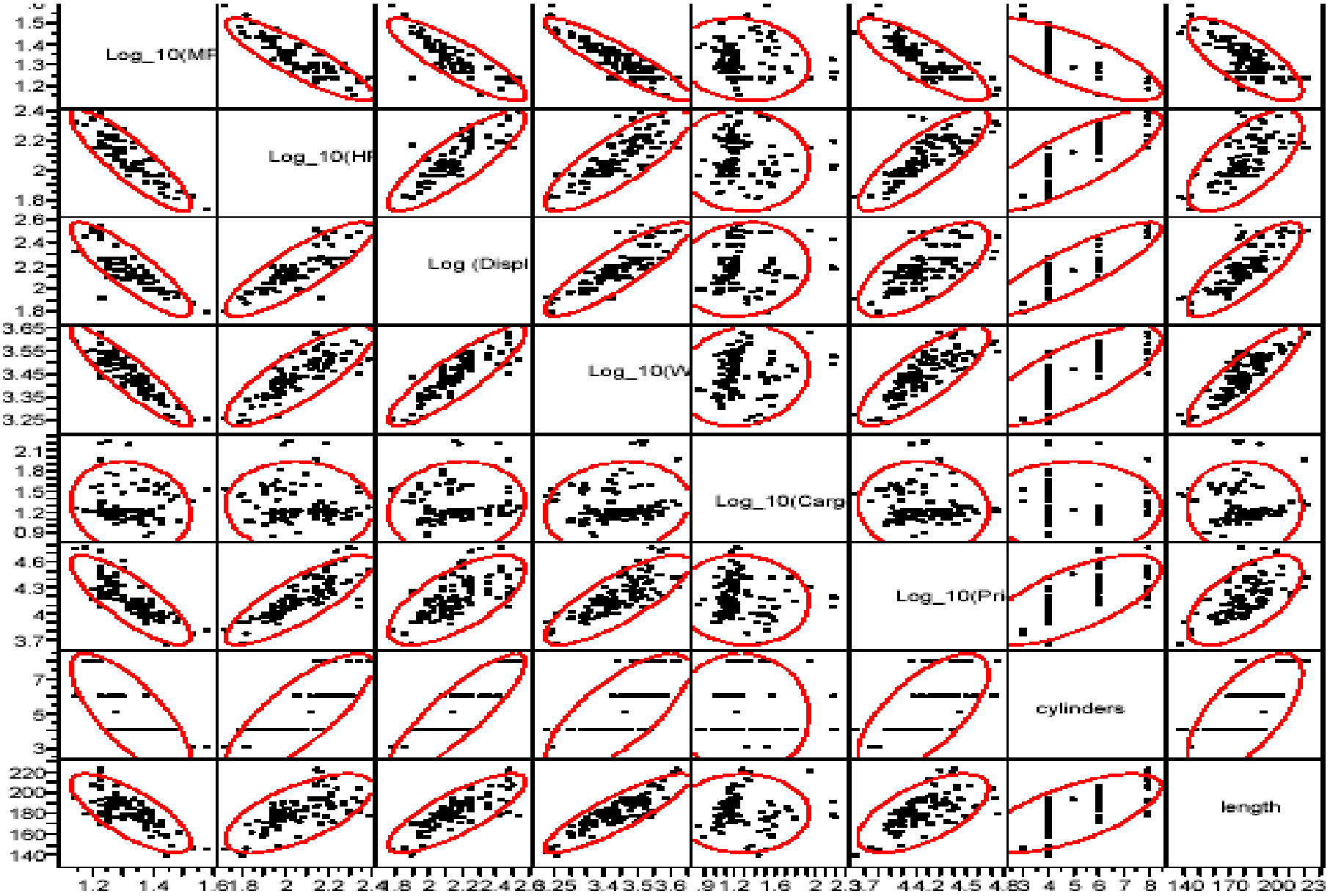
- Use the chosen form of Y variable and of X variables - and perhaps other X-variables as well.

Multivariate Correlations

	Log_10(MPG)	Log_10(HP)	Log (Displ)	Log_10(Wt)	Log_10(Cargo)	Log_10(Price)	cylinders	length
Log_10(MPG)	1.0000	-0.8791	-0.8624	-0.9102	-0.1361	-0.8208	-0.7572	-0.7516
Log_10(HP)	-0.8791	1.0000	0.8530	0.8260	-0.0440	0.8335	0.7976	0.6592
Log (Displ)	-0.8624	0.8530	1.0000	0.8809	0.1060	0.6757	0.8700	0.8073
Log_10(Wt)	-0.9102	0.8260	0.8809	1.0000	0.1449	0.8055	0.7483	0.8768
Log_10(Cargo)	-0.1361	-0.0440	0.1060	0.1449	1.0000	-0.0684	-0.0225	0.0020
Log_10(Price)	-0.8208	0.8335	0.6757	0.8055	-0.0684	1.0000	0.6860	0.6221
cylinder	-0.7572	0.7976	0.8700	0.7483	-0.0225	0.6860	1.0000	0.6724
length	-0.7516	0.6592	0.8073	0.8768	-0.0020	0.6221	0.6724	1.0000

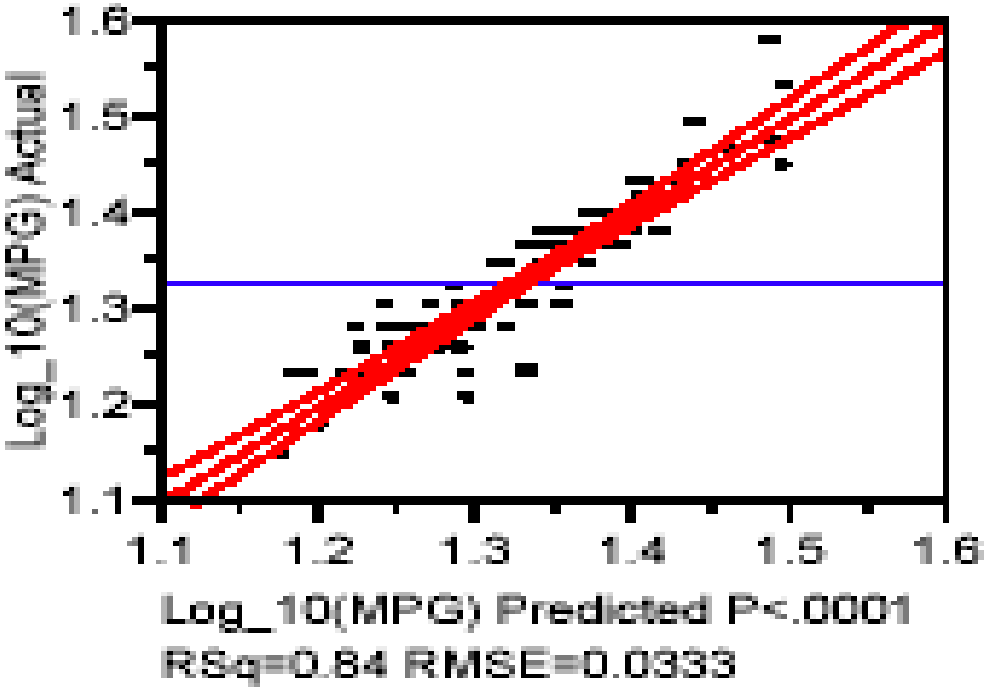
- The biggest correlation is with Log(Wt). Therefore this variable gives the best fitting linear regression.
  - –Note that MPG City = 171.42143 - 18.898537 Log(Weight(lb))
  - –In that analysis, the SSE was 0.1327.
  - –Add Log(HP) to this model as another predictor. This creates a multiple regression with two predictor variables.
  - –MPG City = 3.61 – 0.513 Log(Weight(lb)) –0.25Log(HP)
  - –Choose one (or sometimes more) variables as the most important among the predictor variables. The best single choice is that having the largest correlation with the y-variable.

Scatterplot Matrix

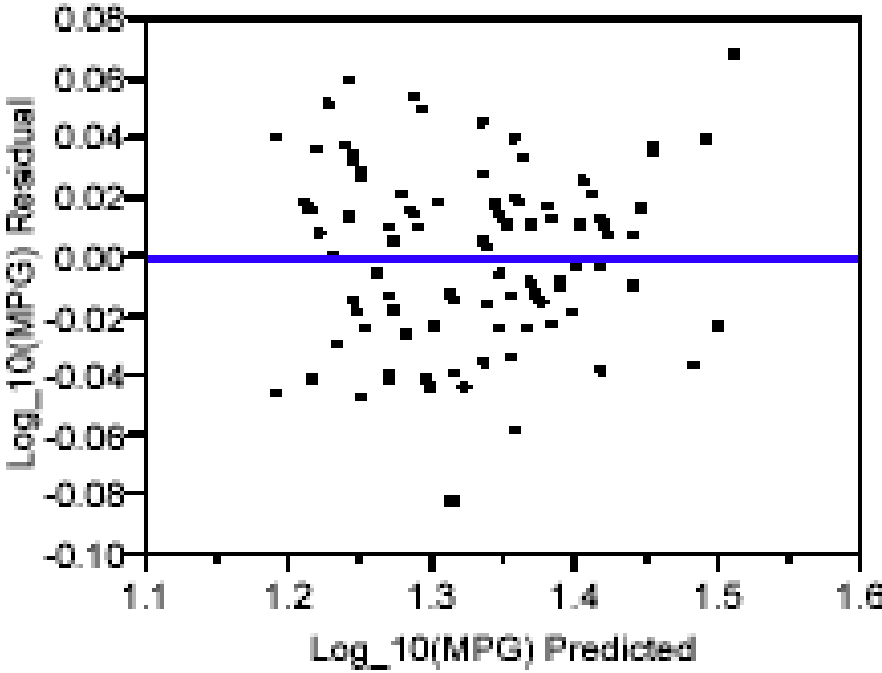


Plots

Actual by Predicted Plot



Residual by Predicted Plot



Software

- MATLAB
  - Many free “toolboxes” on the Web for regression and prediction
  - e.g., see http://lib.stat.cmu.edu/matlab/
    and in particular the CompStats toolbox

- R
  - General purpose statistical computing environment (successor to S)
  - Free (!)
  - Widely used by statisticians, has a huge library of functions and visualization tools

- Commercial tools
  - SAS, Salford Systems, other statistical packages
  - Various data mining packages
  - Often are not progammable: offer a fixed menu of items