



# Simple Linear Regression

## Part 2: Model Assumptions

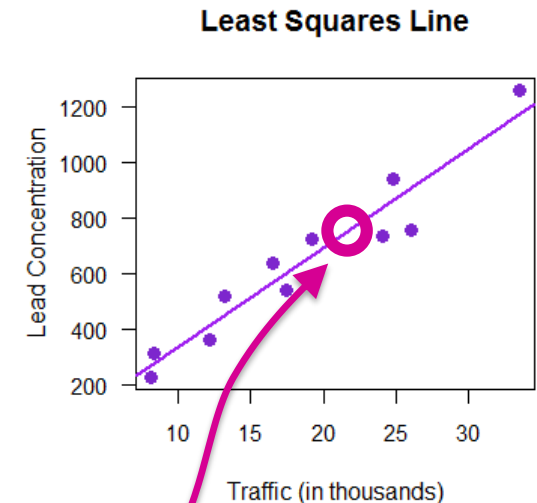
STAT 705: Regression and Analysis of Variance

# Recall

- For the Lead vs. Traffic example
  - estimated intercept = -21
  - estimated slope = 35.7
- You should know how to calculate these values
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \Leftrightarrow \text{Lead} = -21 + 35.7 \times \text{Traffic}$

# Estimated Least Squares Line

- $\text{Lead} = -21 + 35.7 * \text{Traffic}$
- Suppose there is another site that has traffic 22 (thousand vehicles)
- How much lead would we expect to see in the tree bark at this site?
  - $\text{Lead} = -21 + 35.7 * 22 = 764.4$   
micrograms of lead per gram of bark
  - This is the point on the line at  $X = 22$ .



# A Trick Question

- Could we estimate the lead concentration for a site that has 50,000 vehicles?
- We can calculate it
  - $\text{Lead} = -21 + 35.7 \times 50 = 1764$  micrograms per gram of bark
- But does it make sense?
- Short answer: It does not make sense.
  - The traffic values in the data set range from 8.1 to 33.6 (thousand)
  - 50 thousand vehicles is MUCH larger than the largest value in the data
  - We cannot assume the relationship between Traffic and Lead remains the same when Traffic is much larger than the observed values
  - This is EXTRAPOLATION, and it must be avoided

# Model vs. Estimated Line

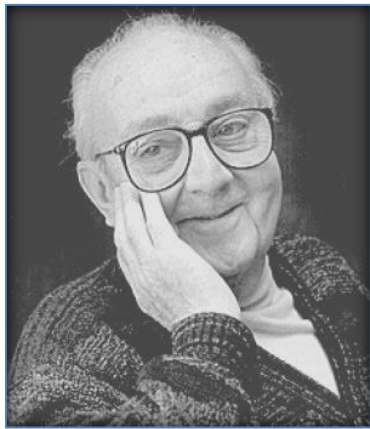
- Regression model:  $Y = \beta_0 + \beta_1 X + \varepsilon$ 
  - Applies to the entire population (e.g., all the sites that could possibly be selected along the highway)
  - $\beta_0$  and  $\beta_1$  are population parameters that describe the “true” relationship between  $X$  and  $Y$  for ALL items in the population
- Estimated regression line:  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ 
  - This is our estimate of the “true” relationship
  - To assess whether or not our estimated line is close to the ‘true’ line, we can compare  $\hat{\beta}_0$  to  $\beta_0$ ,  $\hat{\beta}_1$  to  $\beta_1$ , and  $\hat{Y}$  to  $Y$

# Are There Other Lines?

- Other criteria can be used to construct lines that fit the data
- Least Squares Lines have many desirable statistical properties
- Unless there is a specific reason to use a different method, Least Squares is preferred

# Statistical Model

- Is a conceptualization of a real process
- Is a simplification of a much more complex phenomenon
- The data provide clues about the process
- For some data sets, there might be many “good” models
- For other data sets, finding even one good model can be difficult



*“All models are wrong,  
but some are useful.”*

George E. P. Box  
(1921 – 2013)

# Model Assumptions

- Model:  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Anything that is random has a probability distribution
  - $\beta_0$ ,  $\beta_1$  and  $X_i$  are fixed (not random)
  - $\varepsilon_i$  and  $Y_i$  are random

**ASSUME:  $\varepsilon_i \sim \text{NIID}(0, \sigma^2)$**

- NIID is shorthand for Normally, Identically, and Independently Distributed
- We need to check that this assumption is reasonably satisfied ... or at least not grossly violated!



# Implications of the Assumptions

Model:  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Assume: (1)  $\varepsilon_i \sim \text{NIID}(0, \sigma^2)$  and (2)  $\beta_0, \beta_1$  and  $X_i$  are constants

Recall

- The expected value (mean) of a constant is the constant
- The variance of a constant is 0

Assumption	Implication
$E(\varepsilon_i) = 0$	$E(Y_i) = \beta_0 + \beta_1 X_i$
$\text{Var}(\varepsilon_i) = \sigma^2$	$\text{Var}(Y_i) = \text{Var}(\varepsilon_i) = \sigma^2$
$\varepsilon_i$ 's are independent	$Y_i$ 's are independent
$\varepsilon_i$ 's are normal	$Y_i$ 's are normal



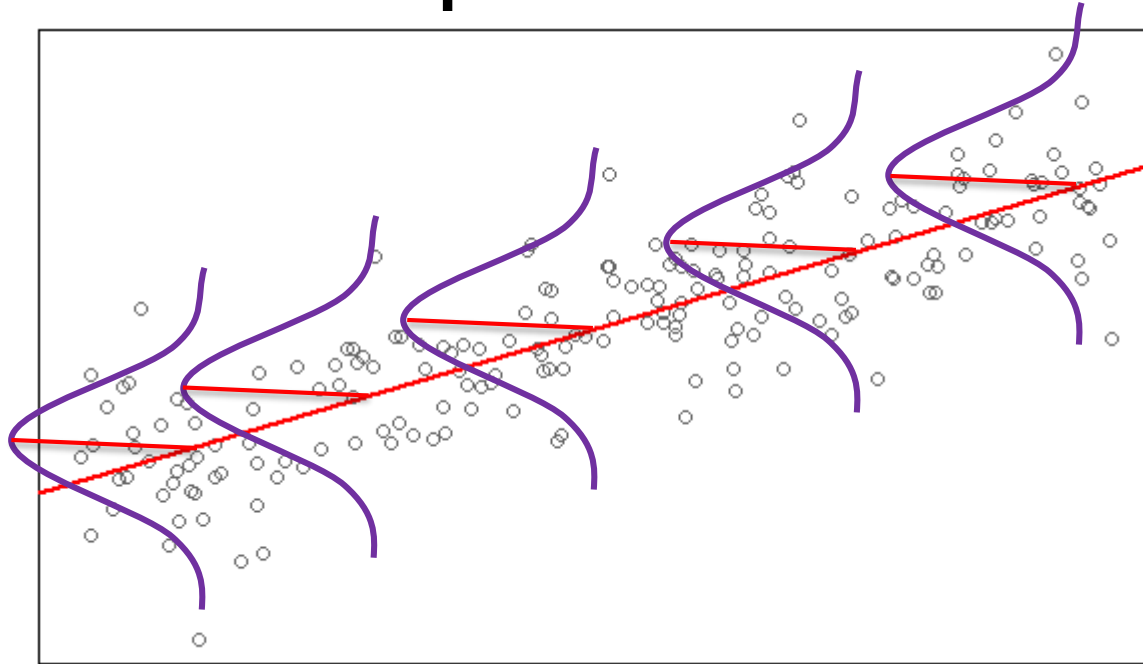
# Implications of Model Assumptions

- In shorthand,

$$\varepsilon_i \sim \text{NIID}(0, \sigma^2) \Rightarrow Y_i \sim \text{NID}(\beta_0 + \beta_1 X_i, \sigma^2)$$

- Notes:
  - “NIID” stands for a normal, independent, distribution (not identical)
  - $Y$  has a probability distribution
  - The mean of  $Y$  depends on  $X$
  - The variance of  $Y$  does not depend on  $X$

# Implications



Mean of  $Y$  is  $\beta_0 + \beta_1 X$  (depends on  $X$ )  
(This is the  $Y$  value on the line for the specific  $X$ )

Variance of  $Y$  is  $\sigma^2$  (does NOT depend on  $X$ )  
(Each normal curve has exactly the same width)

# More Implications

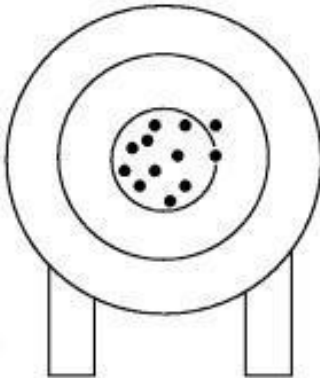
## Gauss-Markov Theorem:

“Under the conditions of the linear regression model, the least squares estimators for  $\beta_0$  and  $\beta_1$  are unbiased and have minimum variance among all unbiased linear estimators.”

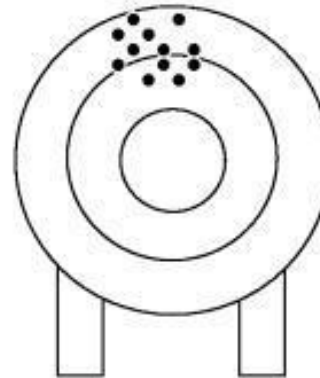
This is the statistical ‘gold standard’ for estimators.

# Unbiased, Minimum Variance

Low Bias, Low Variance

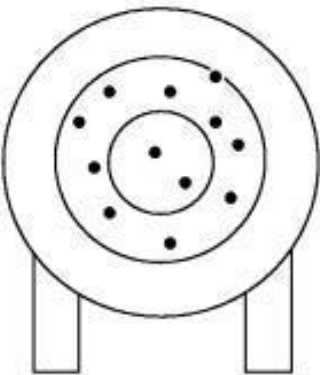


High Bias, Low Variance

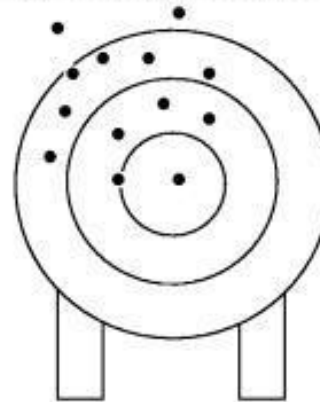


Reliable,  
but not  
accurate

Low Bias, High Variance



High Bias, High Variance



Not reliable,  
not accurate

Reliable and  
accurate

**BEST**

Accurate  
(on average),  
but not reliable

Image source: <http://www.amstat.org/publications/jse/v11n2/martin.html>

# Reasons for Performing Regression

- Get point estimates for parameters
  - Intercept and slope ( $\beta_0$  and  $\beta_1$ )
  - Error variance ( $\sigma^2$ )
- Inference on parameters
  - Confidence intervals (i.e., plausible values)
  - Hypothesis tests
- Estimate the mean of  $Y$  for a given  $X$
- Predict a new value of  $Y$  for a given  $X$

# Point Estimates

- Intercept : The point estimate for  $\beta_0$  is  $\hat{\beta}_0$ 
  - This is the estimate for  $Y$  when  $X = 0$
  - This may be nonsensical. Example: For a location that has zero traffic, do we really expect -21 micrograms of lead?
- Slope : The point estimate for  $\beta_1$  is  $\hat{\beta}_1$ 
  - Average change in  $Y$  for each one-unit increase in  $X$
  - Example: If the traffic increases by one (thousand vehicles) then, on average, we expect the lead to increase by 35.7 micrograms per gram of bark.

# Point Estimate for $\sigma^2$

- $\varepsilon_i \sim \text{NIID}(0, \sigma^2)$
- $\sigma^2$  is the variability in the Y values that is NOT explained by the regression equation
- It is based on the sum of squared errors (residuals)
- Total variability around the line =  $\text{SSE} = \sum_{i=1}^n r_i^2$
- Average variability around the line =  $\text{MSE} = \frac{\text{SSE}}{n-2}$
- Point estimate for  $\sigma^2$  is  $\hat{\sigma}^2 = \text{MSE}$
- For Lead vs. Traffic example,  $\text{MSE} = 7867.2$  (see next slide)



# Calculate MSE for Lead Example

Site (i)	Traffic (X)	Lead (Y)	Est'd Lead	Residual	Resid^2
1	8.1	227	268.17	-41.17	1695.0
2	8.3	312	275.31	36.69	1346.2
3	12.1	362	410.97	-48.97	2398.1
4	13.2	521	450.24	70.76	5007.0
5	16.5	640	568.05	71.95	5176.8
6	17.5	539	603.75	-64.75	4192.6
7	19.2	728	664.44	63.56	4039.9
8	24.8	945	864.36	80.64	6502.8
9	24.1	738	839.37	-101.37	10275.9
10	26.1	759	910.77	-151.77	23034.1
11	33.6	1263	1178.52	84.48	7136.9

$$\text{Est'd Lead} = -21 + 35.7 * \text{Traffic}$$

$$\text{Residual} = \text{Lead} - \text{Est'd Lead}$$

$$\text{Sum this column} \\ \text{SSE} = 70,805.1$$

$$\text{MSE} = 70,805.1 / 9 \\ \text{MSE} = 7867.2$$

# Things You Should Know

- The difference between the regression model and the estimated equation
- The assumptions of the linear regression model
- How to calculate and interpret
  - a point estimate for the slope
  - a point estimate for the intercept
  - the mean of  $Y$  for a given  $X$
  - the predicted value of  $Y$  for a given  $X$
  - a point estimate for the residual variance