3.1 Simple Linear Regression

October 24, 2014

Linear Regression

- Supervised learning
  - Useful for predicting a quantitative response
- Examples:
  1. Is there a relationship between budget and sales?
  2. Which media contribute to sales?
  3. Are these relationships linear?
- Other examples?

Simple Linear Regression

- Predicting quantitative Y based on a single predictor X

$$Y \approx \beta_0 + \beta_1 X.$$

Intercept Slope

β_o and β_1 = Model Coefficients/Parameters
Unknown

Simple Linear Regression

- We can use the training equation to produce estimates for β_o and β_1 and then can predict future outcomes (Y) using this equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

Estimating the Coefficients

- We want to estimate the slope and intercept so it is close as possible to the “true” data or outcome
- Using advertising dataset with 200 different markets
  - Budget
  - Product sales
- Want to obtain coefficient estimates so that the linear model fits the data well and approximates the data well
  - Most common approach is to minimize the least squares

Residual Sum of Squares

- Prediction for Y based on the *i*th value of X

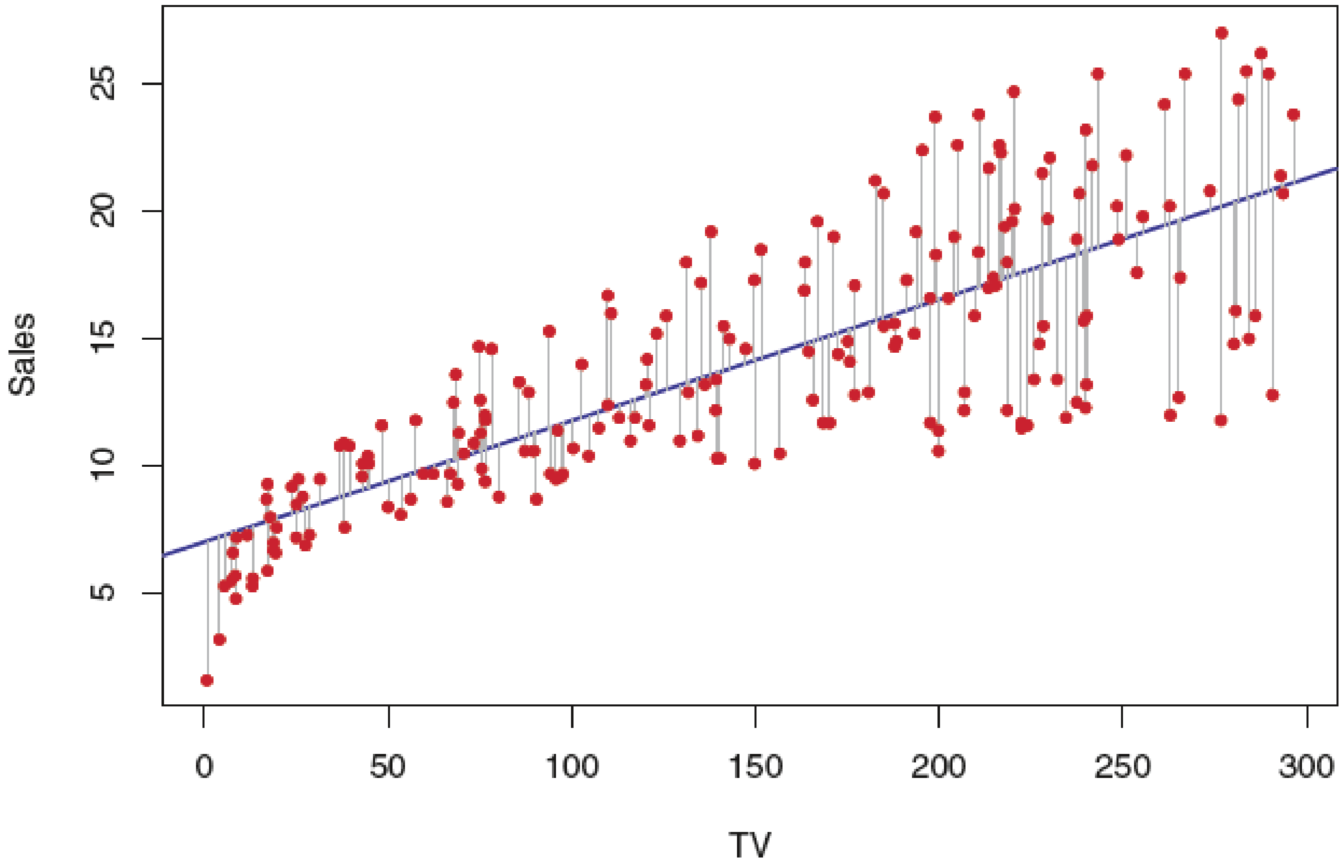
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- e is the difference between *i*th observed response and the *i*th value that is predicted by the mode (Residual)

$$e_i = y_i - \hat{y}_i$$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

Least Squares to minimize the RSS



Accuracy of Coefficient Estimates

True Relationship between X and Y

For an unknown function (*f*) where e is a mean-zero random error

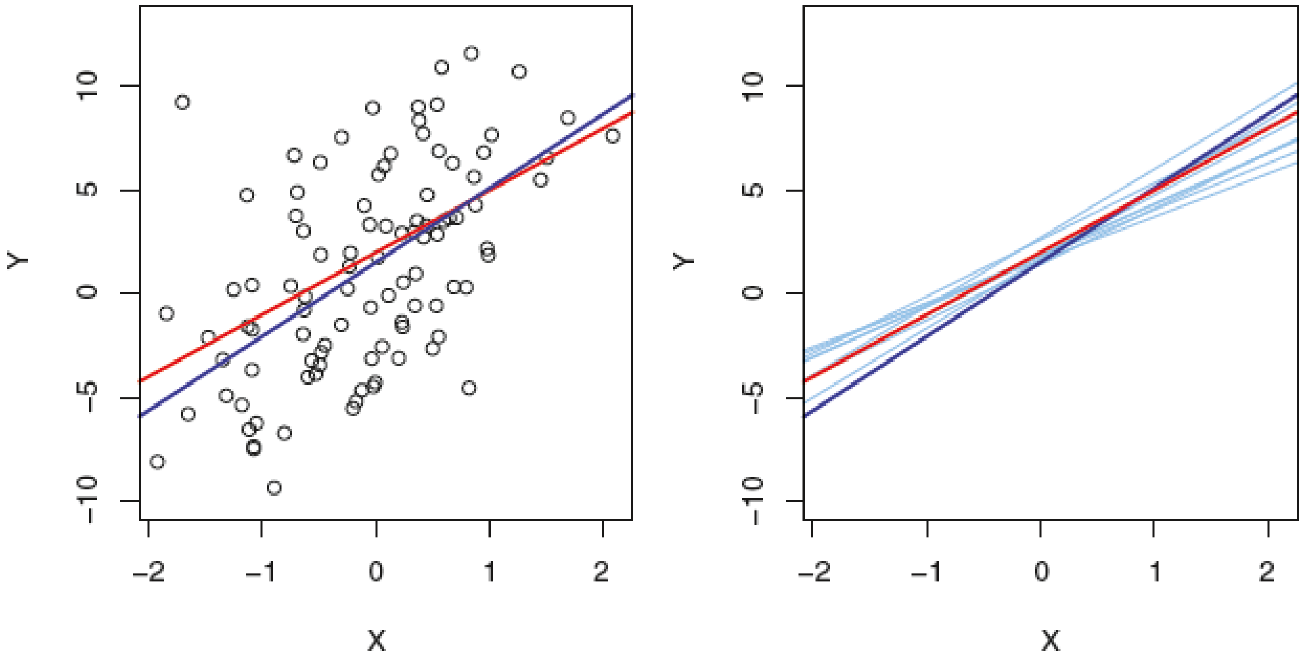
For *f* to be estimated by a linear function:

βo (intercept) = expected Y when X=0

β1 (slope) = average increase in Y with a 1-unit increase in X

Population Regression Line:

$$Y = \beta_0 + \beta_1 X + \epsilon.$$



Red Line = True Relationship f(X) = 2 + 3X (population regression line)
Dark Blue = Least Squares Line

Right Graph = light blue lines that represent 10 different least squares
(on average the least squares are close to the population regression line)

$$Y = 2 + 3X + \epsilon,$$

Standard Error and Confidence Intervals

- Standard Error: Difference between the population mean and the estimate of the population mean

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

- 95% Confidence Intervals: A range that with 95% probability will be a true estimate of a parameter

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

Slope

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0).$$

Intercept

Hypothesis Testing

- Standard errors can be used to perform hypothesis tests

- Null Hypothesis: Ho: $\beta 1 = 0$

  no relationship between X and Y

- Alternative Hypothesis: H1: $\beta o \neq 0$

  Some relationship between X and Y

Hypothesis Testing

- When β1 = 0 the model reduces to Y = βo + ε (X is not associated with Y)
- How do we know the β1 is "far" from zero so we are confident that it is NOT zero?

T-statistic: number of standard deviations the estimate of β1 is away from zero

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

P-value: the probability of observing any value = |t| assuming β1 = 0

Final Model from Advertising Data

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

- Least Squares Model: the number of units sold with the TV advertising budget
- Coefficients of estimates: β0 = 7 and βo = 0.05 are large compared to the Std Error
- Large difference between estimates and std error = large t-statistic
- Therefore, we can reject the null hypothesis that βo = 0 or β1 = 0

Assessing model accuracy

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Assessing model accuracy

- Residual standard error (RSE): average amount the response will deviate from the true regression line
  - Up to the scientist to decide whether the deviation is acceptable
  - In the case of our advertising example the RES = 3.26 or 3,260 units. Is this ok?
- R^2 statistic: proportion of variance explained by the model
  - Will always be between 0 and 1
  - Proportion of variability in Y that can be explained by X
  - Advertising data: R2 is 0.61, is this ok?

Lab

- Boston dataset
- Median house value (medv) for 506 neighborhoods in Boston
- Predict medv using 13 different variables
  - Rm: average number of rooms per house
  - Age: average age of house
  - Lstat: percent of households with low SES

Lab: relationship between medv and lstat

- Coefficients:

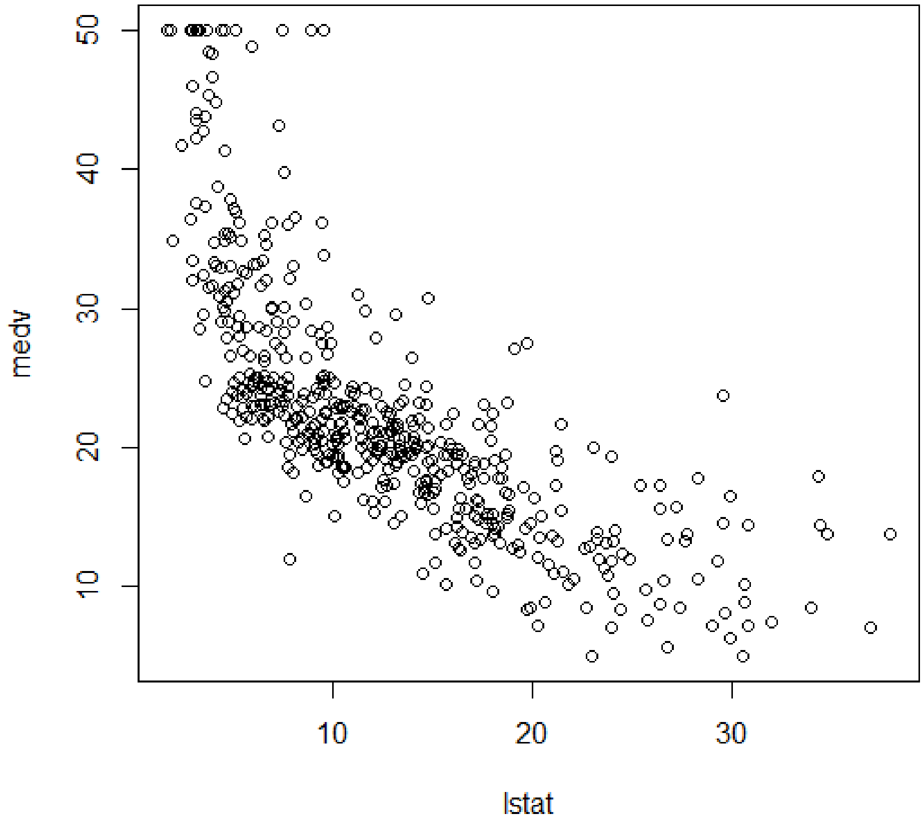
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.55384	0.56263	61.41	<2e-16 ***
lstat	-0.95005	0.03873	-24.53	<2e-16 ***

- Residual standard error: 6.216 on 504 degrees of freedom
- Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432
- F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16

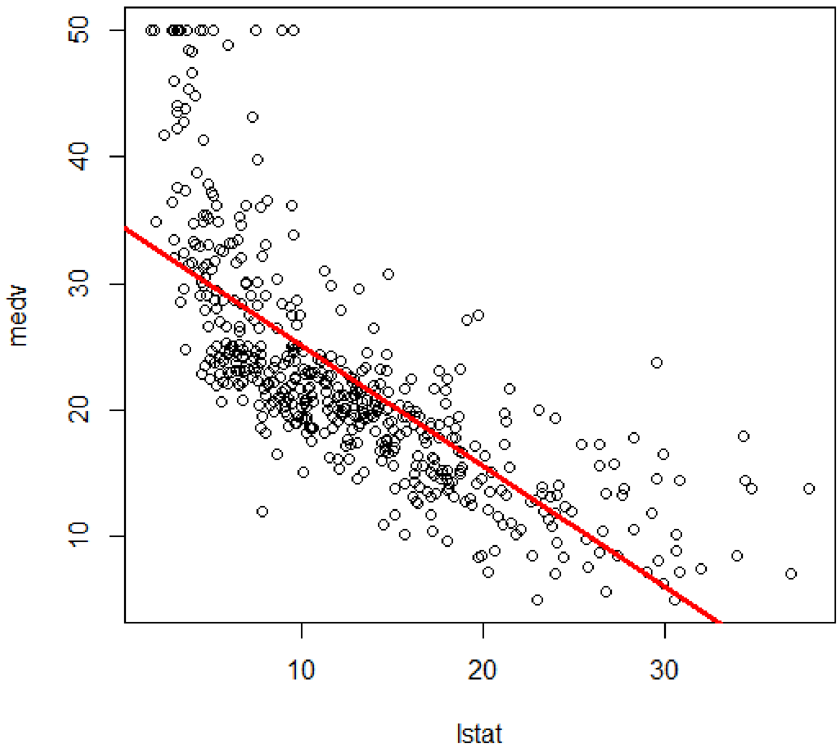
Prediction of the relationship using different values of lstat

	fit (prediction)	lwr	upr
5:	29.80359	29.00741	30.59978
10:	25.05335	24.47413	25.63256
15:	20.30310	19.73159	20.87461

Relationship between medv and lstat



Medv and Lstat with a regression line



Evidence of non-linearity in the relationship