# STAT 480 Statistical Computing Applications

## Unit 3. Linear Regression

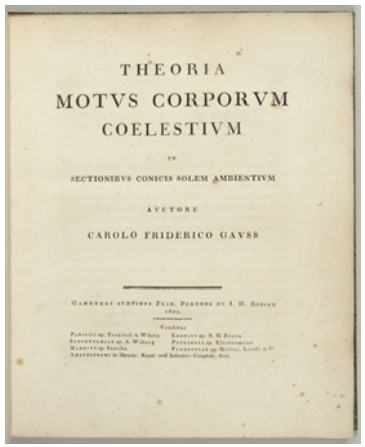
# Lecture 1. Simple Linear Regression

Department of Statistics, Iowa State University

Spring 2019

# Linear Regression

- Linear Regression is an old topic, dating back to Gauss in 1795, and later published in this famous book:


THEORIA
MOTVS CORPORVM
COELESTIVM
IN
SECTIONIBVS CONICIS SOLEM AMBIENTIVM
AVCTORE
CAROLO FRIDERICO GAVSS

- The goal is to present some different perspectives on linear regression that are (hopefully) new.
- We'll start by reviewing the basics.

# Univariate Regression

- Suppose that we have observations $(y_1, y_2, \ldots, y_n)$, and we want to model these as a linear function of $(x_1, x_2, \ldots, x_n)$.
- The univariate linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \;\; i = 1, 2, \ldots, n,$$

  where $\beta_0$ and $\beta_1$ are two unknown constants that represent the intercept and slope, and $\varepsilon$ is the error term.
- The $\beta$'s are also known as coefficients or parameters.
- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, the response variable can be predicted using

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

# Assumptions

Formal assumptions:

1. The relationship between the response $y$ and the regressors is linear, at least approximately.
2. $E(\epsilon_i) = 0$ for all $i$.
3. The errors all have the same variance: $Var(\epsilon_i) = \sigma^2$ for all $i$.
4. The errors are independent of each other.
5. $\epsilon_i$ is normally distributed for all $i$.

# Estimation of the Parameters by Least Squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for $y$ based on $x_i$.
- Let $e_i = y_i - \hat{y}_i$ be the $i$th residual.
- We define the residual sum of squares (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

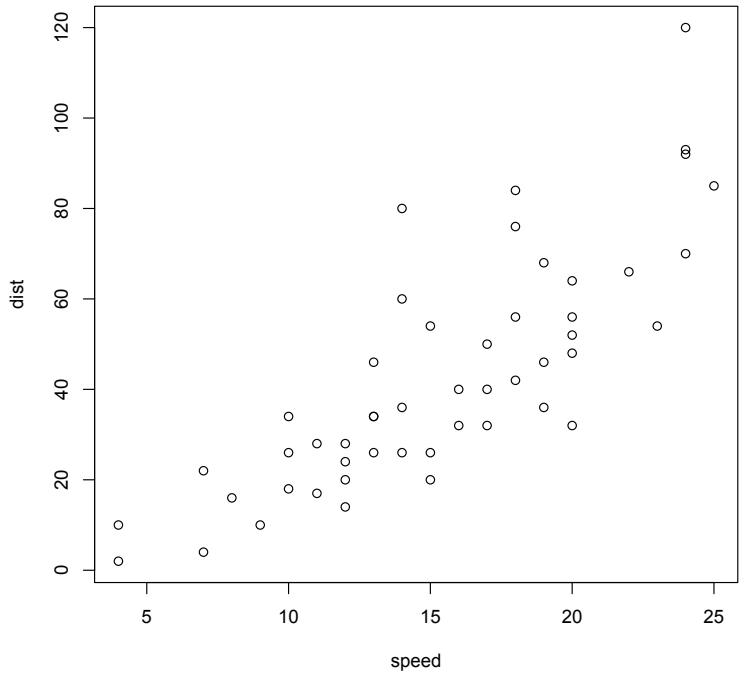
$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ are the sample means.

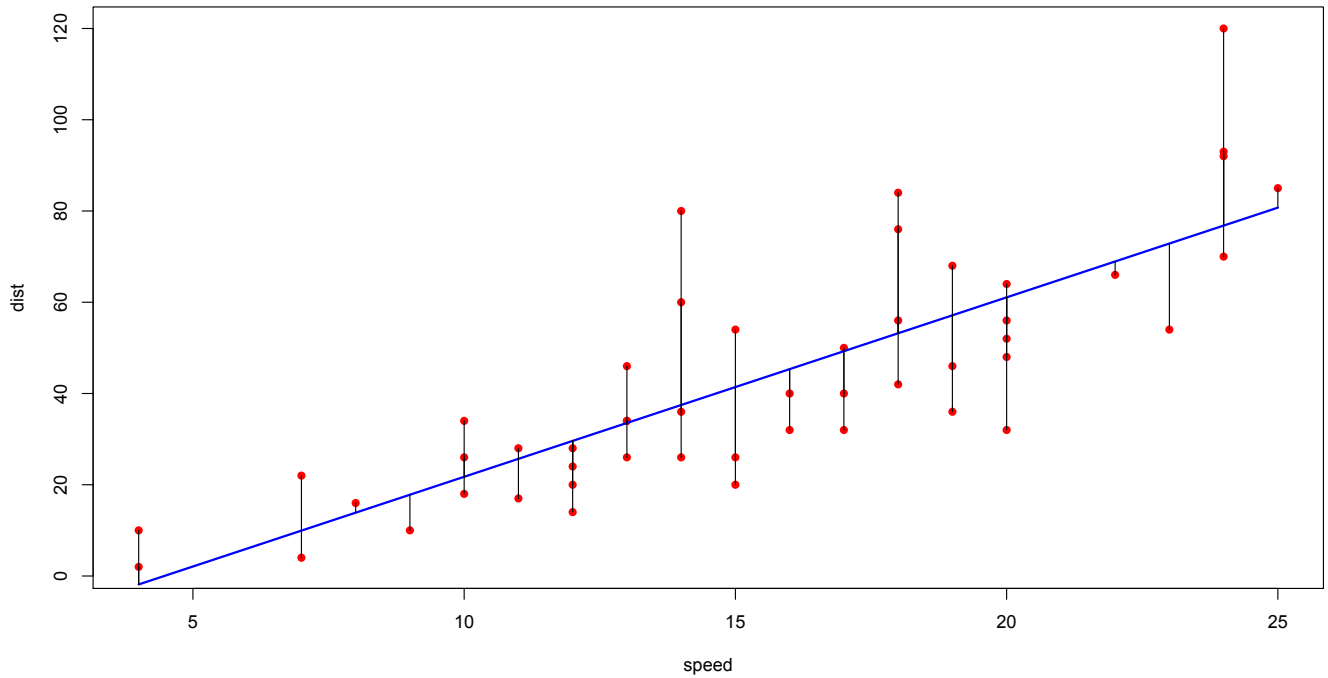
# Linear Regression for the Cars Data

Consider R built-in `cars` data and the questions we might ask:

- Is there a relationship between the speed of cars and the distances taken to stop?
- How strong is the relationship between the speed of cars and the distances taken to stop?
- How accurately can we predict the distances that a car is taken to stop?
- Is the relationship linear?



# Example: Cars Data




The least squares fit for the regression for speed and distance.

In this case, a linear fit captures the essence of the relationship.

# Accuracy of the Coefficient Estimates

- The standard error of an estimator replects how it varies under repeated sampling. Let $\sigma^2 = \mathrm{Var}(\varepsilon)$.

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\mathrm{SE}(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- These standard errors can be used to compute confidence intervals.
- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_j \pm 2 \times \mathrm{SE}(\hat{\beta}_j).$$

# Confidence Intervals

- There is approximately a 95% chance that the interval

$$[\hat{\beta}_1 - 2 \times \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \times \mathrm{SE}(\hat{\beta}_1)]$$

  will contain the true value of $\beta_1$ (under a scenario where we got repeated samples like the present sample).
- For the `cars` data, the 95% confidence interval for $\beta_1$ is $[3.10, 4.76]$.

# Hypothesis Testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

  $H_0$: There is no relationship between $x$ and $y$

  versus the *alternative hypothesis*:

  $H_A$: There is some relationship between $x$ and $y$.

- Mathematically, this corresponds to testing

$$H_0 : \ \beta_1 = 0 \quad v.s. \quad H_A : \ \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$, and $x$ is not associated with $y$.

# Hypothesis Testing (Cont.)

- To test the null hypothesis, we compute a t-statistics, given by

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

- This will have a $t$-distribution with $n - 2$ degress of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p–value.

| | Coefficient | Std. Error | $t$-statistic | p–value |
|---|---|---|---|---|
| Intercept | -17.5791 | 6.7584 | -2.601 | 0.0123 |
| speed | 3.9324 | 0.4155 | 9.464 | $1.49 \times 10^{-12}$ |

# Assess the Overall Accuracy of the Model

- $R^2$ or fraction of variance explained is

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

- It can be shown that in the simple linear regression setting that $R^2 = r^2$, where $r$ is the correlation between $x$ and $y$.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

| | |
|---|---|
| $R^2$ | 0.6511 |
| F–statistic | 89.57 |

# Model Adequacy Checking

- We should consider the validity of the assumptions mentioned before.
- Violations of the assumptions may yield an unstable model in the sense that a different sample could lead to a totally different model with opposite conclusions.
- Graphical analysis of residuals (original or scaled) is a very effective way to investigate the adequacy of the fit.
  - Normal probability plot of residuals
  - Plot of residuals against the fitted values
  - Plot of residuals against each regressor variable
  - Plot of residuals in time series (if time series data were collected)

# Model Adequacy Checking