# Introduction to Linear Regression

Rebecca C. Steorts, Duke University

STA 325, Chapter 3 ISL

# Agenda

- Why linear regression
- Simple Linear Regression
- Estimation of Coefficients
- Accuracy of Coefficients
- Accuracy of Model

# Why Linear Regression

- Linear Regression is a simple approach for supervised learning and predictive and quantiative response.
- It is one of the simplest methods to consider.

# Recall the Advertising data set

Recall the Advertising data, where one is asked to suggest a marketing plan for next year that will result in high product sales.

Questions that we might want to address:

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estiamte the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the realtionship linear?

Linear regression can answer each of these questions

# Simple Linear Regression (SLR)

- SLR is a way for predicting a quantiative response $Y$ on the basis of **one** single predictor $X$.
- Assume the relationship between $X$ and $Y$ is linear.

We write this linear relationship as

$$Y \approx \beta_o + \beta_1 X, \tag{1}$$

where $\approx$ means "approximately modeled as"

- We often say that we are regressing $Y$ onto $X$.

In equation 1, $\beta_o$ and $\beta_1$ represent the slope and the intercept in the linear model.

# The SLR Model

Once we have used our training data to produce estimates $\hat{\beta}_o, \hat{\beta}_1$ for the model coefficients, we can predict future sales on the basis of a particular value of TV advertising by computing

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

where $\hat{y}$ indicates a prediction of $Y$ on the basis of $X = x$.

# Estimating the Coefficients

In practice, $\beta_o, \beta_1$ are unknown.

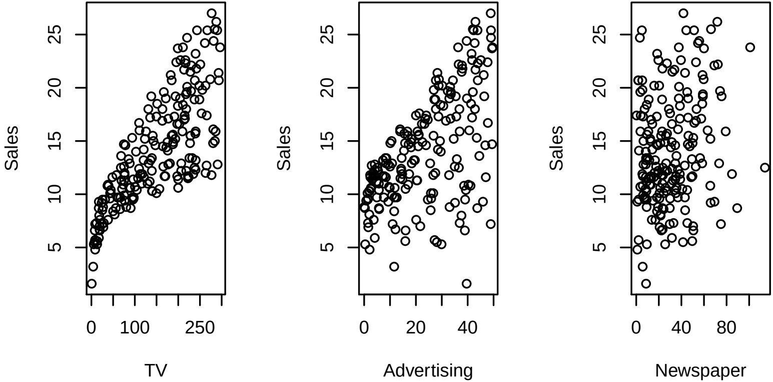
We can use equation 1 to make predictions, but we must use the data to estimate the coefficients.

# Estimating the Coefficients

Let $(x_1, y_1), \ldots, (x_n, y_n)$ represent n obvervation pairs.

# Recall the Advertising dataset and we replot the data set.

```
## pdf
##   2
```



# Estimating the Coefficients

Goal: find an intercept $\hat{\beta}_o$ and a slope $\hat{\beta}_1$ such that the resulting line is as close as possible to the $n = 200$ data points.

- There are a number of ways of measuring closeness.
- Most common approach involves minimizing the least squares criterion (LSE).
- Other approaches can be found in Chapter 6.

# Least Squares Approach

- Let $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i$ be the prediction $Y$ based on the $i$the value of X.
- Then $e_i = y_i - \hat{y}_i$ known as the $i$th **residual**
- The **residual sums of squares (RSS)** is defined as

$$\begin{aligned} \text{RSS} &= e_1^2 + \ldots e_n^2 \qquad (2) \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_o \hat{\beta}_1 x_i)^2 \qquad (3) \end{aligned}$$

- LSE chooses $\beta_o, \beta_1$ as to minimize RSS.
- Using calculus, one can show that the minimizers are

$$\begin{aligned} \hat{\beta}_1 &= \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})} \qquad (4) \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \qquad (5) \end{aligned}$$

# Assessing the Accuracy of the Coefficient Estimates

We assume the true relationship between X and Y takes the form

$$Y = f(X) + \epsilon$$

for some unknown $f$ where $\epsilon$ is a mean zero error term.

If $f$ is approx linear, we can write the relationship as

$$Y = \beta_o + \beta_1 X + \epsilon \tag{6}$$

The above equation defines the **population regression line**

Equation 4 characterizes the **least squares line**

- Here $\beta_o$ is the intercept term—that is, the expected value of Y when $X = 0$, and
- $\beta_1$ is the slope—the average increase in $Y$ associated with a one-unit increase in $X$.

# Simulated Data Set

1. Let the population regression line be $f(X) = 2 + 3X$.
2. Let us simulate from the model

$$f(X) = 2 + 3X + \epsilon$$

where $\epsilon$ is generated from a standard normal distribution.

```r
x <- rnorm(100,0,1)
epsilon <- rnorm(100,0,1)
pop.fun <- 2 + 3*x
sample.fun <- 2 + 3*x + epsilon
```

# Simulation Plots

```
pdf("examples/simulation.pdf",width=7,height=5)
plot(x, sample.fun, ylab="Y", xlab="X")
abline(a=2,b=3,col="red")
lm.fit <- lm(sample.fun ~ x)
abline(a=lm.fit$coeff[1],b=lm.fit$coeff[1],col="blue")
dev.off()
```

```
## pdf
##   2
```

# Simulation Plots

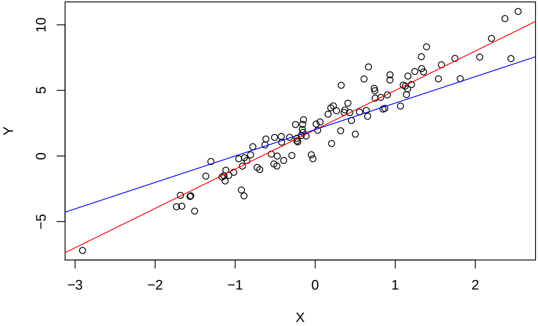


Figure 2: A simulated data set. The red line represents the true relationship, $f(X) = 2 + 3X$. The blue line is the least squares estimate for $f(X)$ based upon the observed data.

# Estimation

- It may be confusing that there are two lines to explain the data above.
- To ease this a bit, let's think about estimating the unknown population mean $\mu$ using the data $y_1, \ldots, y_n$.
- What might be a reasonable estimate for $\mu$?
- What about the sample mean $\bar{y} = \frac{1}{n} \sum_i y_i$?
- This is thought to be a pretty good estimate because it's unbiased.

# Unbiasedness

An estimator $f(Y)$ is **unbiased** for a parameter $\mu$ if $E[f(Y)] = \mu$.

Let's prove the unbiasedness of the sample mean under a linear model.

Recall that $y_i = \beta_o + \beta_1 x_i + \epsilon_i$

Let us define $\mu_i = \beta_o + \beta_1 x_i$.

Then $y_i = \mu_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

Unbiasedness immediately follows:

$$\begin{aligned} E[\bar{y}] &= E[\frac{1}{n}\sum_{i=1}^{n} y_i] && (7) \\ &= \mu_i && (8) \end{aligned}$$

Unbiased estimators tend to not over-estimate or under-estimate the true parameter.

# How accurate is the sample mean of the true parameter?

In order to address this question, we simply calculate the standard error of an estimator $\hat{\mu} = SE(\hat{\mu})^2 = \sigma^2/n$,

where $\sigma$ is the standard deviation of each of the realizations $y_i$.

The standard error tells us the average amount that $\hat{\mu}$ differs from $\mu$.

# How accurate are the regression coeffiecients?

Similarly, we can also look at how close the estimated regression coefficients to the true ones.

To do so, we compute

$$SE(\hat{\beta_o})^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$$

$$SE(\hat{\beta_1})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$

# $\sigma^2$

- Of course, in general, $\sigma^2$ isn't known.
- One way to estimate it is from the data using the **residual standard error**:

$$RSE = \sqrt{RSS/(n-2)}$$

- We can then use the standard errors to compute confidence intervals, compute hypothesis tests, or t-tests.
- Please review this material on your own as you will be expected to know it.

# Material to review

Specifically, be sure to review

- the standard error
- the residual standard error
- hypothesis tests
- confidence intervals
- t-statistic
- p-values

# Assessing the Accuracy of the Model

- Suppose we have rejected the null hypothesis in favor of the alternative hypothesis.

It is then natural to quantify the extent to which the model fits the data using the **residual standard error (RSE)** and the $R^2$ **statistic**.

# The RSE

The RSE is an estimate of the standard of $\epsilon$ or rather the average amount that the response will deviate from the true regression line.

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2}.$$

- The RSE provides an absolute measure of lack of fit of the model to the data.
- But since it is measured in the units of Y , it is not always clear what constitutes a good RSE.

# $R^2$ statistic

$R^2$ takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1, and is independent of the scale of Y.

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

where $TSS = \sum_i (y_i - \bar{y})$ is the **total sum of squares**.

# $R^2$ statistic

- TSS measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.
- In contrast, RSS measures the amount of variability that is left unexplained after performing the regression.
- $R^2$ measures the proportion of variability in Y that can be explained using X

# $R^2$ statistic

- An $R^2$ statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- A number near 0 indicates that the regression did not explain much of the variability in the response
- This might occur because the linear model is wrong, or the inherent error $\sigma^2$ is high, or both.

# Simple Linear Regression

We now follow the lab on SLR in Chapter 3.

# load the MASS and ISLR packages

We investivate the Boston data set, which records medv (median house value) for 506 neighborhoods around Boston.

Goal: predict medv using 13 predictors such as rm (average number of rooms per house), age (average age of houses), and lstat (percent of households with low socioeconomic status).

# Boston data set

```r
# look at variable names
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "
```

```r
# attach the Boston data set
attach(Boston)
```

What types of exploratory data analysis can we do?

# Boston data set

- How do we do linear regression?
- We use the lm() function, with the syntax $\text{lm}(y \sim x, data)$, where
- y is the **reponse** and
- x is a **single predictor** and
- data is the **data** in which these two variables are stored.

# Boston data set

```r
# run the linear regression
lm.fit <- lm(medv~lstat ,data=Boston)
# limited standard output
lm.fit
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Coefficients:
## (Intercept)        lstat
##       34.55        -0.95
```

# Boston data set

```r
# detailed summary output
summary(lm.fit)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## 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 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

# Confidence intervals

```r
confint(lm.fit)
```

```
##                 2.5 %     97.5 %
## (Intercept) 33.448457 35.6592247
## lstat       -1.026148 -0.8739505
```

# Prediction Intervals

The predict() function can be used to produce **confidence intervals** and **prediction intervals** for the prediction of medv for a given value of lstat.

```
predict(lm.fit,data.frame(lstat=(c(5,10,15))),
        interval ="confidence")
```

```
##        fit      lwr      upr
## 1 29.80359 29.00741 30.59978
## 2 25.05335 24.47413 25.63256
## 3 20.30310 19.73159 20.87461
```

# Prediction Intervals

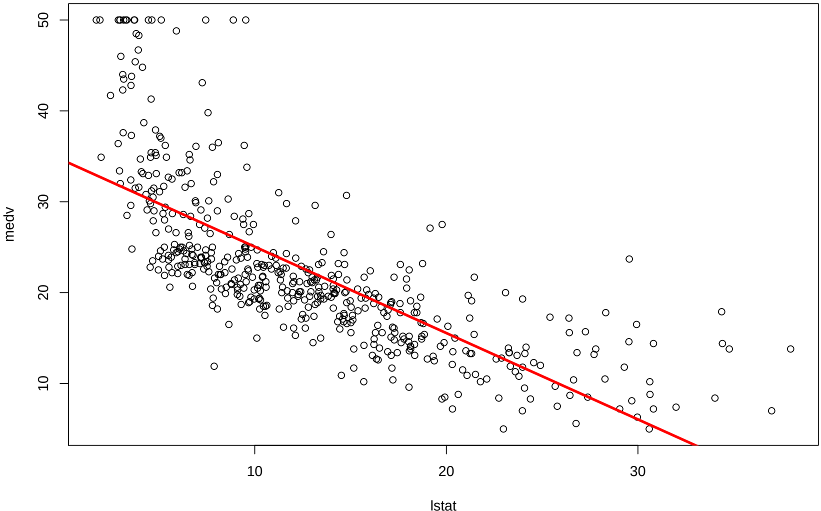
```
r   predict(lm.fit,data.frame(lstat=(c(5,10,15))),
interval ="prediction")
```

```
##         fit       lwr      upr   ## 1 29.80359
17.565675 42.04151   ## 2 25.05335 12.827626 37.27907
## 3 20.30310  8.077742 32.52846
```

- The 95% confidence interval associated with a lstat value of 10 is (24.47, 25.63), and the 95% prediction interval is (12.828, 37.28).
- As expected, the confidence and prediction intervals are centered around the same point (a predicted value of 25.05 for medv when lstat equals 10), but the latter are substantially **wider**.

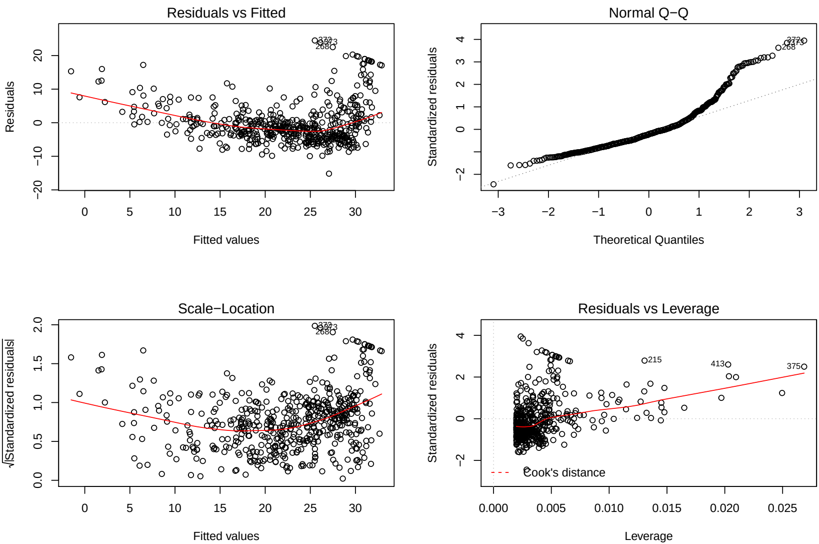
# Regression plot

```
plot(lstat, medv)
# add the regression line
abline(lm.fit,lwd=3,col="red")
```



# Diagnostic plots

```r
par(mfrow=c(2,2))
plot(lm.fit)
```



# Multiple Linear Regression (MLR)

- SLR is a useful approach for predicting a response on the basis of a single predictor variable.
- However, in practice we often have more than one predictor.

# Advertising data set

- In the Advertising data, we have examined the relationship between sales and TV advertising.
- We also have data for the amount of money spent advertising on the radio and in newspapers.
- How can we extend our analysis of the advertising data in order to accommodate these two additional predictors?

# Run many SLR's!

We could run three SLRs. Not a good idea! Why?

1. It is unclear how to make a single prediction of sales given levels of the three advertising media budgets, since each of the budgets is associated with a separate regression equation.
2. Each of the three regression equations ignores the other two media in forming estimates for the regression coefficients.

# MLR

Suppose we have $p$ preditors. Then the MLR takes the form

$$Y = \beta_o + \beta_1 X_1 + \cdots \beta_p X_p + \epsilon, \tag{9}$$

where

- $X_j$ represents the $j$th predictor
- $\beta_j$ quantifies the association between that predictor and the response
- We interpret $\beta_j$ as the average effect on $Y$ of a one unit increase in $X_j$, holding all other predictors fixed.

# Advertising

In the advertising example, the MLR becomes

$$\textit{sales} = \beta_{\textit{o}} + \beta_1 \times \textit{TV} + \beta_2 \times \textit{radio} + \beta_3 \times \textit{newspaper} + \epsilon$$

# Estimating the Regression Coefficients

As was the case in SLR, $\beta_o, \beta_1, \ldots, \beta_p$ are unknown and must be estimated by $\hat{\beta}_o, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

Given the estimated coefficients (found by minimizing the RSS), we can make predictions using

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p.$$

Solving for the LSE is beyond the scope of this class given the calculations are quite tedious and require matrix algebra.

# Important Questions to Keep in Mind

1. Is at least one of the predictors (X) useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

The easiest way to answer many of these questions is by doing quick exploratory analyses, diagnostic plots like we did for linear regression. These all extend for MLR.

For more details regarding MLR and issues that can occur, please see ISLR.

# Summary

- Simple linear regression is one of the most simple tools for analyzing data
- It can be easily extended to more than one feature variable.
- What assumptions do we make from SLR (MLR) ?
- How can we check our assumptions?
- Your homework on this module will allow you to understand linear regression in more detail.