# Ch2 Regression

## Contents

**A Subsection**
- A.1 Sinple Linear Regression (Advertising Data)
- A.2 Questions
- A.3 Regression
- A.4 Regression Approach 1
- A.5 Regression Approach 2
- A.6 Simulation Under perfect case
- A.7 Multiple Linear Regression (Adv data)
- A.8 MLR
- A.9 Pairs()
- A.10 Last row
- A.11 Without Newspaper
- A.12 With only TV and Radio as predictor
- A.13 The Marketing Questions (3.4)
- A.14 1 At least one $X$ useful?
- A.15 2 How good is Model Fit?
- A.16 3 newspaper? (Confounding Effect)

A.17 4 All predictors or just a few?
A.18 5 Effect of each medium?
A.19 6 Prediction Accuracy?
A.20 7 Is the relationship linear?
A.21 8 Is there synergy among the advertising media?
A.22 Boston House Price 1978
A.23 Boston Data
A.24 Median Home Value vs predictors
A.25 Inference vs Prediction
A.26 Advertisement
A.27 Housing
A.28 Flexibility and Interpretability trade-off

---

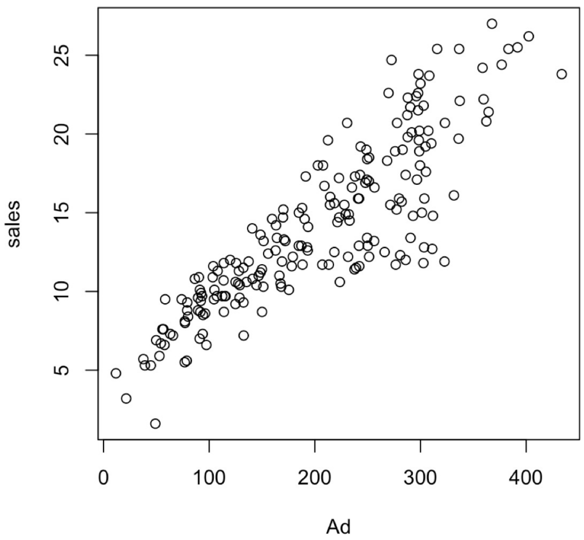
Textbook: James et al. ISLR 2ed.

## A Subsection

[ToC]

## A.1 Sinple Linear Regression (Advertising Data)

```
Advertising.csv from ISLR web site.

Ad = TV + radio + newspaper (Total Spending)
```

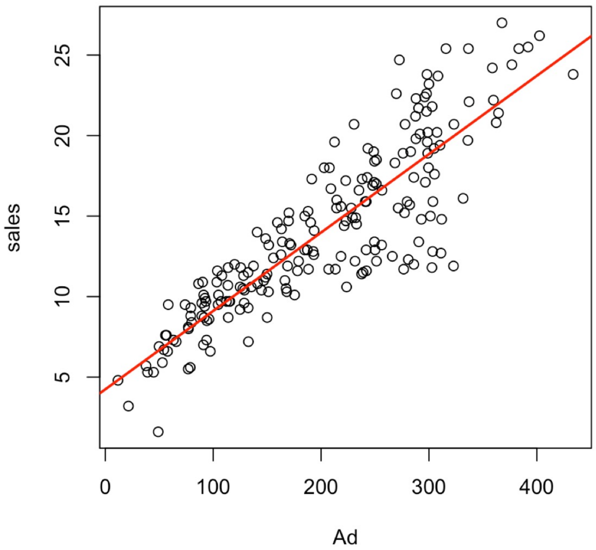


sales
25
20
15
10
5
0
100
200
300
400
Ad

## A.2 Questions

1. Is there a relationship between advertising budget and sales?
2. If so, what is the form of the relationship?
3. How strong is the relationship between advertising budget and sales?
4. How accurately can we estimate the effect?
5. How accurately can we predict future sales?

## A.3 Regression



## A.4 Regression Approach 1

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

- Assumes that $f()$ is linear
- Estimate parameters $\beta_0$ and $\beta_1$ based on

$$\text{RSS} = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

- Formula for $\hat{\beta}_1$ and $\hat{\beta}_0$:

$$\hat{\beta}_1 = r s_y / s_x \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

## A.5 Regression Approach 2

- Assumes that

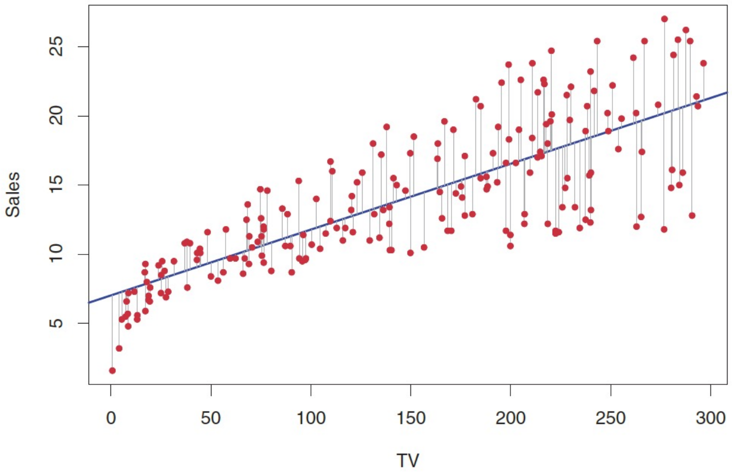
$$Y = \beta_0 + \beta_1 X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

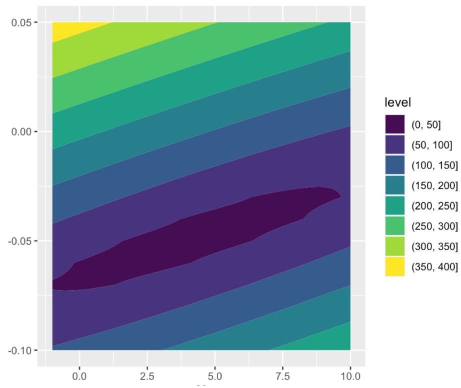
- Estimate parameters $\beta_0$ and $\beta_1$ with best estimators possible. (unbiased, minimum variance)
- Formula for $\hat{\beta}_1$ and $\hat{\beta}_0$:

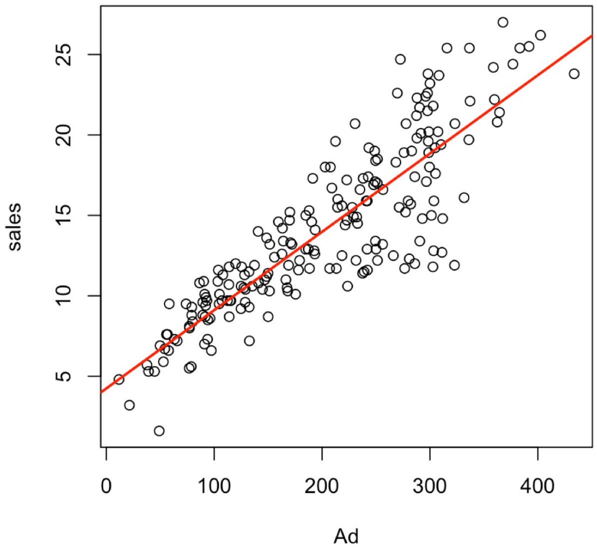
$$\hat{\beta}_1 = r s_y / s_x \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

```
Call:
lm(formula = sales ~ Ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0546 -1.3071  0.1173  1.5961  7.1895 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.243028   0.438525   9.676   <2e-16 ***
Ad          0.048688   0.001982  24.564   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.6 on 198 degrees of freedom
Multiple R-squared:  0.7529,    Adjusted R-squared:  0.7517 
F-statistic: 603.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

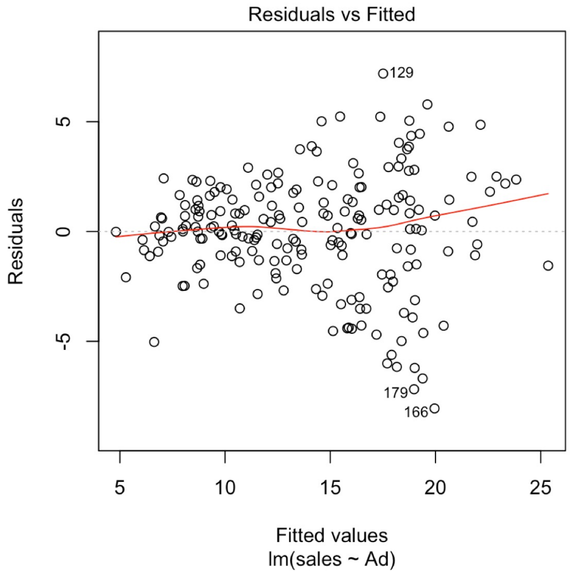
- is assumption true?
- how off is the estimator?
- how accurate is the prediction?


Sales
TV
0
50
100
150
200
250
300
5
10
15
20
25


level
(0, 50]
(50, 100]
(100, 150]
(150, 200]
(200, 250]
(250, 300]
(300, 350]
(350, 400]
0.05
0.00
-0.05
-0.10
0.0
2.5
5.0
7.5
10.0

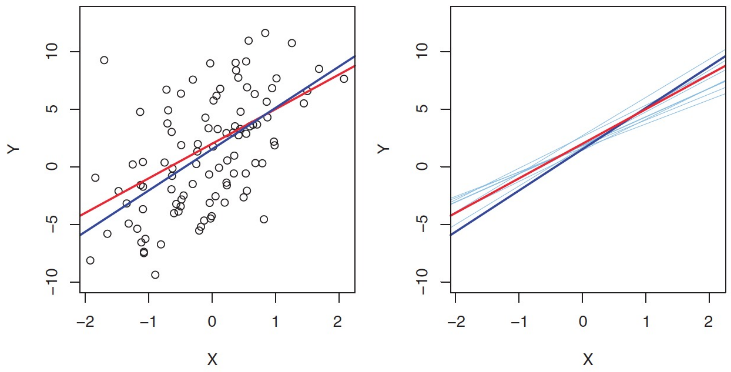


sales
Ad
0
100
200
300
400
5
10
15
20
25

Residuals vs Fitted
Residuals
Fitted values
lm(sales ~ Ad)
129
179
166
5
10
15
20
25
-5
0
5

## A.6 Simulation Under perfect case




```
X = rnorm(30, 3, 5)
Y = 4+1.5* X + rnorm(30, 0, 5)

plot(X, Y, xlim=c(-6, 11), ylim=c(-10, 25))
abline(a=4, b=1.5, col="blue", lwd=2)
```

```
m1 <- lm(Y~X)
abline(m1, col="red")
```

## A.7 Multiple Linear Regression (Adv data)

Multiple Linear Regression (Adv data)

## A.8 MLR

- Want to guess the next $Y$ as accurate as possible

$$\texttt{sales} = \beta_0 + \beta_1\,\texttt{TV} + \beta_2\,\texttt{radio} + \beta_3\,\texttt{newspaper} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

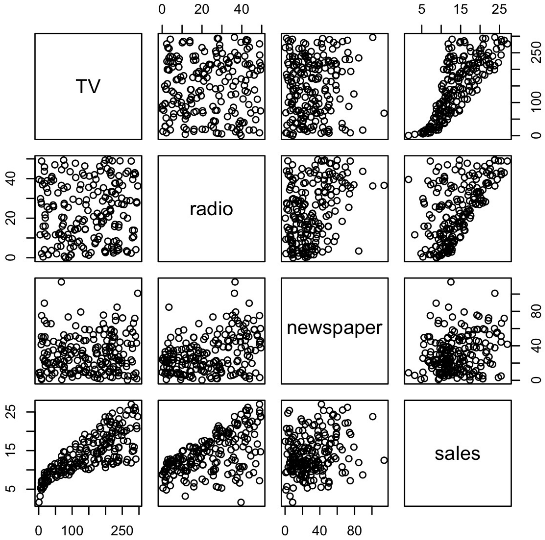
- Estimate parameters by minimizing

$$\text{RSS} = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

- Formula for $\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1, \beta_2, \beta_3)'$:

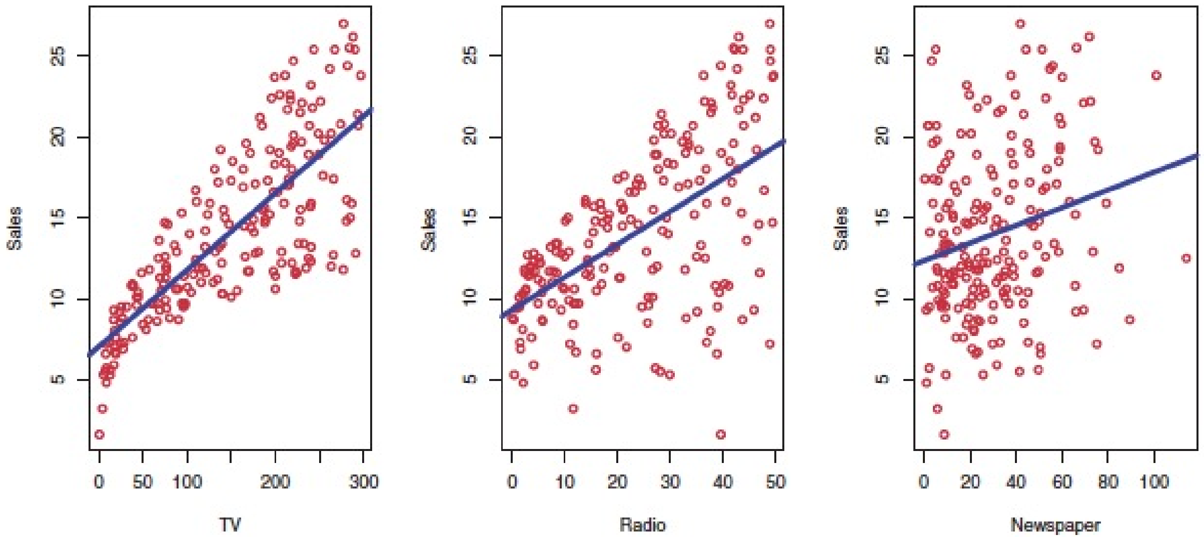
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

## A.9 Pairs()



TV
radio
newspaper
sales
0
20
40
5
15
25
250
100
0
40
20
0
80
40
0
25
15
5
0
100
250
0
40
80

## A.10 Last row



Sales
25
20
15
10
5
0
50
100
200
300
TV
Radio
10
20
30
40
50
Newspaper
60
80
100

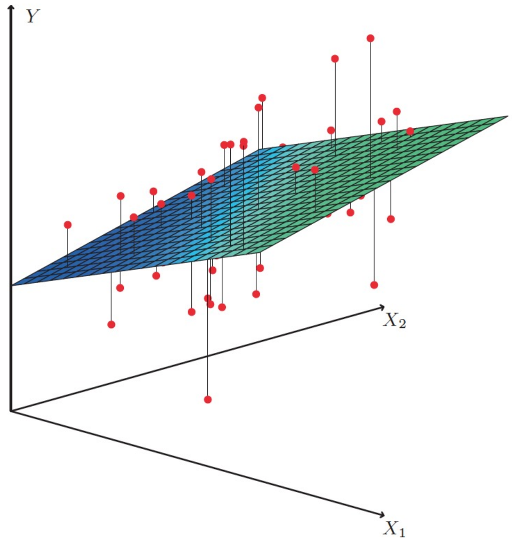
```
Model2 <- lm(sales ~ TV + radio + newspaper)
summary(Model2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

## A.11 Without Newspaper

```
lm(formula = sales ~ TV + radio)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

## A.12 With only TV and Radio as predictor

It's like fitting plane in 3-d space



$Y$
$X_2$
$X_1$

## A.13 The Marketing Questions (3.4)

1. Is there a relationship between advertising sales and budget?
2. How strong is the relationship?
3. Is it important to advertise in newspaper?
4. Is all predictor important, or just a subset?
5. Which media contribute to most to the sales? How much?
6. How accurately can we predict future sales?
7. Is the relationship linear?
8. Is there synergy among the advertising media?

## A.14 1 At least one $X$ useful?

- In SLR, we only need to test $\beta_1 = 0$.
- Now we have to test $\beta_1 = \beta_2 = \beta_3 = 0$.
- Use $F$-statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

```
summary(Model2)

# Residual standard error: 1.686 on 196 degrees of freedom
# Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
# F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

- We can test SUBSET of parameters ($\beta_2 = \beta_3 = 0$) by

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

  Where $RSS_0$ is the RSS from the model using $\beta_2 = \beta_3 = 0$, and $q$ is the number of suppressed parameters.

- Why test as a whole when you can do the t-test individually? (important when $p$ is large)

## A.15 2 How good is Model Fit?

- Coefficient of Determination

$$\mathrm{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad \mathrm{RSS} = \sum_{i=1}^{n}(Y_i - \hat{Y})^2, \qquad R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

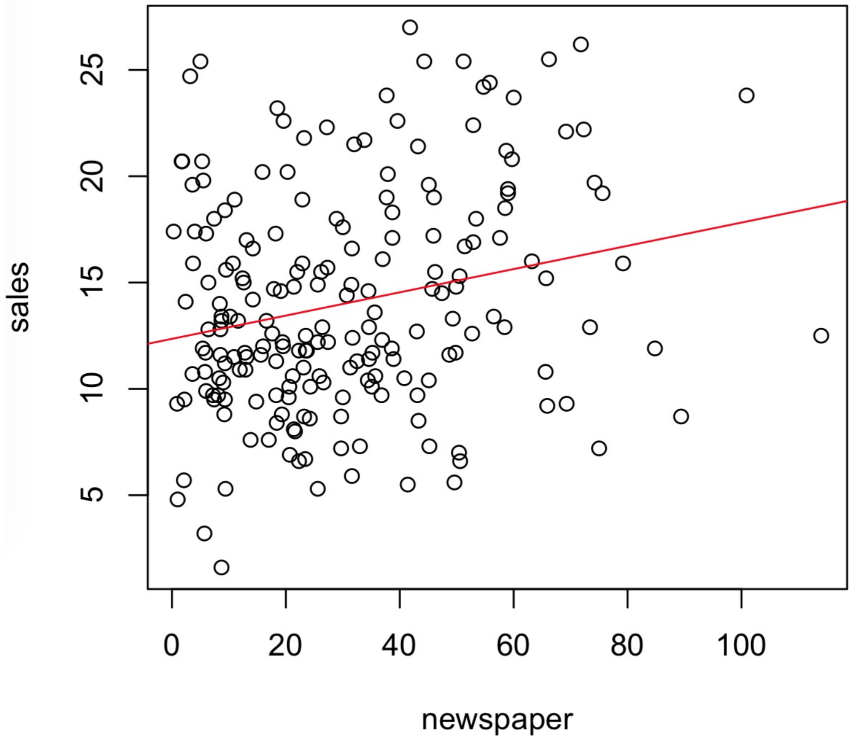
- $R^2 = .89719$ without `newspaper`
- $R^2 = .8972$ with `newspaper`
- MSE (RSE in ISLR) estimates $\sigma^2$ and represents irreducible error.
- With $p$ predictors,

$$MSE = \sqrt{\frac{1}{n-p-1}RSS}$$

## A.16 3 newspaper? (Confounding Effect)

```
Model3 <- lm(sales ~ newspaper)
summary(Model3)

Coefficients:
#            Estimate Std. Error t value Pr(>|t|)
#(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
#newspaper    0.05469    0.01658    3.30  0.00115 **
#---
#Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 5.092 on 198 degrees of freedom
#Multiple R-squared:  0.05212,Adjusted R-squared:  0.04733
#F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

```
plot(newspaper, sales)
abline(Model3, col="red")   # plot reg line from Model3



cor(Adv)   # correlation matrix of each column

#            TV         radio      newspaper  sales
# TV         1.00000000 0.05480866 0.05664787 0.7822244
# radio      0.05480866 1.00000000 0.35410375 0.5762226
# newspaper  0.05664787 0.35410375 1.00000000 0.2282990
# sales      0.78222442 0.57622257 0.22829903 1.0000000
```



sales
newspaper
0
20
40
60
80
100
5
10
15
20
25

- Mutiple Reg suggests `newspaper` has no effect
- In simple regression, `newspaper` gets credit through `radio` because of the correlation.
- Many examples of confounding variables (lurking variables) (shark attack vs ice cream sales, num of cavity vs vocabulary score)

## A.17 4 All predictors or just a few?

- Have to try out many models, and use some kind of criteria to pick the best
- Mallow's $C_p$, AIC, BIC, Adjusted $R^2$. (more in Ch6)
- There's $2^p$ models with $p$ predictors. $2^3 = 8, 2^{10} = 1024, 2^{30} = 1,073,741,824$.
- Forward, Backward, Mixed selection

## A.18 5 Effect of each medium?

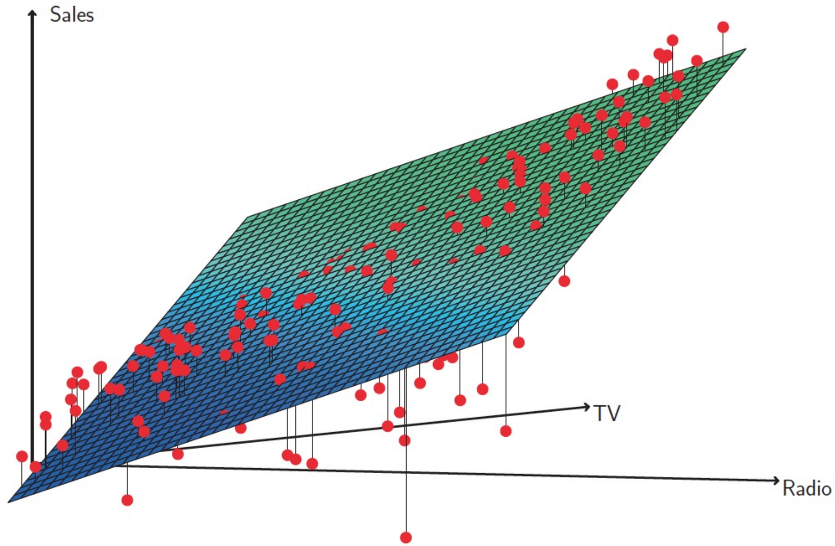
We can construct CI for parameters.

For the Advertising data, the 95% CI

```
(0.043, 0.049) for TV,
(0.172, 0.206) for radio,
(-0.013, 0.011) for newspaper.
```

## A.19 6 Prediction Accuracy?

- We can get $\hat{f}(X)$ using estimated $\beta_i$.
- There could be model bias
- Get CI for parameters, and PI for predictions



Sales
TV
Radio

```
newAdv <- data.frame(TV=c(50, 60), radio=c(20, 10), newspaper=c(0, 0))
newAdv

#  TV radio newspaper
#1 50    20         0
#2 60    10         0

predict(Model2, newdata=newAdv, interval="confidence")
#       fit      lwr      upr
#1 8.997722 8.515752 9.479692
#2 7.570068 7.099337 8.040800

predict(Model2, newdata=newAdv, interval="prediction")

#       fit      lwr      upr
#1 8.997722 5.638898 12.35655
#2 7.570068 4.212838 10.92730
```

## A.20 7 Is the relationship linear?

If the relationships are linear, then the residual plots should display no pattern. Needs transformation?

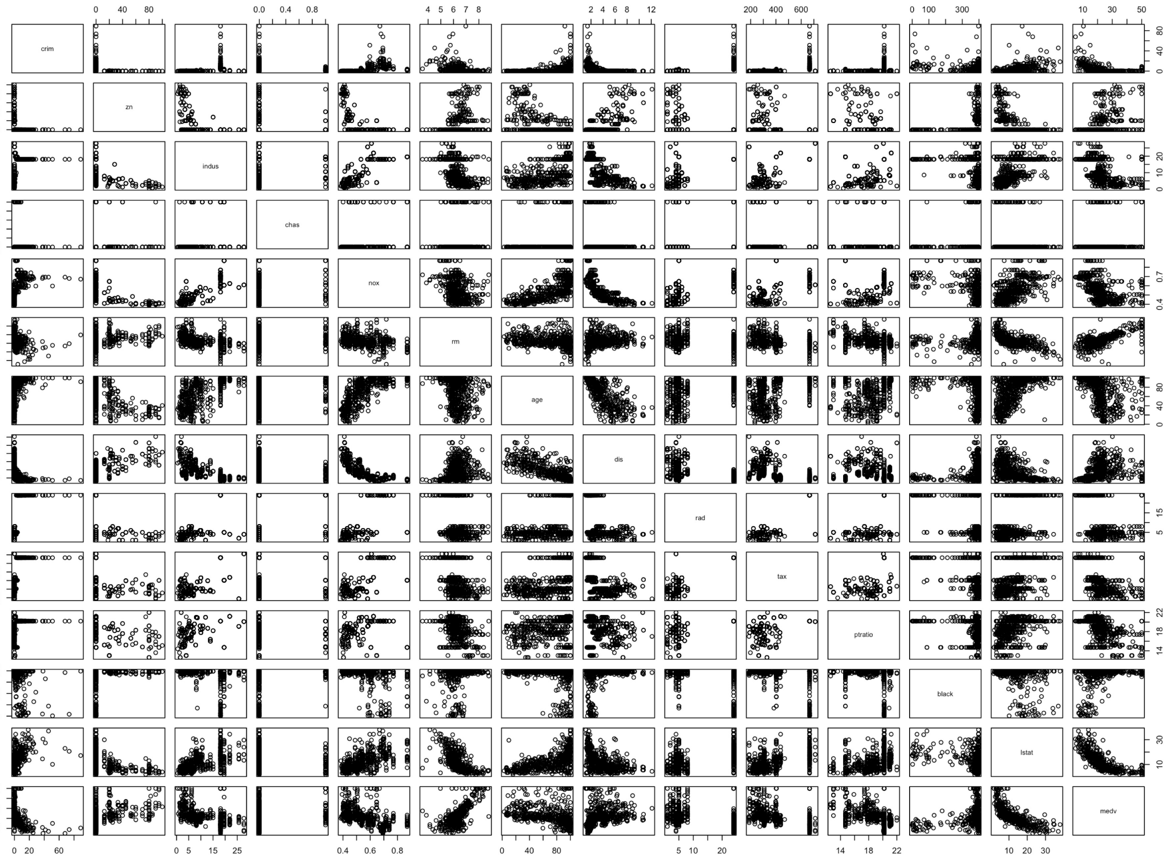
## A.21 8 Is there synergy among the advertising media?

The standard linear regression model assumes an additive relationship between the predictors and the response. Including an interaction term in the model results in a substantial increase in $R^2$, from around 90% to almost 97

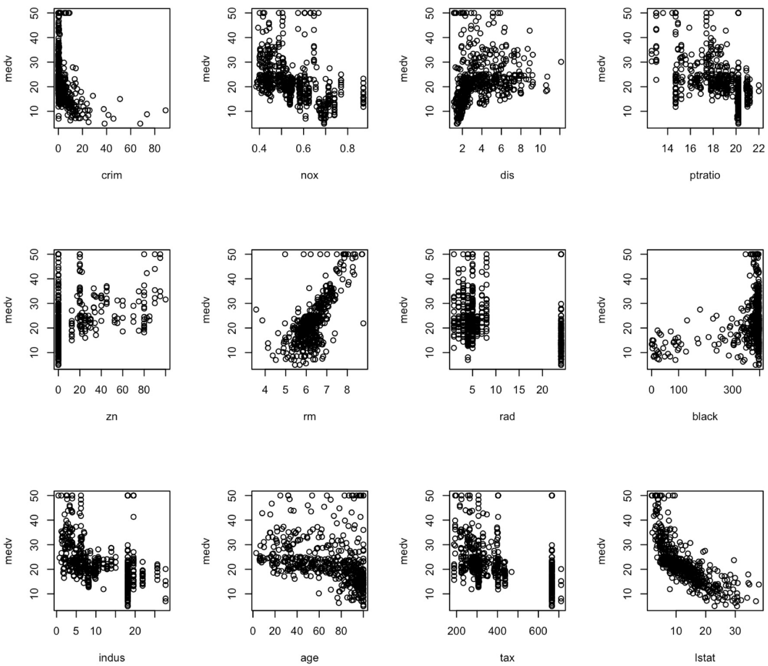
## A.22 Boston House Price 1978

Boston House Price 1978

## A.23 Boston Data


crim
zn
indus
chas
nox
rm
age
dis
rad
tax
ptratio
black
lstat
medv

### A.24 Median Home Value vs predictors


medv
crim
nox
dis
ptratio
zn
rm
rad
black
indus
age
tax
lstat

## A.25 Inference vs Prediction

Inference vs Prediction

## A.26 Advertisement

- Prediction
  Who is more likely to be next customer?

- Inference
  Which media contribute to sales?
  Which media generate the biggest boost in sales?
  How much increase in sales is associated with a given increase in TV advertising?

## A.27 Housing

- Prediction
  Is THIS house overpriced, or underpriced?

- Inference
  What contributes to the value of a house? How can I increase them? If I have X amount of dollars, where should I spend them?

## A.28 Flexibility and Interpretability trade-off