Statistical Learning

Lecture 01d

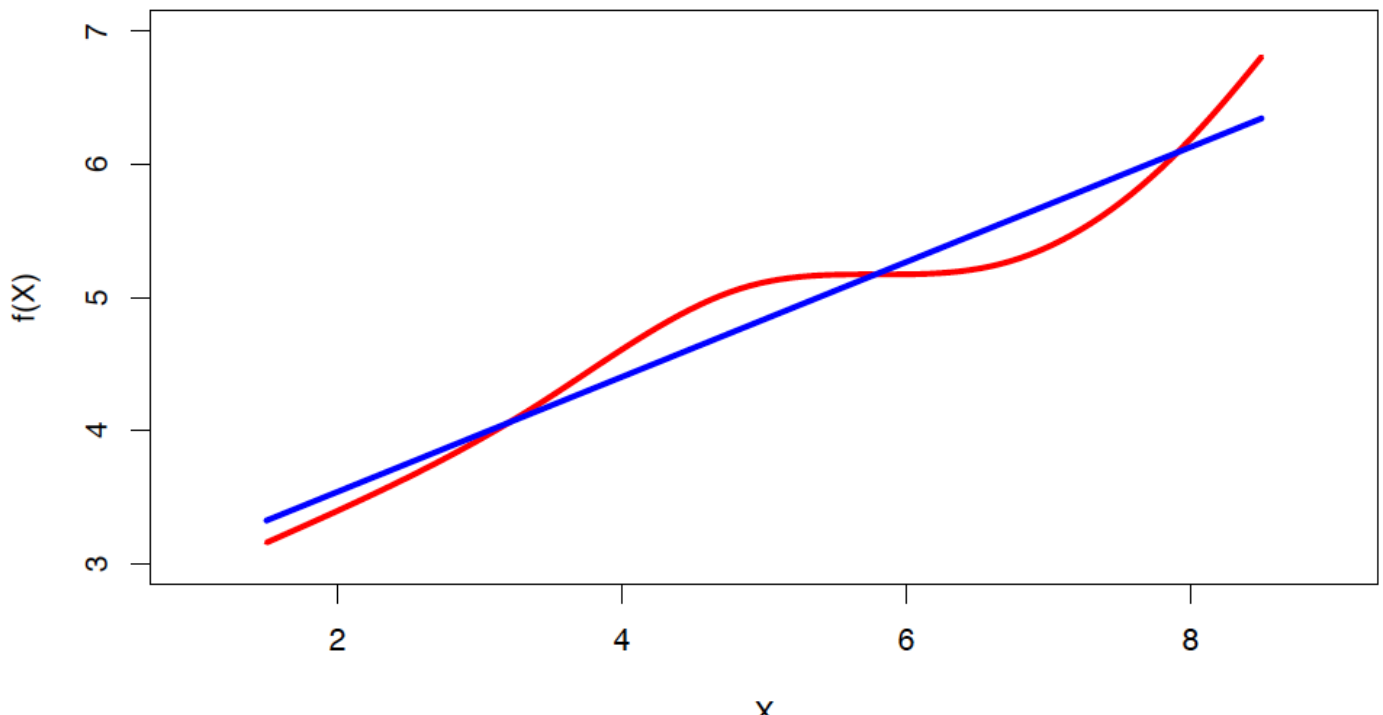
ANU - RSFAS

Last Updated: Thu Feb 24 09:49:10 2022

Linear Regression - Review

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots, X_p$ is linear.

- True regression functions are never linear!



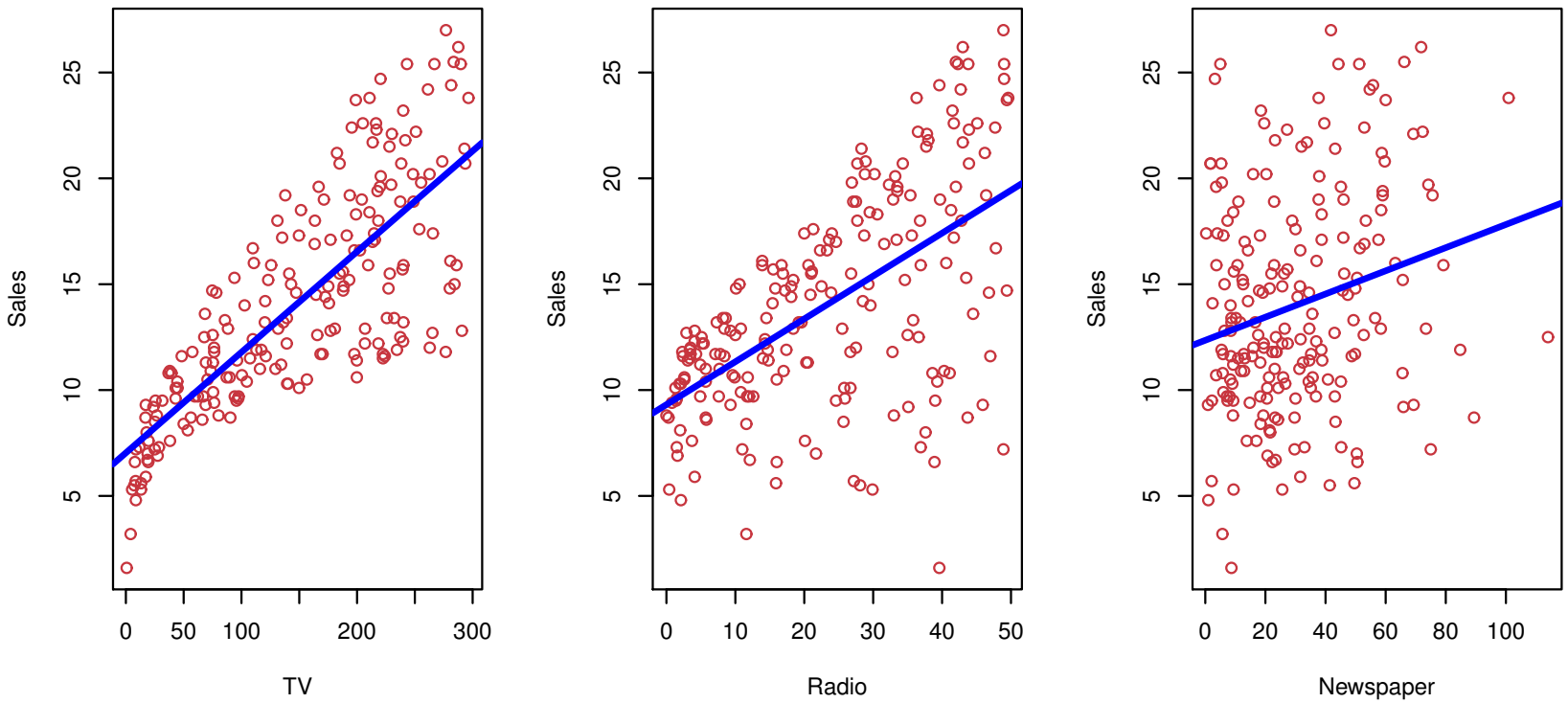
- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Linear Regression for the Advertising Data

Questions we might ask:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Advertising Data



- Shown are **sales** (in thousands of units) vs **TV, radio and newspaper** (budgets in thousands of USD), with a blue linear-regression line fit **separately** to each.

Simple linear regression using a single predictor X.

- We assume a model

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

where β_0 and β_1 are two unknown constants that represent the intercept and slope, also known as coefficients or parameters, and ϵ is the error term. And

$$\epsilon_i \overset{\text{iid}}{\sim} \text{normal}(0, \sigma^2)$$

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales by

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

Estimation of the Parameters

- There are many approaches to estimation (least-squares, maximum-likelihood, Bayesian).
- Here we will consider least-squares:
- Let $e_i = y_i - \hat{y}_i$. Then consider the Residual Sum of Squares:

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

- Minimizing this function wrt to $\hat{\beta}_0$ and $\hat{\beta}_1$ we find:

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

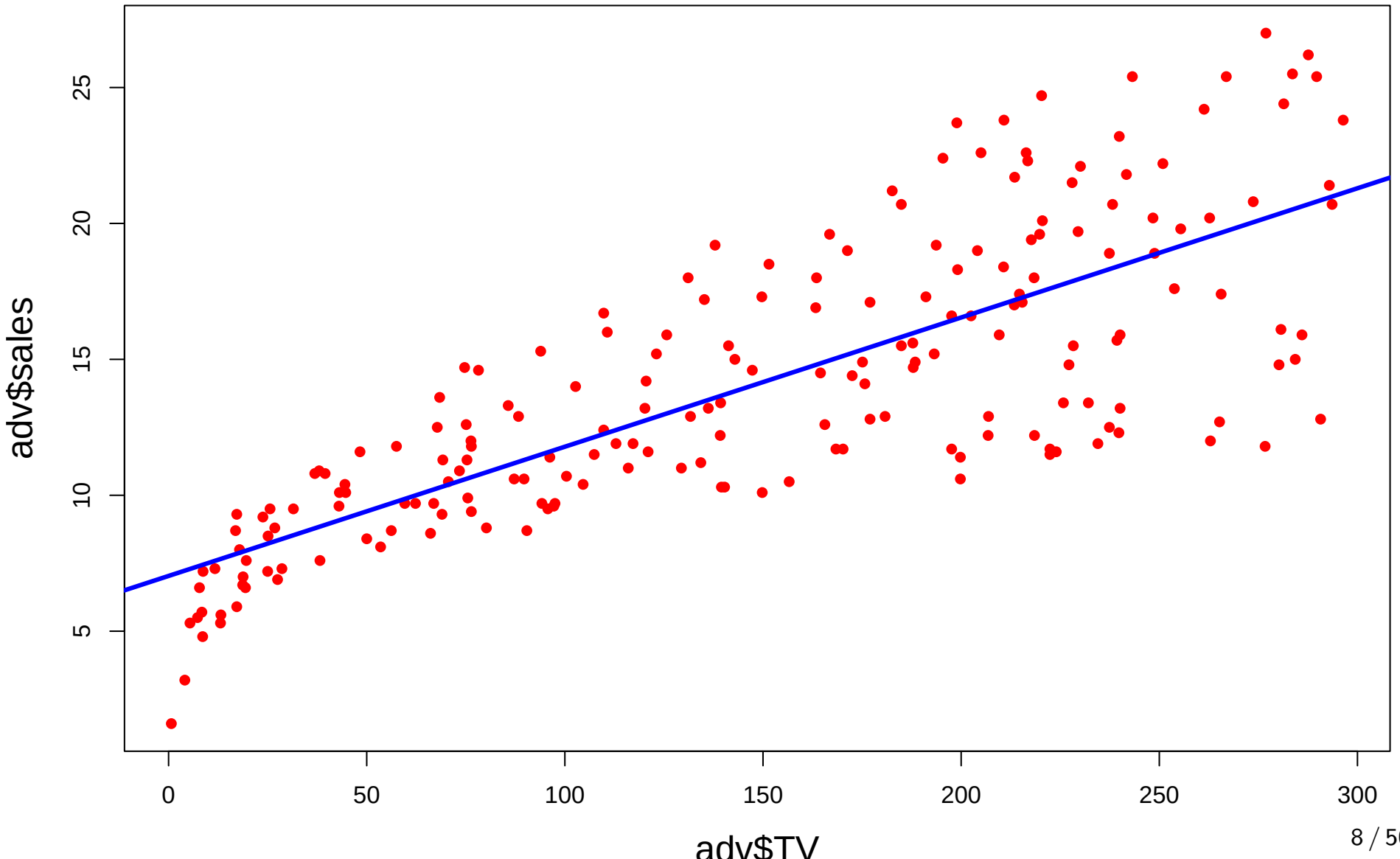
Advertising Data

- The advertising data are available at https://www.statlearning.com/resources-first-edition.

```
adv <- read.csv("Advertising.csv",
                header=TRUE, row.names=1)
head(adv)
```

```
##      TV radio newspaper sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2
```

```r
mod.lm <- lm(sales ~ TV, data=adv)
plot(adv$TV, adv$sales, pch=16, col="red", cex.lab=1.5)
abline(mod.lm, col="blue", lwd=3)
```



Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$
\begin{aligned}
SE(\hat{\beta}_1)^2 &= \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
SE(\hat{\beta}_0)^2 &= \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]
\end{aligned}
$$

Where

$$
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - (p+1)}
$$

```
summary(mod.lm)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = adv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
n <- nrow(adv)
mod.lm$coef
```

```
## (Intercept)          TV
##  7.03259355  0.04753664
```

```
sigma.sq.hat <- sum(mod.lm$res^2)/(n-2)
sqrt(sigma.sq.hat)
```

```
## [1] 3.258656
```

- These standard errors can be used to compute confidence intervals - to assess the uncertainty!

$$\hat{\beta}_1 \pm 1.97 \cdot SE(\hat{\beta}_1)$$

```
qt(0.975, n-2)
```

```
## [1] 1.972017
```

- A decent approximation

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

- Interpretation: Over repeated sampling of the data, we expect 95% of the confidence interval we create to contain the true value β_1.

- For the advertising data, the 95% confidence interval for β_1 is:

```
confint(mod.lm, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

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 versus the alternative hypothesis:

$$\begin{aligned} \mathrm{H}_0 : &\quad \text{There is no relationship between } X \text{ and } Y \\ \mathrm{H}_A : &\quad \text{There is is some relationship between between } X \text{ and } Y \end{aligned}$$

- Mathematically, this corresponds to testing (two sided alternative):

$$\begin{aligned} \mathrm{H}_0 : &\quad \beta_1 = 0 \\ \mathrm{H}_A : &\quad \beta_1 \neq 0 \end{aligned}$$

- If $\beta_1 = 0$ then $Y = \beta_0 + \epsilon$. So there is no relationship between X and Y.

- We can test test the null hypothesis, we compute a t-statistic, given by

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

- This will have a t-distribution with $n - (p + 1)$ degrees 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.

Assessing the Overall Accuracy of the Model

- We compute the Residual Sum of Squares Error:

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- We compute the Residual Standard Error:

$$\text{RSE} = \hat{\sigma} = \sqrt{\frac{1}{n-(p+1)}\text{RSS}}$$

- We compute the Total Sum of Squares:

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- This leads to calculating the R^2:

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

Note: For simple linear regression (one X), R^2 is just the square of the sample correlation, so r^2.

- See **multiple R-squared** in the previous table: $R^2 = 0.612$.

Extend the model - Mutliple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.

Interpreting Regression Coefficients

- The ideal scenario is when the predictors are uncorrelated – a balanced design:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as "a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed", are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous – when X_j changes, everything else changes.
- Claims of causality should be avoided for observational data.

Causality

- Causality is an active area of research in statistics.
  "More has been learned about causal inference in the last few decades than the sum total of everything that has been learned about it in all prior recorded history" - *Gary King*

- Two main approaches:
  - Donald Rubin (rebalance data)
  - Judea Pearl (need additional structure conveyed by a graph)

The woes of (interpreting) regression coefficients

"Data Analysis and Regression" Mosteller and Tukey 1977

- a regression coefficient β_j estimates the expected change in Y per unit change in X_j, with all other predictors held fixed. But predictors usually change together!
  "The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively" - *Fred Mosteller and John Tukey, paraphrasing George Box*
  "Essentially, all models are wrong, but some are useful" - *George Box*

Back to Advertising Data

```
mod.lm2 <- lm(sales ~ TV + radio+ newspaper, data=adv)
cor(adv)
```

```
##                    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
```

```
summary(mod.lm2)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = adv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## 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
```

Some Important Questions

- Is at least one of the predictors $X_1, X_2, \cdots, X_p$ useful in predicting the response?

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

F-statistic: 570.3 on 3 and 196 DF, p-value: $<$ 2.2e-16

We can reject the NULL hypothesis that none of the covariates are important.

- How well does the model fit the data?

Multiple R-squared: 0.8972

Deciding on the Important Variables

- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- However we often can't examine all possible models, since they are 2^p of them; for example when p $= 40$ there are over a billion models!
- Instead we need an automated approach that searches through a subset of them – forward selection and backward selection

Forward Selection

- Begin with the null model – a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS (or some other criterion).
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Backward Selection

- Start with all variables in the model.
- Remove the variable with the largest p-value – that is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Model Selection – continued

- Later we discuss more systematic criteria for choosing an "optimal" member in the path of models produced by forward or backward stepwise selection.

- These include: Mallow's Cp, Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R2 and Cross-validation (CV).

Extensions of the Linear Model

- Suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for **TV** should increase as **radio** increases.
- In this situation, given a fixed budget of $100,000, spending half on **radio** and half on **TV** may increase sales more than allocating the entire amount to either **TV** or to **radio**.
- In marketing, this is known as a synergy effect, and in statistics it is referred to as an interaction effect.

$$\text{sales} = \beta_0 + \beta_1 \text{ TV} + \beta_2 \text{ radio} + \beta_3 \text{ TV} \times \text{radio} + \epsilon$$

```
mod.lm3 <- lm(sales ~ TV + radio + TV:radio, data=adv)
summary(mod.lm3)[[4]]
```

```
##                Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept) 6.750220203 0.2478713699 27.232755 1.541461e-68
## TV          0.019101074 0.0015041455 12.698953 2.363605e-27
## radio       0.028860340 0.0089052729  3.240815 1.400461e-03
## TV:radio    0.001086495 0.0000524204 20.726564 2.757681e-51
```

Thoughts on the Output

- The results in this table suggests that interactions are important.
- The p-value for the interaction term **TV** $\times$ **radio** is extremely low, indicating that there is strong evidence for

$$H_A : \beta_3 \neq 0$$

- The R^2 for the interaction model is 96.8% (not shown in table), compared to only 89.7% for the model that predicts **sales** using **TV** and **radio** without an interaction term.

Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not.

- The hierarchy principle:
  If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

- The rationale for this principle is that interactions are hard to interpret in a model without main effects – their meaning is changed.

Interpretation

- The coefficient estimates in the table suggest that an increase in **TV** advertising of \$1,000 is associated with increased sales of

$$(\hat{\beta}_1 + \hat{\beta}_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio units}$$

- An increase in **radio** advertising of \$1,000 will be associated with an increase in sales of

$$(\hat{\beta}_2 + \hat{\beta}_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV units}$$

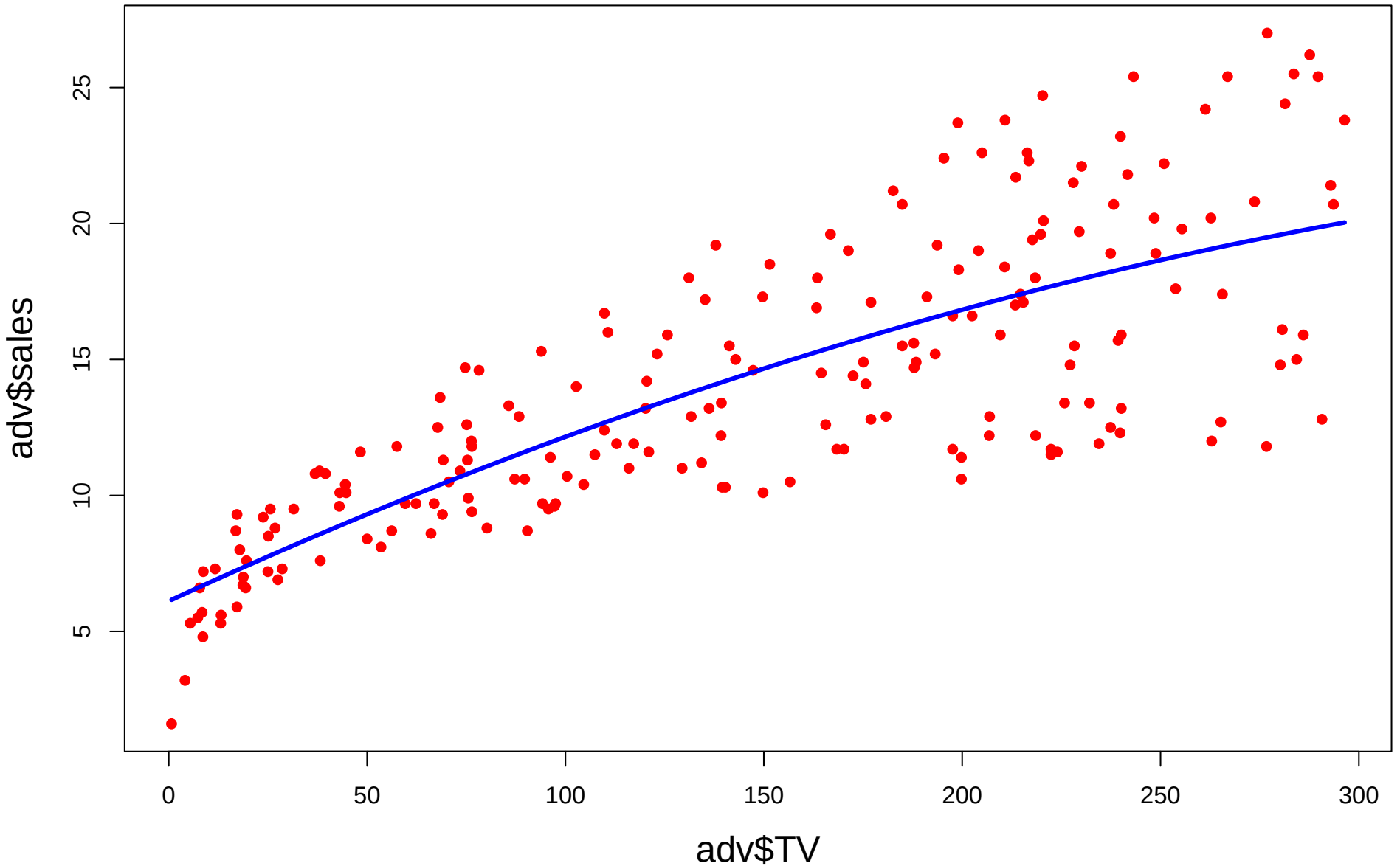
Non-linear effects of Predictors

$$\text{sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{TV}^2 + \epsilon$$

```
mod.lm4 <- lm(sales ~ TV + I(TV^2), data=adv)
summary(mod.lm4)[[4]]
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  6.114120e+00 0.6592224346  9.274745 3.226267e-17
## TV           6.726593e-02 0.0105944044  6.349194 1.461837e-09
## I(TV^2)     -6.846934e-05 0.0000355783 -1.924469 5.573659e-02
```

```
plot(adv$TV, adv$sales, pch=16, col="red", cex.lab=1.5)
sort.TV <- sort(adv$TV)
fit.lm <- mod.lm4$coef[1] + mod.lm4$coef[2]*sort.TV +
  mod.lm4$coef[3]*sort.TV^2
lines(sort.TV, fit.lm, type="l", lwd=3, col="blue")
```


adv$sales
5
10
15
20
25
0
50
100
150
200
250
300
adv$TV

Regression Fun with Matrices

- Let's consider the following simple linear regression model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i &\overset{\text{iid}}{\sim} \text{normal}(0, \sigma^2) \\ i &= 1, \ldots, n \end{aligned}$$

- Let's stack the these:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \vdots &= \vdots \\ Y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

Regression Fun with Matrices

- We can write this as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

- $\boldsymbol{y}_{n\times 1}$ response vector
- $\boldsymbol{X}_{n\times\{p+1\}}$ covariate (design) matrix
- $\beta_{\{p+1\}\times 1}$ coefficient vector
- $\boldsymbol{\epsilon}_{n\times 1}$ error vector

- The Residual Sum of Squares is:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 &= \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)(Y_i - \beta_0 - \beta_1 x_i) \\ &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)\end{aligned}$$

- Now that we have our model let's estimate the regression coefficients:

$$\begin{aligned} RSS &= \boldsymbol{y}'\boldsymbol{y} - \beta'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\beta + \beta'\boldsymbol{X}'\boldsymbol{X}\beta \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\beta'\boldsymbol{X}'\boldsymbol{y} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta \end{aligned}$$

- Now let's differentiate with respect to β:

$$\frac{dRSS}{d\beta} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\beta$$

Let's examine this - Some Math Rules

- Consider the vector $\boldsymbol{x} = \{_c\ x_i\}_{i=1}^{k}$:

$$\frac{\partial}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_k} \end{bmatrix}$$

- Let $\boldsymbol{a}$ be a vector:

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{a}'\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{a}) = \boldsymbol{a}$$

$$\boldsymbol{a}'\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{a}$$

- Consider the vector $\boldsymbol{y}' = [y_1, y_2, \ldots, y_p]$:

$$\frac{\partial \boldsymbol{y}'}{\partial \boldsymbol{x}} = \left\{ m \ \frac{\partial y_j}{\partial x_i} \right\}_{i=1,j=1}^{k \quad p}, \qquad \text{a } k \times p \text{ matrix}$$

$$\frac{\partial \boldsymbol{x}'}{\partial \boldsymbol{x}} = \mathbb{I}, \ \text{identity matrix}$$

- Let $\boldsymbol{A}$ be a matrix:

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{A}) = \frac{\partial \boldsymbol{x}'}{\partial \boldsymbol{x}}\boldsymbol{A} = \mathbb{I}\boldsymbol{A} = \boldsymbol{A}$$

$$\frac{\partial}{\partial x}(\boldsymbol{x}'\boldsymbol{A}') = \boldsymbol{A}'$$

- Consider the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ (inner products):

$$\frac{\partial \boldsymbol{u}'\boldsymbol{v}}{\partial x} = \frac{\partial \boldsymbol{u}'}{\partial x}\boldsymbol{v} + \frac{\partial \boldsymbol{v}'}{\partial x}\boldsymbol{u}$$

- Quadratic forms. Let $\boldsymbol{u}' = \boldsymbol{x}'$ and let $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{x}$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) &= \frac{\partial \boldsymbol{u}'}{\partial \boldsymbol{x}}\boldsymbol{v} + \frac{\partial \boldsymbol{v}'}{\partial \boldsymbol{x}}\boldsymbol{u} \\
&= \frac{\partial \boldsymbol{x}'}{\partial \boldsymbol{x}}\boldsymbol{A}\boldsymbol{x} + \frac{\partial (\boldsymbol{A}\boldsymbol{x})'}{\partial \boldsymbol{x}}\boldsymbol{x} \\
&= \frac{\partial \boldsymbol{x}'}{\partial \boldsymbol{x}}\boldsymbol{A}\boldsymbol{x} + \frac{\partial (\boldsymbol{x}'\boldsymbol{A}')}{\partial \boldsymbol{x}}\boldsymbol{x} \\
&= \frac{\partial \boldsymbol{x}'}{\partial \boldsymbol{x}}\boldsymbol{A}\boldsymbol{x} + \frac{\partial \boldsymbol{x}'}{\partial \boldsymbol{x}}\boldsymbol{A}'\boldsymbol{x} \\
&= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}'\boldsymbol{x}
\end{aligned}$$

When $\boldsymbol{A}$ is symmetric ($\boldsymbol{A}' = \boldsymbol{A}$) we have: $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}$.

$$RSS = \boldsymbol{y}'\boldsymbol{y} - 2\beta'\boldsymbol{X}'\boldsymbol{y} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta$$

- Let $\boldsymbol{a} = \boldsymbol{X}'\boldsymbol{y}$ and Let $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$. Now let's differentiate with respect to β:

$$\frac{dRSS}{d\beta} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\beta$$

- Set this equal to zero and solve (assuming $\boldsymbol{X}'\boldsymbol{X}$ is invertible):

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

- Let's get the expected value of $\hat{\beta}$:

$$\begin{aligned} E[\hat{\beta}] &= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{y}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta \\ &= \beta \end{aligned}$$

The estimator is unbiased!

- Now let's get the variance (actually covariance matrix) of $\hat{\beta}$:

$$
\begin{aligned}
V(\hat{\beta}) &= V((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= V(\boldsymbol{A}\boldsymbol{y}) = \boldsymbol{A}V(\boldsymbol{y})\boldsymbol{A}' \\
&= \boldsymbol{A}\left[\mathbb{I}\sigma^2\right]\boldsymbol{A}' \\
&= \boldsymbol{A}\boldsymbol{A}'\sigma^2 \\
&= \left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)'\sigma^2 \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right)'\sigma^2 \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}
\end{aligned}
$$

Note: $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is symmetric.

- Let's consider a prediction based on $\boldsymbol{x}_0$ a $\{(p+1)\times 1\}$ vector:

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$$

- Note:

$$E\left[\hat{y}_0\right] = E\left[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right] = \boldsymbol{x}_0'\boldsymbol{\beta}$$

$$V\left[\hat{y}_0\right] = V\left[\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}\right] = \sigma^2\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0$$

$$\text{Bias}(\hat{y}_0) = E\left[\hat{y}_0\right] - \boldsymbol{x}_0'\boldsymbol{\beta} = 0$$

$$\begin{aligned} E\left[(y_0-\hat{y}_0)^2\right] &= V(y_0) + V(\hat{y}_0) + [\text{Bias}(\hat{y}_0)]^2 \\ &= \sigma^2 + \sigma^2\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0 + 0^2 \end{aligned}$$

Advertising Data

```
y <- matrix(adv$sales, ncol=1)
y[1:3,]
```

```
## [1] 22.1 10.4  9.3
```

```
n <- length(y)

X <- as.matrix(cbind(rep(1, n), adv[,1:3]))
colnames(X); k <- ncol(X)
```

```
## [1] "rep(1, n)" "TV"         "radio"      "newspaper"
```

```
colnames(X)[1] <- "int"
X[1:3,]
```

```
##   int    TV radio newspaper
## 1   1 230.1  37.8      69.2
## 2   1  44.5  39.3      45.1
## 3   1  17.2  45.9      69.3
```

```
beta.hat <- solve(t(X)%*%X)%*%t(X)%*%y
beta.hat
```

```
##                    [,1]
## int          2.938889369
## TV           0.045764645
## radio        0.188530017
## newspaper -0.001037493
```

```
sigma.sq.hat <- t(y-X%*%beta.hat)%*%(y-X%*%beta.hat)/(n-k)
sigma.hat <- sqrt(sigma.sq.hat)
sigma.hat
```

```
##          [,1]
## [1,] 1.68551
```

```r
beta.hat.se <- sqrt(diag(c(sigma.sq.hat)*solve(t(X)%*%X)))
beta.hat.se
```

```
##         int          TV       radio   newspaper
## 0.311908236 0.001394897 0.008611234 0.005871010
```

```r
summary(mod.lm2)[[4]]
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## radio        0.188530017 0.008611234 21.8934961 1.505339e-54
## newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
```