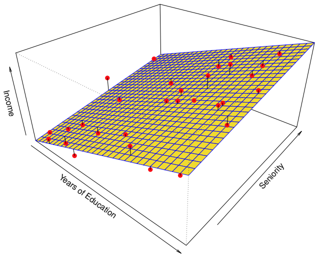
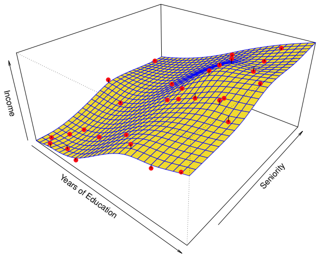
Linear Regression

Damek Davis
School of ORIE, Cornell University
ORIE 4740 Lec 2–3 (Jan 27, Feb 1)

Announcements

Recap





What kind of models are these?

A. Linear (left), Nonlinear (right)

B. Nonlinear (left), Linear (right)

Recap

1. Our goal is to model the relationship between predictor variable and a response variable.

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

2. Our goal is to put data into similar groups or to find a good or representation of data.

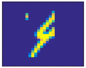
What kind of learning are we doing?

A. 1 = Supervised Learning, 2 = Unsupervised Learning

B. 1 = Unsupervised learning, 2 = Supervised Learning

Regression or Classification?

You want to build a model that determines whether the following images are fours or eights:

This is a

A. Regression Task

B. Classification Task

What to try next?

You have a training set with one predictor variable and one response variable. You fit a model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

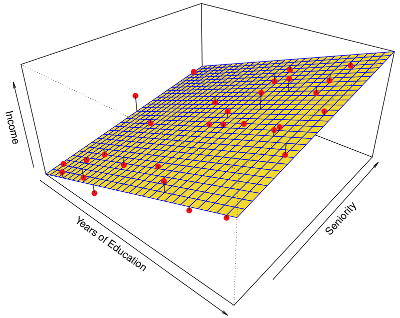
and get perfect training error. On the the other hand, you are astonished to find out that your model performs really poorly on the test data. Which model should you try next?

Choose one:

A. $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

B. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$

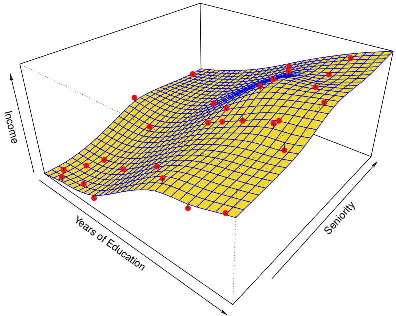
Recap: Supervised Learning



$$y = \underbrace{f(X_1, X_2)}_{\text{model}} + \underbrace{\epsilon}_{\text{error}}$$

$$y \approx \hat{f}(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

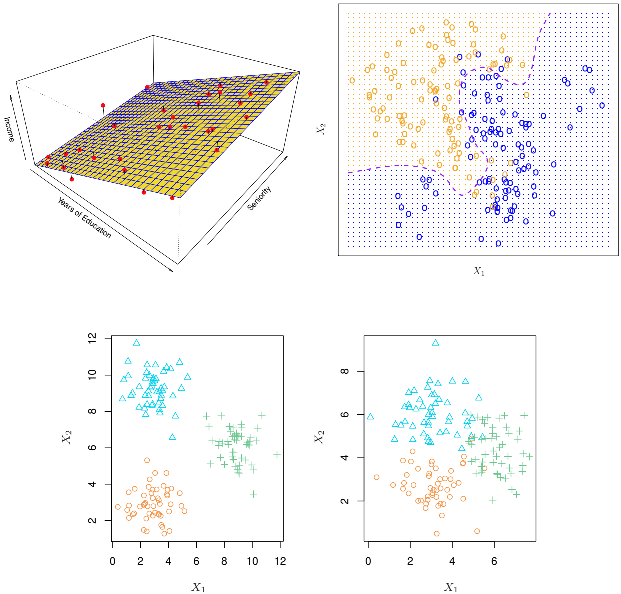
Recap: Supervised Learning



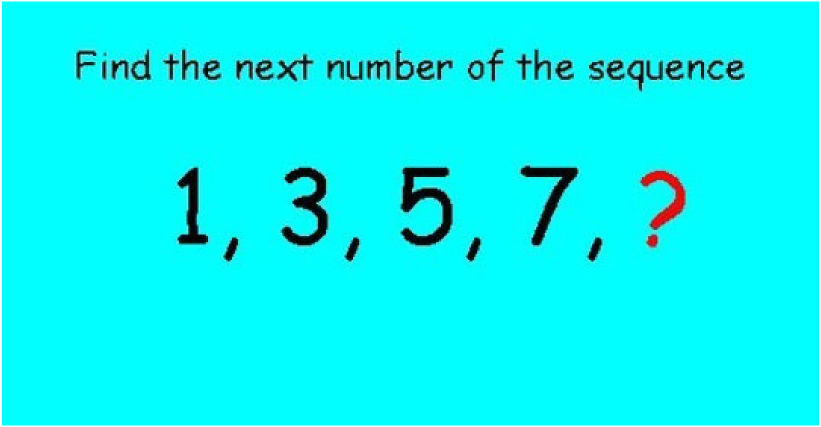
$$y = \underbrace{f(X_1, X_2)}_{\text{model}} + \underbrace{\epsilon}_{\text{error}}$$

$$\cancel{y \approx \hat{f}(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2}$$

Recap: Regression & Classification, Supervised & Unsupervised

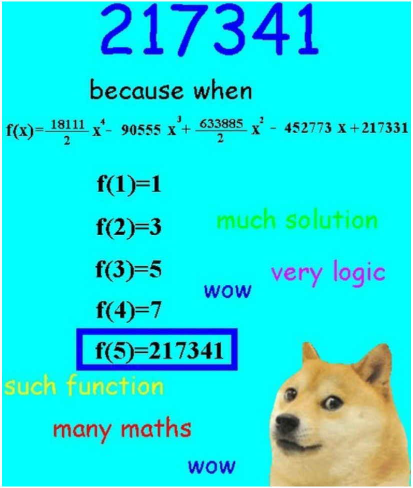


Recap: Training vs Testing Error



- training data = $\{(1,1),(2,3),(3,5),(5,7)\}$,
- $\hat{f}(i) = i$th odd number. Perfect training MSE

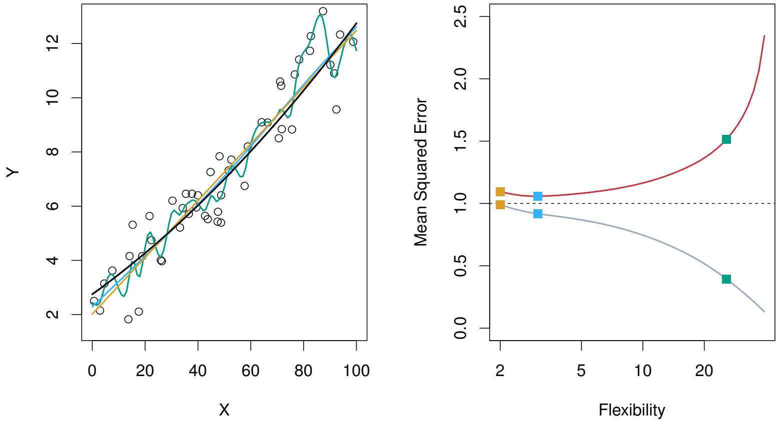
Recap: Training vs Testing Error



- testing data = $\{(5, 27341)\}$
- Testing MSE = $(9 - 27341)^2 = 747,038,224$

Recap: Flexibility vs. Interpretability

Which model will perform best on test data?

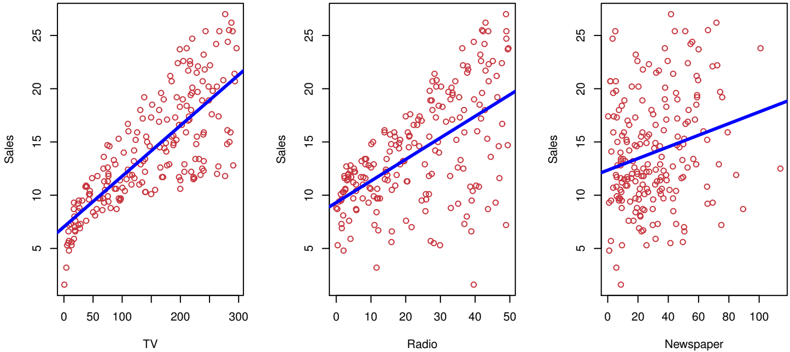


(training data, test error)

Linear Regression

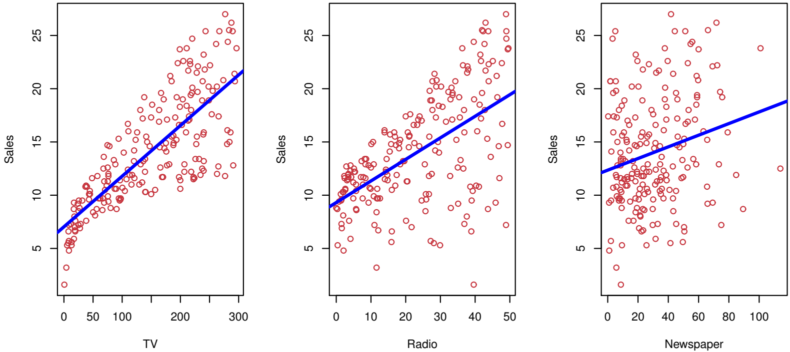
Reference: Chapter 3 of ISLR

Predicting Sales



- **Goal:** predict sales as a function of budget on TV, Radio, and Newspaper.

Predicting Sales

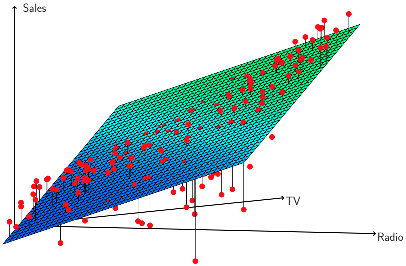


- **Goal:** predict sales as a function of budget on TV, Radio, and Newspaper.
- Independent predictions ignore relations between budgets, so may suggest using a more flexible model.

$$
\begin{aligned}
Y &= f(X) + \epsilon \\
&= \underbrace{f(X_1, X_2, X_3)}_{\text{model}} + \underbrace{\epsilon}_{\text{error}}
\end{aligned}
$$

Advertising Example

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75.0	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1.0	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24.0	4.0	17.4
23.8	35.1	65.9	9.2
97.5	7.6	7.2	9.7
204.1	32.9	46.0	19.0



Assume a linear model:

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i),$$

$$i = 1, 2, \ldots, n$$

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i), \quad i = 1, 2, \ldots, n$$

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i), \quad i = 1, 2, \ldots, n$$

Questions of interest:

- **Modeling**: Why linear regression?

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i),$$
$$i = 1, 2, \ldots, n$$

Questions of interest:

- **Modeling**: Why linear regression?
- **Estimation**: How to estimate $\beta_0, \ldots, \beta_3$?

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i), \quad i = 1, 2, \ldots, n$$

Questions of interest:

- **Modeling**: Why linear regression?
- **Estimation**: How to estimate $\beta_0, \ldots, \beta_3$?
- **Evaluation**: How strong is the relationship between advertising & sales?
- **Evaluation**: Which media contribute more to sales?

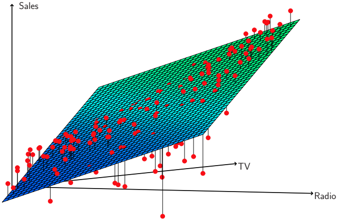
Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i), \quad i = 1, 2, \ldots, n$$

Questions of interest:

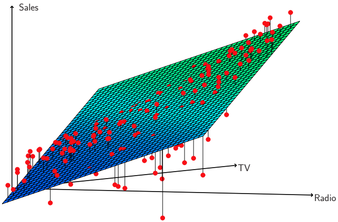
- **Modeling**: Why linear regression?
- **Estimation**: How to estimate $\beta_0, \ldots, \beta_3$?
- **Evaluation**: How strong is the relationship between advertising & sales?
- **Evaluation**: Which media contribute more to sales?
- **Evaluation**: How accurate is the prediction of our model?
- **Evaluation**: Is the relationship really linear?

Why Linear Regression?



- Real data/system is rarely strictly linear

Why Linear Regression?

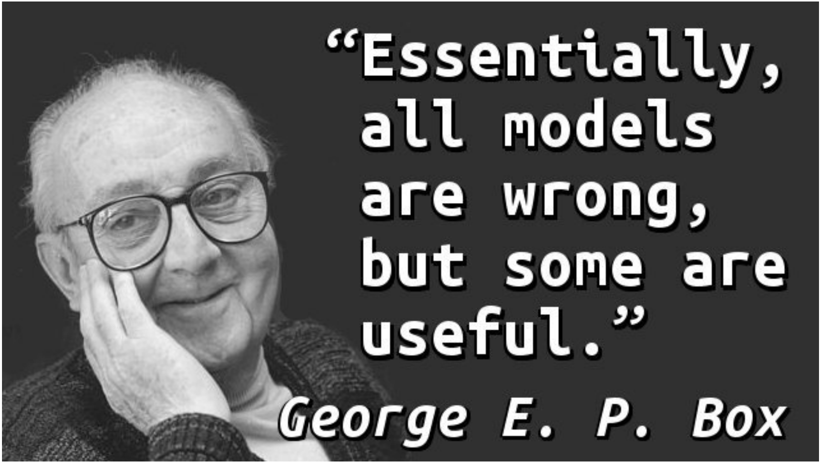


- Real data/system is rarely strictly linear

$$Y = \underbrace{f(X)}_{\text{explained part}} + \underbrace{\epsilon}_{\text{unexplained part}}$$

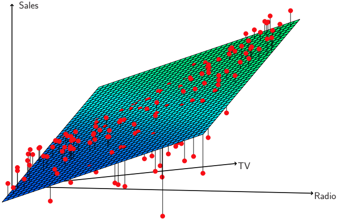
Example: Two homes: same characteristics, but different valuations.

All Models Wrong/Some Models are More Correct



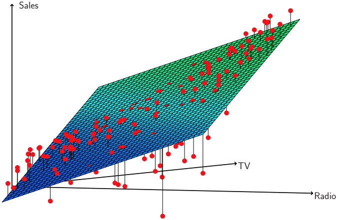
https://commons.wikimedia.org/wiki/File:GeorgeEPBox.jpg

Why Linear Regression



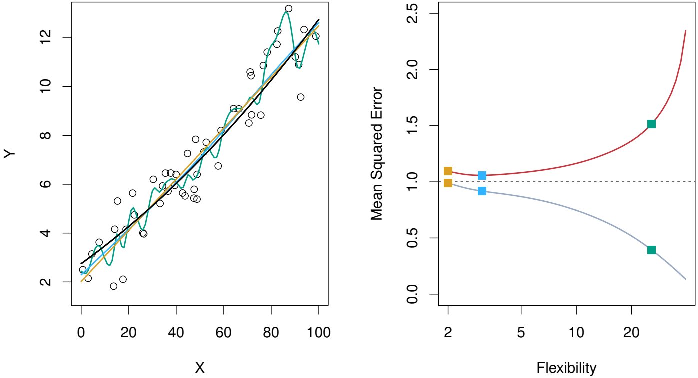
- Real data/system is rarely strictly linear
- Linear regression is simple and easy to interpret.

Why Linear Regression



- Real data/system is rarely strictly linear
- Linear regression is simple and easy to interpret.
- Building block for more sophisticated methods (Generalized linear models/logistic regression, Sparse linear models/LASSO)
- Not “too flexible”

Not too flexible



- Less likely to overfit (not fitting the noise)
- Often a good (first) approximation
- Work well on a new data point

Linear Regression: Estimation

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad i = 1, 2, \ldots, n$$

Linear Regression: Estimation

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad i = 1, 2, \ldots, n$$

In matrix form:

$$y \approx X\beta$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p+1)}, \beta \in \mathbb{R}^{p+1}$.

Linear Regression: Estimation

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad i = 1, 2, \ldots, n$$

In matrix form:

$$y \approx X\beta$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times (p+1)}, \beta \in \mathbb{R}^{p+1}$.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Linear Regression: Estimation

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad i = 1, 2, \ldots, n$$

In matrix form:

$$y \approx X\beta$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p+1)}, \beta \in \mathbb{R}^{p+1}$.

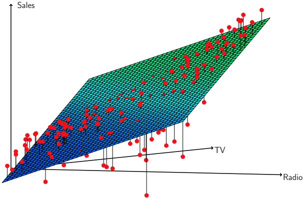
Least squares approach: Find β that minimize the sum of squared residuals

$$\text{RSS} \triangleq \sum_{i=1}^{n}(\underbrace{y_i}_{\text{True Response}} - \underbrace{(\beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip})}_{\text{Predicted Response}})^2 = (y - X\beta)^\top (y - X\beta).$$

Linear Regression: Estimation

Least Squares Sol'n: Want to find the $\beta \in \mathbb{R}^{p+1}$ that minimizes

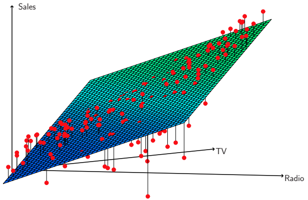
$$\text{RSS} = (y - X\beta)^\top (y - X\beta).$$



Linear Regression: Estimation

Least Squares Sol'n: Want to find the $\beta \in \mathbb{R}^{p+1}$ that minimizes

$$\text{RSS} = (y - X\beta)^\top (y - X\beta).$$



- From your linear algebra class, you know that

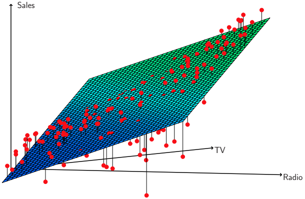
$$\boxed{\hat{\beta} = (X^\top X)^{-1} X^\top y,}$$

assuming that the matrix $X^\top X$ is invertible.

Linear Regression: Estimation

Least Squares Sol'n: Want to find the $\beta \in \mathbb{R}^{p+1}$ that minimizes

$$\mathrm{RSS} = (y - X\beta)^\top (y - X\beta).$$



- From your linear algebra class, you know that

$$\boxed{\hat{\beta} = (X^\top X)^{-1} X^\top y,}$$

  assuming that the matrix $X^\top X$ is invertible.

- **Notation.** We put a ˆ (hat) over β to indicate it was estimated from data.

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i)$$

- Compute least square solution $\hat{\beta} = (X^\top X)^{-1} X^\top y$,

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i)$$

- Compute least square solution $\hat{\beta} = (X^\top X)^{-1} X^\top y$,

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i)$$

- Compute least square solution $\hat{\beta} = (X^\top X)^{-1} X^\top y$,

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

- Interpretation:

For every extra \$1000 spent on radio, we increase sales by 189 units, holding TV and Newspaper fixed.

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times (\text{radio}(i) + 1000) + \beta_3 \times \text{newspaper}(i)$$

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i)$$

- Compute least square solution $\hat{\beta} = (X^\top X)^{-1} X^\top y$,

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

All other budgets being fixed, which would you recommend?

A. increase newspaper budget

B. decrease newspaper budget

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i)$$

- Compute least square solution $\hat{\beta} = (X^\top X)^{-1} X^\top y$,

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

- Residuals: actual $-$ predicted $= (y_i - X_i \hat{\beta})$

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i)$$

- Compute least square solution $\hat{\beta} = (X^\top X)^{-1} X^\top y$,

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

From the table, we see that the the model tends to...

A. overestimate sales

B. underestimate sales

Advertising Example

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i)$$

- Compute least square solution $\hat{\beta} = (X^\top X)^{-1} X^\top y$,

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

- Using the model: given a new x_{new}, predict $y_{\text{new}} = x_{\text{new}}^\top \hat{\beta}$.

Recap: Linear Regression

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad i = 1, 2, \ldots, n$$

In matrix form:

$$y \approx X\beta$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p+1)}, \beta \in \mathbb{R}^{p+1}$.

What is the first column of X?

A. $\beta_0 \mathbf{1}_n$

B. $\mathbf{1}_n$

Recap: Linear Regression

Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p+1)}$, find $\beta \in \mathbb{R}^{p+1}$ s.t.

$$y \approx X\beta.$$

- Least squares approach: find β that minimizes

$$\text{RSS} \triangleq \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = (y - X\beta)^\top (y - X\beta).$$

In some cases, the least squares solution $\hat{\beta}$ is:

A. $\hat{\beta} = X^{-1}y$

B. $\hat{\beta} = (X^\top X)^{-1}X^\top y$

C. $\hat{\beta} = (X^\top X)^{-1}y$

Recap: Linear Regression

Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p+1)}$, find $\beta \in \mathbb{R}^{p+1}$ s.t.

$$y \approx X\beta.$$

- Least squares approach: find β that minimizes

$$\text{RSS} \triangleq \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = (y - X\beta)^\top (y - X\beta).$$

- Least squares solution:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

assuming $X^\top X$ invertible.

Recap: Linear Regression

Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p+1)}$, find $\beta \in \mathbb{R}^{p+1}$ s.t.

$$y \approx X\beta.$$

- Least squares approach: find β that minimizes

$$\text{RSS} \triangleq \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = (y - X\beta)^\top (y - X\beta).$$

- Least squares solution:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

assuming $X^\top X$ invertible.

Given a new x_{n+1}, how do we predict the its response $\hat{y}_{n+1}$?

A. predict $\hat{y}_{n+1} = x_{n+1}^\top \hat{\beta}$

B. predict $\hat{y}_{n+1} = x_{n+1}^\top \hat{\beta} + \epsilon_{n+1}$

Recap: Linear Regression

Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times(p+1)}$, find $\beta \in \mathbb{R}^{p+1}$ s.t.

$$y \approx X\beta.$$

- Least squares approach: find β that minimizes

$$\mathrm{RSS} \triangleq \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = (y - X\beta)^\top (y - X\beta).$$

- Least squares solution:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

assuming $X^\top X$ invertible.

- Using the model: given a new x_{n+1}, predict $\hat{y}_{n+1} = x_{n+1}^\top \hat{\beta}$.

- Why linear regression?
- How to estimate $\beta_0, \ldots, \beta_3$?

- Why linear regression?
- How to estimate $\beta_0, \ldots, \beta_3$?
- Model evaluation:
  - How accurate is the prediction of our model?
  - How strong is the relationship between advertising and sales?
  - Which media contribute to sales?
  - Is the relationship linear?

- Why linear regression?
- How to estimate $\beta_0, \ldots, \beta_3$?
- Model evaluation:
  - How accurate is the prediction of our model?
  - How strong is the relationship between advertising and sales?
  - Which media contribute to sales?
  - Is the relationship linear?

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

Model Evaluation

- Accuracy of the coefficient estimate $\hat{\beta}_j$

- Accuracy of the model

Accuracy of Coefficient Estimates

Assume that the true model is indeed linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

How to estimate

$$|\underbrace{\beta_j}_{\textit{True}} - \underbrace{\hat{\beta}_j}_{\textit{Estimated}}|?$$

Accuracy of Coefficient Estimates

Assume that the true model is indeed linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

How to estimate

$$|\underbrace{\beta_j}_{\textit{True}} - \underbrace{\hat{\beta}_j}_{\textit{Estimated}}|?$$

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{RSS}/(n-p-1) \cdot [(X^\top X)^{-1}]_{ii}}$$

	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	

Accuracy of Coefficient Estimates

How to estimate

$$|\underbrace{\beta_j}_{\textit{True}} - \underbrace{\hat{\beta}_j}_{\textit{Estimated}}|?$$

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{RSS}/(n-p-1) \cdot [(X^\top X)^{-1}]_{jj}}.$$

- We use it to construct confidence intervals. For example,

$$\left[\hat{\beta}_j - 2 \cdot \mathrm{SE}(\hat{\beta}_j), \hat{\beta}_j + 2 \cdot \mathrm{SE}(\hat{\beta}_j)\right]$$

  has a 95% chance of containing the true β_j.
  (A “95% confidence interval”.)

Accuracy of Coefficient Estimates

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j)$$

- An estimate of $|\hat{\beta}_j - \beta_j|$

Accuracy of Coefficient Estimates

Standard Error (SE) of each estimate:

$$\text{SE}(\hat{\beta}_j)$$

- An estimate of $|\hat{\beta}_j - \beta_j|$
- Can be used to construct confidence intervals.

Accuracy of Coefficient Estimates

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j)$$

- An estimate of $|\hat{\beta}_j - \beta_j|$
- Can be used to construct confidence intervals.
- Can be used to perform hypothesis tests:

Null: $H_0 : \beta_j = 0$ (no relationship b/w y and x_j with other vars fixed)

versus

Alt: $H_1 : \beta_j \neq 0$ (some relationship b/w y and x_j with other vars fixed)

	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	

Accuracy of Coefficient Estimates

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j)$$

- An estimate of $|\hat{\beta}_j - \beta_j|$
- Can be used to construct confidence intervals.
- Can be used to perform hypothesis tests:

Null: $H_0 : \beta_j = 0$ (no relationship b/w y and x_j with other vars fixed)

versus

Alt: $H_1 : \beta_j \neq 0$ (some relationship b/w y and x_j with other vars fixed)

	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	

- The smaller p-value, the more evidence against H_0 (i.e., small p-value $\implies$ evidence of some relationship).

p-value: further reading

- Kareem Carr: Don't know what a P-VALUE is?
  https://twitter.com/kareem_carr/status/1312783404975493122
- xkcd
  https://xkcd.com/882/
- Kim and Heejung. *Three common misuses of P values*

Accuracy of Model

How well does the linear model fit the training data?

Accuracy of Model

How well does the linear model fit the training data?

- Residual Sum of Squares (RSS):

$$\text{RSS} \triangleq \sum_{i=1}^{n} \Big(y_i - \underbrace{\hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots \hat{\beta}_p x_{ip}}_{\hat{y}_i}\Big)^2 = \sum_{i=1}^{n} \big(y_i - \hat{y}_i\big)^2.$$

Intuition: how well we fit the data.

Accuracy of Model

How well does the linear model fit the training data?

- Residual Sum of Squares (RSS):

$$\mathrm{RSS} \triangleq \sum_{i=1}^{n} \big(y_i - \underbrace{\hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots \hat{\beta}_p x_{ip}}_{\hat{y}_i}\big)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Intuition: how well we fit the data.

- A similar quantity: Residual Standard Error (RSE)

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

  - Standard deviation of residual.

Accuracy of Model

How well does the linear model fit the training data?

- Residual Sum of Squares (RSS):

$$\text{RSS} \triangleq \sum_{i=1}^{n} \big(y_i - \underbrace{\hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots \hat{\beta}_p x_{ip}}_{\hat{y}_i}\big)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Intuition: how well we fit the data.

- A similar quantity: Residual Standard Error (RSE)

$$\text{RSE} \triangleq \sqrt{\frac{1}{n-p-1}\text{RSS}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}.$$

  - Standard deviation of residual.

RSS/RSE provides an absolute measure of lack of fit of the model to the data.

Accuracy of Model

R^2 statistics: A more useful measure.

Accuracy of Model

R^2 statistics: A more useful measure.

- TSS $\triangleq \sum_{i=1}^{n}(y_i - \bar{y})^2$: Total Sum of Squares
  (total amount of variability of the response variable)
  (here $\bar{y} \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average of the response values y_i.)

- RSS $\triangleq \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$: Residual Sum of Squares
  (amount of variability unexplained by the linear model)

Accuracy of Model

R^2 statistics: A more useful measure.

- TSS $\triangleq \sum_{i=1}^{n}(y_i - \bar{y})^2$: Total Sum of Squares
  (total amount of variability of the response variable)
  (here $\bar{y} \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average of the response values y_i.)

- RSS $\triangleq \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$: Residual Sum of Squares
  (amount of variability unexplained by the linear model)

- $$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = \frac{\text{total var.} - \text{unexplained var.}}{\text{total var.}} = \frac{\text{explained var.}}{\text{total var.}}$$

Proportion of variability of the response explained by the linear model

TSS vs RSS

R^2 statistics: A more useful measure.

- TSS $\triangleq \sum_{i=1}^{n}(y_i - \bar{y})^2$: Total Sum of Squares
  (total amount of variability of the response variable)
  (here $\bar{y} \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average of the response values y_i.)
  (the “no-model" error, since we could predict using avg)

- RSS $\triangleq \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$: Residual Sum of Squares
  (amount of variability unexplained by the linear model)

What can we conclude about TSS and RSS?

A. TSS $\geq$ RSS

B. TSS $\leq$ RSS

C. No relationship in general.

Range of the R^2 Statistic

- Residual Sum of Squares: $\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$.
- Residual Standard Error: $\text{RSE} \triangleq \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}$.
- R^2 statistics: $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = \frac{\text{explained var.}}{\text{total var.}}$

What can we conclude about R^2?

A. $-1 \leq R^2 \leq 1$

B. $0 \leq R^2 \leq 2$

C. $0 \leq R^2 \leq 1$

Range of the R^2 Statistic

- Residual Sum of Squares: RSS $\triangleq \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$.
- Residual Standard Error: RSE $\triangleq \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}$.
- R^2 statistics: $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = \frac{\text{explained var.}}{\text{total var.}}$

If we train two models, which one explains more of the data, one for which

A. $R^2 = 1$; or

B. $R^2 \approx 0$?

Recap: Model Evaluation—Accuracy of $\hat{\beta}_j$

Assume that the true model is indeed linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{RSS}/(n-p-1) \cdot [(X^\top X)^{-1}]_{ii}}.$$

- An estimate of $|\hat{\beta}_j - \beta_j|$

Recap: Model Evaluation—Accuracy of $\hat{\beta}_j$

Assume that the true model is indeed linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{RSS}/(n-p-1) \cdot [(X^\top X)^{-1}]_{jj}}.$$

- An estimate of $|\hat{\beta}_j - \beta_j|$
- Can be used to construct confidence intervals. E.g.,

$$\left[\hat{\beta}_j - 2 \cdot \mathrm{SE}(\hat{\beta}_j), \hat{\beta}_j + 2 \cdot \mathrm{SE}(\hat{\beta}_j)\right]$$

  has a 95% chance of containing the true β_j.
  (A “95% confidence interval”.)

Recap: Model Evaluation—Accuracy of $\hat{\beta}_j$

Assume that the true model is indeed linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{RSS}/(n-p-1) \cdot [(X^\top X)^{-1}]_{jj}}.$$

- An estimate of $|\hat{\beta}_j - \beta_j|$
- Can be used to construct confidence intervals.

Recap: Model Evaluation—Accuracy of $\hat{\beta}_j$

Assume that the true model is indeed linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{RSS}/(n-p-1) \cdot [(X^\top X)^{-1}]_{jj}}.$$

- An estimate of $|\hat{\beta}_j - \beta_j|$
- Can be used to construct confidence intervals.

Consider the sales model

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i), \quad i = 1, 2, \ldots, n$$

Which could you answer with a confidence interval for β_2:

A. Is there a relationship between sales and radio?

B. Is there a relationship between radio and newspaper?

	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	

Recap: Model Evaluation—Accuracy of $\hat{\beta}_j$

Assume that the true model is indeed linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{RSS}/(n-p-1) \cdot [(X^\top X)^{-1}]_{jj}}.$$

- An estimate of $|\hat{\beta}_j - \beta_j|$
- Can be used to construct confidence intervals.

- Can be used to perform hypothesis tests:

Null: $H_0 : \beta_j = 0$ (no relationship b/w y and x_j)

versus

Alternative: $H_1 : \beta_j \neq 0$ (some relationship b/w y and x_j)

	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	

Recap: Model Evaluation—Accuracy of $\hat{\beta}_j$

Assume that the true model is indeed linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Standard Error (SE) of each estimate:

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{RSS}/(n-p-1) \cdot [(X^\top X)^{-1}]_{jj}}.$$

- An estimate of $|\hat{\beta}_j - \beta_j|$
- Can be used to construct confidence intervals.

- Can be used to perform hypothesis tests:

Null: $H_0 : \beta_j = 0$ (no relationship b/w y and x_j)

versus

Alternative: $H_1 : \beta_j \neq 0$ (some relationship b/w y and x_j)

	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	

- The smaller p-value, the more evidence against H_0 (i.e., small p-value $\Longrightarrow$ evidence of some relationship).

Recap: Model Evaluation—Accuracy of the Model?

- Accuracy of Model:
  - Residual Sum of Squares: $\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$.
  - Residual Standard Error: $\text{RSE} \triangleq \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}$.
  - Total Sum of Squares: $\text{TSS} \triangleq \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$
    where $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.
  - R^2 statistics: $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}}$

Recap: Model Evaluation—Accuracy of the Model?

- Accuracy of Model:
  - Residual Sum of Squares: $\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$.
  - Residual Standard Error: $\text{RSE} \triangleq \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}$.
  - Total Sum of Squares: $\text{TSS} \triangleq \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$
    where $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.
  - R^2 statistics: $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}}$

Which of the following does the R^2 statistic measure?

A. the total amount of variability that is unexplained after performing the regression.

B. the proportion of variability in response that is explained by performing regression.

Accuracy of Model

- Residual Sum of Squares: $\mathrm{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$.
- Residual Standard Error: $\mathrm{RSE} \triangleq \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}$.
- R^2 statistics: $R^2 = \frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}} = \frac{\text{explained var.}}{\text{total var.}}$

Accuracy of Model

- Residual Sum of Squares: RSS $\triangleq \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$.
- Residual Standard Error: RSE $\triangleq \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$.
- R^2 statistics: $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = \frac{\text{explained var.}}{\text{total var.}}$

Advertising example:

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

Accuracy of Model

- RSS and R^2 can be used to perform hypothesis test of the whole model:

$$\text{Null: } H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

(no relationship b/w y and all predictors x_j)

versus

$$\text{Alternative: } H_1 : \text{at least one } \beta_j \text{ is non-zero}$$

(some relationship b/w y and the predictors)

- Done by computing the F-statistic (details omitted; cf. ISLR pp75-76)

Accuracy of Model

- RSS and R^2 can be used to perform hypothesis test of the whole model:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \text{ is non-zero}$$

Accuracy of Model

- RSS and R^2 can be used to perform hypothesis test of the whole model:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \text{ is non-zero}$$

Advertising example:

```
> lm.fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(lm.fit)
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
```

- The smaller p-value, the more evidence of some relationship.

Summary: Model Evaluation Techniques

- Accuracy of each coefficient estimate $\hat{\beta}_j$
  - SE($\hat{\beta}_j$)
  - Confidence intervals and hypothesis testing for each $\hat{\beta}_j$

- Accuracy of the model
  - RSS and RSE
  - R^2 statistic
  - Hypothesis testing for the whole linear model

Summary: Model Evaluation Techniques

- Accuracy of each coefficient estimate $\hat{\beta}_j$
  - SE($\hat{\beta}_j$)
  - Confidence intervals and hypothesis testing for each $\hat{\beta}_j$

- Accuracy of the model
  - RSS and RSE
  - R^2 statistic
  - Hypothesis testing for the whole linear model

Will use them to answer:

- How accurate is the prediction of our model?
- How strong is the relationship between advertising and sales?
- Which media contribute to sales?
- Is the relationship linear?

The Advertising Example

```
> Advertising = read.csv("Advertising.csv", header = T)
> fit = lm(Sales~TV+Radio+Newspaper, data = Advertising)
> summary(fit)

Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = Advertising)

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

> confint(fit, level=0.95)
                  2.5 %     97.5 %
(Intercept)  2.32376228 3.55401646
TV           0.04301371 0.04851558
Radio        0.17154745 0.20551259
Newspaper   -0.01261595 0.01054097
```

Model Evaluation: Advertising Example

How accurate is the prediction of our model?

	Coefficient	SE	t-statistics	p-value	95% conf. int.
Intercept	2.939	0.3119	9.42	<0.0001	[2.323, 3.554]
TV	0.046	0.0014	32.81	<0.0001	[0.043, 0.049]
radio	0.189	0.0086	21.89	<0.0001	[0.172, 0.206]
newspaper	-0.001	0.0059	-0.18	0.8599	[−0.013, 0.011]

Model Evaluation: Advertising Example

How strong is the relationship between advertising and sales?

Model Evaluation: Advertising Example

How strong is the relationship between advertising and sales?

- Hypothesis test:

$$\text{Null: } H_0 : \beta_{\text{TV}} = \beta_{\text{radio}} = \beta_{\text{newspaper}} = 0$$

vs.

$$\text{Alternative: } H_1 : \text{at least one of the } \beta\text{'s is non-zero}$$

p-value < 0.0001.

Model Evaluation: Advertising Example

How strong is the relationship between advertising and sales?

- Hypothesis test:

$$\text{Null: } H_0 : \beta_{\text{TV}} = \beta_{\text{radio}} = \beta_{\text{newspaper}} = 0$$

vs.

$$\text{Alternative: } H_1 : \text{at least one of the } \beta\text{'s is non-zero}$$

  p-value < 0.0001.

- Strong evidence of a relationship given all of the modeling assumptions.

Model Evaluation: Advertising Example

How strong is the relationship between advertising and sales?

- Hypothesis test:

$$\text{Null: } H_0 : \beta_{\text{TV}} = \beta_{\text{radio}} = \beta_{\text{newspaper}} = 0$$

vs.

$$\text{Alternative: } H_1 : \text{at least one of the } \beta\text{'s is non-zero}$$

  p-value < 0.0001.

- Strong evidence of a relationship given all of the modeling assumptions.
- RSE = 1681. $\bar{y} = \text{mean}(y_i) = 14022$.

Model Evaluation: Advertising Example

How strong is the relationship between advertising and sales?

- Hypothesis test:

$$\text{Null: } H_0 : \beta_{\text{TV}} = \beta_{\text{radio}} = \beta_{\text{newspaper}} = 0$$

$$\text{vs.}$$

$$\text{Alternative: } H_1 : \text{at least one of the } \beta\text{'s is non-zero}$$

  p-value < 0.0001.

- Strong evidence of a relationship given all of the modeling assumptions.
- RSE = 1681. $\bar{y} = \text{mean}(y_i) = 14022$.
- $R^2 = 0.8972$.

Model Evaluation: Advertising Example

Which media contribute to sales?

	Coefficient	SE	t-statistics	p-value	95% conf. int.
Intercept	2.939	0.3119	9.42	<0.0001	[2.323, 3.554]
TV	0.046	0.0014	32.81	<0.0001	[0.043, 0.049]
radio	0.189	0.0086	21.89	<0.0001	[0.172, 0.206]
newspaper	-0.001	0.0059	-0.18	0.8599	[-0.013, 0.011]

Model Evaluation: Advertising Example

Which media contribute to sales?

	Coefficient	SE	t-statistics	p-value	95% conf. int.
Intercept	2.939	0.3119	9.42	<0.0001	[2.323, 3.554]
TV	0.046	0.0014	32.81	<0.0001	[0.043, 0.049]
radio	0.189	0.0086	21.89	<0.0001	[0.172, 0.206]
newspaper	-0.001	0.0059	-0.18	0.8599	[−0.013, 0.011]

- In the presence of TV and Radio, Newspaper is not significant.

Model Evaluation: Advertising Example

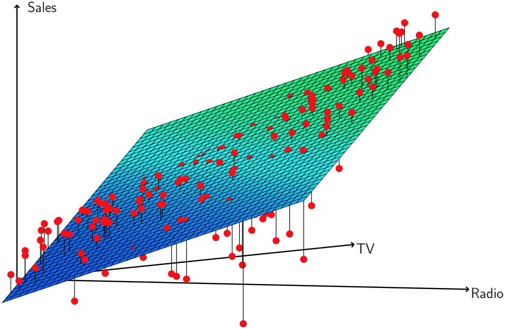
Which media contribute to sales?

	Coefficient	SE	t-statistics	p-value	95% conf. int.
Intercept	2.939	0.3119	9.42	<0.0001	[2.323, 3.554]
TV	0.046	0.0014	32.81	<0.0001	[0.043, 0.049]
radio	0.189	0.0086	21.89	<0.0001	[0.172, 0.206]
newspaper	-0.001	0.0059	-0.18	0.8599	[$-$0.013, 0.011]

- In the presence of TV and Radio, Newspaper is not significant.
- Newspaper on its own may be significant.

Model Evaluation: Advertising Example

Is the relationship linear?
Or is it better to use a nonlinear model?



More on this later.

Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
The digits were taken from the MNIST Dataset.
Slides based on Yudong Chen’s slides.
Some images due to Machine learning @ Berkeley Group

Running a linear regression in **R** we get the following output:

```
Call: lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.31384 -0.67054  0.01942  0.62198  2.35304

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01030    0.06014  -0.171    0.864
x1           1.02598    0.07512  13.658   <2e-16 ***
x2           0.07300    0.08409   0.868    0.387
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 0.3079 on 97 degrees of freedom
Multiple R-squared:  0.8892,    Adjusted R-squared:  0.8881
F-statistic: 790.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

Based on this output we can say that

- **A** The predictor $x1$ can be dropped from the linear model, since it does not help to predict y in the presence of the 2nd predictor.
- **B** The predictor $x2$ can be dropped from the linear model, since it does not help to predict y in the presence of the 1st predictor.
- **C** Both the predictors $x1$ and $x2$ can be dropped from the linear model since they have no relationship with y.

```
Call: lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.31384 -0.67054  0.01942  0.62198  2.35304

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01030    0.06014  -0.171    0.864
x1           1.02598    0.07512  13.658   <2e-16 ***
x2           0.07300    0.08409   0.868    0.387
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 0.3079 on 97 degrees of freedom
Multiple R-squared:  0.8892,    Adjusted R-squared:  0.8881
F-statistic: 790.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

Which of the following is **wrong**

A There is **no** strong evidence that the intercept coefficient is non-zero.

B There is a 88.92% chance that there is some linear relationship between the response and the two predictors.

C A 95% confidence interval for $x2$ is: $0.073 \pm 2 \times 0.084$

Appendix: Useful Math (Optional)

Block Matrix Operation

You learned the rules of matrix multiplication. Multiplying 2 two-by-two matrices, for example, is done as follows

$$\begin{bmatrix} a_{11} & a_{11} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{11} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}.$$

often in statistics and data-mining we deal with block matrices and vectors. This simply means that we construct bigger matrices using smaller matrices as building blocks. Block matrix multiplication can be expressed as follows:

$$\begin{bmatrix} A_{11} & A_{11} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{11} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix},$$

where we have simply replaced scalars in the first example with matrices. As you can see the rules of matrix muliplication apply in the same exact way, except that scalar multiplication has been replaced with matrix multiplication. Note that the formula for block matrix multiplication above is extremely general. You are free to partition the original matrix into 4 blocks of any dimension, as long as the dimensions of internal matrix multiplications agree.

Matrix Algebra and Calculus

Example: $g(w) = (y - Xw)^\top (y - Xw)$, where y and w are vectors and X is a matrix of appropriate dimension.

1. Verify that

$$g(w) = y^\top y - 2y^\top Xw + w^\top X^\top Xw,$$

(since $y^\top Xw = w^\top X^\top y$, Why?)

2. Taking a derivative of a scalar with respect to a vector:

$$\frac{\mathrm{d}g(w)}{\mathrm{d}w} = -2X^\top (y - Xw).$$

Compare this with the scalar case: if $g(w) = (b - aw)^2$, where a, b, w are scalars, then by chain rule

$$\frac{\mathrm{d}g(w)}{\mathrm{d}w} = -2a(b - aw).$$

Rule of Thumb: Taking derivative w.r.t. a vector has similar forms as the scalar case, as long as the dimensions of matrix multiplications agree.

Matrix Algebra and Calculus II

Exercise: Compute the following

- $\frac{\mathrm{d}(y^\top y)}{\mathrm{d}w}$
- $\frac{\mathrm{d}(y^\top Xw)}{\mathrm{d}w}$
- $\frac{\mathrm{d}(w^\top X^\top Xw)}{\mathrm{d}w}$

Hint: each of your answers should be a vector of the same dimension as w.