Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for Civil & Environmental Engineers

In-Class Material: Class 05

Linear Regression (ISL Chapter 3)

Given: Sample data set $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$

Goal: Find the l_____ relationship btw. two random variables X and Y, i.e. $Y = \beta_0 + \beta_1 X$

→ "(Simple) Linear Regression" Analysis

Important questions to address:

1. Is there a relationship between X and Y?
2. How strong is the relationship?
3. Which of X_i contributes to Y?
4. How can we estimate the effect of each X_i on Y?
5. How can we predict Y in future?
6. Is the relationship linear?
7. Is there synergy among $X_i's$?

1. Simple Linear Regression

(a) Assumption: There is approximately a l_____ relationship between X and Y

$Y \approx \beta_0 + \beta_1 X$ where X is a "p_____" variable and Y is a "r_____"

(b) Estimation of model coefficients β_0 (**"intercept"**) and β_1 (**"slope"**)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$

Note: slope β_1 represents the average effect of increasing the response by unit increase in the predictor

(c) "Best" estimates on $\hat{\beta}_0$ and $\hat{\beta}_1$: the values minimizing the sum of the square of the residual $e_i = y_i - \hat{y}_i$ between the ith observed response value and the ith prediction by the linear model $\hat{\beta}_0 + \hat{\beta}_1 x$ ("least square estimator")

Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize RSS → Solve the following equations for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = 2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

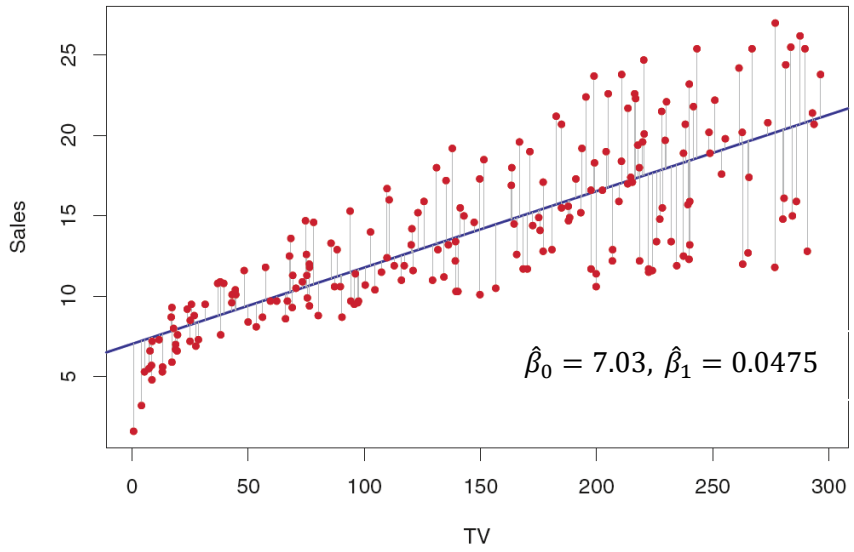
Instructor: Junho Song
junhosong@snu.ac.kr

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

As a result,

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$




```
library(MASS) # MASS library
fix(Boston) # Load 'Boston' data to the editor
attach(Boston)
lm.fit =lm(medv~lstat)
# Fit a simple linear regression with medv as the response and lsat as the
   predictor
lm.fit # Show some basic information about the model
summary(lm.fit) # show more detailed information about the model
plot(lstat,medv) # scatter plot
abline(lm.fit) # least squares regression line
plot(predict(lm.fit), residuals(lm.fit))
# residuals from a linear regression
```

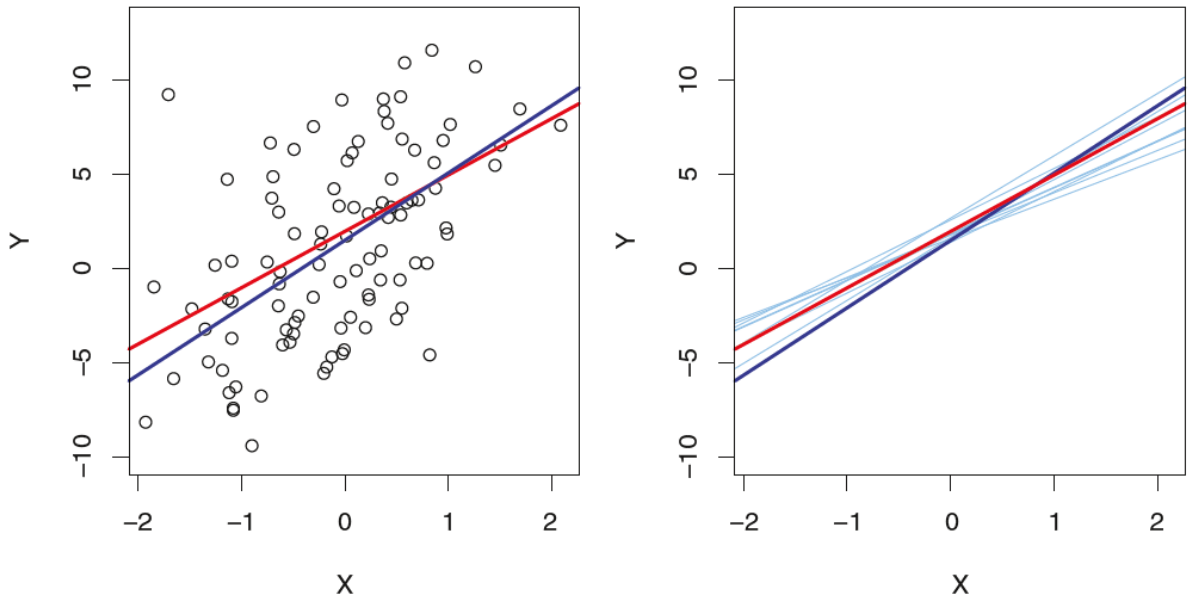
2. Assessing the Accuracy of the Estimated Model Coefficients

(a) "Population" regression line: best linear approximation to the true relationship between X and Y

$Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$ where ϵ is a mean-zero random error term.

Assumption: The error term is independent of X, and the errors ϵ_i for different observations are uncorrelated with each other and share *common* variance σ^2

Instructor: Junho Song
junhosong@snu.ac.kr



In the figure, red line: true model, dark blue: least square line, and light blue: least square lines from different samples

(b) **"Standard error"** (SE) from statistical uncertainty in the coefficients β_0 and β_1

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

In general, σ^2 is not known, but can be estimated by the **"residual standard error"** (RSE), i.e. the average deviation of the response from the true regression line:

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2/(n-2)}$$

(c) **"Confidence interval"** on β_0 and β_1

Assuming β_0 and β_1 follow *t*-distribution, $(1-\alpha)\times 100(\%)$ confidence interval on the model coefficients are given as

$$\left[\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\cdot SE(\hat{\beta}_0),\ \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\cdot SE(\hat{\beta}_0)\right], \quad \left[\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\cdot SE(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\cdot SE(\hat{\beta}_1)\right]$$

(d) **"Confidence interval"** on the estimated conditional mean $y'(x) = \mathrm{E}[y|x] = \hat{\beta}_0 + \hat{\beta}_1 x$

Due to the statistical uncertainty in the parameters $\hat{\beta}_0$ and $\hat{\beta}$,

$$y'(x) \sim N\left(\hat{\beta}_0 + \hat{\beta}_1 x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$

Instructor: Junho Song
junhosong@snu.ac.kr

Since y' also follows *t*-distribution when σ is estimated from the data, $(1-\alpha)\times 100(\%)$ confidence interval on the conditional mean is given as

$$\left[y'(x) - t_{\frac{\alpha}{2},n-2}\cdot RSE\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}},\quad y'(x) + t_{\frac{\alpha}{2},n-2}\cdot RSE\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]$$

(e) **"Prediction interval"** on $y(x)=\beta_0+\beta_1 x+\epsilon$

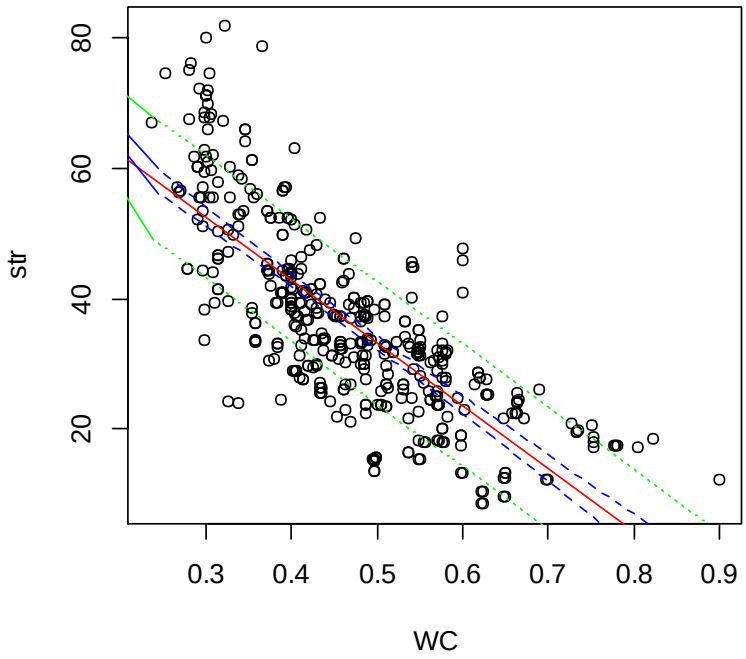
Now additionally including the uncertainty of the residual ϵ, the prediction interval on the dependent variable y at x is derived as

$$\left[y'(x) - t_{\frac{\alpha}{2},n-1}\cdot RSE\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, y'(x) + t_{\frac{\alpha}{2},n-1}\cdot RSE\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]$$

```
confint(lm.fit) # Obtain a confidence interval for coefficient estimates
confint(lm.fit, level = 0.99) # 99%
predict(lm.fit, data.frame(lstat=c(5,10,15)), interval ="confidence")
predict(lm.fit, data.frame(lstat=c(5,10,15)), interval ="prediction")
# Obtain confidence interval and prediction interval for the prediction of
medv for a given value of lsat
# can use the option 'level =' to specifice level (default 95%)
```

CEE Example: Linear regression of concrete strengths (str) on water/cement ratio (WC) (data provided by Prof. Juhyuk Moon at Multi-scale Structural Materials Lab)

- Blue: Confidence interval (95%)
- Green: Prediction interval (68%)



Instructor: Junho Song
junhosong@snu.ac.kr

(f) Hypothesis test

Null Hypothesis H_0: There is no relationship between X and Y, $\beta_1 = 0$
Alternative Hypothesis H_1: There is some relationship between X and Y, $\beta_1 \neq 0$

Computing *t*-statistic to reach one of the following conclusions:
- **Reject** H_0 in favor of H_1 because of sufficient evidence in the data or
- **Fail to reject** H_0 because of insufficient evidence in the data

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

(g) p-value (significance probability)

If the p-value, i.e. $P(T > |t|)$ is smaller than α, "**level of significance**," H_0 is rejected despite the "probability of wrong rejection" α

If the computed p-value is small, we can infer that there is an association between the predictor and the response.

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

3. Assessing the Accuracy of the Model

(a) Residual standard error (RSE)

An estimate of the standard deviation of ϵ → a measure of the l_____ of fit of the model to the data

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

(b) R^2 statistic

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

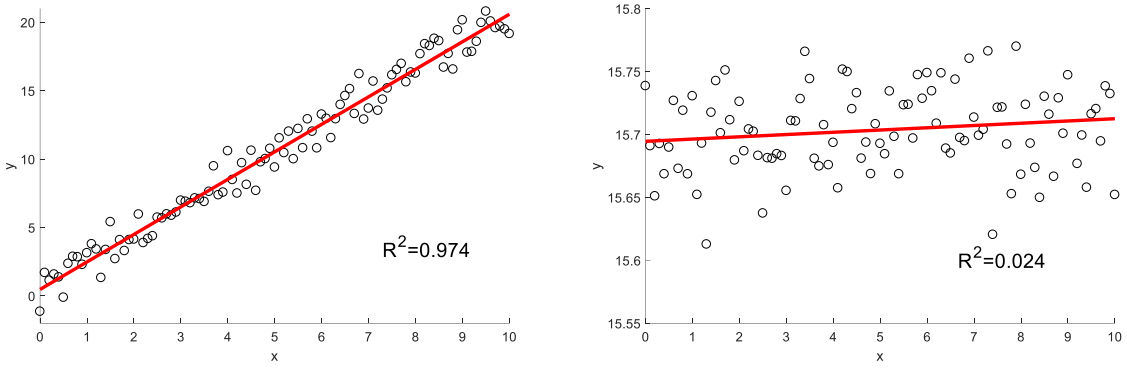
where TSS (Total Sum of Squares) = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

TSS: Amount of variability inherent in the response before the regression is performed
RSS: Amount of variability that is left unexplained after performing the regression
TSS-RSS: Amount of variability in the response that is removed by linear regression

Instructor: Junho Song
junhosong@snu.ac.kr

$R^2 \cong 0$: No reduction of variability by regression (weak linear relationship)

$R^2 \cong 1$: Large reduction of variability by regression (strong linear relationship)



Note: In case of simple linear regression, $R^2 = Cor(X,Y)^2$ (Ref. Sup 01)

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \text{ (sample correlation)}$$

```
summary(lm.fit) #can check RSE and R-squared
(cor(lstat,medv))^2 # square of correlation coefficient
```

Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

M1586.002500 Information Engineering for Civil & Environmental Engineers

In-Class Material: Class 06

Linear Regression (ISL Chapter 3)

Given: Sample data set $\{(x_1, y_1), (x_1, y_1), \dots, (x_n, y_n)\}$

Question: If we have more than one predictor, how can we extend our analysis?

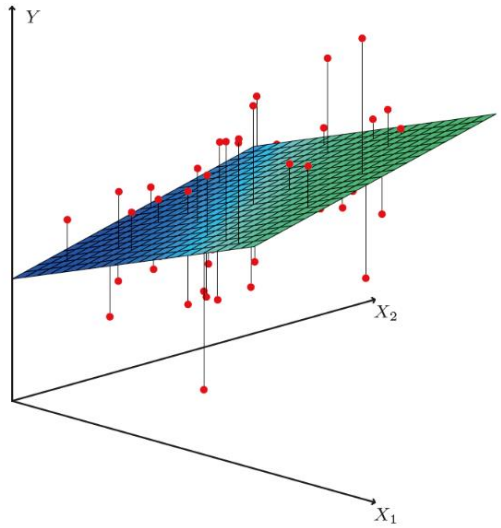
→ "*M*_____ Linear Regression" Analysis, i.e. $Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

1. Multiple Linear Regression

(a) If we have more than one predictor, approach of fitting a separate simple linear regression model for each predictor is **not entirely satisfactory**

- Unclear how to make a prediction when multiple predictors are available
- Ignoring the other (correlated) predictors → misleading estimates (see next page)

Alternative: $Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ where $\{X_1, X_2, \dots, X_p\}$ are predictor variables and Y is a response.



(b) Estimation of model coefficients $\{\beta_0, \beta_1, \dots, \beta_p\}$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ where $\hat{y}$ indicates a prediction of Y on the basis of $X_i = x_i$, $i = 1, \dots, p$

(c) Residual Sum of Squares (RSS): the same approach as the simple linear regression

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

Find $\{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\}$ that minimize RSS → Using matrix algebra (Ref. Sup 01)

Instructor: Junho Song
junhosong@snu.ac.kr

Note: It may look counterintuitive but actually makes sense for the multiple regression to suggest **no relationship** between X_i and Y while the simple linear regression implies the **opposite.**

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

This is due to significant correlation between predictors (risk of simple linear regression)

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

2. [Question 1] Is there a relationship between the response and (at least one of the) predictors?

(a) Verify whether *all* of the regression coefficients are zero?

Null Hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
Alternative Hypothesis H_1: at least one β_j is non-zero

Computing *F*-statistic:

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

where, as with simple linear regression, $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, and $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

If linear model assumptions are correct → $\mathrm{E}[\mathrm{RSS}/(n-p-1)] = \sigma^2$
If H_0 is true → $\mathrm{E}[(\mathrm{TSS} - \mathrm{RSS})/p] = \sigma^2$

Therefore, if H_0 is true, $F \approx 1$, and if H_1 is true, $F \gg 1$.

Reject H_0 in favor of H_1 when $F > F_{\alpha,p,n-p-1}$
Fail to reject H_0 when $F < F_{\alpha,p,n-p-1}$
($F_{\alpha,p,n-p-1}$ can be calculated using the *F* distribution)

Instructor: Junho Song
junhosong@snu.ac.kr

(b) Verify whether *some* of the regression coefficients are zero?

Null Hypothesis H_0: $\beta_{p+q-1} = \beta_{p+q-2} = \cdots = \beta_p = 0$ (q coefficients)

Computing *F*-statistic:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

where, $\mathrm{RSS}_0 = \sum_{i=1}^{n}(y_i - \hat{y}_{i0})^2$ where $\hat{y}_{i0} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_q x_q$

3. [Question 2] Do all the predictors help to explain or only a subset?

→ "Deciding on Important Variables" (extensively studied in Ch. 6)

Variable Selection: Determining which predictors are associated with the response (select one of 2^p models)

(1) Forward selection
Begin with the null model (contains an intercept but no predictors)
Continue the followings until the stopping rule is satisfied:
- Fit p simple linear regressions
- Add to the null model the variable that results in the lowest RSS

(2) Backward selection
Begin with the all variables in the model
Continue the followings until the stopping rule is satisfied:
- Remove the variable with the largest p-value
- Fit the new $(p-1)$-variable model

(3) Mixed selection
Begin with the null model
Continue the followings until the stopping rule is satisfied:
- Fit p simple linear regressions
- Add to the null model the variable that results in the lowest RSS
- If the p-value of a variable rise above a threshold, remove that variable

4. [Question 3] How well does the model fit the data?

(a) Residual standard error

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

Note: the same equation as simple regression ($p = 1$)

Instructor: Junho Song
junhosong@snu.ac.kr

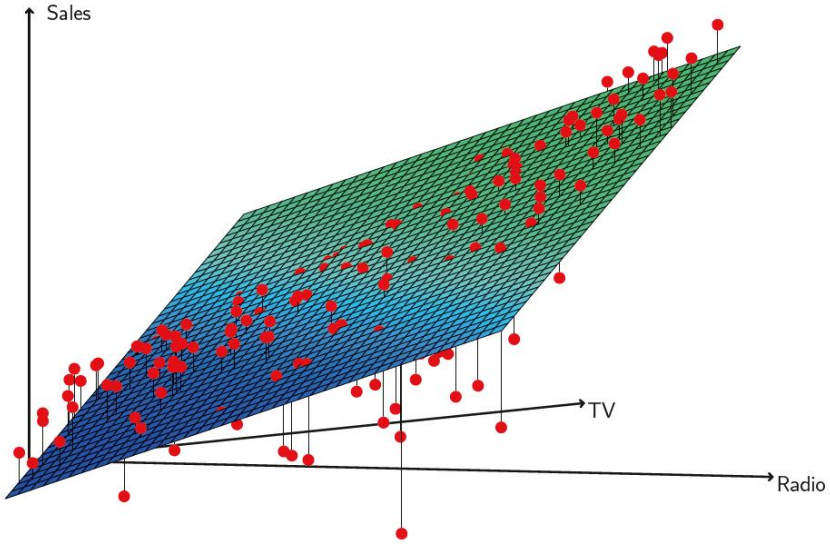
(b) R^2 statistic: the same as simple linear regression

R^2 will **always increase** when more variables are added to the model
→ Tiny increase in R^2 provides an evidence that newly added variable can be dropped from the model

(c) Plot the data

Can reveal problems that are not visible from numerical statistics

- least square regression overestimate when money is spent exclusively on either TV or radio, and underestimate when split
- This nonlinear pattern suggests the existence of *synergy* or *interaction effect* (See Section 3.3.2 of ISL)




```
library(MASS)
install.packages("ISLR")
library(ISLR)

lm.fit1 = lm(medv~lstat+age, data=Boston)
# fit a multiple linear regression with 2 predictors, lstat & age
summary(lm.fit1)

lm.fit2 = lm(medv~., data=Boston)
# fit a multiple linear regression with all predictors
summary(lm.fit2)
summary(lm.fit2)$r.squared # R square
summary(lm.fit2)$sigma # square root of the estimated variance, RSE

# check p-value : excluding predictor which has a high p-value
lm.fit21 = lm(medv~.-age, data=Boston)
# new multiple linear regression excluding age (high p-value)
lm.fit22 = update(lm.fit2, ~.-age)
# update multiple linear regression excluding age
summary(lm.fit21)
summary(lm.fit22) # these two show same results
```

Instructor: Junho Song
junhosong@snu.ac.kr

5. [Question 4] How accurate (or uncertain) are our predictions?

(a) Three types of uncertainties associated with the least-square prediction by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$:

- **Reducible Error**: the inaccuracy in the coefficient estimates $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p\}$
- **Model Bias**: *Potentially* reducible error caused by "linear" model assumption
- **Irreducible Error**: the random error ϵ in $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$

(b) Confidence Interval: quantifies the uncertainty in the estimated conditional mean $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$, caused by the reducible error

(c) Prediction Interval: quantifies the uncertainty in the true value predicted for a particular case, $Y \approx \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p + \epsilon$, caused by both reducible and irreducible error

6. Qualitative Predictors in Linear Regression

(a) What if we wish to include "**categorical**" variables as predictors in the regression model, e.g. gender (female, male), smoking status (smoker, nonsmoker), socioeconomic status (poor, middle, rich)

(b) Strategy: Create indicators or dummy variables that take possible numerical values to represent effects of levels on response

※ If variable has *c* levels, create (*c* – 1) dummy variables, leaving one level as the reference level (see Example 1 and 2 below)

Example 1 (Qualitative predictors with two levels): Investigate differences in credit card balance between males and females in regression model using **Credit** data set

Create dummy variable x_i based on the gender variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Then, the linear regression model equation will be as below

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Now β_0 can be interpreted as the average credit card balance among males, $\beta_0 + \beta_1$ as the average credit card balance among females

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

Instructor: Junho Song
junhosong@snu.ac.kr

The table shows the result of the coefficient estimates and other information associated with the model. The average credit card debt for males is estimated to be $_____, whereas females are estimated to carry $_____ + $_____ = $_____. However, the p-value for the dummy variable is very high, and this indicates that there is no statistical evidence of a difference in average credit card balance between the genders.

The decision to code females as 1 and males as 0 is arbitrary, Instead of a 0/1 coding scheme, what will happen if we create a dummy variable as below? Will it affect the result?

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

Example 2 (Qualitative predictors with more than two levels, continued): With the same data set of Example 1, investigate differences in credit card balance between ethnicity in regression model

There being three levels in ethnicity variable, create two dummy variables x_{i1}, x_{i2}

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

Then, the regression model equation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	–18.69	65.02	–0.287	0.7740
ethnicity[Caucasian]	–12.50	56.68	–0.221	0.8260

The table shows us the estimated balance for the baseline, African American, is $531.00. It is estimated that the Asian category will have $18.69 less debt than the African American category on average, and that the Caucasian category will have $12.50 less debt than the African American category. However, the p-values associated with the coefficient estimates for the two dummy variables are very large, suggesting no statistical evidence of a real difference in credit card balance between the ethnicities.