GBUS 738 Data Mining

Linear Regression

David Svancer – George Mason University School of Business

School of Business

Linear Regression

Linear regression is a supervised learning method for predicting a ***quantitative outcome variable***.

The ***Advertising*** data set contains a company's sales revenue and advertising budgets (in thousands) for 200 markets and serves as the running example in chapter 3 of *An Introduction to Statistical Learning*.

Linear regression can answer the following questions for this data set:

1. *Is there a relationship between advertising budget and sales?*
2. *How strong is the relationship between advertising budget and sales?*
3. *Which advertising types contribute to sales?*
4. *How accurately can we predict future sales?*
5. *Is the relationship linear?*

Sales	TV	Radio	Newspaper
22.1	230.1	37.8	69.2
10.4	44.5	39.3	45.1
9.3	17.2	45.9	69.3
18.5	151.5	41.3	58.5
12.9	180.8	10.8	58.4
…	…	…	…

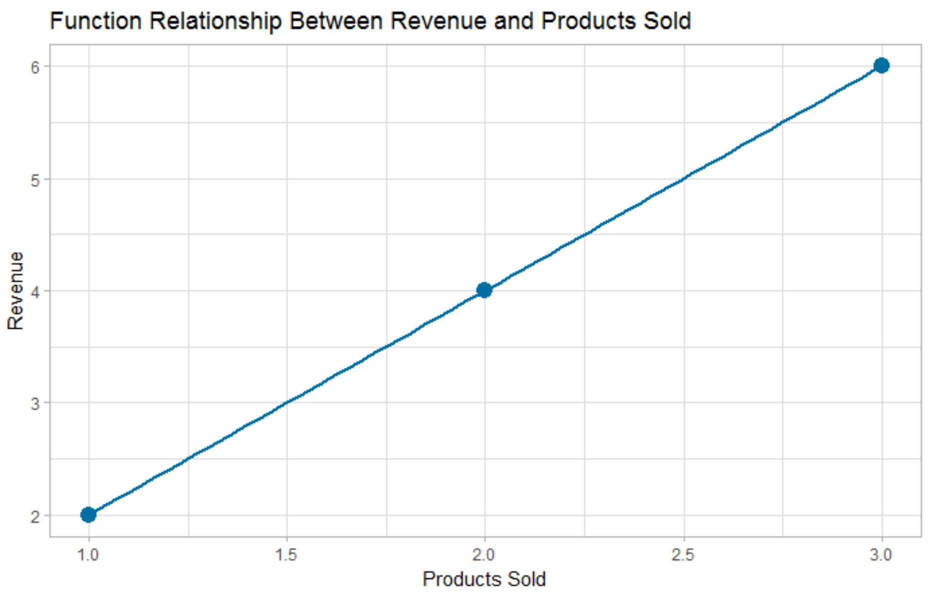
Linear Regression

Functional Relationship Between Variables

Before we get into the details of linear regression, let's review the definition of a functional relationship between variables

- If a company sells a certain product for $2.00, then the number of products sold, and revenue have a **functional** relationship.

Products Sold	Revenue
1	2
2	4
3	6



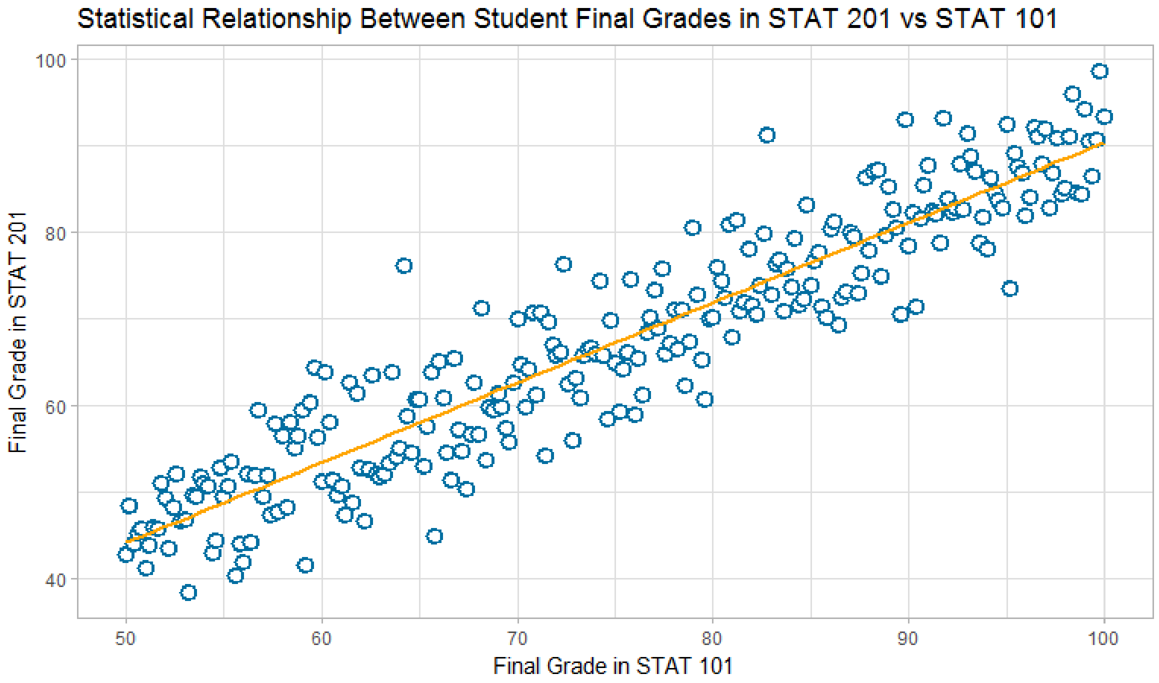
Linear Regression

Statistical Relationship Between Variables

Statistical relationships between variables have two components:

- A **functional** component which represents the expected value (**mean**) of the outcome variable Y given the predictor variable X, usually denoted by $E(Y \mid X = x)$
- A **random** component, which represents random deviations from the functional relationship

The graph on the right displays a **statistical** relationship between a response variable, *Final Grade in STAT 201*, and a predictor, *Final Grade in STAT 101*

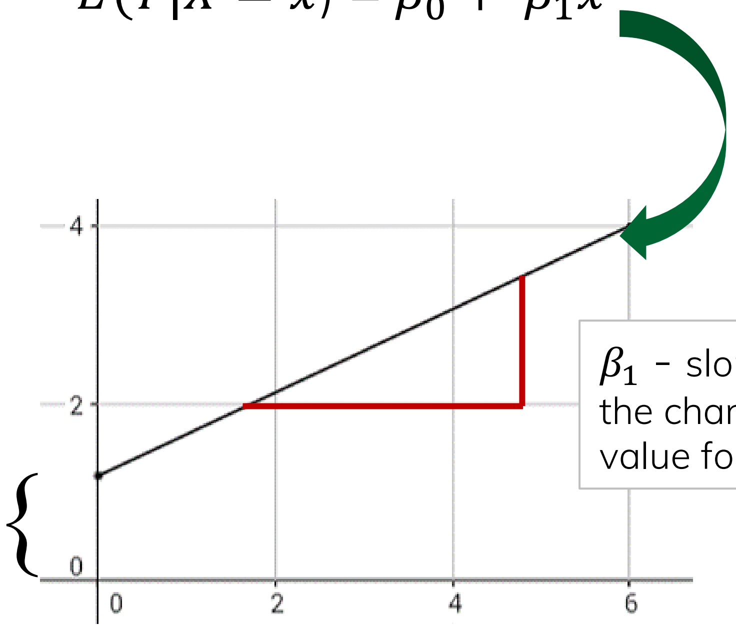


Simple Linear Regression

Functional Component

In **Simple Linear Regression**, we have one predictor variable, and we assume the following functional relationship for the mean of the outcome variable, Y, given a value of the predictor X

$$E(Y|X = x) = \beta_0 + \beta_1 x$$



β_1 - slope of the line represents the change in the mean outcome value for a one unit increase in X

β_0 - Mean outcome value when X = 0

Simple Linear Regression

Adding the Random Component

Simple Linear Regression assumes that each outcome value Y, is a sum of the expected outcome given x (**functional component**) and a random error component, ε

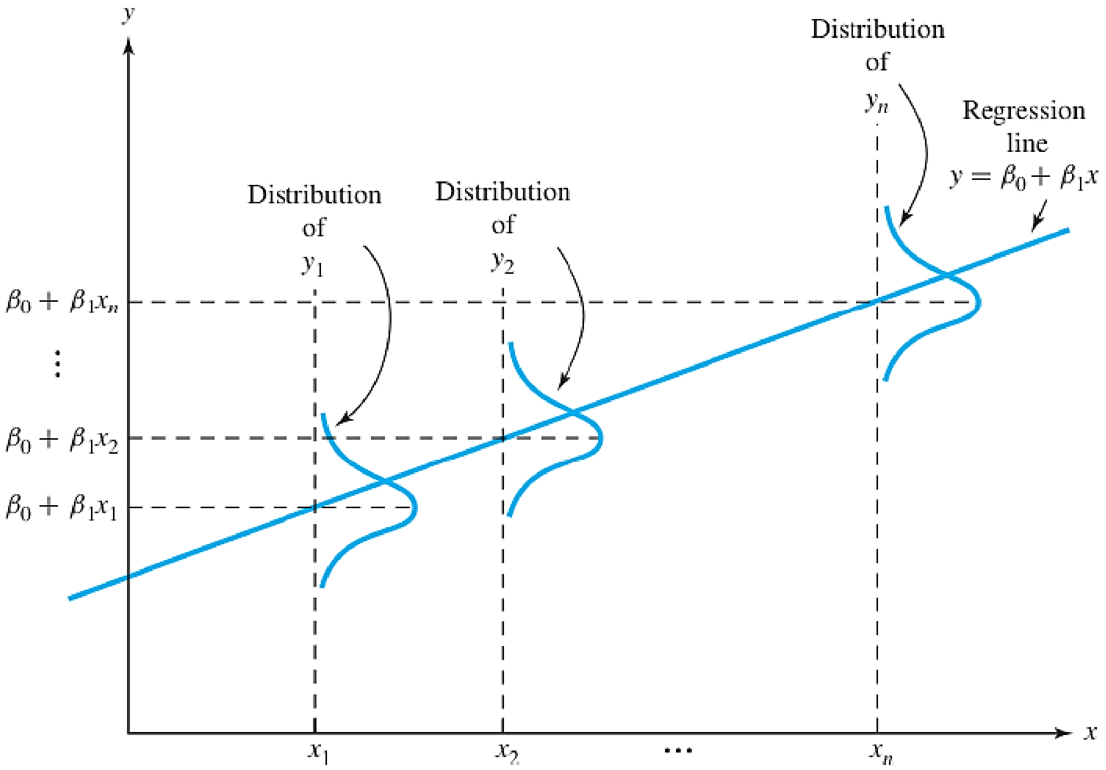


Image source: https://cbmm.mit.edu/sites/default/files/documents/probability_handout.pdf

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Statistical Assumptions

1. $E(Y|\, X = x) = \beta_0 + \beta_1 x$
2. $E(\varepsilon) = 0$
3. $\mathrm{Var}(\varepsilon) = \sigma^2$
4. The error terms are independent
5. Each ε is normally distributed

School of Business

Simple Linear Regression

Estimating the Coefficients

In practice, we do not know the true values of β_0 and β_1

- They must be estimated from our sample data and are denoted as $\widehat{\beta_0}$ and $\widehat{\beta_1}$
- Once we obtain these estimates, we can use our linear model to make predictions
- Predictions are of the form $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$

The most common way to obtain the coefficient estimates is by the **method of least squares**

- We find $\widehat{\beta_0}$ and $\widehat{\beta_1}$ by minimizing the following equation, known as the **Residual Sum of Squares** (RSS)

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2$$

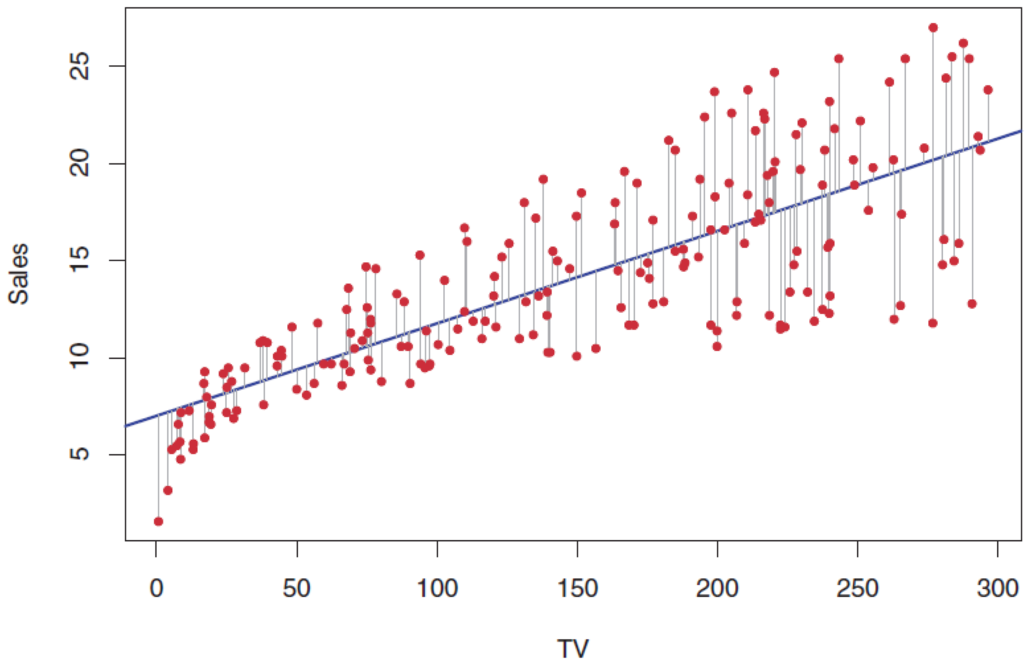


Image source: An Introduction to Statistical Learning with Applications in R
James, Witten, Hastie, Tibshirani

School of Business

Simple Linear Regression

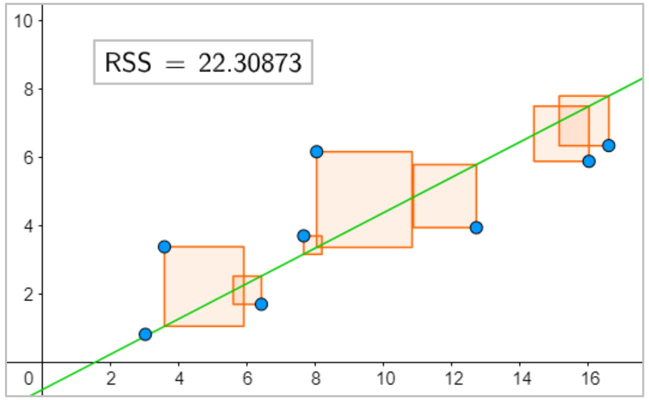
Estimating the Coefficients

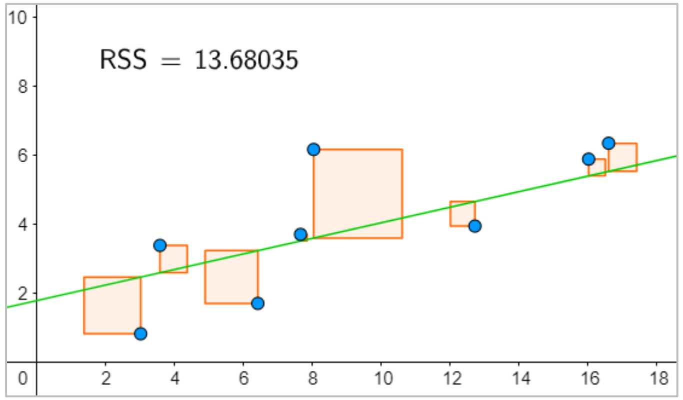
Given a set of n sample data points (x_i, y_i), we can use a system of partial derivatives to determine the values of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ that minimize the RSS

Machine Learning – Gradient Descent Technique

Different estimates of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ produces different values of the RSS (sum of the areas of the orange squares)

Goal: Iterate through data to minimize the area of the orange squares





Simple Linear Regression

Residual Standard Error

Under the assumptions of linear regression, an optimal estimator of σ^2 is

$$\widehat{\sigma^2} = \frac{RSS}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n-2}$$

An estimate of the common standard deviation of the error terms is just $\sqrt{\frac{RSS}{n-2}}$, many textbooks refer to this value as the **RMSE (Root Mean Square Error)**.

In **R**, it is known as the **Residual Standard Error (RSE)**

Roughly speaking, the ***RSE represents the average predictor error of the model***

Sample **R** output with the RSE highlighted
Estimated Regression Line:
Sales = 7.03 + (0.048)TV

```
lm_fit <- lm_model %>% fit(Sales ~ TV, data = advertising)
summary(lm_fit$fit)
```

```
Call:
lm(formula = Sales ~ TV, data = advertising)

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
```

School of Business

Simple Linear Regression

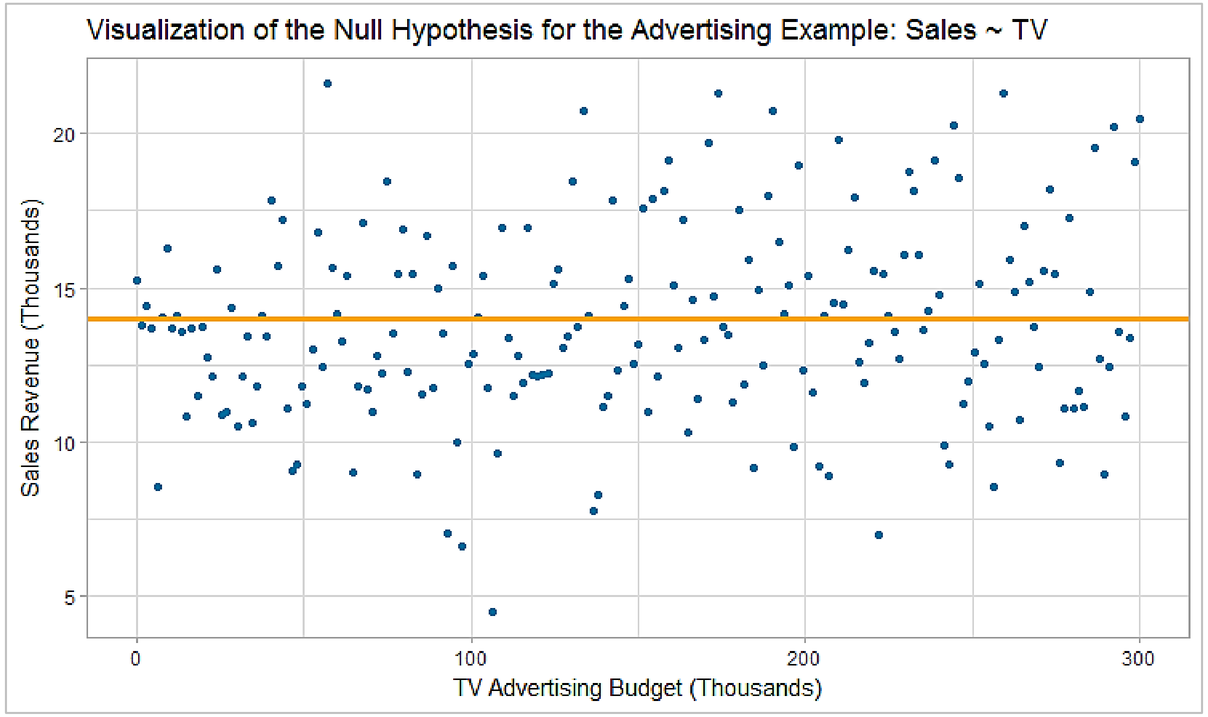
Hypothesis Testing for Model Parameters

In simple linear regression, we are usually interested in knowing the value of the population parameter, β_1. Specifically, whether this value is equal to 0.

If $\beta_1 = 0$, then the true relationship between the outcome and predictor variable is

$$Y = \beta_0 + \varepsilon$$

In other words, there is **no relationship** between the outcome variable, Y, and the predictor variable X.



Simple Linear Regression

Hypothesis Testing for Model Parameters

The hypothesis test of interest in simple linear regression is:

$$H_o\colon \beta_1 = 0 \quad vs \quad H_a\colon \beta_1 \neq 0$$

If H_o **is true**, then the following test statistic follows a **t distribution with *n* – 2** degrees of freedom

$$t = \frac{\widehat{\beta_1}}{\widehat{S.E}\,(\widehat{\beta_1})} = \frac{Estimated\ \beta_1}{Estimated\ Standard\ Error\ of\ \beta_1}$$

In the **R** output on the right, our observed **t statistic is 17.67 with 198 degrees of freedom**.

Interpretation of the p-value in the output: If H_o **is true**, then under random sampling, the probability of observing a t statistic that is either 17.67 or greater or -17.67 or less is **extremely small (<< 0.0001)**

Two things are possible:

1. You just witnessed an ***extremely rare outcome***
2. What you assumed to be true, $\beta_1 = 0$ in this case, is wrong

Sample **R** output with the RSE highlighted
Estimated Regression Line:
Sales = 7.03 + (0.048)TV

```
lm_fit <- lm_model %>% fit(Sales ~ TV, data = advertising)
summary(lm_fit$fit)

Call:
lm(formula = Sales ~ TV, data = advertising)

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
```

GEORGE MASON UNIVERSITY
School of Business

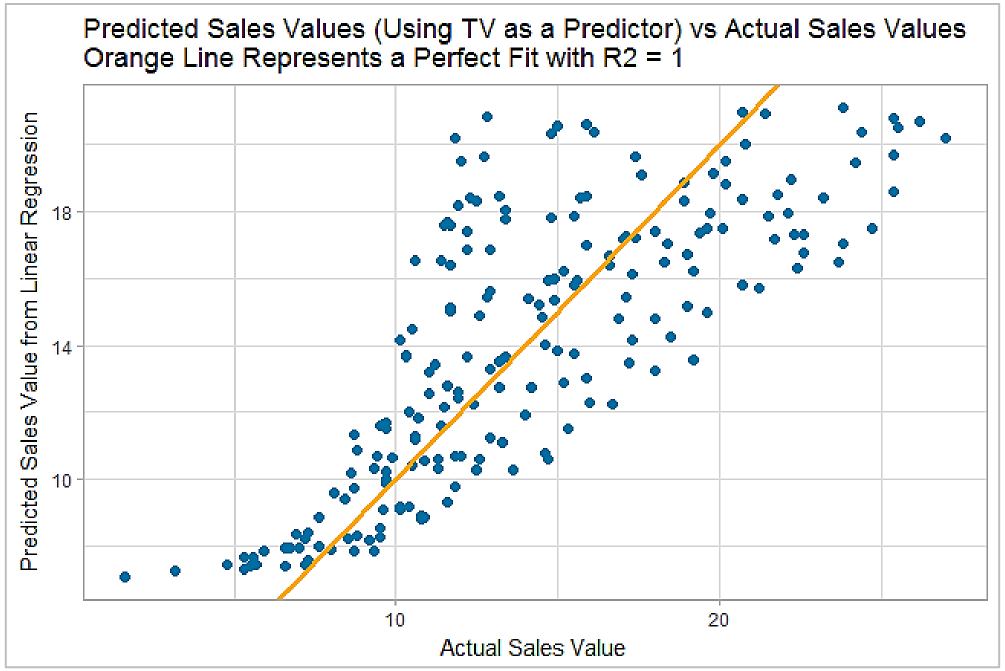
Simple Linear Regression

Assessing the Predictive Power of a Regression Model: R^2

$\boldsymbol{R^2}$ represents the proportion of variability in the outcome values **that is explained by the predictor values**

- Ranges from 0 (worst) to 1 (best)

An equivalent interpretation of R^2 is the ***squared correlation between the observed and predicted values*** in a linear regression model



GEORGE MASON UNIVERSITY
School of Business

Simple Linear Regression

Diagnostic Plots to Access the Assumptions of Linear Regression

Residuals vs Fitted

- Model residuals ($y_i - \hat{y}_i$) vs predicted values $\hat{y}_i$
- You should see: Random scatter about 0 with no trends or patterns
- Assumptions checked include:
  - Expected Value of Y is a linear function of X
  - Common Variance
  - Independent Errors

Normal Q-Q

- You should see: most points fall on the line
- Assumptions checked include:
  - Errors are Normally Distributed

Scale-Location

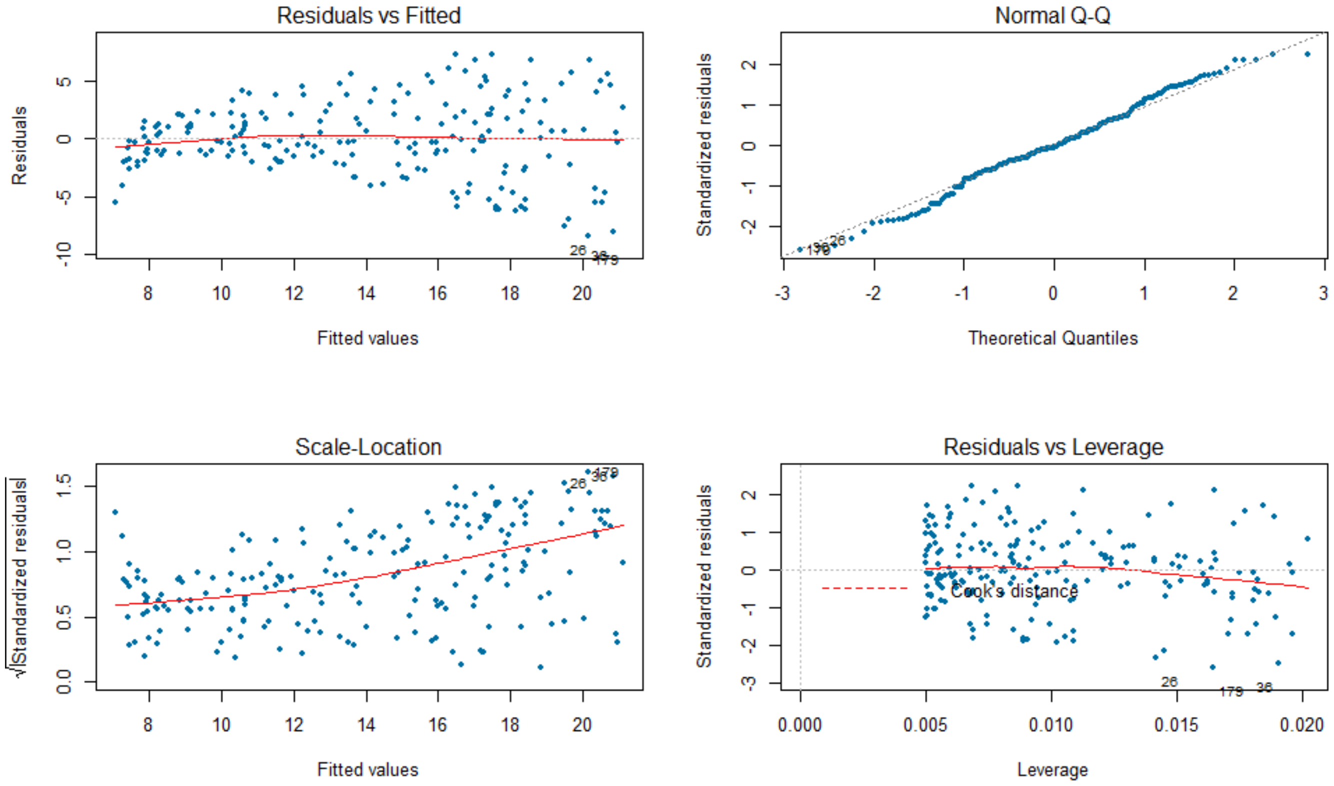
You should see: a **flat** red line

Assumptions checked include:

- Common Variance

Residuals vs Leverage

- Identifies potential outliers and high influence points



Multiple Linear Regression

Extending Linear Regression to Incorporate Multiple Predictors

The simple linear regression model can be extended to allow for multiple predictor variables. Just like in simple linear regression, we are still predicting the value of a numeric outcome variable Y

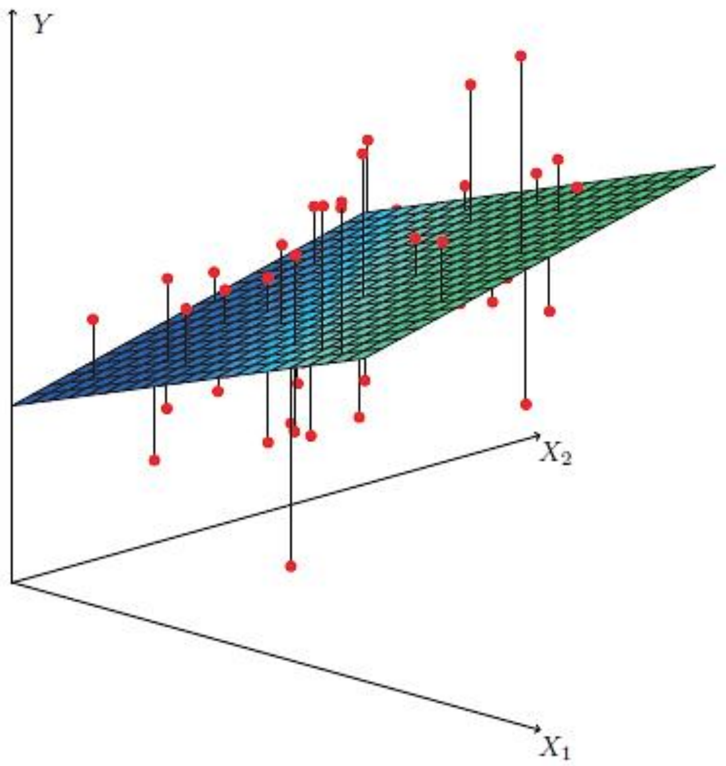
However, we now have **p** predictor variables, where **p** ≥ 2

In this setting we are assuming the following:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

The mean of Y given the predictors is now modeled as a **plane** in multi-dimensional space. The error terms capture the random deviations from this population mean function.

The assumptions from before remain the same



School of Business

Image source: An Introduction to Statistical Learning with Applications in R
James, Witten, Hastie, Tibshirani

Multiple Linear Regression

Estimating the Coefficients and Making Predictions

Just as in Simple Linear Regression, we do not know the true values of $\beta_0,\ \beta_1, \ldots, \beta_p$

- They must be estimated from our sample data and are denoted as $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p}$
- Once we obtain these estimates, we can make predictions of the form $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i,1} + \widehat{\beta_2} x_{i,2} + \ldots + \widehat{\beta_p} x_{i,p}$

We find $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p}$ by minimizing the **Residual Sum of Squares (RSS)**

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i,1} - \widehat{\beta_2} x_{i,2} - \ldots - \widehat{\beta_p} x_{i,p})^2$$

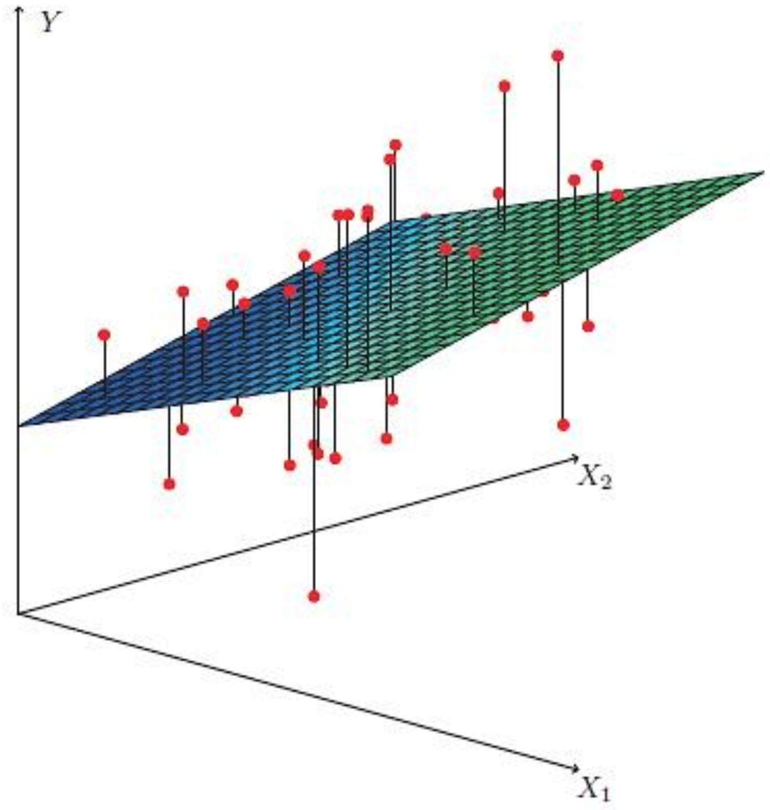


Image source: An Introduction to Statistical Learning with Applications in R
James, Witten, Hastie, Tibshirani

GEORGE MASON UNIVERSITY
School of Business

Multiple Linear Regression

Estimating the Coefficients and Making Predictions for Advertising Data

R will estimate the coefficients for us with the **fit()** function.

On the right, we are estimating the following multiple regression model that predicts mean *Sales* using *TV*, *Radio*, and *Newspaper* budgets:

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

Sales	TV	Radio	Newspaper
22.1	230.1	37.8	69.2
10.4	44.5	39.3	45.1
9.3	17.2	45.9	69.3
18.5	151.5	41.3	58.5
12.9	180.8	10.8	58.4

Our **estimate** of this plane is:

$$\widehat{Sales} = 2.94 + 0.046TV + 0.189Radio - 0.001Newspaper$$

The estimated Sales for the **first** row of the data would be:

$$\hat{y}_i = 2.94 + 0.046(230.1) + 0.189(37.8) - 0.001(69.2) = 20.5$$

Parameter Estimates

```
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
```

GEORGE MASON UNIVERSITY School of Business

Multiple Linear Regression

The F Test

Is at least one predictor variable associated with the outcome variable?

This corresponds to the following hypothesis test for the **Advertising** model:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_{a:} \text{ At least one } \beta_j \text{ is non} - \text{zero}$$

This test is conducted automatically in **R**, and the resulting F statistic and associated *p-value* are displayed at the bottom of the model summary

Our F statistic is **570.3** with a small *p-value* and provides strong evidence ***against*** H_0. We have evidence that ***at least one*** predictor is associated with *Sales*.

```
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
```

Multiple Linear Regression

Partial F-Test

The *t* values and *p*-values displayed by R's **summary()** function have a special interpretation in multiple regression

- They each represent something known as a **partial F-Test**
- **The intuition behind this result is:** *given that I am using TV* and *Radio* as predictors, does *Newspaper* provide increased accuracy to my model? **Answer: No**

```
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
```

GEORGE MASON UNIVERSITY
School of Business

Multiple Linear Regression

Adding Categorical Predictors – Dummy Variable Encoding

Selling Price	Square Footage	City
320,000	1,760	Seattle
410,000	2,100	Auburn
275,000	1,550	Seattle
520,550	2,450	Bellevue
375,000	1,850	Auburn

→

Selling Price	Square Footage	City_Bellevue	City_Seattle
320,000	1,760	0	1
410,000	2,100	0	0
275,000	1,550	0	1
520,550	2,450	1	0
375,000	1,850	0	0

$$Selling\ Price = \beta_0 + \beta_1 Square\ Footage + \beta_2 City\_Bellevue + \beta_3 City\_Seattle$$

Multiple Linear Regression

Accessing Model Fit and Accuracy of Predictions

We can access the model fit in multiple linear regression using the same diagnostic plots and statistics as in Simple Linear Regression:

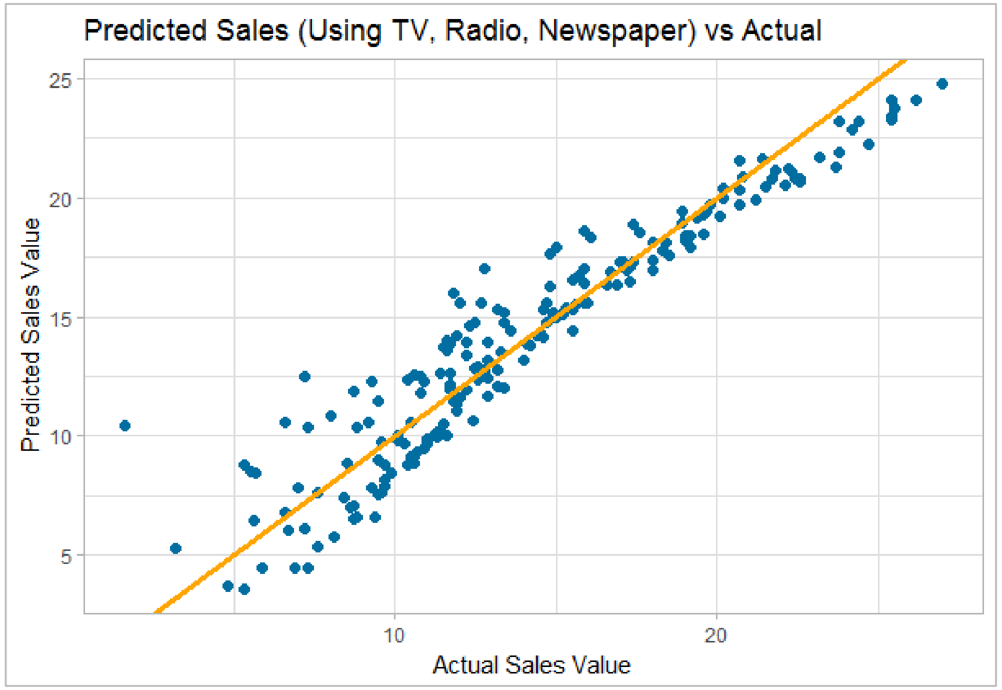
- *RSE , R^2 , Residual Plots*
- *Q-Q Plots, Visualization of R^2 (Predicted vs Actual)*

```
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
```



GEORGE MASON UNIVERSITY
School of Business