Linear Regression

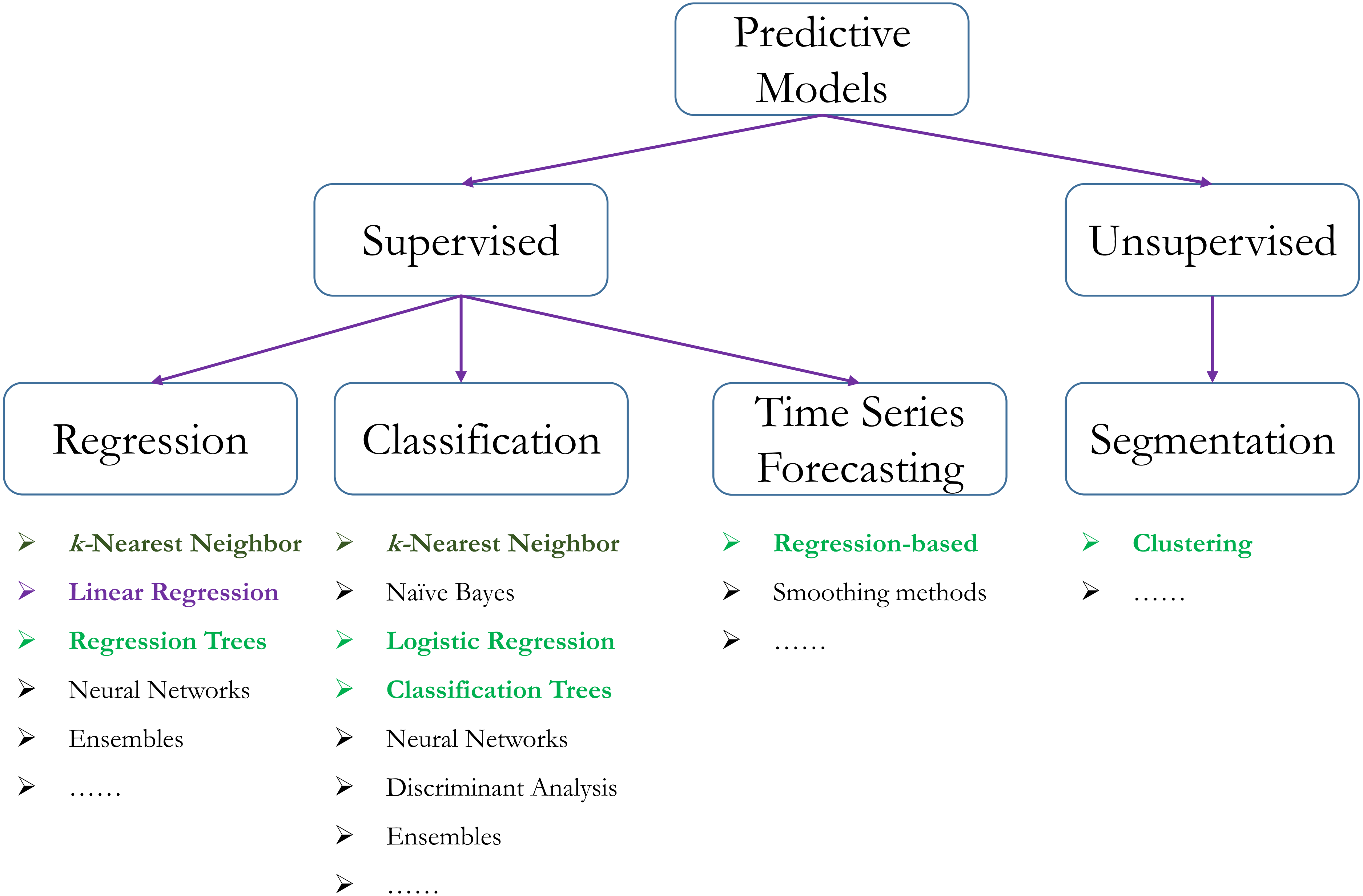
Previous class

- Midterm

Total	Min	P25	Median	Average	P75	Max
/50	19.5	29.0	33.0	33.4	37.8	44.5
/100	39.0	58.0	66.0	66.9	75.5	89.0

Today's class

- Linear Regression
- Application in R/RStudio and Inference


Predictive Models
Supervised
Unsupervised
Regression
Classification
Time Series Forecasting
Segmentation
k-Nearest Neighbor
Linear Regression
Regression Trees
Neural Networks
Ensembles
……
k-Nearest Neighbor
Naïve Bayes
Logistic Regression
Classification Trees
Neural Networks
Discriminant Analysis
Ensembles
……
Regression-based
Smoothing methods
……
Clustering
……

Linear Regression

- Rudimentary model in Supervised Learning
- Predicting a numeric variable
- Many advanced models are extensions of linear regression
- Two forms
    - Simple Linear Regression
    - Multiple Linear Regression

Regression

- Goal: Fit a relationship between
  - numeric output variable Y & set of "p" input variables $X_1, X_2, X_3, \cdots \cdots X_p$
- Output variable Y is also referred as
  - Response / Target / Outcome variable
- Input variables $X_1, X_2, X_3, \cdots \cdots X_p$ are also referred as
  - Predictors / Independent variables / Regressors / Covariates

Linear Regression

- Predict "Y" using a linear combination of predictors $X_1, X_2, X_3, \cdots\cdots X_p$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Noise or Unexplained part

- Information available on both $X's$ & Y
- $\beta_0, \beta_1, \beta_2 \cdots\cdots \beta_p$ are coefficients
- Required to estimate the coefficients
- Underlying estimation process : **Ordinary Least Squares (OLS)**

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

- Estimated values are generally represented by hat $\widehat{\ }$

Types

- Simple Linear Regression (p = 1)

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

- Multiple Linear Regression (p > 1)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Regression modeling includes **estimating coefficients**, and **choosing which predictors ($X's$) to include and in what form**
- E.g., A transformed numerical predictor can be included (E.g., $\log X_1$) in the regression
- Right form depends on domain knowledge, data, required predictive power etc.
- Numerous applications

Example : Real Estate market

- Response : house list price (Y)

- Predictors (X)

  - Square Foot (sqft)
  - Year Built
  - Beds, Bath, Lot Size, Parking Spots
  - Garage (0/1)
  - Zip
  - Crime rate
  - Income
  - Public School Rating
  - ……

$$\text{list price} \approx \beta_0 + \beta_1 \text{ sqft} + \beta_2 \text{ age} + \cdots + \beta_p \text{ school rating}$$

$$\text{list price} \approx 200000 + 34 \text{ sqft} - 27 \text{ age} + \cdots + 72 \text{ school rating}$$

Example : New route Air fare

- Response : fare (Y)
- Predictors (X)
  - Start and End City
  - New Air carriers entering the route
  - Market concentration
  - Start and End City Average Income
  - Start and End City Average Population
  - Distance
  - Vacation route (1/0)
  - ....

$$\text{fare} \approx \beta_0 + \beta_1 \text{ start city} + \beta_2 \text{ end city} + \cdots + \beta_p \text{ distance}$$

$$\text{fare} \approx 200 + 35 \text{ start city} + 25 \text{ end city} + \cdots + 100 \text{ distance}$$

Example : Toyota corolla used car sales

- Response : sale price (Y)

- Predictors (X)
  - Age in months
  - Accumulated km on odometer
  - Fuel type (Petrol, Diesel, CNG)
  - Horsepower
  - Metallic color? (Yes = 1, No = 0)
  - Automatic (Yes = 1, No = 0)
  - Cylinder volume
  - Number of doors
  - ………

$$\text{sale price} = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ km} + \cdots + \beta_p \text{ doors} + \epsilon$$

$$\text{sale price} = 15000 - 34 \text{ age} - 25 \text{ km} + \cdots + 2 \text{ doors} + \epsilon$$

More examples

- Credit card customer activity based on demographics, historical activity
- Vacation expenditures based on frequent flyer data
- Staffing requirements at help desk based on historical data, product and sales information
- Sales in brick & mortar retail store based on labor, traffic, discounts etc.
- Box office revenue of bond movies based on rating and violence

Method

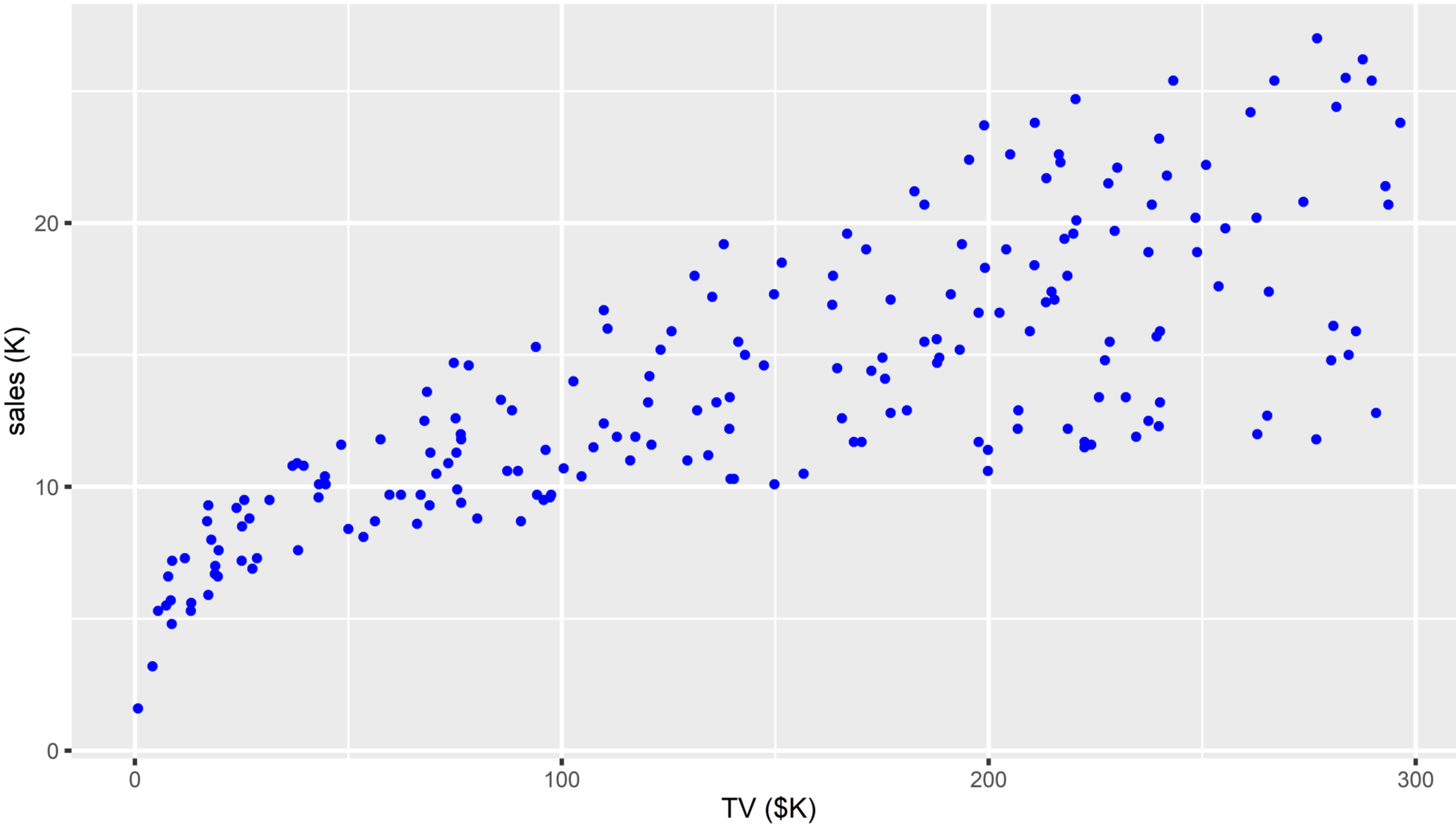
- **Ordinary Least Squares (OLS)**
  - Minimize the sum of squared deviations between outcome (Y) & predicted values ($\hat{Y}$)
- Example
  - Sales of a product in 200 markets
  - sales (K) , tv, radio, newspaper ($K)
  - Response (Y) : sales
  - Predictors (X) : tv, radio, newspaper

Simple Linear Regression

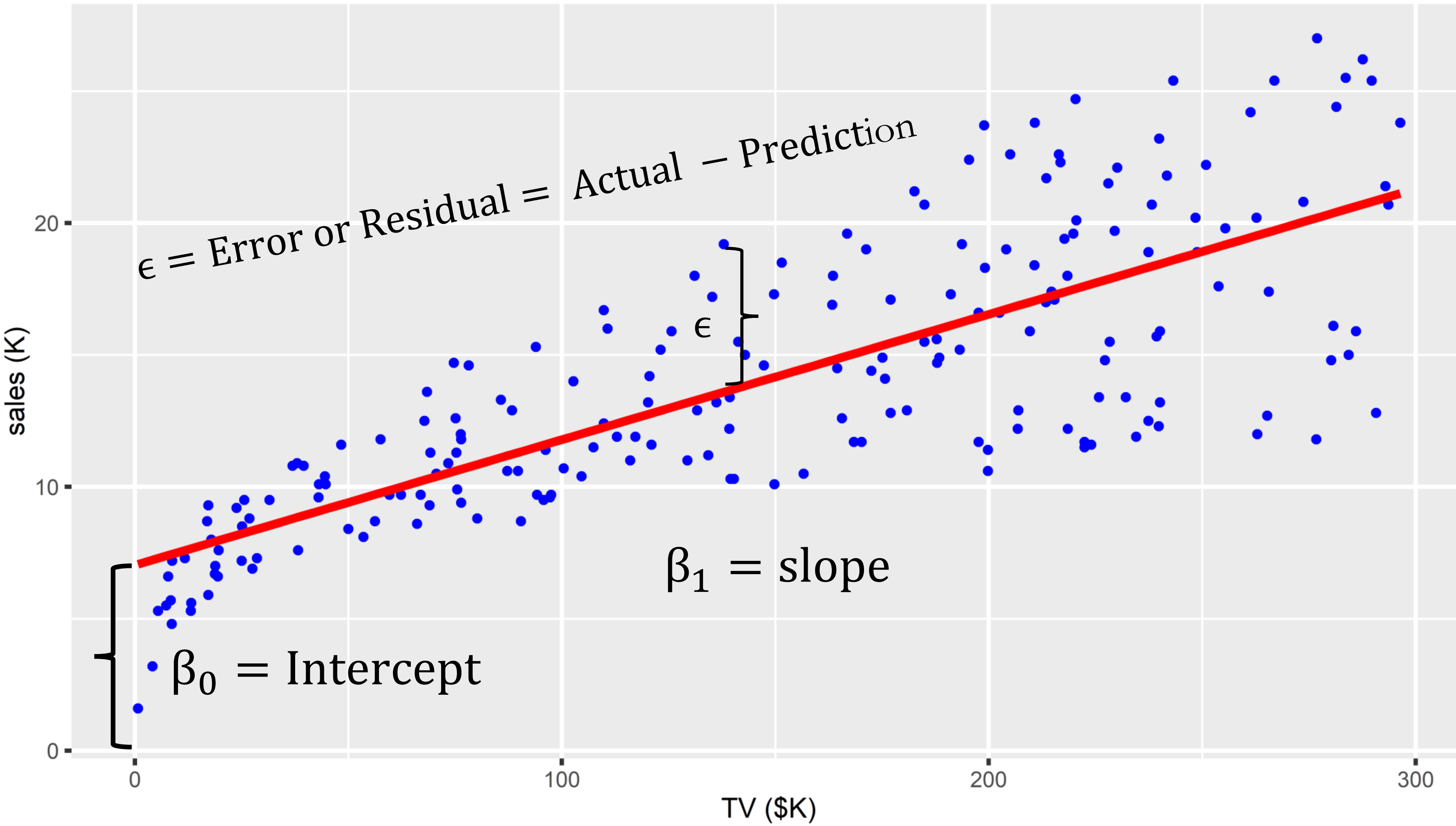
- One predictor
- Sales of a product in 200 markets vs tv expenses ($K)

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad \Rightarrow \quad sales = \beta_0 + \beta_1\, tv + \epsilon$$

Sales vs tv scatter plot



Sales vs tv scatter plot



© Pradeep Pendem

Today's class mandatory steps

- Create a folder name "**g.linear_regression**" within the folder

  "**oba_455_555_ddpm_r/rproject**"
- Download "**linear_regression_code.R**", and all **csv** files from canvas
- Place all downloaded files in

  "**oba_455_555_ddpm_r/rproject/ g.linear_regression**"
- Open RStudio project
- Open "**linear_regression_code.R**" file within RStudio

Is Regression as a whole significant ?

```
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
```

If p-value < 0.05, then at minimum one of the predictor impacts sales

Intercept & slope coefficients

β_0

β_1

```
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
```

Effect of predictors are **insignificant** if you see “.” or no stars

Predictors explanatory power

```
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
```

Multiple R-Square (R^2)
Proportion of variation in sales explained by tv

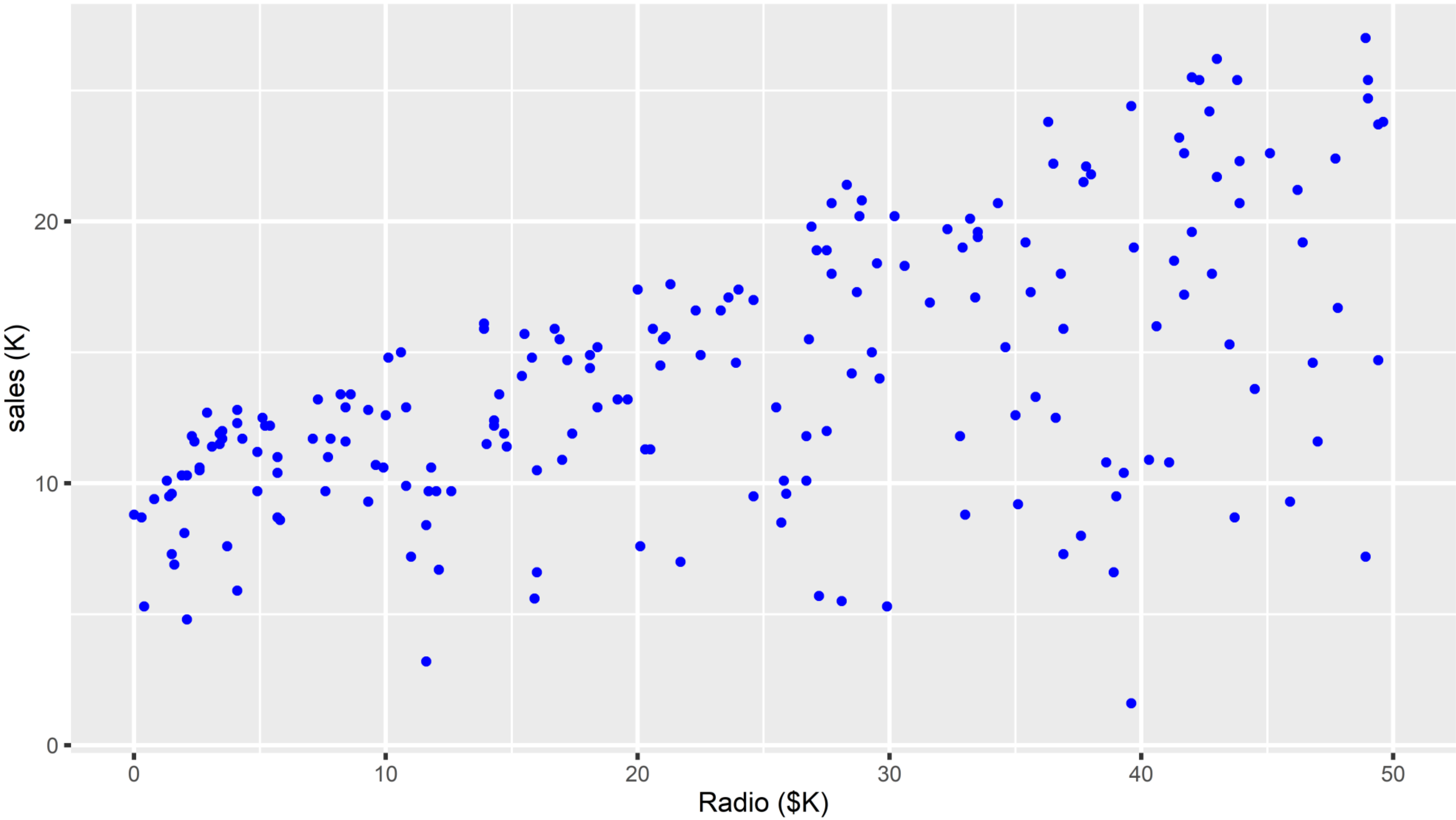
Multiple Linear Regression

- Response : Sales of a product in 200 markets
- Predictors : tv, radio, newspaper (\$K)

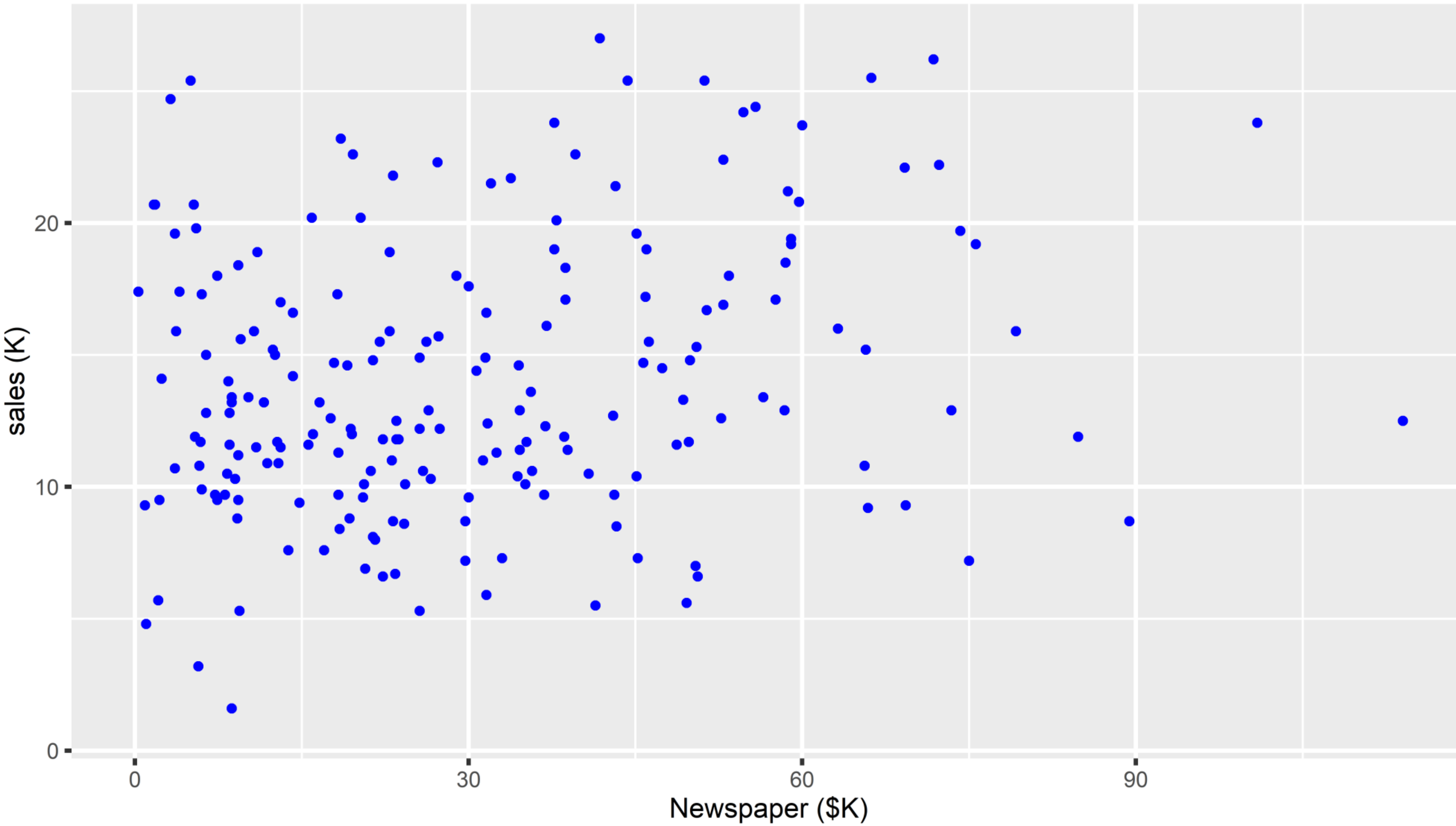
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

$$\text{sales} = \beta_0 + \beta_1\,\text{tv} + \beta_2\,\text{radio} + \beta_3\,\text{newspaper} + \epsilon$$

Sales vs radio scatter plot



Sales vs news paper scatter plot



Multiple linear regression results

```
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
```

Implementation : Toyota corolla used car sales

- Response (Y) - price : Offer price in euros

- Predictors (X)
  - age_08_04 : Age in months as of August 2004
  - km : Accumulated kilometers on odometer
  - fuel_type : Fuel type (Petrol, Diesel, CNG)
  - hp : Horsepower
  - met_color : Metallic color ? (Yes = 1, No = 0)
  - automatic : Automatic (Yes = 1, No = 0)
  - cc : Cylinder volume in cubic centimeters
  - doors : Number of doors
  - quarterly_tax : Quarterly road tax in Euros
  - weight : Weight in Kilograms
- We will use the above selected predictors
- How does the linear regression model looks like?

Multiple Linear Regression model

$$
\begin{aligned}
&\text{price} \\
&= \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ km} \\
&+ \beta_3 \text{ fuel\_type} + \beta_4 \text{ hp} \\
&+ \beta_5 \text{ metcolor} + \beta_6 \text{ automatic} \\
&+ \beta_7 \text{ cc} + \beta_8 \text{ doors} \\
&+ \beta_9 \text{ quarterly tax} + \beta_{10} \text{ weight} \\
&+ \epsilon
\end{aligned}
$$

Is Regression as a whole significant ?

```
Call:
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = toyota)

Residuals:
     Min       1Q   Median       3Q      Max
-11444.0   -755.5    -32.7    755.8   6757.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.326e+03  1.232e+03  -5.948 3.41e-09 ***
age             -1.231e+02  2.596e+00 -47.421  < 2e-16 ***
km              -1.689e-02  1.309e-03 -12.901  < 2e-16 ***
fuel_typeDiesel  6.280e+02  3.758e+02   1.671   0.0949 .
fuel_typePetrol  2.420e+03  3.683e+02   6.571 6.98e-11 ***
hp               2.385e+01  3.466e+00   6.881 8.85e-12 ***
met_color        3.629e+01  7.497e+01   0.484   0.6284
automatic        2.588e+02  1.578e+02   1.640   0.1011
cc              -6.271e-02  9.067e-02  -0.692   0.4893
doors           -7.161e+01  3.966e+01  -1.806   0.0712 .
quarterly_tax    1.231e+01  1.650e+00   7.463 1.46e-13 ***
weight           1.936e+01  1.218e+00  15.894  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1317 on 1424 degrees of freedom
Multiple R-squared:  0.8691, Adjusted R-squared:  0.8681
F-statistic: 859.6 on 11 and 1424 DF,  p-value: < 2.2e-16
```

If p-value < 0.05, then at least one of the predictors impacts price

Significance of individual predictors

```
Call:
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = toyota)

Residuals:
     Min       1Q   Median       3Q      Max
-11444.0   -755.5    -32.7    755.8   6757.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.326e+03  1.232e+03  -5.948 3.41e-09 ***
age             -1.231e+02  2.596e+00 -47.421  < 2e-16 ***
km              -1.689e-02  1.309e-03 -12.901  < 2e-16 ***
fuel_typeDiesel  6.280e+02  3.758e+02   1.671   0.0949 .
fuel_typePetrol  2.420e+03  3.683e+02   6.571 6.98e-11 ***
hp               2.385e+01  3.466e+00   6.881 8.85e-12 ***
met_color        3.629e+01  7.497e+01   0.484   0.6284
automatic        2.588e+02  1.578e+02   1.640   0.1011
cc              -6.271e-02  9.067e-02  -0.692   0.4893
doors           -7.161e+01  3.966e+01  -1.806   0.0712 .
quarterly_tax    1.231e+01  1.650e+00   7.463 1.46e-13 ***
weight           1.936e+01  1.218e+00  15.894  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1317 on 1424 degrees of freedom
Multiple R-squared:  0.8691, Adjusted R-squared:  0.8681
F-statistic: 859.6 on 11 and 1424 DF,  p-value: < 2.2e-16
```

Effect of predictors are **insignificant** if you see "." or no stars

Impact of individual predictors

```
Call:
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = toyota)

Residuals:
     Min       1Q   Median       3Q      Max
-11444.0   -755.5    -32.7    755.8   6757.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.326e+03  1.232e+03  -5.948 3.41e-09 ***
age             -1.231e+02  2.596e+00 -47.421  < 2e-16 ***
km              -1.689e-02  1.309e-03 -12.901  < 2e-16 ***
fuel_typeDiesel  6.280e+02  3.758e+02   1.671   0.0949 .
fuel_typePetrol  2.420e+03  3.683e+02   6.571 6.98e-11 ***
hp               2.385e+01  3.466e+00   6.881 8.85e-12 ***
met_color        3.629e+01  7.497e+01   0.484   0.6284
automatic        2.588e+02  1.578e+02   1.640   0.1011
cc              -6.271e-02  9.067e-02  -0.692   0.4893
doors           -7.161e+01  3.966e+01  -1.806   0.0712 .
quarterly_tax    1.231e+01  1.650e+00   7.463 1.46e-13 ***
weight           1.936e+01  1.218e+00  15.894  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1317 on 1424 degrees of freedom
Multiple R-squared:  0.8691, Adjusted R-squared:  0.8681
F-statistic: 859.6 on 11 and 1424 DF,  p-value: < 2.2e-16
```

Coefficients (All β^s)

Interpreting numeric predictor

```
Call:
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = toyota)

Residuals:
     Min       1Q   Median       3Q      Max
-11444.0   -755.5    -32.7    755.8   6757.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.326e+03  1.232e+03  -5.948 3.41e-09 ***
age             -1.231e+02  2.596e+00 -47.421  < 2e-16 ***
km              -1.689e-02  1.309e-03 -12.901  < 2e-16 ***
fuel_typeDiesel  6.280e+02  3.758e+02   1.671   0.0949 .
fuel_typePetrol  2.420e+03  3.683e+02   6.571 6.98e-11 ***
hp               2.385e+01  3.466e+00   6.881 8.85e-12 ***
met_color        3.629e+01  7.497e+01   0.484   0.6284
automatic        2.588e+02  1.578e+02   1.640   0.1011
cc              -6.271e-02  9.067e-02  -0.692   0.4893
doors           -7.161e+01  3.966e+01  -1.806   0.0712 .
quarterly_tax    1.231e+01  1.650e+00   7.463 1.46e-13 ***
weight           1.936e+01  1.218e+00  15.894  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1317 on 1424 degrees of freedom
Multiple R-squared:  0.8691, Adjusted R-squared:  0.8681
F-statistic: 859.6 on 11 and 1424 DF,  p-value: < 2.2e-16
```

Interpreting character predictor

```
Call:
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = toyota)

Residuals:
     Min       1Q   Median       3Q      Max
-11444.0   -755.5    -32.7    755.8   6757.8

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.326e+03  1.232e+03  -5.948 3.41e-09 ***
age             -1.231e+02  2.596e+00 -47.421  < 2e-16 ***
km              -1.689e-02  1.309e-03 -12.901  < 2e-16 ***
fuel_typeDiesel  6.280e+02  3.758e+02   1.671   0.0949 .
fuel_typePetrol  2.420e+03  3.683e+02   6.571 6.98e-11 ***
hp               2.385e+01  3.466e+00   6.881 8.85e-12 ***
met_color        3.629e+01  7.497e+01   0.484   0.6284
automatic        2.588e+02  1.578e+02   1.640   0.1011
cc              -6.271e-02  9.067e-02  -0.692   0.4893
doors           -7.161e+01  3.966e+01  -1.806   0.0712 .
quarterly_tax    1.231e+01  1.650e+00   7.463 1.46e-13 ***
weight           1.936e+01  1.218e+00  15.894  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1317 on 1424 degrees of freedom
Multiple R-squared:  0.8691,	Adjusted R-squared:  0.8681
F-statistic: 859.6 on 11 and 1424 DF,  p-value: < 2.2e-16
```

What do we see two (of three) levels of **fuel_type** variable?

What is the reference category in the **fuel_type** variable?

Model fit

```
Call:
lm(formula = price_actual ~ age + km + fuel_type + hp + met_color +
    automatic + cc + doors + quarterly_tax + weight, data = toyota)

Residuals:
     Min       1Q   Median       3Q      Max
-11444.0   -755.5    -32.7    755.8   6757.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.326e+03  1.232e+03  -5.948 3.41e-09 ***
age             -1.231e+02  2.596e+00 -47.421  < 2e-16 ***
km              -1.689e-02  1.309e-03 -12.901  < 2e-16 ***
fuel_typeDiesel  6.280e+02  3.758e+02   1.671   0.0949 .
fuel_typePetrol  2.420e+03  3.683e+02   6.571 6.98e-11 ***
hp               2.385e+01  3.466e+00   6.881 8.85e-12 ***
met_color        3.629e+01  7.497e+01   0.484   0.6284
automatic        2.588e+02  1.578e+02   1.640   0.1011
cc              -6.271e-02  9.067e-02  -0.692   0.4893
doors           -7.161e+01  3.966e+01  -1.806   0.0712 .
quarterly_tax    1.231e+01  1.650e+00   7.463 1.46e-13 ***
weight           1.936e+01  1.218e+00  15.894  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1317 on 1424 degrees of freedom
Multiple R-squared:  0.8691  Adjusted R-squared:  0.8681
F-statistic: 859.6 on 11 and 1424 DF,  p-value: < 2.2e-16
```

Multiple R-Square (R^2)

Proportion of variation in price explained by predictors in the model

Model Results and Prediction

- Regression model has been run on the entire data of 1436 observations

- Prediction for the 3 new observations is as follows

```
> predict(toyota.mlr, newobs)
       1        2        3
9439.020 8570.888 9242.667
```

Predictor selection in Linear Regression

- 38 variables in the toyota data
- Numerous variables in the real-world data
- Kitchen-Sink approach
  - Include all the numerous variables in the model
- Problems with Kitchen-Sink approach
  - Expensive and Time consuming
  - Unstable
  - Including uncorrelated predictors (insignificant) can increase the variance of predictions
  - Dropping correlated predictors (significant) can increase the average bias of predictions

How to reduce number of predictors ?

- Domain knowledge
  - Experienced individuals in the industry sometimes can provide a more valuable information than what the can demonstrate
- Computational power
  - Exhaustive search
  - Subset selection algorithms

Exhaustive Search

- Evaluate all combinations of predictors
- For "n" predictors, how many models can you run with different combinations of X's
  - 2^n-1
- Three predictors X_1, X_2 , X_3
  - 7 models
  - $Y \sim X_1, Y \sim X_2, Y \sim X_3, Y \sim X_1 + X_2, Y \sim X_1 + X_3, Y \sim X_2 + X_3, Y \sim X_1 + X_2 + X_3$
- Choose the model based on one of the performance measures
  - High Adjusted R-Square (R^2)
  - Akaike Information Criterion (AIC) , Bayesian Information Criterion (BIC)
  - Mallow's C_p

Subset selection algorithms

- Finding best subset of predictors
- Iterative process
- Computationally inexpensive
- Algorithms
    - Forward selection
    - Backward elimination

Algorithms

- Backward Elimination
  - Step 1 : Run a regression with all the predictor variables
  - Step 2 : Drop the insignificant predictor with the highest p-value
  - Step 3 : Run a regression model with the remaining predictors
  - Step 4 : Repeat steps 2 & 3 until all the predictors are significant
- Forward Selection
  - Step 1 : Run list of regression models with each individual predictor separately
  - Step 2 : Choose the model among the list with highest R^2
  - Step 3 : Run list of regression models by incrementally advancing Step 2 model by adding remaining predictors individually
  - Step 4 : Repeat steps 2 & 3 until all predictors are significant in the model and all exhaustive combinations are executed

Summary

- Advantages
  - Useful for predictions and insights
  - Statistical foundations
  - Appropriate for small or large datasets
- Disadvantages
  - Limited modeling flexibility
  - Statistical assumptions

Next Class

- Model Evaluation and Accuracy measures for Regression

Thank You