Model Evaluation & Accuracy Measures Regression

Previous class

- Linear Regression
- Application in R/RStudio and Inference

Recap of 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
- Estimation methodology : **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 $\widehat{\ }$

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 consider the predictors mentioned above
- How does the linear regression model looks like?

Multiple linear regression model

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

Today’s class mandatory steps

- Create a folder name “**h.model_evaluation_regression**” within the folder “**oba_455_555_ddpm_r/rproject**”
- Download “**model_evaluation_regression _code.R**”, and all **csv** files from canvas
- Place all downloaded files in “**oba_455_555_ddpm_r/rproject/ h. model_evaluation_regression**”
- Open RStudio project
- Open “**model_evaluation_regression_code.R**” file within RStudio

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 βs)

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 
```

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 all 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

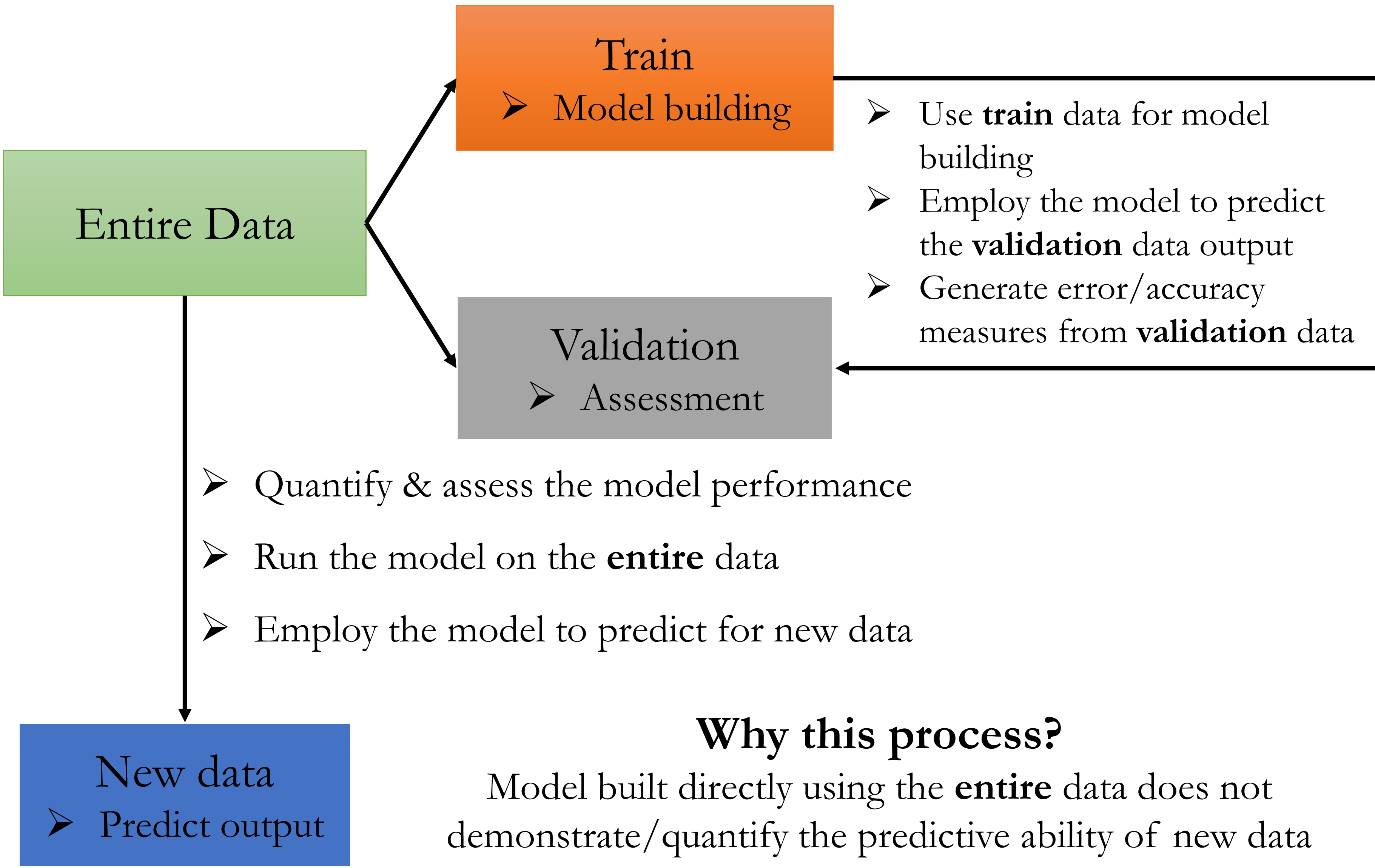
- Advantages
  - Useful for predictions and insights
  - Statistical foundations
  - Appropriate for small or large datasets
- Disadvantages
  - Limited modeling flexibility
  - Statistical assumptions

Managerial challenges

- Up to present we learned to build a regression model and predict for a new observation(s)
- How good are these predictions ?
- Can I deploy this model for making any business decisions ?
- What is the error margin?
- Is there a way to quantify or develop measures for error rate or predictive performance ?

Evaluating Predictive Performance
Regression

Data Partition : Training & Validation



Why this process?

Model built directly using the **entire** data does not demonstrate/quantify the predictive ability of new data

Toyota data partition

- Let's us consider 70-30 partition
- **Train** : Randomly filter 70% of the main data
- **Validation** : Extract remaining 30% of the main data
- 70% training data can differ for each of us. Why?
- Random data generation is machine specific
  - Mac/Windows, 2GHZ/2.4 GHZ, ………
- How to ensure partition process is identical for each of us?
- Seed
  - Set a seed
  - Identical seed is likely to result in same train & validation partition for each of us

Model building on Train data

Is Regression as a whole significant ?

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

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age             -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km              -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp               2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color       -4.311e+00  9.143e+01  -0.047   0.9624
automatic        1.320e+02  1.880e+02   0.702   0.4827
cc              -3.994e-02  9.185e-02  -0.435   0.6638
doors           -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax    8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight           2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,	Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

Regression on train data

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

Significance of individual predictors

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

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age             -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km              -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp               2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color       -4.311e+00  9.143e+01  -0.047   0.9624
automatic        1.320e+02  1.880e+02   0.702   0.4827
cc              -3.994e-02  9.185e-02  -0.435   0.6638
doors           -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax    8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight           2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

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

Impact of individual predictors

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

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age            -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km             -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp              2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color      -4.311e+00  9.143e+01  -0.047   0.9624
automatic       1.320e+02  1.880e+02   0.702   0.4827
cc             -3.994e-02  9.185e-02  -0.435   0.6638
doors          -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax   8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight          2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

Coefficients (All β^s)

Interpreting character predictor

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

Residuals:
     Min       1Q   Median       3Q      Max 
-12352.2   -758.4    -64.0    731.0   6383.4 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age             -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km              -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222    
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp               2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color       -4.311e+00  9.143e+01  -0.047   0.9624    
automatic        1.320e+02  1.880e+02   0.702   0.4827    
cc              -3.994e-02  9.185e-02  -0.435   0.6638    
doors           -1.238e+02  4.824e+01  -2.565   0.0105 *  
quarterly_tax    8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight           2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,	Adjusted R-squared:  0.8736 
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

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

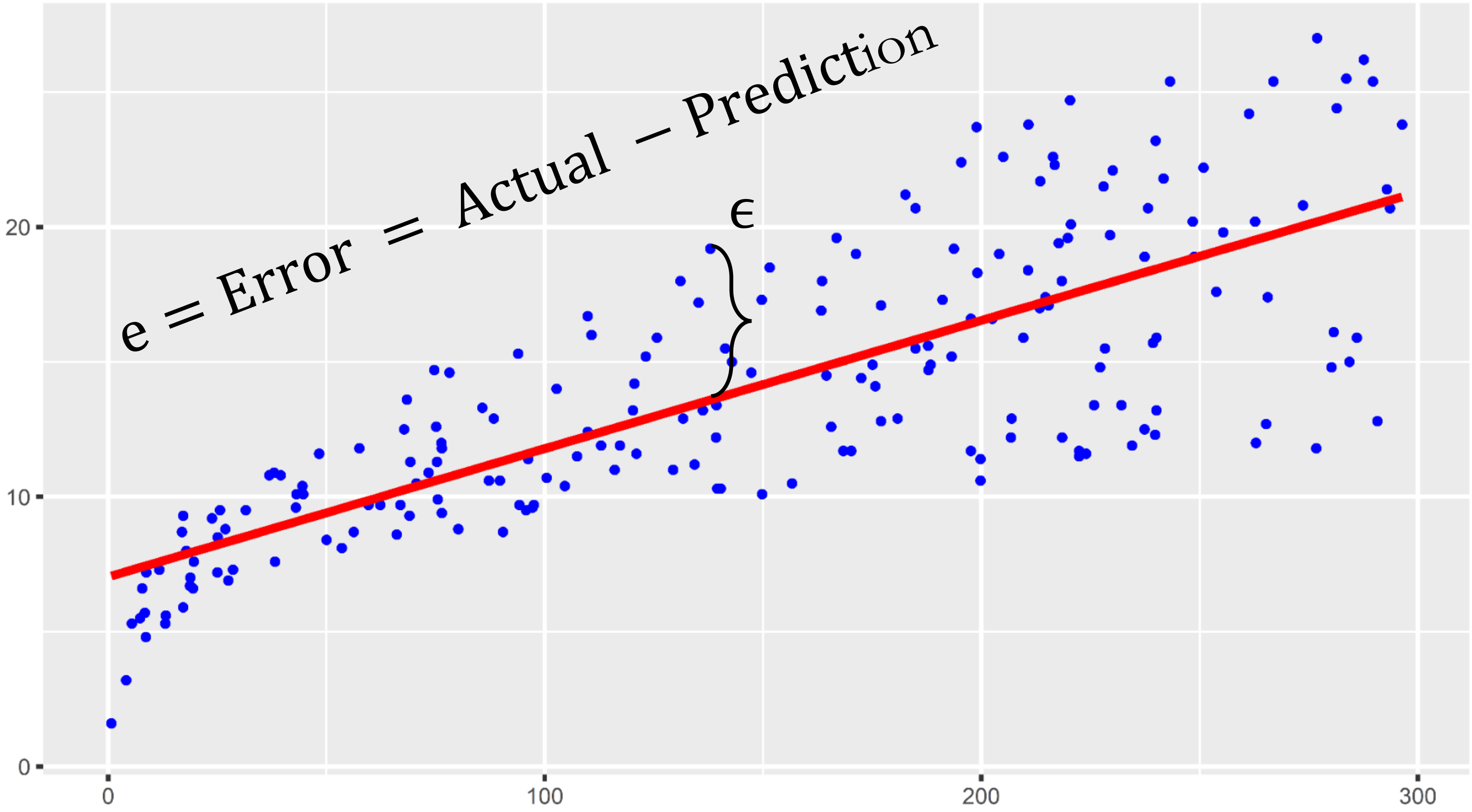
Model fit

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

Residuals:
     Min       1Q   Median       3Q      Max
-12352.2   -758.4    -64.0    731.0   6383.4

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.328e+03  1.514e+03  -6.162 1.04e-09 ***
age             -1.218e+02  3.179e+00 -38.295  < 2e-16 ***
km              -1.774e-02  1.639e-03 -10.825  < 2e-16 ***
fuel_typeDiesel  8.093e+02  5.232e+02   1.547   0.1222
fuel_typePetrol  2.253e+03  5.117e+02   4.404 1.18e-05 ***
hp               2.483e+01  4.130e+00   6.011 2.59e-09 ***
met_color       -4.311e+00  9.143e+01  -0.047   0.9624
automatic        1.320e+02  1.880e+02   0.702   0.4827
cc              -3.994e-02  9.185e-02  -0.435   0.6638
doors           -1.238e+02  4.824e+01  -2.565   0.0105 *
quarterly_tax    8.457e+00  2.031e+00   4.164 3.39e-05 ***
weight           2.175e+01  1.507e+00  14.438  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1326 on 993 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8736
F-statistic: 631.6 on 11 and 993 DF,  p-value: < 2.2e-16
```

Multiple R-Square (R^2) : Proportion of variation in price explained by predictors in the model

Error

- Error (e_i) for each observation i
- Error (e_i) : Difference between actual (Y_i) and predicted outcome ($\hat{Y}_i$)



Error measures for Regression

- Mean Error (ME) : $\frac{1}{n}\sum_{i=1}^{n} e_i$
    - ➢ Indicates on-average predictions are over or under the outcome
- Mean Absolute Error (MAE) : $\frac{1}{n}\sum_{i=1}^{n} |e_i|$
    - ➢ Magnitude of average absolute error
- Mean Percentage Error (MPE) : $\left(\frac{1}{n}\sum_{i=1}^{n} \frac{e_i}{Y_i}\right) * 100$
    - ➢ Measure relative to the size of outcome Y_i
- MAPE (Mean Absolute Percentage Error) : $\left(\frac{1}{n}\sum_{i=1}^{n} \left|\frac{e_i}{Y_i}\right|\right) * 100$
- Root Mean Square Prediction Error (RMSE) : $\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$
    - ➢ Like standard error has same units as outcome Y_i

Error/Accuracy Measures

- We computed error measures for **validation** data
- Can they be computed for **training** data?
- What do the measures infer for each data?

Training

- Goodness-of-fit
- Additional measures - R^2, standard error
- Does not indicate predictive abilities

Validation

- Indicates predictive abilities
- Used to **compare across models** to assess their degree of prediction accuracy

- **Overfitting** can be detected by comparing the error measures between **training** and **validation** data
- Greater the difference in train & validation data error measures, greater the overfitting

Package to compute all error/accuracy measures

- Package "**forecast**"
- Function - **accuracy( )**

Comparing measures from both **train** & **validation** data

Dataset	Mean Error (ME)	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)	Mean % Error (MPE)	Mean Absolute % Error (MAPE)
Validation	-33.243	1317.55	984.833	-1.282	10.104
Training	$1.16 * 10^{-10}$	1317.829	964.804	-1.085	9.498

Next Class

- Model Evaluation and Accuracy measures for Classification

Thank You