

# Weekly Summary Template

Brady Miller

## Table of contents

Tuesday, Feb 16 . . . . .	1
Thursday, Feb 18 . . . . .	17

---

## Tuesday, Feb 16

### ! TIL

Include a *very brief* summary of what you learnt in this class here.

Today, I learnt the following concepts in class:

1. How to make a correlation plot and its interpretation
2. Variance inflation
3. Step-wise Regression

```
# loading in necessary libraries and datasets
library(ISLR2)
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

```
intersect, setdiff, setequal, union
```

```
library(tidyr)
library(purrr)
library(readr)
library(glmnet)
```

Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

```
expand, pack, unpack
```

Loaded glmnet 4.1-6

```
library(caret)
```

Loading required package: ggplot2

Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

```
lift
```

```
library(car)
```

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:purrr':

some

The following object is masked from 'package:dplyr':

recode

```
df <- Boston
attach(Boston)
```

## Correlation Plot

```
# creates new data frame that outputs correlation table with numeric values
R <- df %>%
  keep(is.numeric) %>%
  cor()
R
```

	crim	zn	indus	chas	nox		
crim	1.00000000	-0.20046922	0.40658341	-0.055891582	0.42097171		
zn	-0.20046922	1.00000000	-0.53382819	-0.042696719	-0.51660371		
indus	0.40658341	-0.53382819	1.00000000	0.062938027	0.76365145		
chas	-0.05589158	-0.04269672	0.06293803	1.000000000	0.09120281		
nox	0.42097171	-0.51660371	0.76365145	0.091202807	1.00000000		
rm	-0.21924670	0.31199059	-0.39167585	0.091251225	-0.30218819		
age	0.35273425	-0.56953734	0.64477851	0.086517774	0.73147010		
dis	-0.37967009	0.66440822	-0.70802699	-0.099175780	-0.76923011		
rad	0.62550515	-0.31194783	0.59512927	-0.007368241	0.61144056		
tax	0.58276431	-0.31456332	0.72076018	-0.035586518	0.66802320		
ptratio	0.28994558	-0.39167855	0.38324756	-0.121515174	0.18893268		
lstat	0.45562148	-0.41299457	0.60379972	-0.053929298	0.59087892		
medv	-0.38830461	0.36044534	-0.48372516	0.175260177	-0.42732077		
	rm	age	dis	rad	tax	ptratio	
crim	-0.21924670	0.35273425	-0.37967009	0.625505145	0.58276431	0.2899456	
zn	0.31199059	-0.56953734	0.66440822	-0.311947826	-0.31456332	-0.3916785	
indus	-0.39167585	0.64477851	-0.70802699	0.595129275	0.72076018	0.3832476	
chas	0.09125123	0.08651777	-0.09917578	-0.007368241	-0.03558652	-0.1215152	
nox	-0.30218819	0.73147010	-0.76923011	0.611440563	0.66802320	0.1889327	
rm	1.00000000	-0.24026493	0.20524621	-0.209846668	-0.29204783	-0.3555015	

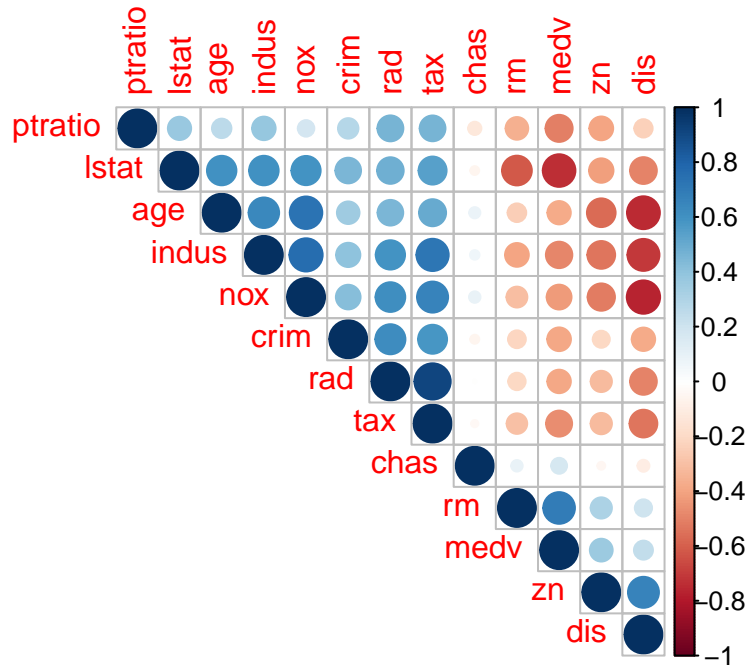
age	-0.24026493	1.00000000	-0.74788054	0.456022452	0.50645559	0.2615150
dis	0.20524621	-0.74788054	1.00000000	-0.494587930	-0.53443158	-0.2324705
rad	-0.20984667	0.45602245	-0.49458793	1.000000000	0.91022819	0.4647412
tax	-0.29204783	0.50645559	-0.53443158	0.910228189	1.00000000	0.4608530
ptratio	-0.35550149	0.26151501	-0.23247054	0.464741179	0.46085304	1.0000000
lstat	-0.61380827	0.60233853	-0.49699583	0.488676335	0.54399341	0.3740443
medv	0.69535995	-0.37695457	0.24992873	-0.381626231	-0.46853593	-0.5077867
	lstat	medv				
crim	0.4556215	-0.3883046				
zn	-0.4129946	0.3604453				
indus	0.6037997	-0.4837252				
chas	-0.0539293	0.1752602				
nox	0.5908789	-0.4273208				
rm	-0.6138083	0.6953599				
age	0.6023385	-0.3769546				
dis	-0.4969958	0.2499287				
rad	0.4886763	-0.3816262				
tax	0.5439934	-0.4685359				
ptratio	0.3740443	-0.5077867				
lstat	1.0000000	-0.7376627				
medv	-0.7376627	1.0000000				

- Correlation can range between  $[-1,1]$ 
  1. Close to 1 = strong positive correlation(as one variable increases, the other variable increases)
  2. Close to -1 = strong negative correlation(as one variable increases, the other variable decreases)
  3. Close to 0 = very weak/no correlation between variables
- Can come up with a  $p$ -value for a correlation value to see if there is a relationship between 2 variables (see if one is a good predictor of another)

```
# creates correlation plot
library(corrplot)
```

corrplot 0.92 loaded

```
corrplot(R, type = 'upper', order = 'hclust')
```



- Red = negative correlation
- Blue = positive correlation
- Can't isolate effect of indus and nox as they are related so won't be able to hold one constant while increasing the other (have very high correlation)
- Tax is negatively correlated with indus (total nuberm of business), so if the number of businesses in neighborhood goes up, then the distance from that given house to a business district decreases/goes down

## Variance Inflation

```
# creating new model that we can use to investigate variance inflation
new_cols <- colnames(df)[-c(5,13)]
model <- lm(medv ~ ., df %>% select(-c(indus, nox, dis)))
summary(model)
```

Call:

```
lm(formula = medv ~ ., data = df %>% select(-c(indus, nox, dis)))
```

Residuals:

```
      Min       1Q   Median       3Q      Max
```

-16.9388 -3.0974 -0.7082 1.8472 28.3443

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	21.655695	4.215323	5.137	4.01e-07	***
crim	-0.091908	0.034722	-2.647	0.008380	**
zn	0.008794	0.012670	0.694	0.487957	
chas	2.952830	0.913519	3.232	0.001309	**
rm	4.100202	0.439135	9.337	< 2e-16	***
age	0.020892	0.012195	1.713	0.087315	.
rad	0.251852	0.067890	3.710	0.000231	***
tax	-0.012434	0.003469	-3.584	0.000371	***
ptratio	-0.886594	0.129206	-6.862	2.03e-11	***
lstat	-0.573951	0.053177	-10.793	< 2e-16	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.084 on 496 degrees of freedom

Multiple R-squared: 0.6999, Adjusted R-squared: 0.6944

F-statistic: 128.5 on 9 and 496 DF, p-value: < 2.2e-16

- Variance inflation = if you have 2 variables that are highly correlated with one another, if you change one variable, then you would expect change in the other variable in the same direction (can't hold that variable constant while increasing the other).
1. variance is used to compute standard errors
  2. if standard error goes up, then  $p$ -value goes up, so it becomes more significant & less likely to reject the null hypothesis
  3. inflating it with sum of other variables
  4. high variance inflation = gets very high, so it becomes very insignificant

```
# inflation variance factor table
library(car)
library(knitr)
library(rmarkdown)
vif_model <- lm(medv ~ ., df)
vif(vif_model) %>% knitr::kable()
```

	x
crim	1.767486
zn	2.298459
indus	3.987181

	x
chas	1.071168
nox	4.369093
rm	1.912532
age	3.088232
dis	3.954037
rad	7.445301
tax	9.002158
ptratio	1.797060
lstat	2.870776

- This table measures the of the amount of multicollinearity in a set of multiple regression variables.
- high inflation factor = anything greater than 2
- low inflation factor = anything less than 2

### Stepwise Regression

- can drop a subset of the variables that don't explain enough the variability in the  $R^2$
- by adding more variables,  $R^2$  goes up → now you can ask what variables leads to the highest increase in  $R^2$  /more significant/good predictors and have low variance inflation

```
# Null model only uses the intercept for the model predicting median house price
null_model <- lm(medv ~ 1,df)
null_model
```

Call:

```
lm(formula = medv ~ 1, data = df)
```

Coefficients:

```
(Intercept)
      22.53
```

```
# Full model utilizes all variables in the prediction of median house price
full_model <- lm(medv ~ .,df)
full_model
```

Call:

```
lm(formula = medv ~ ., data = df)
```

Coefficients:

(Intercept)	crim	zn	indus	chas	nox
41.617270	-0.121389	0.046963	0.013468	2.839993	-18.758022
rm	age	dis	rad	tax	ptratio
3.658119	0.003611	-1.490754	0.289405	-0.012682	-0.937533
lstat					
-0.552019					

Forward Selection

- Starts with no covariates
- Add each covariate 1 by 1 in order of variables importance, increasing  $R^2$

```
# uses step function to run this
forward_model <- step(null_model, direction = 'forward',
                      scope = formula(full_model))
```

Start: AIC=2246.51

medv ~ 1

	Df	Sum of Sq	RSS	AIC
+ lstat	1	23243.9	19472	1851.0
+ rm	1	20654.4	22062	1914.2
+ ptratio	1	11014.3	31702	2097.6
+ indus	1	9995.2	32721	2113.6
+ tax	1	9377.3	33339	2123.1
+ nox	1	7800.1	34916	2146.5
+ crim	1	6440.8	36276	2165.8
+ rad	1	6221.1	36495	2168.9
+ age	1	6069.8	36647	2171.0
+ zn	1	5549.7	37167	2178.1
+ dis	1	2668.2	40048	2215.9
+ chas	1	1312.1	41404	2232.7
<none>			42716	2246.5

Step: AIC=1851.01

medv ~ lstat

	Df	Sum of Sq	RSS	AIC
+ rm	1	4033.1	15439	1735.6
+ ptratio	1	2670.1	16802	1778.4



+ chas	1	786.3	18686	1832.2
+ dis	1	772.4	18700	1832.5
+ age	1	304.3	19168	1845.0
+ tax	1	274.4	19198	1845.8
+ zn	1	160.3	19312	1848.8
+ crim	1	146.9	19325	1849.2
+ indus	1	98.7	19374	1850.4
<none>			19472	1851.0
+ rad	1	25.1	19447	1852.4
+ nox	1	4.8	19468	1852.9

Step: AIC=1735.58

medv ~ lstat + rm

	Df	Sum of Sq	RSS	AIC
+ ptratio	1	1711.32	13728	1678.1
+ chas	1	548.53	14891	1719.3
+ tax	1	425.16	15014	1723.5
+ dis	1	351.15	15088	1725.9
+ crim	1	311.42	15128	1727.3
+ rad	1	180.45	15259	1731.6
+ indus	1	61.09	15378	1735.6
<none>			15439	1735.6
+ zn	1	56.56	15383	1735.7
+ age	1	20.18	15419	1736.9
+ nox	1	14.90	15424	1737.1

Step: AIC=1678.13

medv ~ lstat + rm + ptratio

	Df	Sum of Sq	RSS	AIC
+ dis	1	499.08	13229	1661.4
+ chas	1	377.96	13350	1666.0
+ crim	1	122.52	13606	1675.6
+ age	1	66.24	13662	1677.7
<none>			13728	1678.1
+ tax	1	44.36	13684	1678.5
+ nox	1	24.81	13703	1679.2
+ zn	1	14.96	13713	1679.6
+ rad	1	6.07	13722	1679.9
+ indus	1	0.83	13727	1680.1

Step: AIC=1661.39

medv ~ lstat + rm + ptratio + dis

	Df	Sum of Sq	RSS	AIC
+ nox	1	759.56	12469	1633.5
+ chas	1	267.43	12962	1653.1
+ indus	1	242.65	12986	1654.0
+ tax	1	240.34	12989	1654.1
+ crim	1	233.54	12995	1654.4
+ zn	1	144.81	13084	1657.8
+ age	1	61.36	13168	1661.0
<none>			13229	1661.4
+ rad	1	22.40	13206	1662.5

Step: AIC=1633.47

medv ~ lstat + rm + ptratio + dis + nox

	Df	Sum of Sq	RSS	AIC
+ chas	1	328.27	12141	1622.0
+ zn	1	151.71	12318	1629.3
+ crim	1	141.43	12328	1629.7
+ rad	1	53.48	12416	1633.3
<none>			12469	1633.5
+ indus	1	17.10	12452	1634.8
+ tax	1	10.50	12459	1635.0
+ age	1	0.25	12469	1635.5

Step: AIC=1621.97

medv ~ lstat + rm + ptratio + dis + nox + chas

	Df	Sum of Sq	RSS	AIC
+ zn	1	164.406	11977	1617.1
+ crim	1	116.330	12025	1619.1
+ rad	1	58.556	12082	1621.5
<none>			12141	1622.0
+ indus	1	26.274	12115	1622.9
+ tax	1	4.187	12137	1623.8
+ age	1	2.331	12139	1623.9

Step: AIC=1617.07

medv ~ lstat + rm + ptratio + dis + nox + chas + zn

	Df	Sum of Sq	RSS	AIC
+ crim	1	170.902	11806	1611.8

```

<none>                11977 1617.1
+ tax      1      31.773 11945 1617.7
+ rad      1      28.311 11948 1617.9
+ indus    1      27.377 11949 1617.9
+ age      1       0.071 11977 1619.1

```

Step: AIC=1611.8

```
medv ~ lstat + rm + ptratio + dis + nox + chas + zn + crim
```

```

      Df Sum of Sq  RSS    AIC
+ rad   1   155.006 11651 1607.1
<none>                11806 1611.8
+ indus  1    24.957 11781 1612.7
+ tax    1     1.418 11804 1613.7
+ age    1     0.178 11806 1613.8

```

Step: AIC=1607.11

```
medv ~ lstat + rm + ptratio + dis + nox + chas + zn + crim +
      rad
```

```

      Df Sum of Sq  RSS    AIC
+ tax   1   298.573 11352 1596.0
<none>                11651 1607.1
+ indus  1    44.346 11606 1607.2
+ age    1     0.581 11650 1609.1

```

Step: AIC=1595.98

```
medv ~ lstat + rm + ptratio + dis + nox + chas + zn + crim +
      rad + tax
```

```

      Df Sum of Sq  RSS    AIC
<none>                11352 1596.0
+ age    1     1.6865 11350 1597.9
+ indus  1     1.0784 11351 1597.9

```

```
summary(forward_model)
```

Call:

```
lm(formula = medv ~ lstat + rm + ptratio + dis + nox + chas +
    zn + crim + rad + tax, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-15.1814	-2.7625	-0.6243	1.8448	26.3920

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	41.451747	4.903283	8.454	3.18e-16	***
lstat	-0.546509	0.047442	-11.519	< 2e-16	***
rm	3.672957	0.409127	8.978	< 2e-16	***
ptratio	-0.930961	0.130423	-7.138	3.39e-12	***
dis	-1.515951	0.187675	-8.078	5.08e-15	***
nox	-18.262427	3.565247	-5.122	4.33e-07	***
chas	2.871873	0.862591	3.329	0.000935	***
zn	0.046191	0.013673	3.378	0.000787	***
crim	-0.121665	0.032919	-3.696	0.000244	***
rad	0.283932	0.063945	4.440	1.11e-05	***
tax	-0.012292	0.003407	-3.608	0.000340	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.789 on 495 degrees of freedom

Multiple R-squared: 0.7342, Adjusted R-squared: 0.7289

F-statistic: 136.8 on 10 and 495 DF, p-value: < 2.2e-16

- The lower the AIC value, the better
- Add the variables that have the result in the lowest AIC value at that point
- The AIC next to each variable, says which variable that would be added next would make the AIC value the lowest
- lstat is first one added as it makes the AIC value the smallest out of all the variables
- forward selection -> keep building model by including variables
- at the end, if you add age or indus, the AIC value increases, which means you should stop before adding those variables (don't use those variables in the model)

Backward Selection

- Start with full model
- Remove variables 1 by 1 to see which results in a decrease in the AIC value

```
backward_model <- step(full_model, direction = 'backward',  
                        scope = formula(null_model))
```

Start: AIC=1599.85

medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +

tax + ptratio + lstat

	Df	Sum of Sq	RSS	AIC
- indus	1	1.08	11350	1597.9
- age	1	1.69	11351	1597.9
<none>			11349	1599.8
- chas	1	245.31	11595	1608.7
- tax	1	256.28	11606	1609.2
- zn	1	263.59	11613	1609.5
- crim	1	311.49	11661	1611.6
- rad	1	430.71	11780	1616.7
- nox	1	546.10	11896	1621.6
- ptratio	1	1157.70	12507	1647.0
- dis	1	1258.52	12608	1651.1
- rm	1	1744.36	13094	1670.2
- lstat	1	2733.54	14083	1707.0

Step: AIC=1597.9

medv ~ crim + zn + chas + nox + rm + age + dis + rad + tax +  
ptratio + lstat

	Df	Sum of Sq	RSS	AIC
- age	1	1.69	11352	1596.0
<none>			11350	1597.9
- chas	1	251.21	11602	1607.0
- zn	1	262.99	11614	1607.5
- tax	1	299.68	11650	1609.1
- crim	1	313.07	11664	1609.7
- rad	1	453.61	11804	1615.7
- nox	1	574.23	11925	1620.9
- ptratio	1	1168.01	12518	1645.5
- dis	1	1333.19	12684	1652.1
- rm	1	1750.50	13101	1668.5
- lstat	1	2743.21	14094	1705.4

Step: AIC=1595.98

medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +  
lstat

	Df	Sum of Sq	RSS	AIC
<none>			11352	1596.0
- chas	1	254.21	11606	1605.2
- zn	1	261.75	11614	1605.5

```

- tax      1      298.57 11651 1607.1
- crim     1      313.27 11666 1607.8
- rad      1      452.16 11804 1613.7
- nox      1      601.74 11954 1620.1
- ptratio  1      1168.51 12521 1643.5
- dis      1      1496.35 12848 1656.6
- rm       1      1848.38 13201 1670.3
- lstat    1      3043.23 14395 1714.2

```

```
summary(backward_model)
```

Call:

```
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
    tax + ptratio + lstat, data = df)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-15.1814	-2.7625	-0.6243	1.8448	26.3920

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	41.451747	4.903283	8.454	3.18e-16 ***
crim	-0.121665	0.032919	-3.696	0.000244 ***
zn	0.046191	0.013673	3.378	0.000787 ***
chas	2.871873	0.862591	3.329	0.000935 ***
nox	-18.262427	3.565247	-5.122	4.33e-07 ***
rm	3.672957	0.409127	8.978	< 2e-16 ***
dis	-1.515951	0.187675	-8.078	5.08e-15 ***
rad	0.283932	0.063945	4.440	1.11e-05 ***
tax	-0.012292	0.003407	-3.608	0.000340 ***
ptratio	-0.930961	0.130423	-7.138	3.39e-12 ***
lstat	-0.546509	0.047442	-11.519	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.789 on 495 degrees of freedom

Multiple R-squared: 0.7342, Adjusted R-squared: 0.7289

F-statistic: 136.8 on 10 and 495 DF, p-value: < 2.2e-16

- removes variables that would give it a higher AIC until there are no more of that type
- forward selection and backward selection DON'T always give us the same model

Using both selection methods

- use the full model for both and direction as 'both'

```
selected_model <- step(full_model, direction = 'both',  
                        scope = formula(full_model))
```

Start: AIC=1599.85

```
medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +  
      tax + ptratio + lstat
```

	Df	Sum of Sq	RSS	AIC
- indus	1	1.08	11350	1597.9
- age	1	1.69	11351	1597.9
<none>			11349	1599.8
- chas	1	245.31	11595	1608.7
- tax	1	256.28	11606	1609.2
- zn	1	263.59	11613	1609.5
- crim	1	311.49	11661	1611.6
- rad	1	430.71	11780	1616.7
- nox	1	546.10	11896	1621.6
- ptratio	1	1157.70	12507	1647.0
- dis	1	1258.52	12608	1651.1
- rm	1	1744.36	13094	1670.2
- lstat	1	2733.54	14083	1707.0

Step: AIC=1597.9

```
medv ~ crim + zn + chas + nox + rm + age + dis + rad + tax +  
      ptratio + lstat
```

	Df	Sum of Sq	RSS	AIC
- age	1	1.69	11352	1596.0
<none>			11350	1597.9
+ indus	1	1.08	11349	1599.8
- chas	1	251.21	11602	1607.0
- zn	1	262.99	11614	1607.5
- tax	1	299.68	11650	1609.1
- crim	1	313.07	11664	1609.7
- rad	1	453.61	11804	1615.7
- nox	1	574.23	11925	1620.9
- ptratio	1	1168.01	12518	1645.5
- dis	1	1333.19	12684	1652.1

```
- rm      1    1750.50 13101 1668.5
- lstat   1    2743.21 14094 1705.4
```

Step: AIC=1595.98

```
medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
      lstat
```

	Df	Sum of Sq	RSS	AIC
<none>			11352	1596.0
+ age	1	1.69	11350	1597.9
+ indus	1	1.08	11351	1597.9
- chas	1	254.21	11606	1605.2
- zn	1	261.75	11614	1605.5
- tax	1	298.57	11651	1607.1
- crim	1	313.27	11666	1607.8
- rad	1	452.16	11804	1613.7
- nox	1	601.74	11954	1620.1
- ptratio	1	1168.51	12521	1643.5
- dis	1	1496.35	12848	1656.6
- rm	1	1848.38	13201	1670.3
- lstat	1	3043.23	14395	1714.2

```
summary(selected_model)
```

Call:

```
lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
    tax + ptratio + lstat, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.1814	-2.7625	-0.6243	1.8448	26.3920

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	41.451747	4.903283	8.454	3.18e-16 ***
crim	-0.121665	0.032919	-3.696	0.000244 ***
zn	0.046191	0.013673	3.378	0.000787 ***
chas	2.871873	0.862591	3.329	0.000935 ***
nox	-18.262427	3.565247	-5.122	4.33e-07 ***
rm	3.672957	0.409127	8.978	< 2e-16 ***
dis	-1.515951	0.187675	-8.078	5.08e-15 ***



```
rad          0.283932    0.063945    4.440 1.11e-05 ***
tax          -0.012292    0.003407   -3.608 0.000340 ***
ptratio     -0.930961    0.130423   -7.138 3.39e-12 ***
lstat       -0.546509    0.047442  -11.519 < 2e-16 ***
---
```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.789 on 495 degrees of freedom

Multiple R-squared: 0.7342, Adjusted R-squared: 0.7289

F-statistic: 136.8 on 10 and 495 DF, p-value: < 2.2e-16

## Thursday, Feb 18

### ! TIL

Include a *very brief* summary of what you learnt in this class here.

Today, I learnt the following concepts in class:

1. Item 1
2. Item 2
3. Item 3