THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Multiple Regression 2

Lecture 8

STA 371G

Predicting House prices in the Greater Boston Area

Median house price for each census tract, along with other data.

Predicting House prices in the Greater Boston Area

Median house price for each census tract, along with other data.
The final model:

```
> model <- lm(MEDV ~ CRIME+ZONE+NOX+ROOM+DIST
+                   +RADIAL+TAX+PTRATIO+LSTAT, data=boston)
```

- MEDV: Median Price (response)
- CRIME: Per capita crime rate
- ZONE: Proportion of large lots
- NOX: Nitrogen Oxide concentration
- DIST: Distance to employment centers
- ROOM: Average # of rooms
- RADIAL: Accessibility to highways
- TAX: Tax rate (per $10K)
- PTRATIO: Pupil-to-teacher ratio
- LSTAT: Proportion of “lower status”

Overall Null Hypothesis

Is our model useful? Check the R-squared:

```
> summary(model)$r.squared

[1] 0.7282911
```

Overall Null Hypothesis

Is our model useful? Check the R-squared:

```
> summary(model)$r.squared

[1] 0.7282911
```

Can we be confident that our model will generalize to the population?

Overall Null Hypothesis

Is our model useful? Check the R-squared:

```
> summary(model)$r.squared

[1] 0.7282911
```

Can we be confident that our model will generalize to the population?

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (Data explains nothing!)

Overall Null Hypothesis

Is our model useful? Check the R-squared:

```
> summary(model)$r.squared

[1] 0.7282911
```

Can we be confident that our model will generalize to the population?

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (Data explains nothing!)
$H_1 : \beta_i \neq 0$ for some i (At least one predictor is useful)

Overall Null Hypothesis

Is our model useful? Check the R-squared:

```
> summary(model)$r.squared

[1] 0.7282911
```

Can we be confident that our model will generalize to the population?

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (Data explains nothing!)
$H_1 : \beta_i \neq 0$ for some i (At least one predictor is useful)

or

$H_0 : R^2 = 0$

$H_1 : R^2 > 0$

Overall Null Hypothesis

Check the P-value for the F-statistic in the summary

```
Residual standard error: 96.75 on 496 degrees of freedom
Multiple R-squared:  0.7283,    Adjusted R-squared:  0.7234
F-statistic: 147.7 on 9 and 496 DF,  p-value: < 2.2e-16
```

So we can reject the overall null hypothesis!

Overall Null Hypothesis

Check the P-value for the F-statistic in the summary

```
Residual standard error: 96.75 on 496 degrees of freedom
Multiple R-squared:  0.7283,    Adjusted R-squared:  0.7234
F-statistic: 147.7 on 9 and 496 DF,  p-value: < 2.2e-16
```

So we can reject the overall null hypothesis!
R-squared was already too big to suspect that it is zero and we already knew some predictors are statistically significant.

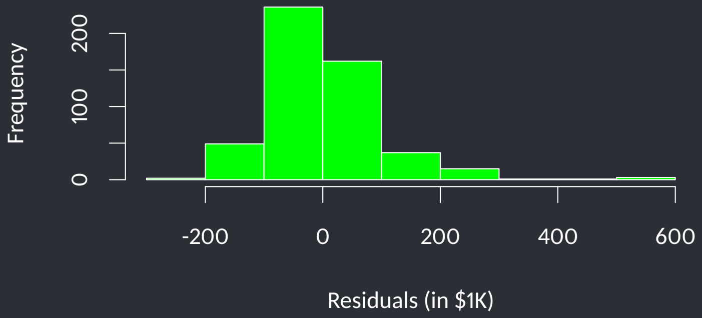
How good are our predictions?

Let’s plot the residuals, i.e., discrepancies between the predictions and the data.

How good are our predictions?

Let's plot the residuals, i.e., discrepancies between the predictions and the data.

```
> hist(model$residuals, col='green',
+   main='', xlab='Residuals (in $1K)', ylab='Frequency')
```



How are we doing with the predictions?

It looks like a normal distribution. Let's look at the mean of the residuals:

How are we doing with the predictions?

It looks like a normal distribution. Let's look at the mean of the residuals:

```
> mean(model$residuals)

[1] -2.028049e-15
```

Virtually zero.

How are we doing with the predictions?

It looks like a normal distribution. Let's look at the mean of the residuals:

```
> mean(model$residuals)

[1] -2.028049e-15
```

Virtually zero.

(It will be always zero since regression minimizes the sum of squared residuals.)

How are we doing with the predictions?

It looks like a normal distribution. Let's look at the mean of the residuals:

```
> mean(model$residuals)

[1] -2.028049e-15
```

Virtually zero.

(It will be always zero since regression minimizes the sum of squared residuals.)

What about the standard deviation?

How are we doing with the predictions?

It looks like a normal distribution. Let's look at the mean of the residuals:

```
> mean(model$residuals)

[1] -2.028049e-15
```

Virtually zero.

(It will be always zero since regression minimizes the sum of squared residuals.)

What about the standard deviation?

```
> sd(model$residuals)

[1] 95.88111
```

How are we doing with the predictions?

It looks like a normal distribution. Let's look at the mean of the residuals:

```
> mean(model$residuals)

[1] -2.028049e-15
```

Virtually zero.

(It will be always zero since regression minimizes the sum of squared residuals.)

What about the standard deviation?

```
> sd(model$residuals)

[1] 95.88111
```

By the 2 standard deviation rule, we could estimate that 95% of the time residuals are in [-$192K, $192K] range.

How are we doing with the predictions?

Can we obtain a similar measure directly from the summary of the regression?

How are we doing with the predictions?

Can we obtain a similar measure directly from the summary of the regression?
It is the Residual standard error!

```
> summary(model)$sigma

[1] 96.74708
```

How are we doing with the predictions?

Can we obtain a similar measure directly from the summary of the regression?
It is the Residual standard error!

```
> summary(model)$sigma

[1] 96.74708
```

```
Residual standard error: 96.75 on 496 degrees of freedom
Multiple R-squared:  0.7283,    Adjusted R-squared:  0.7234
F-statistic: 147.7 on 9 and 496 DF,  p-value: < 2.2e-16
```

Again: regression assumptions

Remember the big four:

1. The residuals are independent.
2. Y is a linear function of Xs (except for the errors).
3. The residuals are normally distributed.
4. The variance of Y is the same for any value of Xs (“homoscedasticity”).

Assumption 1: Independence

Independence: No correlation between residuals.

Assumption 1: Independence

Independence: No correlation between residuals.
Difficult to verify this from plots.

Again: regression assumptions

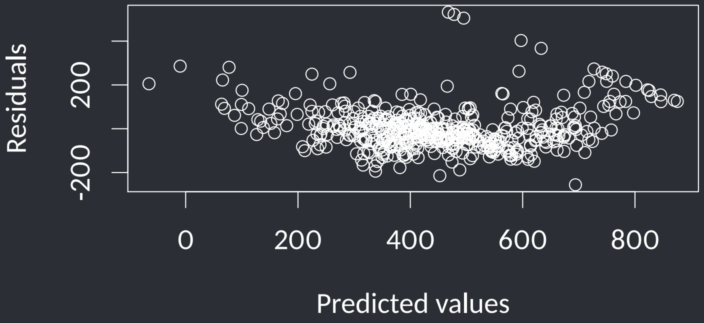
Remember the big four:

1. The residuals are independent.
2. Y is a linear function of Xs (except for the errors).
3. The residuals are normally distributed.
4. The variance of Y is the same for any value of Xs (“homoscedasticity”).

Assumption 2: Linearity

Plot the residuals vs the predicted Y-values and ensure there is no trend:

```
> plot(predict.lm(model), resid(model),
+   xlab='Predicted values', ylab='Residuals')
```



Again: regression assumptions

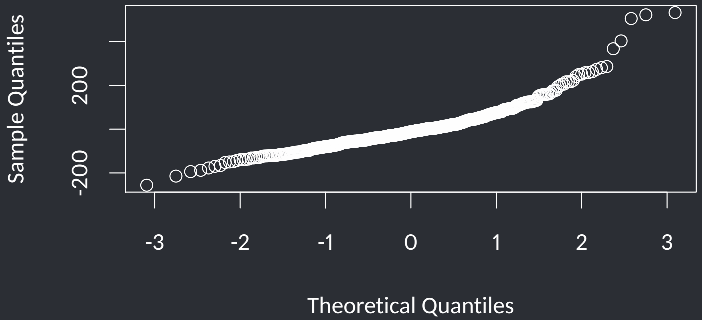
Remember the big four:

1. The residuals are independent.
2. Y is a linear function of Xs (except for the errors).
3. The residuals are normally distributed.
4. The variance of Y is the same for any value of Xs (“homoscedasticity”).

Assumption 3: Normally distributed residuals

Ensure that the Q-Q plot shows a (roughly) straight line:

```
> qqnorm(resid(model), main='')
```



Again: regression assumptions

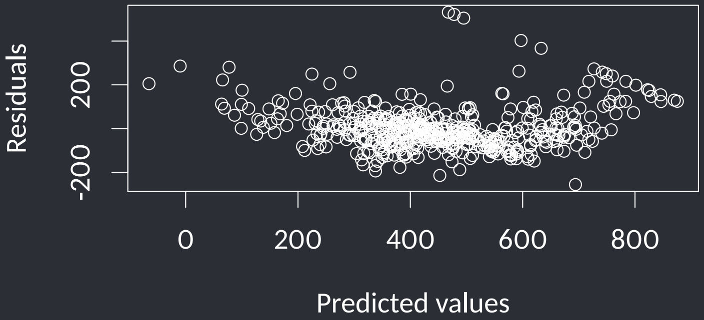
Remember the big four:

1. The residuals are independent.
2. Y is a linear function of Xs (except for the errors).
3. The residuals are normally distributed.
4. The variance of Y is the same for any value of Xs ("homoscedasticity").

Assumption 4: The variance of *Y* is the same across

Look for a (roughly) constant vertical "thickness":

```
> plot(predict.lm(model), resid(model),
+   xlab='Predicted values', ylab='Residuals')
```



We have a model. Then what?

Let's make some predictions.

Model Coefficients

Regression model estimates the coefficients of the predictors.

```
> round(summary(model)$coefficients,2)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   840.07      99.00    8.49        0
CRIME          -2.57       0.66   -3.87        0
ZONE            0.92       0.28    3.34        0
NOX          -346.93      71.81   -4.83        0
ROOM           74.24       8.26    8.99        0
DIST          -31.05       3.78   -8.20        0
RADIAL          6.00       1.29    4.66        0
TAX            -0.27       0.07   -3.87        0
PTRATIO       -19.28       2.63   -7.34        0
LSTAT         -11.07       0.96  -11.56        0
```

Model Coefficients

Let's estimate the median house price in a district, where:

j	Predictor	β_j	X_j	$\beta_j X_j$
0	Intercept	840.07	1	840.07
1	CRIME	-2.57	0.03	-0.0771
2	ZONE	0.92	10	9.2
3	NOX	-346.93	0.5	-173.465
4	ROOM	74.24	4	296.96
5	DIST	-31.05	5	-155.25
6	RADIAL	6	1	6
7	TAX	-0.27	300	-81
8	PTRATIO	-19.28	15	-385.6
9	LSTAT	-11.07	10	-110.7
Price	Estimate	($1000)		342.538

Model Coefficients

Let R do it for us!

```
> predict.lm(model, list(CRIME=0.03, ZONE=10,
+                        NOX=0.5, ROOM=4,
+                        DIST=5,  RADIAL=1,
+                        TAX=300, PTRATIO=15,
+                        LSTAT=10))

       1
343.9552
```

Model Coefficients

Let R do it for us!

```
> predict.lm(model, list(CRIME=0.03, ZONE=10,
+                        NOX=0.5, ROOM=4,
+                        DIST=5,  RADIAL=1,
+                        TAX=300, PTRATIO=15,
+                        LSTAT=10))

       1
343.9552
```

Cool! That was easy!

Model Coefficients

Assume that there are 420 students and 28 teachers in the disctrict (that is why PTRATIO is 15).

Model Coefficients

Assume that there are 420 students and 28 teachers in the disctrict (that is why PTRATIO is 15).

The school board is considering to hire 2 more teachers. How would this affect the house prices in the district?

Model Coefficients

Assume that there are 420 students and 28 teachers in the disctrict (that is why PTRATIO is 15).

The school board is considering to hire 2 more teachers. How would this affect the house prices in the district?

The new PTRATIO will be $420/30 = 14$.

Model Coefficients

Assume that there are 420 students and 28 teachers in the disctrict (that is why PTRATIO is 15).

The school board is considering to hire 2 more teachers. How would this affect the house prices in the district?

The new PTRATIO will be $420/30 = 14$.

```
> predict.lm(model, list(CRIME=0.03, ZONE=10,
+                        NOX=0.5, ROOM=4,
+                        DIST=5,  RADIAL=1,
+                        TAX=300, PTRATIO=14,
+                        LSTAT=10))

       1
363.2349
```

Model Coefficients

Nothing is free. To be able to compansate the new hires, the ISD decides to add $50 more on your tax bill for every $10K of your house price.

Model Coefficients

Nothing is free. To be able to compansate the new hires, the ISD decides to add $50 more on your tax bill for every $10K of your house price.

So, the tax rate increases to 350 per $10K. How would this affect the median house price?

Confidence intervals

We all know our predictions are wrong.
Can we come up with some confidence intervals on our predictions?

Confidence intervals

We all know our predictions are wrong.
Can we come up with some confidence intervals on our predictions?

Remember the two kinds of intervals:

Confidence	Predicting the mean value of Y for a particular set of X values.	Among all the districts whose predictors are as above, what is the mean value of median house price?
Prediction	Predicting Y for a single new case.	If Springfield has the predictors above, what is the median house price in Springfield?

Confidence intervals

```
> predict.lm(model, list(CRIME=0.03, ZONE=10,
+                        NOX=0.5, ROOM=4,
+                        DIST=5,  RADIAL=1,
+                        TAX=350, PTRATIO=14,
+                        LSTAT=10),
+                        interval = 'confidence')

       fit      lwr      upr
1 349.9684 301.9485 397.9883
```

We can also put a confidence intervals on a coefficient to estimate the range of its effect.

```
> confint(model)

                   2.5 %       97.5 %
(Intercept)  645.5520530 1034.5782470
CRIME         -3.8703245   -1.2618439
ZONE           0.3792933    1.4647029
NOX         -488.0175640 -205.8337804
ROOM          58.0099148   90.4751248
DIST         -38.4860994  -23.6129585
RADIAL         3.4693548    8.5311305
TAX           -0.4000457   -0.1306157
PTRATIO      -24.4415728  -14.1179304
LSTAT        -12.9529546   -9.1905075
```

Confidence intervals

Reducing the PTRATIO by one could increase the median house price from $14K to $24K!