THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Multiple Regression 2

Lecture 8

STA 371G

Predicting House prices in 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
```

Are we really sure?

Overall Null Hypothesis

Is our model useful? Check the R-squared:

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

[1] 0.7282911
```

Are we really sure?

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

Overall Null Hypothesis

Is our model useful? Check the R-squared:

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

[1] 0.7282911
```

Are we really sure?

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (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
```

Are we really sure?

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (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

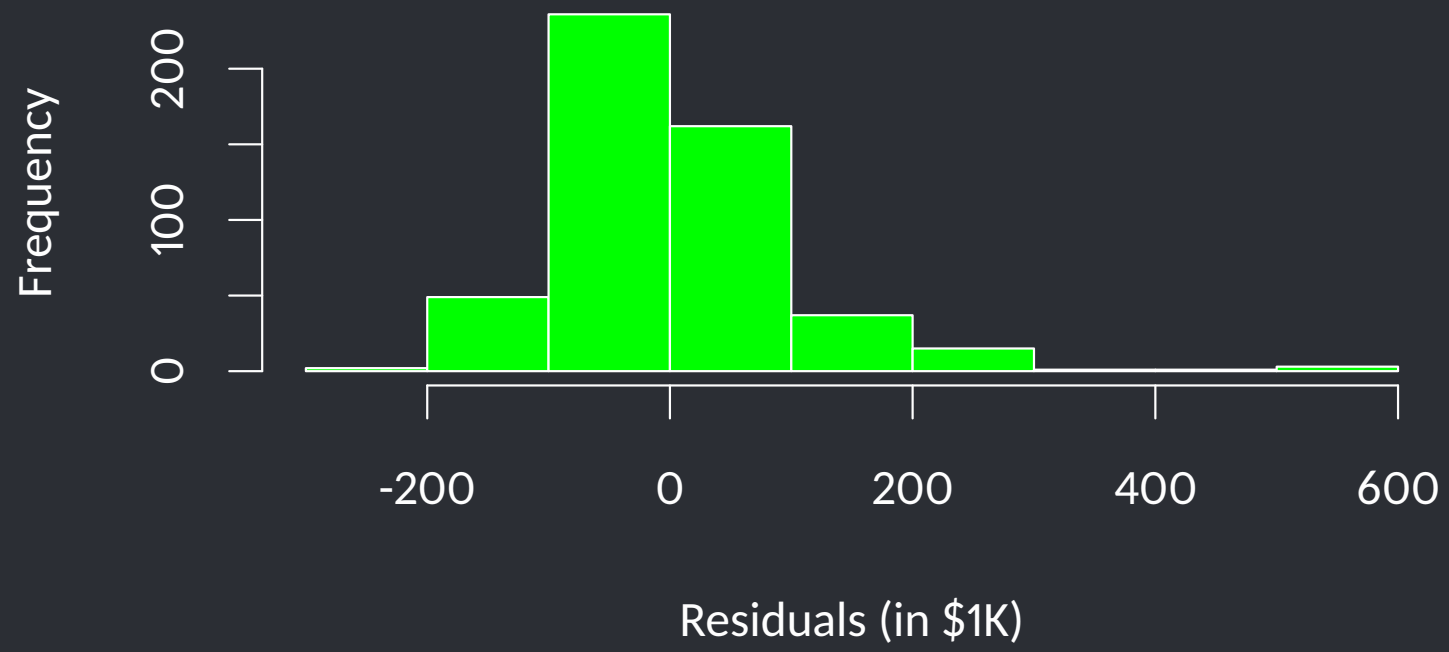
```
Residual standard error: 96.75 on 496 degrees of freedom
Multiple R-squared:  0.7283,    Adjusted R-squared:  0.7234
F-statistic: 147.7 on 9 and 496 DF,  p-value: < 2.2e-16
```

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 do we do with the 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 do we do 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 we allow an intercept and minimize 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 do we do with the predictions?

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

How do we do 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 do we do 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

In multiple regression, we check on five things:

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").
5. No multicollinearity between predictors.

Assumption 1: Independence

Independence: No correlation between residuals.

Assumption 1: Independence

Independence: No correlation between residuals.

Difficult to verify this from plots, use: Durbin-Watson test.

Assumption 1: Independence

Independence: No correlation between residuals.

Difficult to verify this from plots, use: Durbin-Watson test.

H_0: No correlation between residuals (i.e. independent).

H_1: They are not independent

Assumption 1: Independence

Independence: No correlation between residuals.

Difficult to verify this from plots, use: Durbin-Watson test.

H_0: No correlation between residuals (i.e. independent).

H_1: They are not independent

```
> durbinWatsonTest(model)

 lag Autocorrelation D-W Statistic p-value
   1       0.4918641      1.002805       0
 Alternative hypothesis: rho != 0
```

Assumption 1: Independence

Independence: No correlation between residuals.

Difficult to verify this from plots, use: Durbin-Watson test.

H_0: No correlation between residuals (i.e. independent).

H_1: They are not independent

```
> durbinWatsonTest(model)

 lag Autocorrelation D-W Statistic p-value
   1       0.4918641      1.002805       0
 Alternative hypothesis: rho != 0
```

Oops... The model seems to have failed here.

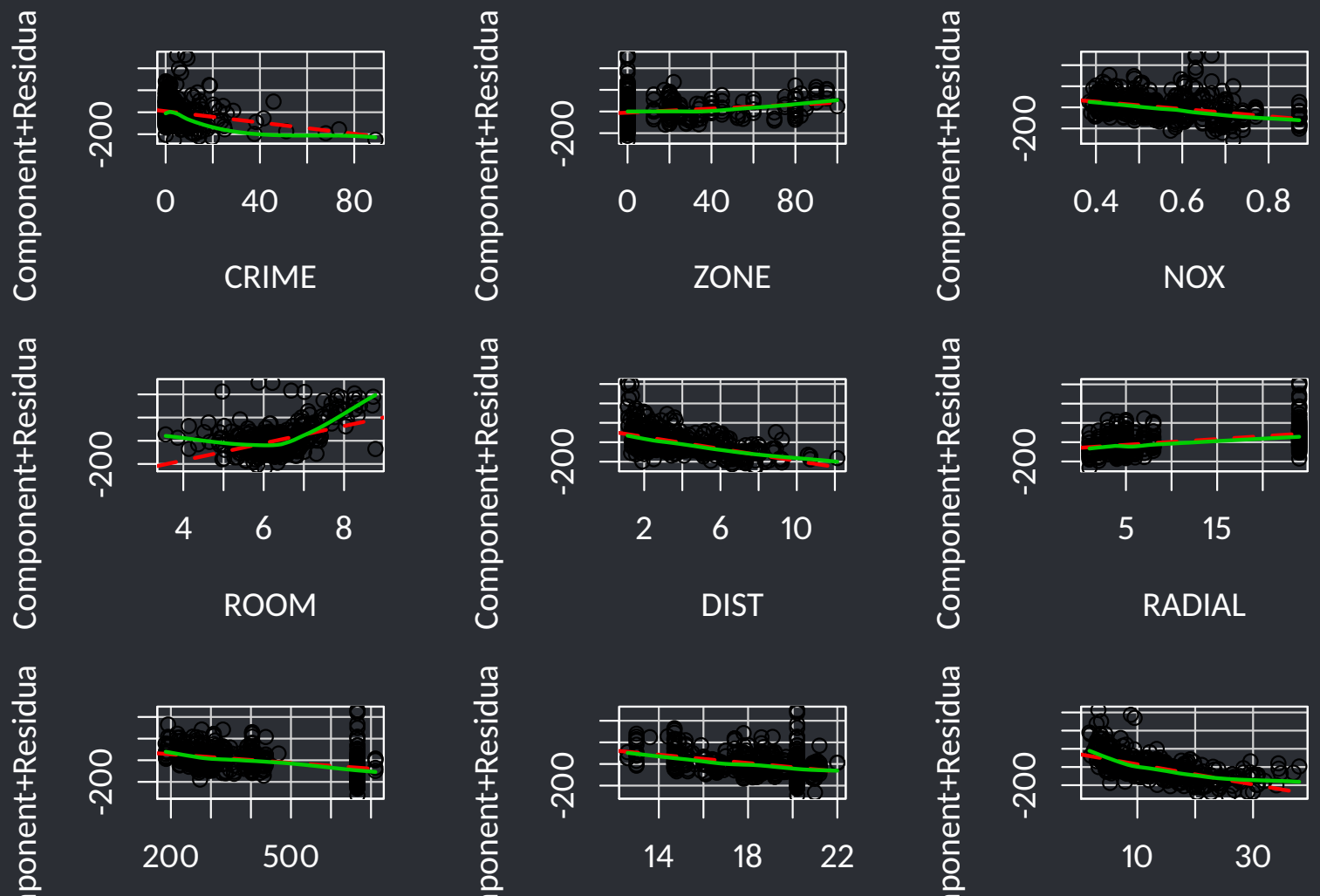
Again: regression assumptions

In multiple regression, we check on five things:

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").
5. No multicollinearity between predictors.

Assumption 2: Linearity

```
> crPlots(model, main='')
```



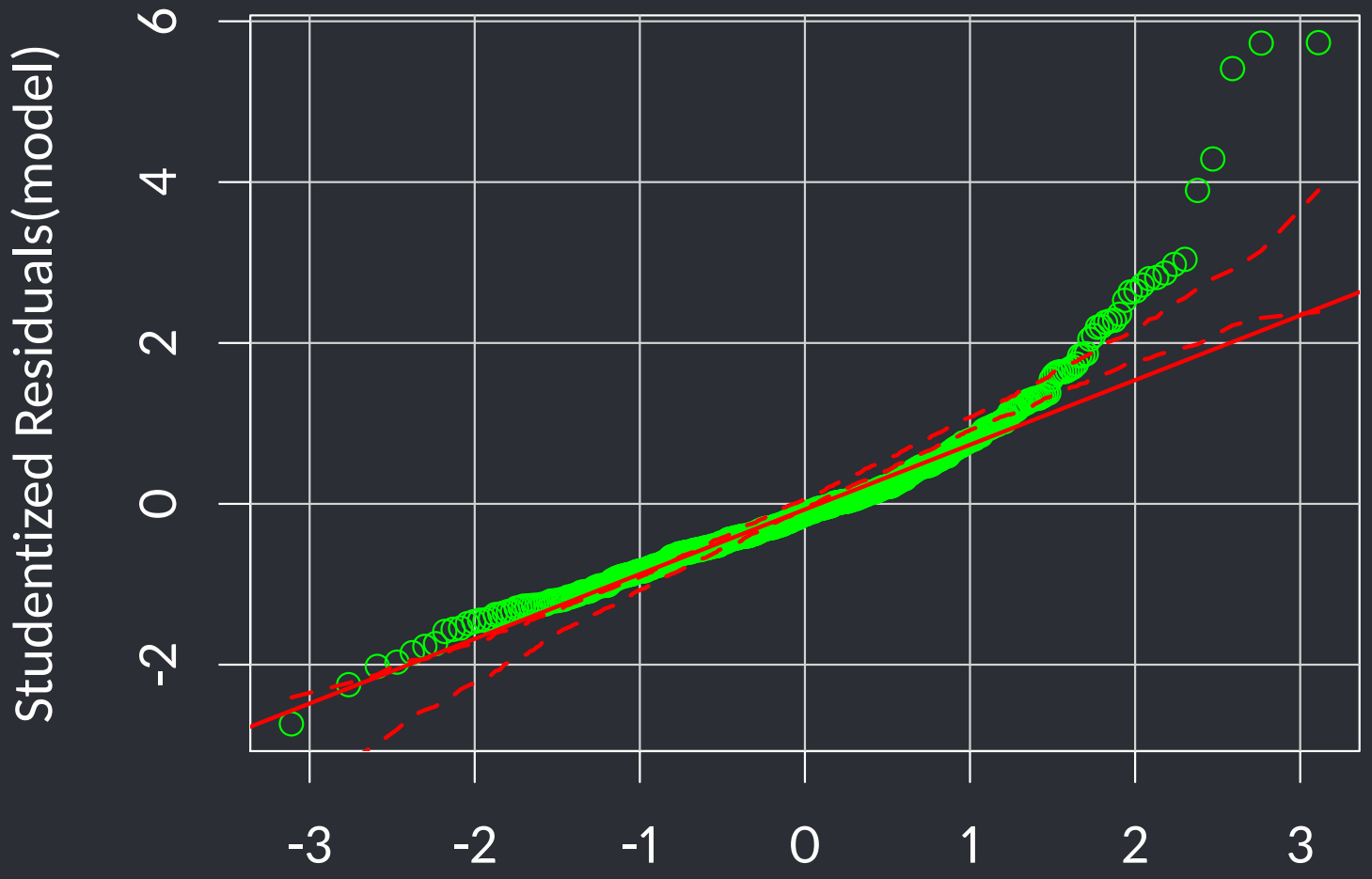
Again: regression assumptions

In multiple regression, we check on five things:

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").
5. No multicollinearity between predictors.

Assumption 3: Normally distributed residuals

```
> qqPlot(model, col='green')
```



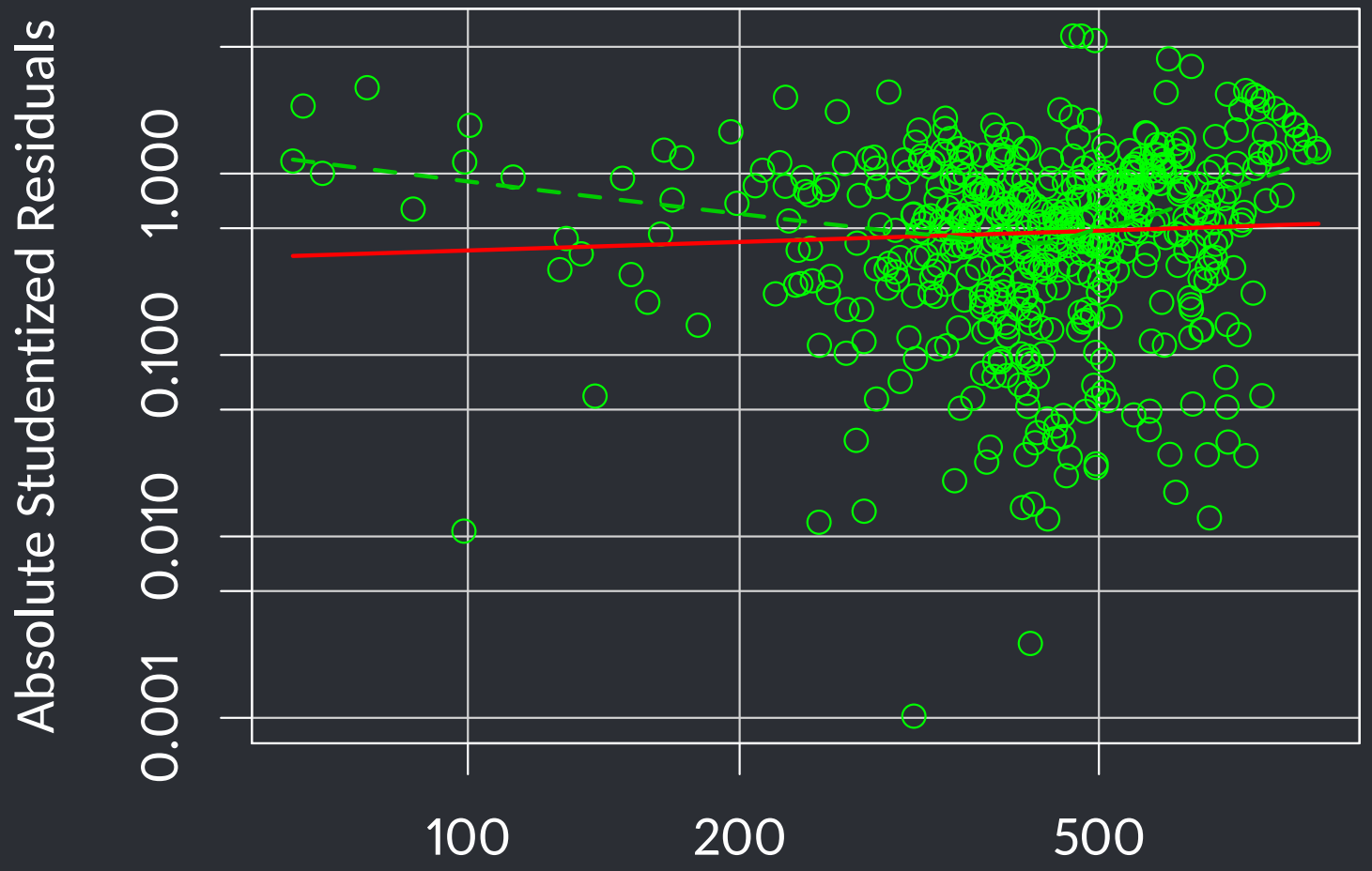
Again: regression assumptions

In multiple regression, we check on five things:

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").
5. No multicollinearity between predictors.

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

```
> spreadLevelPlot(model, col='green', main='')
```



Again: regression assumptions

In multiple regression, we check on five things:

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").
5. No multicollinearity between predictors.

Assumption 5: No multicollinearity

```
> sqrt(vif(model))

   CRIME     ZONE      NOX     ROOM     DIST   RADIAL      T
1.326272 1.496362 1.932860 1.348359 1.851271 2.605272 2.684
   LSTAT
1.588188
```

We have a model. What is next?

Make predictions

Change one of the Xs by one unit

Confidence Intervals: Mean Value and Single case