THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Multiple Regression 2

Lecture 13

STA 371G

Midterm 1 results

- Scores and feedback will be available in Canvas/Quest later today or tomorrow

Midterm 1 results

- Scores and feedback will be available in Canvas/Quest later today or tomorrow
- Approximately: mean = 74%, median = 75%, SD = 15%

Midterm 1 results

- Scores and feedback will be available in Canvas/Quest later today or tomorrow
- Approximately: mean = 74%, median = 75%, SD = 15%
- Since this was our first exam using Quest, you can retake the exam and get up to half credit back starting Thursday 5 PM

Midterm 1 results

- Scores and feedback will be available in Canvas/Quest later today or tomorrow
- Approximately: mean = 74%, median = 75%, SD = 15%
- Since this was our first exam using Quest, you can retake the exam and get up to half credit back starting Thursday 5 PM
- Example: if you got a 60%, and you get a 100% on the retake, your new Midterm 1 score will be 80%

Midterm 1 results

- Scores and feedback will be available in Canvas/Quest later today or tomorrow
- Approximately: mean = 74%, median = 75%, SD = 15%
- Since this was our first exam using Quest, you can retake the exam and get up to half credit back starting Thursday 5 PM
- Example: if you got a 60%, and you get a 100% on the retake, your new Midterm 1 score will be 80%
- Your retake submission is due one week later (Thursday, March 8 at 5 PM)

Predicting median house prices in Boston

```
> 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? Let's look at R^2:

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

[1] 0.7282911
```

Overall Null Hypothesis

Is our model useful? Let's look at R^2:

```
> 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? Let's look at R^2:

```
> 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? Let's look at R^2:

```
> 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!)
$H_A : \beta_i \neq 0$ for some i (At least one predictor is useful)

Overall Null Hypothesis

Is our model useful? Let's look at R^2:

```
> 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!)
$H_A : \beta_i \neq 0$ for some i (At least one predictor is useful)
or
$H_0 : R^2 = 0 \qquad H_A : R^2 > 0$

Overall Null Hypothesis

Check the *p*-value in the summary:

```
Residual standard error: 96750 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 in the summary:

```
Residual standard error: 96750 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^2 was already too big

to suspect that it was actually zero in the population—and we already knew some predictors are statistically significant!)

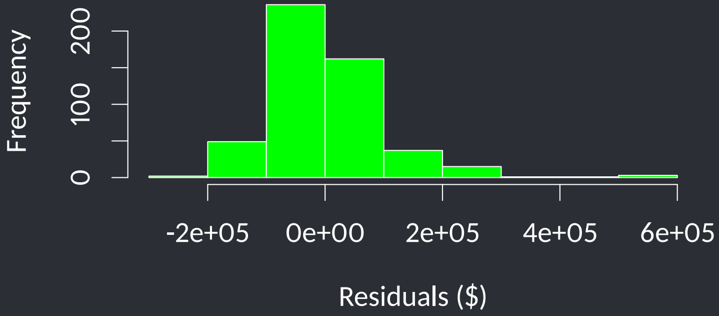
How good are our predictions?

Let's plot the residuals (the discrepancies between the predicted and actual home prices).

How good are our predictions?

Let’s plot the residuals (the discrepancies between the predicted and actual home prices).

```
> hist(resid(model), col='green',
+   main='', xlab='Residuals ($)', ylab='Frequency')
```



How good are our predictions?

Let’s look at the mean of the residuals:

How good are our predictions?

Let’s look at the mean of the residuals:

```
> mean(resid(model))

[1] 7.125761e-12
```

It will be always zero since regression minimizes the sum of squared residuals!

How good are our predictions?

Let’s look at the mean of the residuals:

```
> mean(resid(model))

[1] 7.125761e-12
```

It will be always zero since regression minimizes the sum of squared residuals! Now let’s look at the standard deviation:

How good are our predictions?

Let's look at the mean of the residuals:

```
> mean(resid(model))

[1] 7.125761e-12
```

It will be always zero since regression minimizes the sum of squared residuals! Now let's look at the standard deviation:

```
> sd(resid(model))

[1] 95881.11
```

How good are our predictions?

Let's look at the mean of the residuals:

```
> mean(resid(model))

[1] 7.125761e-12
```

It will be always zero since regression minimizes the sum of squared residuals! Now let's look at the standard deviation:

```
> sd(resid(model))

[1] 95881.11
```

Since the residuals are roughly normal, by the 2 SD rule about 95% of the time predictions will be off by less than $191762.

How good are our predictions?

The residual standard error provides a similar measure directly from the summary of the regression:

```
> summary(model)$sigma

[1] 96747.08
```

How good are our predictions?

The residual standard error provides a similar measure directly from the summary of the regression:

```
> summary(model)$sigma

[1] 96747.08
```

```
Residual standard error: 96750 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 — we have to think this through; can't use a plot here.

Assumption 1: Independence

Independence: No correlation between residuals — we have to think this through; can't use a plot here. Remember that we want the cases to be independent, not the variables. We very much hope that Y will be dependent on the X's (otherwise, why are we trying to predict Y based on the X's?).

Again: regression assumptions

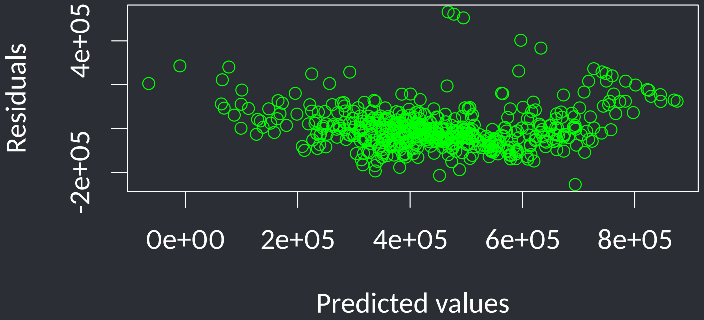
Remember the big four:

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(model), resid(model), col="green",
+   xlab="Predicted values", ylab="Residuals")
```



Again: regression assumptions

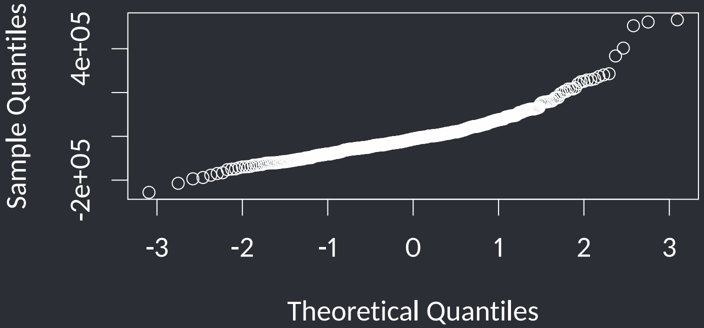
Remember the big four:

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

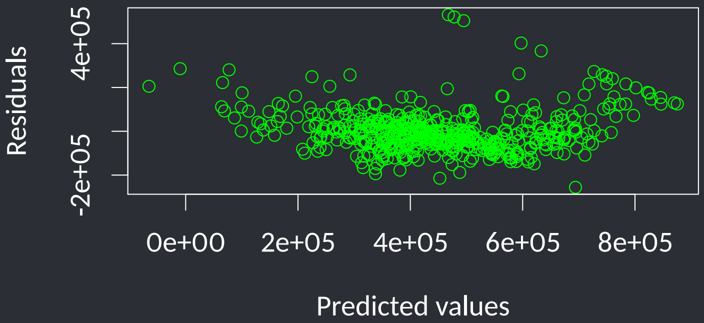
Remember the big four:

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(model), resid(model), col="green",
+   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"). ✓

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"). ✓

w00t w00t!

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"). ✓

w00t w00t! This model meets the assumptions for regression, so it is safe to interpret p-values and confidence intervals.

We have a model. Now what?

Let's make some predictions.

Making predictions

The regression model estimates the coefficients of the predictors:

```
> round(summary(model)$coefficients, 4)

               Estimate Std. Error  t value Pr(>|t|)
(Intercept)  840065.1500 99001.0317   8.4854    0e+00
CRIME         -2566.0842   663.8172  -3.8656    1e-04
ZONE            921.9981   276.2196   3.3379    9e-04
NOX         -346925.6722 71811.3231  -4.8311    0e+00
ROOM          74242.5198  8261.8840   8.9861    0e+00
DIST         -31049.5290  3784.9798  -8.2034    0e+00
RADIAL         6000.2426  1288.1421   4.6581    0e+00
TAX            -265.3307    68.5657  -3.8697    1e-04
PTRATIO      -19279.7516  2627.2042  -7.3385    0e+00
LSTAT        -11071.7310   957.4835 -11.5634    0e+00
```

Making predictions

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

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			342.538

Making predictions

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
343955.2
```

Model coefficients

Suppose there are 420 students and 28 teachers in the disctrict ($\text{PTRATIO} = 420/28 = 15$).

Model coefficients

Suppose there are 420 students and 28 teachers in the disctrict (PTRATIO $= 420/28 = 15$).

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

Model coefficients

Suppose there are 420 students and 28 teachers in the disctrict (PTRATIO $= 420/28 = 15$).

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

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

Model coefficients

Suppose there are 420 students and 28 teachers in the disctrict (PTRATIO $= 420/28 = 15$).

The school board is considering hiring 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
363234.9
```

Model coefficients

To be able to compansate the new hires, the school district must add $50 more on your tax bill for every $10K of your house price.

Model coefficients

To be able to compansate the new hires, the school district must add $50 more on your tax bill for every $10K of your house price.

So, the tax rate increases to $50 per $10K. How would this affect the median house price?

Confidence intervals

We know our predictions are not exactly right.

Can we come up with some confidence intervals on our predictions?

Confidence intervals

We know our predictions are not exactly right.
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 349968.4 301948.5 397988.3
```

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

```
> confint(model)

                    2.5 %       97.5 %
(Intercept)  645552.0530 1034578.2470
CRIME         -3870.3245   -1261.8439
ZONE            379.2933    1464.7029
NOX         -488017.5640 -205833.7804
ROOM          58009.9148   90475.1248
DIST         -38486.0994  -23612.9585
RADIAL         3469.3548    8531.1305
TAX            -400.0457    -130.6157
PTRATIO      -24441.5728  -14117.9304
LSTAT        -12952.9546   -9190.5075
```

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

```
> confint(model)

                     2.5 %       97.5 %
(Intercept)  645552.0530 1034578.2470
CRIME          -3870.3245   -1261.8439
ZONE             379.2933    1464.7029
NOX         -488017.5640 -205833.7804
ROOM           58009.9148   90475.1248
DIST          -38486.0994  -23612.9585
RADIAL          3469.3548    8531.1305
TAX             -400.0457    -130.6157
PTRATIO       -24441.5728  -14117.9304
LSTAT         -12952.9546   -9190.5075
```

We are 95% confident that the effect of reducing the pupil/teacher ratio (PTRATIO) by 1 on median house price is between $14K and $24K!