THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Multiple Regression

Lecture 12

STA 371G

How do you know how much a house is worth?

How do you know how much a house is worth?

Zillow? How do they know?

How do you know how much a house is worth?

Zillow? How do they know?

- Square feet
- Year built
- # of rooms
- Distance to downtown
- Crime rate
- ...

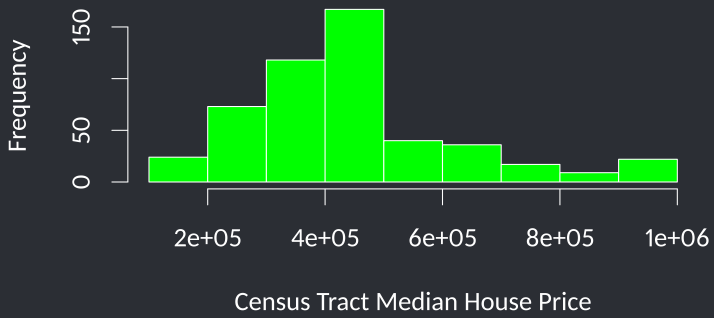
Boston house price data (by census tract, 1970)

- MEDV: Median Price (response)
- LON: Longitude
- LAT: Latitude
- CRIME: Per capita crime rate
- ZONE: Proportion of large lots
- INDUS: Proportion of non-retail business acres
- NOX: Nitrogen Oxide concentration
- ROOM: Average # of rooms
- AGE: Proportion of built before 1940
- DIST: Distance to employment centers
- RADIAL: Accessibility to highways
- TAX: Tax rate (per $10K)
- PTRATIO: Pupil-to-teacher ratio
- LSTAT: Proportion of "lower status"

Can you guess the top three factors?

Distribution of house prices (MEDV)

```
> hist(boston$MEDV, col='green',
+   main='', xlab='Census Tract Median House Price')
```



Multiple Regression Model

We model the median price in a census tract (y_i = median price in ith tract) as a linear function of multiple predictors, plus some error.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{13} x_{i13} + \epsilon_i$$

	β_0	β_1	β_2	...	β_{13}	
		LAT	**LON**	**...**	**LSTAT**	**error**
y_1	1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,13}$	ϵ_1
y_2	1	$x_{2,1}$	$x_{2,2}$	...	$x_{2,13}$	ϵ_2
...	...	...	...	...	...	...

Multiple Regression Model

We model the median price in a census tract (y_i = median price in ith tract) as a linear function of multiple predictors, plus some error.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{13} x_{i13} + \epsilon_i$$

	β_0	β_1	β_2	...	β_{13}	
		LAT	**LON**	**...**	**LSTAT**	**error**
y_1	1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,13}$	ϵ_1
y_2	1	$x_{2,1}$	$x_{2,2}$	...	$x_{2,13}$	ϵ_2
...	...	...	...	...	...	...

We find $\hat{\beta}_0, \ldots, \hat{\beta}_{13}$ to minimize the residuals $(y_i - \hat{y}_i)$

```
> model <- lm(MEDV ~ LON+LAT+CRIME+ZONE+INDUS+NOX+ROOM+AGE+DIST
+                   +RADIAL+TAX+PTRATIO+LSTAT, data=boston)
> summary(model$residuals)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-258106  -57337  -13642       0   39614  531268

> summary(model)$r.squared

[1] 0.7305487

> summary(model)$adj.r.squared

[1] 0.7234291
```

This is a high R^2 compared to the prior examples!

Keep an eye on the Adjusted-R^2...

Here is how the predictors contribute to the estimation:

```
> round(summary(model)$coefficients,3)

                  Estimate  Std. Error t value Pr(>|t|)
(Intercept) -10815106.841 6202196.194  -1.744    0.082
LON           -100538.327   68540.103  -1.467    0.143
LAT            105813.832   75439.531   1.403    0.161
CRIME           -2497.915     665.762  -3.752    0.000
ZONE              920.725     282.649   3.257    0.001
INDUS             447.859    1267.151   0.353    0.724
NOX           -320021.023   82010.164  -3.902    0.000
ROOM            72906.394    8529.875   8.547    0.000
AGE               167.225     273.353   0.612    0.541
DIST           -27489.610    4295.791  -6.399    0.000
RADIAL           6274.465    1362.945   4.604    0.000
TAX              -286.853      76.087  -3.770    0.000
PTRATIO        -18304.240    2801.930  -6.533    0.000
LSTAT          -11416.450    1022.127 -11.169    0.000
```

Here is how the predictors contribute to the estimation:

```
> round(summary(model)$coefficients,3)

                Estimate  Std. Error t value Pr(>|t|)
(Intercept) -10815106.841 6202196.194  -1.744    0.082
LON           -100538.327   68540.103  -1.467    0.143
LAT            105813.832   75439.531   1.403    0.161
CRIME           -2497.915     665.762  -3.752    0.000
ZONE              920.725     282.649   3.257    0.001
INDUS             447.859    1267.151   0.353    0.724
NOX           -320021.023   82010.164  -3.902    0.000
ROOM            72906.394    8529.875   8.547    0.000
AGE               167.225     273.353   0.612    0.541
DIST           -27489.610    4295.791  -6.399    0.000
RADIAL           6274.465    1362.945   4.604    0.000
TAX              -286.853      76.087  -3.770    0.000
PTRATIO        -18304.240    2801.930  -6.533    0.000
LSTAT          -11416.450    1022.127 -11.169    0.000
```

Intercept, INDUS, AGE, LAT and LON seem to be statistically insignificant. Should we omit them altogether?

Let’s interpret the ROOM coefficient 72906.39. The interpretation is:

- *Holding the other predictors (X variables) constant*, our predicted home price increases by \$72906.39 for each additional room.

Let's interpret the ROOM coefficient 72906.39. The interpretation is:

- *Holding the other predictors (X variables) constant*, our predicted home price increases by $72906.39 for each additional room.
- Two census tracts that are otherwise identical (but where one census tract has 1 more room on average than the other) will differ by $72906.39 in their predicted home price.

A p-value of predictor i tests the null hypothesis that $\beta_i = 0$; i.e., that predictor i has no contribution to predicting Y independent above and beyond the other predictors

Omitting other predictors might increase the significance (decrease the p-value) of a statistically insignificant predictor.

```
> model_red <- lm(MEDV ~ LON+LAT+INDUS+AGE, data=boston)
> round(summary(model_red)$coefficients,3)

                 Estimate  Std. Error t value Pr(>|t|)
(Intercept) -54327833.818 8559058.342  -6.347    0.000
LON           -709317.003   92858.638  -7.639    0.000
LAT            107180.101  111629.613   0.960    0.337
INDUS          -11817.533    1305.467  -9.052    0.000
AGE              -235.769     324.422  -0.727    0.468

> summary(model_red)$r.squared

[1] 0.3203884
```

LON and INDUS look like a big deal now, although they do not explain as much with $R^2 = 0.32$.

Let’s start omiting one by one.

INDUS has been omitted.

```
> model <- lm(MEDV ~ LON+LAT+CRIME+ZONE+NOX+ROOM+AGE+DIST
+                    +RADIAL+TAX+PTRATIO+LSTAT, data=boston)
> summary(model)$r.squared

[1] 0.7304803

> summary(model)$adj.r.squared

[1] 0.72392
```

R^2 has not changed too much, Adjusted-R^2 has increased a bit.

```
> round(summary(model)$coefficients,3)

                Estimate  Std. Error t value Pr(>|t|)
(Intercept) -11078358.838 6151842.621  -1.801    0.072
LON           -104687.251   67467.423  -1.552    0.121
LAT            104977.262   75335.440   1.393    0.164
CRIME           -2504.206     664.933  -3.766    0.000
ZONE              907.905     280.062   3.242    0.001
NOX           -311362.657   78196.317  -3.982    0.000
ROOM            72586.523    8474.196   8.566    0.000
AGE               170.953     272.907   0.626    0.531
DIST           -27725.142    4240.019  -6.539    0.000
RADIAL           6136.632    1304.802   4.703    0.000
TAX              -275.379      68.753  -4.005    0.000
PTRATIO        -18137.431    2759.443  -6.573    0.000
LSTAT          -11391.407    1018.762 -11.182    0.000
```

AGE still seems insignificant.

AGE has been omitted.

```
> model <- lm(MEDV ~ LON+LAT+CRIME+ZONE+NOX+ROOM+DIST
+                   +RADIAL+TAX+PTRATIO+LSTAT, data=boston)
> summary(model)$r.squared

[1] 0.7302658

> summary(model)$adj.r.squared

[1] 0.7242596
```

R^2 is again about the same, and Adjusted-R^2 has increased a bit.

```
> round(summary(model)$coefficients,3)

                Estimate  Std. Error t value Pr(>|t|)
(Intercept) -10647181.257 6109452.440  -1.743    0.082
LON            -97363.810   66405.890  -1.466    0.143
LAT            107052.256   75216.281   1.423    0.155
CRIME           -2512.856     664.380  -3.782    0.000
ZONE              891.239     278.624   3.199    0.001
NOX           -300532.234   76214.057  -3.943    0.000
ROOM            73744.325    8265.087   8.922    0.000
DIST           -28594.368    4004.063  -7.141    0.000
RADIAL           6088.931    1301.777   4.677    0.000
TAX              -273.672      68.657  -3.986    0.000
PTRATIO        -18104.114    2757.233  -6.566    0.000
LSTAT          -11177.763     959.387 -11.651    0.000
```

LAT is next.

LAT has been omitted.

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

[1] 0.7291597

> summary(model)$adj.r.squared

[1] 0.7236882
```

Both R^2 and Adjusted-R^2 have reduced. But still not too bad.

```
> round(summary(model)$coefficients,3)

                Estimate  Std. Error t value Pr(>|t|)
(Intercept) -5072210.579 4693368.926  -1.081    0.280
LON           -82749.985   65675.201  -1.260    0.208
CRIME          -2507.401     665.057  -3.770    0.000
ZONE             874.164     278.654   3.137    0.002
NOX          -318434.670   75246.742  -4.232    0.000
ROOM           73594.882    8272.978   8.896    0.000
DIST          -29692.314    3933.116  -7.549    0.000
RADIAL          5853.970    1292.603   4.529    0.000
TAX             -271.753      68.714  -3.955    0.000
PTRATIO       -18211.873    2759.048  -6.601    0.000
LSTAT         -11062.388     956.946 -11.560    0.000
```

Bye LON...

LON has been omitted.

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

[1] 0.7282911

> summary(model)$adj.r.squared

[1] 0.7233609
```

Both R^2 and Adjusted-R^2 have reduced. But that's OK.

```
> round(summary(model)$coefficients,3)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  840065.150  99001.032   8.485    0.000
CRIME         -2566.084    663.817  -3.866    0.000
ZONE            921.998    276.220   3.338    0.001
NOX         -346925.672  71811.323  -4.831    0.000
ROOM          74242.520   8261.884   8.986    0.000
DIST         -31049.529   3784.980  -8.203    0.000
RADIAL         6000.243   1288.142   4.658    0.000
TAX            -265.331     68.566  -3.870    0.000
PTRATIO      -19279.752   2627.204  -7.339    0.000
LSTAT        -11071.731    957.483 -11.563    0.000
```

Notice what happened to the intercept. LON (and perhaps the others) was acting like an intercept!

When to omit, when to keep?

It is usually good to omit statistically insignificant variables, because:

- The model gets simpler
- Insignificant variables may lead to incorrect interpretations (as in LON)
- When the data set is small, we can read too much into the impact of insignificant variables

When to omit, when to keep?

We keep a variable in the model, even if it is statistically insignificant, when:

- We are testing a hypothesis on the variable
- The variable has a big effect, although it is statistically insignificant
- It is an expected control variable (e.g. age in medical studies, race in sociological studies etc.)
- It is included in a higher order term (more on this later)

“Most important” predictors

How to identify which predictors have “more significant” effect on the response?

Parameter estimate?

“Most important” predictors

How to identify which predictors have “more significant” effect on the response?

- Parameter estimate?
- *p*-value?

“Most important” predictors

How to identify which predictors have “more significant” effect on the response?

- Parameter estimate?
- *p*-value?
- t score?

"Most important" predictors

How to identify which predictors have "more significant" effect on the response?

- Parameter estimate?
- *p*-value?
- t score? ✓

Which ones seem to be the most important?

```
> round(summary(model)$coefficients,3)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)  840065.150  99001.032   8.485    0.000
CRIME         -2566.084    663.817  -3.866    0.000
ZONE            921.998    276.220   3.338    0.001
NOX         -346925.672  71811.323  -4.831    0.000
ROOM          74242.520   8261.884   8.986    0.000
DIST         -31049.529   3784.980  -8.203    0.000
RADIAL         6000.243   1288.142   4.658    0.000
TAX            -265.331     68.566  -3.870    0.000
PTRATIO      -19279.752   2627.204  -7.339    0.000
LSTAT        -11071.731    957.483 -11.563    0.000
```