THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Multiple Regression

Lecture 11

STA 371G

How do you know how much to pay for a house?

How do you know how much to pay for a house?

Zillow? How do they know?

How do you know how much to pay for a house?

Zillow? How do they know?

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

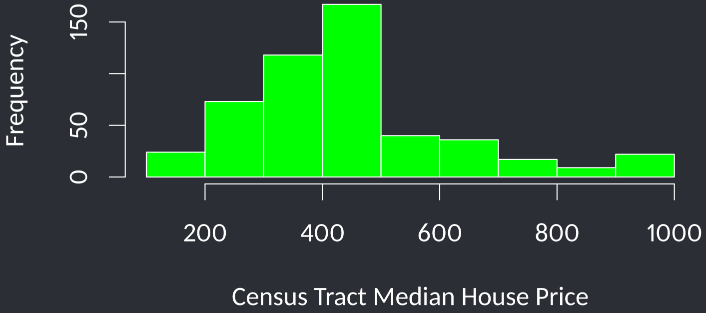
Boston house price data (by census tract, 1970)

- 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.
-258.10   -57.34   -13.64     0.00    39.61   531.30

> 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) -10815.107   6202.196  -1.744    0.082
LON           -100.538     68.540  -1.467    0.143
LAT            105.814     75.440   1.403    0.161
CRIME           -2.498      0.666  -3.752    0.000
ZONE             0.921      0.283   3.257    0.001
INDUS            0.448      1.267   0.353    0.724
NOX           -320.021     82.010  -3.902    0.000
ROOM            72.906      8.530   8.547    0.000
AGE              0.167      0.273   0.612    0.541
DIST           -27.490      4.296  -6.399    0.000
RADIAL           6.274      1.363   4.604    0.000
TAX             -0.287      0.076  -3.770    0.000
PTRATIO        -18.304      2.802  -6.533    0.000
LSTAT          -11.416      1.022 -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) -10815.107   6202.196  -1.744    0.082
LON           -100.538     68.540  -1.467    0.143
LAT            105.814     75.440   1.403    0.161
CRIME           -2.498      0.666  -3.752    0.000
ZONE             0.921      0.283   3.257    0.001
INDUS            0.448      1.267   0.353    0.724
NOX           -320.021     82.010  -3.902    0.000
ROOM            72.906      8.530   8.547    0.000
AGE              0.167      0.273   0.612    0.541
DIST           -27.490      4.296  -6.399    0.000
RADIAL           6.274      1.363   4.604    0.000
TAX             -0.287      0.076  -3.770    0.000
PTRATIO        -18.304      2.802  -6.533    0.000
LSTAT          -11.416      1.022 -11.169    0.000
```

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

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) -54327.834   8559.058  -6.347    0.000
LON           -709.317     92.859  -7.639    0.000
LAT            107.180    111.630   0.960    0.337
INDUS          -11.818      1.305  -9.052    0.000
AGE             -0.236      0.324  -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) -11078.359   6151.843  -1.801    0.072
LON           -104.687     67.467  -1.552    0.121
LAT            104.977     75.335   1.393    0.164
CRIME           -2.504      0.665  -3.766    0.000
ZONE             0.908      0.280   3.242    0.001
NOX           -311.363     78.196  -3.982    0.000
ROOM            72.587      8.474   8.566    0.000
AGE              0.171      0.273   0.626    0.531
DIST           -27.725      4.240  -6.539    0.000
RADIAL           6.137      1.305   4.703    0.000
TAX             -0.275      0.069  -4.005    0.000
PTRATIO        -18.137      2.759  -6.573    0.000
LSTAT          -11.391      1.019 -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) -10647.181   6109.452  -1.743    0.082
LON            -97.364     66.406  -1.466    0.143
LAT            107.052     75.216   1.423    0.155
CRIME           -2.513      0.664  -3.782    0.000
ZONE             0.891      0.279   3.199    0.001
NOX           -300.532     76.214  -3.943    0.000
ROOM            73.744      8.265   8.922    0.000
DIST           -28.594      4.004  -7.141    0.000
RADIAL           6.089      1.302   4.677    0.000
TAX             -0.274      0.069  -3.986    0.000
PTRATIO        -18.104      2.757  -6.566    0.000
LSTAT          -11.178      0.959 -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) -5072.211   4693.369  -1.081    0.280
LON           -82.750     65.675  -1.260    0.208
CRIME          -2.507      0.665  -3.770    0.000
ZONE            0.874      0.279   3.137    0.002
NOX          -318.435     75.247  -4.232    0.000
ROOM           73.595      8.273   8.896    0.000
DIST          -29.692      3.933  -7.549    0.000
RADIAL          5.854      1.293   4.529    0.000
TAX            -0.272      0.069  -3.955    0.000
PTRATIO       -18.212      2.759  -6.601    0.000
LSTAT         -11.062      0.957 -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)  840.065     99.001   8.485    0.000
CRIME         -2.566      0.664  -3.866    0.000
ZONE           0.922      0.276   3.338    0.001
NOX         -346.926     71.811  -4.831    0.000
ROOM          74.243      8.262   8.986    0.000
DIST         -31.050      3.785  -8.203    0.000
RADIAL         6.000      1.288   4.658    0.000
TAX           -0.265      0.069  -3.870    0.000
PTRATIO      -19.280      2.627  -7.339    0.000
LSTAT        -11.072      0.957 -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)   840.065     99.001   8.485    0.000
CRIME          -2.566      0.664  -3.866    0.000
ZONE            0.922      0.276   3.338    0.001
NOX          -346.926     71.811  -4.831    0.000
ROOM           74.243      8.262   8.986    0.000
DIST          -31.050      3.785  -8.203    0.000
RADIAL          6.000      1.288   4.658    0.000
TAX            -0.265      0.069  -3.870    0.000
PTRATIO       -19.280      2.627  -7.339    0.000
LSTAT         -11.072      0.957 -11.563    0.000
```

- Reminder to keep up with the readings in Perusall
- The readings often have technical discussions (e.g., matrix algebra, ANOVA tables) that you don’t need to worry about (we’ll talk about it in class if you need to know it)