Introduction to Social Data Analytics

Class 11

Today: Regression in Stata

By the end of today's lecture, you should be able to:

- Conduct basic regression analysis in Stata using `reg`
- Explain why one must be careful with linear form assumptions and out of sample extrapolation
- Distinguish causal effects from correlations between variables, and describe how naive regression is useful
- Analyze regression results and interpret key elements such as coefficient estimates and variance
- Construct a best fit line in a scatterplot and identify the slope, intercept, and residuals

Suppose we want to know how Y varies with X in a population.

We might model the outcome, Y, as a function of the predictor, X:

$$Y_i = f(X_i)$$

If we assume the relationship is linear, we can write our model as:

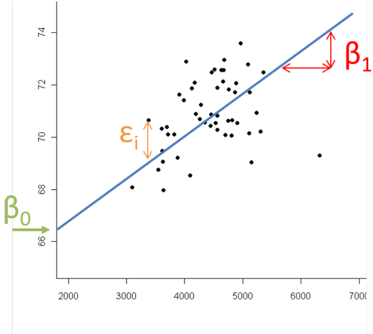
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Our job is to estimate β_0 and β_1 using data.

Let's break down the linear model.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Y-value = Intercept + Slope * X-value + error



We use 'hats' to denote coefficient estimates.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Often we don't have data for the entire population, so we cannot calculate the exact population parameters β_0 and β_1.

We use 'hats' to denote coefficient estimates.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Often we don't have data for the entire population, so we cannot calculate the exact population parameters β_0 and β_1.

Instead, we use a representative sample to estimate them:

$\hat{\beta}_0$: estimate of the intercept, β_0
$\hat{\beta}_1$: estimate of the coefficient, β_1

We use ‘hats’ to denote coefficient estimates.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Often we don’t have data for the entire population, so we cannot calculate the exact population parameters β_0 and β_1.

Instead, we use a representative sample to estimate them:

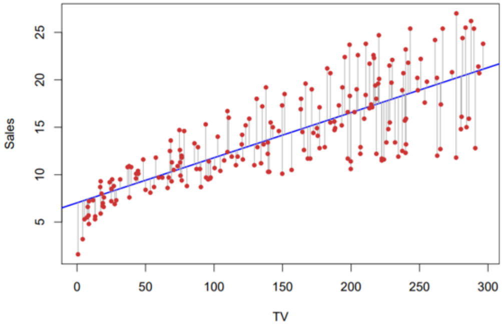
$\hat{\beta}_0$: estimate of the intercept, β_0
$\hat{\beta}_1$: estimate of the coefficient, β_1

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the parameter estimates that best fit the sample data.

How do we estimate β_0 and β_1?

Primary tool: Ordinary Least Squares (OLS)

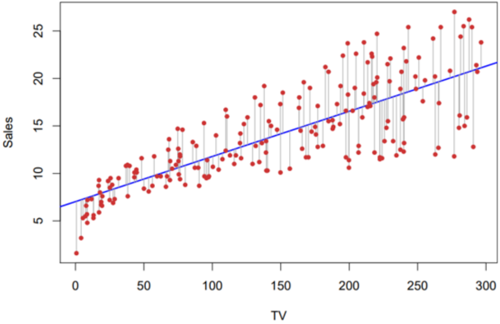
Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of ε_i^2.



How do we estimate β_0 and β_1?

Primary tool: Ordinary Least Squares (OLS)

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of ε_i^2.



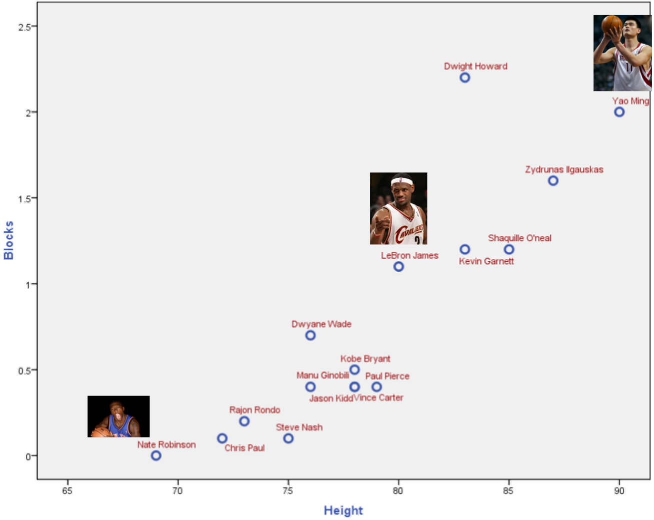
Computing $\hat{\beta}_0$ and $\hat{\beta}_1$ manually can take a very long time...but regression in Stata takes only a few seconds!

We might ask...

How are height (X) and average blocks per game (Y) for NBA players related?

Name	height	weight	age	rebound	blocks
Nate Robinson	[illegible]	180	23	3.1	0
Chris Paul	72	[illegible]	22	4	0.1
Rajon Rondo	73	[illegible]	[illegible]	4.2	0.2
Steve Nash	75	178	[illegible]	[illegible]	0.1
Dwyane Wade	76	216	25	[illegible]	0.7
Jason Kidd	76	210	34	6.5	[illegible]
Vince Carter	78	220	30	6	0.4
Kobe Bryant	78	205	29	6.3	0.5
Manu Ginobili	78	205	30	4.8	0.4
Paul Pierce	79	235	30	5.1	0.4
LeBron James	80	250	23	7.9	[illegible]
Dwight Howard	83	265	22	14.2	[illegible]
Kevin Garnett	83	253	31	9.2	[illegible]
Shaquille O'neal	85	325	35	10.6	1.2
Zydrunas Ilgauskas	87	260	32	[illegible]	1.6
Yao Ming	90	310	27	[illegible]	2

Scatter plot of the data

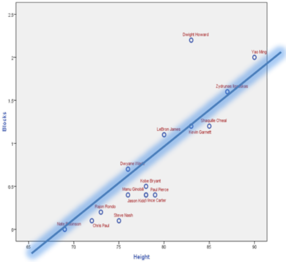


Regression *can* tell us...

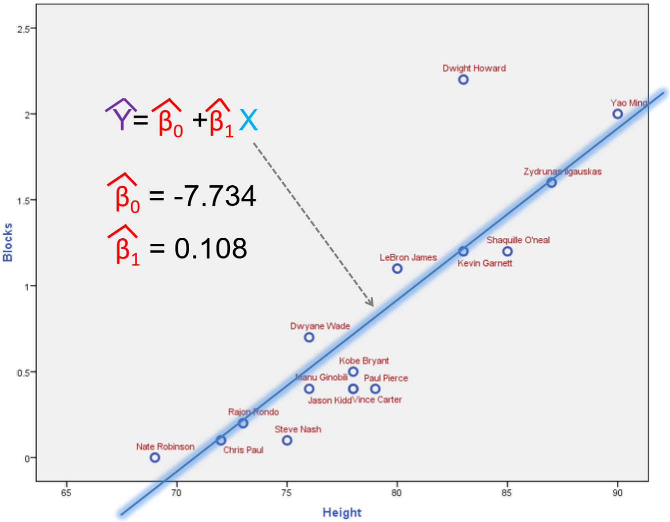
the ‘best fit’ line for the data

the equation for that line

a *predicted* Y for any value of X within the population represented



Linear regression allows us to construct a line of best fit



How do we interpret the slope estimate, $\hat{\beta}_1$?

[Y] changes by [$\hat{\beta}_1$] [Y units] for every one [X unit] increase in [X]...

...on average, all else equal.

In our NBA example:

- $\hat{\beta}_1 = 0.108$
- Y = "number of blocks per game"
- X = "height in inches"

How do we interpret the slope estimate, $\hat{\beta}_1$?

[Y] changes by [$\hat{\beta}_1$] [Y units] for every one [X unit] increase in [X]...

...on average, all else equal.

In our NBA example:

- $\hat{\beta}_1 = 0.108$
- Y = "number of blocks per game"
- X = "height in inches"

"An NBA player's blocks per game increases by 0.108 blocks for every one inch increase in height on average, all else equal."

How do we interpret the intercept estimate, $\hat{\beta}_0$?

When [X] is zero [X units], [Y] is [$\hat{\beta}_0$] [Y units]...

...on average, all else equal.

In our NBA example:

- $\hat{\beta}_0$ = -7.734
- Y = "number of blocks per game"
- X = "height in inches"

How do we interpret the intercept estimate, $\hat{\beta}_0$?

When [X] is zero [X units], [Y] is [$\hat{\beta}_0$] [Y units]...

...on average, all else equal.

In our NBA example:

- $\hat{\beta}_0$ = -7.734
- Y = "number of blocks per game"
- X = "height in inches"

"When an NBA player's height is zero inches, his blocks per game is -7.734 blocks on average, all else equal."

If Tony Parker is 74" tall, what is his predicted blocks/game?

$$\hat{Y}_i = -7.734 + 0.108X_i$$

If Tony Parker is 74" tall, what is his predicted blocks/game?

$$\hat{Y}_i = -7.734 + 0.108X_i$$
$$\hat{Y}_i = -7.734 + 0.108(74) = 0.258$$

If Tony Parker is 74" tall, what is his predicted blocks/game?

$$\hat{Y}_i = -7.734 + 0.108X_i$$
$$\hat{Y}_i = -7.734 + 0.108(74) = 0.258$$

If his actual blocks/game average was 0.1, what's the model error (residual)?

If Tony Parker is 74" tall, what is his predicted blocks/game?

$\hat{Y}_i = -7.734 + 0.108 X_i$
$\hat{Y}_i = -7.734 + 0.108(74) = 0.258$

If his actual blocks/game average was 0.1, what's the model error (residual)?

$Error_i = Actual_i - Predicted_i$
$\varepsilon_i = Y_i - \hat{Y}_i$

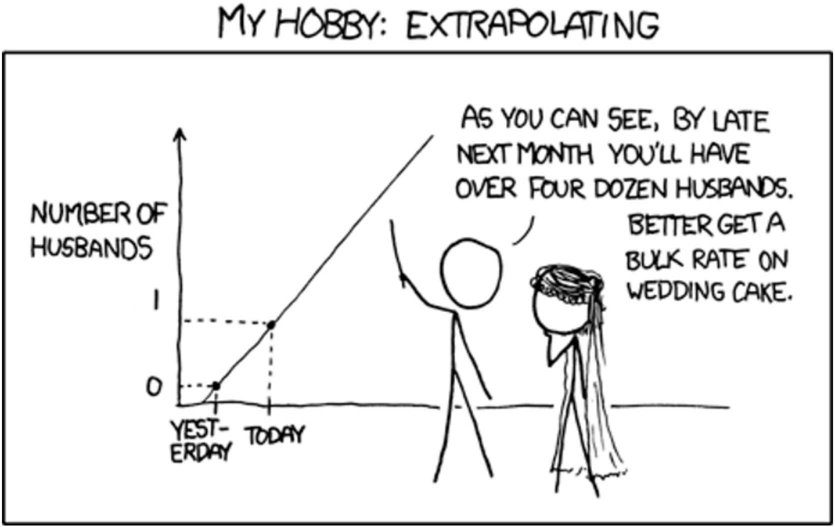
If Tony Parker is 74" tall, what is his predicted blocks/game?

$\hat{Y}_i = -7.734 + 0.108X_i$
$\hat{Y}_i = -7.734 + 0.108(74) = 0.258$

If his actual blocks/game average was 0.1, what's the model error (residual)?

$Error_i = Actual_i - Predicted_i$
$\varepsilon_i = Y_i - \hat{Y}_i$
$\varepsilon_i = 0.1 - 0.258 = -0.158$

A cautionary tale: out-of-sample extrapolation



Can we (accurately) do the following with our model:

- Predict blocks for an NBA player 68 inches in height or shorter?

Can we (accurately) do the following with our model:

- Predict blocks for an NBA player 68 inches in height or shorter?

  No, 68 inches is outside the domain of our data.

Can we (accurately) do the following with our model:

- Predict blocks for an NBA player 68 inches in height or shorter?

  No, 68 inches is outside the domain of our data.

- Predict blocks for a *college* basketball player of 75 inches?

Can we (accurately) do the following with our model:

- Predict blocks for an NBA player 68 inches in height or shorter?

  No, 68 inches is outside the domain of our data.

- Predict blocks for a *college* basketball player of 75 inches?

  No, our results are valid only for NBA players.

Can we (accurately) do the following with our model:

- Predict blocks for an NBA player 68 inches in height or shorter?

  No, 68 inches is outside the domain of our data.

- Predict blocks for a *college* basketball player of 75 inches?

  No, our results are valid only for NBA players.

- Predict how many blocks an NBA player would get if he wore shoes that raised his height by 5 inches?

Can we (accurately) do the following with our model:

- Predict blocks for an NBA player 68 inches in height or shorter?

  No, 68 inches is outside the domain of our data.

- Predict blocks for a *college* basketball player of 75 inches?

  No, our results are valid only for NBA players.

- Predict how many blocks an NBA player would get if he wore shoes that raised his height by 5 inches?

  No, our model estimates apply only to *natural* height.

Can we (accurately) do the following with our model:

- Predict blocks for an NBA player 68 inches in height or shorter?

  No, 68 inches is outside the domain of our data.

- Predict blocks for a *college* basketball player of 75 inches?

  No, our results are valid only for NBA players.

- Predict how many blocks an NBA player would get if he wore shoes that raised his height by 5 inches?

  No, our model estimates apply only to *natural* height.

Be careful with extrapolation and external validity!

Open class11.do in Stata

```
. regress blocks height

      Source |       SS           df       MS      Number of obs   =        16
-------------+----------------------------------   F(1, 14)        =     47.94
       Model |  5.54515291         1  5.54515291   Prob > F        =    0.0000
    Residual |  1.61922234        14  .115658738   R-squared       =    0.7740
-------------+----------------------------------   Adj R-squared   =    0.7578
       Total |  7.16437525        15  .477625016   Root MSE        =    .34009

------------------------------------------------------------------------------
      blocks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .1079611   .0155919     6.92   0.000     .0745198    .1414025
       _cons |  -7.734183   1.232749    -6.27   0.000    -10.37817     -5.0902
------------------------------------------------------------------------------
```

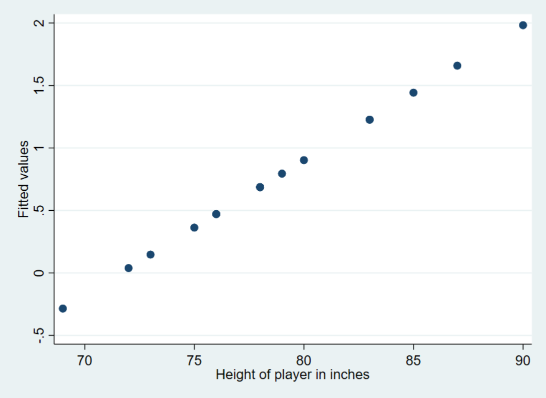
Open class11.do in Stata

```
. regress blocks height

      Source |       SS           df       MS      Number of obs   =        16
-------------+----------------------------------   F(1, 14)        =     47.94
       Model |  5.54515291         1  5.54515291   Prob > F        =    0.0000
    Residual |  1.61922234        14  .115658738   R-squared       =    0.7740
-------------+----------------------------------   Adj R-squared   =    0.7578
       Total |  7.16437525        15  .477625016   Root MSE        =    .34009

------------------------------------------------------------------------------
      blocks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .1079611   .0155919     6.92   0.000     .0745198    .1414025
       _cons |  -7.734183   1.232749    -6.27   0.000    -10.37817     -5.0902
------------------------------------------------------------------------------
```

- What does the Coef. tell us?
- What does the Std. Err. tell us?
- Are the coefficients statistically significant? Are they economically significant?

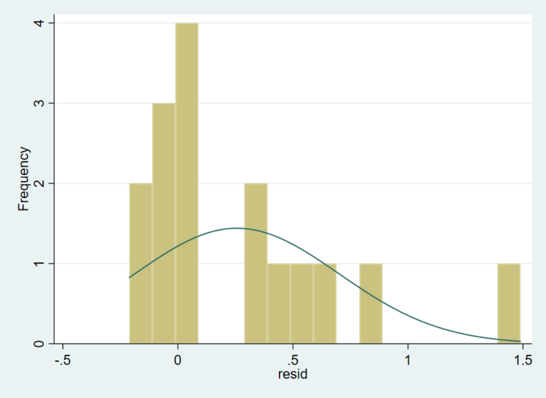
Generating predicted values

```
predict yhat
scatter yhat height
```

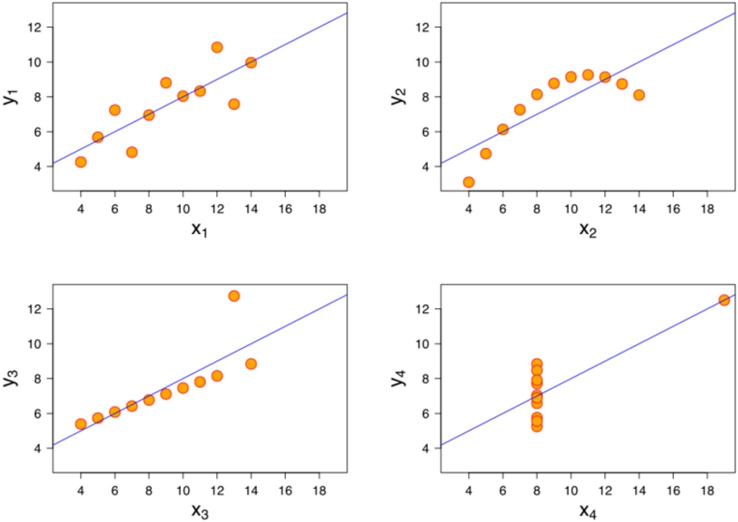


Generating residuals

```
predict resid, residuals
histogram resid, width(0.1) frequency normal
```



A bonus cautionary tale: Anscombe's Quartet



A bonus cautionary tale: Anscombe's Quartet

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

Le

learned: **Regression output does not tell the full story!**