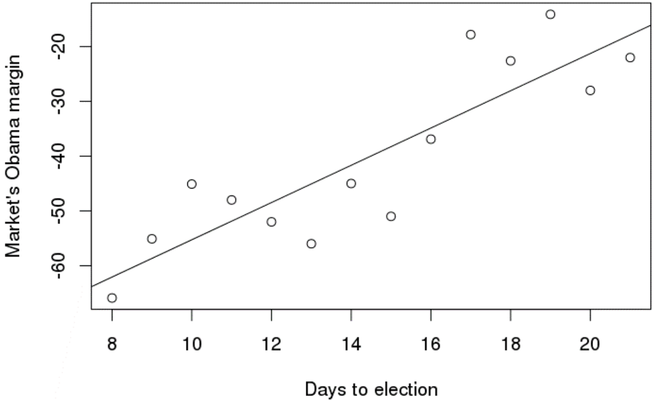
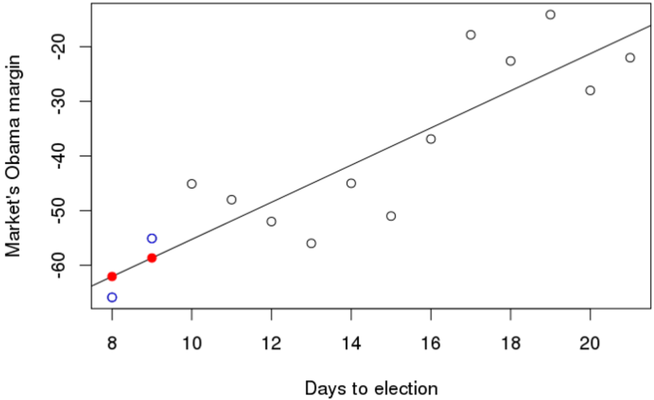
Prediction and Multivariate Regression

Within-Sample Prediction



Within-Sample Prediction



A Refresher: Out-of-Sample Prediction

- Fit a model: $Y_i = X_i$, using a dataset

A Refresher: Out-of-Sample Prediction

- Fit a model: $Y_i = X_i$, using a dataset
- Use the coefficients to predict $\hat{y}_i$ for a new observation i

A Refresher: Out-of-Sample Prediction

- Fit a model: $Y_i = X_i$, using a dataset
- Use the coefficients to predict $\hat{y}_i$ for a new observation i
- For instance, if I was interested in what would happen if my sleep $=$ 14, I could enter 14 into the model and obtain a predicted value $\hat{y}_i$

A Refresher: Out-of-Sample Prediction

- Fit a model: $Y_i = X_i$, using a dataset
- Use the coefficients to predict $\hat{y}_i$ for a new observation i
- For instance, if I was interested in what would happen if my sleep $= 14$, I could enter 14 into the model and obtain a predicted value $\hat{y}_i$
- Remember we **cannot** compare predicted value $\hat{y}_i$ to an observed value y_i

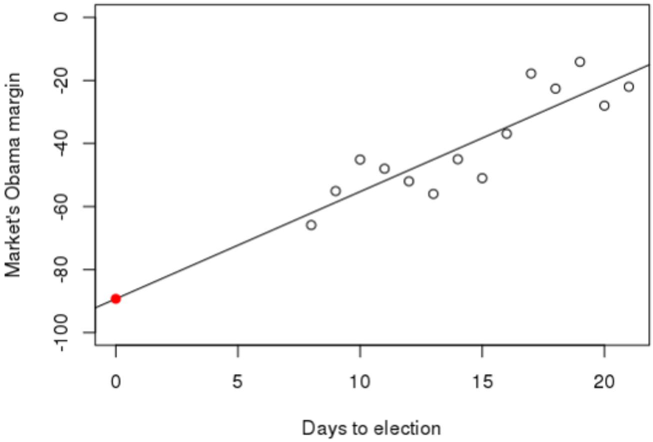
A Refresher: Out-of-Sample Prediction

- Fit a model: $Y_i = X_i$, using a dataset
- Use the coefficients to predict $\hat{y}_i$ for a new observation i
- For instance, if I was interested in what would happen if my sleep $=$ 14, I could enter 14 into the model and obtain a predicted value $\hat{y}_i$
- Remember we **cannot** compare predicted value $\hat{y}_i$ to an observed value y_i
- Useful for prediction beyond the range of existing data

A Refresher: Out-of-Sample Prediction

- Fit a model: $Y_i = X_i$, using a dataset
- Use the coefficients to predict $\hat{y}_i$ for a new observation i
- For instance, if I was interested in what would happen if my sleep $= 14$, I could enter 14 into the model and obtain a predicted value $\hat{y}_i$
- Remember we **cannot** compare predicted value $\hat{y}_i$ to an observed value y_i
- Useful for prediction beyond the range of existing data
  - For instance, it's 2 weeks before the election, and I want to predict the margin if DaysLeft=0

Out-of-Sample Prediction



Out-of-Sample Prediction as *forecasting*

- We can also predict outcomes for an entirely different set of data

Out-of-Sample Prediction as *forecasting*

- We can also predict outcomes for an entirely different set of data
- Let's say we've fit a regression on the average polling margin of victory (by state) the day before the election on actual outcomes (by state), using 2012 data:

Out-of-Sample Prediction as *forecasting*

- We can also predict outcomes for an entirely different set of data
- Let's say we've fit a regression on the average polling margin of victory (by state) the day before the election on actual outcomes (by state), using 2012 data:

Actual $Margin_{State\ i}$ = 0.1 + 1.01 * Average Polling $Margin_{State\ i}$

Out-of-Sample Prediction as *forecasting*

- We can also predict outcomes for an entirely different set of data
- Let's say we've fit a regression on the average polling margin of victory (by state) the day before the election on actual outcomes (by state), using 2012 data:

Actual $Margin_{State\ i} = 0.1 + 1.01$ * Average Polling $Margin_{State\ i}$

- We can predict the results in 2016 using what we know about the 2012 relationship

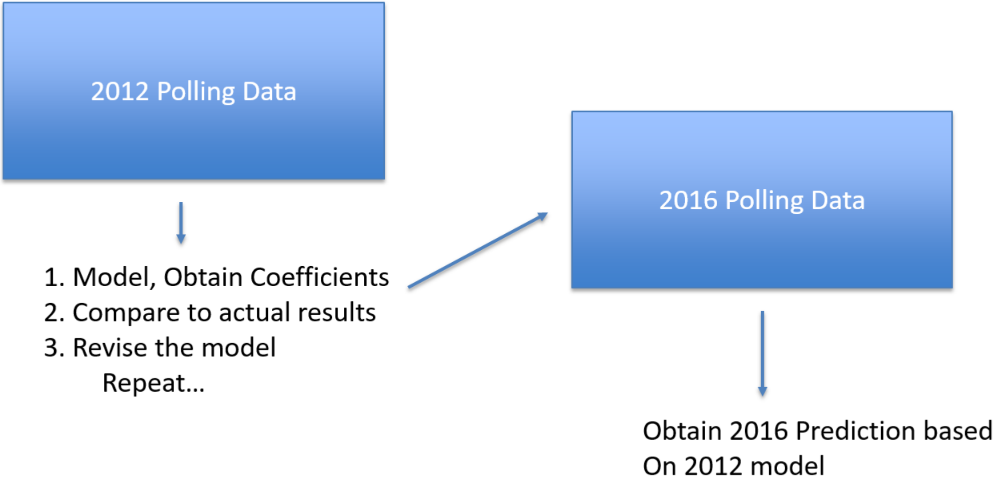
The Process of Out-of-Sample Prediction

2012 Polling Data

1. Model, Obtain Coefficients
2. Compare to actual results
3. Revise the model

   Repeat...

The Process of Out-of-Sample Prediction

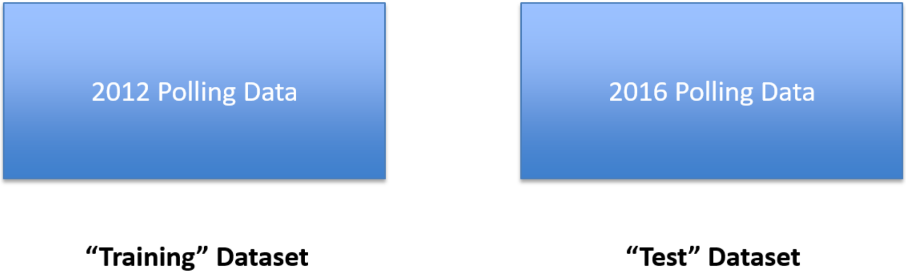


Out-of-Sample Prediction Assumptions

- Key Assumption: The relationship between the independent variable(s) and the dependent variable are the same in each dataset.

Out-of-Sample Prediction Assumptions

- Key Assumption: The relationship between the independent variable(s) and the dependent variable are the same in each dataset.



Multivariate Regression

Models of the world are nearly always more complex than a univariate model.

Multivariate Regression

Models of the world are nearly always more complex than a univariate model.

Consider elections:

Multivariate Regression

Models of the world are nearly always more complex than a univariate model.

Consider elections:

- Even the scion of prediction (Nate Silver) uses multiple data sources to make predictions

Multivariate Regression

Models of the world are nearly always more complex than a univariate model.

Consider elections:

- Even the scion of prediction (Nate Silver) uses multiple data sources to make predictions
- Single variable models don't offer sufficient control for confounds (unless applied to an experiment)

Multivariate Regression

Models of the world are nearly always more complex than a univariate model.

Consider elections:

- Even the scion of prediction (Nate Silver) uses multiple data sources to make predictions
- Single variable models don't offer sufficient control for confounds (unless applied to an experiment)

How do we predict an election from the following: polling margins, betting market margins, the state of the economy, and incumbency?

Multivariate Regression

Models of the world are nearly always more complex than a univariate model.

Consider elections:

- Even the scion of prediction (Nate Silver) uses multiple data sources to make predictions
- Single variable models don't offer sufficient control for confounds (unless applied to an experiment)

How do we predict an election from the following: polling margins, betting market margins, the state of the economy, and incumbency?

- With a model!

Specifiying a Multivariate Regression

$$Margin\,of\,Victory_i =$$
$$Poll\,Margin_i + Betting\,Margin_i + Economy_i + Incumbent_i$$

Specifiying a Multivariate Regression

$$Margin\,of\,Victory_i = \\ Poll\,Margin_i + Betting\,Margin_i + Economy_i + Incumbent_i$$

- In this model we are estimating a **single** intercept and **multiple** slopes.

Specifiying a Multivariate Regression

$$Margin\,of\,Victory_i = \\ Poll\,Margin_i + Betting\,Margin_i + Economy_i + Incumbent_i$$

- In this model we are estimating a **single** intercept and **multiple** slopes.

How does this work?

Specifiying a Multivariate Regression

$$Margin\,of\,Victory_i =$$
$$Poll\,Margin_i + Betting\,Margin_i + Economy_i + Incumbent_i$$

- In this model we are estimating a **single** intercept and **multiple** slopes.

How does this work?

- Let's start with a univariate model where we are interested in the relationship between an index of national income (X) and the strength of democracy (Y)

Specifiying a Multivariate Regression

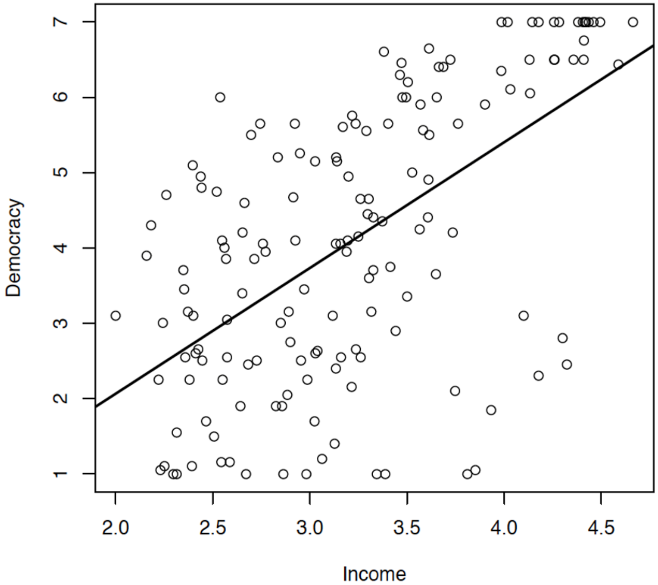
$$Margin\,of\,Victory_i = \\ Poll\,Margin_i + Betting\,Margin_i + Economy_i + Incumbent_i$$

- In this model we are estimating a **single** intercept and **multiple** slopes.

How does this work?

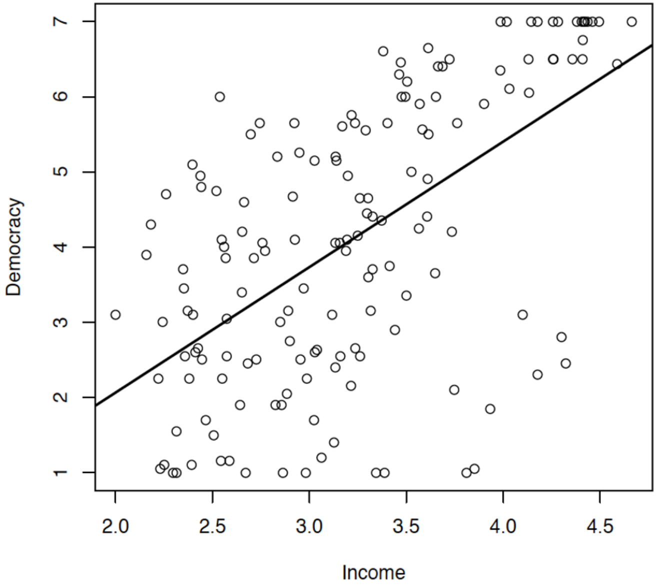
- Let's start with a univariate model where we are interested in the relationship between an index of national income (X) and the strength of democracy (Y)
- We specify the following model:
$Democracy\,Index_{Country\,i} = Income\,Index_{Country\,i}$

Visualizing a Multivariate Regression



Democracy
= -1.26 + 1.6 (Income)

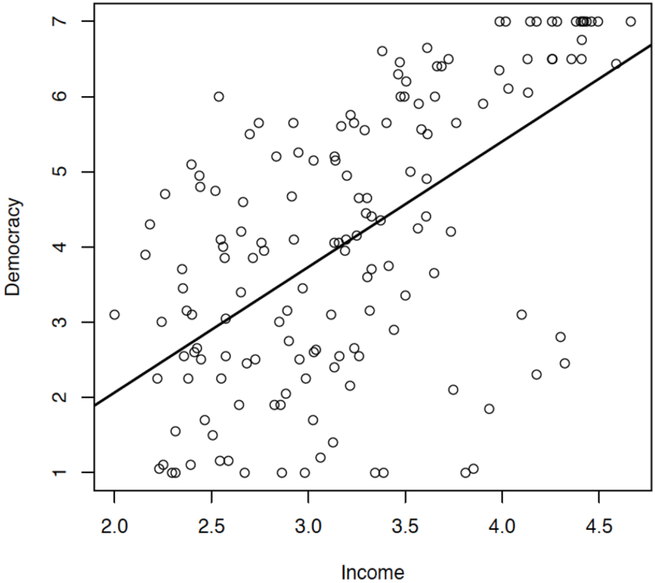
Visualizing a Multivariate Regression



Democracy
= -1.26 + 1.6 (Income)

Not a great fit;
We have more information on countries that we could use to predict democratic strength

Visualizing a Multivariate Regression

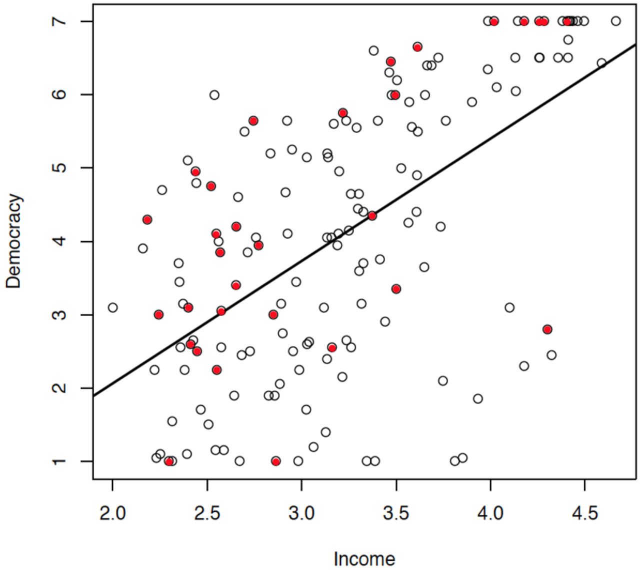


Democracy
= -1.26 + 1.6 (Income)

Not a great fit;
We have more information on countries that we could use to predict democratic strength

For instance, countries that were British colonies may be more likely to have strong democracies.

Visualizing a Multivariate Regression



Democracy

= -1.26 + 1.6 (Income)

Not a great fit;
We have more information on countries that we could use to predict democratic strength

For instance, countries that were British colonies may be more likely to have strong democracies.

We want to "control" for this

Income (X) and Democratic Strength (Y)

- How do we calculate the effect of Income (X) on Democratic Strength (Y), taking British Colonial Status into account?

Income (X) and Democratic Strength (Y)

- How do we calculate the effect of Income (X) on Democratic Strength (Y), taking British Colonial Status into account?
  - We must add a second independent variable to our model! We will call this second variable for Colonial Status (Z)

Income (X) and Democratic Strength (Y)

- How do we calculate the effect of Income (X) on Democratic Strength (Y), taking British Colonial Status into account?
  - We must add a second independent variable to our model! We will call this second variable for Colonial Status (Z)
  - Now when estimating Democratic strength we will have two independent variables to the right in the equation.

Income (X) and Democratic Strength (Y)

- How do we calculate the effect of Income (X) on Democratic Strength (Y), taking British Colonial Status into account?
    - We must add a second independent variable to our model! We will call this second variable for Colonial Status (Z)
    - Now when estimating Democratic strength we will have two independent variables to the right in the equation.

Univariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\mu}_i$

Income (X) and Democratic Strength (Y)

- How do we calculate the effect of Income (X) on Democratic Strength (Y), taking British Colonial Status into account?
  - We must add a second independent variable to our model! We will call this second variable for Colonial Status (Z)
  - Now when estimating Democratic strength we will have two independent variables to the right in the equation.

Univariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\mu}_i$

Multivariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{\mu}_i$

The Intercept and Multivariate Regression

Univariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\mu}_i$

Multivariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{\mu}_i$
  - Note: we are dealing with two different kinds of independent variables: continuous (Income) and binary (colonial status).

The Intercept and Multivariate Regression

Univariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\mu}_i$

Multivariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{\mu}_i$
  - Note: we are dealing with two different kinds of independent variables: continuous (Income) and binary (colonial status).
- $\hat{\alpha}$ is part of both models, but it means different things

The Intercept and Multivariate Regression

Univariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\mu}_i$

Multivariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{\mu}_i$
  - Note: we are dealing with two different kinds of independent variables: continuous (Income) and binary (colonial status).
- $\hat{\alpha}$ is part of both models, but it means different things
  - In the univariate model, it is the estimated value of $\hat{Y}_i$ when X is zero.

The Intercept and Multivariate Regression

Univariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\mu}_i$

Multivariate:

- $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i + \hat{\mu}_i$
  - Note: we are dealing with two different kinds of independent variables: continuous (Income) and binary (colonial status).
- $\hat{\alpha}$ is part of both models, but it means different things
  - In the univariate model, it is the estimated value of $\hat{Y}_i$ when X is zero.
  - In the multivariate model, it is the estimated value of $\hat{Y}_i$ when X *and* Z are zero.

Binary Variables

- $\hat{y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 Z_1$

Binary Variables

- $\hat{y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 Z_1$

We run the model and get the following estimates for $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$:

- $\hat{y} = -1.5 + 1.7X_1 + .58Z_1$

Binary Variables

- $\hat{y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 Z_1$

We run the model and get the following estimates for $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$:

- $\hat{y} = -1.5 + 1.7X_1 + .58Z_1$

Non-British colonies: Assume income (X) of 3.5 and colonial status (z) of 0:

- $\hat{y} = -1.5 + 1.7(3.5) + .58(0)$

Binary Variables

- $\hat{y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 Z_1$

We run the model and get the following estimates for $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$:

- $\hat{y} = -1.5 + 1.7X_1 + .58Z_1$

Non-British colonies: Assume income (X) of 3.5 and colonial status (z) of 0:

- $\hat{y} = -1.5 + 1.7(3.5) + .58(0)$
- Estimated value of democracy: 4.45

Binary Variables

- $\hat{y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 Z_1$

We run the model and get the following estimates for $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$:

- $\hat{y} = -1.5 + 1.7X_1 + .58Z_1$

Non-British colonies: Assume income (X) of 3.5 and colonial status (z) of 0:

- $\hat{y} = -1.5 + 1.7(3.5) + .58(0)$
- Estimated value of democracy: 4.45

Former British colonies: Assume income (X) of 3.5 and colonial status (z) of 1:

- $\hat{y} = -1.5 + 1.7(3.5) + .58(1)$

Binary Variables

- $\hat{y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 Z_1$

We run the model and get the following estimates for $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$:

- $\hat{y} = -1.5 + 1.7X_1 + .58Z_1$

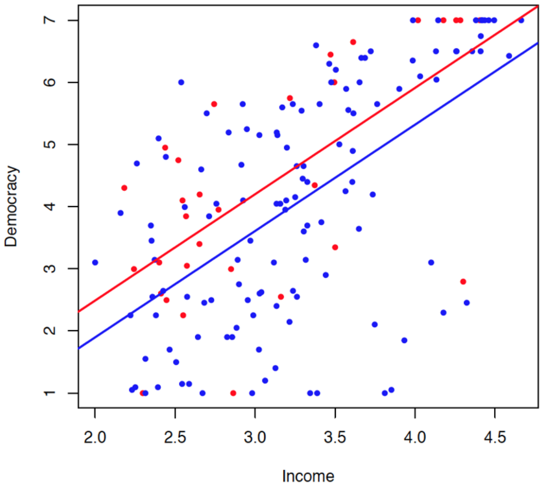
Non-British colonies: Assume income (X) of 3.5 and colonial status (z) of 0:

- $\hat{y} = -1.5 + 1.7(3.5) + .58(0)$
- Estimated value of democracy: 4.45

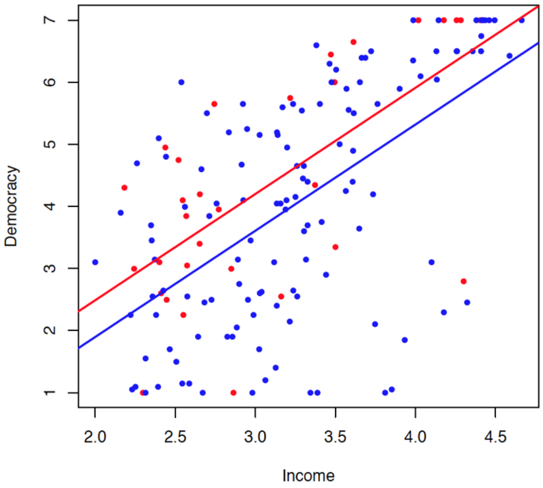
Former British colonies: Assume income (X) of 3.5 and colonial status (z) of 1:

- $\hat{y} = -1.5 + 1.7(3.5) + .58(1)$
- Estimated value of democracy: 5.03

Visualize



Visualize



We are fitting 2 lines with the same slope but different intercepts.

Binary Indpendent Variables

- The slope for a binary variable β will always be multiplied by *either* 0 or 1.

Binary Indpendent Variables

- The slope for a binary variable β will always be multiplied by *either* 0 or 1.
- Recall, for a continuous variable β will always be multiplied by the value of X_i

Multivariate Regression: Interpretation

Consider: $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + ... + u_i$

Multivariate Regression: Interpretation

Consider: $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + ... + u_i$

The following interpretation always holds:

Multivariate Regression: Interpretation

Consider: $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + ... + u_i$

The following interpretation always holds:

- α is the intercept (when $X = 0$ and $Z = 0$)

Multivariate Regression: Interpretation

Consider: $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + ... + u_i$

The following interpretation always holds:

- α is the intercept (when $X = 0$ and $Z = 0$)
- β_1 tells us how many units Y is expected to increase if X increases by one unit and Z does not change.

Multivariate Regression: Interpretation

Consider: $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + ... + u_i$

The following interpretation always holds:

- α is the intercept (when $X = 0$ and $Z = 0$)
- β_1 tells us how many units Y is expected to increase if X increases by one unit and Z does not change.
- β_2 tells us how many units Y is expected to increase if Z increases by one unit and X does not change.

Multivariate Regression: Interpretation

Consider: $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + ... + u_i$

The following interpretation always holds:

- α is the intercept (when $X = 0$ and $Z = 0$)
- β_1 tells us how many units Y is expected to increase if X increases by one unit and Z does not change.
- β_2 tells us how many units Y is expected to increase if Z increases by one unit and X does not change.
- Continues for any number of additional independent variables.

Examples of How to Interpret Multivariate Regression

Suicide Interpretation Example

Consider an example where we are predicting `suicides` (count) with two predictors: `percent_male` (0-100), and `percent_religious` (0-100).

Suicide Interpretation Example

Consider an example where we are predicting `suicides` (count) with two predictors: `percent_male` (0-100), and `percent_religious` (0-100).

	Estimate
(Intercept)	505.00
percent_male	21.01
percent_religious	-10.12

Suicide Interpretation Example

Consider an example where we are predicting `suicides` (count) with two predictors: `percent_male` (0-100), and `percent_religious` (0-100).

	Estimate
(Intercept)	505.00
percent_male	21.01
percent_religious	-10.12

How many suicides would we predict if a country's population is 52% male and 15% religious?

Suicide Interpretation Example

Consider an example where we are predicting `suicides` (count) with two predictors: `percent_male` (0-100), and `percent_religious` (0-100).

	Estimate
(Intercept)	505.00
percent_male	21.01
percent_religious	-10.12

How many suicides would we predict if a country's population is 52% male and 15% religious?

$$\hat{Y} = 505.00 + 21.01 * (52) - 10.12 * (15)$$

Suicide Interpretation Example

Consider an example where we are predicting `suicides` (count) with two predictors: `percent_male` (0-100), and `percent_religious` (0-100).

	Estimate
(Intercept)	505.00
percent_male	21.01
percent_religious	-10.12

How many suicides would we predict if a country's population is 52% male and 15% religious?

$$\hat{Y} = 505.00 + 21.01 * (52) - 10.12 * (15)$$

$$\hat{Y} = 1445.72$$

Retirement Interpretation Example

We are predicting `retirement_savings` (in dollars) with three predictors: `age` (0-100), `annual_income` (in dollars), and `years_of_education` (0-22)

Retirement Interpretation Example

We are predicting `retirement_savings` (in dollars) with three predictors: `age` (0-100), `annual_income` (in dollars), and `years_of_education` (0-22)

	Estimate
(Intercept)	89230.15
age	3671.00
annual_income	0.75
years_of_education	1482.00

Retirement Interpretation Example

We are predicting `retirement_savings` (in dollars) with three predictors: `age` (0-100), `annual_income` (in dollars), and `years_of_education` (0-22)

	Estimate
(Intercept)	89230.15
age	3671.00
annual_income	0.75
years_of_education	1482.00

What amount of retirement savings would we expect from someone who is 34, makes $51,981/year, and who has a BA (16 years of education)?

Retirement Interpretation Example

We are predicting `retirement_savings` (in dollars) with three predictors: `age` (0-100), `annual_income` (in dollars), and `years_of_education` (0-22)

	Estimate
(Intercept)	89230.15
age	3671.00
annual_income	0.75
years_of_education	1482.00

What amount of retirement savings would we expect from someone who is 34, makes $51,981/year, and who has a BA (16 years of education)?

$$\hat{Y} = 89230.15 + 3671.00 * (34) + 0.75 * (51981) + 1482 * (16)$$

Retirement Interpretation Example

We are predicting `retirement_savings` (in dollars) with three predictors: age (0-100), `annual_income` (in dollars), and `years_of_education` (0-22)

	Estimate
(Intercept)	89230.15
age	3671.00
annual_income	0.75
years_of_education	1482.00

What amount of retirement savings would we expect from someone who is 34, makes $51,981/year, and who has a BA (16 years of education)?

$$\hat{Y} = 89230.15 + 3671.00 * (34) + 0.75 * (51981) + 1482 * (16)$$

$$\hat{Y} = 276741.9$$

Life Expectancy Interpretation Example

We are predicting `life_expectancy` (in years) with four predictors: `body_mass_index` (0-??), `average_daily_steps`, `systolic_blood_pressure` (90-180), and `average_sleep_duration_hours` (0-24)

Life Expectancy Interpretation Example

We are predicting life_expectancy (in years) with four predictors: body_mass_index (0-??), average_daily_steps, systolic_blood_pressure (90-180), and average_sleep_duration_hours (0-24)

	Estimate
(Intercept)	30.88
body_mass_index	-0.005
average_daily_steps	0.01
systolic_blood_pressure	-0.0019
average_sleep_duration_hours	0.725

Life Expectancy Interpretation Example

We are predicting `life_expectancy` (in years) with four predictors: `body_mass_index` (0-??), `average_daily_steps`, `systolic_blood_pressure` (90-180), and `average_sleep_duration_hours` (0-24)

	Estimate
(Intercept)	30.88
body_mass_index	-0.005
average_daily_steps	0.01
systolic_blood_pressure	-0.0019
average_sleep_duration_hours	0.725

How long would we expect someone with the following profile to live: `body_mass_index` = 28, `average_daily_steps` = 4,000, `systolic_blood_pressure` = 140, and `average_sleep_duration_hours` = 7.2.

Life Expectancy Interpretation Example

We are predicting `life_expectancy` (in years) with four predictors: `body_mass_index` (0-??), `average_daily_steps`, `systolic_blood_pressure` (90-180), and `average_sleep_duration_hours` (0-24)

	Estimate
(Intercept)	30.88
body_mass_index	-0.005
average_daily_steps	0.01
systolic_blood_pressure	-0.0019
average_sleep_duration_hours	0.725

How long would we expect someone with the following profile to live: `body_mass_index` = 28, `average_daily_steps` = 4,000, `systolic_blood_pressure` = 140, and `average_sleep_duration_hours` = 7.2.

$$\hat{Y} = 30.88 + -0.005 * (28) + 0.01 * (4000) + -0.0019 * (140) + 0.725 * (7.2)$$

Life Expectancy Interpretation Example

We are predicting `life_expectancy` (in years) with four predictors: `body_mass_index` (0-??), `average_daily_steps`, `systolic_blood_pressure` (90-180), and `average_sleep_duration_hours` (0-24)

	Estimate
(Intercept)	30.88
body_mass_index	-0.005
average_daily_steps	0.01
systolic_blood_pressure	-0.0019
average_sleep_duration_hours	0.725

How long would we expect someone with the following profile to live: `body_mass_index` = 28, `average_daily_steps` = 4,000, `systolic_blood_pressure` = 140, and `average_sleep_duration_hours` = 7.2.

$$\hat{Y} = 30.88 + -0.005 * (28) + 0.01 * (4000) + -0.0019 * (140) + 0.725 * (7.2)$$

$$\hat{Y} = 75.69$$

Republican Vote Share Interpretation Example

We are predicting `republican_vote_share` (a proportion) with four predictors: `voter_turnout_percentage` (0-100), `public_approval_rate` (0-100), `incumbant` (0 or 1), and `campaign_spending_millions` (0-??)

Republican Vote Share Interpretation Example

We are predicting `republican_vote_share` (a proportion) with four predictors: `voter_turnout_percentage` (0-100), `public_approval_rate` (0-100), `incumbant` (0 or 1), and `campaign_spending_millions` (0-??)

	Estimate
(Intercept)	10.76
voter_turnout_percentage	-4.00
public_approval_rate	1.04
incumbent (binary)	10.32
campaign_spending_millions	.00001

Republican Vote Share Interpretation Example

We are predicting `republican_vote_share` (a proportion) with four predictors: `voter_turnout_percentage` (0-100), `public_approval_rate` (0-100), `incumbant` (0 or 1), and `campaign_spending_millions` (0-??)

	Estimate
(Intercept)	10.76
voter_turnout_percentage	-4.00
public_approval_rate	1.04
incumbent (binary)	10.32
campaign_spending_millions	.00001

What would we expect the republican_vote_share to be if `voter_turnout_percentage` =61, `public_approval_rate` =21, `incumbant` =1 and `campaign_spending_millions` = 24,234,909

Republican Vote Share Interpretation Example

We are predicting republican_vote_share (a proportion) with four predictors: voter_turnout_percentage (0-100), public_approval_rate (0-100), incumbant (0 or 1), and campaign_spending_millions (0-??)

	Estimate
(Intercept)	10.76
voter_turnout_percentage	-4.00
public_approval_rate	1.04
incumbent (binary)	10.32
campaign_spending_millions	.00001

What would we expect the republican_vote_share to be if voter_turnout_percentage =61, public_approval_rate =21, incumbant =1 and campaign_spending_millions = 24,234,909

$$\hat{Y} = 10.76 + -0.03 * (61) + 1.04 * (21) + 10.32 * (1) + 0.0008 * (24234909)$$

Republican Vote Share Interpretation Example

We are predicting `republican_vote_share` (a proportion) with four predictors: `voter_turnout_percentage` (0-100), `public_approval_rate` (0-100), `incumbant` (0 or 1), and `campaign_spending_millions` (0-??)

	Estimate
(Intercept)	10.76
voter_turnout_percentage	-4.00
public_approval_rate	1.04
incumbent (binary)	10.32
campaign_spending_millions	.00001

What would we expect the republican_vote_share to be if `voter_turnout_percentage` =61, `public_approval_rate` =21, `incumbant` =1 and `campaign_spending_millions` = 24,234,909

$$\hat{Y} = 10.76 + -0.03 * (61) + 1.04 * (21) + 10.32 * (1) + 0.0008 * (24234909)$$

$$\hat{Y} = 41.27$$

Categorical Indepdent Variables

We often want to estimate the effects of categorical variables with more than two values.

Categorical Indepdent Variables

We often want to estimate the effects of categorical variables with more than two values.

- Consider an experiment where we randomized people to receive a basic income each month in cash, a basic income each month on a debit card, or to not receive any funds (control).

Categorical Indepdent Variables

We often want to estimate the effects of categorical variables with more than two values.

- Consider an experiment where we randomized people to receive a basic income each month in cash, a basic income each month on a debit card, or to not receive any funds (control).
- This experiment was designed to test of a universal basic income improves student achievement.

Categorical Indepdent Variables

We often want to estimate the effects of categorical variables with more than two values.

- Consider an experiment where we randomized people to receive a basic income each month in cash, a basic income each month on a debit card, or to not receive any funds (control).
- This experiment was designed to test of a universal basic income improves student achievement.
- We want to know how the the two payment conditions compare to the control.

Categorical Indepdent Variables

We often want to estimate the effects of categorical variables with more than two values.

- Consider an experiment where we randomized people to receive a basic income each month in cash, a basic income each month on a debit card, or to not receive any funds (control).
- This experiment was designed to test of a universal basic income improves student achievement.
- We want to know how the the two payment conditions compare to the control.
- We can do this with a regression model.

Categorical Indepdent Variables

We often want to estimate the effects of categorical variables with more than two values.

- Consider an experiment where we randomized people to receive a basic income each month in cash, a basic income each month on a debit card, or to not receive any funds (control).
- This experiment was designed to test of a universal basic income improves student achievement.
- We want to know how the the two payment conditions compare to the control.
- We can do this with a regression model.

Student Achievement Example

- In this experiment we have an independent variable called `conditions` that can take the values of `control`, `cash` or `card`.

Student Achievement Example

- In this experiment we have an independent variable called `conditions` that can take the values of `control`, `cash` or `card`.
- Our dependent variable is called `student_achievement`.

Student Achievement Example

- In this experiment we have an independent variable called `conditions` that can take the values of `control`, `cash` or `card`.
- Our dependent variable is called `student_achievement`.

We run the following model:

Student Achievement Example

- In this experiment we have an independent variable called `conditions` that can take the values of `control`, `cash` or `card`.
- Our dependent variable is called `student_achievement`.

We run the following model:

- $Y = \alpha + \beta_1 conditions_i + u_i$

Student Achievement Example

- In this experiment we have an independent variable called `conditions` that can take the values of `control`, `cash` or `card`.
- Our dependent variable is called `student_achievement`.

We run the following model:

- $Y = \alpha + \beta_1 conditions_i + u_i$

This corresponds to the following R command:

Student Achievement Example

- In this experiment we have an independent variable called `conditions` that can take the values of `control`, `cash` or `card`.
- Our dependent variable is called `student_achievement`.

We run the following model:

- $Y = \alpha + \beta_1 conditions_i + u_i$

This corresponds to the following R command:

- `lm(student_achievement ~ conditions, data = income)`

Student Achievement Example

- In this experiment we have an independent variable called `conditions` that can take the values of `control`, `cash` or `card`.
- Our dependent variable is called `student_achievement`.

We run the following model:

- $Y = \alpha + \beta_1 conditions_i + u_i$

This corresponds to the following R command:

- `lm(student_achievement ~ conditions, data = income)`

We will get the following output:

Student Achievement Example

- In this experiment we have an independent variable called `conditions` that can take the values of `control`, `cash` or `card`.
- Our dependent variable is called `student_achievement`.

We run the following model:

- $Y = \alpha + \beta_1 conditions_i + u_i$

This corresponds to the following R command:

- `lm(student_achievement ~ conditions, data = income)`

We will get the following output:

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

Student Achievement Example

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

How do we interpret this?

Student Achievement Example

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

How do we interpret this?

1. When you have a categorical independent variable, R will drop one of the values. This is called the reference category.

Student Achievement Example

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

How do we interpret this?

1. When you have a categorical independent variable, R will drop one of the values. This is called the reference category.
2. R will estimate an effect for all other remaining values AND an intercept.

Student Achievement Example

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

How do we interpret this?

1. When you have a categorical independent variable, R will drop one of the values. This is called the reference category.
2. R will estimate an effect for all other remaining values AND an intercept.
3. The interpretation is that the intercept corresponds to estimated effect of the reference category. The other coefficients are the effect **relative** to the reference category.

Student Achievement Example

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

How do we interpret this?

1. When you have a categorical independent variable, R will drop one of the values. This is called the reference category.
2. R will estimate an effect for all other remaining values AND an intercept.
3. The interpretation is that the intercept corresponds to estimated effect of the reference category. The other coefficients are the effect **relative** to the reference category.

Student Achievement Example

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

- Participants were randomly assigned to a *single condition*
- It is not possible for someone to be in the control condition and another condition.

Student Achievement Example

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

- Participants were randomly assigned to a *single condition*
- It is not possible for someone to be in the control condition and another condition.

Consider someone in the cash condition:

$$2.21 = 4.22 + -2.01 * (1) + 1.04 * (0)$$

Student Achievement Example

(Intercept)	conditionsCash	conditionsCard
4.22	-2.01	1.04

- Participants were randomly assigned to a *single condition*
- It is not possible for someone to be in the control condition and another condition.

Consider someone in the cash condition:

$$2.21 = 4.22 + -2.01 * (1) + 1.04 * (0)$$

Consider someone in the control condition:

$$4.22 = 4.22 + -2.01 * (0) + 1.04 * (0)$$