Bivariate Regression I: Conceptual Overview and Estimation

Dr. Michael Fix
mfix@gsu.edu

Georgia State University

2 February 2023

Intro to Inference

- Population: $Y_i = \beta_0 + X_i\beta_1 + u_i$
  - Note a minor notational change from last week in that I am now using β_0 instead of α
- When $u_i \sim N(0, \sigma^2)$, our estimators $\hat{\beta}_0$ (or b_0) and $\hat{\beta}_1$ (or b_1) are defined:
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$
- $\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$

The Key Point

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables.

Due to (*inter alia*):

- **Sampling variability**: Random samples from a population $\rightarrow$ slightly different $\hat{\beta}_0$s and $\hat{\beta}_1$s.
- **Random variability in X**: In cases where X is also a random variable...
- **Intrinsic variability in Y**: Because $Y_i = \mu + u_i$.

Utility of $\hat{\beta}_0$ and $\hat{\beta}_1$

- Remember that $\hat{\beta}_0$ and $\hat{\beta}_1$ (like all estimators) are point estimates.
- Alone, point estimates border on useless.
- What else do we need?

Thinking about Variance

- X is fixed (by assumption or nature)
- Y has both systematic and random variation
  - Systematic (related to X) is what we seek to explain
  - Random goes into the error term, u_i, and we assume:
  - $u_i \sim i.i.d.N(0, \sigma^2)$
  - Or, we can define the stochastic variation in Y as
  - $Var(Y|X, \beta) = \sigma^2$

Thinking about Variance

- Combining the above with the assumption that X is "fixed" (something we will return to later in the course), we can derived estimates of the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$
- $Var(\hat{\beta}_0) = \frac{\sum X_i^2}{N\sum(X_i - \bar{X})^2}\sigma^2$
- $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$
- $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{X}}{\sum(X_i - \bar{X})^2}\sigma^2$
- Note: you can find proofs for these online or in many texts if you are interested.

Important Implications

1. Variance of both estimates β_0 and β_1 is directly proportional to σ^2
2. Variance of both estimates is inversely proportional to $\sum(X_i - \bar{X})$
3. As N increases, the variability of our estimates will go down
4. The covariance of the two estimates depends on the sign of X

OLS is BLUE

- Under a set of specific assumptions, the OLS estimator is ideal for estimating β_0 and β_1
- Specifically, the OLS estimator is **BLUE**:
  - **B**est (minimum variance)
  - **L**inear
  - **U**nbiased
  - **E**stimator
- Unbiasedness and minimum variance can be shown via formal proof

Gauss-Markov Theorem

- Imagine:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

- Rewrite:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

- k are "weights":

$$\hat{\beta}_1 = \sum k_i Y_i$$

- where $k_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$

Gauss-Markov (continued)

- Alternative (non-LS) estimator:

$$\tilde{\beta}_1 = \sum w_i Y_i$$

- Unbiasedness requires $E(\tilde{\beta}_1) = \beta_1$:

$$\begin{aligned} E(\tilde{\beta}_1) &= \sum w_i E(Y_i) \\ &= \sum w_i(\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i X_i \end{aligned}$$

- Thus, $\tilde{\beta}_1$ is only unbaised if $\sum w_i = 0$ and $\sum w_i X_i = 1$

Gauss-Markov (continued)

- Variance:

$$\begin{aligned}
\text{Var}(\tilde{\beta}_1) &= \text{Var}\left(\sum w_i Y_i\right) \\
&= \sigma^2 \sum w_i^2 \\
&= \sigma^2 \sum \left[w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} + \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 \\
&= \sigma^2 \sum \left[w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 + \sigma^2 \left[\frac{1}{\sum (X_i - \bar{X})^2} \right]
\end{aligned}$$

Gauss-Markov (continued)

- Because $\sigma^2 \left[\frac{1}{\sum(X_i - \bar{X})^2}\right]$ is a constant, $\min[\text{Var}(\tilde{\beta}_1)]$ minimizes

$$\sum \left[w_i - \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2} \right]^2$$

- Minimized at:

$$w_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$$

- implying:

$$\begin{aligned} \text{Var}(\tilde{\beta}_1) &= \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \\ &= \text{Var}(\hat{\beta}_1) \end{aligned}$$

Classical Hypothesis Testing — Quick Review

- Declare a null hypothesis: H_0
- Assuming that H_0 is true, calculate the likelihood of obtaining our sample value
- Set a threshold for significance
  - This value is the probability of getting your sample statistic given H_0 is true that you are willing to accept
  - The value is known by the Greek letter α
  - The generic is $\alpha = 5\%$ but it should be based on the context of the study and data
  - This value sets the critical value

Classical Hypothesis Testing — Quick Review

- Compare the sample value to H_0
- If the sample value is above (or below) the critical value we can *reject* H_0
- Note that we are not confirming H_A but instead rejecting H_0
- Instead of utilizing a critical point every time we can compare α to the p-value
- We can reject H_0 if $p \leq \alpha$
- p-values are also useful as they allow us to see how close or far from the threshold α an estimate lies
    - Note: a p-value is simply the probability that we would get our sample value given that the null hypothesis is true

Assumptions and Implications

- As noted above, we assume our error term is normally distributed ($u_i \sim N(0, \sigma^2)$)
- This implies that since $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables that are functions of u_i:

$$\hat{\beta}_0 \sim N(\beta_0, Var(\hat{\beta}_0))$$
$$\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$$

Z-Score

- This should also make inference easy as the Z-score for the βs should be:

$$z_{\hat{\beta}_1} = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\text{Var}(\hat{\beta}_1)}} = \frac{(\hat{\beta}_1 - \beta_1)}{\text{s.e.}(\hat{\beta}_1)}$$

- Note $z_{\hat{\beta}_1} \sim N(0, 1)$

A Problem

- The formula for $z_{\hat{\beta}_1}$ requires us to calculate s.e.$(\hat{\beta}_1)$
- This requires us to know $\hat{\sigma^2}$ (the true population error variance)

Solution

- Instead we can use the estimated variance of the errors, $\hat{\sigma}^2$
- $\hat{\sigma}^2$ is an unbiased estimator of σ^2 (see text for proof)
- We can then calculate:

$$\widehat{\text{s.e.}(\hat{\beta}_1)} = \sqrt{\widehat{\text{Var}(\hat{\beta}_1)}} = \sqrt{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}} = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

Solution

- While this does allow for inference, it has one further implication:

$$t_{\hat{\beta}_1} \equiv \frac{(\hat{\beta}_1 - \beta_1)}{\widehat{\text{s.e.}(\hat{\beta}_1)}} = \frac{(\hat{\beta}_1 - \beta_1)}{\frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}}$$

$$= \frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum(X_i - \bar{X})^2}}{\hat{\sigma}}$$

$$\sim t_{N-k}$$

Predicted Values

- Point prediction:

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_k$$

- Y_k is unbiased:

$$\begin{aligned}
\mathrm{E}(\hat{Y}_k) &= \mathrm{E}(\hat{\beta}_0 + \hat{\beta}_1 X_k) \\
&= \mathrm{E}(\hat{\beta}_0) + X_k \mathrm{E}(\hat{\beta}_1) \\
&= \beta_0 + \beta_1 X_k \\
&= \mathrm{E}(Y_k)
\end{aligned}$$

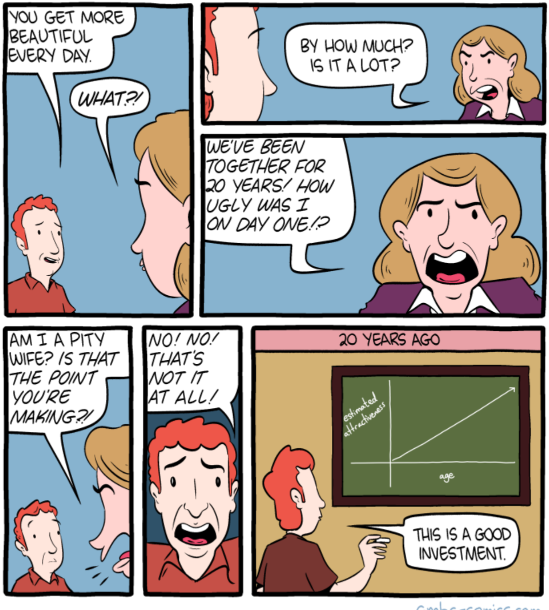
Predicted Values

- Variability:

$$
\begin{aligned}
\mathrm{Var}(\hat{Y}_k) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_k) \\
&= \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \sigma^2 + \left[\frac{\sigma^2}{\sum (X_i - \bar{X})^2} \right] X_k^2 + 2 \left[\frac{-\bar{X}}{\sum (X_i - \bar{X})^2} \sigma^2 \right] X_k \\
&= \sigma^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]
\end{aligned}
$$

- This means that $\mathrm{Var}(\hat{Y}_k)$:
  - Decreases in N
  - Decreases in $\mathrm{Var}(X)$
  - Increases in $|X - \bar{X}|$

Out of Sample Predictions



Let's use a toy model

```r
### Load necessary packages ----
# Use install.packages() if you do not have this package
library(tidyverse) # Data manipulation
library(stargazer) # Creates nice regression output tables

### Load your data ----
# We are using V-Dem version 12
my_data <- readRDS("data/vdem12.rds")

# Let's change names of some of these variables for the sake of simplicity
# I am also subsetting it to only US
us_data <- my_data |>
  filter(country_name == "United States of America") |>
  rename(democracy = v2x_polyarchy, gdp_per_capita = e_gdppc)

### Bivariate OLS ----
# Fit simple linear regression model
my_model <- lm(democracy ~ gdp_per_capita,
               data = us_data,
               x = TRUE, # see arguments in function help page
               y = TRUE) # TRUE allow us to have these values in the list object

# View model summary
summary(my_model)

stargazer(my_model, type = "text")
```

Model output

```
> # View model summary
> summary(my_model)

Call:
lm(formula = democracy ~ gdp_per_capita, data = us_data, x = TRUE,
    y = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-0.240151 -0.043865 -0.007221  0.057909  0.140415

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.3324666  0.0057544   57.78   <2e-16 ***
gdp_per_capita 0.0118020  0.0002537   46.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06302 on 229 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9043,    Adjusted R-squared:  0.9039
F-statistic:  2165 on 1 and 229 DF,  p-value: < 2.2e-16
```

Let's look at y, $\hat{y}$, and residuals

```
# my_model is a list object - which means that it has multiple objects contained
# within an object
names(my_model)

# Get y and y-hat: create a data frame and change column names
y_yhat <- as.data.frame(cbind(my_model$y, my_model$fitted.values, my_model$residuals))
colnames(y_yhat) <- c("My Y", "My Y Hat", "My Residuals")

# Let's look at the first 10 rows
# remember u_i = y - y_hat
y_yhat[1:10, ]
```

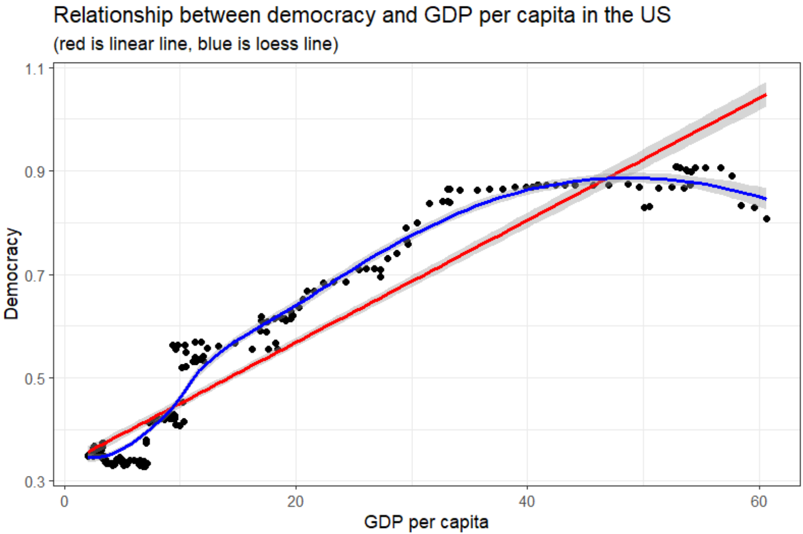
```
> # Let's look at the first 10 rows
> # remember u_i = y - y_hat
> y_yhat[1:10, ]
    My Y  My Y Hat My Residuals
1  0.350 0.3566961 -0.006696131
2  0.349 0.3564365 -0.007436487
3  0.348 0.3567197 -0.008719735
4  0.353 0.3572626 -0.004262626
5  0.353 0.3581360 -0.005135973
6  0.353 0.3592100 -0.006209955
7  0.352 0.3600715 -0.008071500
8  0.354 0.3605790 -0.006578986
9  0.358 0.3608740 -0.002874035
10 0.363 0.3614523  0.001547667
```

Let's use plots for closer examination!

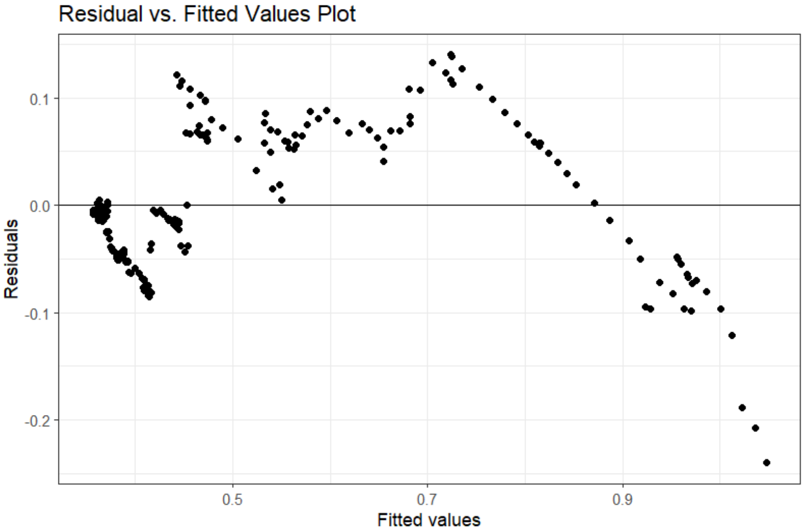
```
### Let's use graphs ----
# Plot the relationship between democracy and GDP per capita
us_data |>
  ggplot(aes(x = gdp_per_capita, y = democracy)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  geom_smooth(color = "blue") +
  theme_bw() +
  labs(x = "GDP per capita", y = "Democracy",
       title = "Relationship between democracy and GDP per capita in the US",
       subtitle = "(red is linear line, blue is loess line)")

# Residual plot -- Fitted values vs residuals
# This plot will be super useful for homoskedasticity assumption
my_model |>
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw() +
  labs(x = "Fitted values", y = "Residuals",
       title = "Residual vs. Fitted Values Plot")

# Histogram of these residuals
hist(my_model$residuals,
     xlab = "Residuals",
     ylab = "Frequency",
     main = "Distribution of residuals")
```


Relationship between democracy and GDP per capita in the US
(red is linear line, blue is loess line)
Democracy
GDP per capita
1.1
0.9
0.7
0.5
0.3
0
20
40
60

Residual vs fitted values plot



Histogram of residuals