

# ProbSet 2, January 23

Gideon Gordon [gideongordon2029@u.northwestern.edu](mailto:gideongordon2029@u.northwestern.edu)

**Due:** January 23, 2026

**Submission:** <https://canvas.northwestern.edu/courses/245562/assignments/1676715>

## 0.1 Problem 1

*Define, in your own words, the conditional expectation function and the best linear predictor.*

*How are these two ideas related, and in what ways are they different?*

The Conditional Expectation Function (CEF) is the actual chart of the expected value of Y conditional on for every possible value of X. I usually imagine it (at least when only one conditioning variable is involved) as a sort of mountain range, with a histogram of possible values of Y at every cross-section. The mean of the distribution of Y for a given value X is the output of the function; the value of X is the input.

The CEF is the best predictor, period, of Y given X. This is true in two ways. Firstly, the CEF is on average right; on average the difference between the Y-value predicted by the CEF and the actual Y-value is 0. That is, the CEF's error is *on average* zero. This does not mean there will not be variations from the CEF's predictions, only that these variations will not be systematic. Secondly, the CEF's error has no correlation with X.

However, the CEF, while it exists theoretically, may have no practical value at all. I can try to find a CEF for the relationship between people with the first name Gerard and the air

[pollution in Alaska](#). It won't mean anything about whether there's a real relationship, but it will exist. This also goes for the BLP.

The BLP is a linear approximation of the CEF. It will only rarely coincide directly with the CEF (it might if the CEF is also linear). The BLP preserves crucial properties from the CEF (an average error of zero, and no correlation between the error and X), but also gives us a bit more traction on understanding what, in general terms, the data is doing. There are other nonlinear approximations of the CEF possible, and the CEF will probably not be linear, but for many purposes, all we want to know for a first cut at the data is if the line goes up or down. If that's all we need, the BLP is a good thing to look for.

The CEF cannot be directly calculated from data in the real world. There are many possible approximations of a CEF compatible with a finite number of points on the X-Y plane (or for that matter higher-dimensional grids of variables). The BLP actually can be directly calculated, and there's only one BLP for a given scatterplot, since we've arbitrarily limited the form to be linear. This means that for most of what we do with regressions, the BLP is what we'll be using.

## 0.2 Problem 2

*Consider the multivariate BLP:*

$$m(X) = _0 + _1X_1 + _2X_2 + \dots + _kX_k$$

*Explain what each of the symbols used in the expression above would usually mean.*

$m$  here refers to an expected value (or a mean of values).  $X$  is the input for the function  $m$ , and represents a vector of  $X$ s of different values; the values of each element in  $X$  are given by the different  $X$ s on the right side of the formula. The  $s$  are each a coefficient with which their respective  $X$ s are multiplied, with the exception of the intercept  $_0$ . There can be

any number of terms in this BLP added linearly together, until  $_kX_k$ . The symbol  $k$  is used to mark the last element in a vector, and can be any natural number (that is, any natural number higher than the rest of the elements).

---

### 0.3 Problem 3

*Explain the regularity conditions for the BLP to exist, and for each, give an example of a situation in which it would fail.*

First condition: The variance of the distribution must be finite. Some distributions, like the Cauchy distribution that Professor Seawright mentioned in class, violate this condition. Because I had no idea what you actually use a Cauchy distribution to model, I looked it up on Wikipedia. Apparently objects that are spinning often produce observations with a Cauchy distribution.

Second condition: None of the predictors are perfectly collinear. If we measured how tall someone is in inches and how tall they are in centimeters, having both these variables in the equation would break everything.

Third condition: X and Y must be jointly distributed. So if we have a situation where Y is perfectly independent of X, there will be no BLP, just a big blob of truly random dots on a scatterplot. Two rolls of a die are completely independent of one another.

On this last point, it is difficult to actually prove one way or another that two things are truly independent. It could be that, if I rolled dice infinite times, I would eventually discover some minuscule systematic effect of the value of my first die on my second die's outcome. These are assumptions we have to make.

---

## 0.4 Problem 4

Let's consider the widely studied relationship between wealth and democracy. This problem will guide you through an analysis using the *Quality of Governance* dataset, helping you contrast the Conditional Expectation Function (CEF) with the Best Linear Predictor (BLP).

**Data Loading:** Run the following code to load the data. If you encounter issues with the `rqog` package, use the alternative CSV file provided.

```
# Option 1: Using the rqog package (preferred)
# devtools::install_github("ropengov/rqog")
# library(rqog)

# qogts <- read_qog(which_data = "standard", data_type = "time-series")

#Option 2: Alternative if rqog doesn't work (uncomment and run)
qogts <- read.csv("https://github.com/jnseawright/ps405/raw/refs/heads/main/Data/qog_sam

# Clean the data for analysis
library(dplyr)
qog_clean <- qogts %>%
  select(wdi_gdpcapppcon2017, vdem_libdem) %>%
  filter(!is.na(wdi_gdpcapppcon2017), !is.na(vdem_libdem)) %>%
  rename(gdp_pc = wdi_gdpcapppcon2017, democracy = vdem_libdem)
```

### 0.4.1 4a.

Create a visualization of the relationship between wealth (`wdi_gdpcapppcon2017` or `gdp_pc` in the cleaned data) and democracy (`vdem_libdem` or `democracy`). Your plot should include:

1. The raw data points (use transparency if there are many observations)
2. A LOESS curve (to approximate the CEF)

3. A linear regression line (estimate of the BLP)

4. Clear labels, titles, and a legend

```
# Your code for 4a here

library(ggplot2)

# Create the plot with both LOESS (CEF approximation) and linear (BLP) fits
dem_wealth_plot <- ggplot(qog_clean, aes(x = gdp_pc, y = democracy)) +
  geom_point(alpha = 0.3, color = "gray50") + # Raw data
  geom_smooth(method = "loess", se = TRUE, color = "blue",
              aes(color = "LOESS (CEF approx)"), size = 1.2) +
  geom_smooth(method = "lm", se = TRUE, color = "red",
              aes(color = "Linear (BLP)"), size = 1.2) +
  scale_color_manual(values = c("LOESS (CEF approx)" = "blue",
                                "Linear (BLP)" = "red")) +
  labs(title = "Wealth and Democracy: CEF vs. BLP",
       x = "GDP per capita (constant 2017 USD)",
       y = "Liberal Democracy Score (VDem)",
       color = "Fit Type") +
  theme_minimal() +
  theme(legend.position = "bottom")

# Display the plot
dem_wealth_plot
```

**Questions for 4a:** 1. *Describe what each curve (LOESS and linear) suggests about the relationship between wealth and democracy.* The linear approximation suggests that there's a direct positive relationship between wealth and democracy. The LOESS curve seems to

suggest that the relationship between wealth and democracy plateaus at a certain point.

2. *Which fit seems more appropriate for these data and why?* The LOESS curve better captures a key point in the data, which is that in practice, increasing your GDP per capita past about 60,000 USD does not do much to improve your chances at democracy. There is a point of diminishing returns. But the peak visible in the curve is not until around 60,000 USD, so for the vast majority of countries on Earth, the linear relationship between wealth and democracy is probably more relevant.

3. *Based on the LOESS curve, does the relationship appear to be linear throughout the range of GDP values?* No. The upper GDP values farther to the right provide very little marginal increase in GDP per capita. (I should say that though the LOESS curve slopes downward towards the right-hand side of the graph, this is a bit of an illusion caused by the smaller number of cases with high GDP per capita. The 95% confidence interval shadow better captures the plateau, as the upper bound of the confidence interval remains fairly straight though the lower bound extends downward to capture what are most likely to be high-income low-democracy petro-dictatorships).

---

#### 0.4.2 4b.

*Fit the empirical approximation of the Best Linear Predictor connecting democracy and wealth. Report and interpret the coefficients.*

```
# Fit the linear model (BLP)  
blp_model <- lm(democracy ~ gdp_pc, data = qog_clean)  
  
# Display model summary  
summary(blp_model)
```

```
# Alternative: Using modelsummary for nicer output

library(modelsummary)

#modelsummary(blp_model, stars = TRUE, output = "markdown")
```

**Questions for 4b:** 1. Interpret the intercept and slope coefficients in substantive terms.

The intercept is the predicted value of democracy (on the V-dem score) when GDP per capita is 0. The slope of `gdp_pc` is how much the v-dem score increases for every one dollar increase in GDP.

2. What is the predicted democracy score for a country with \$20,000 GDP per capita? Show your calculation.

```
20000*(5.993*10^{-6}) + 0.3027
```

```
[1] 0.42256
```

We get a V-dem score of 0.42 for a country with \$20,000 GDP per capita.

3. Calculate and interpret R-squared. What does it tell us about this BLP?

R-squared is given in the summary statistics as 0.2391. The R-squared is the proportion of the variation explained by the model, which in this case is about 23% — so our model leaves about 77% of the variation in democracy across countries “unexplained.” (I use the term unexplained because I don’t know a better term, but really, we haven’t explained democracy simply by showing a correlation).

---

#### 0.4.3 4c.

Now let’s compare the BLP to a simple approximation of the CEF using grouped means:

```

# Create wealth groups

qog_clean <- qog_clean %>%
  mutate(wealth_group = case_when(
    gdp_pc < 10000 ~ "Low (<$10K)" ,
    gdp_pc >= 10000 & gdp_pc <= 30000 ~ "Medium ($10K-$30K)" ,
    gdp_pc > 30000 ~ "High (>$30K)"
  ))
}

# Calculate group means (simple CEF approximation)

group_means <- qog_clean %>%
  group_by(wealth_group) %>%
  summarize(
    mean_democracy = mean(democracy, na.rm = TRUE) ,
    mean_gdp = mean(gdp_pc, na.rm = TRUE) ,
    n = n()
  )

# Display group means

group_means

# Create comparison plot

library(ggplot2)

# Generate predictions from BLP for plotting

blp_predictions <- data.frame(
  gdp_pc = seq(min(qog_clean$gdp_pc, na.rm = TRUE) ,
                max(qog_clean$gdp_pc, na.rm = TRUE) ,

```

```

        length.out = 100)
)

blp_predictions$democracy_pred <- predict(blp_model, newdata = blp_predictions)

# Create the comparison visualization

comparison_plot <- ggplot(qog_clean, aes(x = gdp_pc, y = democracy)) +
  geom_point(alpha = 0.2, color = "gray50") +
  # BLP line
  geom_line(data = blp_predictions,
            aes(x = gdp_pc, y = democracy_pred, color = "BLP"),
            size = 1.5) +
  # Group means (simple CEF approximation)
  geom_point(data = group_means,
             aes(x = mean_gdp, y = mean_democracy, color = "Group Means (CEF approx)"),
             size = 4, shape = 17) +
  # Vertical lines at group boundaries
  geom_vline(xintercept = c(10000, 30000), linetype = "dashed", alpha = 0.5) +
  scale_color_manual(values = c("BLP" = "red",
                               "Group Means (CEF approx)" = "darkgreen")) +
  labs(title = "Comparing BLP to Grouped Means (Simple CEF)",
       x = "GDP per capita (constant 2017 USD)",
       y = "Liberal Democracy Score",
       color = "Estimate Type") +
  theme_minimal() +
  theme(legend.position = "bottom")

comparison_plot

```

**Questions for 4c:** 1. How well does the BLP approximate the grouped means (our simple CEF approximation)? 2. In which wealth range does the BLP fit best? Where does it fit worst? The BLP is an okay approximation of the grouped means. At lower levels of GDP it overestimates democracy, and at higher levels of GDP it actually underestimates democracy. The BLP fits the group means best at the mid-range from about 10,000 USD per capita to about 30,000, and it's farthest from correct in the highest category.

3. Discuss: Under what conditions might the BLP be a poor approximation of the true CEF for these data? The problem seems to be that the underlying data is roughly logarithmic, with a rapid increase earlier on and a plateau later. (The plateau we know about from the earlier LOESS curve, but we don't see it in the group means because there's no separate group mean for 60,000 USD plus). The group means suggest that the BLP is a better approximation of the true CEF when it's closer to the mean value of X, but it does a worse job farther from the mean value of X. I would add based on the LOESS curve that the BLP is best at the points where it intercepts the CEF, which occurs in this case near the center of the data.

The BLP is therefore a poor approximation at extreme values, where the non-linearity of the CEF produces outcomes the linear model cannot account for. And moving along the data's long tail to the right, the BLP gets less and less accurate.

4. Calculate the mean squared error (MSE) for both the BLP and the grouped means approach (treating group means as predictions for all observations in that group). Which has lower MSE?

```
# Calculate MSE for BLP  
  
blp_mse <- mean(residuals(blp_model)^2)  
  
# Calculate MSE for grouped means approach  
  
qog_with_group_preds <- qog_clean %>%
```

```

left_join(select(group_means, wealth_group, mean_democracy), by = "wealth_group") %>%
  mutate(group_residual = democracy - mean_democracy)
group_mse <- mean(qog_with_group_preds$group_residual^2, na.rm = TRUE)

cat("BLP MSE:", round(blp_mse, 4), "\n")
cat("Grouped Means MSE:", round(group_mse, 4), "\n")
cat("Difference (BLP - Grouped):", round(blp_mse - group_mse, 4))

```

The MSE for the grouped means is lower. So, even an extremely rudimentary way to model the CEF might produce less error than our BLP.

---

#### 0.4.4 4d.

*Modernization theory in political science suggests that democracy increases with wealth but at a decreasing rate (diminishing returns).*

```

# Let's explore a non-linear specification
# Option 1: Polynomial (quadratic) model
poly_model <- lm(democracy ~ poly(gdp_pc, 2, raw = TRUE), data = qog_clean)
summary(poly_model)

# Option 2: Log transformation (common for diminishing returns)
log_model <- lm(democracy ~ log(gdp_pc), data = qog_clean)
summary(log_model)

# Compare models
library(modelsummary)
model_comparison <- list(

```

```

"Linear (BLP)" = blp_model,
"Quadratic" = poly_model,
"Log-Linear" = log_model

)

modelsummary(model_comparison, stars = TRUE, output = "markdown")

# Create comparison plot
library(patchwork)

# Generate predictions from all models
comparison_data <- data.frame(
  gdp_pc = seq(min(qog_clean$gdp_pc), max(qog_clean$gdp_pc), length.out = 200)
)

comparison_data$linear_pred <- predict(blp_model, newdata = comparison_data)
comparison_data$quadratic_pred <- predict(poly_model, newdata = comparison_data)
comparison_data$log_pred <- predict(log_model, newdata = comparison_data)

# Reshape for plotting
library(tidyr)
comparison_long <- comparison_data %>%
  pivot_longer(cols = -gdp_pc, names_to = "model", values_to = "prediction") %>%
  mutate(model = factor(model,
    levels = c("linear_pred", "quadratic_pred", "log_pred"),
    labels = c("Linear (BLP)", "Quadratic", "Log-Linear")))

```

```

# Plot all models together

model_comparison_plot <- ggplot(qog_clean, aes(x = gdp_pc, y = democracy)) +
  geom_point(alpha = 0.1, color = "gray50") +
  geom_line(data = comparison_long,
    aes(x = gdp_pc, y = prediction, color = model),
    size = 1.2) +
  scale_color_brewer(palette = "Set1") +
  labs(title = "Comparing Linear and Non-Linear Specifications",
    x = "GDP per capita",
    y = "Democracy Score",
    color = "Model") +
  theme_minimal() +
  theme(legend.position = "bottom")

model_comparison_plot

```

**Questions for 4d:**

1. Does the LOESS curve from 4a support the modernization theory prediction of diminishing returns? More or less. As I noted looking at the LOESS curve in 4a, the curve itself seems to decline to the right, but we also know that this is a result of a few very wealthy authoritarian outliers. Democracy scores increase (though unevenly) until about \$60,000 and then stop increasing.
2. Compare the linear (BLP), quadratic, and log-linear models. Which seems to best capture the relationship suggested by the LOESS curve? The BLP massively overpredicts democracy in very wealthy countries; the quadratic function seems to imply that democracy actually declines once a country gets “too wealthy”. The log-linear model seems to intuitively capture the relationship best, but...

```

#poly_mse = mean(residuals(poly_model)^2)
#log_mse = mean(residuals(log_model)^2)
#cat("BLP MSE:", round(blp_mse, 4), "\n")
#cat("Polynomial MSE:", round(poly_mse, 4), "\n")
#cat("Logarithmic MSE:", round(log_mse, 4), "\n")

```

...when I did the math, I did find that the polynomial model had the lowest MSE (0.045, versus 0.0498 for the BLP and 0.0487 for the logarithmic model). This I suppose is a cautionary tale in the difference between a model being useful for describing data precisely and a model being useful for making policy.

*3. What are the trade-offs between using a simple linear model (BLP) versus a more flexible specification?*

On a more math-y side of things. The more flexible specifications reduce error, which is good in theory, but they run the risk of overfitting. This, I think, is the problem with the polynomial specification above, which reduces error by matching the data available very neatly, but suggests a relationship which makes very little intuitive sense.

The BLP has higher error than other specifications, so it will be in general “wronger,” but only if our end goal is simply to describe the data. If the goal is to understand an underlying relationship, having a model which flattens out irregularities might get us closer to the truth.

On a science communication side, flexible specifications run the risk of giving false precision. A BLP is helpful, in my view, because it’s so clearly an abstraction.

*4. If you were writing a paper on wealth and democracy, which model would you choose and why? Consider both statistical fit and substantive interpretability.*

I would probably choose the logarithmic model. This is for a few reasons.

1. It fits well enough without overcorrecting, and as I said for the BLP, it smooths out

some of the outliers.

2. It has some theoretical grounding. I have theoretical reasons to think that a logarithmic relationship applies and good reasons (from my outside knowledge) to ignore certain outliers that would otherwise change the shape.
3. A human being can look at the graph and usefully interpret it without doing much additional brain-work.

On the other hand, in an actual paper, if I wanted to be complete, I would provide a lot more than just one graph. So, for example, I would want to show what the scatterplot would look like excluding rentier states, and separately what it looks like excluding European states. In both cases I would show both the basic BLP to give a broad sense of the relationship and the logarithmic model.