Problem Set 2

Due Date: January 23, 2026
Submission: https://canvas.northwestern.edu/courses/245562/assignments/1676715

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 provides the best available prediction of the value of an output variable (Y) given the values of one or more input variables ((X)). This function may be linear, but it may also be highly nonlinear and complex. The best linear predictor is a model that reduces the conditional expectation function to a linear equation of the variables in (X), as in:

$$\text{BLP}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k$$

When the CEF is linear, the BLP and the CEF will be identical. When they are not, the BLP will be the best possible linear simplification of the CEF, which may be useful but also may be an important substantive distortion, depending on the relations in question.

Problem 2

Consider the multivariate BLP:

$$m(\mathbb{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

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

The $\mathbb{X}$ represents an array of k different input variables. The $m()$ represents a function that transforms that array into a predicted mean of the output variable. β_0 represents the intercept of the linear equation. β_1 through β_k represent the slopes associated with each of the k individual input variables. X_1 through X_k represent the individual inputs. The subscript numbers count across input variables with subscript zero representing the intercept, and subscripts 1 through k representing substantive input variables. The $+$ symbols are, of course, addition, and are one of the key reasons this is called a linear equation. Finally the $\cdots$ symbol represents the uncertain number of additional input variables between 2 and k.

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.

From our slides, we have the following conditions:

- $E[Y^2] < \infty$
- $E||\mathbb{X}||^2 < \infty$
- The columns of $\mathbb{X}$ are linearly independent.

The first condition means that the population variance of the output variable must be finite, unlike the Cauchy distribution we discussed in class.

The second condition means that the absolute magnitude of the input variable matrix must also have finite variance; again, a Cauchy distribution would violate this assumption.

The third condition means that no variable in $\mathbb{X}$ can be constructed by a linear equation made of columns of $\mathbb{X}$. For example, if it were true that $X_4 = -5 + 0.5X_1 - 0.5X_2$, the third condition would be violated.

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")
```

```
## Skipping install of 'rqog' from a github remote, the SHA1 (5412c245) has not changed since last insta
##   Use `force = TRUE` to force installation
```

```
library(rqog)
qogts <- read_qog(which_data = "standard", data_type = "time-series")
```

```
## Local file not found.
## Downloading QoG qog_std_ts_jan23.csv data
## from http://www.qogdata.pol.gu.se/data/qog_std_ts_jan23.csv
## in file: C:\Users\jws780\AppData\Local\Temp\RtmpYNJueX/rqog/qog_std_ts_jan23.csv
```

```
## Reading cache file C:\Users\jws780\AppData\Local\Temp\RtmpYNJueX/rqog/qog_std_ts_jan23.csv
```

```
# 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_sample.csv")

# Clean the data for analysis
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
qog_clean <- qogts %>%
  select(wdi_gdpcappppcon2017, vdem_libdem) %>%
  filter(!is.na(wdi_gdpcappppcon2017), !is.na(vdem_libdem)) %>%
  rename(gdp_pc = wdi_gdpcappppcon2017, democracy = vdem_libdem)
```

4a.

*Create a visualization of the relationship between wealth (`wdi_gdpcappppcon2017` 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*

It's possible that I inadvertently provided a full solution to this problem below.

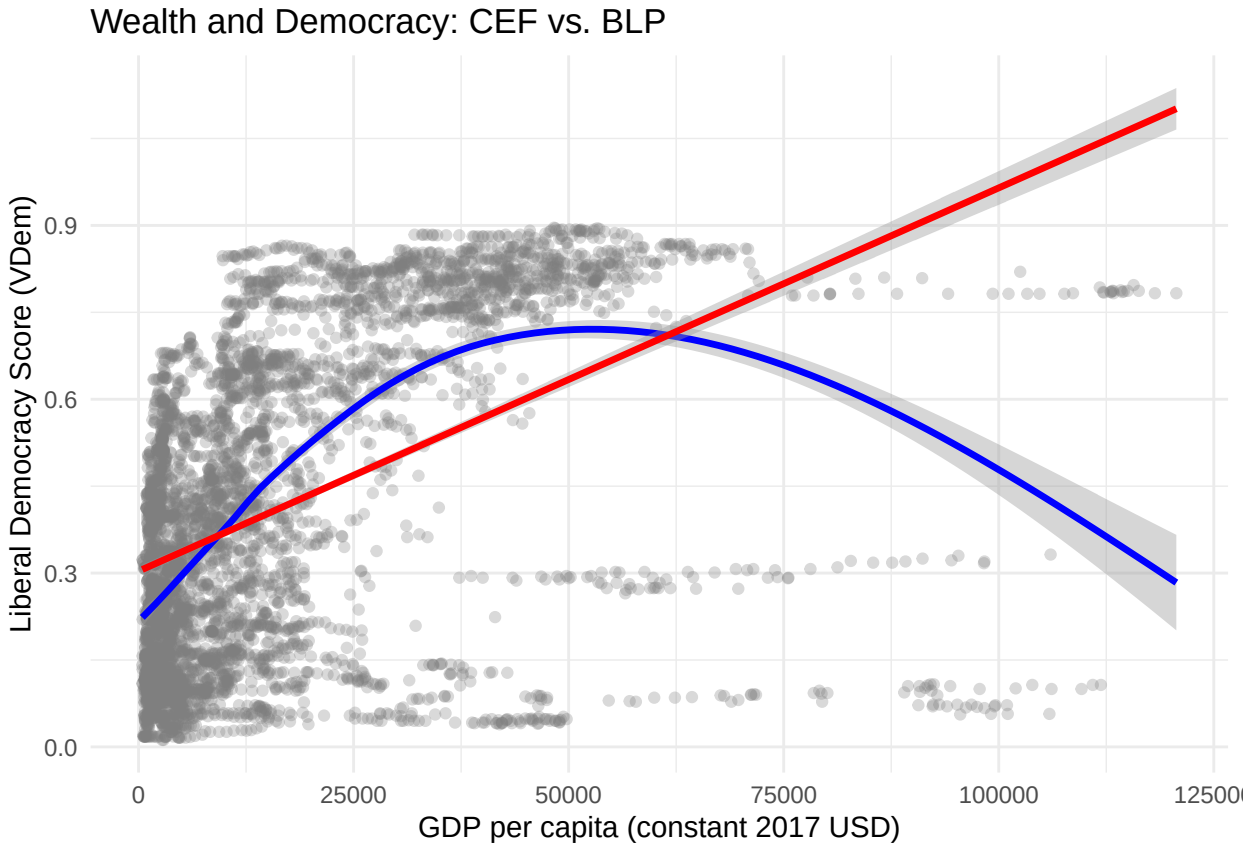
```
# Your code for 4a here
library(ggplot2)

# 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")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# Display the plot
dem_wealth_plot
```

```
## Ignoring unknown labels:
## * colour : "Fit Type"
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: No shared levels found between `names(values)` of the manual scale and the
## data's colour values.
```



Questions for 4a: *1. Describe what each curve (LOESS and linear) suggests about the relationship between wealth and democracy.*

As we talked about in class, the linear relation suggests a modernization-theory relationship where wealth increases the level of democracy, whereas the LOESS curve suggests a nonlinear relationship where democracy increases for moderate levels of wealth before falling on average at the highest levels.

2. Which fit seems more appropriate for these data and why?

Neither fit seems spectacular to me. The linear model departs the range of the data entirely after about 75000 dollars per capita, while the LOESS curve seems to heavily emphasize the experience of a small number of wealthy autocracies. If I had to pick just one, I would choose the LOESS because it at least doesn't give out-of-bounds predictions, but we can do better than either.

3. Based on the LOESS curve, does the relationship appear to be linear throughout the range of GDP values?

The LOESS curve seems very strongly nonlinear.

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)
```

```
##
```

```
## Call:
## lm(formula = democracy ~ gdp_pc, data = qog_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9467 -0.1878  0.0151  0.1967  0.4790
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.033e-01  4.431e-03   68.45   <2e-16 ***
## gdp_pc      6.614e-06  1.733e-07   38.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2354 on 5028 degrees of freedom
## Multiple R-squared:  0.2246, Adjusted R-squared:  0.2244
## F-statistic:  1456 on 1 and 5028 DF,  p-value: < 2.2e-16
```

```
# Alternative: Using modelsummary for nicer output
library(modelsummary)
modelsummary(blp_model, stars = TRUE, output = "markdown")
```

	(1)
(Intercept)	0.303***
	(0.004)
gdp_pc	0.000***
	(0.000)
Num.Obs.	5030
R2	0.225
R2 Adj.	0.224
AIC	-274.1
BIC	-254.5
Log.Lik.	140.054
F	1456.267
RMSE	0.24

- $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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

The intercept represents the average level of democracy of a hypothetical collection of countries with a per capita GDP of 0. That incoherent collection would have an estimated score of 0.303. This isn't very substantively meaningful, but it is mathematically important to anchor the model.

The slope for per capita GDP is 6 times ten to the negative sixth power. This means that every extra dollar of per capita GDP is associated with an increase on average in democracy score of 0.000006. This may not seem like a large change, but GDP increases by substantial multiples of dollars. For example, from 2022 to 2023, Brazil's per capita GDP grew by 1000 dollars. This would be associated with a change of almost 0.01 on a democracy score that ranges only from 0 to 1, looking just at one year's economic change. Hence, this is a reasonably substantial positive slope.

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

3.033e-01 + 20000*6.614e-06 = 0.43558

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

I guess the R-squared is reported above as 0.225. It tells us that the BLP fits about a quarter of the underlying variation of the data on Y.

4c.

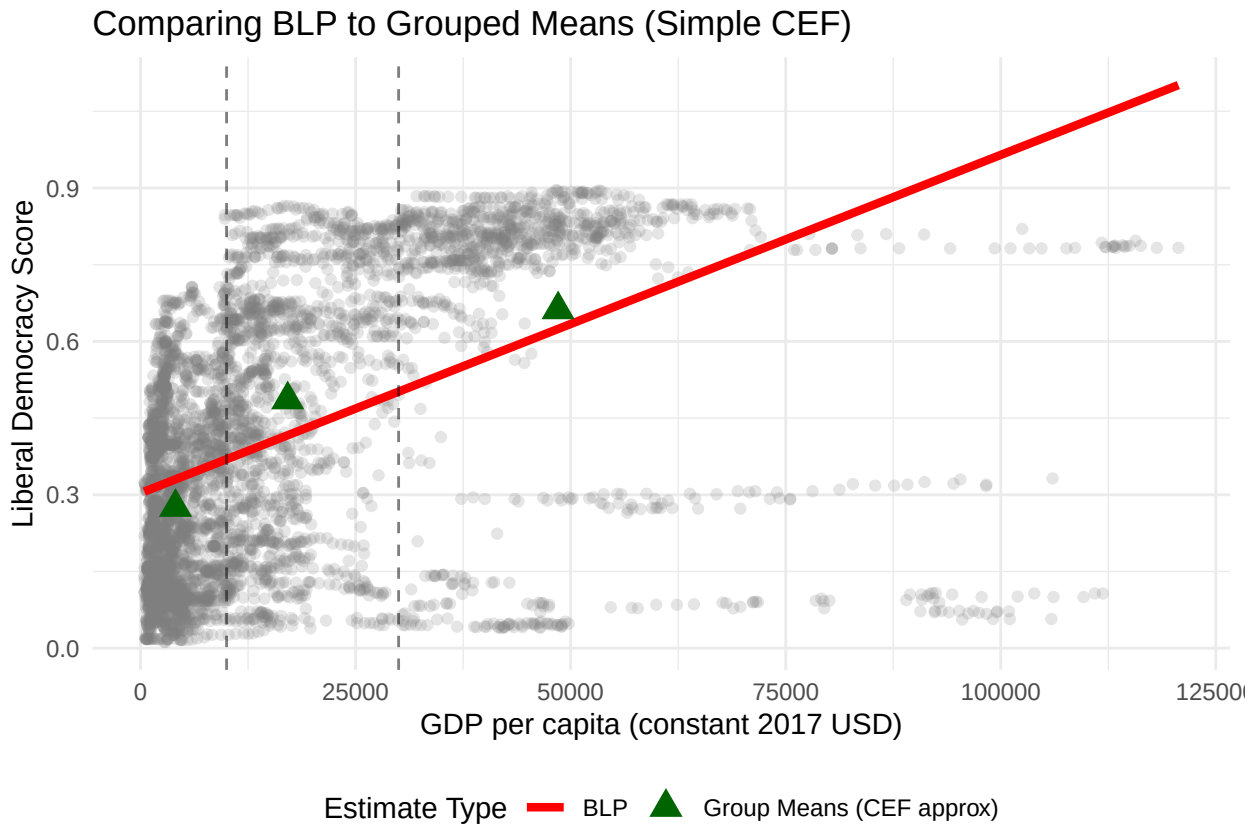
Now let's compare the BLP to a simple approximation of the CEF using grouped means:

```
# 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
```

```
## # A tibble: 3 x 4
##   wealth_group       mean_democracy mean_gdp     n
##   <chr>                       <dbl>    <dbl> <int>
## 1 High (>$30K)                0.662   48544.  1032
## 2 Low (<$10K)                 0.276    4064.  2553
## 3 Medium ($10K-$30K)          0.486   17107.  1445
```

```
# 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)?*

It's a decent approximation, although there are important errors between the CEF approximation and the BLP. I would say this shows a moderate-quality fit.

2. In which wealth range does the BLP fit best? Where does it fit worst?

The BLP fits best among wealthy countries, and worst among middle-income countries.

3. Discuss: Under what conditions might the BLP be a poor approximation of the true CEF for these data?

To the extent that we take seriously the possibility of nonlinearity, the problem of fit in middle-income countries may suggest that the BLP we have estimated might be a bad approximation.

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")
```

```
## BLP MSE: 0.0554
```

```
cat("Grouped Means MSE:", round(group_mse, 4), "\n")
```

```
## Grouped Means MSE: 0.0476
```

```
cat("Difference (BLP - Grouped):", round(blp_mse - group_mse, 4))
```

```
## Difference (BLP - Grouped): 0.0078
```

The grouped means actually have a lower MSE, which really does argue that we're doing things wrong here.

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)
```

```
##
## Call:
## lm(formula = democracy ~ poly(gdp_pc, 2, raw = TRUE), data = qog_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66966 -0.14693  0.02872  0.15392  0.70934
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.162e-01  4.854e-03   44.54   <2e-16 ***
## poly(gdp_pc, 2, raw = TRUE)1  1.784e-05  3.821e-07   46.69   <2e-16 ***
## poly(gdp_pc, 2, raw = TRUE)2 -1.577e-10  4.888e-12  -32.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2143 on 5027 degrees of freedom
## Multiple R-squared:  0.3575, Adjusted R-squared:  0.3573
## F-statistic:  1399 on 2 and 5027 DF,  p-value: < 2.2e-16
```

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

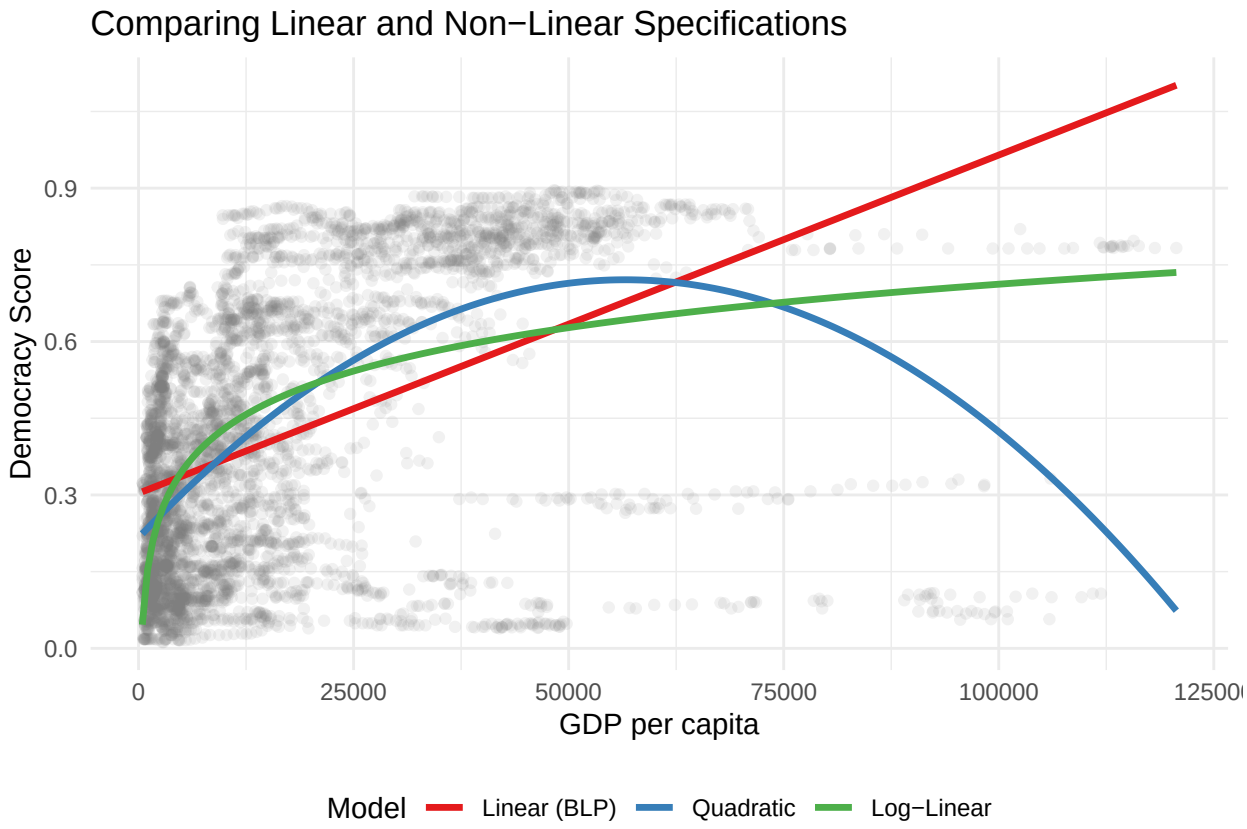
```
##
## Call:
## lm(formula = democracy ~ log(gdp_pc), data = qog_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66217 -0.15618  0.02932  0.18754  0.41968
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.698504   0.023554  -29.66   <2e-16 ***
## log(gdp_pc)  0.122529   0.002568   47.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2218 on 5028 degrees of freedom
## Multiple R-squared:  0.3116, Adjusted R-squared:  0.3115
## F-statistic:  2276 on 1 and 5028 DF,  p-value: < 2.2e-16
```

```r
# 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")
```

	Linear (BLP)	Quadratic	Log-Linear
(Intercept)	0.303***	0.216***	-0.699***
	(0.004)	(0.005)	(0.024)
gdp_pc	0.000***		
	(0.000)		
poly(gdp_pc, 2, raw = TRUE)1		0.000***	
		(0.000)	
poly(gdp_pc, 2, raw = TRUE)2		-0.000***	
		(0.000)	
log(gdp_pc)			0.123***
			(0.003)
Num.Obs.	5030	5030	5030
R2	0.225	0.358	0.312
R2 Adj.	0.224	0.357	0.312
AIC	-274.1	-1218.2	-873.1
BIC	-254.5	-1192.1	-853.5
Log.Lik.	140.054	613.111	439.554
F	1456.267	1398.805	2276.306
RMSE	0.24	0.21	0.22

- $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

```r
# 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?*

It certainly does.

2. Compare the linear (BLP), quadratic, and log-linear models. Which seems to best capture the relationship suggested by the LOESS curve?

The quadratic curve is closest, although the log-linear model may have advantages on the left side.

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

Interpretability, simplicity, and fewer semi-arbitrary choices of models by the researcher, on the one hand, versus fidelity to the actual relationships involved, on the other.

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 certainly choose either the log-linear or the quadratic model. That choice would keep me up at nights. The linear model obviously does not capture the CEF in question and violates the data structure, but the choice between the other two is much more complex and would require attention to additional statistical details.