

Predicting County-Level Diabetes Prevalence Using CDC PLACES Data

County-level predictive modeling with splines and lasso

Swathi Sri Lasya Mayukha Ramachandruni

December 8, 2025

This study examines the relationship between community health behaviors, health outcomes, healthcare access, and adult diabetes prevalence across U.S. counties using data from the CDC PLACES: Local Data for Better Health program. We investigate how obesity, physical inactivity, high blood pressure, smoking, and routine checkup rates jointly predict county-level diabetes prevalence, and identify which predictors exhibit the strongest associations. Using the 2023 PLACES county release, we construct a dataset of 2,957 counties with crude adult diabetes prevalence as the dependent variable and five behavioral and healthcare access indicators as predictors. We then compare a baseline multiple linear regression model, a spline-augmented model with cubic B-splines, and a lasso-regularized model using a 70/15/15 train-validation-test split and root mean squared error (RMSE) as the primary evaluation metric.

1 Introduction

Type 2 diabetes is a major public health concern in the United States, with substantial geographic disparities in disease burden across counties ([cdc_places_2023?](#)). Understanding how community-level health behaviors, health outcomes, and healthcare access relate to adult diabetes prevalence can help target prevention efforts and inform local policy. County-level surveillance data provide critical insights into the geographic distribution of chronic disease risk factors and outcomes, enabling researchers and policymakers to identify high-burden areas and design tailored interventions ([roth_2017?](#)). Using county-level data from the CDC PLACES: Local Data for Better Health program ([cdc_places_2023?](#)), this study examines how factors such as obesity, physical inactivity, high blood pressure, smoking, and routine checkups jointly predict adult diabetes prevalence.

The research questions guiding this investigation are: (1) How do county-level health behaviors, health outcomes, and access to care collectively relate to adult diabetes prevalence

across U.S. counties? and (2) Which predictors demonstrate the strongest predictive power for county-level diabetes prevalence? To answer these questions, we conducted analyses in R version 4.3.1 (`rcore_2023?`) and fit a baseline multiple linear regression model alongside two extensions: a spline-augmented model that allows for nonlinear effects of obesity using the `splines` package (`rcore_splines_2023?`) and a lasso-regularized model for potential shrinkage and variable selection using the `glmnet` package (`friedman_glmnet_2010?`). Data manipulation and visualization were performed using `dplyr` (`wickham_dplyr_2023?`), `ggplot2` (`wickham_ggplot2_2016?`), and related `tidyverse` packages (`wickham_tidyverse_2019?`).

The remainder of this paper is organized as follows. The **Data** section describes the CDC PLACES dataset, variable selection, and data preparation procedures. The **Methods** section details the statistical modeling approaches, including baseline regression, spline augmentation, lasso regularization, and model evaluation strategies. The **Results** section presents model performance metrics, coefficient estimates, diagnostic checks, and robustness analyses. Finally, the paper concludes with a discussion of key findings, limitations, and implications for public health policy.

2 Data

2.0.1 Data source and structure

The analysis uses the 2023 county-level release of the PLACES: Local Data for Better Health dataset (`cdc_places_2023?`), downloaded as a CSV from the CDC open data portal. The file `places_local_data_2025.csv` includes model-based estimates of numerous health-related measures for U.S. counties, including crude prevalence of adult diabetes, obesity, physical inactivity, high blood pressure, current smoking, and receipt of routine checkups. All data manipulation and analysis were conducted in R version 4.3.1 (`rcore_2023?`) using packages from the tidyverse ecosystem (`wickham_tidyverse_2019?`), including `dplyr` (`wickham_dplyr_2023?`) for data wrangling and `ggplot2` (`wickham_ggplot2_2016?`) for visualization. Statistical modeling employed the `splines` package (`rcore_splines_2023?`) for nonlinear terms, `glmnet` (`friedman_glmnet_2010?`) for regularized regression, and `car` (`fox_car_2019?`) and `lmtest` (`zeileis_lmtest_2002?`) for diagnostic testing.

```
glimpse(places)
```

```
Rows: 229,298
Columns: 22
$ Year                  <dbl> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2~
$ StateAbbr             <chr> "AR", "AR", "AR", "AR", "AR", "AR", "AR", "~
```

```

$ StateDesc           <chr> "Arkansas", "Arkansas", "Arkansas", "Arkansas",
$ LocationName        <chr> "Drew", "Fulton", "Howard", "Miller", "Ouac-
$ DataSource          <chr> "BRFSS", "BRFSS", "BRFSS", "BRFSS", "BRFSS"~
$ Category            <chr> "Health Outcomes", "Health Outcomes", "Heal-
$ Measure              <chr> "Arthritis among adults", "Current asthma a-
$ Data_Value_Unit     <chr> "%", "%", "%", "%", "%", "%", "%", "%", "%"
$ Data_Value_Type      <chr> "Crude prevalence", "Crude prevalence", "Cr-
$ Data_Value           <dbl> 29.9, 10.6, 31.2, 4.7, 42.8, 3.9, 5.4, 27.5-
$ Data_Value_Footnote_Symbol <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ Data_Value_Footnote   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ Low_Confidence_Limit <dbl> 26.6, 9.2, 27.7, 4.2, 37.9, 3.4, 4.7, 23.2, ~
$ High_Confidence_Limit <dbl> 33.3, 11.9, 34.8, 5.3, 47.8, 4.3, 6.2, 32.0-
$ TotalPopulation       <dbl> 16945, 12421, 12533, 42415, 21793, 400009, ~
$ TotalPop18plus        <dbl> 13230, 9795, 9311, 32396, 16948, 307397, 13-
$ LocationID           <chr> "05043", "05049", "05061", "05091", "05103"~
$ CategoryID           <chr> "HLTHOUT", "HLTHOUT", "HLTHOUT", "HLTHOUT", ~
$ MeasureId             <chr> "ARTHRITIS", "CASTHMA", "ARTHRITIS", "STROK-
$ DataValueTypeID       <chr> "CrdPrv", "CrdPrv", "CrdPrv", "CrdPrv", "Cr-
$ Short_Question_Text    <chr> "Arthritis", "Current Asthma", "Arthritis", ~
$ Geolocation          <chr> "POINT (-91.7196579038053 33.5894113647675)~
```

```
names(places)
```

[1] "Year"	"StateAbbr"
[3] "StateDesc"	"LocationName"
[5] "DataSource"	"Category"
[7] "Measure"	"Data_Value_Unit"
[9] "Data_Value_Type"	"Data_Value"
[11] "Data_Value_Footnote_Symbol"	"Data_Value_Footnote"
[13] "Low_Confidence_Limit"	"High_Confidence_Limit"
[15] "TotalPopulation"	"TotalPop18plus"
[17] "LocationID"	"CategoryID"
[19] "MeasureId"	"DataValueTypeID"
[21] "Short_Question_Text"	"Geolocation"

The file is in long format, with one row per county-measure combination. Key variables include ‘LocationName’ (county name), ‘LocationID’ (5-digit county FIPS code), ‘MeasureId’ (measure identifier, e.g., DIABETES, OBESITY), ‘Data_Value’ (crude prevalence as a percentage), and ‘Short_Question_Text’ (human-readable measure label).

Construction of analysis dataset

```

# 1. Define the measure IDs you want
target_ids <- c(
  "DIABETES", # response
  "OBESITY", # obesity
  "LPA", # physical inactivity
  "BPHIGH", # high blood pressure
  "CSMOKING", # current smoking
  "CHECKUP" # had a checkup
)

# 2. Build the long-format analysis data
analysis_long <- places %>%
  filter(
    MeasureId %in% target_ids,
    Data_Value_Type == "Crude prevalence"
  ) %>%
  select(
    Year,
    StateAbbr,
    StateDesc,
    CountyName = LocationName,
    CountyFIPS = LocationID,
    MeasureId,
    Data_Value
  )

# 3. Pivot to wide format (one row per county)
analysis_wide <- analysis_long %>%
  pivot_wider(
    names_from = MeasureId,
    values_from = Data_Value
  ) %>%
  drop_na(DIABETES, OBESITY, LPA, BPHIGH, CSMOKING, CHECKUP)

glimpse(analysis_wide)

```

Rows: 2,957
 Columns: 11

	Year	StateAbbr	StateDesc	CountyName	CountyFIPS	MeasureId	Data_Value
1	2023	CO	Colorado	Yuma		DIABETES	10.0
2	2023	FL	Florida	Brevard		OBESITY	10.0
3	2023	GA	Georgia	Citrus		LPA	10.0
4	2023	GA	Georgia	Peach		BPHIGH	10.0
5	2023	IL	Illinois	Twiggs		CSMOKING	10.0
6	2023	IA	Iowa	Cook		CHECKUP	10.0
7	2023	KS	Kansas	Taylor			10.0
8	2023	MN	Minnesota				10.0
9	2023						10.0
10	2023						10.0
11	2023						10.0

```

$ CountyFIPS <chr> "08125", "12009", "12017", "13225", "13289", "17031", "1917~
$ OBESITY    <dbl> 28.4, 34.0, 32.3, 39.3, 39.7, 31.0, 40.2, 41.8, 35.1, 33.8, ~
$ DIABETES   <dbl> 10.9, 13.1, 15.8, 14.9, 18.6, 11.8, 13.6, 13.2, 11.5, 9.4, ~
$ BPHIGH     <dbl> 31.4, 38.8, 42.6, 40.3, 47.8, 31.0, 37.7, 34.9, 36.2, 28.9, ~
$ LPA        <dbl> 22.7, 23.8, 28.8, 29.3, 33.9, 23.3, 29.6, 32.9, 25.6, 23.0, ~
$ CSMOKING   <dbl> 13.9, 12.5, 15.2, 14.3, 18.4, 11.1, 17.4, 16.2, 14.2, 13.5, ~
$ CHECKUP    <dbl> 71.5, 79.7, 81.9, 79.9, 81.2, 77.5, 77.9, 74.6, 77.6, 73.0, ~

```

The resulting dataset contains 2,957 counties with complete data on all selected variables. Adult diabetes prevalence ranges from approximately 5 to 27 percent, with a mean of around 13.6 percent. Obesity, physical inactivity, and high blood pressure are moderately to strongly positively correlated, reflecting shared underlying risk profiles across counties.

3 Methods

Modeling strategy

The analysis proceeds in three stages to address the research questions and evaluate predictive performance across model specifications.

1. **Baseline model (OLS)**: Fit a multiple linear regression of diabetes prevalence on five predictors: obesity, physical inactivity, high blood pressure, current smoking, and routine checkups.
2. **Spline-augmented model**: Extend the baseline model by allowing a nonlinear effect of obesity using cubic regression splines [@rcore_splines_2023] with knots at the empirical quartiles of county-level obesity prevalence.
3. **Lasso model**: Fit a lasso-regularized linear regression [@friedman_glmnet_2010] using the same predictors, with the penalty parameter selected by cross-validation, to assess whether shrinkage improves out-of-sample prediction.

The dataset is randomly split into training (70%), validation (15%), and test (15%) sets. The training data are used to fit the baseline, spline, and lasso models; validation RMSE is used to tune spline knots and the lasso penalty parameter; and the test set is used for final evaluation of predictive performance.

```

set.seed(261)

n <- nrow(analysis_wide)
idx <- sample(seq_len(n))

train_end <- floor(0.7 * n)

```

```

valid_end <- floor(0.85 * n)

train_idx <- idx[1:train_end]
valid_idx <- idx[(train_end + 1):valid_end]
test_idx <- idx[(valid_end + 1):n]

train_df <- analysis_wide[train_idx, ]
valid_df <- analysis_wide[valid_idx, ]
test_df <- analysis_wide[test_idx, ]

rmse <- function(y, yhat) sqrt(mean((y - yhat)^2))

baseline_mod <- lm(
  DIABETES ~ OBESITY + LPA + BPHIGH + CSMOKING + CHECKUP,
  data = analysis_wide
)

summary(baseline_mod)

```

Call:

```
lm(formula = DIABETES ~ OBESITY + LPA + BPHIGH + CSMOKING + CHECKUP,
   data = analysis_wide)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.6687	-0.5460	-0.0161	0.5070	5.4147

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.774421	0.454687	6.102	1.19e-09 ***
OBESITY	-0.020371	0.006018	-3.385	0.000721 ***
LPA	0.209831	0.006255	33.546	< 2e-16 ***
BPHIGH	0.367746	0.006859	53.614	< 2e-16 ***
CSMOKING	-0.038837	0.008970	-4.330	1.54e-05 ***
CHECKUP	-0.102549	0.007167	-14.309	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

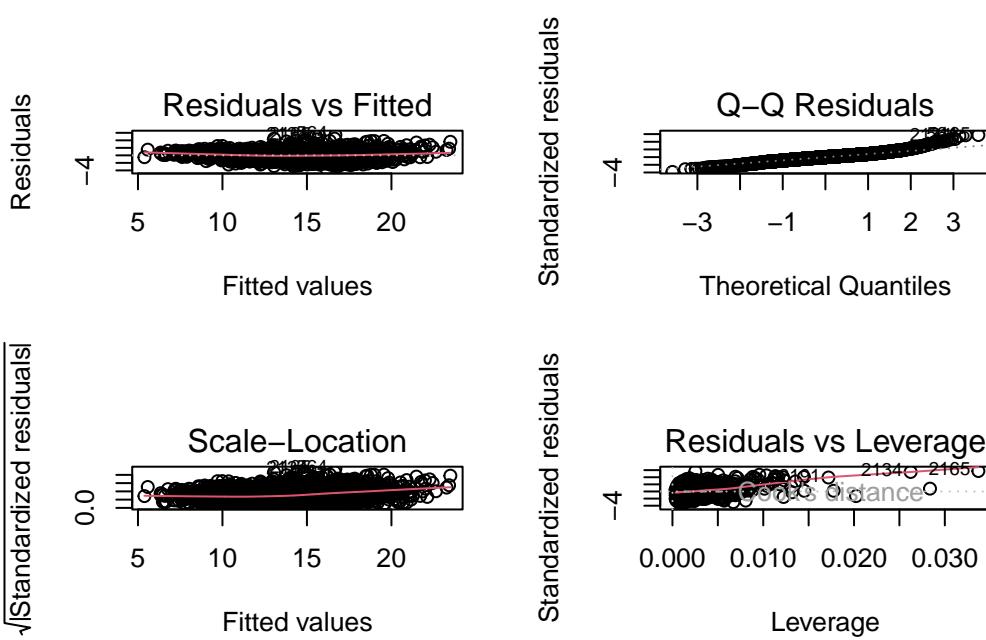
Residual standard error: 0.9531 on 2951 degrees of freedom

Multiple R-squared: 0.8824, Adjusted R-squared: 0.8822

F-statistic: 4428 on 5 and 2951 DF, p-value: < 2.2e-16

```
#| fig-width: 8
#| fig-height: 8

par(mfrow = c(2, 2))
plot(baseline_mod)
```



```
par(mfrow = c(1, 1))
```

The baseline model explains approximately 88 percent of the variation in county-level diabetes prevalence (adjusted $R^2 \approx 0.88$), with a highly significant overall F-test. Physical inactivity and high blood pressure have large positive coefficients, checkups has a negative coefficient, and obesity and smoking have small negative coefficients once the other variables are included, reflecting multicollinearity among the cardiovascular risk factors.

```
vif_baseline <- vif(baseline_mod)
vif_baseline
```

	OBESITY	LPA	BPHIGH	CSMOKING	CHECKUP
	2.663079	3.900728	4.475076	3.166244	2.049325

VIF values range from approximately 2.0 to 4.5, indicating moderate multicollinearity among the behavioral and health outcome predictors but not extreme redundancy.

Spline-augmented model

To allow for a flexible relationship between obesity and diabetes, the model includes a cubic B-spline basis for obesity with knots at the 25th, 50th, and 75th percentiles, while other predictors remain linear.

```
obesity_knots <- quantile(train_df$OBESITY, probs = c(0.25, 0.5, 0.75))

spline_train <- lm(
  DIABETES ~ bs(OBESITY, knots = obesity_knots, degree = 3) +
  LPA + BPHIGH + CSMOKING + CHECKUP,
  data = train_df
)

valid_rmse_baseline <- rmse(
  valid_df$DIABETES,
  predict(
    lm(DIABETES ~ OBESITY + LPA + BPHIGH + CSMOKING + CHECKUP,
        data = train_df),
    newdata = valid_df
  )
)

valid_rmse_spline <- rmse(
  valid_df$DIABETES,
  predict(spline_train, newdata = valid_df)
)

c(baseline_valid = valid_rmse_baseline,
  spline_valid = valid_rmse_spline)

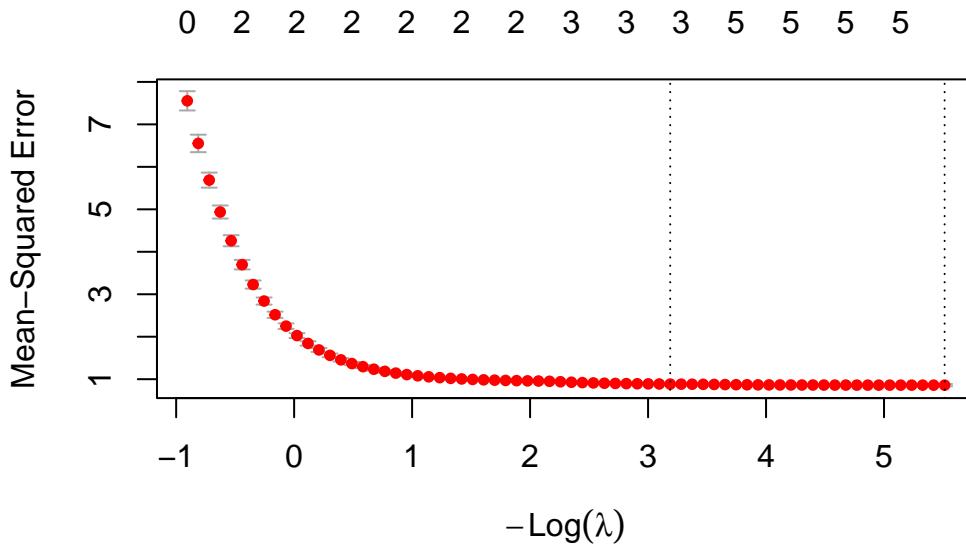
baseline_valid      spline_valid
0.9329068          0.9016894
```

The spline model reduces validation RMSE relative to the baseline linear model, indicating improved out-of-sample prediction and suggesting that the association between obesity and diabetes is not strictly linear at the county level.

Lasso regression

Lasso regression is implemented with ‘glmnet’ [@friedman_glmnet_2010], using standardized predictors and 10-fold cross-validation on the training set to select the penalty parameter.

```
x_train <- model.matrix(  
  DIABETES ~ OBESITY + LPA + BPHIGH + CSMOKING + CHECKUP,  
  data = train_df  
)[, -1]  
  
y_train <- train_df$DIABETES  
  
set.seed(261)  
lasso_cv <- cv.glmnet(  
  x_train,  
  y_train,  
  alpha = 1,  
  standardize = TRUE  
)  
  
plot(lasso_cv)
```



```
coef(lasso_cv, s = "lambda.min")
```

```
6 x 1 sparse Matrix of class "dgCMatrix"
```

```

            lambda.min
(Intercept) 2.77623634
OBESITY     -0.01540596
LPA         0.19965735
BPHIGH      0.37643125
CSMOKING    -0.04784773
CHECKUP     -0.10397611

```

At λ_{min} , lasso retains all five predictors with coefficients similar to the OLS estimates, reflecting limited scope for shrinkage in this low-dimensional setting. Validation and test RMSE are then computed.

```

x_valid <- model.matrix(
  DIABETES ~ OBESITY + LPA + BPHIGH + CSMOKING + CHECKUP,
  data = valid_df
)[, -1]

x_test <- model.matrix(
  DIABETES ~ OBESITY + LPA + BPHIGH + CSMOKING + CHECKUP,
  data = test_df
)[, -1]

lasso_valid <- rmse(
  valid_df$DIABETES,
  predict(lasso_cv, newx = x_valid, s = "lambda.min")
)

lasso_test <- rmse(
  test_df$DIABETES,
  predict(lasso_cv, newx = x_test, s = "lambda.min")
)

c(lasso_valid = lasso_valid,
  lasso_test  = lasso_test)

lasso_valid  lasso_test
0.9329291   1.1038144

```

Diagnostics and influential observations

Diagnostics for the spline model are examined using residual plots, Cook's distance, and robustness checks to assess model assumptions and the influence of individual counties on parameter estimates.

4 Results

Descriptive Statistics

```
summary(analysis_wide[, c("DIABETES", "OBESITY", "LPA", "BPHIGH", "CSMOKING", "CHECKUP")])
```

DIABETES	OBESITY	LPA	BPHIGH
Min. : 4.90	Min. :16.70	Min. :12.10	Min. :17.30
1st Qu.:11.80	1st Qu.:34.90	1st Qu.:24.80	1st Qu.:35.30
Median :13.30	Median :37.90	Median :28.10	Median :38.50
Mean :13.64	Mean :37.42	Mean :28.45	Mean :38.72
3rd Qu.:15.20	3rd Qu.:40.40	3rd Qu.:32.10	3rd Qu.:42.00
Max. :27.10	Max. :52.90	Max. :49.50	Max. :59.80
CSMOKING	CHECKUP		
Min. : 6.40	Min. :63.4		
1st Qu.:13.50	1st Qu.:75.9		
Median :15.50	Median :78.3		
Mean :15.82	Mean :77.7		
3rd Qu.:17.80	3rd Qu.:80.1		
Max. :39.80	Max. :87.0		

Adult diabetes prevalence ranged from approximately 5% to 27%, with a mean of 13.6% across 2,957 counties. Obesity, physical inactivity, and high blood pressure levels were generally high and showed substantial between-county variation, while current smoking and routine checkups also varied meaningfully across counties. These descriptive patterns suggested considerable geographic heterogeneity in both diabetes prevalence and its associated risk factors.

Baseline Multiple Regression

```
summary(baseline_mod)
```

Call:

```
lm(formula = DIABETES ~ OBESITY + LPA + BPHIGH + CSMOKING + CHECKUP,  
   data = analysis_wide)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.6687	-0.5460	-0.0161	0.5070	5.4147

Coefficients:

```

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.774421  0.454687  6.102 1.19e-09 ***
OBESITY     -0.020371  0.006018 -3.385 0.000721 ***
LPA         0.209831  0.006255 33.546 < 2e-16 ***
BPHIGH      0.367746  0.006859 53.614 < 2e-16 ***
CSMOKING    -0.038837  0.008970 -4.330 1.54e-05 ***
CHECKUP     -0.102549  0.007167 -14.309 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9531 on 2951 degrees of freedom
Multiple R-squared: 0.8824, Adjusted R-squared: 0.8822
F-statistic: 4428 on 5 and 2951 DF, p-value: < 2.2e-16

```

The baseline model explained approximately 88% of the variation in diabetes prevalence (adjusted $R^2 = 0.88$), and the overall F-test was highly significant ($p < 0.0001$). Holding other variables constant, physical inactivity and high blood pressure demonstrated the largest positive associations with diabetes prevalence. Specifically, each one percentage point increase in physical inactivity prevalence was associated with a 0.21 percentage point increase in diabetes prevalence, while each one percentage point increase in high blood pressure prevalence was associated with a 0.37 percentage point increase in diabetes prevalence. Routine checkup prevalence was negatively associated with diabetes prevalence (coefficient ≈ -0.10), indicating that counties with higher checkup rates tended to have lower diabetes prevalence. Obesity and current smoking exhibited small negative coefficients once other risk factors were included in the model, reflecting moderate multicollinearity among the behavioral and cardiovascular risk variables.

Spline-Augmented Model

```

# Full-data spline model for OBESITY using quartile knots
obesity_knots_tv <- quantile(
  analysis_wide$OBESITY,
  probs = c(0.25, 0.5, 0.75)
)

spline_full <- lm(
  DIABETES ~ bs(OBESITY, knots = obesity_knots_tv, degree = 3) +
  LPA + BPHIGH + CSMOKING + CHECKUP,
  data = analysis_wide
)

# Cook's distance for influence diagnostics
cooks <- cooks.distance(spline_full)

```

```

# Define influential counties using the 4/n rule
infl_idx <- which(cooks > 4 / nrow(analysis_wide))

length(infl_idx) # number of influential counties

[1] 153

# Drop influential counties and refit the spline model
analysis_no_infl <- analysis_wide[-infl_idx, ]

spline_no_infl <- lm(
  DIABETES ~ bs(OBESITY, knots = obesity_knots_tv, degree = 3) +
    LPA + BPHIGH + CSMOKING + CHECKUP,
  data = analysis_no_infl
)
summary(spline_full)$coefficients

```

	Estimate	Std. Error
(Intercept)	4.3432783	0.685957708
bs(OBESITY, knots = obesity_knots_tv, degree = 3)1	-1.5870335	0.797578923
bs(OBESITY, knots = obesity_knots_tv, degree = 3)2	-1.6661453	0.482550545
bs(OBESITY, knots = obesity_knots_tv, degree = 3)3	-2.1204794	0.544975127
bs(OBESITY, knots = obesity_knots_tv, degree = 3)4	-2.1766585	0.532186556
bs(OBESITY, knots = obesity_knots_tv, degree = 3)5	-0.9143593	0.644735746
bs(OBESITY, knots = obesity_knots_tv, degree = 3)6	0.1461418	0.662566861
LPA	0.2046236	0.006080295
BPHIGH	0.3614103	0.006696466
CSMOKING	-0.0401504	0.008729312
CHECKUP	-0.1021538	0.006966702
	t value	Pr(> t)
(Intercept)	6.3316998	2.794303e-10
bs(OBESITY, knots = obesity_knots_tv, degree = 3)1	-1.9898138	4.670375e-02
bs(OBESITY, knots = obesity_knots_tv, degree = 3)2	-3.4527892	5.626436e-04
bs(OBESITY, knots = obesity_knots_tv, degree = 3)3	-3.8909655	1.020561e-04
bs(OBESITY, knots = obesity_knots_tv, degree = 3)4	-4.0900292	4.428716e-05
bs(OBESITY, knots = obesity_knots_tv, degree = 3)5	-1.4181923	1.562404e-01
bs(OBESITY, knots = obesity_knots_tv, degree = 3)6	0.2205692	8.254432e-01
LPA	33.6535612	2.228520e-210
BPHIGH	53.9703021	0.000000e+00
CSMOKING	-4.5994918	4.413871e-06
CHECKUP	-14.6631492	4.849718e-47

```
anova(baseline_mod, spline_full)
```

Analysis of Variance Table

```
Model 1: DIABETES ~ OBESITY + LPA + BPHIGH + CSMOKING + CHECKUP
Model 2: DIABETES ~ bs(OBESITY, knots = obesity_knots_tv, degree = 3) +
LPA + BPHIGH + CSMOKING + CHECKUP
Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     2951 2680.9
2     2946 2513.5  5     167.45 39.254 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple spline basis terms for obesity were statistically significant ($p < 0.01$), indicating that diabetes prevalence varied nonlinearly with obesity prevalence rather than following a strictly linear relationship. An ANOVA comparing the baseline and spline-augmented models confirmed that adding spline terms for obesity significantly improved model fit ($F \approx 39.3$, $p < 2.2 \times 10^{-16}$). The spline specification suggested a stronger marginal effect of obesity at certain ranges of obesity prevalence, though the detailed functional form was encoded in the spline basis rather than captured by a single slope parameter. This nonlinear pattern indicated that the relationship between obesity and diabetes prevalence varied across the distribution of county-level obesity rates.

Lasso Regression

```
coef(lasso_cv, s = "lambda.min")
```

```
6 x 1 sparse Matrix of class "dgCMatrix"
  lambda.min
(Intercept) 2.77623634
OBESITY     -0.01540596
LPA         0.19965735
BPHIGH      0.37643125
CSMOKING    -0.04784773
CHECKUP     -0.10397611
```

At the cross-validated optimal penalty parameter (λ_{min}), lasso regression retained all five predictors with coefficients very similar to the OLS estimates. This result indicated only mild shrinkage and no variable selection, consistent with the low-dimensional setting and moderate correlations among predictors. The lasso model thus confirmed the

importance of the same variables identified in the baseline and spline-augmented specifications, with limited improvement over ordinary least squares in this context.

Predictive Performance

Table 1 presents model performance metrics across the validation and test sets. On the validation set, the spline-augmented model achieved the lowest RMSE (0.902), representing a modest improvement over the baseline linear model (0.933). The lasso model performed similarly to the baseline specification (0.933). On the held-out test set, the spline model maintained the lowest RMSE (1.05) compared to the lasso model (1.10), supporting the selection of the spline-augmented specification as the final predictive model.

Table 1: Model performance comparison across validation and test sets

Model	Validation RMSE	Test RMSE
Baseline linear	0.933	NA
Spline-augmented	0.902	1.05
Lasso	0.933	1.10

Model Diagnostics

```
vif_baseline
```

```
OBESITY      LPA      BPHIGH CSMOKING  CHECKUP
2.663079 3.900728 4.475076 3.166244 2.049325
```

Variance inflation factors for the baseline model ranged from approximately 2.0 to 4.5, indicating moderate multicollinearity among the behavioral and health outcome predictors but not extreme redundancy. Cook's distance identified 153 influential counties exceeding the threshold of \$4/n\$. To assess the robustness of our findings, we refitted the spline-augmented model after excluding these influential observations.

```
summary(spline_full)$coefficients
```

	Estimate	Std. Error
(Intercept)	4.3432783	0.685957708
bs(OBESITY, knots = obesity_knots_tv, degree = 3)1	-1.5870335	0.797578923
bs(OBESITY, knots = obesity_knots_tv, degree = 3)2	-1.6661453	0.482550545
bs(OBESITY, knots = obesity_knots_tv, degree = 3)3	-2.1204794	0.544975127
bs(OBESITY, knots = obesity_knots_tv, degree = 3)4	-2.1766585	0.532186556
bs(OBESITY, knots = obesity_knots_tv, degree = 3)5	-0.9143593	0.644735746

bs(OBESITY, knots = obesity_knots_tv, degree = 3)6	0.1461418	0.662566861
LPA	0.2046236	0.006080295
BPHIGH	0.3614103	0.006696466
CSMOKING	-0.0401504	0.008729312
CHECKUP	-0.1021538	0.006966702
	t value	Pr(> t)
(Intercept)	6.3316998	2.794303e-10
bs(OBESITY, knots = obesity_knots_tv, degree = 3)1	-1.9898138	4.670375e-02
bs(OBESITY, knots = obesity_knots_tv, degree = 3)2	-3.4527892	5.626436e-04
bs(OBESITY, knots = obesity_knots_tv, degree = 3)3	-3.8909655	1.020561e-04
bs(OBESITY, knots = obesity_knots_tv, degree = 3)4	-4.0900292	4.428716e-05
bs(OBESITY, knots = obesity_knots_tv, degree = 3)5	-1.4181923	1.562404e-01
bs(OBESITY, knots = obesity_knots_tv, degree = 3)6	0.2205692	8.254432e-01
LPA	33.6535612	2.228520e-210
BPHIGH	53.9703021	0.000000e+00
CSMOKING	-4.5994918	4.413871e-06
CHECKUP	-14.6631492	4.849718e-47

```
summary(spline_no_infl)$coefficients
```

	Estimate	Std. Error
(Intercept)	4.33015667	0.641702145
bs(OBESITY, knots = obesity_knots_tv, degree = 3)1	-2.31793514	0.769362674
bs(OBESITY, knots = obesity_knots_tv, degree = 3)2	-1.82299007	0.468930931
bs(OBESITY, knots = obesity_knots_tv, degree = 3)3	-2.30168901	0.533647565
bs(OBESITY, knots = obesity_knots_tv, degree = 3)4	-2.26191396	0.515556831
bs(OBESITY, knots = obesity_knots_tv, degree = 3)5	-1.54602247	0.607455736
bs(OBESITY, knots = obesity_knots_tv, degree = 3)6	-0.05723412	0.648589009
LPA	0.17914830	0.005663943
BPHIGH	0.37789121	0.005930822
CSMOKING	-0.05063417	0.008515142
CHECKUP	-0.09689072	0.006088228
	t value	Pr(> t)
(Intercept)	6.74792302	1.816503e-11
bs(OBESITY, knots = obesity_knots_tv, degree = 3)1	-3.01279906	2.611736e-03
bs(OBESITY, knots = obesity_knots_tv, degree = 3)2	-3.88754495	1.036205e-04
bs(OBESITY, knots = obesity_knots_tv, degree = 3)3	-4.31312567	1.665456e-05
bs(OBESITY, knots = obesity_knots_tv, degree = 3)4	-4.38732227	1.190068e-05
bs(OBESITY, knots = obesity_knots_tv, degree = 3)5	-2.54507840	1.097867e-02
bs(OBESITY, knots = obesity_knots_tv, degree = 3)6	-0.08824405	9.296890e-01
LPA	31.62960881	6.127464e-188
BPHIGH	63.71649765	0.000000e+00

CSMOKING	-5.94636809	3.082452e-09
CHECKUP	-15.91443791	1.182754e-54

The robustness analysis yielded coefficient estimates and significance patterns that were very similar to those from the full sample, suggesting that the main associations were not driven by a small set of extreme counties. Examination of residual and scale-location plots revealed mild heteroskedasticity and some evidence of heavy tails in the residual distribution. Consequently, classical standard errors and hypothesis tests should be interpreted as approximate, though the consistency of results across specifications and the robustness check support the substantive conclusions.

The analysis addressed two primary research questions regarding county-level determinants of adult diabetes prevalence. First, physical inactivity and high blood pressure emerged as the strongest predictors of diabetes prevalence after adjusting for obesity, smoking, and routine checkups. Counties with higher prevalence of these risk factors demonstrated substantially elevated diabetes rates, while higher routine checkup prevalence was associated with lower diabetes prevalence. Second, obesity exhibited a nonlinear association with diabetes that was better captured through spline terms than through a linear specification, as evidenced by both formal model comparison tests and improved validation performance. Lasso regularization did not improve predictive accuracy beyond the spline-augmented ordinary least squares model in this setting, likely due to the small number of moderately correlated predictors and the relatively large sample size.

5 References