

# **Regression case study**

Thomas Lumley

9(10) July 2025

# In this session

Shorter lecture section, presenting a worked example

Longer exercise section, for you to do one

# NHANES data example

Data on blood pressure and diet from the US NHANES health survey.

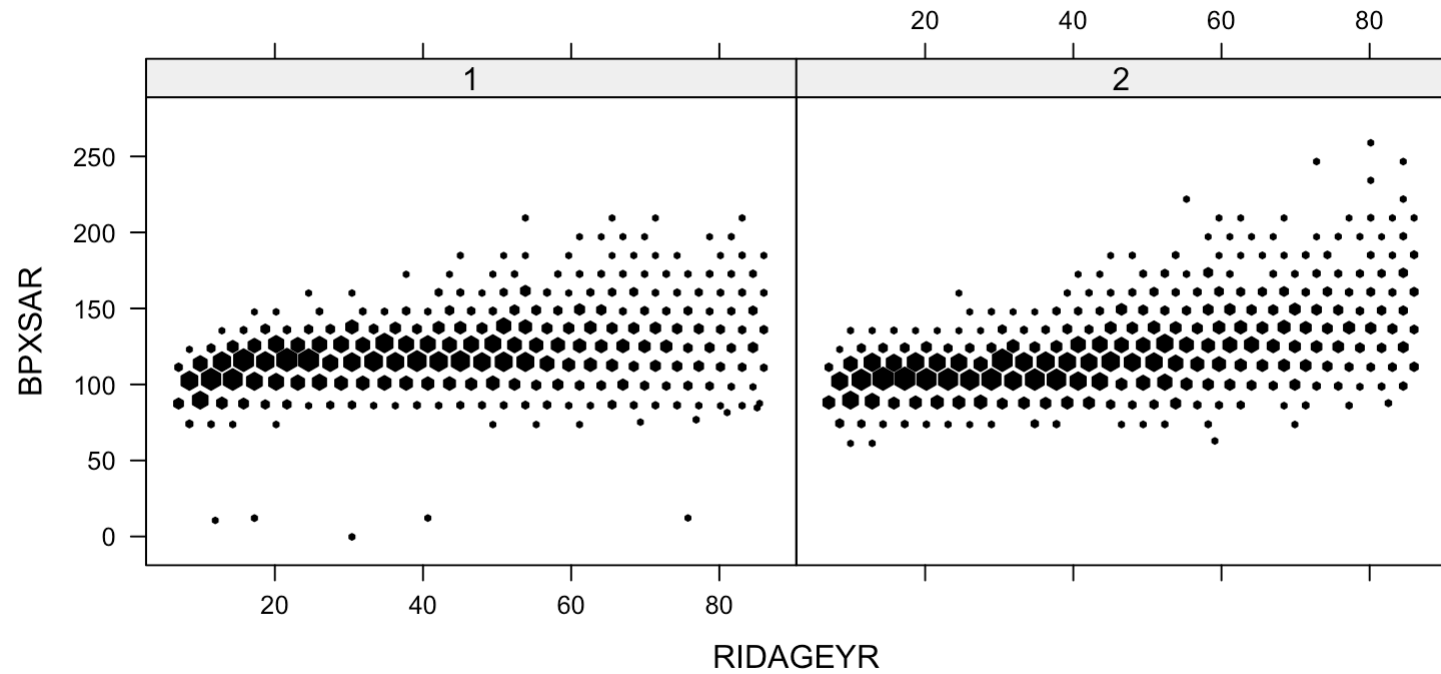
Complex four-stage survey, but public-use data approximates by two-stage design.

Already done: select blood pressure, BMI, age, gender, race/ethnicity, dietary sodium, potassium from various NHANES files for two two-year waves.

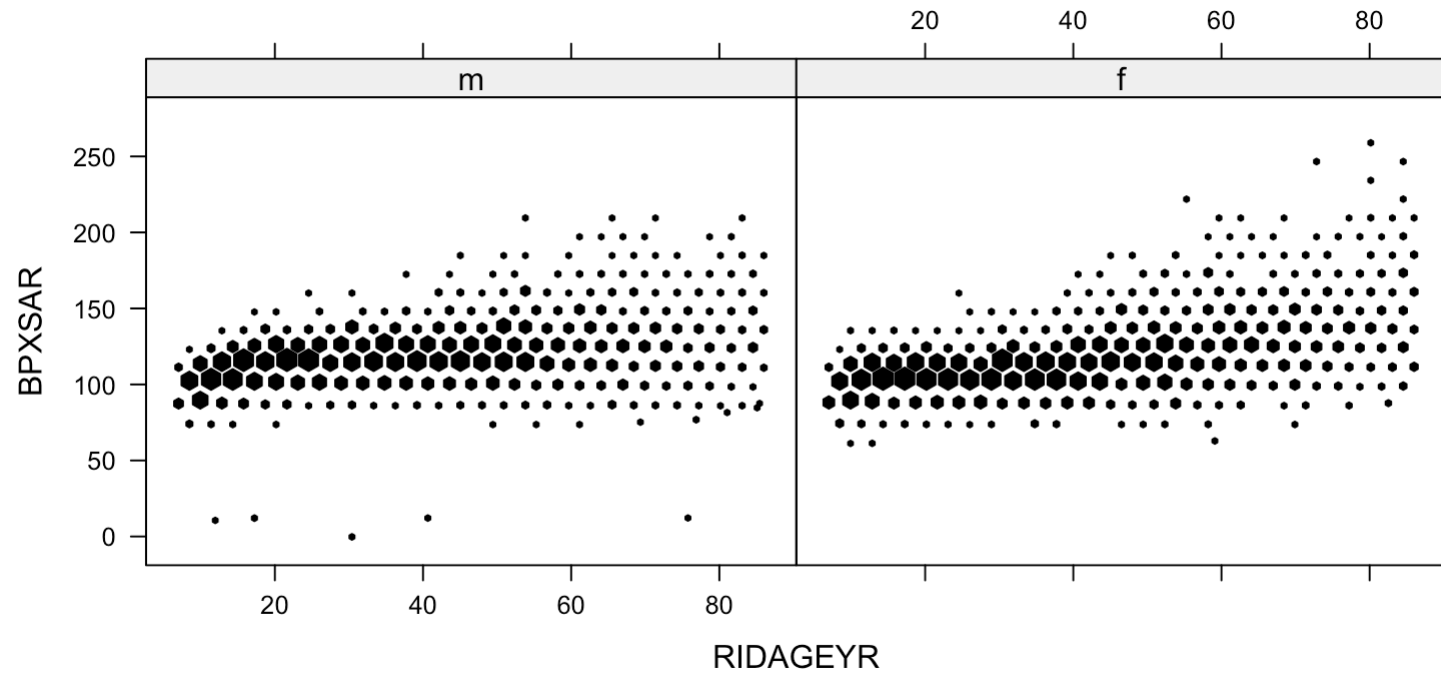
# Survey definition

```
nhanes$fouryearwt <- nhanes$WTDRD1/2
nhanesdes <- svydesign(id=~SDMVPSU, strata=~SDMVSTRA,
  weights=~fouryearwt, nest=TRUE
  data=subset(nhanes, !is.na(WTDRD1)))
nhanesdes <- update(nhanesdes, sodium=DR1TSODI/1000
  potassium=DR1TPOTA/1000)
nhanesdes <- update(nhanesdes, namol = sodium/23,
  kmol= potassium/23)
nhanesdes <- update(nhanesdes, htn = (BPXSAR>140) | (BPXDAR>90))
```

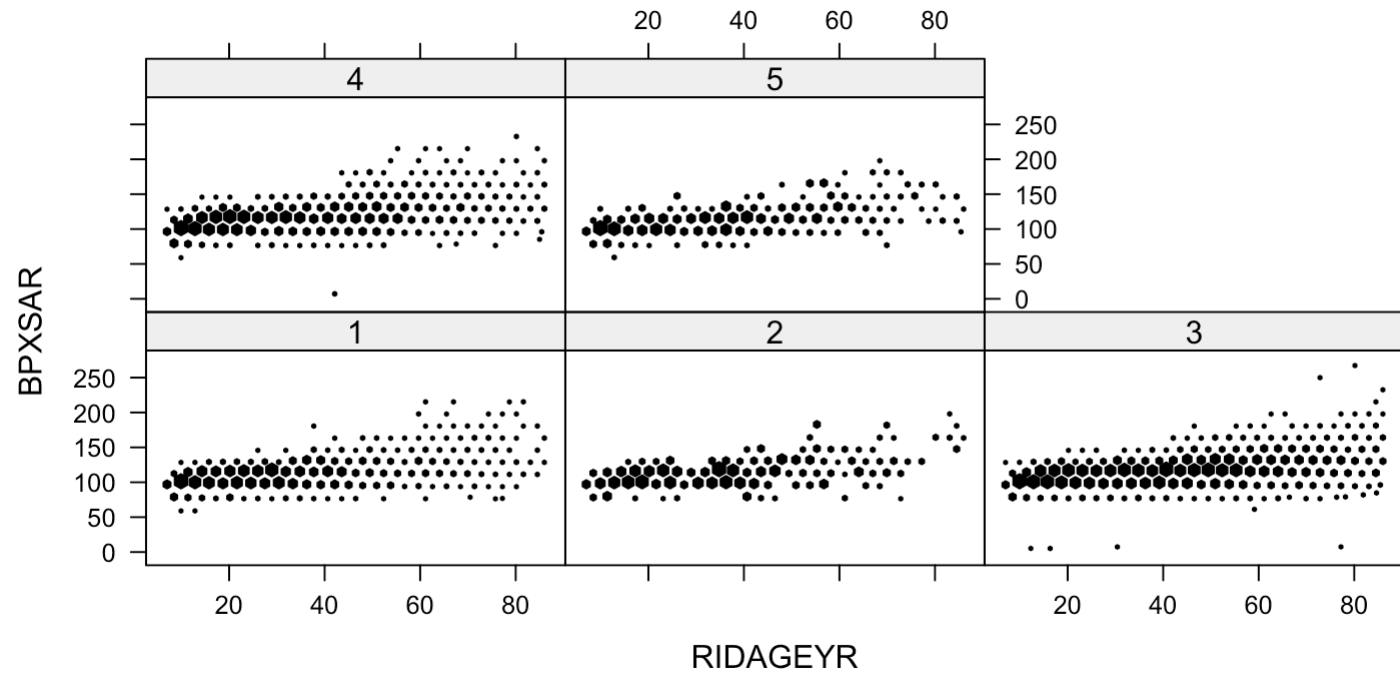
```
svycoplot(BPX SAR~RIDAGEYR | factor(RIAGENDR), design=nhanesdes, xbins=30)
```



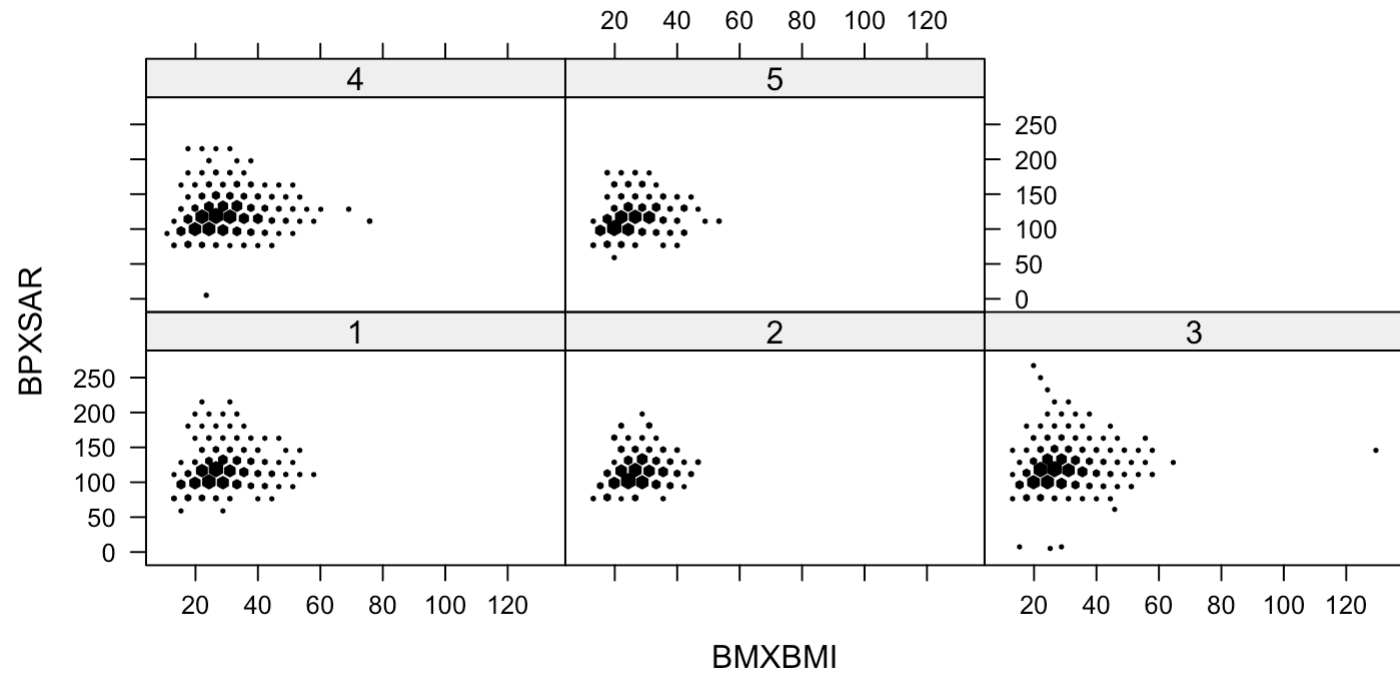
```
svyplot(BPX SAR ~ RIDAGEYR | factor(RIAGENDR, labels=c("m", "f")), design=nhanesdes, xbins=30)
```



```
svycoplot(BPX SAR~RIDAGEYR | factor(RIDRETH1), design=nhanesdes, xbins=30)
```



```
svycoplot(BPX SAR~BMXBMI | factor(RIDRETH1), design=nhanesdes, xbins=30)
```





# Linear regression example

## Unadjusted model

```
coef(summary(model<-svyglm(BPX SAR~sodium+potassium,  
  design=nhanesdes)))
```

	##	Estimate	Std. Error	t value	Pr(> t )
	## (Intercept)	120.3899	0.7105	169.436	1.039e-43
	## sodium	-0.6907	0.1658	-4.166	2.685e-04
	## potassium	0.7750	0.2655	2.919	6.853e-03

# Age/sex adjusted

```
coef(summary(model<-svyglm(BPX SAR~RIAGENDR+RIDAGEYR+sodium+potassium,
design=nhanesdes)))
```

##		Estimate	Std. Error	t value	Pr(> t )
##	(Intercept)	105.8284	1.22299	86.532	1.583e-33
##	RIAGENDR	-3.3004	0.37878	-8.713	3.437e-09
##	RIDAGEYR	0.4976	0.01149	43.298	9.159e-26
##	sodium	0.5943	0.16008	3.712	9.855e-04
##	potassium	-1.0884	0.18169	-5.990	2.524e-06

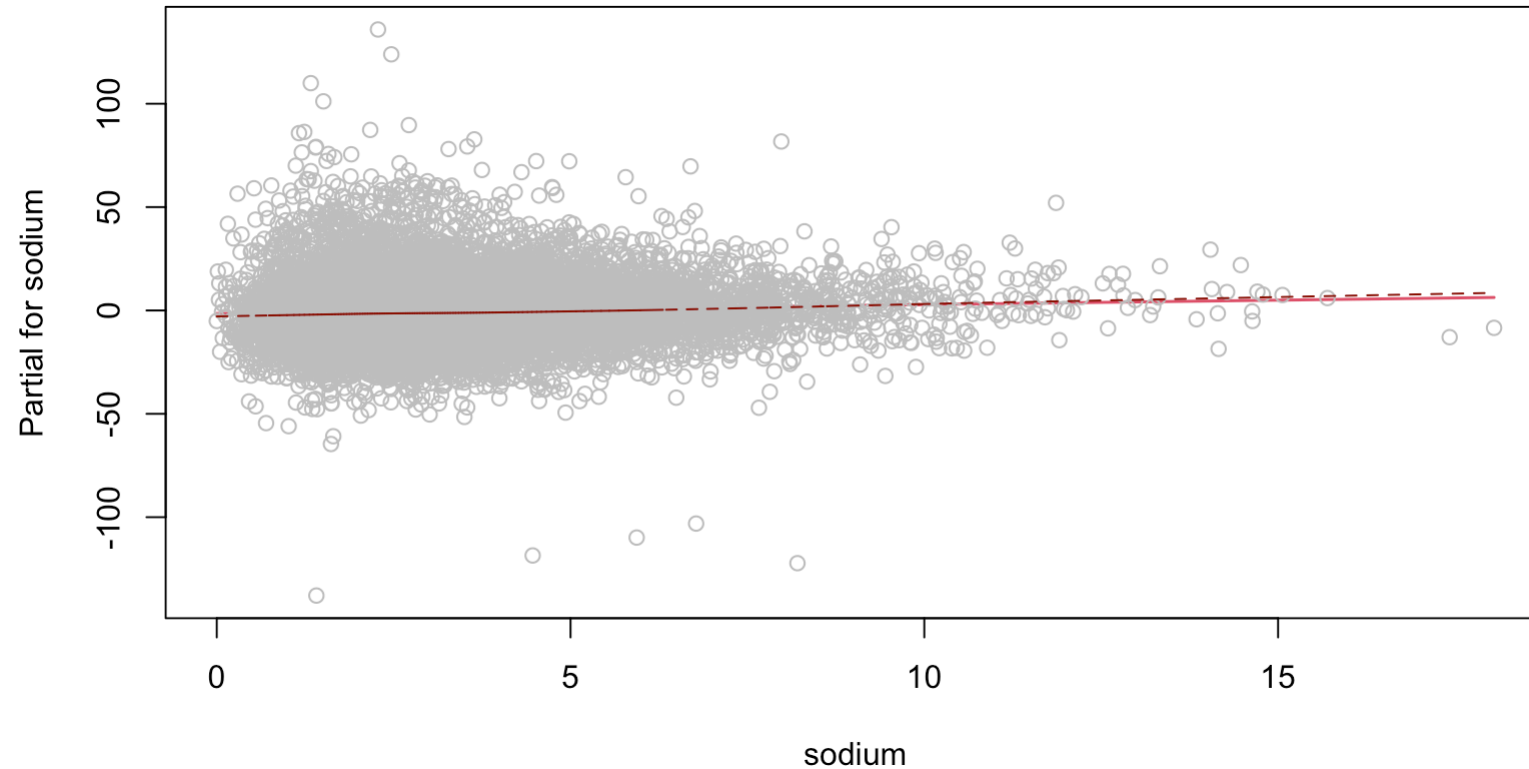
# More adjusted

```
coef(summary(model<-svyglm(BPX SAR~RIAGENDR+RIDAGEYR+factor(RIDRETH1)
+BMXBMI+sodium+potassium,
design=nhanesdes)))
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	97.0462	1.38921	69.8573	2.354e-26
## RIAGENDR	-3.3705	0.38507	-8.7529	1.891e-08
## RIDAGEYR	0.4651	0.01145	40.6221	1.905e-21
## factor(RIDRETH1)2	0.2377	1.35465	0.1755	8.624e-01
## factor(RIDRETH1)3	-0.5100	0.62820	-0.8119	4.260e-01
## factor(RIDRETH1)4	3.0297	0.64396	4.7049	1.207e-04
## factor(RIDRETH1)5	1.2947	0.88675	1.4600	1.591e-01
## BMXBMI	0.3710	0.03806	9.7478	3.024e-09
## sodium	0.4288	0.16190	2.6486	1.502e-02
## potassium	-0.8499	0.17133	-4.9606	6.578e-05

# Relationship is very weak: nonlinear?

```
termplot(model, terms=5, partial=TRUE, smooth=panel.smooth)
```



# Perhaps age is nonlinear?

```
library(splines)
model2<-svyglm(BPX SAR~RIAGENDR*ns(RIDAGEYR,4)+factor(RIDRETH1)
               +BMXBMI+sodium+potassium,
               design=nhanesdes)
coef(summary(model2))[c("sodium","potassium"),]
```

##		Estimate	Std. Error	t value	Pr(> t )
##	sodium	0.3082	0.1567	1.966	0.0694138
##	potassium	-0.7229	0.1636	-4.418	0.0005839

No real change. Weak association may be true or due to measurement error.

# Some tests

```
AIC(model,model2)
```

```
##      eff.p    AIC deltabar  
## [1,] 21.24 114804    2.124  
## [2,] 34.36 114248    2.021
```

```
regTermTest(model2, ~sodium+potassium)
```

```
## Wald test for sodium potassium  
## in svyglm(formula = BPXSAR ~ RIAGENDR * ns(RIDAGEYR, 4) + factor(RIDRETH1) +  
##      BMXBMI + sodium + potassium, design = nhanesdes)  
## F =  9.77 on  2 and 14 df: p= 0.0022
```

```
regTermTest(model2, ~factor(RIDRETH1),method="Wald")
```

```
## Wald test for factor(RIDRETH1)
## in svyglm(formula = BPXSAR ~ RIAGENDR * ns(RIDAGEYR, 4) + factor(RIDRETH1) +
##      BMXBMI + sodium + potassium, design = nhanesdes)
## F = 14.84 on 4 and 14 df: p= 0.000061
```

```
regTermTest(model2, ~factor(RIDRETH1),method="LRT")
```

```
## Working (Rao-Scott+F) LRT for factor(RIDRETH1)
## in svyglm(formula = BPXSAR ~ RIAGENDR * ns(RIDAGEYR, 4) + factor(RIDRETH1) +
##      BMXBMI + sodium + potassium, design = nhanesdes)
## Working 2logLR = 40.9 p= 0.00067
## (scale factors: 1.5 1.3 0.75 0.46 ); denominator df= 14
```

# Now, hypertension

```
model_htn<-svyglm(htn~RIAGENDR*ns(RIDAGEYR,4)+factor(RIDRETH1)  
                  +BMXBMI+sodium+potassium,  
                  design=nhanesdes, family=quasibinomial())
```



```
coef(summary(model_htn))
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	-13.80003	3.44255	-4.0087	0.001294
## RIAGENDR	3.82846	1.82753	2.0949	0.054850
## ns(RIDAGEYR, 4)1	11.78098	3.38031	3.4852	0.003641
## ns(RIDAGEYR, 4)2	6.46797	2.40486	2.6895	0.017614
## ns(RIDAGEYR, 4)3	24.46024	6.62715	3.6909	0.002420
## ns(RIDAGEYR, 4)4	3.60093	1.76480	2.0404	0.060636
## factor(RIDRETH1)2	0.10221	0.32679	0.3128	0.759073
## factor(RIDRETH1)3	-0.07610	0.14923	-0.5099	0.618050
## factor(RIDRETH1)4	0.42435	0.15785	2.6883	0.017657
## factor(RIDRETH1)5	0.47596	0.20787	2.2896	0.038090
## BMXBMI	0.03347	0.00829	4.0372	0.001223
## sodium	0.03142	0.03947	0.7961	0.439292
## potassium	-0.05121	0.04918	-1.0413	0.315408
## RIAGENDR:ns(RIDAGEYR, 4)1	-4.81717	1.85668	-2.5945	0.021203
## RIAGENDR:ns(RIDAGEYR, 4)2	-0.54762	1.35303	-0.4047	0.691789
## RIAGENDR:ns(RIDAGEYR, 4)3	-11.12228	3.53841	-3.1433	0.007187
## RIAGENDR:ns(RIDAGEYR, 4)4	1.87404	1.16184	1.6130	0.129053