nhanes_regression

July 25, 2024

1 Linear and logistic regression modeling - case study using NHANES data

[4]:
```
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import statsmodels.api as sm
import numpy as np
```

[5]:
```
da = pd.read_csv("nhanes_2015_2016.csv")

vars = ["BPXSY1", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "BMXBMI",␣
 ↪"SMQ020"]
da = da[vars].dropna()
```

1.1 Linear regression

Focus: Regression models in which systolic blood pressure (SBP) is the outcome (dependent) variable. SBP is an important indicator of cardiovascular health. It tends to increase with age, is greater for overweight people (i.e. people with greater body mass index or BMI), and also differs among demographic groups, for example among gender and ethnic groups.

Objective: Model SBP using linear regression.

1.1.1 Interpreting regression parameters in a basic model

Simple linear regression model with only one covariate, age, predicting SBP.

[6]:
```
model = sm.OLS.from_formula("BPXSY1 ~ RIDAGEYR", data=da)
result = model.fit()
result.summary()
```

[6]:

Dep. Variable:	BPXSY1	**R-squared:**	0.207
Model:	OLS	**Adj. R-squared:**	0.207
Method:	Least Squares	**F-statistic:**	1333.
Date:	Thu, 25 Jul 2024	**Prob (F-statistic):**	2.09e-259
Time:	17:09:16	**Log-Likelihood:**	-21530.
No. Observations:	5102	**AIC:**	4.306e+04
Df Residuals:	5100	**BIC:**	4.308e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	102.0935	0.685	149.120	0.000	100.751	103.436
RIDAGEYR	0.4759	0.013	36.504	0.000	0.450	0.501

Omnibus:	690.261	**Durbin-Watson:**	2.039
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1505.999
Skew:	0.810	**Prob(JB):**	0.00
Kurtosis:	5.112	**Cond. No.**	156.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

This fitted model implies that when comparing two people whose ages differ by one year, the older person will on average have 0.48 units higher SBP than the younger person. This difference is statistically significant, based on the p-value shown under the column labeled `P>|t|`. This means that there is strong evidence that there is a real association between between systolic blood pressure and age in this population.

SBP is measured in units of *millimeters of mercury*, expressed *mm/Hg*. In order to better understand the meaning of the estimated regression parameter 0.48, we can look at the standard deviation of SBP:

[7]:
```
da.BPXSY1.std()
```

[7]:
```
18.486559500781865
```

The standard deviation of around 18.5 describes the *unexplained variation* in systolic blood pressure values. It is large compared to the regression slope of 0.48, which describes the average difference between blood pressure values for two people whose ages differ by one year. It is important to note that the regression effect accumulates with age. Comparing a 40 year-old person to a 60 year-old person, there is a 20 year difference in age, which translates into a $20 \cdot 0.48 = 9.6$ unit difference in average SBP between these two people. This difference is around half of one standard deviation. Thus, while there is a substantial tendency for blood pressure to increase with age, there is also a great deal of variation among people with the same age – we should not be surprised to find, say, a 40 year old with greater blood pressure than a 60 year old.

1.1.2 R-squared and correlation

[8]:
```
cc = da[["BPXSY1", "RIDAGEYR"]].corr()
print(cc.BPXSY1.RIDAGEYR**2)
```

```
0.2071545962518702
```

The primary summary statistic for assessing the strength of a predictive relationship in a linear regression model is the *R-squared*, which is shown to be 0.207 in the regression output above. This means that 21% of the variation in SBP is explained by age.

1.1.3 Adding a second variable to the linear model

[9]:
```
da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
```

Fit the linear model in which age and gender are both serving as predictors of systolic blood pressure:

[10]:
```
model = sm.OLS.from_formula("BPXSY1 ~ RIDAGEYR + RIAGENDRx", data=da)
result = model.fit()
result.summary()
```

[10]:

Dep. Variable:	BPXSY1	**R-squared:**	0.215
Model:	OLS	**Adj. R-squared:**	0.214
Method:	Least Squares	**F-statistic:**	697.4
Date:	Thu, 25 Jul 2024	**Prob (F-statistic):**	1.87e-268
Time:	17:09:19	**Log-Likelihood:**	-21505.
No. Observations:	5102	**AIC:**	4.302e+04
Df Residuals:	5099	**BIC:**	4.304e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	100.6305	0.712	141.257	0.000	99.234	102.027
RIAGENDRx[T.Male]	3.2322	0.459	7.040	0.000	2.332	4.132
RIDAGEYR	0.4739	0.013	36.518	0.000	0.448	0.499

Omnibus:	706.732	**Durbin-Watson:**	2.036
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1582.730
Skew:	0.818	**Prob(JB):**	0.00
Kurtosis:	5.184	**Cond. No.**	168.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The model that was fit above uses both age and gender to explain the variation in SBP. It finds that two people with the same gender whose ages differ by one year tend to have blood pressure values differing by 0.47 units. Note that thisis essentially the same age parameter (0.48) that we found above in the model based on age alone. This model also shows us that comparing a male and a female person of the same age, the male will on average have 3.23 units greater SBP than the female.

We noted above that the regression coefficient for age did not change by much when we added gender to the model.

Confirming below that gender and age are nearly uncorrelated in this data set (the correlation of around -0.02 is negligible). Thus, it is expected that when we add gender to the model, the age coefficient is unaffected.

[11]:
```
da[["RIDAGEYR", "RIAGENDR"]].corr()
```

[11]:
```
           RIDAGEYR  RIAGENDR
RIDAGEYR   1.000000 -0.021398
RIAGENDR -0.021398  1.000000
```

1.1.4 Categorical variables and reference levels

The female level of the gender variable is the reference level. Female blood pressure is the default, and the coefficient for males (3.23) shifts the blood pressure by a certain amount for males only.

1.1.5 A linear model with three variables - BMI

Adding a third variable, body mass index (BMI), to the model predicting SBP.

[12]:
```
model = sm.OLS.from_formula("BPXSY1 ~ RIDAGEYR + BMXBMI + RIAGENDRx", data=da)
result = model.fit()
result.summary()
```

[12]:

Dep. Variable:	BPXSY1	**R-squared:**	0.228
Model:	OLS	**Adj. R-squared:**	0.228
Method:	Least Squares	**F-statistic:**	502.0
Date:	Thu, 25 Jul 2024	**Prob (F-statistic):**	8.54e-286
Time:	17:09:21	**Log-Likelihood:**	-21461.
No. Observations:	5102	**AIC:**	4.293e+04
Df Residuals:	5098	**BIC:**	4.296e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	91.5840	1.198	76.456	0.000	89.236	93.932
RIAGENDRx[T.Male]	3.5783	0.457	7.833	0.000	2.683	4.474
RIDAGEYR	0.4709	0.013	36.582	0.000	0.446	0.496
BMXBMI	0.3060	0.033	9.351	0.000	0.242	0.370

Omnibus:	752.325	**Durbin-Watson:**	2.040
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1776.087
Skew:	0.847	**Prob(JB):**	0.00
Kurtosis:	5.343	**Cond. No.**	316.

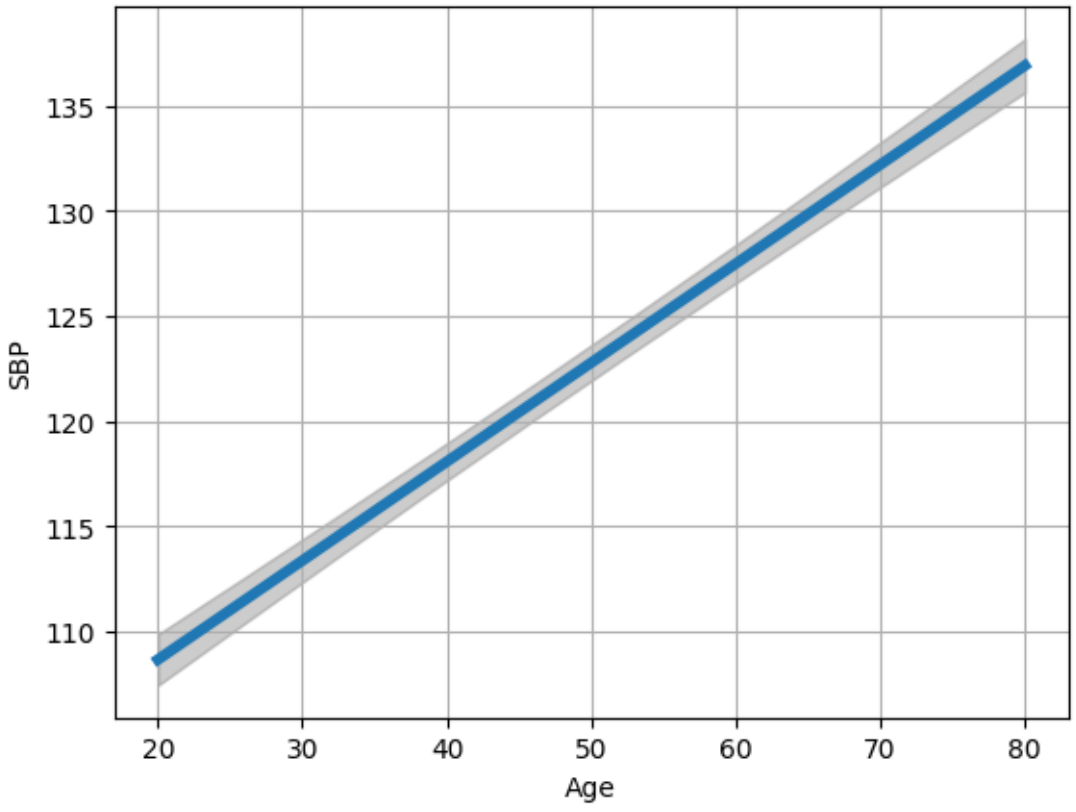
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

BMI is positively associated with SBP. Given two subjects with the same gender and age, and whose BMI differs by 1 unit, the person with greater BMI will have, on average, 0.31 units greater systolic blood pressure (SBP).

1.1.6 Visualization of regression models

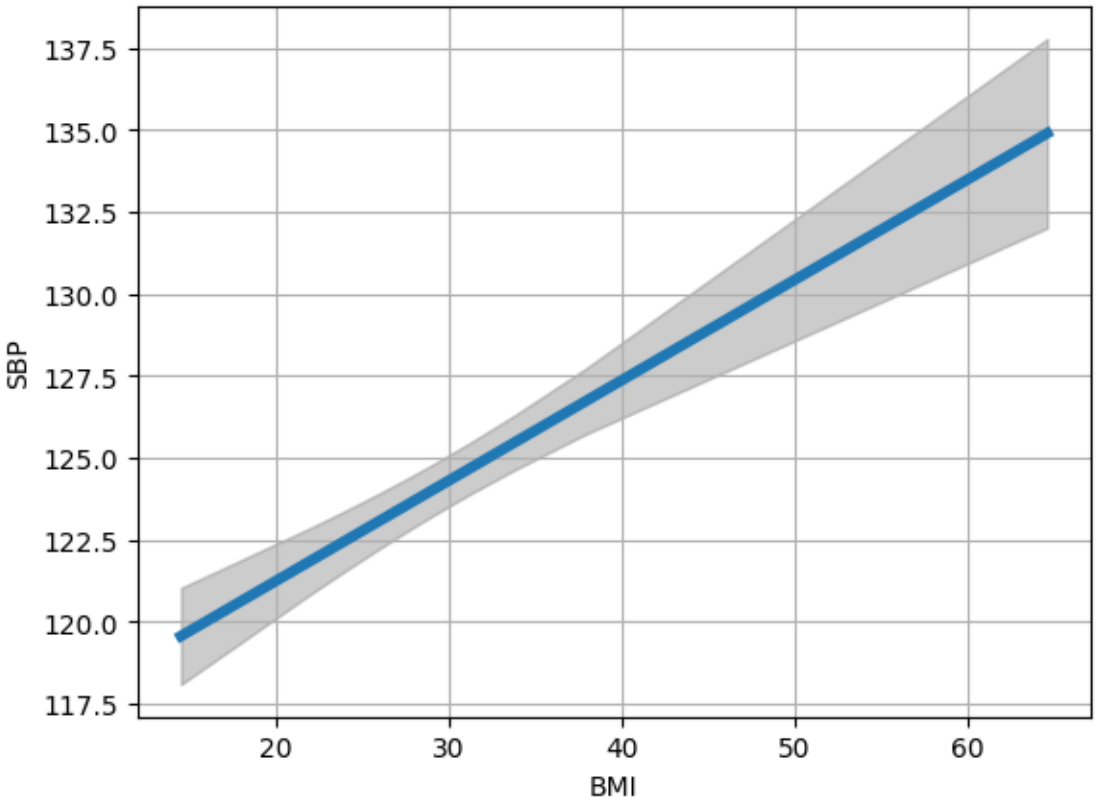
The graphs below show the relationship between expected SBP and age for women with BMI equal to 25.

[18]:
```
from statsmodels.sandbox.predict_functional import predict_functional

values = {"RIAGENDRx": "Female", "RIAGENDR": 1, "BMXBMI": 25,
          "DMDEDUC2": 1, "RIDRETH1": 1, "SMQ020": 1}

# The returned values are the predicted values (pr), the confidence bands (cb),
# and the function values (fv).
pr, cb, fv = predict_functional(result, "RIDAGEYR",
                values=values, ci_method="simultaneous")

ax = sns.lineplot(x=fv, y=pr, lw=4)
ax.grid(True)
ax.fill_between(fv, cb[:, 0], cb[:, 1], color='grey', alpha=0.4)
ax.set_xlabel("Age")
ax.set_ylabel("SBP");
```



Looking at the relationship between expected SBP and age for women of age 50.

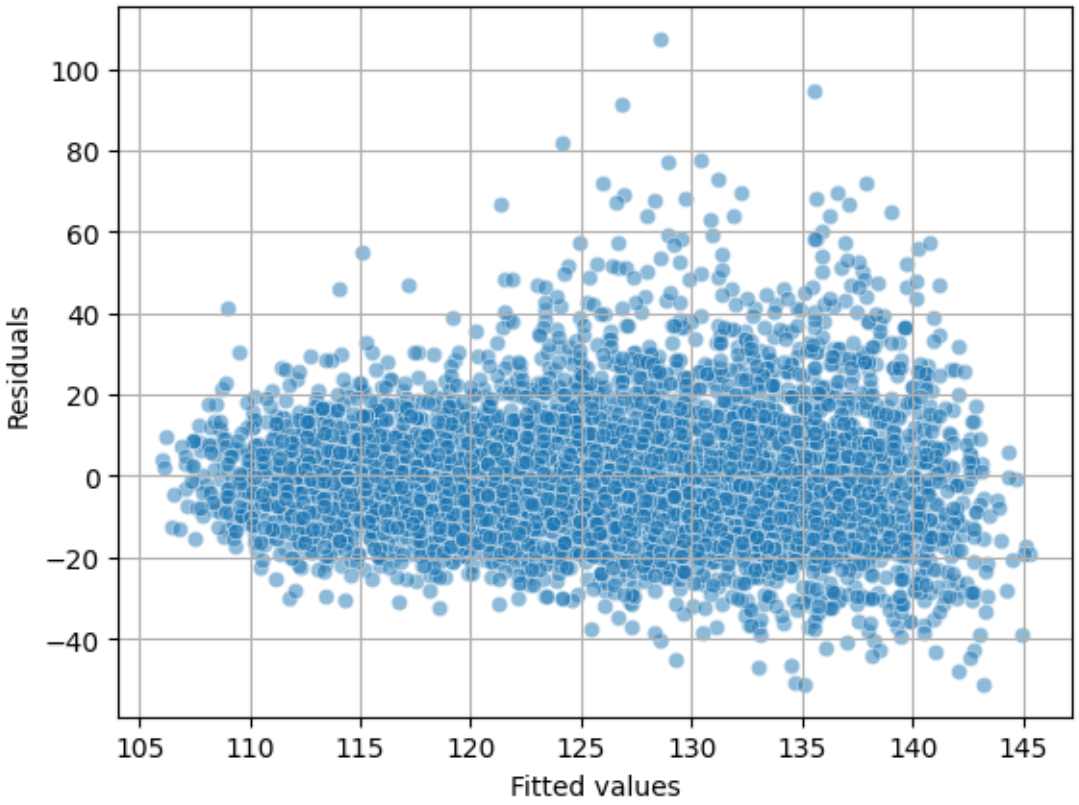
[19]:
```
del values["BMXBMI"]
values["RIDAGEYR"] = 50
pr, cb, fv = predict_functional(result, "BMXBMI",
                values=values, ci_method="simultaneous")

ax = sns.lineplot(x=fv, y=pr, lw=4)
ax.grid(True)
ax.fill_between(fv, cb[:, 0], cb[:, 1], color='grey', alpha=0.4)
ax.set_xlabel("BMI")
ax.set_ylabel("SBP");
```



The error band for BMI is notably wider than the error band for age, indicating that there is less certainty about the relationship between BMI and SBP compared to the relationship between age and SBP.

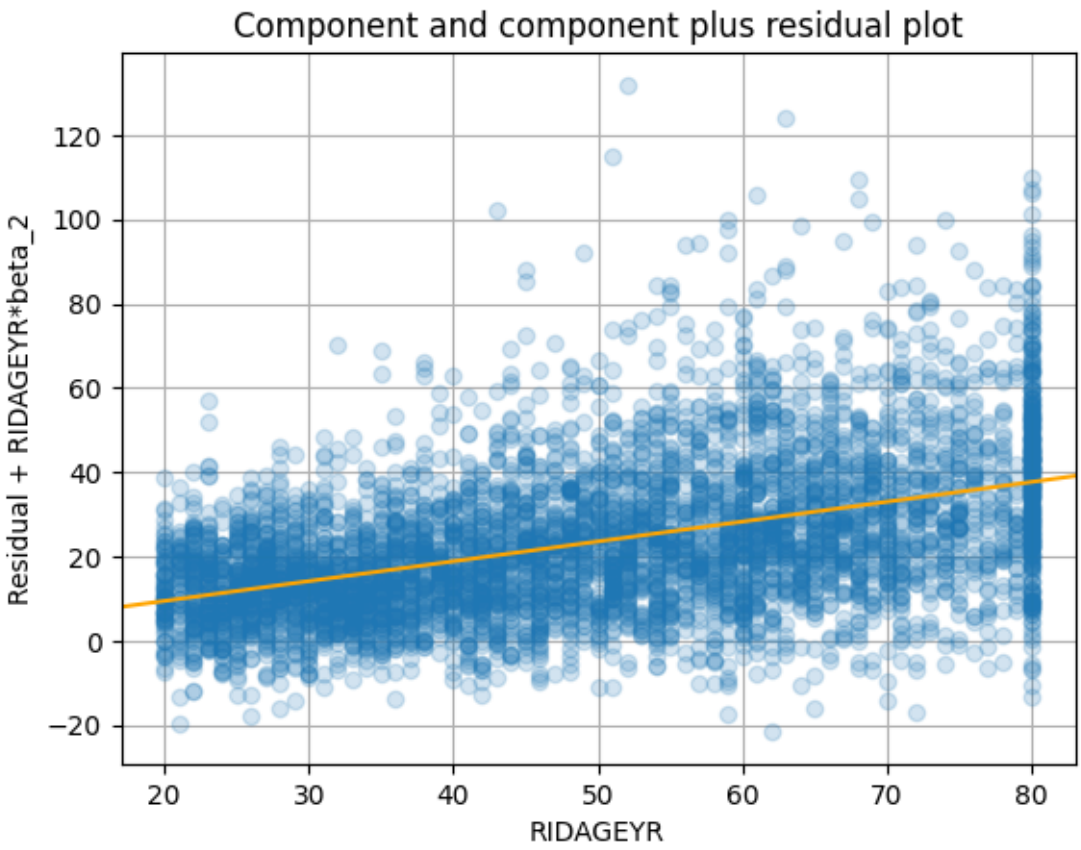
Below is the plot of residuals on fitted values for the NHANES data. It appears that we have a modestly increasing mean/variance relationship. That is, the scatter around the mean blood pressure is greater when the mean blood pressure itself is greater.

[20]:
```
pp = sns.scatterplot(x=result.fittedvalues, y=result.resid,alpha=0.5)
pp.grid(True)
pp.set_xlabel("Fitted values")
pp.set_ylabel("Residuals");
```



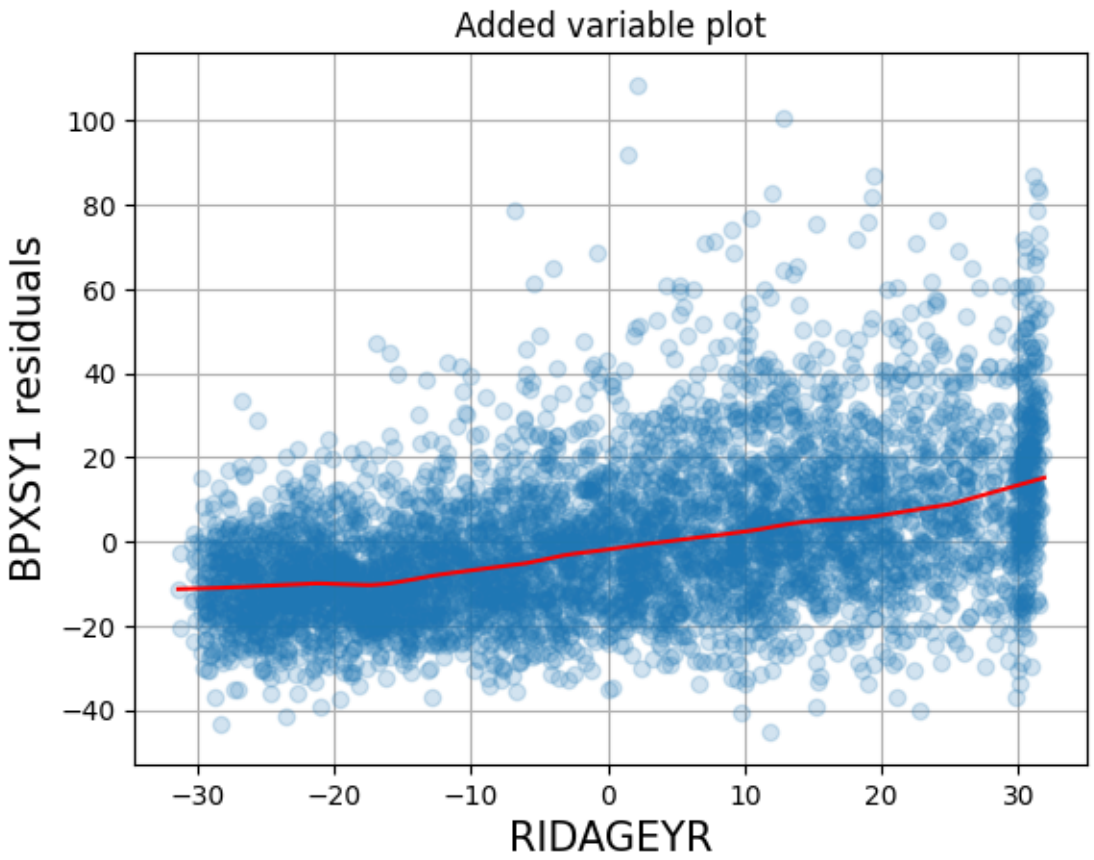
The plot below implies that when BMI and gender are held fixed, the average blood pressures of an 80 and 18 year old differ by around 30 mm/Hg. This plot also shows that the deviations from the mean are somewhat smaller at the low end of the range compared to the high end of the range.

[21]:
```
from statsmodels.graphics.regressionplots import plot_ccpr

ax = plt.axes()
plot_ccpr(result, "RIDAGEYR", ax)
ax.lines[0].set_alpha(0.2)
ax.lines[1].set_color('orange')
ax.grid(True);
```



Below is an added variable plot for age as a predictor of SBP. The red line is an estimte of the relationship between age and blood pressure. This implies that blood pressure increases slightly more slowly for people in theie 20s and early 30s, then begins increasing faster after that point.

[23]:
```
from statsmodels.graphics.regressionplots import add_lowess

model = sm.GLM.from_formula("BPXSY1 ~ RIDAGEYR + BMXBMI + RIAGENDRx", data=da)
result = model.fit()
result.summary()

fig = result.plot_added_variable("RIDAGEYR")
ax = fig.get_axes()[0]
ax.lines[0].set_alpha(0.2)
add_lowess(ax)
ax.grid(True);
```



1.2 Logistic regression

Below is a version of the variable in which smoking and non-smoking are coded as 1 and 0, respectively, and rare responses like *don't know* and *refused to answer* are coded as missing values.

[24]:
```
da["smq"] = da.SMQ020.replace({2: 0, 7: np.nan, 9: np.nan})
```

Looking at the odds of smoking for women and men separately:

[25]:
```
c = pd.crosstab(da.RIAGENDRx, da.smq).apply(lambda x: x/x.sum(), axis=1)
c["odds"] = c.loc[:, 1] / c.loc[:, 0]
c
```

[25]:

smq	0.0	1.0	odds
RIAGENDRx			
Female	0.680197	0.319803	0.470162
Male	0.467453	0.532547	1.139252

The probability that a woman has ever smoked is substantially lower than the probability that a man has ever smoked (32% versus 53%). This is reflected in the odds for a woman smoking being

much less than 1 (around 0.47), while the odds for a man smoking is around 1.14.

The odds ratio for smoking, comparing males to females, is around 2.4. In other words, a man has around 2.4 times greater odds of smoking than a woman.

[26]:
```
c.odds.Male / c.odds.Female
```

[26]:
```
2.423105552613186
```

Below is the log odds for smoking history status of females and males in the NHANES data. The negative log odds for females reflects that fact that substantially less than 50% of females have a history of smoking. The log odds for males is closer to 0, consistent with around half of males having a history of smoking.

[27]:
```
c["logodds"] = np.log(c.odds)
c
```

[27]:

smq	0.0	1.0	odds	logodds
RIAGENDRx				
Female	0.680197	0.319803	0.470162	-0.754679
Male	0.467453	0.532547	1.139252	0.130371

1.2.1 A basic logistic regression model

The dependent variable of this initial model is smoking status, and the only covariate is gender.

[28]:
```
model = sm.GLM.from_formula("smq ~ RIAGENDRx", family=sm.families.Binomial(),␣
↪data=da)
result = model.fit()
result.summary()
```

[28]:

Dep. Variable:	smq	**No. Observations:**	5094
Model:	GLM	**Df Residuals:**	5092
Model Family:	Binomial	**Df Model:**	1
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3350.6
Date:	Thu, 25 Jul 2024	**Deviance:**	6701.2
Time:	17:47:44	**Pearson chi2:**	5.09e+03
No. Iterations:	4	**Pseudo R-squ. (CS):**	0.04557
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.7547	0.042	-18.071	0.000	-0.837	-0.673
RIAGENDRx[T.Male]	0.8851	0.058	15.227	0.000	0.771	0.999

1.2.2 Adding additional covariates

Below is a logistic regression for smoking status using age (RIDAGEYR) and gender as covariates.

[30]:
```
model = sm.GLM.from_formula("smq ~ RIDAGEYR + RIAGENDRx", family=sm.families.
  ↪Binomial(), data=da)
result = model.fit()
result.summary()
```

[30]:

Dep. Variable:	smq	**No. Observations:**	5094
Model:	GLM	**Df Residuals:**	5091
Model Family:	Binomial	**Df Model:**	2
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3296.6
Date:	Thu, 25 Jul 2024	**Deviance:**	6593.2
Time:	17:49:09	**Pearson chi2:**	5.10e+03
No. Iterations:	4	**Pseudo R-squ. (CS):**	0.06558
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-1.6166	0.095	-16.985	0.000	-1.803	-1.430
RIAGENDRx[T.Male]	0.8920	0.059	15.170	0.000	0.777	1.007
RIDAGEYR	0.0172	0.002	10.289	0.000	0.014	0.021

Adding age to the model leads to a very small shift in the gender parameter (it changed from 0.885 to 0.892). This fitted model suggests that older people are more likely to have a history of smoking than younger people. The log odds for smoking increases by 0.017 for each year of age. The greater prevalence of smoking history among older people could be partly due to the definition of smoking status that we are using here – an older person has had more time to smoke 99 cigarettes than a younger person.

1.2.3 A logistic regression model with three predictors

Including educational attainment as a predictor:

[32]:
```
da["DMDEDUC2x"] = da.DMDEDUC2.replace({1: "lt9", 2: "x9_11", 3: "HS", 4:␣
  ↪"SomeCollege",
                                       5: "College", 7: np.nan, 9: np.nan})

model = sm.GLM.from_formula("smq ~ RIDAGEYR + RIAGENDRx + DMDEDUC2x", family=sm.
  ↪families.Binomial(), data=da)
result = model.fit()
result.summary()
```

[32]:

Dep. Variable:	smq	**No. Observations:**	5093
Model:	GLM	**Df Residuals:**	5086
Model Family:	Binomial	**Df Model:**	6
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3201.2
Date:	Thu, 25 Jul 2024	**Deviance:**	6402.4
Time:	17:54:54	**Pearson chi2:**	5.10e+03
No. Iterations:	4	**Pseudo R-squ. (CS):**	0.09976
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-2.3060	0.114	-20.174	0.000	-2.530	-2.082
RIAGENDRx[T.Male]	0.9096	0.060	15.118	0.000	0.792	1.028
DMDEDUC2x[T.HS]	0.9434	0.090	10.521	0.000	0.768	1.119
DMDEDUC2x[T.SomeCollege]	0.8322	0.084	9.865	0.000	0.667	0.998
DMDEDUC2x[T.lt9]	0.2662	0.109	2.438	0.015	0.052	0.480
DMDEDUC2x[T.x9_11]	1.0986	0.107	10.296	0.000	0.889	1.308
RIDAGEYR	0.0183	0.002	10.582	0.000	0.015	0.022

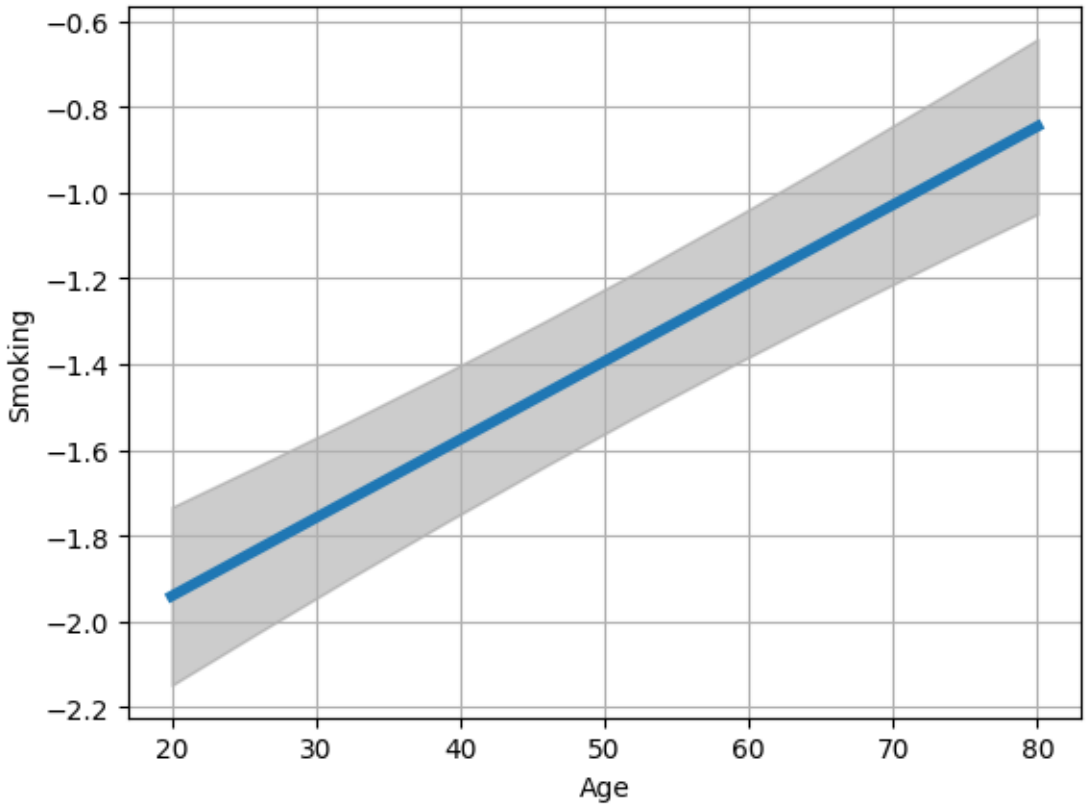
People with a college degree have the lowest rate of smoking, followed by people with less than 9 years of schooling, then (after a large gap) people with some college, then people with a high school degree (and no college), and finally (with the greatest rate of smoking), people with 9-11 years of schooling. The overall story here is that smoking rates are much lower for people who graduated from college or did not start high school, presumably for very different reasons. On the other hand, people with some high school, people who completed high school, and people who began but did not complete college have much higher rates of smoking.

1.2.4 Visualization of the fitted models

Focusing on how the smoking rate varies with age, and restricting the population to female college graduates.

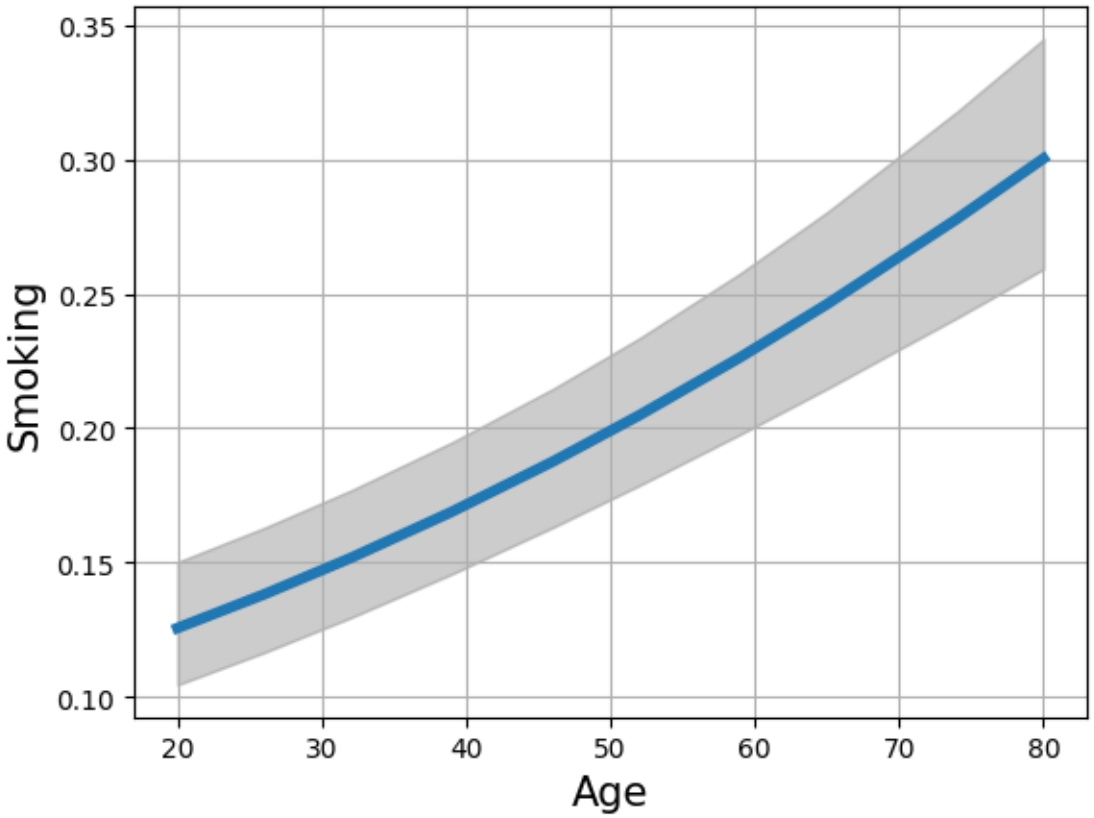
The following plot shows the fitted log odds (or logit) probability for the smoking outcome as a function of age. The grey band is a simultaneous 95% simultaneous confidence band.

[33]:
```
values = {"RIAGENDRx": "Female", "RIAGENDR": 1, "BMXBMI": 25,
          "DMDEDUC2": 1, "RIDRETH1": 1, "SMQ020": 1,
          "DMDEDUC2x": "College", "BPXSY1": 120}

pr, cb, fv = predict_functional(result, "RIDAGEYR",
                values=values, ci_method="simultaneous")

ax = sns.lineplot(x=fv, y=pr, lw=4)
ax.fill_between(fv, cb[:, 0], cb[:, 1], color='grey', alpha=0.4)
ax.set_xlabel("Age")
ax.set_ylabel("Smoking")
ax.grid(True);
```



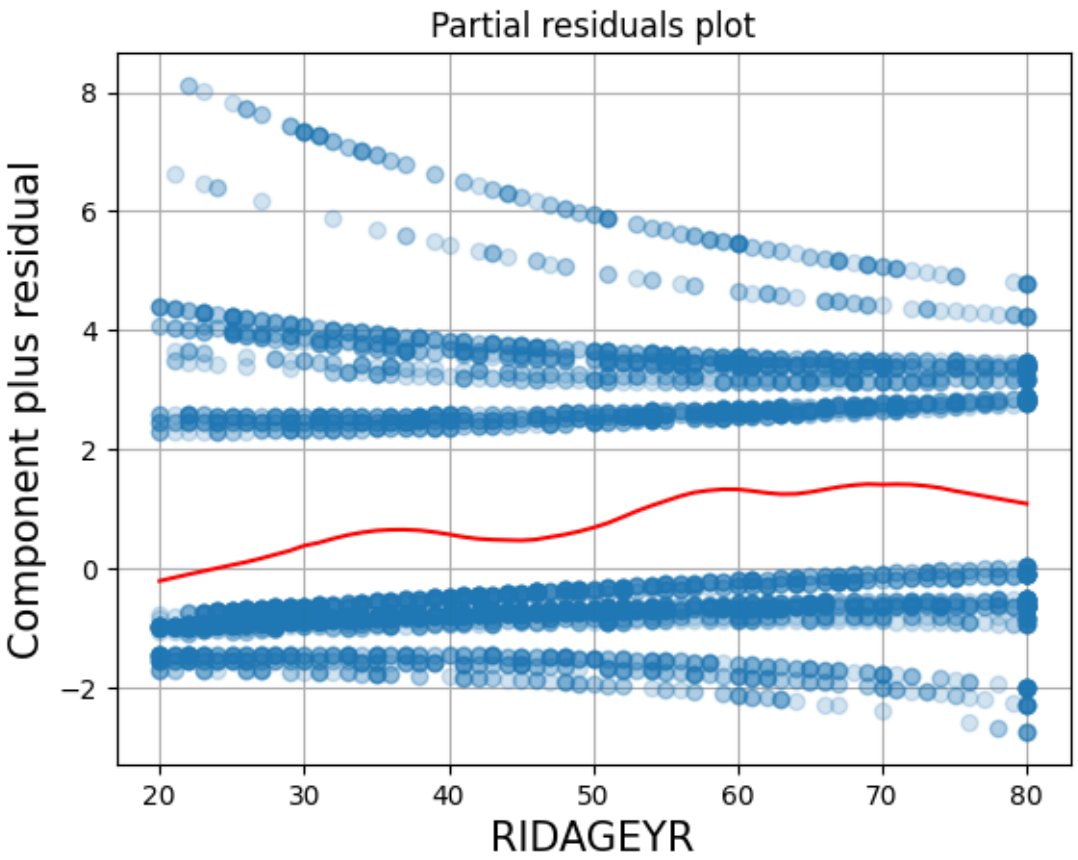
Displaying the same plot in terms of probabilities instead of in terms of log odds:

[34]:
```
pr1 = 1 / (1 + np.exp(-pr))
cb1 = 1 / (1 + np.exp(-cb))
ax = sns.lineplot(x=fv, y=pr1, lw=4)
ax.fill_between(fv, cb1[:, 0], cb1[:, 1], color='grey', alpha=0.4)
ax.set_xlabel("Age", size=15)
ax.set_ylabel("Smoking", size=15)
ax.grid(True);
```

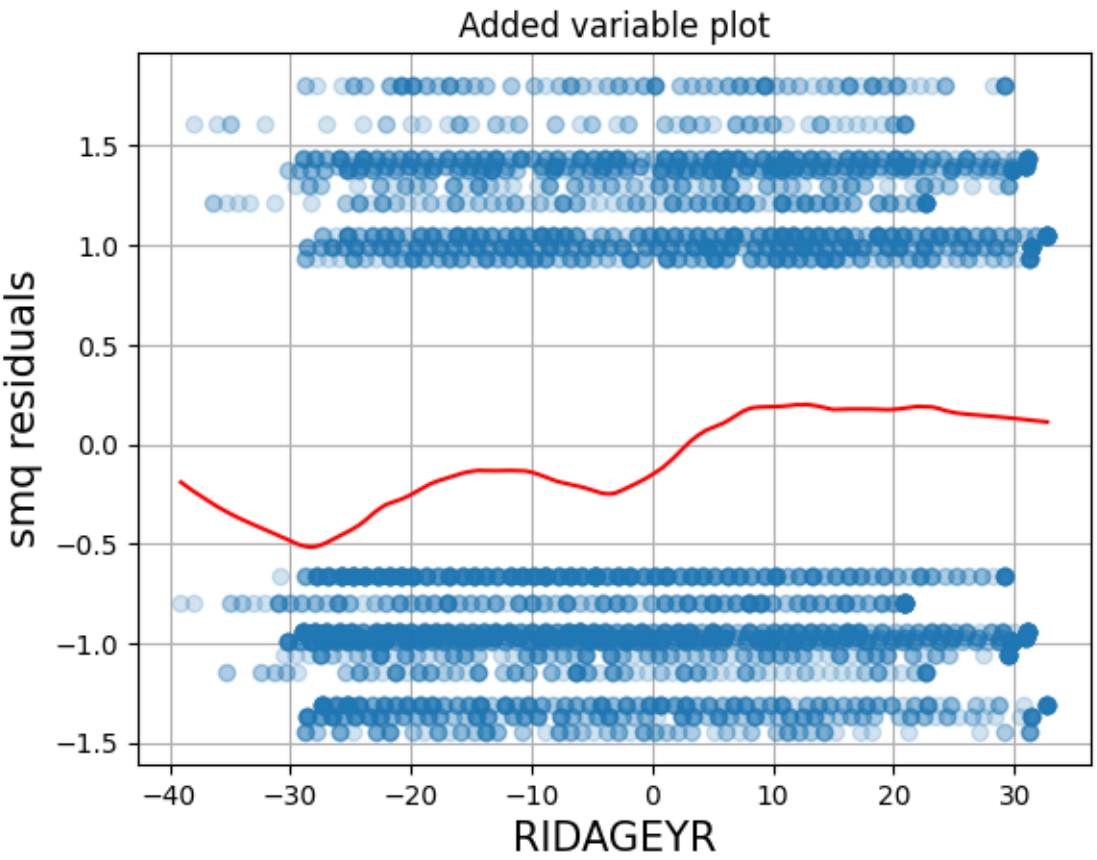


Partial residual plots, added variable plots, and CERES plots:

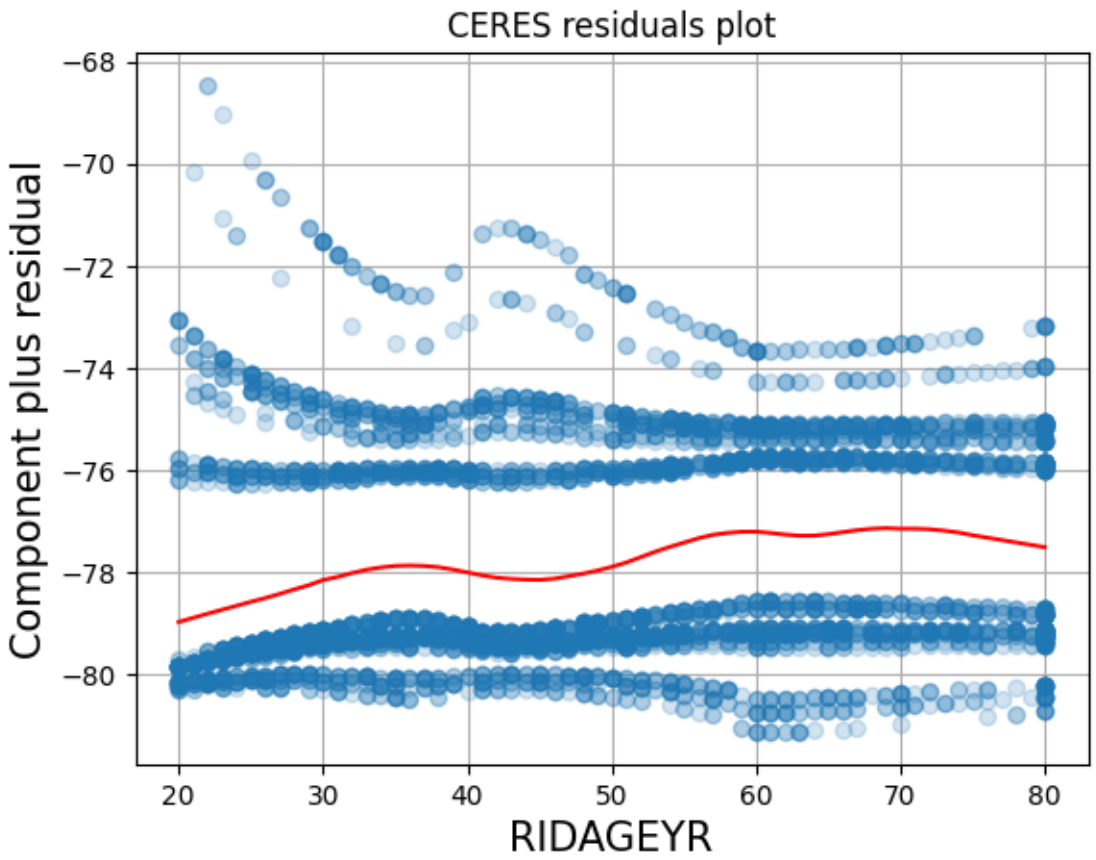
[35]:
```
fig = result.plot_partial_residuals("RIDAGEYR")
ax = fig.get_axes()[0]
ax.lines[0].set_alpha(0.2)
add_lowess(ax)
ax.grid(True);
```



[36]:
```
fig = result.plot_added_variable("RIDAGEYR")
ax = fig.get_axes()[0]
ax.lines[0].set_alpha(0.2)
add_lowess(ax)
ax.grid(True);
```



[37]:
```
fig = result.plot_ceres_residuals("RIDAGEYR")
ax = fig.get_axes()[0]
ax.lines[0].set_alpha(0.2)
add_lowess(ax)
ax.grid(True);
```



To identify major discrepancies between the fitted model and the population. The plots below suggest that smoking rates may rise slightly faster for people between the ages of 20 and 35, and again for people between the ages of 50 and 60, with a period of minimal increase between these age intervals.

2 Multilevel and marginal modeling, a case study with the NHANES data

[40]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import statsmodels.api as sm
import numpy as np
```

[43]:
```
da = pd.read_csv("nhanes_2015_2016.csv")

vars = ["BPXSY1", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "BMXBMI",
        "SMQ020", "SDMVSTRA", "SDMVPSU"]
da = da[vars].dropna()
```

2.1 Clustering structure in NHANES

We have access to "masked variance units" (MVUs), which are formed by combining subregions of different counties into artificial groups that are not geographically contiguous.

The MVU identifiers can be obtained by combining the `SDMVSTRA` and `SDMVPSU` identifiers, which we do next:

[45]:
```
da["group"] = 10*da.SDMVSTRA + da.SDMVPSU
```

2.2 Intraclass correlation

Similarity among observations within a cluster can be measured using the Intraclass Correlation Coefficient or ICC.

First looking at the ICC for systolic blood pressure:

[46]:
```
model = sm.GEE.from_formula("BPXSY1 ~ 1", groups="group",
           cov_struct=sm.cov_struct.Exchangeable(), data=da)
result = model.fit()
print(result.cov_struct.summary())
```

```
The correlation between two observations in the same cluster is 0.030
```

The estimated ICC is 0.03, which is small but not negligible.

Calculating the ICC for a number of different variables to get a more systematic view of the ICC values induced by clustering:

[48]:
```
da["smq"] = da.SMQ020.replace({2: 0, 7: np.nan, 9: np.nan})

for v in ["BPXSY1", "RIDAGEYR", "BMXBMI", "smq", "SDMVSTRA"]:
    model = sm.GEE.from_formula(v + " ~ 1", groups="group",
           cov_struct=sm.cov_struct.Exchangeable(), data=da)
    result = model.fit()
    print(v, result.cov_struct.summary())
```

```
BPXSY1 The correlation between two observations in the same cluster is 0.030
RIDAGEYR The correlation between two observations in the same cluster is 0.035
BMXBMI The correlation between two observations in the same cluster is 0.039
smq The correlation between two observations in the same cluster is 0.026
SDMVSTRA The correlation between two observations in the same cluster is 0.959
```

2.3 Conditional intraclass correlation

We know that older people have higher SBP than younger people. Also, some clusters may contain a slightly older or younger set of people than others. Thus, by controlling for age, the ICC might become smaller. This is shown in the next analysis:

[49]:
```
model = sm.GEE.from_formula("BPXSY1 ~ RIDAGEYR", groups="group",
           cov_struct=sm.cov_struct.Exchangeable(), data=da)
```

```
result = model.fit()
print(result.cov_struct.summary())
```

```
The correlation between two observations in the same cluster is 0.019
```

The ICC for SBP drops from 0.03 to 0.02. Assessing whether it drops even further when we add additional covariates that we know to be predictive of blood pressure.

[50]:
```
da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})

model = sm.GEE.from_formula("BPXSY1 ~ RIDAGEYR + RIAGENDRx + BMXBMI +␣
 ↪C(RIDRETH1)",
           groups="group",
           cov_struct=sm.cov_struct.Exchangeable(), data=da)
result = model.fit()
print(result.cov_struct.summary())
```

```
The correlation between two observations in the same cluster is 0.013
```

We see that the ICC has further reduced, to 0.013, due to controlling for these additional factors including ethnicity.

2.4 Multilevel models

Using multilevel modeling as an alternative way to accommodate dependence in clustered data.

Each cluster has a random effect that is shared by all observations in that cluster. For example, if SBP tends to be around 0.5 units higher in one cluster, then the random effect for that cluster would be 0.5, and it would add to the predicted SBP for every observation in the cluster.

[55]:
```
model = sm.MixedLM.from_formula("BPXSY1 ~ RIDAGEYR + RIAGENDRx + BMXBMI +␣
 ↪C(RIDRETH1)",
           groups="group", data=da)
result = model.fit()
result.summary()
```

[55]:

The variance for groups is estimated to be 3.615. This means that if we were to choose two groups at random, their random effects would differ on average by around 2.69. This is a sizable shift, comparable to the difference between females and males, or to around 6 years of aging.

Model:	MixedLM	Dependent Variable:	BPXSY1
No. Observations:	5102	Method:	REML
No. Groups:	30	Scale:	256.6952
Min. group size:	106	Log-Likelihood:	-21409.8702
Max. group size:	226	Converged:	Yes
Mean group size:	170.1		

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Intercept	92.173	1.402	65.752	0.000	89.426	94.921
RIAGENDRx[T.Male]	3.650	0.452	8.084	0.000	2.765	4.535
C(RIDRETH1)[T.2]	0.153	0.887	0.172	0.863	-1.586	1.891
C(RIDRETH1)[T.3]	-2.238	0.758	-2.954	0.003	-3.723	-0.753
C(RIDRETH1)[T.4]	3.098	0.836	3.707	0.000	1.460	4.737
C(RIDRETH1)[T.5]	-0.439	0.878	-0.500	0.617	-2.161	1.282
RIDAGEYR	0.474	0.013	36.482	0.000	0.449	0.500
BMXBMI	0.280	0.033	8.404	0.000	0.215	0.346
group Var	3.615	0.085				