Linear Regression with NHANES Data

This tutorial will be taking an excerpt from the NHANES case study provided in this week and reviewing the linear regression portion. We will cover model parameters such as coefficients, r-squared, and correlation. Additionally, we will construct models utilzing more than one predictor, introduce how categorical variables are handled, and generate visualizations of our models.

As with our previous work, we will be using the Pandas library for data management, the Numpy library for numerical calculations, and the Statsmodels library for statistical modeling.

We begin by importing the libraries that we will be using:

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

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

In [ ]:
```
# Drop unused columns, drop rows with any missing values.
vars = ["BPXSY1", "RIDAGEYR", "RIAGENDR", "RIDRETH1", "DMDEDUC2", "BMXBMI", "SMQ020"]
da = da[vars].dropna()
```

In [ ]:
```
da.head()
```

Linear regression

Simple Linear Regression with One Covariate

In [ ]:
```
### OLS Model of BPXSY1 with RIDAGEYR
model = sm.OLS.from_formula("BPXSY1 ~ RIDAGEYR", data=da)
result = model.fit()
result.summary()
```

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

R-squared and correlation

The primary summary statistic for assessing the strength of a predictive relationship in a 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. Note that this value is exactly the same as the squared Pearson correlation coefficient between SBP and age, as shown below.

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

Adding a Second Predictor

Now we will add gender to our initial model so we have two predictors, age and gender.

In [ ]:
```
# Create a labeled version of the gender variable
da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
```

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

The syntax `RIDAGEYR + RIAGENDRx` in the cell above does not mean that these two variables are literally added together. Instead, it means that these variables are both included in the model as predictors of blood pressure (`BPXSY1`).

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, which is essentially the same gender parameter that we found above in the model based on age alone. This model also shows us that comparing a man and a woman of the same age, the man will on average have 3.23 units greater SBP.

It is very important to emphasize that the age coefficient of 0.47 is only meaningful when comparing two people of the same gender, and the gender coefficient of 3.23 is only meaningful when comparing two people of the same age. Moreover, these effects are additive, meaning that if we compare, say, a 50 year old man to a 40 year old woman, the man's blood pressure will on average be around 3.23 + 10*0.47 = 7.93 units higher, with the first term in this sum being attributable to gender, and the second term being attributable to age.

We noted above that the regression coefficient for age did not change by much when we added gender to the model. It is important to note however that in general, the estimated coefficient of a variable in a regression model will change when other variables are added or removed. We see here that a coefficient is nearly unchanged if any variables that are added to or removed from the model are approximately uncorrelated with the other covariates that are already in the model.

Below we confirm that gender and age are nearly uncorrelated in this data set (the correlation of around -0.02 is negligible):

In [ ]:
```
# We need to use the original, numerical version of the gender
# variable to calculate the correlation coefficient.
da[["RIDAGEYR", "RIAGENDR"]].corr()
```

A model with three variables

Next we add a third variable, body mass index (BMI), to the model predicting SBP. BMI is a measure that is used to assess if a person has healthy weight given their height. BMXBMI is the NHANES variable containing the BMI value for each subject.

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

Not surprisingly, 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). Also note that after adding BMI to the model, the coefficient for gender

became somewhat greater. This is due to the fact that the three covariates in the model, age, gender, and BMI, are mutually correlated, as shown next:

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

Although the correlations among these three variables are not strong, they are sufficient to induce fairly substantial differences in the regression coefficients (e.g. the gender coefficient changes from 3.23 to 3.58). In this example, the gender effect becomes larger after we control for BMI - we can take this to mean that BMI was masking part of the association between gender and blood pressure. In other settings, including additional covariates can reduce the association between a covariate and an outcome.

Visualization of the Fitted Models

In this section we demonstrate some graphing techniques that can be used to gain a better understanding of a regression model that has been fit to data.

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

# Fix certain variables at reference values.  Not all of these
# variables are used here, but we provide them with a value anyway
# to prevent a warning message from appearing.
values = {"RIAGENDRx": "Female", "RIAGENDR": 1, "BMXBMI": 25,
          "DMDEDUC2": 1, "RIDRETH1": 1, "SMQ020": 1}

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

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

The analogous plot for BMI is shown next. Here we fix the gender as "female" and the age at 50, so we are looking at the relationship between expected SBP and age for women of age 50.

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

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

Below we show 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.

In [ ]:
```
pp = sns.scatterplot(result.fittedvalues, result.resid)
pp.set_xlabel("Fitted values")
_ = pp.set_ylabel("Residuals")
```

A "component plus residual plot" or "partial residual plot" is intended to show how the data would look if all but one covariate could be fixed at reference values. By controlling the values of these covariates, all remaining variation is due either to the "focus variable" (the one variable that is left unfixed, and is plotted on the horizontal axis), or to sources of variation that are unexplained by any of our covariates.

For example, the partial residual plot below shows how age (horizontal axis) and SBP (vertical axis) would be related if gender and BMI were fixed. Note that the origin of the vertical axis in these plots is not meaningful (we are not implying that anyone's blood pressure would be negative), but the differences along the vertical axis are meaningful. This plot 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, as discussed above, that the deviations from the mean are somewhat smaller at the low end of the range compared to the high end of the range. We also see that at the high end of the range, the deviations from the mean are somewhat right-skewed, with exceptionally high SBP values being more common than exceptionally low SBP values.

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

ax = plt.axes()
plot_ccpr(result, "RIDAGEYR", ax)
_ = ax.lines[0].set_alpha(0.2) # Reduce overplotting with transparency
```

Next we have a partial residual plot that shows how BMI (horizontal axis) and SBP (vertical axis) would be related if gender and age were fixed. Compared to the plot above, we see here that age is more uniformly distributed than BMI. Also, it appears that there is more scatter in the partial residuals for BMI compared to what we saw above for age. Thus there seems to be less information about SBP in BMI, although a trend certainly exists.

In [ ]:
```
ax = plt.axes()
plot_ccpr(result, "BMXBMI", ax)
_ = ax.lines[0].set_alpha(0.2)
```