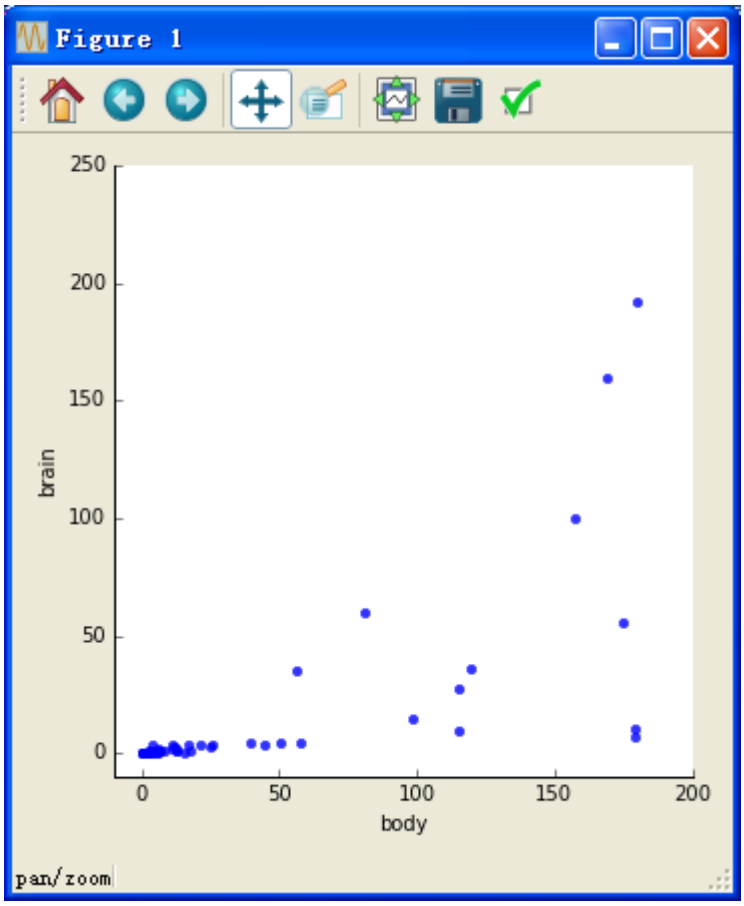
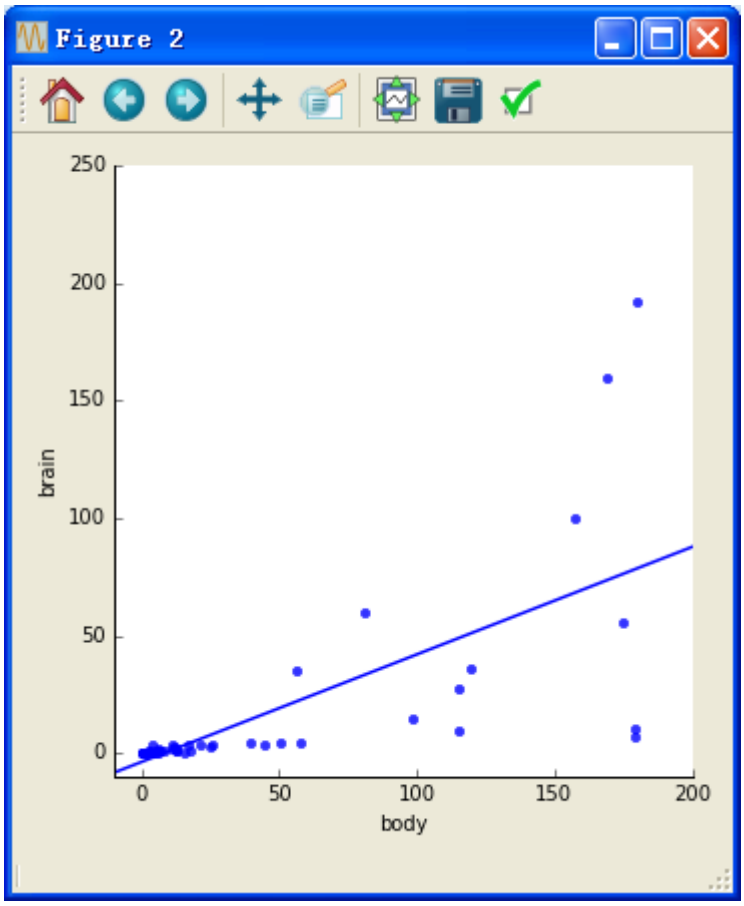
案例:bias/variance 的权衡

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib

# 62 种哺乳动物的平均脑重/体重数据集

url =
'http://people.sc.fsu.edu/~jburkardt/ datasets/regression/x01.txt'

df = pd.read_table(url,sep='\s+',skiprows=33,
                   names=['id','brain','body'],index_col='id')

df.head()
     brain   body
id
1     3.385   44.5
2     0.480   15.5
3     1.350    8.1
4   465.000  423.0
5    36.330  119.5

# 取一个子集,其中体重小于 200
df = df[df.body < 200]
df.shape
(51, 2)

# 创建散点图
sns.lmplot(x='body',y='brain',data=df,ci=None,fit_reg=False)

plt.xlim(-10,200)
plt.ylim(-10,250)
```

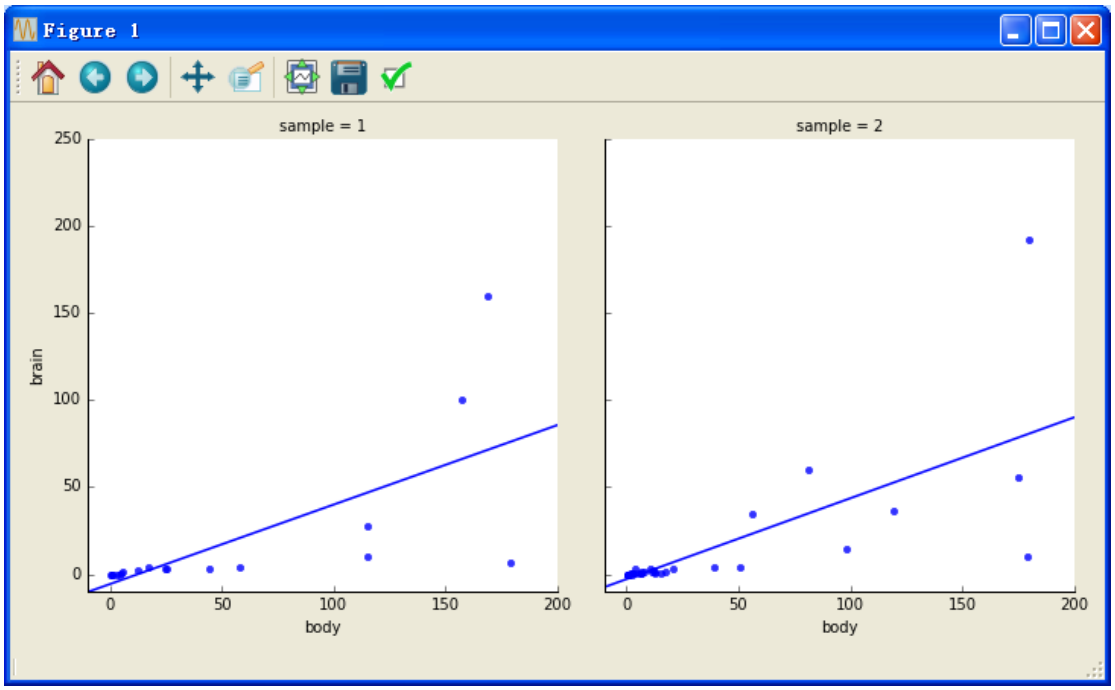



```
# 似乎有关系.目前先假设有正相关.
# 线性回归.利用 seaborn 执行一次多项式回归并绘图
sns.lmplot(x='body',y='brain',data=df,ci=None)

plt.xlim(-10,200)
plt.ylim(-10,250)
```

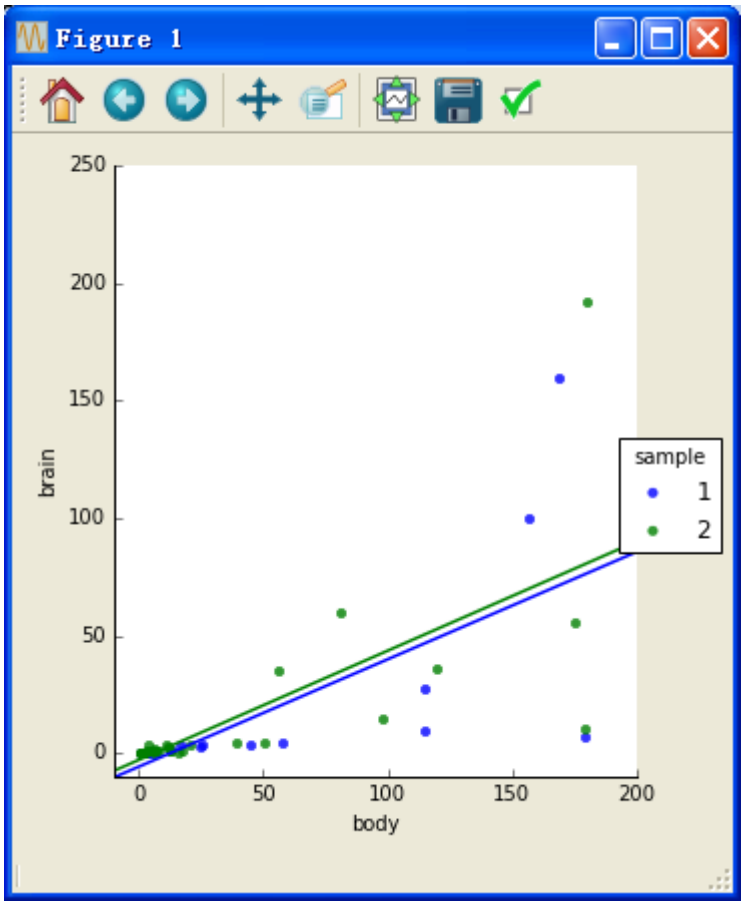



```
# 假设有一种哺乳动物平均体重为 100,预测其平均脑重(而不是直接测量)
# 根据直线,可预测脑重大约 45.

# 显然直线拟合的不是太好,可能并非最佳模型.即,线性回归模型是高偏差的
# 但线性回归一般具有低方差.我们通过实验来证明.

# 将数据集随机划分为两个样本

np.random.seed(12345)

# 将每一行随机划入 sample1 或 sample2
df['sample'] = np.random.randint(1,3,len(df))

df.head()
     brain   body  sample
id
1    3.385   44.5       1
2    0.480   15.5       2
3    1.350    8.1       2
5   36.330  119.5       2
6   27.660  115.0       1
```

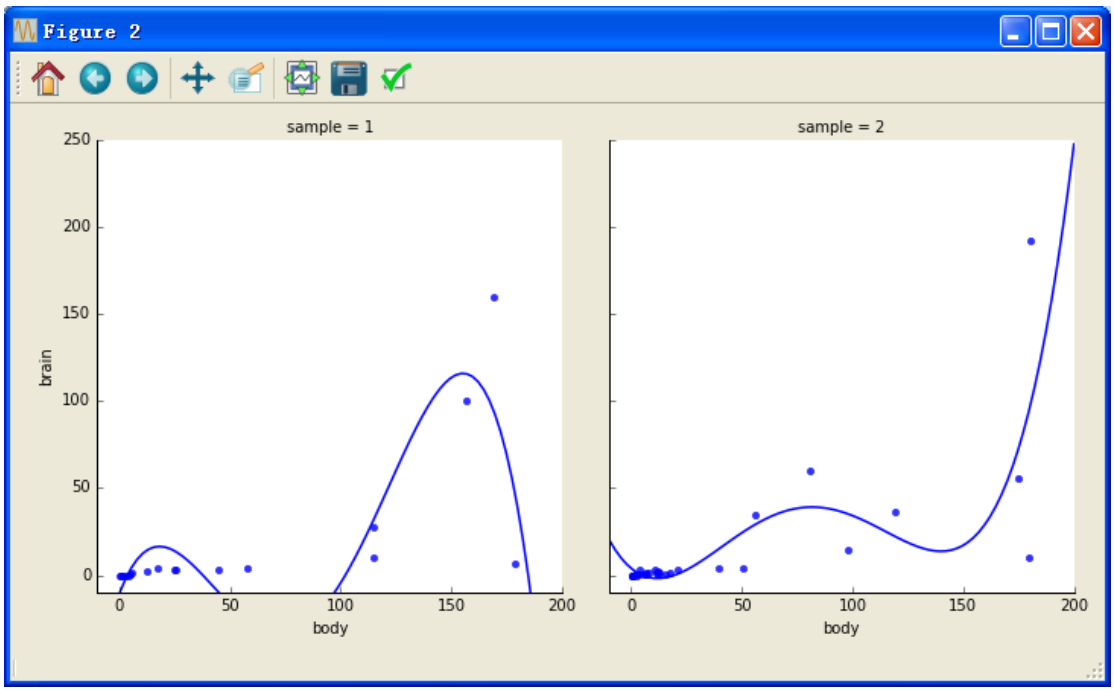
```
# 比较两个样本,似乎很不一样
df.groupby('sample')[['brain','body']].mean()
           brain       body
sample
1       18.113778  52.068889
2       13.323364  34.669091

# 下面绘制两个图:左图使用 sample1 数据,右图使用 sample2 数据
sns.lmplot(x='body',y='brain',data=df,ci=None,col='sample')

plt.xlim(-10,200)
plt.ylim(-10,250)
```



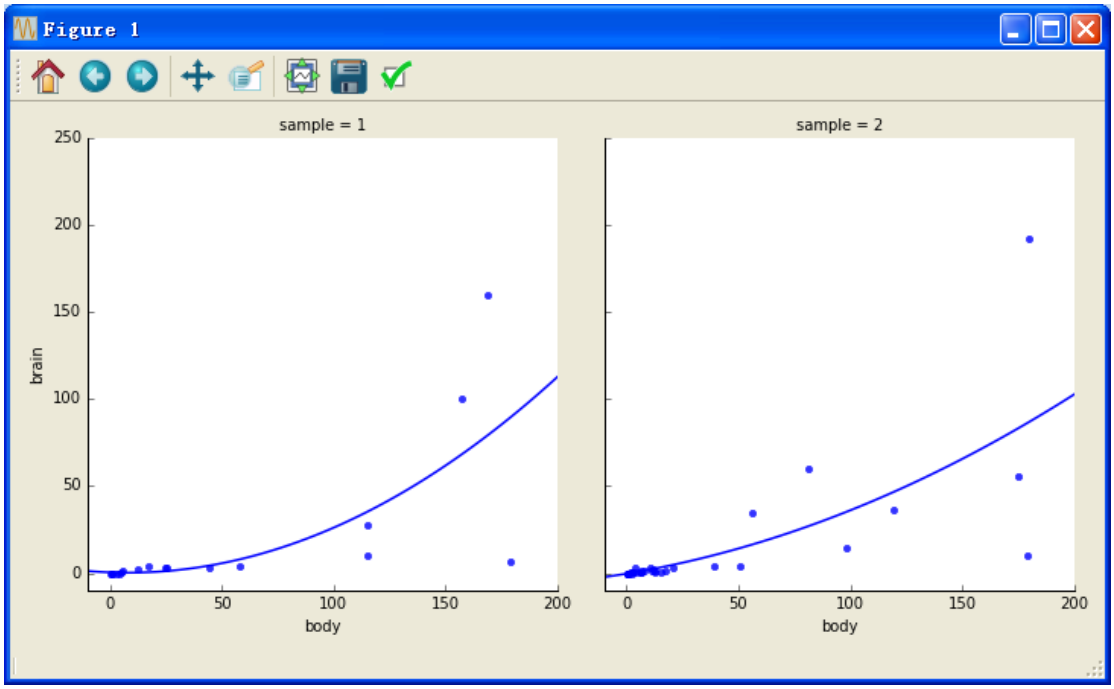

```
# 尽管没有一个训练数据是共有的,但两根直线看上去几乎一样.
# 将两根直线放在同一个图中看得更清楚(用不同颜色来分隔样本数据)

sns.lmplot(x='body',y='brain',data=df,ci=None,hue='sample')

plt.xlim(-10,200)
plt.ylim(-10,250)
```




```
# 可见两直线确实非常相似,尽管使用的是不同训练集.说明线性回归模型具有低方差.

# 下面增加模型复杂度,看看能否学到更多?
# 利用 seaborn 拟合四次多项式.曲线能更好地拟合数据.
sns.lmplot(x='body',y='brain',data=df,ci=None,col='sample',order=
4)

plt.xlim(-10,200)
plt.ylim(-10,250)
```




```
# 对来自同一总体的两个不同样本,四次多项式看上去很不一样,说明模型依赖于样本
# 这意味着模型的高方差. 根据不同样本,对体重 100 分别预测脑重 40 和 0
# 但这个模型是低偏差的,因为拟合数据很好
# 四次多项式模型没有发现很明显存在的正相关.第一个样本导致模型最后向下走,而第二
# 个样本导致模型最后向上走.所以四次多项式模型是不可预料的,根据不同训练集,表现
# 可能迥异.

# 我们的目标是权衡偏差/方差.
# 希望可以构造一个比线性模型少一点偏差,比四次多项式模型少一点方差的模型
# 试试二次多项式
sns.lmplot(x='body',y='brain',data=df,ci=None,col='sample',order=
2)

plt.xlim(-10,200)
plt.ylim(-10,250)
```



这个图较好地平衡了偏差和方差:在不同样本上模型相似,且拟合样本也不错