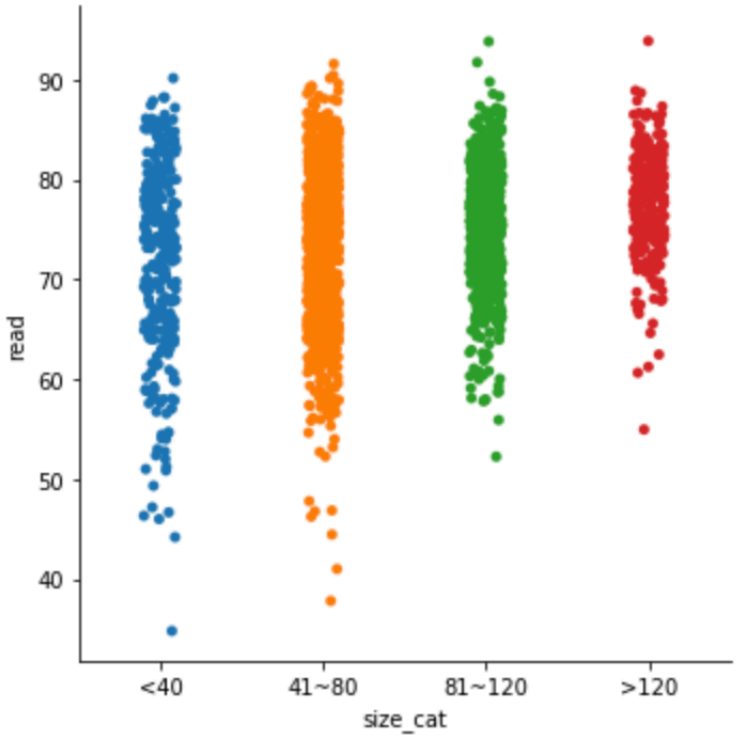
2021 中级计量经济学作业 4 参考答案

1. 课本 574 页习题 28.2

   问题简述：以色列的学校要求每班人数不可大于 40 人。一旦多于 40，则会分成两个班。此时，每个班的人数便会一下子减少许多。利用这个政策，可以考察班级大小是否会对学生的成绩产生影响。因此，入学人数 size 若是多于 40，则每个班的人数会变小。

   首先直观上可以先看一下在不同的 size 上，read 是否变化很大



   可以看到，size<40 时，read 偏低的值相对较多。read 最低的值对应的班级人数是 38。

```
    rd read size, z0(40)
```

```
. rd read size, z0(40)
Two variables specified; treatment is
assumed to jump from zero to one at Z=40.

 Assignment variable Z is size
 Treatment variable X_T unspecified
 Outcome variable y is read

Estimating for bandwidth 7.490981019337976
Estimating for bandwidth 3.745490509668988
Estimating for bandwidth 14.98196203867595
```

read	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwald	3.062374	3.763187	0.81	0.416	-4.313336	10.43808
lwald50	.9665485	6.159906	0.16	0.875	-11.10665	13.03974
lwald200	4.282076	2.465659	1.74	0.082	-.5505267	9.114679

断点在 40 时，可以看到，班级变小（也即 size 超过 40）时，成绩会越好，但结果是否显著和带宽有关。带宽的选择往往会对结果产生影响。如果数据呈现比较强的非线性，那么带宽应当选择小一些。由此处也可以看到，断点回归的结果也会和 hyperparameter 的选择有很强的关系（和动态面板类似）。

断点在 40 的倍数时，比如 80 时，结果如下：

```
    rd read size, z0(80)
```

```
. rd read size, z0(80)
Two variables specified; treatment is
assumed to jump from zero to one at Z=80.

 Assignment variable Z is size
 Treatment variable X_T unspecified
 Outcome variable y is read

Estimating for bandwidth 7.94315674033172
Estimating for bandwidth 3.97157837016586
Estimating for bandwidth 15.88631348066344
```

read	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwald	-1.935336	1.661789	-1.16	0.244	-5.192383	1.321711
lwald50	-5.176138	2.394469	-2.16	0.031	-9.869211	-.4830644
lwald200	-.2143482	1.235631	-0.17	0.862	-2.63614	2.207444

这个结果产生的原因可能是：因为在 40 时已做了分班，如果断点只设置在 80，那么在 80 时分班的效果会和 40 时的混在一起，因此实际上不能观察到分班结果了。

评论：一般的，计量模型越复杂，hyperparameter 的可操作性就越强，比如断点回归里的带宽，动态面板里的滞后工具变量阶数。在倾向得分匹配里也类似。而这些往往

没有“最好”的选择。所以一般的做法是查阅论文，看看计量学家的建议是什么，以及做同一类问题的其他学者用什么样的方法和 hyperparameter。另一方面，对于同一个问题，总会用各种模型从各个不同的角度考察，也是为了减少出错的概率。

2. housing.csv 是美国加州 1990 年的房价数据。请分别用线性回归和决策树模型（也可以用任意你觉得合理、有效的其他模型，或者 Ensembling）对房价做出预测。注意以下几个问题：

    1. 设定模型好坏判断标准 (evaluation metrics)
    2. 数据划分 training, validation 以及 test(train-validation-test split)
    3. 防止模型过拟合 (overfitting)
    4. 可自行创建新的自变量 (feature engineering)

参考答案见 housing-example.html 文件。