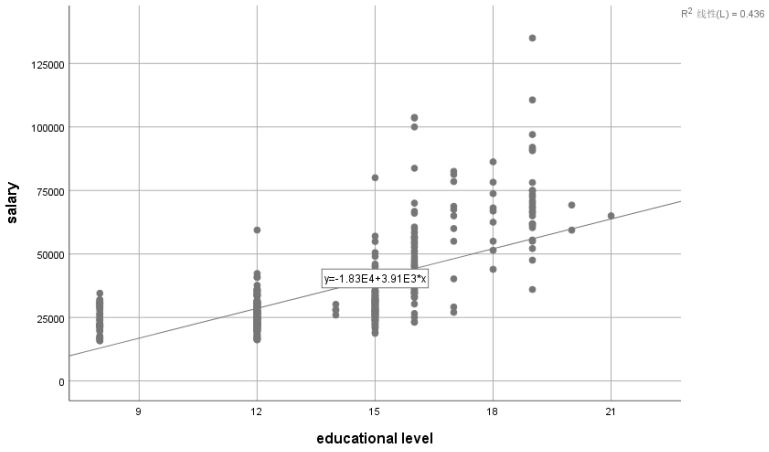
当前工资的影响因素及其预测

对于Employee data数据集，其中重要的变量有当前工资、受教育水平、开始工资、教育年限、工作年限、民族等。

【分析一】

当前的工资(salary)与教育水平(educ: educational level)两者可能相关，对此我们将对其相关程度进行探索。不过即使教育水平与当前工资有显著相关，也可能是由于教育水平不同的其开始工资(salbegin: beginning salary)不同造成的，对此我们将控制开始工资，再次探索教育水平与当前工资的关系。

先做一个当前工资与教育水平的散点图如下，可知这两个变量间有线性的关系趋势。



对当前工资和教育水平做Pearson相关可知相关性 $r = 0.661$，$p < 0.001$。则当前工资与教育水平存在相关关系，且为显著的高相关。可能教育水平越高，当前工资越高。

控制开始工资，对当前工资和教育水平做偏相关可知当前工资和教育水平相关性 $r = 0.281$，$p < 0.001$。即在控制开始工资的条件下，当前工资与教育水平为显著的低相关。可知，开始工资影响着教育水平与当前工资的相关关系,可能导致了教育水平与当前工资的高相关。

【分析二】

工作经验与教育年限一般也会与工资有一定关系。我们将尝试建立工作经验和教育年限对工资的回归方程模型，并评估该回归方程模型的有效性，以及工作经验和教育年限对当前工资的预测性，并比较这些变量对于工资的预测效果。此外，该数据集还含有民族变量，因此，我们也将试着探索一下工作经验和教育年限对当前工资的预测效果在少数民族和非少数民族之间是否具有差异性。

表1. 当前工资salary、教育年限、工作经验的平均值、标准差

	M	*SD*
salary	34419.6	17075.7
教育年限	13.5	2.9
工作经验	95.9	104.6

表2. 当前工资、教育年限与工作经验的相关分析表

	salary	教育年限	工作经验
salary	-		
教育年限	.661**	-	
工作经验	-.097*	-.252**	-

* p<0.05 , ** p<0.01.

可知教育年限与当前工资为显著的正的高相关，工作经验与当前工资为显著的负的低相关，工作经验与教育年限为显著的负的低相关。

（1）对教育年限、工作经验和当前工资做回归分析可得回归方程模型为：

工资 = -20978.30 + 4020.34*教育年限 + 12.07*工作经验

（2）ANOVA分析发现 F（2，471）=186.132，p<0.001，即该回归方程具有有效性和统计学意义。教育年限和工作经验与工资的多元相关系数 R=0.664，变异系数 R^2=0.441，说明当前工资的变异中44.1%可由教育年限和工作经验来解释。

（3） 教育年限的非标准化偏回归系数B=4020.34，标准化系数B=0.679，p <0.001，显著。工作经验非标准化偏回归系数B=12.07，标准化系数B=5.810，p=0.038，显著。因为教育年限的标准化系数大于工作经验的标准化系数，所以教育年限的预测效果更好一些。

（4）当前工资、教育年限、民族与工作经验的相关分析表如下：

	salary	教育年限	工作经验	民族
salary	-			
教育年限	.661**	-		
工作经验	-.097*	-.252**	-	
民族	-.177**	-.133**	.145**	-

* p<0.05 , ** p<0.01.

可知教育年限与当前工资为显著的正的高相关，工作经验与当前工资为显著的负的低相关，工作经验与教育年限为显著的负的低相关，民族与当前工资为显著的负的低相关，民族与教育年限为显著的负的低相关，民族与当前工资为显著的正的低相关。

ANOVA分析发现 F（3，470）=128.87，p<0.001，该回归模型有效。R=0.672，R^2=0.0.451，说明当前工资的变异中45.1%可由教育年限、工作经验和民族来解释。

工资的回归方程模型为：

工资 = -19424.23 + 14.03*教育年限 + 3958.88*工作经验 - 4158.56*民族

当民族为0，即民族为非少数民族时：

现有工资= -19424.23 + 14.03*教育年限 + 3958.88*工作经验

当民族为1，即民族为少数民族时：

现有工资 = -23582.79 + 14.03*教育年限 + 3958.88*工作经验

两个模型的回归系数一样，但截距不同，说明教育年限和工作经验为零时，员工的工资不同，即教育年限和工作经验对少数民族和非少数民族的预测不同。