

逻辑回归的代码实现与模型结果理解

Pandas、statsmodels

小胖

目录

ONE 初探数据

变量种类、数据可视化

TWO 搭建模型

构建模型、模型参数的稳定性

THREE 理解模型结果

发生比、边际效应

初探数据

数据简介

数据的变量说明

数据来自美国加州大学欧文分校

美国个人收入的普查数据

预测变量是年收入分类

预测变量



变量名	变量类型	说明
age	数值型变量	年龄
workclass	类别型变量	工作类型，如公务员、私企职工等
fnlwgt	数值型变量	抽样权重。（普查时使用的变量，与建模分析无关）
education	类别型变量	学历，如本科、研究生等
education_num	数值型变量	受教育年限
marital-status	类别型变量	婚姻状况
occupation	类别型变量	所在行业
relationship	类别型变量	家庭角色，比如丈夫、妻子等
race	类别型变量	种族
sex	类别型变量	性别
capital_gain	数值型变量	该年度投资收益
capital_loss	数值型变量	该年度投资损失
hours_per_week	数值型变量	每星期工作时间
native_country	类别型变量	出生国家
label	类别型变量	年收入分类，分为两类：“>50K”和“≤50K”

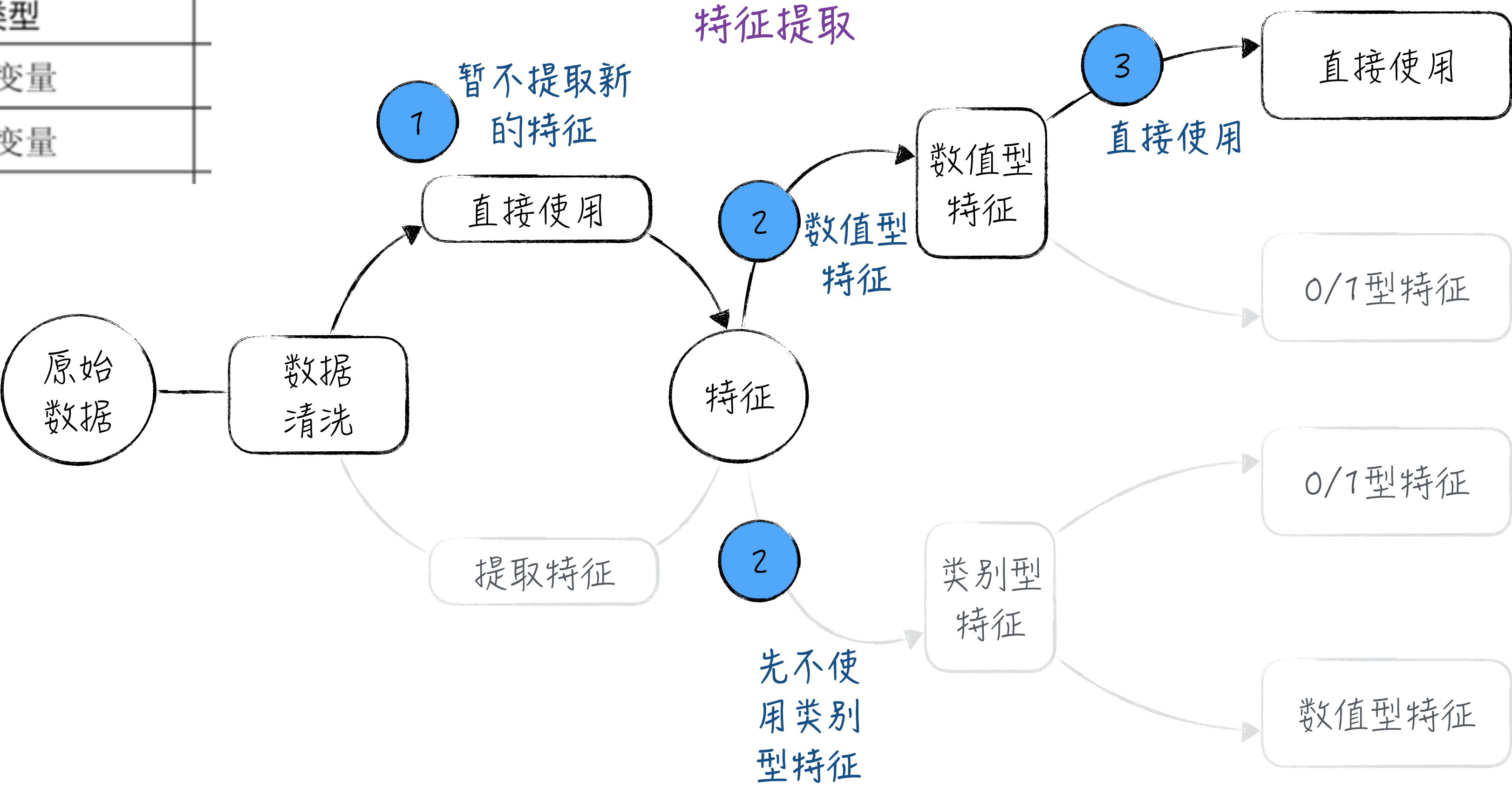
初探数据

变量种类

变量名	变量类型
age	数值型变量
workclass	类别型变量

两种类型的特征

- 数值型特征：可直接使用
- 类别型特征：转换后只用



初探数据

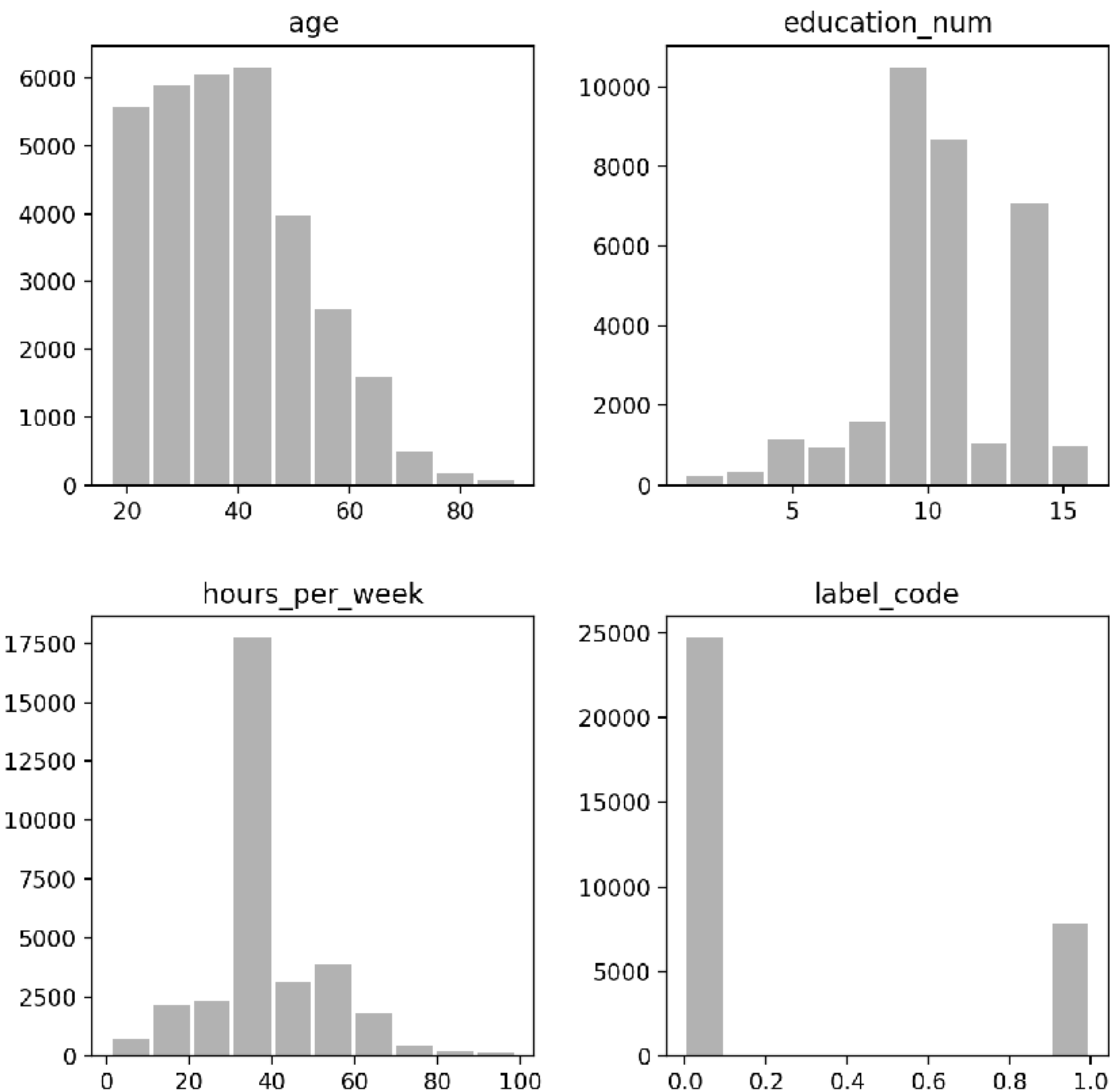
数据可视化

为了聚焦于模型，只使用数值型特征：

- age
- education_num
- capital_in
- capital_loss
- hours_per_week

使用数据可视化，得到对数据的直观印象

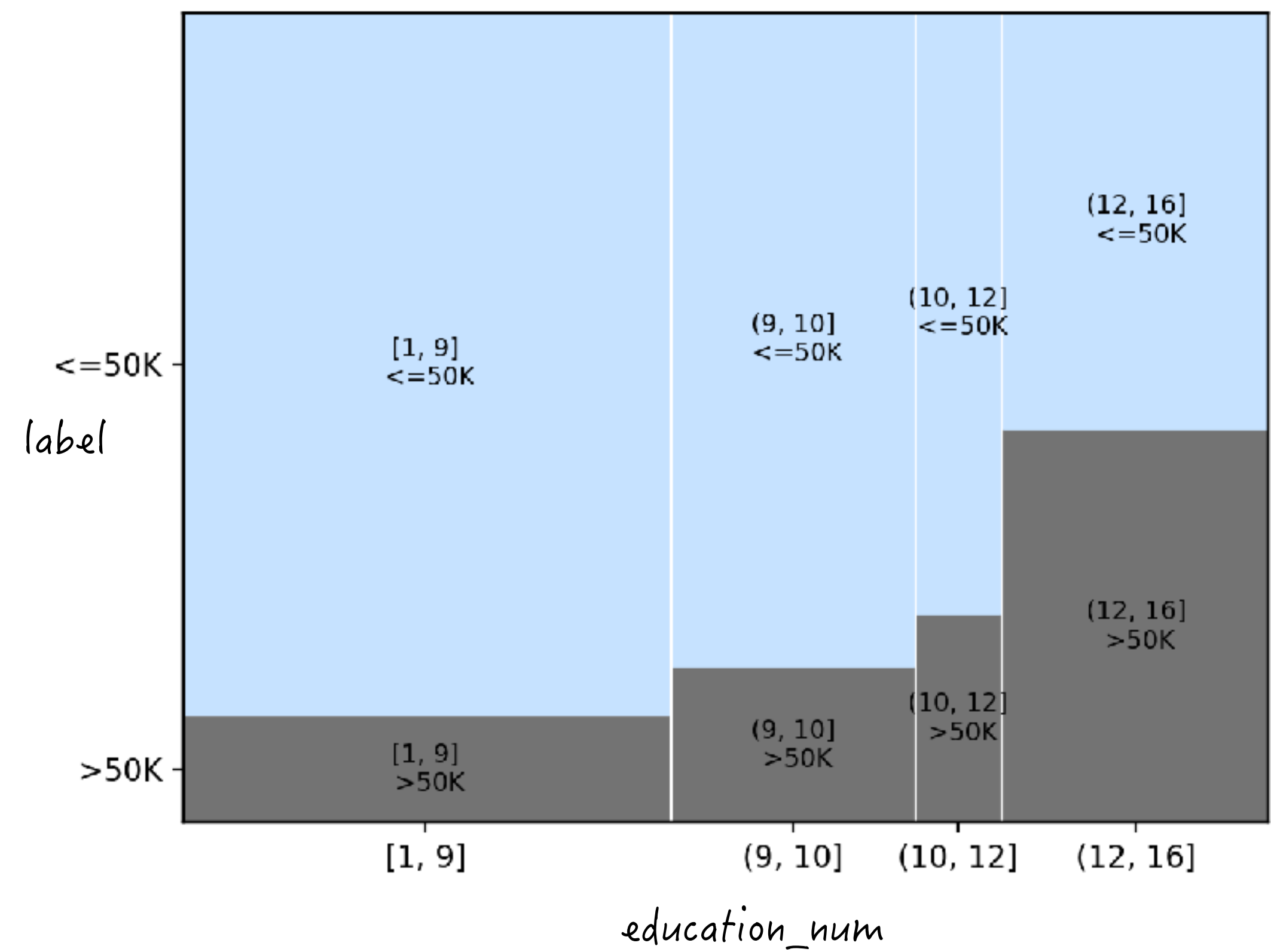
各变量的直方图



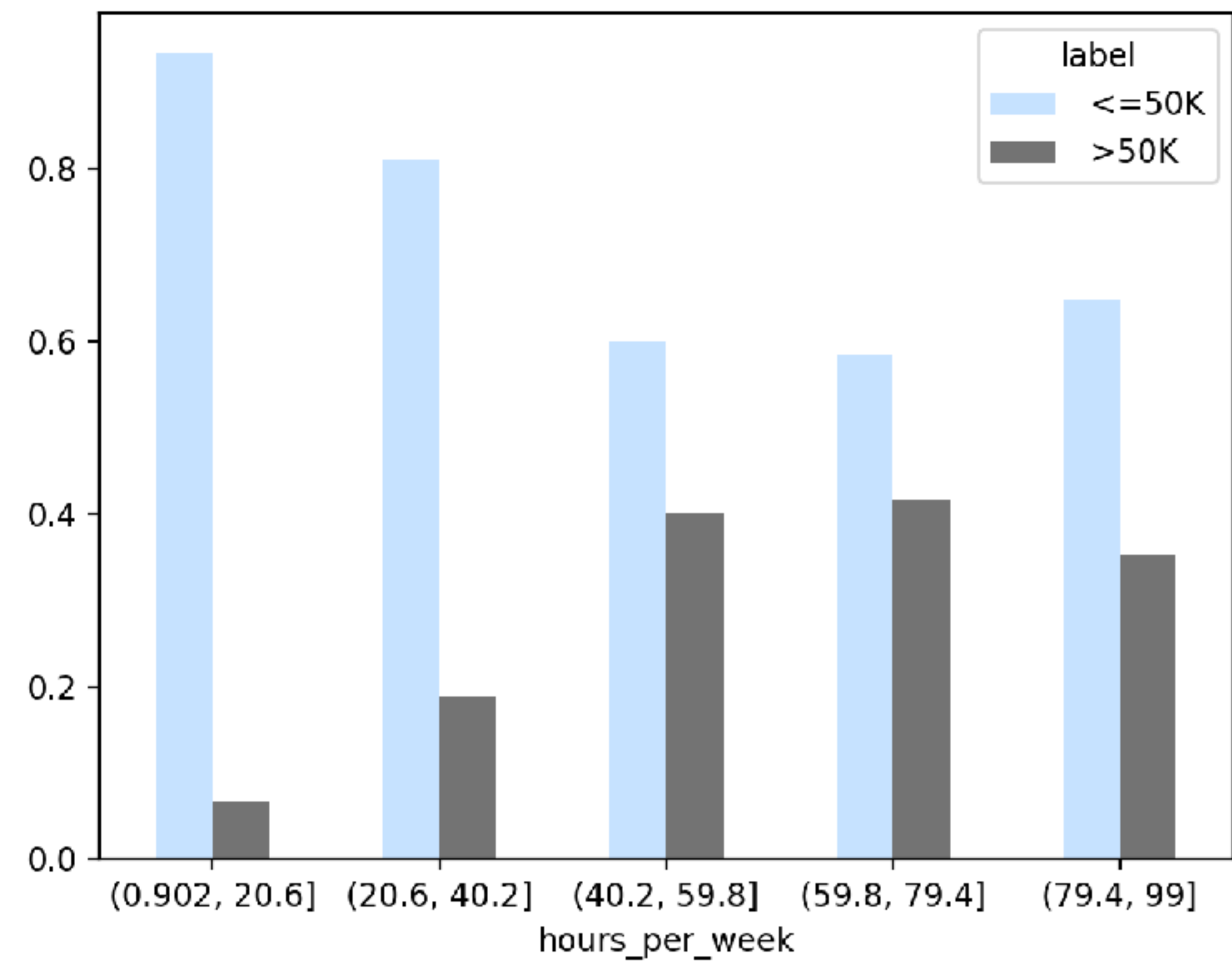
初探数据

数据可视化

education_num和label的交叉报表



hours_per_week和label的交叉报表



目录

ONE 初探数据

变量种类、数据可视化

TWO 搭建模型

构建模型、模型参数的稳定性

THREE 理解模型结果

发生比、边际效应

搭建模型

构建模型

将数据分为训练集和测试集

使用statsmodels, 定义模型

分析模型参数的稳定性、假设检验

使用模型做预测

这行代码表示, 检验的假设为: 变量
education_num的系数等于0; 并非
education_num=0

检验假设education_num的系数等于0:

```
print re.f_test("education_num=0")  
<F test: F=array([[ 1783.4276255]]), p=0.0, df_denom=26042, df_num=1>
```

P-value小于0.05。拒绝education_num
的系数等于0这个假设, 即它的系数
是显著的

检验假设education_num的系数等于0.32和hours_per_week的系数等于0.04同时成立:

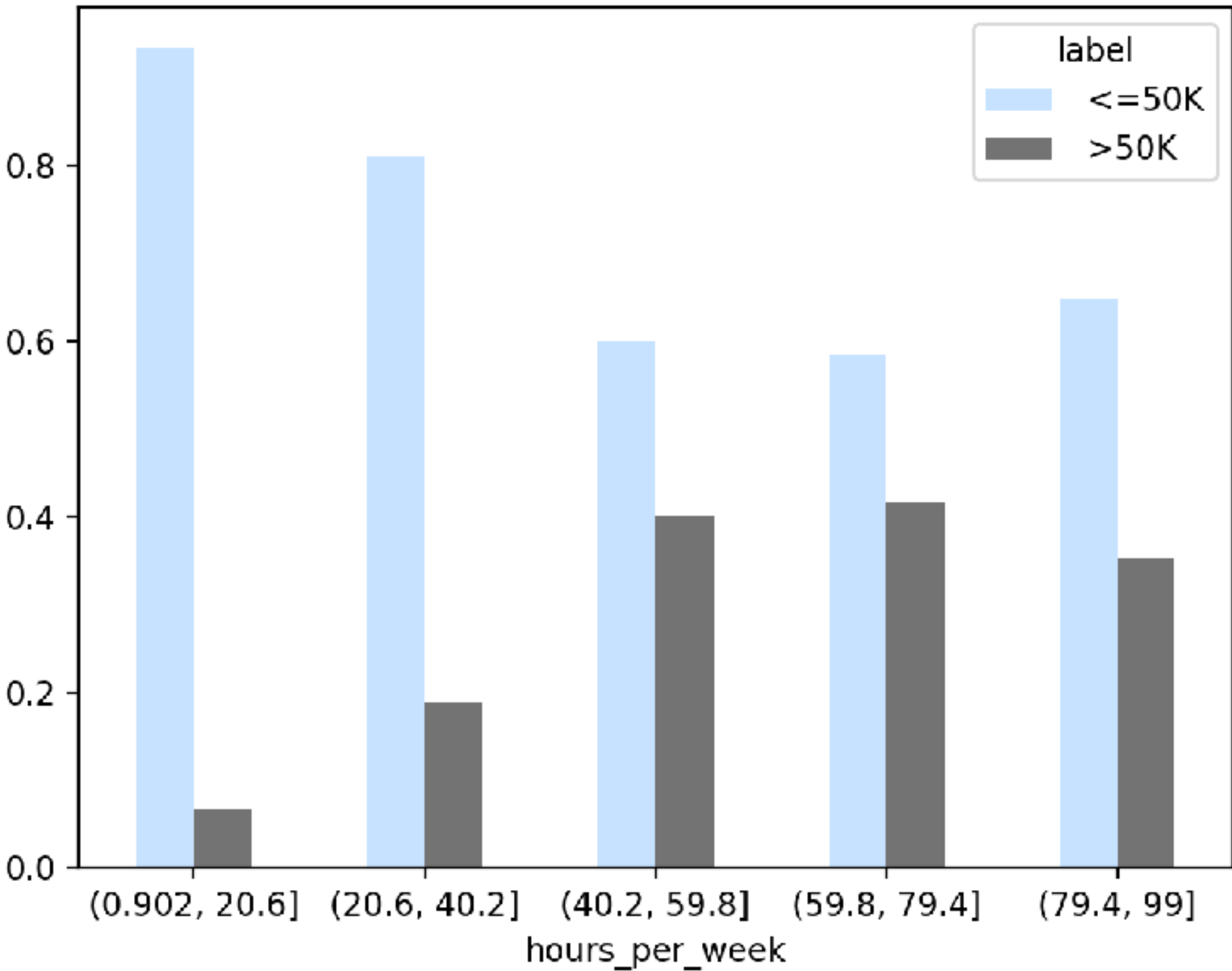
```
print "检验假设education_num的系数等于0.32和hours_per_week的系数等于0.04同时成立: "  
<F test: F=array([[ 0.01940236]]), p=0.980784667777, df_denom=26042, df_num=2>
```

P-value大于0.05。不能拒绝这两
个假设同时成立

搭建模型

模型参数的稳定性

模型结论与实际数据
并不完全相符



Logit Regression Results

=====

Dep. Variable: label_code No. Observations: 1

Model: Logit Df Residuals: 1

Method: MLE Df Model: 1

Date: Sun, 07 May 2017 Pseudo R-squ.: 0.000

Time: 20:01:50 Log-Likelihood: 0.000

converged: True LL-Null: 0.000

LLR p-value: 0.000

	coef	std err	z	P> z	[0.025	0.975]
Intercept	-8.2970	0.128	-64.623	0.000	-8.549	-8.045
age	0.0435	0.001	31.726	0.000	0.041	0.046
education_num	0.3215	0.008	42.231	0.000	0.307	0.336
capital_gain	0.0003	1.07e-05	29.650	0.000	0.000	0.000
capital_loss	0.0007	3.64e-05	20.055	0.000	0.001	0.001
hours_per_week	0.0399	0.001	26.995	0.000	0.037	0.043

1

hours_per_week
的系数大于0

参数估
计值

参数都
显著

参数估计值的置
信区间

目录

ONE 初探数据

变量种类、数据可视化

TWO 搭建模型

构建模型、模型参数的稳定性

THREE 理解模型结果

发生比、边际效应

理解模型结果

发生比

在线性回归里，模型参数表示了各个变量对结果的影响幅度

x变动1, y变动a

$$y_i = ax_i + bz_i + c$$

不论z的取值如何，变量x的边际效应都是a

在逻辑回归里，模型参数又代表了什么呢？

$$P(y_i = 1) = \frac{1}{1 + e^{-(ax_i + bz_i + c)}}$$

理解模型结果

发生比

在逻辑回归里，模型参数又代表了什么呢？

$$P(y_i = 1) = \frac{1}{1 + e^{-(ax_i + bz_i + c)}}$$

$$odds = \frac{P(y_i = 1)}{1 - P(y_i = 1)}$$

odds被称为发生比

$$\ln \frac{P(y_i = 1)}{1 - P(y_i = 1)} = ax_i + bz_i + c$$

$$\ln odds(x_i = k + 1) - \ln odds(x_i = k) = a$$

$$\frac{odds(x_i = k + 1)}{odds(x_i = k)} = e^a$$

x增加1时，相应的发生比变为之前的 e^a 倍

理解模型结果

边际效应

在逻辑回归里，模型变量的边际效应又是怎样的呢？

$$P(y_i = 1) = \frac{1}{1 + e^{-(ax_i + bz_i + c)}}$$

变量x对模型结果P的影响如何呢？

$$\ln \frac{P(y_i = 1)}{1 - P(y_i = 1)} = ax_i + bz_i + c$$

$$\frac{1}{P} \frac{\partial P}{\partial x} + \frac{1}{1 - P} \frac{\partial P}{\partial x} = a$$

$$\frac{\partial P}{\partial x} = aP(1 - P)$$

变量x的边际效应并不恒定，
常利用训练数据的平均边际效应来衡量

理解模型结果

发生比与边际效应

变量对事件发生比的影响

	2.5%	97.5%	OR
Intercept	0.000194	0.000321	0.000249
age	1.041611	1.047218	1.044411
education_num	1.358725	1.399879	1.379149
capital_gain	1.000298	1.000340	1.000319
capital_loss	1.000659	1.000802	1.000731
hours_per_week	1.037733	1.043769	1.040746

估计值的上界

估计值的下界

每星期工作时间增加1

年收入大于50K的发生比增加4.07%

变量的边际效应

Logit Marginal Effects						
Dep. Variable:	label_code					
Method:	dydx					
At:	overall					
	dy/dx	std err	z	P> z	[95.0% Conf. Int.]	
age	0.0056	0.000	33.563	0.000	0.005 0.006	
education_num	0.0413	0.001	47.313	0.000	0.040 0.043	
capital_gain	4.09e-05	1.3e-06	31.500	0.000	3.84e-05 4.34e-05	
capital_loss	9.372e-05	4.54e-06	20.648	0.000	8.48e-05 0.000	
hours_per_week	0.0051	0.000	28.167	0.000	0.005 0.005	

边际效应

边际效应的置信区间

THANK YOU

—