Задача 3

Условие

Набор данных: r23i\_os26c.sav, собранный в процессе исследования Russia Longitudinal Monitoring Survey of HSE

Объясняемая перменная: *salary* – зарплата за последний месяц

Регресссоры:

* *sh5* пол
* *s\_age* возраст
* *s\_marst*  семейное положение
* *s\_diplom*  образование
* *status* тип населенного пункта
* *sj13.2* средняя зарплата за последний год
* *sj6.2* средняя рабочая неделя

1. Постройте линейную регрессию зарплаты на все параметры, которые Вы выделили  
   из данных мониторинга. Не забудьте оценить коэффициент вздутия дисперсии VIF.
2. Поэкспериментируйте с функциями вещественных параметров: используйте  
   логарифм и степени (хотя бы от 0.1 до 2 с шагом 0.1).
3. Выделите наилучшие модели из построенных: по значимости параметров,  
   включённых в зависимости, и по объяснённому с помощью построенных  
   зависимостей разбросу adjusted R2 – R2adj.
4. Сделайте вывод о том, какие индивиды получают наибольшую зарплату.
5. Оцените регрессии для подмножества индивидов: а) городские жители, не состоявшие в браке; б) разведенные женщины, без высшего образования

Решение

Считаем данные из r23i\_os26c.sav, удалим NA, приведем данные к удобному виду:

* *sex –* пол (0 – женский, 1 - мужской)
* *age –* возраст нормализованный по формуле: *salary - mean(salary)) / sqrt(var(salary))*
* *wed1* - (1 - в браке, 0 - не в браке)
* *wed2 –* (1 – разведен или вдовец, 0 – не разведен или вдовец)
* *wed3 –* (1 состоял в браке, 0 - никогда не состоял в браке)
* *higher\_educ –* наличие высшего образования(1 – есть высшее образование, 0 – нет высшего образования)
* *city\_status –* (1 – проживает в городе, 0 – не проживает в городе)
* *working\_hours* - среднее количество рабочих часов в неделю нормализованное по формуле *salary - mean(salary)) / sqrt(var(salary))*

1. Построим линейную регрессию зарплаты на все параметры, выделенные из данных мониторинга. Оценим коэффицент вздутия дисперсии VIF

*model1 = lm(data = data\_normalized, salary ~ sex + age + wed1 + wed2 + wed3 + higher\_educ + city\_status + working\_hours)*  - данная модель показала плохой коэффицент R2 ~ 0,1539, хороший VIF и низкие p-значения у перменных wed1, wed2. В связи с этим попробуем убрать из модели данные переменные

*model1 = lm(data = data\_normalized, salary ~ sex + age + wed3 + higher\_educ + city\_status + working\_hours) –* данная модель показала практически идентичный R2 ~ 0,1538, хороший VIF и отличную p-статистику. В дальнейшем будем использовать именно ее.

1. Используем в модели логарифмы и степени

Логарифмы и степени введем только для переменных *age* и *working\_hours,* т.к. все остальные переменные принимают значения 0/1

**Логарифмы**:

*model1 = lm(data = data\_normalized, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))*

Переменные *working\_hours* и *I(log(working\_hours))* имеют плохой VIF (>5), уберем их из модели

Модели с хорошим VIF:

*model1 = lm(data = data\_normalized, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(log(age)))*

Значение R2 ~ 0.179, плохая p-статистика для *wed3* и *log(age)*

*model2 = lm(data = data\_normalized, salary ~ sex + age + wed3 + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))*

Значение R2 ~ 0.2082, плохая p-статистика для *wed3* и *log(age)*

*model3 = lm(data = data\_normalized, salary ~ sex + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))*

Значение R2 ~ 0.1924, отличная p-статистика

Из вышеприведенных моделей с хорошим VIF наилучший R2 (0,2082) имеет *model2,* но у нее очень плохая p-статистика для *wed3* и *log(age). model3* имеет чуть более низкий R2 (0,1924), но идеальную p-статистику.

**Степени:**

Степень в интервале от 0,1 до 2 включительно с шагом 0,1 будет задаваться переменной

current\_pow

Рассмотрим модель для *current\_pow = 1*:

*model1 = lm(data = data\_normalized, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow)),* она имеет

очень плохой VIF и очень плохую p-статистику. Регрессоры *working\_hours* и *age* имеют

нибольший VIF, уберем их из модели.

*model1 = lm(data = data\_normalized, salary ~ sex + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))*, эта модель показывает себя гораздо

лучше: она имеет хороший VIF, и неплохую p-статистику за исключение регрессора *wed3*,

R2 ~ 0,196. Уберем из модели регрессор *wed3.*

*model1 = lm(data = data\_normalized, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow)),* R2 данной модели не меняется по

сравнению с предыдущей, VIF все еще очень хороший, p-статистика очень хорошая,

значит, регрессор *wed3* можно исключить как незначительный.

Проделывая те же действия для моделей с другими степенями мы можем заметить

следующие тенденции:

* Для степеней в пределе от 0,1 до 1,1 включительно (степень 1 не считаем, т.к. это исходный датасет) значение R2 растет с повышением *current\_pow*
* После *current\_pow* = 1,2 включительно значение R2 начинает падать, VIF и p-статистика не будут меняться

1. Выделим наилучшие модели из построенных: по значимости параметров, включённых в зависимости, и по объяснённому с помощью построенных зависимостей разбросу adjusted R2 – R2\_adj

Наилучшими являются модели со степенями 0.9, 1.1 и 1.2, все они показывают отличный VIF и хорошую p-статистику. Данные модели имеют практически одинаковый разброс

R2 – R2\_adj . Лучшей является модель с *current\_pow* = 1.1, она имеет наивысший показатель R2, равный 0,2118

1. Согласно модели, которую мы сочли наболее эффективной, самую большую зарплату получают молодые мужчины, живущие в городах, имеющие высшее образование и работающие большое число часов в неделю. На заработную плату семейное положение не оказывает значительного эффекта.
2. Оценим регрессии для подмножества индивидов, указанных в варианте: не вступавшие в брак мужчины без высшего образования; городские жители, мужчины, состоящие в браке.

**Множество мужчин без высшего образования, не вступавших в брак:**

Имеем следующую модель: *model1 = lm(data = data\_picked, salary ~ city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))*, R2 ~ 0,1746, все регрессоры являются

значимыми. Согласно этой модели среди данной группы наибольшую зарплату имеют молодые люди, проживающие в городах и много работающие.

**Множество мужчин, городских жителей, состоящих в браке:**

Имеем следующую модель: *model1 = lm(data = data\_picked, salary ~ higher\_educ + I(working\_hours^current\_pow) + I(age^current\_pow)),* R2 ~ 1858, регрессоры *higher\_educ* и *I(age^current\_pow)* являются самыми значимыми, *I(working\_hours^current pow)* - второй по значимости. Согласно этой модели среди данной группы наибольшую зарплату имеют молодые люди, имеющие высшее образование и много работающие.

Выводы

Из всей выборки самую болшую заработную плату имеют молодые мужчины с высшим образованием, много работающие и проживающие в городах, семейное положение на доход влияния не оказывает.

Среди мужчин без высшего образования, не вступавших в брак, больше всего зарабатывают молодые люди, проживающие в городах и много работающие.

Среди мужчин, живущих в городах и состоящих в браке, больше всего зарабатывают молодые люди, имеющие высшее образование и много работающие.



















