بسم الله الرحمن الرحیم

پروژه ۴ درس مبانی علوم داده

دکتر بهرک و دکتر یعقوب زاده

محمد امین توانایی - ۸۱۰۱۰۱۳۹۶

سید علی تهامی - ۸۱۰۱۰۱۳۹۷

مهدی وجهی - ۸۱۰۱۰۱۵۵۸

سوالات

سوال ۱

به دلیل این که مقیاس و حدود این اعداد مشخص نیست نمی توان درباره خوب و بد بودن آنها حرفی زد و تنها می توانیم آنها را با یکدیگر مقایسه کنیم و همانطور که انتظار داریم مدل روی داده های آموزشی بهتر از داده های آزمایشی عمل می کند و در تمامی معیار ها مقادیر کمتری می گیرد که یعنی بهتر است.

سوال ۲

R_squared : در واقع یک شاخص است که نشان میدهد تغییرات متغیر وابسته(خروجی) چقدر توسط متغیر ورودی (مستقل) توضیح داده میشوند.

برای محاسبه R Squared لازم است چند چیز را متوجه باشیم.

۱.unexplained variance : فاصله مقادیر واقعی از مقدار پیشبینی شده بتوان دو

۲.total variance : فاصله مقدار واقعی از میانگین مقدار واقعی

حال اثبات میشود که همیشه unexplained variance از total variance کوچکتر می باشد و R Squared همیشه بین صفر و یک است و هر چقدر نسبت unexplained variance به total variance کمتر باشد نشان میدهد پراکندگی نسبت به پیشبینی از پراکندگی نسبت به میانگین بسیار کمتر است و مدل خوبی فیت شده است.

R Squared به صورت زیر محاسبه میشود:

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

هرچه R Squared به یک نزدیکتر باشد یعنی تغییرات متغیر وابسته توسط متغیر مستقل به خوبی پیشبینی شده است.

حال یکی از بدی های این مدل اینست که هر چه predictor های بیشتری به مدل اضافه کنیم هر چند نتواند آن داده را به خوبی پیش بینی کند باعث افزایش مقدار R Squared می شود در حالی که Adj R Squared با اضافه شدن پیش بینی کننده جدید تنها زمانی افزایش می‌یابد که عبارت جدید مدل را بالاتر از آنچه که با احتمال به دست می‌آید افزایش دهد و زمانی کاهش می‌یابد که یک پیش‌بینی‌کننده مدل را کمتر از آنچه به طور تصادفی پیش‌بینی می‌شود، افزایش دهد.

سوال ۳

Hypothesis Testing

هنگامی که یک آزمون آماری را انجام می دهید، یک مقدار p به شما کمک می کند تا اهمیت نتایج خود را در رابطه با فرضیه صفر تعیین کنید.فرضیه صفر (H0) بیان می کند که هیچ رابطه ای بین دو متغیر مورد مطالعه وجود ندارد (یک متغیر بر دیگری تاثیر نمی گذارد). بیان می کند که نتایج به دلیل شانس هستند و در حمایت از ایده مورد بررسی قابل توجه نیستند. بنابراین، فرضیه صفر فرض را بر این می‌گذارد که هر چیزی که بخواهید ثابت کنید، اتفاق نیفتاده است.فرضیه جایگزین (Ha یا H1) فرضیه ای است که در صورت نادرست بودن فرضیه صفر، آن را باور خواهید کرد.Ha بیان می‌کند که متغیر مستقل بر متغیر وابسته تأثیر می‌گذارد و نتایج در حمایت از نظریه مورد بررسی معنادار هستند (یعنی نتایج به دلیل شانس تصادفی نیستند).مقدار p یا مقدار احتمال، عددی است که نشان می‌دهد چقدر احتمال دارد که داده‌های شما به‌طور تصادفی رخ داده باشند (یعنی فرضیه صفر درست است).هرچه مقدار p کوچکتر باشد، احتمال وقوع نتایج به صورت تصادفی کمتر است و شواهد قوی‌تری مبنی بر رد فرضیه صفر وجود دارد.

Summary statistic ای که در قسمت ۴ بدست آمده نشان دهنده اینست که bmi و bp دارای p value های بهتری هستند و بهتر دیابت را توضیح میدهند.

سوال ۴

در این قسمت با استفاده از کلاس OLS، مدل رگرسیون خطی چندگانه را روی داده‌های y_train و x_train فیت می‌کنیم.

خروجی ایجاد شده توسط دستور summary، بسیار مفصل است و مباحث مربوط به وجود هم‌خطی و نرمال بودن باقی‌مانده‌ها، توسط آزمون‌های durbin watson و jarque bera صورت گرفته است که مشخص است فرض نرمال بودن با توجه به بزرگ بودن مقدار احتمال ($borP(BJ)$=0.785) رد نمی‌شود.

اهمیت هر ویژگی در مجموعه داده‌ها بر اساس نتایج به‌دست‌آمده از روش کمترین مربعات عادی (OLS) ارزیابی می‌شود. این روش آماری به تعیین قدرت و اهمیت رابطه بین متغیر وابسته (وضعیت دیابتی) و هر متغیر مستقل (ویژگی‌ها) کمک می‌کند.

تحلیل شامل بررسی coefficients و مقادیر p-value برای هر ویژگی خواهد بود. Coefficients بالاتر نشان‌دهنده تأثیر قوی‌تر بر وضعیت دیابتی است، در حالی که مقدار p پایین‌تر نشان می‌دهد که ویژگی از نظر آماری معنی‌دار است.

ویژگی‌هایی با مقادیر p پایین و ضرایب مطلق coefficients به عنوان مهم‌تر در نظر گرفته می‌شوند زیرا تأثیر بیشتری بر پیشرفت دیابت دارند.

در قسمتی از خروجی، یک جدول تحلیل واریانس (ANOVA) را مشاهده می‌کنید که به ضرایب و همچنین مقدار احتمال برای معنی‌داری هر یک از ضرایب مدل رگرسیونی پرداخته است. به غیر از عرض از مبدا (Constant) که دارای مقدار احتمال بزرگتر از ۵٪ است، بقیه متغیرها می‌توانند در مدل حضور داشته باشند. از طرفی بزرگ بودن مقدار F-statistic و همینطور کوچک بودن مقدار احتمال (Prob F-statistic) نشان از

مناسب بودن مدل است. R-squared (ضریب تعیین): نسبت تغییرات متغیر وابسته‌ای که قابل پیش‌بینی/توضیح داده شده است.

Adj. R-squared (ضریب تعیین تعدیل‌شده): فرم تغییریافته‌ای از R-squared که برای تعداد متغیرهای مستقل در مدل تعدیل شده است. ارزش ضریب تعیین تعدیل‌شده با افزودن متغیرهای اضافی که واقعاً به بهبود مدل کمک می‌کنند، افزایش می‌یابد.

F-statistic : نسبت خطای میانگین مربعات مدل به خطای میانگین مربعات باقی‌مانده‌ها. این نسبت اهمیت کلی مدل را مشخص می‌کند.

coef (ضرایب) :ضرایب متغیرهای مستقل و جمله ثابت در معادله.

T: نسبت تفاوت بین مقدار تخمین‌زده‌شده و مقدار فرضی یک پارامتر، به خطای استاندارد.

پیش‌بینی مقادیر:

از جدول نتایج، ضریب x و جمله ثابت را یادداشت می‌کنیم. این مقادیر در معادله اصلی جایگزین شده و خط رگرسیون با استفاده از کتابخانه matplotlib ترسیم می‌شود.