Journal of Water & Wastewater Science & Engineering (jwwse)
Summer 2017, Vol. 2, No. 2.
DOI:10.22112/jwwse.2017.89374.1024

علوم و مهندسی آب و فاضلاب

نشریه علوم و مهندسی آب و فاضلاب
سال دوم، شماره ۲، تابستان ۱۳۹۶

Using and Development of Regression Models for Predicting Pipes Failure Rate in Water Distribution Networks

Sonia Ghasemnejad[1] and Homayoun Motiei[2*]
1- M.Sc., Faculty of Water and Environment, Shahid Beheshti University, Tehran, Iran.
2- Assistant Professor, Faculty of Water and Environment, Shahid Beheshti University, Tehran, Iran.
* Corresponding author, Email:H_Motiei@sbu.ac.ir

Received: 13/6/2017
Revised: 24/11/2017
Accepted: 29/11/2017

Abstract

Pipes failure events in water distribution networks leads to leakage of water. Failures cause the loss of significant fresh water and investments losses. The most important parameters of pipe failues are: material, age, length, diameter and hydraulic pressure. In this paper four statistical methods have used for analyzing pipe incidents, with the goal of estimation of failure probability in future, with finding the most influencing parameters on the incidents. The statistical regression models used in this research are linear, exponential, Poisson, and Logistic regression models. For evaluation of these models, the data of a region in the District 1 of the Tehran Water and Wastewater Company with more than 48500 consumers, total pipe length of 582702 meter, different materials and diameters were used. The results demonstrated that the logistic model has a better performance than the others to predict the future events with a higher probability.

Keywords: Events, Leakage, Pipe failure, Regression model, Water distribution networks.

کاربرد وتوسعه مدل‌های رگرسیونی برای پیش‌بینی میزان شکست لوله‌های شبکه توزیع آب شهری – مورد مطالعاتی ناحیه یک منطقه یک تهران

سونیا قاسم‌نژاد[1] و همایون مطیعی[2*]
۱-دانش‌آموخته کارشناسی ارشد عمران - آب، دانشکده آب و محیط زیست، دانشگاه شهید بهشتی، تهران، ایران.
۲- نویسنده مسئول، استادیار دانشکده آب و محیط زیست - دانشگاه شهید بهشتی، تهران، ایران.
* نویسنده مسئول، ایمیل: h_motiei@sbu.ac.ir

تاریخ دریافت: ۱۳۹۶/۳/۲۳
تاریخ اصلاح: ۱۳۹۶/۹/۳
تاریخ پذیرش: ۱۳۹۶/۹/۸

چکیده

شکست لوله‌ها در شبکه‌های توزیع آب شهری، باعث نشت جریان از شبکه شده و نه‌تنها باعث هدررفت مقادیر قابل‌توجهی از آب تصفیه شده می‌شود، بلکه سبب اتلاف سرمایه‌های مادی نیز می‌شود. جنس، سن، طول، قطر و فشار هیدرولیکی از مهم‌ترین متغیرهای تاثیرگذار در شکست لوله‌ها هستند. در این مقاله، از چهار روش آماری، برای تحلیل این متغیرها در شکست لوله‌ها استفاده شده است که هدف، یافتن معادلات لازم برای تخمین احتمال شکست لوله‌ها در آینده و تعیین پارامترهایی است که بیشترین تأثیر را بر روی احتمال شکست دارند. این چهار مدل رگرسیونی آماری عبارتند از مدل‌های رگرسیون خطی، نمایی، پواسون و لجستیک. به‌منظور ارزیابی روش‌های ارائه شده از داده‌های جمع‌آوری شده حوادث لوله‌ها در شبکه توزیع آب ناحیه ۱ از منطقه ۱ آب و فاضلاب شهر تهران با تعداد بیش از ۴۸۵۰۰ مشترک و ۵۸۲۷۰۲ متر طول کل لوله و متشکل از لوله‌هایی با جنس و قطرهای مختلف، استفاده شد. نتایج نشان دادند که از میان مدل‌های آماری بررسی شده، مدل رگرسیون لجستیک عملکرد بهتری داشته و با احتمال بالاتری می‌تواند حوادث آینده را پیش‌بینی کند.

کلمات کلیدی: شبکه‌های آب شهری، مدل رگرسیون، شکست لوله، نشت، حوادث و اتفاقات.

۱- مقدمه

خرابی و شکست لوله‌ها در شبکه توزیع آب، چالش‌های زیادی را برای شرکت‌های آب و فاضلاب در سراسر جهان ایجاد می‌کند. خوردگی[۱] و پوسیدگی لوله‌ها منجر به شکست لوله‌ها، نشت جریان، کاهش ظرفیت حمل آب، افزایش هزینه‌های تعمیرات، کاهش ظرفیت آتش‌نشانی و آلودگی آب می‌شود. Kropp and Herz (2005) نشان دادند که برای یک لوله با شرایط محیطی ثابت، گذر زمان لوله را فرسوده و کارایی آن را کاهش می‌دهد. یکی از راهکارها و کلیدهای مهم مدیریت بهینه بهره‌برداری، اتخاذ استراتژی‌های صحیح نوسازی و بازسازی در شبکه‌های توزیع آب شهری، پیش‌بینی نرخ شکست لوله‌ها و ارزیابی قابلیت کاربری آن‌ها است. در نتیجه لازم است بررسی کاملی از حوادث و اتفاقات شبکه، عوامل مؤثر در ایجاد حادثه و تأثیر مشخصات لوله‌ها در تعداد حوادث و اتفاقات شبکه داشت. در سال ۲۰۰۰ به‌طور متوسط ۷۰۰ شکست لوله در روز در کانادا و آمریکا روی داده که هزینه‌ای معادل با ۱۰ میلیارد دلار کانادا در سال برای دولت‌های این دو کشور در برداشته است (Kabir et al., 2015). در تاکید اهمیت موضوع در شبکه‌های توزیع آب ایران، در سال ۱۳۷۶، حدود یک میلیون حادثه در سامانه‌های توزیع آب کشور رخ داده که بیش از ۲۰ درصد از کل درآمدهای شرکت آب و فاضلاب کشور را برای تعمیر، بازسازی و اصلاح به خود اختصاص داده است که حدود ۳۰ درصد این حوادث روی لوله‌های شبکه توزیع بوده است (بیگی، ۱۳۷۸). در کلان شهر تهران براساس وضعیت موجود تلفات آب بین ۲۵ تا ۳۰ درصد حجم کل آب تامین شده بوده که یکی از عوامل اصلی آن پوسیدگی و شکست لوله‌ها می‌باشد. قبل از اتخاذ هر تصمیمی برای تعمیر و تغییر لوله‌های فرسوده، نیاز است که درک بهتری از مکانیسم شکست و عوامل ایجاد شکست، به‌وجود بیاید. عواملی نظیر جنس لوله، محیط قرار گرفتن لوله و ویژگی‌های عملکردی سیستم می‌توانند روی احتمال شکست لوله‌ها تأثیر گذارند (جلیلی قاضی‌زاده و همکاران، ۱۳۸۶). روش‌های مختلف مدل‌سازی فیزیکی، مدل‌سازی توصیفی و مدل‌سازی آماری برای تحلیل شکست لوله‌ها، قابلیت اعتماد و عمر باقیمانده آن‌ها ایجاد شده است (Nishiyama and Filion, 2013)

هدف از این تحقیق، مقایسه‌ی مدل‌های آماری مختلف مناسب برای شناخت الگوی شکست لوله‌ها است. مدل‌های به کار رفته در این تحقیق عبارتند از مدل خطی[۲]، مدل نمایی[۳]، مدل خطی تعمیم یافته پواسون[۴] و مدل خطی تعمیم‌یافته لجستیک[۵]، (Kettler and Goulter,1985; Shamir and Howard,1979). روش بکار رفته در این مقاله بدین‌صورت انجام گرفت که ابتدا کلیه مدل‌های آماری رگرسیونی از نظر تئوری مورد مطالعه قرار گرفت و سپس با جمع‌آوری آمار و اطلاعات شبکه منطقه یک تهران مدل‌های آماری بر اطلاعات این منطقه فرمول‌بندی و برای کلیه متغیرهای فوق براساس این مدل‌ها روابطی تهیه شد که کلیه متغیرها در این روابط وجود داشته باشند. به‌دلیل تاثیر و اهمیت زیاد متغیرهایی نظیر قطر، طول، جنس، سن و فشار داخلی لوله که تاثیر بالایی در شکست لوله‌ها دارند این متغیرها انتخاب شدند.

۲- مواد و روش‌ها

روش رگرسیون از روش‌های آماری برای بررسی و مدل‌سازی ارتباط بین متغیرها است. تحلیل رگرسیونی، پرکاربردترین روش در بین تکنیک‌های آماری است (تابش و همکاران، ۱۳۸۷). آنچه در رگرسیون حائز اهمیت می‌باشد یافتن معادله‌ای است که بیان‌گر رابطه بین متغیرها بوده و از این معادله، پیش‌بینی یا برآوردهای لازم انجام گیرد. در بخش‌های زیر روش‌های رگرسیونی توضیح داده شده‌اند.

۲-۱- مدل رگرسیون خطی

در این مدل فرض بر آن است که متغیر Y تابعی از متغیر توصیفی X_i است (Montgomery et al., 1992):

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \qquad (۱)$$

که β_i و β_0: مقادیر ثابت یا پارامترهای رگرسیون هستند که تخمین زده می‌شوند. ε: مقدار خطا است با این فرض که خطاها با میانگین صفر و واریانس نامعلوم، توزیع نرمال داشته و مستقل باشند.

رابطه خطی بین تعداد شکست‌های یک قطعه از لوله و سن لوله، اولین بار توسط Kettler and Goulter (1985) مطرح شد (معادله ۲).

$$N = k_0 \times A \qquad (۲)$$

که N: تعداد شکست‌ها در هر قطعه از لوله در سال، k_0: پارامتر رگرسیون و A: سن لوله در اولین شکست هستند. این معادله حاصل از بررسی داده‌های شکست در طی ۱۰ سال در شهر Winnipeg کانادا است. آن‌ها همچنین دریافتند که رابطه‌ای خطی منفی بین قطر لوله‌ها و نرخ شکست وجود داشته که نشان می‌دهد قطر لوله‌ها با نرخ شکست رابطه عکس دارد.

در تحقیق حاضر، با تغییراتی که در فرم ابتدایی رابطه (۱) اعمال شد، عواملی مانند جنس، طول، فشار داخلی و قطر لوله به معادله اولیه که فقط شامل زمان بود اضافه شدند. مدل اصلاح شده در رابطه (۳) آمده است.

$$N=\beta_0+\beta_1(L)+\beta_2(P)+\beta_3(D)+\beta_4(Age)+ \beta_5(AC)+\beta_6(DI)+\beta_7(GCI)+\beta_8(PE)+\beta_9(PVC) \qquad (۳)$$

که N: تعداد شکست‌ها، P: فشار داخل لوله، D: قطر لوله، Age: سن لوله، AC: آزبست، DI: داکتایل، GCI: چدن، PE: پلی‌اتیلن و PVC: لوله‌ی پی‌وی‌سی هستند.

۲-۲- مدل رگرسیون نمایی

فرم کلی یک مدل رگرسیون غیرخطی عبارت است از (Kettler and Goulter, 1985):

$$y = f(x,\beta) + \varepsilon \qquad (۴)$$

که y: متغیر وابسته، $f(x,\beta)$: یک تابع غیرخطی با پارامترهای β_0، β_1 و ... و ε: میزان خطای باقی‌مانده هستند. بزرگ‌ترین مزیت مدل‌های غیرخطی این است که می‌توانند طیف وسیعی از توابع را برازش دهند.

Shamir and Howard (1979) تحلیل رگرسیون غیرخطی را برای یافتن رابطه نمایی بین سن لوله و شکست‌های آن به‌کار بردند. مدل ارائه شده آن‌ها در معادله (۵) نشان داده شده است.

$$N(t) = N(t_0) \times e^{A(t+g)} \qquad (۵)$$

که $N(t)$: تعداد شکست‌ها در واحد طول در سال، $N(t_0)$: تعداد شکست‌ها در واحد طول در سال نصب لوله، t: زمان بین شکستگی یک سال تا سال شکستگی قبلی، g: سن لوله در زمان t و A: ضریب نرخ شکستگی در yr^{-1}است.

Walski and Pellica (1982) با اضافه کردن دو پارامتر مؤثر دیگر در مدل فوق آن را به‌صورت معادله (۶) به‌دست آوردند:

$$N(t) = C_1 \times C_2 \times N(t_0) \times e^{A(t+g)} \qquad (۶)$$

که C_1: بیانگر تأثیر شکست‌های قبلی لوله، براساس مشاهداتی است که لوله‌ای که قبلاً دچار حداقل یک شکست شده باشد، احتمال شکست مجدد آن بیشتر است و C_2: بیانگر تفاوت نرخ شکست در لوله‌های چدنی با قطر متفاوت هستند.

۲-۳- مدل خطی تعمیم‌یافته‌ی پواسون

این مدل پاسخ میانگین یک توزیع شرطی خاص را به یک تابع پیش‌بینی مرتبط می‌کند، که مبتنی بر یک تابع توزیع احتمال[۶] (PDF) فرضی برای داده‌های گسسته (شمارشی) و همچنین یک تابع اتصال بوده و پارامترهای تابع توزیع احتمال را به متغیرهای موجود ارتباط می‌دهند (Agresti and Kateri, 2011). اگر $x_i = [x_{1i}, x_{2i}, \ldots, x_{ni}]$ بردار متغیرهای کمکی از قطعه i-ام سیستم (i=1,2,…,m) باشد، تعداد شکست‌های قطعه i با y_i نمایش داده می‌شود. سیستم مورد نظر در این تحقیق، سیستم توزیع آب و قطعات آن قطعات لوله‌ها خواهند بود. Guikema and Davidson (2006) مثال‌هایی از روش پواسون را برای تخمین خطر شکست در یک شبکه زیرساخت به‌کار بردند.

یک مدل رگرسیونی برمبنای توزیع پواسون، مشخص می‌کند که میانگین شرطی آمارها در یک تابع پیوسته $\mu(\vec{x}_i,\vec{\beta})$ از مقادیر متغیرهای مستقل، طبق معادله (۷) می‌باشد که در آن بردار n×1 از پارامترهای رگرسیون است (Cameron and Trivedi, 1998).

$$E(y_i|x_i) = \mu(\vec{x}_i, \beta) \qquad (۷)$$

با داشتن تابع x_i تابع y_i و برای مقادیر صحیح مثبت y_i، روابط (۸) و (۹) به‌دست می‌آید.

$$f(y_i|x_i) = \frac{\mu_i^{y_i} \times e^{-\mu_i}}{y_i!} \qquad (۸)$$

$$E(y_i|x_i] = \exp(\vec{x}_i, \vec{\beta}) \qquad (۹)$$

در مدل تعمیم‌یافته خطی پواسون فرض بر این است که میانگین شرطی و واریانس شرطی، برابرند.

$$\mu_i = exp(\vec{x}_i, \vec{\beta}) \quad (۱۰)$$

مدل خطی تعمیم یافته‌ی پواسون در اینجا به شکل معادلات (۱۱) تا (۱۳) است.

$$P\ (Y = y|\vec{x}) = e^{-\mu}\ \frac{\mu^{y}}{y!} \quad (۱۱)$$

$$\mu = E(Y|\vec{x}) \quad (۱۲)$$

$$Log(\mu) = \beta_0 + \beta_1(D) + \beta_2(AC) + \beta_3(GCI) + \beta_4(DI) + \beta_5(PE) + \beta_6(PVC) + \beta_7(L) + \beta_8(Age) + \beta_9(P) \quad (۱۳)$$

که Y: تعداد شکست‌ها، β_i: پارامترهای رگرسیون و متغیرهای مستقل مطابق آنچه در بخش قبل توضیح داده شد هستند.

۲-۴- مدل خطی تعمیم‌یافته‌ی لجستیک (رگرسیون لجستیک)

در بسیاری از موارد هدف بررسی وقوع یا عدم وقوع شکست در یک لوله در یک بازه زمانی خاص است و نه تعداد شکست‌ها در آن لوله در کل. متغیر وابسته یک متغیر باینری است که وقتی حداقل یک شکست در یک بازه زمانی، در لوله‌ای رخ دهد مقدار متغیر پاسخ ۱ خواهد بود. متغیرهای مستقل می‌توانند مقدار یک را با احتمال P و مقدار صفر با احتمال (P-۱) داشته باشند.

یک تابع رگرسیون لجستیک که تبدیل یافته P به لجیت[۷] آن است در معادله (۱۴) آمده است.

$$P = \frac{e^{\alpha+\beta 1x1+\beta 2x2+\cdots+\beta ixi}}{1 + e^{(\alpha+\beta 1x1+\beta 2x2+\cdots+\beta ixi)}} \quad (۱۴)$$

که α: پارامتر ثابت رگرسیون، β_i: ضرایب رگرسیون برای متغیرهای توصیفی و x_i: متغیرهای مستقل هستند. یک فرم جایگزین برای این مدل، مدل لجیت است که در آن تابع اتصال یک تابع لجیت است. این تابع لجیت در معادله (۱۵) نشان داده شده است.

$$Logit[P(x)] = Log[\frac{p(x)}{1-p(x)}] = \alpha + \beta_1(x_1) + \beta_2(x_2) + \cdots + \beta_i(x_i) \quad (۱۵)$$

معادله (۱۶) مدل لجستیک تعمیم‌یافته خطی پیشنهاد شده است.

$$Logit[P(x)] = Log\left[\frac{p(x)}{1-p(x)} =\right] \alpha + \beta_0(D) + \beta_1(AC) + \beta_2(GCI) + \beta_3(DI) + \beta_4(PE) + \beta_5(PVC) + \beta_6(Age) + \beta_7(L) + \beta_8(P) \quad (۱۶)$$

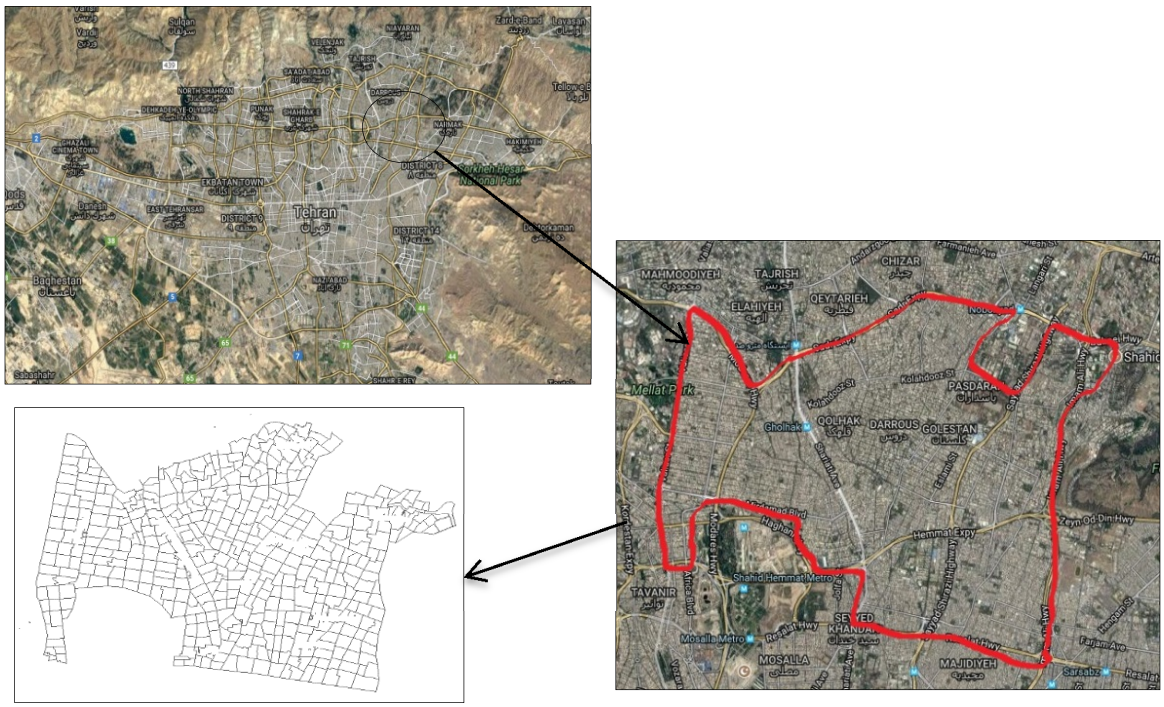
که $P(x)$: احتمال وقوع شکست، $1-P(x)$: احتمال عدم وقوع شکست، α: عرض از مبدا و β_i: پارامترهای رگرسیونی هستند که تخمین زده می‌شوند.

۲-۵- نرم‌افزار R

در این تحقیق برای مدل‌سازی آماری از نرم‌افزار R که یک نرم‌افزار متن‌باز برای محاسبات آماری و پردازش داده‌ها است استفاده شد (Maindonald, 2008; Everitt and Hothorn, 2010). این نرم‌افزار دارای توانایی‌های گسترده‌ای از محاسبات آماری نظیر مدل‌سازی خطی و غیرخطی، آزمون‌های کلاسیک آماری، تحلیل سری‌های زمانی، رده‌بندی و خوشه‌بندی بوده و دارای قابلیت بالایی در ترسیم اشکال گرافیکی و نمودارها است. در این نرم‌افزار کاربر علاوه بر استفاده از توابع موجود در آن، قابلیت ایجاد توابع جدید را داشته و می‌تواند به توابع قبلی اضافه کند (موسوی ندوشنی، ۱۳۹۱). اطلاعات بیشتر در مورد نحوه کارکرد و قابلیت‌های نرم‌افزار در سایت http://cran.r-project.org قابل دسترس است.

۳- منطقه مورد مطالعه

پایلوت مورد مطالعه، ناحیه ۱ از نواحی سه‌گانه منطقه شمیرانات تهران است. شکل ۱ نشان‌دهنده محدوده پایلوت مورد مطالعه است. تعداد مشترکین این ناحیه بالغ بر ۴۸۵۰۰ نفر و طول کل لوله‌های این ناحیه در حدود ۵۸۲۷۰۲ متر و دارای حوادث ثبت شده بالایی است. به‌عنوان مثال، تعداد حوادث در یک دوره زمانی از تیرماه ۱۳۸۳ تا آذرماه ۱۳۸۶ بیش از ۶۵۰۰۰ مورد بوده که از این میان سال ۱۳۸۶به‌تنهایی بیش از ۲۵۰۰۰ مورد ثبت شده است. به‌دلیل محدودیت‌های موجود در جمع‌آوری اطلاعات، در نهایت پارامترهای جنس، قطر، طول، فشار و سن لوله برای بررسی در این تحقیق انتخاب شدند. در منطقه مورد مطالعه عمدتاً از لوله‌هایی با جنس‌های داکتایل، آزبست، چدن، پلی‌اتیلن و PVC استفاده شده است و میانگین سن لوله‌ها حدود ۳۰ سال است. خلاصه‌ای از داده‌های ورودی در جدول ۱ آمده است.



شکل ۱- محدوده‌ی پایلوت مورد مطالعه، ناحیه ۱ از منطقه ۱ شبکه آب تهران

جدول ۱- متغیر های ورودی پایلوت و واحد آن‌ها

متغیر	توضیحات	واحد
NB	تعداد شکست	-
D	قطر لوله	میلی‌متر
AC	سیمان آزبست	-
DI	داکتایل	-
GCI	چدن	-
PE	پلی اتیلن	-
Age	سن لوله	سال

۴- نتایج و بحث

با تحلیل داده‌های موجود با روش‌های رگرسیونی، رابطه میان پارامترهای تأثیرگذار و نرخ حوادث به‌وقوع پیوسته، به‌دست آمده است. برای تحلیل مدل‌های حاصل از هریک از روش‌ها، دو رویکرد کلی مورد استفاده قرار گرفت. در ابتدا مدل به تمام داده‌های موجود برازش یافت. سپس در یک مرحله از ۷۰٪ داده‌ها به‌عنوان داده‌های مدل‌سازی و از ۳۰٪ باقیمانده، به‌عنوان داده‌های آزمون استفاده شد.

۴-۱- نتایج مدل خطی

در این بخش از یک رگرسیون خطی گام‌به‌گام بنابر مقادیر p-value استفاده و سطح معناداری روی ۰/۰۵ تعریف شد. داده‌ها به مدل خطی برازش داده شدند و p-value برای متغیرهای مختلف محاسبه شد. متغیرهایی که سطح معناداری پایینی داشتند (با p-value بزرگ‌تر از ۰/۱) در برازش داده‌ها به مرحله بعدی حذف شدند، به‌عبارتی ارتباط معناداری بین آن‌ها و تعداد شکست‌ها وجود ندارد. این چرخه تا زمانی ادامه می‌یابد که تمام متغیرهای موجود در مدل سطح معناداری بالایی داشته باشند (p-value کوچک‌تر از ۰/۰۰۱). خلاصه‌ای از شش مرحله انجام شده در جدول ۲ آمده است. در نهایت مناسب‌ترین حالت، حالتی است که در معادله (۱۷) نشان داده شده است.

$$NB = 1.024 + 0.00073L - 1.293\,DI \qquad (۱۷)$$

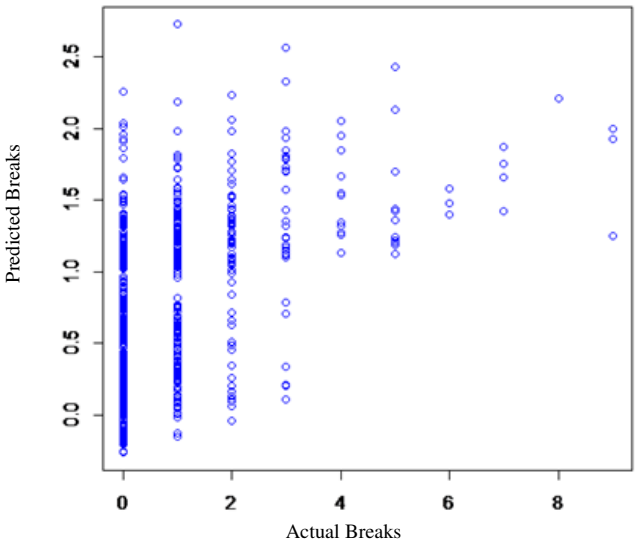
بالاترین مقدار p-value برای پارامترهای موجود در معادله (۱۷) مربوط به متغیر L و مقدار آن برابر $۸/۲۵\times۱۰^{-۱۶}$ بوده و برای متغیر DI از این مقدار نیز کمتر است.

برای اطمینان از نتایج، باید دقت مدل‌های پیش‌بینی‌کننده با استفاده از مقادیر میانگین خطای مربعات (MSE) ارزیابی شود (جدول ۳). برای سنجش دقت مدل پیش‌بینی کننده، داده‌های موجود به دو دسته تقسیم شدند، به این‌صورت که

جدول ۲- نتایج آزمون معناداری پارامترها در مدل خطی

	حالت ۱	حالت ۲	حالت ۳	حالت ٤	حالت ٥	حالت ٦
عرض از مبدا	****	****	****	****	****	****
L	****	****	****	****	****	****
P	*	*	*	.	.	–
D	.	–	–	–	–	–
Age	.	.	–	–	–	–
AC	.	.	.	–	–	–
DI	****	****	****	****	****	****
GCI	.	.	.	.	–	–
PE	NA	–	–	–	–	–
R^2	۰/۱۹۱۳	۰/۱۹۰۶	۰/۱۹۰۲	۰/۱۸۹۱	۰/۱۸۸۱	۰/۱۸۵۶
درجه آزادی	۷۷۱	۷۷۲	۷۷۳	۷۷۴	۷۷۵	۷۷۶

****	p-value برای پارامتر مورد نظر بین ۰ تا ۰/۰۰۱ است.
***	p-value برای پارامتر مورد نظر بین ۰/۰۰۱ تا ۰/۰۱ است.
**	p-value برای پارامتر مورد نظر بین ۰/۰۱ تا ۰/۰۵ است.
*	p-value برای پارامتر مورد نظر بین ۰/۰۵ تا ۰/۱ است.
.	p-value برای پارامتر مورد نظر بین ۰/۱ تا ۱ است.
–	از پارامتر مورد نظر در مدل استفاده نشده است.



شکل ۲- نمودار تعداد شکست‌های پیش‌بینی شده در مقابل تعداد واقعی در مدل خطی

۷۰٪ از لوله‌های هر جنس برای مدل‌سازی و ۳۰٪ باقیمانده به آزمون اختصاص یافت. شکل ۲ نشان‌دهنده تعداد شکست‌های پیش‌بینی شده در مقابل تعداد واقعی آن‌ها در سال است. میانگین مربعات خطاها (MSE) برای نرخ شکست‌های غیرصفر ۲/۳۷ و برای نرخ شکست‌های صفر ۰/۶۵ است (جدول ۳).

۴-۲- نتایج مدل نمایی

برازش یک مدل رگرسیون غیرخطی، به مقادیر اولیه برای پارامترهای مدل نیاز دارد. استفاده از مقادیر اولیه مناسبی که به مقادیر صحیح پارامترها نزدیک باشد، مشکلات هم‌گرایی را به حداقل می‌رساند و یک انتخاب نامناسب منجر به دستیابی

به یک مدل غیر بهینه می‌شود (Montgomery and Peck, 1992). انتخاب مقادیر اولیه، با استفاده از آزمون و خطا انجام شد و سپس این مقادیر به‌عنوان مقادیر اولیه پارامترها استفاده شد. مدل رگرسیون نمایی در معادله (۱۸) آمده است:

$$NB = exp\,(0.00092L - 0.13P - 0.0049D - 0.79\,AC - 0.895DI + 0.905GCI + 0.86\,(PE)) \quad (۱۸)$$

بیشترین مقدار p-value برای پارامترهای موجود در معادله (۱۸) مربوط به طول لوله است و جنس لوله چدن داکتایل با کمترین مقدار p-value، بالاترین سطح معناداری از لحاظ آماری را دارد. در این معادله NB: تعداد شکست‌ها در ۳/۵ سال، L: طول لوله، P: فشار داخلی لوله، AC: جنس لوله آزبست، DI: لوله با جنس داکتایل، GCI: لوله چدنی و PE: جنس لوله‌ی پلی‌اتیلن است. براساس پارامترهای برآورد شده در معادله (۱۸) برای متغیرهای موجود، با افزایش طول لوله تعداد شکست‌ها افزایش یافته است. همچنین بین قطر لوله و تعداد شکست‌های آن رابطه‌ی عکس وجود دارد و استفاده از جنس داکتایل با کاهش معناداری تعداد حوادث همراه بوده است. مقدار میانگین مربعات خطاها، برای شکست‌های صفر و غیر صفر، به‌ترتیب برابر ۰/۹۵ و ۴/۱۷ است (جدول ۳).

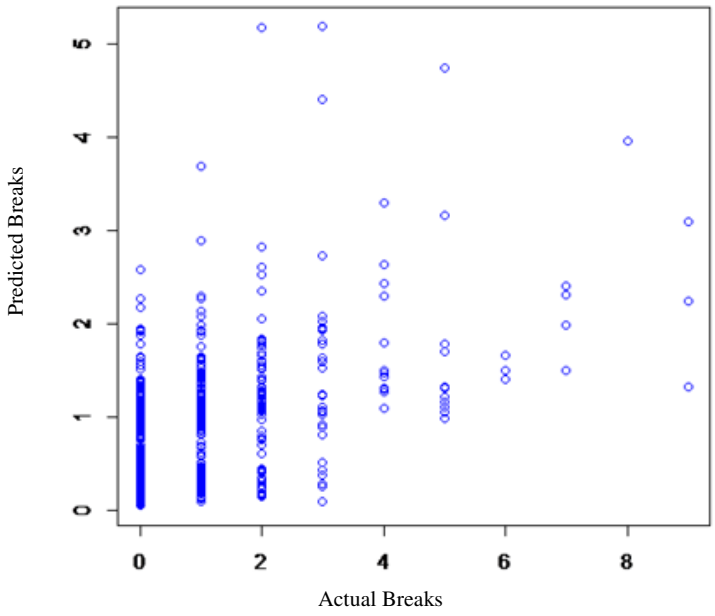
۴-۳- نتایج مدل پواسون

در این مدل نیز مشابه مدل‌های خطی و نمایی، مدل‌سازی با ۹ متغیر ابتدایی آغاز شده و سپس در طی عملیاتی گام به گام، در هر مرحله متغیری که سطح معناداری پایینی داشت، حذف شده و برازش مجدداً با متغیرهای باقیمانده انجام شد. مدل به‌دست آمده در معادله (۱۹) ارائه شده است.

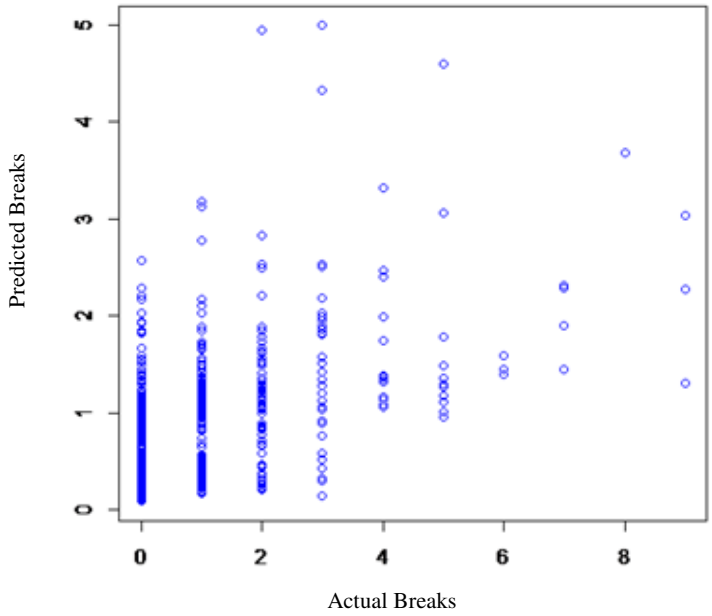
$$Log\mu = 0.7158 + 0.00089\,L - 0.144\,P - 0.0034\,D - 1.585\,DI \quad (۱۹)$$

متغیرهای موجود در رابطه (۱۹) عبارتند از: طول لوله (L)، فشار داخلی لوله (P)، قطر لوله (D) و جنس لوله داکتایل. با تعریف سطح معناداری روی ۰/۰۵، متغیرهای فشار و جنس لوله داکتایل بالاترین سطح معناداری را از لحاظ آماری دارند. در این مدل تعداد شکست‌ها با افزایش طول لوله افزایش و با افزایش قطر، کاهش می‌یابند. هم‌چنین استفاده از لوله‌های داکتایل منجر به کاهش تعداد شکست‌ها می‌شود.

تحلیل مدل‌سازی و آزمون درمورد این مدل انجام شد و نتایج نشان‌دهنده این مطلب است که مدل پواسون، در پیش‌بینی تعداد شکست‌های صفر بهتر از شکست‌های غیرصفر عمل می‌کند. شکل ۴ تعداد شکست‌های مشاهداتی در مقابل مقادیر پیش‌بینی شده شکست‌ها است. مقدار میانگین مربعات خطاها،



شکل ۳- نمودار تعداد شکست‌های پیش‌بینی شده در مقابل تعداد واقعی در مدل نمایی



شکل ۴- نمودار تعداد شکست‌های پیش‌بینی شده در مقابل تعداد واقعی در مدل پواسون

برای شکست‌های صفر و غیرصفر، به‌ترتیب برابر ۰/۵۵ و ۲/۳ است (جدول ۳).

در جدول ۴ آماره‌های نکویی برازش، برای مدل‌های خطی، نمایی و پواسون برای رکورد ۳/۵ ساله داده‌های پایلوت نشان داده شده است.

با در نظر گرفتن مقادیر AIC[۸] ، انحراف[۹] و GCV[۱۰]، مدلی که کمترین مقادیر را در این سه آماره دارد، مدل مناسب‌تر است. از سوی دیگر، با درنظر گرفتن لگاریتم احتمال LL[۱۱]، مدل با بیشترین مقدار این آماره مدل بهتری است. براساس نتایج موجود به‌نظر می‌رسد که کیفیت مدل خطی و نمایی در پیش‌بینی‌ها تقریباً مشابه و مدل پواسون داده‌ها را به‌شکل بهتری برازش می‌دهد. مقایسه میانگین مربعات خطاها برای تعداد شکست‌های صفر و غیرصفر نشان می‌دهد که هر سه مدل تعداد شکست‌های صفر را بهتر از شکست‌های غیرصفر پیش‌بینی می‌کنند (جدول ۳). همچنین، نتایج مدل نمایی و پواسون بسیار به یکدیگر نزدیک هستند. به‌طور کلی، تا این‌جا مدل‌های به‌کار رفته در این تحقیق، مدل‌های پیش‌بینی‌کننده مناسبی برای تعداد شکست‌های غیرصفر نیستند. برای رفع مشکل ناتوانی مدل‌های قبلی در پیش‌بینی تعداد شکست‌های غیرصفر، از مدل لجستیک استفاده شد.

جدول ۳- میانگین مربعات خطاها برای مدل‌های رگرسیون خطی و نمایی و پواسون

مدل	MSE	MSE برای شکست صفر	MSE برای شکست غیرصفر
رگرسیون خطی	۱/۴۱۴	۰/۶۵	۲/۳۷
رگرسیون نمایی	۱/۳۰۵	۰/۵۴	۲/۲۸
رگرسیون پواسون	۱/۳۱۲	۰/۵۵	۲/۳

جدول ۴- خلاصه آماره‌های نکویی برازش برای مدل‌های رگرسیون خطی و نمایی و پواسون

نوع مدل	Log Likelihood (LL)	AIC	Deviance	GCV
رگرسیون خطی	-۱۲۴۰/۴۱	۲۴۸۸/۸۳	۱۱۰۱/۸۸	۱/۴۲
رگرسیون نمایی	-۱۲۰۹/۰۸	۲۴۳۴/۱۴	۱۰۱۶/۷	۱/۳۲
رگرسیون تعمیم‌یافته خطی پواسون	-۹۰۰/۷۵	۱۸۱۱/۵	۹۹۱/۳۶	۱/۳۲

۴-۴- نتایج مدل لجستیک

برخلاف سه مدل قبلی که به‌خاطر وجود تعداد شکست‌های صفر زیاد در داده‌ها، مدل مناسبی نبودند، مدل لجستیک به‌طور ویژه‌ای برای کار با داده‌های باینری ۰ و ۱، در مواقعی که تعداد زیادی عدد صفر وجود دارد، طراحی شده است. چنانچه یک لوله در طی بازه زمانی ثبت داده‌های جمع‌آوری شده، حداقل یک شکست را تجربه کرده باشد، متغیر پاسخ صفر و یک برای آن، مقدار یک خواهد بود. داده‌های موجود با چهار حالت لجستیک برازش شدند که از میان آن‌ها حالت چهارم مناسب‌ترین مدل است. نتایج برازش نشان می‌دهد که متغیرهای طول، سن، جنس لوله چدن داکتایل و چدنی از لحاظ آماری، در سطح ۰/۰۵ معنادارند.

در طی مدل‌سازی، متغیرهای دیگر تا قبل از رسیدن به حالت چهارم، متغیرهای مستقل دیگر اگر دارای وضعیت معناداری نبودند حذف شدند. نتایج معناداری در برازش مدل کلی در جدول ۵ آمده است. در معادله (۲۰) مدل نهایی با پارامترهای براورد شده آمده است.

همان‌طور که دیده می‌شود با افزایش طول و سن لوله، تعداد شکست‌های مدل افزایش می‌یابد، همچنین استفاده ازلوله‌هایی با جنس چدن داکتایل باعث کاهش تعداد شکست‌ها و برعکس،

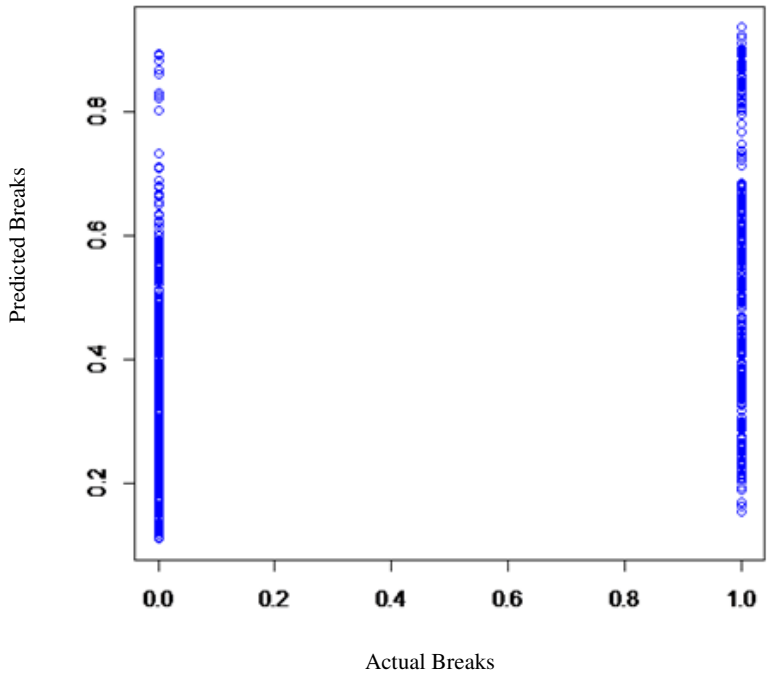
$$Logit\ (P(x)) = 0.0009L\ -\ 0.2P\ +\ 0.025Age\ -\ 0.76DI + 1.38GCI \qquad (۲۰)$$

استفاده از لوله‌های چدنی باعث افزایش شکست‌ها می‌گردد. با توجه به جدول ۷، انحراف این مدل ۹۳۶/۶۰ و درجه آزادی ۷۷۴ و AIC برابر ۹۴۶ برای همین درجه آزادی است. مقادیر پارامترهای رگرسیون نشان می‌دهد که در لوله‌های با طول بیشتر، احتمال تجربه‌ی شکست بالاتر است. هم‌چنین با توجه به شکل ۵ دیده می‌شود که مدل کلی در پیش‌بینی نرخ شکست‌های صفر بهتر عمل می‌کند.

جدول ۵- نتایج آزمون معناداری پارامترها در مدل خطی تعمیم‌یافته لجستیک

	حالت ۱	حالت ۲	حالت ۳	حالت ۴
عرض از مبدا		-	-	-
L	****	****	****	****
P	**	**	**	****
D	*	*	*	-
Age	**	***	***	***
AC	*	*	-	-
DI	****	****	****	****
GCI	**	**	****	****
PE	NA	-	-	-
Residual Deviance	۹۳۲/۷۱	۹۳۳/۱۴	۹۳۴/۰۱	۹۳۶/۰۶
AIC	۹۴۸/۷۱	۹۴۷/۱۴	۹۴۶	۹۴۶

****	p-value برای پارامتر مورد نظر بین ۰ تا ۰/۰۰۱ است.
***	p-value برای پارامتر مورد نظر بین ۰/۰۰۱ تا ۰/۰۱ است.
**	p-value برای پارامتر مورد نظر بین ۰/۰۱ تا ۰/۰۵ است.
*	p-value برای پارامتر مورد نظر بین ۰/۰۵ تا ۰/۱ است.
.	p-value برای پارامتر مورد نظر بین ۰/۱ تا ۱ است.
-	از پارامتر مورد نظر در مدل استفاده نشده است.



شکل ۵- نمودار تعداد شکست‌های پیش‌بینی شده در مقابل تعداد واقعی در مدل لجستیک

۵- نتیجه‌گیری

بررسی نتایج برازش داده‌ها با مدل خطی نشان می‌دهد که مدل مناسبی برای پیش‌بینی نرخ شکست در شبکه‌ی توزیع آب نیست. مدل نمایی ارائه شده در رابطه (۱۸)، نشان می‌دهد که با افزایش طول لوله، تعداد شکست‌ها افزایش می‌یابد و با افزایش مقادیر متغیرهای قطر، تعداد شکست‌ها کاهش می‌یابد. هم‌چنین استفاده از لوله چدن داکتایل باعث کاهش شکست‌ها و برعکس استفاده از لوله‌های با جنس سیمان آزبست، چدن و پلی‌اتیلن، رابطه مستقیم با افزایش تعداد شکست‌ها دارد. مقادیر MSE (میانگین مربعات خطاها) در دو مدل نمایی و خطی تعمیم‌یافته پواسون بسیار نزدیک و کمتر از مدل خطی هستند. بنابراین می‌توان نتیجه گرفت که مدل پواسون، در مقایسه با دو مدل نمایی و خطی مدل مناسب‌تری در پیش‌بینی شکست‌ها است.

مدل رگرسیون لجستیک، با توجه به مقدار پایین MSE برای لوله‌های با شکست صفر، نتایج بسیار خوبی ارائه می‌دهد. در مجموع می‌توان گفت مدل لجستیک، مدل مناسبتری در تخمین و پیش‌بینی احتمال شکست لوله‌های سیستم توزیع پایلوت مورد نظر در بازه زمانی سه سال و نیم بوده است. در سه مدل نمایی، پواسون و لجستیک، نتایج فشار با میزان شکست در پایلوت مورد مطالعه هم‌خوانی نداشته که این موضوع می‌تواند ناشی از دو علت باشد:

اول- داده‌های موجود در مورد فشار کافی نبوده و نیاز به اطلاعات بیشتر بوده و یا مقادیر آن‌ها به اندازه داده‌های موجود در سایر پارامترها نبوده است؛

دوم- تاثیر پارامترهای دیگر نظیر نو بودن لوله‌ها و یا افزایش قطر لوله‌ها بر پارامتر فشار تاثیر داشته است.

۶- تشکر و قدردانی

از شرکت آب و فاضلاب منطقه ۱ تهران برای دریافت داده‌ها و از آقای دکتر موسوی ندوشنی عضو هیئت علمی و آقای آهنی دانشجوی دکترای منابع آب دانشگاه شهید بهشتی برای همکاری در این تحقیق تشکر می‌شود.

۷- پی‌نوشت‌ها

1- Corrosion
2- Time linear model
3- Exponential model

4- Poisson generalized linear model
5- Logistic generalized linear model
6- Probability distribution function
7- LOGIT
8- AKAIKE Information Criterion
9- Deviance
10- Generalized Cross Validation
11- Log likelihood
12- Mean Squared Error

۸- مراجع

بیگی، ف.، (۱۳۷۸)، «آسیب‌شناسی شبکه‌های توزیع آب شهری»، *فصلنامه آب و محیط زیست*، ۳۷، ۱۰-۱۶.

تابش، م.، آقایی، آ.، و ابریشمی، ج.، (۱۳۸۷)، «بررسی نقش عوامل موثر بر فراوانی حوادث در لوله‌های اصلی آبرسانی با استفاده از الگوی رگرسیونی ترکیبی"، *نشریه دانشکده فنی دانشگاه تهران*، ۴۲(۶)، ۶۹۱-۷۰۳.

جلیلی قاضی‌زاده، م.، حنیفی یزدی، س.ح.، و راستی اردکانی، ر.، (۱۳۸۷)، «ارائه روابط پیش بینی وقوع حوادث در شبکه‌های توزیع آب شهری»، *دومین همایش ملی آب و فاضلاب با رویکرد بهره‌برداری*، تهران.

موسوی ندوشنی، س.س.، (۱۳۹۱)، *آشنایی با زبان آماری R*، انتشارات دانشگاه شهید عباسپور.

Agresti, A., and Kateri, M., (2011), *Categorical data analysis*, Springer.

Cameron, A.C., and Trivedi, P.K. (1998), "Regression analysis of count data", 53, Cambridge University.

Everitt, B., and Hothorn, T., (2010), *A handbook of statistical analyses using R*, Second Edition, CRC Taylor and Francis Groups.

Kabir, G., Tesfamariam, S., Francisque, A., and R. Sadiq, (2015), "Evaluating risk of water mains failure using a Bayesian belief network model", *European Journal of Operational Resources*, 240(1), 220-234.

Kettler, A., and Goulter, I., (1985), "An analysis of pipe breakage in urban water distribution networks", *Canadian Journal of Civil Engineering*, 12(2), 286-293.

Kropp, I., Gat, Y.L., and Poulton, M., (2009), "Application of a failure forecast model at the strategic asset management planning level", In: *Proceedings of LESAM 2009*, Miami, USA.

Maindonald, J.H., (2008), *Using R for data analysis and graphics introduction, code and commentary*, Centre for Mathematics and Its Applications, Australian National University.

Montgomery, D.C., Peck, E.A., and Vining, G.G., (2012), *Introduction to linear regression analysis*, 821, John Wiley & Sons.

Nishiyama, M., and Filion, Y., (2013), "Review of statistical water main break prediction models", *Canadian Journal of Civil Engineering*, 40(10), 972-979.

Shamir, U., and Howard, C., (1979), "An analytical approach to scheduling pipe replacement", *Journal of American Water Works Association*, 71(5), 248-258