

Analysis of Indiana's Extension of Graduated Driver Licensing Law

BY: CAROL SADEK

Client: Yudan Wang

Nov 29, 2018

Abstract

The Graduated Driver Licensing Law passed in Indiana in July 2015 put restrictions on drivers ages 18, 19, and 20 until they received their full license at least 180 days after their first license. In this paper, we use logistic and linear regression models as well as survival analysis to analyze the time to first crash of novice drivers. We also use linear models to analyze the number of licenses issued before and after the law was passed. From the logistic regression model, we find that age at license, cohort, and their interaction all have a significant effect on whether or not an individual crashes. From the linear regression model, we find that the same variables as well as month and year all have a significant effect on time to first crash. From the survival analysis, we find that the new law has improved the time to first crash of 16, 17, and 18 year-olds. However, time to first crash decreases in drivers ages 19 and 20 after the law is passed. Finally, from the last linear regression model, we find that a seasonal effect has a significant influence on the number of licenses issued, whereas, the passage of the new law does not.

1 Introduction

In July 1, 2015, the state of Indiana passed a law that extended the Graduated Driver License (GDL) to drivers ages 18 to 20. Before this law was passed, drivers who wished to obtain an unrestricted full driver's license at ages 18, 19, or 20, could do so without first obtaining a probationary license. Under the new law, the probationary license allows 18, 19, and 20 year-olds to drive unsupervised, but under some restrictions. These drivers may receive the probationary license after 180 days of receiving their learner's permit. The probationary license restrictions include:

- Drivers must hold a permit for at least 180 days and complete at least 50 hours of supervised driving practice before receiving probationary license
- Drivers may not use communication devices while driving
- Drivers may not drive between 10 p.m. and 5 a.m. for 180 days after receiving the license
- Drivers may not drive during the following hours until they turn 18 and after 180 days of full licensure:
 - Saturday and Sunday, between 1 a.m. and 5 a.m.
 - Sunday through Thursday, after 11 p.m.
 - Monday through Friday, before 5 a.m.
- Drivers may not have passengers for 180 days after probationary license, unless accompanied in the front seat by a licensed instructor

Note that before July 1, 2015, the Graduated Driver License law applied to 16 and 17 year-olds, however, drivers 18 years and older were able to obtain unrestricted full licenses. After the law was passed, the probationary license with the above restrictions was required for 18, 19, and 20 year-olds as well. The biggest changes resulting from extending the GDL law are:

- 18 to 20 year-olds are not permitted to use communication devices
- 18 to 20 year-olds must complete 50 hours of supervised training under a permit before receiving a probationary license
- 16 year-olds may obtain their probationary license 90 days (instead of 180 days) after their 16th birthday if they completed the permit requirements and taken a driver's education course

We are interested in how this law has affected the crashes of novice (under 21 years old) drivers. We define the “new” cohort as drivers who received their first full license between the ages of 16 and 21 after July 1, 2015 and received their full license at least 180 days after receiving their permit. These are the individuals that were required to obtain a permit before full license under the new law. We define the “old” cohort as drivers between the ages of 16 and 21 who received their first license, either permit or full, prior to the law.

In this report we will answer the following questions:

- Do drivers in the new cohort have a longer crash course, or time period between full licensure and first crash?
- Has the licensing rate been affected by the new law?

To answer these questions, we perform logistic and linear regressions to predict the crash course. Since our data only covers six years and many of our subjects do not have a first crash within those six years, we also use survival analysis to analyze the time to first crash of novice drivers. To analyze the licensing rate, we use a logistic regression for each novice age group.

2 Data

Three datasets were used in this analysis: `all.sas7bdat`, `learnerdriver1.sas7bdat`, and `updated_crash_rates_and_licensing_counts.xlsx`.

The dataset `all.sas7bdat` contains 2,356,392 observations and 30 variables including dates of crashes (`COLLDTE`) and associated driver's license numbers (`dv1`). Each observation is associated with a crash. Multiple drivers can be associated with a crash, and therefore, rows in this dataset may refer to the same crash with different drivers. The variables and their descriptions can be found in Appendix I in Table 10.

The dataset `learnerdriver1.sas7bdat` contains 510,820 observations and 14 variables including the driver's license number (`dv1`) and the dates of the first two licenses (`licensedate21` and `licensedate22`). Each observation is associated with a unique driver who received a license between January 1, 2012 and April 13, 2018 before turning 21. A description of all the variables can be found in Appendix 1 in Table 9.

These two datasets were merged on the driver's license variable `dv1` such that there was one observation per unique driver's license in the `learnerdriver1` dataset. Table 11 in Appendix I describes the columns in the merged dataset `learnerdriver_crashes.sas7bdat`. All individuals in the merged dataset have received full licenses either after receiving their permit, after receiving a different full license, or as their first license. Thus, the age at full license is calculated as the number of years between date at full license and date of birth.

2.1 Data Cleaning

The `all.sas7bdat` contained many duplicate rows. To clean this dataset, we removed all rows with identical driver's license numbers (`dv1`) and collision date (`COLLDTE`) combinations. We also removed driver's license values in ("0", "UNKNOWN", "X", "UNLICENSED", "XXXXXXXXXX", "NONE", "NOLICENSE", "UNK", "NA", "U"). We then merged `all.sas7bdat` with `learnerdriver1.sas7bdat` by driver's license (`dv1`).

2.2 Data Exploration

Figure 1 shows the number of crashes by driver type from January 2013 to December 2017 by quarter. Table 1 shows the equations for line of best fit for each age's crash count with period indicating the number of quarter-years since the first quarter of 2012 and cc_i representing the number of crashes by drivers age i .

Line of Best Fit	Intercept (t -ratio, p -value)	period (t -ratio, p -value)	r^2	RMSE
$cc_{16} = 640.8 + 18.62 \times \text{period}$	$t = 11.83, p < 0.0001$	$t = 4.73, p = 0.0002$	0.55	101.5
$cc_{17} = 278.9 + 10.84 \times \text{period}$	$t = 7.21, p < 0.0001$	$t = 3.86, p = 0.0011$	0.45	72.45
$cc_{18} = 144.5 + 6.027 \times \text{period}$	$t = 6.77, p < 0.0001$	$t = 3.89, p = 0.0011$	0.46	39.95
$cc_{19} = 73.2 + 1.82 \times \text{period}$	$t = 11.15, p < 0.0001$	$t = 3.82, p = 0.0013$	0.45	12.30
$cc_{20} = 36.16 + 1.2555 \times \text{period}$	$t = 6.15, p < 0.0001$	$t = 2.94, p = 0.0088$	0.32	11.02

Table 1. Lines of best fit for crash count by age over time period measured in quarter-years since the first quarter of 2012.

Note that the slopes are all positive, indicating that the number of crashes increases over time for novice drivers, with the largest increase occurring in 16 year-olds. While the p -values indicate that period is a very significant predictor of number of crashes for all age groups, the r^2 values indicate that only 55% of the variance is explained by the 16 year-old crash count line of best fit and even less variance is explained by the other four lines of best fit. Regardless, the increase in number of crashes over time might be due to the economy boost after 2012 and resulting decrease in gas prices as seen in Figure 2, which leads to more drivers on the road. The equation for the line of best fit of experienced drivers, ages 25 to 34 is

$$\text{experienced crash count} = 13219 - 62.66 \times \text{period}$$

with $\text{RMSE} = 1093$ and $r^2 = 0.11$ indicating a decrease in the number of crashes of experienced drivers over time with period being a very significant predictor ($t = -3.45, p = 0.0008$). In order to account for trends due to factors such as gas prices, economy changes, and weather, we create a ratio of novice to experienced driver crashes, where experienced drivers are 25 to 34 years old.

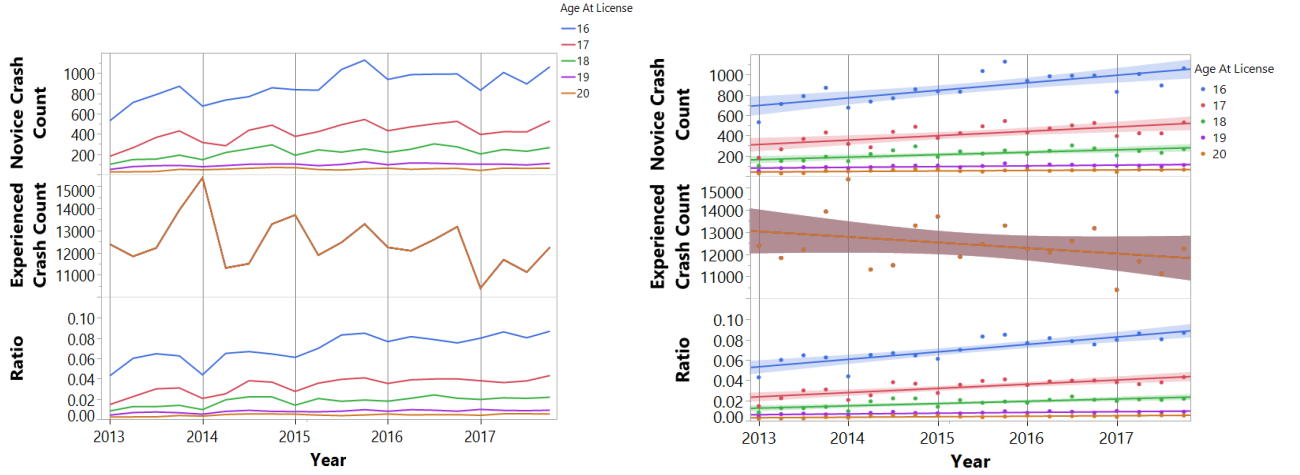


Figure 1. Number and ratio of crashes of novice and experienced drivers over length of data collection time period. The right plot shows the lines of best fit with confidence intervals.

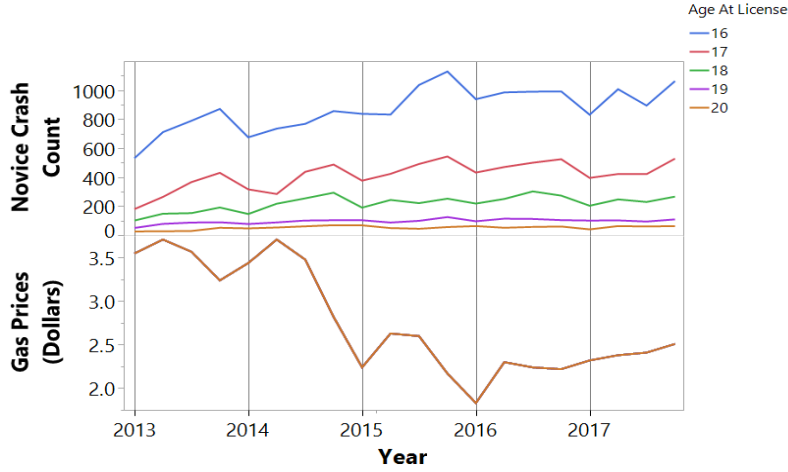


Figure 2. Gas prices in dollars over time compared with number of novice crashes by age.

3 Models and Results

3.1 Unpaired t -Tests for Ratio of Crashes

In order to compare the number of crashes of novice drivers with a control group, we aggregated the crash data by quarter years for novice drivers (drivers who received licenses at ages 16 to 20) and for experienced drivers (drivers who received licenses at ages 25 to 34). Figure 3 shows histograms and normal quantile plots of $\frac{\# \text{ novice crashes}}{\# \text{ experienced crashes}}$ by age at license and cohort. The y -axes in the normal quantile plots show the ratio of number of novice crashes to number of experienced crashes, and the x -axes show the expected normal scores for each y value. Assuming the sample data follows a normal distribution, we would expect the sample points to fall on the red diagonal straight line seen in each plot below. The red bounds are the Lilliefors 95% confidence limits. If the sample data comes from a normal distribution, we would expect the points to not leave the bounds more than 5% of the time.

Note that the sample points do not leave the Lilliefors bounds in any of the normal quantile plots below. These plots show a fairly normal distribution for all ratios with most sample points lying close to the diagonal red line. Using the Shapiro-Wilk test of normality with $\alpha = 0.05$, we

fail to reject the assumption of normality for all ratios as seen in Table 2. This implies that there is not strong evidence against normality.

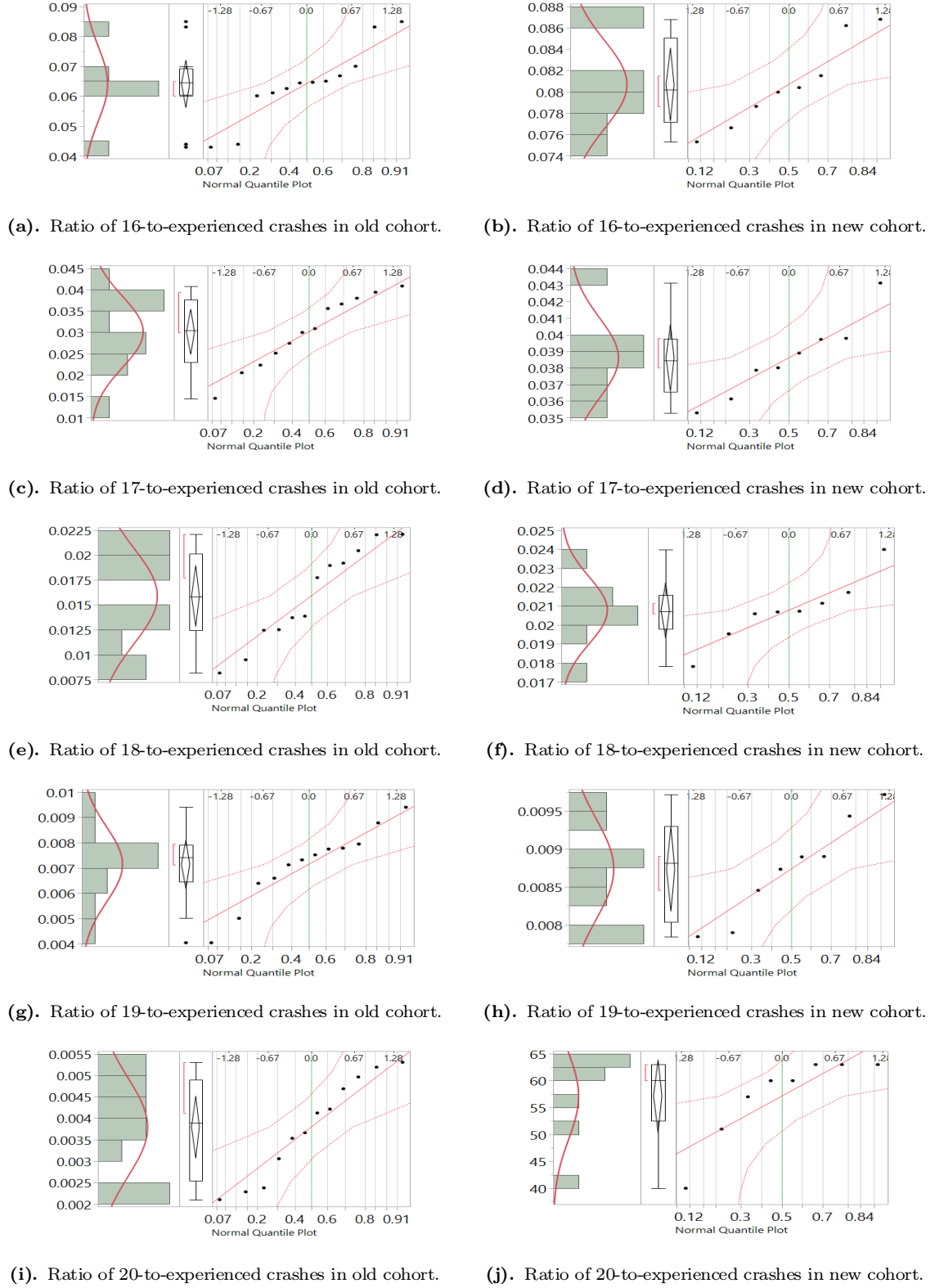


Figure 3. Histograms and Normal Quantile Plots for ratio of number of novice crashes to number of experienced (25-34 year-olds) crashes. The upper x-axis shows the normal quantile scale and the lower x-axis shows the empirical cumulative probability for each y value. The dashed red lines are the Lilliefors confidence bounds.

Model for Old Cohorts	W	Prob< W	Model for New Cohorts	W	Prob< W
(a) 16-to-exp crashes	0.9055	0.1870	(b) 16-to-exp crashes	0.9303	0.5190
(c) 17-to-exp crashes	0.9517	0.6624	(d) 17-to-exp crashes	0.9556	0.7674
(e) 18-to-exp crashes	0.9280	0.3591	(f) 18-to-exp crashes	0.9437	0.6480
(g) 19-to-exp crashes	0.9458	0.5773	(h) 19-to-exp crashes	0.9456	0.6667
(i) 20-to-exp crashes	0.9275	0.3544	(j) 20-to-exp crashes	0.9058	0.3253

Table 2. Shapiro-Wilk Goodness-of-Fit test for normality show that at the $\alpha = 0.01$ level, we fail to reject the assumption of normality.

Under the assumption of normality of the ratios, we performed an unpaired t -test to analyze the difference in means between the old and new cohorts. Our null hypothesis states that the mean of the ratio of novice-to-experienced crashes in the old cohort is equal to the ratio of novice-to-experienced crashes in the new cohort.

Assuming unequal variances in the two cohorts, the p -values in Table 3 show that, with $\alpha = 0.05$, there is a very significant difference in means for 16, 17, 18, 19, and 20 year-old ratios between the old and the new cohort. We fail to reject the null for all ages combined. Note that we expect the number of novice crashes in the new cohort to be less than the number of crashes in the old cohort. Thus, we also perform a one-sided t -test, shown in Table 3. Again, at the $\alpha = 0.05$ level, we find a very significant difference in means with larger means in the new cohort for each novice-to-experienced-driver ratio, but not for all ages combined.

p -values	t -ratio	$H_N: \mu_O = \mu_N$	$H_N: \mu_O \geq \mu_N$
16 year-olds	$t = -4.25$	Prob $> t = 0.0008^*$	Prob $< t = 0.0004^*$
17 year-olds	$t = -3.33$	Prob $> t = 0.0052^*$	Prob $< t = 0.0026^*$
18 year-olds	$t = -3.22$	Prob $> t = 0.0058^*$	Prob $< t = 0.0028^*$
19 year-olds	$t = -3.10$	Prob $> t = 0.0067^*$	Prob $< t = 0.0033^*$
20 year-olds	$t = -1.71$	Prob $> t = 0.0279^*$	Prob $< t = 0.0139^*$
All ages	$t = -2.44$	Prob $> t = 0.0276$	Prob $< t = 0.1138$

Table 3. Unpaired t -tests for ratio of novice to experienced crashes by cohort.

3.2 Logistic and Linear Regressions for Crash Course

In this section, we use a logistic and linear regression to answer the following question: Do drivers in the new cohort have a longer time period between full licensure and first crash as compared with drivers in the old cohort? We performed a logistic regression to predict the probability that an individual does not crash after receiving their full license using their age at full licensure and their cohort as predictors. Both predictors and their interaction have a very significant effect ($p < 0.0001$) on the occurrence of a crash. Figure 4 shows that the probability of no crash is higher in the new cohort for drivers receiving their licenses at ages 15, 16, 17, 18, or 19. However, drivers receiving their licenses at age 20 in the old cohort have a higher probability of no crash. The maximum likelihood estimates are shown in Table 4.

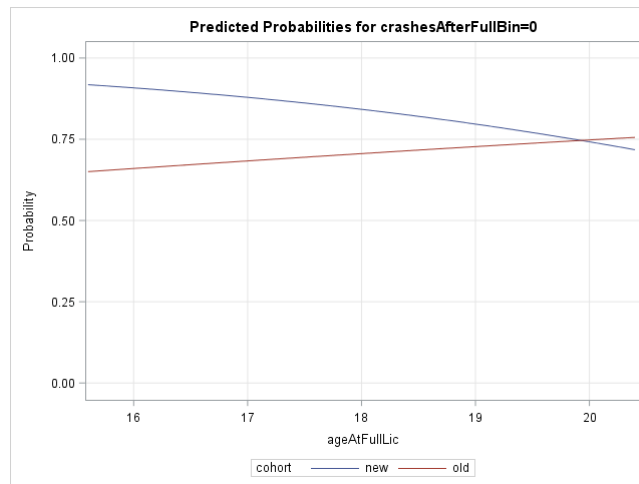


Figure 4. Logistic regression model predicting no crash over course of study.

Predictor		Estimate	Standard Error	Pr>ChiSq
Intercept		3.10	0.0479	<0.0001
AgeAtFullLicense		-0.10	0.0027	<0.0001
cohort	new	4.13	0.0479	<0.0001
AgeAtFullLicense*cohort	new	-0.21	0.0027	<0.0001

Table 4. Maximum likelihood estimates of logistic regression model predicting likelihood of novice non-crashes.

We performed a linear regression to predict the amount of time until the first crash after full licensure. We used age at licensure, cohort, year, and month of first crash as predictors. Cohort was coded 0 if “old” and 1 if “new”. We found that age at licensure ($t = 135.93, p = <0.0001$), cohort ($t = -182.67, p < 0.0001$), year of crash ($t = 259.80, p < 0.0001$), and month of crash ($t = 51.18, p < 0.0001$) all have a very significant linear effect on time to first crash. The residuals (actual time to crash – expected time to crash) histogram and normal quantile plots are shown in Figure 5. Note that the residuals appear normally distributed, and therefore, this is an appropriate analysis for the data. Table 5 shows the results of the linear regression model. The estimated regression equation is

$$\text{timeToCrash} = -4.3376 + 0.2214 \times \text{ageAtFullLicense} - 1.2782 \times \text{cohort} + 0.508 \times \text{yearOfCrash} + 0.0376 \times \text{monthOfCrash}$$

Note that time to crash is positively correlated with age at license as expected. This implies, drivers who receive their licenses when older are likely to have a longer time to crash. Cohort is negatively correlated with time to crash, signifying that when cohort=“new”, time to crash is shorter.

Predictor	Estimate	Standard Error	t -value	$\text{Pr}> t $
Intercept	-4.3376	0.03191	-135.94	<0.0001
AgeAtFullLicense	0.22134	0.00163	135.93	<0.0001
cohort	-1.2782	0.00700	-182.67	<0.0001
yearOfCrash	0.5080	0.00196	259.80	<0.0001
monthOfCrash	0.0376	0.00073	51.18	<0.0001

Table 5. Linear regression estimates predicting time to first crash after full licensure for novice drivers.

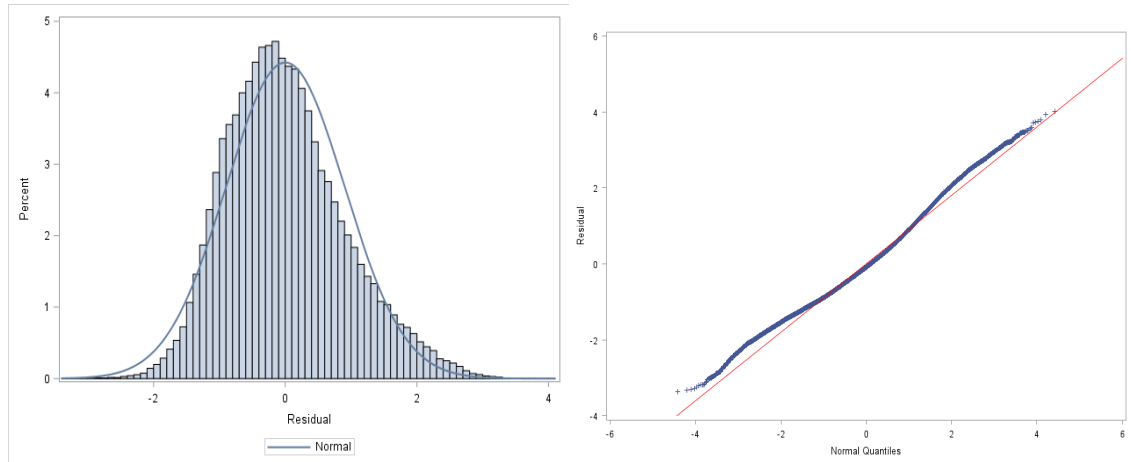


Figure 5. Residual histogram and normal quantile plots for linear regression to predict time to crash.

3.3 Survival Analysis for Crash Course

In this section, we use survival analysis to further study the effect of the law on the length of time to a driver's first crash.

Survival analysis is used to study the time to an event occurrence in longitudinal studies. This analysis allows for the use of censored observations, subjects that have either not yet experienced the event within the data collection period or will never experience it. There are two variables used in survival analysis: censor, a binary variable that indicates whether or not the event occurred, and time to occurrence, a variable that measure the time to the event or time until the end of the data collection for individuals who have not yet experienced the event. For the linear regression model, we only used the subjects who experienced the event, a crash. However, subjects who have not yet crashed are also of interest.

We use the hazard function $h(t_{ij})$ to measure whether and when an event occurs. This function is the conditional probability that subject i will experience the event in time period j given that he/she has not already experienced it. The discrete-time hazard function can be expressed as

$$h(t_{ij}) = \Pr[T_i = j | T_i \geq j]$$

where T_i is the time period when subject i experiences the event. We can estimate this conditional probability using the ratio of the number of individuals who experienced the event in time period j to the number of individuals at risk during that time period:

$$\hat{h}(t_j) = \frac{(n \text{ events})_j}{(n \text{ at risk})_j}.$$

The larger the hazard in time period j , the higher the risk for an individual to experience the event in that time period. We can also use the survival function

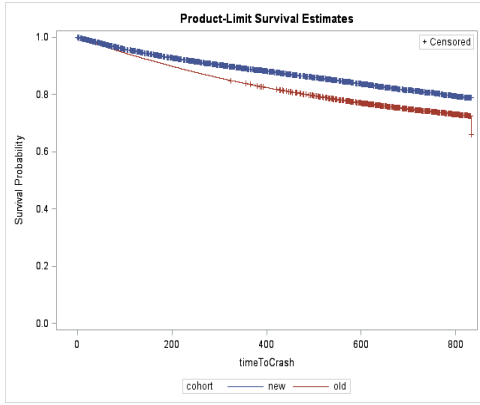
$$S(t_{ij}) = \Pr[T_i > j]$$

to calculate the probability that individual i will survive, or not experience the event, up to or beyond time period j . We can estimate the survival probability by using the probability of survival from the previous time period and multiplying with one minus the risk that the event will occur in the current time period:

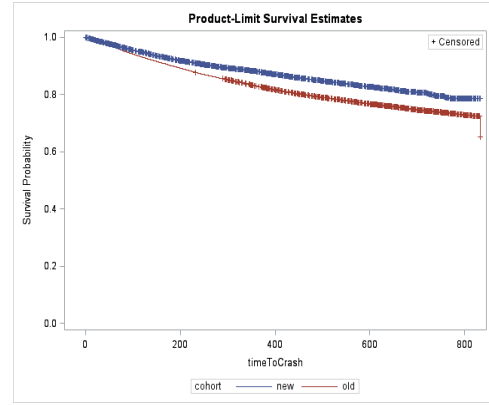
$$\hat{S}(t_j) = \hat{S}(t_{j-1})[1 - \hat{h}(t_j)].$$

In this analysis, we identify the event as a driver's first crash after receiving their first full license on record. Note that many subjects do not have a first crash within the scope of the data collection. The beginning of time is January 1, 2012, the earliest crash record in our dataset. The metric used is days between first full license and first crash after full license. We censor subjects who have not had a crash after their first full license before the censoring date, April 13, 2018, the last day of data collection. Note that this is non-informative censoring, meaning the subjects who have not experienced the event at the end of the collection period are representative of the population who would have experienced it if the study had been extended. We also call it right-censoring, meaning the event has not yet occurred or will never occur for censored subjects.

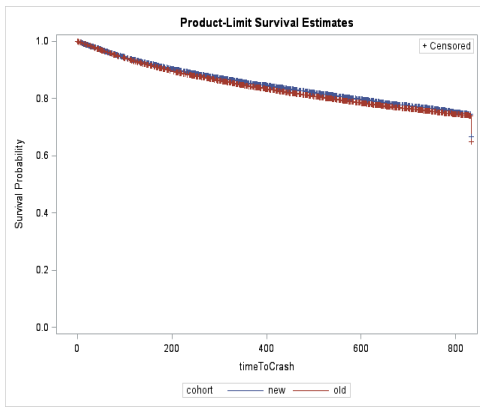
To perform this analysis, we used the Kaplan-Meier product limit survival estimate in Figure 6 on the length of time in days between full licensure and first crash after full licensure. The Kaplan-Meier analysis allows us to use subjects studied for different lengths of time. The follow-up time for the variable `timeToCrash` was set to 833, the maximum number of days an individual from the new cohort could be followed. If an individual did not have his/her first crash after full licensure before April 13, 2018, he/she was given a `censor` value of 1. Otherwise, `censor`=0.



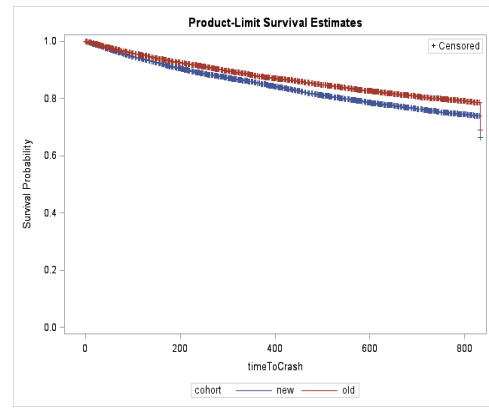
(a). Time to first crash of 16 year-olds.



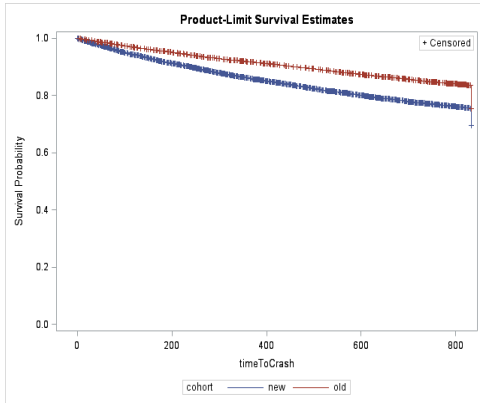
(b). Time to first crash of 17 year-olds.



(c). Time to first crash of 18 year-olds.



(d). Time to first crash of 19 year-olds.



(e). Time to first crash of 20 year-olds.

Model	χ^2	$\text{Pr} > \chi^2$
Time to first crash, drivers licensed at 16	571.42	<0.0001
Time to first crash, drivers licensed at 17	228.38	<0.0001
Time to first crash, drivers licensed at 18	5.56	0.0183
Time to first crash, drivers licensed at 19	40.35	<0.0001
Time to first crash, drivers licensed at 20	162.04	<0.0001

(f). Table of p -values for log-rank test of equality.

Figure 6. Survival Plots for old (licensed before the law, red line) and new (licensed after the law, blue line) cohorts by age at full license.

Figure 6 shows that drivers in the new cohort, indicated by the blue lines, who received their full license at ages 16, 17 or 18, had a delayed time until their first crash as compared with the old cohort, indicated by the red lines. However, drivers in the new cohort who received their full license at ages 19 or 20 appear to have a shorter time to crash than their old cohort. This is likely due to the fact that drivers labeled as part of the old cohort include drivers that received their full license after the GDL law was passed, but who did not have a permit for at least 180 days before

receiving their full driver license. The p -values for the log-rank test are shown in Figure 6d. Note that the survival distributions are statistically different for all age groups at the $\alpha = 0.01$ level, and only the 18 year-old survival distributions are not significantly different at the $\alpha = 0.05$ level.

3.4 Linear Regression for License Counts

Figure 7 shows the number of licenses issued for all novices ages 16 to 20 over the course of the study. Note that there is a spike in number of licenses issued at the time the GDL is passed. Figure 8 shows the number of licenses issued for each of the novice age groups. Note that the large spike in licenses issued is most apparent in the 16 year-old age group. This is most likely due to the fact that under the new law, 16 year-olds were allowed to receive their full licenses after 90 days instead of 180 if they passed a driver's education course. Note that there appears to be a seasonal effect, as seen in Figure 8b, with a larger number of licenses issued during the summer months (June, July, and August) and fewer during the winter months.

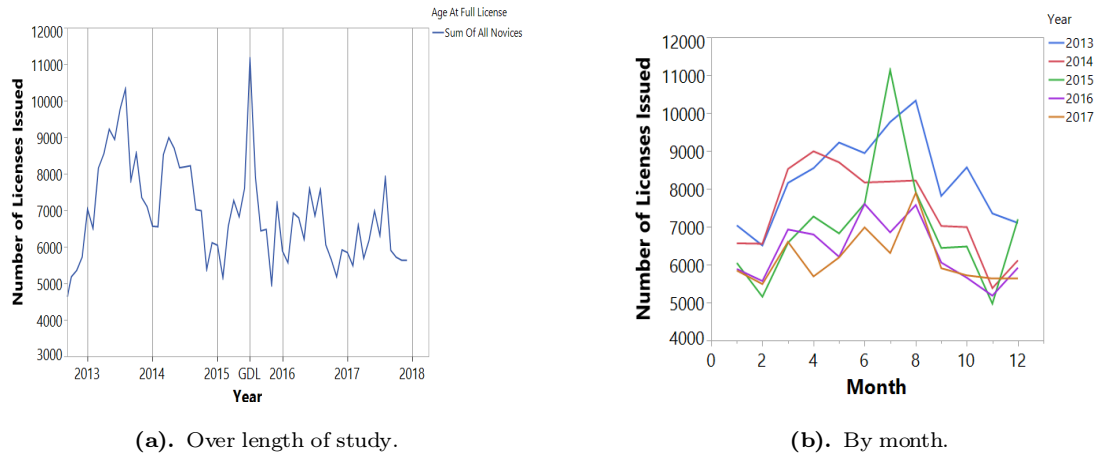


Figure 7. Number of full licenses issued for 16 to 20 year olds.

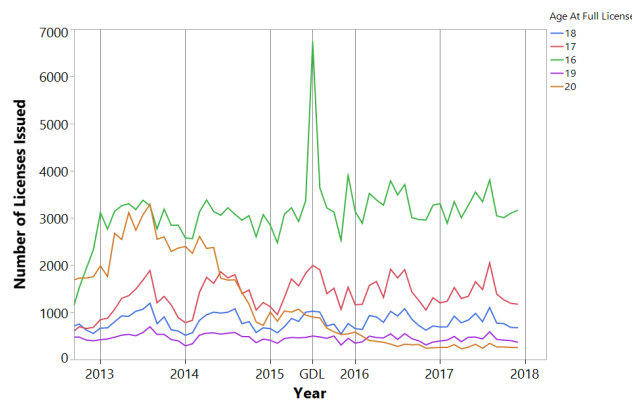


Figure 8. Number of full licenses issued for 16 to 20 year olds by age and by quarter year.

We performed a linear regression to predict the number of licenses using age at full licensure, cohort, year, and three seasonal dummy variables (s_1 , s_2 , s_3) associated with quarter-year as specified in Table 6.

$s1$	Jan-March
$s2$	Apr-June
$s3$	Jul-Sept
intercept	Oct-Dec

Table 6. Seasonal dummy variable definitions. Each variable is given a value of 1 if within the range of months and 0 otherwise.

Table 7 shows that at the $\alpha=0.05$ level, only the intercept, which is the fourth seasonal dummy variable, the age at full license, and the second seasonal dummy variable had significant linear effect on the number of full licenses issued. The estimated regression equation is

$$\text{numLicIssued} = 9359.02 - 447.25 \times \text{ageAtLicense} - 90.01 \times \text{cohort} - 15.85 \times \text{year} + 74.04 \times s1 + 291.49 \times s2 + 232.91 \times s3.$$

Predictor	Parameter Estimate	Standard Error	t -ratio	p -value
intercept	9359.02	603.75	15.50	<0.0001
age	-447.25	32.87	-13.61	<0.0001
cohort	-90.01	190.54	-0.47	0.64
year	-15.85	59.70	-0.27	0.79
$s1$	74.04	140.26	0.53	0.60
$s2$	291.49	140.26	2.08	0.04
$s3$	232.91	125.87	1.85	0.07

Table 7. Linear regression to predict number of licenses issued for all novice ages combined.

We performed a linear regression to predict the number of licenses for each age individually using cohort, year, and the same three seasonal dummy variables ($s1, s2, s3$) as predictors. Table 8 shows the results by age. Note that cohort has a significant linear effect only for drivers receiving their licenses at age 20. The intercept, or the fourth quarter-year dummy variable is the only one that has significant linear effect on the number of full licenses issued for all age groups.

Predictor / Age	16	17
intercept	$t = 10.17, p < .0001$	$t = 8.75, p < .0001$
cohort	$t = 0.52, p = 0.61$	$t = -1.21, p = 0.23$
year	$t = 1.51, p = 0.14$	$t = 3.23, p = 0.002$
$s1$	$t = 0.23, p = 0.82$	$t = 0.85, p = 0.40$
$s2$	$t = 1.31, p = 0.20$	$t = 3.03, p = 0.004$
$s3$	$t = 0.75, p = 0.46$	$t = 3.22, p = 0.002$

Predictor / Age	18	19	20
intercept	$t = 17.49, p < .0001$	$t = 22.37, p < .0001$	$t = 15.47, p < .0001$
cohort	$t = 0.31, p = 0.75$	$t = -0.27, p = 0.79$	$t = -2.05, p = 0.045$
year	$t = -0.35, p = 0.73$	$t = -1.28, p = 0.21$	$t = -5.43, p < .0001$
$s1$	$t = 0.43, p = 0.67$	$t = 0.22, p = 0.83$	$t = 2.27, p = 0.03$
$s2$	$t = 5.17, p < .0001$	$t = 3.70, p = 0.001$	$t = 2.64, p = 0.01$
$s3$	$t = 6.08, p < .0001$	$t = 5.31, p < .0001$	$t = 2.01, p = 0.05$

Table 8. Linear regression to predict number of licenses issued by age.

Figure 9 shows the residual histogram and normal quantile plot for this regression model. Note that the residuals do not appear to follow a normal distribution. Adjusting the seasonality variable such that $s1$ started in December instead of January and modifying the other two variables such that each one represented 3 months provided similar residual plots, appearing to not well follow a normal distribution. Transforming the data using a log transform as well as a square root transform also provided similar residual plots.

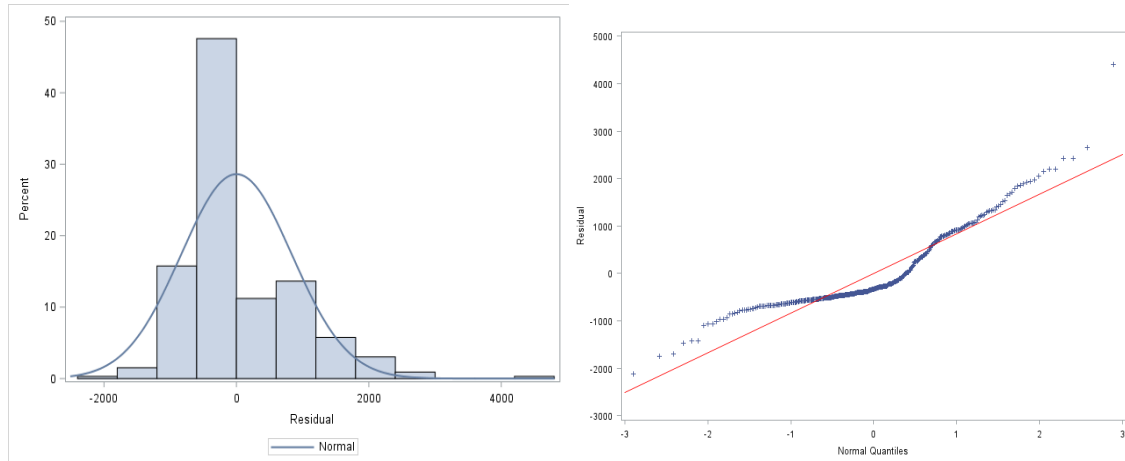


Figure 9. Residual histogram and normal quantile plots for linear regression of number of licenses issued to all novices.

4 Conclusion

The survival analysis model showed that the new Graduated Driver's License law had a significant effect on the time to first crash for drivers ages 16, 17, and 18. However, drivers who received their full licenses at ages 19 or 20 showed the opposite result. This may be due to our definition of the cohort variable and missing permit information for the older novices. Our logistic and linear models for crash course showed similar results with cohort and ages at license having significant effects on crash course.

Our licensing linear regression model showed that the variable that influenced the number of licenses issued to novice drivers across all age groups was one of the seasonal effects. It also showed that the cohort variable did not have a significant effect on this number. However, the residuals do not appear to be normally distributed, and therefore, a non-linear regression might be more suited to answer this question.

5 References

- Augustine, Glenn. "How a New Indiana Law May Save Teens' Lives." *Indianapolis Star*, IndyStar, 5 July 2015, <https://www.indystar.com/story/opinion/columnists/2015/07/02/new-state-law-may-save-teens-lives/29628509/>
- Curry, AE, Foss, RD, Williams, AF. Graduated driver licensing for older novice drivers: Critical analysis of the issues. *American Journal of Preventive Medicine*. 2017; Volume 6.
- Bloyd, Kyle. "ISP Hopes New Driving Laws Reduce Crashes." *WISH*, WISH, 1 July 2015, <https://www.wishtv.com/news/local-news/isp-hopes-new-driving-laws-reduce-crashes/1116324009>

Bureau of Motor Vehicles. “Probationary Driver’s License”. <https://www.in.gov/bmv/2583.htm>

Fell, JC, Jones K, Romano E, Voas R. An evaluation of graduated driver licensing effects on fatal crash involvements of young drivers in the United States. *Traffic Inj Prev.* 2011; 12(5): 423-31.

Indiana Graduated Driver’s Licensing System. https://www.in.gov/bmv/files/Indiana_Graduated_Drivers_License_System.pdf

Masten, SV and Robert, DF. Long-term effect of the North Carolina graduated driver licensing system on licensed driver crash incidence: A 5-year survival analysis. *Science Direct*, 2010. 42(6): 1647-1652.

Singer, JD and Willett, JB. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, 2003.

Tan, WD, Gan, FF, Chang, TC. Using normal quantile plot to select an appropriate transformation to achieve normality. *Science Direct*, 2004. 45(3): 609-619.

6 Appendix I

The dataset `learnerdriver1.sas7bdat` contains all the licenses received by drivers younger than 21 years old. The variables are listed in Table 9.

Variable Name	Type	Description
<code>dvl</code>	char	Driver’s license number
<code>name</code>	char	Name associated with <code>dvl</code>
<code>dob2</code>	mmddyy10	Date of birth
<code>licenseage2</code>	float	Age at second license
<code>licensedate21</code>	mmddyy10	Date of first license associated with <code>dvl</code> in data
<code>licensedate22</code>	mmddyy10	Date of second license associated with <code>dvl</code> in data
<code>licenseage1n</code>	int	<code>int(licenseage1)</code>
<code>licenseyear1</code>	int	<code>year(licensedate21)</code>
<code>licenseage2n</code>	int	<code>int(licenseage2)</code>
<code>licenseyear2</code>	int	<code>year(licensedate22)</code>
<code>learnerdriver</code>	int	1 if first license is “LEARNER PERMIT” and second license is in (“OPERATOR”, “OPERATOR (4 YR)”, “OPERATOR (4 Year) NonProb”) <p>2 if age at first license is in [18, 21) and first license is in (“OPERATOR”, “OPERATOR (4 YR)”, “OPERATOR (4 Year) NonProb”) but not “LEARNER PERMIT”</p>
<code>cohort</code>	char	“old” if <code>licenseage1n</code> <18 and <code>licensedate21</code> <’30JUN2015’ or <code>learnerdriver</code> =1, driver is between 18 and 21, and received second license after at least 180 days or <code>learnerdriver</code> =2 <p>“new” otherwise</p>
<code>inter</code>	int	1 if first license is “LEARNER PERMIT” and second license is in (“OPERATOR”, “OPERATOR (4 YR)”, “OPERATOR (4 Year) NonProb”) and time between first two licenses is at least 180 days <p>0 otherwise</p>
<code>interdate</code>	mmddyy10	180 days after first full license in data

Table 9. Description of variables in `learnerdriver1.sas7bdat`, a dataset with all licenses received at ages younger than 21 between January 1, 2012 and July 12, 2018.

The dataset `all.sas7bdat` contains all the variables listed in Table 10 as well as all the variables in Table 9 except `learnerdriver`. This dataset is created by merging the data from `alldrivers.sas7bdat`, which contains all the crashes between January 1, 2012 and April 13, 2018, and `learnerdriver1.sas7bdat`, which contains all the licenses received by drivers younger than 21 years old. The code used to create this dataset is in `redefine_intermediate_license_10-8.sas`.

Variable Name	Type	Description
MSTRRECNBRTXT	char	State Repository Unique Identifier; nine digit number printed on top of each crash report
UNITNMB	int	Identifies the unit number
UNITTYPECODE	char	Indicates the type of unit involved
VEHICLENMB	int	Indicates the vehicle number
OCCUPSNMB	int	Number of occupants
personnmb	int	Identifies the person number
agenmb	int	Calculates the individual's age using Date of Birth and Date of Collision
gendercdo	char	Gender code: "M", "F", "U" or " "
posinvehcde	int	Position in or on the vehicle 01 – Front Left 02 – Front Center 03 – Front Right 04 – Rear Left 05 – Rear Center 06 – Rear Right 07 – Bed/Sleeper/Third Seat Row 08 – Outside Front 09 – Outside Left 10 – Outside Right 11 – Outside Rear 12 – Bed/Sleeper/Third Seat Row Left 13 – Bed/Sleeper/Third Seat Row Center 14 – Bed/Sleeper/Third Seat Row Right
COLLDTE	date9	Date of crash
COLLDAYWEEKCDE	int	Day of week when crash occurred 1 – Sunday 2 – Monday 3 – Tuesday 4 – Wednesday 5 – Thursday 6 – Friday 7 – Saturday
COLLTIMETXT	char	Actual local time when crash occurred
COLLTIMEAMPMTXT	char	AM/PM indicator ("AM", "PM", " ")
collmon	int	Month of collision (1, 2, ..., 12)
collyr	int	Year of collision
comp	int	1 if <code>agenmb</code> is between 25 and 34, inclusive 0 otherwise
intercrash	int	0 if no crash occurs 1 if <code>COLLDTE</code> is between <code>licensedate22</code> and <code>interdate</code> 2 if crash occurs but not within interim time period

Table 10. Description of variables in `all.sas7bdat`, a dataset with all crashes occurring between January 1, 2012 and April 13, 2018.

Table 11 shows the descriptions of variables in `learnerdriver_crashes.sas7bdat`. This dataset was created by merging `all.sas7bdat` and `learnerdriver1.sas7bdat` for novice drivers. Each observation is associated with a unique driver's license (`dv1`).

Variable Name	Type	Description
<code>dv1</code>	char	Driver's license number (unique identifier)
<code>dob</code>	mmddyy10	Date of birth
<code>cohort</code>	char	"old" or "new" from <code>learnerdriver1.sas7bdat</code>
<code>ageAtFullLic_f</code>	float	Floating point age at full license
<code>ageAtFullLic</code>	int	Integer age at full license
<code>fullLicDTE</code>	date9	Date full license received
<code>permitDTE</code>	date9	Date permit received
<code>totNumCrashes</code>	int	Total number of crashes in entire time span of data collection
<code>numCrashesB4Full</code>	int	Number of crashes before full license received
<code>numCrashesAfterFull</code>	int	Number of crashes after full license received
<code>firstCrashDTE</code>	date9	Date of first crash
<code>firstCrashAfterFullDTE</code>	date9	Date of first crash after full license
<code>numCrashesDurInter</code>	int	Number of crashes after full license before <i>interdate</i>
<code>firstCrashDurInterDTE</code>	date9	Date of first crash during intermediate time period
<code>numCrashesAfterInter</code>	int	Number of crashes after end of intermediate time period
<code>firstCrashAfterInterDTE</code>	date9	Date of first crash after end of intermediate time period

Table 11. Description of variables in `learnerdriver_crashes.sas7bdat`, the merged dataset used in time to first crash models.