A ZERO-INFLATED ANALYSIS OF THE 2020 TOUR DE FRANCE

By Gian Carlo Di-Luvi*,

*University of British Columbia**

The Tour de France is the most prestigious cycling race in the world. Tadej Pogačar won the 2020 Tour on the last possible stage after surprisingly beating favorite Primož Roglič—who had led the Tour for 11 stages. But was Pogačar's victory something we should have foreseen? In this report, we propose a zero-inflated model based on the Tweedie distribution to study how unlikely, or not, Pogačar's victory was. Specifically, we model the time difference of each rider to the race leader and train the model on data from the first 19 stages of the Tour. We then predict the results of stage 20—in which Pogačar won the Tour—finding that although the model predicted Roglič to be the Tour's winner, Pogačar had a 23% estimated probability of winning as well.

1. Introduction. The Tour de France is cycling's best known and most prestigious race. In its current format, multiple teams of riders compete in 21 races (called stages) spred across 23 days. Many competitions take place at the Tour each year. In the points classification, points are given to stage winners and during intermediate sprints; in the mountains classification, riders who arrive first to mountain summits get points. The most important competition, however, is the general classification. Succinctly, the general classification is won by the rider who completes the Tour in the least amount of time. The rider who wins the general classification is said to have won the Tour, and in this report we refer to the race leader as the rider who is currently leading the general classification.

The 2020 Tour de France saw 175 riders organized in 22 teams compete for the general classification. The favorite to win the Tour was Slovenian Primož Roglič, who is currently considered to be the best cyclist in the world.[1] Roglič led the Tour for 11 stages but lost the lead to compatriot and youngster Tadej Pogačar in stage 20. Tradition dictates the last stage (i.e. stage 21) to be ceremonial: cyclists ride to Paris while celebrating that the Tour is over, and no positions are contested. Pogačar thus maintained his lead and won the Tour. But Pogačar not only beat Roglič in the last possible stage—he did it against all odds and by a significant time margin of

[1]See https://www.uci.org/road/rankings.

59 seconds (after being down 57 seconds in stage 19).

In this report, we explore whether Pogačar's victory was truly as unpredictable as people took it to be, or if there is evidence to suggest his chances of winning were underestimated. For this purpose, we develop a zero-inflated model based on the Tweedie distribution to study the time difference of each rider to the race leader. The Tweedie distribution has been studied extensively in the actuarial science and rainfall literatures, where zero-inflated phenomena occur frequently (see Jørgensen and Paes De Souza, 1994; Withers and Nadarajah, 2011, and references therein). It has also been used to model dollar-denominated outcomes (Lauderdale, 2012) and even to predict cyanobacterial biovolume from a Bayesian perspective (Haakonsson et al., 2020). We train this model on data scraped from the Tour's official website (`letour.fr`).

6: Normalized residuals of the compound Poisson-Gamma models, withcolor by response function. The vast majority of values are well within the±3 bound. The models tend to overestimate when fitted values are smalland underestimate when they are large.overcome this would be to restrict the values of the parameter associatedwith stage number to be non-negative, which would allow that covariateto linearly influence the response. Given the observed positive associationbetween stage number and time difference, this might be reasonable. Therest of the parameters can be scaled using another response function—whichwould result in a non-linear model—or alternatively they can be similarlyrestricted. Whether optimization methods are guaranteed to converge in thissetting is also left for future work. Another alternative would be to use aBayesian approach and directly incorporate the parameter restrictions intotheir prior distributions. Unfortunately, popular probabilistic programminglanguages such asStando not have a built-in Tweedie regression family,which would require the user to develop code to approximate the density.Another limitation is that the model assumes independence between differ-ent stages of the same rider. This can be partly justified by the fact that riderscan almost arbitrarily lose or win time difference at any given stage. However,

Section 2 describes the data collection and summarises the data set. In Section 3 we do an exploratory data analysis to inform modeling. We describe the model and some modifications in Section 4, and then show the results of fitting the model in Section 5. Section 6 contains concluding remarks and directions for future work.

2. Data description. The data contain information of each cyclist and stage of the 2020 Tour de France. A summary of all the variables included in the data set can be found in Table 1. The data are public and were scraped from the Tour's official website[2] using `Python` and the 2020 Tour de France Wikipedia webpage[3] by hand. They were then combined into a single data set containing 3,390 observations and 23 variables.

Each row corresponds to an observation and each column to a variable. The observations are at the "rider by stage" level, i.e., each observation corresponds to a rider at a given stage. Some variables, such as the stage winner's country, are constant across stages. The number of riders is not the same at each stage because some riders leave the race due to injuries. Specifically, 175 riders started the tour but only 146 finished it. There are five different stage types in this data set: flat, medium mountain, hilly, mountain, and mountain time trial. Other stage types exist in the sport but were not included in this year's Tour de France (e.g. team time trial). The cumulative time is determined as the sum of the times at each stage, minus the total bonus seconds, plus the total penalty seconds. Bonus points are awarded to the first three riders to finish each stage; penalties are impossed when cyclists break rules, such as receiving food close to the finish line. Finally, race leader is used interchangeably with general classification leader because that is the most important event of the Tour, even though there are many races going on at the same time.

3. Exploratory data analysis. There are many ways to predict the winner of the Tour with the information of the first 19 stages. In this section we investigate two potential variables, cumulative time and time difference to leader. The former is not amenable to statistical model due to the combinatorial restrictions that need to be accounted for. We thus explore the latter in more detail and in preparation for modeling in Section 4. It should be noted that stage number plays an important role because, as the race progresses, the riders become more tired.

To address the statistical question, we could directly work with the rank at each stage. However, working with this variable leads to many statistical complications and complex combinatorial restrictions. We instead explore variables that are not exactly the rank, but that can be used to determine it and are more amenable to statistical modelling. Figure 1 shows the cumula-

[2] `letour.fr`.

[3] `https://en.wikipedia.org/wiki/2020_Tour_de_France`.

Variable	Description	Type of variable
rank	rider's rank at that stage	integer
rider	rider's name	string
rider_number	rider's bib number	integer
team	rider's team	string
time	rider's time in that stage (hms format)	string
bonus	rider's bonus seconds in that stage	float
penalty	rider's penalty seconds in that stage	float
stage	stage number	integer
date	date of stage	date
distance	distance of stage in km	float
origin	origin of stage	string
destination	destination of stage	string
stage_type	type of stage (mountain, flat, etc.)	string
winner_country	country of rider who won stage	string
general	bib number of rider leading the race	integer
points	bib number of rider leading the points race	integer
mountains	bib number of rider leading the mountains race	integer
young	bib number of rider leading the young race	integer
stage_winner	bib number of rider that won the stage	integer
time_seconds	rider's time in seconds in that stage	float
cum_time	rider's cumulative time at that stage (including bonus and penalty seconds)	float
gc_rank	rider's rank in the general classification	integer
timediff	rider's time difference in seconds to race leader	float

TABLE 1
Variables included in the data set.

tive time of riders by stage. Cumulative time seems to increase linearly by stage, which suggests that a linear regression model might be appropriate. There are, however, some issues with this variable. First, the variance of the observations increases with stage. This is due to the fact that cumulative time is determined by adding individual stage times. The more times are added, the more variability there will be. Second, the observations are not independent: clearly a rider's cumulative time has to increase from one stage to the next. This also imposes a non-trivial combinatorial constraint that makes cumulative time not amenable to statistical modeling.

One alternative to cumulative time is the time difference of every rider to the race leader at every stage. A visual inspection of Figure 2, which shows how time differences vary through stage, suggests a linear relationship between stage and time difference. As was the case of cumulative time, the variance of the time difference also increases with stage—for similar reasons. Although time difference also resulst in a combinatorially complicated structure, it is not nearly as complicated as the one induced by the cumulative

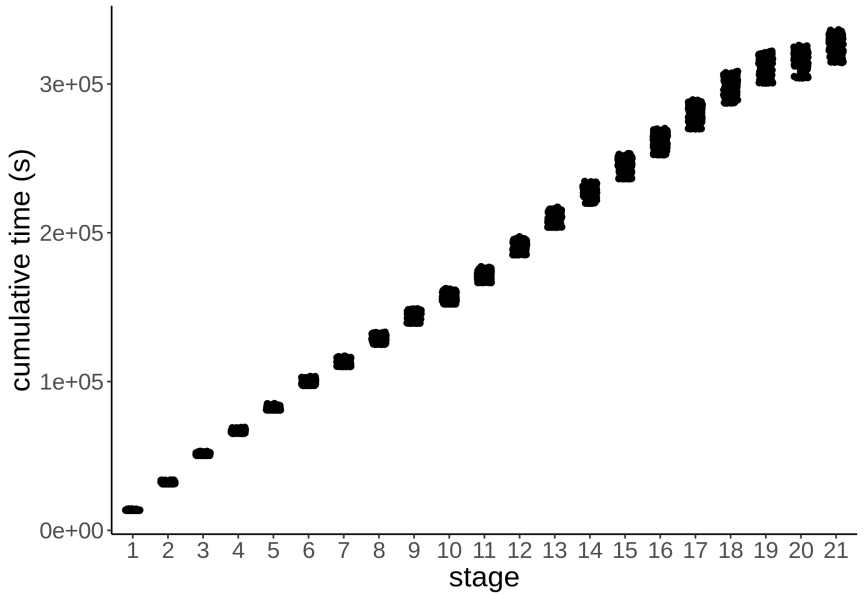

Fig 1: Scatterplot of cumulative time by stage. Each point corresponds to a rider's cumulative time at the corresponding stage.

time. The model proposed in Section 4 addresses these issues.

Figures 3 and 4 suggest that time differences vary between the top contenders[4] of the Tour, and slightly between the teams as well. The model proposed in Section 4 takes these two variables into account.

4. Statistical methodologies. In this section, a zero-inflated probability model is proposed to study the time differences of riders. The model is then modified to account for combinatorial restrictions. Two response functions are proposed to fit the model to the data.

4.1. *Tweedie distribution.* There are two issues when dealing with the time differences. First, there is a non-zero probability of the time difference being zero. Indeed, at each stage there will be at least one observation equal to zero—namely, the leader of the race. Second, there *has* to be at least one observation equal to zero at each stage.

To deal with the first issue, we propose modeling the time differences with

[4]These were determined by looking up the team leaders and favorites prior to the race.

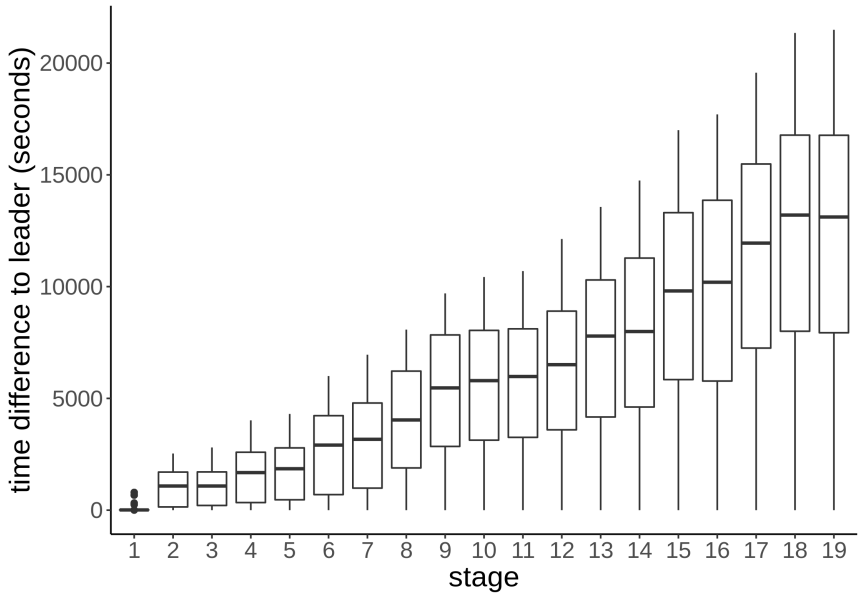

Fig 2: Boxplots of time difference to leader for every stage. The variability in the time difference increases as stages progress.

a *compound Poisson-Gamma* model (Tweedie, 1984; Jørgensen, 1987), which is defined as follows. Let $N \sim \mathsf{Poiss}(\lambda)$ and define

$$Y \mid N = \begin{cases} 0, & N = 0, \\ \gamma_{N,b}, & N > 0, \end{cases}$$

where $\gamma_{N,b} \sim \mathsf{Gamma}(N, b)$. The marginal distribution of Y is a compound Poisson-Gamma distribution with parameters λ and b, denoted $Y \sim \mathsf{CPG}(\lambda, b)$. Observe that Y has a non-zero probability of being zero. Specifically,

$$P(Y = 0) = P(N = 0) = \exp\{-\lambda\}.$$

When $N > 0$, $\gamma_{N,b}$ can be thought of as adding N i.i.d. $\mathsf{Exp}(b)$ distributions. More generally, one can add N i.i.d. $\mathsf{Gamma}(a, b)$ distributions (see Withers and Nadarajah, 2011). For the purposes of this report, however, we assume the simpler case of adding exponential distributions. This has the benefit of having one less parameter to fit without losing much flexibility. In that case, it is easy to prove that

$$\mathbb{E}[Y] = \frac{\lambda}{b}.$$

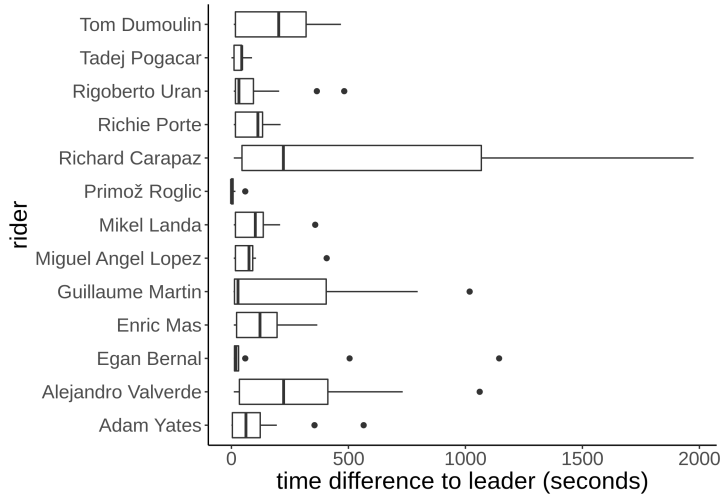

Fig 3: Boxplots of time difference to leader for every top contender. Contenders seem to vary between them.

The marginal density of Y can be calculated in closed form.

PROPOSITION 1. *Let* $Y \sim \mathsf{CPG}(\lambda, b)$. *Then*

$$p(y;\lambda,b) = \exp\{-\lambda\}\delta(y) + b\lambda\exp\{-\lambda - by\}\sum_{n=0}^{\infty}\frac{(by\lambda)^n}{(n+1)!n!}. \tag{1}$$

where $\delta(y)$ *is the Dirac* δ*-function.*

The proof follows from the law of total probability. (See Öztürk, 1981, for example). The sum in Eqn. 1 converges quickly due to the denominator growing much faster than the numerator (Withers and Nadarajah, 2011).

What makes the compound Poisson-Gamma attractive from a regression perspective is that it is an *exponential dispersion model*. A random variable Y is said to be an exponential dispersion model if its density can be expressed as

$$p(y;\theta,\sigma^2) = \kappa(\sigma^2, y)\exp\left\{\frac{\theta y - A(\theta)}{\sigma^2}\right\}.$$

The mean of an exponential dispersion model is $\mu = A'(\theta)$, and we denote $Y \sim \mathsf{ED}(\mu, \sigma^2)$. Furthermore, if $Y \sim \mathsf{ED}(\mu, \sigma^2)$ then $\mathrm{Var}(Y) = \sigma^2 V(\mu)$,

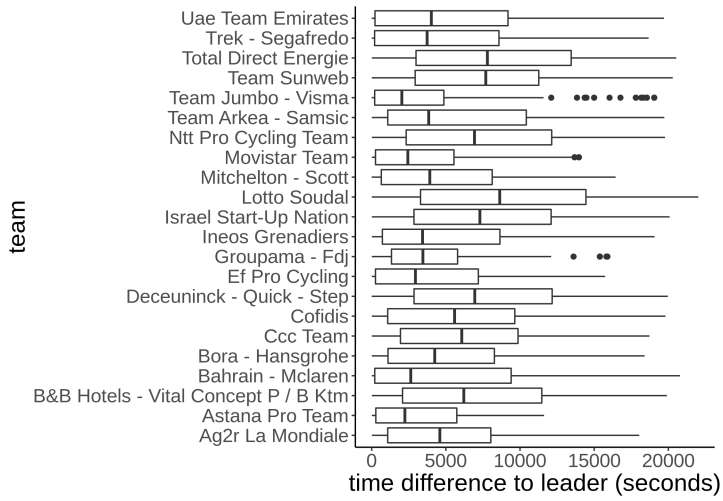

Fig 4: Boxplots of time difference to leader for every team. Contenders seem to vary more between them than teams.

where $V(\mu) = A''((A')^{-1}(\mu))$ is the variance function of Y. A special case of exponential dispersion models is the *Tweedie family of distributions*. A random variable is said to follow a Tweedie distribution with power parameter p if $Y \sim \mathsf{ED}(\mu, \sigma^2)$ and $V(\mu) = \mu^p$ for some $p \in \mathbb{R}$, in which case we denote $Y \sim \mathsf{Tweedie}(\mu, \sigma^2, p)$. Compound Gamma-Poisson models are Tweedie distributions.

PROPOSITION 2. *If* $Y \sim \mathsf{CPG}(\lambda, b)$ *then* $Y \sim \mathsf{Tweedie}(\mu, \sigma^2, p = 1.5)$ *with*

$$\mu = \frac{\lambda}{b},$$
$$\sigma^2 = \frac{2}{\sqrt{\lambda b}}.$$

A proof can be found in (Jørgensen and Paes De Souza, 1994) by substituting $\alpha = -1$ in their analysis. Smyth (1996) shows that, for the case of a $\mathsf{Tweedie}(\mu, \sigma^2, p = 1.5)$ distribution,

$$\theta = -2\mu^{-1/2}, \quad A(\theta) = -4\theta^{-1}. \tag{2}$$

In summary, by modeling the time difference of riders to the race leader with a compound Poisson-Gamma model, we actually end up with a Tweedie distribution.

4.2. *Random component.* Maximum-likelihood estimates (MLE) for the parameters of a Tweedie distribution can be approximated realiably (Öztürk, 1981; Withers and Nadarajah, 2011). This, combined with its close relationship to exponential families, means that it is amenable to be used in generalized linear models.

In our case, let Y_n^k denote the time difference of rider n to the race leader at stage k, $n = 1, ..., N$, $k = 1, ...K$. We will assume that $Y_n^k \overset{\text{i.i.d.}}{\sim} \mathsf{Tweedie}(\mu_n^k, \sigma^2, p = 1.5)$. For presentation purposes, we assume there are the same number of riders by stage to avoid using N_k. This has no impact on the analysis.

We also assume independence both between riders and within stages for a given rider. Between riders, each rider is assumed to try to race in as little time as possible regardless of other riders. Within riders, it would be intuitive to assume that a given rider's time difference in a stage has a lot of influence on the next stage. This is not necessarily so, however: it is not uncommon for riders to launch attacks and reduce their time difference considerably, or alternatively to break down in a climb and lose plenty of time.[5] Finally, it is also possible for two riders to have the same time difference, and even for two riders to have a time difference of zero. Specifically, final sprints at some stages can result in riders having the same time. Although it is unlikely that two riders will be tied as leaders, it is not technically impossible.[6]

The likelihood of the data under this setting can be written down as

$$p(y_1^1, ..., y_N^K; \mu, \sigma^2) = \prod_{k=1}^{K} \prod_{n=1}^{N} p(y_n^k; \mu_n^k, \sigma^2),$$

where $\mu = (\mu_1^1, ..., \mu_N^K)^\top$ and $p(y_n^k; \mu_n^k, \sigma^2)$ is the Tweedie likelihood. However, as we mentioned before, there is an additional combinatorial restriction: at each stage, there has to be at least one rider with a time difference of zero. We overcome this by adding a multiplicative factor to the likelihood that ensures this restriction is satisfied:

$$p(y_1^1, ..., y_N^K; \mu, \sigma^2) = \prod_{k=1}^{K} \left[\prod_{n=1}^{N} p(y_n^k; \mu_n^k, \sigma^2) \right] \mathbb{1}_{\{0\}}(y_{(N)}^k),$$

[5]This happended to the defending champion, Egan Bernal, while ascending the Grand Colombier in stage 15. He lost 7 minutes on his 59 seconds gap and subsequently abandoned the race.

[6]In 1989, Greg LeMond won the Tour with a time difference of a mere 8 seconds.

where $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 otherwise and $y^k_{(N)} = \min\{y^k_1, ..., y^k_N\}$. The additional factor is 1 when each stage has an observation that is zero and 0 otherwise. The log-likelihood is more illuminating:

$$\log p(y^1_1, ..., y^K_N;\, \mu, \sigma^2) = \sum_{k=1}^{K} \sum_{n=1}^{N} \log p(y^k_n;\, \mu^k_n, \sigma^2) + \sum_{k=1}^{K} \log \mathbb{1}_{\{0\}}(y^k_{(N)}). \tag{3}$$

The first term in Eqn. 3 is the log-likelihood of i.i.d. Tweedie observations. The second term can be thought of as a penalty: it takes the value of 0 when each stage has an observation that is zero, and the value of $-\infty$ otherwise. In the case a stage does not have an observation that is zero, the MLE is not defined. Else, the MLE is that of the Tweedie distribution.

Note that our data satisfy the combinatorial constraint by construction. Indeed, that is the reason we considered it in the first place. Hence, the log-likelihood of the model parameters—which will be introduced briefly—given our data is not going to have the additional penaly term, which is going to equal zero. This means that we can do regular MLE without worrying about the combinatorial constraint.

Finally, it should be noted that the variance of $Y \sim \mathsf{Tweedie}(\mu, \sigma^2, p = 1.5)$ can be written as

$$\mathrm{Var}(Y) = \sigma^2 \mu^{1.5}.$$

In our setting, this will also account for the non-constant variance.

4.3. *Systematic component.* Based on the exploratory data analysis of Section 3, we will use stage, distance, team, and rider as predictors. The last two variables are categorical, and so it is necessary to create indicator variables for each category (minus one reference category). Because there are around 150 riders and 22 teams, this would entail having $150 + 22 = 172$ covariates. Even though the training data consist of over 3,000 observations, we consider the risk of overfitting to be too large. On the other hand, however, it is necessary to produce predictions at the rider level to be able to answer the main research question.

To remedy this, we instead include indicator variables only for the 13 top contenders of the Tour, and aggregate the other riders into an "Other" category. Not only does this strike a balance between having predictions at the rider level and not overfitting, but it is also justified in the context of

the Tour: not all riders in the Tour are actively trying to win the general classification. Some are aiming to win some of the other races in the Tour (points and mountains), while some are there only to support the team leader.

The linear predictor will thus be

$$\eta_n^k = \beta_0 + \beta^\top x_n^k, \quad \beta = (\beta_1, ..., \beta_{36})^\top,$$

where β_1 corresponds to the stage number; β_2 corresponds to the stage distance; β_3 through β_{23} correspond to the indicator variables of the different teams; and β_{24} through β_{36} correspond to the indicator variables of the different contenders, or favorite riders.

4.4. *Response function.* Motivated by Figure 2, we would like to equate the linear predictor with the mean of the response variable μ_n^k to leverage the linear relationship between the stage number and the time differences. However, μ_n^k has to be non-negative, and the linear predictor need not be positive. Indeed, its sign depends on the values of β. Hence, using an identity response function is not immediatly possible. We leave exploring ways to enforce a linear relationship between stage number and time differences for future work.

One alternative is to use a log response function,

$$\log \mu_n^k = \eta_n^k, \quad n = 1, ..., N,\ k = 1, ..., K. \tag{4}$$

Under the log response function, the covariates have a multiplicative effect on the mean of the response, which makes the model easier to interpret. Another alternative is to use the (scaled) canonical response which, following Eqn. 2, for this case is given by

$$(\mu_n^k)^{-1/2} = \eta_n^k, \quad n = 1, ..., N,\ k = 1, ..., K. \tag{5}$$

Under the canonical response, the log-likelihood is concave and thus iterative optimization methods have convergence guarantees.

5. Results. We fit two models to the data. Both model the time difference as a Tweedie distribution, but differ in the response function used. Specifically, one model uses the log response function (4) and the other model uses the canonical response function (5). The code to generate all figures and fit the models can be found at https://github.com/GiankDiluvi/tdf-2020.

Figure 5 shows a comparison of the true time differences against the fitted and predicted values of both models. Regardless of the response function,

the compound Gamma-Poisson models capture the increase in variability as a function of stage. However, variability is underestimated at earlier stages and overestimated at later stages, most likely as a result of not using the identity link. Both models tend to estimate the time difference as either too large or too small. This is probably due to the contender dummy variables: the model can make more accurate predictions for the contenders, who will usually have smaller time differences. The other riders are differentiated only by team, and so their predictions are much larger and closer between each other. One potential way to remedy this would be to further split the "Others" group into two or more categories reflecting the performance of the riders (e.g. high, medium, and low performers). In any case, predicted values seem to be well within the expected range.

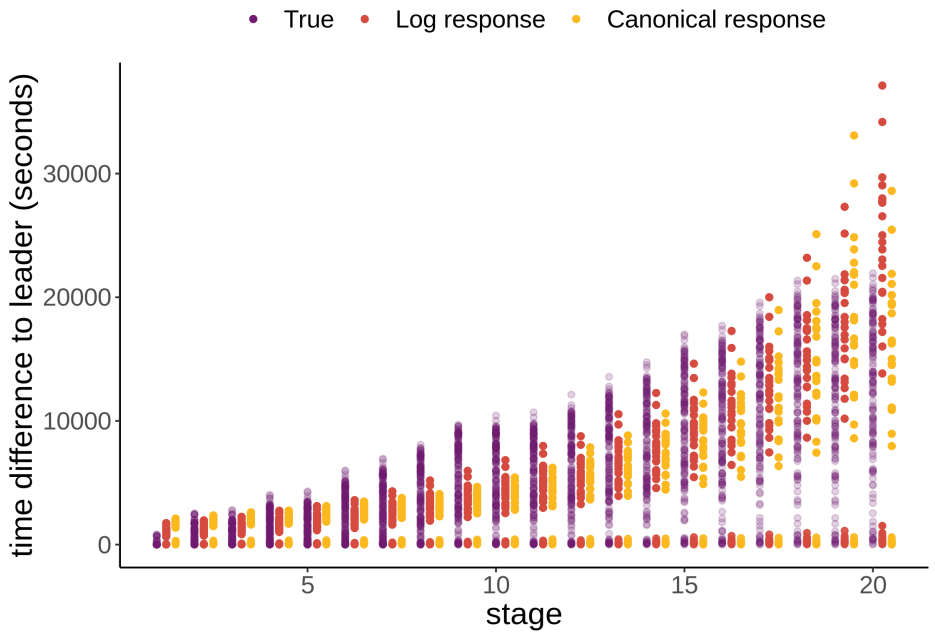

Fig 5: Comparison of fitted (stages 1 to 19) and predicted (stage 20) values of both models, with color by response function, against the true values. The compound Poisson-Gamma model successfully captures the increasing variability of observations.

To answer the research question, Table 2 shows both model's predictions for the top 10 riders. Neither model was able to predict Tadej Pogačar as the Tour's winner, although both placed him in second place. However, both

models assigned Pogačar a large probability of winning (i.e. of his time difference being zero, which can be calculated from the estimated parameters). Regarding the probabilities of winning, the model with the canonical response tends to assign larger probabilities in general, which means that the underlying Tweedie distribution has more mass concentrated on zero. With the exception of 10th place, both models seem to have recovered the top contenders, albeit not in exact order.

Rider	**True ranking**	**Log response function**		**Canonical response function**	
		Ranking	**Probability of winning**	**Ranking**	**Probability of winning**
Tadej Pogačar	1	2	0.228	2	0.545
Primož Roglič	2	1	0.552	1	0.841
Richie Porte	3	7	0.111	6	0.373
Mikel Landa	4	6	0.112	7	0.373
Enric Mas	5	8	0.095	8	0.334
Miguel A. López	6	3	0.182	3	0.475
Tom Dumoulin	7	10	0.052	9	0.241
Rigoberto Uran	8	4	0.159	4	0.409
Adam Yates	9	5	0.152	5	0.394
Damiano Caruso	10	67	<0.0001	67	<0.0001

TABLE 2

Top 10 riders of the 2020 Tour de France. For each rider, its true ranking and predicted ranking are shown, along with the predicted probability of winning for both models. The model that uses the canonical response function tends to assign larger probabilities to the time difference being zero.

Figure 6 shows the normalized residuals for both models. The vast majority of normalized residuals lie within 3 units from the origin. However, some normalized residuals for small fitted values are below the -3 bound for both models. The figure also suggests that the models overestimate when fitted values are large and underestimate when they are small, which is in line with Figure 5. Again, this is likely due to the response functions not being the identity. The coefficients of the model, along with the corresponding P-values, are shown in the Appendix.

Both models generally agree on their results. Even though neither managed to predict Pogačar's victory, they both assigned non-trivial probabilities to his victory. Specifically, the model with the log response function assigns Pogačar a 23% probability of winning the Tour.

6. Conclusion. We proposed a zero-inflated generalized linear model based on the Tweedie distribution to model the time difference of each rider

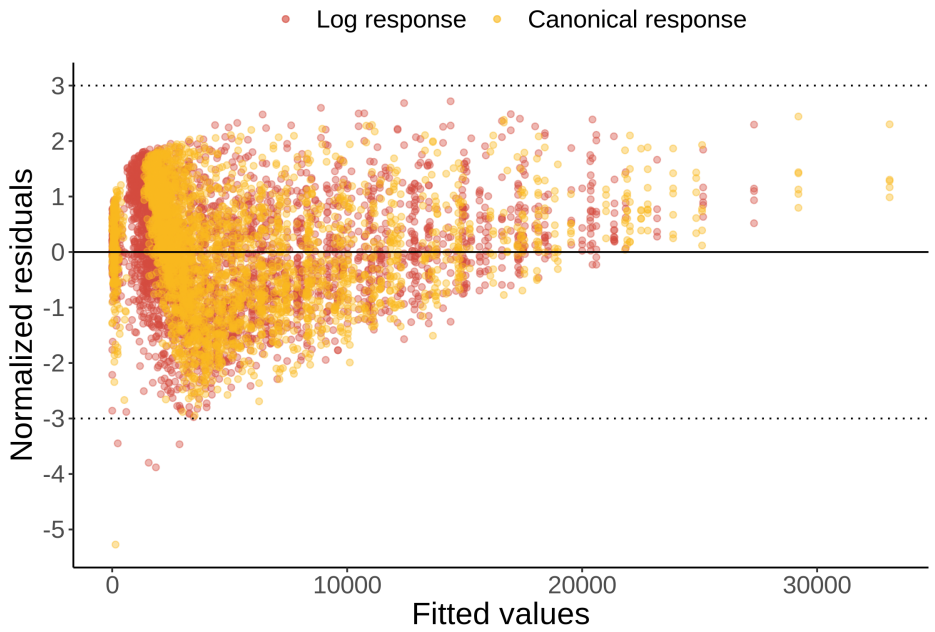

Fig 6: Normalized residuals of the compound Poisson-Gamma models, with color by response function. The vast majority of values are well within the $\pm$ 3 bound. The models tend to overestimate when fitted values are small and underestimate when they are large.

of the 2020 Tour de France to the Tour's leader at each stage. We suggested two possible response functions—logarithm and canonical—to associate the linear predictor—which included the stage number, distance, rider's team, and favorite riders—with the mean of the time difference, which led to two models. After fitting these two models, we found that both predicted favorite Primož Roglič to be the winner of the Tour. However, they both also assigned relatively large probabilities of winning to Tadej Pogačar, the actual winner of the Tour. In that sense, Pogačar's victory should not have been as much a surprise as it was.

One limitation of the model is that it cannot capture the linear relationship between the time difference and the stage number because the identity link would result in values of the linear predictor that are not valid for the response mean. The impact of this limitation was discussed in previous sections, and boils down to an underestimation of the variability in the response for early stages, and an overestimation for later ones. One possible way to

overcome this would be to restrict the values of the parameter associated with stage number to be non-negative, which would allow that covariate to linearly influence the response. Given the observed positive association between stage number and time difference, this might be reasonable. The rest of the parameters can be scaled using another response function—which would result in a non-linear model—or alternatively they can be similarly restricted. Whether optimization methods are guaranteed to converge in this setting is also left for future work. Another alternative would be to use a Bayesian approach and directly incorporate the parameter restrictions into their prior distributions. Unfortunately, popular probabilistic programming languages such as Stan do not have a built-in Tweedie regression family, which would require the user to develop code to approximate the density.

Another limitation is that the model assumes independence between different stages of the same rider. This can be partly justified by the fact that riders can almost arbitrarily lose or win time difference at any given stage. However, it would be interesting to properly account for this longitudinal structure. Whether the gains in accuracy, if any, offset the increase in complexity is also left for future work.

Finally, a variable that was not included in the model and that might have had a significant impact is the type of stage. Some riders specialize in certain types of stages (such as mountain stages), and this can have a significant effect on their performance. However, the 20th stage in this year's Tour de France was the only time trial[7] of the whole Tour. Hence, there were no data on which to train the model for this stage type. Future research can include results from other Tours in the training data set, although how to account for missing and new riders is not clear. (Specifically, this year's Tour was Pogačar's first.)

[7]Where riders race individually against the clock, usually in shorter routes that allow them to output more power throughout the stage.

References.

Haakonsson, S., Rodríguez, M. A., Carballo, C., del Carmen Pérez, M., Arocena, R. and Bonilla, S. (2020). Predicting cyanobacterial biovolume from water temperature and conductivity using a Bayesian compound Poisson-Gamma model. *Water Research* 115710.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)* **49** 127–145.

Jørgensen, B. and Paes De Souza, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal* **1994** 69–93.

Lauderdale, B. E. (2012). Compound Poisson—Gamma Regression Models for Dollar Outcomes That Are Sometimes Zero. *Political Analysis* 387–399.

Öztürk, A. (1981). On the study of a probability distribution for precipitation totals. *Journal of Applied Meteorology* **20** 1499–1505.

Smyth, G. K. (1996). Regression modelling of quantity data with exact zeroes. *Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management* 572-580.

R Core Team (2020). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Tweedie, M. C. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference* **579** 579–604.

Withers, C. and Nadarajah, S. (2011). On the compound Poisson-gamma distribution. *Kybernetika* **47** 15–37.

APPENDIX A: COEFFICIENTS AND P-VALUES

Figures 7 and 8 show the coefficient estimates and P-values for both models. In both models stage and distance are significant at the 0.05 level, some teams are also significant, and most contenders are not. Both models were fit using R 4.0.3 (R Core Team, 2020) in an ASUS ROG GL552VX computer with 64-bit Linux Mint 19.3, 8 GB of RAM memory, and an Intel Core i7-6700HQ CPU at 2.60GHz.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.4379300	0.3661776	6.658	3.28e-11
stage	0.1537548	0.0017909	85.852	< 2e-16
distance	-0.0011738	0.0005539	-2.119	0.034153
teamAstana Pro Team	-0.3864309	0.0687834	-5.618	2.10e-08
teamB&B Hotels - Vital Concept P / B Ktm	0.3102654	0.0599302	5.177	2.40e-07
teamBahrain - Mclaren	0.1238670	0.0650762	1.903	0.057080
teamBora - Hansgrohe	0.0035822	0.0630280	0.057	0.954680
teamCcc Team	0.1586706	0.0615669	2.577	0.010007
teamCofidis	0.3055901	0.0635110	4.812	1.57e-06
teamDeceuninck - Quick - Step	0.3182378	0.0594423	5.354	9.25e-08
teamEf Pro Cycling	-0.1691317	0.0660756	-2.560	0.010525
teamGroupama - Fdj	-0.2399547	0.0666645	-3.599	0.000324
teamIneos Grenadiers	0.2056217	0.0647272	3.177	0.001504
teamIsrael Start-Up Nation	0.3550768	0.0595031	5.967	2.69e-09
teamLotto Soudal	0.6002505	0.0633244	9.479	< 2e-16
teamMitchelton - Scott	0.1831051	0.0645441	2.837	0.004585
teamMovistar Team	-0.1106192	0.0672371	-1.645	0.100029
teamNtt Pro Cycling Team	0.3771788	0.0620467	6.079	1.36e-09
teamTeam Arkea - Samsic	0.1013535	0.0623668	1.625	0.104240
teamTeam Jumbo - Visma	-0.1332511	0.0674485	-1.976	0.048290
teamTeam Sunweb	0.3065779	0.0595168	5.151	2.75e-07
teamTotal Direct Energie	0.5176496	0.0601176	8.611	< 2e-16
teamTrek - Segafredo	0.0570666	0.0644631	0.885	0.376086
teamUae Team Emirates	0.2651569	0.0650856	4.074	4.74e-05
contenderAlejandro Valverde	1.4829139	0.4368429	3.395	0.000696
contenderEgan Bernal	0.5562726	0.4971861	1.119	0.263295
contenderEnric Mas	0.7443833	0.4694709	1.586	0.112938
contenderGuillaume Martin	0.6257442	0.4546145	1.376	0.168790
contenderMiguel Angel Lopez	0.3691141	0.5074108	0.727	0.467008
contenderMikel Landa	0.3580139	0.4771518	0.750	0.453123
contenderOther	4.4511209	0.3494673	12.737	< 2e-16
contenderPrimož Roglic	-1.9894197	0.7104141	-2.800	0.005136
contenderRichard Carapaz	1.6488717	0.4199015	3.927	8.80e-05
contenderRichie Porte	0.4383972	0.4763351	0.920	0.357460
contenderRigoberto Uran	0.3032488	0.4974214	0.610	0.542144
contenderTadej Pogacar	-0.5654485	0.5264305	-1.074	0.282854
contenderTom Dumoulin	1.2150522	0.4485637	2.709	0.006791

Fig 7: Coefficient estimates, standard error, t-statistics, and P-values for the model with the log response function.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.201e-01	1.881e-02	6.385	1.97e-10
stage	-8.995e-04	1.390e-05	-64.701	< 2e-16
distance	-1.009e-05	4.311e-06	-2.340	0.019336
teamAstana Pro Team	2.488e-03	5.365e-04	4.638	3.67e-06
teamB&B Hotels - Vital Concept P / B Ktm	-1.560e-03	3.902e-04	-3.999	6.51e-05
teamBahrain - Mclaren	-5.511e-04	4.360e-04	-1.264	0.206262
teamBora - Hansgrohe	-8.641e-05	4.346e-04	-0.199	0.842411
teamCcc Team	-8.705e-04	4.101e-04	-2.123	0.033855
teamCofidis	-1.528e-03	4.073e-04	-3.751	0.000179
teamDeceuninck - Quick - Step	-1.673e-03	3.831e-04	-4.367	1.30e-05
teamEf Pro Cycling	8.736e-04	4.748e-04	1.840	0.065836
teamGroupama - Fdj	1.853e-03	5.148e-04	3.600	0.000323
teamIneos Grenadiers	-9.272e-04	4.285e-04	-2.164	0.030561
teamIsrael Start-Up Nation	-1.825e-03	3.832e-04	-4.762	2.01e-06
teamLotto Soudal	-2.799e-03	3.855e-04	-7.262	4.81e-13
teamMitchelton - Scott	-8.748e-04	4.332e-04	-2.019	0.043537
teamMovistar Team	7.854e-04	4.799e-04	1.637	0.101800
teamNtt Pro Cycling Team	-1.954e-03	3.971e-04	-4.920	9.11e-07
teamTeam Arkea - Samsic	-4.441e-04	4.226e-04	-1.051	0.293368
teamTeam Jumbo - Visma	8.335e-04	4.815e-04	1.731	0.083541
teamTeam Sunweb	-1.542e-03	3.859e-04	-3.997	6.58e-05
teamTotal Direct Energie	-2.445e-03	3.797e-04	-6.441	1.37e-10
teamTrek - Segafredo	-4.145e-04	4.364e-04	-0.950	0.342267
teamUae Team Emirates	-1.399e-03	4.237e-04	-3.302	0.000971
contenderAlejandro Valverde	-4.705e-02	2.052e-02	-2.293	0.021926
contenderEgan Bernal	-1.924e-02	2.443e-02	-0.788	0.431048
contenderEnric Mas	-1.680e-02	2.403e-02	-0.699	0.484643
contenderGuillaume Martin	-3.500e-02	2.141e-02	-1.634	0.102303
contenderMiguel Angel Lopez	2.198e-02	3.192e-02	0.688	0.491255
contenderMikel Landa	-5.957e-03	2.557e-02	-0.233	0.815817
contenderOther	-9.307e-02	1.880e-02	-4.952	7.75e-07
contenderPrimož Roglic	4.390e-01	2.161e-01	2.031	0.042311
contenderRichard Carapaz	-5.988e-02	1.958e-02	-3.058	0.002248
contenderRichie Porte	-5.923e-03	2.560e-02	-0.231	0.817035
contenderRigoberto Uran	2.508e-03	2.739e-02	0.092	0.927031
contenderTadej Pogacar	5.469e-02	3.939e-02	1.388	0.165098
contenderTom Dumoulin	-3.660e-02	2.150e-02	-1.702	0.088776

Fig 8: Coefficient estimates, standard error, t-statistics, and P-values for the model with the canonical response function.

Department of Statistics
University of British Columbia
Vancouver, BC, Canada V6T 1Z4
E-mail: gian.diluvi@stat.ubc.ca