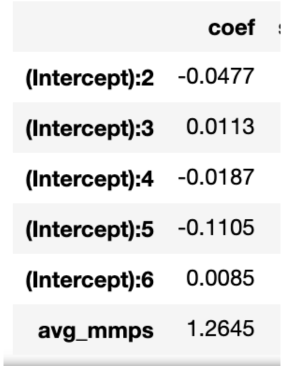
1. (Baseline Model) In this question we build a basic model for predicting race winners. Construct a feature avg mmps containing the average value of mmps from all strictly prior races for that dog. The definition of mmps is described in Live Lecture 11. You can directly use the parameters defining mmps from the lecture. Fit a conditional multinomial logit model of the form

$$\texttt{twinner} \sim \texttt{avg mmps}.$$

Here twinner is equal to winner for races with a unique winner, and is a randomly chosen winner in the other cases (see the conditional logit model notebook for code that creates the twinner column, and for the mlogit function that fits a conditional multinomial logit model).

   (a) Fit the above model on races between July 1st, 2019 and January 31st, 2020, and report your coefficients.



	coef
(Intercept):2	-0.0477
(Intercept):3	0.0113
(Intercept):4	-0.0187
(Intercept):5	-0.1105
(Intercept):6	0.0085
avg_mmps	1.2645

   (b) Report your out-of-sample Brier score using the races on and after February 1st, 2020. This is computed using all of your forecasted probabilities, the twinner column, and the Brier score loss function from sklearn.

```
Brier Score:  0.14007
```

   (c) Submit your results to this problem (i.e., 1 only) in a single PDF on gradescope (listed under HW 8 Check-in - 1). You will also resubmit your solutions to this problem when you submit the full miniproject (listed under HW 8).

2. (Building a Speed Model) In this question we will build a linear model to better predict dog speeds in upcoming races. The outputs of this model can then be used as a feature in our multinomial logit models.

   (a) Fit a linear model of the form

   $$\texttt{mmps} \sim \texttt{mmps\_ema}$$

   where mmps is the modified speed computed for a given dog in the current race, and mmps ema is an exponentially weighted moving average of mmps using data for that dog from strictly prior races.

      i. Fit the above model on races between July 1st, 2019 and November 30th, 2019, and report your coefficients.

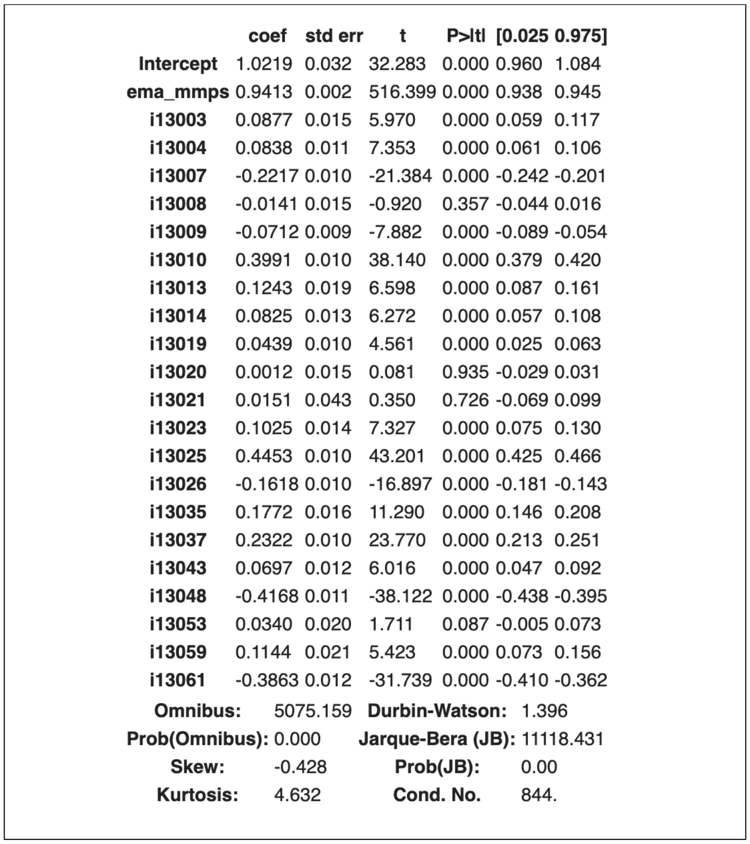
      ```
      Intercept:  0.8223
      ema_mmps:  0.9530
      ```

      ii. Report your out-of-sample average square loss (for predicting mmps) using the races between December 1st, 2019, and January 31st, 2020, inclusive.

      ```
      MSE: 0.06873
      ```

   (b) Improve your mmps prediction model in the preceding part by also incorporating the stadium id.

      i. Fit the above model on races between July 1st, 2019 and November 30th, 2019, and report your coefficients.



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.0219	0.032	32.283	0.000	0.960	1.084
ema_mmps	0.9413	0.002	516.399	0.000	0.938	0.945
i13003	0.0877	0.015	5.970	0.000	0.059	0.117
i13004	0.0838	0.011	7.353	0.000	0.061	0.106
i13007	-0.2217	0.010	-21.384	0.000	-0.242	-0.201
i13008	-0.0141	0.015	-0.920	0.357	-0.044	0.016
i13009	-0.0712	0.009	-7.882	0.000	-0.089	-0.054
i13010	0.3991	0.010	38.140	0.000	0.379	0.420
i13013	0.1243	0.019	6.598	0.000	0.087	0.161
i13014	0.0825	0.013	6.272	0.000	0.057	0.108
i13019	0.0439	0.010	4.561	0.000	0.025	0.063
i13020	0.0012	0.015	0.081	0.935	-0.029	0.031
i13021	0.0151	0.043	0.350	0.726	-0.069	0.099
i13023	0.1025	0.014	7.327	0.000	0.075	0.130
i13025	0.4453	0.010	43.201	0.000	0.425	0.466
i13026	-0.1618	0.010	-16.897	0.000	-0.181	-0.143
i13035	0.1772	0.016	11.290	0.000	0.146	0.208
i13037	0.2322	0.010	23.770	0.000	0.213	0.251
i13043	0.0697	0.012	6.016	0.000	0.047	0.092
i13048	-0.4168	0.011	-38.122	0.000	-0.438	-0.395
i13053	0.0340	0.020	1.711	0.087	-0.005	0.073
i13059	0.1144	0.021	5.423	0.000	0.073	0.156
i13061	-0.3863	0.012	-31.739	0.000	-0.410	-0.362

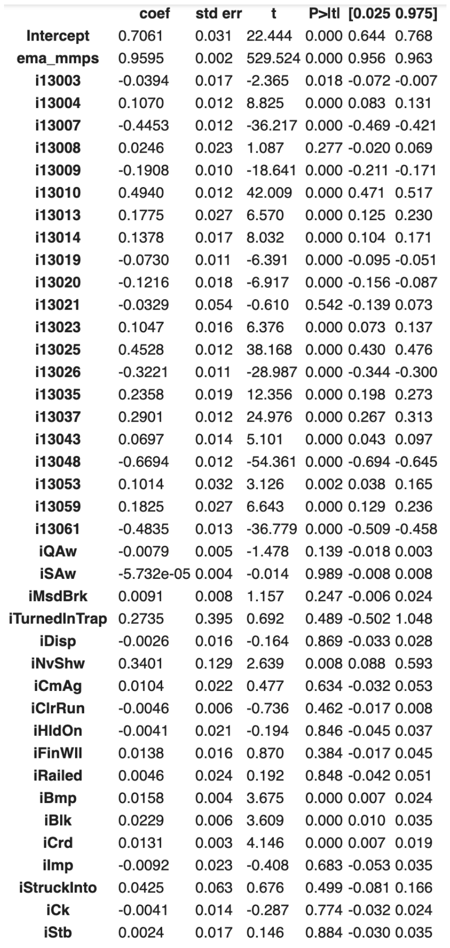
Omnibus:	5075.159	**Durbin-Watson:**	1.396
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11118.431
Skew:	-0.428	**Prob(JB):**	0.00
Kurtosis:	4.632	**Cond. No.**	844.

ii. Report your out-of-sample average square loss (for predicting mmps) using the races between December 1st, 2019, and January 31st, 2020, inclusive.

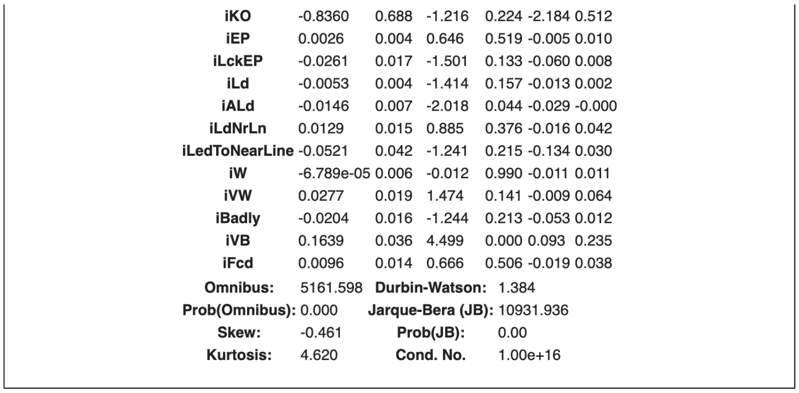
```
MSE: 0.06479
```

3. (Incorporating Comments) The comment column of our data includes useful information about what events happened to each dog during the course of the race. In this question we will incorporate the comment information into the mmps prediction model we built in the previous part.

   (a) Fit the above model on races between July 1st, 2019 and November 30th, 2019, and report your coefficients.



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.7061	0.031	22.444	0.000	0.644	0.768
ema_mmps	0.9595	0.002	529.524	0.000	0.956	0.963
i13003	-0.0394	0.017	-2.365	0.018	-0.072	-0.007
i13004	0.1070	0.012	8.825	0.000	0.083	0.131
i13007	-0.4453	0.012	-36.217	0.000	-0.469	-0.421
i13008	0.0246	0.023	1.087	0.277	-0.020	0.069
i13009	-0.1908	0.010	-18.641	0.000	-0.211	-0.171
i13010	0.4940	0.012	42.009	0.000	0.471	0.517
i13013	0.1775	0.027	6.570	0.000	0.125	0.230
i13014	0.1378	0.017	8.032	0.000	0.104	0.171
i13019	-0.0730	0.011	-6.391	0.000	-0.095	-0.051
i13020	-0.1216	0.018	-6.917	0.000	-0.156	-0.087
i13021	-0.0329	0.054	-0.610	0.542	-0.139	0.073
i13023	0.1047	0.016	6.376	0.000	0.073	0.137
i13025	0.4528	0.012	38.168	0.000	0.430	0.476
i13026	-0.3221	0.011	-28.987	0.000	-0.344	-0.300
i13035	0.2358	0.019	12.356	0.000	0.198	0.273
i13037	0.2901	0.012	24.976	0.000	0.267	0.313
i13043	0.0697	0.014	5.101	0.000	0.043	0.097
i13048	-0.6694	0.012	-54.361	0.000	-0.694	-0.645
i13053	0.1014	0.032	3.126	0.002	0.038	0.165
i13059	0.1825	0.027	6.643	0.000	0.129	0.236
i13061	-0.4835	0.013	-36.779	0.000	-0.509	-0.458
iQAw	-0.0079	0.005	-1.478	0.139	-0.018	0.003
iSAw	-5.732e-05	0.004	-0.014	0.989	-0.008	0.008
iMsdBrk	0.0091	0.008	1.157	0.247	-0.006	0.024
iTurnedInTrap	0.2735	0.395	0.692	0.489	-0.502	1.048
iDisp	-0.0026	0.016	-0.164	0.869	-0.033	0.028
iNvShw	0.3401	0.129	2.639	0.008	0.088	0.593
iCmAg	0.0104	0.022	0.477	0.634	-0.032	0.053
iClrRun	-0.0046	0.006	-0.736	0.462	-0.017	0.008
iHldOn	-0.0041	0.021	-0.194	0.846	-0.045	0.037
iFinWll	0.0138	0.016	0.870	0.384	-0.017	0.045
iRailed	0.0046	0.024	0.192	0.848	-0.042	0.051
iBmp	0.0158	0.004	3.675	0.000	0.007	0.024
iBlk	0.0229	0.006	3.609	0.000	0.010	0.035
iCrd	0.0131	0.003	4.146	0.000	0.007	0.019
iImp	-0.0092	0.023	-0.408	0.683	-0.053	0.035
iStruckInto	0.0425	0.063	0.676	0.499	-0.081	0.166
iCk	-0.0041	0.014	-0.287	0.774	-0.032	0.024
iStb	0.0024	0.017	0.146	0.884	-0.030	0.035



iKO	-0.8360	0.688	-1.216	0.224	-2.184	0.512
iEP	0.0026	0.004	0.646	0.519	-0.005	0.010
iLckEP	-0.0261	0.017	-1.501	0.133	-0.060	0.008
iLd	-0.0053	0.004	-1.414	0.157	-0.013	0.002
iALd	-0.0146	0.007	-2.018	0.044	-0.029	-0.000
iLdNrLn	0.0129	0.015	0.885	0.376	-0.016	0.042
iLedToNearLine	-0.0521	0.042	-1.241	0.215	-0.134	0.030
iW	-6.789e-05	0.006	-0.012	0.990	-0.011	0.011
iVW	0.0277	0.019	1.474	0.141	-0.009	0.064
iBadly	-0.0204	0.016	-1.244	0.213	-0.053	0.012
iVB	0.1639	0.036	4.499	0.000	0.093	0.235
iFcd	0.0096	0.014	0.666	0.506	-0.019	0.038

Omnibus:	5161.598	**Durbin-Watson:**	1.384
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	10931.936
Skew:	-0.461	**Prob(JB):**	0.00
Kurtosis:	4.620	**Cond. No.**	1.00e+16

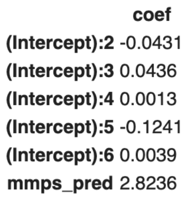
(b) Report your out-of-sample average square loss (for predicting mmps) using the races between December 1st, 2019, and January 31st, 2020, inclusive.

```
MSE: 0.06283
```

4. (Improving the Baseline) In this final problem, we improve on our baseline model from the first question.

   (a) Build an improved multinomial logit model for twinner by adding the forecasts of our mmps prediction model as a feature.

      i. Fit the above model on races strictly before February 1st, 2020, and report your coefficients.



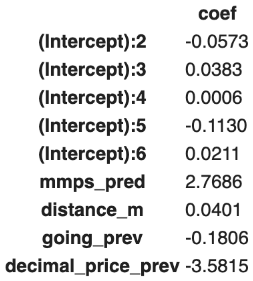
	coef
(Intercept):2	-0.0431
(Intercept):3	0.0436
(Intercept):4	0.0013
(Intercept):5	-0.1241
(Intercept):6	0.0039
mmps_pred	2.8236

      ii. Report your out-of-sample Brier score using the races on and after February 1st, 2020. This is computed using all of your forecasted probabilities, the twinner column, and the Brier score loss function from sklearn.

```
Brier Score:  0.13894
```

   (b) Improve your model from the previous part in some way. You can do this by improving your mmps prediction model, or by adding features to the multinomial logit model. Note that you can use stadium id, kg, distance m, race grade, and box from the current race, and going, decimal price from strictly prior races in your fits.

      i. Fit the above model on races strictly before February 1st, 2020, and report your coefficients.



	coef
(Intercept):2	-0.0573
(Intercept):3	0.0383
(Intercept):4	0.0006
(Intercept):5	-0.1130
(Intercept):6	0.0211
mmps_pred	2.7686
distance_m	0.0401
going_prev	-0.1806
decimal_price_prev	-3.5815

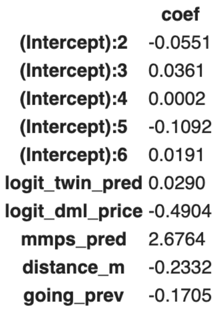
Factors used: `mmps_pred, distance_m, going_prev, decimal_price_prev`

      ii. Report your out-of-sample Brier score using the races on and after February 1st, 2020. This is computed using all of your forecasted probabilities, the twinner column, and the Brier score loss function from sklearn.

```
Brier Score:  0.13873
```

(c) Take your final model from the previous part, and fit a combined Benter-style model (i.e., use the logits of your forecast, and the logit of the market implied probabilities as the two features in a conditional multinomial logit model).

i. Fit the above model on races between July 1st, 2019 and January 31st, 2020, and report your coefficients.



	coef
(Intercept):2	-0.0551
(Intercept):3	0.0361
(Intercept):4	0.0002
(Intercept):5	-0.1092
(Intercept):6	0.0191
logit_twin_pred	0.0290
logit_dml_price	-0.4904
mmps_pred	2.6764
distance_m	-0.2332
going_prev	-0.1705

Factors used: `logit(twin_pred), logit(decimal_price_prev), mmps_pred, distance_m, going_prev`

ii. Report your out-of-sample Brier score using the races on and after February 1st, 2020. This is computed using all of your forecasted probabilities, the twinner column, and the Brier score loss function from sklearn.

```
Brier Score:  0.13798
```