```
* Dr. Steven S. Vickner;
* MSBA 635 - Data Analytics II;

* print data;
*xbeta=take your relevant data and multiply it by beta
*logit/probit regression example
*you take the car or the bus, it's the only choices
*auto is the dependent variable 1=auto, 0=bus. Dependent variable will be either 1 or 0.
*some kind of a choice model is a logit/probit
*dummy variables on the right hand side of the wine case study
*choice=1 if auto, otherwise choice=0
proc print data=tmp1.transport;
run;
```

The SAS System 17:14 Wednesday, January 16, 2019 1

Obs	autotime	bustime	dtime	auto
1	52.9	4.4	-4.85	0
2	4.1	28.5	2.44	0
3	4.1	86.9	8.28	1
4	56.2	31.6	-2.46	0
5	51.8	20.2	-3.16	0
6	0.2	91.2	9.10	1
7	27.6	79.7	5.21	1
8	89.9	2.2	-8.77	0
9	41.5	24.5	-1.70	0
10	95.0	43.5	-5.15	0
11	99.1	8.4	-9.07	0
12	18.5	84.0	6.55	1
13	82.0	38.0	-4.40	1
14	8.6	1.6	-0.70	0
15	22.5	74.1	5.16	1
16	51.4	83.8	3.24	1
17	81.0	19.2	-6.18	0
18	51.0	85.0	3.40	1
19	62.2	90.1	2.79	1
20	95.1	22.2	-7.29	0
21	41.6	91.5	4.99	1

```
* display data attributes;

proc contents data=tmp1.transport;
run;
```


The CONTENTS Procedure

Data Set Name	TMP1.TRANSPORT	Observations	21
Member Type	DATA	Variables	4
Engine	V9	Indexes	0
Created	06/19/2010 08:07:02	Observation Length	32
Last Modified	06/19/2010 08:07:02	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	4096
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	126
Obs in First Data Page	21
Number of Data Set Repairs	0
Filename	C:\Users\nxnguy01\Desktop\transport.sas7bdat
Release Created	9.0202M0
Host Created	XP_PRO

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
4	auto	Num	8	= 1 if auto chosen
1	autotime	Num	8	commute time via auto, minutes
2	bustime	Num	8	commute time via bus, minutes
3	dtime	Num	8	= (bus time - auto time)/10, 10 minute units

```
* produce frequencies;
*look at your response data here, 11 took the bus, 10 took automobile

proc freq data=tmp1.transport;
tables auto;
run;
```


The FREQ Procedure

= 1 if auto chosen

auto	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	11	52.38	11	52.38
1	10	47.62	21	100.00

```
* estimate linear probability model using proc reg;
*linear probability model "auto" is a bunch of 1s and 0s, you're not supposed to run
regression on this but you do it anyway. Not supposed to run if dependent variable a 1 or
0. Should use logit/probit
*LPM=Linear Probability Model
options nolabel;
proc reg data=tmp1.transport;
model auto = dtime;
output out=lpmout p=phat_lpm;
run;
quit;
```


The REG Procedure
Model: MODEL1
Dependent Variable: auto

Number of Observations Read	21
Number of Observations Used	21

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.20218	3.20218	29.88	<.0001
Error	19	2.03591	0.10715		
Corrected Total	20	5.23810			

Root MSE	0.32734	R-Square	0.6113
Dependent Mean	0.47619	Adj R-Sq	0.5909
Coeff Var	68.74200		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.48480	0.07145	6.79	<.0001
dtime	1	0.07031	0.01286	5.47	<.0001

```
* print data;
*some highlight probabilities are over 1, can't have that, there's also negative
probabilities. Can't have this, they must be between 0 and 1. That's why the linear
probability model here has been abandoned

proc print data=work.lpmout;
run;
```

The SAS System 17:14 Wednesday, January 16, 2019 5

Obs	autotime	bustime	dtime	auto	phat_lpm
1	52.9	4.4	-4.85	0	0.14379
2	4.1	28.5	2.44	0	0.65635
3	4.1	86.9	8.28	1	**1.06696**
4	56.2	31.6	-2.46	0	0.31183
5	51.8	20.2	-3.16	0	0.26262
6	0.2	91.2	9.10	1	**1.12462**
7	27.6	79.7	5.21	1	0.85111
8	89.9	2.2	-8.77	0	**-0.13182**
9	41.5	24.5	-1.70	0	0.36527
10	95.0	43.5	-5.15	0	0.12270
11	99.1	8.4	-9.07	0	**-0.15292**
12	18.5	84.0	6.55	1	0.94533
13	82.0	38.0	-4.40	1	0.17543
14	8.6	1.6	-0.70	0	0.43558
15	22.5	74.1	5.16	1	0.84759
16	51.4	83.8	3.24	1	0.71260
17	81.0	19.2	-6.18	0	0.05028
18	51.0	85.0	3.40	1	0.72385
19	62.2	90.1	2.79	1	0.68096
20	95.1	22.2	-7.29	0	**-0.02776**
21	41.6	91.5	4.99	1	0.83564

```
* estimate probit using proc qlim;
*now probit using 3 different procs will be run. Proc qlim, proc logistic, proc probit
*this one here is the probit model
*the QLIM procedure (qualitative limited procedure)
*non-linear model so no ANOVA or R-square.
*we don't have a normal variability in the model because the dependent variables are 0 or
1.
*xbeta=data*beta
options nolabel;
proc qlim data=tmp1.transport;
model auto = dtime / discrete;
output out=probitout xbeta marginal;
run;
quit;
```


The QLIM Procedure

Discrete Response Profile of auto

Index	Value	Total Frequency
1	0	11
2	1	10

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	auto
Number of Observations	21
Log Likelihood	-6.16516
Maximum Absolute Gradient	9.85945E-7
Number of Iterations	9
Optimization Method	Quasi-Newton
AIC	16.33032
Schwarz Criterion	18.41936

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	16.734	2 * (LogL - LogL0)
Upper Bound of R (U)	29.065	- 2 * LogL0
Aldrich-Nelson	0.4435	R / (R+N)
Cragg-Uhler 1	0.5493	1 - exp(-R/N)
Cragg-Uhler 2	0.7329	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.6948	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.5497	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.5758	R / U
Veall-Zimmermann	0.7639	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.7352	

N = # of observations, K = # of regressors

Algorithm converged.

```
*here the dtime variable estimate=0.299990
*0.0035<0.5 level so it is significant
```

*the longer it takes you to get to work on the bus the more likely you will choose auto to get to work is what this QLIM is saying

The QLIM Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-0.064434	0.399244	-0.16	0.8718
dtime	1	0.299990	0.102869	2.92	0.0035

```
* print probitout data;
*-4.85, 0 says bus-auto=negative. This says its faster to go on the bus by 48.5 minutes by car
*2.44,0 says it takes 24.4 minutes longer to take the bus. You take the bus for a variety of reasons irrespective of the time it takes you to get there vs. using a car
*this is about choice based on relative commuting times not anything else
*dtime=Bus time - auto time/10
*9.10,1 get to work by car 91 minutes faster than by bus
*-4.40,1 faster to get to work by 44 minutes but you chose your car
*2.44,0 and -4.40,1 are misclassified based on the data.
*look at the relative commute time and the choices

proc print data=work.probitout;
run;
```


Obs	autotime	bustime	dtime	auto	Xbeta_ auto	Meff_P1_ dtime	Meff_P2_ dtime
1	52.9	4.4	**-4.85**	**0**	-1.51938	-0.03773	0.03773
2	4.1	28.5	**2.44**	**0**	0.66754	-0.09578	0.09578
3	4.1	86.9	8.28	1	2.41948	-0.00641	0.00641
4	56.2	31.6	-2.46	0	-0.80241	-0.08674	0.08674
5	51.8	20.2	-3.16	0	-1.01240	-0.07169	0.07169
6	0.2	91.2	**9.10**	**1**	2.66547	-0.00343	0.00343
7	27.6	79.7	5.21	1	1.49851	-0.03894	0.03894
8	89.9	2.2	-8.77	0	-2.69534	-0.00317	0.00317
9	41.5	24.5	-1.70	0	-0.57442	-0.10148	0.10148
10	95.0	43.5	-5.15	0	-1.60938	-0.03278	0.03278
11	99.1	8.4	-9.07	0	-2.78534	-0.00247	0.00247
12	18.5	84.0	6.55	1	1.90050	-0.01967	0.01967
13	82.0	38.0	**-4.40**	**1**	-1.38439	-0.04590	0.04590
14	8.6	1.6	-0.70	0	-0.27443	-0.11526	0.11526
15	22.5	74.1	5.16	1	1.48351	-0.03982	0.03982
16	51.4	83.8	3.24	1	0.90753	-0.07928	0.07928
17	81.0	19.2	-6.18	0	-1.91837	-0.01901	0.01901
18	51.0	85.0	3.40	1	0.95553	-0.07581	0.07581
19	62.2	90.1	2.79	1	0.77254	-0.08880	0.08880
20	95.1	22.2	-7.29	0	-2.25136	-0.00949	0.00949
21	41.6	91.5	4.99	1	1.43252	-0.04290	0.04290

```
* create predicted probabilities and their classification;
*phat=predictive probabilities
*if predictive probability>50% then the individual is taking the car, if<50% then taking
the bus
*xbeta= intercept+coefficient*dtime. The dtime is the parameter estimate
*we want the predictive probabilities
*lets you compare predictive classifications with the actual

data transport2;
set work.probitout;
phat = probnorm(xbeta_auto);
phat_classify = (phat >= .5);
run;

* sort transport2 data;

proc sort data=work.transport2 out=trn2sorted;
by dtime;
run;

* print trn2sorted data;
*11 will take the bus based on this model
*-4.40,1 is misclassified
*10 classified as taking the automobile. The 1 under auto doesn't match the 0
phat_classify
*we predicted that they took the bus when they took the automobile
*these are the classifications based on the probabilities

proc print data=work.trn2sorted;
run;
```


Obs	autotime	bustime	dtime	auto	Xbeta_ auto	Meff_P1_ dtime	Meff_P2_ dtime	phat	phat_ classify
1	99.1	8.4	-9.07	0	-2.78534	-0.00247	0.00247	0.00267	0
2	89.9	2.2	-8.77	0	-2.69534	-0.00317	0.00317	0.00352	0
3	95.1	22.2	-7.29	0	-2.25136	-0.00949	0.00949	0.01218	0
4	81.0	19.2	-6.18	0	-1.91837	-0.01901	0.01901	0.02753	0
5	95.0	43.5	-5.15	0	-1.60938	-0.03278	0.03278	0.05377	0
6	52.9	4.4	-4.85	0	-1.51938	-0.03773	0.03773	0.06433	0
7	82.0	38.0	**-4.40**	**1**	-1.38439	-0.04590	0.04590	0.08312	**0**
8	51.8	20.2	-3.16	0	-1.01240	-0.07169	0.07169	0.15567	0
9	56.2	31.6	-2.46	0	-0.80241	-0.08674	0.08674	0.21116	0
10	41.5	24.5	-1.70	0	-0.57442	-0.10148	0.10148	0.28284	0
11	8.6	1.6	-0.70	0	-0.27443	-0.11526	0.11526	0.39188	0
12	4.1	28.5	2.44	0	0.66754	-0.09578	0.09578	0.74779	1
13	62.2	90.1	2.79	1	0.77254	-0.08880	0.08880	0.78010	1
14	51.4	83.8	3.24	1	0.90753	-0.07928	0.07928	0.81794	1
15	51.0	85.0	3.40	1	0.95553	-0.07581	0.07581	0.83035	1
16	41.6	91.5	4.99	1	1.43252	-0.04290	0.04290	0.92400	1
17	22.5	74.1	5.16	1	1.48351	-0.03982	0.03982	0.93103	1
18	27.6	79.7	5.21	1	1.49851	-0.03894	0.03894	0.93300	1
19	18.5	84.0	6.55	1	1.90050	-0.01967	0.01967	0.97132	1
20	4.1	86.9	8.28	1	2.41948	-0.00641	0.00641	0.99223	1
21	0.2	91.2	9.10	1	2.66547	-0.00343	0.00343	0.99616	1

```
* produce frequencies;
*0, 1 column at top is predicted, 0, 1 rows at the side are actual
*there is one explanatory variable here
*what did my model predict, and what actually happened?
*10 bus riders, 9 car drivers

proc freq data=work.trn2sorted;
tables auto*phat_classify;
run;
*
```


The FREQ Procedure

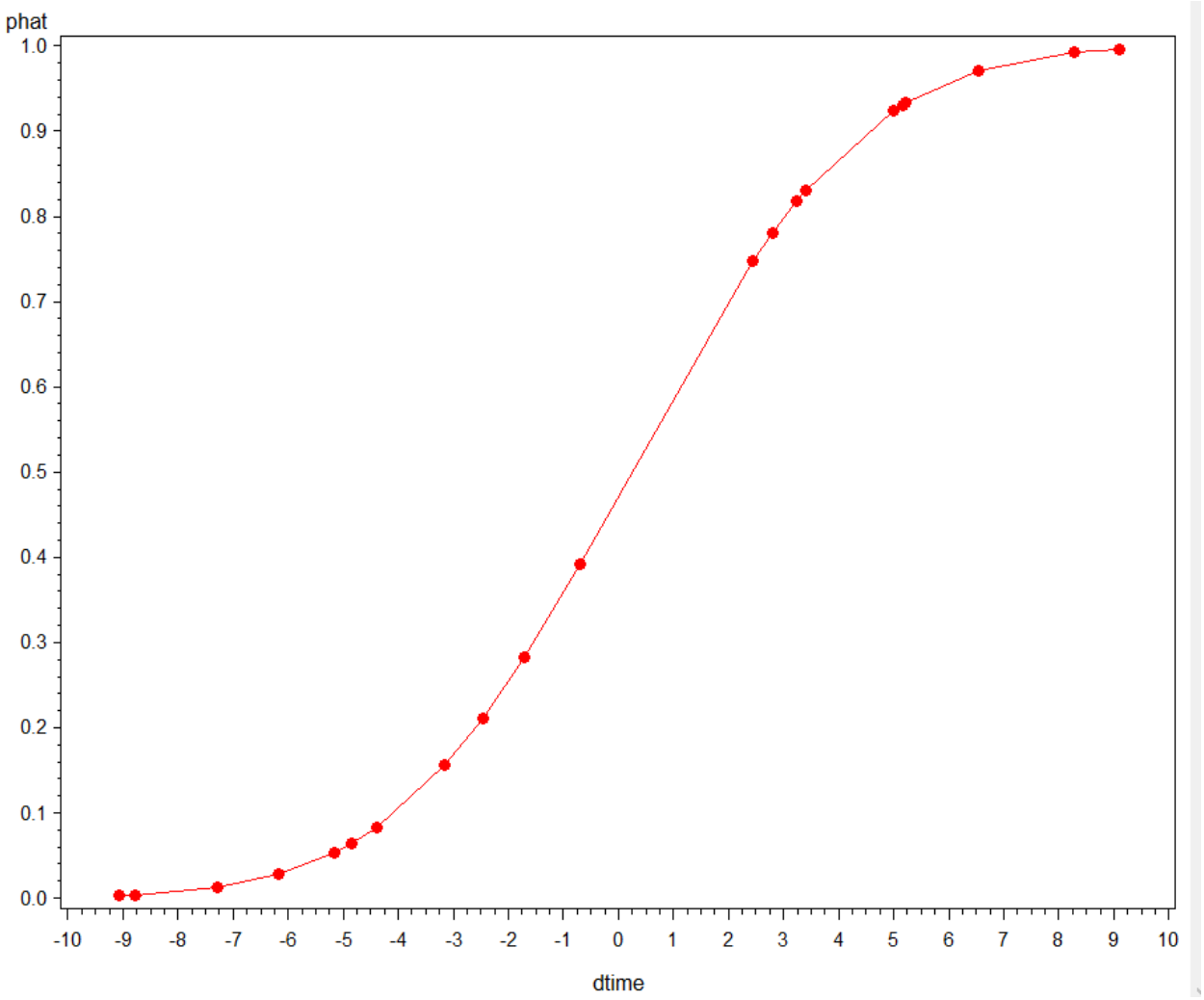
Table of auto by phat_classify

auto phat_classify

Frequency Percent Row Pct Col Pct	0	1	Total
0	10 47.62 90.91 90.91	1 4.76 9.09 10.00	11 52.38
1	1 4.76 10.00 9.09	9 42.86 90.00 90.00	10 47.62
Total	11 52.38	10 47.62	21 100.00

```
* plot trn2sorted data;
*you can get to work 90 minutes faster on bus than in a car. Therefore more likely to
choose bus then to choose car. The negatives mean bus is faster, positives is car is
faster
*the faster it is to take the bus to work (-10—1) you have lower probability that you
will take the car.
*(0-10) means you have higher probability to take a car

symbol1 value=dot color=red interpol=join;
proc gplot data=work.trn2sorted;
plot phat*dtime=1;
run;
```



```
* produce means;
*look at the probit procedure, marginal effect P1 dtime (Meff_P1_dtime), if you increase
commuter time by 10 minutes (1 unit increase) what is the change in probability when that
happens? What is the marginal effect when this happens? This is what the Meff_P1_dtime is
measuring
*the average estimated increase in the probability of choosing an automobile giving a 1
unit increase in travel time = 0.0484069 so roughly 5 minutes or 5 percentage points
*AKA what is the increase in probability of choosing a car when commute time goes up by
10 minutes is nearly 5 percentage points
*change of probability/change in commuter time
*this is a handy way to calculate that average response, probability of choosing
automobile increases by 5% with 1 unit (10 min) increase in commute time

options nolabel;
proc means data=work.trn2sorted;
var meff_p1_dtime meff_p2_dtime;
run;
```


The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Meff_P1_dtime	21	-0.0484069	0.0364573	-0.1152559	-0.0024738
Meff_P2_dtime	21	0.0484069	0.0364573	0.0024738	0.1152559

```
* estimate logit using proc qlim;
*marginal is put in so you can go to the "*produce means" and run it across your whole
data set
*the marginal effects is rate of change in vertical given change in the horizontal (see
the previous graph). Get the 21 points and average the 21 slopes

options nolabel;
proc qlim data=tmp1.transport;
model auto = dtime / discrete (d=logit);
output out=logitout marginal;
run;
quit;
```


The QLIM Procedure

Discrete Response Profile of auto

Index	Value	Total Frequency
1	0	11
2	1	10

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	auto
Number of Observations	21
Log Likelihood	-6.16604
Maximum Absolute Gradient	2.51183E-7
Number of Iterations	7
Optimization Method	Quasi-Newton
AIC	16.33208
Schwarz Criterion	18.42113

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	16.732	2 * (LogL - LogL0)
Upper Bound of R (U)	29.065	- 2 * LogL0
Aldrich-Nelson	0.4435	R / (R+N)
Cragg-Uhler 1	0.5492	1 - exp(-R/N)
Cragg-Uhler 2	0.7329	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.6947	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.5497	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.5757	R / U
Veall-Zimmermann	0.7639	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.8969	

N = # of observations, K = # of regressors

Algorithm converged.

The QLIM Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-0.237575	0.750477	-0.32	0.7516
dtime	1	0.531098	0.206425	2.57	0.0101

```
* produce means;
*says that if you increase by 1 unit you increase percentage by nearly 5 percentage
points
*logits and probits will give you almost the same result
*if you increase your commuting time relative to the car, probability of choosing the car
goes up 4.6%
*the means procedure for both match almost exactly

options nolabel;
proc means data=work.logitout;
var meff_p1_dtime meff_p2_dtime;
run;
```


The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Meff_P1_dtime	21	-0.0461540	0.0371525	-0.1211742	-0.0033454
Meff_P2_dtime	21	0.0461540	0.0371525	0.0033454	0.1211742

```
* estimate logit using proc logistic;
*instead of running a probit you are running a logit
*
options nolabel;
proc logistic data=tmp1.transport descending;
model auto = dtime;
run;
quit;
```


The LOGISTIC Procedure

Model Information

Data Set	TMP1.TRANSPORT
Response Variable	auto
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	21
Number of Observations Used	21

Response Profile

Ordered Value	auto	Total Frequency
1	1	10
2	0	11

Probability modeled is auto=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	31.065	16.332
SC	32.109	18.421
-2 Log L	29.065	12.332

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.7325	1	<.0001
Score	12.8378	1	0.0003
Wald	6.6196	1	0.0101

*the same slope as the 0.53 about in the probit model

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.2376	0.7505	0.1002	0.7516
dtime	1	0.5311	0.2064	6.6196	0.0101

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
dtime	1.701	1.135	2.549

Association of Predicted Probabilities and Observed Responses

Percent Concordant	95.5	Somers' D	0.909
Percent Discordant	4.5	Gamma	0.909
Percent Tied	0.0	Tau-a	0.476
Pairs	110	c	0.955

```
* estimate probit using proc logistic;
*use proc logistic to obtain a probit model
*logit has heavier tails than the normal

options nolabel;
proc logistic data=tmp1.transport descending;
model auto = dtime / link=probit;
run;
quit;
```


The LOGISTIC Procedure

Model Information

Data Set	TMP1.TRANSPORT
Response Variable	auto
Number of Response Levels	2
Model	binary probit
Optimization Technique	Fisher's scoring

Number of Observations Read	21
Number of Observations Used	21

Response Profile

Ordered Value	auto	Total Frequency
1	1	10
2	0	11

Probability modeled is auto=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	31.065	16.330
SC	32.109	18.419
-2 Log L	29.065	12.330

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.7342	1	<.0001
Score	12.8378	1	0.0003
Wald	8.4994	1	0.0036

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0645	0.4007	0.0259	0.8722
dtime	1	0.3000	0.1029	8.4994	0.0036

Association of Predicted Probabilities and Observed Responses

Percent Concordant	95.5	Somers' D	0.909
Percent Discordant	4.5	Gamma	0.909
Percent Tied	0.0	Tau-a	0.476
Pairs	110	c	0.955

```
* sort data descending;
*if using proc probit you need to sort your data a certain way going into the proc probit
Have the 1s first and the 0s second
*if you don't it predicts the compliments instead

proc sort data=tmp1.transport out=trnsorted;
by descending auto;
run;


* print data;

proc print data=work.trnsorted;
run;
```


Obs	autotime	bustime	dtime	auto
1	4.1	86.9	8.28	1
2	0.2	91.2	9.10	1
3	27.6	79.7	5.21	1
4	18.5	84.0	6.55	1
5	82.0	38.0	-4.40	1
6	22.5	74.1	5.16	1
7	51.4	83.8	3.24	1
8	51.0	85.0	3.40	1
9	62.2	90.1	2.79	1
10	41.6	91.5	4.99	1
11	52.9	4.4	-4.85	0
12	4.1	28.5	2.44	0
13	56.2	31.6	-2.46	0
14	51.8	20.2	-3.16	0
15	89.9	2.2	-8.77	0
16	41.5	24.5	-1.70	0
17	95.0	43.5	-5.15	0
18	99.1	8.4	-9.07	0
19	8.6	1.6	-0.70	0
20	81.0	19.2	-6.18	0
21	95.1	22.2	-7.29	0

```
* estimate probit using proc probit;

options nolabel;
proc probit data=work.trnsorted order=data;
model auto = dtime;
run;
quit;
```


The Probit Procedure

Model Information

Data Set	WORK.TRNSORTED
Dependent Variable	auto
Number of Observations	21
Name of Distribution	Normal
Log Likelihood	-6.16515849

Number of Observations Read	21
Number of Observations Used	21

Class Level Information

Name	Levels	Values
auto	2	1 0

Response Profile

Ordered Value	auto	Total Frequency
1	1	10
2	0	11

PROC PROBIT is modeling the probabilities of levels of auto having LOWER Ordered Values in the response profile table.

Algorithm converged.

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
dtime	1	8.5047	0.0035

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.0644	0.3992	-0.8469	0.7181	0.03	0.8718
dtime	1	0.3000	0.1029	0.0984	0.5016	8.50	0.0035

The Probit Procedure

Probit Model in Terms of Tolerance Distribution

MU	SIGMA
0.21478651	3.33344679

Estimated Covariance Matrix
for Tolerance Parameters

	MU	SIGMA
MU	1.778164	0.096266
SIGMA	0.096266	1.306559