Problem Set 3
C528
Predictive Analytics for Business Strategy

1. Use the Amazon worksheet in PS3.xlsx to answer the following questions:
    a. Show and discuss the effects of the new content on viewing time for Amazon around the 80th percentile.

```
. qreg ViewTime NewContent, quantile(80)
Iteration  1:  WLS sum of weighted deviations =  356882.76

Iteration  1: sum of abs. weighted deviations =   358707.4
Iteration  2: sum of abs. weighted deviations =   324635.4
Iteration  3: sum of abs. weighted deviations =   275216.8

.8 Quantile regression                          Number of obs =     19,593
  Raw sum of deviations 304436.8 (about 157)
  Min sum of deviations 275216.8                Pseudo R2     =     0.0960

------------------------------------------------------------------------------
    ViewTime | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  NewContent |          46   1.078877    42.64   0.000     43.88531    48.11469
       _cons |         134   .7641455   175.36   0.000     132.5022    135.4978
------------------------------------------------------------------------------
```

    New content increases viewing time by 46 minutes for customers in the 80th percentile. As discussed in class, the customers in high percentiles are already viewing content, so new content could help with customer retention, i.e., if Amazon is concerned about losing customers who have large viewing times due to running out of content to watch, new content could help retain those customers by preventing them from running out of content.

    b. Is Amazon likely most interested in this quantile? Explain why or why not.

    Amazon is most likely not interested in this quantile. They would likely be more interested in losing customers who already don't view much content. Anecdotally, this would be me (Thomas) because I have Prime but have maybe watched one movie on Prime Video this year. Amazon would likely be more concerned with losing a customer like me, so adding new content could be attractive enough that if I were considering dropping my Prime subscription, the new content could be just enough to keep me as a subscriber. Another way of putting it would be that the 80th percentile is the right tail, but Amazon is likely more interested in the left tail.

2. How would you test for heterogeneity in treatment effects? Explain what you are looking for and how it works.
    Heterogeneity of treatment effects can be tested using the ratio between the variances of the treated and untreated groups. If the variances are equal

(homogenous) the ratio of the variances would equal approximately 1. Very extreme ratios (e.g., 17,000 or 0.00001) indicate the variances in the treatment groups do not equal 1 and instead differ from each other. If the variances in the treatment groups differ there is heterogeneity in treatment effects.

Running sdtest ViewTime, by(NewContent) shows that we have a ratio of 0.4454 with a p-value of 0, therefore, we reject our null hypothesis that there is no treatment effect heterogeneity.

```
. sdtest ViewTime, by(NewContent)

Variance ratio test
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	9,764	99.3388	.4010946	39.63334	98.55257	100.125
1	9,829	130.3579	.5990375	59.38937	129.1837	131.5322
Combined	19,593	114.8998	.3775335	52.84525	114.1598	115.6398

```
    ratio = sd(0) / sd(1)                                     f =   0.4454
H0: ratio = 1                           Degrees of freedom = 9763, 9828

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.0000         2*Pr(F < f) = 0.0000          Pr(F > f) = 1.0000
```

3. Use the PassFail worksheet in PS3.xlsx to answer the following questions:
    a. Give and explain two reasons for using a logit model over a linear probability model.

    In the context of predicting a binary outcome such as passing a test (Y = Pass, where 0 is fail and 1 is pass) based on a continuous variable like hours studied (X = Hours studied), a logit model offers several advantages over a linear probability model:

    1. Bounded Probabilities: The logit model ensures that all predicted probabilities fall between 0 and 1, which is important when we are predicting a probability. For example, let's say a student studies for 10 hours and our linear model predicts a probability of 1.2 of passing the test - this doesn't make sense as probabilities should be between 0 and 1. However, the logit model, because it applies a logistic function to the linear model's output, always returns a probability between 0 and 1, regardless of the value of the input.

    2. Interpretation of the Data-Generating Process: The logit model can offer a more realistic representation of the data-generating process when dealing with binary outcomes. In real-world scenarios, the effect of a predictor on the likelihood of an outcome is often not linear. For example, when studying for a test, the first few hours studied might significantly increase the probability of passing the test, while additional hours studied might not have as large an effect. This type of non-

linear relationship is captured well by the logit model but is missed in a linear probability model. This is because in a logit model, a unit change in the predictor variable changes the odds of the outcome by a certain factor, not by a constant amount.

b. Give and explain one benefit of using a logit model over a probit model.

Robust to Extreme Observations: The logistic distribution, which underpins the logit model, has heavier tails compared to the normal distribution used in the probit model. This makes the logit model more robust to outliers or extreme observations, which are often present in real-world datasets. For instance, in predicting the likelihood of passing a test based on hours studied, if there are a few students who studied an extraordinarily high number of hours, the logit model would be less affected by these extreme observations than the probit model.

c. Estimate a logit model of Pass on Hours Studied (showing how you arrived at the answers).

   i. Show and discuss the meaning of the results.

      The marginal effect is the change in the likelihood of a student passing the test for every additional hour spent studying, assuming all other variables (such as innate ability and prior experience) remain unchanged. In other words, it quantifies how much the probability of a successful outcome (passing the test) shifts when the study time increases by one hour, given that all other contributing factors are held constant. In this case, each additional hour studied increases the individual's preparedness to pass by 11.7%, all else equal.

```
. logit Pass HoursStudied

Iteration 0:   log likelihood = -59.854953
Iteration 1:   log likelihood = -46.312071
Iteration 2:   log likelihood = -45.127787
Iteration 3:   log likelihood = -45.113648
Iteration 4:   log likelihood = -45.113637
Iteration 5:   log likelihood = -45.113637

Logistic regression                                     Number of obs =    105
                                                        LR chi2(1)    =  29.48
                                                        Prob > chi2   = 0.0000
Log likelihood = -45.113637                             Pseudo R2     = 0.2463

------------------------------------------------------------------------------
        Pass | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
HoursStudied |   .1173602   .0260097     4.51   0.000     .0663822    .1683382
       _cons |  -3.806059   .7414874    -5.13   0.000    -5.259347   -2.35277
------------------------------------------------------------------------------
```

   ii. What is the marginal effect of increasing from 5 hours to 6 hours?

There is a 0.434% increase in the probability of passing going from 5 to 6 hours studied.

```
. mfx compute, at(HoursStudied = 5)

Marginal effects after logit
      y  = Pr(Pass) (predict)
         =  .03844741
```

variable	dy/dx	Std. err.	z	P>|z|	[95% C.I.]		X
HoursS~d	.0043387	.00167	2.60	0.009	.001072	.007606	5

iii. What is the marginal effect of increasing from 14 hours to 15 hours?

There is a 1.08% increase in the probability of passing going from 14 to 15 hours studied.

```
. mfx compute, at(HoursStudied = 14)

Marginal effects after logit
      y  = Pr(Pass) (predict)
         =  .10312116
```

variable	dy/dx	Std. err.	z	P>|z|	[95% C.I.]		X
HoursS~d	.0108543	.0023	4.73	0.000	.006355	.015353	14