part A

(a). plug into all these values into the model

note : $(\ln y, x)$ refer to $\%\Delta y = (100\beta_1)\Delta x$

$H=141$, $W=3542$ $L=205$, $V=0.25$, $T=0.75$, $A=0.2$, $P=0.1$,

$B=0.05$ $C=0.1$ D_i $(i=1\cdots6)=0$.

$\ln$ Price ≈ 7.878

Price ≈ 2637

(b) Holding all other terms contant, a car with V8 & power steering in 1955 costs $(-4.4\%)+(1\%)+(8.8\%)=5.4\%$.

(c) The increase from 56' to 57' is $D_3-D_2 = 1.9\%-(-1.5\%)=3.4\%$.

56' to 59' is $D_5-D_2 = 4.4\%-(-1.5\%)=5.9\%$.

(d). 100 pounds = 0.1 t-pound.

To exclude the joint effect of additional weight from the hard top, let $\ln \text{Price} = \beta_W(W+0.1T)+\beta_T T+\cdots$

$\ln \text{Price} = 0.1\beta_W T+\beta_T T+\beta_W \cdot W$

$0.023 = b_t+0.1\times0.249$ $\Rightarrow b_t=-0.099$.

The hard top reduces the cost by 0.19%.

(e). $F = \frac{(0.922 - 0.919)/16}{(1-0.922)/(570-16)} \approx 3.55$ > critical value,

Hence reject the null hypothesis.

(f). $F = \frac{(e'e - \sum_{t=1}^{T} e_t' e_t)/(m-(k+1))}{\sum_{t=1}^{T} e_t' e_t/(n-m)}$

$$= \frac{[1425 - (104 + \cdots + 211)]/(70-16)}{[104 + \cdots + 211]/(570-70)} \approx 1.175 < \text{critical value}$$

$$\frac{165/54}{1300/500}$$

∴ Cannot reject the null hypothesis.

part B

1.

(a) only "mhispan" is not significant.

Most covariates increases the prob of the dead of a baby within a year, except age, education, foreign born.

(b) $LR = -2 \frac{\max L_r}{\max L_u} = -2(l_r - l_u)$.

$LR \sim \chi_{k-1} = \chi_{10}$

From the regression output, $l_r = -27627$, $l_u = -27049$.

$LR = 1156$

prob > chi2(10) = 0.

Hence reject the hypothesis that all coefficients are insignificant.

Assume that when the predicted prob > 0.5, it's a "good" prediction. Sum up such predictions. Their share is 0.99. Thus the percentage of "correct prediction" is 99%.

(c)., see outputs

(d). see outputs

The absolute values of APE are slightly larger than marginal effect at means

扫描全能王 创建

APE is prefered since it give weights to the whole distribution

(e) The average effect for alcohol : 0.0016

" " " tobacco : 0.0039

They are basically same as in (c). (c) : $\hat{\beta}_{alcohol} = 0.0015$

$\hat{\beta}_{tobacco} = 0.0035$

$\hat{\beta}_{tobacco}$ is away from the one in (c). Treating binary variables as discret variables is more appropriate.

2.

(a) See output.

LR = 1151   prob > chi2(10) = 0.

Reject the hypothesis which all coefficients are zero.

APE : $\hat{\beta}_{tobacco} = 0.0039$

$\hat{\beta}_{alcohol} = 0.0014$

(b) The APE are very close to each other. Parameters in the logit model differ from those in Q1 because the form of the function changed.

(c) ML estimator $Pr(dead = 1 | X_1 \cdots X_n) = F(aX_1^2 + bX_1 + cX_2 \cdots\cdots)$

The function minimizes when $X_1 = \frac{b}{-2a} \approx 28.37$ yr

Delta method in asymptotic variance:

if $\hat{\beta} \sim N(\beta, V)$ and $c(\cdot)$ is differentiable.

$$c(\hat{\beta}) \sim N\left(c(\beta), \left[\frac{dc(b)}{db_1}, \dots, \frac{dc(b)}{db_k}\right] V \left[\frac{dc(b)}{db_1}, \dots \frac{dc(b)}{db_k}\right]'\right)$$

$$\sqrt{n}\,(c(\hat{\beta}) - c(\beta)) \xrightarrow{d} N\left[0,\ n\left[\frac{dc(b)}{db_1}, \dots, \frac{dc(b)}{db_k}\right] \cdot V \cdot \left[\frac{dc(b)}{db_1}, \dots, \frac{dc(b)}{db_k}\right]'\right)$$

$$AVAR(c(\hat{\beta})) = \left[\frac{dc(b)}{db_1}, \dots, \frac{dc(b)}{db_k}\right] \cdot V \cdot \left[\frac{dc(b)}{db_1}, \dots, \frac{dc(b)}{db_k}\right]' \Bigg|_{b=\hat{\beta}}$$

where $V = \hat{V} = AVAR(\hat{\beta})$

CI: $[X_1 - 1.96\sqrt{AVAR(c(\hat{\beta}))},\ X_1 + 1.96\sqrt{AVAR(c(\hat{\beta}))}]$

use "nlcom" command to compute X_1 & CI & std.err.

CI: [26.28, 30.47].

(d). $Pr(y=1|X) = F(X\hat{\beta})$

plug into these conditions, $\hat{Pr}(y=1|X) = 0.0107$ (also by delta method).

CI: [ ]

(similar process in (c)).

(e). marginal effect: $\frac{dPr(y=1|x)}{dx} = f(x\beta)\cdot\beta$.

For obs at mean, the marginal effect is smaller than obs with specific conditions. $\because$ They are in different parts of the distribution. Effects at means are more fitted to the whole set

AVAR: $\frac{\partial f(X\hat{\beta})\cdot\hat{\beta}}{\partial\beta}\cdot[n\,\mathfrak{I}(\hat{\beta})]^{-1}\cdot\left[\frac{\partial f(X\hat{\beta})\cdot\hat{\beta}}{\partial\beta}\right]$

(4). Estimators from the LPM model differs from probit & logit models.

OLS is not a good method to estimate $Pr(y=1|x)$ since it only has two values. The marginal effect differs from 2(e) especially in extreme values.

The predict prob is 0.01138. $AVAR(c(\hat{\beta})) = 1.903\times10^{-7}$.

CI is: $[0.0105, 0.0122]$.

3. let L_1 be the likelihood of the full model.

L_0 be the likelihood of the "constant-only" model.

χ^2 is defined as $2(L_1 - L_0)$.

R is defined as $1-\frac{L_1}{L_0}$.

扫描全能王 创建

part C

(a) see output.

(b) use nlcom to calculate $AVAR[C(\hat{\beta})]$

see output

(c) Bootstrap std. errors are larger.

(d) see output.

part D

1. $\hat{\beta} = -266.03$ std. error is 4.76.

Assumption: smoking or not is independent with things can also decide the birth weight

2. $\hat{\beta}_{tobacco} = -231.98$ std. error is 4.72

CI: [-241.24, -222.71]

The effect of tobacco decreases. It makes sence since ppl who smoke during pregnancy tend to have other similar preference which would also effect the birth wieght.

when the "tobacco" indicator is randomly assigned, the $\hat{\beta}_{tobacco}$ truly reflects the causal effect

3. (a) The two knernel density distribution are similar when the bandwidth is large.

$h = \frac{2 \cdot sd(x)}{N^{\frac{1}{5}}}$ ← optimal bandwidth. $h \approx 45.3669$

(stata output).

(b) Ppl who smoke has lighter babies on expecation values.

The effect of tobacco looks constant around the mean value, while at the right tail its effect decreases...

see output. use the default estimator. and optimal bandwidth because it's optimal.

4. $\hat{\beta}_{tobacco} = -227.92$ (interaction terms: race # medue.

$|\hat{\beta}_{tobacco}|$ decreases as adding more interaction terms.

In this case, non-linear function may be more suitable.

It's semiparametric since the model allows interaction between explantary variables.

The benefit of this model is, we consider more internal effect, decreasing the possibility of overestimation. The disadvantage is, if there exists too many variables.

5. use the default kernel function, and optimal bandwidth which stata calculated.

The relationship between smoking and education is non-linear. The less education the mother has, the higher probability of being a smoker during pregnancy is.

However, on the right side of the graph, the effect is negative.

A possible explantion is the tax of cigar has different effect on different people.

The effect is not a causal effect since there aren't contol variables.

The distribution of "cigar" has a fat tail on its left side while there're few observations in other groups. Therefore, the kernel regression may not be a good choice.