Statistics and Econometrics for Travel Behavior

Matthew Wigginton Conway
School of Geographical Sciences and Urban Planning, Arizona State University

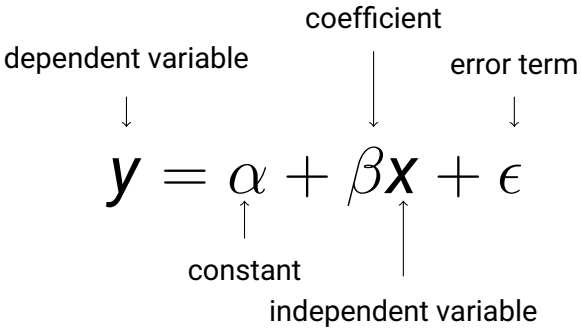
August 27, 2020
Version: 1.1

Table of contents

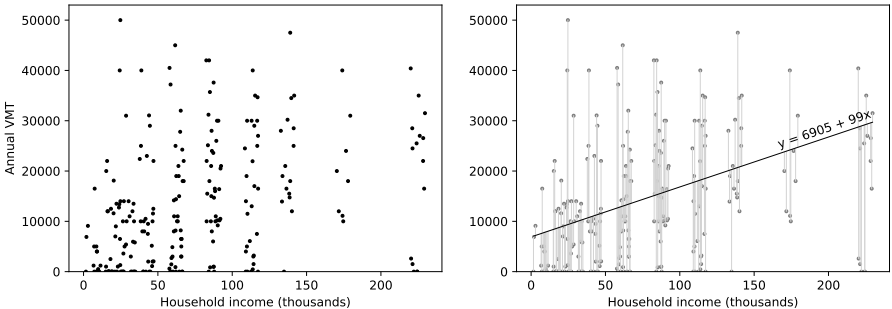
Linear regression

- Estimate the associations between dependent and independent variables
- Predict or explain the variation in the dependent variable
- For instance, what is the relationship between income and annual vehicle miles traveled

Linear regression: the math



Linear regression: example



Data: 2017 NHTS

Linear regression: example

annual vehicle miles traveled

$$y = 6905 + 99x$$

income (thousands)

Data: 2017 NHTS

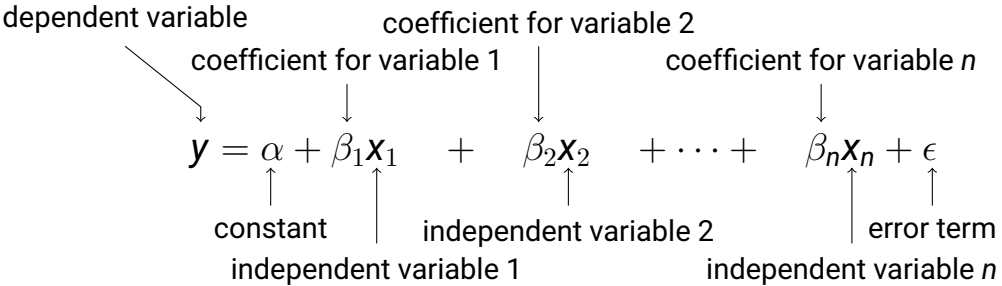
Linear regression: table form

	Coefficient	Std. err.	*t*-value	*p*-value	95% Conf.	Int.
Constant	6905	1659	4.161	0.0	3637	10173
Income (thousands)	99	17	5.788	0.0	66	133

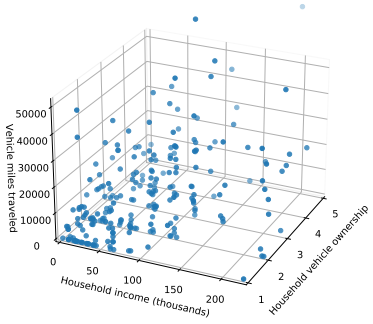
Dependent variable	Annual vehicle miles traveled		
R^2	0.12	Adjusted R^2	0.12
Sample size	250		

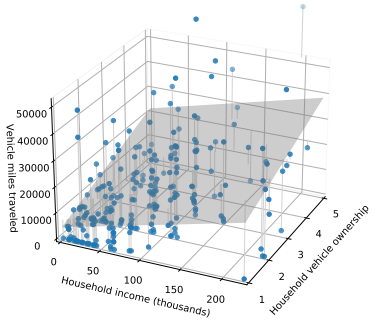
Data: 2017 NHTS

Multiple linear regression



Multiple linear regression





Data: 2017 NHTS

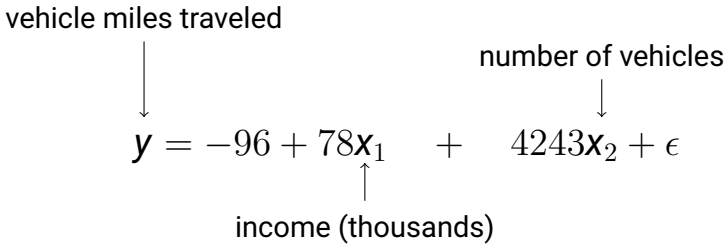
Multiple linear regression

	Coefficient	Std. err.	*t*-value	*p*-value	95% Conf.	Int.
Constant	-96	2187	-0.044	0.965	-4403	4211
Income (thousands)	78	17	4.575	0.000	45	112
Number of vehicles	4243	908	4.675	0.000	2455	6030

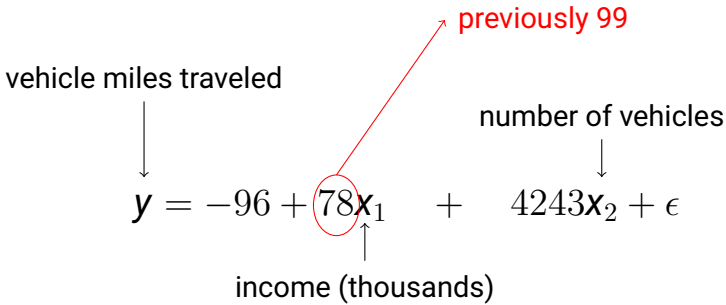
Dependent variable	Annual vehicle miles traveled		
R^2	0.19	Adjusted R^2	0.18
Sample size	250		

Data: 2017 NHTS

Multiple linear regression



Multiple linear regression



Data: 2017 NHTS

Control variables

- Every coefficient in a multiple regression model is the change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant
- The income coefficient now represents the association between increased income and driving, at a constant level of car ownership
- Often many variables in the model are control variables
- The association with control variables is not of interest, but they protect the coefficients of interest from bias

Handling categorical variables in a regression

- What if we want to include a variable with categories in a regression?
- For example, maybe the region where people live is associated with how much they drive
- What if we coded them as Northeast = 1, Midwest = 2, West = 3, South = 4?

Handling categorical variables in a regression

- What if we want to include a variable with categories in a regression?
- For example, maybe the region where people live is associated with how much they drive
- What if we coded them as Northeast = 1, Midwest = 2, West = 3, South = 4?
- Assumes that driving in the Northeast < Midwest < West < South, or vice versa
- Assumes that the difference from Northeast to Midwest is the same as Midwest to South

Handling categorical variables in a regression

- The solution is to dummy variables
- We create one variable for each category
- Set this variable to 1 if the observation is in that category, 0 otherwise
- Exclude one category from the regression
- Coefficients for other categories represent differences from excluded category

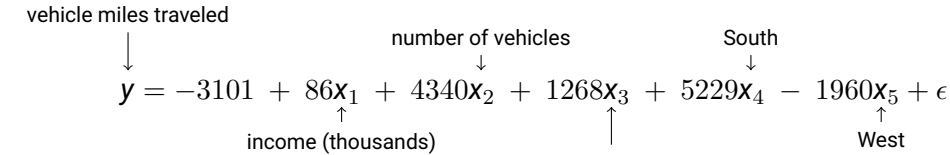
Handling categorical variables in a regression

	Coefficient	Std. err.	*t*-value	*p*-value	95% Conf.	Int.
Constant	-3101	3109	-0.997	0.320	-9226	3024
Income (thousands)	86	17	4.997	0.000	52	120
Number of vehicles	4340	900	4.822	0.000	2567	6112
Region: Midwest	1268	3586	0.353	0.724	-5796	8332
Region: South	5229	2865	1.825	0.069	-415	10873
Region: West	-1960	3163	-0.619	0.536	-8191	4272

Dependent variable	Annual vehicle miles traveled		
R^2	0.23	Adjusted R^2	0.21
Sample size	250		

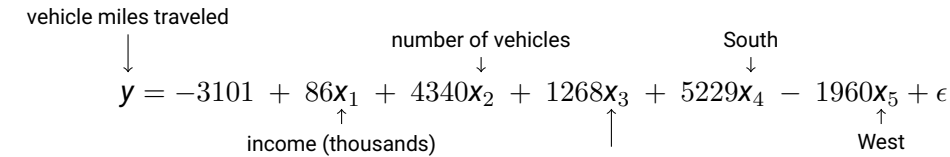
Data: 2017 NHTS

Handling categorical variables in a regression



Data: 2017 NHTS

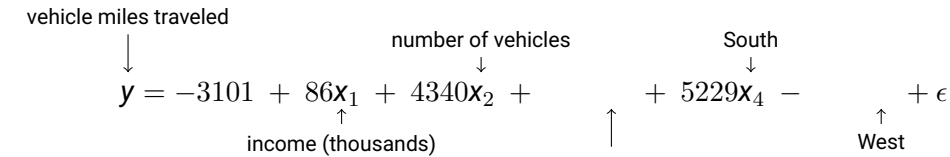
Handling categorical variables in a regression



Data: 2017 NHTS

What is the estimated vehicle miles traveled for a household in the South?

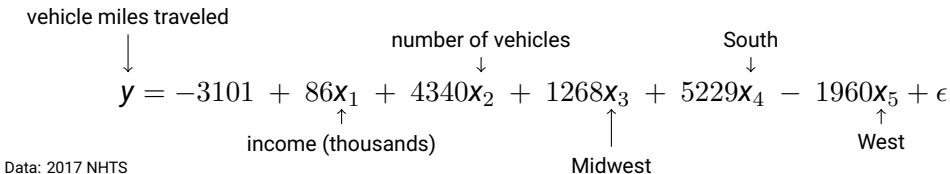
Handling categorical variables in a regression



Data: 2017 NHTS

What is the estimated vehicle miles traveled for a household in the South?

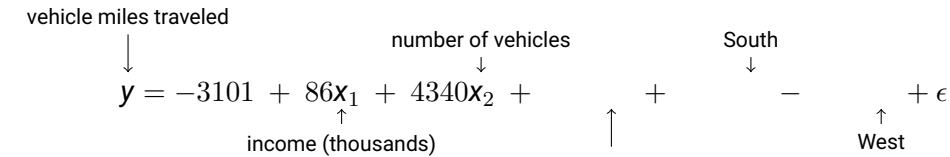
Handling categorical variables in a regression



Data: 2017 NHTS

What is the estimated vehicle miles traveled for a household in the Northeast?

Handling categorical variables in a regression



Data: 2017 NHTS

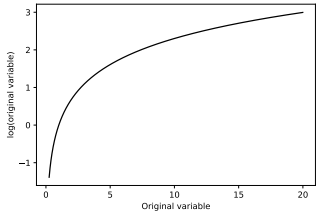
What is the estimated vehicle miles traveled for a household in the Northeast?

Variable transformations

- Linear regression assumes linear relationships between independent and dependent variables
  - i.e. a one-unit change in the independent variable associated with a constant change in the dependent variable
- Sometimes we might think the effect is non-linear
- To handle this, we transform either the dependent or independent variables

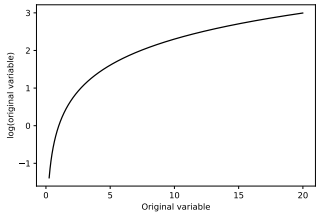
Variable transformations: the logarithm

- The logarithm is a common variable transformation
- As the original variable gets larger, the change in the logarithm gets smaller



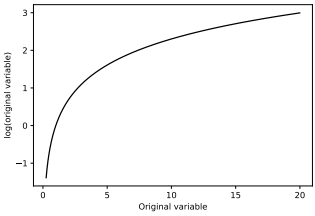
Variable transformations: logged dependent variable

- We log dependent variables when the effects of independent variables get larger at as the dependent variable does
- Maybe an additional bedroom is more valuable in a $1,000,000 home than a $100,000 home
- Coefficient interpretation: a one-unit increase in the independent variable is associated with an $100(e^{\beta} - 1)$% increase in the dependent variable (Ford 2018)



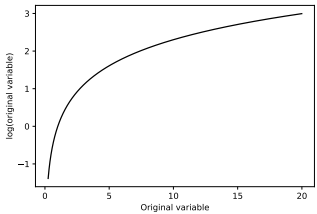
Variable transformations: logged dependent variable

- We log dependent variables when the effects of independent variables get larger at as the dependent variable does
- Maybe an additional bedroom is more valuable in a $1,000,000 home than a $100,000 home
- Coefficient interpretation: a one-unit increase in the independent variable is associated with an $100(e^{\beta} - 1)$% increase in the dependent variable (Ford 2018)
  - when β is small, $e^{\beta} - 1 \approx \beta$



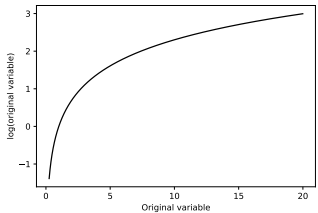
Variable transformations: logged independent variable

- We log independent variables when we think the association attenuates at larger values of the independent variable
- Maybe at high incomes, additional income matters less to driving, because people already go everywhere they want to
- Coefficient interpretation: a 1% increase in the independent variable is associated with a $\frac{\beta}{100}$ absolute increase in the dependent variable (Ford 2018)



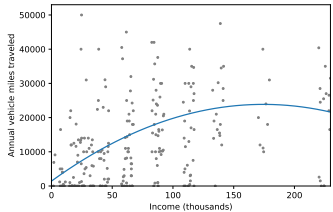
Variable transformations: logged dependent and independent variable

- We log dependent and independent variables when we think the relationship is multiplicative rather than additive
- The coefficient is then an elasticity
- ...the percent change in the dependent variable for a 1% change in the independent variable



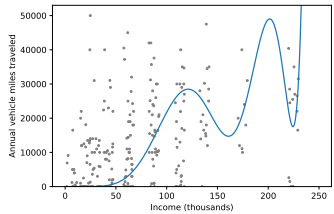
Other variable transformations

- Power functions allow modelling U-shaped functions
  - Powers beyond 2 are questionable
- Piecewise regression allows different coefficient in different ranges
- Categorization estimates same value in each range



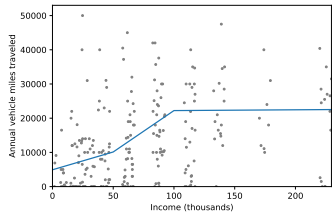
Other variable transformations

- Power functions allow modelling U-shaped functions
  - Powers beyond 2 are questionable
- Piecewise regression allows different coefficient in different ranges
- Categorization estimates same value in each range



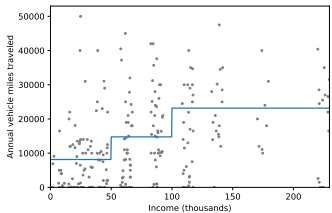
Other variable transformations

- Power functions allow modelling U-shaped functions
  - Powers beyond 2 are questionable
- Piecewise regression allows different coefficient in different ranges
- Categorization estimates same value in each range



Other variable transformations

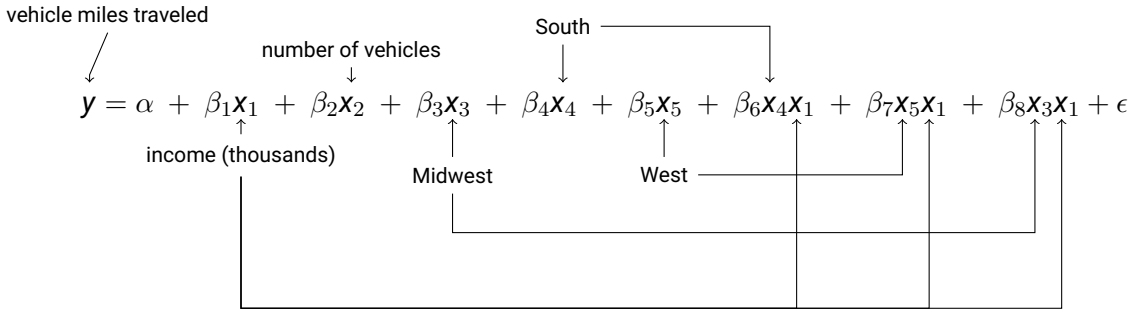
- Power functions allow modelling U-shaped functions
  - Powers beyond 2 are questionable
- Piecewise regression allows different coefficient in different ranges
- Categorization estimates same value in each range



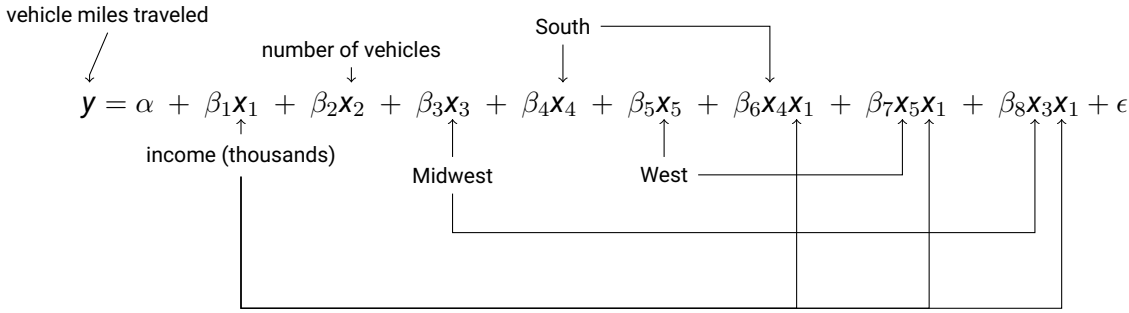
Interaction terms

- What if we think the relationship between two variables is dependent on the value of a third variable?
- Maybe the effect of income differs across regions of the country
- We can use an interaction term
- ...which means multiplying the two variables together
- Most often done with at least one dummy variable, easier to interpret

Interaction terms

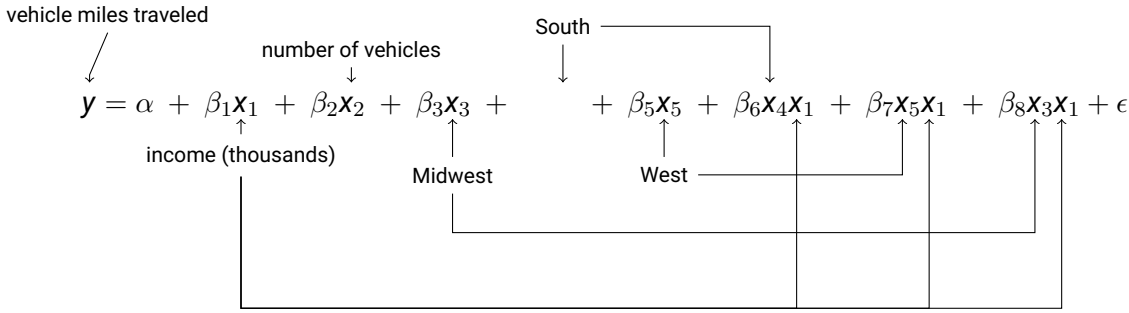


Interaction terms



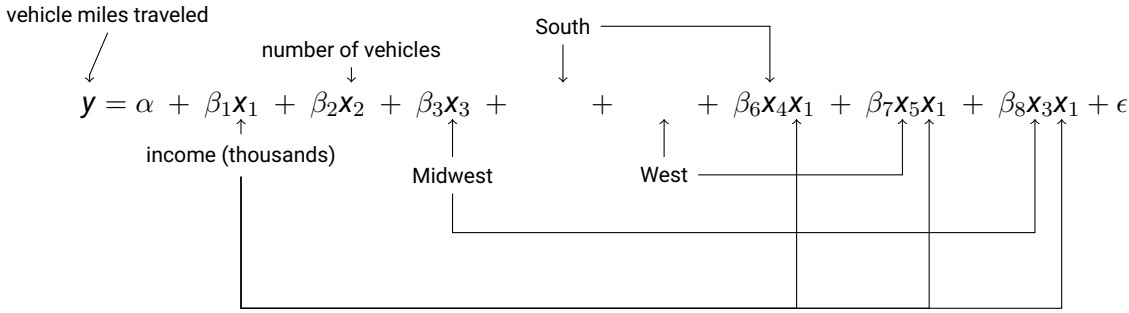
What is the predicted VMT of a respondent in the Midwest?

Interaction terms



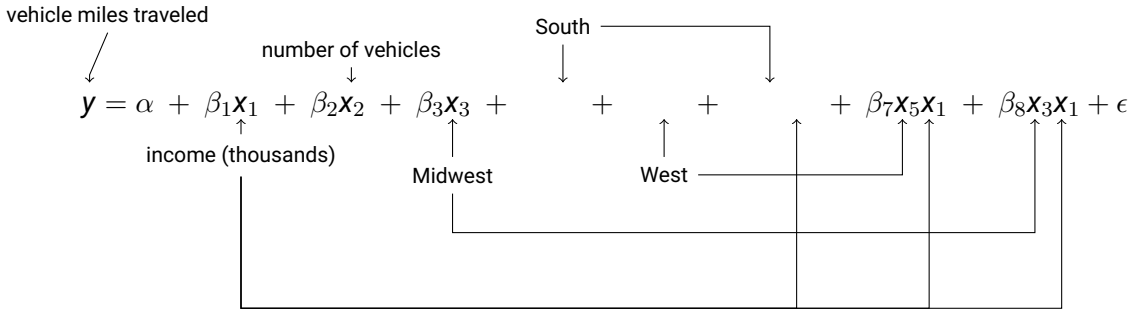
What is the predicted VMT of a respondent in the Midwest?

Interaction terms



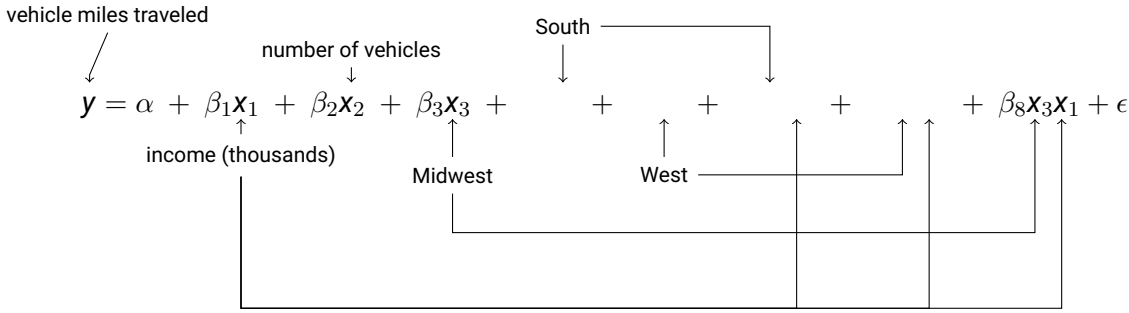
What is the predicted VMT of a respondent in the Midwest?

Interaction terms



What is the predicted VMT of a respondent in the Midwest?

Interaction terms



What is the predicted VMT of a respondent in the Midwest?

Interaction terms

	Coefficient	Std. err.	*t*-value	*p*-value	95% Conf.	Int.
Constant	-2291	4649	-0.493	0.623	-11449	6867
Income (thousands)	74	50	1.484	0.139	-24	172
Number of vehicles	4393	903	4.867	0.000	2615	6170
Region: Midwest	-4598	6167	-0.746	0.457	-16746	7550
Region: South	3937	5019	0.784	0.434	-5950	13823
Region: West	-235	5644	-0.042	0.967	-11353	10883
Income (West)	-15	57	-0.255	0.799	-127	98
Income (South)	17	55	0.313	0.755	-92	126
Income (Midwest)	91	73	1.249	0.213	-53	235

Dependent variable	Annual vehicle miles traveled		
R^2	0.24	Adjusted R^2	0.21
Sample size	250		

Data: 2017 NHTS

Pitfalls of linear regression: assumption of linearity

- Linear regression assumes relationships are linear
- Transformations can help

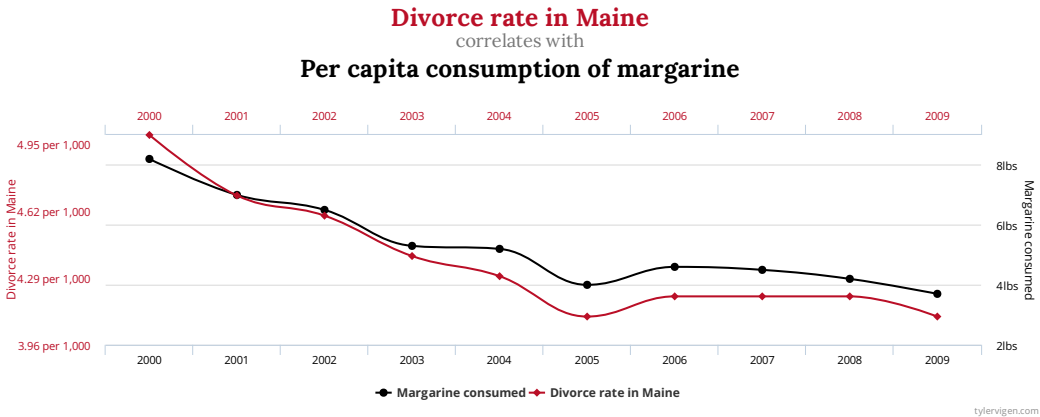
Pitfalls of linear regression: assumption of linearity

- Linear regression assumes relationships are linear
- Transformations can help
- ...but not always clear which one to use

Pitfalls of linear regression: assumption of linearity

- Linear regression assumes relationships are linear
- Transformations can help
- ...but not always clear which one to use
- Best defense: does the model make sense?

Pitfalls of regression: correlation does not imply causation



Pitfalls of regression: omitted variable bias

- Every estimated association in the model depends on all the independent variables in the model
- If important variables are omitted, the remaining coefficients will be biased
- If we didn't have income in our model, the coefficient for number of vehicles would likely be larger
- ...because it would capture some of the income effect

Pitfalls of regression: omitted variable bias

- Every estimated association in the model depends on all the independent variables in the model
- If important variables are omitted, the remaining coefficients will be biased
- If we didn't have income in our model, the coefficient for number of vehicles would likely be larger
- ...because it would capture some of the income effect
- Best defense: are there obvious variables missing from the model?

Pitfalls of regression: multicollinearity

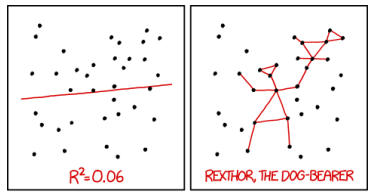
- When variables are highly correlated, the regression cannot differentiate between them
- For example, population density and intersection density might be highly correlated
- The result is that the regression will not be able to determine which variable the independent variable is associated with, and both may become insignificant

Pitfalls of regression: multicollinearity

- The usual diagnostic is the Variance Inflation Factor (VIF) for each coefficient
- Measured how much coefficient variance (square of the standard error) is increased due to multicollinearity
- Usual thresholds for concern are 4, 5 or 10, but depends on sample size and other factors (see O'Brien 2007)

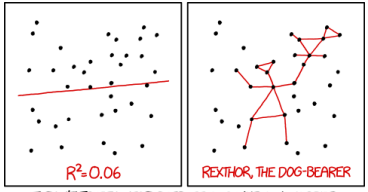
Evaluating the fit of regression models: R^2

- proportion of variation in dependent variable explained by independent variables
- Ranges from 0–1
- Higher is better



Evaluating the fit of regression models: Adjusted R^2

- Penalizes R^2 for the number of variables in the model
- When adding variables to a model, normal R^2 cannot go down
- But adjusted R^2 can
- Helps prevent overfitting



Evaluating the fit of regression models: log-likelihood

- The likelihood is the probability of observing the data given the coefficients
- Maximum likelihood estimation finds coefficients to maximize this probability
- Usually the logarithm of likelihood is used because it makes math easier
- Since probabilities range from 0–1, log-likelihoods range from $-\infty$–0
- Higher (closer to zero) is better
- Can't compare log-likelihoods of models with different dependent variables or samples
- A likelihood-ratio test is often used to determine if a more complex model is better than a nested simpler model

Evaluating the fit of regression models: log-likelihood

- Authors may present up to three log-likelihood values
  - Log-likelihood at convergence (may be notated LL(β)): Log-likelihood of the full model with all coefficients
  - Log-likelihood at constant(s) (LL(C)): Log-likelihood with only the constant in the model (or only the alternative specific constants in multinomial model)
  - Log-likelihood at zero/null (LL(0)): Log-likelhood with all coefficients including the constant at zero
- Models try to maximize LL(β)
- The difference from LL(C) or LL(0) and LL(β) is measure of how much better the model does than a model with no predictive power

Evaluating the fit of regression models: AIC and BIC

- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) also measure the fit of the model
- Transformations of the likelihood that penalize additional parameters, like Adjusted R^2
- Smaller values are better (unlike log-likelihood)
- Can't compare across different dependent variables or datasets

Evaluating the fit of regression models: AIC and BIC

not to be confused with

AOC

Bic

Evaluating the fit of regression models: AIC and BIC

- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) also measure the fit of the model
- Transformations of the likelihood that penalize additional parameters, like Adjusted R^2
- Smaller values are better (unlike log-likelihood)
- Can't compare across different dependent variables or datasets

Evaluating the fit of regression models: pseudo-R^2

- Transformation of the log-likelihood function to make it behave “more like” an R^2
- Still can’t be compared across different dependent variables or datasets (“FAQ: What Are Pseudo R-Squareds?” 2011)
- Many types of pseudo-R^2’s available

Evaluating the fit of regression models: summary

Measure	Higher or lower is better	Penalty for number of terms	Comparable across different data
R^2	Higher is better	No	Yes
Adjusted R^2	Higher is better	Yes	Yes
Log-likelihood	Higher (closer to zero) is better	No	No
Akaike information criterion	Lower is better	Yes	No
Bayesian information criterion	Lower is better	Yes	No
pseudo-R^2	Higher is better	Maybe	No

Evaluating the fit of regression models: why bother?

- Goodness of fit is important if you want to predict outcomes
- But when explaining relationships, it usually doesn't matter very much
- Exception is when the predictive power is low because of omitted variables

Hypothesis testing

- A way to evaluate whether a statistic is likely to be true if re-calculated with a different sample
- Tests a hypothesis, for instance that the average income is greater than $50,000

Hypothesis testing: intuitive explanation

- Suppose you want to know if people prefer raspberry or pistachio ice cream
- So, you run a survey
- If you ask three random people, and two prefer pistachio, does that mean that most people prefer pistachio?

Hypothesis testing: intuitive explanation

- Suppose you want to know if people prefer raspberry or pistachio ice cream
- So, you run a survey
- If you ask three random people, and two prefer pistachio, does that mean that most people prefer pistachio?
- If you asked three more random people, do you think you would get the same result?

Hypothesis testing: intuitive explanation

- Suppose you want to know if people prefer raspberry or pistachio ice cream
- So, you run a survey
- If you ask three random people, and two prefer pistachio, does that mean that most people prefer pistachio?
- If you asked three more random people, do you think you would get the same result?
- What about if you asked 3000 people, and 2000 said they preferred pistachio?

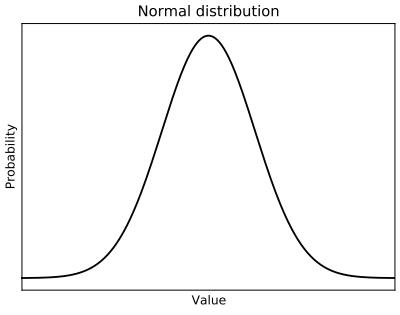
Hypothesis testing

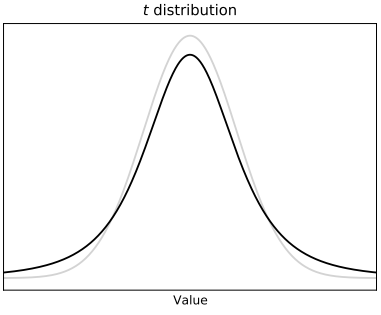
- Hypothesis testing is a mathematical way to express that intuition
- Create a null and an alternative hypothesis
- Alternative hypothesis is what you believe may be true
- Null hypothesis: that your alternative hypothesis is wrong
- The null hypothesis is H_0 and the alternative is H_a or H_1

Hypothesis testing

- Estimate the sampling distribution if the null hypothesis is true
- Compare survey result to sampling distribution
- If it is unlikely that survey result is from sampling distribution, reject the null hypothesis

Sampling distributions

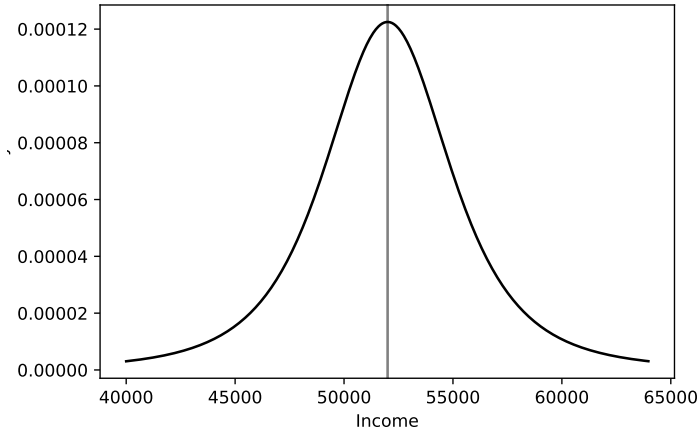




A concrete example

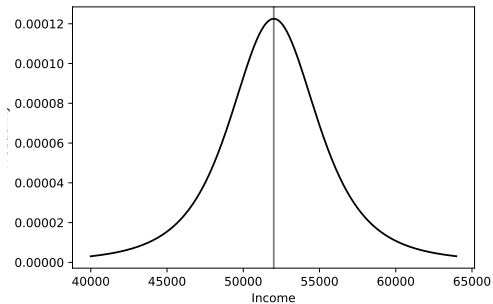
- Suppose you want to know if the average income in Tempe is more than $50,000
- So you survey 100 people
- and get an average of $52,000
- Does this mean the average income is more than $50,000?
- If we did the survey again, would we get a different answer?

A concrete example



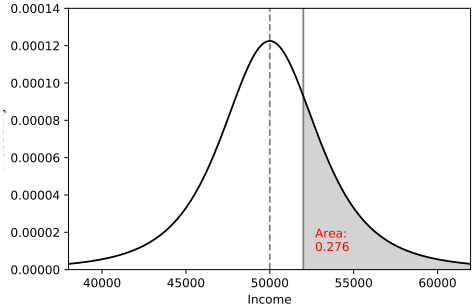
Adding some math

- The shaded area represents the probability of getting a value of $52,000 or more if the null hypothesis is true
- We can integrate over this area to determine this probability, known as the *p*-value
- In this case it is 28%
- How sure are we that the true average is more than $50,000 now?



Adding some math

- A hypothesis test compares the *p*-value with a critical value
- This critical value is often 0.05, but 0.1 and 0.01 are also used
- If the *p*-value is less than the critical value, we reject the null hypothesis and accept the alternative hypothesis
- Can we reject the null hypothesis in this example, with a critical value of 0.05?
- Which critical values are more conservative?



Aside: William Gosset and the Student's *t* distribution

- The *t* distribution was devised by William Gosset
- Gosset published under the pseudonym *Student* because he couldn't publish under his own name at the request of his employer

Aside: William Gosset and the Student's *t* distribution

- The *t* distribution was devised by William Gosset
- Gosset published under the pseudonym *Student* because he couldn't publish under his own name at the request of his employer
- ...Guinness Brewing
- ...which was developing statistics to gain an edge in the beer industry

Hypothesis tests of regression coefficients

- We can put *anything* in the model and get an estimated coefficient
- But how do we know if this coefficient is real or due to the specific sample we're using?

Hypothesis tests of regression coefficients

- We can put *anything* in the model and get an estimated coefficient
- But how do we know if this coefficient is real or due to the specific sample we're using?
- ...with a hypothesis test

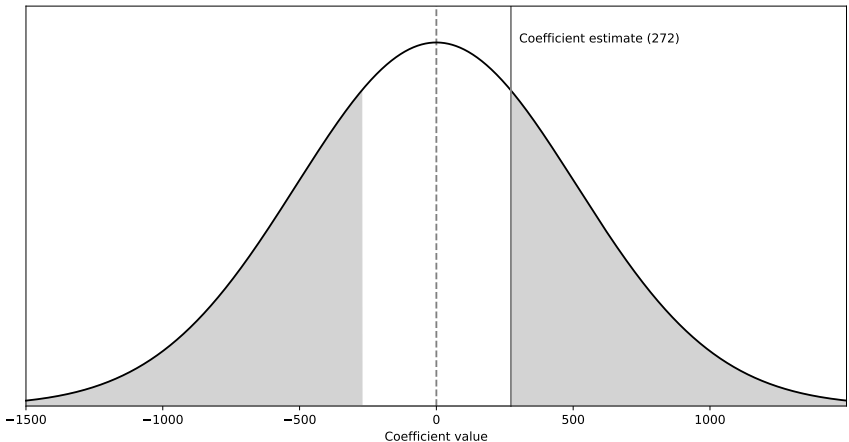
Hypothesis tests of regression coefficients

	Coefficient	Std. err.	*t*-value	*p*-value	95% Conf.	Int.
Constant	-1253	3091	-0.405	0.685	-7341	4834
Income (thousands)	79	17	4.593	0.000	45	112
Number of vehicles	4249	909	4.674	0.000	2458	6039
Day of week	272	513	0.531	0.596	-739	1284

Dependent variable	Annual vehicle miles traveled		
R^2	0.19	Adjusted R^2	0.18
Sample size	250		

Data: 2017 NHTS

Hypothesis tests of regression coefficients



Hypothesis tests of regression coefficients

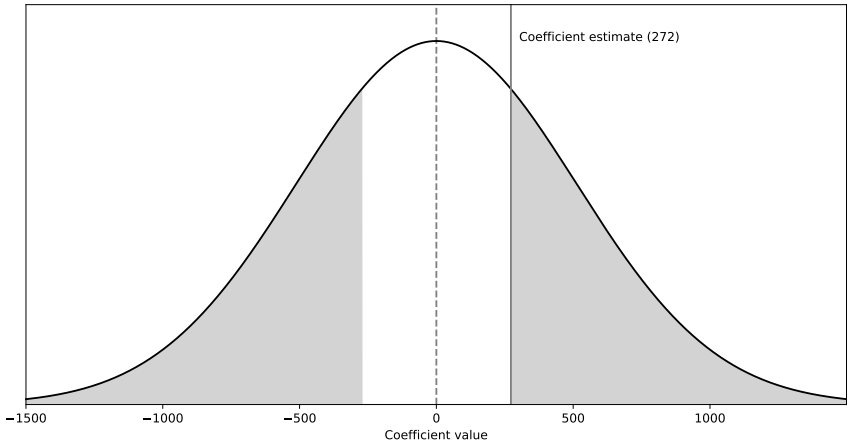
	Coefficient	Std. err.	*t*-value	*p*-value	95% Conf.	Int.
Constant	-1253	3091	-0.405	0.685	-7341	4834
Income (thousands)	79	17	4.593	0.000	45	112
Number of vehicles	4249	909	4.674	0.000	2458	6039
Day of week	272	513	0.531	0.596	-739	1284

Dependent variable	Annual vehicle miles traveled		
R^2	0.19	Adjusted R^2	0.18
Sample size	250		

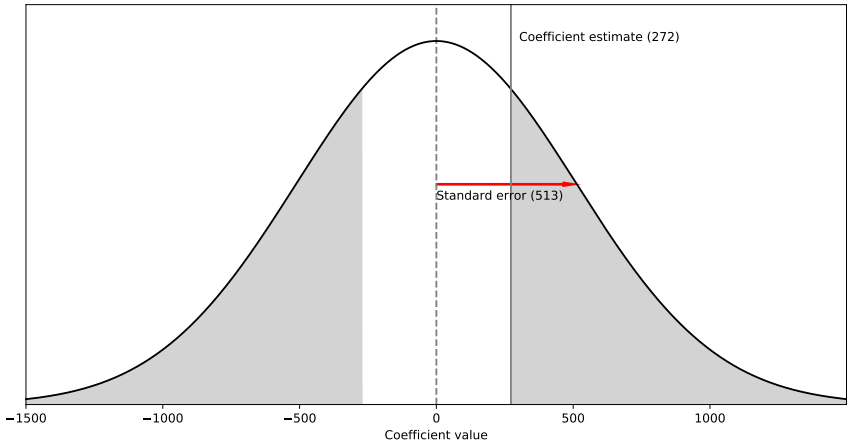
Standard errors, t-scores, and z-scores

- Not every paper reports a *p*-value
- Some report standard errors, *t*-scores, or *z*-scores instead
  - These all tell us the same information—how unlikely the observed coefficient would be if the null hypothesis were true
- Standard error is the spread of the sampling distribution for the hypothesis test
- *t*/*z* score is the number of standard errors between 0 and the coefficient

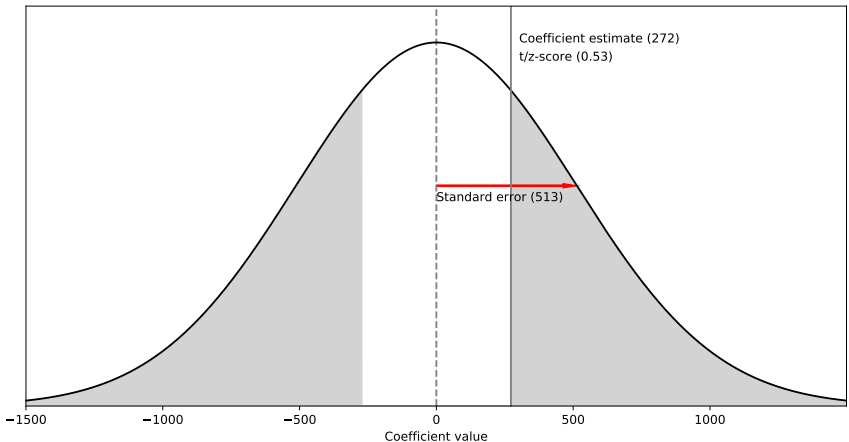
Standard errors, t-scores, and z-scores



Standard errors, t-scores, and z-scores

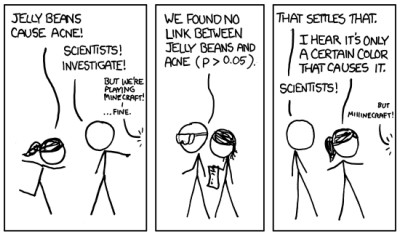


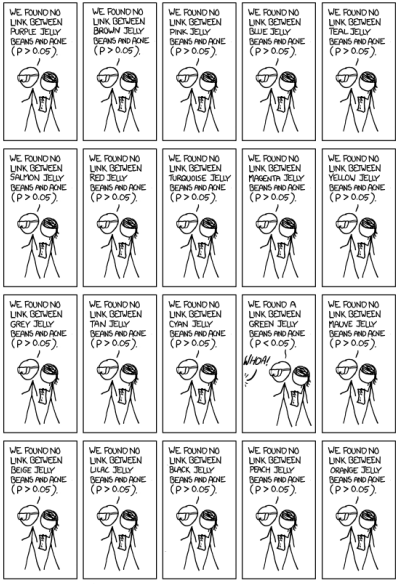
Standard errors, t-scores, and z-scores

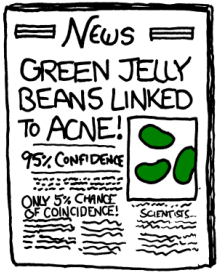


Multiple testing bias

- With a single hypothesis test, there is a small chance that we will reject the null hypothesis when it is true
- With many hypothesis tests, this effect compounds
- This can be used to *p*-hack: testing many different relationships to find one that is significant







Multiple testing bias in regression

- Multiple testing is a particular concern in regression
- ...because there are so many hypothesis tests
- This can result in intentional or unintentional p-hacking
- It is particularly a concern when authors have tested many different model forms

Multiple testing bias in regression

- Multiple testing is a particular concern in regression
- ...because there are so many hypothesis tests
- This can result in intentional or unintentional p-hacking
- It is particularly a concern when authors have tested many different model forms
- Exercise: `https://projects.fivethirtyeight.com/p-hacking/`

Exclusion of insignificant variables and stepwise selection

- Two common methods raise particular concern about multiple testing
- Many papers exclude insignificant variables from their final models
  - Can bias the remaining coefficients because important control variables may be excluded
- Stepwise selection selects variable for the model one at a time based on their significance
  - "[S]tepwise methods tend to capitalize outrageously on sampling error" (Thompson 1995)
  - i.e. stepwise selection is likely to include many variables that are significant by chance
- Both of these techniques were considered state of the art at some point, and their shortcomings have become apparent more recently

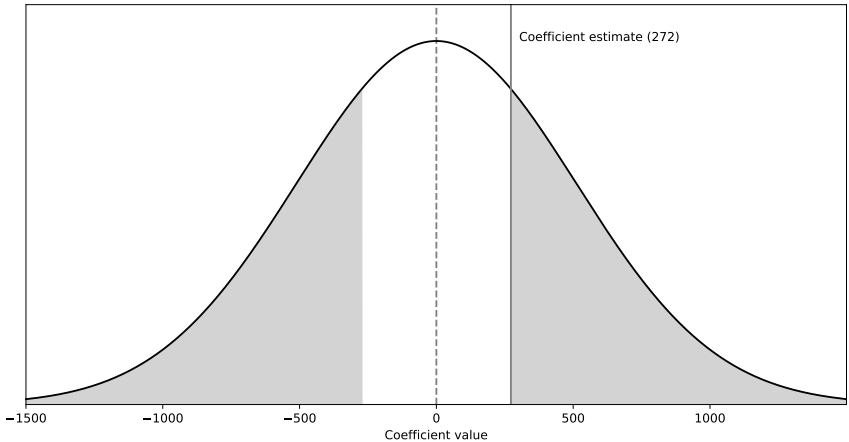
Non-sampling error

- Non-sampling error occurs when the sample is not representative
- Surveying graduate students at ASU would not give you a good idea of the average income in Tempe
- Can also occur when people do not respond to your survey
- Non-sampling error is part of the concern about the Census citizenship question

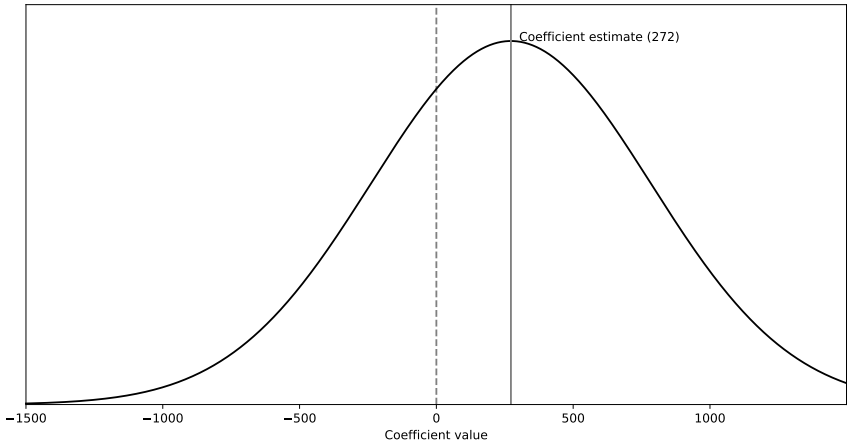
Confidence intervals

- Alternative to hypothesis tests
- A confidence interval is a range around the coefficient that is x% likely to contain the true coefficient
  - Usually 95%
- Does not require a null hypothesis

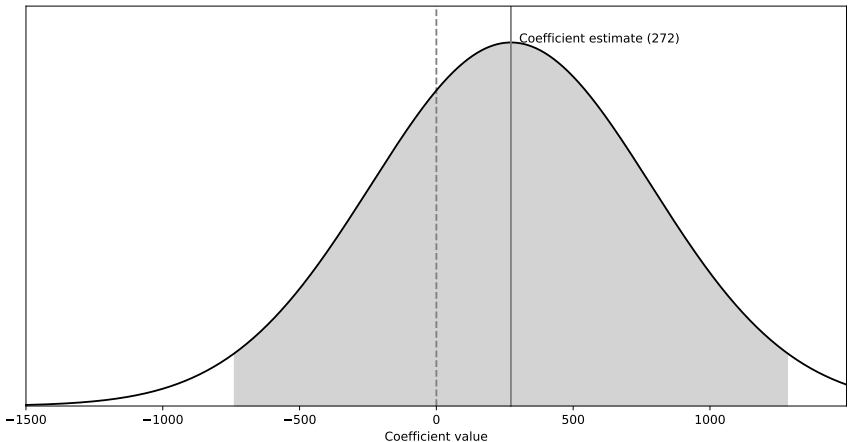
Confidence intervals



Confidence intervals



Confidence intervals



Discrete choice models

- Linear regression is great, but how do we predict a discrete choice?
- For example, the choice of car or bus?

Applications of discrete choice models

- Mode choice
- Destination choice
- Route choice
- Residential location choice
- Vehicle choice

Measurement levels

- Nominal:

Measurement levels

- Nominal: categories with no order (e.g. colors)

Measurement levels

- Nominal: categories with no order (e.g. colors)
- Ordinal:

Measurement levels

- Nominal: categories with no order (e.g. colors)
- Ordinal: ordered categories with no information about distance between them (e.g. much less, less, same, more, much more)

Measurement levels

- Nominal: categories with no order (e.g. colors)
- Ordinal: ordered categories with no information about distance between them (e.g. much less, less, same, more, much more)
- Interval:

Measurement levels

- Nominal: categories with no order (e.g. colors)
- Ordinal: ordered categories with no information about distance between them (e.g. much less, less, same, more, much more)
- Interval: distances between items are meaningful, ratios are not (no meaningful zero; e.g. temperature)

Measurement levels

- Nominal: categories with no order (e.g. colors)
- Ordinal: ordered categories with no information about distance between them (e.g. much less, less, same, more, much more)
- Interval: distances between items are meaningful, ratios are not (no meaningful zero; e.g. temperature)
- Ratio:

Measurement levels

- Nominal: categories with no order (e.g. colors)
- Ordinal: ordered categories with no information about distance between them (e.g. much less, less, same, more, much more)
- Interval: distances between items are meaningful, ratios are not (no meaningful zero; e.g. temperature)
- Ratio: distances between items and ratios are meaningful (e.g. income)

Measurement levels

- Nominal: categories with no order (e.g. colors)
- Ordinal: ordered categories with no information about distance between them (e.g. much less, less, same, more, much more)
- Interval: distances between items are meaningful, ratios are not (no meaningful zero; e.g. temperature)
- Ratio: distances between items and ratios are meaningful (e.g. income)

Binary outcomes

- Simpler case is a binary outcome
- To bike or not to bike
- To carpool or not to carpool
- ...and so on

One option: the linear probability model

- Uses linear regression to predict binary outcome
- Code outcomes as 0 or 1
- Interpret coefficients as change in probability

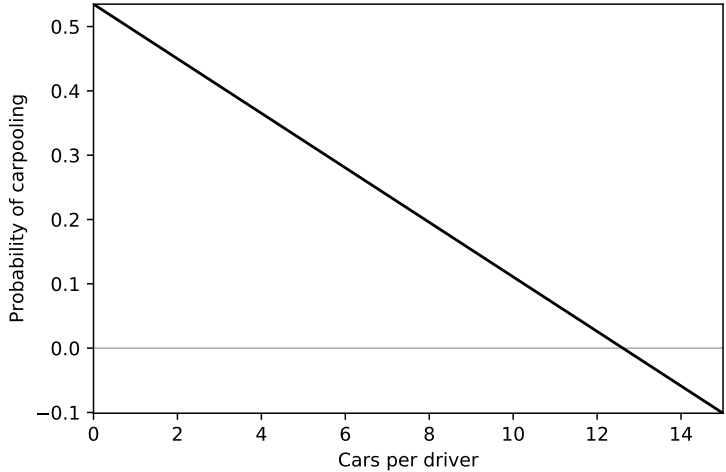
One option: the linear probability model

	Coefficient	Std. err.	*t*-value	*p*-value	95% Conf.	Int.
Constant	0.535	0.012	43.622	0.0	0.511	0.559
Cars per driver	-0.042	0.009	-4.540	0.0	-0.061	-0.024

Dependent variable	Carpool on trip		
R^2	0.00	Adjusted R^2	0.00
Sample size	10000		

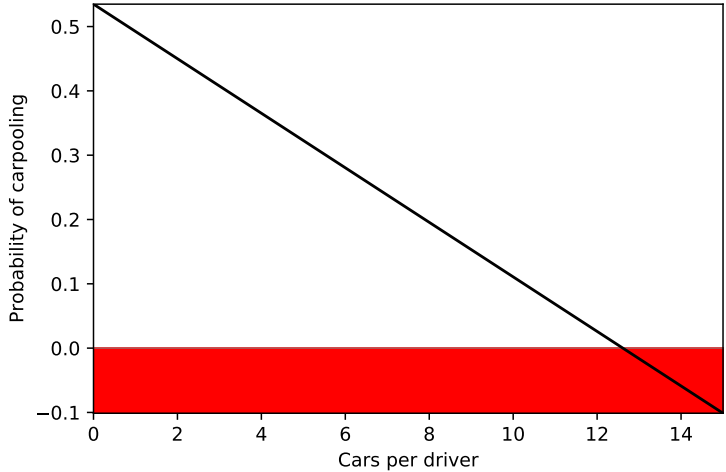
Data: 2017 NHTS

One option: the linear probability model



Data: 2017 NHTS

One option: the linear probability model

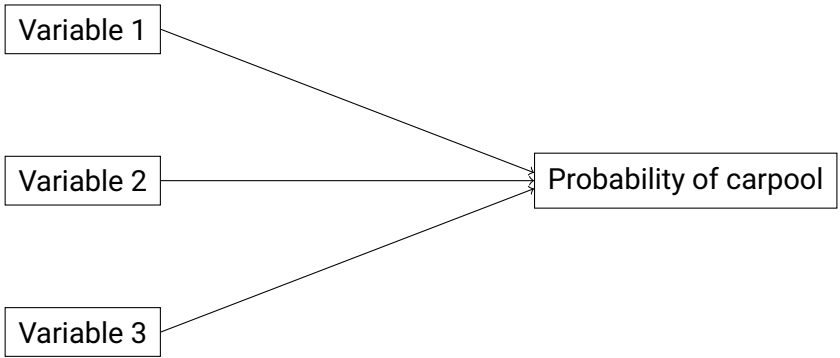


Data: 2017 NHTS

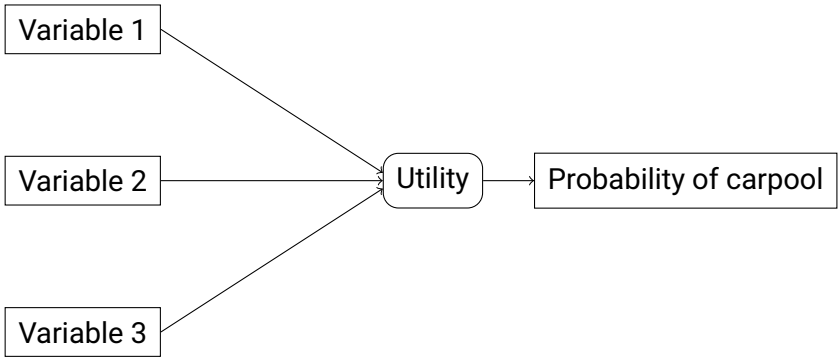
Alternate option: random utility models

- Assume that the probability of choosing an option is a function of utility
- Utility is a concept from economics, basically means value
- Utility is usually a linear function—like linear regression!
- A transformation is applied to utility to model probability
- Utility is a latent variable—we don't measure it, we infer it

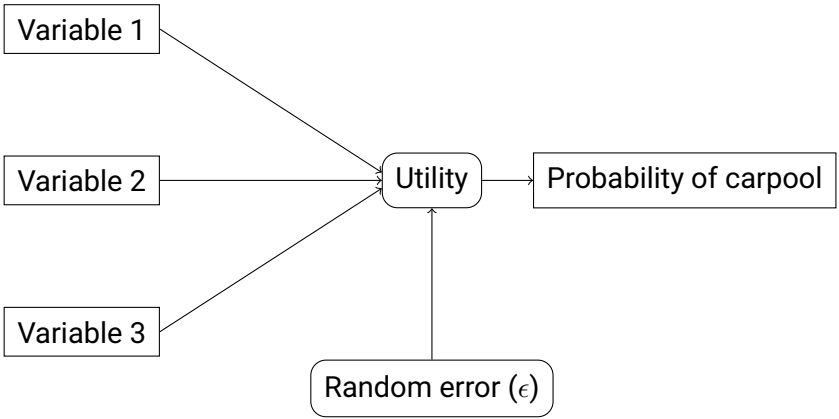
Alternate option: random utility models



Alternate option: random utility models

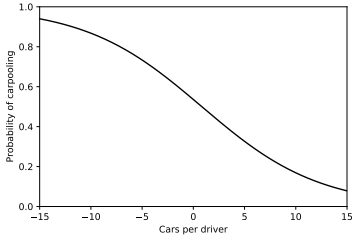


Alternate option: random utility models



The logit model

- Transforms probability to utility using a logit function
- $U = \log(\frac{p}{1-p})$



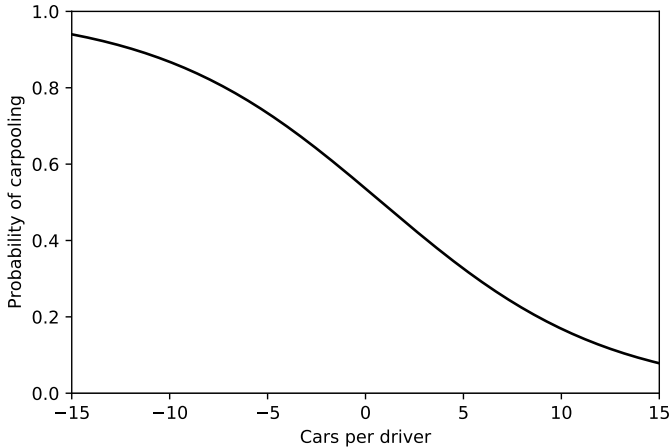
The logit model

	Coefficient	Std. err.	*t*-value	*p*-value	95% Conf.	Int.
Constant	0.144	0.050	2.865	0.004	0.045	0.243
Cars per driver	-0.174	0.039	-4.501	0.000	-0.250	-0.098

Dependent variable	Carpool on trip
Pseudo-R^2	0.00
Sample size	10000

Data: 2017 NHTS

The logit model



Data: 2017 NHTS

Interpreting the logit model: odds ratios

- Coefficients in a logistic regression are hard to interpret
- Often presented as an odds ratio, exponentiation of coefficient
- ...the ratio of the odds of an event occurring when the variable increases by one
- Odds is the probability of the event divided by the probability of not the event
- For example, if the event occurs once and does not occur four times, the odds are one-to-four or $\frac{1}{4}$

Interpreting the logit model: odds ratios

- Coefficients in a logistic regression are hard to interpret
- Often presented as an odds ratio, exponentiation of coefficient
- ...the ratio of the odds of an event occurring when the variable increases by one
- Odds is the probability of the event divided by the probability of not the event
- For example, if the event occurs once and does not occur four times, the odds are one-to-four or $\frac{1}{4}$
- ...what is the probability in this case?

Interpreting the logit model: odds ratios

- Odds ratio greater than one, increase in outcome probability
- Odds ratio = 1, no relationship
- Odds ratio less than one, decrease in outcome probability
- The odds ratio is *not* the ratio of probabilities, but is similar when the probability of the outcome is small

Logit model: example

Variable	Odds Ratio	*p*-Value	95% Conf Int.	
			Low	High
Household size (base: household size 1)				
Household size 2	1.380 ***	0.001	1.146	1.661
Household size 3	0.913	0.492	0.704	1.184
Household size 4+	0.687 **	0.008	0.521	0.905
Worker	1.412 ***	0.000	1.344	1.483
Homeowner	0.752 ***	0.000	0.713	0.793
Male	1.135 ***	0.000	1.100	1.171
Presence of children				
Children 0–12	1.000	0.998	0.928	1.078
Children 13–17	0.980	0.642	0.901	1.066
Income, household size 1 (base: $24,999 or less)				
$25,000–$50,000	1.270	0.110	0.947	1.702
$50,000–$100,000	2.083 ***	0.000	1.630	2.662
More than $100,000	3.500 ***	0.000	2.745	4.463
Income, household size 2 (base: $24,999 or less)				
$25,000–$50,000	0.920	0.550	0.701	1.208
$50,000–$100,000	1.456 **	0.001	1.156	1.835
More than $100,000	2.867 ***	0.000	2.301	3.573
Income, household size 3 (base: $24,999 or less)				
$25,000–$50,000	0.956	0.778	0.700	1.306
$50,000–$100,000	1.190	0.204	0.909	1.558
More than $100,000	2.602 ***	0.000	2.021	3.349
Income, household size 4+ (base: $24,999 or less)				
$25,000–$50,000	0.832	0.255	0.607	1.141
$50,000–$100,000	1.135	0.350	0.870	1.481
More than $100,000	2.647 ***	0.000	2.060	3.401

Conway, Salon, and King (2018)

Logit model: example

Neighborhood preference, walkability, walking, and obesity

	Any walk trips		Any non-discretionary walk trips		Any discretionary walk trips		Obesity	
	Est. odds	95% Confidence interval	Est. odds	95% Confidence interval	Est. odds	95% Confidence interval	Est. odds	95% Confidence interval
Age	0.97	0.96–0.99	0.97	0.96–0.99	0.98	0.96–1.00		
Race[a]							0.57	0.38–0.87
Gender[b]							0.64	0.47–0.87
Household income[c]							0.94	0.88–1.00
Total vehicles in the household	0.68	0.56–0.83	0.67	0.55–0.82				
# Licensed drivers/ vehicle			0.67	0.44–1.03	0.56	0.29–1.07		
Neighborhood preference quartile[d]								
2nd quartile	1.67	0.89–3.12	1.39	0.71–2.74	1.97	0.72–5.39	0.88	0.60–1.30
3rd quartile	3.23[e]	1.93–5.17	3.49[e]	2.02–5.74	2.14	0.79–5.79	0.68[e]	0.46–0.96
4th quartile	4.81[e]	2.98–7.32	5.05[e]	3.02–7.95	4.45[e]	1.79–10.34	0.38[e]	0.23–0.61
Walkability quartile								
2nd quartile	0.72	0.39–1.31	0.80	0.42–1.52	0.60	0.20–1.76	0.77	0.50–1.19
3rd quartile	1.11	0.64–1.91	1.20	0.67–2.16	1.28	0.52–3.11	0.90	0.57–1.40
4th quartile	1.62	0.95–2.76	1.72	0.97–3.03	2.14	0.91–5.00	0.89	0.55–1.42

Frank et al. (2007)

Marginal effects

- An alternate measure of the association between an independent and dependent variable is the marginal effect
- This is the average amount the dependent variable changes for a unit change in the input variable
- In linear regression, this is the same as the coefficient
- But in other types of regression it isn't
- No need to record marginal effects unless coefficients/odds ratios are not reported

Other transformations: probit, tobit, etc.

- There are a lot of other similar models with different transformation functions
- These represent different assumptions for the distribution of the random error term ϵ
- Logit is by far the most common as the math is easier

More than two outcomes

- The logit models we've seen so far can only model two outcomes (e.g. carpool/not carpool)
- But what if we have more outcomes (e.g. bike/drive/walk)?
- In these cases we use a multinomial logit model

The multinomial logit model

Income

Travel time by walk

Travel time by transit

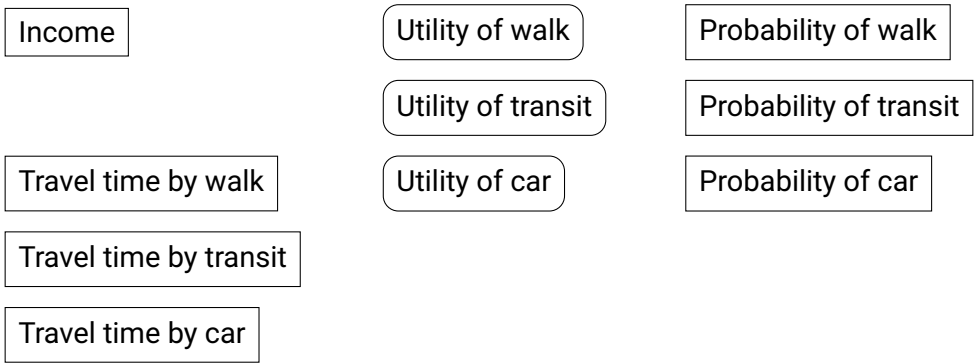
Travel time by car

Probability of walk

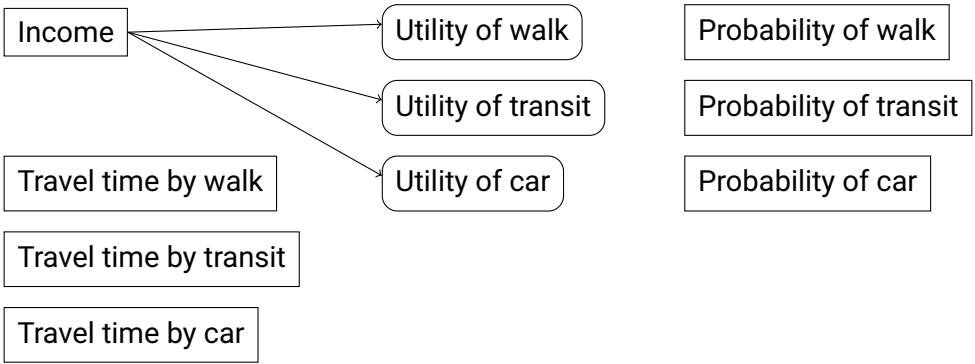
Probability of transit

Probability of car

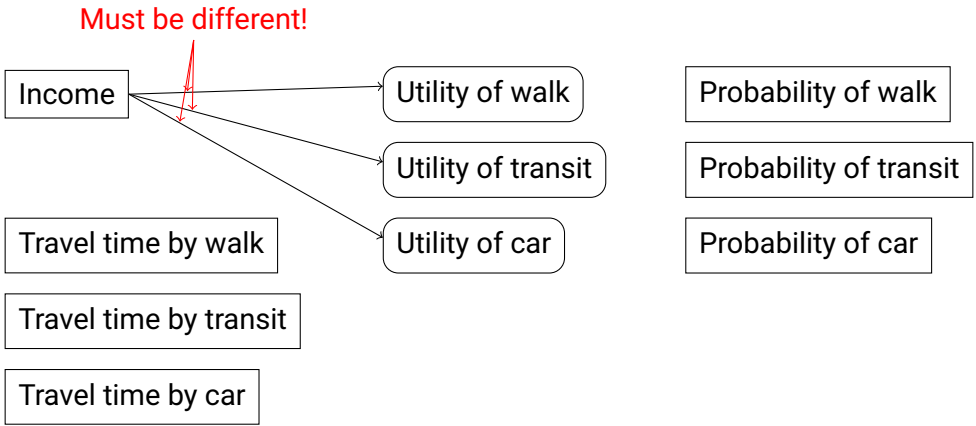
The multinomial logit model



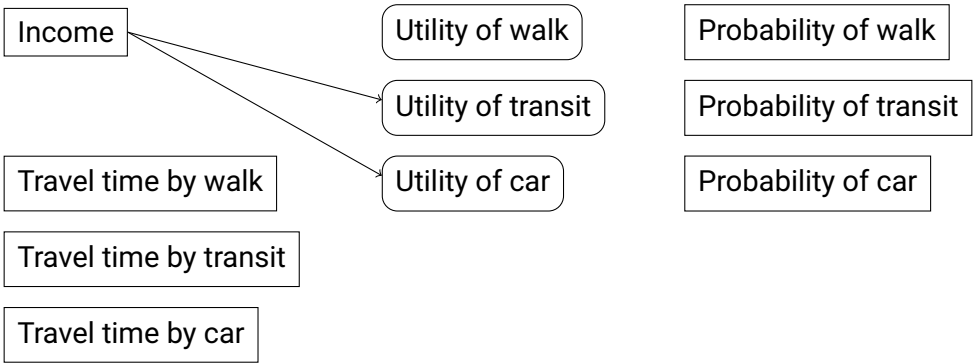
The multinomial logit model



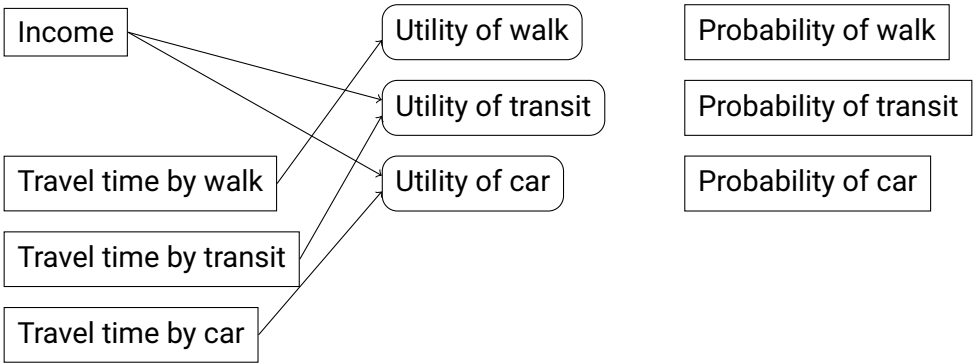
The multinomial logit model



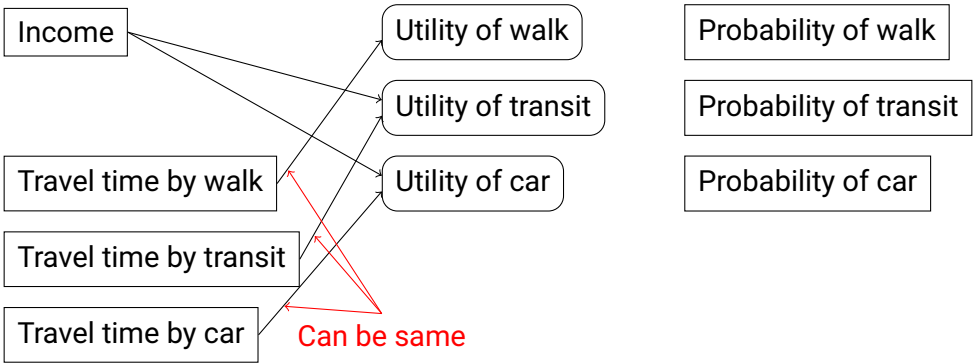
The multinomial logit model



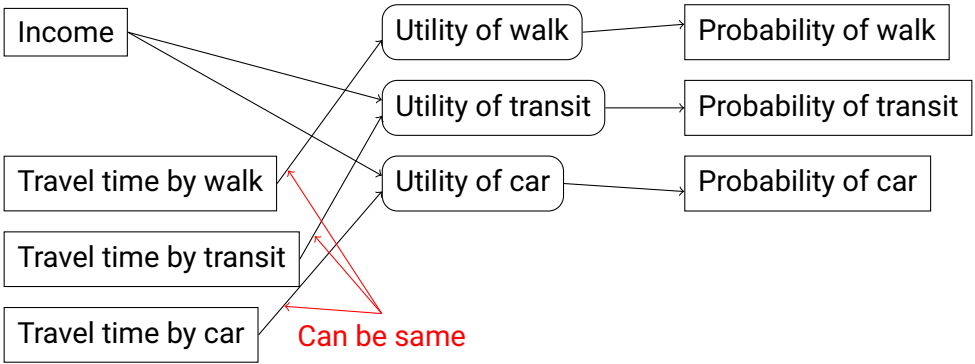
The multinomial logit model



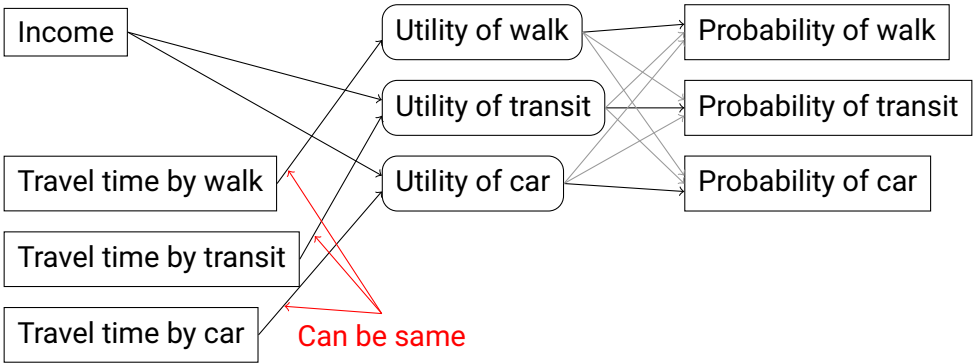
The multinomial logit model



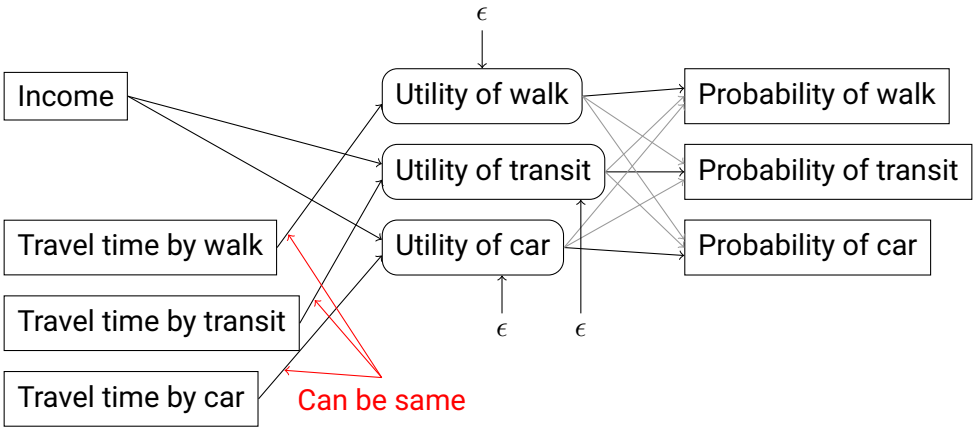
The multinomial logit model



The multinomial logit model



The multinomial logit model



Multinomial logit model: mathematical form

$$
\begin{aligned}
U_{walk} &= \alpha_{walk} + \beta_{income,walk} x_{income} + \beta_{time} x_{time,walk} + \epsilon_{walk} \\
U_{transit} &= \alpha_{transit} + \beta_{income,transit} x_{income} + \beta_{time} x_{time,transit} + \epsilon_{transit} \\
U_{car} &= \alpha_{car} + \beta_{income,car} x_{income} + \beta_{time} x_{time,car} + \epsilon_{car}
\end{aligned}
$$

Multinomial logit model: mathematical form

$$
\begin{aligned}
U_{walk} &= & & \beta_{time} x_{time,walk} & + \epsilon_{walk} \\
U_{transit} &= \alpha_{transit} & + \beta_{income,transit} x_{income} & + \beta_{time} x_{time,transit} & + \epsilon_{transit} \\
U_{car} &= \alpha_{car} & + \beta_{income,car} x_{income} & + \beta_{time} x_{time,car} & + \epsilon_{car}
\end{aligned}
$$

Multinomial logit model: mathematical form

$$
\begin{aligned}
U_{walk} &= && && \beta_{time} x_{time,walk} && + \epsilon_{walk} \\
U_{transit} &= \alpha_{transit} && + \beta_{income,transit} x_{income} && + \beta_{time} x_{time,transit} && + \epsilon_{transit} \\
U_{car} &= \alpha_{car} && + \beta_{income,car} x_{income} && + \beta_{time} x_{time,car} && + \epsilon_{car}
\end{aligned}
$$

$$
p_{walk} = \frac{e^{U_{walk}}}{\sum_{m \in \{walk, transit, car\}} e^{U_m}}
$$

Multinomial logit model: results

	Value	Std err	t-test	p-value
Car				
Alternative specific constant	1.58	0.37	4.31	0.0
Income (thousands CHF/month)	-0.15	0.04	-3.42	0.0
Transit				
Alternative specific constant	1.36	0.37	3.65	0.0
Income (thousands CHF/month)	-0.16	0.05	-3.46	0.0
Travel time	-0.01	0.00	-11.77	0.0

Data: Bierlaire (2018)

Multinomial logit model: example

TABLE 4 Model Estimation Results

Model	Variables	Car/Van Pool		Transit		Non-motor	
		β-coef	t-ratio	β-coef	t-ratio	β-coef	t-ratio
Model 1: Model including only demographic and socio-economic variables (BD)	Constant	4.772	3.77	5.904	4.62	4.787	3.01
	Age of commuter	-0.011	-1.41	0.010	1.21	-0.028	-1.92
	Household Income (<$7500) dummy	-0.595	-2.18	-0.266	-1.09	0.811	2.18
	Lifecycle (Couple with children) dummy	-0.020	-0.12	0.050	0.27	-0.899	-2.06
	Number of vehicles per adult	-1.210	-5.37	-1.782	-8.54	-1.548	-4.88
	Occupation (Low paid blue collar) dummy	0.441	2.16	-0.132	-0.53	-0.066	-0.13
	Number of adults (+18)	0.095	0.65	-0.392	-2.48	-0.507	-1.87
	Employed outside home	-3.954	-3.89	-3.973	-3.86	-3.077	-2.86

Kuppam, Pendyala, and Rahman (1999)

Factor analysis

- Factor analysis is a way to reduce many variables into a smaller number
- For instance, survey questions that are likely to be correlated

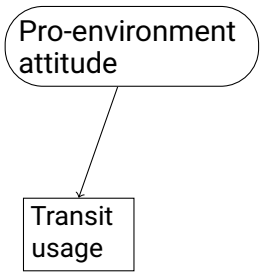
Factor analysis: Likert scales

Please rate the amount you agree or disagree with the following statements:

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Environmental protection costs too much	☐	☐	☐	☐	☐
Environmental protection is good for California's economy	☐	☐	☐	☐	☐
Environmentalism hurts minority and small businesses	☐	☐	☐	☐	☐
I am not comfortable riding with strangers	☐	☐	☐	☐	☐
Stricter vehicle smog control laws should be introduced and enforced	☐	☐	☐	☐	☐
Whoever causes environmental damage should repair the damage	☐	☐	☐	☐	☐
Vehicle emissions increase the need for health care	☐	☐	☐	☐	☐
I feel that I am wasting time when I have to wait	☐	☐	☐	☐	☐
We should provide incentives to people who use electric . . . vehicles	☐	☐	☐	☐	☐
We should raise the price of gasoline to reduce congestion and air pollution	☐	☐	☐	☐	☐

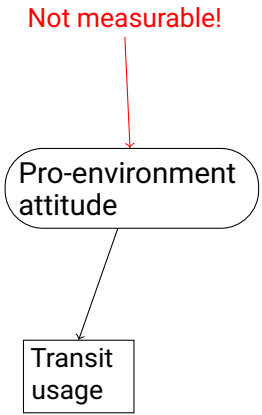
Based on Kitamura, Mokhtarian, and Laidet (1997)

Factor analysis



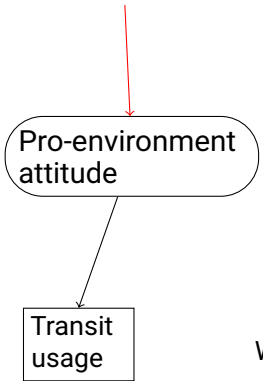
Based on Kitamura, Mokhtarian, and Laidet (1997)

Factor analysis



Based on Kitamura, Mokhtarian, and Laidet (1997)

Factor analysis



Environmental protection costs too much

Environmental protection is good for California's economy

Environmentalism hurts minority and small businesses

I am not comfortable riding with strangers

Stricter vehicle smog control laws should introduced and enforced

Whoever causes environmental damage should repair the damage

Vehicle emissions increase the need for health care

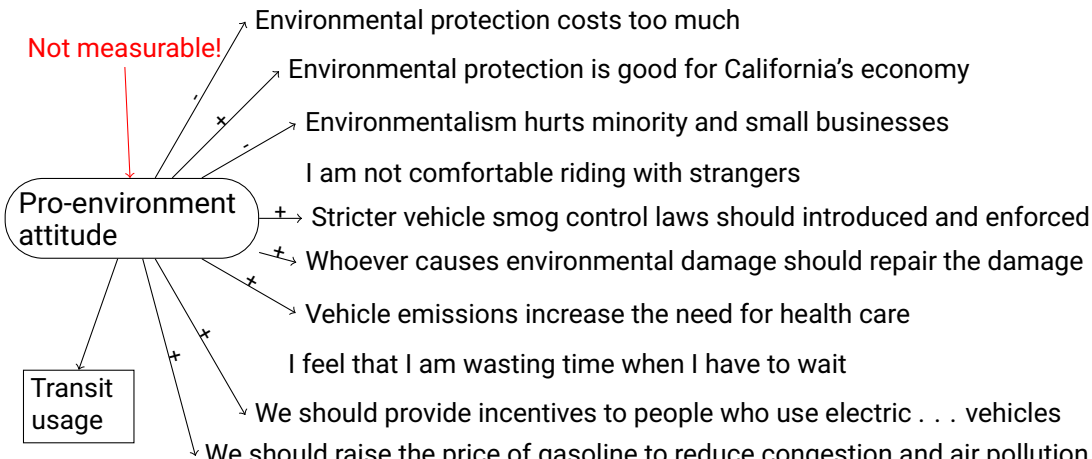
I feel that I am wasting time when I have to wait

We should provide incentives to people who use electric . . . vehicles

We should raise the price of gasoline to reduce congestion and air pollution

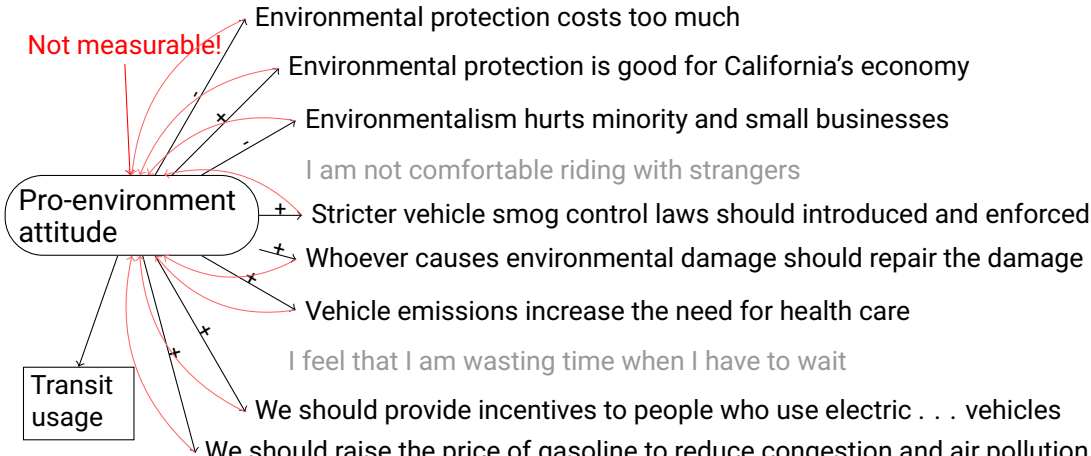
Based on Kitamura, Mokhtarian, and Laidet (1997)

Factor analysis



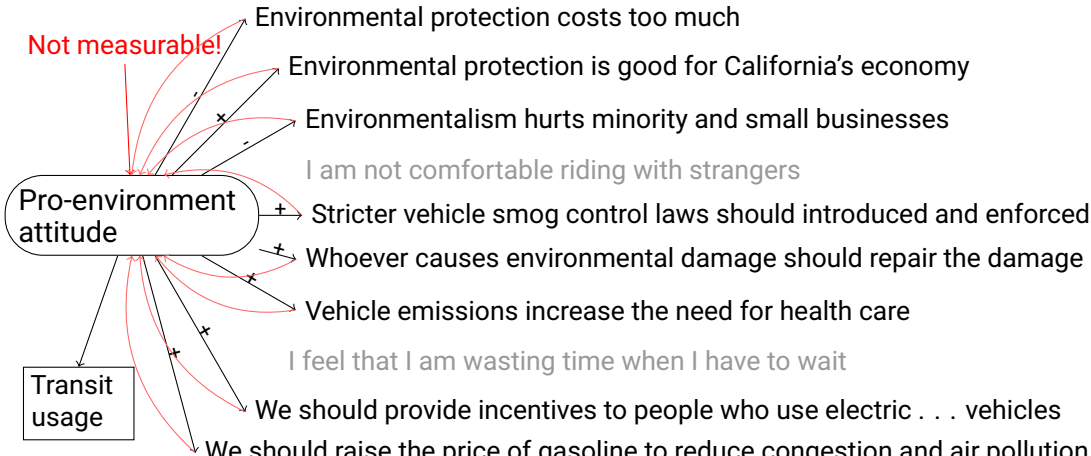
Based on Kitamura, Mokhtarian, and Laidet (1997)

Factor analysis



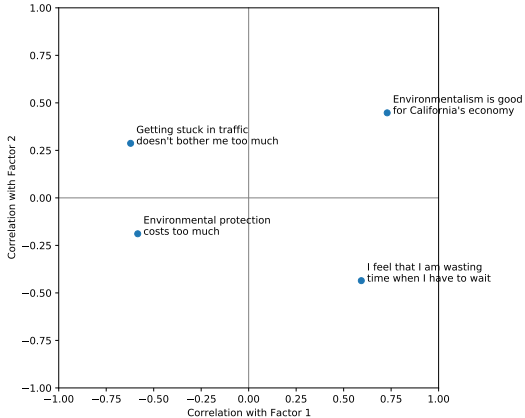
Based on Kitamura, Mokhtarian, and Laidet (1997)

Factor analysis



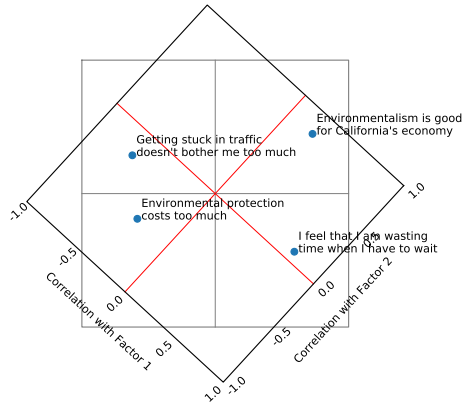
Based on Kitamura, Mokhtarian, and Laidet (1997)

Factor analysis: unrotated factors



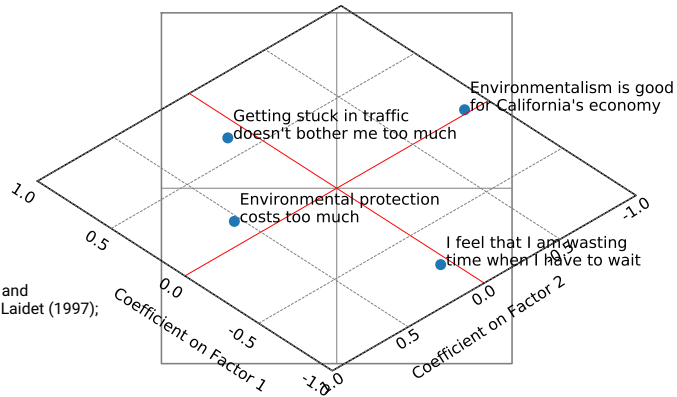
Adapted from Kline (1994) and Kitamura, Mokhtarian, and Laidet (1997); synthetic data.

Factor analysis: orthogonal rotation



Adapted from Kline (1994) and Kitamura, Mokhtarian, and Laidet (1997); synthetic data.

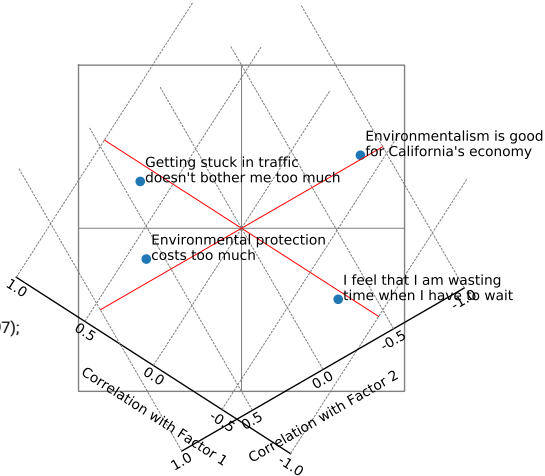
Factor analysis: oblique rotation, pattern matrix (regression coefficients)



Adapted from Kline (1994) and Kitamura, Mokhtarian, and Laidet (1997);

synthetic data.

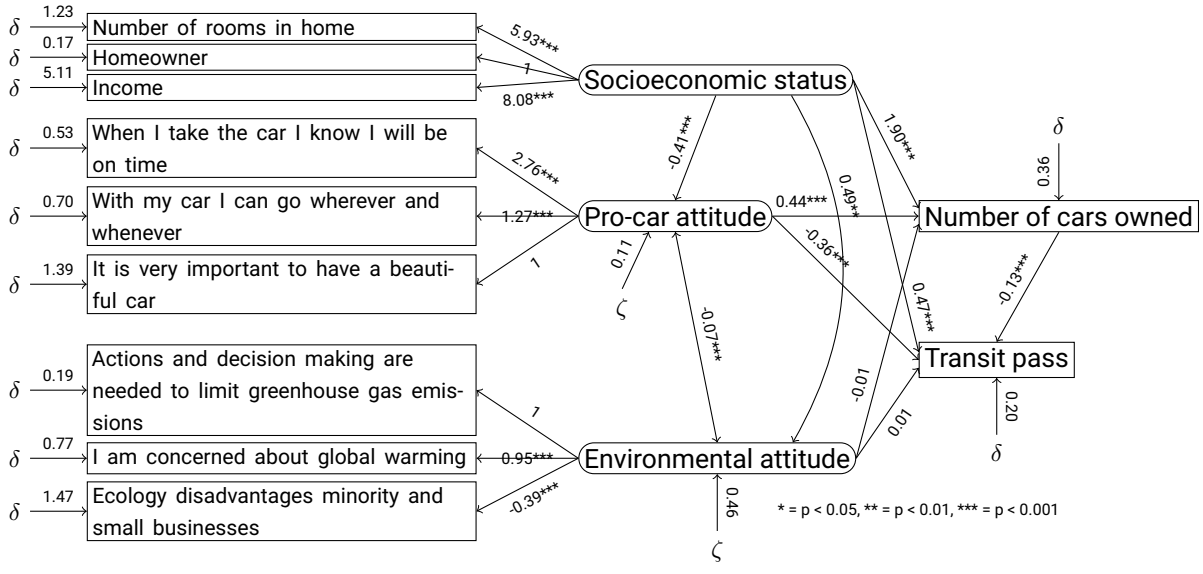
Factor analysis: oblique rotation, structure matrix (correlations)



Adapted from Kline (1994) and Kitamura, Mokhtarian, and Laidet (1997);

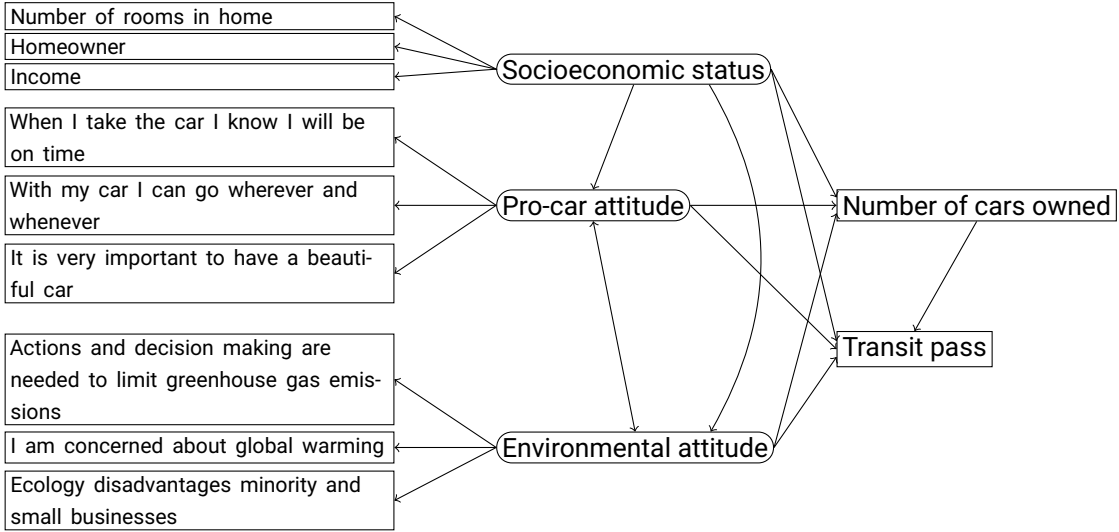
synthetic data.

Structural equation models



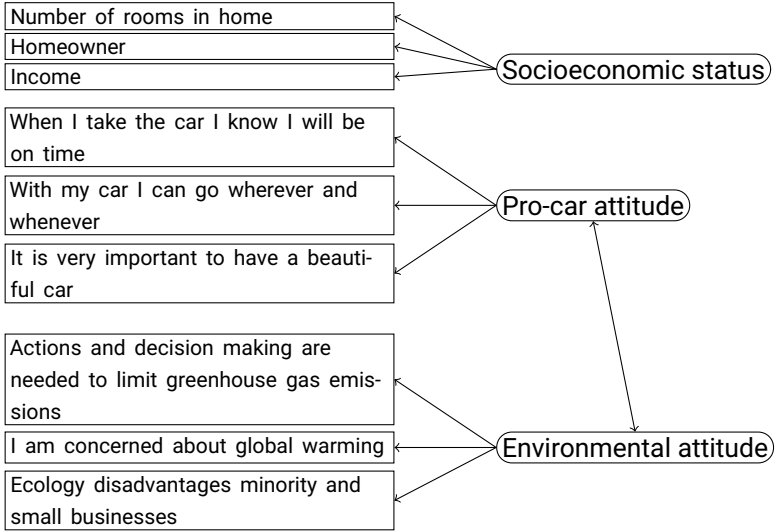
Data: Bierlaire (2018)

Structural equation models



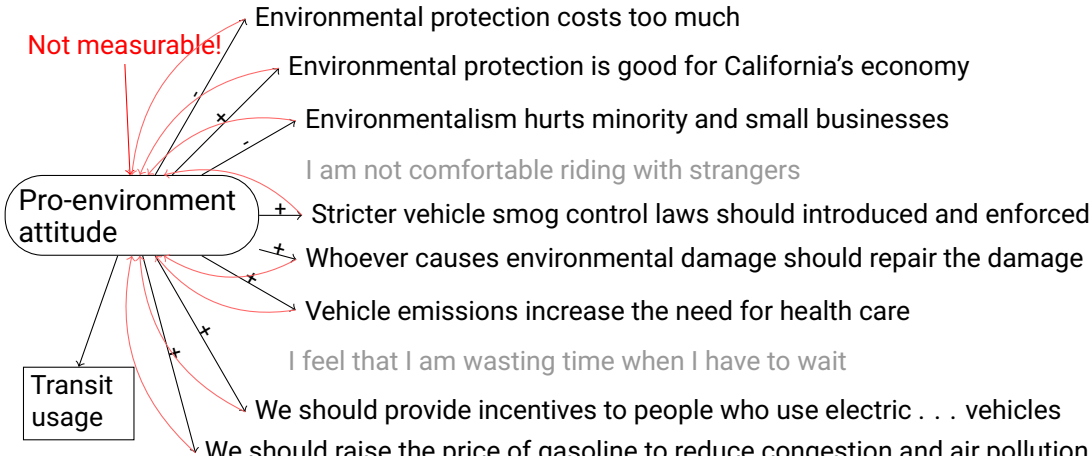
Data: Bierlaire (2018)

Structural equation models



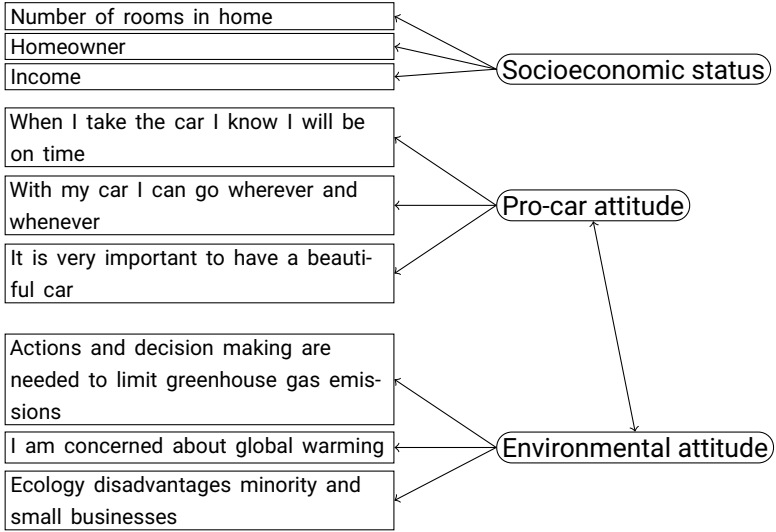
Data: Bierlaire (2018)

Structural equation models



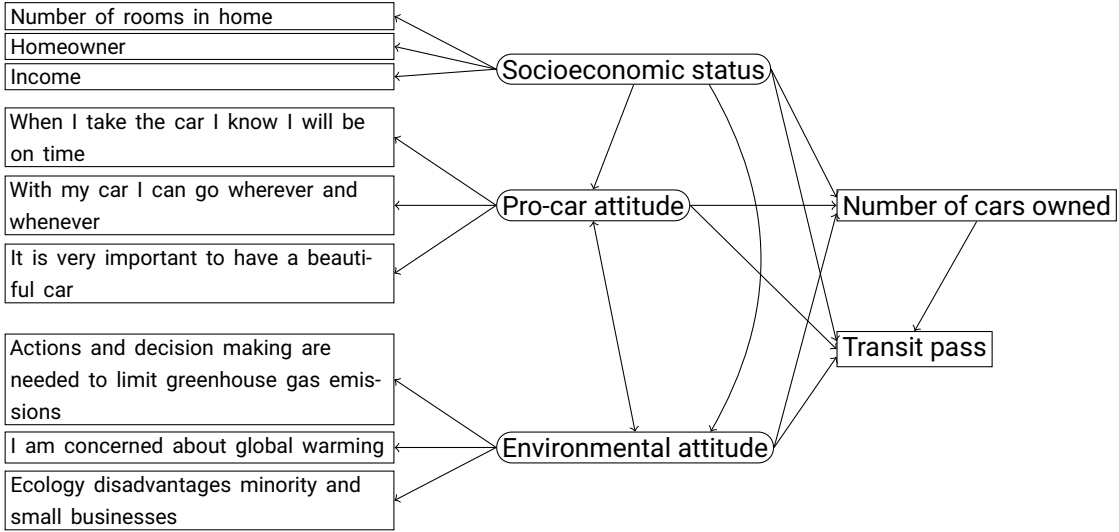
Based on Kitamura, Mokhtarian, and Laidet (1997)

Structural equation models



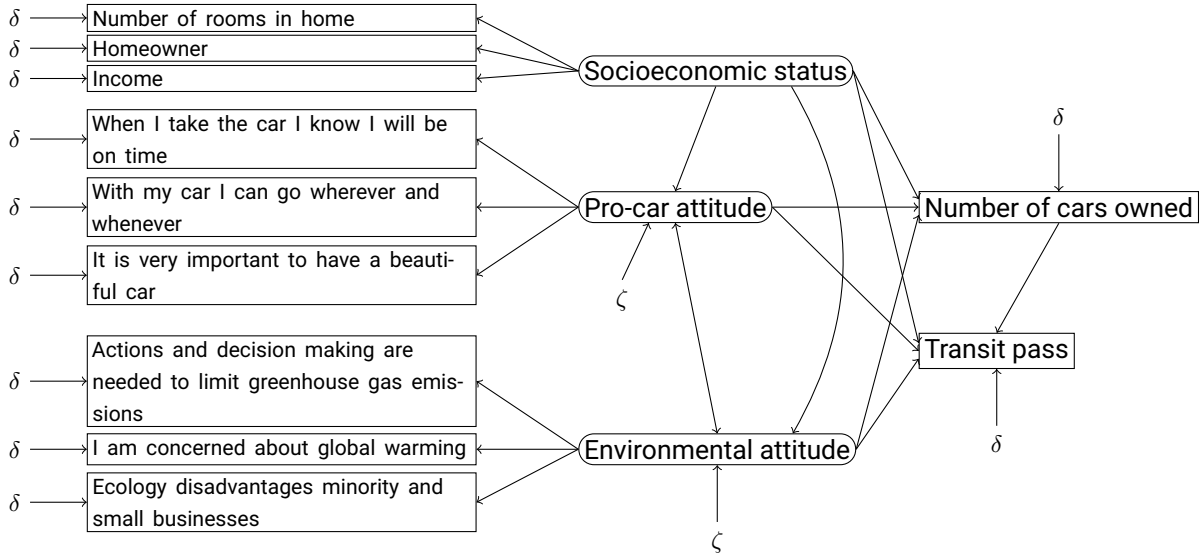
Data: Bierlaire (2018)

Structural equation models



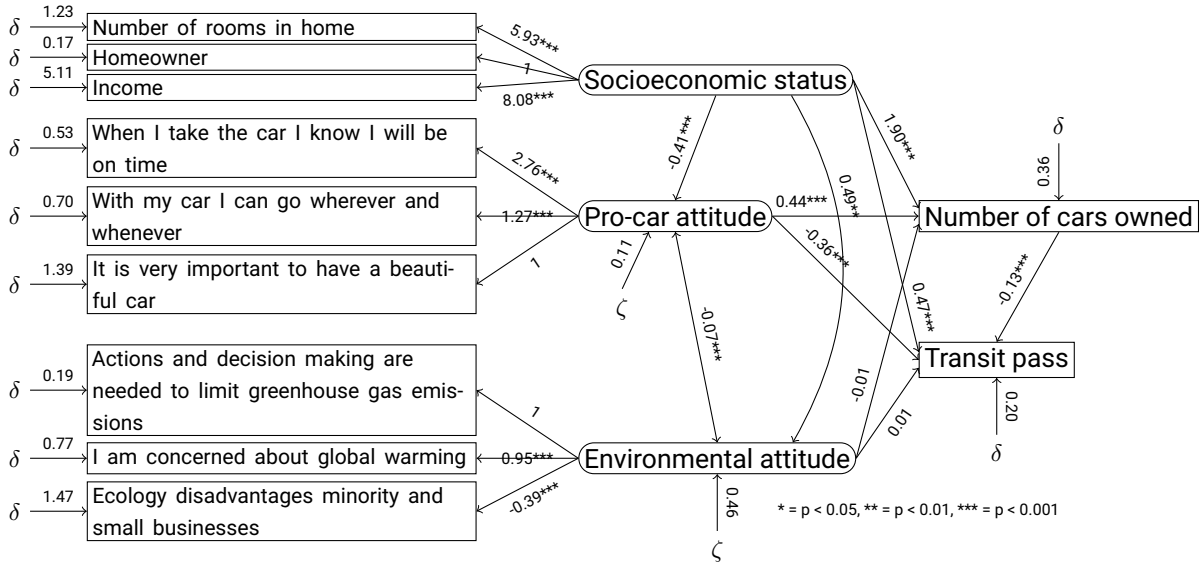
Data: Bierlaire (2018)

Structural equation models



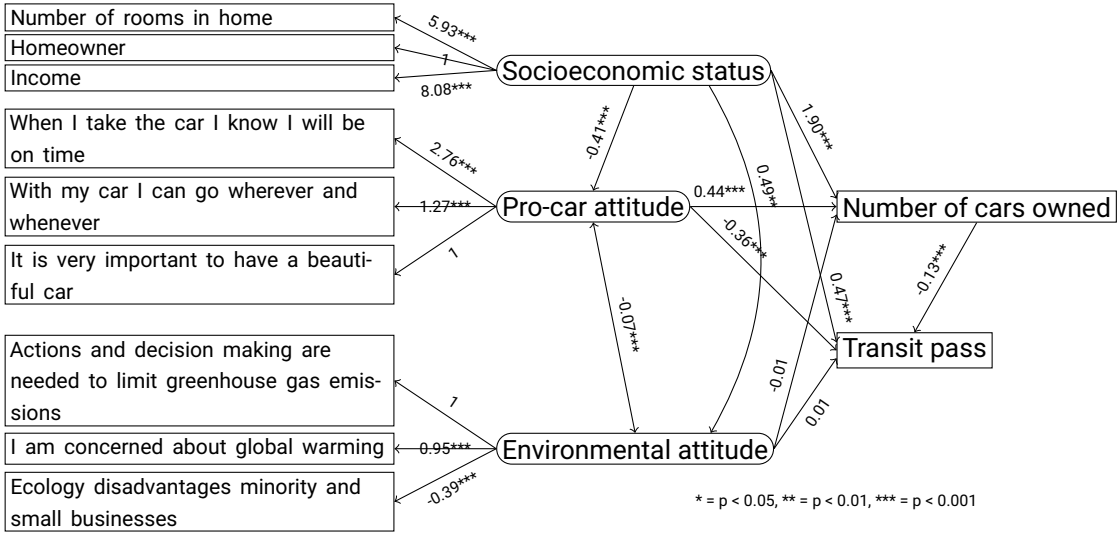
Data: Bierlaire (2018)

Structural equation models



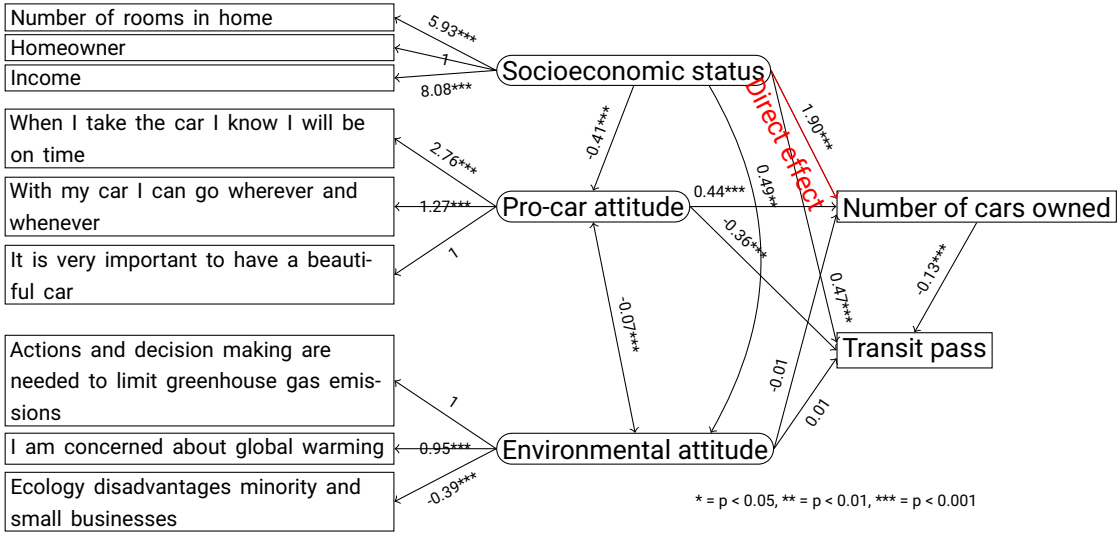
Data: Bierlaire (2018)

Structural equation models



Data: Bierlaire (2018)

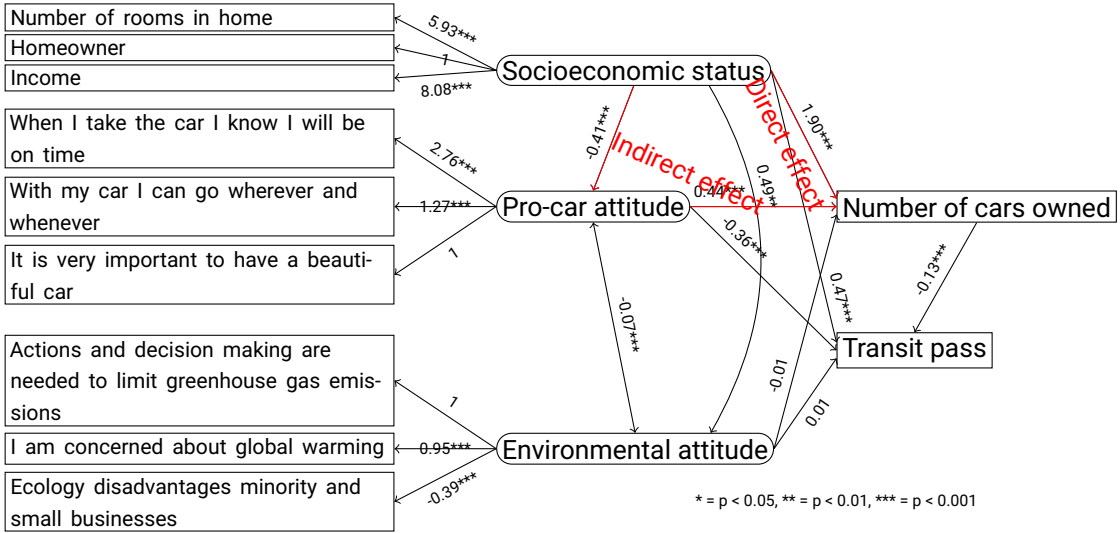
Structural equation models



* = p < 0.05, ** = p < 0.01, *** = p < 0.001

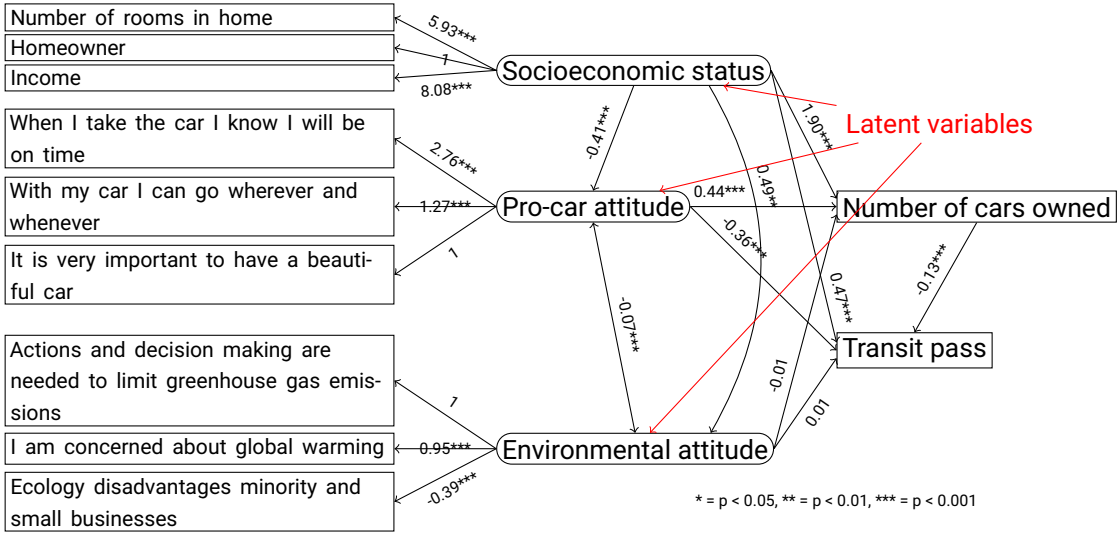
Data: Bierlaire (2018)

Structural equation models



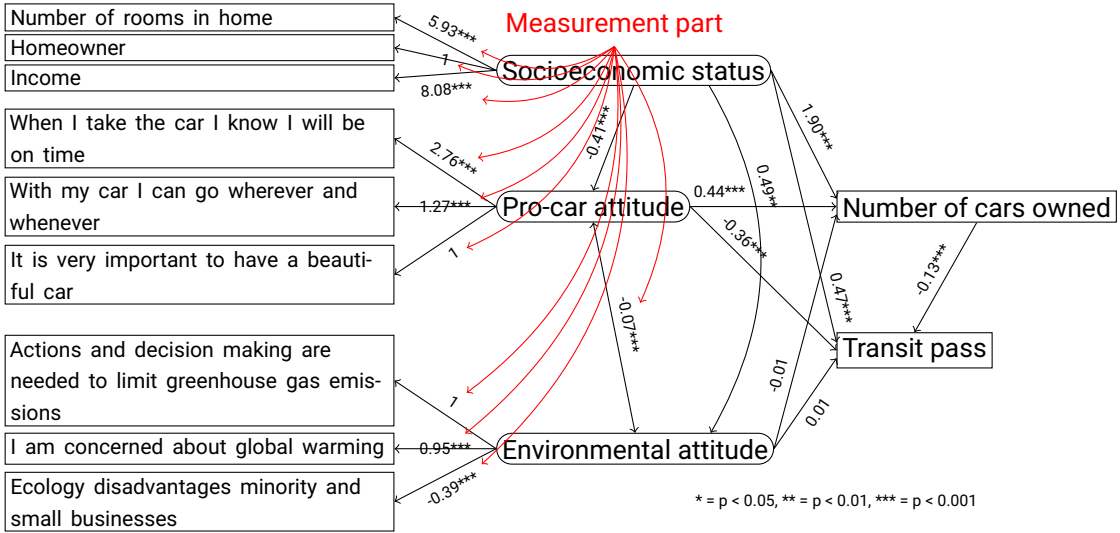
Data: Bierlaire (2018)

Structural equation models



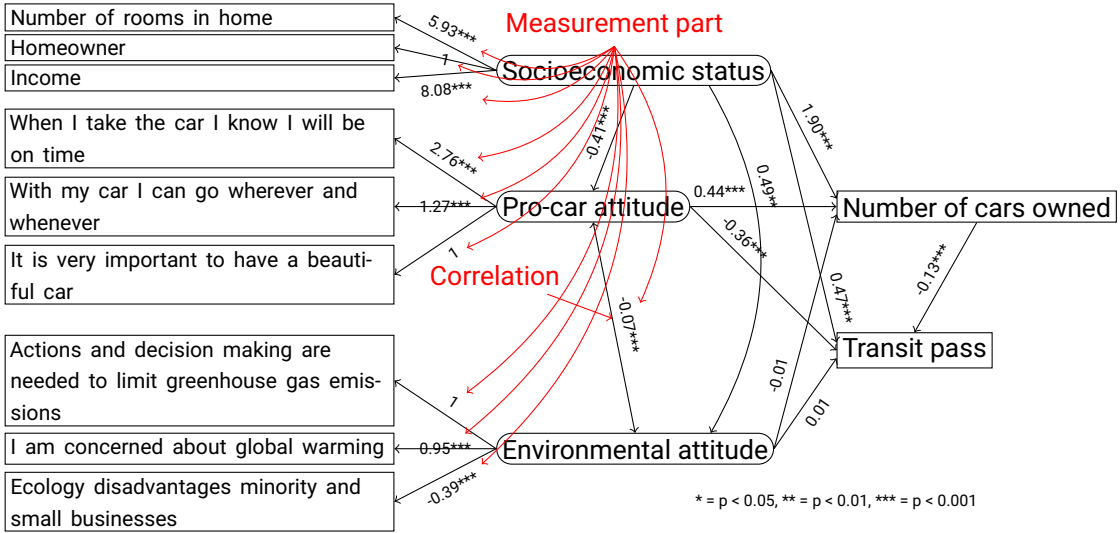
Data: Bierlaire (2018)

Structural equation models

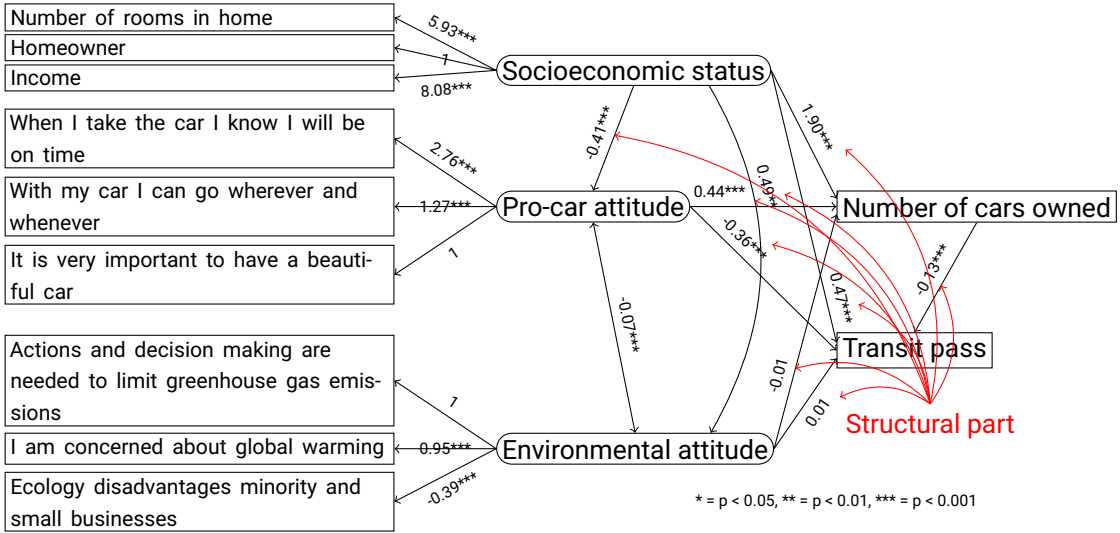


Data: Bierlaire (2018)

Structural equation models



Structural equation models



Data: Bierlaire (2018)

Structural equation models

- SEMs need a strong basis in theory
- Correlation (still) does not imply causation (even when there are arrows!)
  - SEMs test whether data is consistent with theoretical causal structures (Bollen and Pearl 2013)
- Authors should test multiple SEMs (Bowen and Guo 2012)
- There are many ways to evaluate model fit (see Bowen and Guo 2012)

Structural equation models: table form

Standardized direct effects of travel attitudes on travel-related decisions.

From → / To ↓	Male head			Female head		
	Pro-driving	Pro-alternatives	Pro-accessibility	Pro-Driving	Pro-alternatives	Pro-accessibility
Model I (weekday)						
Distance to the city center	0.111[b]		−0.112[b]	0.098[c]		−0.090[b]
Distance to transit		−0.074[c]		0.089[c]		−0.079
Non-work destination accessibility	0.109[c]		−0.199[a]			
Isolated bicycle lane	0.121[b]	0.154[b]			−0.152[b]	
Commute distance (M)				0.091[c]		
Commute distance (F)			0.070	0.120[a]		
Car ownership	0.073[c]					
Household VMT					−0.050	
Total travel time (M)		−0.146[a]				
Total travel time (F)		−0.161[a]				
Share of travel by car (M)				−0.107[c]	0.068	
Share of travel by car (F)				0.073		
Share of travel by non-motorized modes (M)	−0.094					
Share of travel by non-motorized modes (F)						

Guan and Wang (2019)

Latent and observed variables

- Structural equation models consist of latent and observed variables
- An observed variable is one that come from outside the model—like a survey question
- A latent variable is a variable that is determined inside the model—like an attitudinal factor
- Every model has observed variables, some don't have latent variables

Exogenous and endogenous variables

- An exogenous variable is one that is not influenced by any other variable in the model
  - *i.e.* no arrows pointing to the variable
- An endogenous variable is a variable that is influenced by at least one other variable in the model
  - *i.e.* at least one arrow pointing to the variable
  - there may or may not be arrows pointing away from the variable
- All SEMs have at least one endogenous variable, almost all have exogenous variables

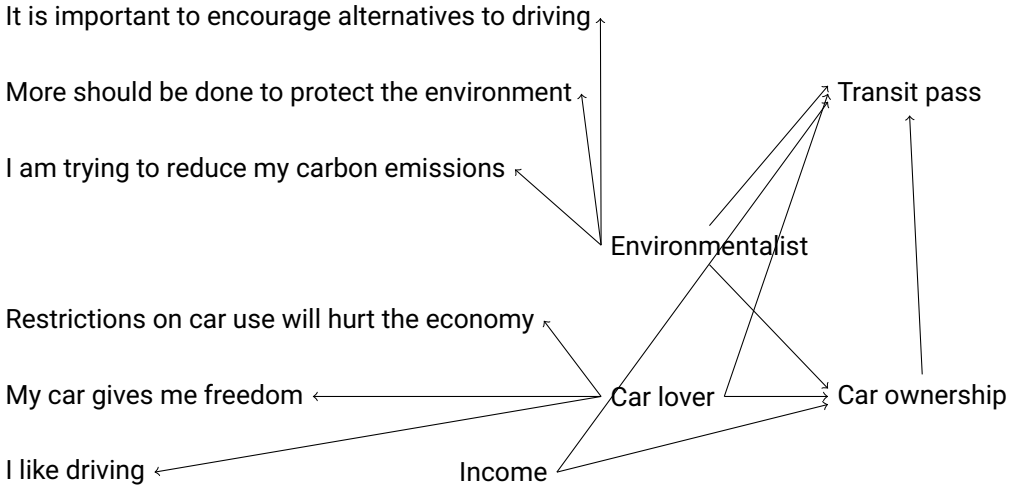
Latent and observed vs. endogenous and exogenous variables

- Latent and observed have nothing to do with endogenous or exogenous
- All of these types of variables are possible
  - observed, endogenous
  - latent, endogenous
  - observed, exogenous
  - latent, exogenous

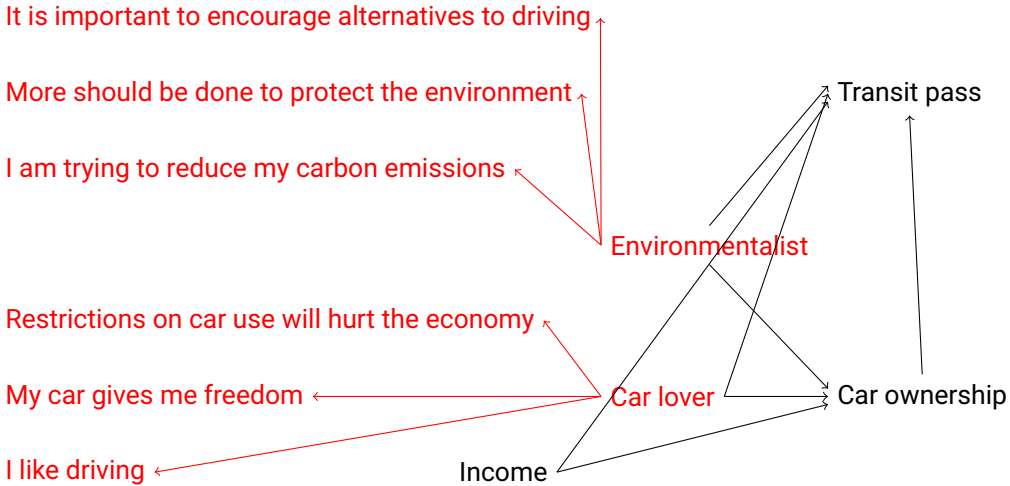
Exploratory factor analysis, confirmatory factor analysis, and SEMs

- Exploratory factor analysis is a separate method from SEMs
- Confirmatory factor analysis can be (and usually is) part of the SEM
- Most authors use an exploratory factor analysis to identify the structure before using a confirmatory factor analysis in their SEM

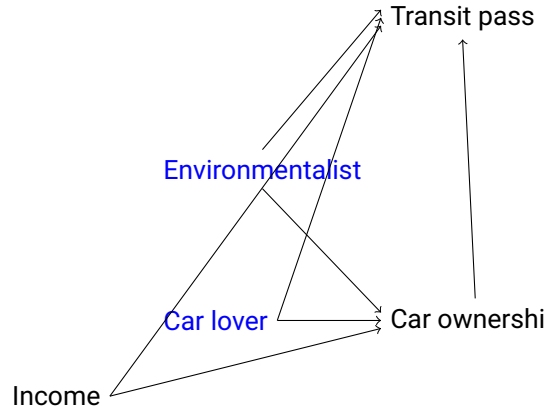
Confirmatory factor analysis in an SEM



Confirmatory factor analysis in an SEM



Exploratory factor analysis in an SEM



Exploratory factor analysis in an SEM

- If factors are found with an exploratory factor analysis, there are calculated before the SEM is fit
- Thus, from the perspective of the SEM, these are observed variables
- ...even though they were latent variables in the exploratory factor analysis

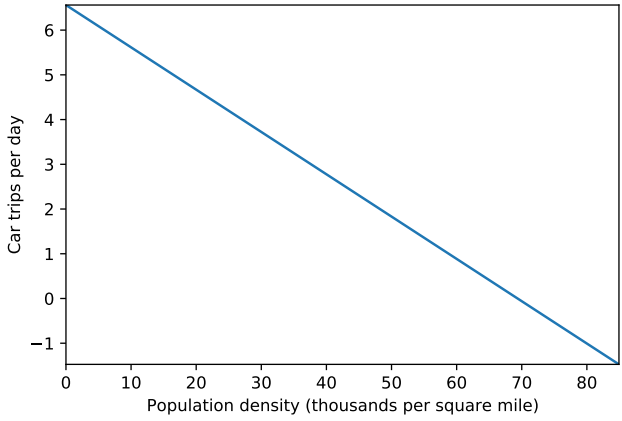
Exploratory factor analysis in an SEM

- If factors are found with an exploratory factor analysis, there are calculated before the SEM is fit
- Thus, from the perspective of the SEM, these are observed variables
- ...even though they were latent variables in the exploratory factor analysis
- (but not everyone agrees with me on this)

Count models

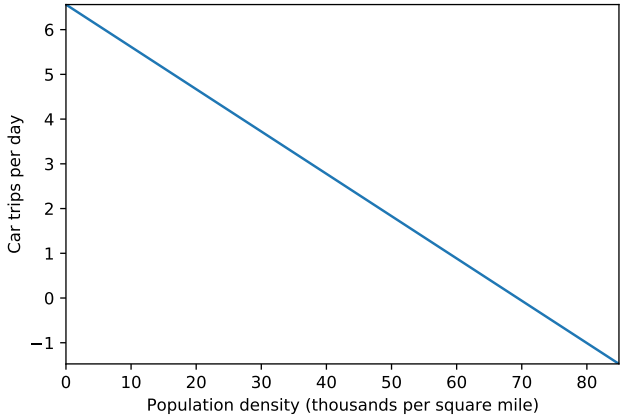
- Used when modeling counts of items
- Predicts probability of integer outcomes
- ...therefore, is a discrete choice model

Count models: why not linear regression?



Data: 2017 NHTS

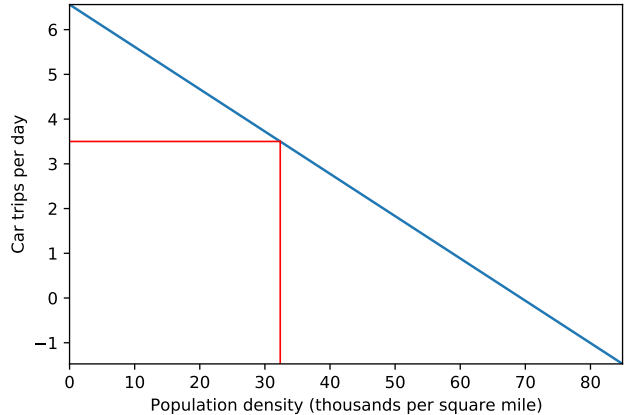
Count models: why not linear regression?



What is the predicted vehicle ownership in a neighborhood with 32,000 people per square mile?

Data: 2017 NHTS

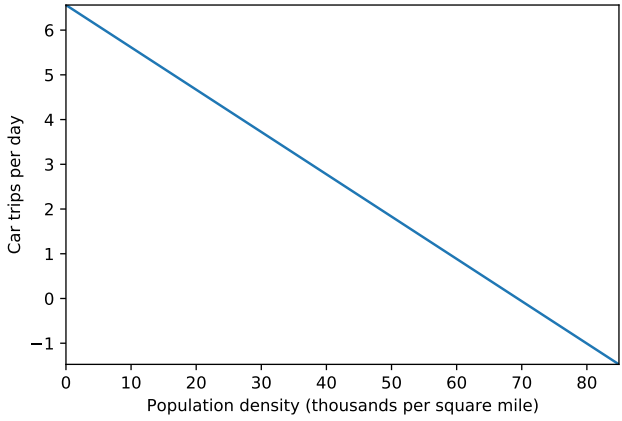
Count models: why not linear regression?



What is the predicted vehicle ownership in a neighborhood with 32,000 people per square mile?

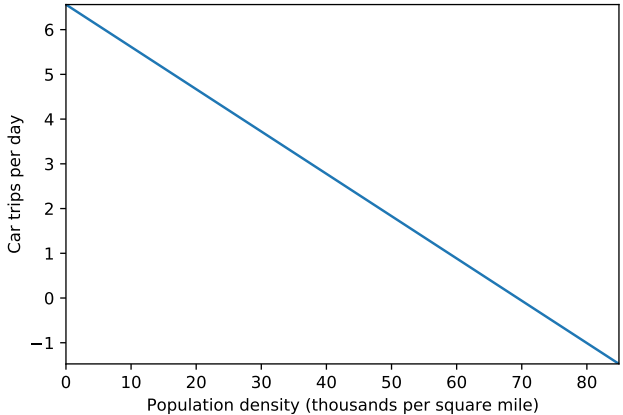
Data: 2017 NHTS

Count models: why not linear regression?



Data: 2017 NHTS

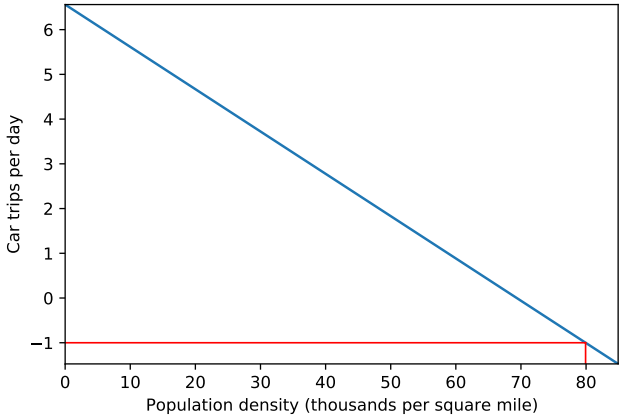
Count models: why not linear regression?



What is the predicted car ownership in a neighborhood with 71,000 people per square mile?

Data: 2017 NHTS

Count models: why not linear regression?



What is the predicted car ownership in a neighborhood with 71,000 people per square mile?

Data: 2017 NHTS

Poisson regression

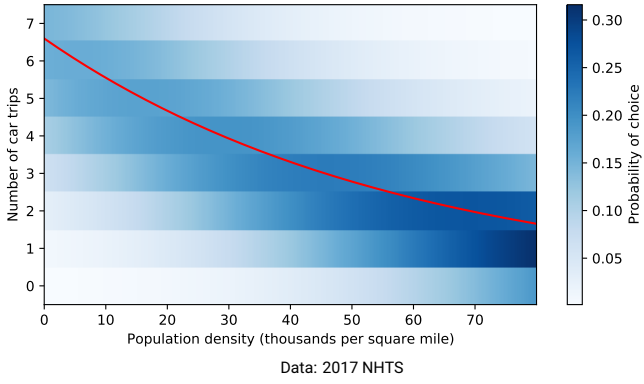
- Poisson regression solves these problems by modeling counts using a Poisson distribution
- Poisson distribution is discrete, defined only for nonnegative integers
- The mean of the distribution is modeled as $\mu = e^{\alpha+\beta_1 x_1+\cdots+\beta_n x_n}$
- This determines the probabilities for all possible counts
- Thus, Poisson is a multiplicative model
- e^{β} is the incidence rate ratio, ratio of expected counts when variable increase by one

Poisson regression: example

	Coefficient	Incidence rate ratio	Std. err.	p-value
Constant	1.887		0.001	0.0
Population density (thousands per square mile)	-0.017	0.983	0.0	0.0

Data: 2017 NHTS

Poisson regression: example



Negative binomial regression

- Poisson regressions assumes that the standard deviation of the error term is the same as the mean
- This is a result of the derivation of the Poisson distribution from the binomial distribution
- Assumes that the model predicts perfectly

Negative binomial regression

- Poisson regressions assumes that the standard deviation of the error term is the same as the mean
- This is a result of the derivation of the Poisson distribution from the binomial distribution
- Assumes that the model predicts perfectly
- ...which is never true

Negative binomial regression

- Negative binomial regression replaces the Poisson distribution with the negative binomial
- Adds a parameter α that models the standard deviation of the error term
- When $\alpha = 0$ or $\ln\alpha = 1$, equivalent to Poisson
- When $\alpha > 0$ or $\ln\alpha > 1$, errors are overdispersed
- When $\alpha < 0$ or $\ln\alpha < 1$, not consistent with negative binomial assumptions
- Basically, negative binomial makes estimates less certain to account for errors in prediction

Negative binomial regression: example

	Coefficient	Incidence rate ratio	Std. err.	p-value
Constant	1.893		0.003	0.0
Population density (thousands per square mile)	-0.019	0.981	0.0	0.0
α	0.675		0.004	0.0

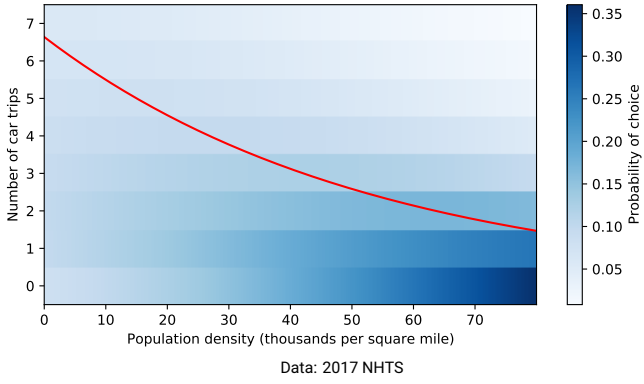
Data: 2017 NHTS

Negative binomial regression: example

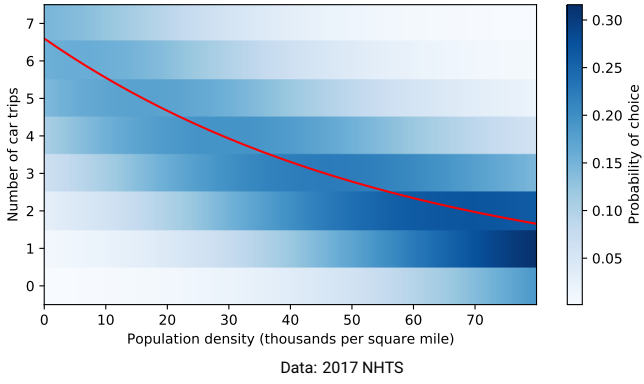
	Coefficient	Incidence rate ratio	Std. err.	p-value
Constant	1.893		0.003	0.0
Population density (thousands per square mile)	-0.019	0.981	0.0	0.0
α	0.675		0.004	0.0

Data: 2017 NHTS

Negative binomial regression: example



Poisson regression: example



Zero-inflated models

- Sometimes count data have more zeros than expected
- One solution is to use zero-inflated models
- These models hypothesize two reasons for a zero
  1. Structural: that observation must always* be zero (e.g. households without cars)
  2. Chance: no events were observed by chance (e.g. households that decided not to go out by car on the travel day)

Zero-inflated models

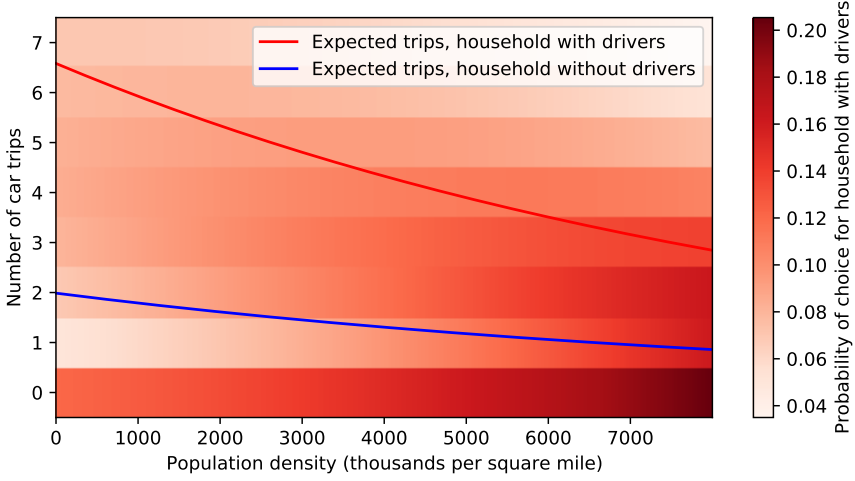
- Zero-inflated models estimate a binary model for the structural zeros, and a count model for the remaining observations
- Models do not have to use the same variables
- Zero-inflated Poisson and zero-inflated negative binomial are both available

Zero-inflated models: example

	Coefficient	Incidence rate ratio	Std. err.	p-value
Zero-inflation model				
Constant	-2.234		0.012	0.0
No driver in household	3.220	25.025	0.037	0.0
Count model				
Constant	1.99	7.283	0.003	0.0
Population density (thousands per square mile)	-0.010	0.990	0.0	0.0
α	0.355		0.003	0.0

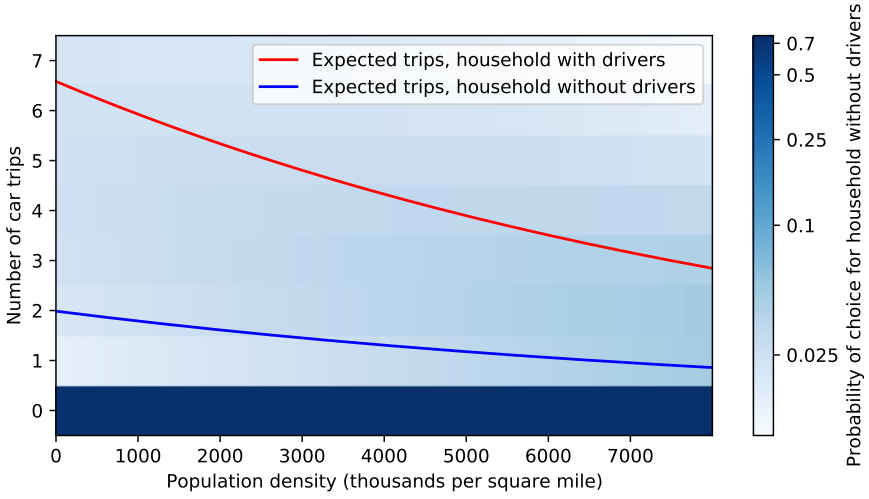
Data: 2017 NHTS

Zero-inflated models: example



Data: 2017 NHTS

Zero-inflated models: example



Data: 2017 NHTS

Zero-inflated τ models

- In a zero-inflated τ (tau) model, there is not a separate set of coefficients for the binary model
- Rather, a single coefficient τ is estimated that scales the coefficients from the count model to predict the structural zeros
- Assumes same relative contribution of variables to count and structural zeros

Interpreting count models

- Much like interpreting linear regression with a logged dependent variable
- When incidence rate ratios are presented, they are easy to interpret
- Otherwise, positive coefficients mean larger counts, negative means smaller
- Zero-inflated models: must account for effects from both count and zero-inflation model

Interpreting count models: example

	Schoolchildren		
Variable	Elementary	Middle	High
Constant	1.19 (0.85, 1.66) 0.99		
General density[a]	0.94*** (0.91, 0.96) -4.83	0.93*** (0.89, 0.97) -3.56	1.08*** (1.03, 1.13) 3.09
Home vs. work/ school[a]	1.01 (0.94, 1.10) 0.37	1.05 (0.95, 1.16) 0.88	1.11** (1.00, 1.23) 2.04
Intersection vs. population density[a]	0.93** (0.87, 0.99) -2.17	1.24*** (1.12, 1.38) 4.03	1.23*** (1.08, 1.40) 3.06
Home in CBD	0.37*** (0.26, 0.55) -5.07		
Land-use mix	1.06 (0.84, 1.34) 0.48	0.83 (0.61, 1.13) -1.21	0.50*** (0.34, 0.72) -3.69

Salon et al. (2019)

Ordered probit model

- Model used for ordered but not interval or ratio scaled data
- Like linear regression, but with thresholds for the different categories

Ordered probit: example

Ordered probit model for change in driving

	Coefficient	*p*-value
Constant[a]	1.508	0.000
Current age	−0.006	0.014
Currently working	0.155	0.065
Current # kids <18 years	0.070	0.051
Limits on driving	−0.678	0.000
Change in income	0.000	0.000
# groceries within 1600 m	−0.014	0.048
# pharmacies within 1600 m	−0.028	0.041
# theaters within 400 m	−0.703	0.055
Change in accessibility factor	−0.269	0.000
Change in safety factor	−0.088	0.000
Car dependent	0.115	0.000
Pro-bike/walk	−0.070	0.020
Threshold parameter—1	0.543	0.000
Threshold parameter—2	2.142	0.000
Threshold parameter—3	2.589	0.000
N	1490	
Log-likelihood at 0	−2378.038	
Log-likelihood at constant	−1954.785	
Log-likelihood at convergence	−1869.302	
Pseudo-*R* square	0.214	
Adjusted pseudo-*R* square	0.209	

Handy, Cao, and Mokhtarian (2005)

Ordered probit: example

$$y^* = 1.51 - 0.006x_{\text{age}} + 0.16x_{\text{working}} + 0.07x_{\text{children}} - 0.68x_{\text{limitsondriving}} \cdots + \epsilon$$

Based on Handy, Cao, and Mokhtarian (2005)

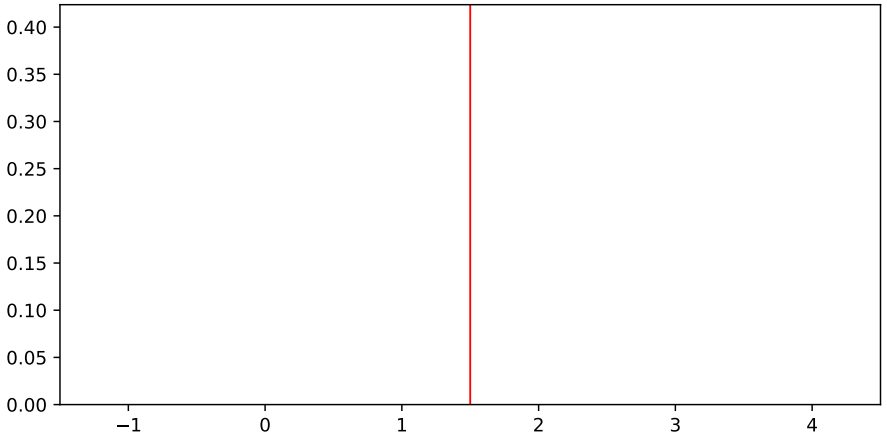
Ordered probit: example

$$y^* = 1.51 - 0.006x_{\text{age}} + 0.16x_{\text{working}} + 0.07x_{\text{children}} - 0.68x_{\text{limitsondriving}} \cdots + \epsilon$$

$$y = \begin{cases} \text{a lot less now} & \text{if } y^* \leq 0 \\ \text{a little less now} & \text{if } 0 < y^* \leq 0.54 \\ \text{about the same} & \text{if } 0.54 < y^* \leq 2.14 \\ \text{a little more now} & \text{if } 2.14 < y^* \leq 2.59 \\ \text{a lot more now} & \text{if } 2.59 < y^* \end{cases}$$

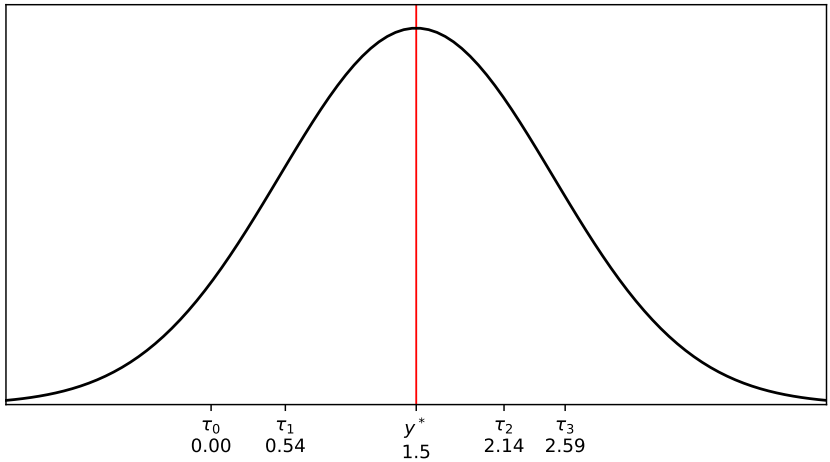
Based on Handy, Cao, and Mokhtarian (2005)

Ordered probit: example



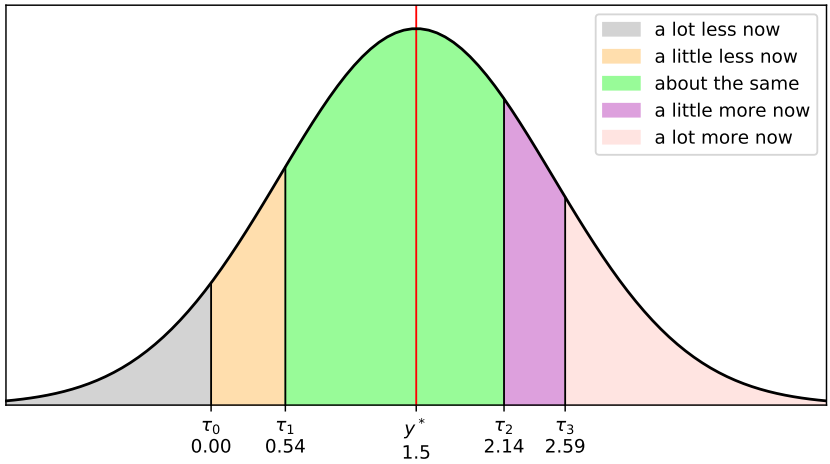
Based on Handy, Cao, and Mokhtarian (2005)

Ordered probit: example



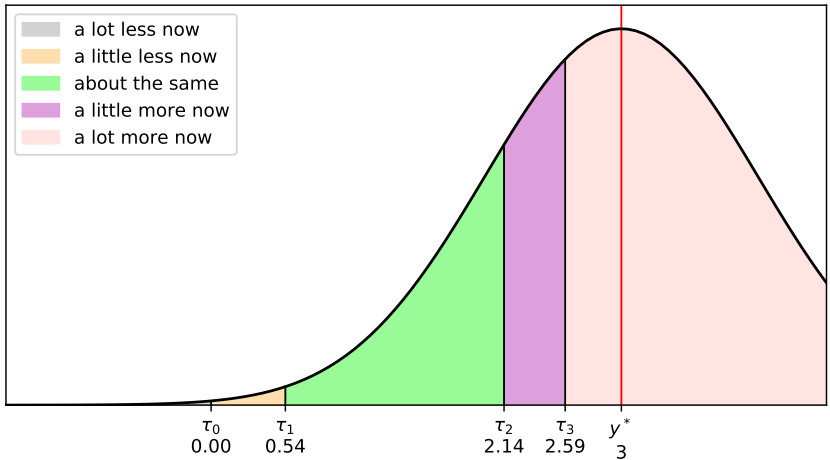
Based on Handy, Cao, and Mokhtarian (2005)

Ordered probit: example



Based on Handy, Cao, and Mokhtarian (2005)

Ordered probit: example

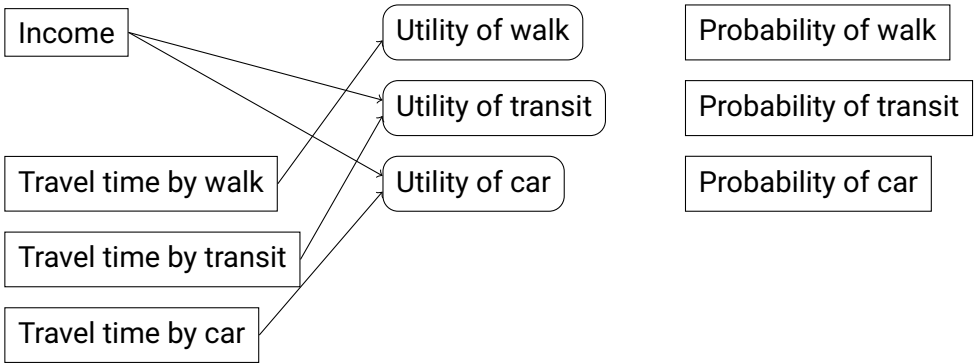


Based on Handy, Cao, and Mokhtarian (2005)

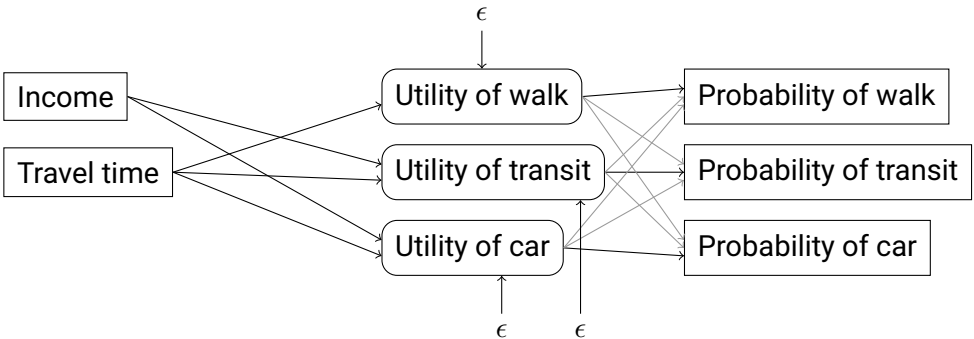
Integrated choice and latent variable models

- Integrated Choice and Latent Variable models are an extension of multinomial logit models
- They allow including attitudes in the model specification
- The attitudes are included as another latent variable, since we don't model them directly
- These latent variables are informed by attitudinal indicators
- Which indicators go with which latent variables generally based on factor analysis
- Also called hybrid choice models

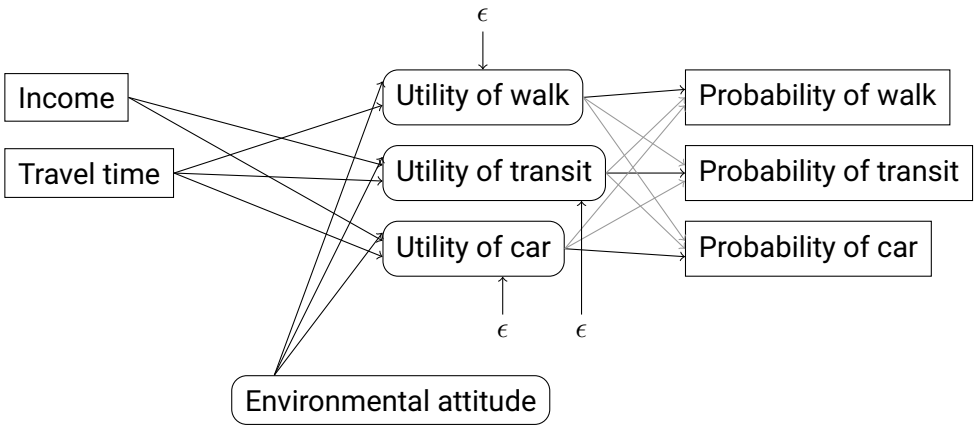
Integrated choice and latent variable models



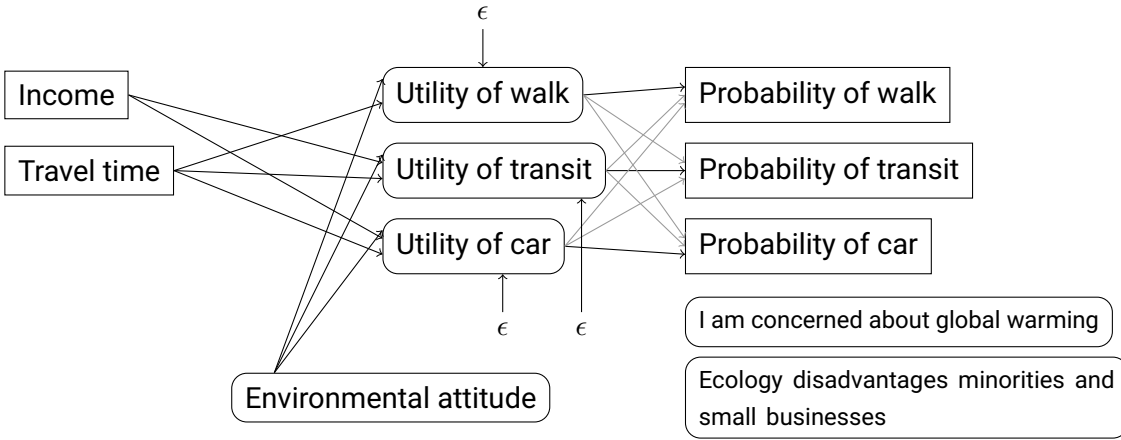
Integrated choice and latent variable models



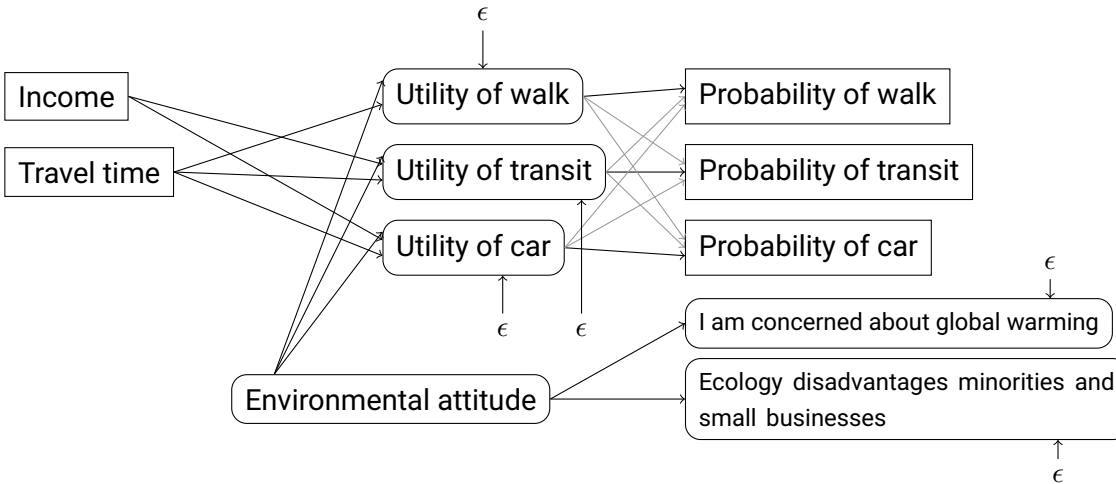
Integrated choice and latent variable models



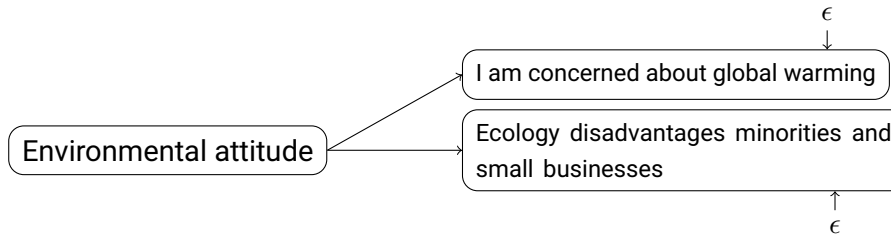
Integrated choice and latent variable models



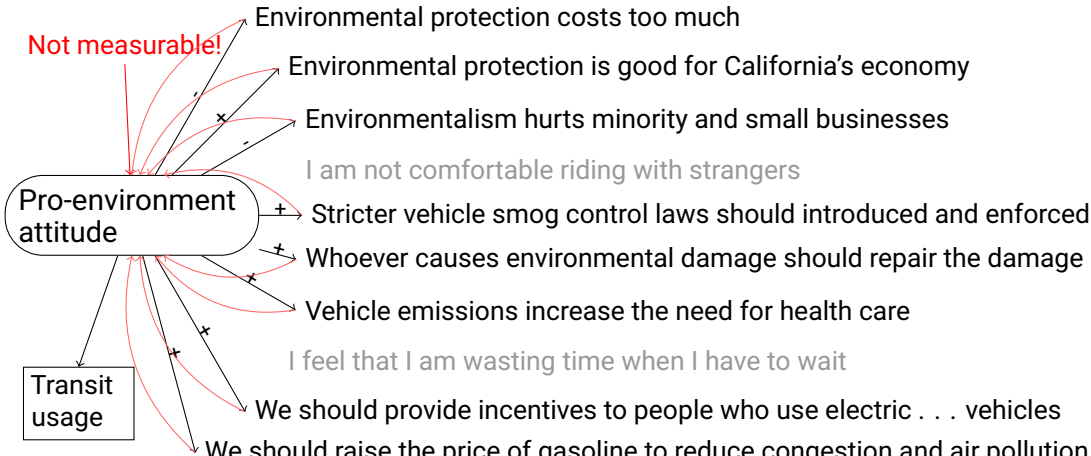
Integrated choice and latent variable models



Integrated choice and latent variable models

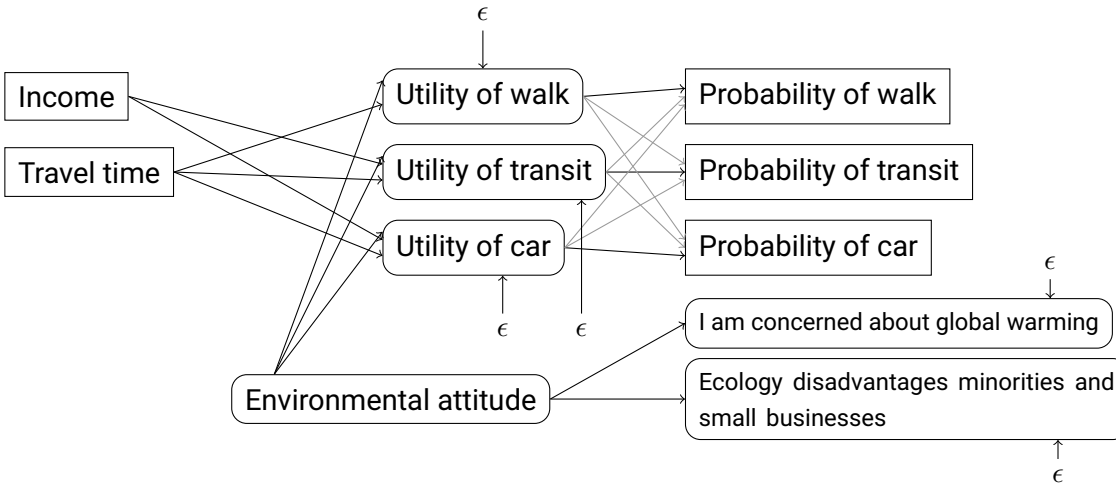


Integrated choice and latent variable models

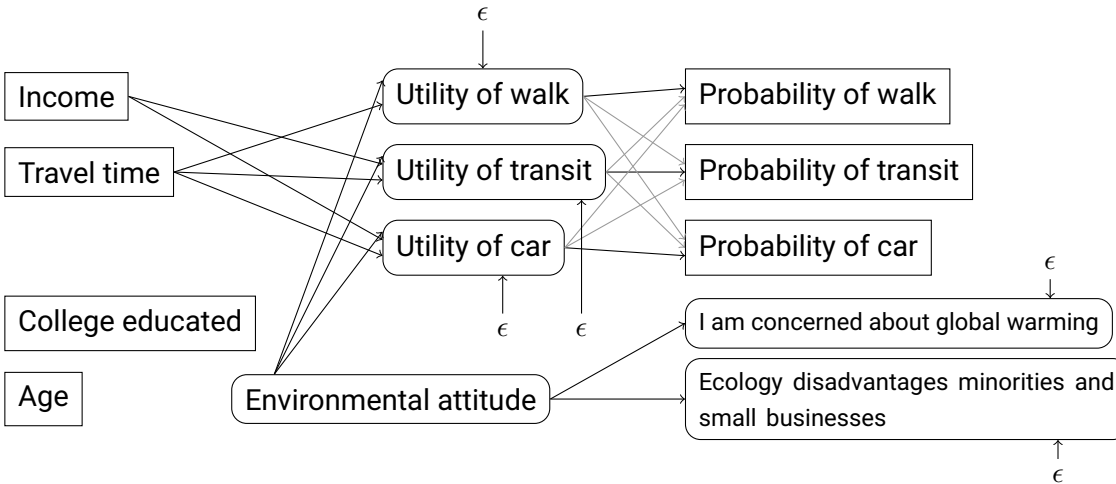


Based on Kitamura, Mokhtarian, and Laidet (1997)

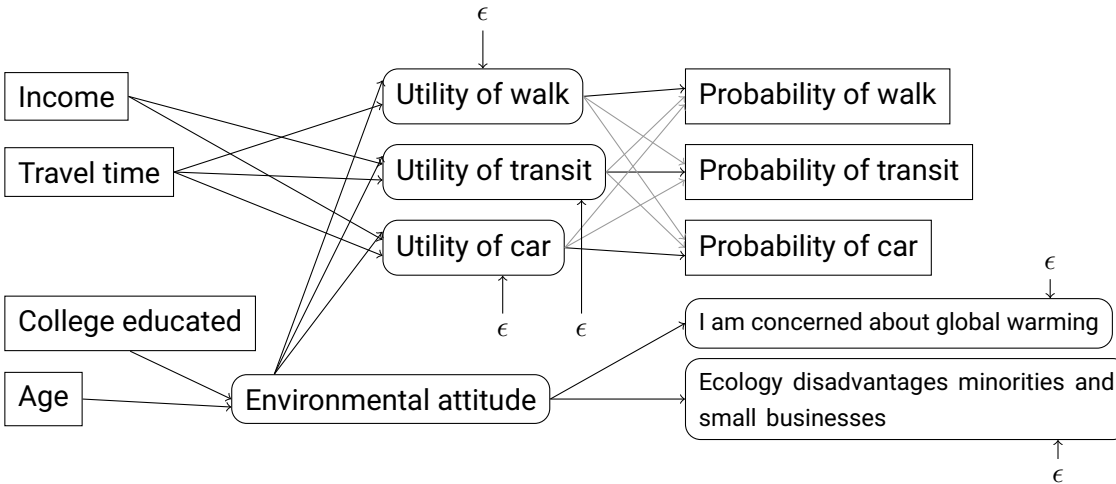
Integrated choice and latent variable models



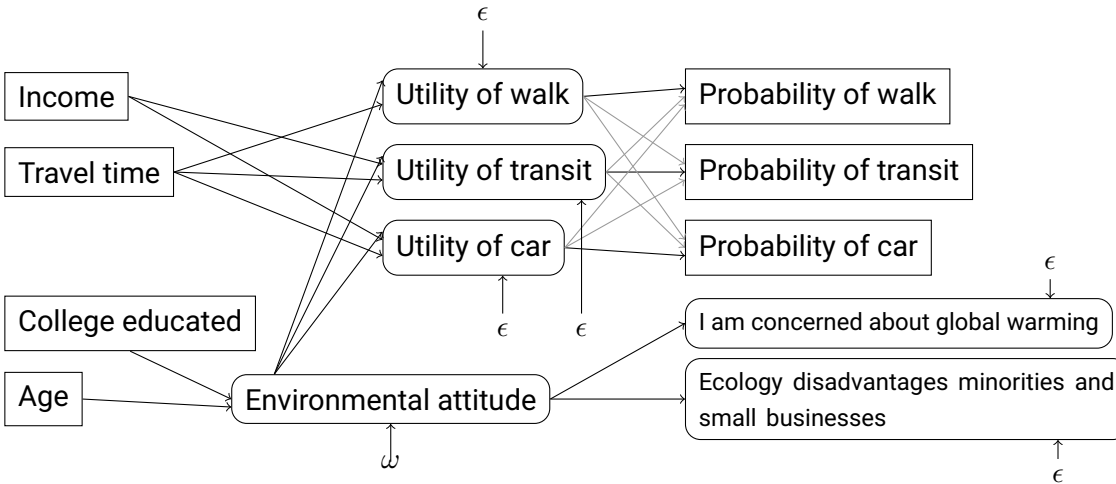
Integrated choice and latent variable models



Integrated choice and latent variable models



Integrated choice and latent variable models



Integrated choice and latent variable models: the math

$$
\begin{aligned}
U_{walk} &= && && \beta_{time} x_{time,walk} && + \beta_{walk,environment} x_{*} && + \epsilon_{walk} \\
U_{transit} &= \alpha_{transit} && + \beta_{income,transit} x_{income} && + \beta_{time} x_{time,transit} && + \beta_{transit,environment} x_{*} && + \epsilon_{transit} \\
U_{car} &= \alpha_{car} && + \beta_{income,car} x_{income} && + \beta_{time} x_{time,car} && && + \epsilon_{car}
\end{aligned}
$$

Integrated choice and latent variable models: the math

Latent attitude

$$
\begin{aligned}
U_{walk} &= & & \beta_{time} x_{time,walk} & +\beta_{walk,environment} x_* & +\epsilon_{walk} \\
U_{transit} &= \alpha_{transit} & +\beta_{income,transit} x_{income} & +\beta_{time} x_{time,transit} & +\beta_{transit,environment} x_* & +\epsilon_{transit} \\
U_{car} &= \alpha_{car} & +\beta_{income,car} x_{income} & +\beta_{time} x_{time,car} & & +\epsilon_{car} \\
x_* &= \alpha_{x*} & +\beta_{x*,age} x_{age} & +\beta_{x*,college} x_{college} & +\sigma\omega &
\end{aligned}
$$

Integrated choice and latent variable models: the math

Latent attitude

$$
\begin{aligned}
U_{walk} &= & & \beta_{time} x_{time,walk} & +\beta_{walk,environment} x* & +\epsilon_{walk} \\
U_{transit} &= \alpha_{transit} & +\beta_{income,transit} x_{income} & +\beta_{time} x_{time,transit} & +\beta_{transit,environment} x* & +\epsilon_{transit} \\
U_{car} &= \alpha_{car} & +\beta_{income,car} x_{income} & +\beta_{time} x_{time,car} & & +\epsilon_{car} \\
x* &= \alpha_{x*} & +\beta_{x*,age} x_{age} & +\beta_{x*,college} x_{college} & +\sigma\omega & \\
I_{globalwarming} & & = \alpha_{globalwarming} & +\beta_{x*,globalwarming} x* & \epsilon_{globalwarming} & \\
I_{economy} & & = \alpha_{economy} & +\beta_{x*,economy} x* & \epsilon_{economy} &
\end{aligned}
$$

Random variable

Integrated choice and latent variable models: choice model

	Value	Std err	t-test	p-value
Travel time (hours)	-0.60	0.01	-61.34	0.00
Walk				
Environmental attitude	-0.15	0.17	-0.89	0.37
Public transit				
Alternative specific constant	0.26	0.60	0.44	0.66
Income (tens of thousands of Swiss francs)	-1.42	0.10	-14.76	0.00
Environmental attitude	0.12	0.08	1.56	0.12
Car				
Alternative specific constant	0.85	0.57	1.49	0.14
Income (tens of thousands of Swiss francs)	-1.27	0.09	-13.71	0.00

Data: Bierlaire (2018)

Integrated choice and latent variable models: measurement models

	Value	Std err	t-test	p-value
Agreement with "I am concerned about global warming"				
Constant	0	–	–	–
Environmental attitude	1	–	–	–
Agreement with "Ecology disadvantages minorities and small businesses"				
Constant	5.78	0.17	34.45	0.00
Environmental attitude	-0.81	0.05	-17.36	0.00

Data: Bierlaire (2018)

Integrated choice and latent variable models: measurement models

Fixed

	Value	Std err	t-test	p-value
Agreement with "I am concerned about global warming"				
Constant	0	–	–	–
Environmental attitude	1	–	–	–
Agreement with "Ecology disadvantages minorities and small businesses"				
Constant	5.78	0.17	34.45	0.00
Environmental attitude	-0.81	0.05	-17.36	0.00

Data: Bierlaire (2018)

Integrated choice and latent variable models: measurement models

Fixed

	Value	Std err	t-test	p-value
Agreement with "I am concerned about global warming"				
Constant	0	–	–	–
Environmental attitude	1	–	–	–
Agreement with "Ecology disadvantages minorities and small businesses"				
Constant	5.78	0.17	34.45	0.00
Environmental attitude	-0.81	0.05	-17.36	0.00

Data: Bierlaire (2018)

Integrated choice and latent variable models: latent variable model

	Value	Std err	t-test	p-value
Constant	3.60	0.01	245.57	0.00
Age (10s of years)	-0.02	0.00	-9.12	0.00
College	0.32	0.01	27.89	0.00
σ	-4×10^{-6}	0.00	-0.01	0.99

Data: Bierlaire (2018)

Integrated choice and latent variable models: latent variable model

	Value	Std err	t-test	p-value
Constant	3.60	0.01	245.57	0.00
Age (10s of years)	-0.02	0.00	-9.12	0.00
College	0.32	0.01	27.89	0.00
σ	-4×10^{-6}	0.00	-0.01	0.99

Data: Bierlaire (2018)

Minimal influence of random variable

Integrated choice and latent variable models: real-world example

Parameter	Estimate	Std err	*t*-stat	*p* value
$\beta_{\text{drop-off}}$	0.481	0.283	1.70	0.09
β_{van}	0.422	0.490	0.86	0.39
β_{taxi}	2.58	0.321	8.02	0.00
β_{metro}	2.22	0.371	5.99	0.00
β_{bus}	2.04	1.06	1.93	0.05
Parking cost/income, park	−0.189	0.0435	−4.34	0.00
Leisure, park	0.278	0.146	1.90	0.06
Heavy, park	0.348	0.150	2.32	0.02
With family, park	0.390	0.157	2.47	0.01
Terminal roofed parking, park	0.352	0.144	2.44	0.01
Frequency, park	0.129	0.0357	3.61	0.00
Travel cost/income, van	−0.0165	0.0059	−2.80	0.01
Travel time, van	−0.00841	0.00524	−1.60	0.11
Heavy, van	0.912	0.171	5.32	0.00
With friend, van	0.739	0.221	3.34	0.00
With family, van	0.601	0.246	2.44	0.01
Frequency, van	0.0681	0.0394	1.73	0.08

Parameter	Estimate	Std err	*t*-stat	*p* value
Travel cost/income, taxi	−0.0153	0.0036	−4.24	0.00
Travel time, taxi	−0.0016	0.0033	−0.48	0.63
Frequency, taxi	0.081	0.0346	2.34	0.02
Travel cost, metro	−0.0942	0.0985	−0.96	0.34
Travel time, metro	−0.004	0.003	−1.33	0.18
Heavy, metro	−1.09	0.146	−7.46	0.00
Family, metro	−0.675	0.169	−4.00	0.00
Frequency, metro	0.0936	0.0363	2.58	0.01
Travel cost, bus	−0.135	0.0661	−2.05	0.04
Travel time, bus	−0.0217	0.0107	−2.02	0.04
Leisure, bus	0.850	0.383	2.22	0.03
Heavy, bus	−0.967	0.406	−2.38	0.02
PBE, park	0.173	0.101	1.72	0.09
NCH × PBE, van	−0.0829	0.0406	−2.04	0.04
NCH × PBE, taxi	−0.120	0.0294	−4.09	0.00
NCH × PBE, metro	−0.126	0.0375	−3.36	0.00
NCH × PBE, bus	−0.129	0.0968	−1.33	0.18

Yazdanpanah and Hadji Hosseinlou (2017)

Integrated choice and latent variable models: real-world example

Parameter	Estimate	Std err	t-Stat	p Value
Structural equations				
b_{PBE}	0.917	0.0671	13.66	0.00
Age ≤ 30	0.36	0.630	5.71	0.00
Female	0.111	0.0588	1.89	0.06
Married	0.230	0.0574	4.01	0.00
Not lived abroad	0.442	0.0558	7.92	0.00
Education < university	0.138	0.0786	1.75	0.08
Income < 3 million IRR	0.117	0.0592	1.97	0.05
σ_{ω}	0.906	0.0410	22.12	0.00
Measurement equations				
α_1	0	N.A	N.A	
α_2	0.194	0.0958	2.03	
α_3	2.11	0.0577	36.60	
λ_1	1	N.A	N.A	
λ_2	0.825	0.0622	13.25	
λ_3	0.346	0.0346	10.00	
σ_{v1}	0.793	0.0433	18.33	
σ_{v2}	0.804	0.0310	25.99	
σ_{v3}	1.06	0.0184	57.62	

Question	Mean	Stdv
(1) I think the streets of my residential area are standard and proper	1.49	1.24
(2) I think traffic flow at my residential area is fluent and convenient	1.42	1.13
(3) I think my residential area has good access to main streets and highways	2.63	1.11

(b)

Question	Strongly Disagree%	Disagree%	Neither Disagree nor Agree%	Agree%	Strongly Agree%	Total N
Q1.	25.3	33.2	15	20.1	6.4	359
Q2.	23.5	37	15	22.8	1.7	359
Q3.	6.1	12.8	12	50.4	18.7	359

Yazdanpanah and Hadji Hosseinlou (2017)

Integrated choice and latent variable models: why not just use factor analysis

1. ICLV models allow finding the combination of indicators that best fits for this dependent variable
2. When doing prediction, the attitudinal indicators do not need to be forecast

Bierlaire, Michel. 2018. “Mode Choice in Switzerland (Optima)”: 8. https://transp-or.epfl.ch/documents/technicalReports/CS_OptimaDescription.pdf.

Bollen, Kenneth A., and Judea Pearl. 2013. “Eight Myths about Causality and Structural Equation Models.” In Handbook of Causal Analysis for Social Research. Dordrecht: Springer.

Bowen, Natasha K, and Shenyang Guo. 2012. Structural Equation Modeling. Pocket Guides to Social Work Research Methods. Oxford, UK: Oxford University Press.

Conway, Matthew Wigginton, Deborah Salon, and David A King. 2018. “Trends in Taxi Use and the Advent of Ridehailing, 1995–2017: Evidence from the US National Household Travel Survey.” Urban Science 2, no. 3 (September 1): 79. doi:10.3390/urbansci2030079.

Ford, Clay. 2018. “Interpreting Log Transformations in a Linear Model.” Accessed September 2, 2019. https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/.

Frank, Lawrence Douglas, Brian E Saelens, Ken E Powell, and James E Chapman. 2007. “Stepping towards Causation: Do Built Environments or Neighborhood and Travel Preferences Explain Physical Activity, Driving, and Obesity?” Social Science & Medicine 65, no. 9 (November): 1898–1914. doi:10.1016/j.socscimed.2007.05.053.

Guan, X., and D. Wang. 2019. “Residential Self-Selection in the Built Environment-Travel Behavior Connection: Whose Self-Selection?” Transportation Research Part D: Transport and Environment 67:16–32. doi:10.1016/j.trd.2018.10.015.

Handy, Susan, Xinyu Cao, and Patricia Mokhtarian. 2005. “Correlation or Causality between the Built Environment and Travel Behavior? Evidence from Northern California.” Transportation Research Part D: Transport and Environment 10, no. 6 (November): 427–444. ISSN: 13619209. doi:10.1016/j.trd.2005.05.002.

Kitamura, Ryuichi, Patricia L Mokhtarian, and Laura Laidet. 1997. “A Micro-Analysis of Land Use and Travel in Five Neighborhoods in the San Francisco Bay Area.” Transportation 24, no. 2 (January 1): 125–158. doi:10.1023/A:1017959825565.

Kline, Paul. 1994. An Easy Guide to Factor Analysis . Psychology Press. ISBN: 978-0-415-09490-0.

Kuppam, A, Ram M Pendyala, and Shela Rahman. 1999. “Analysis of the Role of Traveler Attitudes and Perceptions in Explaining Mode-Choice Behavior.” … Record: Journal of the … 1676 (January 1): 68–76. doi:10.3141/1676-09.

O’Brien, Robert M. 2007. “A Caution Regarding Rules of Thumb for Variance Inflation Factors.” Quality & Quantity 41, no. 5 (March 13): 673–690. doi:10.1007/s11135-006-9018-6.

Salon, Deborah, Matthew Wigginton Conway, Kailai Wang, and Nathaniel Roth. 2019. "Heterogeneity in the Relationship between Biking and the Built Environment." Journal of Transport and Land Use 12, no. 1 (January 28): 99–126. doi:10.5198/jtlu.2019.1350.

Thompson, Bruce. 1995. "Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply Here: A Guidelines Editorial." Educational and Psychological Measurement 55, no. 4 (January 1): 525–534. doi:10.1177/0013164495055004001.

"FAQ: What Are Pseudo R-Squareds?" 2011. Accessed August 30, 2019. https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/.

Yazdanpanah, M., and M. Hadji Hosseinlou. 2017. "The Impact of Perception toward the Built Environment in Airport Access Mode Choice Using Hybrid Choice Modeling." Journal of Advanced Transportation 2017. doi:10.1155/2017/8268701.

License

Software used for example models: `statsmodels`, `biogeme`, `lavaan`