Appendix II: HILDA Data

HILDA data for pension adequacy - detailed analysis notes

Description of data analysis process and examples of relevant STATA code.

Data from the Household Income and Labour Dynamics in Australia (HILDA) survey were analysed using the statistical package STATA to examine spending patterns among Australians over the age of 65. STATA code is presented in *blue italics*.

Box A1

HILDA Survey

The Household, Income and Labour Dynamics in Australia (HILDA) Survey is a household-based panel study which began in 2001. It has the following key features:

- It collects information about economic and subjective well-being, labour market dynamics and family dynamics.
- Special questionnaire modules are included each wave.
- The wave 1 panel consisted of 7,682 households and 19,914 individuals. In wave 11 this was topped up with an additional 2,153 households and 5,477 individuals.
- Interviews are conducted annually with all adult members of each household.
- The panel members are followed over time.
- The funding has been guaranteed for 18 waves, though the survey is designed to continue for longer than this.
- Academic and other researchers can apply to use the General Release datasets for their research.

The variables from HILDA that were imported into STATA for this analysis were:

Variable	Description
nhhiage	Age last birthday at date of interview
nhhtype	Household type – Single, couple without children etc.
nhhpers	Number of in-scope persons in household
nftsex	Gender
nhstenr	Home tenure. Own, Rent or live rent free
nhhra	ASGC 2001 Remoteness area
nhhmsr	ASGC 2001 Major Statistical Region

nhifditp	Household financial year disposable income [Imputed]
nhwnwip	Household net worth (imputed)
nhwhmeqp	Household wealth: Home equity
nxpgroci	Household weekly expenditure on all groceries [imputed]
nxpfoodi	Household weekly expenditure on groceries for food and drink [imputed]
nhxyalci	Household annual expenditure - Alcohol ($) [imputed]
nhxycigi	Household annual expenditure - Cigarettes and tobacco ($) [imputed]
nhxypbti	Household annual expenditure - Public transport and taxis ($) [imputed]
nxposmli	Household weekly expenditure on meals outside the home [imputed]
nhxymvfi	Household annual expenditure - Motor vehicle fuel ($) [imputed]
nhxymcfi	Household annual expenditure - Mens clothing and footwear ($) [imputed]
nhxywcfi	Household annual expenditure - Womens clothing and footwear ($) [imputed]
nhxyteli	Household annual expenditure - Telephone rent, calls and internet charges [imputed]
nhxyphii	Household annual expenditure - Private health insurance ($) [imputed]
nhxyoii	Household annual expenditure - Other insurance (home/contents/ motor vehicle) [imputed]
nhxyhlpi	Household annual expenditure - Fees paid to health practitioners ($) [imputed]
nhxyphmi	Household annual expenditure - Medicines, prescriptions, pharmaceuticals
nhxyutli	Household annual expenditure – Electricity bills, gas bills and other heating [imputed]
nhxyhmri	Household annual expenditure - Home repairs/renovations/ maintenance ($) [Imputed]
nhxymvri	Household annual expenditure - Motor vehicle repairs/ maintenance ($) [imputed]
nhsmg	Mortgage usual repayments $ per month
nhsrnt	Rent usual payments $ per month

Initial data filtering to leave only records of interest

Data were filtered to leave only those aged over 65 with annual income per person less than $65,000. $65,000 chosen to maximise the accuracy of models to lower income pensioners. It's close to triple the income of single full pensioners so should include enough individuals to show how spending changes as income increases above pension rate. Removing higher income individuals eliminates outliers and concentrates predictive power to the income brackets we are interested in.

These records were further filtered to leave only those in single or couple households who are either renting or living in their own home in order to remove the relatively few records of pensioners living with family or friends or in free accommodation. There were not enough of such records to independently analyse their expenditure.

Expenditure data in HILDA are self-reported and not generated through expenditure diaries. As a result, they contain anomalies and inaccuracies2. Many of these are undetectable and we must rely on statistical power to overcome them (assuming the errors are unbiased) but some can be picked up visually or computationally.

All raw variables were plotted in both scatter plots and bar graphs to determine distributions, detect outliers and examine patterns. New variables were then created eliminating outliers and removing negative values (used in HILDA to explain missing values).

New variables created from the above:

Generate new independent variables

- *gen gender=nftsex if nftsex>0*
- *gen housing=nhstenr if nhstenr>0*
- *gen location=nhhra if nhhra>0*
- *gen age=nhhiage if nhhiage>0*
- *gen inc=nhifditp/nhhpers/52 if nhifditp/nhhpers>18000*
- *gen nonhomeassets=nhwnwip - nhwhmeqp if nhwnwip>=0 & nhwhmeqp>=0*

Generate new expense (dependent) variables

- *gen groceries=nxpgroca/nhhpers if nxpgroca>0 & nxpgroca<500 // total groceries*
- *gen food=nxpfood/nhhpers if nxpfood>0 & nxpfood<400 //food*
- *gen alcohol=nxpalca/nhhpers if nxpalca>0 & nxpalca<250 //alcohol*
- *gen cigarettes=nxpciga/nhhpers if nxpciga>0 & nxpciga<250 //cigarettes*
- *gen pubtrans=nxppubta/nhhpers if nxppubta >0 & nxppubta<100 //Public transport and taxis (bill payer)*
- *gen usepubtrans = nxppubta >0 & nxppubta<100 //User of public transport (1) non-user (0) and taxis (bill payer)*
- *gen mealsoh=nxposml/nhhpers if nxposml>=0 & nxposml<250 //meals outside the home*
- *gen mealsob=nxpwmeoa/nhhpers if nxpwmeoa>=0 & nxpwmeoa<250 //Meals eaten out (bill payer)*
- *gen vfuel=nxpmvfa/nhhpers*12/52 if nxpmvfa>=0 & nxpmvfa<10000 //Motor vehicle fuel (bill payer)*
- *gen clothingm=nxpmcfa/nhhpers*12/52 if nxpmcfa>0 & nxpmcfa<300 //Mens clothing and footware (bill payer)*
- *gen clothingw=nxpwcfa/nhhpers*12/52 if nxpwcfa>0 & nxpwcfa<300 //Womens clothing and footware (bill payer)*
- *gen telint=nxptelia/nhhpers*12/52 if nxptelia>0 & nxptelia<500 //Telephone and internet (bill payer)*
- *gen phealthi=nxpphia/nhhpers/52 if nxpphia>=0 & nxpphia<10000 //Private health insurance (bill payer)*
- *gen otherins=nxpoia/nhhpers/52 if nxpoia>0 & nxpoia<6000 //Other insurance (home/contents/motor vehicle) (bill payer)*
- *gen healthp=nxphltpa/nhhpers/52 if nxphltpa>0 & nxphltpa<5000 //Fees paid to health practitioners (bill payer)*
- *gen pharma=nxpphrma/nhhpers/52 if nxpphrma>0 & nxpphrma<500 //Medicines, prescriptions, pharmaceuticals, alternative medicines (bill payer)*

- *gen utilities=nxputila/nhhpers/52 if nxputila>0 & nxputila<6000 //Electricity bills, gas bills and other heating fuel (bill payer)*
- *gen hrepairs=nxphmrna/nhhpers/52 if nxphmrna>0 & nxphmrna<10000 //Home repairs/renovations/maintenance (bill payer)*
- *gen vrepairs=nxpmvra/nhhpers/52 if nxpmvra>0 & nxpmvra<5000 //Motor vehicle repairs/maintenance (bill payer)*
- *gen mortgage=nhsmg/nhhpers*12/52 if nhsmg>0*
- *gen rent=nhsrnt/nhhpers*12/52 if nhsrnt>0*

Generate expenditure categories

egen exptotal=rowtotal(food alcohol cigarettes pubtrans mealsoh vfuel clothingm clothingw telint phealthi otherins healthp pharma utilities hrepairs vrepairs rent)

egen extransport=rowtotal(vfuel pubtrans vrepairs)

egen exfood=rowtotal(food mealsoh)

egen exhealth=rowtotal(pharma healthp phealthi)

vegen exhousing=rowtotal(hrepairs mortgage rent utilities)

Summary (STATA: sum, d) statistics were then generated for the new variables to examine skewness and kurtosis and they were graphed using scatter plots and box plots with confidence intervals to visually assess distributions, relationships and the effectiveness of filtering outliers.

Defining full pensioners

The HILDA survey does not include a flag for full age pensioner. As a result, we identified full pensioners by how much Age Pension they receive.

20 September 2015	**Couples**	**Singles**
Base	$594.30	788.40
Pension Supplement	$48.60	$64.50
Energy Supplement	$10.60	14.10
Total	**$653.50**	**$867.00**

Rent assistance is paid and reported as part of the pension payment. Adding maximum rent assistance to the above figures in order to capture full pension renters and putting in a buffer to capture people who round down a little leads to:

Full single pensioner

generate fspensioner = nbncapa>800 & nbncapa<1000 if nbncapa<. & nhhtype==24

Full couple pensioner

generate fcpensioner = nbncapa>590 & nbncapa<775 if nbncapa<. & nhhtype==1

All full pensioners

generate fullpensioner = fspensioner==1 | fcpensioner==1

These boundaries were chosen to capture individuals reporting on their base pension and on those reporting totals (including rounding up to nearest $10). The data show clear spikes both at the base rate of full pensions and at the total including supplements (Figure A1). This will result in some part pensioners who also receive CRA being classified as full pensioners but this was unavoidable and, in terms of the results of the analysis, is likely to have very little impact.

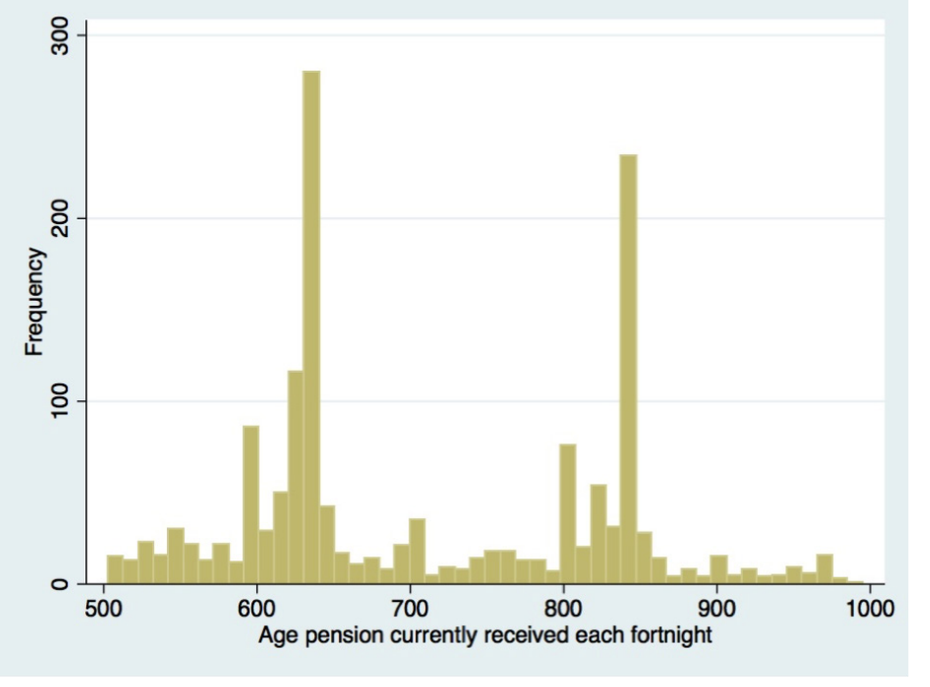

Figure A1. Amount of Age Pension received

Models

Generalised Linear Models (GLMs) were constructed for expenditure variables that did not have a disproportionate number of zeros. Dependent variables initially included were income (inc), gender, location, region, house tenure (housing), age, and assets other than main residence (nonhomeassets). Also included were interaction terms between all of these (except location and region) and quadratic terms for income and age. Least significant variables were then removed one by one while rerunning the models and comparing Akaike Information Criterion (AIC) scores for each of the models and leaving variables out if the AIC dropped or remained the same; or returning them to the model if the AIC rose. Dependent variables for which GLMs were used were: groceries, food, pharmaceuticals, health practitioners, housing, meals eaten out, utilities and telephone and internet.

For expenditure data (dependent variables) with many zeros a hurdle model was constructed. Variables with many zeros can present challenges for many statistical models and can distort results. However, the zeros in these data are informative and cannot be omitted. One such example is motor vehicle fuel. Clearly, there are many individuals who do not own cars and if we removed the zeros we would be artificially inflating the predicted expenditure on petrol for the average pensioner. The hurdle model calculates the probability of expenditure being greater than zero and, then, if greater than zero, predicts the mean expenditure as above. The final predicted value is the amount multiplied by the probability. Dependent variables for which hurdle models were used were: private health insurance, vehicle fuel, cigarettes, alcohol, public transport, home repairs, and vehicle repairs.

Specific variable notes

Gender

Expenditure data in HILDA are collected at a household level. This means that gender analyses can only be meaningfully done for single person households as couples share their data.

Region

HILDA records region in five categories: major city, inner regional, outer regional, remote and very remote. There were insufficient data for the last two categories and, as a result, all of our analyses only consider major city, inner regional and outer regional.

Model code:

Below is the STATA code used for each model that contributed to the findings in the report.

Food

```
glm food inc incsq i.gender i.housing i.location, family(gaussian) link(log)

predict pmexfood

rvfplot2

twoway (scatter pmexfood inc if inc<1000) (qfitci pmexfood inc if housing==2) (qfitci
pmexfood inc if housing==1) if inc<1000, ytitle(Expenditure ($ per week)) xtitle(Income ($
per week)) by(, title(Expenditure on food - Singles over 65) note(, size(zero))) by(gender)

margins if fspensioner==1, at(inc=(480 520)) by(gender housing) vsquish
```

STATA output for food GLM

```
Generalized linear models                    No. of obs = 1,595
Optimization: ML                             Residual df = 1,577
                                             Scale parameter = 1183.111
Deviance = 1865766.635                       (1/df) Deviance = 1183.111
Pearson = 1865766.635                        (1/df) Pearson = 1183.111

Variance function: V(u) = 1 [Gaussian]       AIC = 9.925001
                                             BIC = 1854137
Link function: g(u) = ln(u) [Log]
Log likelihood = -7897.188437
```

food	*Coef.*	*Std. Err.*	*z*	*P>|z|*	*[95% Conf. Interval]*	
inc	*0.001*	*0.000*	*3.78*	*0*	*0.001*	*0.002*
incsq	*0.000*	*0.000*	*-2.86*	*0.004*	*0.000*	*0.000*
2.gender	*0.013*	*0.023*	*0.54*	*0.592*	*-0.033*	*0.058*
2.housing	*-0.101*	*0.035*	*-2.84*	*0.005*	*-0.170*	*-0.031*
Location						
19	*-0.046*	*0.039*	*-1.18*	*0.239*	*-0.123*	*0.031*
21	*-0.123*	*0.041*	*-2.97*	*0.003*	*-0.204*	*-0.042*
29	*-0.138*	*0.049*	*-2.82*	*0.005*	*-0.233*	*-0.042*
31	*-0.173*	*0.053*	*-3.26*	*0.001*	*-0.277*	*-0.069*
39	*-0.078*	*0.046*	*-1.71*	*0.087*	*-0.168*	*0.011*
41	*-0.005*	*0.052*	*-0.09*	*0.927*	*-0.106*	*0.097*
49	*-0.176*	*0.073*	*-2.41*	*0.016*	*-0.319*	*-0.033*
51	*-0.044*	*0.049*	*-0.91*	*0.361*	*-0.140*	*0.051*
59	*-0.181*	*0.071*	*-2.56*	*0.011*	*-0.320*	*-0.042*
61	*-0.240*	*0.076*	*-3.17*	*0.002*	*-0.389*	*-0.092*
71	*-0.221*	*0.234*	*-0.95*	*0.344*	*-0.680*	*0.237*
81	*-0.087*	*0.079*	*-1.11*	*0.266*	*-0.242*	*0.067*
_cons	***3.880***	***0.116***	***33.39***	***0***	***3.652***	***4.108***

The production of figures and the margins command were the same for all variables and will not be repeated below.

Housing

glm exhousing inc i.nhhtype i.housing i.gender i.region i.location age genloc agehouse ageloc houseloc locinc incsq inccub loginc agesq, family(gaussian) link(log)

Meals eaten out

glm mealsoh inc age i.nhhtype i.housing i.gender i.location ageloc incsq, family(gaussian) link(identity)

Utilities

glm utilities inc i.nhhtype i.housing i.location agesq, family(gaussian) link(log)

Telephone and internet

glm telint inc i.nhhtype nonhomeassets houseloc, family(gaussian) link(identity)

Pharmaceuticals

glm pharma inc i.gender i.housing i.location agehouse, family(gaussian) link(identity)

Health practitioners

glm healthp inc i.gender age i.housing i.location inccub, family(gaussian) link(identity)

Private health insurance

churdle linear phealthi i.housing i.gender inc inccub i.region, ll(0) select(i.region i.housing i.gender inc nonhomeassets)

Vehicle fuel

churdle linear vfuel i.region i.housing i.gender i.nhhtype age inc, ll(0) select(i.nhhtype i.region i.housing i.gender age inc nonhomeassets)

Cigarettes

churdle linear cigarettes age i.gender i.housing inc, ll(0) select(i.gender i.housing nonhomeassets)

Alcohol

churdle linear alcohol age i.gender i.housing i.region, ll(0) select(i.gender i.housing nonhomeassets i.region)

Public transport

churdle exponential pubtrans i.housing i.gender age inc incsq agesq, ll(0) select(i.gender i.housing inc)

Home repairs

churdle exponential hrepairs i.housing i.gender age inc genloc genage ageinc locinc, ll(0) select(i.housing inc)

Vehicle repairs

churdle exponential vrepairs i.housing i.gender age inc, ll(0) select(age i.gender i.housing inc)

Summary of modelling and cautionary notes on interpretation

	"Essentially, all models are wrong, but some are useful" ***George Box - statistician***	

As expected, these models explain only a small amount of the substantial variation present in the expenditure variables. There are a vast number of factors that influence household expenditure to the point that every household expenditure profile is unique2. However, all of the models presented are statistically significant and tease out the mean relationship between expenditure variables and our segments of the HILDA survey population.

Importantly, the purpose of these models is not to explain variation in expenditure, but to predict changes in expenditure between different segments of the survey population. While the relationships and trends presented in this report are statistically robust, specific numbers may not be (in part due to the low accuracy of individual estimates of expenditure from the HILDA survey respondents2) and certainly no generalisation can be applied to individual cases. For example, we can predict, on average how much a male pensioner living alone in rental accommodation spends on alcohol but we certainly would not suggest that this average predicted figure tells us anything about any individual single male renters, for whom the range of expenditure on alcohol is enormous.

The key findings we present in the body of the report regarding relative expenditure between different groups and relative slopes of expenditure curves are robust. However, caution should be used when extrapolating more from the results than that.

Appendix II References

1. Saunders, P. et al. Development of Indicative Budget Standards for Australia. (Commonwealth of Australia, 1998).

2. Wilkins, R. & Sun, C. Assessing the Quality of the Expenditure Data Collected in the Self-Completion Questionnaire. (2010). at <https://www.melbourneinstitute.com/downloads/hilda/Bibliography/HILDA_Discussion_Papers/hdps110.pdf>