

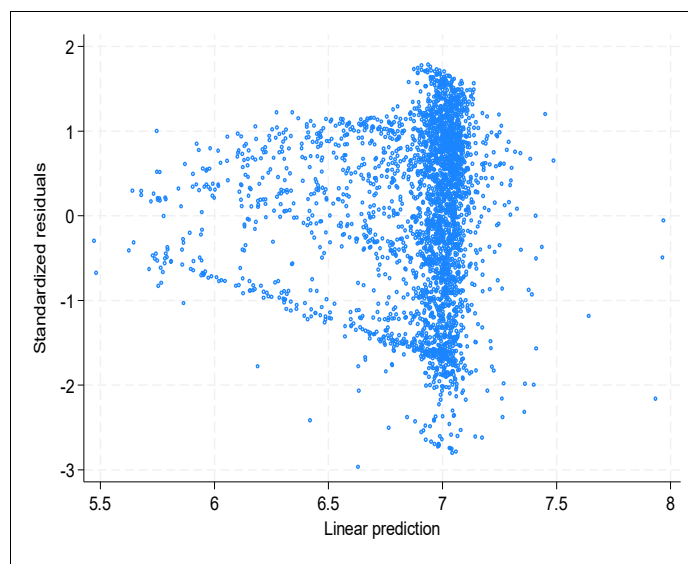
Exercise #3 – Solution

Linear Regression – Diagnostics

This lab is designed to help you recall how to check the overall and case-by-case fit of the model to the data. Remember we do this to validate our model but also to ensure that we learn about the associations, and what influences them, as we go.

Begin by running the model with `-intvl_ln-` as the outcome and `-p_rct-`, `-hdsiz-`, `-pyear_ct-` and `-pyear_sq-`.

1. Evaluate the assumption of homoscedasticity both graphically and statistically. What do you conclude? Explain the pattern of observations.



The standardized residuals show an apparent “fan” (or “cone”) pattern with larger variance at higher predicted values than at low predicted values. In reality, this is not easy to assess in a plot with so many points. The sheer volume of points with linear predictor values around 7 will naturally lead to a larger spread in the residuals in that range, without necessarily corresponding to . One option to explore this further is to stratify the points into suitable intervals based on the fitted values, and then compute standard deviations of the standardized residuals in each of those intervals; we omit the details. The presence of heteroscedasticity is confirmed by different statistical tests (of which we show the BPCW test).

```
. estat hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
```

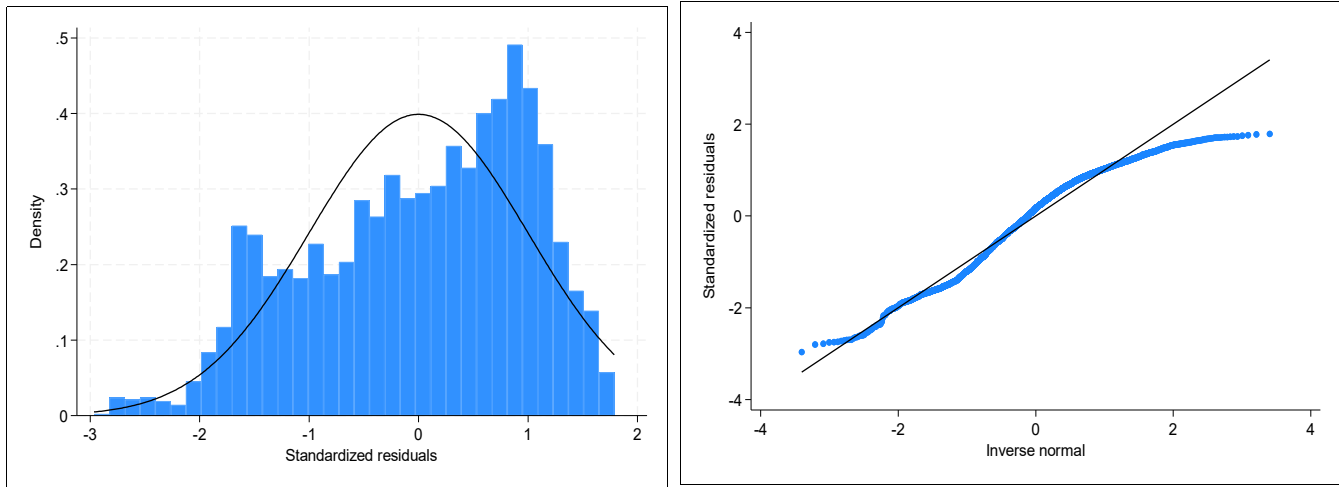
```
Ho: Constant variance
```

```
Variables: fitted values of intvl_ln
```

```
chi2(1)      =      52.46
```

```
Prob > chi2  =      0.0000
```

2. Evaluate the assumption of normality both graphically (using both a histogram and a normal probability plot) and statistically. What is your interpretation?



Both the histogram and the normal probability plot show departures from normality for the residuals (e.g., their distribution is left-skewed). This is confirmed by the Shapiro-Wilk test for normality being highly significant (note that when the significance assessment is so clear, any approximation related to the reference distribution and the residuals not being independent can be ignored).

```
. swilk stdres
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
stdres	2,987	0.96157	65.570	10.793	0.00000

3. Correcting problems of overall fit: It appears that we have problems with the overall fit. Often if we can correct one of the problems, that will correct both problems. So let's look and see what we might do to obtain somewhat more normally-distributed residuals.

- (a) First, start with the original outcome variable `-intvl-` and carry out a Box-Cox analysis using the `-boxcox-` command to see if there is a better transformation than “ln”. Note that we have already transformed Y to `lnY`, so we should go back and rerun the model with `-intvl-` as the outcome for this question. What does this approach suggest?

```
. boxcox intvl p_rct hdsiz e pyear_ct pyear_sq
... output omitted
```

Log likelihood = -24913.103

Number of obs	=	2,987
LR chi2(4)	=	225.25
Prob > chi2	=	0.000

intvl	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/theta	.1645652	.0177373	9.28	0.000	.1298008 .1993296

The Box Cox procedure suggests that the transformation which will improve the compliance with model assumptions (normality and homoscedasticity of errors) is $y^{0.165}$. However, applying this transformation does not produce a model which is much more satisfactory in terms of either normality or homoscedasticity.

```
. estat hettest
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
```

```
Ho: Constant variance
```

```
Variables: fitted values of intvl_bc
```

```
chi2(1) = 83.25
```

```
Prob > chi2 = 0.0000
```

```
. swilk stdres
```

```
Shapiro-Wilk W test for normal data
```

Variable	Obs	W	V	z	Prob>z
stdres	2,987	0.97480	42.999	9.704	0.00000

(b) Second, try applying the Box Cox method to the variable `intvl_ln`. In this instance we should obtain a $\theta = 1.898$ so in approximate terms we should model $-\text{intvl_ln}^{-2}$. Does this help?

In this case, we are attempting a double transformation. First the data are log transformed and then squared. Unfortunately, it does not work either (results not shown). It would in any case have been difficult to interpret the resulting transformation.

4. Subject by subject analysis of fit

(a) We first look for poor fitting subjects (potential outliers). These will have standardized residual values >2 or <-2 . Are there any?

There are no standardized residuals above 2, but very many below -2. The 10 most extreme residuals below -2 are shown in the listing below.

```
. sort stdres
```

```
. list id intvl_ln p_rct hdsiz p_year fit stdres delres in 1/10, noobs clean
```

id	intvl_ln	p_rct	hdsiz	p_year	fit	stdres	delres
2489	3.526361	0	86	2000	6.6292	-2.963959	-2.967837
2357	4.110874	3	51	1995	7.042443	-2.799768	-2.802985
723	4.143135	1	24	1992	7.058354	-2.783901	-2.787059
2477	4.158883	1	44	1992	7.040209	-2.751371	-2.754408
668	4.158883	0	45	1993	7.03561	-2.74717	-2.750192

144	4.158883	3	84	1992	7.030485	-2.742186	-2.745189
445	4.127134	1	91	1995	6.979589	-2.724562	-2.727502
689	4.143135	1	78	1991	6.98419	-2.71317	-2.716069
231	4.158883	0	31	1996	6.983579	-2.698138	-2.700985
1473	4.158883	3	65	1990	6.981791	-2.695966	-2.698805

The outlier test based on the deletion residuals indicates that deletion residuals numerically greater than 4.31 could be considered as significant outliers for this dataset. Based on this test, there are no significant outliers. Note however, that the test is based on the model assumptions being met, and that is clearly not the case here.

- (b) Lets now look for high leverage subjects; these could have big impact on the model. Look for cases with extreme values. Are there many? What are their characteristics?

The leverage threshold value is = 0.00335 (or 0.0050 if you consider the more conservative threshold). There are quite a few observations with leverage values above the threshold. In general, they are either quite large herds or ones with large values of p_rct.

```
. sort lev
. list id intvl_ln p_rct hdsiz p_year fit stdres lev in -23/-14, noobs clean
```

id	intvl_ln	p_rct	hdsiz	p_year	fit	stdres	lev
223	6.814543	0	366	2002	6.129432	.6600047	.0179965
222	5.749393	2	435	2000	6.339141	-.5702394	.0252234
221	5.47227	7	450	1999	6.491442	-.9855473	.0254031
2843	7.056175	55	280	1997	7.435472	-.3669779	.0264391
2901	5.313206	56	301	1998	7.361374	-1.982958	.027724
220	6.426488	65	299	1995	7.640965	-1.181592	.037215
62	5.231109	66	401	2001	7.104988	-1.826807	.0410773
2075	7.463937	74	105	1993	7.964081	-.4905893	.0528041
436	5.752573	86	169	1989	7.933545	-2.159655	.0705684
2065	7.914252	102	472	1991	7.968274	-.0542788	.0972719

```
list id intvl_ln p_rct hdsiz p_year fit stdres delres in 1/10, noobs clean
```

- (c) We now move to identifying subjects that actually have a large influence on the model (using either or both of Cook's D or Dfits). Are there many? Do any in particular stand out?

Leverage is only a measure of "potential influence". However, the same pattern holds for observations with large Cook's D or Dfits ... they are generally large herds or have high values of -p_rct-. The threshold values for Cook's D and dfits are 0.00134 and 0.0818, respectively. The listing below includes observations at larger cutoffs (subjectively set after inspecting the variables):

```
. sort dfit
. list id intvl_ln p_rct hdsiz p_year stdres fit cook dfit if (cook>0.005 |
abs(dfit)>0.15) & dfit~=., noobs clean
```

id	intvl_ln	p_rct	hdsiz	p_year	stdres	fit	cook	dfit
436	5.752573	86	169	1989	-2.159655	7.933545	.0708258	-.5954536
62	5.231109	66	401	2001	-1.826807	7.104988	.0285913	-.3782442
2901	5.313206	56	301	1998	-1.982958	7.361374	.0224246	-.3350123
220	6.426488	65	299	1995	-1.181592	7.640965	.0107933	-.2323223

1159	4.941642	27	46	1991	-2.315489	7.358566	.0076161	-.1952855
2651	4.343805	1	221	1992	-2.427726	6.87963	.006738	-.1836994
64	5.31812	30	83	1993	-1.994868	7.39961	.0062445	-.176787
60	4.477337	4	250	1999	-2.064783	6.63304	.0056838	-.1686724
2308	4.41884	18	190	1993	-2.605849	7.143147	.0053296	-.1634015
142	4.779123	24	107	1991	-2.377164	7.263377	.0053231	-.1632694
221	5.47227	7	450	1999	-.9855473	6.491442	.0050634	-.1591132
2205	5.241747	26	158	1989	-1.824712	7.146574	.004603	-.151767
12	8.701679	24	267	1990	1.557775	7.077463	.0045312	.1505549
4	8.699348	42	120	1990	1.202291	7.449911	.0046325	.1522036

- (d) Finally, because `-p_rct-` is our main exposure of interest, we might see if any of the observations have a particularly large influence on the estimate of that coefficient. Again, what do you find?

As expected, the observations which have the most influence on the coefficient for `-p_rct-` are herds which had high values for this predictor. The five largest `dbetas` for this predictor had more than 30 reactors (although some herds with larger number of reactors had smaller `dbetas`).

```
. sort db_rct

. list id intvl_ln p_rct hdsz p_year stdres fit db_rct if abs(db_rct)>0.14 &
db_rct~=., noobs clean
```

id	intvl_ln	p_rct	hdsz	p_year	stdres	fit	db_rct
436	5.752573	86	169	1989	-2.159655	7.933545	-.5800523
62	5.231109	66	401	2001	-1.826807	7.104988	-.2870666
2901	5.313206	56	301	1998	-1.982958	7.361374	-.2746634
220	6.426488	65	299	1995	-1.181592	7.640965	-.2015166
1159	4.941642	27	46	1991	-2.315489	7.358566	-.1879207
64	5.31812	30	83	1993	-1.994868	7.39961	-.1696355
142	4.779123	24	107	1991	-2.377164	7.263377	-.1461776
390	5.777652	33	115	1993	-1.565816	7.410426	-.1410908
4	8.699348	42	120	1990	1.202291	7.449911	.1448241

- (e) Try refitting the model but leave out either the observations which have the biggest impact on overall fit, or the ones which most influence the estimate of the coefficient for `-p_rct-`. Does this make much difference?

Neither of these re-fits are very different from the original model.

Conclusions

At this point you might be quite despondent (probably verging on suicidal) about not being able to fix the problems of heteroscedasticity and non-normality. Three more advanced approaches that we could take to solving this problem are as follows.

1. Use robust standard errors which are less susceptible to violations of the major assumptions. (These will be introduced when we talk about clustered data.)

2. The data are really “time-to-event” data so survival analysis methods may be more applicable. (These are discussed later in the semester in VHM 812.)

3. We could categorize the outcome into 2 levels (short vs long) and fit a logistic model or into >2 levels and fit some sort of multinomial model. The former will be covered in this semester, the latter will not. However this approach entails the loss of a lot of potentially useful information.

Finally, we could also consider going back to Ireland and regenerate data that did not have the structural problem of the relationship between “year” and “interval” built in.