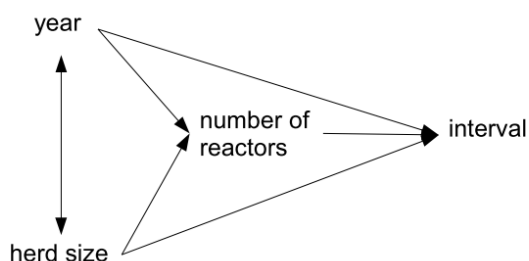


Linear Regression Exercise #2

Linear Regression – Model building

1. Draw a causal diagram showing how you think the three predictors might relate to the outcome and to each other.

Year and herd size are related because, in general, herds are getting larger over time. Whether this is presented as a causal relationship (one way arrow from year to herd size), or just as an association (two headed arrow) does not matter for this exercise. Both year and herd size may be related to number of reactors because the effectiveness of the control program maybe changing over the years and larger herds are more likely to have more reactors. All three factors may be related to the length of the interval between TB episodes.



2. Fit models with `-p_year-` as the only predictor. First fit it as a continuous variable and then as a categorical variable.

```
regress intvl_ln p_year /* continuous predictor */
```

Source	SS	df	MS	
Model	147.547346	1	147.547346	Number of obs = 3000
Residual	3376.80672	2998	1.12635314	F(1, 2998) = 131.00
Total	3524.35407	2999	1.17517641	Prob > F = 0.0000
				R-squared = 0.0419
				Adj R-squared = -0.0415
				Root MSE = 1.0613

intvl_ln	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
p_year	-.0478439	.0041802	-11.45	0.000	-.0560402 -.0396475
_cons	102.3055	8.336632	12.27	0.000	85.95939 118.6516

```
. regress intvl_ln i.p_year /* categorical predictor */
```

Source	SS	df	MS	
Model	260.386099	18	14.4658944	Number of obs = 3000
Residual	3263.96797	2981	1.09492384	F(18, 2981) = 13.21
Total	3524.35407	2999	1.17517641	Prob > F = 0.0000
				R-squared = 0.0739
				Adj R-squared = 0.0683
				Root MSE = 1.0464

intvl_ln	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
p_year					
1990	.021196	.0788123	0.27	0.788	-.133336 .1757279
1991	-.0087745	.0819222	-0.11	0.915	-.1694042 .1518552
1992	.1245631	.0853603	1.46	0.145	-.042808 .2919342
1993	.2724067	.094199	2.89	0.004	.087705 .4571084
1994	.2170947	.1062763	2.04	0.041	.0087123 .4254771
1995	.0149642	.1104147	0.14	0.892	-.2015325 .2314609
1996	.1282697	.1058112	1.21	0.226	-.0792007 .3357402
1997	-.0235158	.117306	-0.20	0.841	-.2535248 .2064932

1998		-.3849102	.1184682	-3.25	0.001	-.6171979	-.1526225
1999		-.3533293	.1046926	-3.37	0.001	-.5586064	-.1480521
2000		-.2175166	.1131298	-1.92	0.055	-.439337	.0043037
2001		-.146469	.1345969	-1.09	0.277	-.4103811	.1174432
2002		-.4536966	.1307225	-3.47	0.001	-.7100121	-.197381
2003		-.5163069	.1339179	-3.86	0.000	-.7788878	-.253726
2004		-.6939645	.1509123	-4.60	0.000	-.9898673	-.3980616
2005		-.9484242	.1696501	-5.59	0.000	-1.281067	-.615781
2006		-1.380731	.1696501	-8.14	0.000	-1.713374	-1.048087
2007		-1.910497	.6073476	-3.15	0.002	-3.10136	-.7196344
_cons		6.957966	.0624222	111.47	0.000	6.835571	7.080361

(a) Interpret the two intercepts.

The intercept from the model with p_year as a continuous predictor is the estimated log-interval in the year 0 AD (i.e., the year Christ was born). Clearly this is an illogical value to try to estimate when all of the data come from 1989-2007. The intercept from the model with p_year as a categorical predictor is the average log-interval in 1989. This is a more reasonable thing to estimate.

(b) How could you improve the “interpretability” of the intercept from the first model? Make the necessary change(s) to do this.

You could rescale p_year by subtracting 1989 (i.e., observations with $p_year=1989$ will become $p_year=0$) so that the intercept becomes the estimated value for that year (7.144)

intvl_ln	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pyear_1989	-.0478439	.0041802	-11.45	0.000	-.0560402	-.0396475
_cons	7.144043	.0294498	242.58	0.000	7.0863	7.201787

(c) Which is better, the original model (with $-p_year-$ as a continuous predictor) or this new model?

The model with $-p_year-$ as a categorical variable has a large R^2 , but this is achieved at a price. First, there are a lot of coefficients which are perhaps each of less interest. Second, there certainly appears to be a pattern (coefficients becoming increasingly negative), but the set of individual coefficients don't really describe that pattern). You could compute the F test statistic to compare these models (because the linear model is a submodel of the model using p_year as categorical).

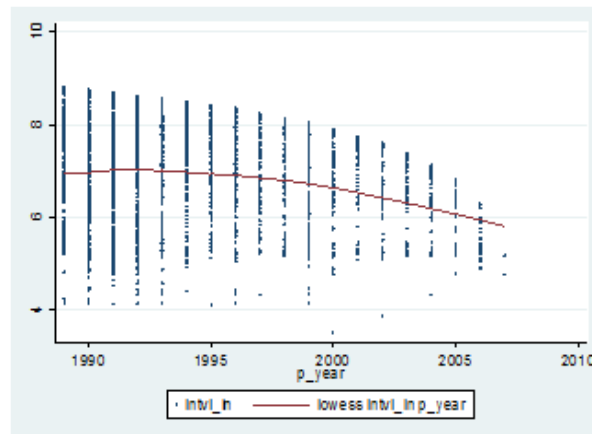
```
. scalar F=(3376.80672-3263.96797)/(2998-2981)/1.09492384
```

```
lack-of-fit F-test: F=6.0621326 P-value=4.551e-14
```

The tested null hypothesis (H_0) is that the simple model (p_year as continuous) shows the same fit as the more complex model (i.e., p_year as categorical). The very low p -value gives clear evidence against H_0 and therefore shows that the simple linear regression model doesn't fit the data as well as the categorical model. Perhaps the assumption of linearity inherent in the model with $-p_year-$ as a continuous variable needs to be checked.

3. We need to determine if the relationship between `-p_year-` and `intvl_ln-` is linear.

(a) Evaluate this by fitting a scatter plot with a lowess smoothed curve through the points.



The scatterplot and lowess smoothed curve clearly show a non-linear relationship. The annual change (reduction) in the length of the interval was much greater after 2000 than before.

(b) Also evaluate it by fitting both linear and quadratic terms to the model.

```
. generate pyear_sq = p_year^2
```

```
. regress intvl_ln p_year pyear_sq
```

Source	SS	df	MS	Number of obs =	3000
Model	222.485518	2	111.242759	F(2, 2997)	= 100.97
Residual	3301.86855	2997	1.10172457	Prob > F	= 0.0000
Total	3524.35407	2999	1.17517641	R-squared	= 0.0631
				Adj R-squared	= 0.0625
				Root MSE	= 1.0496

intvl_ln	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
p_year	30.95684	3.759349	8.23	0.000	23.58568 38.32801
pyear_sq	-.0077659	.0009416	-8.25	0.000	-.0096122 -.0059196
_cons	-30843.28	3752.19	-8.22	0.000	-38200.41 -23486.16

The quadratic term is highly significant. It is also worth noting that the predictive ability of the model (as indicated by the R^2) has gone up substantially (from the model with the single linear term). Clearly, the quadratic model fits better than the linear one. Note that the coefficient for the linear term has changed completely, and is now impossible to interpret meaningfully on its own.

4. Compute the VIFs from the model with the quadratic term. Do these signal a problem with the model?

```
. estat vif
```

```
Variable |          VIF          1/VIF
-----+-----
p_year   | 826858.31      0.000001
pyear_sq | 826858.31      0.000001
-----+-----
Mean VIF | 826858.31
```

The VIF values are huge (remember that anything over 10 is an indicator of substantial collinearity). Although the estimation and model fit don't seem to be affected of this, it is advisable to reparametrize to avoid a parameter (the intercept) being an extreme extrapolation.

- (a) Explore how the VIFs change after centring `-year-` before squaring it. (Call the new variable `pyear_ct-`.)

Centring can be understood specifically as subtracting the mean value of the predictor, or less specifically as subtracting a “central” value of the predictor, such as the approximate mean or the median or some other value in the center of the predictor distribution. For `-p_year-`, the mean equals 1994.3 and the median equals 1993, so subtracting either 1994 or 1993 could work fine. Here will instead center by subtracting the midpoint (1998) of the range 1989-2007. Despite being a bit off the arithmetic mean, centring by this value will both improve the interpretation of the coefficients (the intercept now corresponding to 1998 as we aim for) and reduce the VIFs.

```
. generate pyear_ct=p_year-1998
```

```
. replace pyear_sq=pyear_ct^2
(3,000 real changes made)
```

```
. regress intvl_ln pyear_ct pyear_sq
```

```
Source |          SS          df           MS      Number of obs   =      3,000
-----+-----
Model   | 222.485518            2    111.242759      F(2, 2997)       =    100.97
Residual| 3301.86855          2,997    1.10172457      Prob > F         =     0.0000
-----+-----
Total   | 3524.35407          2,999    1.17517641      R-squared        =     0.0631
                                           Adj R-squared    =     0.0625
                                           Root MSE       =     1.0496
```

```
-----+-----
intvl_ln |          Coef.      Std. Err.      t    P>|t|      [95% Conf. Interval]
-----+-----
pyear_ct |   -0.0757948      .0053458    -14.18  0.000    -0.0862767    -0.065313
pyear_sq |   -0.0077659      .0009416     -8.25  0.000    -0.0096122    -0.0059196
   _cons |    6.883048      .0319912    215.15  0.000     6.820321     6.945775
-----+-----
```

```
. estat vif
```

Variable	VIF	1/VIF
pyear_ct	1.67	0.598086
pyear_ctsq	1.67	0.598086
Mean VIF	1.67	

Although the main purpose of centring is to improve the interpretability of the constant, here it helps to reduce the problem of collinearity between the linear and quadratic terms. Two things to note: 1) the linear term is now more reasonable (-0.076), and 2) the quadratic term is exactly the same in both models.

5. Is the year (fit as both linear and quadratic terms) a confounder for the effect of herd size (-hdsiz-)?

```
. regress intvl_ln hdsiz /* model without year */
```

Source	SS	df	MS	Number of obs	=	2987
Model	7.24137996	1	7.24137996	F(1, 2985)	=	6.17
Residual	3504.58457	2985	1.17406518	Prob > F	=	0.0131
Total	3511.82595	2986	1.1760971	R-squared	=	0.0021
				Adj R-squared	=	0.0017
				Root MSE	=	1.0835

intvl_ln	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hdsiz	-.0010355	.000417	-2.48	0.013	-.0018531 - .000218
_cons	6.954334	.0324156	214.54	0.000	6.890774 7.017893

```
. regress intvl_ln hdsiz pyear_ct pyear_sq /* model with year */
```

Source	SS	df	MS	Number of obs	=	2,987
Model	223.892476	3	74.6308254	F(3, 2983)	=	67.71
Residual	3287.93347	2,983	1.10222376	Prob > F	=	0.0000
Total	3511.82595	2,986	1.1760971	R-squared	=	0.0638
				Adj R-squared	=	0.0628
				Root MSE	=	1.0499

intvl_ln	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hdsiz	-.0002863	.000412	-0.69	0.487	-.0010941 .0005216
pyear_ct	-.0754918	.0053892	-14.01	0.000	-.0860586 -.0649249
pyear_sq	-.0078035	.0009457	-8.25	0.000	-.0096579 -.0059491
_cons	6.903749	.0433748	159.16	0.000	6.818701 6.988796

```
. display "percent change=" abs(-.0002863-(-.0010355))/abs(-.0010355)*100
percent change=72.351521
```

```
. * added assessment of association hdsiz vs year (should really be adjusted for herds)
. regress hdsiz p_year
```

Source	SS	df	MS	Number of obs =	2987
Model	248014.759	1	248014.759	F(1, 2985) =	113.81
Residual	6504857.89	2985	2179.18187	Prob > F =	0.0000
Total	6752872.65	2986	2261.51127	R-squared =	0.0367
				Adj R-squared =	0.0364
				Root MSE =	46.682

hdsiz	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
p_year	1.965962	.1842821	10.67	0.000	1.604629 2.327295
_cons	-3859.25	367.5183	-10.50	0.000	-4579.865 -3138.635

Adding year to the model makes a substantial change in the estimate of the effect of herd size (it goes from -0.0010 to 0.0003, and becomes non-significant). Whether you consider year as a true confounder depends on whether you think that “year” is a cause of “herd size”. Given that herds are getting larger over time, I think this is a reasonable way to view the relationship.

- We want to evaluate an interaction, but to simplify this process, we are going to convert -p_year- into a dichotomous variable. There are many ways to do this, but the command that I used was....

```
. egen pyear_c2=cut(p_year), at(0 1999 2999) icodes
```

- Is there interaction between -p_rct- and this new dichotomous variable for year?

```
. regress intvl_ln pyear_c2##c.p_rct, vsquish
```

Source	SS	df	MS	Number of obs =	3000
Model	159.565065	3	53.188355	F(3, 2996) =	47.36
Residual	3364.789	2996	1.12309379	Prob > F =	0.0000
Total	3524.35407	2999	1.17517641	R-squared =	0.0453
				Adj R-squared =	0.0443
				Root MSE =	1.0598

intvl_ln	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
1.pyear_c2	-.4645391	.0577574	-8.04	0.000	-.5777874 -.3512909
p_rct	.0128656	.0036582	3.52	0.000	.0056927 .0200385
1.pyear_c2#c.p_rct					
1	-.0176227	.0081487	-2.16	0.031	-.0336003 -.001645
_cons	6.963248	.0253014	275.21	0.000	6.913638 7.012858

Yes. The interaction term is weakly significant ($P=0.031$), suggesting that the effect of $-p_rct-$ was different in the early years compared to the later years.

(b) What was the effect of p_rct in the early years (1989-1998)?

The model could be written as follows (using shortened variable names for simplicity)

$$Y = \beta_0 + \beta_1(pyc2) + \beta_2(p_rct) + \beta_3(pyc2 * p_rct)$$

When the coefficients are added, this becomes the following for early episodes:

$$Y = 6.96 - 0.465(0) + 0.013(p_rct) - 0.017(0 * p_rct) = 6.96 + 0.013(p_rct)$$

which means that for each additional reactor, the log interval was expected to get 0.013 log-days longer.

(c) What was the effect of p_rct in the later years (1999 – 2007)?

For later episodes:

$$Y = 6.96 - 0.465(1) + 0.013(p_rct) - 0.017(1 * p_rct) = 6.495 + 0.013(p_rct) - 0.017(p_rct) = 6.495 - 0.004(p_rct)$$

which means that for each additional reactor, the log interval was expected to get 0.004 log-days shorter.

7. To simplify things, we will carry on using the dichotomous version of $-p_year-$ ($pyear_c2$)

(a) If your main outcome of interest is p_rct , should either of the other two predictors be included in a regression model? Why?

Yes – they are both potential confounders. Also, $-pyear_c2-$ must be included because it interacts with $-p_rct-$

(b) What is your final model? What do you conclude about the relationship between p_rct and $intvl_ln$?

```
. regress intvl_ln hdsiz pyear_c2##c.p_rct, vsquish
```

Source	SS	df	MS	Number of obs =	2987
Model	164.348406	4	41.0871015	F(4, 2982) =	36.60
Residual	3347.47754	2982	1.12256122	Prob > F =	0.0000
Total	3511.82595	2986	1.1760971	R-squared =	0.0468
				Adj R-squared =	0.0455
				Root MSE =	1.0595

intvl_ln	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]

hdsizes		-.000811	.0004466	-1.82	0.070	-.0016867	.0000648
1.pyear_c2		-.4590911	.0580284	-7.91	0.000	-.5728707	-.3453114
p_rct		.0149895	.0038641	3.88	0.000	.0074129	.0225661
pyear_c2#c.p_rct							
1		-.0166078	.0081597	-2.04	0.042	-.0326071	-.0006085
_cons		7.003838	.0331564	211.24	0.000	6.938826	7.06885

The final model is shown above. The effect of $-p_rct$ is very similar to what is shown in question 6 so the calculations won't be repeated. The margins and marginplot commands were used to generate the graph below showing the relationship between p_rct and $-\ln$ for the two categories of $pyear_c2$. Since there is one graph of each herd size the graph was created for a median herd size value (50).

```
.margins pyear_c2, at(hdsizes=50 p_rct=(0(10)40)) expression(exp(predict(xb)))
marginplot
```

