Further topics in linear regression

Part 1

David Barron

Hilary Term 2018

Introduction

Introduction

This class will meet weekly for 16 weeks. In odd-numbered weeks I will introduce a substantive topic in quantitative research methods in a 2-hour session. In even-numbered weeks, I will run "labs", where the focus will be on putting the theory into practice using R. These sessions will be 1 hour.

You can find the handouts (including the R code used to generate them) on my personal Weblearn site: https://weblearn.ox.ac.uk/x/MbYn1T. Datasets that are used in demonstrations and in the labs are also here. These should be available to anyone with an Oxford single sign on username and password, but let me know if you have any problems.

The class will be assessed by means of an assignment, but that won't be due in until after the end of Trinity Term. You will be required to carry out an analysis of a dataset of your choice using any of the methods that we have covered in the course.

Overview of weeks 1-4

- Review of multiple regression
- Modelling
  - Dummy variables
  - Interactions
- Regression diagnostics
  - Normality of residuals
  - Collinearity
  - Model selection
  - Outliers
  - Heteroskedasticity
  - Linearity
- Sample selection bias

Regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i,$$

for $i = 1, \ldots, n$ sampled observations. $\epsilon_i \sim \text{NID}(0, \sigma^2)$.

Fitted model

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki};$$

$$y_i = \hat{y}_i + e_i,$$

where b_j are estimates of the corresponding β_j, and the e_i are residuals.

Ordinary Least Squares (OLS) estimates of b_j are those that minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Modelling

Dummy Variables

What are dummy variables?

Often we want to use explanatory variables in regressions that are categorical. To do this, we have to use *dummy variables*. Which category a particular observation falls in to is identified by a series of binary (0/1) variables, one fewer variables than there are categories. That's because there is always one category that does not give us any additional information: if someone isn't a man, they must be a woman and hence we only need a variable identifying whether someone is a man (dummy variable $= 1$) or isn't a man (dummy variable $= 0$). How, though, do we interpret the parameter estimates associated with dummy variables?

Simulated data

In this example, we have a categorical variable with 4 categories and a continuous variable that are related to a dependent variable in the following way.

$$y = -.7x_1 - .2x_2 + .3x_3 + .9x_4 + .4x_c + \epsilon(0, 2)$$

We first perform a regression of y on the continuous variable, x_c, only.

Regression with continuous variable only

```
Call:
lm(formula = y ~ xc, data = dta)

Residuals:
   Min     1Q Median     3Q    Max
 -6.15  -1.47   0.01   1.37   6.66

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1095     0.0648    1.69    0.092
xc            0.3753     0.0126   29.73   <2e-16

Residual standard error: 2.05 on 998 degrees of freedom
Multiple R-squared:  0.47,  Adjusted R-squared:  0.469
F-statistic:  884 on 1 and 998 DF,  p-value: <2e-16
```

First category excluded

```
Call:
lm(formula = y ~ xfac + xc, data = dta)

Residuals:
    Min      1Q  Median      3Q     Max
-5.270 -1.399  0.013   1.361   6.473

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.8110     0.1233   -6.58  7.7e-11
xfacB         0.6967     0.1743    4.00  6.9e-05
xfacC         1.2551     0.1743    7.20  1.2e-12
xfacD         1.7307     0.1748    9.90  < 2e-16
xc            0.3852     0.0121   31.94  < 2e-16

Residual standard error: 1.95 on 995 degrees of freedom
Multiple R-squared:  0.522, Adjusted R-squared:  0.52
F-statistic:  272 on 4 and 995 DF,  p-value: <2e-16
```

Last category excluded

```
Call:
lm(formula = y ~ xfac + xc, data = dta)

Residuals:
    Min      1Q  Median      3Q     Max
-5.270 -1.399  0.013   1.361   6.473

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9198     0.1236    7.44  2.2e-13
xfacA        -1.7307     0.1748   -9.90  < 2e-16
xfacB        -1.0340     0.1750   -5.91  4.7e-09
xfacC        -0.4756     0.1745   -2.73   0.0065
xc            0.3852     0.0121   31.94  < 2e-16

Residual standard error: 1.95 on 995 degrees of freedom
Multiple R-squared:  0.522, Adjusted R-squared:  0.52
F-statistic:  272 on 4 and 995 DF,  p-value: <2e-16
```

What is the relationship between the two?

Category	*A* excluded	*D* excluded
A	-0.81	$0.92 - 1.73 = -0.81$
B	$-0.81 + 0.70 = -0.11$	$0.92 - 1.03 = -0.11$
C	$-0.81 + 1.26 = 0.44$	$0.92 - 0.48 = 0.44$
D	$-0.81 + 1.73 = .92$	0.92

Parameter estimates give how much that category differs from the *excluded category*.

Interpreting t-values

Because parameter estimates depend on the arbitrary choice of excluded category, you can't interpret the $t-$values associated with each estimate in the usual way. To determine whether a dummy variable is statistically signficant, it is conventional to use an $F-$ test, using the formula:

$$\frac{(RSS_r - RSS_c)/p}{RSS_c/(n-k-p-1)},$$

where RSS_r is the residual sum of squares (RSS) from the regression without dummy variables, RSS_c is the SSR from the complete model, p is the number of extra parameters in the complete model, n is the sample size, k is the number of variable in the restricted model (not counting the constant).

Example

We can get the numbers we want by using the `anova` function in `R`.

```
anova(cont, xf1)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
998	4192	NA	NA	NA	NA
995	3778	3	414	36.3	0

You can simply read off the test from this, but you might want to check the formula above using the RSS numbers.

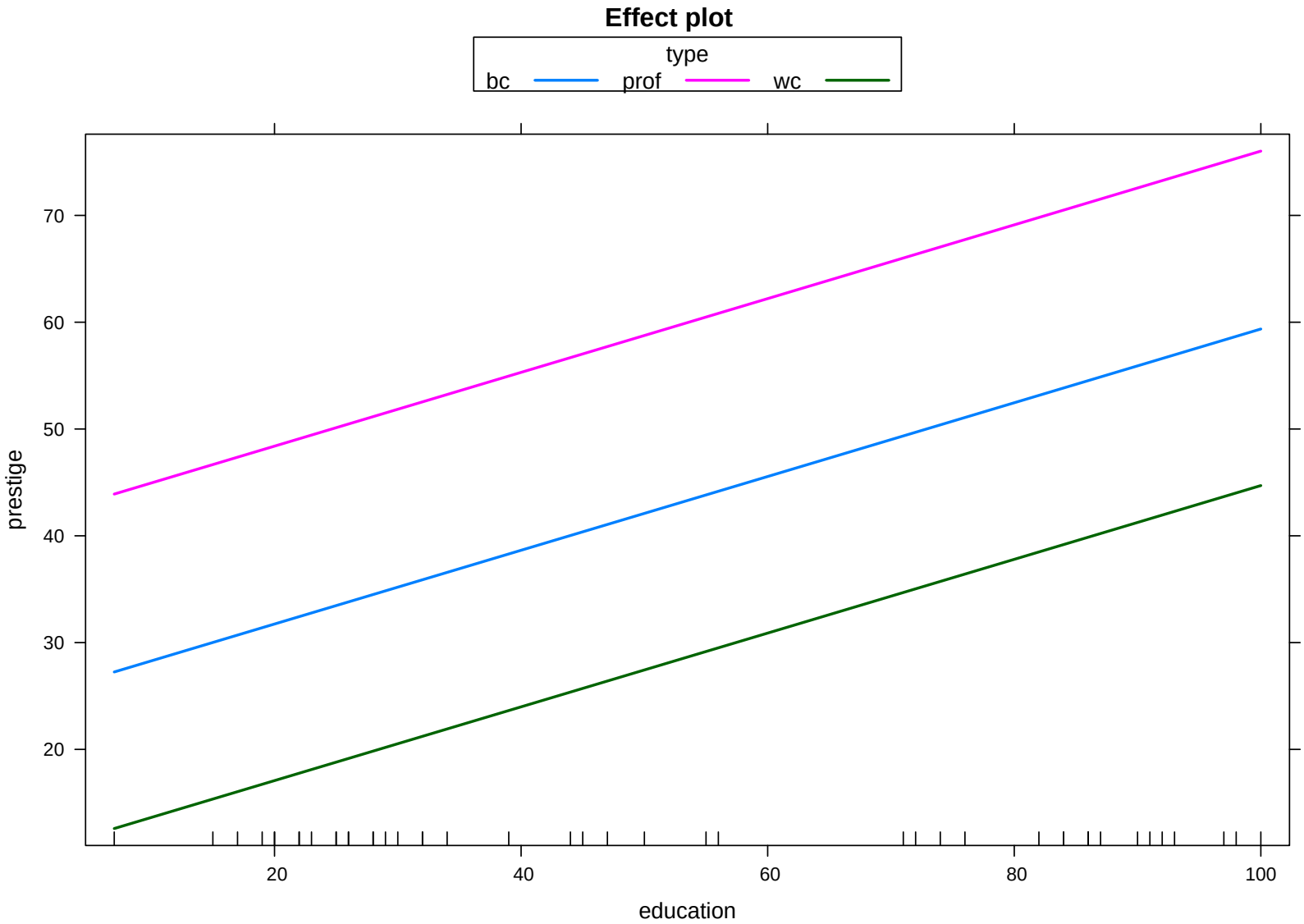
Duncan's occupational prestige data

This example uses data on occupational prestige. The outcome variable is the percentage of survey respondents who rated an occupation's prestige **excellent** or **good**. The explanatory variables are *income*, which is the percentage of males in the occupation earning $3500 or more in 1950; *education*, the percentage of males in the occupation in 1950 who were high school graduates; and *type*, which is a factor distinguishing occupations that are *professions*, *white collar* or *blue collar*.

Regression output

```
Call:
lm(formula = prestige ~ income + education + type, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-14.89  -5.74  -1.75    5.44  28.97

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1850     3.7138   -0.05   0.9605
income        0.5975     0.0894    6.69  5.1e-08
education     0.3453     0.1136    3.04   0.0042
typeprof     16.6575     6.9930    2.38   0.0221
typewc      -14.6611     6.1088   -2.40   0.0211

Residual standard error: 9.74 on 40 degrees of freedom
Multiple R-squared:  0.913, Adjusted R-squared:  0.904
F-statistic:  105 on 4 and 40 DF,  p-value: <2e-16
```

Effect plot


Effect plot
type
bc
prof
wc
prestige
70
60
50
40
30
20
20
40
60
80
100
education

Interpretation

You can see that dummy variables have the effect of shifting estimated regression lines up or down. The lines are parallel to each other. Here we can see that occupational prestige increases with education and that at all levels of education, estimated prestige is lowest for white collar jobs and highest for professional occupations.

Interactions

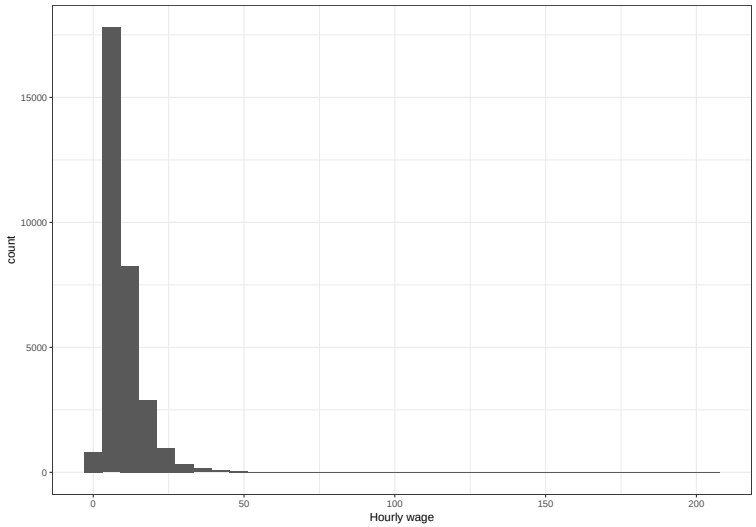
Motivation

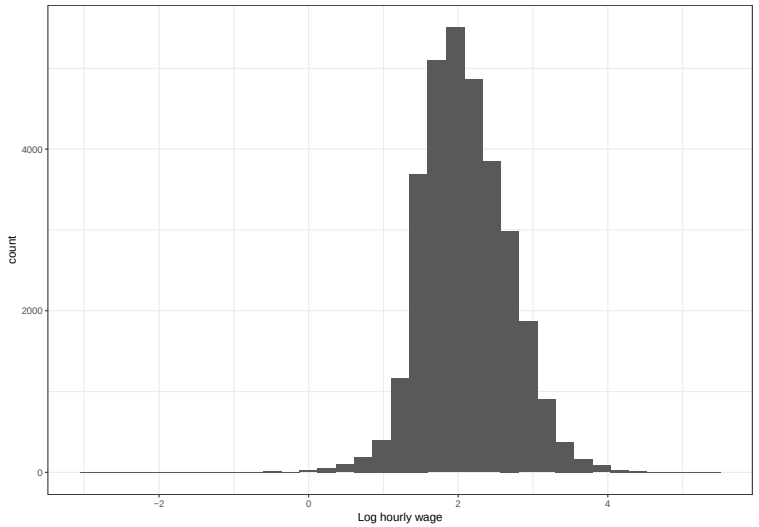
The standard linear regression model implies that the size of the effect of any given explanatory variable on the outcome variable is the same at all values of the other explanatory variables. What do we do if we think that is not true? For example, the effect of marital status and number of children on wages may be different for men and women. The standard way of incorporating such interactions is to multiply two variables together:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

Example: Labour Force Survey data

Data from the UK Labour Force Survey gives information about wages as well as age, gender, marital status and number of children. Wages are tranformed to an hourly basis, and then logged because otherwise they would be very skewed.





Results without interactions

```
Call:
lm(formula = Loghourpay ~ sex + age + allchildren + married,
    data = lfs)

Residuals:
   Min     1Q Median     3Q    Max
-5.111 -0.377 -0.039  0.362  3.012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.625512   0.011926  136.30   <2e-16
sexmale     0.252642   0.006236   40.52   <2e-16
age         0.007007   0.000306   22.92   <2e-16
allchildren 0.009708   0.003519    2.76   0.0058
marriedyes  0.113604   0.007628   14.89   <2e-16

Residual standard error: 0.552 on 31405 degrees of freedom
  (32149 observations deleted due to missingness)
Multiple R-squared:  0.0932,    Adjusted R-squared:  0.0931
F-statistic:  807 on 4 and 31405 DF,  p-value: <2e-16
```

All of these estimates are statistically signficant, although the overall model fit is pitiful!

Results with interactions

```
Call:
lm(formula = Loghourpay ~ sex + age + allchildren + married +
    sex:married + sex:allchildren, data = lfs)

Residuals:
   Min     1Q Median     3Q    Max
-5.070 -0.372 -0.041  0.356  2.964

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.714088   0.012730  134.65  < 2e-16
sexmale              0.099962   0.010108    9.89  < 2e-16
age                  0.006595   0.000305   21.64  < 2e-16
allchildren         -0.016965   0.004836   -3.51  0.00045
marriedyes           0.019848   0.009579    2.07  0.03827
sexmale:marriedyes   0.206089   0.013061   15.78  < 2e-16
sexmale:allchildren  0.044640   0.006665    6.70  2.2e-11

Residual standard error: 0.549 on 31403 degrees of freedom
  (32149 observations deleted due to missingness)
Multiple R-squared:  0.104, Adjusted R-squared:  0.104
F-statistic:  605 on 6 and 31403 DF,  p-value: <2e-16
```

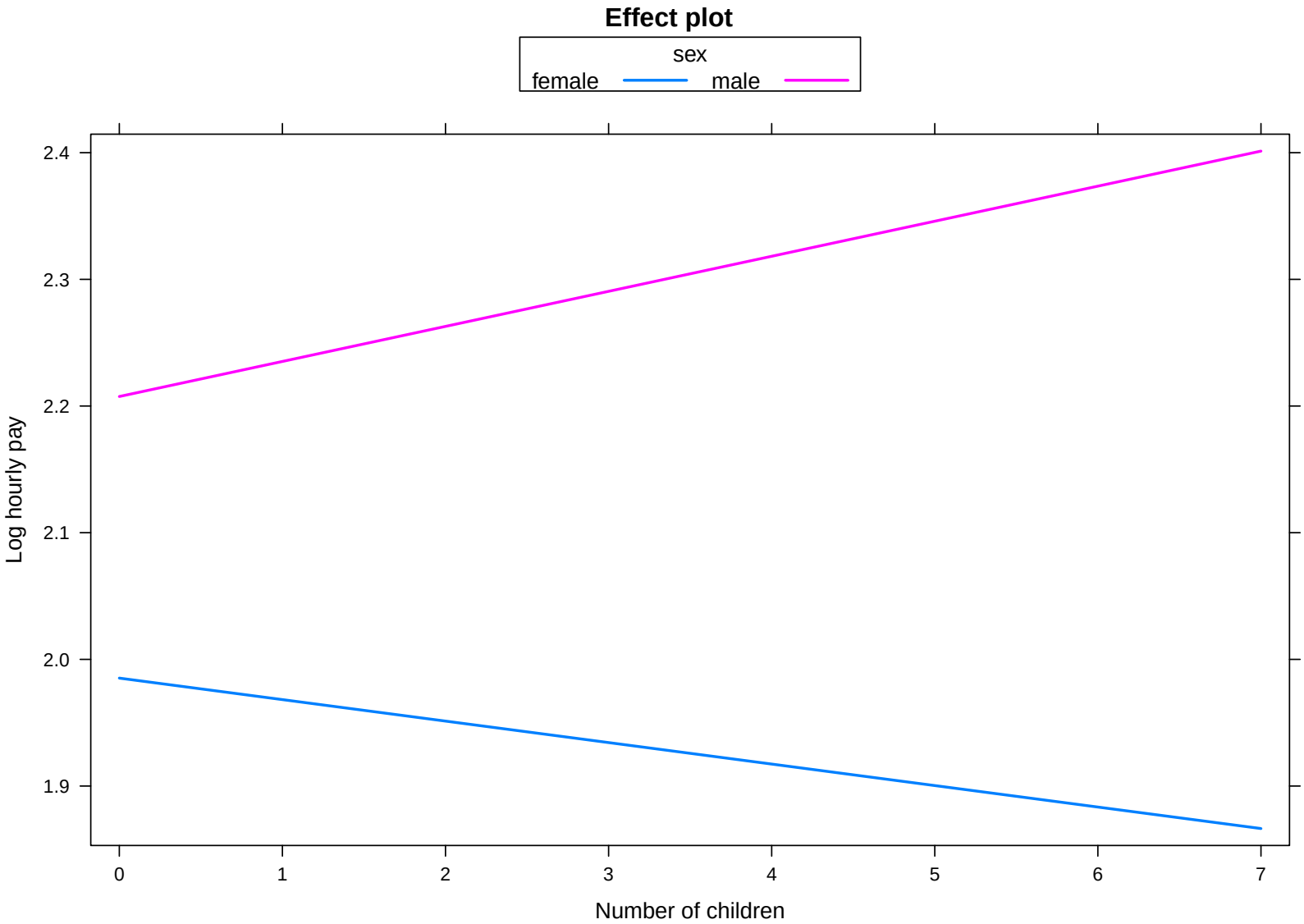
Interpretation of interactions

There are two additional parameter estimates, representing the interaction of sex and marital status, and sex and number of children, respectively. We can work out the effect of being married on log hourly wages for men and women as follows:

Category	No interaction	With interaction
Unmarried women	1.626	1.714
Married women	1.739	1.734
Unmarried men	1.878	1.814
Married men	1.992	2.04

You can see that in the first column the difference between being married and unmarried is the same for men and women, but in the second column the differences are much bigger for men than for women.

Interpretation of interactions 2



Regression diagnostics

Normality of residuals

Assumption

The standard assumption of linear regression is that the errors are normally distributed. If they are not, you will still get unbiased estimates of the regression parameters. However, the estimates will not (necessarily) be as efficient as they could be (i.e., standard errors will be larger than they need to be). Hypothesis testing (which relies on us knowing the sampling distribution of estimates) also depends on normality assumption being met.

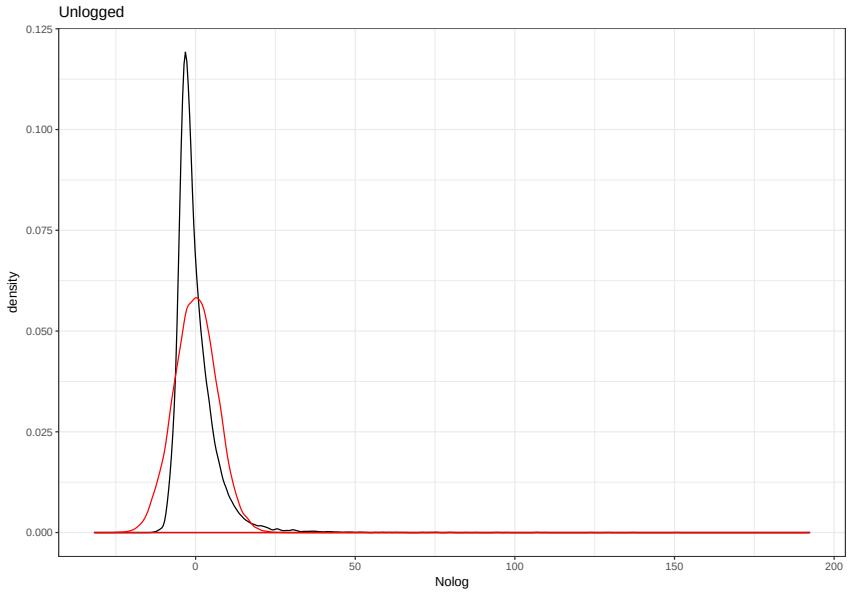
Example

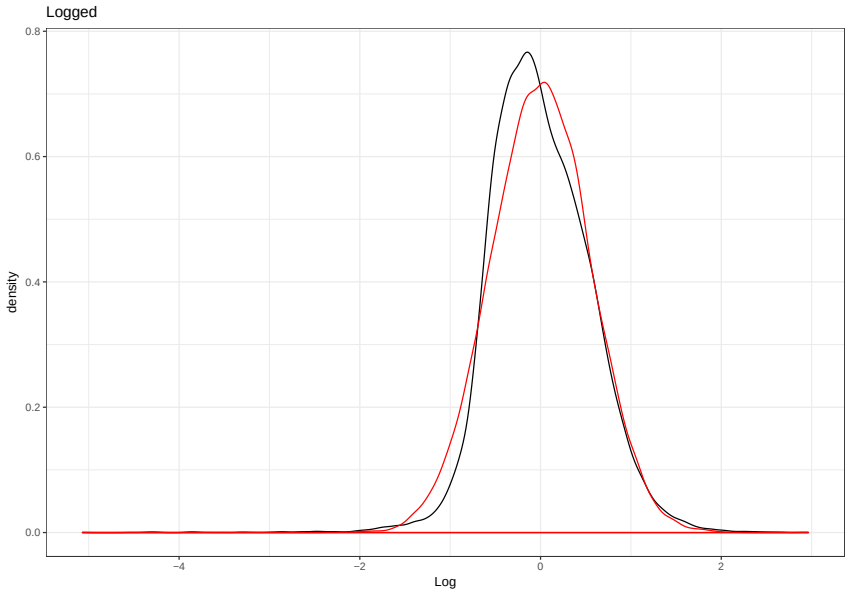
As an example, look at the Labour Force Survey data again but do the regression without taking logs of hourly pay.

```
Call:
lm(formula = hourpay0 ~ sex * married + age + sex * allchildren,
    data = lfs)

Residuals:
   Min     1Q Median     3Q    Max
-13.05  -3.79  -1.64   2.11 192.34

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.88800    0.15596   37.75  < 2e-16
sexmale              1.02923    0.12383    8.31  < 2e-16
marriedyes           0.00190    0.11736    0.02     0.99
age                  0.06402    0.00373   17.14  < 2e-16
allchildren         -0.05126    0.05924   -0.87     0.39
sexmale:marriedyes   2.34017    0.16001   14.62  < 2e-16
sexmale:allchildren  0.43559    0.08166    5.33  9.7e-08

Residual standard error: 6.73 on 31403 degrees of freedom
  (32149 observations deleted due to missingness)
Multiple R-squared:  0.0764,    Adjusted R-squared:  0.0763
F-statistic:  433 on 6 and 31403 DF,  p-value: <2e-16
```

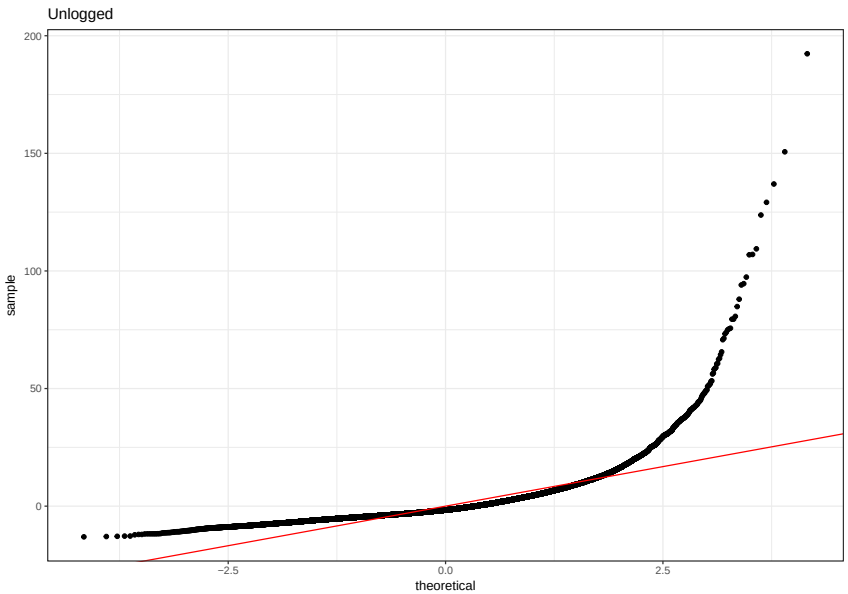
Density plot

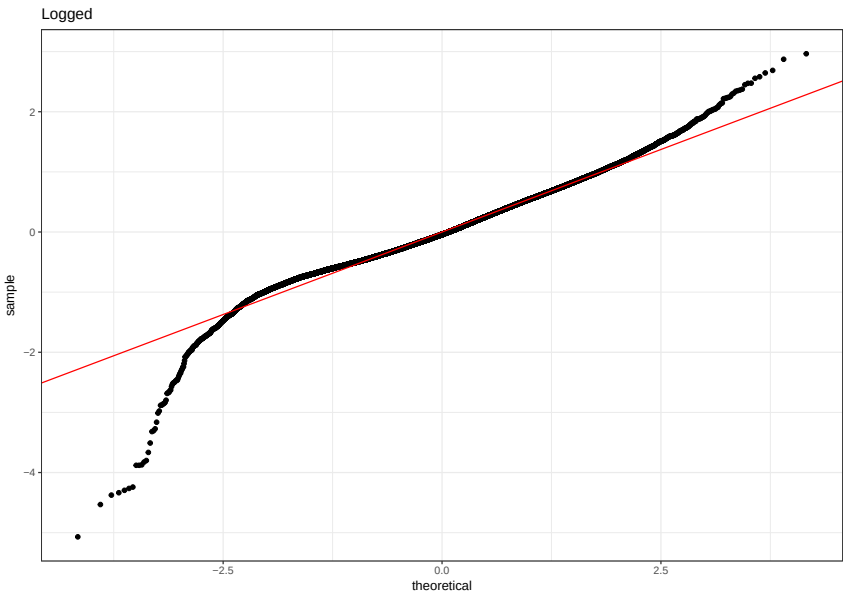
A density plot is a kind of "continuous histogram". You can compare the distribution of residuals with a normal distribution with the same mean and standard deviation.





QQ-plot


Unlogged
sample
theoretical
200
150
100
50
0
−2.5
0.0
2.5


Logged
sample
theoretical
2
0
−2
−4
−2.5
0.0
2.5

Multi-collinearity

Definition

(Multi-)collinearity is the problem of two or more explanatory variables not being independent of each other. Strictly speaking, this is not a violation of the assumptions of the linear regression model, but when collinearity becomes very high, estimated standard errors become very high and in some circumstances regression parameter estimates can be difficult to obtain. One way to measure collinearity relies on R_i^2, the proportion of the variance of the i th explanatory variable that is associated with the other explanatory variables in the model. That is, if the regression model is

$$y = b_0 + b_1 x_1 + b_2 x_x + \cdots + b_k x_k + e,$$

then we regress one explanatory variable on the others:

$$x_1 = c_0 + c_2 x_2 + \cdots + c_k x_k + e$$

and find the R^2 of this second regression.

Variance inflation factor and tolerance

More commonly, two statistics that are derived from R_i^2 are reported.

- **Tolerance** $= (1 - R_i^2)$;
- **Variance inflation factor** $= 1/(1 - R_i^2)$.

The VIF is thus the reciprocal of the tolerance. The VIF (or its square root) is the most commonly reported statistic because it is the impact on the estimated variance (or standard error) of parameter estimates that we are usually most concerned about:

$$\sigma^2(b_i) = \frac{\sigma_\epsilon^2}{\sum x_i^2} \times \text{VIF}$$

A common rule of thumb is that a VIF of 10 or above is a source of concern. However, treat such rules with caution, as it is possible to make matters worse by using common "solutions."

Example

Calculate the VIF for the Labour Force Survey regression above:

```
                   VIF
sex               2.66
age               1.40
allchildren       2.24
married           2.31
sex:married       3.69
sex:allchildren   2.71
```

You can see that all these VIFs are quite small, so (despite there being two interaction effects, where collinearity can sometimes be a problem), we don't have any concerns about this. What do we do if there is evidence of high collinearity, though?

Solutions?

In many cases, there is no straightforward solution; if variables are highly collinear, that's just the way the world is and you can't change it no matter how inconvenient it may be. For example, it might be difficult to separate the impact of age and years of experience on wages. It is increasing the risk of failing to reject a null hypothesis even if it is false, so if estimates are significant anyway, you're OK. If you need to reduce the impact of collinearity, there are a few possibilities.

- Collect more data. This reduces standard errors, but it may not be practical.
- Combine two or more explanatory variables into a single indicator. Only an option in (rare) cases where this would make theoretical sense.
- Remove one or more variables from the regression. This is very risky, and introduces the broader question of how to select the "best" regression model.

Misspecification bias

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9440 -0.6652 -0.0507  0.7705  1.8246

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.725      0.414    8.99  1.3e-09
x1             0.280      0.173    1.62     0.12
x2            -1.080      0.177   -6.11  1.6e-06

Residual standard error: 1.02 on 27 degrees of freedom
Multiple R-squared:  0.746, Adjusted R-squared:  0.727
F-statistic: 39.6 on 2 and 27 DF,  p-value: 9.3e-09
```

```
Call:
lm(formula = y ~ x2)

Residuals:
   Min     1Q Median     3Q    Max
-2.437 -0.577  0.138  0.677  1.792

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3458     0.3514    9.52  2.8e-10
x2           -0.8393     0.0986   -8.51  3.0e-09

Residual standard error: 1.05 on 28 degrees of freedom
Multiple R-squared:  0.721, Adjusted R-squared:  0.711
F-statistic: 72.4 on 1 and 28 DF,  p-value: 2.99e-09
```

Note

This is based on simulated data with $b_0 = 4$, $b_1 = 0.5$ and $b_2 = -1.3$. The two explanatory variables are strongly correlated. Removing x_1 from the analysis because it is not statistically significant introduces bias in the estimate of b_2.

Things to look out for

- Large change in the parameter estimate of b_2 across the two regressions.
- Large change in the R^2 across the two regressions.
- **Most important** is your theory; make decisions based on theory, not by blindly following some statistical "rule."
- Consider using one of the step-wise regression methods as an aid to model building.
    - These are particularly appropriate when you are building models with the primary purpose of prediction

Stepwise regression

The basic idea of stepwise regression is to identify a subset of potential explanatory variables that explain as much variance as possible in the outcome variable as parsimoniously as possible. There are two possible approaches:

- We start with a minimal model and add variables until there is no improvement in fit;
- We start with all possible variables and remove them until there is no deterioration in fit.

The criterion most commonly used to assess fit is the Akaike information criterion (AIC), which is smaller the better fitting the model, taking account of the number of parameters being estimated.

Example

Data on credit histories of 1,319 applicants for credit cards. The outcome variable is the number of major negative reports. *Age* in years; *Income* in US dollars/10,000; *Share* is ratio of monthly credit card expenditure to yearly income; *Owner* is a factor, whether a home owner; *Dependents* is number of dependents; *Months* at current address.

```
Call:
lm(formula = reports ~ age + income + share + owner + dependents +
    months, data = CreditCard)

Residuals:
   Min     1Q Median     3Q    Max
-1.027 -0.575 -0.415 -0.074 13.416

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.474032   0.139941    3.39  0.00073
age          0.004181   0.004345    0.96  0.33617
income       0.006233   0.024055    0.26  0.79557
share       -2.158597   0.389605   -5.54  3.6e-08
owneryes    -0.238853   0.083708   -2.85  0.00439
dependents   0.024947   0.031925    0.78  0.43469
months       0.000929   0.000617    1.50  0.13270

Residual standard error: 1.33 on 1312 degrees of freedom
Multiple R-squared:  0.033, Adjusted R-squared:  0.0286
F-statistic: 7.47 on 6 and 1312 DF,  p-value: 6.88e-08
```

Backwards elimination

```
Call:
lm(formula = reports ~ share + owner + months, data = CreditCard)

Residuals:
   Min     1Q Median     3Q    Max
-1.049 -0.568 -0.420 -0.088 13.422

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.628971   0.061066   10.30  < 2e-16
share       -2.230742   0.386298   -5.77  9.6e-09
owneryes    -0.188593   0.075701   -2.49    0.013
months       0.001155   0.000568    2.03    0.042

Residual standard error: 1.33 on 1315 degrees of freedom
Multiple R-squared:  0.0315,    Adjusted R-squared:  0.0293
F-statistic: 14.2 on 3 and 1315 DF,  p-value: 3.88e-09
```

Forwards addition

```
Call:
lm(formula = reports ~ share + owner + months, data = CreditCard)

Residuals:
   Min     1Q Median     3Q    Max
-1.049 -0.568 -0.420 -0.088 13.422

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.628971   0.061066   10.30  < 2e-16
share       -2.230742   0.386298   -5.77  9.6e-09
owneryes    -0.188593   0.075701   -2.49    0.013
months       0.001155   0.000568    2.03    0.042

Residual standard error: 1.33 on 1315 degrees of freedom
Multiple R-squared:  0.0315,    Adjusted R-squared:  0.0293
F-statistic: 14.2 on 3 and 1315 DF,  p-value: 3.88e-09
```

In this case, both methods give the same answer, which adds to our confidence. This isn't always the case. Care needs to be taken using stepwise methods; there is no substitute for thinking!