Statistical models in R exercise solutions

Applied Statistics and Data Science Group
with contributions from Yue Liu and Kim Dill-McFarland
U. of British Columbia

version March 14, 2019

Contents

Setup

We will be working with the same data and packages as in the notes and main.R files.

```
# Suite of packages for data manipulation and visualization
library(tidyverse)
# puts output of statistical models in a nice data frame
library(broom)
# Split, process, and recombine data
library(plyr)
# Fit linear and generalized linear mixed-effects models
library(lme4)
# various functions, including Anova
library(car)
```

```r
# least-square means
library(lsmeans)
## generalized linear model fitter
## Also has quine data  set
library(MASS)

# Data set libraries
## Fruitfly longevity, size, and sexual activity
library(faraway)
## Life expectancy, GDP and population by country
library(gapminder)
## A variety of data sets; we will use the plasma data
library(HSAUR3)
```

Experimental design

1. Discuss with following in pairs
    - What are some advantages and disadvantages of using a balanced experimental design?
        - **Advantages**: the test will have larger statistical power; t-test statistic is less susceptible to small departures from the assumption of equal variances (homoscedasticity)
        - **Disadvantages**: forced removal of relevant data to acheive balance
    - Give an example of when a balanced design might not be possible.
        - Missing a sampling from 1 or more participants in a longitudinal study; missing data in general; natural populations are not equal sizes
    - There are 3 undergraduates assisting you with your experiment that assess the addiction potential of Saturday morning cartoons in rats. You need to run the experiments every Saturday, but one of your undergraduate assistants can only help out 2 Saturdays a month, while the other two undergraduate assistants can be there every Saturday. Rat behaviour is sensitive to handler. What should you do? (source)
        - Randomize handlers (2) across all time points
2. *True or false.* A completely randomized design offers no control for lurking variables (a variable that is not included as an explanatory or response variable in the analysis).
    - FALSE. Although it does not control perfectly for lurking variables, a randomized design offers some control for lurking variables.

1-way ANOVA

1. Using ANOVA, test if fruit fly longevity is effected by size (as measured by thorax length). What are your null and alternate hypotheses? What can you conclude from these results?

Null Hypothesis, H_0: Body size has no effect on the population mean longevity of male fruit flies.
Alternative Hypothesis, H_A: Body size has an effect on population mean longevity of male fruit flies.

```r
# create an ANOVA "model" object
fruitfly_model <- aov(longevity ~ thorax,
                              data = fruitfly_2groups)

# view output of aov() as a nice dataframe using tidy() from the broom package
tidy(fruitfly_model)
```

```
## # A tibble: 2 x 6
##   term         df  sumsq meansq statistic  p.value
##   <chr>     <dbl>  <dbl>  <dbl>     <dbl>    <dbl>
## 1 thorax        1  9267.  9267.      42.3  6.87e-9
## 2 Residuals    78 17098.   219.      NA   NA
```

Conclusion: Given that p is much much smaller than the commonly used threshold for rejecting the null hypothesis, p < 0.05, we can reject our null hypothesis that body size has no effect on the population mean longevity of male fruit flies, and accept the alternative hypothesis that body size **does** has an effect on population mean longevity of male fruit flies.

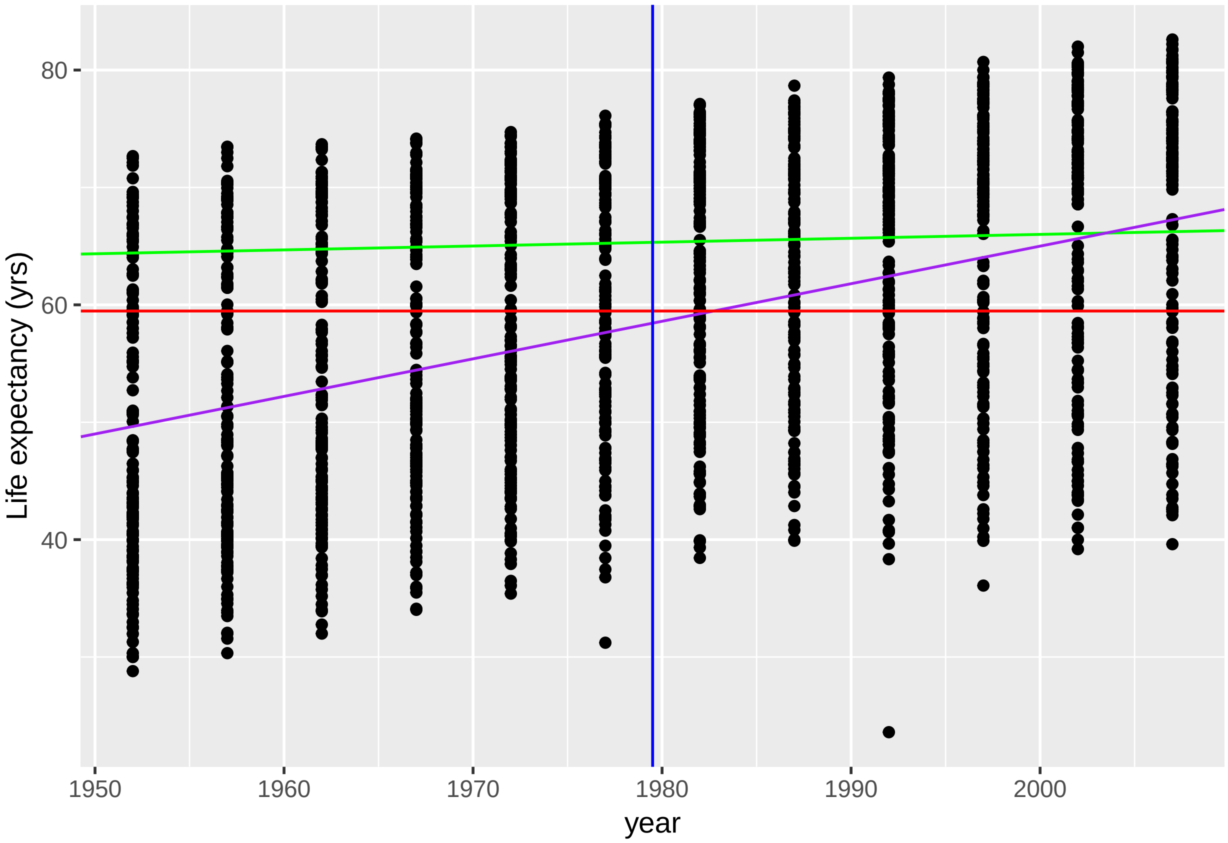
ANOVA

Determine whether the following statements are *true or false*?

1. ANOVA tests the null hypothesis that the sample means are all equal?
   - FALSE. ANOVA tests the equality of the popula1on means.
2. We use ANOVA to compare the variances of the population?
   - FALSE. We use ANOVA to compare the popula1on means.
3. A one-way ANOVA is equivalent to a *t*-test when there are 2 groups to be compared.
   - TRUE. Two groups can be represented as a factor with 2 levels.
4. In rejecting the null hypothesis, one can conclude that all the population means are different from one another?
   - FALSE. We can only conclude that there are at least 2 different popula1on means. We cannot conclude that they are not all equal.

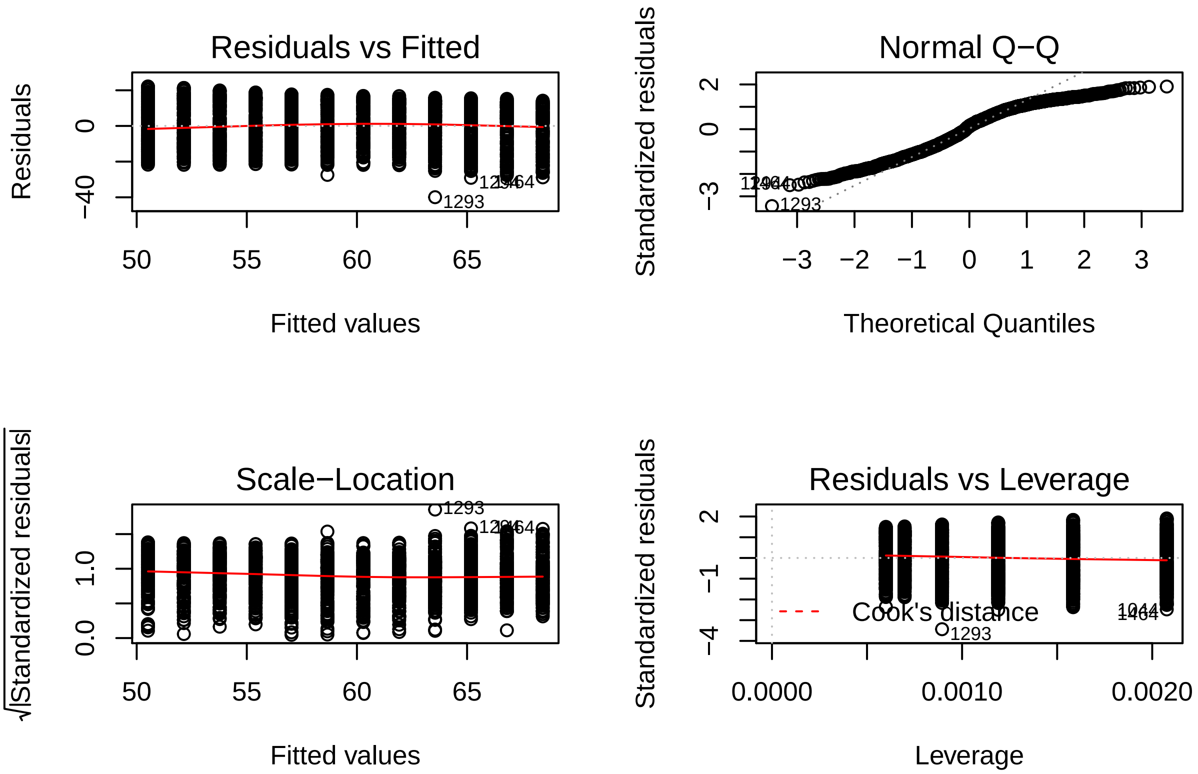
Questions courtesy of Dr. Gabriela Cohen Freue's DSCI 562 course (UBC)

Best fit lines



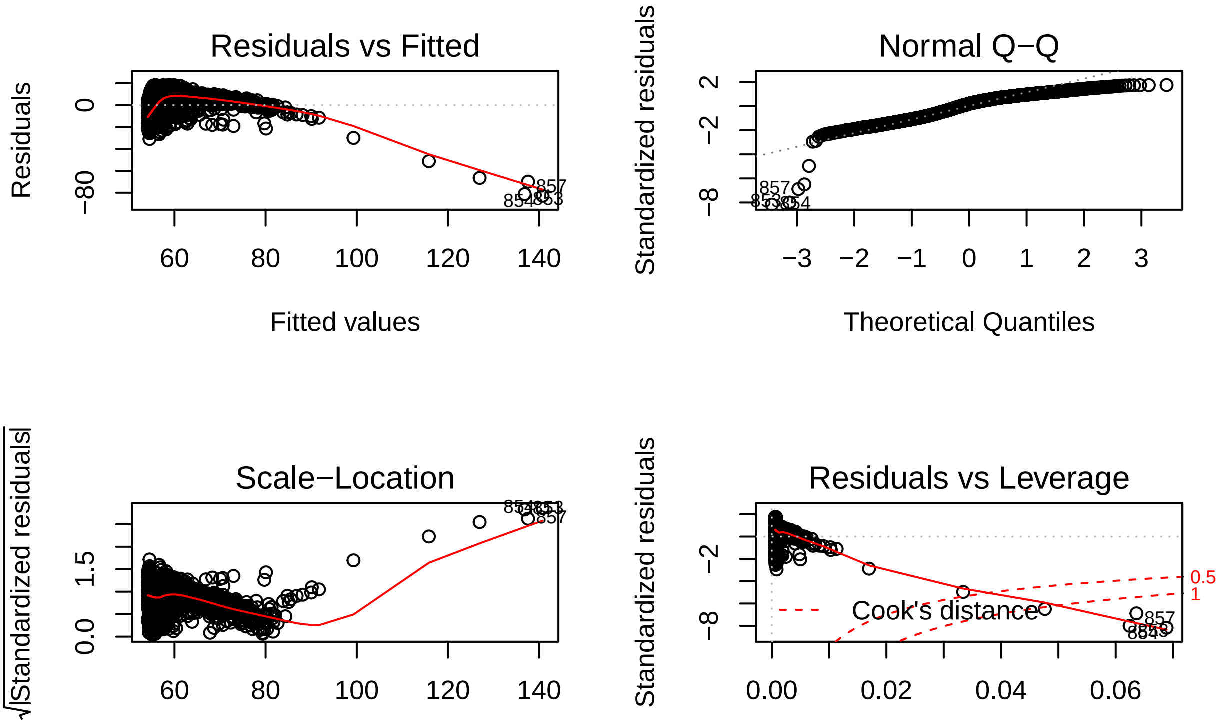
1. Which line best describes the data?
   - It's difficult (if not impossible) to tell just by looking.
2. The red one is a horizontal line at the overall mean life expectancy. It seems a reasonable model, but what is missing?
   - Missing information on the changing yearly mean as it uses just the overall data mean.

Linear models



1. Looking at the summary plots above, do you feel that our model can be extrapolated to a much wider `year` range? Why or why not?
   - Extrapolation is unlikely to be accurate as the residuals increase at both extremes of the current data, *e.g.* the model is less and less of a good fit for very distant or very recent years.
2. Fit a linear model of life expectancy as a function of per-capita gdp. Using the summary table and diagnostic plots, discuss whether or not you think this is a good fit for these data.

```r
lifeExp_model <- lm(lifeExp ~ gdpPercap,
                    data = gapminder)
# Set plot frame to 2 by 2
par(mfrow=c(2,2))
# Create diagnostic plots
plot(lifeExp_model)
```

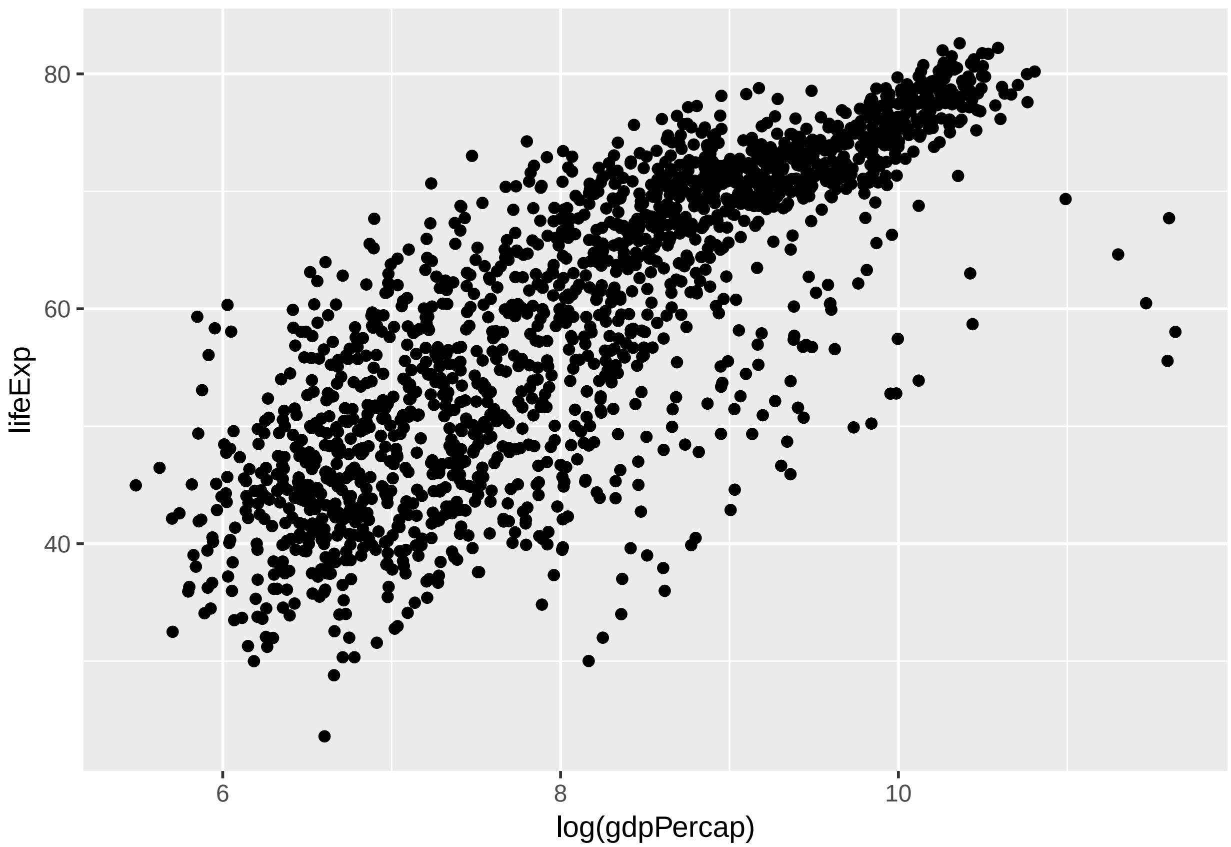


Tranforming predictors

1. Find a function that makes the plot more "linear" and fit a model of life expectancy as a function of the transformed per-capita gdp. Is it a better model?
   - Go back to your original `gdpPercap` vs. `lifeExp` plot and think about what function creates a similar trend.

A log transformation is one option for improving the linear fit.

```
gapminder %>%
  ggplot(aes(x = log(gdpPercap), y = lifeExp)) +
  geom_point()
```

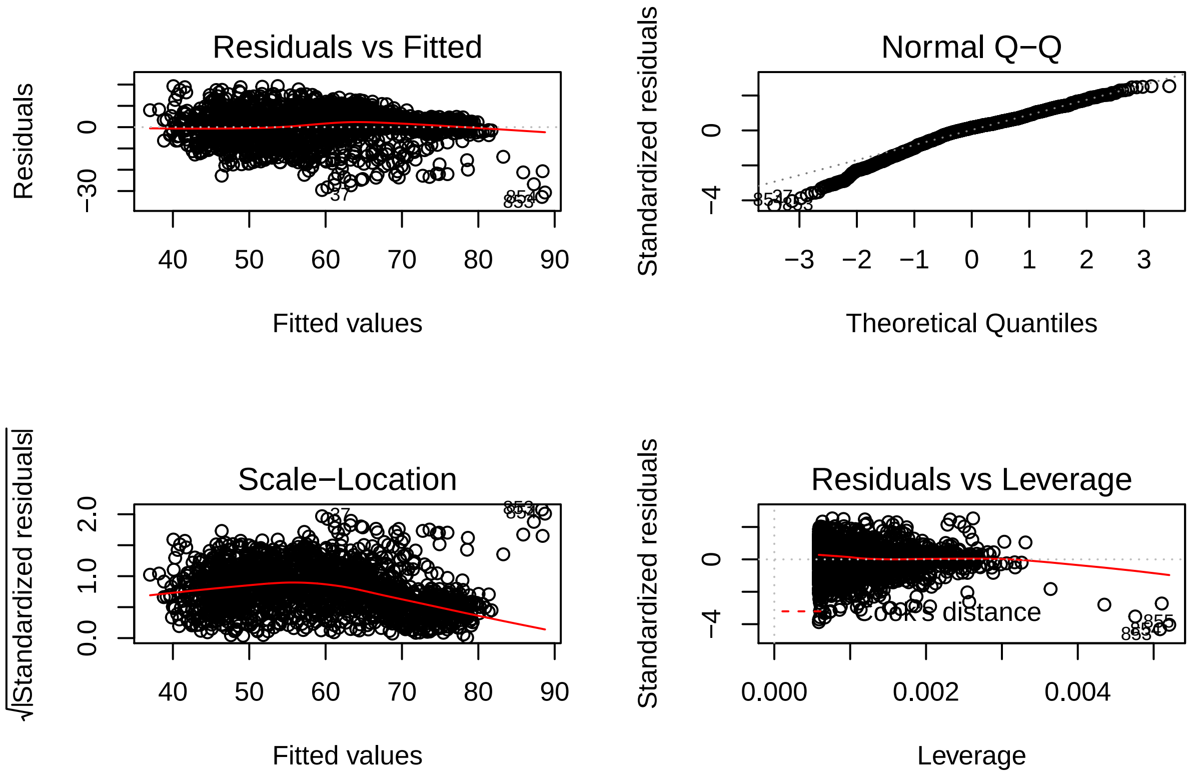


When we model these data, we see that our model fit is much improved!

```
#Fit model and view results
lifeExp_model <- lm(lifeExp ~ log(gdpPercap),
                    data = gapminder)
summary(lifeExp_model)
```

```
##
## Call:
## lm(formula = lifeExp ~ log(gdpPercap), data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.778  -4.204   1.212   4.658  19.285
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -9.1009     1.2277  -7.413 1.93e-13 ***
## log(gdpPercap)     8.4051     0.1488  56.500  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.62 on 1702 degrees of freedom
## Multiple R-squared:  0.6522, Adjusted R-squared:  0.652
## F-statistic:  3192 on 1 and 1702 DF,  p-value: < 2.2e-16
```

```
# Set plot frame to 2 by 2
par(mfrow=c(2,2))
# Create diagnostic plots
plot(lifeExp_model)
```



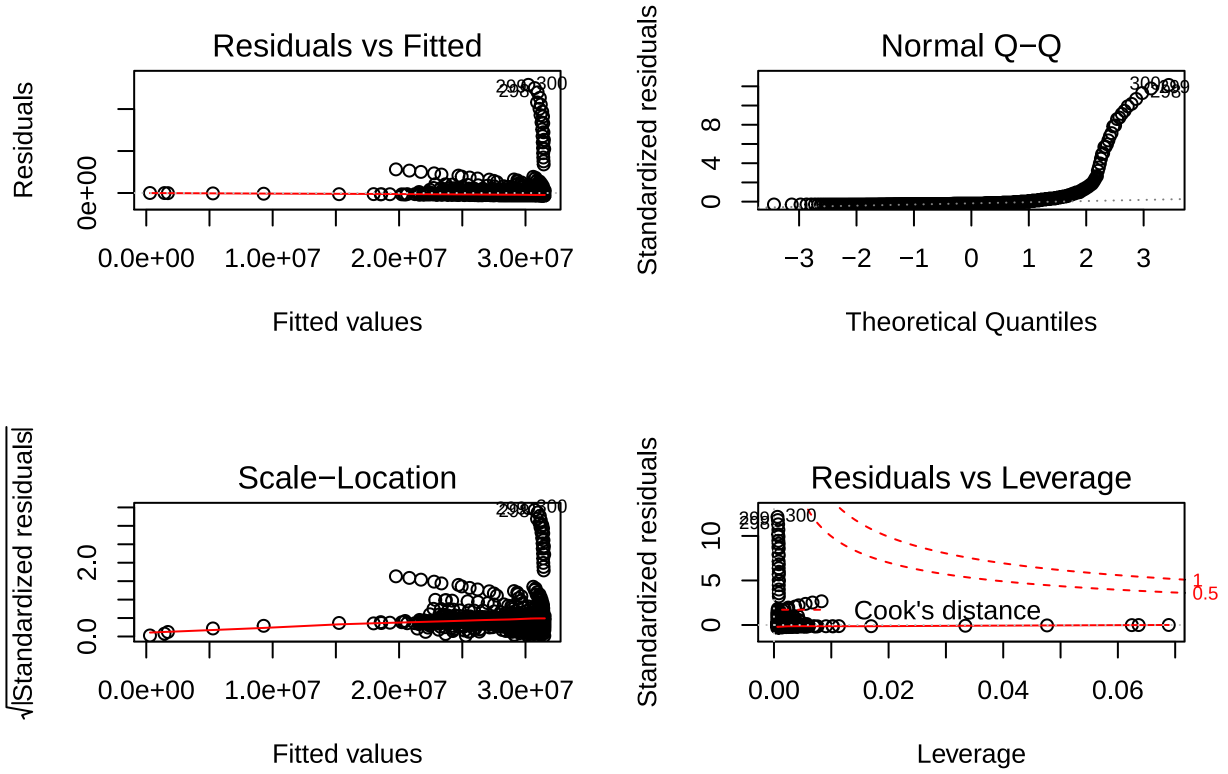
Multiple linear regression

1. So far, we have worked with `lifeExp` as our independent variable. Now, in small groups, try to produce a model of population (`pop`) using one or more of the variables available in `gapminder`.

```
pop_model <- lm(pop ~ gdpPercap,
                     data = gapminder)
summary(pop_model)
```

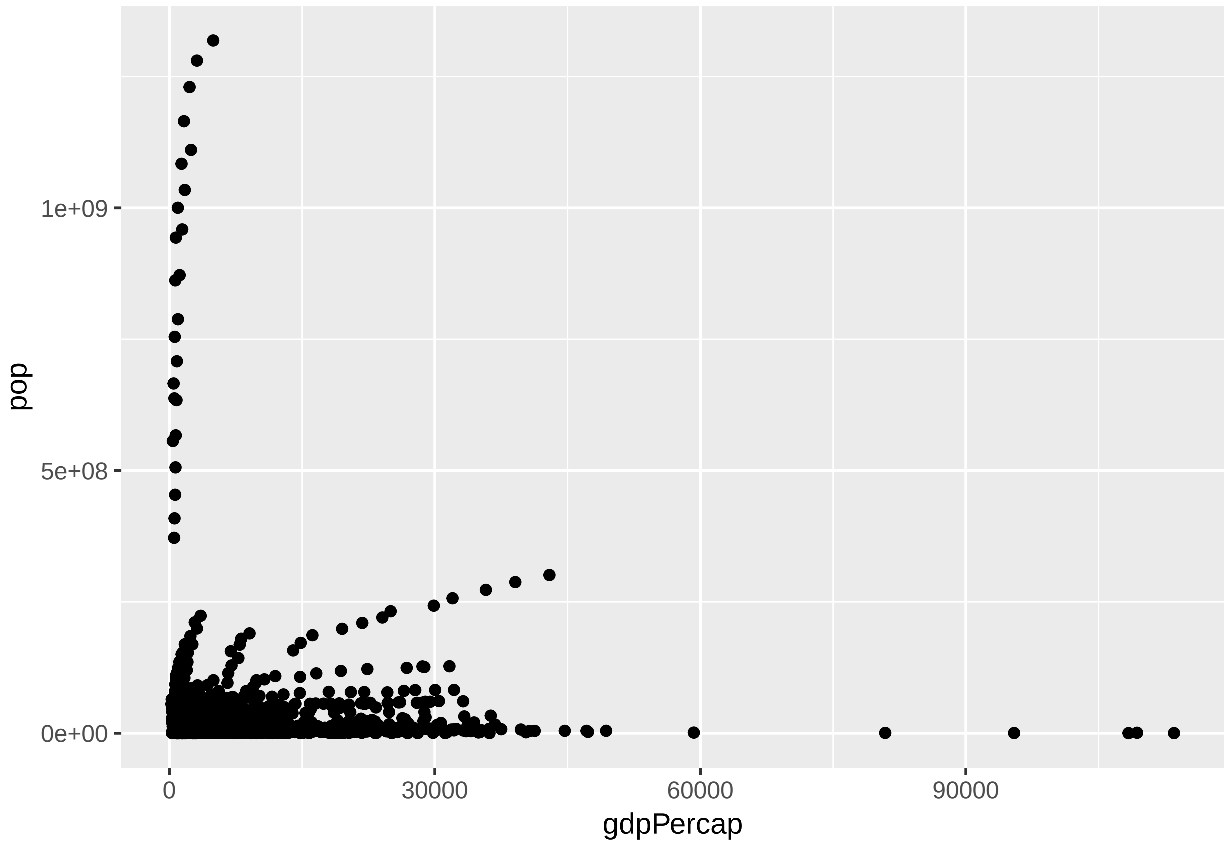
```
##
## Call:
## lm(formula = pop ~ gdpPercap, data = gapminder)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
##  -31291781  -27331437  -22315453  -10091438 1288459870
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31590402.5  3187212.3   9.912   <2e-16 ***
## gdpPercap       -275.7      261.0  -1.056    0.291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106200000 on 1702 degrees of freedom
## Multiple R-squared:  0.0006553,  Adjusted R-squared:  6.818e-05
## F-statistic: 1.116 on 1 and 1702 DF,  p-value: 0.2909
```

```
# Create diagnostic plots
par(mfrow=c(2,2))
plot(pop_model)
```



Clearly this is a poor model for these data (population vs. GDP) and serves as a gentle reminder to always plot your data first! If we do so now, we see that there is high variance as well as a number of potential outliers. In truth, these data would require a much more sophisticated model

```
gapminder %>%
  ggplot(aes(x = gdpPercap, y = pop)) +
  geom_point()
```



LME

1. Using the `sleepstudy` dataset, fit an LME on Reaction against Days, grouped by Subject.

```
sleep_model <- lmer(Reaction ~ Days + (Days | Subject),
                data=sleepstudy)
summary(sleep_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4633  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 611.90   24.737
##           Days         35.08    5.923   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  251.405      6.824  36.843
## Days          10.467      1.546   6.771
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.138
```

2. What is the intercept and slope of subject #310 in the model from question 1?

```r
coeffs <- coef(sleep_model)
coeffs$Subject["310", ]
```

```
##     (Intercept)     Days
## 310    212.4488 5.017706
```

3. CHALLENGE. Using the Teams dataset from the Lahman package, fit a model on runs (`R`) from the variables 'walks' (`BB`) and 'Hits' (`H`), grouped by team (`teamID`).
    - *Hint*: wrap the scale function around each predictor variable.

```r
library(Lahman)
```

```
##
## Attaching package: 'Lahman'
```

```
## The following object is masked from 'package:carData':
##
##     Salaries
```

```r
Lahman_model <- lmer(R ~ scale(BB) + scale(H) +
               (scale(BB) + scale(H) | teamID),
           data=Teams)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.0043881
## (tol = 0.002, component 1)
```

```r
summary(Lahman_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: R ~ scale(BB) + scale(H) + (scale(BB) + scale(H) | teamID)
##    Data: Teams
##
## REML criterion at convergence: 31768.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2212 -0.6223 -0.0791  0.4741  6.7306
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  teamID   (Intercept) 11571.5  107.57
##           scale(BB)     309.1   17.58   -0.58
##           scale(H)      331.9   18.22    0.49 -0.21
##  Residual              3608.4   60.07
## Number of obs: 2835, groups:  teamID, 149
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  772.493     10.662  72.453
## scale(BB)     26.065      2.980   8.748
## scale(H)     123.954      3.321  37.322
##
## Correlation of Fixed Effects:
##           (Intr) sc(BB)
## scale(BB) -0.189
## scale(H)   0.469 -0.438
## convergence code: 0
## Model failed to converge with max|grad| = 0.0043881 (tol = 0.002, component 1)
```

Logistic GLM

1. In the plasma data (from the `HSAUR3` package), use logistic regression to estimate the probabilities of ESR > 20, given the level of fibrinogen in the blood.

```
logit.model.esr <- glm(ESR ~ fibrinogen,
                       data = plasma,
                       family = binomial)
summary(logit.model.esr)
```

```
##
## Call:
## glm(formula = ESR ~ fibrinogen, family = binomial, data = plasma)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9298  -0.5399  -0.4382  -0.3356   2.4794
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.8451     2.7703  -2.471   0.0135 *
## fibrinogen    1.8271     0.9009   2.028   0.0425 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.885  on 31  degrees of freedom
## Residual deviance: 24.840  on 30  degrees of freedom
## AIC: 28.84
##
## Number of Fisher Scoring iterations: 5
```

The parameter for *fibrinogen* is significantly larger than 0. This suggests that the odds of having ESR greather than 20 mm/hour increases with the fibrinogen level in the blood. This is also visible in the fitted probabilities:

```
lsmeans(logit.model.esr, ~ fibrinogen, type = "response",
        at = list(
            fibrinogen = seq(from = 2, to = 4, by = 0.2)
        )
)
```

```
##  fibrinogen   prob     SE  df asymp.LCL asymp.UCL
##         2.0 0.0395 0.0400 Inf   0.00519     0.245
##         2.2 0.0560 0.0476 Inf   0.01004     0.257
##         2.4 0.0787 0.0551 Inf   0.01892     0.275
##         2.6 0.1096 0.0623 Inf   0.03405     0.301
##         2.8 0.1507 0.0704 Inf   0.05692     0.343
##         3.0 0.2036 0.0832 Inf   0.08557     0.411
##         3.2 0.2693 0.1056 Inf   0.11397     0.513
##         3.4 0.3468 0.1394 Inf   0.13719     0.639
##         3.6 0.4335 0.1794 Inf   0.15455     0.762
##         3.8 0.5244 0.2165 Inf   0.16746     0.858
##         4.0 0.6138 0.2416 Inf   0.17737     0.921
##
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale
```

2. Using the `womensrole` data set from the `HSAUR3` package, try to fit a logistic regression to the agreement with the statement, given the years of education and the respondent's sex (also attributed as `gender` in these data).

With `tidyr`, it is easy to get that into the form we are more familiar with:

```
wr.df <- gather(womensrole, key = response, value = freq,
                agree, disagree, factor_key = TRUE)
```

Note, however, that the factor `response` has agree as first level, which means that `glm` will model the probability for *disagree*, not *agree*!

```
wr.mod <- glm(response ~ gender + education,
              data = wr.df,
              family = binomial,
              weights = freq)
summary(wr.mod)
```

```
##
## Call:
## glm(formula = response ~ gender + education, family = binomial,
##     data = wr.df, weights = freq)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -20.782   -4.169    0.000    4.154   17.662
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.50937    0.18389 -13.646   <2e-16 ***
## genderFemale  0.01145    0.08415   0.136    0.892
## education     0.27062    0.01541  17.560   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3736.0  on 76  degrees of freedom
## Residual deviance: 3348.3  on 74  degrees of freedom
## AIC: 3354.3
##
## Number of Fisher Scoring iterations: 5
```

Or you can also fit logistic regression with a 2 dimensional response of the form (success, failure).

```r
wr.mod.2 <- glm(cbind(agree, disagree) ~ gender + education,
                data = womensrole,
                family = binomial)
summary(wr.mod.2)
```

```
##
## Call:
## glm(formula = cbind(agree, disagree) ~ gender + education, family = binomial,
##     data = womensrole)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.72544  -0.86302  -0.06525   0.84340   3.13315
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.50937    0.18389  13.646   <2e-16 ***
## genderFemale -0.01145    0.08415  -0.136    0.892
## education    -0.27062    0.01541 -17.560   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 451.722  on 40  degrees of freedom
## Residual deviance:  64.007  on 38  degrees of freedom
## AIC: 208.07
##
## Number of Fisher Scoring iterations: 4
```

We can also look at the change in the probability as a function of the years of education:

```
lsmeans(wr.mod.2, ~ education,
        type = "response",
        at = list(education = 0:20)
)
```

```
##  education   prob      SE  df asymp.LCL asymp.UCL
##          0 0.9244 0.01247 Inf    0.8960    0.9455
##          1 0.9032 0.01430 Inf    0.8713    0.9278
##          2 0.8768 0.01606 Inf    0.8417    0.9050
##          3 0.8445 0.01759 Inf    0.8068    0.8759
##          4 0.8055 0.01871 Inf    0.7662    0.8396
##          5 0.7596 0.01920 Inf    0.7200    0.7952
##          6 0.7068 0.01890 Inf    0.6684    0.7424
##          7 0.6478 0.01777 Inf    0.6122    0.6818
##          8 0.5839 0.01590 Inf    0.5524    0.6147
##          9 0.5170 0.01362 Inf    0.4903    0.5436
##         10 0.4495 0.01145 Inf    0.4272    0.4721
##         11 0.3839 0.00999 Inf    0.3645    0.4036
##         12 0.3222 0.00953 Inf    0.3038    0.3411
##         13 0.2661 0.00976 Inf    0.2474    0.2857
##         14 0.2167 0.01013 Inf    0.1975    0.2372
##         15 0.1743 0.01028 Inf    0.1550    0.1954
##         16 0.1387 0.01008 Inf    0.1201    0.1596
##         17 0.1094 0.00955 Inf    0.0920    0.1296
##         18 0.0857 0.00879 Inf    0.0700    0.1046
##         19 0.0667 0.00788 Inf    0.0528    0.0839
##         20 0.0517 0.00693 Inf    0.0397    0.0671
##
## Results are averaged over the levels of: gender
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale
```

Poisson GLM

1. Check which *covariates* have a significant effect on the response in the model fitted with the Poisson family and with the quasi-Poisson family and compare the results. What do you observe?

Similar to logistic regression, we can also check the significance of the covariates using the `Anova` function:

```
polyps_model1 <- glm(number ~ treat + age,
                    data = polyps,
                    family = poisson)
summary(polyps_model1)
```

```
## 
## Call:
## glm(formula = number ~ treat + age, family = poisson, data = polyps)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.2212  -3.0536  -0.1802   1.4459   5.8301  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  4.529024   0.146872   30.84  < 2e-16 ***
## treatdrug   -1.359083   0.117643  -11.55  < 2e-16 ***
## age         -0.038830   0.005955   -6.52 7.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 378.66  on 19  degrees of freedom
## Residual deviance: 179.54  on 17  degrees of freedom
## AIC: 273.88
## 
## Number of Fisher Scoring iterations: 5
```

```
polyps_model2 <- glm(number ~ treat + age,
                     data = polyps,
                     family = quasipoisson)
summary(polyps_model2)
```

```
## 
## Call:
## glm(formula = number ~ treat + age, family = quasipoisson, data = polyps)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.2212  -3.0536  -0.1802   1.4459   5.8301  
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.52902    0.48106   9.415 3.72e-08 ***
## treatdrug   -1.35908    0.38533  -3.527  0.00259 ** 
## age         -0.03883    0.01951  -1.991  0.06284 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.72805)
##
##     Null deviance: 378.66  on 19  degrees of freedom
## Residual deviance: 179.54  on 17  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
Anova(polyps_model1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: number
##       LR Chisq Df Pr(>Chisq)
## treat  169.311  1  < 2.2e-16 ***
## age     49.018  1  2.536e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(polyps_model2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: number
##       LR Chisq Df Pr(>Chisq)
## treat  15.7821  1  7.107e-05 ***
## age     4.5692  1    0.03255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quasi-poisson vs. negative binomial GLM

1. Fit the above model with a Quasi-Poisson family and check the over-dispersion in that fit. Is there a difference in the significance of any terms compared to the NB model? Would a Poisson model be appropriate as well?

```
quine_model_q <- glm(Days ~ Sex * (Age + Eth * Lrn),
                     data = quine,
                     family = quasipoisson)
summary(quine_model_q)
```

```
##
## Call:
## glm(formula = Days ~ Sex * (Age + Eth * Lrn), family = quasipoisson,
##     data = quine)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.3924  -2.4923  -0.5745   1.5507   8.5631
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.89964    0.28113  10.314  < 2e-16 ***
## SexM            -0.39728    0.38999  -1.019  0.31020
## AgeF1           -0.48941    0.31188  -1.569  0.11899
## AgeF2           -0.19579    0.34387  -0.569  0.57007
## AgeF3           -0.21627    0.31725  -0.682  0.49662
## EthN            -0.08595    0.27661  -0.311  0.75650
## LrnSL            0.73625    0.30990   2.376  0.01895 *
## EthN:LrnSL      -1.27702    0.41039  -3.112  0.00228 **
## SexM:AgeF1      -0.33326    0.51247  -0.650  0.51662
## SexM:AgeF2       0.82205    0.43483   1.890  0.06088 .
## SexM:AgeF3       1.23318    0.43783   2.817  0.00560 **
## SexM:EthN       -0.40470    0.36726  -1.102  0.27250
## SexM:LrnSL      -0.39846    0.43055  -0.925  0.35641
## SexM:EthN:LrnSL  1.86724    0.57423   3.252  0.00146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.53364)
##
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 1393.5  on 132  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
Anova(quine_model_q)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Days
##             LR Chisq Df Pr(>Chisq)
## Sex           1.4457  1  0.2292133
## Age          17.0958  3  0.0006754 ***
## Eth          14.0792  1  0.0001753 ***
## Lrn           6.8149  1  0.0090403 **
## Eth:Lrn       1.5446  1  0.2139309
## Sex:Age      15.1022  3  0.0017314 **
## Sex:Eth       1.8346  1  0.1755855
## Sex:Lrn       0.8270  1  0.3631425
## Sex:Eth:Lrn  10.8139  1  0.0010074 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compared to

```r
quine_model <- glm.nb(Days ~ Sex * (Age + Eth * Lrn),
                       data = quine)
summary(quine_model)
```

```
##
## Call:
## glm.nb(formula = Days ~ Sex * (Age + Eth * Lrn), data = quine,
##     init.theta = 1.597990733, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8950  -0.8827  -0.2299   0.5669   2.1071
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.01919    0.29706  10.163  < 2e-16 ***
## SexM             -0.47541    0.39550  -1.202 0.229355
## AgeF1            -0.70887    0.32321  -2.193 0.028290 *
## AgeF2            -0.61486    0.37141  -1.655 0.097826 .
## AgeF3            -0.34235    0.32717  -1.046 0.295388
## EthN             -0.07312    0.26539  -0.276 0.782908
## LrnSL             0.94358    0.32246   2.926 0.003432 **
## EthN:LrnSL       -1.35849    0.37719  -3.602 0.000316 ***
## SexM:AgeF1       -0.01486    0.46225  -0.032 0.974353
## SexM:AgeF2        1.24328    0.46134   2.695 0.007040 **
## SexM:AgeF3        1.49319    0.45337   3.294 0.000989 ***
## SexM:EthN        -0.60586    0.36896  -1.642 0.100572
## SexM:LrnSL       -0.70467    0.46536  -1.514 0.129966
## SexM:EthN:LrnSL   2.11991    0.58056   3.651 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.598) family taken to be 1)
##
##     Null deviance: 234.56  on 145  degrees of freedom
## Residual deviance: 167.56  on 132  degrees of freedom
## AIC: 1093
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.598
##           Std. Err.:  0.213
##
##  2 x log-likelihood:  -1063.025
```

```
Anova(quine_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Days
##             LR Chisq Df Pr(>Chisq)
## Sex           0.9284  1  0.3352783
## Age          14.9609  3  0.0018503 **
## Eth          16.9573  1  3.823e-05 ***
## Lrn           5.6903  1  0.0170588 *
## Eth:Lrn       2.5726  1  0.1087268
## Sex:Age      19.8297  3  0.0001841 ***
## Sex:Eth       0.6547  1  0.4184372
## Sex:Lrn       1.4965  1  0.2212106
## Sex:Eth:Lrn  12.9647  1  0.0003174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The standard errors of the parameters are slightly larger in the Quasi-Poisson model compared to the NB model. This results in the interaction between sex and age "F1" to be not significant at the $\alpha = 0.05$ level anymore. A Poisson model would not be appropriate for this data, since the quasi-poisson model sestimated an overdispersion of 10.5.