Homework 6

Context

This assignment reinforces ideas in Linear Models (topic_linear_models.html).

Due date and submission

Due: December 3 at 11:59pm.

Please submit (via courseworks) the web address of the GitHub repo containing your work for this assignment; git commits after the due date will cause the assignment to be considered late.

R Markdown documents included as part of your solutions must not install packages, and should only load the packages necessary for your submission to knit.

Points

Problem	**Points**
Problem 0	20
Problem 1	–
Problem 2	40
Problem 3	40
Optional survey	No points

Problem 0

This “problem” focuses on structure of your submission, especially the use git and GitHub for reproducibility, R Projects to organize your work, R Markdown to write reproducible reports, relative paths to load data from local files, and reasonable naming structures for your files. To that end:

- create a public GitHub repo + local R Project; we suggest naming this repo / directory `p8105_hw6_YOURUNI` (e.g. `p8105_hw6_ajg2202` for Jeff), but that’s not required
- create a single .Rmd file named `p8105_hw6_YOURUNI.Rmd` that renders to `github_document`
- create a subdirectory to store the local data files used in the assignment, and use relative paths to access these data files
- submit a link to your repo via Courseworks

Your solutions to Problems 1 and 2 should be implemented in your .Rmd file, and your git commit history should reflect the process you used to solve these Problems.

For this Problem, we will assess adherence to the instructions above regarding repo structure, git commit history, and whether we are able to knit your .Rmd to ensure that your work is reproducible. Adherence to appropriate styling and clarity of code will be assessed in Problems 1+ using the style rubric (homework_style_rubric.html).

This homework includes figures; the readability of your embedded plots (e.g. font sizes, axis labels, titles) will be assessed in Problems 1+.

Problem 1

For this problem, we’ll use the 2017 Central Park weather data that we’ve seen elsewhere. The code chunk below (adapted from the course website) will download these data.

```
weather_df =
  rnoaa::meteo_pull_monitors(
    c("USW00094728"),
    var = c("PRCP", "TMIN", "TMAX"),
    date_min = "2017-01-01",
    date_max = "2017-12-31") %>%
  mutate(
    name = recode(id, USW00094728 = "CentralPark_NY"),
    tmin = tmin / 10,
    tmax = tmax / 10) %>%
  select(name, id, everything())
```

The boostrap is helpful when you’d like to perform inference for a parameter / value / summary that doesn’t have an easy-to-write-down distribution in the usual repeated sampling framework. We’ll focus on a simple linear regression with `tmax` as the response and `tmin` as the predictor, and are interested in the distribution of two quantities estimated from these data:

- $\hat{r}^2$
- $\log(\hat{\beta}_0 * \hat{\beta}_1)$

Use 5000 bootstrap samples and, for each bootstrap sample, produce estimates of these two quantities. Plot the distribution of your estimates, and describe these in words. Using the 5000 bootstrap estimates, identify the 2.5% and 97.5% quantiles to provide a 95% confidence interval for $\hat{r}^2$ and $\log(\hat{\beta}_0 * \hat{\beta}_1)$. Note:

`broom::glance()` is helpful for extracting $\hat{r}^2$ from a fitted regression, and `broom::tidy()` (with some additional wrangling) should help in computing $\log(\hat{\beta}_0 * \hat{\beta}_1)$.

Problem 2

The *Washington Post* has gathered data on homicides in 50 large U.S. cities and made the data available through a GitHub repository here (https://github.com/washingtonpost/data-homicides). You can read their accompanying article here (https://www.washingtonpost.com/graphics/2018/investigations/where-murders-go-unsolved/).

Create a `city_state` variable (e.g. "Baltimore, MD"), and a binary variable indicating whether the homicide is solved. Omit cities Dallas, TX; Phoenix, AZ; and Kansas City, MO – these don't report victim race. Also omit Tulsa, AL – this is a data entry mistake. For this problem, limit your analysis those for whom `victim_race` is `white` or `black`. Be sure that `victim_age` is numeric.

For the city of Baltimore, MD, use the `glm` function to fit a logistic regression with resolved vs unresolved as the outcome and victim age, sex and race as predictors. Save the output of `glm` as an R object; apply the `broom::tidy` to this object; and obtain the estimate and confidence interval of the adjusted **odds ratio** for solving homicides comparing male victims to female victims keeping all other variables fixed.

Now run `glm` for each of the cities in your dataset, and extract the adjusted odds ratio (and CI) for solving homicides comparing male victims to female victims. Do this within a "tidy" pipeline, making use of `purrr::map`, list columns, and `unnest` as necessary to create a dataframe with estimated ORs and CIs for each city.

Create a plot that shows the estimated ORs and CIs for each city. Organize cities according to estimated OR, and comment on the plot.

Problem 3

In this problem, you will analyze data gathered to understand the effects of several variables on a child's birthweight. This dataset, available here (data/birthweight.csv), consists of roughly 4000 children and includes the following variables:

- `babysex` : baby's sex (male = 1, female = 2)
- `bhead` : baby's head circumference at birth (centimeters)
- `blength` : baby's length at birth (centimeteres)
- `bwt` : baby's birth weight (grams)
- `delwt` : mother's weight at delivery (pounds)
- `fincome` : family monthly income (in hundreds, rounded)
- `frace` : father's race (1 = White, 2 = Black, 3 = Asian, 4 = Puerto Rican, 8 = Other, 9 = Unknown)
- `gaweeks` : gestational age in weeks
- `malform` : presence of malformations that could affect weight (0 = absent, 1 = present)
- `menarche` : mother's age at menarche (years)
- `mheigth` : mother's height (inches)
- `momage` : mother's age at delivery (years)
- `mrace` : mother's race (1 = White, 2 = Black, 3 = Asian, 4 = Puerto Rican, 8 = Other)
- `parity` : number of live births prior to this pregnancy
- `pnumlbw` : previous number of low birth weight babies

- `pnumgsa` : number of prior small for gestational age babies
- `ppbmi` : mother's pre-pregnancy BMI
- `ppwt` : mother's pre-pregnancy weight (pounds)
- `smoken` : average number of cigarettes smoked per day during pregnancy
- `wtgain` : mother's weight gain during pregnancy (pounds)

Load and clean the data for regression analysis (i.e. convert numeric to factor where appropriate, check for missing data, etc.).

Propose a regression model for birthweight. This model may be based on a hypothesized structure for the factors that underly birthweight, on a data-driven model-building process, or a combination of the two. Describe your modeling process and show a plot of model residuals against fitted values – use `add_predictions` and `add_residuals` in making this plot.

Compare your model to two others:

- One using length at birth and gestational age as predictors (main effects only)
- One using head circumference, length, sex, and all interactions (including the three-way interaction) between these

Make this comparison in terms of the cross-validated prediction error; use `crossv_mc` and functions in `purrr` as appropriate.

Note that although we expect your model to be reasonable, model building itself is not a main idea of the course and we don't necessarily expect your model to be "optimal".

Optional post-assignment survey

If you'd like, a you can complete this short survey (https://forms.gle/b4hS6r9sL92VRoXA8) after you've finished the assignment.