Homework 7

STAT3530 Applied Linear Models

Due Tuesday, November 3, in class.

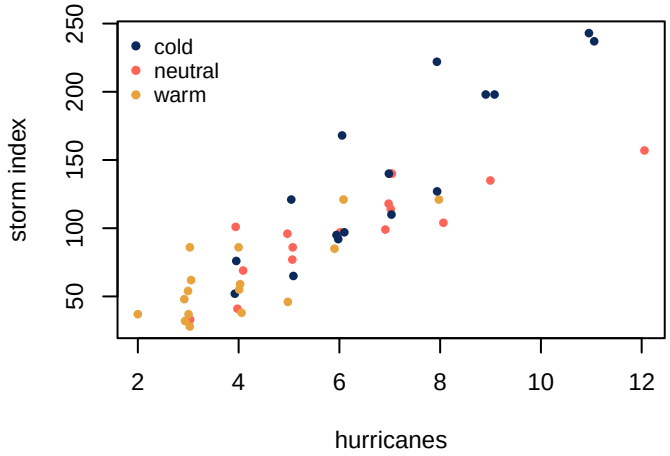
You may collaborate, but write up your own solutions. Submit a paper copy of this document with handwritten solutions below each prompt. Please attach any plots or output you are asked to produce and label them by problem number.

El Niño and the Atlantic hurricane season

Annual Atlantic hurricane seasons from 1950 to 1997. For each season, `hurricanes` is the number of hurricanes and `stormindex` is an index of the overall intensity of the season. The variable `elnino` classifies the year as `cold`, `neutral`, or `warm` according to sea surface temperature in the eastern Pacific.

Here we ask whether season intensity differs in El Niño (warm) or La Niña (cold) years after adjusting for the number of hurricanes in a season.

```
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/seasons.RData"))
```



1. Check the encoding of `elnino` in R, write the corresponding indicator variable(s), and identify the reference level.

2. Write a model for the data without interaction terms, and express the (null) hypothesis of interest in terms of model coefficients.

3. Fit the model and test the hypothesis. Report the F statistic, numerator and denominator degrees of freedom, and p-value for the test, and interpret the result in context.

4. Overlay the parallel lines for the fitted model on the scatterplot of the data. Attach your plot. Do interaction terms seem reasonable to consider here? Explain.

5. Rewrite the model with interaction terms included, and express the (null) hypothesis of interest in terms of the coefficients of this updated model.

6. Fit the model and test the hypothesis. Report the F statistic, numerator and denominator degrees of freedom, and p-value for the test, and interpret the result in context.

7. Test for an interaction between season type (El Niño state) and number of hurricanes. Write the hypotheses in terms of model coefficients, report the F statistic, numerator and denominator degrees of freedom, and p-value, and interpret the result in context.

8. Using the model that best fits the data, estimate, for a season with 6 hurricanes, the difference in mean intensity between El Niño years and neutral years, and the difference in mean intensity between La Niña years and neutral years. Provide 95% confidence intervals.

Light and meadowfoam flowering

Meadowfoam is a small plant grown for seed oil. In a controlled growth chamber, researchers crossed six light intensities (`intensity`, in μmol/m^2/sec) with two choices of when to switch the light treatment on (`timing`): at the point of photoperiodic floral induction (`late`) or 24 days before it (`early`). Two plants were grown in each of the twelve cells, and `flowers` records the average number of flowers per plant.

```
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/meadowfoam.RData"))
```

9. Plot `flowers` against `intensity`, distinguishing the two timings. Attach the plot. Which model does the plot suggest: `intensity` only, `intensity` and `timing`, or `intensity` and `timing` and the interaction? Explain in one sentence.

10. Test whether the relationship between light intensity and mean flowering depends on timing.
    i. Write the model appropriate for the test.

    ii. Write the (null) hypothesis of interest.

    iii. Report the F statistic, numerator and denominator degrees of freedom, and p-value.

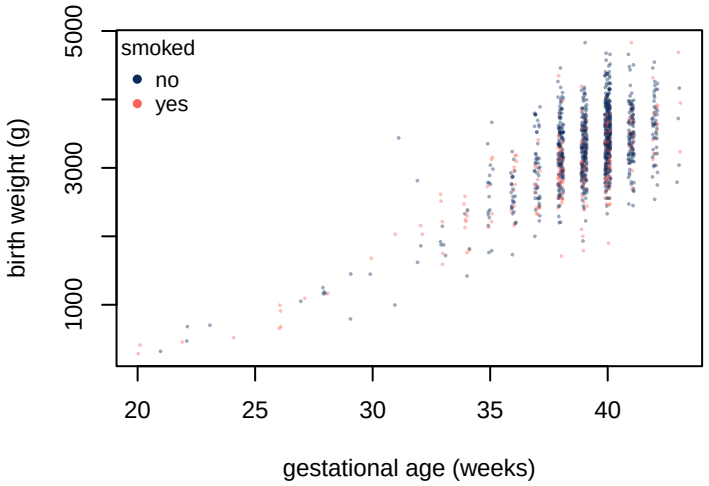
    iv. Interpret the result of the test in context.

11. A biologist reads your answer to (10) and concludes that light intensity affects flowering identically under the two timings. Are they correct?

Smoking and birth weight

The data for this problem are birth weight in grams (`grams`) and gestational age in weeks (`gestate`) for 1115 births in Philadelphia, along with `smoke`, whether the mother smoked during pregnancy (846 `no`, 269 `yes`). Babies born earlier are smaller, and babies born to mothers who smoked are smaller still. Here we ask whether the second deficit is the same size at every gestational age.

```
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/phbirths.RData"))
```



12. Test for an interaction between smoking and gestational age using the appropriate linear model.

    i. Write the model and hypothesis of interest.

    ii. Report the F statistic, numerator and denominator degrees of freedom, and p-value.

    iii. On this evidence, does the relationship between birth weight and gestational age depend on whether the mother smoked? Answer in one sentence.

13. Overlay the lines from the fitted model in (12) on the data. Does the model appear well-specified? Explain in one sentence.

14. Ignoring smoking status, test for curvature by comparing a simple linear regression against a cubic spline with four basis functions. (Linear functions are expressible in the basis representation, so an F test works here.)

    i. Write the R formulae for the models under comparison.

    ii. Report the F statistic, numerator and denominator degrees of freedom, and p-value.

    iii. Interpret the result in context.

15. Now fit a parallel *curve* model with a 4-df cubic spline expansion of gestational age and smoking status as predictors entering additively. Overlay the fitted curves on the data and attach your plot.

```
fit.add <- lm(grams ~ bs(gestate, df = 4) + smoke, data = phbirths)
```

16. Fit a model with separate curves for each smoking status by adding an interaction between the basis expansion and the smoking indicator. Overlay the fitted curves on the data and attach your plot.

```
fit.int <- lm(grams ~ bs(gestate, df = 4) * smoke, data = phbirths)
```

17. Test whether the curves differ by smoking status.

    i. Write the model in (16).

    ii. Express the (null) hypothesis of interest in terms of model coefficients.

iii. Report the F statistic, numerator and denominator degrees of freedom, and p-value.

iv. On this evidence, does the relationship between birth weight and gestational age depend on whether the mother smoked?

18. [Optional extra credit] Under the separate curves model, the smoking deficit is no longer a single number: it is the vertical distance between the two curves, which depends on gestational age. Write codes to construct a sampling distribution for that distance at a fixed gestational age t.

    i. Compute point estimates and 95% confidence intervals for the values specified below.

gestational age	estimated gap (g)	95% interval
34 weeks	_____	_____
38 weeks	_____	_____
42 weeks	_____	_____

    ii. Now recompute on a grid spanning the range of gestational ages in the data and plot a curve of the estimated differences with pointwise 95% uncertainty bands.