STAT-S632

Assignment 3

John Koo

```
# packages, etc.
import::from(magrittr, `%>%`, `%<>%`)
dp <- loadNamespace('dplyr')
library(ggplot2)
import::from(GGally, ggpairs)
theme_set(theme_bw())
```

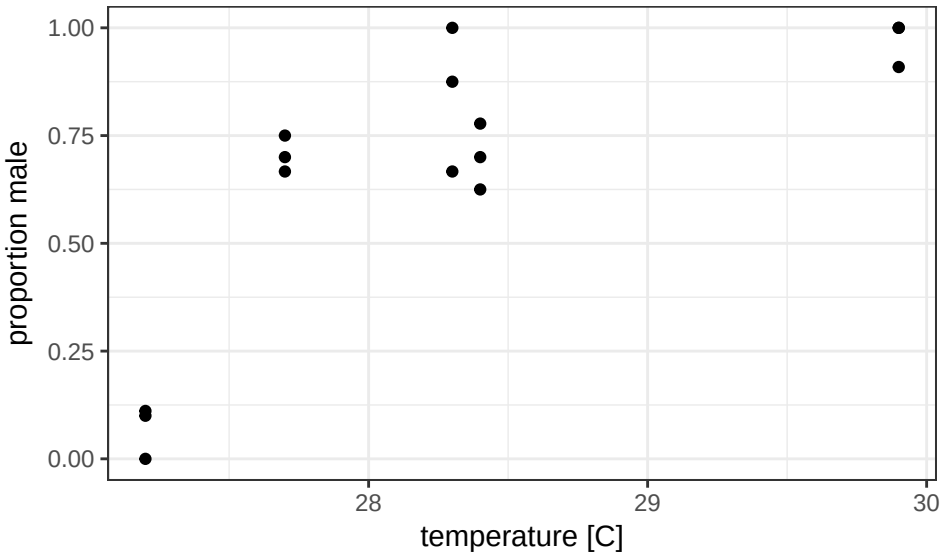
Exercise 3.2

```
# get the data
turtles.df <- faraway::turtle %>%
  dp$mutate(turtles = male + female,
            prop.male = male / turtles)
summary(turtles.df)
```

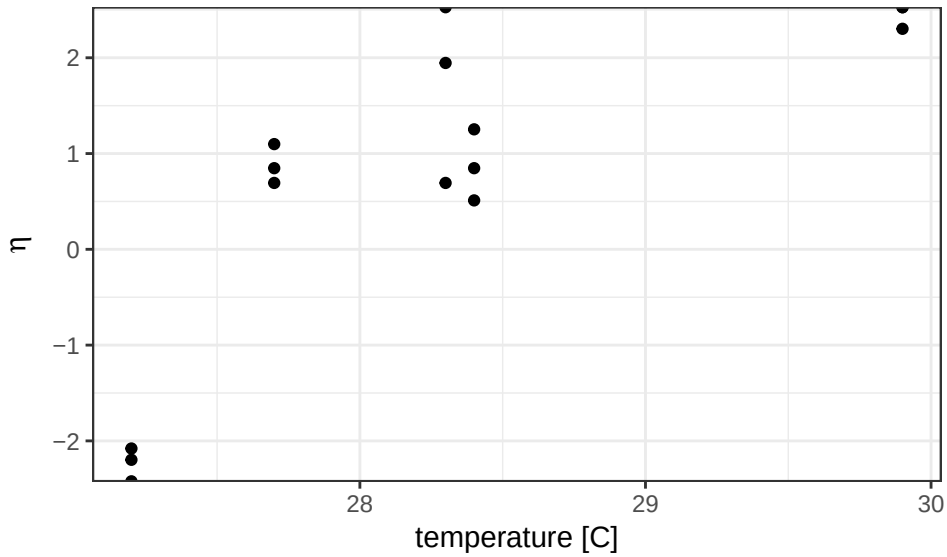
```
      temp           male            female     turtles
 Min.   :27.2   Min.   : 0.000   Min.   :0   Min.   : 6.000
 1st Qu.:27.7   1st Qu.: 4.500   1st Qu.:1   1st Qu.: 8.000
 Median :28.3   Median : 7.000   Median :2   Median : 9.000
 Mean   :28.3   Mean   : 6.067   Mean   :3   Mean   : 9.067
 3rd Qu.:28.4   3rd Qu.: 7.500   3rd Qu.:3   3rd Qu.:10.000
 Max.   :29.9   Max.   :13.000   Max.   :9   Max.   :13.000
   prop.male
 Min.   :0.0000
 1st Qu.:0.6458
 Median :0.7000
 Mean   :0.6588
 3rd Qu.:0.8920
 Max.   :1.0000
```

Part a

```
# p vs temp
ggplot(turtles.df) +
  geom_point(aes(x = temp, y = prop.male)) +
  labs(x = 'temperature [C]', y = 'proportion male')
```




```
# eta vs temp
ggplot(turtles.df) +
  geom_point(aes(x = temp, y = log(prop.male / (1 - prop.male)))) +
  labs(x = 'temperature [C]', y = expression(eta))
```



We can see that the proportion of male turtles increases with temperature. However, it does not look like a logit is a good fit on these data. In particular, when we transform the p_is to η_is ($\eta = \log \frac{p}{1-p}$), we do not get a linear relationship between η and temperature.

Part b

```
# fit the model
binom.mod <- glm(cbind(male, female) ~ temp,
                 data = turtles.df, family = binomial)
summary(binom.mod)
```

```
Call:
glm(formula = cbind(male, female) ~ temp, family = binomial,
    data = turtles.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0721  -1.0292  -0.2714   0.8087   2.5550

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
temp          2.2110     0.4309   5.132 2.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.508  on 14  degrees of freedom
Residual deviance: 24.942  on 13  degrees of freedom
AIC: 53.836

Number of Fisher Scoring iterations: 5
```

```
pchisq(binom.mod$deviance, binom.mod$df.residual, lower = FALSE)
```

```
[1] 0.02348863
```

```
pchisq(summary(binom.mod)$null.deviance,
       summary(binom.mod)$df.null,
       lower = FALSE)
```

```
[1] 2.912592e-10
```

The residual deviance suggests that the model is not a good fit for these data.

Part c

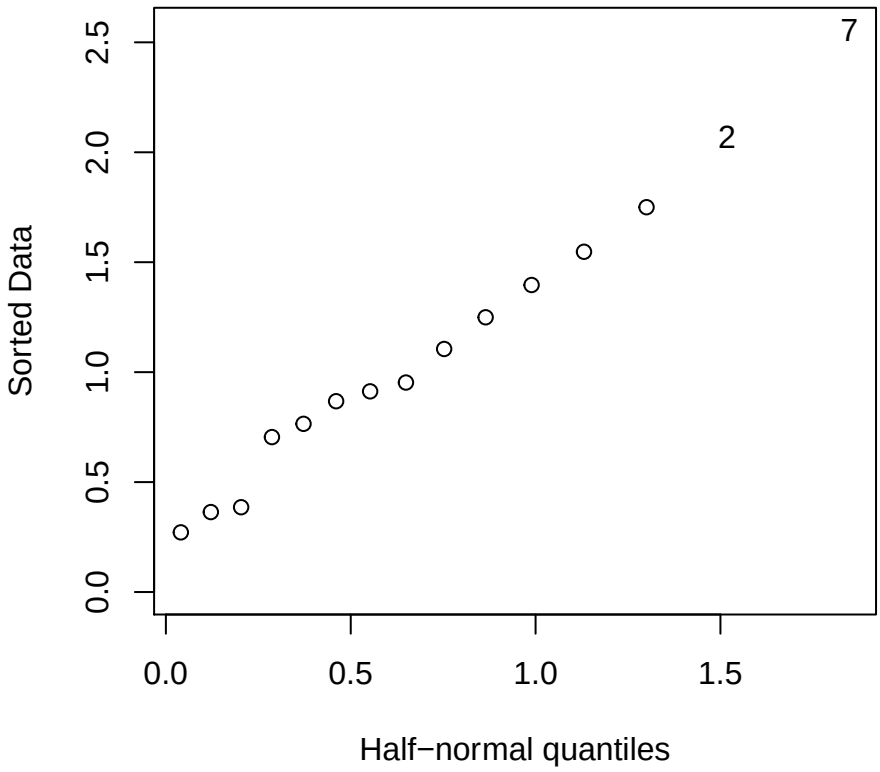
```
# number of observations for each sub-sample
summary(turtles.df$turtles)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.000   8.000   9.000   9.067  10.000  13.000
```

Using our rule of thumb of $m_i > 5$, we can say that the data are not sparse.

Part d

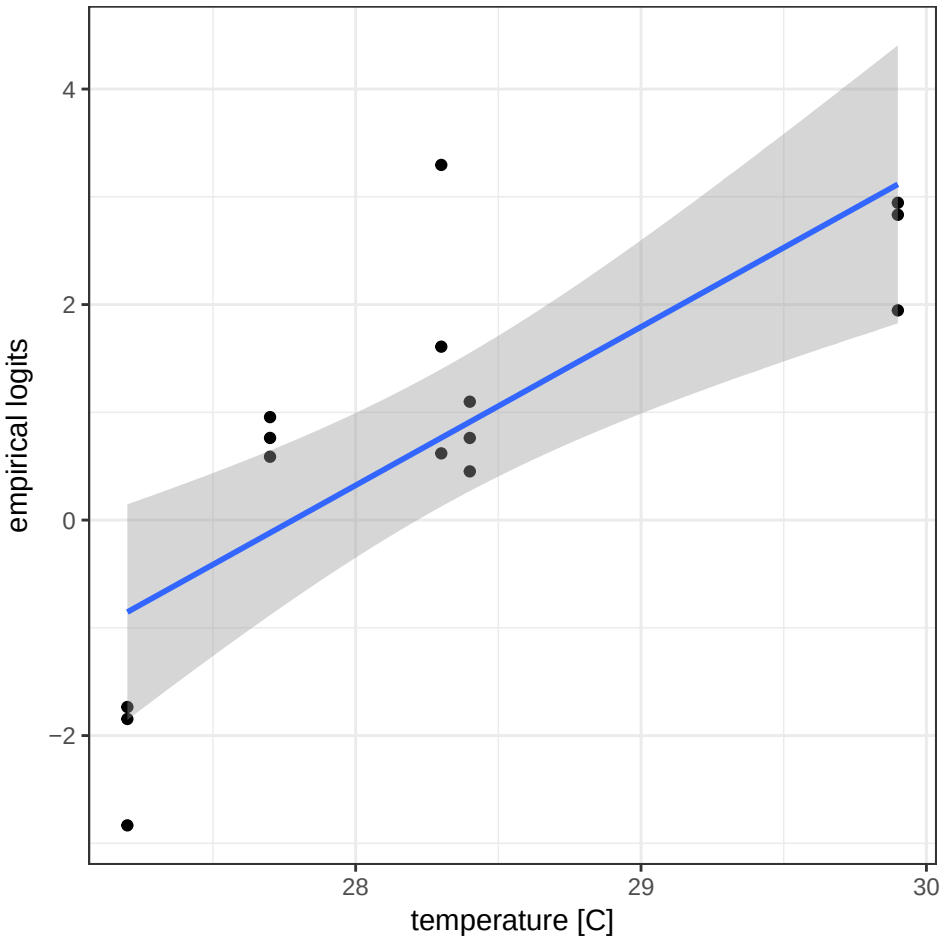
```
# half normal plot
faraway::halfnorm(residuals(binom.mod))
```



Based on the half-normal plot, we have no reason to believe that the data contain outliers.

Part e

```
# empirical log plot
ggplot(turtles.df) +
  geom_point(aes(x = temp, y = log((male + .5) / (turtles - male + .5)))) +
  labs(x = 'temperature [C]', y = 'empirical logits') +
  stat_smooth(aes(x = temp, y = log((male + .5) / (turtles - male + .5))),
              method = 'lm')
```



Plotting the empirical logits vs the temperature suggests that the relationship is not linear. It appears that a concave curve is more appropriate. A transformation or higher order term might be appropriate.

Part f

```
# model with quadratic term
quad.binom.mod <- glm(cbind(male, female) ~ temp + I(temp ** 2),
                      data = turtles.df, family = binomial)
# wald tests
summary(quad.binom.mod)
```

```
Call:
glm(formula = cbind(male, female) ~ temp + I(temp^2), family = binomial,
    data = turtles.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6703  -0.8875  -0.4194   0.9481   2.2198

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -677.5950   268.7984  -2.521   0.0117 *
```

```
temp          45.9173    18.9169   2.427   0.0152 *
I(temp^2)     -0.7745     0.3327  -2.328   0.0199 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.508  on 14  degrees of freedom
Residual deviance: 20.256  on 12  degrees of freedom
AIC: 51.15

Number of Fisher Scoring iterations: 4
```

```r
# check for model fit
pchisq(quad.binom.mod$deviance, quad.binom.mod$df.residual, lower = FALSE)
```

```
[1] 0.06239194
```

```r
# LR test
anova(binom.mod, quad.binom.mod, test = 'Chi')
```

```
Analysis of Deviance Table

Model 1: cbind(male, female) ~ temp
Model 2: cbind(male, female) ~ temp + I(temp^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        13     24.942
2        12     20.256  1   4.6863   0.0304 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using a significance level of $\alpha = 0.05$, we can say that the quadratic term is significant and the model (kinda) fits the data.

Part g

```r
# sample variance
var(turtles.df$male)
```

```
[1] 12.35238
```

```r
# sample variance assuming binomial
turtles.df %>%
  dp$group_by(temp) %>%
  dp$summarise(var.male = sum(male) * sum(female) / sum(turtles)) %>%
  dp$ungroup() %>%
  .$var.male %>%
  sum()  # assuming independence
```

```
[1] 16.87077
```

```r
# dispersion parameter estimate
s2 <- sum(residuals(quad.binom.mod, type = 'pearson') ** 2) /
  quad.binom.mod$df.residual

# F tests
```

```
drop1(quad.binom.mod, scale = s2, test = 'F')
```

```
Single term deletions

Model:
cbind(male, female) ~ temp + I(temp^2)

scale:  1.438774

          Df Deviance    AIC F value Pr(>F)
<none>         20.256 51.150
temp       1   25.366 52.702  3.0271 0.1074
I(temp^2)  1   24.942 52.407  2.7762 0.1215
```

Comparing the sample variance to the sample variance assuming a binomial model, as well as the estimate for the dispersion parameter, suggests that there is some overdispersion.

Part h

```
# aggregated data
agg.turtles.df <- turtles.df %>%
  dp$group_by(temp) %>%
  dp$summarise_all(sum) %>%
  dp$ungroup()

# model
combin.binom.mod <- glm(cbind(male, female) ~ temp,
                        data = agg.turtles.df, family = binomial)
# wald tests
summary(combin.binom.mod)
```

```
Call:
glm(formula = cbind(male, female) ~ temp, family = binomial,
    data = agg.turtles.df)

Deviance Residuals:
     1       2       3       4       5
-2.224   2.248   1.239  -1.382  -1.191

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
temp          2.2110     0.4309   5.132 2.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64.429  on 4  degrees of freedom
Residual deviance: 14.863  on 3  degrees of freedom
AIC: 33.542
```

```
Number of Fisher Scoring iterations: 5
```

```r
# check for model fit
pchisq(combin.binom.mod$deviance, combin.binom.mod$df.residual, lower = FALSE)
```

```
[1] 0.001937595
```

As before, the residual deviance suggests that the model does not fit the data. The summary outputs for both this model and the model from part (b) are identical.