ENV 710

binomial logistic regression
or log-linear models

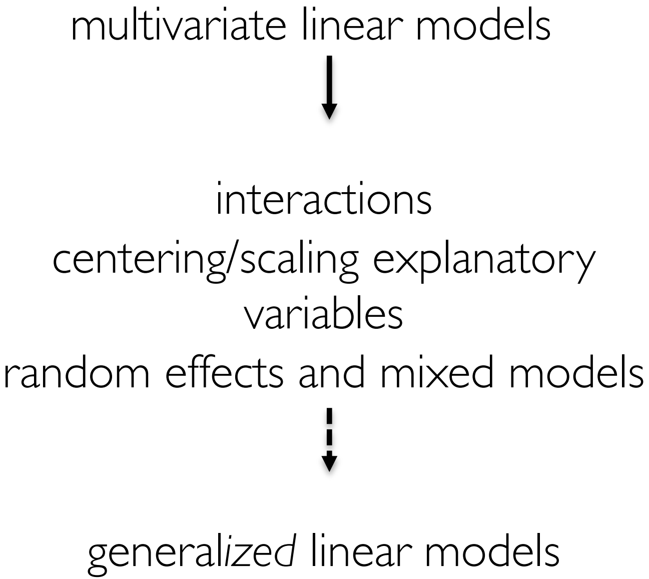
Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

roadmap

- download: `pheno.csv`
- load packages:

```
COUNT
DHARMa
boot
vcdExtra
ResourceSelection
ggplot2
pscl
```

where we are



multivariate linear models

interactions
centering/scaling explanatory variables
random effects and mixed models

generali*zed* linear models

logistic regression

- binary response variable
  - success or failure (1 or 0)
  - model the probability of success (π)

$$\pi = \beta_0 + \beta_1 X_1$$

- model log odds

$$logit(\pi) = log[\pi/(1-\pi)]$$

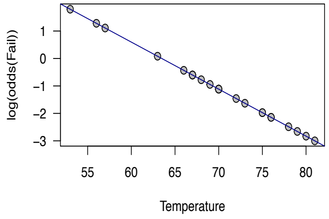
$$logit(\pi) = \beta_0 + \beta_1 X_k + ... + \beta_k X_k$$

- calculate probabilities

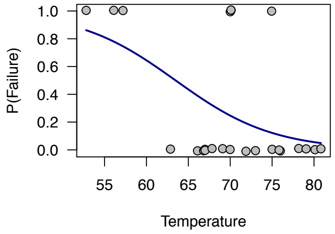
$$\pi = \frac{exp(\beta_0 + \beta_1 X_k + ... + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_k + ... + \beta_k X_k)}$$

goal is to model the probability of 'success'

transform response to be linearly related to the IV's through the logit transform



back transform to probability using the inverse logit



logistic regression for binomial counts

- binomial count = sum of independent binary responses
- Y ~ binomial(n, π), with population proportion, π, and n trials
- observed proportion of 1's is Y/n – binomial proportion
- continuous proportions of amounts cannot be modeled
  - e.g., the proportion of fat that is saturated fat
  - numerator and denominator are not integers, and no n is involved
  - when response variable is a continuous proportion, use ordinary regression methods

- model how population proportion depends on the explanatory variables through a nonlinear link function

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- interpretation of model is same as for binary logistic regression, which is a special case of the binomial model in which all n_i's are 1

e.g., Titanic

```
   survive cases age sex class
1        1     1   0   0     1
2       13    13   0   0     2
3       14    31   0   0     3
4        5     5   0   1     1
5       11    11   0   1     2
6       13    48   0   1     3
7      140   144   1   0     1
8       80    93   1   0     2
9       76   165   1   0     3
10      57   175   1   1     1
11      14   168   1   1     2
12      75   462   1   1     3
```

binomial regression

$Y = x$ out of n trials

```
  pclass survived    sex    age surv
1    1st survived female  29.00    1
2    1st survived   male   0.91    1
3    1st     died female   2.00    0
4    1st     died   male  30.00    0
5    1st     died female  25.00    0
6    1st survived   male  48.00    1
```

logistic regression

$Y = 1$ or 0

I – Phenology

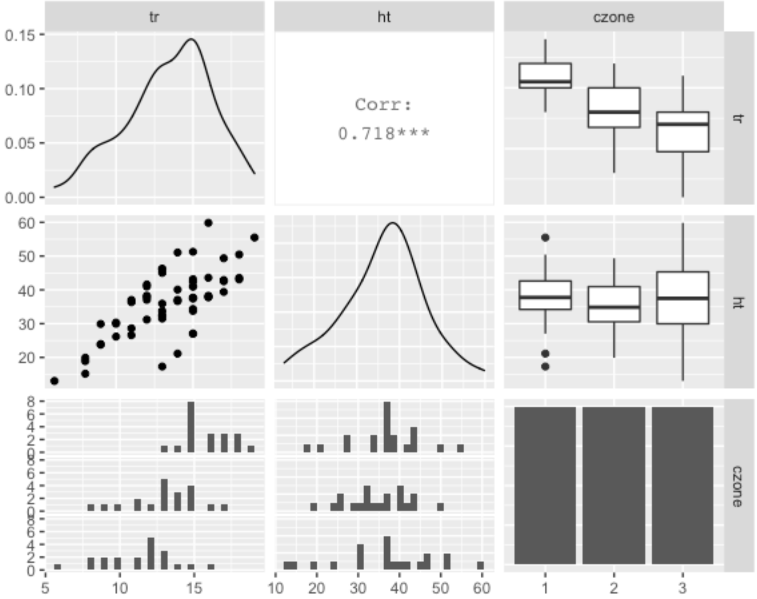
Do minimum nighttime temperature and height of trees influence proportion of trees that flower?

- 20 samples of 20 trees surveyed in each zone
- `czone`: 3 climate zones (hi, med, low nighttime temps)
- `height`: average tree height of each sample of trees
- `n`: number of trees sampled

I – Phenology

Do minimum nighttime temperature and height of trees influence proportion of trees that flower?

```
   tr      ht     czone   n
1  13    18.09      1     20
2  13    25.31      1     20
3  15    26.39      1     20
4  15    29.98      1     20
5  15    30.61      1     20
6  15    33.27      1     20
  ...     ...      ...    ...
```



```
glm(cbind(tr, n - tr) ~ factor(czone) * ht,
          family = binomial, data = trees)
```

```
 b0 <- glm(cbind(tr, n-tr)~factor(czone)*ht, family=binomial,
             data=trees)
```

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.326077   0.515127  -0.633  0.52673
factor(czone)2    -1.403413   0.733661  -1.913  0.05576 .
factor(czone)3    -0.937770   0.616617  -1.521  0.12830
ht                 0.047032   0.014231   3.305  0.00095 ***
factor(czone)2:ht  0.021752   0.020764   1.048  0.29482
factor(czone)3:ht -0.004873   0.016824  -0.290  0.77208
---

    Null deviance: 114.2732  on 59  degrees of freedom
Residual deviance:   5.8137  on 54  degrees of freedom
AIC: 211.19
```

```
b1<-update(b0,.~.-factor(czone):ht)
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.38963    0.27057  -1.440     0.15
factor(czone)2 -0.66316    0.16544  -4.008 6.11e-05 ***
factor(czone)3 -1.11444    0.16438  -6.779 1.21e-11 ***
ht              0.04884    0.00683   7.152 8.58e-13 ***
---

    Null deviance: 114.2732  on 59  degrees of freedom
Residual deviance:   8.1554  on 56  degrees of freedom
AIC: 209.53
```

I – Phenology

```
Compare models

lrtest(b0, b1)
Likelihood ratio test

Model 1: cbind(tr, n - tr) ~ factor(czone) * ht
Model 2: cbind(tr, n - tr) ~ factor(czone) + ht
  #Df   LogLik Df  Chisq Pr(>Chisq)
1   6  -99.595
2   4 -100.766 -2 2.3417     0.3101
```

```
Check model fit

pchisq(b1$deviance, b1$df.residual, lower.tail = F)
1
```

```
Check for overdispersion

d2 = sum(residuals(b1,"pearson")^2)
 disp = d2/df.residual(b1)
  phi = sqrt(disp)

[1] 0.3801822
```

```
DHARMa::testDispersion(b1)

DHARMa nonparametric dispersion test via sd of
residuals fitted vs.
simulated

data:  simulationOutput
dispersion = 0.67906, p-value = 0.024
alternative hypothesis: two.sided
```

I – Phenology

Coefficients as log-odds

```
coef(b1)
Intercept)   factor(czone)2  factor(czone)3      ht
-0.38962645  -0.66315645      -1.11443652     0.04884507
```

Coefficients as odds

```
exp(coef(b1))
(Intercept) factor(czone)2   factor(czone)3    ht
 0.6773098   0.5152225          0.3281001   1.0500577
```

Coefficients as probability

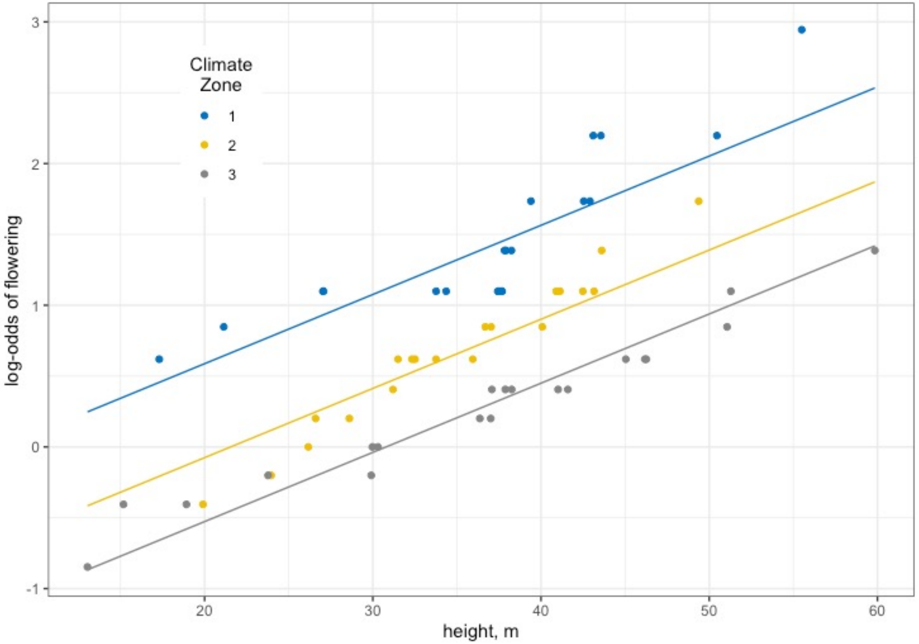
```
inv.logit(coef(b1))
(Intercept) factor(czone)2 factor(czone)3        ht
 0.4038072      0.3400309      0.2470447      0.5122088
```

I – Phenology

Coefficients as log-odds

```
coef(b1)
Intercept)    factor(czone)2  factor(czone)3      ht
-0.38962645  -0.66315645       -1.11443652     0.048845
```

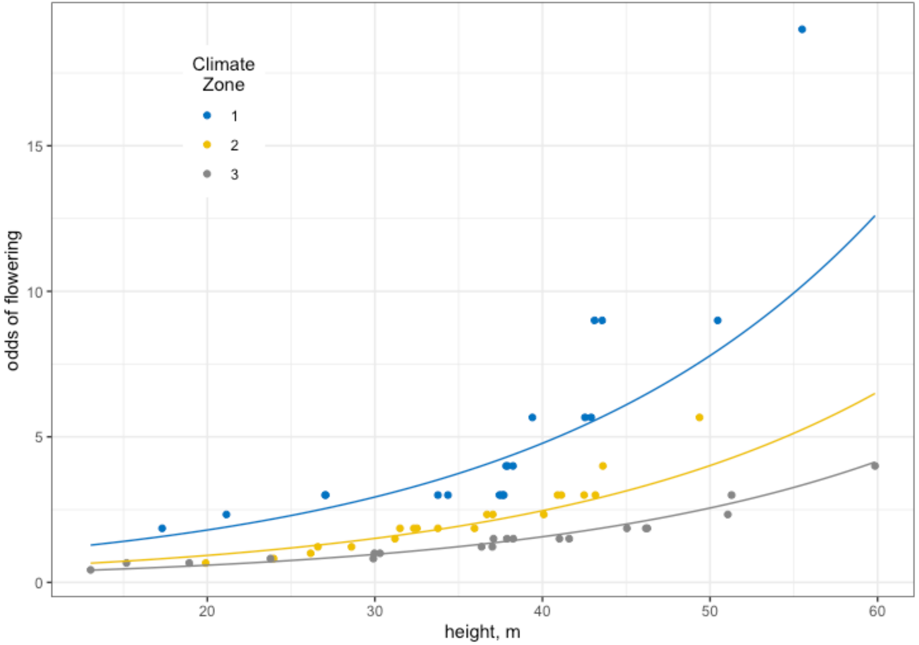
- `Intercept` is log-odds of flowering at czone 1 when tree height is 0
- `czone 2` is the difference in the log-odds of flowering for czone 2 compared to czone 1: log-odds of flowering decreases by -0.66 from czone 1 to 2.
- log-odds of flowering in czone 2 is -0.389+-0.663 = -1.052.
- `ht` is the change in log-odds of flowering for every additional meter of tree height.



I – Phenology

```
Coefficients as odds

exp(coef(b1))
(Intercept)  factor(czone)2    factor(czone)3     ht
 0.6773098    0.5152225           0.3281001    1.0500577
```

- `Intercept` is the odds of flowering at czone 1 when tree height is 0
- `czone  2` is the odds ratio of flowering between czone 2 and czone 1
- mean odds of flowering in czone 2 is 0.349 `[exp(-0.38962645+-0.66315645)]`
- `ht`, the odds of flowering increase by a factor of 1.05 for every meter, or approximately 5% increase in odds of flowering for each additional unit of tree height



I – Phenology

```
Coefficients as probability

inv.logit(coef(b1))
(Intercept)  factor(czone)2  factor(czone)3          ht
 0.4038072       0.3400309       0.2470447     0.5122088
```

- log-odds coefficients can be converted into probabilities with the `inv.logit()`, but can't be directly interpreted because the relationship is not linear
- best to evaluate probability at specific values of predictors

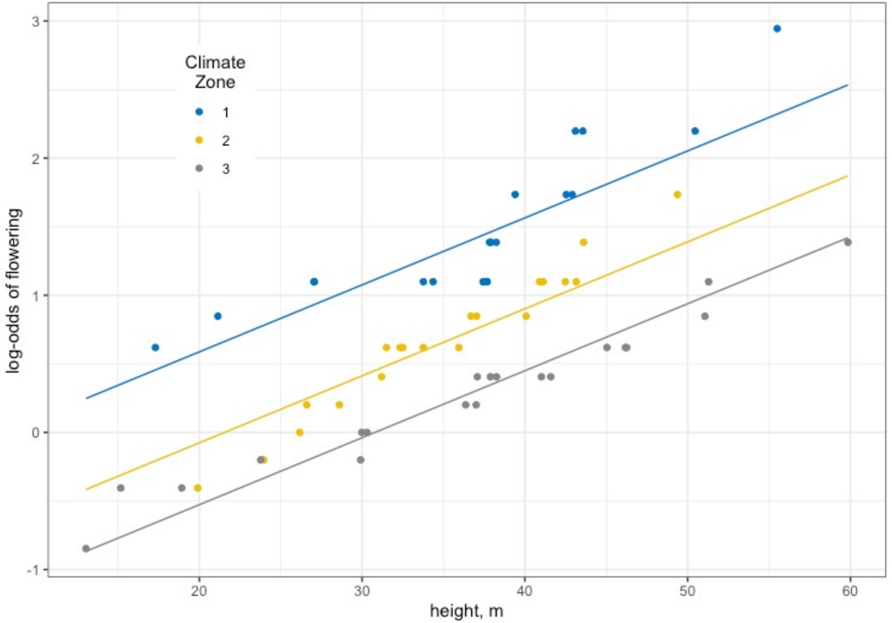
$$p = \frac{1}{(1 + 1/e^{\beta})}$$

```
Probability of tree flowering in climate
zone 2 at 15, 35, and 50 m of height

cfs <- coef(b1)
inv.logit(cfs[1] + cfs[2] + cfs[3]*0 +
          cfs[4]*c(15, 35, 50))

[1] 0.4206497 0.6585400 0.8005077
```

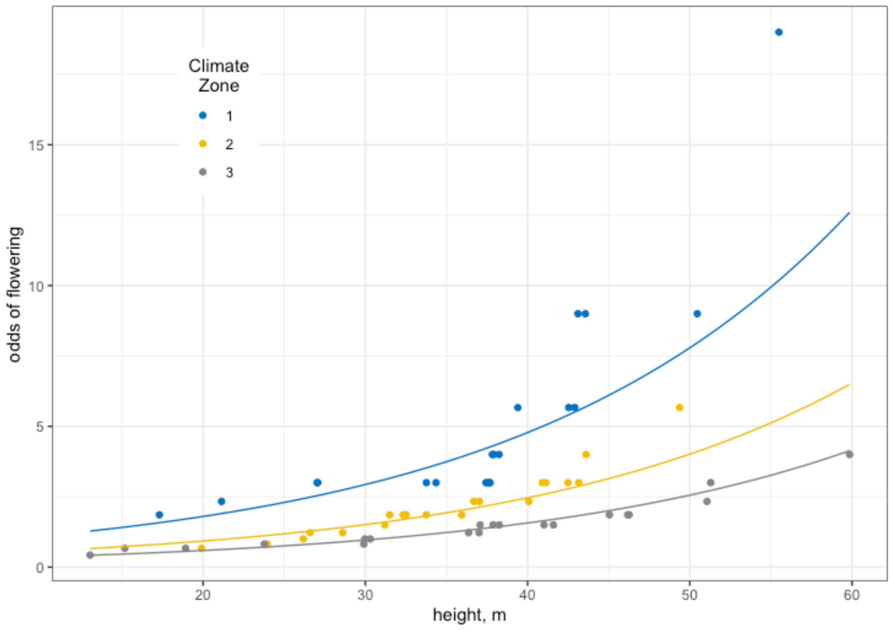
I – Phenology

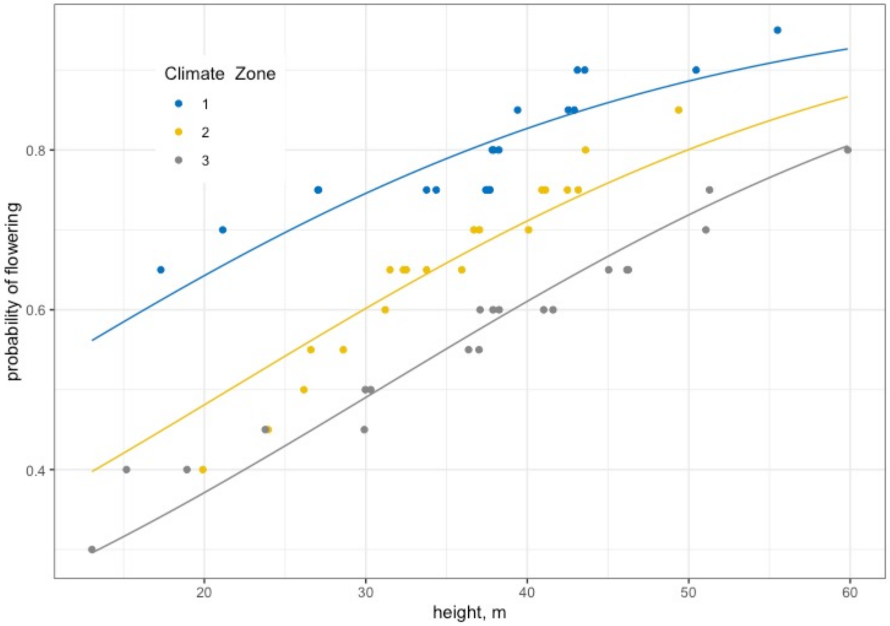
"I'll do algebra, I'll do trigonometry, I'll even do statistics…
But graphing is where I draw the line."

```
jcoPalette <- c("#0073C2FF", "#EFC000FF", "#868686FF", "#CD534CFF",
"#7AA6DCFF", "#003C67FF", "#8F7700FF", "#3B3B3BFF", "#A73030FF",
"#4A6990FF")


b <- coef(b1)

ggplot(data = trees, aes(x = ht, y = log(tr/(n-tr)))) +
  geom_point(aes(color = czone)) +
     stat_function(fun = function(x){b[1]+b[4]*x}, colour =
jcoPalette[1]) +
     stat_function(fun = function(x){b[1]+b[2]+b[4]*x}, colour =
jcoPalette[2]) +
     stat_function(fun = function(x){b[1]+b[3]+b[4]*x}, colour =
jcoPalette[3]) +
  theme_bw() + scale_colour_manual("Climate\n  Zone",
values=jcoPalette) +
  theme(legend.position = c(0.2, 0.8)) +
  labs(x = "height, m", y = "log-odds of flowering")
```



I – Phenology


Climate Zone
● 1
● 2
● 3
odds of flowering
0
5
10
15
20
30
40
50
60
height, m


Climate Zone
● 1
● 2
● 3
probability of flowering
0.4
0.6
0.8
20
30
40
50
60
height, m

I – Phenology

```
## plot as odds with ggplot
ggplot(data = trees, aes(x = ht, y = tr/(n-tr))) +
  geom_point(aes(color = czone)) +
     stat_function(fun = function(x){exp(b[1]+b[4]*x)},
                   colour = jcoPalette[1]) +
     stat_function(fun = function(x){exp(b[1]+b[2]+b[4]*x)},
                   colour = jcoPalette[2]) +
     stat_function(fun = function(x){exp(b[1]+b[3]+b[4]*x)},
                   colour = jcoPalette[3]) +
  theme_bw() + scale_colour_manual("Climate\n  Zone", values=jcoPalette) +
  theme(legend.position = c(0.2, 0.8)) +
  labs(x = "height, m", y = "odds of flowering")

## plot as probabilities with ggplot
ggplot(data = trees, aes(x = ht, y = tr/n)) +
  geom_point(aes(color = czone)) +
     stat_function(fun = function(x){inv.logit(b[1]+b[4]*x)},
                   colour = jcoPalette[1]) +
     stat_function(fun = function(x){inv.logit(b[1]+b[2]+b[4]*x)},
                   colour = jcoPalette[2]) +
     stat_function(fun = function(x){inv.logit(b[1]+b[3]+b[4]*x)},
                   colour = jcoPalette[3]) +
  theme_bw() + scale_colour_manual("Climate\  Zone", values=jcoPalette) +
  theme(legend.position = c(0.2, 0.8)) +
  labs(x = "height, m", y = "probability of flowering")
```

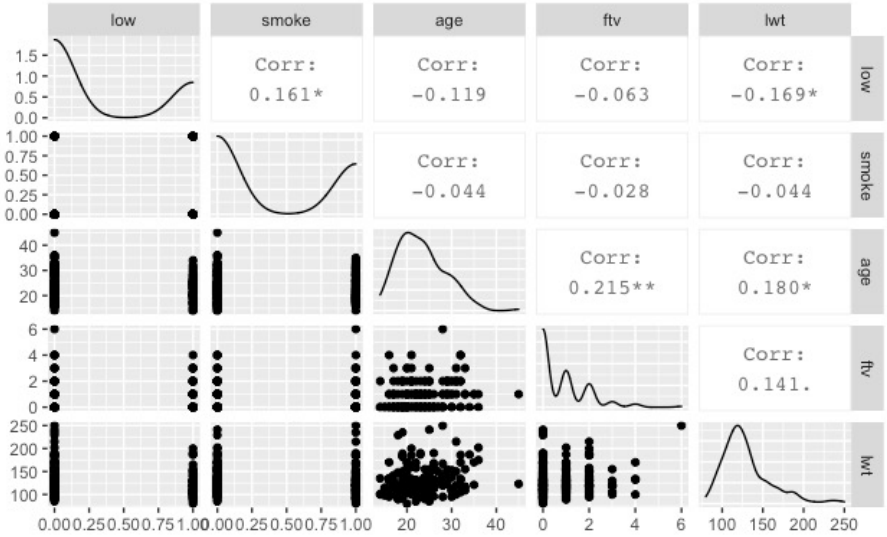
2 – Low birth weight

What variables result in low birth weight?

- low: 1=low birthweight baby; 0=normal weight
- smoke: 1=history of mother smoking;
  0=mother nonsmoker
- age: age of mother: 14-45
- lwt: weight (lbs) at last menstrual period: 80-250 lbs
- ftv: number of physician visits in 1st trimester: 0-6

Build model, reduce it, check its goodness of fit, and interpret the results

```
require(COUNT)
data(lbw)
 lbw$lwt <- as.numeric(lbw$lwt)
  ggpairs(lbw, columns = c("low", "smoke", "age",
          "ftv", "lwt"))
```



2 – Low birth weight

```
lr1 <- glm(low ~ factor(smoke) + age + ftv + lwt,
             family = "binomial", data = lbw)
 lr2 <- update(lr1, .~.-ftv)
  lr3 <- update(lr2, .~.-age)
  summary(lr3)

Call:
glm(formula = low ~ factor(smoke) + lwt,
    family = "binomial", data = lbw)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.619201   0.795870   0.778   0.4366
factor(smoke)1  0.676579   0.324685   2.084   0.0372 *
lwt            -0.013301   0.006088  -2.185   0.0289 *
---

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 224.36  on 186  degrees of freedom
AIC: 230.36
```

```
lrtest(lr1, lr2, lr3)
Likelihood ratio test

Model 1: low ~ factor(smoke) + age + ftv + lwt
Model 2: low ~ factor(smoke) + age + lwt
Model 3: low ~ factor(smoke) + lwt
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -111.42
2   4 -111.45 -1 0.0645     0.7995
3   3 -112.18 -1 1.4634     0.2264
```

```
pchisq(lr3$deviance, lr3$df.residual,
lower.tail = F)
[1] 0.02868237
```

```
> pR2(lgr4)
fitting null model for pseudo-r2
      llh         llhNull           G2
-112.17946472 -117.33599810   10.31306675
  McFadden         r2ML            r2CU
0.04394673     0.05310445      0.07468003
```

2 – Low birth weight

```
lr1 <- glm(low ~ factor(smoke) + age + ftv + lwt,
            family = "binomial", data = lbw)
 lr2 <- update(lr1, .~.-ftv)
  lr3 <- update(lr2, .~.-age)
  summary(lr3)

Call:
glm(formula = low ~ factor(smoke) + lwt,
    family = "binomial", data = lbw)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.619201   0.795870   0.778   0.4366
factor(smoke)1  0.676579   0.324685   2.084   0.0372 *
lwt            -0.013301   0.006088  -2.185   0.0289 *
---

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 224.36  on 186  degrees of freedom
AIC: 230.36
```

How does smoking affect the odds of having a low birthweight baby?

What is the effect of low mother weight on the odds of having a low birthweight baby?

What is the probability of having a low birthweight baby for a smoker weighing in the lowest 25% of women?

2 – Low birth weight

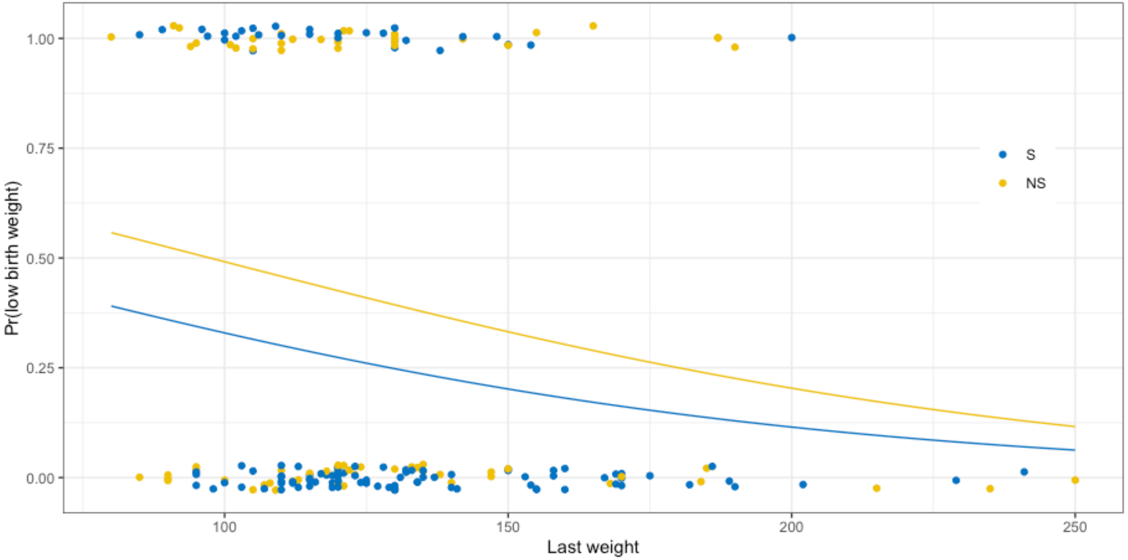
```
lr1 <- glm(low ~ factor(smoke) + age + ftv + lwt,
            family = "binomial", data = lbw)
 lr2 <- update(lr1, .~.-ftv)
  lr3 <- update(lr2, .~.-age)
  summary(lr3)

Call:
glm(formula = low ~ factor(smoke) + lwt,
    family = "binomial", data = lbw)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)  
(Intercept)     0.619201   0.795870   0.778   0.4366  
factor(smoke)1  0.676579   0.324685   2.084   0.0372 *
lwt            -0.013301   0.006088  -2.185   0.0289 *
---

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 224.36  on 186  degrees of freedom
AIC: 230.36
```

Smoking increases the odds of having a low birthweight baby by 1.97 times or 97% (z = 2.08, p = 0.037).

Every additional pound of weight decreases the odds of a woman having a low birthweight child by 1.3% (z = -2.19, p = 0.029).

The probability of having a low birthweight child for a smoker in the 25th percentile of weights is 45.8%.

2 – Low birth weight



```
ggplot(lbw, aes(x = as.numeric(lwt), y = as.numeric(low),
                  color=as.factor(smoke))) +
            stat_function(fun = function(x){inv.logit(lr3c[1]+lr3c[3]*x)},
                    colour = jcoPalette[1]) +
            stat_function(fun = function(x){inv.logit(lr3c[1]+lr3c[2] + lr3c[3]*x)},
                    colour = jcoPalette[2]) +
        geom_point(position=position_jitter(height=0.03, width=0)) +
        xlab("Last weight") + ylab("Pr(low birth weight)") + theme_bw() +
   theme(legend.position=c(0.9, 0.7)) +
   scale_colour_manual("", values=jcoPalette, labels = c("S", "NS"))
```

Questions?