Lecture 1: Intro and Regression Model

- sample means $(\bar{x}, \bar{y})$ and variances (s_x^2, s_y^2)

$$\bar{x} = \frac{\sum_i^n x_i}{n} \qquad \bar{y} = \frac{\sum_i^n y_i}{n}$$
$$s_x^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1} \qquad s_y^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$$
$$\therefore \sum_{i=1}^n (x_i - \bar{x})^2 = s_x^2(n-1) \qquad \sum_{i=1}^n (y_i - \bar{y})^2 = s_y^2(n-1)$$

  - note: $\sum_i^n y_i = n\bar{y}$

- covariance (r_{xy}) and correlation (r_{xy})

$$s_{xy} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

  - note: denominator is standard deviation, not variance

- regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  - note: $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$

- in case we get asked to write RSS (residual sum squared in terms of summary statistics) or RMS (residual mean square)

$$RMS = s^2 = \frac{RSS}{n-2} = \frac{n-1}{n-2}(s_y^2 + \hat{\beta}_1^2 s_x^2 - 2\hat{\beta}_1 r_{xy} s_x s_y)$$

- observational studies: research observe the effect of a treatment without trying to change who is or isn't exposed to it
- experimental studies: researcher introduce an intervention and studies the results

Lecture 2: Residuals

- residual: $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$
- Residual SS (residual sum squared): Res $SS = \sum_i^n e_i^2 = \sum_i^n (y_i - b_0 - b1x_i)^2$ (we want to minimize this)
- how to find b_0 and b_1 - we can get an estimate using

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$
$$\hat{b}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{r_{xy} s_y}{s_x}$$

- we can rewrite the regression line as: $y = \bar{y} + \hat{b}_1(x - \bar{x})$
- using residual plots:
  - e_i against $\hat{y}_i$ or x_i: you want a random cloud pattern
    - → note: do not plot against y_i because they're related to each other
  - e_i against Normal scores (quantile plot): the data points should fall onto the line (that means normal errors is a reasonable assumption)
  - e_i against e_{i-1}: you want random cloud pattern - means there are no serial correlation

Lecture 3: Residuals (continued)

- here we'll start to treat Y as a random variable

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon$$
$$E(Y_i) = E(\beta_0) + x_i E(\beta_1) + E(\varepsilon) = \beta_0 + \beta_1 x_i$$
$$Var(Y_i) = 0 + Var(\varepsilon) = 0 + \sigma^2$$
$$\therefore Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

  - where β_0, β_1, and σ^2 are population parameters (unknown)

Lecture 4: Confidence Interval

- review of CI: $CI = \hat{\theta} \pm c \times se(\hat{\theta})$
  - $\hat{\theta}$ is the estimate of the parameter
  - c is the percentile sampling distribution of the sampling distribution of $\hat{\sigma}$ (i.e. qt(0.975, df))
  - $se(\hat{\theta})$ is an estimator of the standard deviation of
- CI for slope parameter
  - blah blah B_1 is an unbiased estimator of the true parameter β_1
  - because we don't know σ^2, we will approximate it using s^2 (residual mean square)
  - we'll also use $df = n - k$ (where k is the number of parameter in the model - classic case it'll be 2)

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} \approx \sigma^2$$
$$CI = \hat{b}_1 \pm t_{(0.975,\ n-2)} \sqrt{\frac{s^2}{\sum (x_i - \bar{x})^2}}$$

- CI for the intercept parameter
  - same assumption as above, we don't know σ^2 so we'll use s^2 (calculated same way as above)
  - B_0 is also an unbiased estimator of β_0

$$CI = \hat{b}_0 \pm t_{(0.975,\ n-2)} \sqrt{\frac{s^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}}$$

- CI for the expected response
  - suppose we wish to predict the expected value μ_0 of Y for a given value $x = x_0$ - we can use estimator $\hat{\mu}_0 = \bar{Y} + B_1(x_0 - \bar{x})$ (replace with estimates when it comes to calculation time)

$$CI = \hat{\mu}_0 \pm t_{(n-2, 0.975)} \sqrt{\frac{s^2}{n} + \frac{(x_0 - \bar{x})^2 s^2}{\sum (x_i - \bar{x})^2}}$$

Lecture 5: Prediction Interval

- now we wish to predict a particular response, Y_* at $x = x_*$, we say that $\hat{y}_* = \hat{b}_0 + \hat{b}_1 x_*$

$$E(\hat{Y}_* - Y_*) = 0$$
$$Var(\hat{Y}_* - Y_*) = \frac{\sigma^2}{n} + \frac{(x_* - \bar{x})^2 \sigma^2}{\sum (x_i - \bar{x})^2} + \sigma^2$$
$$PI = \hat{y}_* \pm t_{(0.975, n-2)} \times s \times \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Lecture 6: Inference in Regression

- after using some distributions rules, we can say
  - $se(B_i) \sim \chi_{n-2}$
  - $\frac{B_1 - \beta_1}{se(B_1)} = \frac{B_1 - \beta_1}{S/s_x\sqrt{n-1}} \sim t_{n-2}$
- hypothesis testing for β_1 (usually frame in a way like want to prove y is not co-linearly dependent with x - if it is then $B_1 \neq 0$
  - hypothesis: $H_0 : \beta_1 = 0 \quad$ vs $\quad H_1 : \beta_1 \neq 0$
  - under the null (test hypotehsis):

$$\frac{B_1 - \beta_1}{se(B_1)} = \frac{B_1 - \beta_1}{S/s_x\sqrt{n-1}} = \frac{B_1}{S/s_x\sqrt{n-1}} \sim t_{n-2}$$

  - reject if pt(test_stat, df=n-2) is less than α (watch for which tail since this is a 2-sided test) TODO

Lecture 8: Matrix Notation

- $\vec{y}$ is a column vector with each of the response value y_i
- $\mathbf{X}$ is a matrix with 1's in the first column (model bias/intercept) then the x_i values in the 2nd column
- we can get estimates $\hat{b} =< \hat{b}_0, \hat{b}_1 >$ by

$$X^T X \vec{b} = X^T y$$
$$\hat{b} = (X^T X)^{-1} X^T y$$

- model diagnostic statistics:
  - fitted values (vector): $\hat{y} = \mathbf{X}\vec{b}$
  - residuals (vector): $e = y - \hat{y}$
  - sum of residual squares (scalar): $SS(Res) = e^T e$
  - residual mean square: $s^2 = \hat{\sigma}^2 = \frac{e^T e}{n-k} = \frac{SS(Res)}{n-k}$
    - → this is an estimator of σ^2
    - → k is the number of columns in X
- multiple correlation coefficient (aka coefficient of determination - R^2 ratio)

$$R^2 \coloneqq \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SS(Res)}{y^T y - n\bar{y}^2}$$

- when $R^2 = 1$ (its max): the model fits perfectly (bigger = better)
- adding new predictor variable to a model cannot make the model fit worse as measured R^2 (so R^2 cannot decrease by the addition of the new term in the model) - **but $\hat{\sigma}^2$ may not necessarily decrease by the addition of the new term** - all this is to say **you shouldn't use R^2 as a tool when deciding to include a new term or not (or model selection in general)**

- adjusted R^2: this we can use for model selection (bigger is better)

$$\text{adjusted } R^2 = 1 - \frac{\frac{SS(Res)}{n-k}}{\frac{y^T y - n\bar{y}^2}{n-1}} = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)}$$

- by taking into account df, we're able to compare models fitted to different data sets, and models of different sizes for the same data

- variance of $\hat{\beta}$

$$Var(B_i) = \sigma^2 (X^T X)^{-1}_{(i+1,i+1)}$$
$$se(B_i) = \hat{\sigma}\sqrt{(X^T X)^{-1}_{(i+1,i+1)}}$$

- recall: σ^2 estimated by the residual sum squared error ($\hat{\sigma}^2$ or s^2)
- diagonal entries of matrix gives variance and the other entries give covariance (note: matrix indexing starts at 1)
- ex. $se(B_2) = \hat{\sigma}\sqrt{(X^T X)^{-1}_{(3,3)}}$

- CI for parameters under matrix notation:

$$CI = \hat{b}_i \pm t_{n-k,\ 1-\alpha/2} \times \hat{\sigma}\sqrt{(X^T X)^{-1}_{(i+1,i+1)}}$$

- CI for expected response at a given set of explanatory variables value $\mathrm{x}_0 = (1, x_1^*, x_2^*, \ldots, x_p^*)^T$:

$$CI = \hat{\mu}_Y(\mathrm{x}_0) \pm t_{n-k,1-\alpha/2}\hat{\sigma}\sqrt{\mathrm{x}_0^T (X^T X)^{-1}\mathrm{x}_0}$$

- PI for predicted response:

$$CI = \hat{\mu}_Y(\mathrm{x}_0) \pm t_{n-k,1-\alpha/2}\hat{\sigma}\sqrt{\mathrm{x}_0^T (X^T X)^{-1}\mathrm{x}_0 + 1}$$

Lecture 9: Properties of Residuals (Matrix)

- note: errors are random values, residuals is the difference between the fitted values
- let $\mathbf{E}$ be the random vector from which we observe the residuals e
  - important: define the hat matrix

$$P = X(X^T X)^{-1} X^T$$

  - then we have

$$\mathbf{E} = (I_n - P)\varepsilon$$
$$E(\mathbf{E}) = 0 \qquad Var(\mathbf{E}) = (I_n - P)\sigma^2$$

- we can define the residual MS as

$$MS(Res) = \frac{\mathbf{E}^T\mathbf{E}}{n-k}$$
$$E\left(MS(Res)\right) = E\left(\frac{\mathbf{E}^T\mathbf{E}}{n-k}\right) = \sigma^2 \qquad Var(\mathbf{E}_i) = \sigma^2(1 - P_{ii})$$

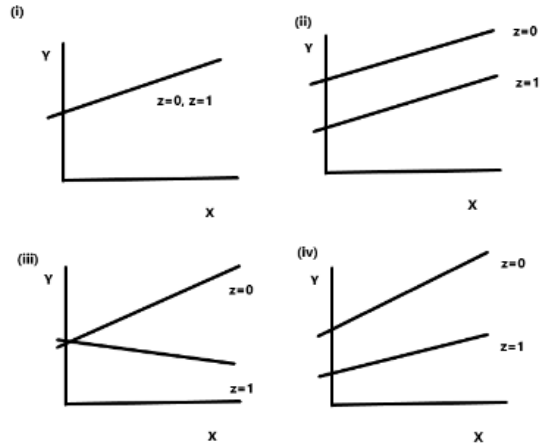
  - last line in the eq above is how you can find the individual variance of a single residual

- standardized individual residual

$$E_i' = \frac{E_i}{\sigma\sqrt{1-P_{ii}}}, \quad i = 1, \ldots, n$$

Lecture 10: Categorical Variables

- motivating example: we have a linear model involving a response variable Y and two predictors x (continuous) and z (binary) - there are 4 scenarios
  1. same line fits Y for both category of z
  2. different lines fits Y for the two categories of z but the lines have the same slope
  3. different lines fit Y for the two categories of z, but they have the same intercept (additive model)
  4. different lines fit Y for the two categories of z, the lines different in both slope and intercept (multiplicative model - includes interaction term)



- parameterising the model

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$
$$Y = \beta_0 + \beta_1 x + \gamma_0 z + \varepsilon \tag{2}$$
$$Y = \beta_0 + \beta_1 x + \gamma_1 zx + \varepsilon \tag{3}$$
$$Y = \beta_0 + \beta_1 x + z(\gamma_0 + \gamma_1 x) + \varepsilon \tag{4}$$

  - in this case $z = 0$ is effectively the "baseline"
  - so for (2), when $z = 1$, it alters the intercept to $B_0 + \gamma_0$ - normal otherwise
  - for (3), when $z = 1$, it alters the slope to $x(\beta_1 + \gamma_1)$
  - for (4), when $z = 1$, it alters both the slope and intercept (combine the two above)

- number of parameters: find number of scenarios and then each scenarios need 2 parameter each
- if we added another to include another binary term - you might have to introduce another term to model when both the events happen

$$z_i = \begin{cases} 0 & \text{if jumper was male} \\ 1 & \text{if jumper was female} \end{cases}$$

$$w_i = \begin{cases} 0 & \text{if distance } i \text{ was not jumped at altitude} \\ 1 & \text{if distance } i \text{ was jumped at altitude} \end{cases}$$

$$Y = \beta_0 + \beta_1 x + z(\gamma_0 + \gamma_1 x) + w(\delta_0 + \delta_1 x) + zw(\alpha_0 + \alpha_1 x)$$

  - the $zw(\alpha_0 + \alpha_1 x)$ allows both intercept and slope to be different for each sex at altitude compared to not at altitude
  - there are 4 possible scenarios (each of which requires 2 parameters) so you will need 4×2 parameters

- when taking CI, remember to add them up the standard error correctly (i.e $se_{\text{male}}(\text{slope}) = \sqrt{se(\text{slope})^2 + se(\text{slope:male})^2}$)
- note: parameters refers to the number of β_j in the model

Lecture 11: More on Categorical Variables

- this lecture did stuff where y was continuous (prices of food) and x was categorical (different cities)
- to encode dummy variables in the linear model we can set

$$x_{1i} = \begin{cases} 1 & \text{if restaurant } i \text{ was in London} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{if restaurant } i \text{ was in NY} \\ 0 & \text{otherwise} \end{cases}$$

  - Boston would be the baseline
  - full model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
  - can also split it into 3 models

$$Y = \begin{cases} \beta_0 + \varepsilon & \text{for Boston} \\ \beta_0 + \beta_1 + \varepsilon & \text{for London} \\ \beta_0 + \beta_2 + \varepsilon & \text{for NY} \end{cases}$$

- we can also test the hypothesis the the meals from NY and LDN were equal

$$H_0 : \beta_1 = \beta_2 \quad \text{vs} \quad H_1 : \beta_1 \neq \beta_2$$
$$\text{test statistic} = \frac{b_1 - b_2}{se(B_1 - B_2)} \sim t_{n-k}$$

- note: if given R output for model selection, focus mostly on R^2 scores
  - R will also give an output for the p-value of the individual parameter - the smaller the better (it means it's unlikely to be 0)

Lecture 12: Quadratic Model and Curve Fitting

- linear model, we means linear in parameters (the β_i are linear)
  - ex. $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$ is linear
  - ex. $Y = \beta_0 + \beta_1 + \beta_3(\beta_2)^x + \varepsilon$ is NOT linear
- quadratic model: we can do something like `lm(formula = Asset s ~ Accounts + I(Accounts^2))` and the equation would look like $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
- when to pick quadratic models: you can take a look at things like adjusted R^2 (not just R^2 - that always increases with more terms), look at if the coefficient for the extra term is significant or not
- if performance is around the same, you prefer the simpler model
- since x and x^2 comes from the same column, they are highly-correlated → colinearity is a concern
  - can make parameter estimates unstable (slight change in data have big impact on the estimates)
  - fix: adjust quadratic term by removing mean number of accounts (793.6) before squaring → `lm(formula = Assets ~Accounts + I((Accounts - 793.6)^2))`
- variance inflation factor (VIF): measures colinearity - the formula is $VIF_i = \frac{1}{1-R_i^2}$
  - R_i^2 is the R^2 when regressing the predictor i (one you're looking at) and regressing it against every other predictor in the model
  - in the case of x and x^2 - their R^2 is just the correlation
  - if $VIF > 10$ then we say colinearity is an issue with covariate x_j
- model fitting: we can also fit models that have higher power relationships (cubes, power of 4, etc) - we can decide which polynomial to choose by this algorithm
  1. set $q = 1$
  2. fit model (of order) q
  3. if 95% CI for β_q includes zero then stop, output model $q-1$ as the answer, if not then increment q by 1 and go back to step 2

Lecture 15: Model Selection

- what is the "best" linear model one can fit for the above variables?
- Mallow's C_p: for a model with design matrix $\mathbf{X}$ having p columns

$$C_p \coloneqq \frac{\text{Res SS}(p)}{\text{Res MS}(full)} - (n - 2p)$$

$$\text{Res SS}(p) = \text{residual sum square from model containing } p \text{ params}$$

$$\text{Res MS}(full) = \text{RMS from the largest model (all predictors included)} = \frac{\text{Res SS}(full)}{n - (\text{all possible predictors})}$$

  - we say that a model is acceptable if $\mathbf{C_p \approx p}$ - also smaller is better if they're both close to p (note: p includes β_0 but not ε)
  - should plot C_p vs p and the values that fall on the linear line are considered good
  - problem: the statistic is subject to sampling variation but its sampling distribution is unknown
- can use residual SS for the null model to compute R^2 and adjusted R^2 for any fitted model (null model is not the same as full model)

$$R^2 = 1 - \frac{\text{Res SS}}{\text{Res SS (Null)}} \qquad adjR^2 = 1 - \frac{\frac{\text{Res SS}}{(n-p)}}{\frac{\text{Res SS (Null)}}{(n-1)}}$$

  - for both, bigger is better
  - you can also find R by looking at `corr(y, y_hat)`
  - look above for another $adjR^2$ formula
- you can also use residual standard error (smaller is better) as well as look at if the new terms are significant (small p-val)
- other model selection methods: forward/backwards selection, or try out every combination and use a train/test set

Chapter 16: Leverage and Influence

- recall that in matrix form, we have

$$\hat{y} = X(X^TX)^{-1}X^Ty = Xb = Py$$
$$b = (X^TX)^{-1}X^Ty$$
$$P = X(X^TX)^{-1}X^T = \text{ the hat matrix}$$

  - (note: the hat matrix is symmetric and idempotent)
  - so if given the hat matrix, you can do $\hat{y}_i = (P \cdot y)_i$
- influential point is a point that has a large impact on the regression - slight difference from being an outlier, a point can be influential without being an outlier
  - basically like asking "how much do the fitted values change if we omit an observation"
- diagonal entries of the hat matrix, P_{ii} gives an indication of potential influence of a point i → diagonal entries are called leverages
  - we say observations with $P_{ii} > 2\frac{k}{n}$ to have high leverage (k here is the number of β - include intercepts)
- Cook's distance: actually measure the influence of a point

$$D_i = \frac{e_i'^2}{k}\left(\frac{P_{ii}}{1-P_{ii}}\right)$$
$$e_i' = \text{the residual between the fitted value including the point vs. fitted value when not including the point}$$

  - we say that a point i is influential if $D_i > 1$
  - do note that this does not always have to be the case, if Cook's distance for 1 point is significantly higher than its peers, we can say that it's influential (online says criteria could also be $D_i > 4/n$)
- some items of note from the webwork
  - influential observation does not need to have extreme residual
  - number of influential observations can be 0,1,2,..., or n-p-1, where p is the number of explanatory variables
  - influential observations sometimes are in the extremes of the space of explanatory variables (to the very right or left of the x axis)
  - influential observations have heavy influence on some $\hat{\beta}$

Lecture 17: Transformations

- primary reason for transforming variables: improve the fit of a model and alleviate violations of assumptions
- example: there's a multiplication model $Y = \beta_0 x_1^{\beta_1} + x_2^{\beta_2}\varepsilon$, take the log and get $\log(y) = \log(\beta_0) + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \log(\varepsilon)$
- interpretation: for a fitted model that looks like $Y = 3.767 - 0.299\log(x)$ (only logged the x variable which is the GDP)
  - means if we increase the log of the GDP by 1 unit we decrease the crowdedness index (Y) by about 0.3 units
  - note: if $\log(GDP) = 9 \to GDP = \$8103.08$ (increasing $\log(GDP)$ by 1 unit is the same as increasing GDP by factor of e)
  - increasing GDP x by m% (say 10%), the change in y is (last line):

$$\hat{y} = 3.7 - 0.299\log(x) \qquad \hat{y}' = 3.7 - 0.299\log(1.1x)$$
$$\therefore\ \hat{y}' - \hat{y} = -0.299\log(1.1x/x) = -0.299\log(1.1)$$

- if a variance about a line fitted seems non-constant, square rooting response var might help
- note: taking transformation is easy but it makes resulting model less interpretable, in general:
  1. if making predictions/estimates for response vals, these should be in the original scale → interval estimates may not be symmetric
  2. parameter estimates are interpreted on the transformed scale, no way to turn them back (i.e β_1 above is made on the log scale) → hard to compare params across models if they're on different transformation scales

Lecture 19: Logistic Regression

- you use this when your response variable is binary
- we can denote the possibilities as 0 and 1 with $P(Y_i = 1) = \pi_i$ - for each observation, the probability of $\{Y_i = 1\}$ could depend on the covariates $x_i = (x_{i1}, \ldots, x_{ip})^T$
- we can also say the $E[Y_i] = \pi_i$
- logit transform:

$$\text{logit}(\pi_i) \coloneqq \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\boldsymbol{\beta}$$
$$\therefore \frac{\pi_i}{1-\pi_i} = e^{x_i^T\boldsymbol{\beta}}$$
$$\pi_i = \frac{e^{x_i^T\boldsymbol{B}}}{1+e^{x_i^T\boldsymbol{\beta}}} = \frac{1}{1+e^{-x_i^T\boldsymbol{\beta}}}$$

  so given the output of an R model, you can plug it into here and get the probability
- example: budworm - the set up for a question can be a little strange

	Dose (in mg)					
	1	2	4	8	16	32
Died	3	7	17	25	32	37

- so for dose = 1 mg, 3 worms died (successes) and 47 lived (failures)
- we could estimate each π_i with the sample proportion (i.e $\pi_{4\text{mg}} = 17/40$ - but this model fits the data exactly, not that interesting → called the **maximal (or saturated) model**

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.41188    0.22916  -6.161 7.22e-10 ***
Dose         0.16473    0.02502   6.585 4.55e-11 ***
```

- so we can predict the estimated death rates at dose = 16 mg

$$\pi_{16\text{mg}} = \frac{e^{-1.412+0.1647\times 16}}{1+e^{-1.412+0.1647\times 16}} = 0.773$$

- residual: for the i-th residual, you can take the difference of observed counts and those predicted by the model (i.e $40 \times \hat{\pi}_i$) or subtract sample propotion vs fitted probability ($\pi_i - \hat{\pi}_i$)
- note: probability (π_i) vs log odds $\left(\frac{\pi_i}{1-\pi_i}\right)$
  - log odds and dose **IS** linear; but the relationship between π_i and dose level **IS NOT** linear
- categorical variables in logistic regression: say above you separate the data further into male and female worms

	Dose (in mg)					
Sex	1	2	4	8	16	32
M	2	5	10	14	18	20
F	1	2	7	11	14	17

  - then fitting the model without interactions we get `intercept = -1.889; sexM = 0.8480; dose - 0.1705`
  - ex: use this model to estimate **odds** (sometimes also called odds ratio) of male budworm mortality at dose level of 2 mg

$$\begin{aligned}\hat{O}_{2,m} &= \frac{\hat{\pi}_{2,m}}{1-\hat{\pi}_{2,m}} \\ &= e^{x_{2,m}^T \boldsymbol{\beta}} \\ &= e^{-1.889+0.848+0.1705(2)} = 0.497\end{aligned}$$

  - you can compare the odds ratio between categories (i.e ratio of male vs female at 2 mg or $\hat{O}_{2,m}/\hat{O}_{2,f}$) → ratio of 1 would suggest mortality rate at that level is independent of sex
  - note that we can do something very similar to above if we decided to model interactions between does and sex

Lecture 21: Model Selection for LR

- idea: compare the likelihood of each model for performance ($\vec{\pi} =< \pi_1, \ldots, \pi_n >$ and $\vec{y} =< y_1, \ldots, y_n >$)

$$L = \prod_{i=1}^{n} \binom{n}{y_i} \pi_i^{y_i} (1-\pi_i)^{n-y_i}$$

$$\begin{aligned} l(\vec{\pi}, \vec{y}) &= \sum_{i=1}^{n} \log\left(\pi_i^{y_i} \cdot (1-\pi_i)^{n_i - y_i} + \log(n_i C r_i)\right) \\ &= \sum_{i=1}^{n} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i}\end{aligned}$$

  - note: L is the likelihood while l is the log-likelihood
  - can ignore the combination term (must ignore for all models then)
  - generally higher is better but think about extra params trade-offs (AIC kinda looks at this trade off for you)
  - in our case, p for the maximal model is 6 (every x has its own π), but the fitted one (without sex) only has 2 (β_0 and β_1)
  - note: if we were to take version with sex, the maximal model has 12 parameters, 3 for the fitted one
- Akaike Information Criterion (AIC): lower AIC is better

$$\begin{aligned} AIC &= -2 \cdot (l(\hat{\pi}; \mathbf{y}) - p) \\ l(\hat{\pi}; \mathbf{y}) &= \text{ log-likelihood of the data} \\ p &= \text{ number of parameters in the model}\end{aligned}$$

- Bayesian Information Criterion (BIC): alternative to AIC, punishes number of parameters more

$$BIC = -2l(\hat{\pi}; \mathbf{y}) + p \log n$$

  - again, lower is better
  - preferable when explanation, rather than prediction, is the aim

Lecture 22: Poisson Regression

- recall the Poisson probability model

$$P(Y=y) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad E[Y] = Var(Y) = \lambda$$

- motivating example: say we have the number of red cards per month

Month:	Aug	Sep	Oct	Nov	Dec
No. matchdays	3	4	3	3	7
No. red cards:	8	3	5	3	10
Month:	Jan	Feb	Mar	Apr	May
No. matchdays	4	4	3	5	2
No. red cards:	7	2	2	4	3

  we can model this as a Poisson because the variable is a count
- null model: we model the # of red card per month, Y, as a Poisson variable with some mean λ
  - assumption: parameter does not depend on the number of matchdays in a month or where the month falls in the season
  - the natural estimator of λ under this model is the mean number of red cards per month, 4.7
  - under the null, the log-likelihood is

$$l(\lambda, \mathbf{y}) = \log\left(\prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}\right)$$

  - our log-likelihood here is -23.226
- maximal (saturated) model: fits a separate Poisson variable for each observation
  - this model might be more reasonable because since number of matches varies by month, more matches might mean more red cards; also number of red cards might differ on time of the season
  - here, we fit $\hat{\lambda}_i = y_i$ (the number of red cards for that month, as seen in the data) for $i = 1, \ldots 10$

$$l(\lambda, \mathbf{y}) = \log\left(\prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}\right)$$

  - we get a log-likelihood of -16.426, bigger than the null
- better model?: we could try and allow the monthly red card rate to depend on the number of matchdays in the month

$$\log(\lambda_i) = x_i^T \beta$$

  In R, we get smt like: `intercept = 0.7390; MatchDays = 0.2023`
  - use the model fitted to estimate the probability of no red cards in a month with three matchdays

$$\hat{\lambda}_i = e^{x_i^T \beta} = e^{0.739+0.2023 x_i} = e^{0.739+0.2023(3)} = 3.8416$$

$$P(Y=0) = \frac{(3.8416)^0 e^{-3.8416}}{0!} = 0.02146$$

- residuals: the definition for the i-th standardised residual is

$$e_i' = \frac{\text{residual}}{\sqrt{s^2}} = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

Lecture 23: Model Selection for Poisson Regression

- (residual) deviance of a fitted model $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_{10})$

$$\begin{aligned} \text{deviance} &= 2\left[l(\lambda^{(s)}; y) - l(\hat{\lambda}; y)\right] \\ l(\lambda^{(s)}; y) &= \text{log likelihood of saturated model} \\ l(\hat{\lambda}; y) &= \text{log-likelihood of null (or proposed) model}\end{aligned}$$

  - we prefer the deviance to the perfect model to be small
  - note: this method can be applied to Logistic Regression as well
- AIC: similar to Logistic Regression, we can also use AIC

$$\begin{aligned} AIC &= -2\left(l(\hat{\lambda}; y) - p\right) \\ l(\hat{\lambda}; y) &= \text{log likelihood of model with } \hat{\lambda} \\ p &= \text{ number of parameter}\end{aligned}$$

  - again, lower is better → pay attention to p for the different models

Random Stuff

- observational vs experimental: observational is where you observe certain variables and try to determine if there is any correlation; experimental is where you control certain variables and try to determine if there is any causality.
- confounding variable: variable that might affect both response and explanatory variables (i.e health records when monitoring death vs iron tablets usage)