

# Review of Weeks 1-4

1 / 20

## General Linear Model

- ▶ Model definition and assumptions

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i, \quad i = 1, \dots, n$$

or in matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- ▶ The errors are independent, normally distributed, with mean zero and constant variance,  $\sigma^2$ .
- ▶ Interpretation of coefficients for numerical predictors

$$\beta_j = \text{effect on mean response of one unit change in } X_j \text{ with all other predictors fixed}$$

2 / 20

## General Linear Model

- ▶ Least squares estimates,  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  minimize

$$Q = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})]^2$$

- ▶ Prediction equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

- ▶ t-tests and confidence intervals

$$t = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \quad \hat{\beta}_j \pm t(\alpha/2, \text{error df}) \times \text{se}(\hat{\beta}_j)$$

3 / 20

## General Linear Model

- ▶ Interaction model (with two numerical predictors)

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \\ &= (\beta_0 + \beta_1 X_1) + (\beta_2 + \beta_3 X_1) X_2 \end{aligned}$$

- ▶ Polynomial regression:

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$$

- ▶ ANOVA table: overall and sequential

$$\begin{aligned} \text{Total SS} &= \text{Model SS} + \text{Error SS} \\ &= \text{SS}(X_1) + \text{SS}(X_2|X_1) + \cdots + \text{SS}(X_k|X_1, \dots, X_{k-1}) \\ &\quad + \text{SSE} \end{aligned}$$

4 / 20

## General Linear Model

- ▶ F-test for comparison of nested models

$$F = \frac{[SSE(\text{reduced}) - SSE(\text{complete})] / (k - p)}{MSE(\text{complete})}$$

- ▶ Coefficient of determination,  $R^2$ , and Multiple correlation,  $R$

$$R^2 = \frac{\text{Model SS}}{\text{Total SS}} = 1 - \frac{\text{Error SS}}{\text{Total SS}}$$

$R = \text{cor}(\mathbf{Y}, \hat{\mathbf{Y}}) =$  correlation between observed and fitted values.

- ▶ Categorical predictors: Dummy/indicator variables
- ▶ Models with numerical and categorical predictors. e.g. Fungal growth. Interaction model allows different slopes (and intercepts).

5 / 20

## Generalized Linear Models

- ▶ Link function relates mean to linear predictor
- ▶ Logistic regression for binary/binomial outcomes (logit link)

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k = \eta$$

- ▶ Model for the odds of success is

$$\begin{aligned} \frac{p}{1-p} &= \exp \{ \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k \} \\ &= (e^{\beta_0}) (e^{\beta_1})^{X_1} \dots (e^{\beta_k})^{X_k} \end{aligned}$$

$e^{\beta_1}$  is the multiplicative effect on the odds of a one unit change in  $X_1$  with the other variables fixed

6 / 20

## Generalized Linear Models

- ▶ Success probability is

$$p = \frac{e^\eta}{1 + e^\eta}$$

- ▶ Wald test and confidence intervals:

$$Z = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \quad \hat{\beta}_j \pm z \times \text{se}(\hat{\beta}_j) \quad e^{\hat{\beta}_j \pm z \times \text{se}(\hat{\beta}_j)}$$

- ▶ Likelihood ratio test: comparison of nested models using the deviance.
- ▶ Compare  $D(\text{reduced}) - D(\text{complete})$  to chisquared reference distribution with df equal to the difference in the number of parameters

7 / 20

## Generalized Linear Models

- ▶ Using the model deviance to test lack-of-fit when there are replicate responses (i.e. binomial data). Rule of thumb:  $np$  and  $n(1-p) \geq 5$
- ▶ Compare deviance (or Pearson statistic) to chisquared reference distribution with df = error or residual df
- ▶ Pearson residuals and Pearson statistic

$$r_i = \frac{Y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}} \quad \chi^2 = \sum_{i=1}^n r_i^2$$

- ▶ Examples: Space shuttle, beetle mortality, attitudes towards women

8 / 20

- ▶ Loglinear model for counts:  $Y \sim \text{Poisson}(\mu)$

$$\log \mu = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k = \eta$$

- ▶ Multiplicative interpretation of coefficients:

$$\begin{aligned}\mu &= \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k) \\ &= (e^{\beta_0})(e^{\beta_1})^{X_1} \dots (e^{\beta_k})^{X_k}\end{aligned}$$

- ▶ Deviance, lack-of-fit, overdispersion, Pearson residuals, Pearson statistic
- ▶ Examples: Textile faults, fish species, car insurance claims, Caesarean births (offset)

- ▶  $R^2$  proportion of total SS explained by the model - does not penalize complexity.  $\text{Max } R^2 \equiv \text{Min SSE}$
- ▶ Adjusted  $R^2$ .  $\text{Max Adjusted } R^2 \equiv \text{Min MSE}$
- ▶ Mallows's  $C_p$ . Mean squared error of fitted values.
- ▶ PRESS. Prediction error sum of squares.
- ▶ ML. No penalty for complexity.
- ▶ AIC. ML + penalty term based on number of parameters.

## Solutions to sample problems

## Problem 1

- ▶ The estimates below are from an analysis of the growth of two fungal isolates. Growth area (in  $\text{cm}^2$ ) was measured each day for a period of 5 days. Two independent replicates were observed for each isolate. Hence, there are a total of 20 observations. The model assumes linear growth for each isolate.

	Estimate	Std. Error
(Intercept)	-10.9	1.02
day	11.9	1.09
isolate2	-1.3	0.65
day:isolate2	2.6	0.58

- ▶ *Comment: The isolate factor is coded as 0 for isolate1 and 1 for isolate2*

- ▶ Let  $Y$  denote the fungal area. Determine the prediction equations for isolates 1 and 2.

$$\text{isolate 1: } \hat{Y} = -10.9 + 11.9\text{Day}$$

$$\begin{aligned} \text{isolate 2: } \hat{Y} &= (-10.9 - 1.3) + (11.9 + 2.6)\text{Day} \\ &= 12.2 + 14.5\text{Day} \end{aligned}$$

- ▶ Fill in the missing values in the ANOVA table below.

	Df	SumSq	MeanSq	F value
time	1	1200	1200	120
isolate	1	40	40	4
time:isolate	1	200	200	20
Residuals	16	160	10	

13 / 20

- ▶ What proportion of the total variation in  $Y$  is explained by the model?

$$R^2 = \frac{\text{Model SS}}{\text{Total SS}} = \frac{1440}{1600} = 0.9$$

- ▶ How would you test the hypothesis of no difference between the slopes for the two isolates? Write down an appropriate test-statistic value, and reference distribution.

From the ANOVA table:  $F = 20$

Compare to F-distribution with  $df_1 = 1$ ,  $df_2 = 16$

14 / 20

## Problem 2

- ▶ This question concerns a study in which male and female budworms were exposed to doses of a chemical toxin for 3 days. The numbers that died at each dose level were recorded. The design was balanced in the sense that 20 male and female budworms were exposed at each dose level.
- ▶ The experimenter initially fit a logistic regression model for the probability of being killed of the form

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \text{LDose} + \beta_2 \text{I(Male)} + \beta_3 \text{LDose} \times \text{I(Male)}$$

where LDose denotes  $\log_2(\text{dose})$  and  $\text{I(Male)}=1$  if the budworm sample was male and 0 if it was female.

15 / 20

- ▶ The Analysis of Deviance table for the model is

### Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	P(> Chi )
NULL			11	124.876	
ldose	1	107.892	10	16.984	0.000
sex	1	10.227	9	6.757	0.001
ldose:sex	1	1.763	8	4.994	0.184

- ▶ Is there a significant dose by sex interaction? (Yes or No). Write down or determine the relevant test statistic for testing this hypothesis based on the ANOVA table.

No.  $D(\text{reduced})-D(\text{complete}) = 1.763$ ,  $P = 0.184$

16 / 20

- ▶ The model fit without interaction is as follows:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-3.4732	0.4685	-7.413	1.23e-13	***
ldose	1.0642	0.1311	8.119	4.70e-16	***
sexM	1.1007	0.3558	3.093	0.00198	**

---

Null deviance: 124.876 on 11 degrees of freedom  
Residual deviance: 6.757 on 9 degrees of freedom

- ▶ Write down the fitted logistic model without interaction

$$\log \frac{\hat{p}}{1 - \hat{p}} = -3.47 + 1.06 \text{LDose} + 1.10 \text{I(Male)}$$

17 / 20

- ▶ Is there evidence of lack of fit for this model? (Yes or No). Write down the relevant test statistic for assessing lack of fit.

No. Residual Deviance = 6.759, less than df=9

- ▶ Compute an approximate 95% confidence interval for the multiplicative effect on the odds of being killed of a one unit increase in LDose.

$$e^{\hat{\beta} \pm 2 \text{se}} = e^{1.06 \pm 2(0.13)} = (e^{0.80}, e^{1.32}) = (2.23, 3.74)$$

18 / 20

## Problem 3

- ▶ In a dilution assay experiment the number of infected organisms is recorded at various dilution levels. The goal is to estimate the density of organisms in the original solution. The experimenter fit a Poisson regression model to the counts, using a log link, and with log concentration as the predictor. A summary of the output is as follows:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	2.83638	0.10833	26.182	0.0000	***
logconc	-0.61572	0.07136	-8.628	0.0000	***

---

Null deviance: 106.075 on 20 degrees of freedom  
Residual deviance: 13.450 on 19 degrees of freedom

19 / 20

- ▶ Let  $\mu_i$  denote the expected count at log concentration level  $i$ , where  $i = 0, 1, 2, 3$  and 4. Write down the equation of the fitted model.

$$\log \hat{\mu}_i = 2.84 - 0.62 \log \text{conc}$$

- ▶ Calculate an approximate 95% confidence interval for the expected count in the original solution; i.e. when  $\log \text{conc} = 0$ .

$$\hat{\mu}_0 = e^{\hat{\beta}_0} : e^{\hat{\beta}_0 \pm 2 \text{se}} = e^{2.84 \pm 2(0.11)} = (13.7, 21.3)$$

- ▶ Based on the model fit, would a count of 5 when  $\log \text{conc} = 4$  be surprising? (Hint: Determine the Pearson residual for this value.)

$$\hat{\mu} = e^{2.84 - 0.62(4)} = 1.43, \quad \frac{Y - \hat{\mu}}{\sqrt{\hat{\mu}}} = \frac{5 - 1.43}{1.43} = 2.99$$

Yes. A count of 5 would be three sd's larger than expected.

20 / 20