Exam 3 review. 6 sheets of notes and a calculator. Monday, Nov. 30. This exam will likely be online.

Types of problems.

From Exam 2: LS CLT 61)

76) Use $\boldsymbol{Z}_n \sim AN_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let a be a constant, let $\boldsymbol{A}$ be a $k \times r$ constant matrix, and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_r\boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_r$,

$$a\boldsymbol{Z}_n \sim AN_r\left(a\boldsymbol{\mu}_n, a^2\boldsymbol{\Sigma}_n\right), \text{ and } \boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\mu}_n + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_n\boldsymbol{A}^T\right),$$

$$\hat{\boldsymbol{\theta}}_n \sim AN_r\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}}{n}\right), \text{ and } \boldsymbol{A}\hat{\boldsymbol{\theta}}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{c}, \frac{\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T}{n}\right).$$

ch. 7

77) Problems with the OLS full model: i) If $n = p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$ regardless of how bad the predictors are. ii) If $n < p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$ or the program fails. iii) Need $n > Jp$ where $J \geq 5$, and preferably $J \geq 10$ for good estimation. If $n < 5p$, the OLS full model overfits.

78)

Label	coef	SE	shorth 95% CI for β_i
Constant=intercept= x_1	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$[\hat{L}_1, \hat{U}_1]$
x_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$[\hat{L}_2, \hat{U}_2]$
$\vdots$			
x_p	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	$[\hat{L}_p, \hat{U}_p]$

The classical OLS large sample 95% CI for β_i is $\hat{\beta}_i \pm 1.96SE(\hat{\beta}_i)$. Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If $0 \in$ CI for β_i, then fail to reject H_0, and conclude x_i is not needed in the MLR model given the other predictors are in the model. If $0 \notin$ CI for β_i, then reject H_0, and conclude x_i is needed in the MLR model.

79) A *model for variable selection* is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). If $S \subseteq I$, then $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. Note that $\boldsymbol{\beta}_E = \boldsymbol{0}$. Let $k_S = a_S - 1 =$ the number of population active nontrivial predictors. Then $k = a - 1$ is the number of active predictors in the candidate submodel I.

80)

I_j	model	x_2	x_3	x_4	x_5	$\hat{\boldsymbol{\beta}}_{I_j,0}$ if $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_j}$
I_2	1		*			$(\hat{\beta}_1, 0, \hat{\beta}_3, 0, 0)^T$
I_3	2		*	*		$(\hat{\beta}_1, 0, \hat{\beta}_3, \hat{\beta}_4, 0)^T$
I_4	3	*	*	*		$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, 0)^T$
I_5	4	*	*	*	*	$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_4)^T = \hat{\boldsymbol{\beta}}_{OLS}$

Model I_{min} is the model, among p candidates, that minimizes C_p if $n \geq 10$, or EBIC if $n < 10p$. Model I_j contains j predictors, $x_1^*, x_2^*, ..., x_j^*$ where $x_1^* = x_1 \equiv 1$, the constant.

81) Variable selection is a search for a subset of predictors that can be deleted without important loss of information if $n \geq 10p$ and such that model I (containing the remaining predictors that were not deleted) is good for prediction if $n < 10p$. Note that the "100%" shorth CI for a β_i that is a component of $\boldsymbol{\beta}_O$ is [0,0].

82) Underfitting occurs if $S \not\subseteq I$ so that $\boldsymbol{x}_I$ is missing important predictors. Underfitting will occur if $\boldsymbol{x}_I$ is $k \times 1$ with $d = k < a_S$. Overfitting occurs if $S \subset I$ with $S \neq I$ or if $n < 5k$.

83) In 80), sometimes TRUE = * and FALSE = blank. The x_i may be replaced by the variable name or letters like a b c d.

I_j	model	x_2	x_3	x_4	x_5
I_2	1	FALSE	TRUE	FALSE	FALSE
I_3	2	FALSE	TRUE	TRUE	FALSE
I_4	3	TRUE	TRUE	TRUE	FALSE
I_5	4	TRUE	TRUE	TRUE	TRUE

84) The out\$cp line gives $C_p(I_2), C_p(I_3), ..., C_p(I_p) = p$ and I_{min} is the I_j with the smallest C_p.

85) Typical bootstrap output for forward selection, lasso, and elastic net is shown below. The SE column is usually omitted except possibly for forward selection. The term "coef" might be replaced by "Estimate." This column gives $\hat{\boldsymbol{\beta}}_{I,0}$ where $I = I_{min}$ for forward selection, $I = L$ for lasso, and $I = EN$ for elastic net. Note that the SE entry is omitted if $\hat{\beta}_i = 0$ so variable x_i was omitted by the variable selection method. In the output below, $\hat{\beta}_2 = \hat{\beta}_3 = 0$. The SE column corresponds to the OLS SE obtained by acting as if the OLS full model contains a constant and the variables not omitted by the variable selection method. The OLS SE is incorrect unless the variables were selected before looking at the data for forward selection.

Label	Estimate or coef	SE	shorth 95% CI for β_i
Constant=intercept= x_1	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$[\hat{L}_1, \hat{U}_1]$
x_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$[\hat{L}_2, \hat{U}_2]$
x_3	0		$[\hat{L}_3, \hat{U}_3]$
x_4	0		$[\hat{L}_4, \hat{U}_4]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	$[\hat{L}_p, \hat{U}_p]$

86) The OLS SE is also accurate for forward selection with C_p if $\boldsymbol{X}^T\boldsymbol{X}/n \to \boldsymbol{V}^{-1} = diag(d_1, ..., d_p)$ where all $d_i > 0$. The diagonal limit matrix will occur if the predictors are orthogonal or if the nontrivial predictors are independent with 0 mean and finite variance.

```
regbootsim3(nruns=500)
$cicov
0.942 0.954 0.950 0.948 0.944 0.946 0.946 0.940 0.938 0.940
$avelen
0.398 0.399 0.397 0.399 2.448 2.448 2.448 2.448 2.448 2.450
$beta
[1] 1 1 0 0
$k
[1] 1
```

87) Simulation output for regression is similar to that shown above. Usually want coverage near 0.95 since nominal 95% CIs are used and tests with nominal $\delta = 0.05$ are used. To suggest that the actual coverage is near the nominal coverage of 0.95, want cov in [0.94,0.96] with 5000 runs, want cov in [0.93,0.97], with 1000 runs, want cov in [0.92,0.98] with 500 runs, and want cov in [0.91,0.99] with 100 runs. Let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1}$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, ..., 1, 0, ..., 0)^T$ with β_1, k ones, and $p - k - 1$ zeros. Then $S = \{1, ..., k + 1\}$ and $E = \{k + 2, ..., p\}$. Note that S corresponds to the first $k + 1$ β_i while E corresponds to the last $p - k + 1$ β_i.

The first 4 numbers are the bootstrap shorth confidence intervals for $\beta_1, \beta_2, \beta_{p-1}$, and β_p. The average lengths of the CIs along with the proportion of times (coverage) the CI for β_i contained β_i are given. The next three numbers test $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$. The prediction region method, hybrid method, and Bickel and Ren methods are used. Hence the fifth interval gives the length of the interval $[0, D_{(c)}]$ where H_0 is rejected if $D_0 > D_{(c)}$ and the fifth "coverage" is the proportion of times the prediction region method test fails to reject H_0. The last three numbers are similar but test $H_0 : \boldsymbol{\beta}_S = (\beta_1, 1, ..., 1)^T$. Hence the last length 2.450 corresponds to the Bickel and Ren method with coverage 0.940. For the output shown, lengths near 2.45 correspond to $\sqrt{\chi^2_2(0.95)}$ where $P(X \leq \chi^2_2(0.95)) = 0.95$ if $X \sim \chi^2_2$.

88) Three bootstrap methods for MLR $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ use $\boldsymbol{Y}^* = \boldsymbol{X}^*\hat{\boldsymbol{\beta}} + \boldsymbol{e}^*$ where the $\boldsymbol{e}^*_i$ are iid and zero mean with respect to the bootstrap distribution. Do the OLS regression of $\boldsymbol{Y}^*$ on $\boldsymbol{X}^*$ to get $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}^*$. Repeat to get the bootstrap sample $\hat{\boldsymbol{\beta}}^*_1, ..., \hat{\boldsymbol{\beta}}^*_B$. For testing $H_0 : \boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$, use the bootstrap sample $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_1, ..., \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_B$. Apply bootstrap confidence regions to the bootstrap sample.

a) *parametric bootstrap*: $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_n\boldsymbol{I}) \sim N_n(\boldsymbol{H}\boldsymbol{Y}, \hat{\sigma}^2_n\boldsymbol{I})$ where **we are not assuming** that the $e_i \sim N(0, \sigma^2)$, and

$$\hat{\sigma}^2_n = MSE = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2$$

where the residuals are from the full OLS model. Hence

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{e}^*$$

where the e^*_i are iid $N(0, MSE)$. Note that $\boldsymbol{X}^* = \boldsymbol{X}$.

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$ if $S \subseteq I$.

b) *residual bootstrap*: Let $\boldsymbol{e}^* = \boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the OLS full model residuals. Then

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W.$$

Hence

$$E(r_i^W) = \frac{1}{n}\sum_{i=1}^{n} r_i = 0, \quad V(r_i^W) = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{n-p}{n} MSE,$$

$$E(\boldsymbol{r}^W) = \boldsymbol{0}, \text{ and } \quad \mathrm{Cov}(\boldsymbol{Y}^*) = \mathrm{Cov}(\boldsymbol{r}^{\mathrm{W}}) = \sigma_{\mathrm{n}}^2 \boldsymbol{I}_{\mathrm{n}}.$$

Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma_n^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and $E(\hat{\boldsymbol{\beta}}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. Under regularity conditions for the OLS MLR model, $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{W}) \sim N_p(\boldsymbol{0}, \boldsymbol{V})$. We conjecture that if $S \subseteq I_j$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$.

c) *nonparametric bootstrap*: Draw a sample on size n cases with replacement from the data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$ to form $(\boldsymbol{Y}^*, \boldsymbol{X}^*)$. For the full model,

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

where $\boldsymbol{r}^W$ has the same bootstrap distribution as in b), and for a submodel I,

$$\boldsymbol{Y}^* = \boldsymbol{X}_I^*\hat{\boldsymbol{\beta}}_{I,OLS} + \boldsymbol{r}_I^W.$$

Under regularity conditions for the OLS MLR model, $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{W}) \sim N_p(\boldsymbol{0}, \boldsymbol{V})$. Hence if $S \subseteq I_j$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$.

89) Let $\boldsymbol{x}_i^T = (1 \;\; \boldsymbol{u}_i^T)$. It is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W} = (W_{ij})$. For $j = 1, ..., p-1$, let W_{ij} denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Then the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n}.$$

Then regression through the origin is used for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. Thus the centered response $Z_i = Y_i - \overline{Y}$ and $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Then $\hat{\boldsymbol{\eta}}$ does not depend on the units of measurement of the predictors. Linear combinations of the $\boldsymbol{u}_i$ can be written as linear combinations of the $\boldsymbol{x}_i$, hence $\hat{\boldsymbol{\beta}}$ can be found from $\hat{\boldsymbol{\eta}}$.

90) Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j \tag{1}$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression $\hat{\boldsymbol{\eta}}_R$, $j = 1$ corresponds to lasso $\hat{\boldsymbol{\eta}}_L$, and $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS_W(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Note that for a $k \times 1$ vector $\boldsymbol{\eta}$, the squared (Euclidean) L_2 norm $\|\boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \sum_{i=1}^k \eta_i^2$ and the L_1 norm $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^k |\eta_i|$.

Lasso and ridge regression have a parameter λ. When $\lambda = 0$, the OLS full model is used. Otherwise, the centered response and scaled nontrivial predictors are used with $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. These methods also use a maximum value λ_M of λ and a grid of M λ values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{M-1} < \lambda_M$ where often $\lambda_1 = 0$. For lasso, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ for $i < M$.

91) The elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{2}$$

where $\lambda_1 = (1 - \alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ with $0 \leq \alpha \leq 1$.

92) Use $\boldsymbol{Z}_n \sim AN_g(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_g(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let a be a constant, let $\boldsymbol{A}$ be a $k \times g$ constant matrix, and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_g\boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_g$,

$$a\boldsymbol{Z}_n \sim AN_g\left(a\boldsymbol{\mu}_n, a^2\boldsymbol{\Sigma}_n\right), \text{ and } \boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\mu}_n + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_n\boldsymbol{A}^T\right),$$

$$\hat{\boldsymbol{\theta}}_n \sim AN_g\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}}{n}\right), \text{ and } \boldsymbol{A}\hat{\boldsymbol{\theta}}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{c}, \frac{\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T}{n}\right).$$

93) Assume $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Let $\boldsymbol{s}_n = (s_{1n}, ..., s_{p-1,n})^T$ where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_i)$ if $\hat{\eta}_i \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 1$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 1$. Then

i) $\hat{\boldsymbol{\eta}}_R = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1n}}{n}n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$.

ii) $\hat{\boldsymbol{\eta}}_L = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\ n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\ \boldsymbol{s}_n$.

iii) $\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\ \left[\frac{\lambda_1}{n}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n\right]$.

94) Assume that the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}$. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,...,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Let $\hat{\boldsymbol{\eta}}_A$ be $\hat{\boldsymbol{\eta}}_{EN}$, $\hat{\boldsymbol{\eta}}_L$, or $\hat{\boldsymbol{\eta}}_R$. Let p be fixed.

i) OLS CLT: $\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

ii) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

iii) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$, *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s\eta}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

iv) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{R} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}).$$

v) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s\eta}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{L} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

ii) and v) are the Lasso CLT, ii) and iv) are the RR CLT, and ii) and iii) are the EN CLT.

95) If $[r_L, r_U]$ is a $100(1-\delta)\%$ PI for a future residual r_f and $\hat{Y}_f = \boldsymbol{x}_f\hat{\boldsymbol{\beta}}_I$, then a $100(1-\delta)\%$ PI for Y_f is $[\hat{Y}_f + r_L, \hat{Y}_f + r_U]$.

ch. 9

96) **Regression** is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $(p-1) \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. In a **1D regression model**, the response Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$.

97) Any regression model that depends on the predictors only through $\boldsymbol{\beta}^T\boldsymbol{x}$ is a 1D model. MLR, LR, PR, single index models $Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e$ and transformation models $Y = t^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e)$ are very important special cases. Given a list of models, be able to tell which are 1D models.

98) If the 1D regression model holds, then $Y \perp\!\!\!\perp \boldsymbol{x}|a + c\boldsymbol{\beta}^T\boldsymbol{x}$ for any constants a and $c \neq 0$. The quantity $a + c\boldsymbol{\beta}^T\boldsymbol{x}$ is called a *sufficient predictor* (**SP**), and a sufficient summary plot is a plot of any SP versus Y. An *estimated sufficient predictor* (**ESP**) is $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T\boldsymbol{x}$ where $\tilde{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant c. An *estimated sufficient summary plot* (ESSP) or **response plot** is a plot of any ESP versus Y. The response plot is a goodness of fit plot and is used to visualize the conditional distribution $Y|SP$, and **this plot should be made for any 1D regression analysis**.

99) Given a response plot, be able to find $\hat{y}$ for a given value of the ESP with "up and over lines." Also be able to add an estimate of $E(y|ESP)$ to the response plot. See Q10, HW11.

100) Know that a response plot suggests that the MLR model is appropriate if the plotted points fall about some line. If the plotted points follow some curve or if outliers are present, then the MLR model is not appropriate. (In particular, if the `tvreg` estimator is used to make the trimmed views, the MLR model is appropriate if the plotted points follow the identity line, but if the plotted points follow a smooth nonlinear function with small variance function, then an additive error single index model (that is not MLR) may be appropriate.)

101) The OLS view or the "best trimmed view" can be used as the response plot. Given several views, be able to pick out the best trimmed view. See HW11. The `tvreg` estimator is a trimmed views estimator.

102) The OLS view works best if there are no strong nonlinearities among the predictors, eg if the predictor distribution is MVN (or EC). To check this condition make a scatterplot matrix of the predictors and a DD plot. In the scatterplot matrix, want the marginal plots to be spherical, box shaped, ellipsoidal or like a line. Outliers or strong curvature is bad. Add scatterplot smoothers such as *lowess* and *slicesmooth* to the response plot for a nonparametric estimate of $E(Y|ESP)$. The nonparametric estimates can be used to suggest simple functions (such as $\log(Y)$) or to check whether the parametric estimator is reasonable.

103) Let the errors e_i be iid, and assume that the regression model $Y_i = g(\boldsymbol{x}_i, \boldsymbol{\eta}, e_i)$ has a unique solution for e_i :

$$e_i = h(\boldsymbol{x}_i, \boldsymbol{\eta}, Y_i).$$

Then the ith residual

$$\hat{e}_i = h(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}}, Y_i)$$

where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$. Given a regression model, be able to find the ith residual.

ch. 10 Let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_s^T)^T$ and $\boldsymbol{x} = (1\boldsymbol{u}^T)^T$.

104) The binary **logistic regression (LR)** model states that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{binomial}(1, \rho(\boldsymbol{x}_\text{i}))$. where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \text{P}(\text{Y} = 1|\boldsymbol{x}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}.$$

Notice that $P(Y = 1|\boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x}) = 1 - P(\text{failure}|\boldsymbol{x})$.

105) Given values of $\boldsymbol{x} = (x_1 = 1, x_2, ..., x_p)^T$ for a LR model, estimate $\rho(\boldsymbol{x})$ with the estimator

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x})}{1 + \exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x})}.$$

106) The **Poisson regression (PR)** model states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i}))$$

where

$$\mu(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i).$$

107) Given values of $\boldsymbol{x} = (x_1 = 1, x_2, ..., x_p)^T$ for a PR model, estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}).$$

Response = Y
Terms = $(X_2, ..., X_p)$
Sequential Analysis of Deviance

Predictor	df	Total Deviance	df	Change Deviance
Ones	$n-1=df_o$	G^2_o		
X_2	$n-2$		1	
X_3	$n-3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
X_p	$n-p=df_{FULL}$	G^2_{FULL}	1	

```
----------------------------------------
Data set = cbrain, Name of Fit = B1
Response      = sex
Terms         = (cephalic size log[size])
Sequential Analysis of Deviance
                    Total              Change
Predictor      df   Deviance     |     df   Deviance
Ones          266   363.820      |
cephalic      265   363.605      |      1   0.214643
size          264   315.793      |      1   47.8121
log[size]     263   305.045      |      1   10.7484
```

108) To perform inference for LR and PR, computer output is needed. Above shows output using symbols and *Arc* output from a real data set with $k = p - 1 = 3$ nontrivial predictors. Know how to use this type of output for the deviance test described below. Assume that the response plot has been made and that the logistic or Poisson regression model fits the data well in that the nonparametric step or lowess estimated mean function follows the estimated model mean function closely. The deviance test is used to test whether $\boldsymbol{\beta} = \mathbf{0}$ or $\boldsymbol{\beta} \neq \mathbf{0}$. If $\boldsymbol{\beta} = \mathbf{0}$, then the predictors are not needed in the GLM model. If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then the estimator $\hat{\rho} = \overline{Y} = \sum_{i=1}^n Y_i/n$ for binary LR and $\hat{\mu} = \overline{Y}$ for PR should be used.

109) The 4 step **deviance test** is

i) $H_o : \boldsymbol{\beta}_s = \mathbf{0}$ $\quad H_A : \boldsymbol{\beta}_s \neq \mathbf{0}$

ii) test statistic $G^2(o|F) = G^2_o - G^2_{FULL}$

iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi^2_k$ has a chi–square distribution with k degrees of freedom. Note that $k = p - 1 = df_o - df_{FULL} = n - 1 - (n - p)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a (LR or PR) GLM relationship between Y and the predictors $X_2, ..., X_k$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a (LR or PR) GLM relationship between Y and the predictors $X_2, ..., X_k$.

Response = Y $\quad$ Terms = $(X_2, ..., X_p)$ $\quad$ (Full Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

Degrees of freedom: n - p = df_{FULL}
Deviance: $D = G^2_{FULL}$

Response = Y Terms = $(X_2, ..., X_r)$ (Reduced Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_r/se(\hat{\beta}_r)$	for Ho: $\beta_r = 0$

Degrees of freedom: n - r = df_{RED}
Deviance: $D = G^2_{RED}$

```
(Full Model) Response = Status, Terms = (Diagonal Bottom Top)
Label      Estimate        Std. Error     Est/SE    p-value
Constant   2360.49         5064.42         0.466     0.6411
Diagonal  -19.8874         37.2830        -0.533     0.5937
Bottom     23.6950         45.5271         0.520     0.6027
Top        19.6464         60.6512         0.324     0.7460

Degrees of freedom:          196
Deviance:                  0.009

(Reduced Model) Response = Status, Terms = (Diagonal)
Label      Estimate        Std. Error     Est/SE    p-value
Constant   989.545         219.032         4.518     0.0000
Diagonal  -7.04376         1.55940        -4.517     0.0000

Degrees of freedom:          198
Deviance:                 21.109
```

110) The output shown above, both in symbols and for a real data set, can be used to perform the change in deviance test. After obtaining an acceptable full model where

$$SP = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP_R = \beta_{R1} + \beta_{R2} x_{R2} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $p-r$ predictors that are in the full model but not the reduced model. For binary logistic regression, the reduced model is $Y_i|\boldsymbol{x}_{Ri} \sim$ independent Binomial$(1, \rho(\boldsymbol{x}_{Ri}))$ while for Poisson regression the reduced model is $Y_i|\boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the response plot looks good. Then we want to test H_o: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}.

111) The 4 step **change in deviance test** is

i) H_o: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$

iii) The p–value $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{p-r}$ has a chi–square distribution with $p-r$ degrees of freedom. Note that p is the number of predictors in the full model while r is the number of predictors in the reduced model. Also notice that $p - r = df_{RED} - df_{FULL} = n - r - (n - p)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

112) If the reduced model is good, then the **EE plot** of $ESP_R = \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line.

113) For LR, an estimated sufficient summary or **response plot** is a plot of the MLE $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus Y with the logistic curve of fitted proportions

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added to the plot along with a step function of observed proportions (eg from slicesmooth). Suppose that the ESP takes on many values (eg the LR model has a continuous predictor) and that $k + 1 << n$. Know that the LR model is good if the step function follows the logistic curve of fitted proportions fairly closely in the response plot. Also **know that you should check that the LR model is good with the response plot before doing inference with the LR model**.

114) For PR, the response plot is a plot of the MLE ESP vs Y with the with the exponential curve of estimated means $\hat{\mu}(ESP) = \exp(ESP)$ added to the plot along with a lowess curve. Suppose that the ESP takes on many values (eg the PR model has a continuous predictor) and that $p << n$. Know that the PR model is good if the lowess follows the exponential curve of estimated means closely in the response plot (except perhaps at the largest values of Y). Also **know that you should check that the PR model is good with the response plot before doing inference with the PR model**.

115) If the response plot suggests that the LR model is good and if the logistic curve of estimated proportions fits the step function of observed proportions better than any

horizontal line, then the deviance test should reject Ho $\boldsymbol{\beta} = \mathbf{0}$ and conclude that $\boldsymbol{\beta} \neq \mathbf{0}$. If a horizontal line fits the step function about as well as the logistic curve fits the step function, then this graphical diagnostic is inconclusive (does not necessarily imply that one should fail to reject Ho), but does suggest that either the LR relationship with the predictors is weak or that there is no LR relationship with the predictors.

116) Given a predictor x, sometimes x is not used by itself in the full LR model. Suppose that Y is binary. Then to decide what functions of x should be in the model, look at the conditional distribution of $x|Y = i$ for $i = 0, 1$. These rules are used if x is an indicator variable or if x is a continuous variable.

distribution of $x\|y = i$	functions of x to include in the full LR model
$x\|y = i$ is an indicator	x
$x\|y = i \sim N(\mu_i, \sigma^2)$	x
$x\|y = i \sim N(\mu_i, \sigma_i^2)$	x and x^2
$x\|y = i$ has a skewed distribution	x and $\log(x)$
$x\|y = i$ has support on (0,1)	$\log(x)$ and $\log(1 - x)$

117) If the response plot suggests that the PR model is good and if the exponential curve (estimated PR mean function) fits the lowess curve (nonparameric estimated mean function) better than any horizontal line, then the deviance test should reject Ho $\boldsymbol{\beta}_s = \mathbf{0}$ and conclude that $\boldsymbol{\beta}_s \neq \mathbf{0}$. If a horizontal line fits the lowess curve about as well as the exponential curve fits the lowess curve, then this graphical diagnostic is inconclusive (does not necessarily imply that one should fail to reject Ho), but does suggest that either the PR relationship with the predictors is weak or that there is no PR relationship with the predictors.

118) Know how to use the weighted forward response plot as a goodness of fit plot for PR (along with the response plot and OD plot) and the weighted residual plot as a lack of fit plot for PR.

119) To build LR and PR models make a scatterplot matrix of the predictors with the response Y on the top or bottom. For LR, also mark the plotted points by a 0 if $Y = 0$ and by + if $Y = 1$. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model.

120) Create a full model and assume that the response plot for the full model is good. Let the number of predictors including a constant be p. Suppose that the Y_i are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i$ = the number of 1's and $N_0 = n - N_1$ = the number of 0's. Rule of thumb: want $k + 1 \leq \min(N_1, N_0)/5$. For PR, want $p \leq n/5$. Rule of thumb: the full model is ok if $G^2 \leq n - p + 3\sqrt{n - p}$.

121) **Variable selection** for LR and PR is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. Find the model I_{min} with the smallest AIC. Let $\Delta(I) = AIC(I) - AIC(I_{min})$. Then find the model I_I with the fewest number of predictors such that $\Delta(I_I) \leq 2$. Then submodel

I_I is the initial submodel to examine. Also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$.

122) Know how to find good models from output. For binary LR let N_i be the number of $Y_i = i$ for $i = 0, 1$. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 12 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ predictors, including a constant. Do not use more predictors than submodel I_I, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Assume the response plot for the full model is good. Then the submodel I is good if
i) the response plot for the submodel looks like the response plot for the full model.
ii) corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) For binary LR want $r_I \leq \min(N_1, N_0)/10$. For PR, want $r_I \leq n/10$.
vi) The plotted points in the VV plot cluster tightly about the identity line.
vii) Want the deviance $G^2(I) \geq G^2(full)$ but close. ($G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance.)
viii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.
xi) Want $G^2(I) \leq n - r_I + 3\sqrt{n - r_I}$.
xii) The OD plot should look good.

123) Variable selection can be bootstrapped much like variable selection for multiple linear regression, using the parametric bootstrap for LR and PR. For PR, use $\boldsymbol{Y}^* = (Y_i^*)$ where $Y_i^* \sim Poisson(\exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})) \sim Poisson(\exp(ESP))$ for $i = 1, ..., n$. Then do a Poisson regression of $\boldsymbol{Y}^*$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}_1^*$. Repeat to get the bootstrap sample $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$.

124) Overdispersion occurs if the actual conditional variance function $V(Y|\boldsymbol{x}) > V_m(Y|\boldsymbol{x})$, the model conditional variance function. The *OD plot* is a plot of the estimated model variance $\hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%.

126) In a *generalized additive model* (GAM), $Y \perp\!\!\!\perp \boldsymbol{x}|AP$ where the *additive predictor* $AP = \alpha + \sum_{j=2}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated additive predictor* EAP $= \hat{\alpha} + \sum_{j=2}^{p} \hat{S}_j(\boldsymbol{x}_j)$. A *response plot* is a plot of EAP versus Y.

127) A GLM is a special case of a GAM with $\alpha = \beta_1$ and $S_j(x_j) = \beta_j x_j$. Replace SP by AP to get models and ESP by EAP to get plots. For the binomial GAM, $Y_1, ..., Y_n$ are independent with $Y|AP_i \sim$ binomial($\mathrm{m_i}, \rho(\mathrm{AP_i})$). For the *Poisson regression* GAM, $Y_1, ..., Y_n$ are independent random variables with $Y|AP \sim$ Poisson(exp(AP)).

128) Check a GLM by fitting the corresponding GAM and making the EE plot of EAP versus ESP. With a plot check that each $\hat{S}_j$ is linear. If not, add terms like $\log(x)$, x_j^2 or x_j^3 to the GLM.