1. You may use one sheet of notes and a calculator. If you do not have a calculator work out expressions as far as you can.
2. Distributions are at the end of the exam.
3. Most parts do not depend on earlier parts of the problem, so if you are stuck please move on to the next section. Partial credit will be given if an answer does depend on an earlier part that was incorrect and the later problem was worked correctly after accounting for the previous error.
4. You do not have to re-derive well known results unless asked to, but if you do use specific results from class or Theorems please state them where used in your explanations.
5. The amount of space is not always an indication of the expected length of an answer. In general brief answers to all questions are better than more detailed responses to half the problems, so please use your time wisely.
6. Partial credit will be given where appropriate, although no points will be given for simply restating the problem and points may be taken off for rambling inconsistent answers.
7. Please cross out any work that you do not want to be graded, so that there is one solution. If you need additional space, you may use the backs of pages, but please indicate that you have done so.
8. Grades may be curved upwards if the mean is lower than expected.
9. In signing your name below, you agree to abide by the Duke Community Standard and will not receive or give help from/to anyone else.

Printed Name: Aasha Reddy

Signature: Aasha Reddy

Score:

1. Circle True or False (do not write in an answer)

   (a) [ **True or False**]: In generalized linear models, the model that has the smallest deviance is optimal.

   not always the case

   (b) [ **True or False**]: In generalized linear models, we apply the link function to the response.

   (c) [ **True or False**]: A large $R^2$ value (say 0.95) is evidence that the model is adequate.

   (d) [ **True or False**]: We can still use OLS even if the variances of the observations are not constant.

   (e) [ **True or False**]: For logistic regression, the residual deviance has an approximate (asymptotic) Chi-squared distribution with $p$ degrees of freedom (here there are $p$ coefficients in the linear predictor).

   $n-p$

   (f) [ **True or False**]: Values with leverage close to 1, will almost always be declared outliers by the outlier test.

   (g) [ **True or False**]: The `PowerTransform` function may be used to transform the predictors and response to find the optimal transformations for normality of residuals and linearity of predictors in the regressions function.

   → only the response, not the predictors

   (h) [ **True or False**]: The absolute value/magnitude of the coefficients is a reliable measure of the importance of a variable.

   → need to look at other measures as well

   (i) [ **True or False**]: The default link function for logistic regression is the log, which is why the coeffients are called log odds-ratios.

   → it is the logit not the log

   (j) [ **True or False**]: For adding a binary variable to a regression model, we should apply a logit transform to the predictor so that the coefficients may be unconstrained.

   → apply logit to a binary response, not a binary predictor

(k) [ **True or False**]: If we have a binary predictor in a model, we should use binary regression.

↳ response

(l) [ **True or False**]: For logistic regression, to obtain a 95 % confidence interval for the log odds ratio, we compute the confidence interval using the asymptotic normal distribution for the estimated coefficients, and then apply the inverse logit transform to the lower and upper values of the confidence interval.

↳ coefficients $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \dots \log\left(\frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi}{1-\pi}}\right) = \frac{e^{\beta_0} e^{\beta_1 x_A}}{e^{\beta_0} e^{\beta_1 x_B}} = \log\left(e^{\beta_1(A-B)}\right)$ are already log odds ratios $\beta_1(A-B)$

(m) [ **True or False**] Transforming predictors will have no impact on Cook's distance in practice.

↳ transforming predictors changes the results

(n) [ **True or False**] By increasing our sample size we can reduce the prediction error of the model.

$Y^* = \hat{f}(x) = x_*^T \hat{\beta} - \hat{f}(x^*) + \varepsilon^*$

(o) [ **True or False**] A value for Cook's Distance greater than one implies that the case had a large leverage value.

$\frac{isr^2}{p} \cdot \frac{h_{ii}}{(1-h_{ii})}$

(p) [ **True or False**] Observations with large studentized residuals (based on an appropriate test) but small Cook's distance may be removed from an analysis to avoid influencing the estimate of the variance as there is evidence that the they have a different mean from the regression function.

• large studentized resid => outlier

(q) [ **True or False**] Underdispersion in logistic regression occurs when the residual deviance is much smaller than the residual degrees of freedom and is evidence of lack of fit.

$\frac{\text{residual dev}}{df} < 1$

(r) [ **True or False**] For constructing a prediction interval for a single new point based on linear regression model with $n$ observations and $p$ coefficients, we use a $t$ distribution $n - p$ degrees of freedom.

(s) [ **True or False**] We do not need to worry about the bias term for the Mean Squared prediction error when using Ordinary Least Squares, as OLS is unbiased.

(t) [ **True or False**] For adding a factor that has $J$ levels to a normal regression model (with no other predictors), the change in residual sum of squares would have an $F^2$ distribution with $J$ and $n - J$ degrees of freedom.

↳ adds $J-1$ predictors

2. Fill in the blanks: **Failure to write the response in the space provided may result in a zero grade during scanning**

(a) The link function in generalized linear models relates the expected value of the response to the predictors.

(b) Cook's distance is a measure of potential influence.

(c) Bonferroni corrected t-tests can be used to test if points are outliers.

(d) When applying the Bonferonni correction for testing outliers we compare the p-values to $\frac{\alpha}{n}$ so that the overall error rate is no greater than 0.05.

(e) We use a $\chi^2$ distribution with $J-1$ degrees of freedom for testing whether we need to add a factor with $J$ levels in normal linear regression with $p$ other coefficients and a sample size $n$.

(f) We use a student-t distribution with $n-p$ degrees of freedom to construct prediction intervals in linear regression with $n$ observations and $p$ coefficients.

$Y^* = f(x^*) + \epsilon^*$

3. For predicting a new case $Y^*$, the prediction MSE is given by

$$\mathsf{E}[(Y^* - \widehat{f(x^*)})^2] = \mathsf{Var}(\widehat{f(x^*)}) + [\mathsf{Bias}(\widehat{f(x^*)})]^2 + \mathsf{Var}(\epsilon^*)$$

(a) Suppose your naive estimate of $f(x)$ is always 42. **Circle** the best option. $\hat{f}(x) = 42$

i. [**True or False**] The $\mathsf{Var}(\widehat{f(x^*)})$ will always be greater than zero

Variance will always be 0.

ii. The $[\mathsf{Bias}(\widehat{f(x^*)})]^2$ will be (a) less than, (b) greater than, (c) equal, (d) not enough information than $\mathsf{Var}(\widehat{f(x^*)})$.

↳ bias could = 0 → because bias $= f(x^*) - E(\hat{f}(x^*))$

$= f(x^*) - 42$, but what if $f(x^*) = 42$

→ Then bias = 0 = var

(b) You decide to use a much more sophisticated estimator. **Circle** the best option

i. The term $\mathsf{Var}(\widehat{f(x^*)})$ will (a) decrease, (b) increase, (c) stay unchanged, (d) cannot tell.

↳ must be greater than 0.

ii. The term $[\mathsf{Bias}(\widehat{f(x^*)})]^2$ will (a) decrease, (b) increase, (c) stay unchanged, (d) cannot tell.

↳ depends on if we achieve better functional form

iii. The term $\mathsf{Var}(\epsilon^*)$ will (a) decrease, (b) increase, (c) stay unchanged, (d) cannot tell.

$Y^* = f(X^*) + \epsilon^*$

$\hat{f}(x)$ always 42

so $\hat{f}(X^*) = 42$ as well

↳ variance = 0

• bias always ≥ 0c

$E\left[E(\hat{f}(X^*) - \hat{f}(X^*))^2\right]$

$E(42) - 42$

$(42 - 42)\,\sigma^2$

$E(0) = 0$

4. Some reasearchers collected a prostate cancer dataset with the following variables

**weight** prostate weight

**cavol** cancer volume

**age** age of man

**svi** seminal vesicle invasion (0 or 1)

**gleason** Gleason score (6, 7, 8, 9)

**psa** prostate specific antigen

The researcher wanted to see if `svi` was associated with the other predictors and obtained the following summary:

```
> svi.1 = glm(svi ~  log(psa) +  factor(gleason) + age, data=Prostate, family=binomial)
> summary(svi.1)

Call:
glm(formula = svi ~ log(psa) + factor(gleason) + age, family = binomial,
    data = Prostate)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.63661  -0.48619  -0.00005   0.00000   2.31866

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -2.873e+01  2.381e+03  -0.012   0.9904
log(psa)          2.302e+00  6.503e-01   3.540   0.0004 ***
factor(gleason)7  1.868e+01  2.381e+03   0.008   0.9937
factor(gleason)8  6.921e-01  1.789e+04   0.000   1.0000
factor(gleason)9  1.931e+01  2.381e+03   0.008   0.9935
age               3.428e-02  5.360e-02   0.640   0.5224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 101.353  on 96  degrees of freedom
Residual deviance:  50.875  on 91  degrees of freedom
AIC: 62.875

Number of Fisher Scoring iterations: 19

> anova(svi.1, test="Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: svi

Terms added sequentially (first to last)
```

```
                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            96    101.353
log(psa)         1   40.571        95     60.781 1.896e-10 ***
factor(gleason)  3    9.493        92     51.289   0.02341 *
age              1    0.413        91     50.875   0.52037
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) **True or False** After accounting for psa, there is evidence at the $\alpha = 0.005$ level that svi is related to the Gleason score.

(b) **True or False** After accounting for psa and Gleason score, there is evidence at the $\alpha = 0.005$ level that svi is related to the Gleason score.

(c) **True or False** There is evidence that model shows lack of fit based on the residual deviance.

$\frac{50.875}{91} < 1$

(d) **True or False** The Chi-squared test in the analysis of deviance to compare the null model to the fitted model would have 5 degrees of freedom.

$96 - 91 = 5$

(e) **True or False** Based on the output, we should accept the null model as the residual deviance under the null model (101.353) is close its expected value of 96 and there is no evidence of lack of fit:

```
> pchisq(101.353, 96, lower=FALSE)
[1] 0.3345873
```

(f) **True or False** The model with log(psa) and Gleason score has a residual deviance of 51.289.

5. The researcher ultimately decided to fit a simpler model:

```
> svi.2 = glm(svi ~  log(psa), data=Prostate, family=binomial)
> summary(svi.2)$coef

             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -7.898462   1.727365 -4.572550 4.818251e-06
log(psa)     2.249090   0.541584  4.152801 3.284307e-05

> confint(svi.2)

                 2.5 %    97.5 %
(Intercept) -11.861542 -5.009048
log(psa)      1.334727  3.480495
```

(a) Provide a sentence interpreting the confidence interval for the intercept in terms of odds for the researcher.

$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \log(psa)$ $\hookrightarrow \log(psa) = 0$

$\frac{\pi}{1-\pi_i} = e^{\beta_0}$ when $psa = 1$

$\hookrightarrow e^{(-11.861542,\ -5.009048)}$

$= (0.000007,\ 0.00667)$

At a prostate specific antigen equal to 1, we are 95% confident that the odds of seminal vesicle invasion are expected to be between 0.000007 and 0.00667

(b) Show that the odds ratio of seminal vesicle invasion with a 10% increase in psa levels is $1.10^{\beta_{psa}}$.

$$\frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_i}{1-\pi_i}} = \frac{e^{\beta_0}(1.1 * psa)^{\beta_1 psa}}{e^{\beta_0}(psa)^{\beta_1 psa}}$$

$$= \frac{(1.1)^{\beta_1 psa} * (psa)^{\beta_1 psa}}{(psa)^{\beta_1 psa}}$$

$$= (1.1)^{\beta_1 psa}$$

(c) Calculate a 95% confidence interval for $1.10^{\beta_{psa}}$

$(1.1)^{(1.334727,\ 3.480495)}$

$= (1.136,\ 1.393)$

(d) Provide a sentence interpreting the confidence interval for the researcher.

We are 95% confident that the odds ratio of a seminal vesicle invasion will be between 1.136 and 1.393 for a 10% increase in PSA levels.

Useful Distributions

Multivariate Normal

$$\begin{aligned}\mathbf{Y} \mid \boldsymbol{\mu}, \mathbf{V} &\sim \mathsf{N}(\boldsymbol{\mu}, \mathbf{V}) \\ p(\mathbf{Y}) &= (2\pi)^{-n/2}|V|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(Y-\boldsymbol{\mu})\} \quad \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^n, \mathbf{V} > 0\end{aligned}$$

Poisson

$$\begin{aligned}Y \mid \lambda &\sim \text{Poi}(\lambda) \\ p(y) &= \frac{y^\lambda e^{-\lambda}}{y!}\end{aligned}$$

Bernoulli

$$\begin{aligned}Y \mid \lambda &\sim \mathsf{Ber}(\pi) \\ p(y) &= \pi^y(1-\pi)^{1-y}\end{aligned}$$

Gamma

$$\begin{aligned}Y \mid a, b &\sim \mathsf{Gamma}(a, b) \\ p(y) &= \frac{b^a}{\Gamma(a)}y^{a-1}e^{-by} \text{ for } a > 0, b > 0, y > 0 \\ \mathsf{E}[Y] &= a/b \\ \mathsf{Var}[Y] &= a/b^2\end{aligned}$$

Generalized Inverse Gaussian

$$\begin{aligned}Y \mid \mu, \nu, \xi &\sim \mathsf{InvGaussian}(\mu, \nu, \xi) \\ p(y) &= \frac{1}{2}K_\mu(\sqrt{\nu\xi})\left(\frac{\nu}{\xi}\right)^{\mu/2}y^{\mu-1}\exp\left\{-\frac{1}{2}\left(\nu y + \frac{\xi}{y}\right)\right\} \quad y > 0, \nu > 0, \xi > 0, \mu \in \mathbb{R}\end{aligned}$$

$K_\mu(\cdot)$ is a modified Bessel function of the second kind

Double Exponential

$$\begin{aligned}Y \mid \lambda &\sim \mathsf{DE}(\mu, \lambda) \\ p(y) &= \frac{\lambda}{2}\exp\left(-\lambda|Y-\mu|\right) \qquad y \in \mathbb{R}, \lambda > 0, \mu \in \mathbb{R}\end{aligned}$$

Student-t

$$\begin{aligned}Y \mid \mu, \sigma^2 &\sim \mathsf{St}(\nu, \mu, \sigma^2) \\ p(y) &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\sigma^2\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2} \qquad y \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0\end{aligned}$$

Multivariate Student-t

$$\mathbf{Y} \mid \boldsymbol{\mu}, \mathcal{S} \sim \mathsf{St}(\nu, \boldsymbol{\mu}, \mathcal{S})$$

$$p(\mathbf{Y}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\nu\pi)^{p/2}|\mathcal{S}|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{1}{\nu}(\mathbf{Y}-\boldsymbol{\mu})^T\mathcal{S}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right)^{-(\nu+p)/2} \qquad \mathbf{Y}, \boldsymbol{\mu} \in \mathbb{R}^p, \mathcal{S} > 0$$