STA521 Midterm I October 6, 2018

Notes: Add Name before opening exam!

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. When time is up or you have completed the exam, please turn in to the proctor, collect your belongings and leave the room quietly. There is a class after this so please respect their time.
9. Grades may be curved upwards if the mean is lower than expected.
10. 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:______________________________

Signature:______________________________

Score:______________________________

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

   (a) [**True or False**]**:** When comparing a logistic regression model to the null model, we should select the model that has the smallest deviance.

   (b) [**True or False**]**:** For logistic regression, we apply the logit transformation to the response.

   (c) [**True or False**]**:** A small value for the residual deviance in Gaussian regression is an indication of a good model.

   (d) [**True or False**]**:** For normal and binary regression, the variance of an observation is assumed to be constant.

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

   (f) [**True or False**]**:** If we fit a model with $\log(Y)$ and Y, R^2 can be used to select which transformation is better (assuming that we have the same predictors in both).

   (g) [**True or False**]**:** BoxTidwell is used to find transformations of the response and predictors so that their joint distribution is close to normal.

   (h) [**True or False**]**:** Coefficients with small absolute values are not important and those variables can be dropped from the model.

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

   (j) [**True or False**]**:** For adding a factor with J levels to a regression model, `R` adds J dummy or indicator variables.

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

   (l) [**True or False**]**:** For testing whether a case is an outlier, the degrees of freedom for the t-test are the same as the residual degrees of freedom, $n - p$, where p is the number of coefficients in the mean function, including the intercept.

   (m) [**True or False**] A case with a large Cook's distance will still be influential if we only transform the response as the leverage will not change.

   (n) [**True or False**] If we add more predictors or increase our sample size we can reduce the irreducible error of the model.

   (o) [**True or False**] A value for Cook's Distance greater than one implies that a point is also an outlier.

   (p) [**True or False**] Observations with high leverage should removed from an analysis to avoid influential cases.

   (q) [**True or False**] The residual deviance of the saturated model in logistic regression is zero.

   (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 $n - p - 1$ degrees of freedom.

   (s) [**True or False**] Ordinary Least Squares can be used to find estimates of the mean function in logistic regression.

   (t) [**True or False**] BoxCox is used to find the link function for Gaussian linear regression.

2. Fill in the blanks:

   (a) For linear regression with n observations and p coefficients, the studentized residuals have a __________________________ distribution with _______________ degrees of freedom.

   (b) For logistic regression with n observations and p coefficients, the coefficients have a _______________ distribution with _______________ degrees of freedom for large sample sizes.

   (c) The change in deviance in a logistic or Poisson regression model has a _______________ distribution with ______________________ degrees of freedom for large samples (assuming the model is correct).

   (d) For testing whether we need to add a factor with J levels in logistic regression with a sample size n we use a _______________ distribution with degrees of freedom ______________________________.

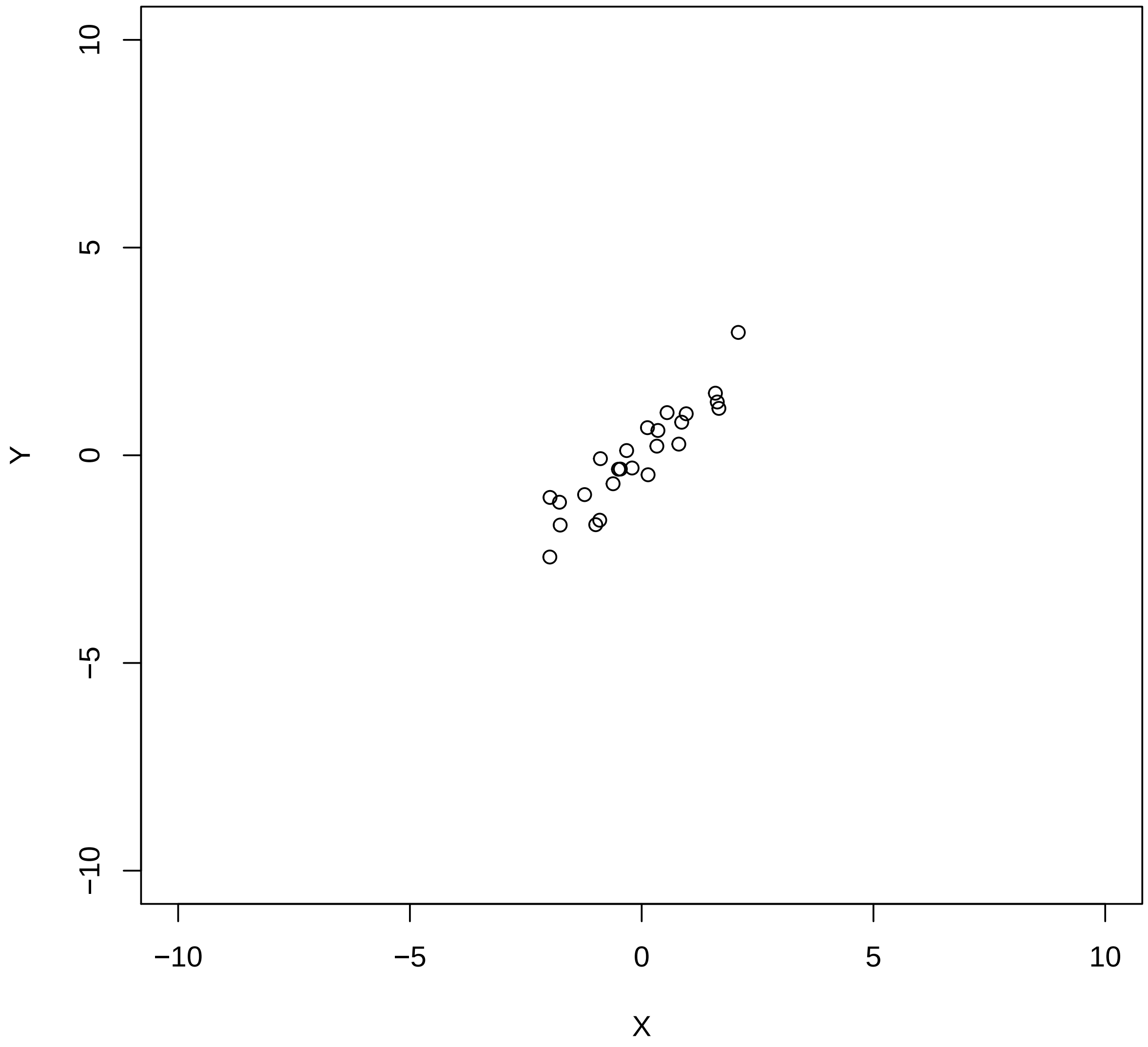
   (e) For linear regression with n observations and p coefficients, we use a _______________ distribution with _______________ degrees of freedom to construct confidence intervals for the coefficients.

3. In the plot below, add points and clearly label them with the corresponding letters for the following situations:

   (a) an outlier that is not influential

   (b) actually influential, but not an outlier

   (c) high leverage, but helps to reduce the variance of $\hat{beta}_1$ in the simple linear regression.



4. The following problem considers modeling the volume of prostate cancer tumors as a function of the 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

(a) A linear regression model was fit to the data with the response `cavol` leading to the following residual plots. In the last plot the horizontal line is the absolute value of the t-quantile such that $P(t > 3.6) = .025/n$ for 90 degrees of freedom. From these plots we can conclude that:

i. there is(are) _______________ outlier(s).

ii. there is(are) _______________ potentially influential observation.

iii. there is(are) _______________ actually influential observation.

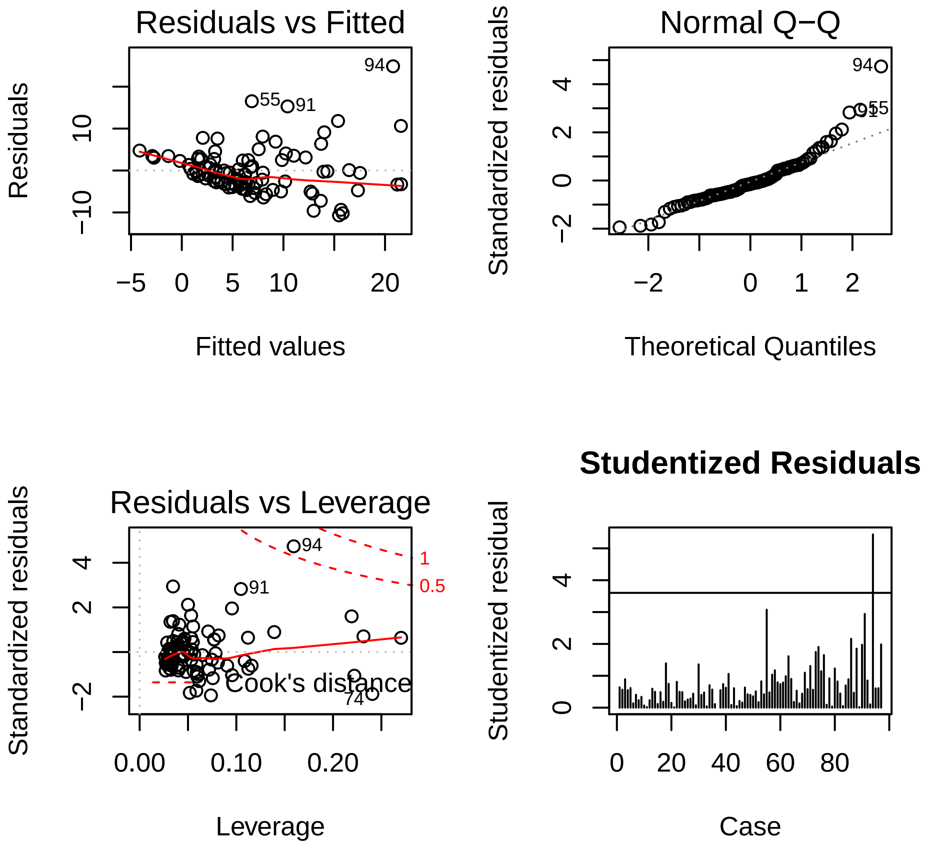
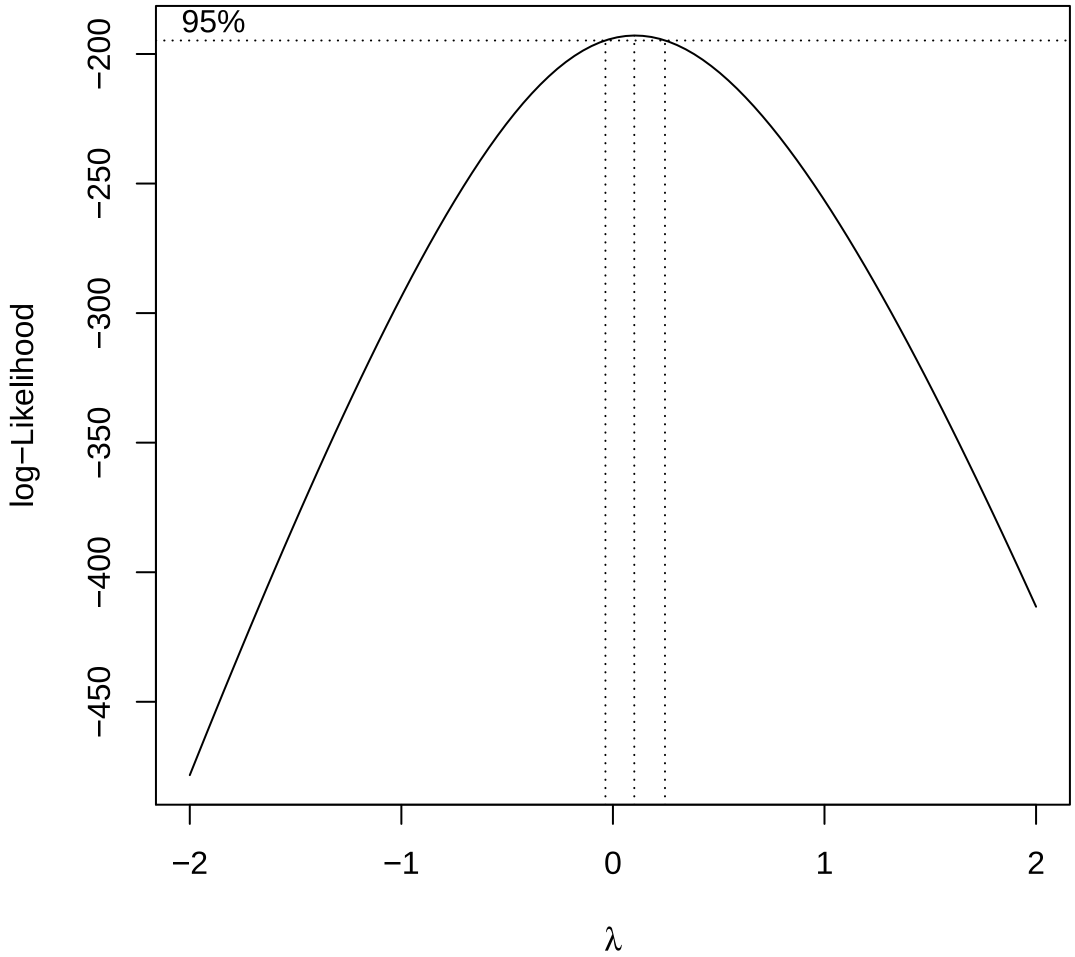


Figure 1: Residual plots from the Prostate regression for `cavol`.



(b) The researcher decided to try to transform the response, but could not remember how to interpret the plot above. What recommendation for transforming the response would you give to the researcher? Explain.

(c) The researcher obtained the following model summaries

```
p.lm = lm(log(cavol) ~ GScore + age + svi +  log(psa), data=Prostate)
summary(p.lm)
```

```
##
## Call:
## lm(formula = log(cavol) ~ GScore + age + svi + log(psa), data = Prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7995 -0.5370  0.0000  0.5997  1.8160
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.06034    0.70767  -1.498   0.1375
## GScore7      0.34129    0.19836   1.721   0.0888 .
## GScore8      0.49516    0.80104   0.618   0.5380
## GScore9      0.67477    0.39097   1.726   0.0878 .
## age          0.01014    0.01121   0.904   0.3685
## svi          0.42544    0.23683   1.796   0.0758 .
## log(psa)     0.57857    0.08911   6.493 4.51e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7783 on 90 degrees of freedom
## Multiple R-squared:  0.5912,Adjusted R-squared:  0.564
## F-statistic:  21.7 on 6 and 90 DF,  p-value: 1.277e-15
```

```
confint(p.lm)
```

```
##                   2.5 %     97.5 %
## (Intercept) -2.46625559 0.34557933
## GScore7     -0.05278976 0.73537334
## GScore8     -1.09625515 2.08657059
## GScore9     -0.10196910 1.45150145
## age         -0.01214279 0.03241642
## svi         -0.04505877 0.89594653
## log(psa)     0.40154005 0.75560194
```

```
anova(p.lm)
```

```
## Analysis of Variance Table
##
## Response: log(cavol)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## GScore     3 34.478 11.4926 18.9746 1.276e-09 ***
## age        1  1.035  1.0348  1.7085    0.1945
## svi        1 17.801 17.8012 29.3903 4.919e-07 ***
## log(psa)   1 25.534 25.5337 42.1569 4.509e-09 ***
## Residuals 90 54.511  0.6057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i. **True or False** Based on this output, we could remove all variables except `log(psa)`.

ii. **True or False** Based on the R^2 value there is evidence of lack of fit.

iii. **True or False** Keeping the other variables constant, a one-unit increase in `psa` would lead to a 40 to 76 percent increase in the cancer volume. Explain.

iv. **True or False** For constructing a prediction interval, we would use a Student t distribution with 90 degrees of freedom.

v. **True or False** If the `psa` levels increase by 10%, we expect that the median cancer volume will increase by about 4 to 8 percent (holding the other variables constant).

vi. **True or False** The median cancer volume for tumors with seminal vesicle invasion is about 4.4% lower than tumors that do not have seminal vesicle invasion (holding the other variables constant).

vii. **True or False** To construct an added-variable plot to remove the effects of the other predictors from `svi` we should use a logistic regression as `svi` is binary.

viii. **True or False** The regression model explains about 59% of the variation in cancer volume.

5. 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.

(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.

(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:

```r
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.

6. 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)

## Waiting for profiling to be done...

##                 2.5 %    97.5 %
## (Intercept) -11.861541 -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.

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

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

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