STAT 526 HW 3

P0

Name

Bowen Zheng

P1

Consider a logit and probit model for binary data with one predictor. Show that if the predictor variable equals the negative value of the intercept (β_0) divided by the slope (β_1) then the probability is 1/2. This value of the predictor is often denoted the LD50. Find the LD50 for the complementary log-log link, i.e., g(p) = log(- log(1 - p)).

Let $X = \frac{-\beta_0}{\beta_1}$.

For logit

We know

$$g(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Then:

$$\begin{aligned} \beta_0 + X\beta_1 &= \beta_0 + \frac{-\beta_0}{\beta_1}\beta_1 \\ &= 0 \\ &= log(\frac{p_i}{1-p_i}) \\ &\implies p_i = 0.5 \end{aligned}$$

For Probit

We know

$$g(p) = \Phi^{-1}(p) = \beta_0 + \beta_1 x$$

Then:

$$\begin{aligned}\beta_0 + X\beta_1 &= \beta_0 + \frac{-\beta_0}{\beta_1}\beta_1 \\ &= 0 \\ &= \Phi^{-1}(p_i) \\ \implies p_i &= 0.5\end{aligned}$$

Now Log-log

We can try to find the log log one by starting with $p_i = 0.5$.

We know the link function is

$$g(p) = \ln(-\ln(1-p)) = \beta_0 + \beta_1 x$$

Then:

$$\begin{aligned}\ln(-\ln(1-0.5)) &= \ln(-\ln(0.5)) \\ &= \ln(\ln 2) \\ \implies \ln(\ln 2) &= \beta_0 + \beta_1 x \\ \implies x &= \frac{\ln(\ln 2) - \beta_0}{\beta_1}\end{aligned}$$

P2

Explain why $\sum Y_i = \sum \hat{p}_i$ in logistic regression

We know that the log likelihood (from the slides in chapter 6) is

$$L(\beta) = \sum_{i=1}^{n} [Y_i(x_i^T \beta) - \log(1 + \exp(x_i^T \beta))]$$

Then, we take the derivative with respect to β and knowing $p_i = \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}$, we find

$$\begin{aligned}
\overrightarrow{0} = \frac{\partial}{\partial \beta} L(\beta) &= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} [Y_i(x_i^T \beta) - \log(1 + \exp(x_i^T \beta))] \\
&= \sum_{i=1}^{n} [Y_i x_i - \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} x_i] \\
&= \sum_{i=1}^{n} [Y_i x_i - p_i x_i] \\
&= \sum_{i=1}^{n} [(Y_i - p_i) x_i]
\end{aligned}$$

In the model with intercept, we know for β_0, the column of $x_i = 1$, and we find:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} [(Y_i - p_i)1] \\
&= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} p_i \\
\implies \sum_{i=1}^{n} Y_i &= \sum_{i=1}^{n} p_i
\end{aligned}$$

P3 Faraway Chapter 2 Exercise #5

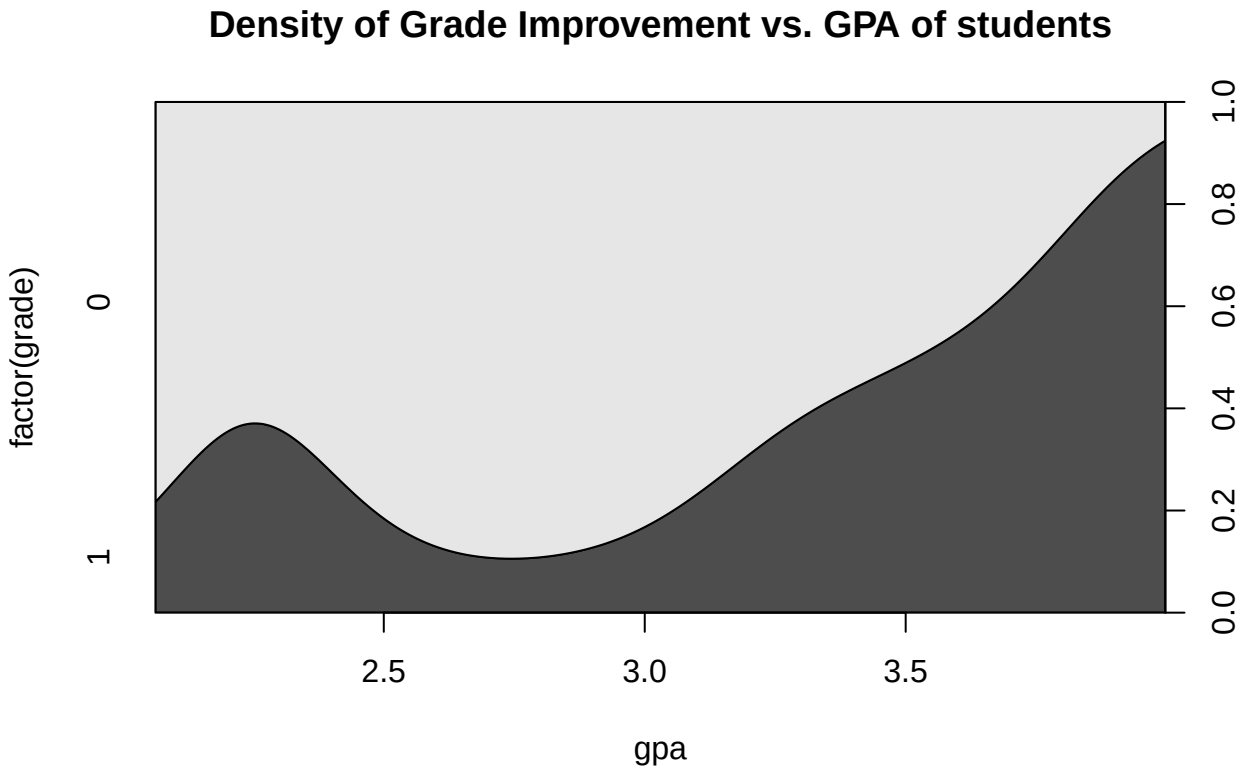
> A study was conducted to determine the effectiveness of a new teaching method in economics. The data may be found in the dataset spector. Write a report on how well the new method works. You should include suitable graphical depictions of the data, diagnostics on your chosen model and an interpretation of the effects of the predictors on the response.

We begin with looking at our data.

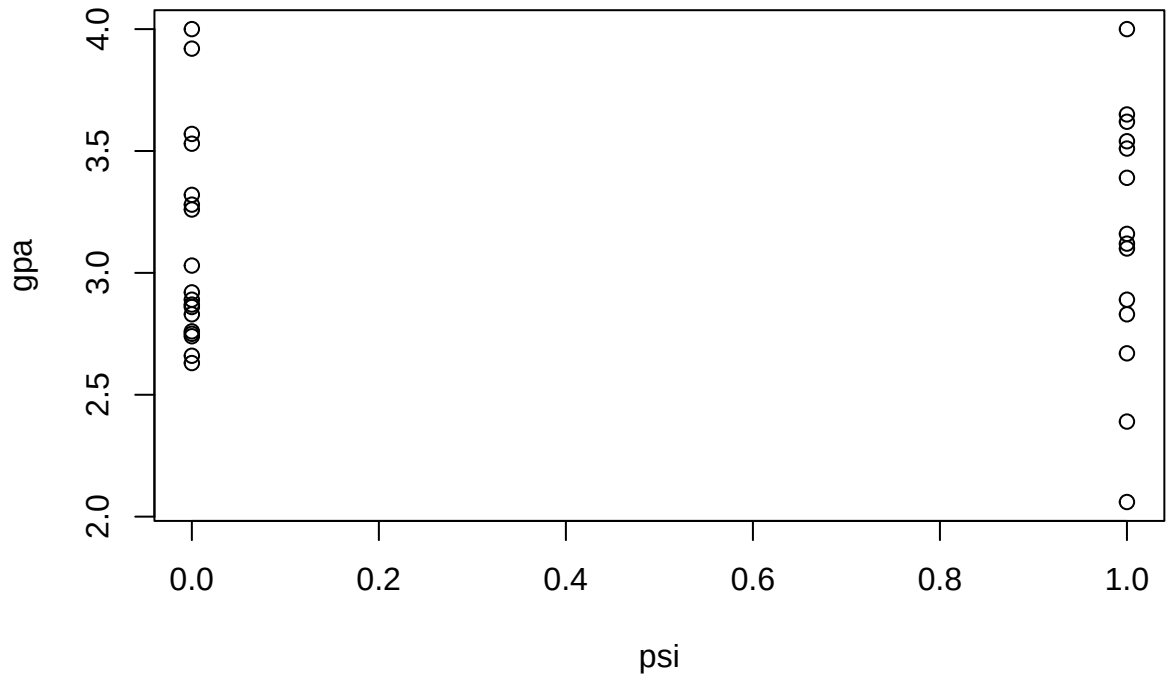
```
library(faraway)
data(spector)
attach(spector)
head(spector)
```

```
##   grade psi tuce  gpa
## 1     0   0   20 2.66
## 2     0   0   22 2.89
## 3     0   0   24 3.28
## 4     0   0   12 2.92
## 5     1   0   21 4.00
## 6     0   0   17 2.86
```

```
cdplot(factor(grade) ~ gpa, data = spector, main = "Density of Grade Improvement vs. GPA of students")
```



```r
plot(gpa ~ psi)
```



Looks promising. We fit simple logistic model using glm

```r
# grade is the response, gpa, tuce, and psi are predictors
model1 <- glm(grade ~ gpa + tuce + psi, family = binomial, data = spector)
summary(model1)
```

```
##
## Call:
## glm(formula = grade ~ gpa + tuce + psi, family = binomial, data = spector)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.02135    4.93127  -2.641  0.00828 **
## gpa           2.82611    1.26293   2.238  0.02524 *
## tuce          0.09516    0.14155   0.672  0.50143
## psi           2.37869    1.06456   2.234  0.02545 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 25.779  on 28  degrees of freedom
```

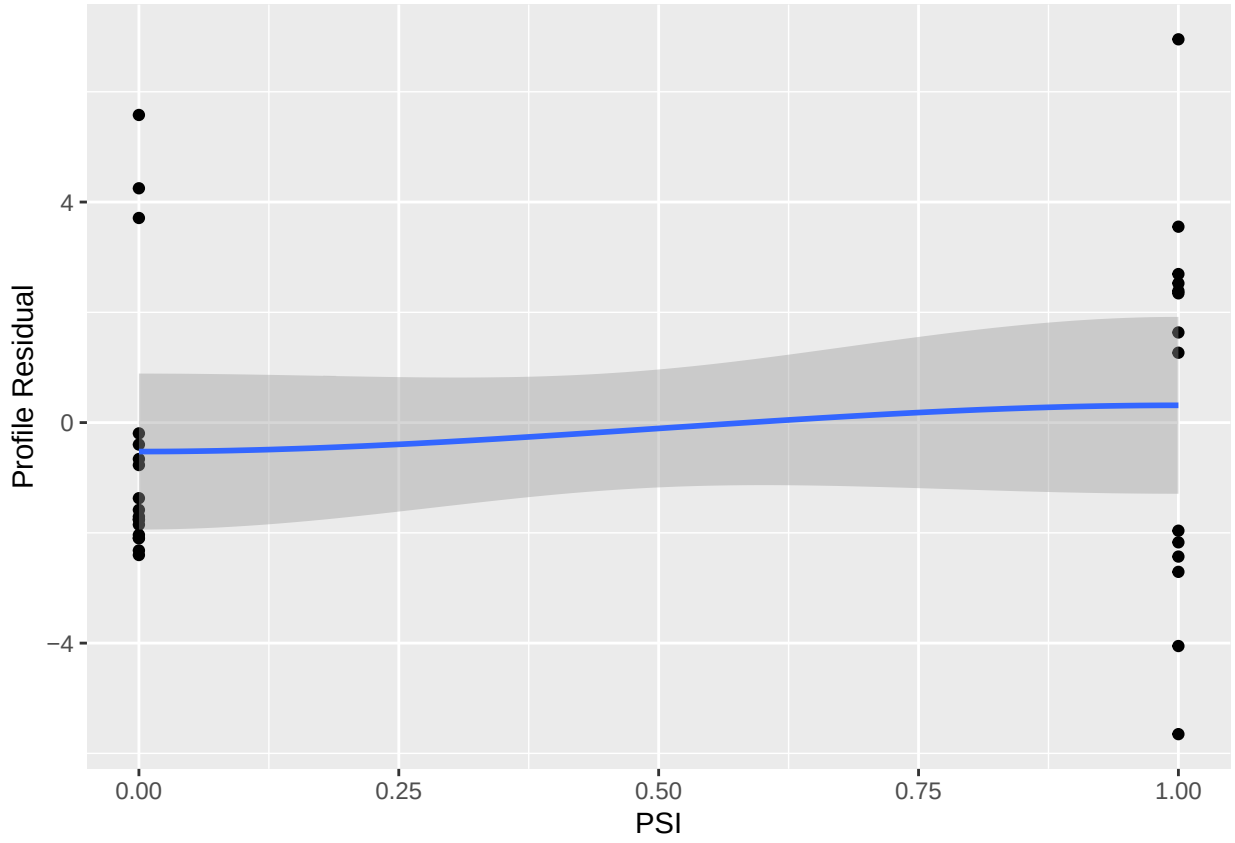
```
## AIC: 33.779
##
## Number of Fisher Scoring iterations: 5
```

```
## Obtaining the odds ratio for an increase of psi
exp(coef(model1)[4])
```

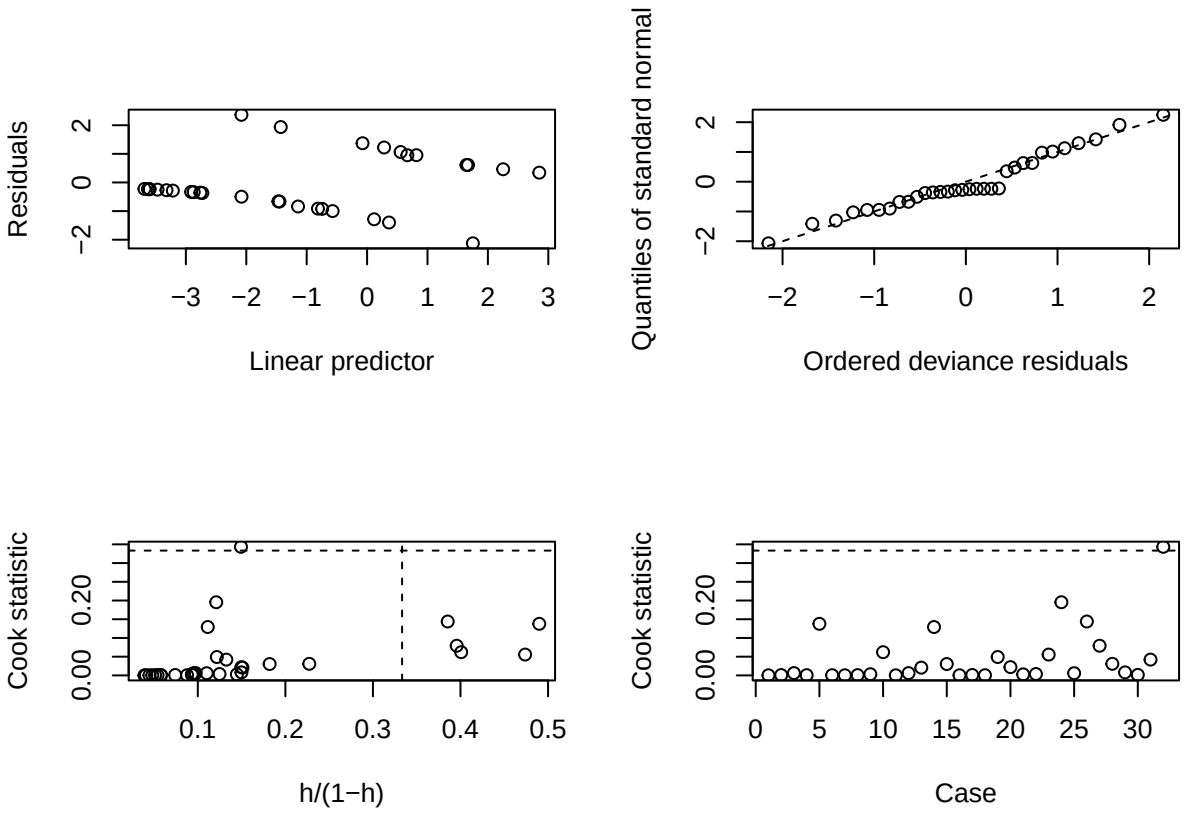
```
##      psi
## 10.79073
```

Using Professor Song's code, we will be assessing the diagnostics of our model:

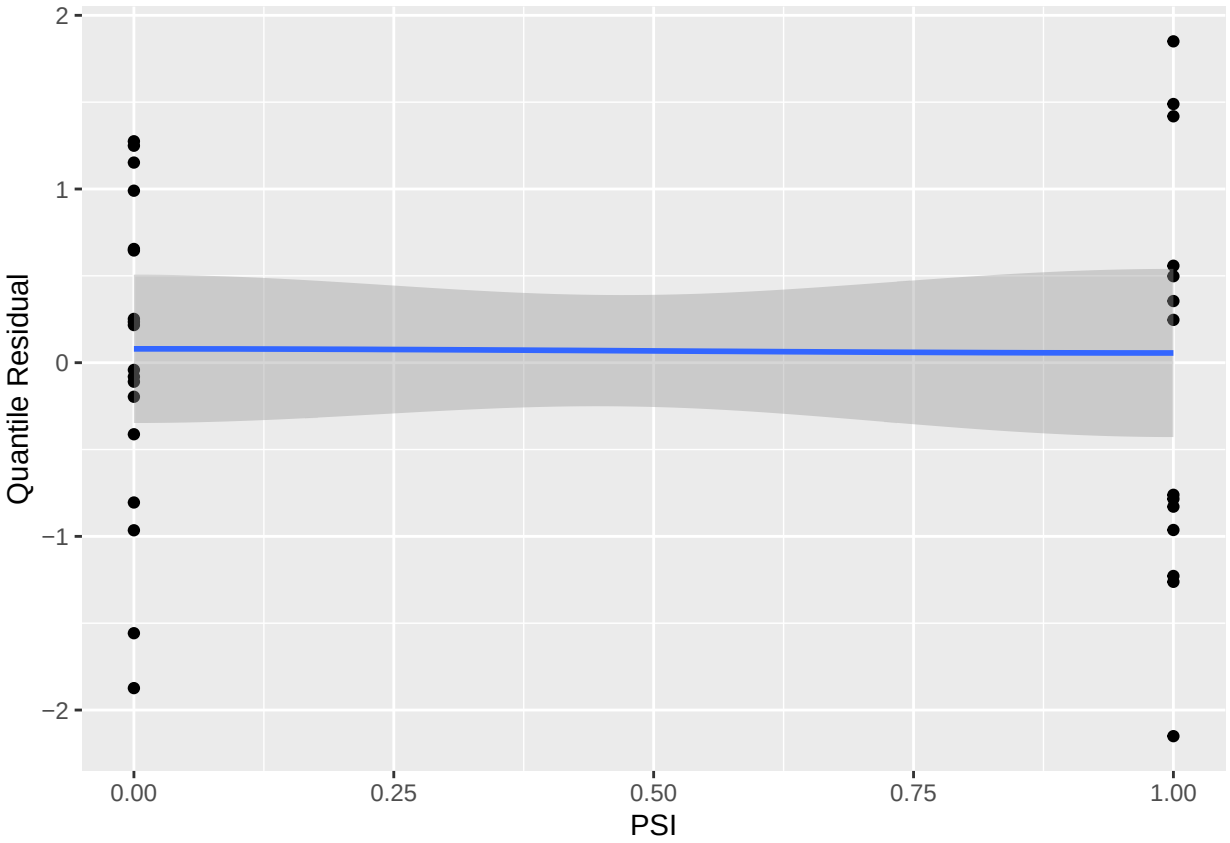
```
require(ggplot2)
##Generating a partial residual plot
qplot(psi,residuals(model1,type="partial"),geom=c('point','smooth'),
      xlab="PSI",ylab="Profile Residual")
```



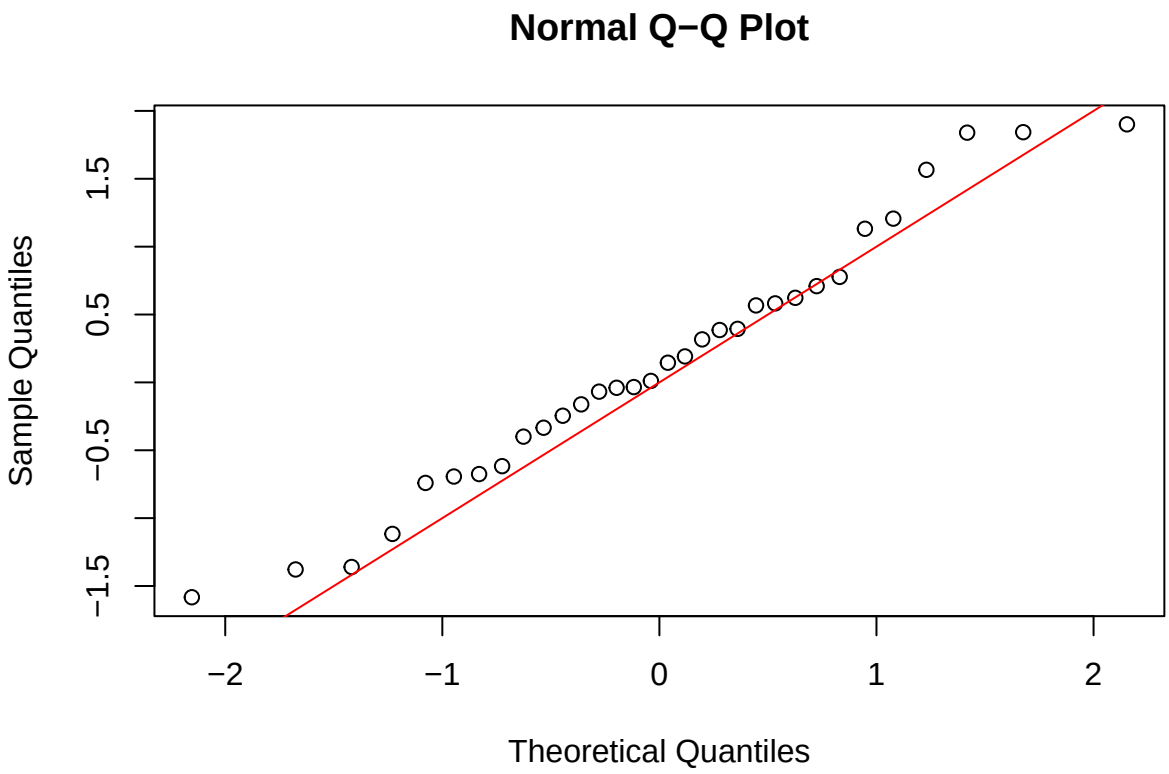

```
##Generating set of diagnostic plots
library(boot)
glm.diag.plots(model1)
```




```
##Consider looking at randomized quantile residuals
library(statmod)
##Residuals plot versus psi
qplot(psi,qres.binom(model1),geom=c('point','smooth'),
      xlab="PSI",ylab="Quantile Residual")
```




```
##QQplot
qqnorm(qres.binom(model1))
abline(a=0,b=1,col="red")
```



Unfortunately, many of the plots are not very helpful for diagnostics of logistic regression. Even with our quantile residual, it is hard to assess the model, but it generally looks bunched up kind of normally around some mean, so ideally, the assumption holds. The plots showing leverage and Cooks distance look good, not many if at all any, extreme outliers.

Our results are that the coefficient for psi is 2.338 which is about 10.79 times higher odds of grade improvement by those who do get the new teaching method (psi) versus who that don't, given that we keep the other variables constant. The significance for (psi) is below the 0.05 threshold, so we do not reject that it is statistically significant. We confirm the method should be effective.

P4 Simulation

For this problem, you will demonstrate that the deviance and Pearson X^2 goodness of fit measures are not accurate when there are no replicates.For each setting, you will generate N = 1000 data sets, fit each using the correct binary logistic regression model, and save the deviance and Pearson X^2 measures. You will then compare the distribution of each measure against the appropriate asymptotic X^2 density. You will use the following model in all simulations: y|x ~ Bernoulli(p) where x ~ N(0, 1) and logit(p)= 0.35 + x. Consider the following three settings: 1. n = 50 cases 2. n = 200 cases 3. n = 800 cases For each setting and goodness-of-fit measure, create a histogram and overlay the appropriate X^2 density. Summarize what you find, making sure to address whether the distributions of deviance and Pearson X^2 better fit the X^2 density as the sample size increases.

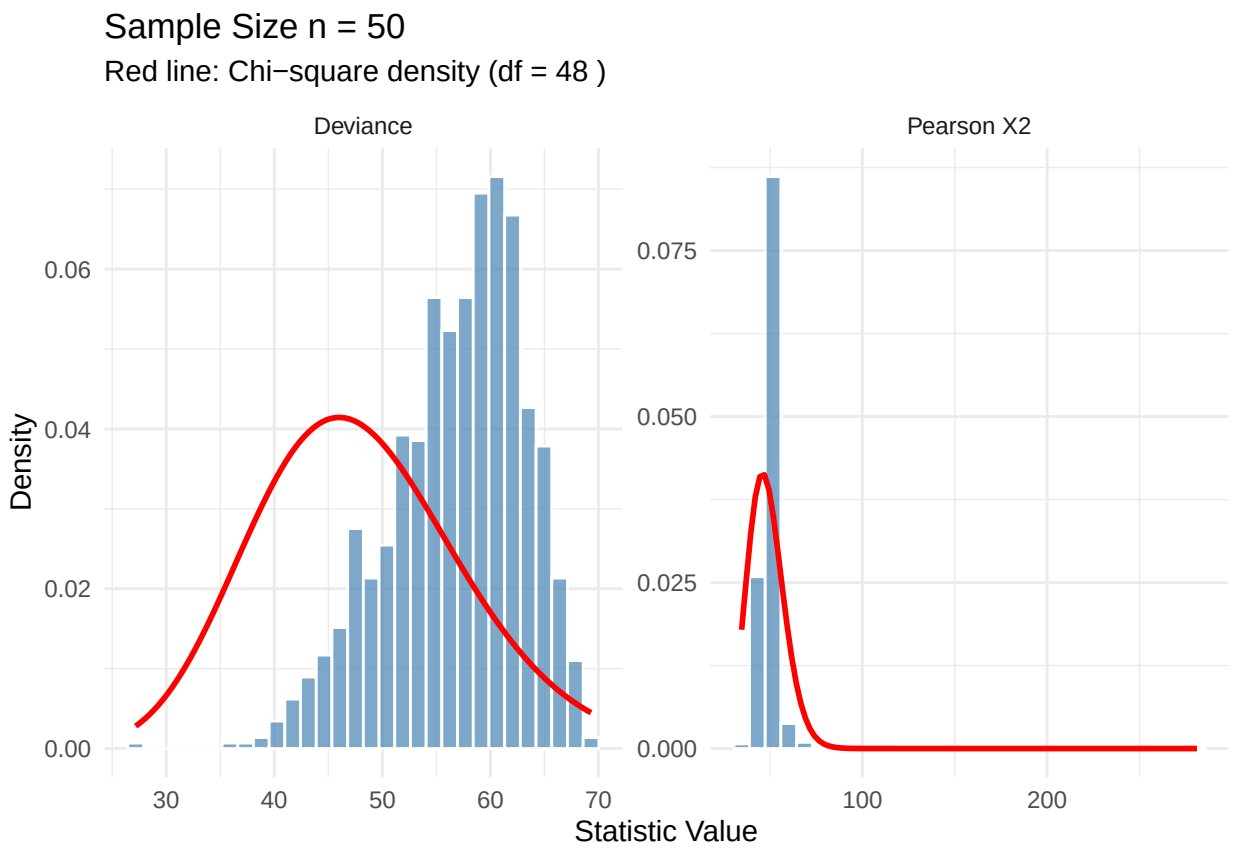
Ok first the simulation

```
set.seed(42)
N_data <- 1000
sample_sizes <- c(50, 200, 800)
run_simulation <- function(n) {
  deviances <- numeric(N_data)
  pearsons <- numeric(N_data)

  for (i in 1:N_data) {
    # 1. Generate Data
    x <- rnorm(n, 0, 1)
    logit_p <- 0.35 + x
    p <- exp(logit_p) / (1 + exp(logit_p))
    y <- rbinom(n, 1, p)
    # 2. Fit Model
    model <- glm(y ~ x, family = binomial)
    # 3. Save Measures
    deviances[i] <- deviance(model)
    pearsons[i] <- sum(residuals(model, type = "pearson")^2)
  }

  # Collect Data together / organize it all
  df <- data.frame(
    Value = c(deviances, pearsons),
    Measure = rep(c("Deviance", "Pearson X2"), each = N_data)
  )

  # Create Plot
  df_theory <- n - 2
  # Fancy Plot Code From Google
  p_plot <- ggplot(df, aes(x = Value)) +
    geom_histogram(aes(y = ..density..), bins = 30, fill = "steelblue", color = "white", alpha = 0.7) +
    stat_function(fun = dchisq, args = list(df = df_theory), color = "red", size = 1) +
    facet_wrap(~Measure, scales = "free") +
    labs(title = paste("Sample Size n =", n),
         subtitle = paste("Red line: Chi-square density (df =", df_theory, ")"),
         x = "Statistic Value", y = "Density") +
    theme_minimal()
```

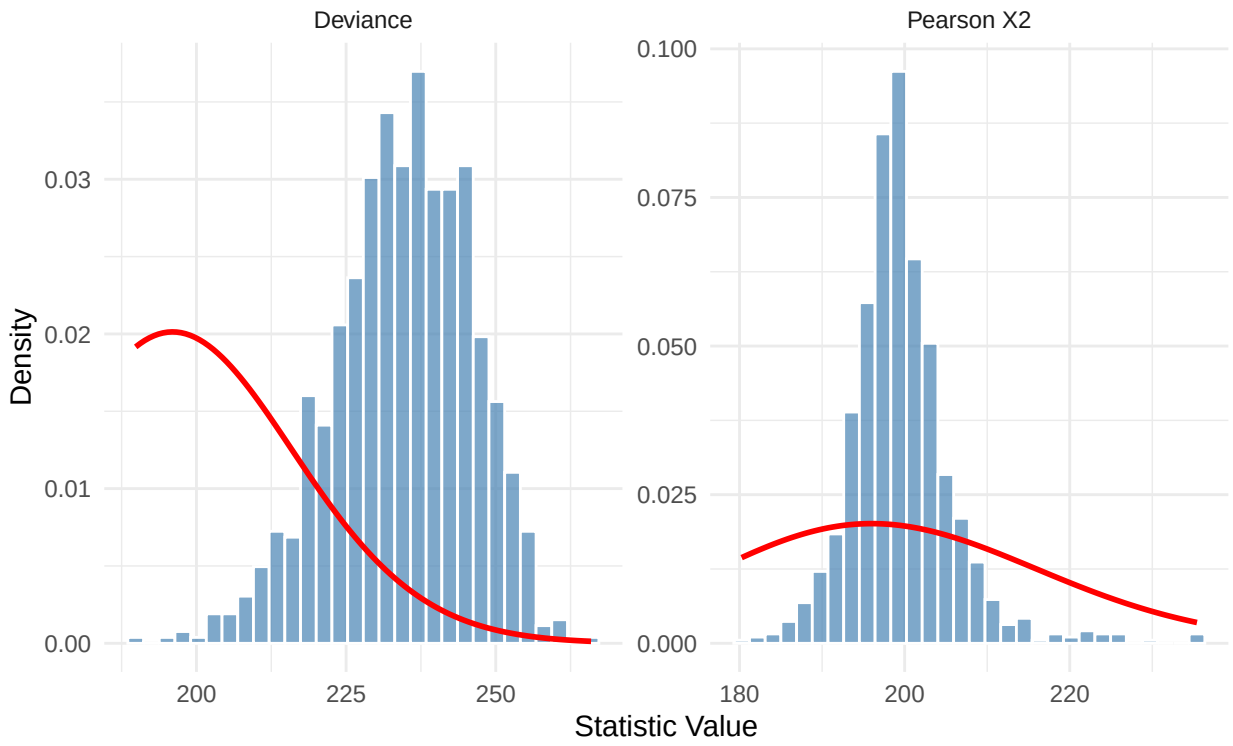
```
  return(p_plot)
}
```

```
plots <- lapply(sample_sizes, run_simulation)
plots
```

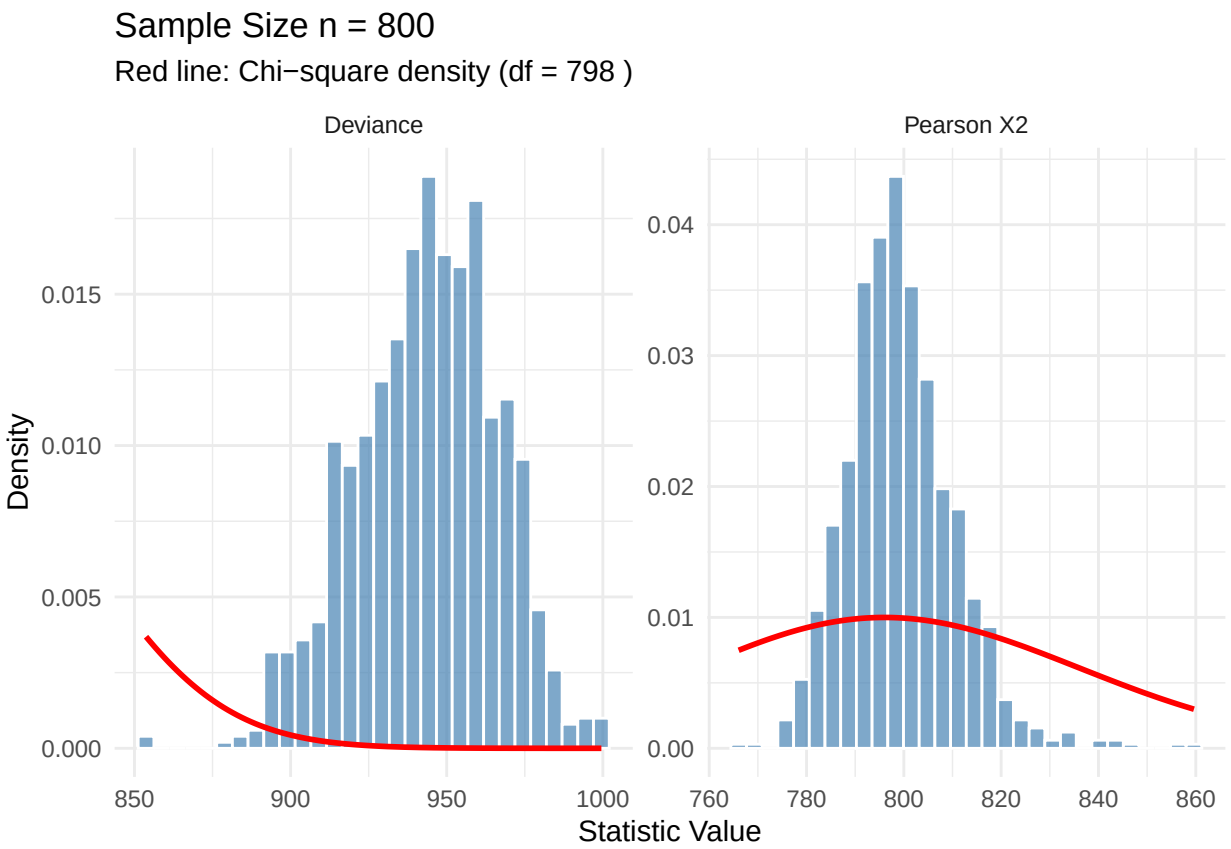
```
## [[1]]
```




```
##
## [[2]]
```


Sample Size n = 200
Red line: Chi−square density (df = 198)
Deviance
Pearson X2
Density
0.03
0.02
0.01
0.00
0.100
0.075
0.050
0.025
0.000
200
225
250
180
200
220
Statistic Value

```
##
## [[3]]
```



We can look at each graphs theoretical density in red versus our actual densities from simulation in blue. They do not match at all in all 3 cases. It seems clear that these would not be good goodness of fit tests for our logistic regression model.

Consistently, chi-squared distribution density seems to perform better / be more of a realistic density.

P5

1.

Write the assumptions of the model, and the expression of the log-likelihood.

We have a logistic model that is trying to estimate binary satisfied or not based on influence of renters, type of property and contact between renters. The first assumption would be that the response variable is binary, that we have independent observations, that the logit link function (default in R) fits the relationship between predictors and mean, that there is linear outcome from the explanatory variables, and finally that the variance (or dispersion of binomial family taken to be 1) is p(1-p). One quick initial issue is that situations with high contact between renters may not be very independent.

We can calculate the log-likelihood. n_i is the number of people living in each housing situation and y_i is the number of satisfied people in each housing situation.

$$L(\beta) = \prod_{i=1}^{24} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

We take the log

$$\ell(\beta) = \sum_{i=1}^{24} \left[\ln \binom{n_i}{y_i} + y_i \ln(p_i) + (n_i - y_i) \ln(1-p_i) \right]$$

2.

According to the fitted model, what percentage of renters, who have low influence on management, live in apartment, and have high contact between neighbors, are highly satisfied?

We can take the model and create a prediction for this. We are given all the coefficients so we can calculate it.

$$\eta = -0.6551 + 0(\text{Low Infl}) - 0.5285(\text{TypeApartment}) + 0.3130(\text{ContHigh}) = -0.8706$$

From our linear estimate, now we can convert to p.

```
# Use Inverse Logit Function in R
inv.logit(-0.8706)
```

```
## [1] 0.2951295
```

Thus, $p = 29.51\%$ which means that we predict that 29.51% of renters who are low influence on management, live in apartment, an dhave high contact between neighbors, are highly satisfied.

3.

> Do people who live in apartments have a significantly different probability of satisfaction than people who live in atriums? The correlation between the respective coefficients is 0.494.

We need to use the Wald Test from chapter 4 to test for our coefficients. The difference is $d = -0.5285 - (-0.4872) = -0.0413$.

We are given standard errors and the correlation, so we can calculate variance and covariance. We find $Var(\hat{\beta}_{TypeApartment}) = (0.1295)^2 = 0.01677$, $Var(\hat{\beta}_{TypeAtrium}) = (0.1728)^2 = 0.02986$, and $Cov(\hat{\beta}_{TypeApartment}, \hat{\beta}_{TypeAtrium}) = Corr \cdot SE_{Apt} \cdot SE_{Atr} = 0.494 \cdot 0.1295 \cdot 0.1728 = 0.01105$

We can use the formula:

$$SE(d) = \sqrt{Var(\hat{\beta}_{Apt}) + Var(\hat{\beta}_{Atr}) - 2 \cdot Cov(\hat{\beta}_{Apt}, \hat{\beta}_{Atr})}$$

Plugging things in again, we find:

$$SE(d) = \sqrt{0.01677 + 0.02986 - 2(0.01105)} \approx 0.1566$$

Now we can perform the Wald test. From chapter 4, we plug in our values and find:

$$Z = \frac{\text{Difference}}{SE(d)} = \frac{-0.0413}{0.1566} \approx -0.2637$$

It should be approximately normal(0,1). $p \approx 0.792$ for the two tailed test so we fail to reject that the two coefficients are different. The differences in probability for satisfaction is not statistically significant in this case between atrium and apartment.

4.

> Estimate the odds ratio of high satisfaction for groups with high contact among neighbors over groups with low contact among neighbors, using a 95% confidence interval.

The odds ratio would just be taking e to the coefficients value as the power.

$$\text{OR} = e^{0.3130} \approx 1.3675$$

Profile likelihood CI would likely be better in this case. However, we do not have access to be able to calculate the likelihood ratio. We will have to use the Wald CI.

We know $\hat{\beta}_j = 0.3130$, $se(\hat{\beta}_j) = 0.1077$, and $Z_{\alpha/2} = Z_{0.5/2} = 1.96$.

$$\text{CI}_{\beta_j} = \hat{\beta}_j \pm (Z_{\alpha/2} \times se(\hat{\beta}_j))$$

We find that the lower bound is $0.3130 - 0.2111 = 0.1019$ and $e^{0.1019} \approx 1.1073$. Upper bound is $0.3130 + 0.2111 = 0.5241$ and $e^{0.5241} \approx 1.6889$. Our 95% CI is (1.107, 1.689).