621 Homework 2

Sam Reeves

A Modern Approach to Regression with R

2.1

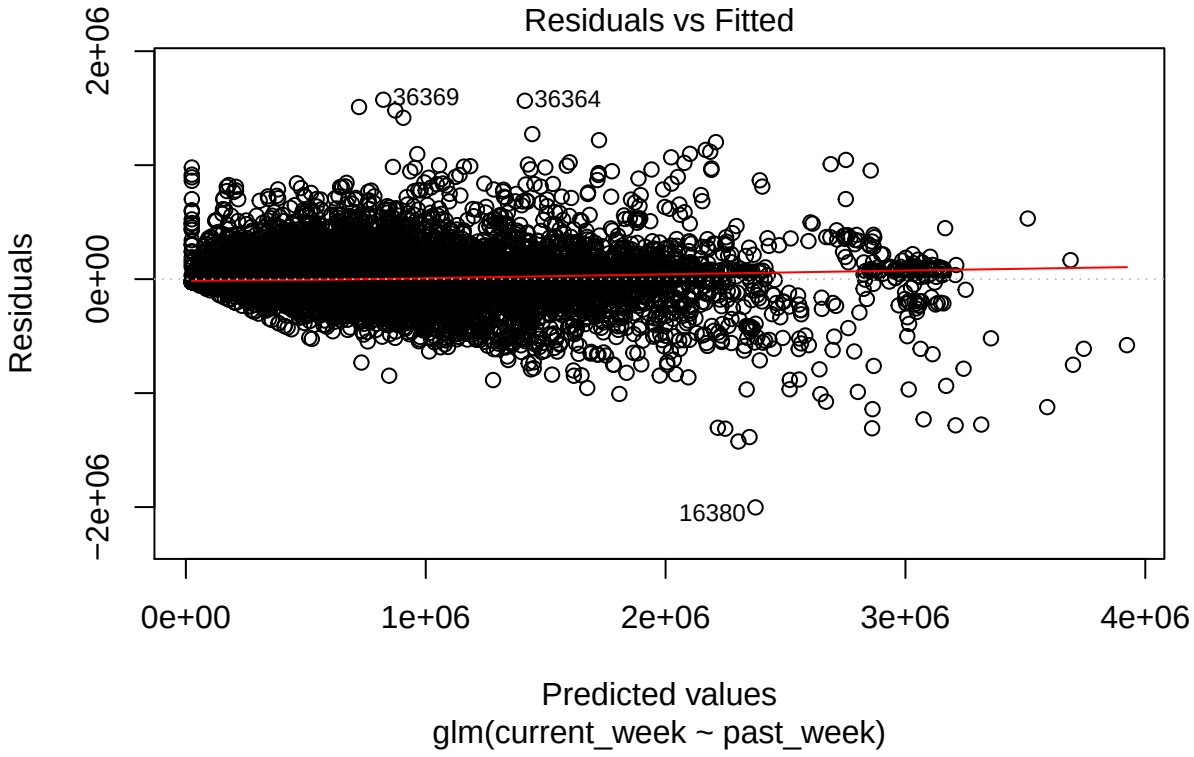
The web site www.playbill.com provides weekly reports on the box office ticket sales for plays on Broadway in New York. We shall consider the data for the week October 11–17, 2004 (referred to below as the current week). The data are in the form of the gross box office results for the current week and the gross box office results for the previous week (i.e., October 3–10, 2004). The data, plotted in Figure 2.6 , are available on the book web site in the file playbill.csv.

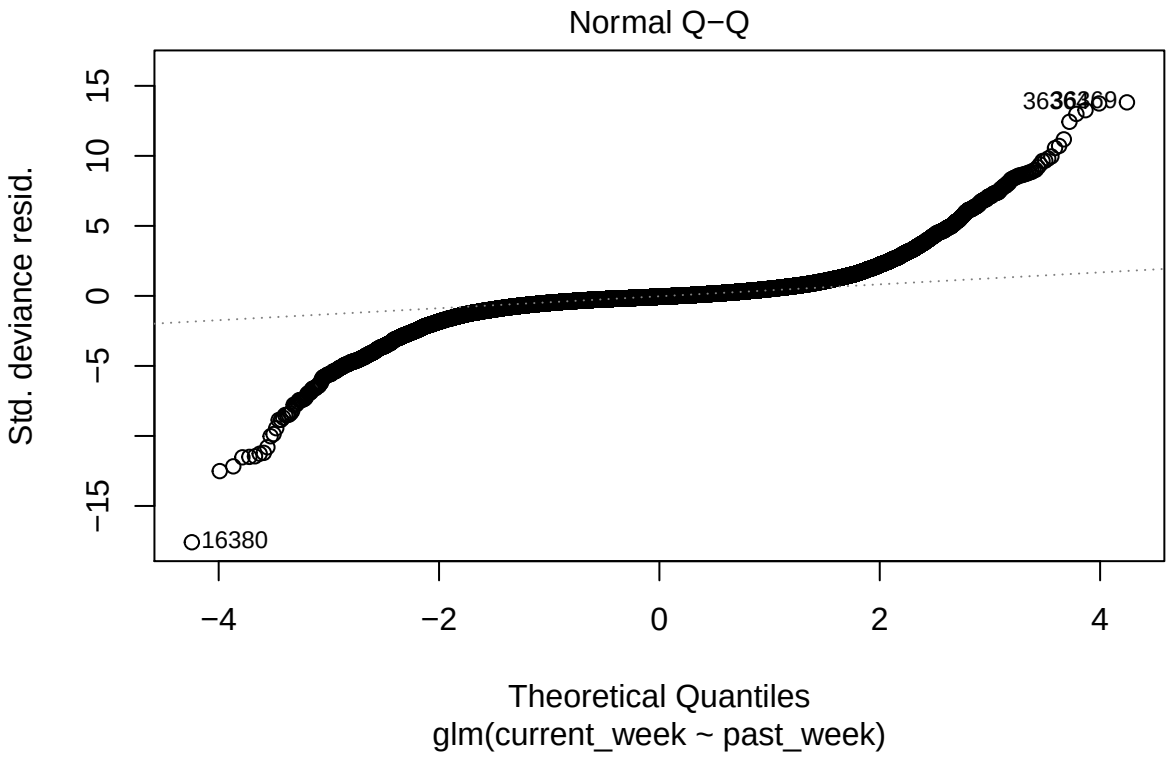
Fit the following model to the data: $Y = \beta_0 + \beta_1 x + \epsilon$ where Y is the gross box office results for the current week (in $) and x is the gross box office results for the previous week (in $). Complete the following tasks:

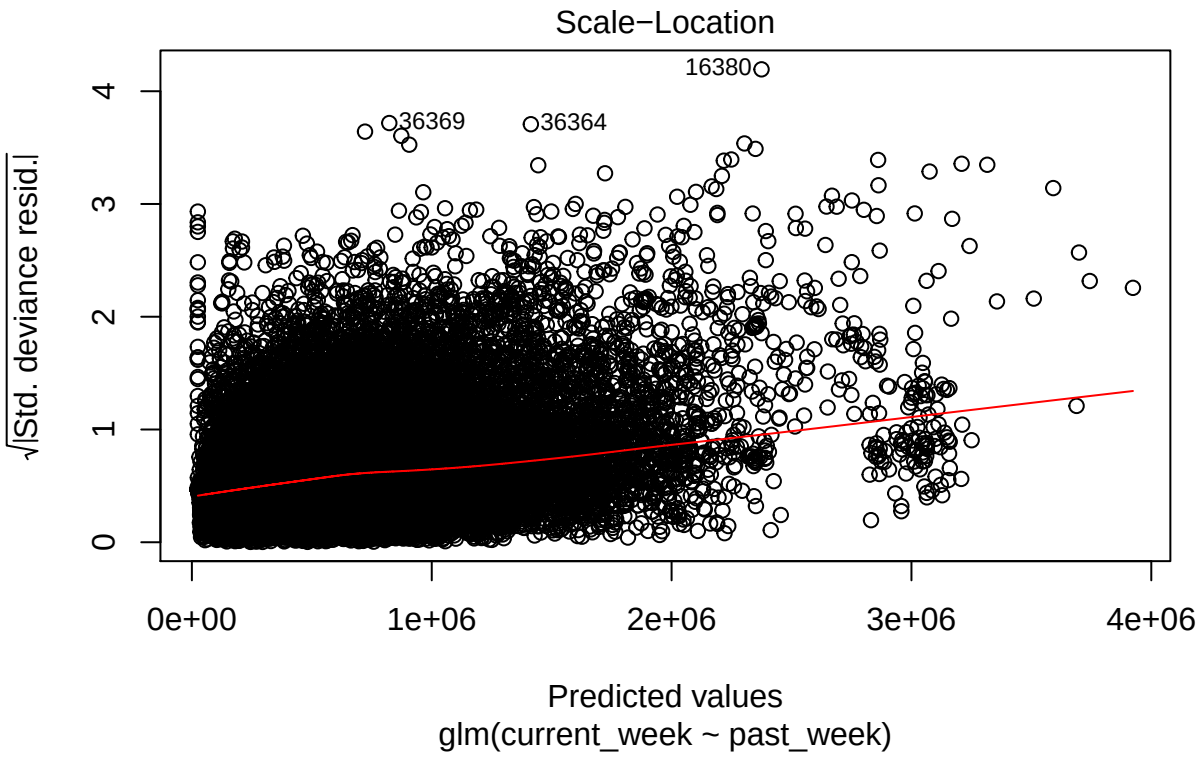
```
pb <- read.csv('playbill.csv')

summary(lm1 <- glm(current_week ~ past_week,
                   data = pb))
```

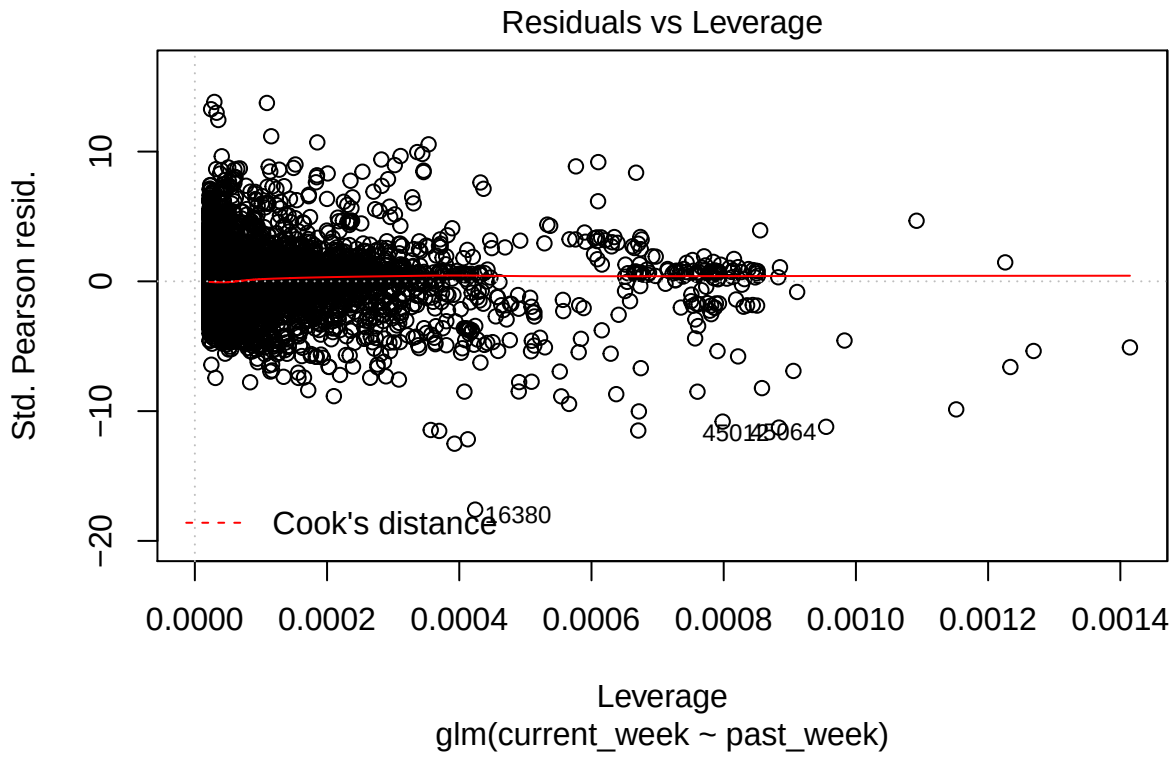
```
##
## Call:
## glm(formula = current_week ~ past_week, data = pb)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2002747    -35777     -6625     29735   1574083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.484e+04  8.809e+02    28.2   <2e-16 ***
## past_week   9.647e-01  1.225e-03   787.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 12967761840)
##
##     Null deviance: 8.6338e+15  on 45455  degrees of freedom
## Residual deviance: 5.8944e+14  on 45454  degrees of freedom
## AIC: 1187479
##
## Number of Fisher Scoring iterations: 2
```

```
plot(lm1)
```


Residuals vs Fitted
36369
36364
16380
Residuals
2e+06
0e+00
−2e+06
0e+00
1e+06
2e+06
3e+06
4e+06
Predicted values
glm(current_week ~ past_week)


Normal Q−Q
Std. deviance resid.
15
10
5
0
−5
−15
16380
−4
−2
0
2
4
Theoretical Quantiles
glm(current_week ~ past_week)


Scale−Location
16380
36369
36364
√|Std. deviance resid.|
0
1
2
3
4
0e+00
1e+06
2e+06
3e+06
4e+06
Predicted values
glm(current_week ~ past_week)



(a) Find a 95% confidence interval for the slope of the regression model, b 1 . Is 1 a plausible value for β_1 ? Give a reason to support your answer.

```
confint(lm1, 'past_week', level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
##     2.5 %    97.5 %
## 0.9622568 0.9670579
```

Although 1 is close to our values, it doesn't fall within the 95% confidence interval.

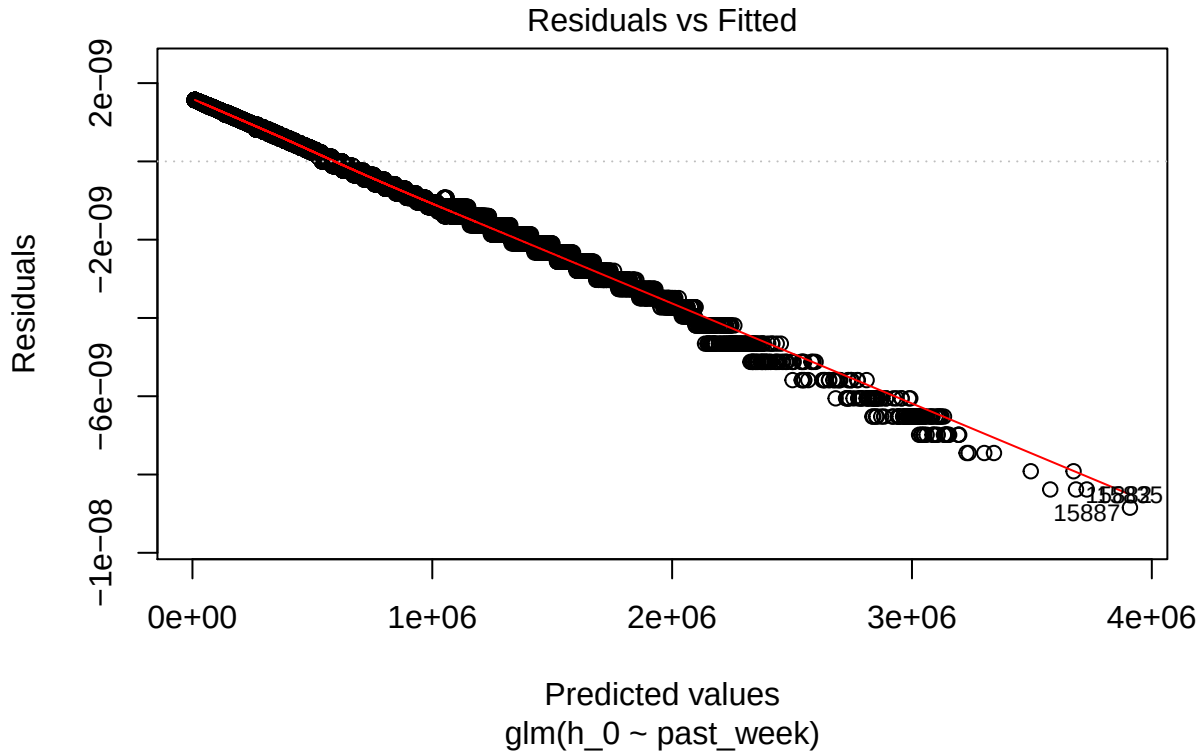
(b) Test the null hypothesis $H_0 : \beta_0 = 10000$ against a two-sided alternative. Interpret your result.

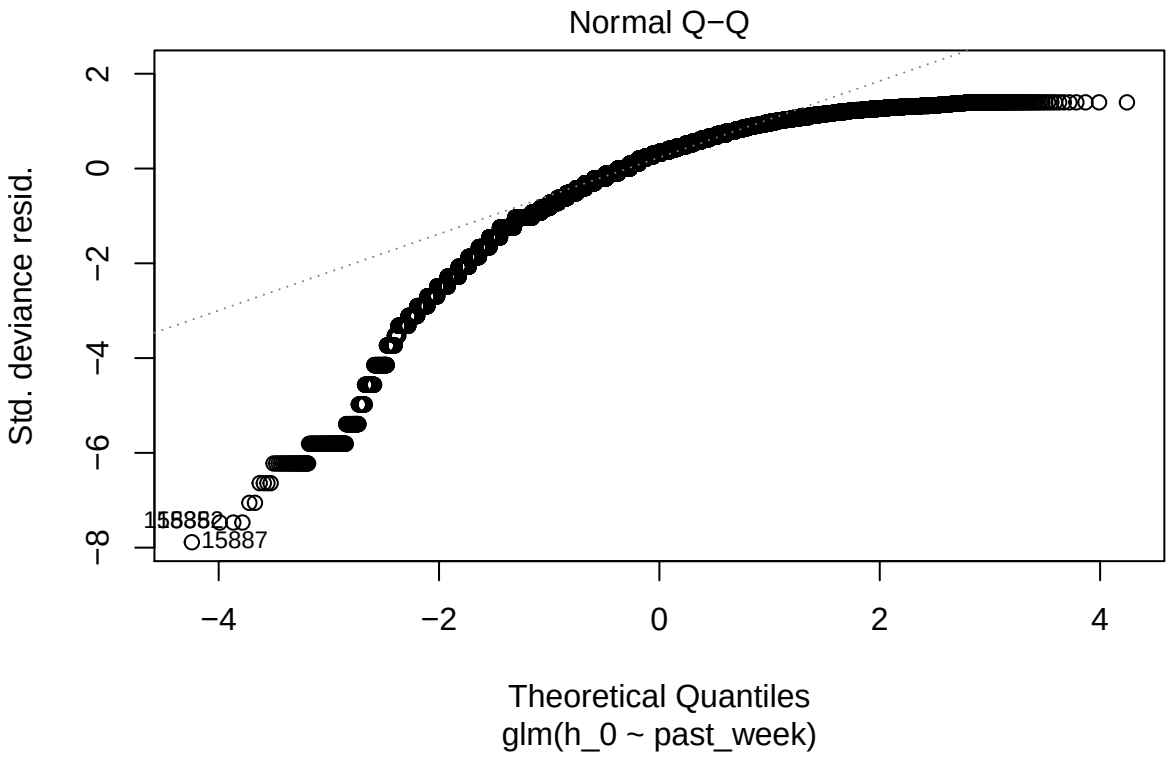
```
h_0 <- 10000 + 0.9647 * pb$past_week
pb <- cbind(pb, h_0)

summary(lm2 <- glm(h_0 ~ past_week,
                   data = pb))
```

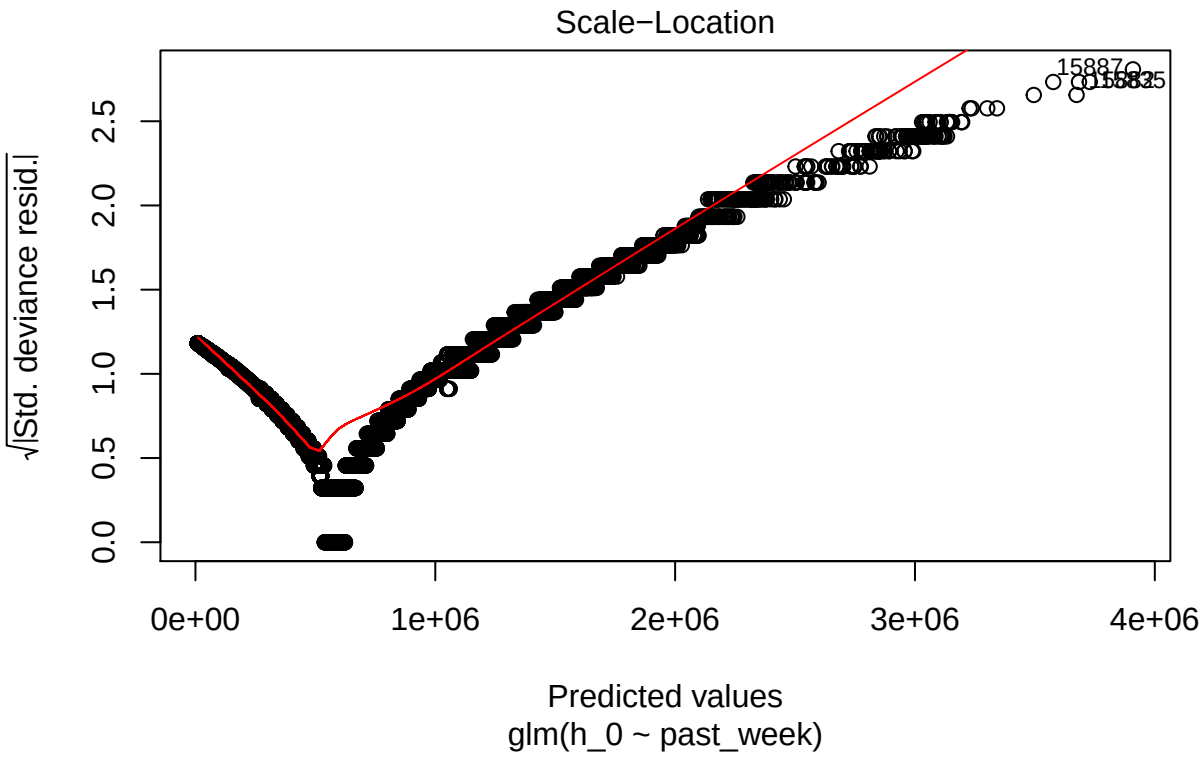
```
##
## Call:
## glm(formula = h_0 ~ past_week, data = pb)
```

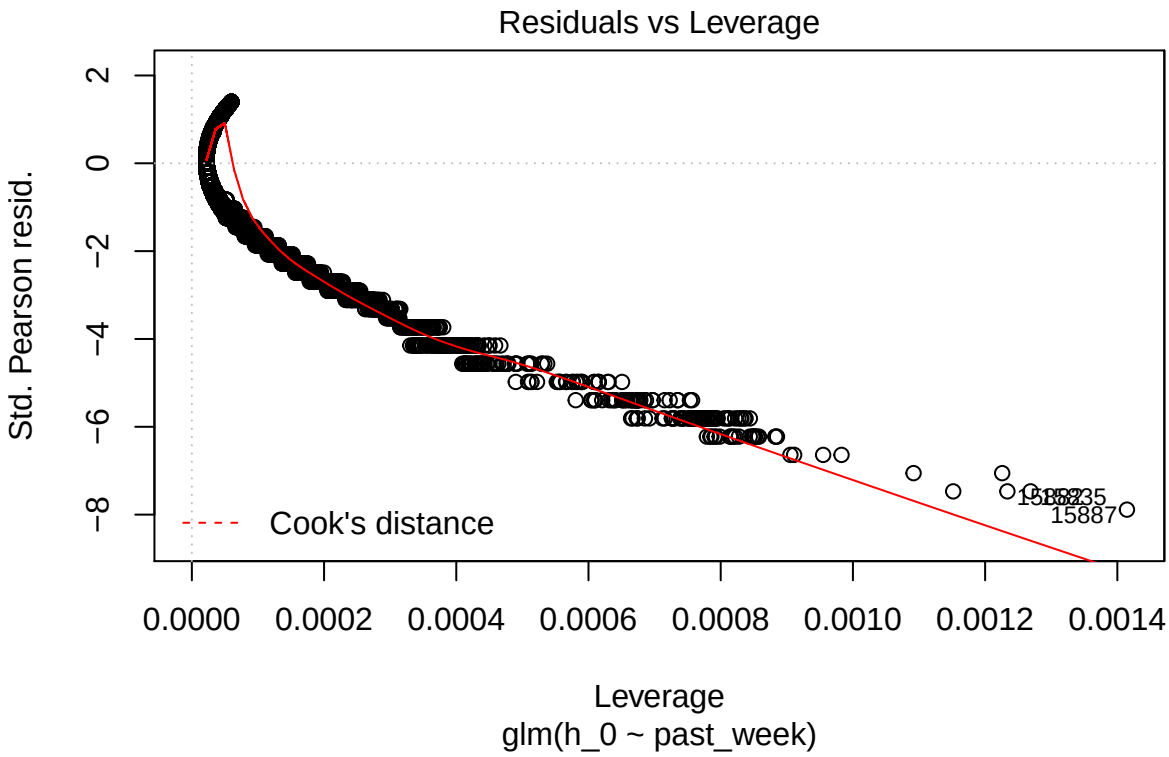
```
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -8.848e-09  -3.492e-10   3.492e-10   8.731e-10   1.568e-09
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 1.000e+04  8.684e-12 1.151e+15   <2e-16 ***
## past_week   9.647e-01  1.207e-17 7.989e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.260429e-18)
##
##     Null deviance: 8.0451e+15  on 45455  degrees of freedom
## Residual deviance: 5.7292e-14  on 45454  degrees of freedom
## AIC: -1744470
##
## Number of Fisher Scoring iterations: 1
```

```
plot(lm2)
```




Normal Q−Q
Std. deviance resid.
2
0
−2
−4
−6
−8
15887
−4
−2
0
2
4
Theoretical Quantiles
glm(h_0 ~ past_week)


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
0e+00
1e+06
2e+06
3e+06
4e+06
Predicted values
glm(h_0 ~ past_week)



(c) Use the fitted regression model to estimate the gross box office results for the current week (in $) for a production with $400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office results for the current week (in $) for a production with $400,000 in gross box office the previous week. Is $450,000 a feasible value for the gross box office results in the current week, for a production with $400,000 in gross box office the previous week? Give a reason to support your answer.

(d) Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this week's gross box office results. Comment on the appropriateness of this rule.

2.2

A story by James R. Hagerty entitled With Buyers Sidelined, Home Prices Slide published in the Thursday October 25, 2007 edition of the Wall Street Journal contained data on so-called fundamental housing indicators in major real estate markets across the US. The author argues that... prices are generally falling and overdue loan payments are piling up . Thus, we shall consider data presented in the article on Y = Percentage change in average price from July 2006 to July 2007 (based on the S&P/Case-Shiller national housing index); and x = Percentage of mortgage loans 30 days or more overdue in latest quarter (based on data from Equifax and Moody's). The data are available on the book web site in the file indicators.txt. Fit the following model to the data: $Y = \beta_0 + \beta_1 x + \epsilon$. Complete the following tasks:

(a) Find a 95% confidence interval for the slope of the regression model, b 1 . On the basis of this confidence interval decide whether there is evidence of a significant negative linear association.

(b) Use the fitted regression model to estimate $\epsilon(Y|X = 4)$. Find a 95% confidence interval for $\epsilon(Y|X = 4)$. Is 0% a feasible value for $\epsilon(Y|X = 4)$? Give a reason to support your answer.

Linear Models with R

2.4

The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with lpsa as the response and lcavol as the predictor. Record the residual standard error and the R^2 . Now add lweight, svi, lbph, age, lcp, pgg45 and gleason to the model one at a time. For each model record the residual standard error and the R 2 . Plot the trends in these two statistics.

2.5

Using the prostate data, plot lpsa against lcavol. Fit the regressions of lpsa on lcavol and lcavol on lpsa. Display both regression lines on the plot. At what point do the two lines intersect?