Regression Part 2

```r
# Read in data
> nfl = read.csv(file = "~/data/FCI-NFL-2014.csv")
> meta = read.csv(file = "~/data/NFL-Meta-Data.csv")

# Merge meta data into nfl data frame
> nfl = merge(nfl, meta, by = "team")

# Create age of stadium variable
> nfl$ageStadium = 2014 - nfl$yearOpened

# Create log of outcome
nfl$Lfci = log(nfl$fci)

# Create log coachYrswTeam
> nfl$LcoachYrswTeam = log(nfl$coachYrswTeam + 1)

# Fit regression model
> lm.a = lm(Lfci ~ ageStadium + I(ageStadium ^ 2) + LcoachYrswTeam, data = nfl)
```

Fortify the Model

```
> library(ggplot2)

> out.a = fortify(lm.a)

# Add row number
> out.a$ID = 1:32

> head(out.a)

      Lfci ageStadium I(ageStadium^2) LcoachYrswTeam       .hat     .sigma     .cooksd
1 6.077849          8              64      0.6931472 0.06496391 0.1344354 0.010445889
2 6.080688         22             484      1.9459101 0.07770947 0.1351090 0.006869880
3 6.304924         16             256      1.9459101 0.06550076 0.1342658 0.011748575
4 5.948166         41            1681      0.6931472 0.12144820 0.1355817 0.004566732
5 6.003344         18             324      1.3862944 0.04112109 0.1333362 0.011231506
6 6.391515         90            8100      0.6931472 0.92028011 0.1310104 5.713349583
   .fitted      .resid  .stdresid ID
1 6.177924 -0.10007515 -0.7754982  1
2 6.153881 -0.07319251 -0.5710860  2
3 6.199289  0.10563544  0.8188209  3
4 5.993637 -0.04547114 -0.3635138  4
5 6.137099 -0.13375552 -1.0235253  5
6 6.444532 -0.05301678 -1.4070150  6
```

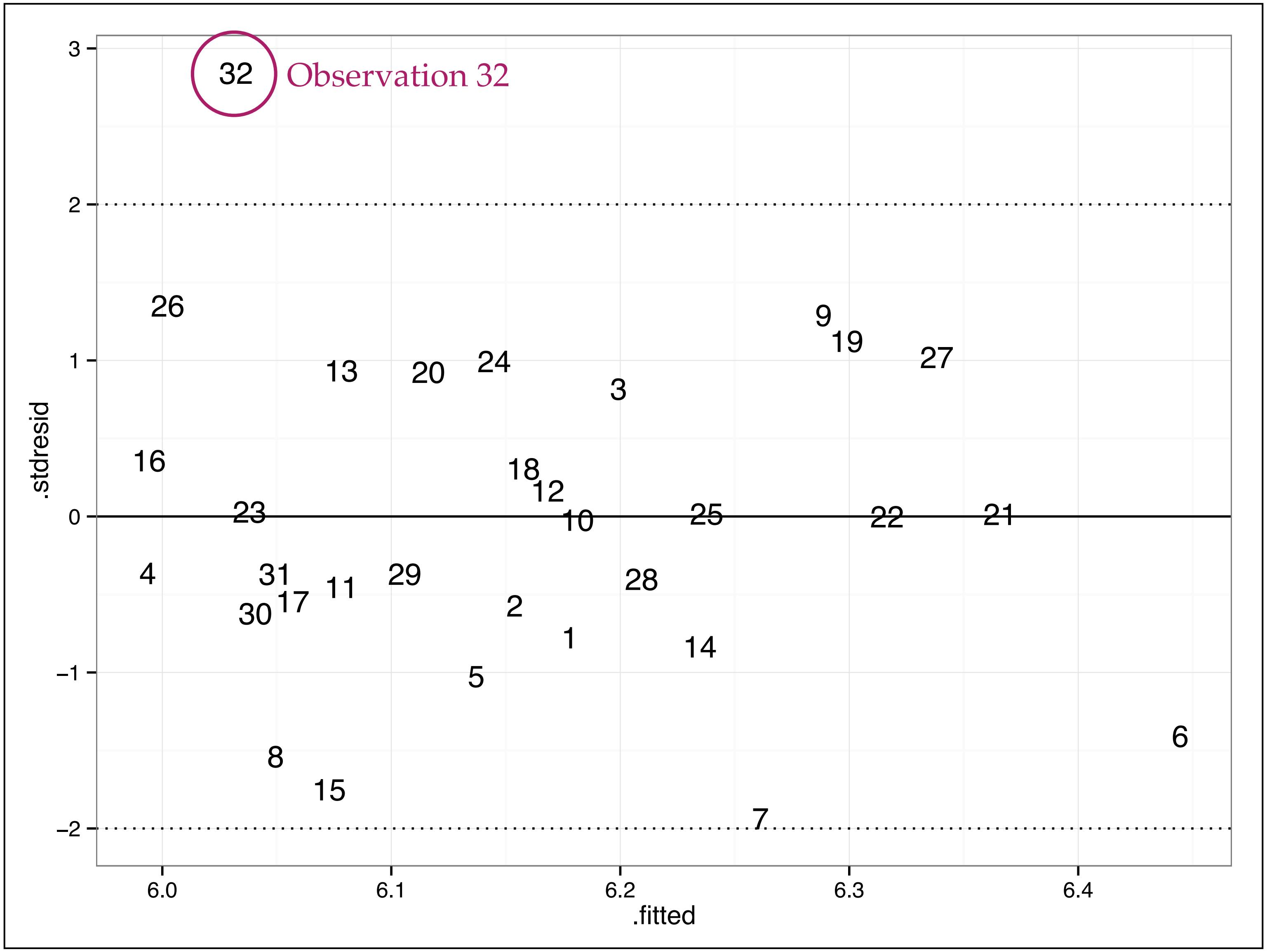
Outlying Cases

Cases are considered regression outliers if their studentized residual is more than two standard errors from the mean. For large samples ($n > 200$) they are considered regression outliers if their studentized residual is more than four standard errors from the mean.

Outliers have an unusually high or low Y-value relative to their predicted value.

```
# Which rows have .stdresid > 2
> out.a[out.a$.stdresid > 2, ]

       Lfci ageStadium I(ageStadium^2) LcoachYrswTeam      .hat    .sigma   .cooksd .fitted
32 6.392771         17             289              0 0.0973017 0.1146119 0.2178943 6.03222
     .resid .stdresid ID
32 0.360551  2.843571 32
```

Residual Plot



Cases with High Leverage Values

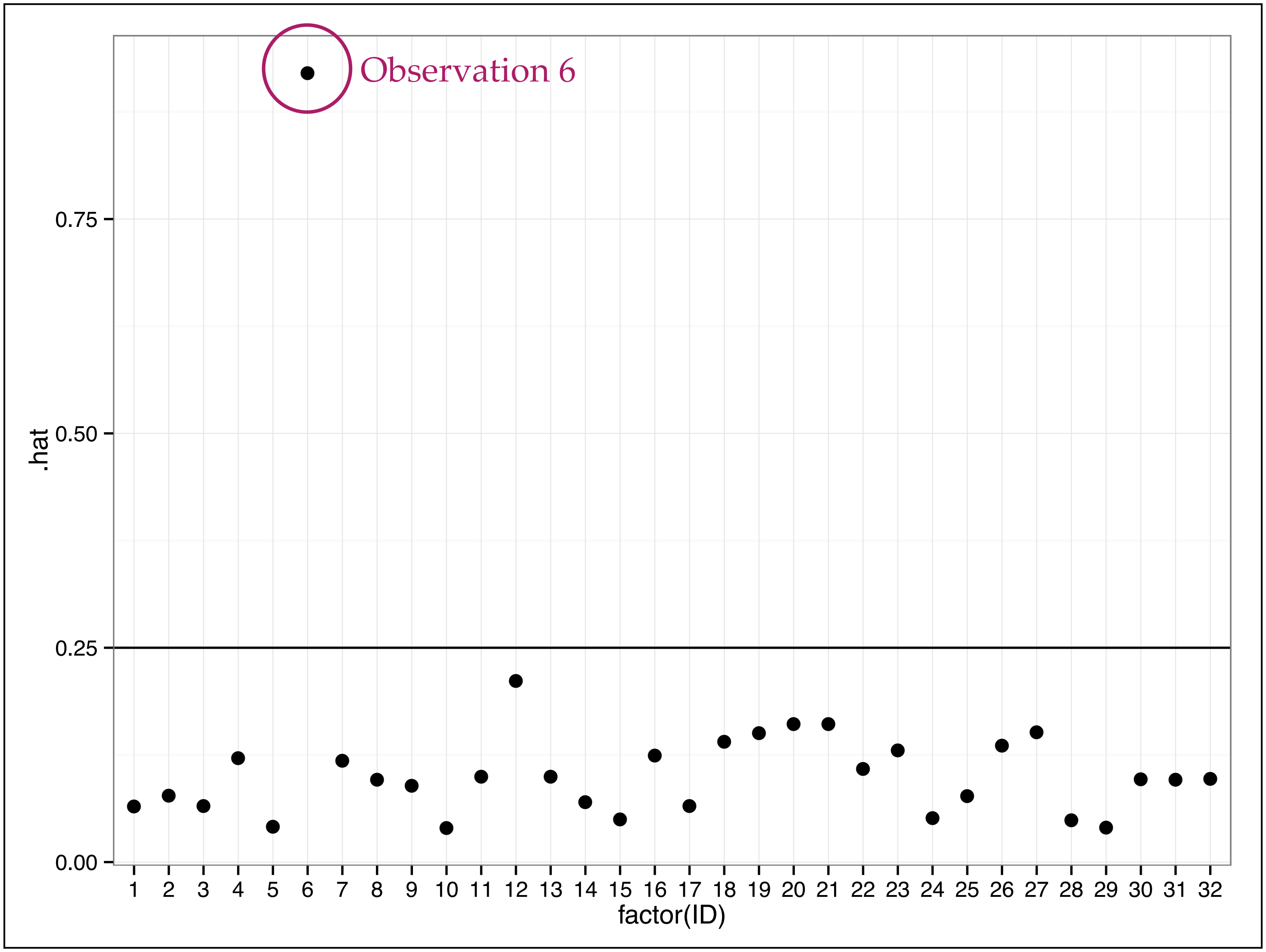
Cases have high leverage values if their hat value is greater than 2 x 4/32.

Leverage indicates an observations' distance from the centroid. Observations with high leverage have an unusual value (outlying?) in the *X*-space

Removing observations with high leverage will generally not have an effect on the regression coefficient estimates, but will affect the SEs of the coefficients and model summary measures (e.g., RMSE, R^2)

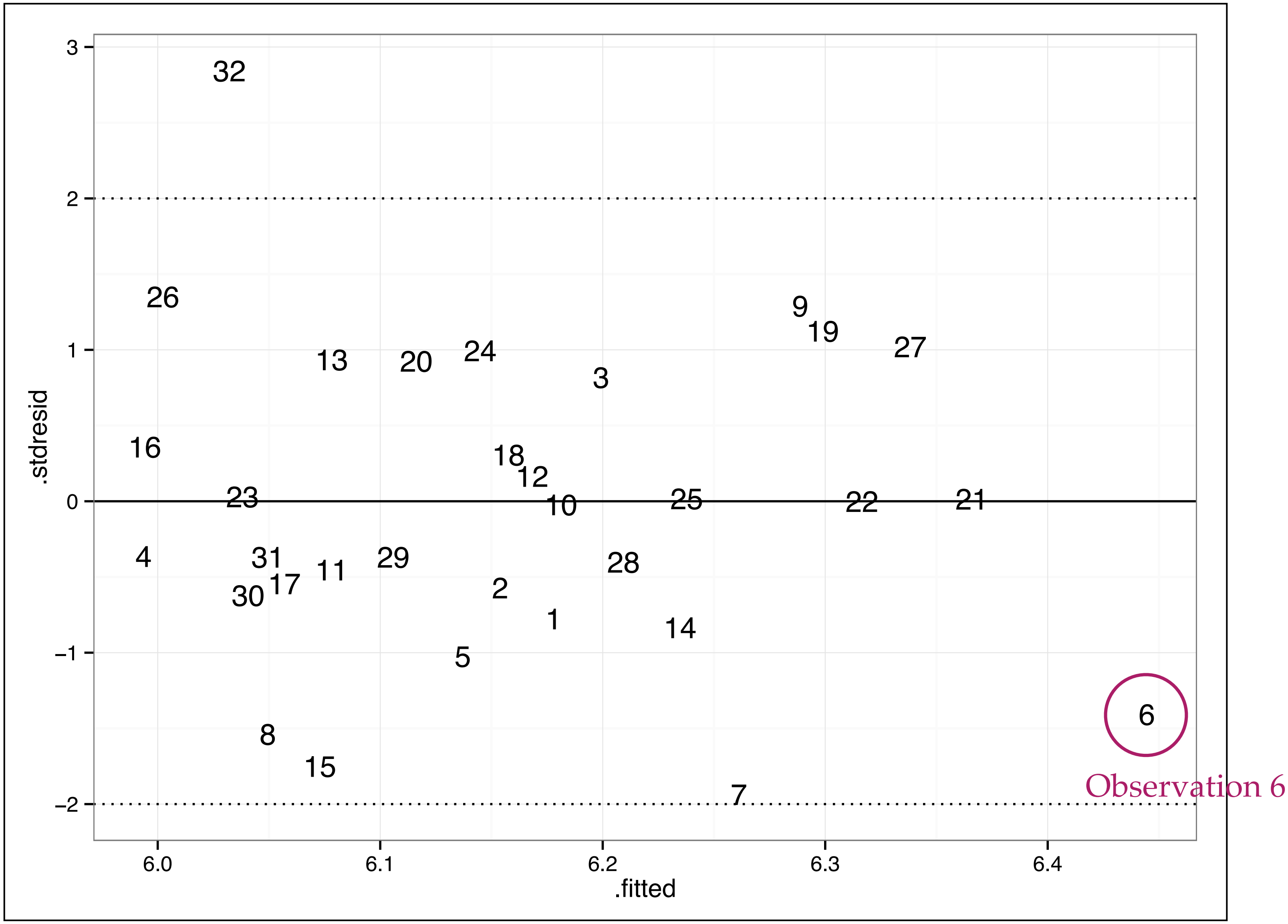
```
> bad.hat = 2 * 4 / 32
> bad.hat

[1] 0.25

# Which rows are greater than bad.hat
> out.a[out.a$.hat > bad.hat, ]

      Lfci ageStadium I(ageStadium^2) LcoachYrswTeam      .hat    .sigma .cooksd  .fitted
6 6.391515         90            8100      0.6931472 0.9202801 0.1310104 5.71335 6.444532
       .resid .stdresid ID
6 -0.05301678 -1.407015  6
```

Case Plot of Leverage



$$2 \times \frac{4}{32} = 0.25$$

Observation 6, which has a high leverage point, is unusually large relative to the other observations in the predictor-space.



Influential Observations

Influential observations are observations that unduly influence the regression coefficients. Removing these observations can have drastic impact on the size of the estimates. are outliers and have high

We quantify influence via Cook's Distance

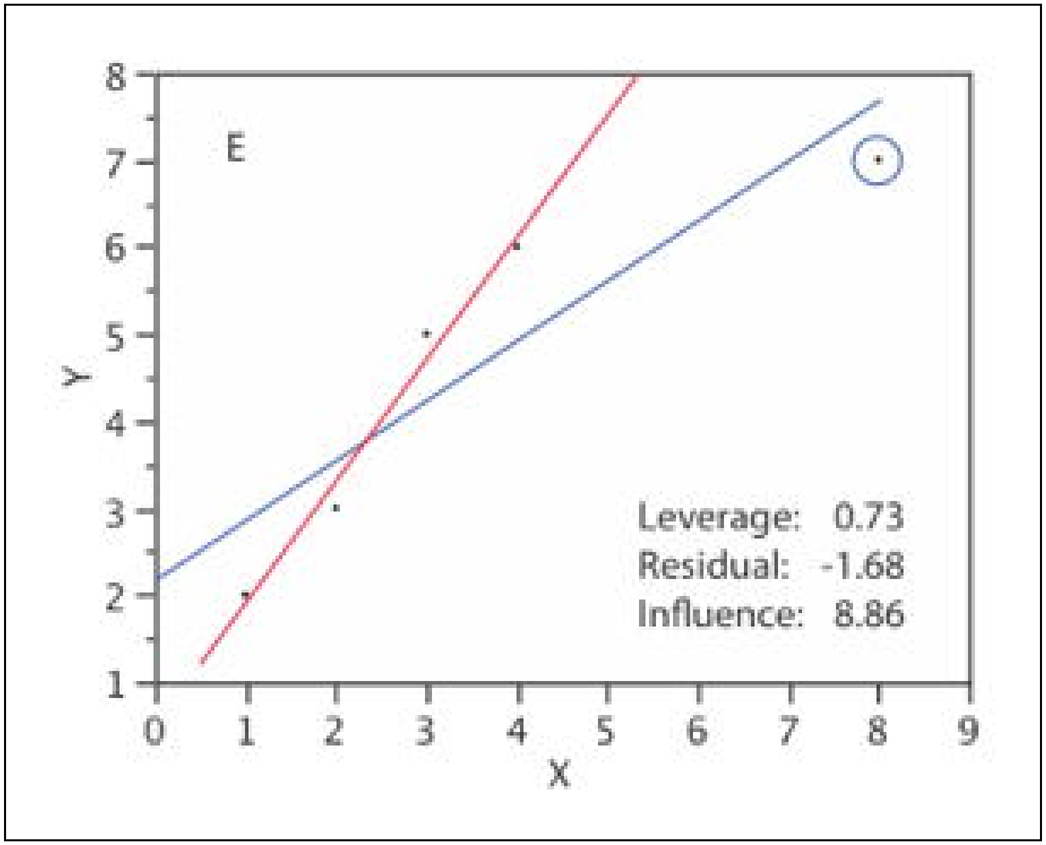
$$D_i = \frac{e_i^2}{p\ \mathrm{MSE}} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right]$$

where h_{ii} is the observation's leverage value and e_i is the observation's residual

Influence is the combination of "outlyingness" and leverage.

The following plots are from OnlineStatBook (http://onlinestatbook.com/2/regression/influential.html)

The blue line shows the regression line for the whole dataset.
The red line shows regression line if the circled observation is removed.

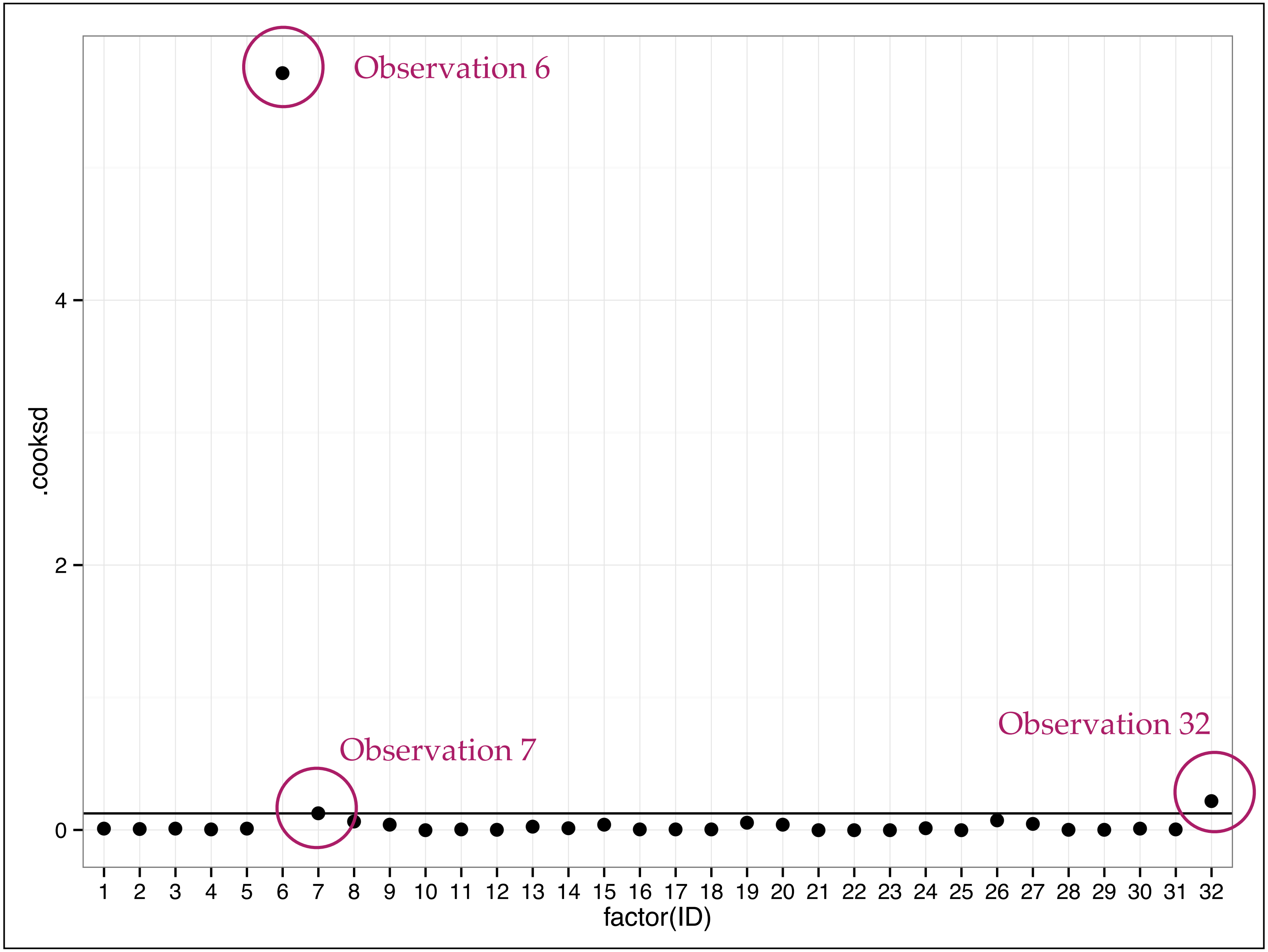


This observation is an influential observation.

Observations with a Cook's D value $> 4/n$ are considered influential.

```
> bad.d = 4 / 32
> out.a[out.a$.cooksd > bad.d, ]


       Lfci ageStadium I(ageStadium^2) LcoachYrswTeam      .hat    .sigma   .cooksd  .fitted
6  6.391515         90            8100      0.6931472 0.9202801 0.1310104 5.7133496 6.444532
7  6.018691         14             196      2.4849066 0.1181983 0.1264856 0.1255299 6.261240
32 6.392771         17             289      0.0000000 0.0973017 0.1146119 0.2178943 6.032220
        .resid .stdresid ID
6  -0.05301678 -1.407015  6
7  -0.24254940 -1.935457  7
32  0.36055095  2.843571 32
```

Observations 6, 7, and 32 appear to be influential, however only observation 6 seems really bad.



Remove Observation 6

```
> lm.a2 = update(lm.a, subset = -c(6))
> display(lm.a2)

                coef.est coef.se
(Intercept)      6.30     0.08
ageStadium      -0.02     0.01
I(ageStadium^2)  0.00     0.00
LcoachYrswTeam   0.07     0.03
---
n = 31, k = 4
residual sd = 0.13, R-Squared = 0.47
```

```
# Results from all of the observations

                coef.est coef.se
(Intercept)      6.23     0.06
ageStadium      -0.01     0.00
I(ageStadium^2)  0.00     0.00
LcoachYrswTeam   0.08     0.03
---
n = 32, k = 4
residual sd = 0.13, R-Squared = 0.46
```

Uncertainty in the Regression Coefficients

```
> mySim = sim(lm.a2, n.sim = 1000)

# Get the simulated 95% CIs for the parameters
> apply(coef(mySim), MARGIN = 2, FUN = quantile, probs = c(0.025, 0.975))

          [,1]         [,2]         [,3]       [,4]
2.5%  6.138610 -0.035974814 0.0001093404 0.01259112
97.5% 6.454682 -0.009001952 0.0005769752 0.12804002
```

apply(*data frame/matrix*, MARGIN = *1 or 2*, FUN = *some function*, ...)

Predicting for the Vikings New Stadium

```
> myData = data.frame(
  ageStadium = 0,
  LcoachYrswTeam =
  )

# Prediction from the fitted model
> predict(lm.a2, newdata = myData, interval = "prediction")

       fit      lwr      upr
1 6.377683 6.075772 6.679595
```

Predicting for the Vikings New Stadium for the First Set of Simulated Coefficients

```
# Get first set of simulated coefficients
> coef(mySim)[1, ]
[1]  6.2059341384 -0.0178104690  0.0002221918  0.1299631865

# Multiply simulated coefficients by vector of x (bx)
> coef(mySim)[1, ] * c(1, 0, 0, 1.098612)

[1] 6.2059341 0.0000000 0.0000000 0.1427791

# Sum the bx values
> sum(coef(mySim)[1, ] * c(1, 0, 0, 1.098612))

[1] 6.348713
```

Write a Function to compute the Sum of bx

```
myFunction = function(x){
    sum(x * c(1, 0, 0, 1.098612))
}
```

```
# Try the function
> myFunction( coef(mySim)[1, ] )

[1] 6.348713
```

Use the Function to Compute the Y-hats

```
> preds = apply(coef(mySim), MARGIN = 1, FUN = myFunction)

> exp(quantile(preds, probs = c(0.025, 0.975)))

    2.5%    97.5% 
513.7383 677.7805 
```

Accounting for **model uncertainty** using a simulation, the predicted FCI for a Vikings game in 2016 will be between $514 and $678.

Prediction Uncertainty for the Vikings New Stadium for the First Set of Simulated Coefficients

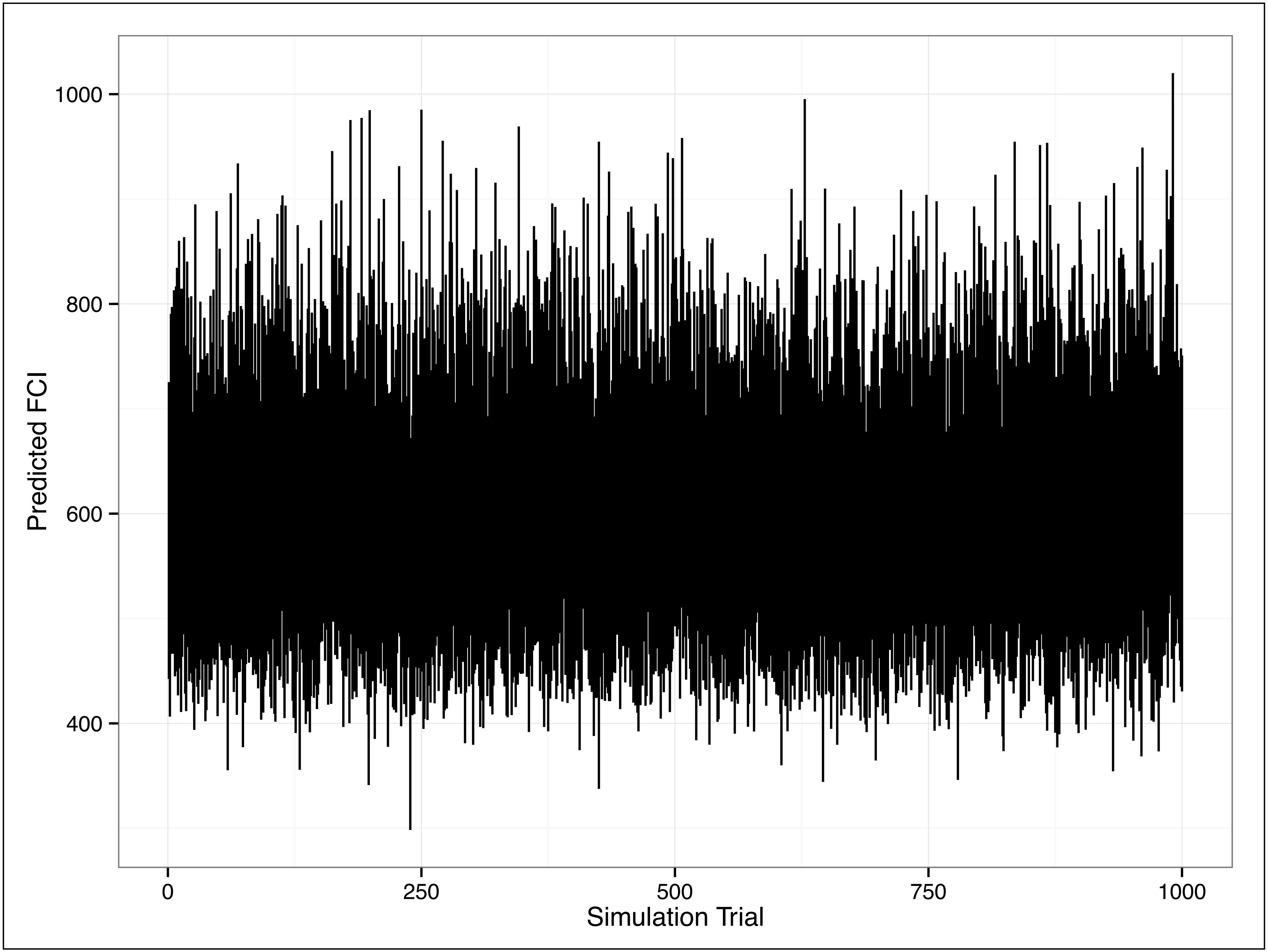
```
# Predicted value
> sum(coef(mySim)[1, ] * c(1, 0, 0, 1.098612))

[1] 6.348713

# Residual standard error
> mySim@sigma[1]
[1] 0.1316461

> sim1 = rnorm(1000, mean = 6.348713, sd = 0.1316461)
> exp(quantile(sim1, probs = c(0.025, 0.975)))

    2.5%    97.5%
440.9169 736.3640
```

Accounting for **prediction uncertainty** using a simulation, the predicted FCI for a Vikings game in 2016 will be between $441 and $736.

Prediction Uncertainty for the Vikings New Stadium for All Sets of Simulated Coefficients

```
> lower = rep(NA, 1000)
> upper = rep(NA, 1000)

> for(i in 1:1000){
    sim.i = rnorm(1000, mean = preds[i], sd = mySim@sigma[i])
    q.i = quantile(sim.i, probs = c(0.025, 0.975))
    lower[i] = q.i[[1]]
    upper[i] = q.i[[2]]
}

> exp(quantile(lower, probs = 0.025))
    2.5%
387.9941

> exp(quantile(upper, probs = 0.975))
   97.5%
915.7992
```

Accounting for **model uncertainty** and **prediction uncertainty** using a simulation, the predicted FCI for a Vikings game in 2016 will be between $388 and $916.


Predicted FCI
1000
800
600
400
0
250
500
750
1000
Simulation Trial

Plot of the Model and Predicted Uncertainty

```
> new = data.frame(
    sim = 1:1000,
    yhat = exp(preds),
    ll = exp(lower),
    ul = exp(upper)
    )

> ggplot(data = new, aes(x = sim, xend = sim, y = ll, yend = ul)) +
    geom_segment() +
    theme_bw() +
    xlab("Simulation Trial") +
    ylab("Predicted FCI")
```