Lecture 4:
Inference in SLR (continued)
Diagnostic approaches in SLR

BMTRY 701
Biostatistical Methods II

A little more in inference of β's

- Confidence interval for β_1
- This follows easily from discussion of t-test
- Recall sampling distribution for slope:

$$\boxed{\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\hat{\beta}_1))}$$

- From this, the 95% CI follows:

$$\boxed{\hat{\beta}_1 \pm t_{0.975,n-2}\hat{\sigma}(\hat{\beta}_1)}$$

Confidence Intervals

- More generally,

$$\boxed{\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\,\hat{\sigma}(\hat{\beta}_1)}$$

- And, the same approach is used for the intercept (if you would care to):

$$\boxed{\hat{\beta}_0 \pm t_{1-\alpha/2,n-2}\,\hat{\sigma}(\hat{\beta}_0)}$$

SENIC data

```
> reg <- lm(data$LOS~ data$BEDS)
> summary(reg)$coefficients
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 8.625364302 0.272058856 31.704038 1.851535e-57
data$BEDS   0.004056636 0.000858405  4.725782 6.765452e-06
```

```
> qt(0.975,111)
[1] 1.981567
```

95% CI for β_1:

0.00406 +/- 1.98*0.000858 = {0.00236, 0.00576}

More meaningful:

- what about the difference in LOS for a 100 bed difference between hospitals?

- Go back to sampling distribution:

$$\boxed{\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\hat{\beta}_1))}$$

- for a 100 unit difference:

$$\boxed{100\hat{\beta}_1 \sim N(100\beta_1, (100^2)\sigma^2(\hat{\beta}_1))}$$

More meaningful:

- So that implies that the CI takes the form

$$\boxed{100\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\{100\hat{\sigma}(\hat{\beta}_1)\}}$$

- Hence, simply multiply the 95% CI limits by 100:

95% CI for 100*β1:

100* 0.00406 +/- 1.98*100*0.000858 = {0.236, 0.576}

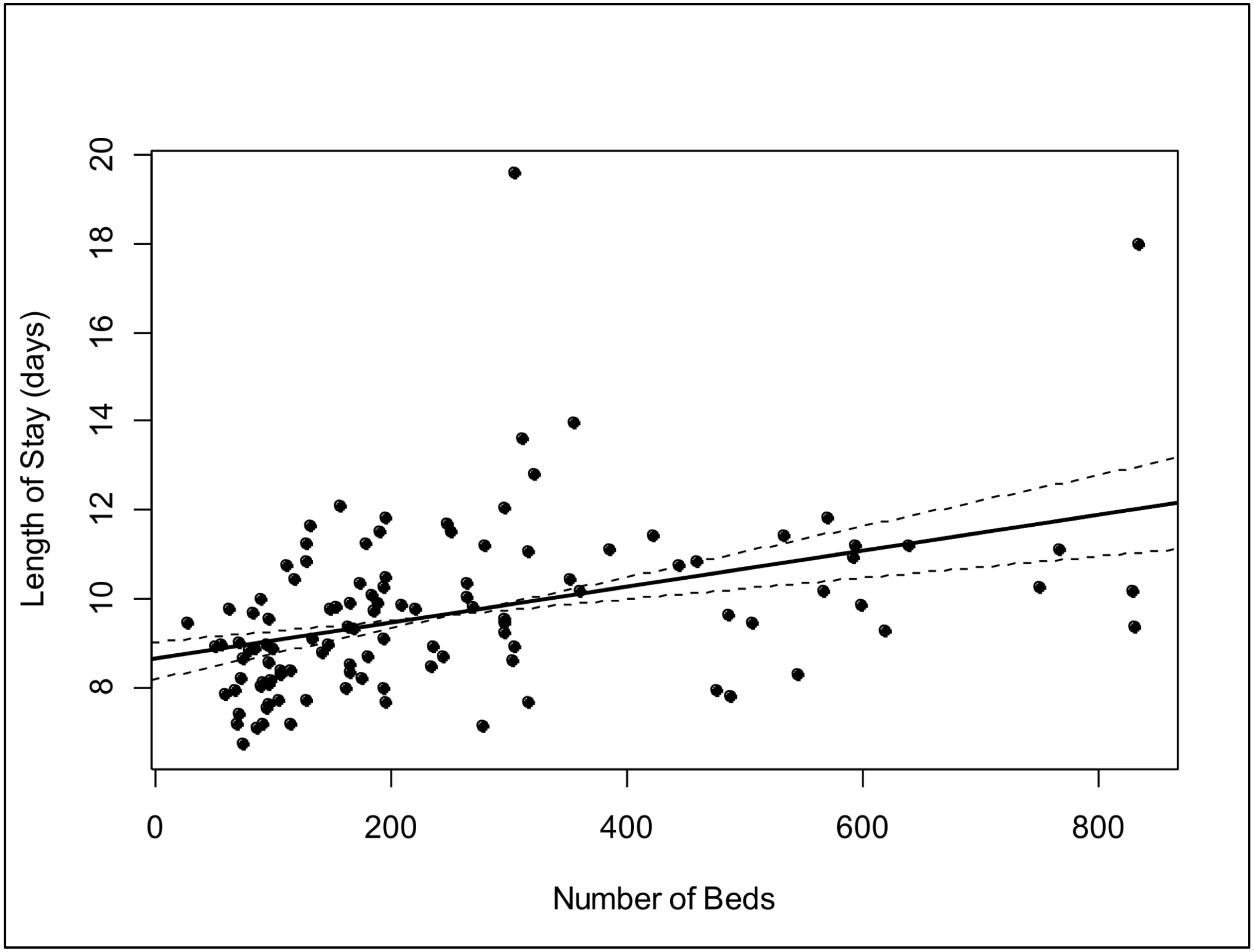
What would this look like?

- Recall that the regression line always goes through the means of X and Y.

- We can add our 95% CI limits of the slope to our scatterplot by using the knowledge that the regression line will go through the means of x and y.

```
> mean(data$LOS)
[1] 9.648319
> mean(data$BEDS)
[1] 252.1681
# use these as x and y values.  then, use each
# of the slopes to find corresponding intercepts
> abline(8.198, 0.00576, lty=2)
> abline(9.055, 0.00236, lty=2)
```

SENIC data: 95% CI for slope



Important implication

- The slope and intercept are NOT independent
- Notice what happens to the intercept if we increase the slope?
- What happens if we decrease the slope?

```
> attributes(summary(reg))
$names
 [1] "call"          "terms"          "residuals"     "coefficients"
 [5] "aliased"       "sigma"          "df"            "r.squared"
 [9] "adj.r.squared" "fstatistic"     "cov.unscaled"

$class
[1] "summary.lm"
```

```
> summary(reg)$cov.unscaled
               (Intercept)     data$BEDS
(Intercept)  2.411664e-02 -6.054327e-05
data$BEDS   -6.054327e-05  2.400909e-07
```

A few comments r.e. inferences

- We assume Y|X ~ Normal
- if this is “seriously” violated, our inferences may not be valid.
- But, no surprise, a large sample size will save us
- Slope and intercept sampling distributions are ***asymptotically normal***

Spread of the X’s

- Recall the estimate of the standard error for the slope:

$$\boxed{\hat{\sigma}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \overline{X})^2}}}$$

- What happens to the standard error when the spread of the X’s is narrow?
- What happens to the standard error when the spread of the X’s is wide?
- (Note: intercept is similarly susceptible)

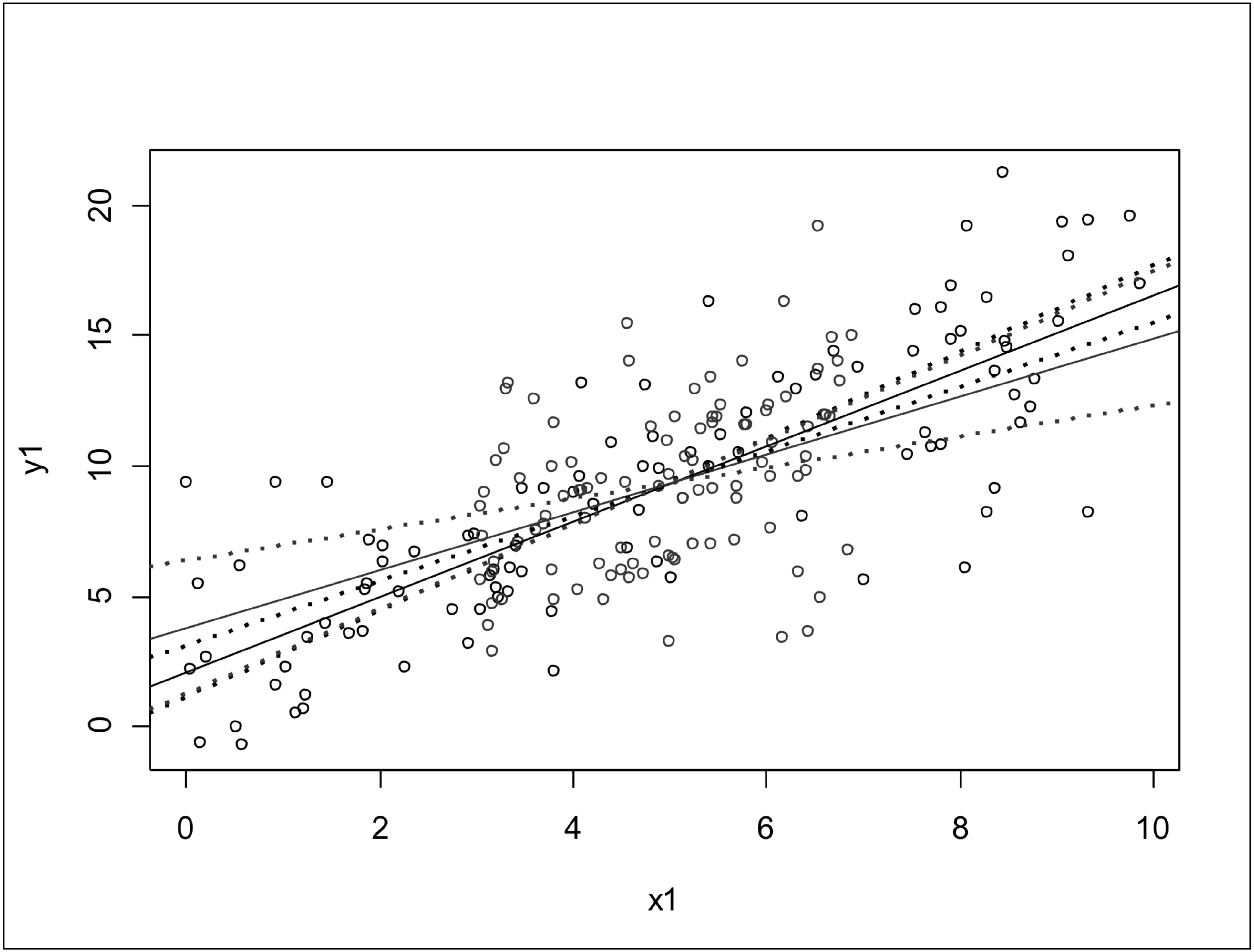
R code

```
##################
# simulate data
x1 <- runif(100,0,10)
x2 <- runif(100,3,7)
e <- rnorm(100,0,3)

y1 <- 2 + 1.5*x1 + e
y2 <- 2 + 1.5*x2 + e
plot(x1, y1)
points(x2, y2, col=2)

# fit regression models
reg1 <- lm(y1 ~ x1)
reg2 <- lm(y2 ~ x2)
abline(reg1)
abline(reg2, col=2)

# compare standard errors
summary(reg1)$coefficients
summary(reg2)$coefficients
```


y1
0
5
10
15
20
x1
0
2
4
6
8
10

Interval Estimation of Y’s

- Recall the model:

$$\boxed{E(Y) = \beta_0 + \beta_1 X}$$

- We might be interested in the mean value for a given value of X.
- This means, for example, “What is true mean LOS when number of beds is 400?”
- It does NOT mean “What is value of LOS when number of beds is 400?”

Mean versus individual

- Keep it straight: can be confusing.
- Using previous results,

$$\hat{Y}_j \sim N(E(Y_j), \sigma^2(\hat{Y}_j))$$

$$where$$

$$\sigma^2(\hat{Y}_j) = \sigma^2 \left[\frac{1}{n} + \frac{(X_j - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right]$$

- We call this the sampling distribution of Yhat.

Interval estimation

- Normality: follows from residuals, slope, and intercept being normal.
- Mean: easily shown by substituting in slope and intercept
- Variance: a little more detail
  - variability depends on distance of X from mean of X
  - Recall plots of 95% CIs
  - variation in slope has greater impact at extremes of X than in the middle
  - We substitute our estimate of MSE and then we have a t-distribution

Example:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.6253643  0.2720589  31.704  < 2e-16 ***
data$BEDS   0.0040566  0.0008584   4.726 6.77e-06 ***
---
Residual standard error: 1.752 on 111 degrees of freedom

> mean(data$BEDS)
[1] 252.1681
> sum( (data$BEDS - mean(data$BEDS))^2)
[1] 4165090
```

$$\boxed{\hat{\sigma}^2(\hat{Y}_j) = 1.752^2\left[\frac{1}{113} + \frac{(400-252.2)^2}{4165090}\right] = 0.0433}$$

Interval estimation

- Use our standard confidence interval approach:

$$\boxed{\hat{Y}_j \pm t_{1-\alpha/2, n-2}\,\hat{\sigma}(\hat{Y}_j)}$$

- Note that what differs is the way the standard error is calculated.
- Otherwise, all of the these tests and intervals follow the same pattern.

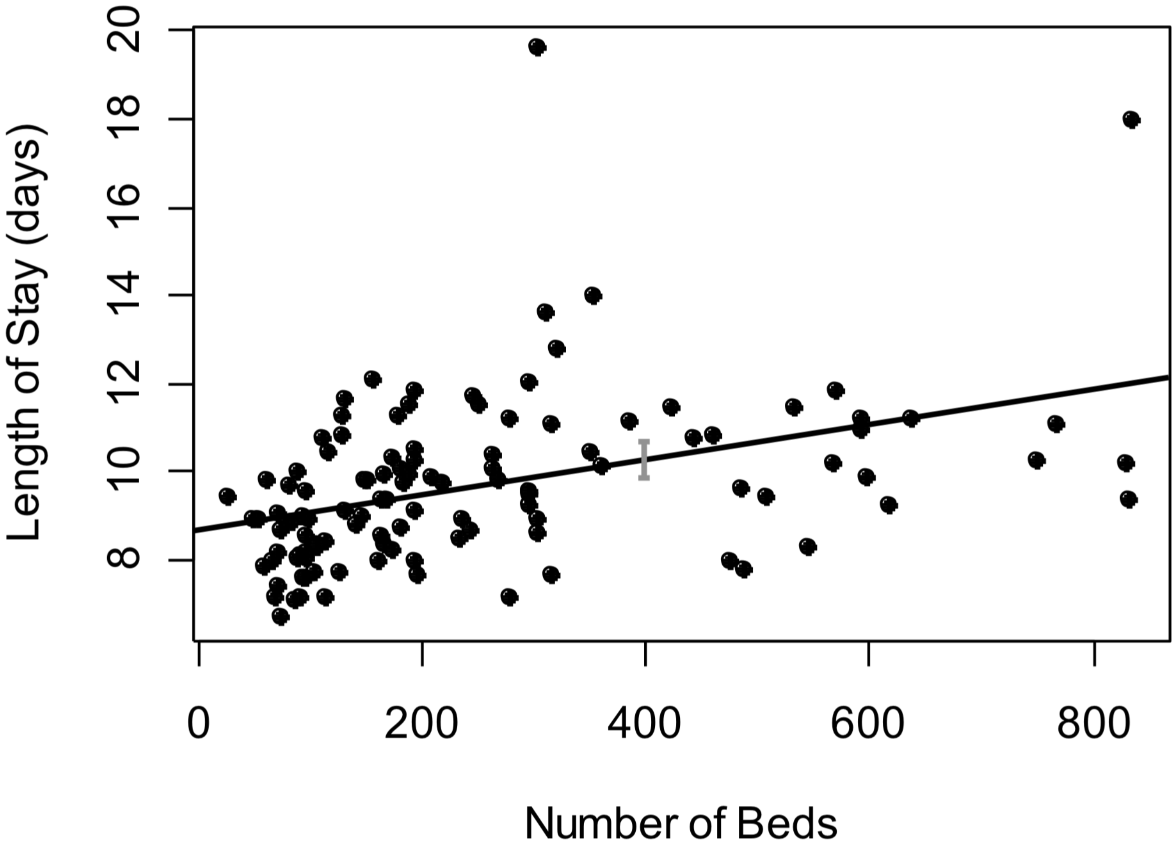
Example:

$$\hat{Y}_j = 8.63 + 0.00406 * 400 = 10.254$$

$$\hat{\sigma}(\hat{Y}_j) = \sqrt{0.0433} = 0.208$$

$95\% CI:$

$$\hat{Y}_j \pm 1.98 * \hat{\sigma}(\hat{Y}_j) = \{9.842, 10.67\}$$



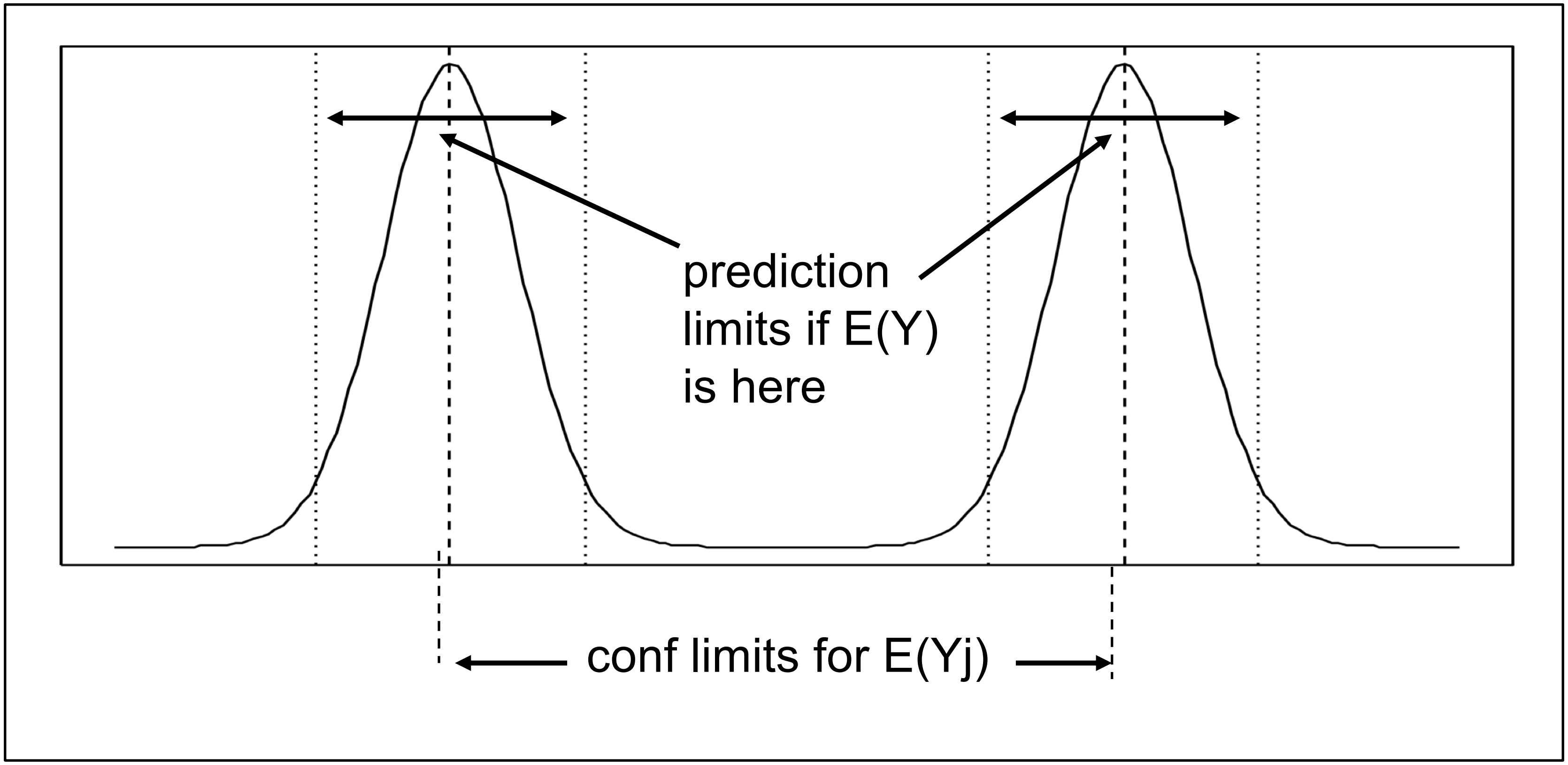
Prediction

- We’d like to know what to expect for NEW observations
- Example: if we added another hospital to our dataset with 400 beds, what is the likely observed mean LOS for that hospital?
- “Prediction interval”
- Intuition:
    - we are making inference about an individual hospital, not the mean of all hospitals
    - it should be wider than the confidence interval for the mean of Y|X

Prediction

- Can be divided into two types of uncertainty

1. Variation in the location of the the distribution of Y|X
2. Variation within the probability distribution of Y.



Added variability in prediction intervals

- Variance of a given Y value, given X is:

- Variance of the sampling distribution of Yhat is:

- So, σ^2(prediction) =

Prediction interval

- Based on the estimate of the variance of the residuals

$$\hat{Y}_j \pm t_{1-\alpha/2, n-2}\,\hat{\sigma}(pred)$$

$$where$$

$$\hat{\sigma}^2(pred) = \hat{\sigma}^2 + \hat{\sigma}^2(\hat{Y}_j)$$

$$= \hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(X_j - \bar{X})^2}{\sum_{i=1}^{n}(X_j - \bar{X})^2}\right]$$

Revisit our example

$$\hat{\sigma}^2(\hat{Y}_j) = 1.752^2\left[\frac{1}{113} + \frac{(400-252.2)^2}{4165090}\right] = 0.0433$$

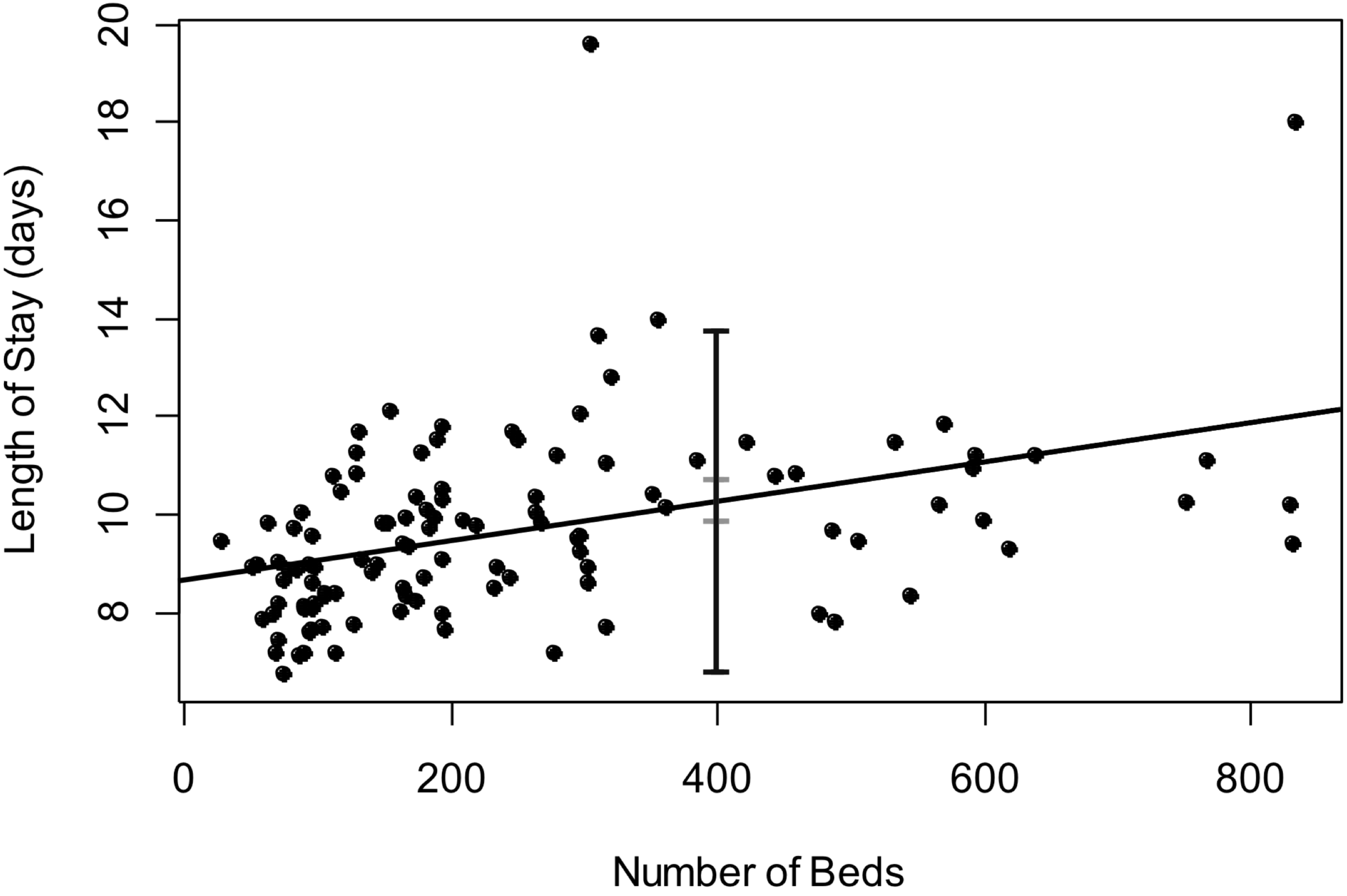
$$\hat{\sigma}^2(pred) = \hat{\sigma}^2 + \hat{\sigma}^2(\hat{Y}_j) = 1.752^2 + 0.0433 = 3.11$$

$$\hat{Y}_j = 8.63 + 0.00406 * 400 = 10.254$$

$$\hat{\sigma}(pred) = \sqrt{3.11} = 1.76$$

95% prediction interval:

$$\hat{Y}_j \pm 1.98 * \hat{\sigma}(pred) = \{6.77, 10.74\}$$



Length of Stay (days)
20
18
16
14
12
10
8
0
200
400
600
800
Number of Beds

Nearer to the mean?

- What about 250 bed hospital?

$$\hat{\sigma}^2(\hat{Y}_j) = 1.752^2\left[\tfrac{1}{113} + \tfrac{(250-252.2)^2}{4165090}\right] = 0.0272$$

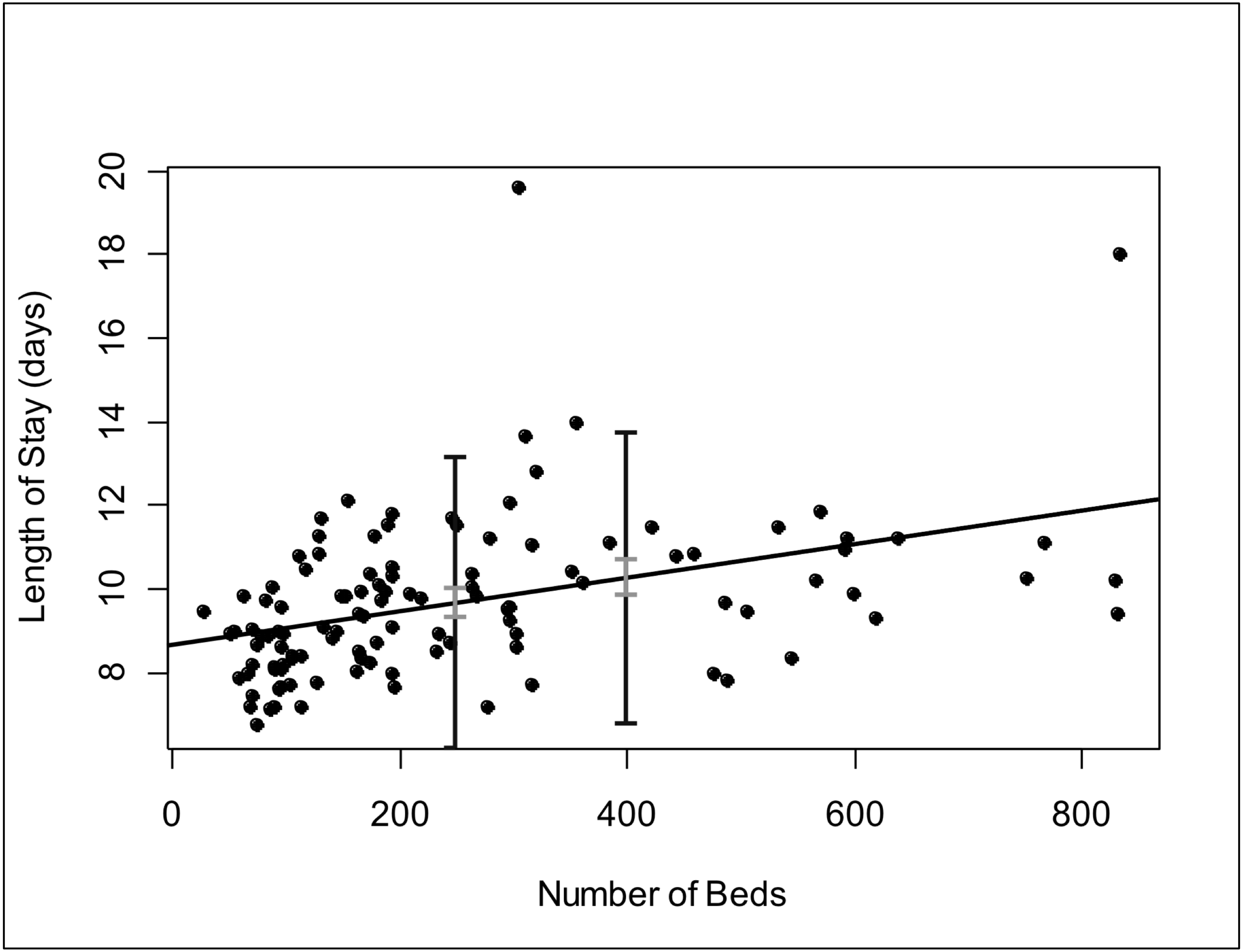
$$\hat{\sigma}^2(pred) = \hat{\sigma}^2 + \hat{\sigma}^2(\hat{Y}_j) = 1.752^2 + 0.0272 = 3.10$$

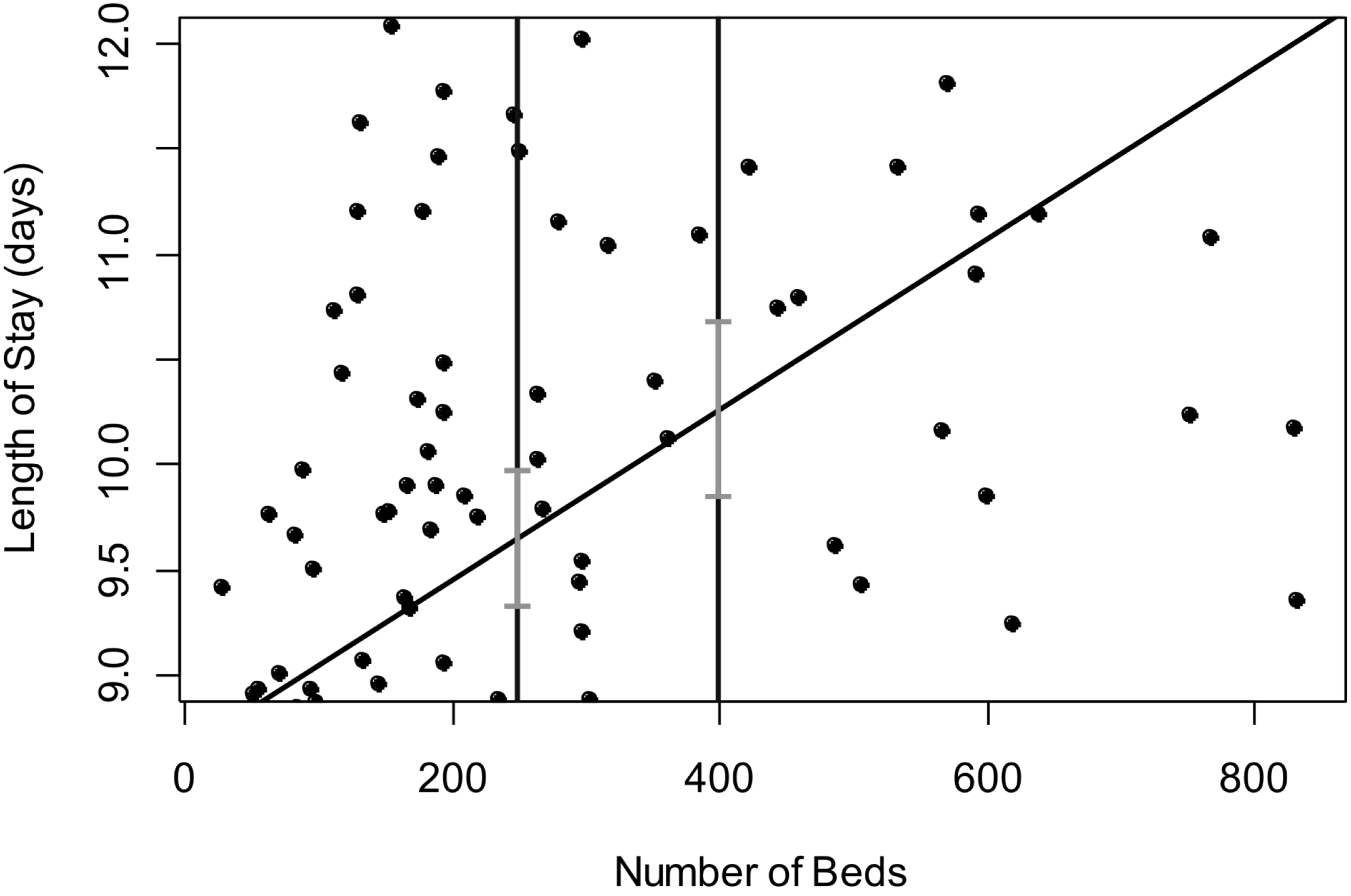
$$\hat{Y}_j = 8.63 + 0.00406 * 250 = 9.645$$

$$\hat{\sigma}(pred) = \sqrt{3.10} = 1.76$$

95% prediction interval:

$$\hat{Y}_j \pm 1.98 * \hat{\sigma}(pred) = \{6.16, 13.13\}$$


Length of Stay (days)
20
18
16
14
12
10
8
0
200
400
600
800
Number of Beds


Length of Stay (days)
12.0
11.0
10.0
9.5
9.0
0
200
400
600
800
Number of Beds

Diagnostics

- We made some assumptions
- Most relate to the residuals
- It is important to check them when possible.

- Recall:
  - residuals are normal
  - variance of residuals is constant over the range of X
  - residuals are independent of one another

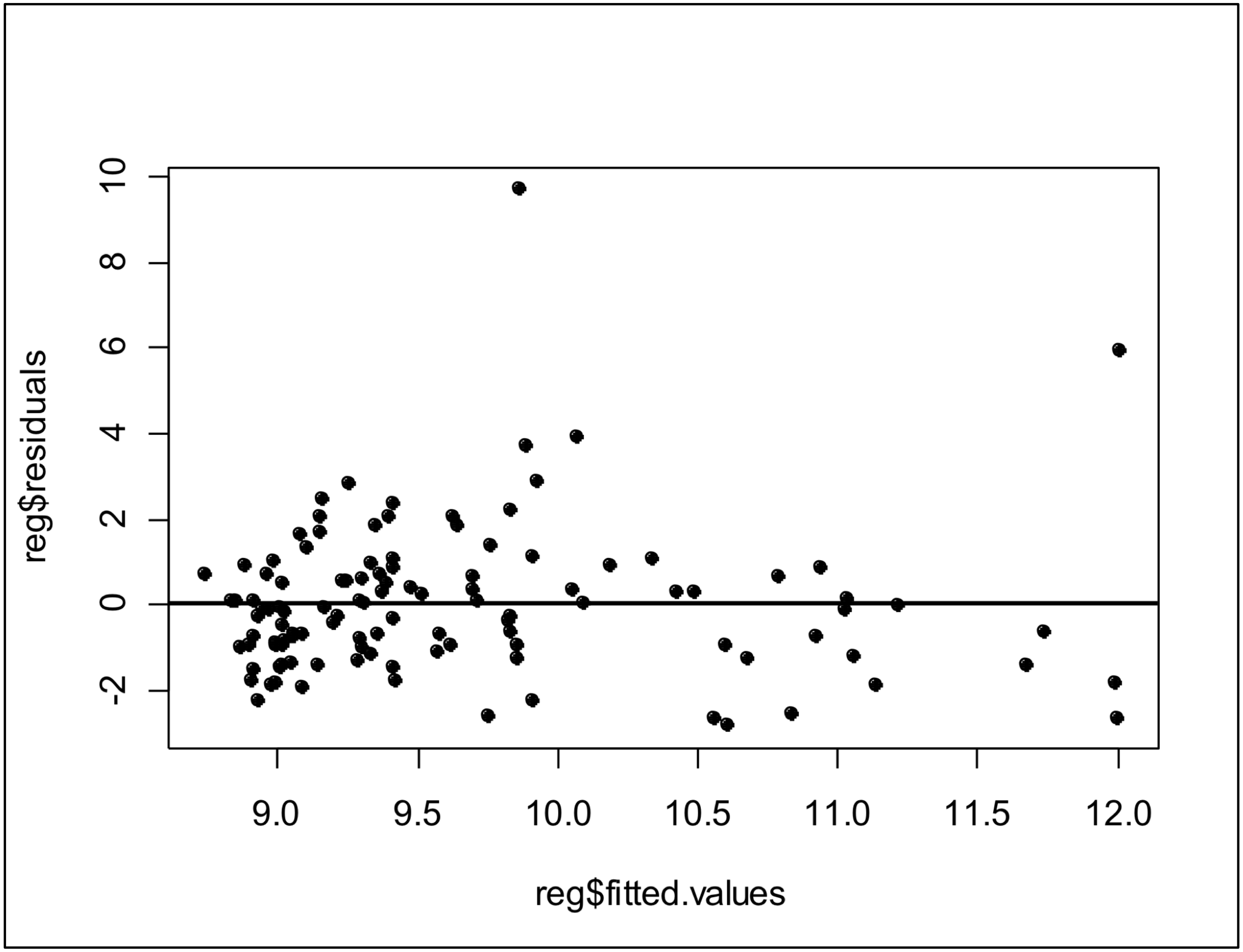
Diagnostic Considerations via Residuals

- The residuals are not normally distributed
- The residuals do not have constant variance
- The model fits all but one or a few outliers
- The regression function is not linear
- The residuals are not independent
- One or more predictors have been omitted from the model

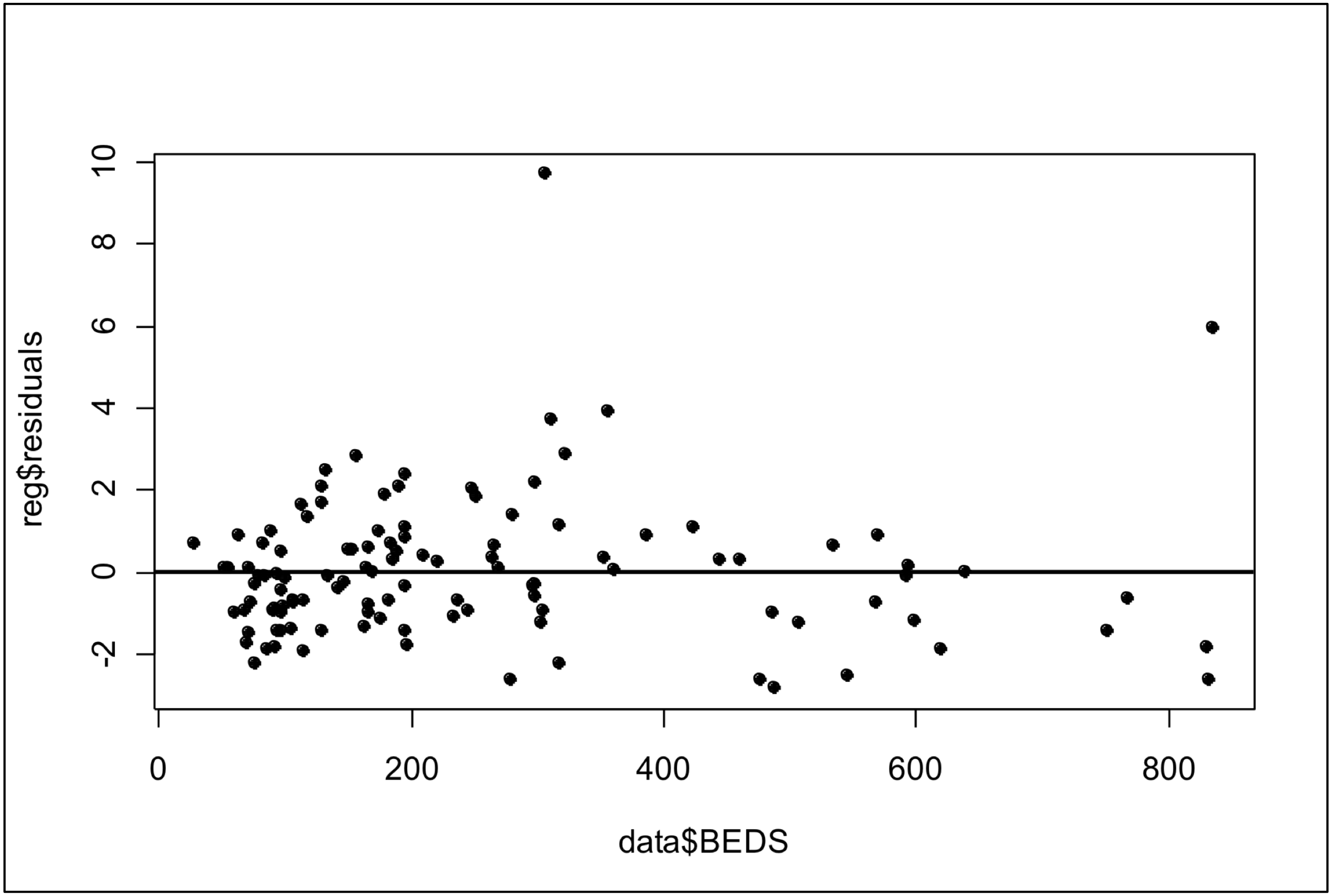
Several flavors of residual plots

- Residuals (y-axis) vs. Fitted values (x-axis)
- Residuals (y-axis) vs. Covariate (x-axis)
- Squared residuals (y-axis) vs. covariate (x-axis)
- Residuals vs. time
- Residuals vs. omitted predictor (MLR)
- Boxplot of residuals
- Normality probability plot of residuals

Classic diagnostic tool: residual plot
What can you see from here?

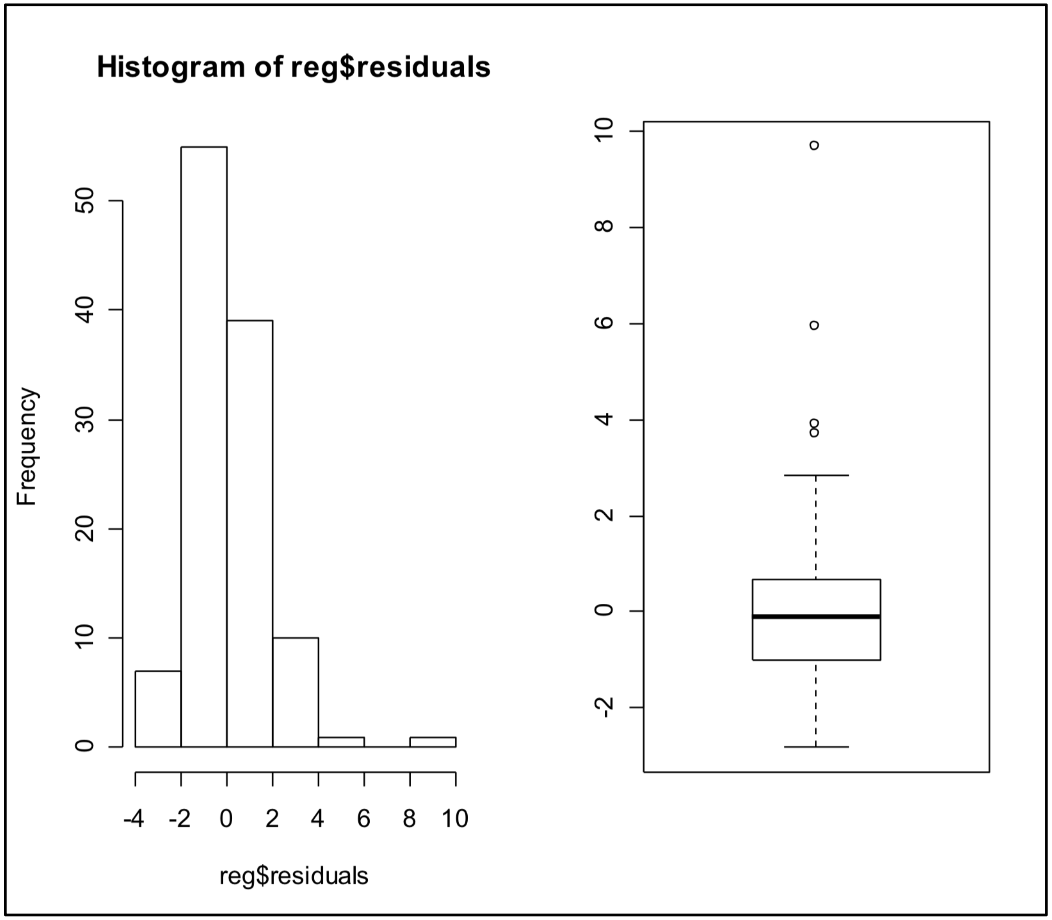


Residuals vs. X



Normality of Residuals

- How to check?
- Look at their distribution!

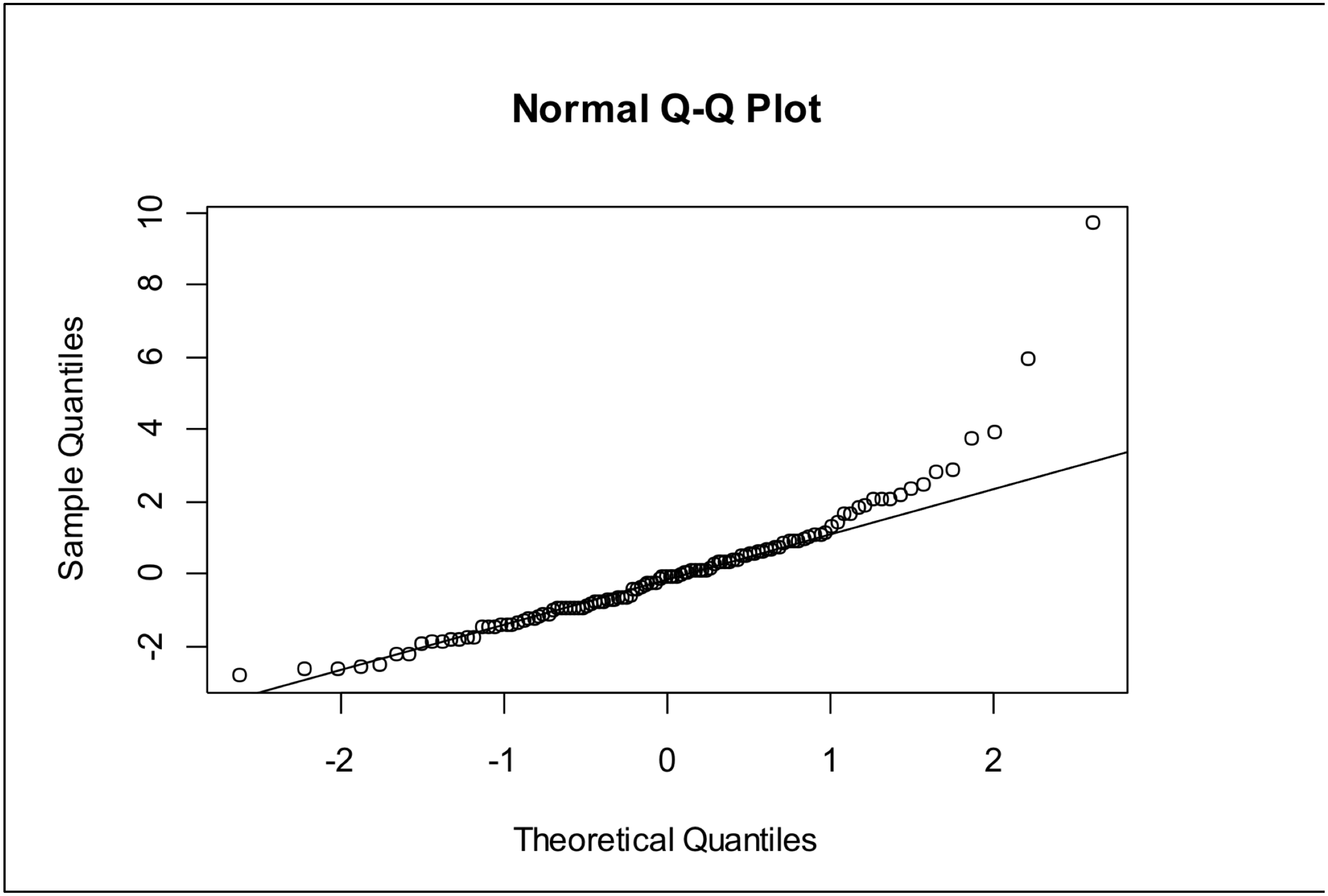


Normal Probability Plots & QQ-Plots

- Small departures from normality (of residuals) is not a big deal
- But, if the residuals have heavy tails than a normal distribution, that implies outliers
- Diagnostic Tools:
    - Normal Probability Plot:
        - plots residuals (x-axis) vs. the cumulative probability p = (i-1/2)/n
        - If residuals are normal, this will be a straight line
    - Quantile-Quantile Plot: (aka QQ-plot)
        - plots quantiles of the standard normal vs. quantiles of the data
        - Should be a straight line if residuals are normal

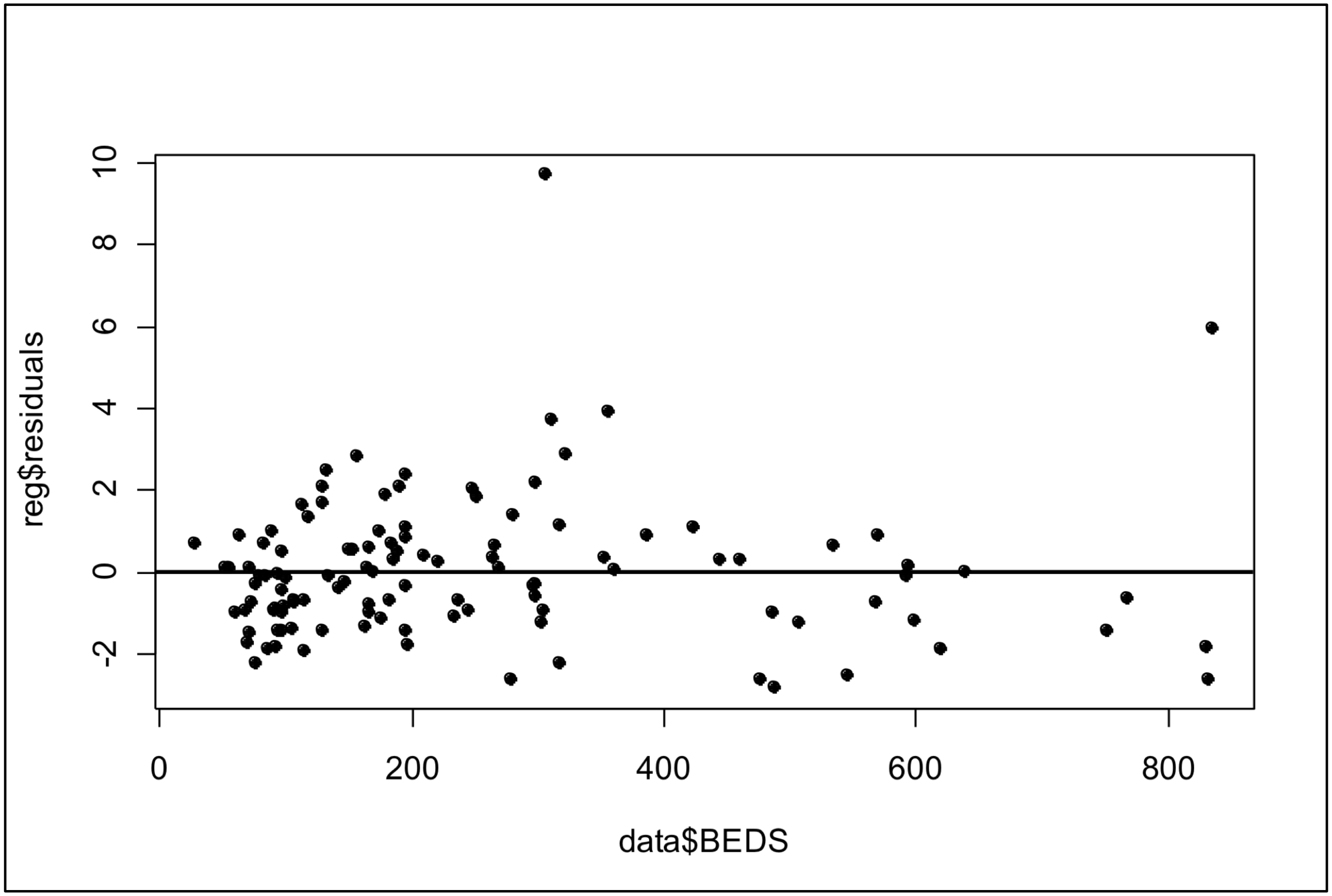
QQ-plot of our regression

```
qqnorm(res)
qqline(res)
```



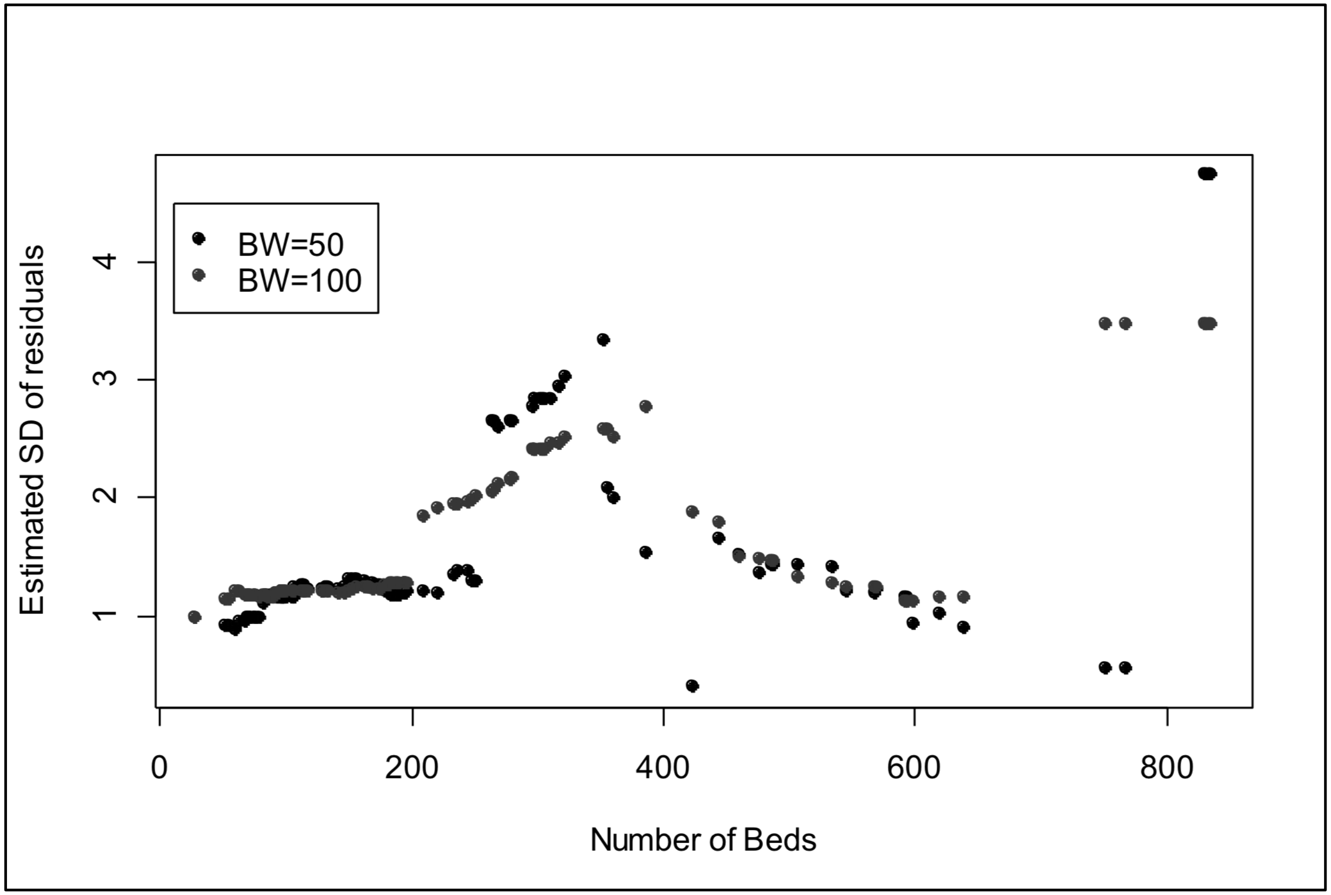
Constant variance assumption

- Is the spread of the residuals around the y=0 line approximately constant over the range of x?



Constant variance assumption

- Is the spread of the residuals around the y=0 line approximately constant over the range of x?



R code

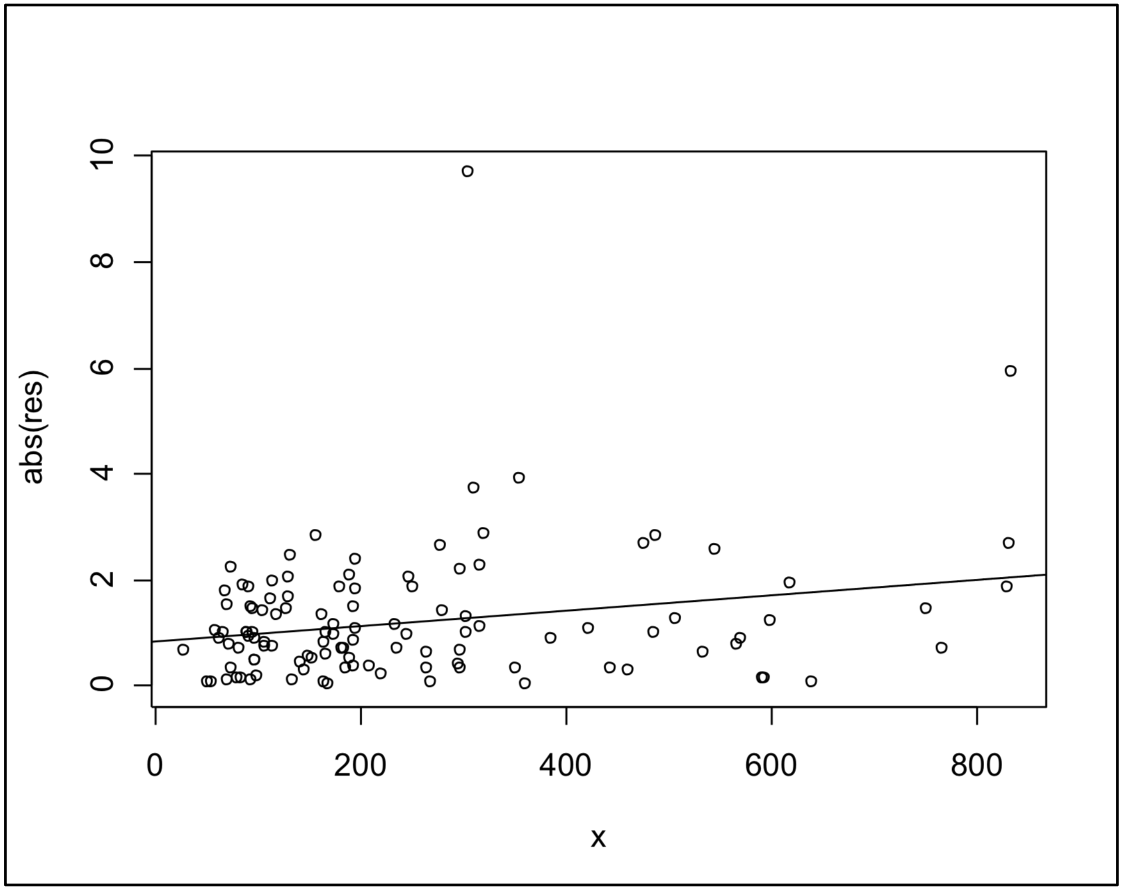
```
res <- reg$residuals
x <- data$BEDS

# estimate variance in bins
# use two different bin widths:  +/-50 and +/-100
n <- nrow(data)
vv50 <- vv100 <- rep(NA, n)
for(i in 1:n) {
   vv50[i] <- var(res[x>x[i]-50 & x<x[i]+50])
   vv100[i] <- var(res[x>x[i]-100 & x<x[i]+100])
}

# plot
plot(x, sqrt(vv50), ylab="Estimated SD of residuals",
   xlab="Number of Beds",  pch=16)
points(x, sqrt(vv100), col=2, pch=16)
legend(10,4.5, c("BW=50","BW=100"), pch=c(16,16),
   col=c(1,2))
```

Another Approach for Constant Variance Test

- Covariate vs. Squared (or Absolute) Residuals
- Tests for “fanning” of data: larger variance as X increases

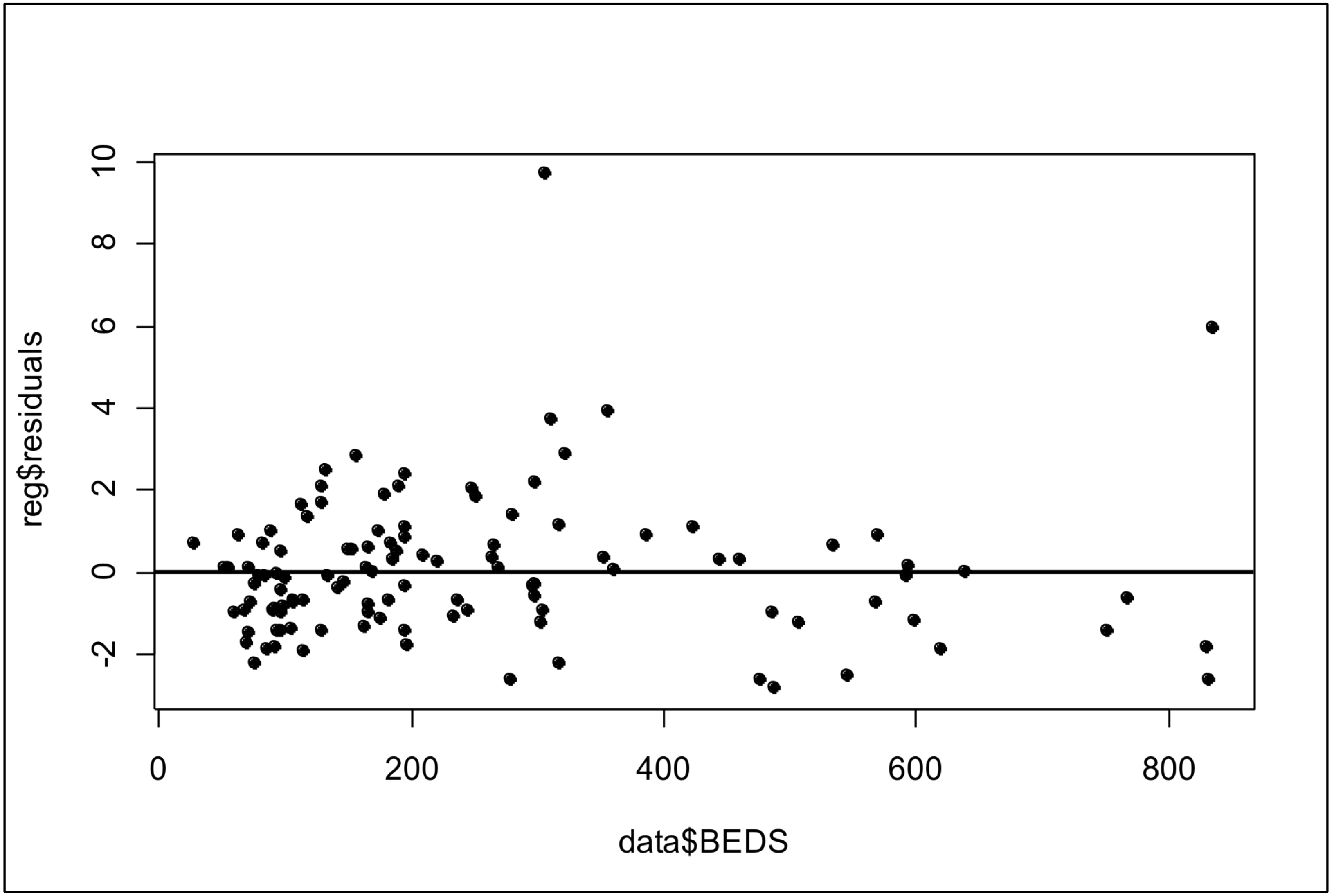


R code

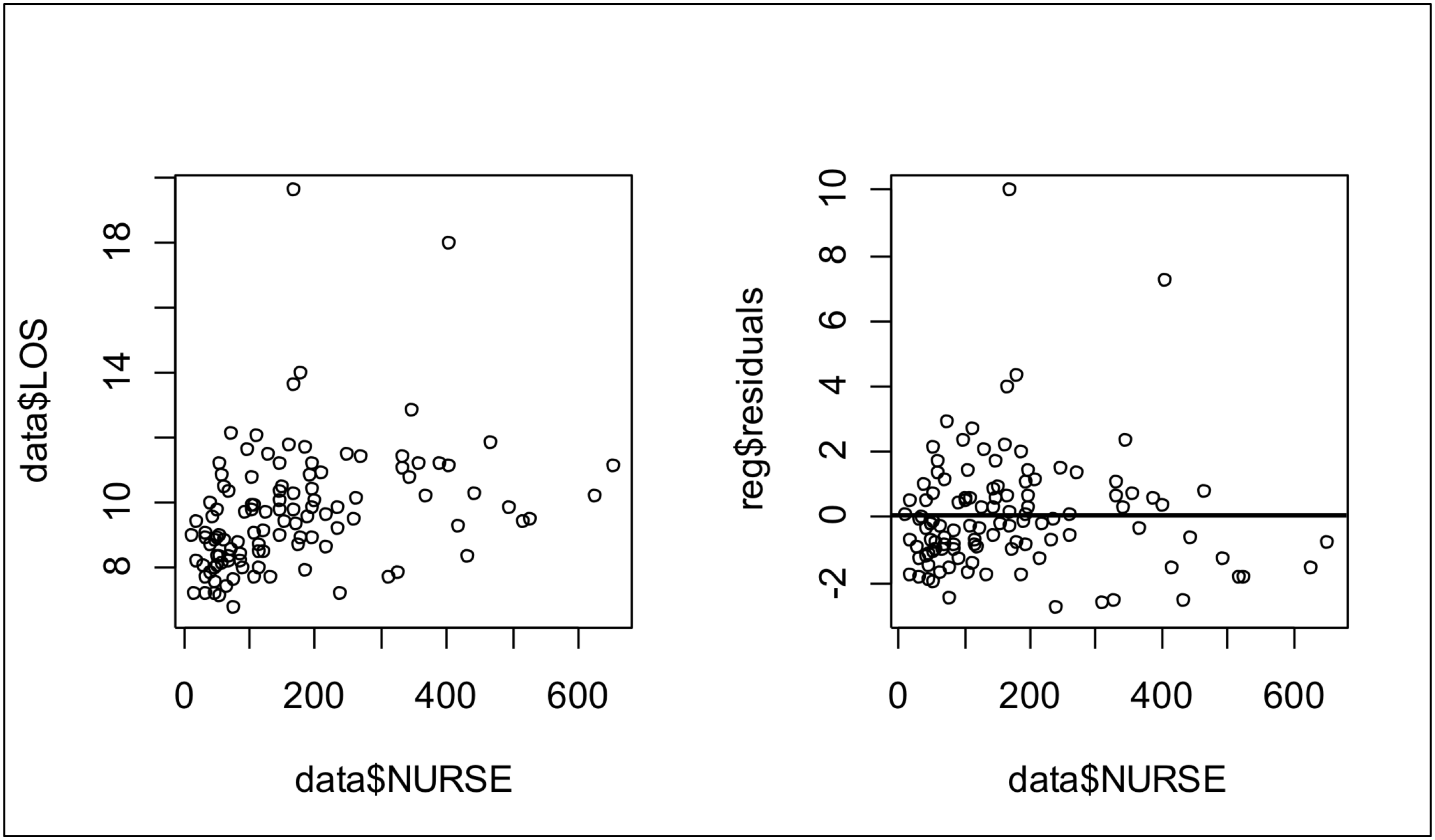
```
# plot x vs. absolute value of resids
plot(x, abs(res))
res.reg <- lm(abs(res) ~ x)
abline(res.reg)
summary(res.reg)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.8533786  0.1891567   4.511 1.61e-05 ***
x           0.0014415  0.0005968   2.415   0.0174 *
```

Lack of linear fit



Example of lack of linear fit



Curvature in the model?

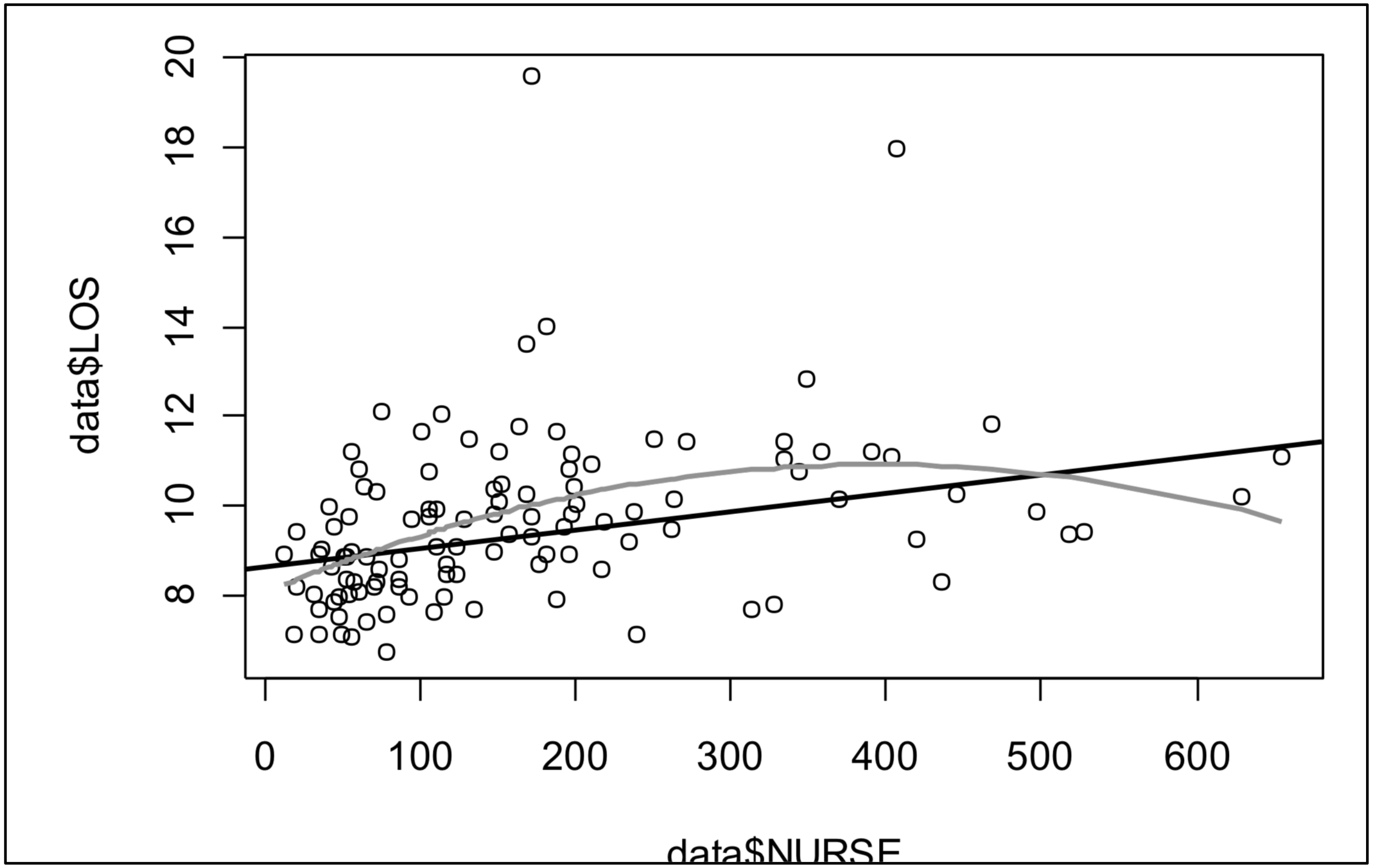
```
> nurse2 <- (data$NURSE)^2
> reg2 <- lm(data$LOS ~ data$NURSE + nurse2)
> summary(reg2)

Call:
lm(formula = data$LOS ~ data$NURSE + nurse2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.3397 -0.9841 -0.2522  0.6164  9.5678 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.038e+00  3.977e-01  20.212  < 2e-16 ***
data$NURSE   1.453e-02  3.846e-03   3.777 0.000258 ***
nurse2      -1.842e-05  6.833e-06  -2.695 0.008136 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

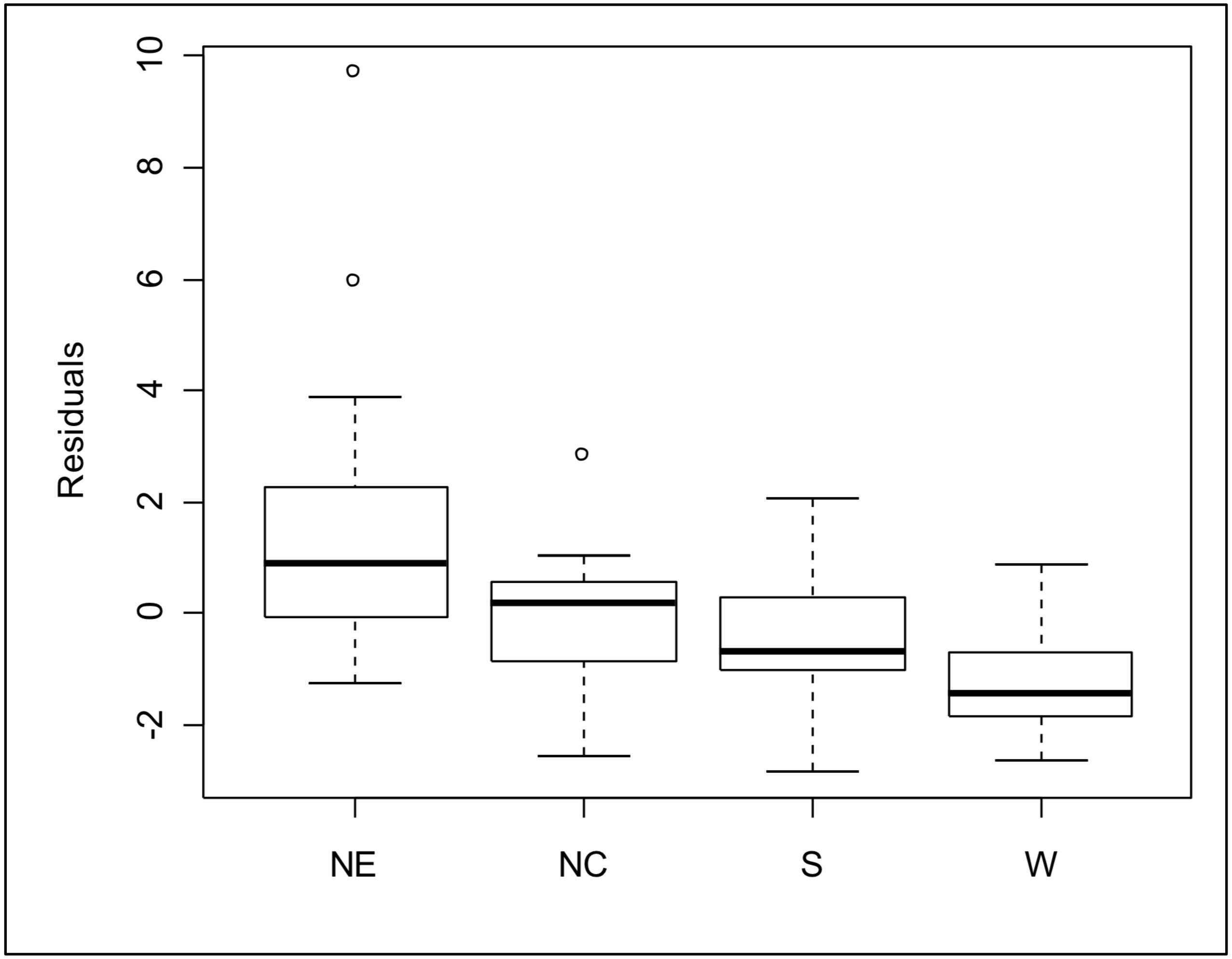
```
> yhat <- reg2$fitted.values
> plot(data$NURSE, data$LOS)
> lines(sort(data$NURSE), yhat[order(data$NURSE)], col=2,
             lwd=2)
> abline(reg, lwd=2)
```



Independence?

- Can be hard to test
- If there is *time* information and it is thought that their may be a time trend, you can try that
- But isn’t time then a predictor?
- yes: if you adjust for time, then you may gain independence
- Example: region. Are residuals independent with respect to region?

Residuals by Region



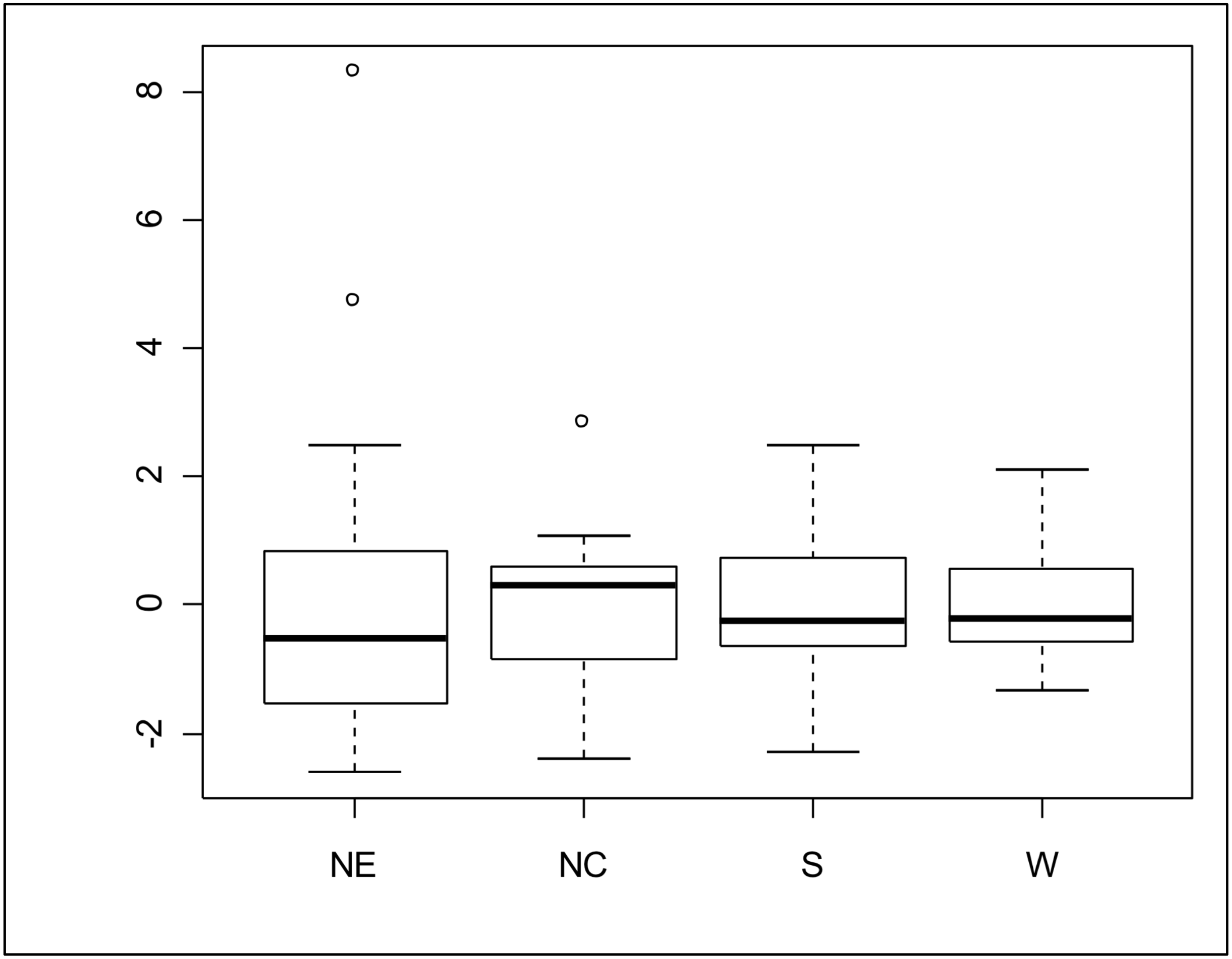
```
par(mar=c(4,5,1,1))
reg <- lm(data$LOS~ data$BEDS)
boxplot(reg$residuals ~ data$REGION, xaxt="n", ylab="Residuals")
axis(1, at=1:4, labels=c("NE","NC","S","W"))
```

Adjust for Region

```
> reg2 <- lm(data$LOS ~ data$BEDS + factor(data$REGION))
> summary(reg2)
…
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          10.1352898  0.3480667  29.119  < 2e-16 ***
data$BEDS             0.0036774  0.0007546   4.873 3.79e-06 ***
factor(data$REGION)2 -1.4805010  0.3944535  -3.753 0.000283 ***
factor(data$REGION)3 -1.8651866  0.3815803  -4.888 3.56e-06 ***
factor(data$REGION)4 -2.7142774  0.4803359  -5.651 1.31e-07 ***
---…
```

```
> boxplot(reg2$residuals ~ data$REGION, xaxt="n")
> axis(1, at=1:4, labels=c("NE","NC","S","W"))
```

Adjust for Region (continued)



So what do you think?

- Is our model ok?

- If not, what violations do you see?

- How might we improve our SLR?