Lecture 12: Regression Diagnostics I

Ailin Zhang

2023-06-07

Regression Diagnostics

- Regression diagnostics are used after fitting to check if a fitted mean function ($E(Y|X)$) and assumptions are consistent with observed data.
- To fit a model with least square estimates and perform statistical analysis rely on the regression assumptions:
  - Assumptions about the model form
  - Assumptions about the errors
  - Assumptions about the predictors
  - Assumptions about the observations
- We will use graphs (major) and numerical rules for detecting and correcting violations of model assumptions.

Agenda

- Revisit Regression Assumptions
- Leverage
- Types of Residuals
- Residual Plots
- Graphical Methods
- Influential Points and Outliers
- Measure of Influence

Revisit: Assumptions about the model form

We assume that the relationship between the response (Y) and the predictors $(X_1, \ldots, X_p)$ is linear. (linearity assumption)

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- For SLR, one can check linearity just by plotting Y against X
- For MLR, it's harder to check the linearity assumption due to high dimensions
  - One can use residual plot to check linearity (to be discussed later)
- Sometimes a non-linear relation can be turned linear by transforming variables.

Revisit: Assumptions about the Errors

The errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are

1. Independent

   Residual plot

2. with mean 0 and

   Since $E(Y|X) = X\beta$

3. constant variance σ^2, and

   Residual plot

4. (optional) normally distributed

   Check a normal probability plot

Assumptions about the Predictors

1. The predictors $X_1, X_2, \ldots, X_p$ are nonrandom fixed values.
2. The predictors $X_1, X_2, \ldots, X_p$ are measured withour error.
   - Hardly satisfied in real life.
   - Prediction are less accurate.
3. The predictors are linearly independent, i.e., no predictor can be expressed as a linear combination of others
   - No unique LS estimates for coefficients if there exist exact collinearity between predictors
   - Fine if there is no strong collinearity
   - Violation of this assumption is called multicollinearity, will be discussed later

Revisit: Assumptions about the observations

All observations are equally reliable and have an approximately equal role in determining the regression results and in influencing conclusions

The Hat (Projection) Matrix H

Recall for MLR:

- $Y = X\beta + \epsilon$, where X is often called the model matrix or the design matrix.
- The sum squares of errors: $(Y - X\hat{\beta})^T(Y - X\hat{\beta})$
- The normal equations [a system of equations whose solution is the Ordinary Least Squares (OLS) estimator of the regression coefficients]:

  $X^T X\hat{\beta} = X^T Y$
- OLS estimator for β:

  $\hat{\beta} = (X^T X)^{-1} X^T Y$
- Predicted Value $\hat{Y}$

  $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$ (where H is called the hat matrix or the projection matrix)

Leverage

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$\hat{Y} = HY$ means every predicted value $\hat{y}_i$ is a linear combination of $y_1, \dots, y_n$

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n$$

and h_{ij} is the (i, j) th element of the matrix H, and is completely determined by the predictors X as $H = X(X^TX)^{-1}X^T$

- h_{ij} is the weight given to y_j in predicting $\hat{y}_i$.
- h_{ii} is the weight given to y_i in predicting $\hat{y}_i$, is called the **leverage** of i-th observation.

Leverage

- If the leverage of i-th observation, h_{ii}, is large (close to 1), then this i-th observation is called a leverage point. It means the prediction of $\hat{y}_i$ depends a lot on the observation y_i itself and relatively less on other observations.
- Recall in SLR, we can solve H analytically:

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

And the leverage in SLR is given by

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Leverage

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Observe that h_{ii} is large when x_i is far from $\bar{x}$ relative to the SD of X.

It further means that the i-th observation is an outlier in the X space.

Other properties:

1. $\frac{1}{n} \leq h_{ii} \leq 1$
2. $\sum h_{ii} = p + 1$
3. Thus, on average, $h_{ii} \approx (p+1)/n$

   We can look for values far from this as rough screen for high leverage points.

Types of Residuals: Raw Residuals

Recall the residual of the i-th observation is defined to be

$$e_i = y_i - \hat{y}_i = \text{observed } y_i - \text{predicted } y_i$$

We call this as (raw) residuals.

Error ϵ's have 0 mean and constant variance σ^2

- Residual e_i also have 0 mean, $E(e_i) = 0$
- Residuals are uncorrelated with each predictor X_k $\text{Cov}(X_k, e) = 0$
- But they have unequal variance: $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$ where h_{ii}=leverage
- Residuals are NOT independent of each other as they must add up to 0

Standardized Residuals

- Raw residuals have different variances $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$, therefore we cannot identify outliers by comparing the magnitude of raw residuals.
- We standardize the i-th residual e_i as

$$z_i = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}}$$

- When we estimate σ by $\hat{\sigma} = \sqrt{MSE}$, we get the standardized residual or internally studentized residuals:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

- r_i has mean zero and standard deviation 1
- Observations w/ large $|r_i|$ (over 2 or 3 or 4) are potential outliers

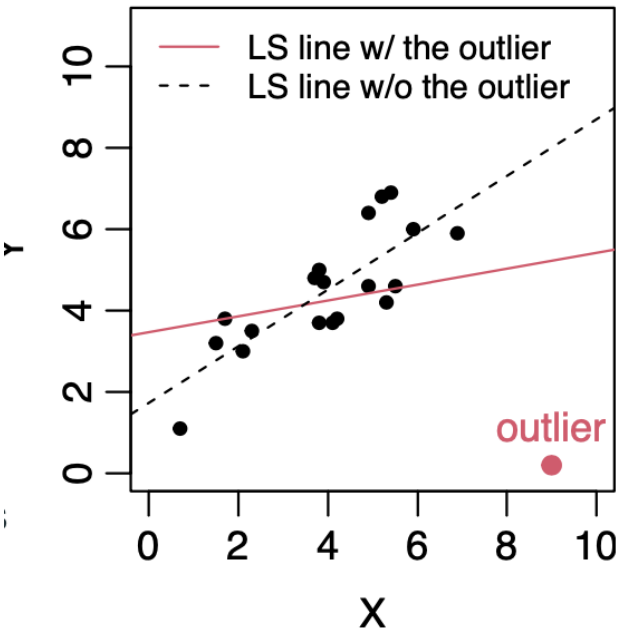
A Drawback of Internally Studentized Residuals

When there exists an outlier, it will

- distort the LS line,
- enlarge the residuals of other points and $\hat{\sigma}^2$
- underestimate the internally studentized residuals of the outlier.

Hence, it's better to estimate σ^2 excluding the outlier.

This is the idea behind **externally studentized** residuals



Studentized Residuals = Externally Studentized Residuals

Externally studentized residuals or studentized residuals are defined as:

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

- e_i is still computed using all the data but
- $\hat{\sigma}_{(i)}$ is computed from the MSE of the model that uses all the data EXCEPT the i-th observation
- The subscript "(i)" means "all but the ith observation".
- Externally studentized residuals r_i^* can be calculated from internally studentized residuals r_i via $r_i^* = r_i\sqrt{\frac{n-p-2}{n-p-1-r_i^2}}$

  If an observation is not an outlier, $r_i^* \approx r_i$.

Comparisons of 3 Types of Residuals

- e_i's add up to 0, r_i's and r_i^*'s do not add up to 0
- e_i's have unequal variance, r_i's and r_i^*'s have variance 1
- r_i^*'s has a t-distribution with df=n-p-2, but r_i does not have a t-distribution.
- With a large enough sample, r_i's and r_i^*'s are approximately $N(0, 1)$
- None of the 3 types of residuals are strictly independent, but the dependence can be ignored with large enough samples

Different Residuals in R

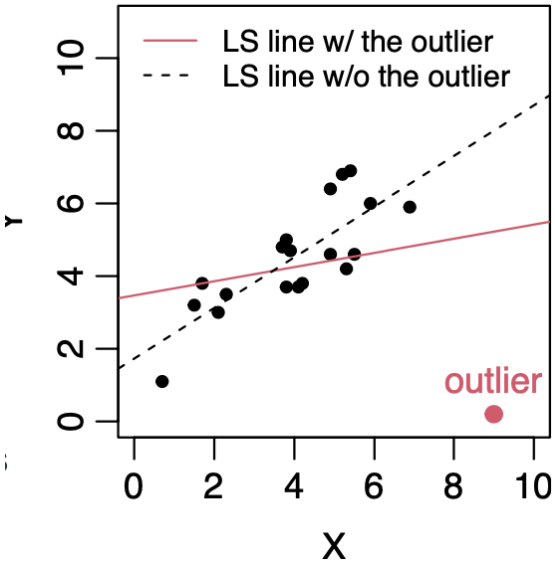
- The (raw) residuals e_i can be obtained like `modelname$res`

```
lm1 = lm(Y~X)
lm1$res
```

- The internally and externally studentized residuals can be obtained using `rstandard()` and `rstudent()` command

```
lm1 = lm(Y~X)
rstandard(lm1)
rstudent(lm1)
```

For the data in the plot below



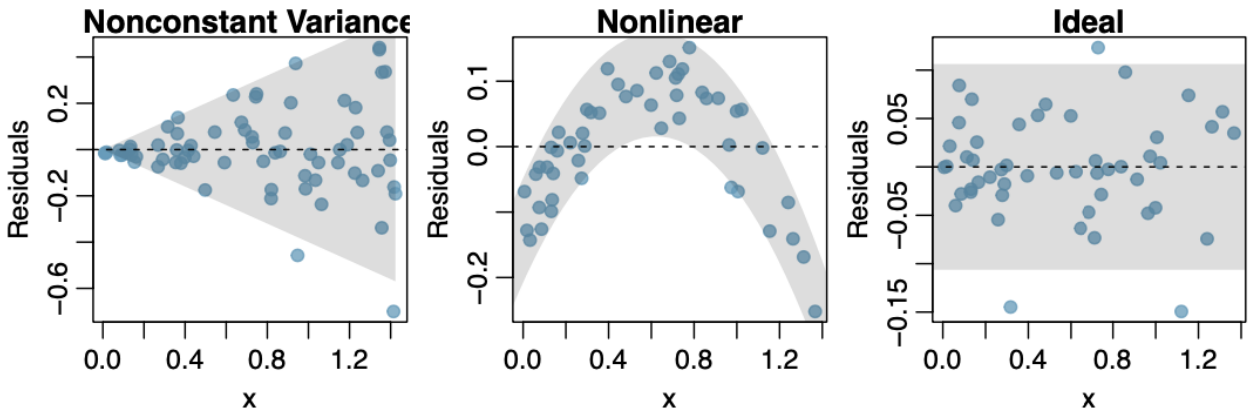

```
lm1 = lm(Y~X)
Raw.Res = round(lm1$res,2)
Int.Res = round(rstandard(lm1),2)
Ext.Res = round(rstudent(lm1),2)
data.frame(X,Y,Raw.Res,Int.Res,Ext.Res)
```

```
##      X   Y Raw.Res Int.Res Ext.Res
## 1  9.0 0.2   -5.02   -3.65   -6.96
## 2  5.9 6.0    1.38    0.84    0.84
## 3  4.9 6.4    1.98    1.19    1.20
## 4  3.9 4.7    0.47    0.28    0.27
## 5  6.9 5.9    1.09    0.69    0.67
## 6  4.1 3.7   -0.57   -0.34   -0.33
## 7  3.7 4.8    0.61    0.37    0.36
```

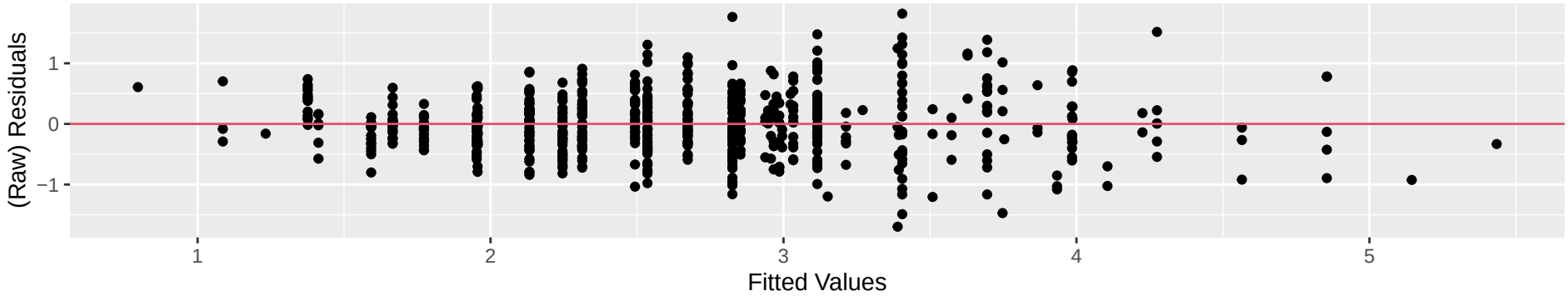
Various Kinds of Residual Plots

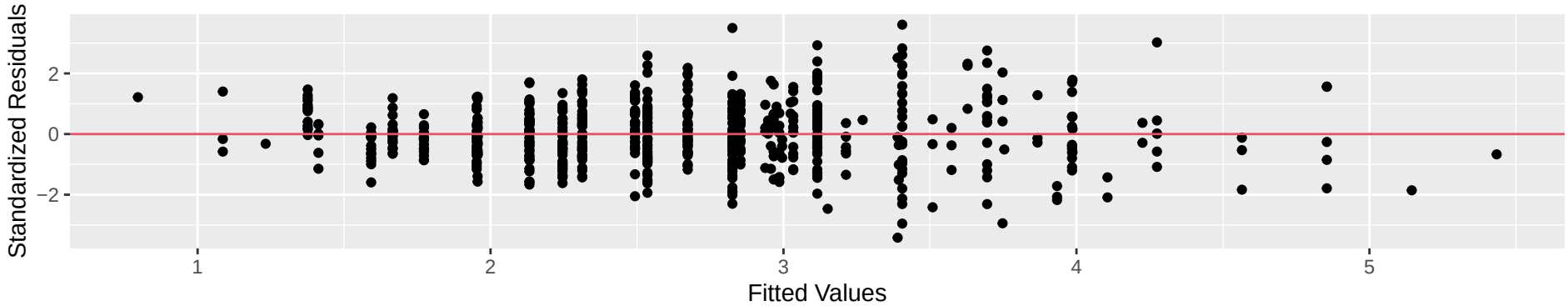
- Residuals v.s. fitted values
- Residuals v.s. each predictor
- Residuals v.s. potential predictors not yet included in the model
- Residuals v.s. several predictors using `ggplot()`
- Residuals v.s. time if the data are collected over time
- Residuals v.s. ...

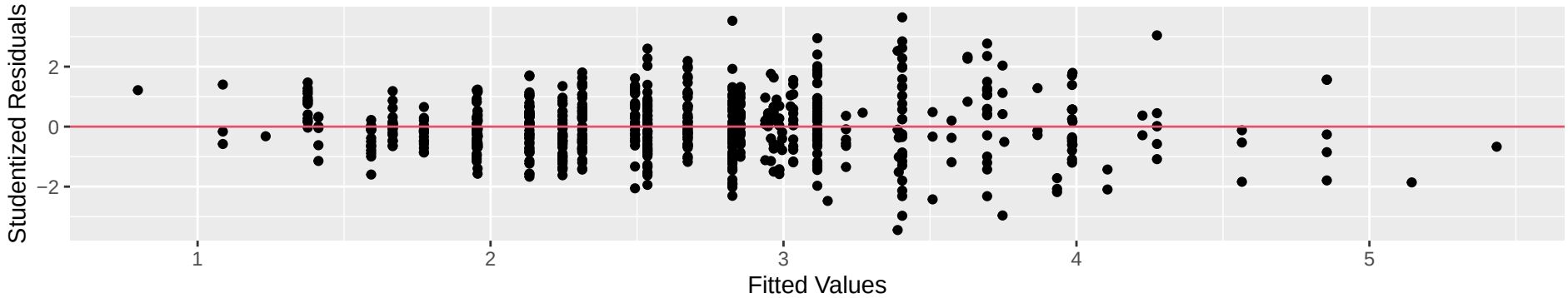
In all the plots above, points should scatter evenly above and below the zero line in a band of constant width.



Example: fev data

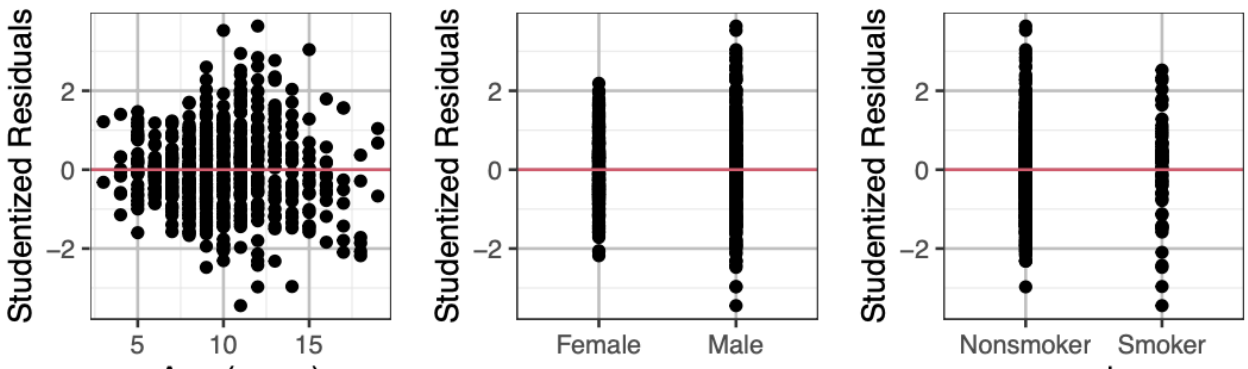






Residuals v.s. Each Predictor

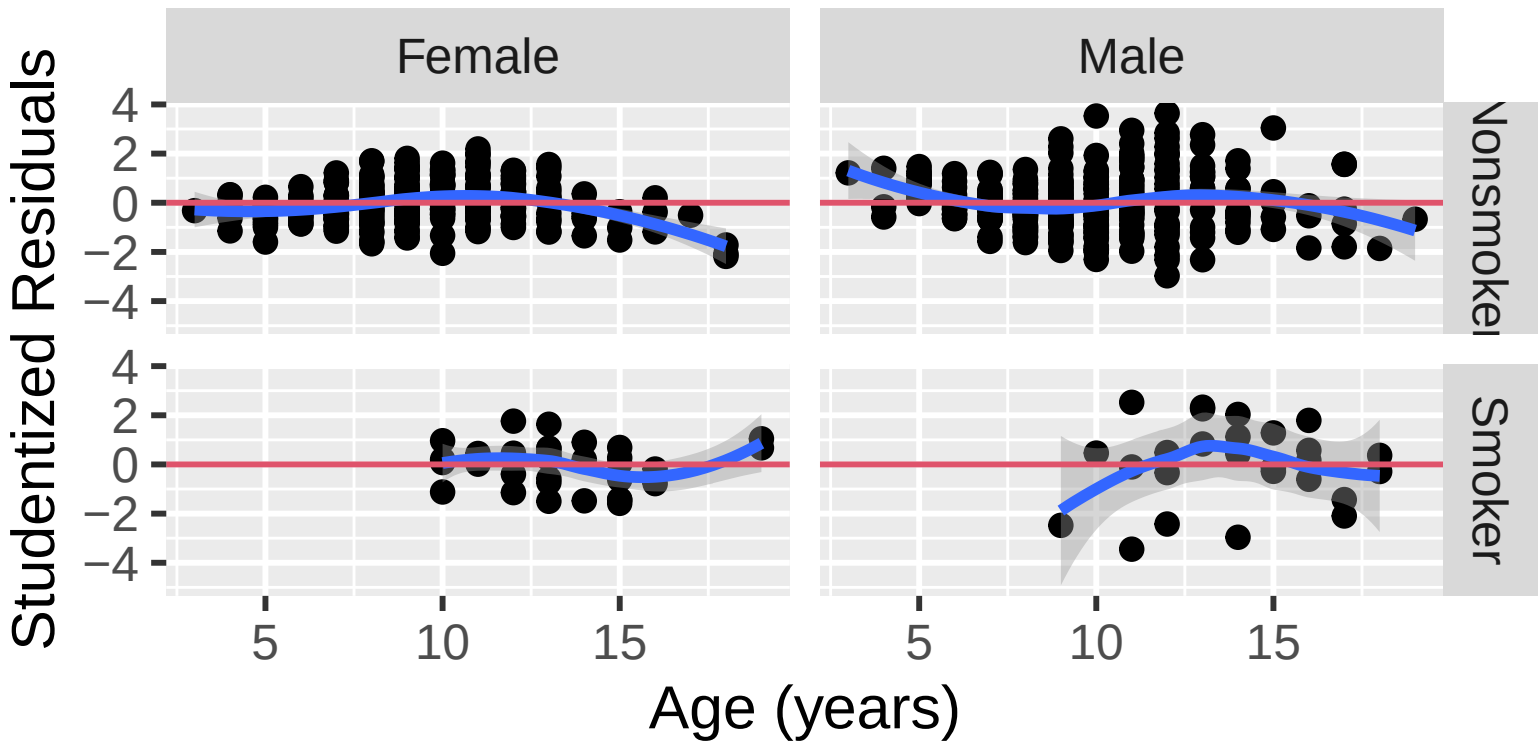
```
lm1 = lm(fev ~ age*smoke + age*sex, data=fevdata)
ggplot(fevdata, aes(x=age, y=rstudent(lm1))) + geom_point() +
  xlab("Age (years)") + ylab("Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
ggplot(fevdata, aes(x=sex, y=rstudent(lm1))) + geom_point() +
  ylab("Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
ggplot(fevdata, aes(x=smoke, y=rstudent(lm1))) + geom_point()
  ylab("Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
```



Residuals v.s. Several Predictors

```
ggplot(fevdata, aes(x=age, y=rstudent(lm1))) +
  geom_point() + facet_grid(smoke~sex) +
  geom_smooth(method='loess') +
  labs(x = "Age (years)", y= "Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

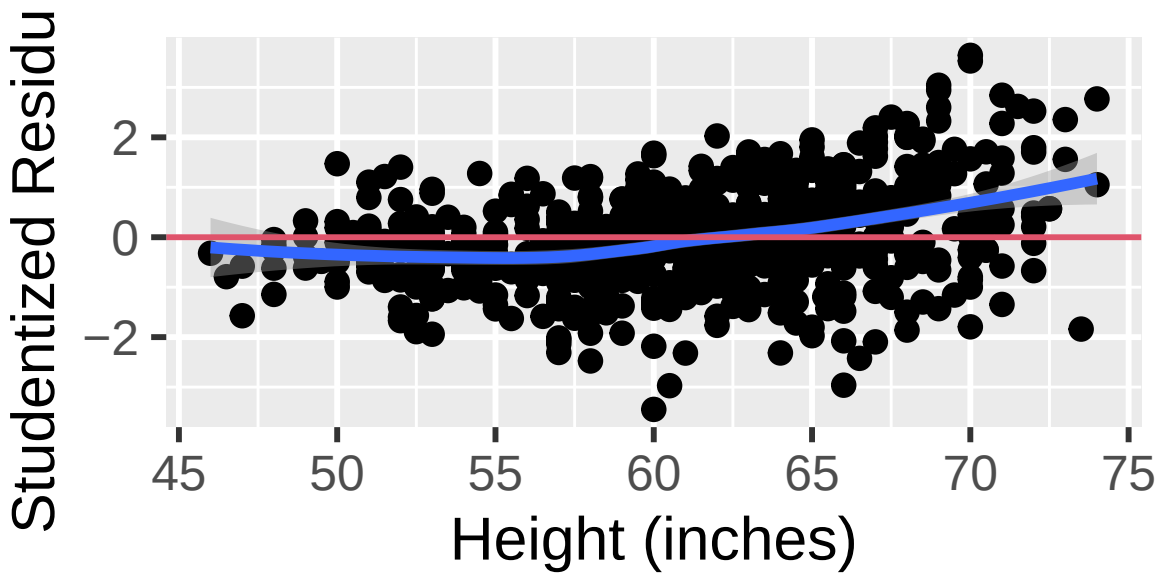


Residuals v.s. Potential Predictors

Recall the model lm1 doesn't not include ht (Height) as a predictor. Let's plot the residuals of lm1 against ht.

```
ggplot(fevdata, aes(x=ht, y=rstudent(lm1))) + geom_point()+geom_smooth(method='loess
  labs(x="Height (inches)", y="Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The residuals clearly have a positive nonlinear relation with height, meaning ht should be included in the model.

Issues Identified So Far

For the model below:

```
lm1 = lm(fev ~ age*smoke + age*sex, data=fevdata)
```

we identified the following issues based on the residual plots

- nonlinearity between age and fev
- variance of noise increases with fitted value
- ht or its transformation should be included