STAT 224 Lecture 10

Chapter 4 Model Diagnostics, Part 1

Yibi Huang

Assumptions of Multiple Regression Models

Assumptions about the Model Form

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

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

- For SLR, one can check linearity just by plotting Y against X
- For MLR, it's harder check the linearity assumption
- Sometimes a non-linear relation can be turned linear by transforming variables.

Assumptions about the Errors

The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are

- independent .... Chapter 8
- with mean 0 and
- common variance σ^2, and .... Chapter 6 & 7
- (optional) normally distributed

Assumptions about the Predictors

1. The predictors $X_1, X_2, \dots, X_p$ are **nonrandom fixed values**

- The assumption more closely fits designed experiments, where X_i's are conditions, dose levels, etc, which can be manipulated and controlled
- Otherwise, the inferences are conditional on the observed data. This subtle distinction will not be of further concern to us from now.

2. The predictors $X_1, X_2, \dots, X_p$ are **measured without error**.

- Never completely satisfied in real life.
- Prediction intervals are less accurate.

Assumptions about the Predictors (2)

3. The predictors are **linearly independent**, i.e., no predictor can be expressed as a linear combination of others
   - Ex: if X_1 = #undergrads, X_2 =#grads, X_3 = #students, then $X_1 + X_2 = X_3$
   - no unique LS estimates for coefficients if there exist exact col-linearity between predictors
   - fine if there is no **strong** collinearity
   - Violation of this assumption is called multicollinearity, will discuss in Ch 9-10.

One hallmark of Multiple Linear Regression Model is that small deviations from these assumptions do not invalidate our conclusions in a major way.

Leverage

MLR Models in Matrix Notation

Recall the MLR model

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + \varepsilon_j.$$

The matrix representation is

$$\begin{array}{rcccccc} & \overbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}^{\mathbf{Y}} & = & \overbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}}^{\mathbf{X}} & \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}^{\boldsymbol{\beta}} & + & \overbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}^{\varepsilon} \\ \text{dimensions:} & [n \times 1] & & [n \times (p+1)] & [(p+1) \times 1] & & [n \times 1] \end{array}$$

This is often written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

for short, and $\mathbf{X}$ is often called the **model matrix** or the **design matrix**.

The Hat Matrix H

The sum of squares $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$ can be written as

$$(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

- The normal equations can be written as:

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$$

- Least squares estimate for $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- Predicted Value $\widehat{\mathbf{Y}}$:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}Y$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, is called the *hat matrix* or the projection matrix

Leverage

$$\underbrace{\begin{pmatrix}\widehat{y_1}\\ \widehat{y_2}\\ \vdots\\ \widehat{y_n}\end{pmatrix}}_{\widehat{\mathbf{Y}}} = \underbrace{\begin{pmatrix}h_{11} & h_{12} & \cdots & h_{1n}\\ h_{21} & h_{22} & \cdots & h_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ h_{n1} & h_{n2} & \cdots & h_{nn}\end{pmatrix}}_{\mathbf{H}} \underbrace{\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}}_{\mathbf{Y}}$$

$\widehat{\mathbf{Y}} = \mathbf{HY}$ means every predicted value $\widehat{y_i}$ is a linear combination of $y_1, \ldots, y_n$

$$\widehat{y_i} = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n,$$

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

- h_{ij} = the weight given to y_j in predicting $\hat{y}_i$.
- h_{ii} = the **weight** given to y_i in predicting $\hat{y}_i$, is called the *leverage* of ith observation, $i = 1, 2, \ldots, n$.

Leverage (2)

- If the leverage of ith observation, h_{ii}, is large (close to 1), then this ith observation is called a **leverage point**. It means the prediction of $\widehat{y_i}$ depends a lot on the observation y_i itself and relatively less on other observations. It further means that the ith observation is an outlier in the X space.

Leverage (2)

- If the leverage of ith observation, h_{ii}, is large (close to 1), then this ith observation is called a **leverage point**. It means the prediction of $\widehat{y_i}$ depends a lot on the observation y_i itself and relatively less on other observations. It further means that the ith observation is an outlier in the X space.

- When there is only a single predictor in the model (SLR) we have
$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$
And the leverage in SLR is given by
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$
Observe that h_{ii} is large when x_i is far from $\bar{x}$ relative to the SD of X, which means *x_i is an outlier in X*.

Properties of the leverage h_{ii} (in MLR with intercept)

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

Properties of (Raw) Residuals

Recall the (raw) residual of the ith observation is defined to be

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

Recall the errors ε's have 0 mean and constant variance σ^2.

- Residuals e_i also have 0 mean, $\mathrm{E}(e_i) = 0$, but
- *unequal* variance $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$, where h_{ii} = leverage

Properties of (Raw) Residuals

Recall we proved on page 25 of the slides `L02.pdf` that

- $\sum_i e_i = 0$ . . . . Residuals add up to 0
- $\mathrm{Cor}(X_k, e) = 0$ . . Residuals are uncorrelated w/ each predictor

Hence residuals have 0 correlation with fitted values:

$$\mathrm{Cor}(\widehat{Y}, e) = 0$$

Properties of (Raw) Residuals

Recall we proved on page 25 of the slides `L02.pdf` that

- $\sum_i e_i = 0$ .... Residuals add up to 0
- $\text{Cor}(X_k, e) = 0$ .. Residuals are uncorrelated w/ each predictor

Hence residuals have 0 correlation with fitted values:

$$\text{Cor}(\widehat{Y}, e) = 0$$

About Independence:

- We assume the errors ε's to be independent of each other
- *Residuals are NOT independent* of each other as they must add up to 0

Standardized Residuals = Internally Studentized Residuals

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

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

- When the unknown σ is estimated by $\sqrt{\text{MSE}}$, we get the **standardized residual** or **internally studentized residuals**

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

- r_i has mean zero and standard deviation 1, but r_i's no longer add up to 0
- Observations w/ large $|r_i|$ (over 2 or 3 or 4) are potential outliers

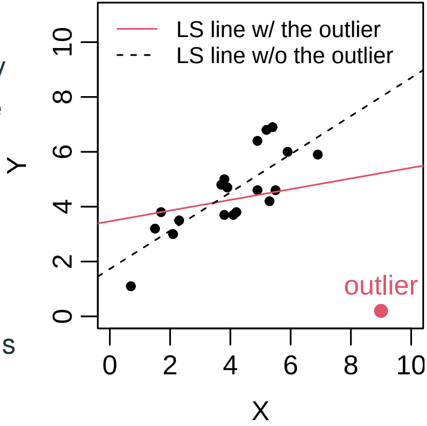
A Drawback of Internally Studentized Residuals

When there exists an outlier, it will

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

Hence, it's better 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:

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

- e_i is still computed using all the data but $\widehat{\sigma}_{(i)}$ is computed from the MSE of the model that uses all the data *EXCEPT the ith observation*
  - The subscript “(i)” means “all but the ith observation”.
- Externally studentized residuals $r_i^\star$ can be calculated from internally studentized residuals r_i via

$$r_i^\star = r_i\sqrt{\frac{n - p - 2}{n - p - 1 - r_i^2}}$$

If an observation is not an outlier, $r_i^\star \approx r_i$. It makes little difference which one we used.

Comparisons of 3 Types of Residuals

Under assumptions of MLR models

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

3 Kinds of Residuals in R

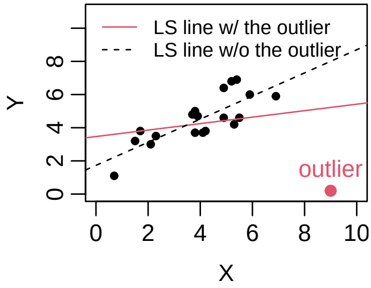
- The (raw) residuals e_i can be obtained like `modelname$res`

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

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

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

For the data in the plot below



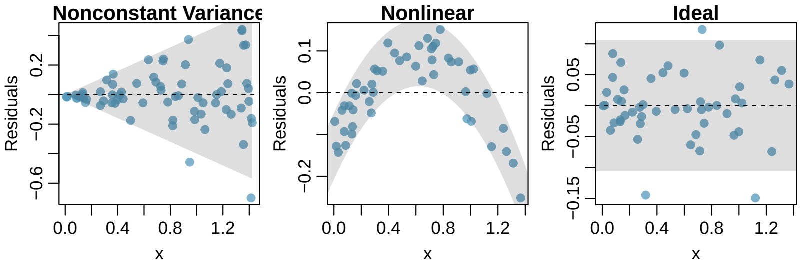
```
lm1 = lm(Y~X)
Raw.Res = round(lm1$res,2)
Int.Res = round(rstandard(lm1),2)
Ext.Res = round(rstudent(lm1),2)
data.frame(X,Y,Raw.Res,Int.Res,Ext.Res)
     X   Y Raw.Res Int.Res Ext.Res
1  9.0 0.2   -5.02   -3.65   -6.96
2  5.9 6.0    1.38    0.84    0.84
3  4.9 6.4    1.98    1.19    1.20
4  3.9 4.7    0.47    0.28    0.27
5  6.9 5.9    1.09    0.69    0.67
```

```
     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
8  1.5 3.2   -0.56   -0.35   -0.35
9  5.5 4.6    0.06    0.04    0.03
10 2.1 3.0   -0.88   -0.54   -0.53
11 1.7 3.8    0.00    0.00    0.00
12 2.3 3.5   -0.42   -0.26   -0.25
13 0.7 1.1   -2.50   -1.65   -1.74
14 3.8 5.0    0.79    0.47    0.46
15 4.9 4.6    0.18    0.11    0.10
16 5.3 4.2   -0.30   -0.18   -0.18
17 3.8 3.7   -0.51   -0.30   -0.30
18 5.4 6.9    2.38    1.44    1.48
19 4.2 3.8   -0.49   -0.29   -0.28
20 5.2 6.8    2.32    1.40    1.44
```

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. . . . (be creative)

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



FEV Lung Capacity Data Revisit

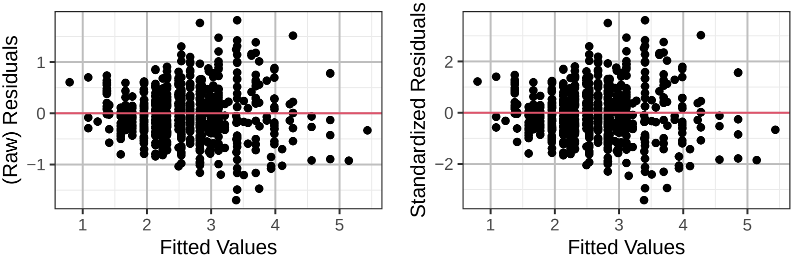
For the FEV lung capacity data,

```r
fevdata = read.table("fevdata.txt", header = TRUE)
fevdata$sex = factor(fevdata$sex, labels=c("Female","Male"))
fevdata$smoke = factor(fevdata$smoke, labels=c("Nonsmoker","Smoker"))
```

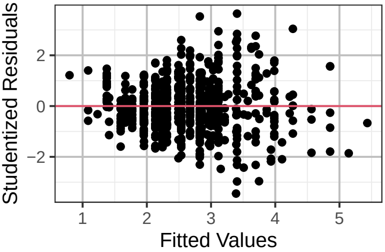
recall we considered the model below with `age*smoke` and `age*sex` interactions.

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

Residuals v.s. Fitted Values



- The residuals can be raw, standardized, or studentized
- Usually, the look of residual plots doesn't depend much on the type of residuals used, if all leverages $h_{ii} \ll 1$ or all are of similar magnitude.
- **Variance increases with fitted values** in all 3 plots

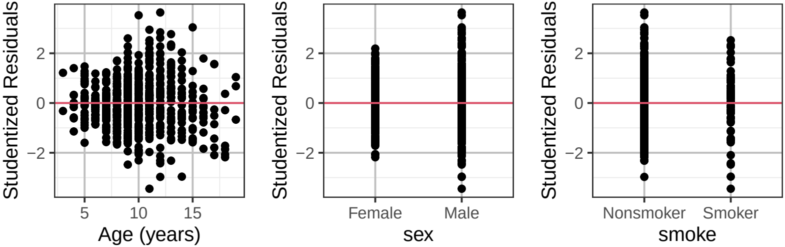


R Codes for The Plots on the Previous Page

```r
ggplot(fevdata, aes(x=lm1$fit, y=lm1$res)) + geom_point() +
  xlab("Fitted Values") + ylab("(Raw) Residuals") +
  geom_hline(yintercept = 0, col=2)
ggplot(fevdata, aes(x=lm1$fit, y=rstandard(lm1))) +
  geom_point() + xlab("Fitted Values") +
  ylab("Standardized Residuals") +
  geom_hline(yintercept = 0, col=2)
ggplot(fevdata, aes(x=lm1$fit, y=rstudent(lm1))) +
  geom_point() + xlab("Fitted Values") +
  ylab("Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
```

Residuals v.s. Each Predictor

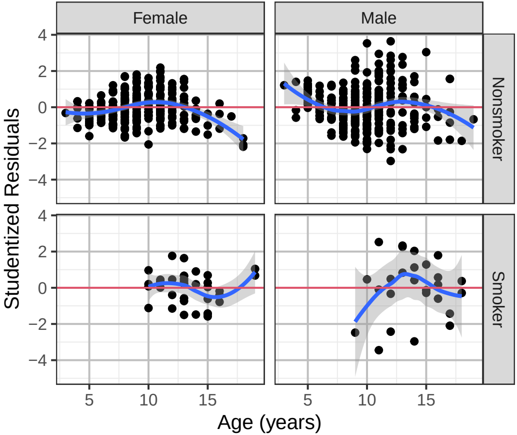
```
ggplot(fevdata, aes(x=age, y=rstudent(lm1))) + geom_point() +
  xlab("Age (years)") + ylab("Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
ggplot(fevdata, aes(x=sex, y=rstudent(lm1))) + geom_point() +
  ylab("Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
ggplot(fevdata,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

- The blue line is the `loess` smoother
- Slight nonlinearity among M & F nonsmokers

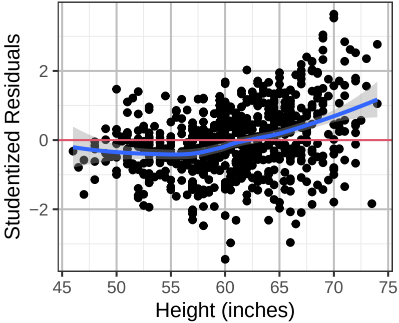


Residuals v.s. Potential Predictors

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

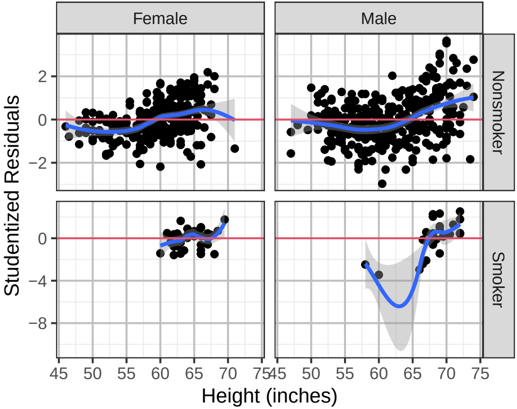
```
ggplot(fevdata, aes(x=ht, y=rstudent(lm1))) +
  geom_point()+geom_smooth(method='loess') +
  labs(x="Height (inches)", y="Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
`geom_smooth()` using formula 'y ~ x'
```

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



Residuals v.s. Potential Predictors (2)

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



Problems Identified So Far

For the model below

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

we found the following problems based on the residual plots

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