MLR: Diagnostic Plots and Multicollinearity

BIOS 6611

CU Anschutz

Week 10

Linear Regression Assumptions Revisited

(Multiple) Linear Regression Assumptions

Existence: For any combination of $X_1, X_2, ..., X_k$, Y is a random variable with a certain probability distribution having finite mean and variance.

Independence: The Y-values are statistically independent of one another.

Linearity: The mean value of Y for each combination of $X_1, X_2, ..., X_k$ is a linear function of $X_1, X_2, ..., X_k$.

Homoscedasticity: The variance of Y ($\sigma^2_{Y|X_1,X_2,...,X_k}$) is the same for any fixed combination of $X_1, X_2, ..., X_k$.

Normal Distribution: For any fixed combination of $X_1, X_2, ..., X_k$, the residuals are normally distributed.

Regression Diagnostic Plots

Regression Diagnostic Plots

Many of the same plots we introduced for simple linear regression can be used for multiple linear regression, with a few additional considerations:

- Y-X scatterplot $\rightarrow$ *partial regression plot*
- Scatterplot of the residuals and $X \rightarrow$ *scatterplot of* $\hat{Y}$ *and residuals*
- Histogram of the residuals
- PP or QQ plot of the residuals

We used **jackknife residuals** in our plots, but other types of residuals can also be used.

Partial Regression Plot

This will replace our Y-X scatterplot from simple linear regression. A **partial regression plot** (also known as a partial plot, added variable plot, or adjusted variable plot) characterizes the relationship between the dependent variable (Y) and an independent variable (X), adjusting for other covariates in the model ($C_1, C_2, ..., C_k$).

We can calculate the partial regression plot by following 3 steps:

1. Perform a regression of Y on $C_1, C_2, ..., C_k$ and save the observed residuals.
2. Perform a regression of X on $C_1, C_2, ..., C_k$ and save the observed residuals.
3. Plot the residuals from step (1) and step (2).

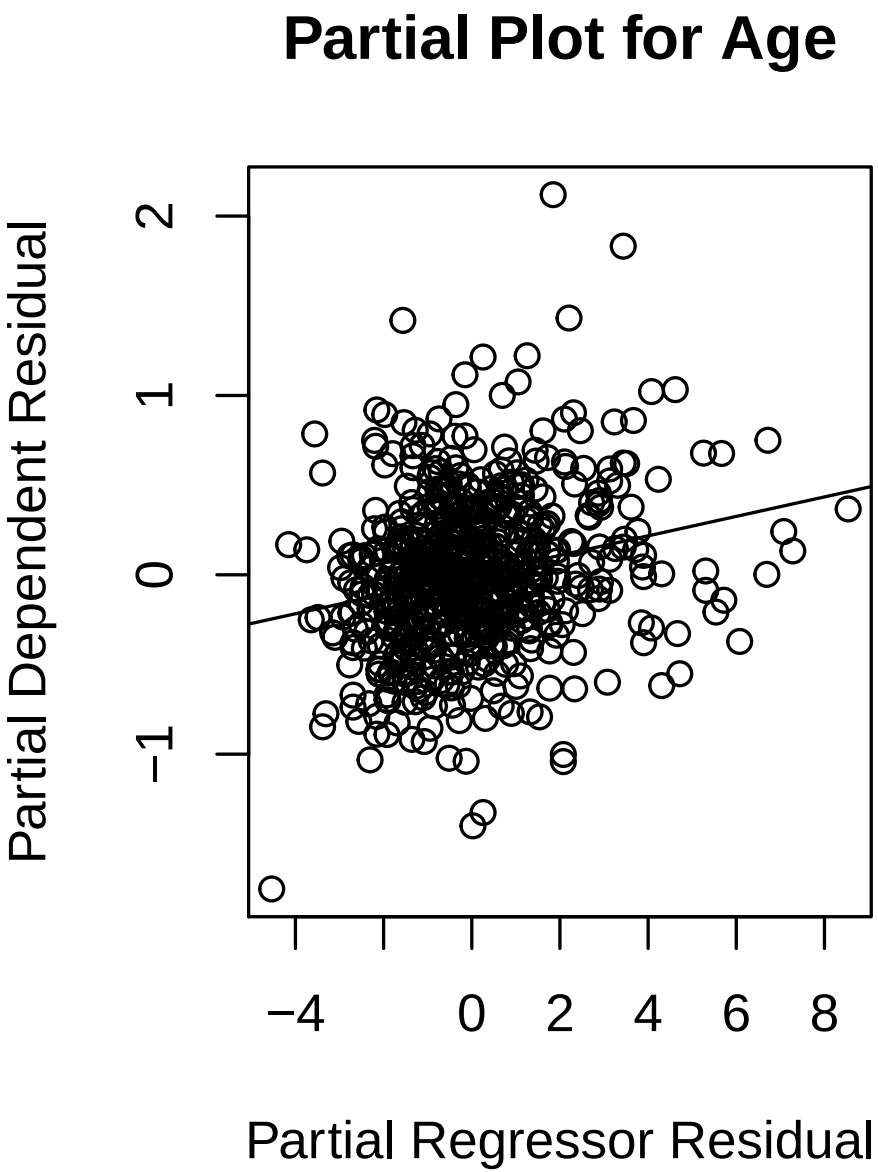
Partial Regression Plot Example

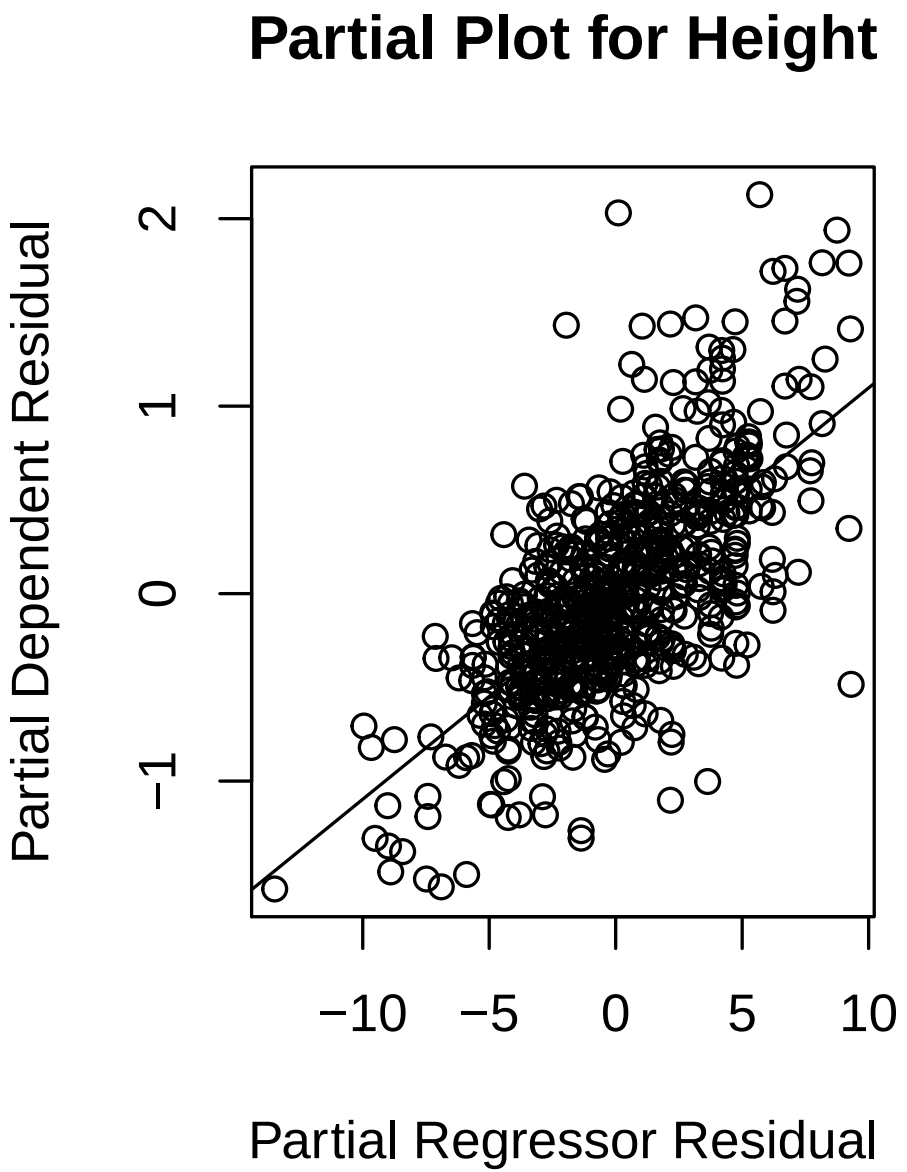
```
fev <- read.csv('FEV_rosner.csv')
mod1 <- glm( fev ~ age + height, data=fev )

# Partial Plot for Age
age_step1 <- glm(fev ~ height, data=fev)
age_step2 <- glm(age ~ height, data=fev)
plot(x=residuals(age_step2), y=residuals(age_step1),
     main='Partial Plot for Age', ylab='Partial Dependent Residual',
     xlab='Partial Regressor Residual')
# Add SLR line to show slope
abline(lm(residuals(age_step1) ~ residuals(age_step2)))

# Partial Plot for Height
height_step1 <- glm(fev ~ age, data=fev)
height_step2 <- glm(height ~ age, data=fev)
plot(x=residuals(height_step2), y=residuals(height_step1),
     main='Partial Plot for Height', ylab='Partial Dependent Residual',
     xlab='Partial Regressor Residual')
# Add SLR line to show slope
abline(lm(residuals(height_step1) ~ residuals(height_step2)))
```

Partial Regression Plot Example





Partial Regression Plot Example

The slope of the partial plot will be the same as the slope of X in the MLR model of Y on $X, C_1, C_2, ..., C_k$. In other words, a simple linear regression of the residuals from step (1) on step (2) will result in the same estimate of $\hat{\beta}_X$ from the MLR model.

```
round(summary(glm( fev ~ age + height, data=fev ))$coefficients,4)
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -4.6105     0.2243 -20.5576        0
## age           0.0543     0.0091   5.9609        0
## height        0.1097     0.0047  23.2628        0
```

```
round(summary(glm(residuals(age_step1) ~ residuals(age_step2)))$coefficients,4)
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.0000     0.0164  0.0000        1
## residuals(age_step2)    0.0543     0.0091  5.9655        0
```

```
round(summary(glm(residuals(height_step1) ~ residuals(height_step2)))$coef,4)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.0000     0.0164  0.0000        1
## residuals(height_step2)    0.1097     0.0047 23.2806        0
```

Scatterplot of $\hat{Y}$ and Residuals

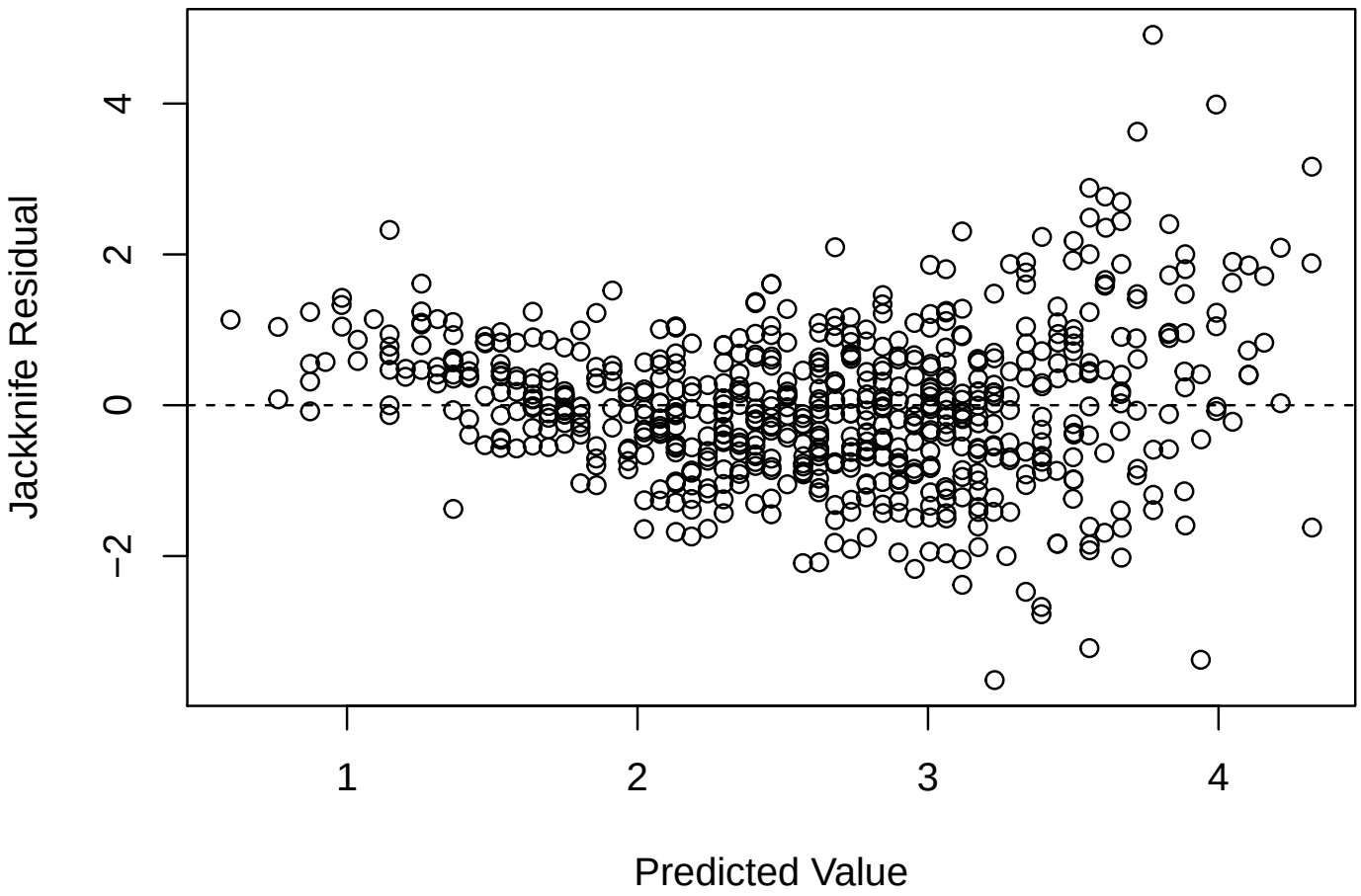
This will replace our scatterplot of the residuals by X. By plotting the scatterplot of our residuals by $\hat{Y}$ we can account for the relationship amongst all our independent variables, $X_1, X_2, ..., X_k$.

We can use this plot to check similar assumptions regarding linearity and homoscedasticity that we evaluated before:

- Do the residuals appear to jump around a residual of 0 for all values of $\hat{Y}$ (*linearity* assumption)
- Do the residuals form a horizontal band around the 0 line? (*homoscedasticity* assumption)
- Do any points seem to be extremely large/small? (potential outliers, more on this later!)

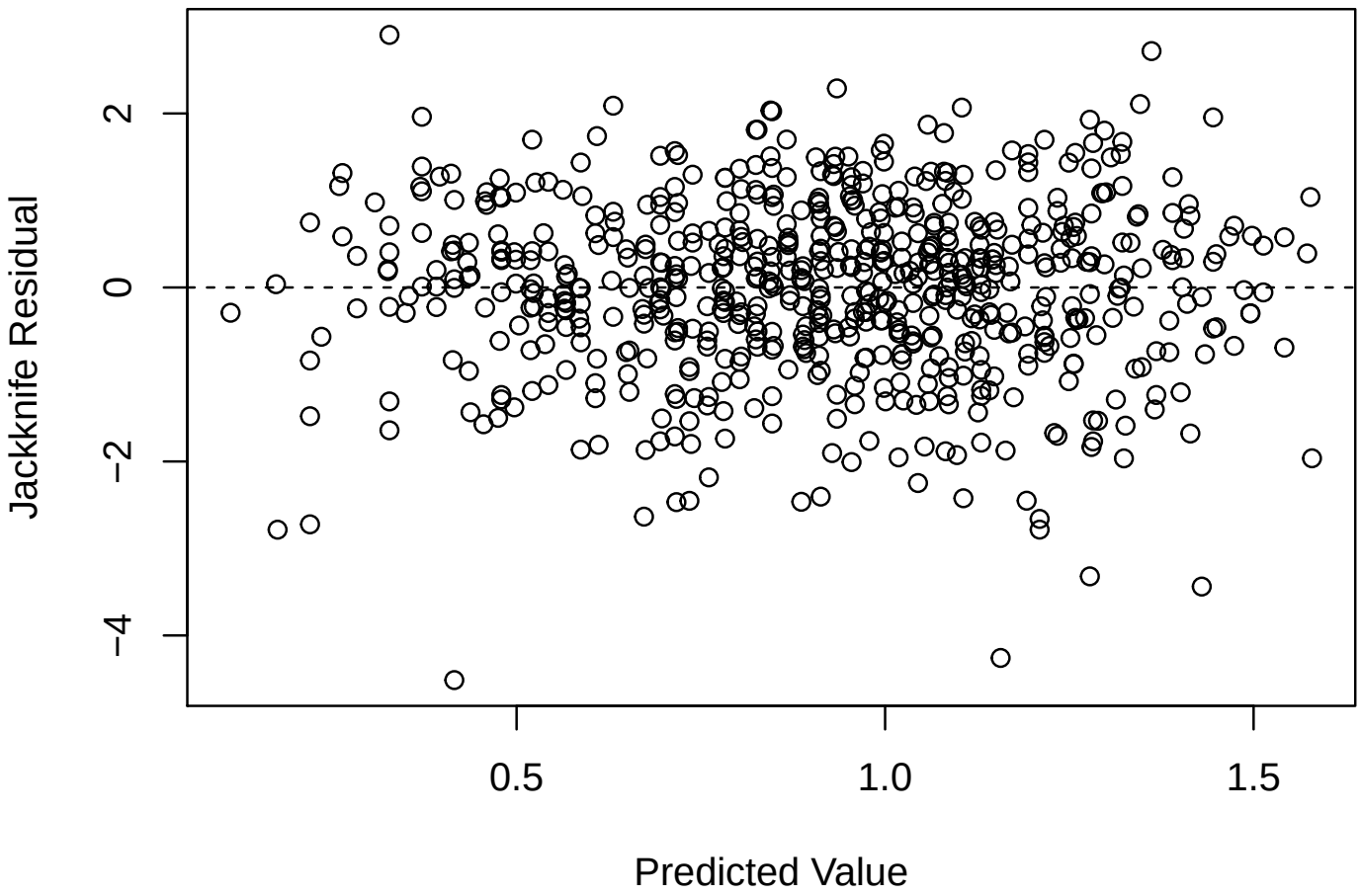
Scatterplot of $\hat{Y}$ and Residuals Example

```
par(mar=c(4.1,4.1,1.1,1.1))
plot(x=predict(mod1), y=rstudent(mod1),
     xlab='Predicted Value', ylab='Jackknife Residual')
abline(h=0, lty=2)
```



Scatterplot of $\hat{Y}$ and Residuals Example

```
mod2 <- glm( log(fev) ~ age + height, data=fev)
plot( x=predict(mod2), y=rstudent(mod2),
     xlab='Predicted Value', ylab='Jackknife Residual')
abline(h=0, lty=2)
```



Multicollinearity

Multicollinearity

Collinearity is where we have **two** explanatory variables with a linear association.

Multicollinearity is where we have **two or more** explanatory variables that are highly linearly related.

In general, this suggests we have highly correlated predictors. At the extreme, one predictor may be a linear combination of other predictors (e.g., $X_3 = 2X_1 - X_2$).

This is problematic for a few reasons:

- It can be difficult to determine the true effect of each predictor on the outcome.
- It can lead to poorly estimated coefficients and standard errors (i.e., misleading p-values or confidence intervals).
- The overall F-test may provide a significant result even if each individual predictor is not significant.

Evaluating Multicollinearity

The **variance inflation factor (VIF)** is often used to measure collinearity in the context of multiple linear regression. It is computed for the j^{th} predictor variable as:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination based on regressing X_j as the outcome on the remaining $k - 1$ predictors.

A rule of thumb is to be concerned with a $VIF > 10$, which corresponds to an $R_j^2 > 0.9$.

Addressing Multicollinearity

If a VIF indicates multicollinearity...

- Consider if it makes sense (interaction or polynomial terms are expected to be correlated without additional transformations)
- Choose variable with largest adjusted R squared in the model
- Create a new variable where appropriate (BMI from height and weight)
- PCA of the variables (for cases with a larger # of covariates)

Multicollinearity Examples

```r
library(car) # load car package for vif()

mod_vif1 <- glm( fev ~ age + height + sex + smoke, data=fev)
vif(mod_vif1)
```

```
##      age   height      sex    smoke
## 3.019010 2.829728 1.060228 1.209564
```

```r
mod_vif2 <- glm( fev ~ age + I(age^2) + height, data=fev)
vif(mod_vif2)
```

```
##       age  I(age^2)    height
## 43.234975 34.050491  3.336106
```

```r
fev$newvar <- 3*fev$height + sqrt(fev$age)
mod_vif3 <- glm( fev ~ age + sex + height + newvar, data=fev)
vif(mod_vif3)
```

```
##          age          sex       height       newvar
## 9.958277e+01 1.064032e+00 1.457695e+05 1.518023e+05
```