Multiple Linear Regression

Author: Nicholas G Reich

Modeling kid's test scores

Setting: a model for cognitive test scores of 434 three- and four-year-old children using characteristics of their mothers. Data are from a survey of adult American women and their children - a subsample from the National Longitudinal Survey of Youth.

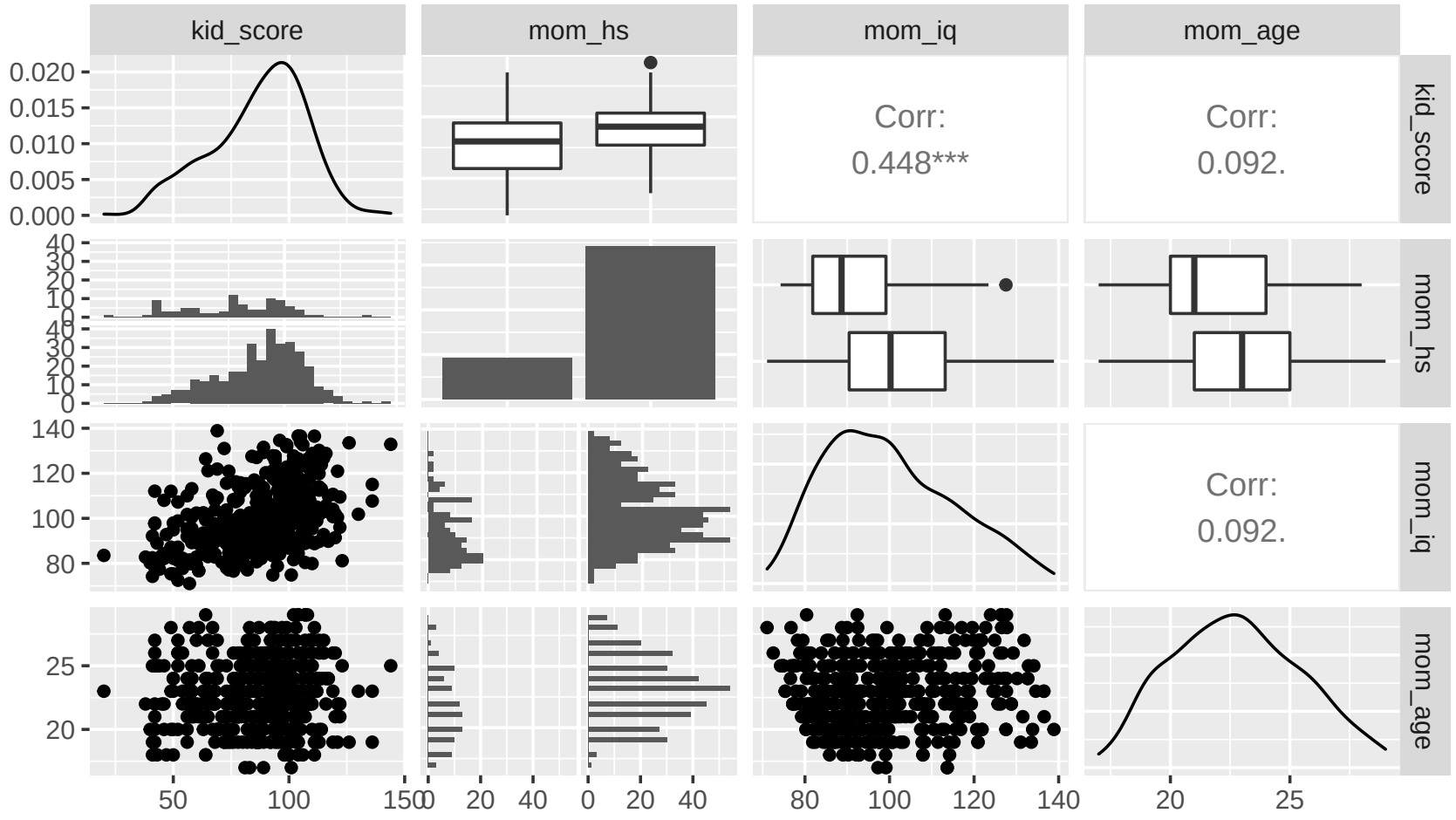
```
library(rstanarm)
data("kidiq")
head(kidiq)
```

```
##   kid_score mom_hs    mom_iq mom_age
## 1        65      1 121.11753      27
## 2        98      1  89.36188      25
## 3        85      1 115.44316      27
## 4        83      1  99.44964      25
## 5       115      1  92.74571      27
## 6        98      0 107.90184      18
```

Gelman, Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (2007) Cambridge University Press.

Exploratory analysis

```
library(GGally)
kidiq$mom_hs <- factor(kidiq$mom_hs, levels=c(0,1), labels=c("no", "yes"))
ggpairs(kidiq)
```



What might we want to know?

What impact does a mom's education have on her child's intelligence?

(as measured by a standardized test)

Does the relationship change when we account for other characteristics about the mom?

Interpreting an SLR coefficient

What is the correct interpretation of the coefficient for mom's high school status?

```
fm <- lm(kid_score ~ mom_hs, data=kidiq)
round(summary(fm)$coef, 3)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   77.548      2.059  37.670        0
## mom_hsyes     11.771      2.322   5.069        0
```

Interpreting an SLR coefficient

What is the correct interpretation of the coefficient for mom's high school status?

```
fm <- lm(kid_score ~ mom_hs, data=kidiq)
round(summary(fm)$coef, 3)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   77.548      2.059  37.670        0
## mom_hsyes     11.771      2.322   5.069        0
```

Kids with mothers who finished high school tend to score on average 11.8 points higher on the IQ test compared to kids whose mothers did not finish high school.

Interpreting uncertainty in an SLR coefficient

```
confint(fm)
```

```
##                 2.5 %   97.5 %
## (Intercept) 73.502246 81.59453
## mom_hsyes    7.206598 16.33592
```

Kids with mothers who finished high school tend to score on average 11.8 points higher on the IQ test compared to kids whose mothers did not finish high school (95% CI: 7.2 - 16.3).

What percent of the variation is explained by a model?

R^2 is a common metric. It is defined as the percent of variation in the outcome described by a model:

$$
\begin{aligned}
R^2 &= \frac{\text{explained variability in } y}{\text{total variability in } y} \\
&= 1 - \frac{\text{unexplained variability in } y}{\text{total variability in } y} \\
&= 1 - \frac{\text{variation of model residuals}}{\text{total variability in } y}
\end{aligned}
$$

What percent of the variation in kid IQ is explained by mom HS?

$$
\begin{aligned}
R^2 &= 1 - \frac{\text{variation of model residuals}}{\text{total variability in } y} \\
&= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}
\end{aligned}
$$

```
1 - sum( resid(fm)^2 ) / sum( (kidiq$kid_score-mean(kidiq$kid_score))^2 )
```

```
## [1] 0.0561294
```

```
summary(fm)$r.squared
```

```
## [1] 0.0561294
```

Interpreting an MLR coefficient

What does this model look like in terms of math?

```
fm1 <- lm(kid_score ~ mom_hs + mom_iq + mom_age, data=kidiq)
```

Interpreting an MLR coefficient

What is the correct interpretation of the coefficient for mom's high school status?

```
coef(fm1)

## (Intercept)   mom_hsyes      mom_iq     mom_age
##  20.9846620   5.6471512   0.5625443   0.2247505

confint(fm1)

##                  2.5 %     97.5 %
## (Intercept)  3.0394352 38.9298887
## mom_hsyes    1.2097371 10.0845653
## mom_iq       0.4433466  0.6817419
## mom_age     -0.4253280  0.8748289
```

Controlling for a mom's age and score on an IQ test, kids with mothers who finished high school scored on average 5.6 points higher on the IQ test compared to kids whose mothers did not finish high school (95% CI: 1.2 - 10.1).
Could also say, **holding a mom's age and IQ test score constant**, kids with mothers...

How much variation is explained now?

$$
\begin{aligned}
R^2 &= 1 - \frac{\text{variation of model residuals}}{\text{total variability in } y} \\
&= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}
\end{aligned}
$$

```
1 - sum( resid(fm1)^2 ) / sum( (kidiq$kid_score-mean(kidiq$kid_score))^2 )
```

```
## [1] 0.2149895
```

```
summary(fm1)$r.squared
```

```
## [1] 0.2149895
```

```
summary(fm1)$adj.r.squared
```

```
## [1] 0.2095127
```

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.056	0.054
Model 2 (Multiple)	0.215	0.210

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2

$$R^2_{adj} = 1 - \frac{\text{variation of model residuals}}{\text{total variability in } y} \cdot \frac{n-1}{n-p-1}$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a (arbitrary, but useful) penalty for the number of predictors included in the model.
- Therefore, one model selection criteria could be to choose models with higher R^2_{adj} over others.

Taking stock

- First model: "mom finishing high school is really important."
- Second model: "meh. mom finishing high school is still important, but not as important after you account for her age and IQ score."
- This is an example of "confounding": predictor variables are correlated with each other and only looking at a subset of them can mask the true association.

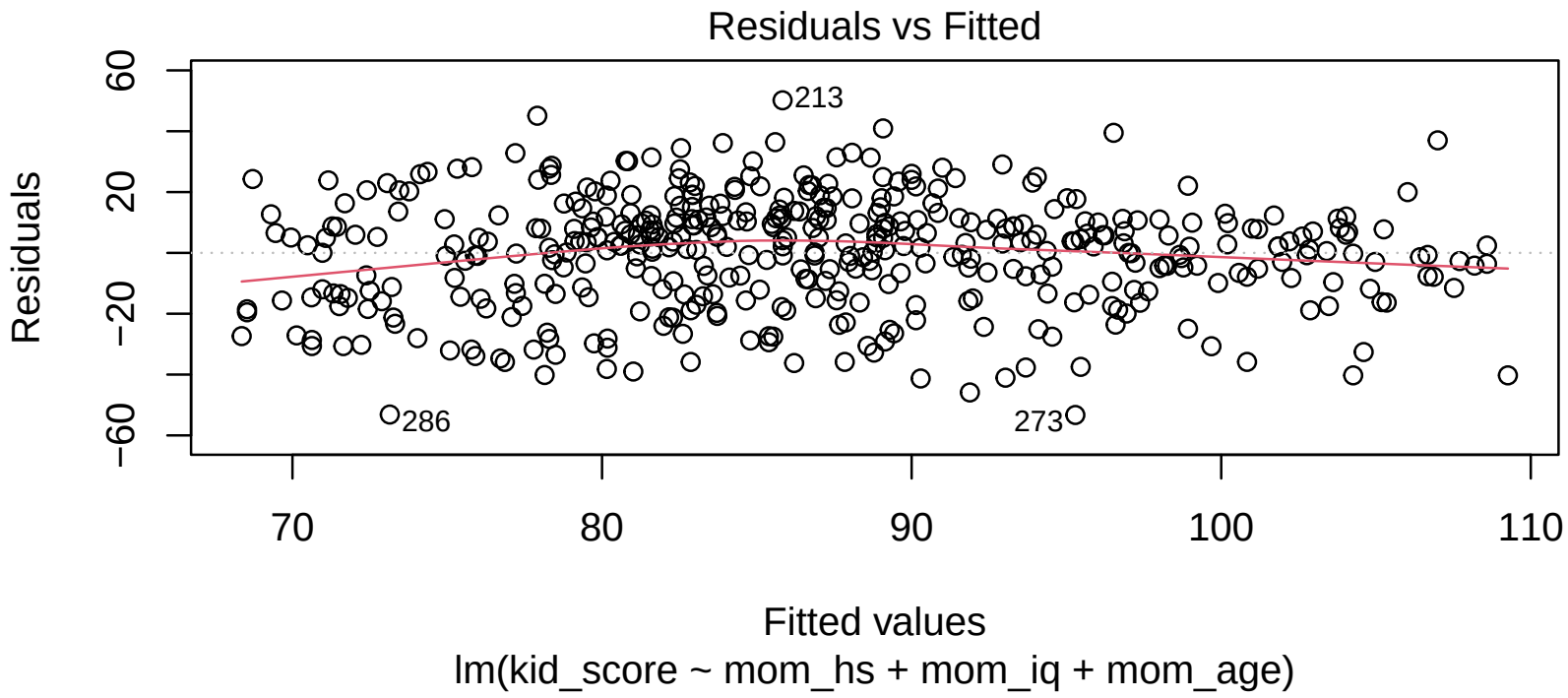
Model diagnostics

1. Check basic model assumptions.
2. Look at fitted relationships.
3. Look at predictions from the fitted model.

1. Check model assumptions

A smooth line through the residuals vs fitted plot should roughly have mean=0.

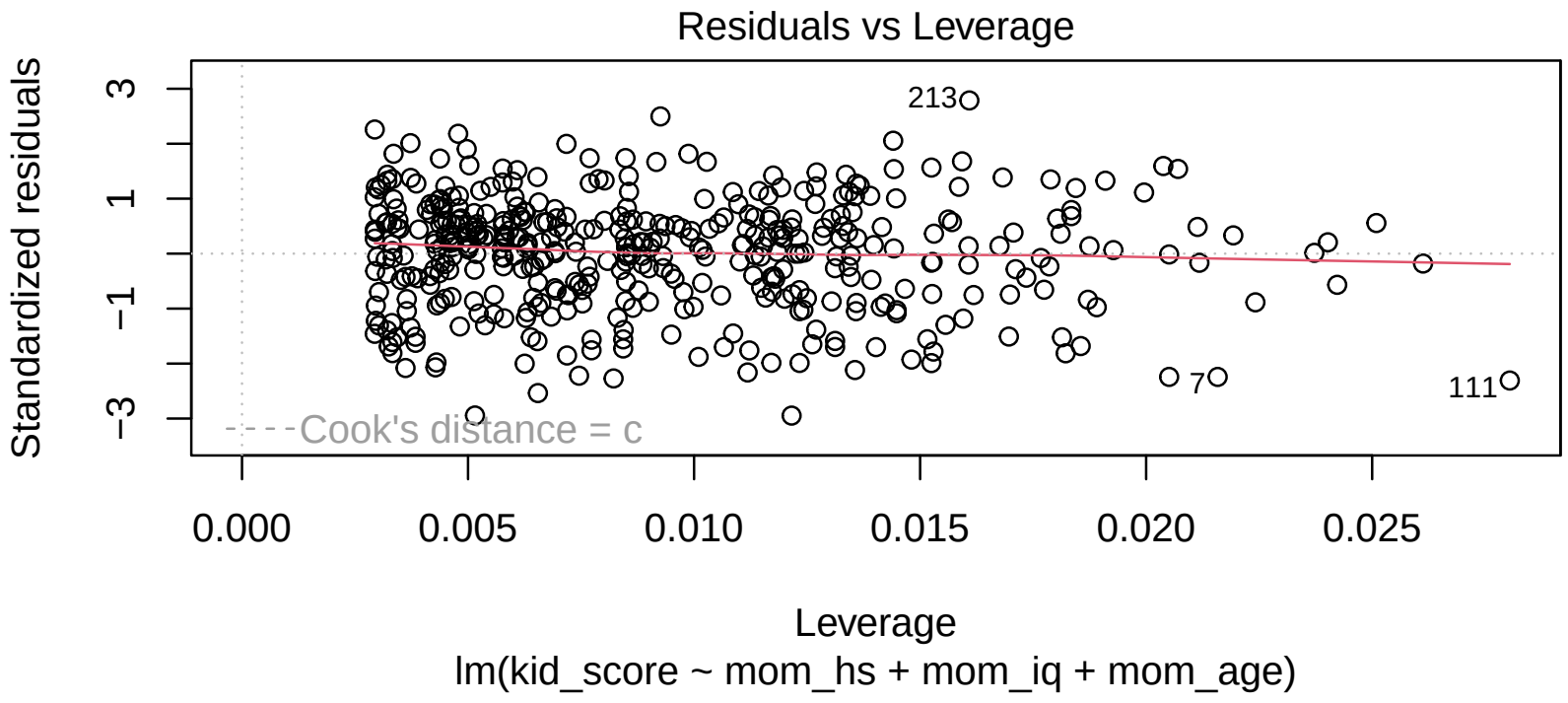
```
plot(fm1, which = 1)
```



1. Check model assumptions

Points that have high residual value (in either direction) and high leverage should be examined.

```
plot(fm1, which = 5)
```



2. Look at fitted relationships

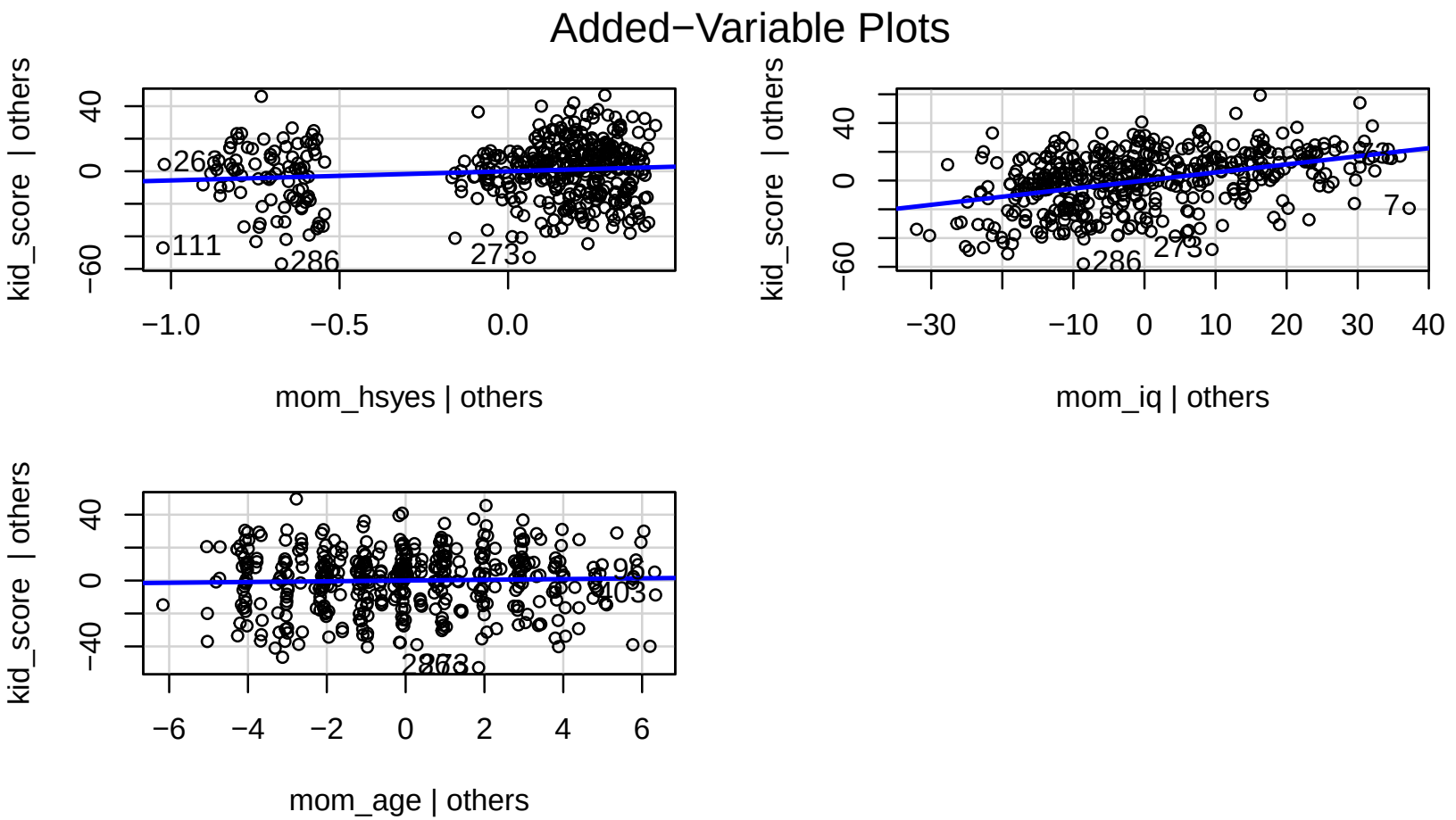
Added variable plots for the outcome Y and one predictor X show residuals from $Y \sim Z$ plotted against the residuals from $X \sim Z$, where Z are the other predictor variables. In this way, the plot shows the part of Y and X that are not explained by a linear relationship on the other variables. By mathematical definition the regression line through this plot has the same slope as the coefficient on X.

2. Added variable plots

```
coef(fm1)

## (Intercept)   mom_hsyes      mom_iq     mom_age
##  20.9846620   5.6471512   0.5625443   0.2247505

car::avPlots(fm1)
```



3. Look at model predictions

Note that predictions can be for the mean value of a particular covariate profile (type="confidence"), or for a new observed data point value (type = "prediction")

```
newdata <- data.frame(mom_hs = "yes", mom_iq = 100, mom_age = 27)
predict(fm1, newdata = newdata, interval = "confidence")

##        fit      lwr      upr
## 1 88.9545 85.74326 92.16575

predict(fm1, newdata = newdata, interval = "prediction")

##        fit      lwr      upr
## 1 88.9545 53.14235 124.7666
```