T15-03 Presentation

Yifan Jin, Stuart Morrison, Derek Ng, Christopher Saad

31/10/2019

Introduction

- Dataset representing prenatal risk factors and child birth weight
- We examine how certain risk factors mothers show while pregnant effect the infant's birth weight with a multiple lienar regression model
- We find that each of the risk factors we examine has a statistically significant effect on a infant's birth weight - ie, putting the infant at risk

The data we use in our analysis

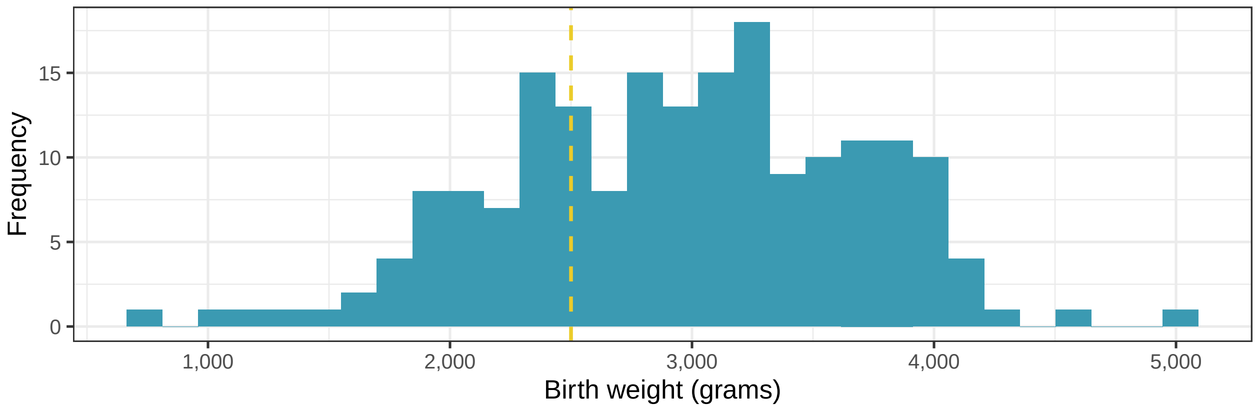
- Our data was originally collected by physicians in 1986 to examine the effects of risk factors on the birth weight of an infant (Hosmer, David W., et al., 2013) - included in the `MASS` package
- Low birth weight is associated with higher infant morality rates and birth defect rates
- It contains observations on 189 different pregnancies

Having a look at the data

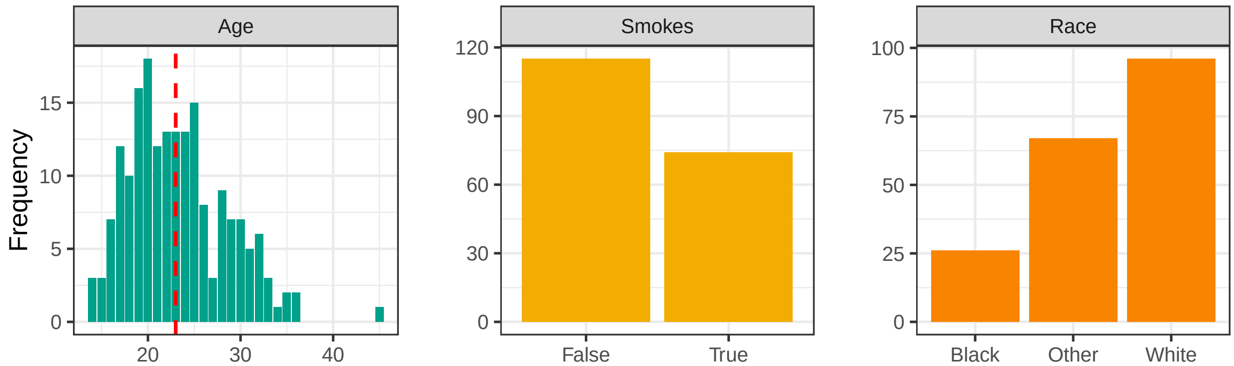
- There are 10 variables included in the data
- The key dependent variable for this analysis is:
  - `birthwt` - a variable measuring the infant's weight at birth
- The risk factors we examine are:
  - the racial background of the mother
  - whether the mother is a smoker
  - whether the mother has had a premature labour before
  - whether the mother experiences hypertension
  - whether the mother experiences physical irritability

The distribution of birth weights

- 31.2% of the infants in our data had a clinically low birth weight



The demographics represented in our data



Data cleaning and verascity

- No variables appear to have erroneous measurements - birth weights and ages are in in the expected range
- We convert categorical variables to factors
- Drop extraneous variables from the data, such as the indicator function `low` that shows whether the infant had a low birthweight

Model selection

- AIC is an indicator of quality of model - the less AIC the better quality we get
- Three approaches:
  - `Backward selection with AIC` - drop variables in full model to get model which has lower AIC
  - `Forward selection with AIC` - add variables in null model to get model which has lower AIC
  - `Backward selection with p-value` - drop the variable with largest p-value in orginal model

Model selection (cont.)

- Result:

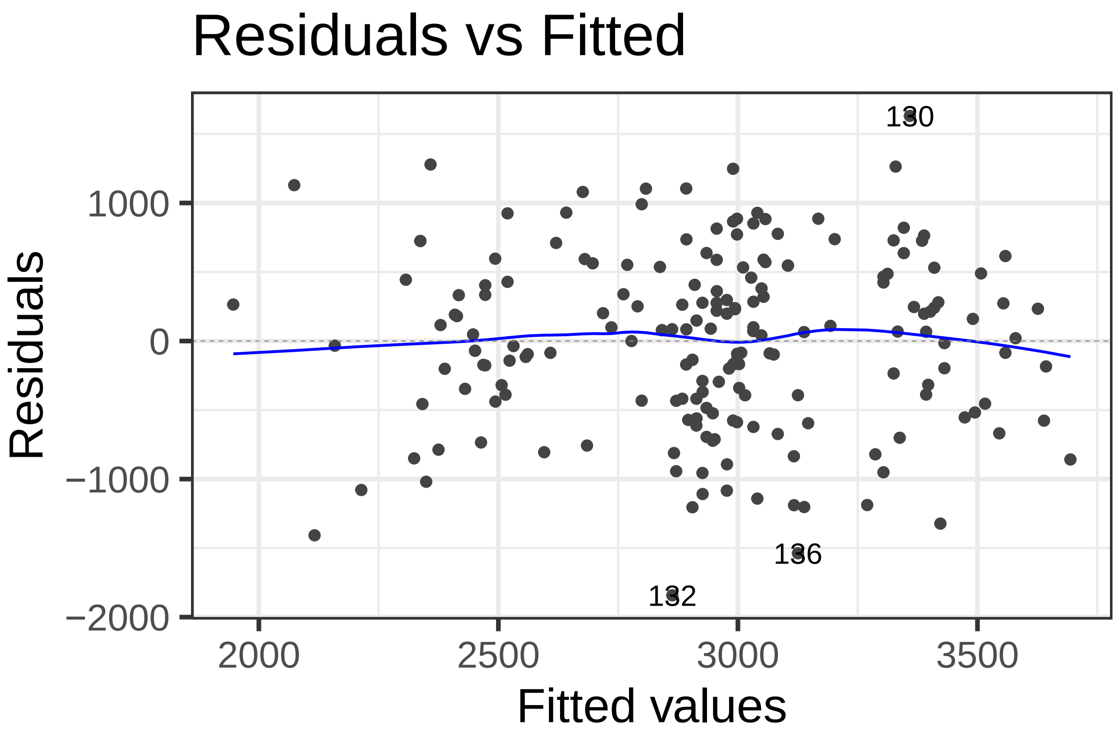
term	estimate	std.error	statistic	p.value
(Intercept)	2837.26	243.68	11.64	0.00
mother_weight	4.24	1.68	2.53	0.01
raceBlack	-475.06	145.60	-3.26	0.00
raceOther	-348.15	112.36	-3.10	0.00
smokesTrue	-356.32	103.44	-3.44	0.00
hypertension	-585.19	199.64	-2.93	0.00
uterine_irr	-525.52	134.68	-3.90	0.00

Assumptions checking

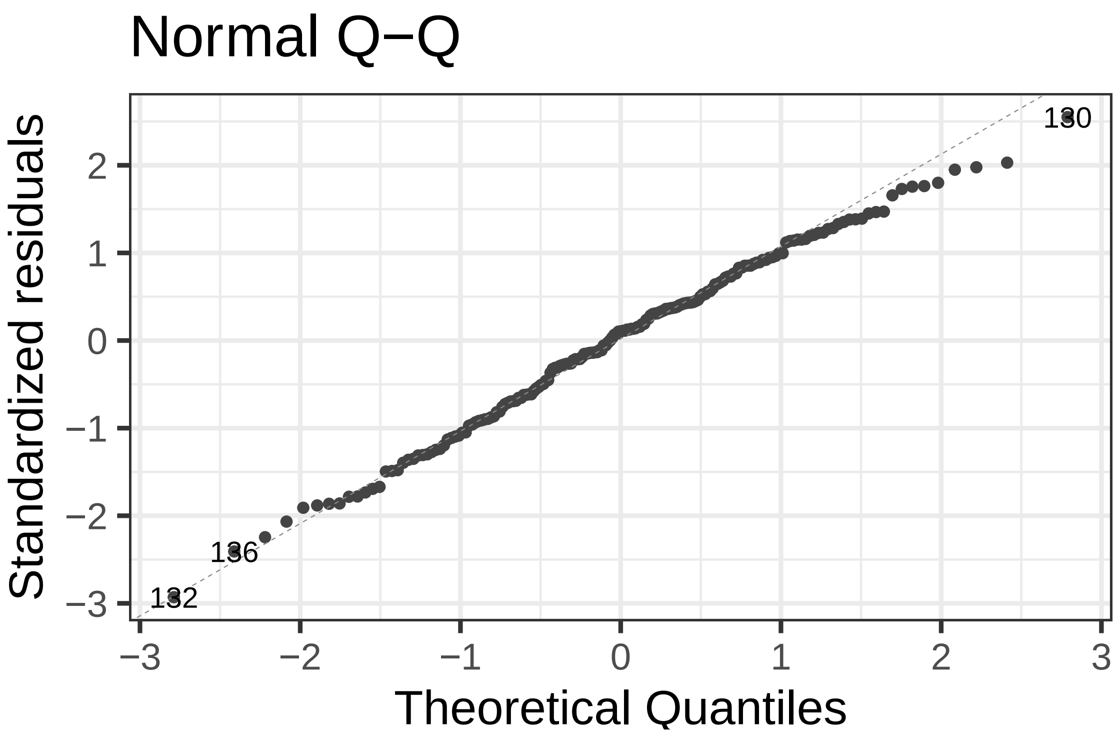
- The residuals ε_i are *iid* $N(0, \sigma^2)$ and there is a linear relationship between y and x.
  - Linearity
  - Independence
  - Homoskedasticity
  - Normality

Linearity, Independence and Homoskedasticity

```
## Warning: package 'ggfortify' was built under R version 3.5.2
```



Normality



Results - Interpretation of Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	2837.26	243.68	11.64	0.00
mother_weight	4.24	1.68	2.53	0.01
raceBlack	-475.06	145.60	-3.26	0.00
raceOther	-348.15	112.36	-3.10	0.00
smokesTrue	-356.32	103.44	-3.44	0.00
hypertension	-585.19	199.64	-2.93	0.00
uterine_irr	-525.52	134.68	-3.90	0.00

Results - In-sample Performance

- Full Model

```
summary(bwt_lm1)$r.squared
```

```
## [1] 0.2467462
```

- Simplified Model

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

```
## [1] 0.2403945
```

Results - Out-of-sample Performance

- Full Model

```
## Linear Regression
##
## 189 samples
##   8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 170, 170, 169, 169, 171, 170, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   663.3933  0.2298755  541.705
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Results - Out-of-sample Performance

- Simplified Model

```
## Linear Regression
##
## 189 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 170, 170, 169, 169, 169, 170, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   657.2103  0.1975652  533.7115
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Conclusion

- We examined how certain risk factors effect the infant's birth weight
- We applied a multiple linear regression model to delineate the effect of each of the factors
- We find that each of the risk factors we examine have a statistically significant effect on a infant's birth weight - ie, putting the infant at risk
- Potential limitation - data was recorded at a charity health centre so our data may not represent the population of mothers