STAT 213
Multi-category Variables

Colin Reimer Dawson

Oberlin College

March 5, 2018

Outline

Outline

Multiple Indicators

A question of interest is how birth weights (`BirthWeightOz`) in North Carolina might differ according to mother's race. In the dataset `NCbirths` (available in `Stat2Data`), the variable `MomRace` codes the mother's "race" as Black, Latinx, "Other"[1], or White. For the fitted model

$$\widehat{\texttt{BirthWeightOz}} = 117.87 + 7.96 \cdot \texttt{ILatinx} + 6.58 \cdot \texttt{IOther} + 7.31 \cdot \texttt{IWhite}$$

the predictors are equal to 1 when the mother identifies with the race in question, and zero otherwise. What does each coefficient tell us about race and birth weights? (Assume that each mother picks one category to identify with.)

[1] "Other" encompasses American Indian, Chinese, Japanese, Hawaiian, Filipino, and Other Asian or Pacific Islander

Reference Coding

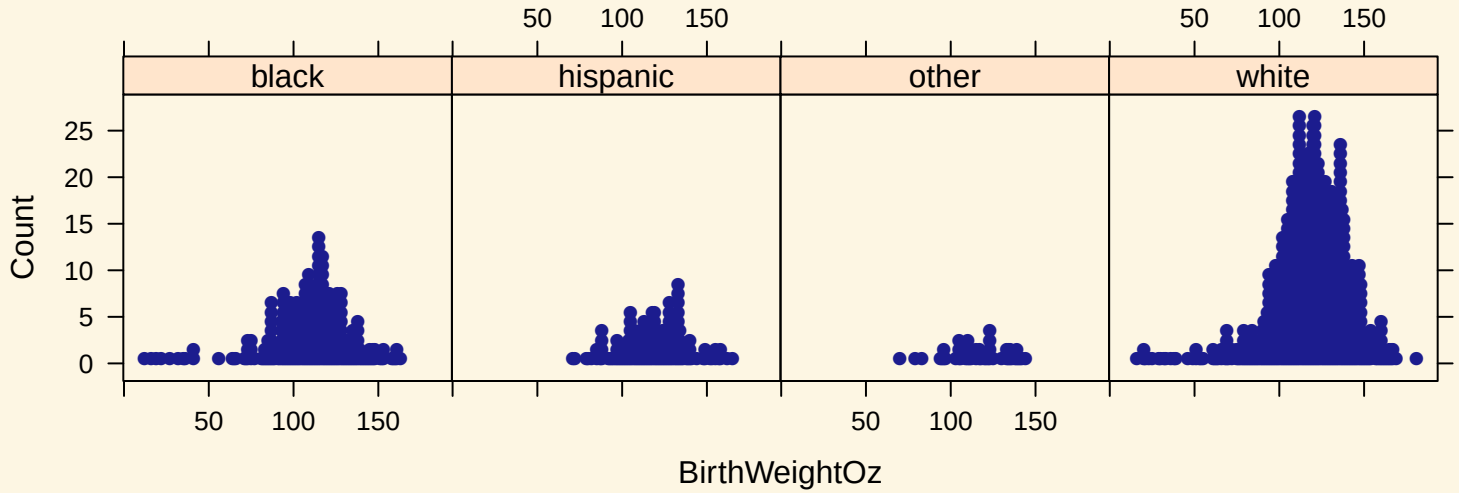
- In this model, one of the categories ("Black" in this case) is chosen as the *reference category*
- Each other category gets a binary indicator variable which is 1 for cases in that category
- All zeroes correspond to the reference category
- The "intercept" is then the prediction for the reference category
- Other coefficients represent differences vs. reference

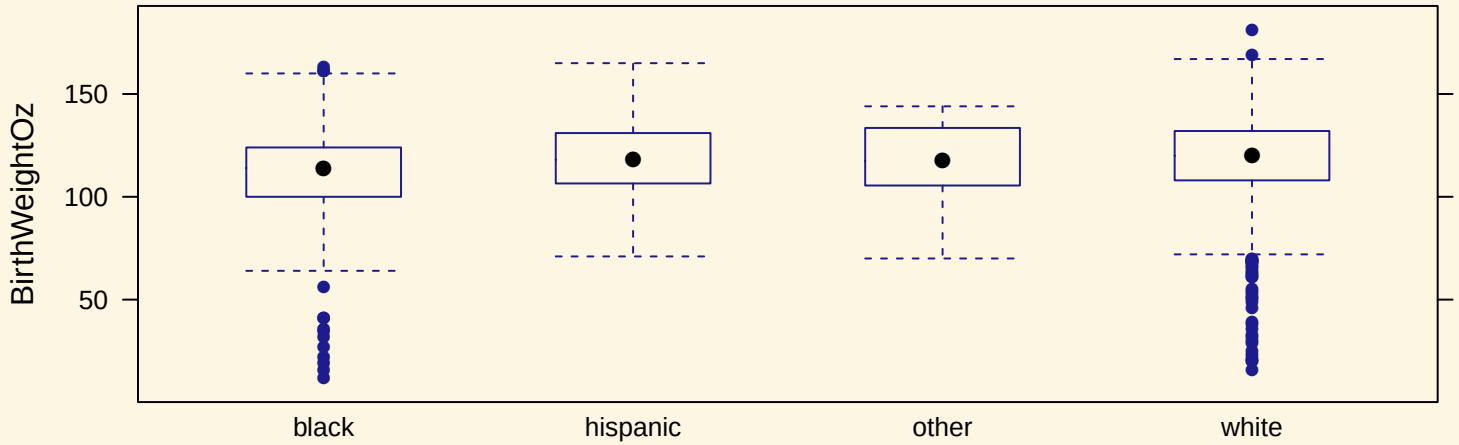
Two Representations

Case	BirthWeightOz	MomRace
1	125	white
2	108	hispanic
3	139	other
4	118	black
5	113	hispanic

Case	BirthWeightOz	IWhite	ILatinx	IOther
1	125	1	0	0
2	108	0	1	0
3	139	0	0	1
4	118	0	0	0
5	113	0	1	0

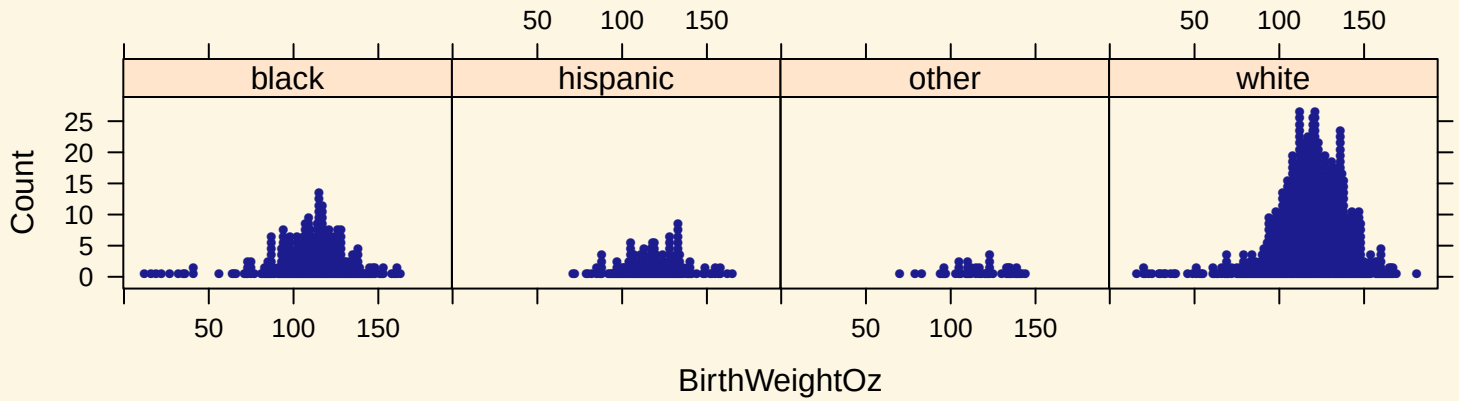
The Data


black
hispanic
other
white
Count
BirthWeightOz
0
5
10
15
20
25
50
100
150

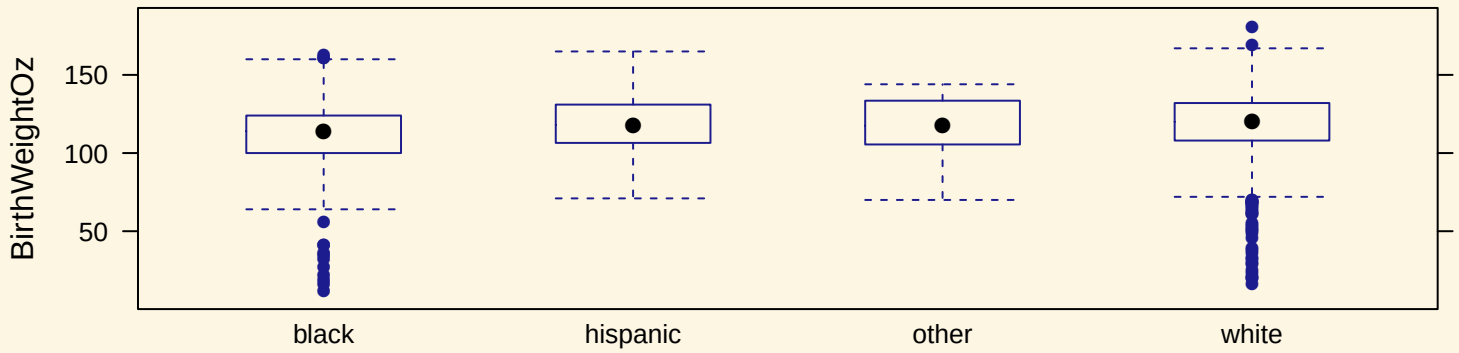

BirthWeightOz
50
100
150
black
hispanic
other
white

For Reference: R Code

```
library(Stat2Data); library(mosaic); data(NCbirths)
dotPlot(~BirthWeightOz | MomRace, data = NCbirths, width = 1, cex = 2)
```




```
bwplot(BirthWeightOz ~ MomRace, data = NCbirths)
```



Testing For Differences

```
bwmodel <- lm(BirthWeightOz ~ MomRace, data = NCbirths)
summary(bwmodel) %>% coefficients() %>% round(digits = 2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	110.56	1.21	91.02	0.00
MomRacehispanic	7.96	2.11	3.77	0.00
MomRaceother	6.58	3.42	1.93	0.05
MomRacewhite	7.31	1.42	5.15	0.00

Individual coefficients only tell us about differences involving the reference group

Testing For Differences

```
anova(bwmodel)

     Analysis of Variance Table

     Response: BirthWeightOz
                 Df Sum Sq Mean Sq F value    Pr(>F)
     MomRace      3  14002  4667.5  9.5282 3.118e-06 ***
     Residuals 1446 708332   489.9
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overall F-test asks "is there evidence that any of the means differ?" (equivalent H_0 is "all non-intercept coefficients are zero")

Evidence vs. Effect Size

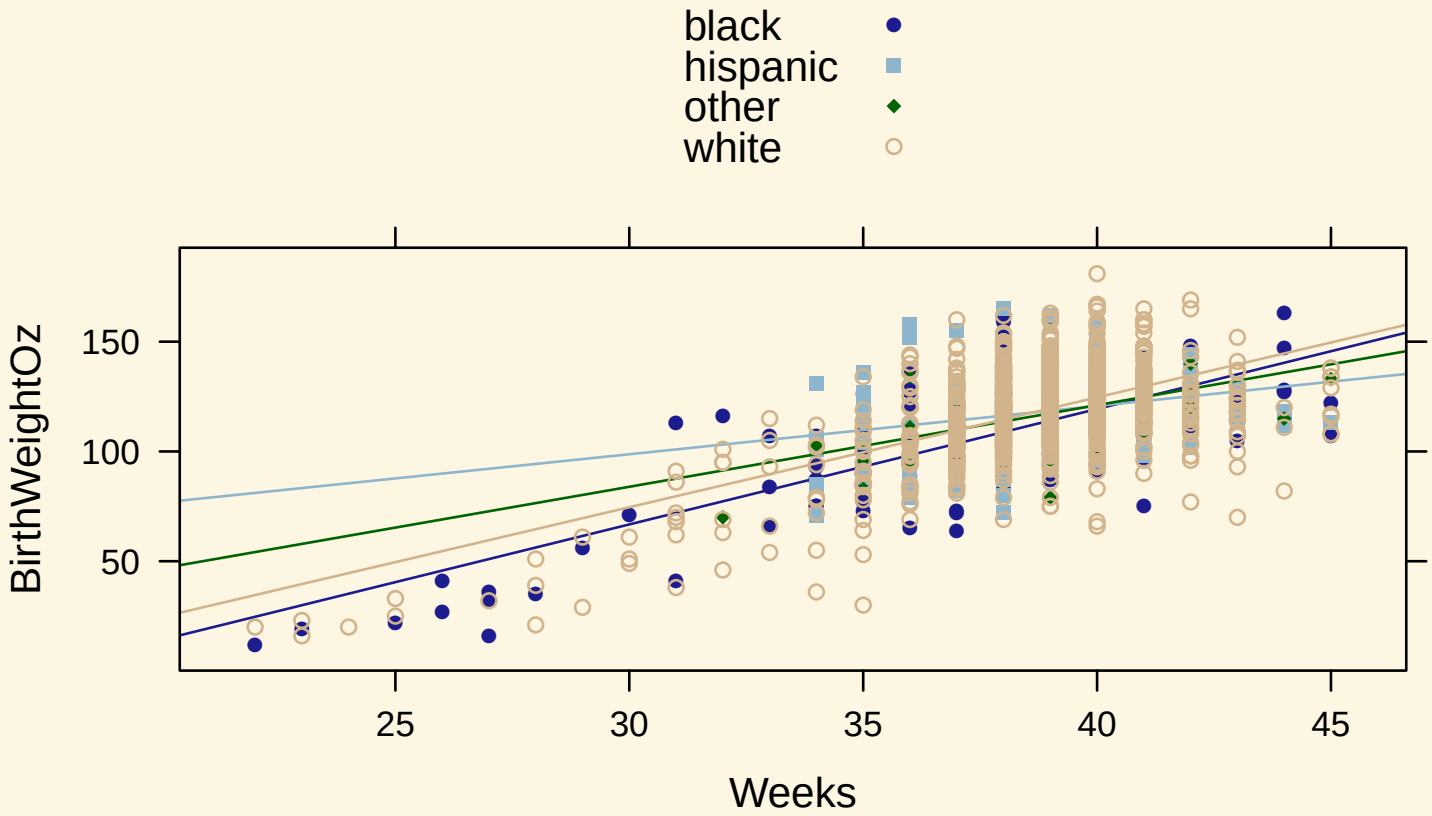
```
rsquared(bwmodel)

     [1] 0.0193849
```

Note that a significant F-test does not imply a large proportion of variance accounted for.

Outline

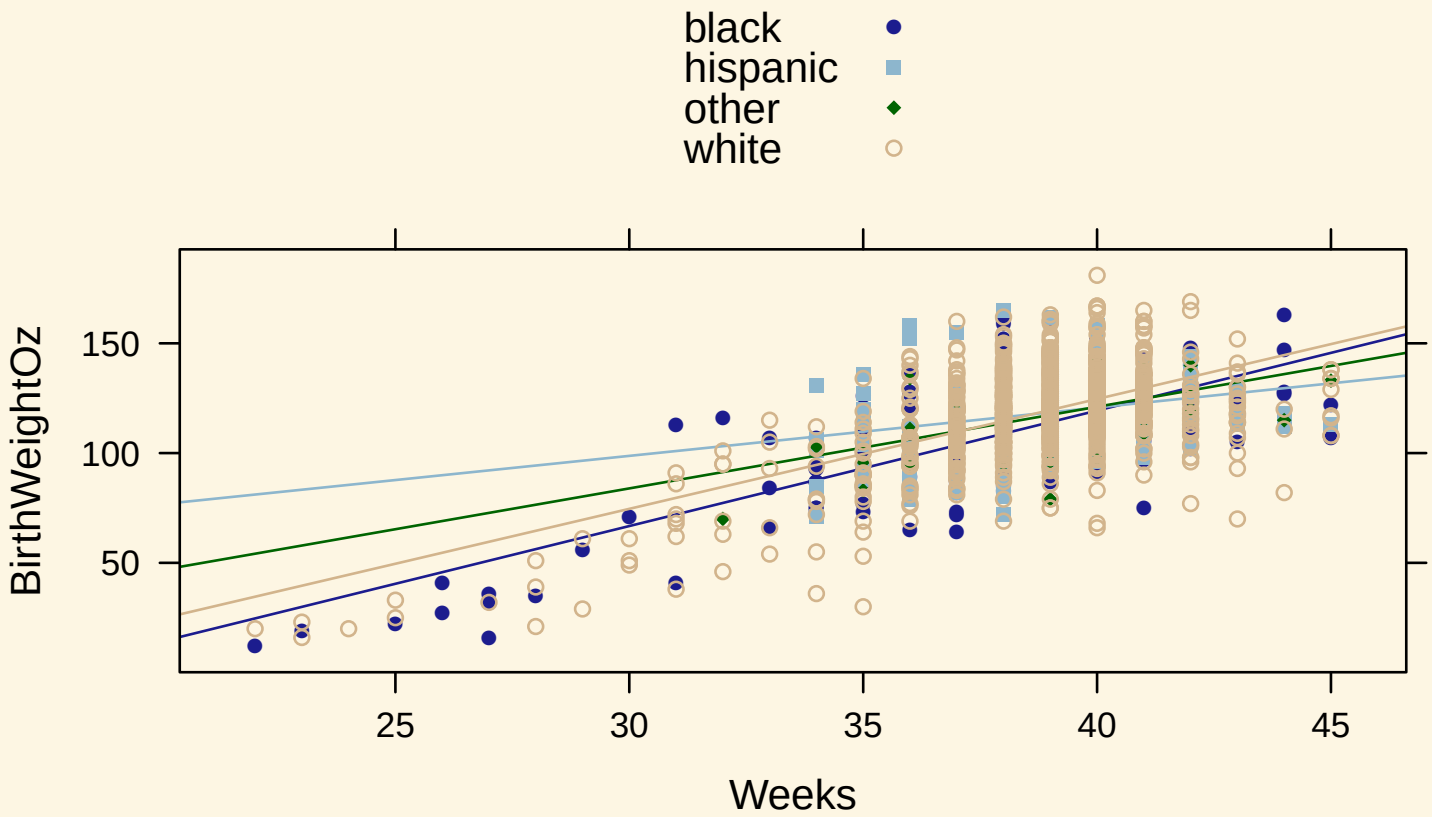
Many Lines Model

Suppose we are interested in the relationship between gestational age (`Weeks`) and birth weight, and whether this *relationship* differs by race.



For Reference: R Code

```
xyplot(BirthWeightOz ~ Weeks, groups = MomRace, data = NCbirths,
       type = c("p", "r"), auto.key = TRUE)
```



Regression Model

$$\widehat{\texttt{BirthWeightOz}} = \beta_0 + \beta_1 \texttt{ILatinx} + \beta_2 \texttt{IOther} + \beta_3 \texttt{IWhite} + \beta_4 \texttt{Weeks} + \beta_5 \texttt{ILatinx} \cdot \texttt{Weeks} + \beta_6 \texttt{IOther} \cdot \texttt{Weeks} + \beta_7 \texttt{IWhite} \cdot \texttt{Weeks}$$

What do these coefficients represent?

Regression Model

$$\widehat{\texttt{BirthWeightOz}} = \beta_0 + \beta_1 \texttt{ILatinx} + \beta_2 \texttt{IOther} + \beta_3 \texttt{IWhite} + \beta_4 \texttt{Weeks} + \beta_5 \texttt{ILatinx} \cdot \texttt{Weeks} + \beta_6 \texttt{IOther} \cdot \texttt{Weeks} + \beta_7 \texttt{IWhite} \cdot \texttt{Weeks}$$

$$\widehat{\texttt{BirthWeightOz}} = \begin{cases} \beta_0 + \beta_4 \texttt{Weeks} & \text{if } \texttt{MomRace} = \texttt{black} \\ (\beta_0 + \beta_1) + (\beta_4 + \beta_5)\texttt{Weeks} & \text{if } \texttt{MomRace} = \texttt{latinx} \\ (\beta_0 + \beta_2) + (\beta_4 + \beta_6)\texttt{Weeks} & \text{if } \texttt{MomRace} = \texttt{other} \\ (\beta_0 + \beta_3) + (\beta_4 + \beta_7)\texttt{Weeks} & \text{if } \texttt{MomRace} = \texttt{white} \end{cases}$$

Testing Differences in Slope

Full Model (many lines):

$$\widehat{\texttt{BirthWeightOz}} = \beta_0 + \beta_1 \texttt{ILatinx} + \beta_2 \texttt{IOther} + \beta_3 \texttt{IWhite} + \beta_4 \texttt{Weeks} + \beta_5 \texttt{ILatinx} \cdot \texttt{Weeks} + \beta_6 \texttt{IOther} \cdot \texttt{Weeks} + \beta_7 \texttt{IWhite} \cdot \texttt{Weeks}$$

Reduced Model (parallel lines):

$$\widehat{\texttt{BirthWeightOz}} = \beta_0 + \beta_1 \texttt{ILatinx} + \beta_2 \texttt{IOther} + \beta_3 \texttt{IWhite} + \beta_4 \texttt{Weeks}$$

Nested test: $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$

Nested Test for Differences in Slope

```
full.model <- lm(BirthWeightOz ~ MomRace * Weeks, data = NCbirths)
reduced.model <- lm(BirthWeightOz ~ MomRace + Weeks, data = NCbirths)
anova(reduced.model, full.model)
```

```
Analysis of Variance Table

Model 1: BirthWeightOz ~ MomRace + Weeks
Model 2: BirthWeightOz ~ MomRace * Weeks
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1444 460307
2   1441 454142  3    6165.8 6.5214 0.0002218 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: There is significant *evidence* that the *relationship* between gestational age and birthweight differs by mother's race

Effect Size

But still...

```
## Even further reduced model
baseline.model <- lm(BirthWeightOz ~ Weeks, data = NCbirths)
## 35% of variability in birth weight accounted for by gest. age
rsquared(baseline.model)
```

```
     [1] 0.3512314
```

```
## This goes up to ~37% if we also take mom's race into account
rsquared(full.model)
```

```
     [1] 0.370946
```

```
## Only an additional ~2% accounted for by mom race when
## controlling for gest. age
rsquared(full.model) - rsquared(baseline.model)
```

```
     [1] 0.01971462
```