Week 8 R Code demo (sample solutions)

Prof Moon is expecting a baby in April. In this demo, we'll explore the NCbirths data to investigate how big we should expect her baby to be!

```
library(tidyverse) # As usual
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#install.packages("Stat2Data")
library(Stat2Data) # This package contains the NCbirths data

# Sometimes, we need to use the data() function to load data from an R package into our R session
data(NCbirths) # After running this line, you'll see it appear in the Environment pane in the top left
glimpse(NCbirths)
```

```
## Rows: 1,450
## Columns: 15
## $ ID            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...
## $ Plural        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ Sex           <int> 1, 2, 1, 1, 1, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1,...
## $ MomAge        <int> 32, 32, 27, 27, 25, 28, 25, 15, 21, 27, 26, 20, 19, 1...
## $ Weeks         <int> 40, 37, 39, 39, 39, 43, 39, 42, 39, 40, 41, 41, 40, 3...
## $ Marital       <int> 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 2, 2, 1, 1, 1,...
## $ RaceMom       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 5, 1, 1,...
## $ HispMom       <fct> N, N, N, N, N, N, N, N, N, N, N, N, N, P, N, M, N, N,...
## $ Gained        <int> 38, 34, 12, 15, 32, 32, 75, 25, 28, 37, 45, 52, 26, 3...
## $ Smoke         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0,...
## $ BirthWeightOz <int> 111, 116, 138, 136, 121, 117, 143, 113, 120, 124, 121...
## $ BirthWeightGm <dbl> 3146.85, 3288.60, 3912.30, 3855.60, 3430.35, 3316.95,...
## $ Low           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ Premie        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ MomRace       <fct> white, white, white, white, white, white, white, whit...
```

R Demo

Q1 - What do we know about these data? (and where can we find out?)

(a) When we load data from a package, where do we go to (hopefully) get more information about the data?

```
#?NCbirths  # Type '?', then the name of the dataset (or you can type the name of a function to get mor
```

(b) How many observations are there in these data?

1450

(c) What does each observation represent?

A birth in North Carolina in 2001

(d) Who contributed these data?

John Holcomb from Cleveland State University

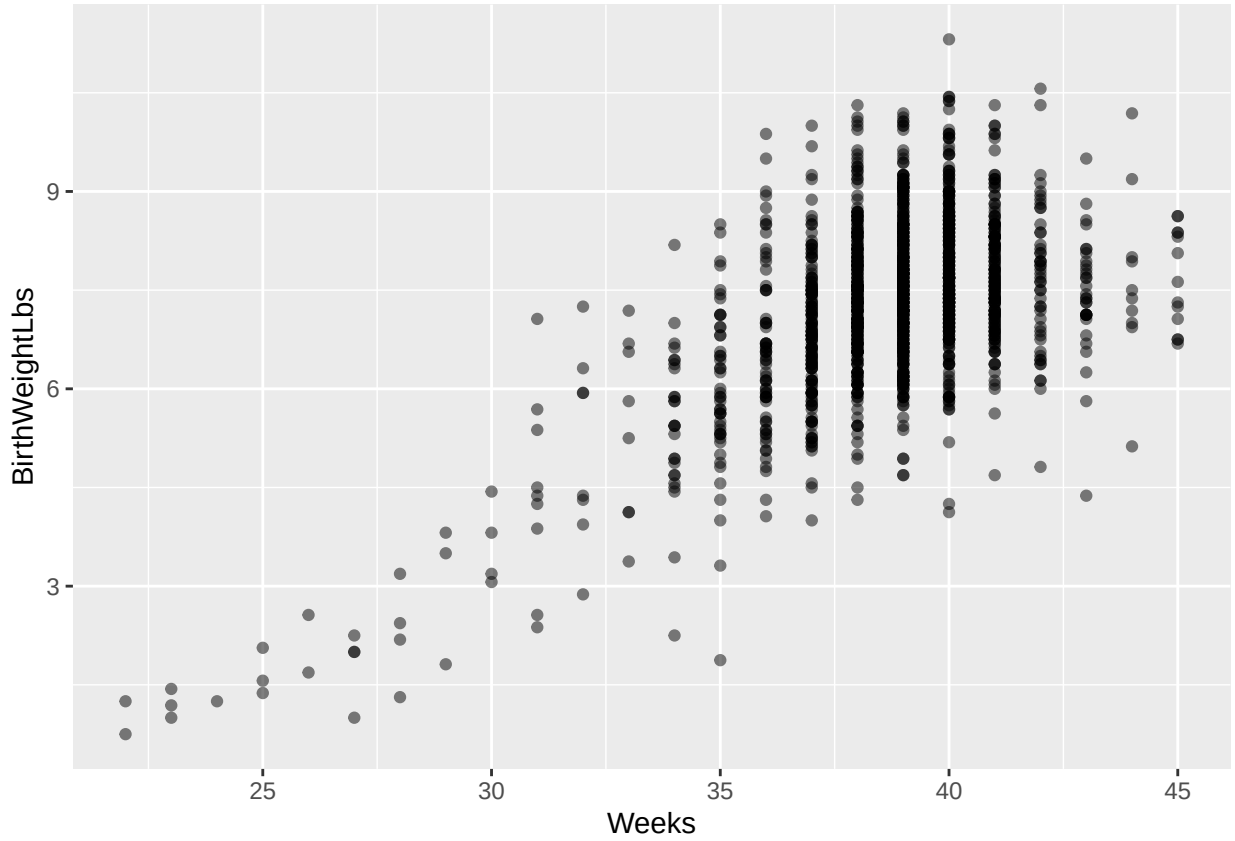
Q2 - Prof Moon is more familiar with measuring babies' weights in pounds, so let's create a new variable called `BirthWeightLbs`. We'll use the fact that there are 16 oz in 1 lb.

```
NCbirths <- NCbirths %>% mutate(BirthWeightLbs = BirthWeightOz / 16)
```

Q3 - Create a visualization to explore the association between the duration of a pregnancy (in weeks) and the baby's weight at birth. Describe this distribution.

```
NCbirths %>% ggplot(aes(x=Weeks, y=BirthWeightLbs)) +
  geom_point(alpha=0.5)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



The association between duration of pregnancy (in weeks) and babies' weight is approximately linear, although the association is much stronger before roughly 32 weeks and much weaker from ~32 weeks to 45 weeks. There are many more observations after ~35 weeks of pregnancy, which makes sense since these correspond to full term births. The association between weeks of pregnancy and birthweight is positive, which again makes sense - babies born later in a pregnancy have more time to grow and gain weight before birth.

Q4 - Prof Moon is currently 34 weeks pregnant, so she is particularly interested in learning about the association between the duration of pregnancy and birthweight for babies born at 34 weeks or later (babies born before this time are quite premature and so their birthweight is not be as relevant). Let's create a new tibble containing only observations for babies born in this range.

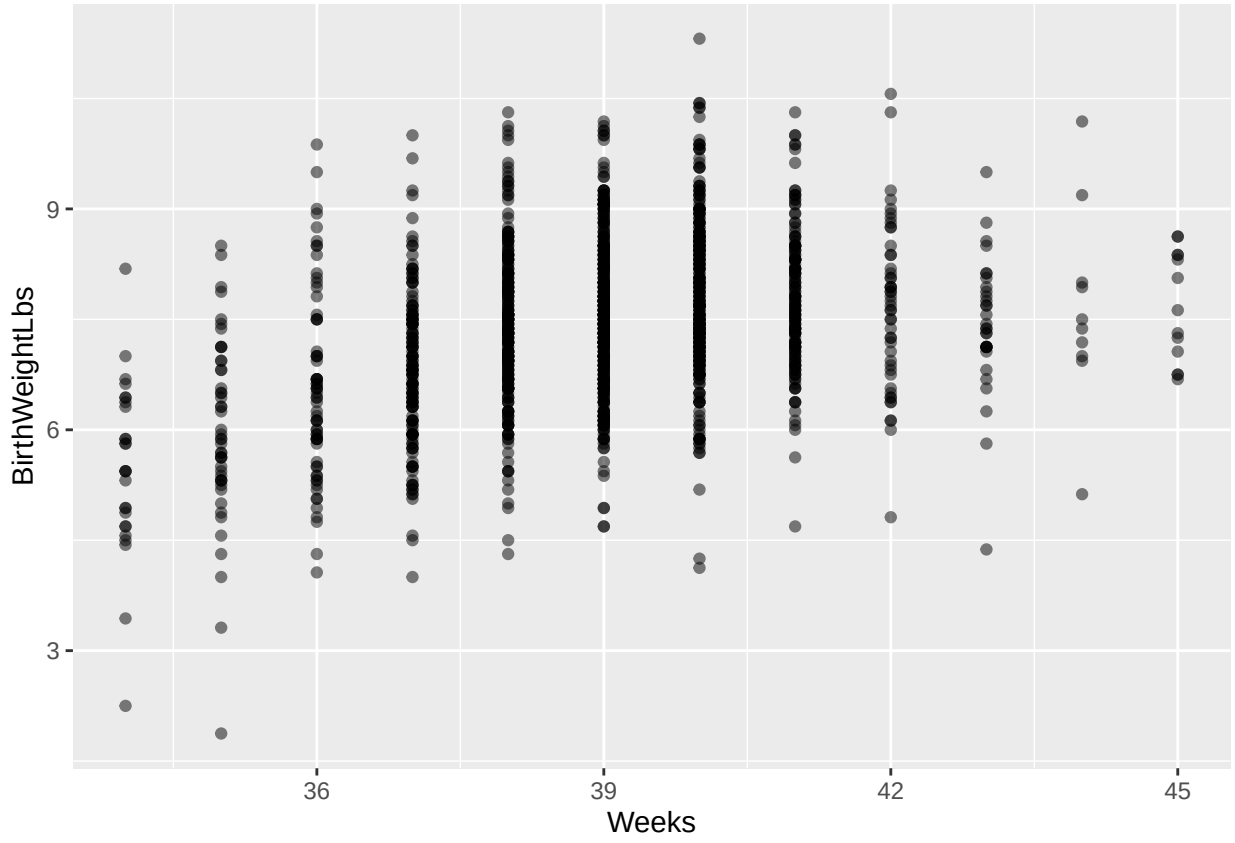
```
NCbirths_34plus <- NCbirths %>% filter(Weeks >= 34)

glimpse(NCbirths_34plus)
```

```
## Rows: 1,398
## Columns: 16
## $ ID            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
## $ Plural        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ Sex           <int> 1, 2, 1, 1, 1, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1...
## $ MomAge        <int> 32, 32, 27, 27, 25, 28, 25, 15, 21, 27, 26, 20, 19, ...
## $ Weeks         <int> 40, 37, 39, 39, 39, 43, 39, 42, 39, 40, 41, 41, 40, ...
## $ Marital       <int> 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 2, 2, 1, 1, 1...
## $ RaceMom       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 5, 1, 1...
```

```
## $ HispMom        <fct> N, N, N, N, N, N, N, N, N, N, N, N, N, P, N, M, N, N...
## $ Gained         <int> 38, 34, 12, 15, 32, 32, 75, 25, 28, 37, 45, 52, 26, ...
## $ Smoke          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0...
## $ BirthWeightOz  <int> 111, 116, 138, 136, 121, 117, 143, 113, 120, 124, 12...
## $ BirthWeightGm  <dbl> 3146.85, 3288.60, 3912.30, 3855.60, 3430.35, 3316.95...
## $ Low            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ Premie         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ MomRace        <fct> white, white, white, white, white, white, white, whi...
## $ BirthWeightLbs <dbl> 6.9375, 7.2500, 8.6250, 8.5000, 7.5625, 7.3125, 8.93...
```

```
NCbirths_34plus %>% ggplot(aes(x=Weeks, y=BirthWeightLbs)) +
  geom_point(alpha=0.5)
```



Part A

In this part, we'll explore the association between the duration of a pregnancy and birthweight (after 34 weeks gestation).

QA3 - Calculate the correlation between the number of weeks of pregnancy and birthweight (focusing only on babies born after 34 weeks gestation). Does the correlation reflect our intuition about the association between these two variables?

```
cor(NCbirths_34plus$Weeks, NCbirths_34plus$BirthWeightLbs)
```

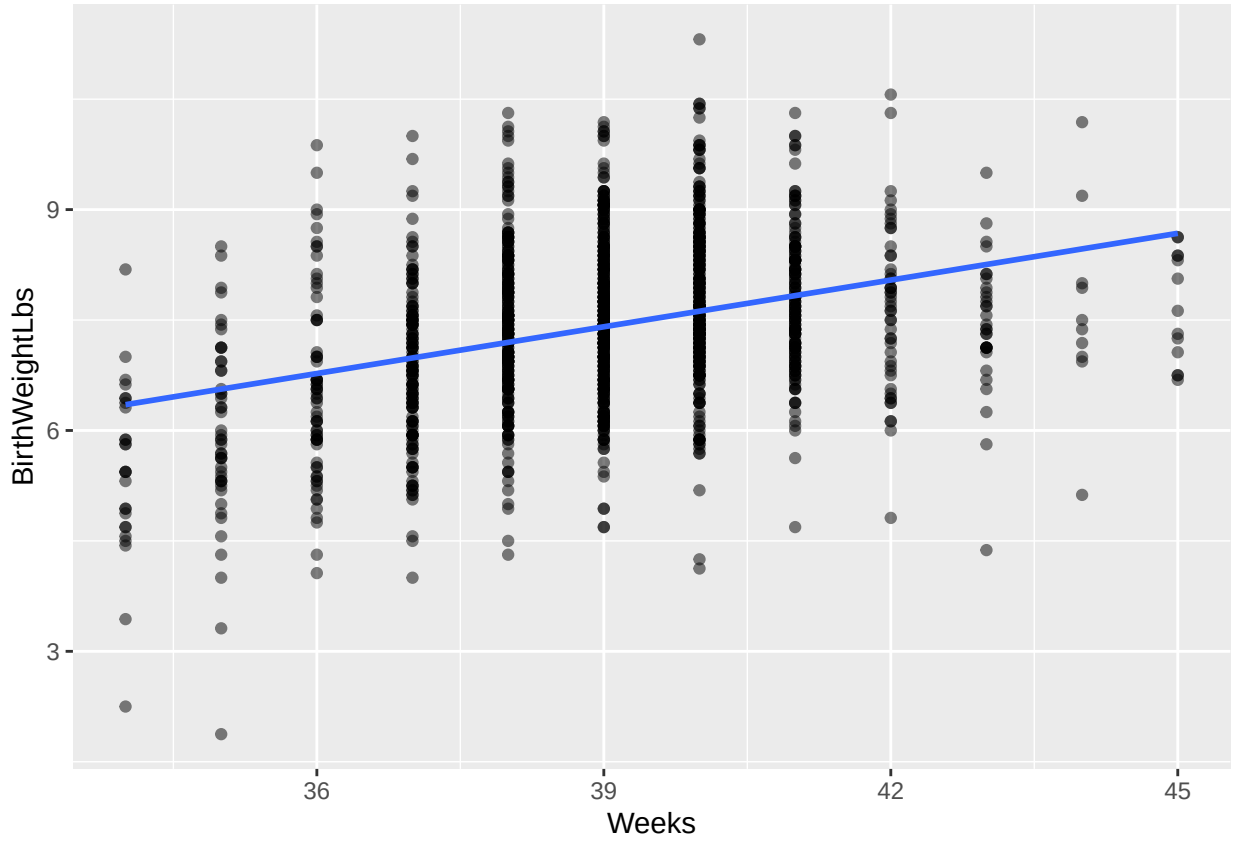
```
## [1] 0.3531734
```

QA4 - Fit a linear regression model to predict the weight of a baby based on the duration of the pregnancy (in weeks), to find the least-squares estimates for the regression parameters. Create a visualization to visualize the association between weeks and birthweight for babies born at or after 34 weeks, adding the fitted regression line to the plot.

```
model1 <- lm(BirthWeightLbs ~ Weeks, data = NCbirths_34plus)
summary(model1)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.8428561 0.58519197 -1.440307 1.500048e-01
## Weeks        0.2115674 0.01499991 14.104575 2.477011e-42
```

```
NCbirths_34plus %>% ggplot(aes(x=Weeks, y=BirthWeightLbs)) +
  geom_point(alpha=0.5) +
  geom_smooth(method="lm", se=FALSE, formula = y~x) # Note that includeing `formula=y~x` is not require
```



QA5 - How can we interpret the estimated slope (beta1-hat)?

The estimated slope $\hat{\beta}_1 = 0.212$, which means that on average, babies born one week later (after 34 weeks) weigh 0.212 lbs more at birth.

QA6 - How can we interpret the estimated intercept (beta0-hat)?

It does not make sense to interpret the estimated intercept $\hat{\beta}_0 = -0.843$. In therory, this value corresponds to the predicted weight of a baby born after 0 weeks of pregnancy, but it is neither possible to talk about babies born after 0 weeks of pregnancy, nor is it possible to have a negative birthweight. However, this estimate is still important as it tells us the height of the fitted regression line.

QA7 - Based on these data, is there evidence that the slope of this linear regression model significantly different from 0?

To answer this question, we need to look at the p-value corresponding to testing $H_0 : \hat{\beta}_1 = 0$ vs $H_A : \hat{\beta}_1 \neq 0$, which is $2.4770112 \times 10^{-42}$. Since this pvalue is very small, we have very strong evidence against the null hypothesis that the slope of this linear regression model is equal to 0.

QA8 - What proportion of the variability in weight is explained by our regression model? What does this suggest?

```
summary(model1)$r.squared
```

```
## [1] 0.1247314
```

```r
cor(NCbirths_34plus$Weeks, NCbirths_34plus$BirthWeightLbs)^2
```

```
## [1] 0.1247314
```

By calculating the R^2 value corresponding to our fitted regression model, we see that only 12.5 % of the variation in birthweight after 34 weeks gestation is explained by our simple linear regression model (with only weeks of gestation as a predictor).

This suggests that there are lots of other sources of variation affecting birthweight, apart from the weeks of gestation (i.e. other factors that influence the weight of a baby at birth). In Module 9, we'll talk about how to build richer linear regression models with *more than one* predictor!

Part B

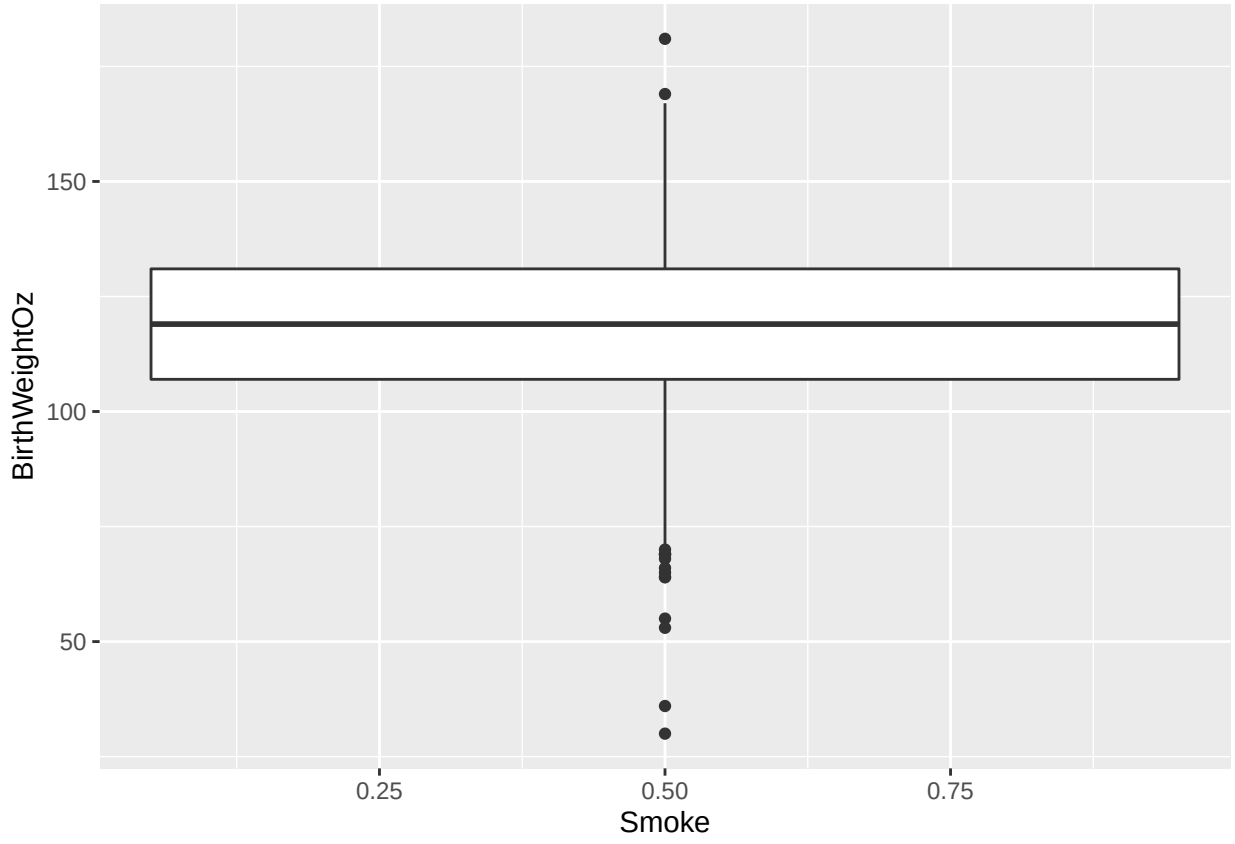
In this part, we'll explore the association between a mothers' smoking status and birthweight in infants born after 34 weeks gestation.

QB1 - Let's create a visualization to explore the association between mothers' smoking status and birthweight (after 34 weeks), and describe what we observe.

```
NCbirths_34plus %>% ggplot(aes(x=Smoke, y=BirthWeightOz)) +
  geom_boxplot()
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```

```
## Warning: Removed 5 rows containing missing values (stat_boxplot).
```



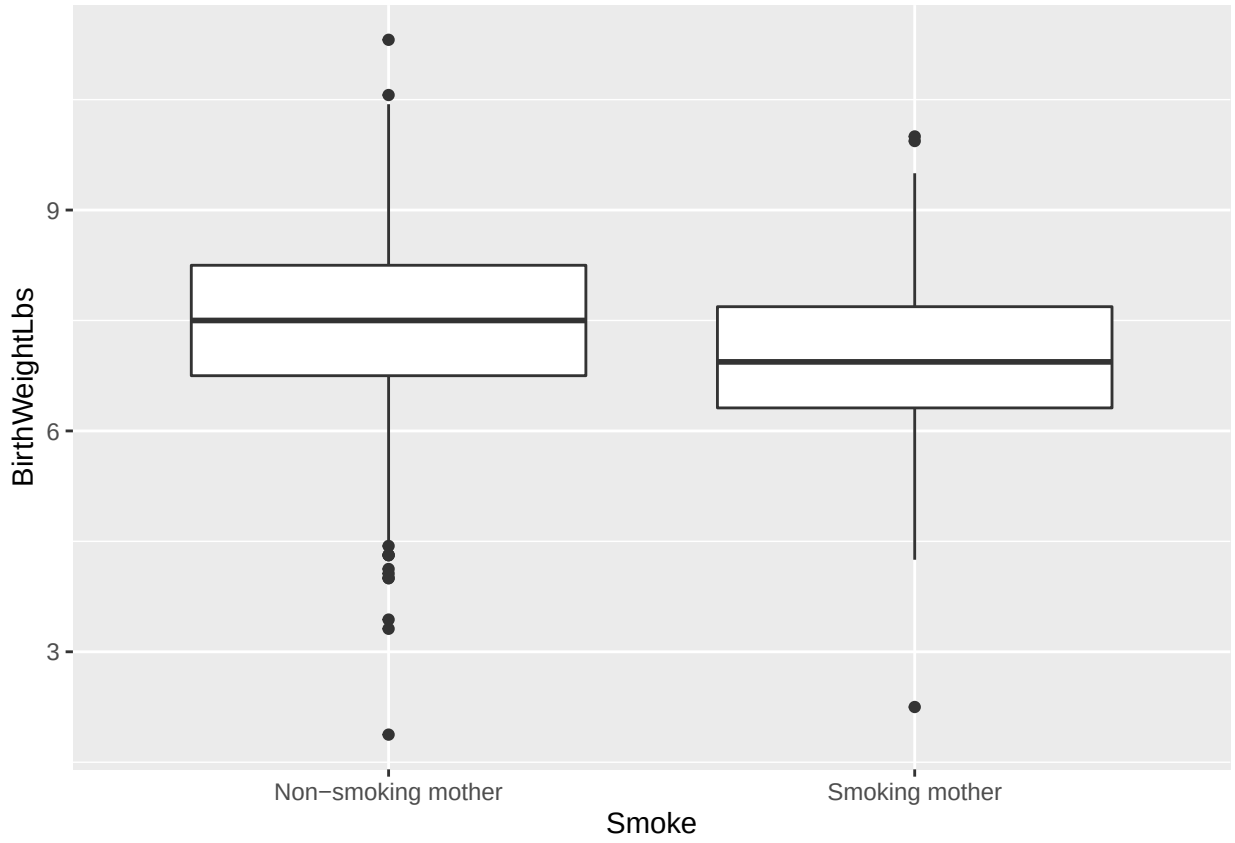

```
## What's weird here?
## 1. We get a warning saying the x aesthetic is continuous (when it should be categorical instead)
## 2. Some observations have missing values so are omitted

NCbirths_34plus <- NCbirths_34plus %>% filter(!is.na(Smoke) & !is.na(BirthWeightLbs)) %>%
  mutate(Smoke = case_when(Smoke == 0 ~ "Non-smoking mother",
                           Smoke == 1 ~ "Smoking mother"))
glimpse(NCbirths_34plus)
```

```
## Rows: 1,393
## Columns: 16
## $ ID             <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
```

```
## $ Plural         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ Sex            <int> 1, 2, 1, 1, 1, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1...
## $ MomAge         <int> 32, 32, 27, 27, 25, 28, 25, 15, 21, 27, 26, 20, 19, ...
## $ Weeks          <int> 40, 37, 39, 39, 39, 43, 39, 42, 39, 40, 41, 41, 40, ...
## $ Marital        <int> 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 2, 1, 2, 2, 1, 1, 1...
## $ RaceMom        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 5, 1, 1...
## $ HispMom        <fct> N, N, N, N, N, N, N, N, N, N, N, N, N, P, N, M, N, N...
## $ Gained         <int> 38, 34, 12, 15, 32, 32, 75, 25, 28, 37, 45, 52, 26, ...
## $ Smoke          <chr> "Non-smoking mother", "Non-smoking mother", "Non-smo...
## $ BirthWeightOz  <int> 111, 116, 138, 136, 121, 117, 143, 113, 120, 124, 12...
## $ BirthWeightGm  <dbl> 3146.85, 3288.60, 3912.30, 3855.60, 3430.35, 3316.95...
## $ Low            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ Premie         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ MomRace        <fct> white, white, white, white, white, white, white, whi...
## $ BirthWeightLbs <dbl> 6.9375, 7.2500, 8.6250, 8.5000, 7.5625, 7.3125, 8.93...
```

```
NCbirths_34plus %>% ggplot(aes(x=Smoke, y=BirthWeightLbs)) +
  geom_boxplot()
```



Babies born to non-smoking mothers tend to have higher birthweights than babies born to mothers who smoke, although there is a lot of overlap between the distribution of birthweights for these two groups. The median birthweight is slightly higher for non-smoking than smoking mothers (~7lbs vs ~6.7bls), and in both cases, the distributions of birthweight are approximately symmetrical.
The range of birthweights is larger for non-smoking mothers than for smoking mothers.

QB2 - Let's fit a linear regression model to predict the birthweight of a baby based on his/her mother's smoking status.

```
model2 <- lm(BirthWeightLbs ~ Smoke, data=NCbirths_34plus)
summary(model2)$coefficients
```

```
##                       Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)           7.469795 0.03343417 223.418002 0.0000e+00
## SmokeSmoking mother -0.481851 0.08890639  -5.419757 7.0262e-08
```

QB3 - What is the baseline level in the model above?

The baseline level is "Non-smoking mothers", because it is the one that does *not* appear in the table of coefficients produced in the previous question.

QB4 - How can we interpret the estimated intercept (beta0-hat)?

The estimated intercept $\hat{\beta}_0 = 7.47$ is the average birthweight for babies born to non-smoking mothers, based on this sample.
Another way of saying this is that it is the best prediction we can make, based on this fitted model, for the average birthweight of babies born to non-smoking mothers.

```
# Checking our work
NCbirths_34plus %>% group_by(Smoke) %>%
  summarise(mean = mean(BirthWeightLbs))
```

```
## # A tibble: 2 x 2
##   Smoke               mean
## * <chr>              <dbl>
## 1 Non-smoking mother  7.47
## 2 Smoking mother      6.99
```

```
# The numbers here match up with the values we get from the fitted regression model :)
```

QB5 - How can we interpret the estimated slope (beta1-hat)?

The estimated slope $\hat{\beta}_1 = -0.48$ is the average difference between the birthweight of babies born to smoking mothers vs non-smoking mothers based on this sample. In other words, babies born to smoking mothers are, on average, 0.48 pounds less heavy at birth than babies born to non-smoking mothers, based on this fitted model.

QB6 - Based on these data, is there a difference between the mean weight of babies born to smoking vs non-smoking mothers?

To answer this question, we need to look at the p-value corresponding to testing $H_0 : \hat{\beta}_1 = 0$ vs $H_A : \hat{\beta}_1 \neq 0$, which is 7.0261998×10^{-8}. Since this pvalue is very small, we have very strong evidence against the null hypothesis that the slope of this linear regression model is equal to 0.

Conclusion

Where do we go from here?

In the Week 9 module, we'll look at how to build richer linerar regression models with more than 1 predictor. For example, it's clear based on these examples that there are many factors that affect an infants babies' birthweight, and looking at these potential predictors one at a time is not enough to make good predictions.