Module 7: Linear regression

Jianhui Gao

July 18, 2025

Linear regression in R

```
library(tidyverse) #ggplot2, dplyr, etc.
library(reshape2) #need this for melt()
library(knitr) #need this for kable
library(MASS) #contains dataset
```

Load the birthwt data. This data contains 189 observations, 9 predictors, and an outcome, birthweight, available both as a continuous measure and a binary indicator for low birth weight.

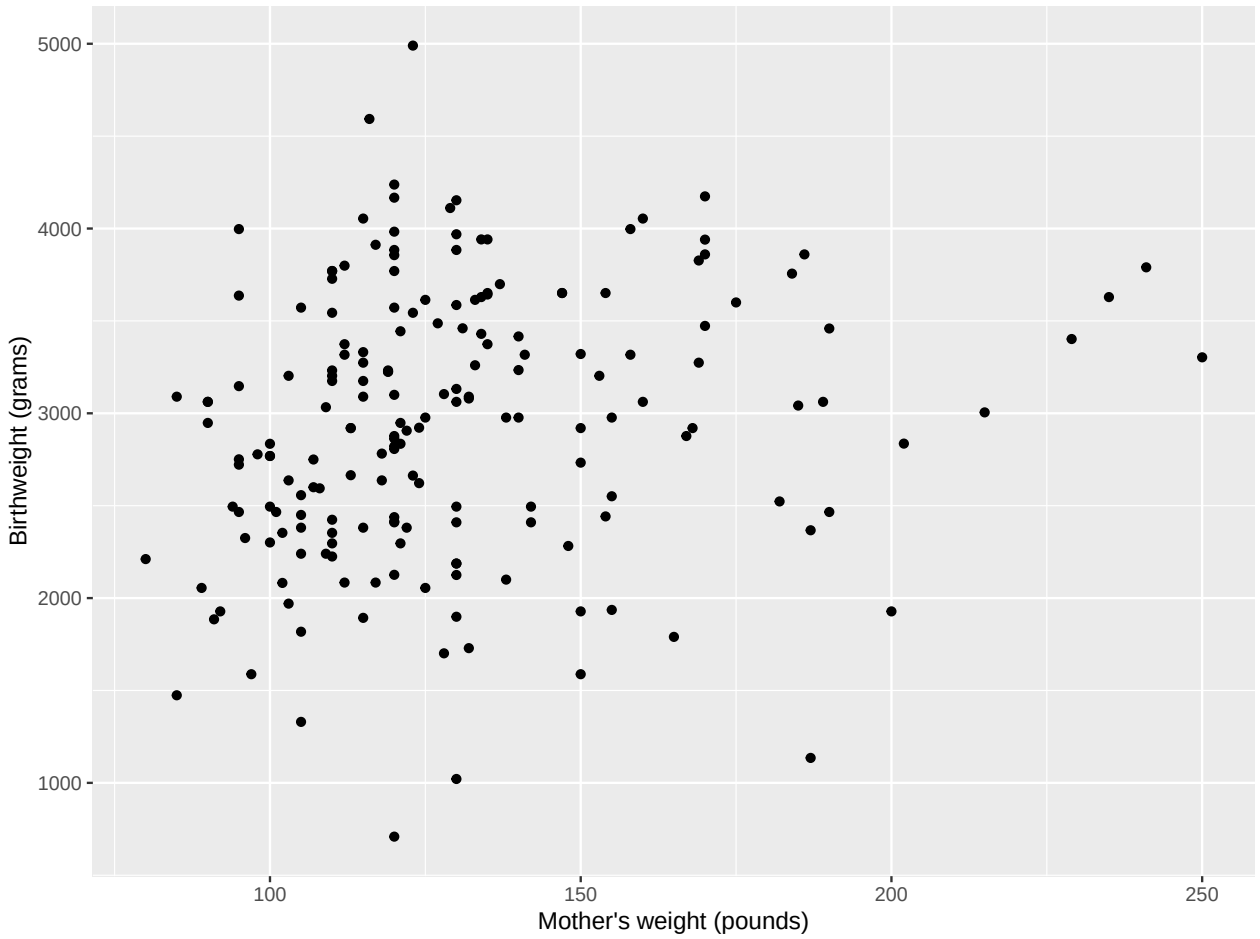
```
data(birthwt)
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

1. Plot a scatterplot of birthweight (bwt) and mother's weight (lwt).

Solution

```
birthwt %>%
ggplot(aes(x = lwt, y = bwt)) +
  geom_point() +
  labs(x = "Mother's weight (pounds)", y = "Birthweight (grams)")
```



2. Use OLS to fit the regression of birthweight on mother's weight.

Solution

```
fit <- lm(bwt ~ lwt, data = birthwt)
summary(fit)
```

```
##
## Call:
## lm(formula = bwt ~ lwt, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2192.12  -497.97    -3.84   508.32  2075.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2369.624    228.493  10.371   <2e-16 ***
## lwt            4.429      1.713   2.585   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 718.4 on 187 degrees of freedom
## Multiple R-squared:  0.0345, Adjusted R-squared:  0.02933
## F-statistic: 6.681 on 1 and 187 DF,  p-value: 0.0105
```

3. Extract the following: estimated coefficients, standard errors, variance-covariance matrix, and confidence intervals.

Solution

```
# Estimated coefficients.
coefficients(fit)
```

```
## (Intercept)         lwt
## 2369.623518    4.429108
```

```
# Standard errors.
summary(fit)$coeff[, 2]
```

```
## (Intercept)         lwt
##  228.493206    1.713494
```

```
# Variance-covariance matrix.
vcov(fit)
```

```
##             (Intercept)         lwt
## (Intercept)  52209.1453 -381.144214
## lwt           -381.1442    2.936061
```

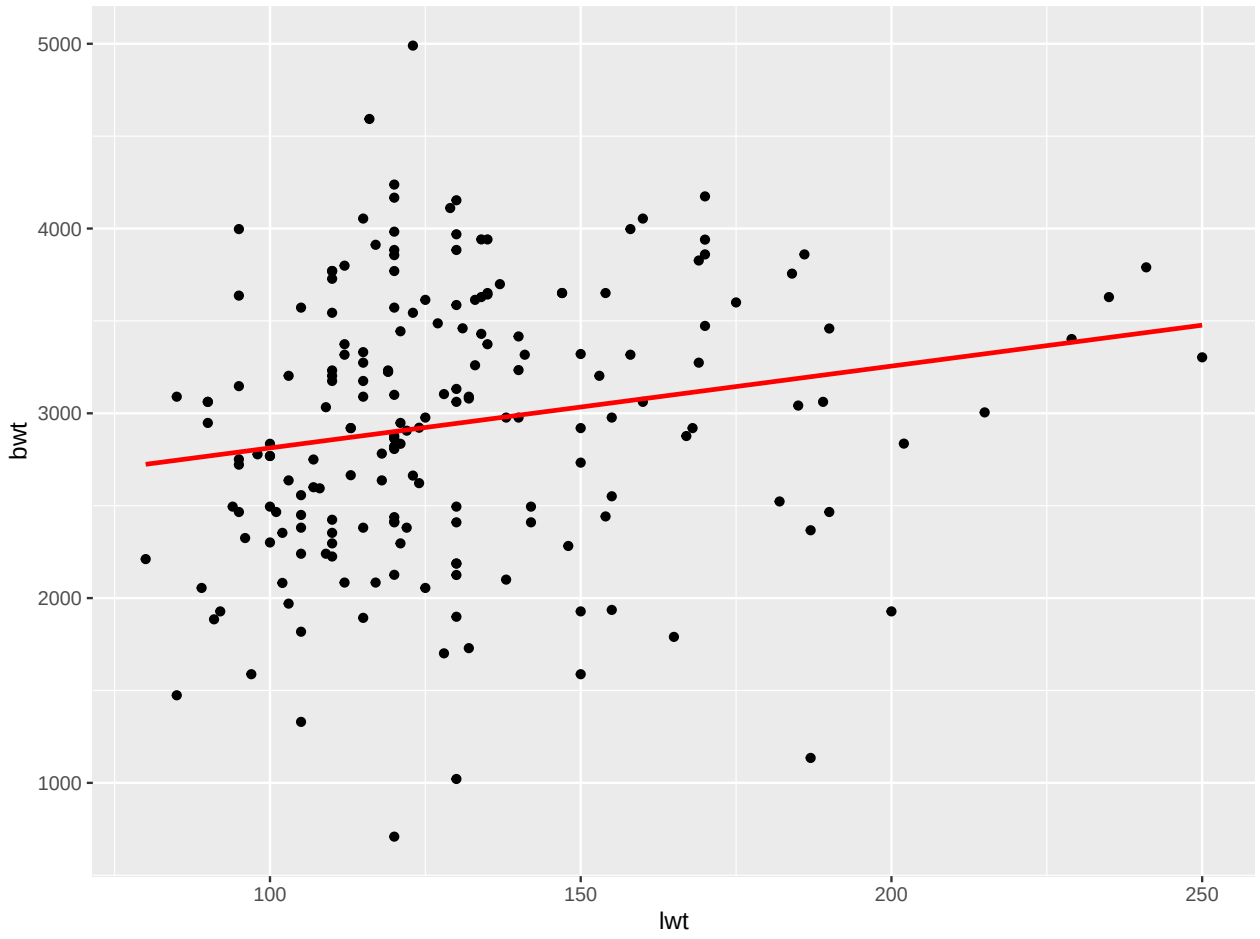
```
# Confidence intervals.
confint(fit)
```

```
##                    2.5 %     97.5 %
## (Intercept) 1918.867879 2820.37916
## lwt            1.048845    7.80937
```

4. Plot the regression line and interpret the intercept and slope

Solution

```
birthwt %>%
ggplot(aes(x = lwt, y = bwt)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red", se = FALSE)
```




```
# Or we can manually add the regression line with geom_abline()
birthwt %>%
ggplot(aes(x = lwt, y = bwt)) +
  geom_point() +
  geom_abline(slope = fit$coefficients[2],
              intercept = fit$coefficients[1], col = "red")
```

5. Does the interpretation of the intercept make sense? How might we change this?

Solution

No, the intercept does not make sense because it implies that if a mother has zero pounds of weight, the expected birthweight is 2369 grams, which is not realistic.

We can center the mother's weight variable to make the intercept more interpretable. This means we will subtract the mean of mother's weight from each observation of mother's weight.

```
birthwt <- birthwt %>% mutate(lwt_star = lwt - mean(lwt))
fit.new <- lm(bwt ~ lwt_star, data = birthwt)
summary(fit.new)
```

```
##
## Call:
## lm(formula = bwt ~ lwt_star, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2192.12  -497.97    -3.84   508.32  2075.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2944.587     52.259  56.346   <2e-16 ***
## lwt_star       4.429      1.713   2.585   0.0105 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 718.4 on 187 degrees of freedom
## Multiple R-squared:  0.0345, Adjusted R-squared:  0.02933
## F-statistic: 6.681 on 1 and 187 DF,  p-value: 0.0105
```

6. Now, we want to fit a model that includes race, mother's age, and smoking status in the model. Race takes on value 1 for white, 2 for black, and 3 for other. Mother's age is continuous. Smoking status is binary. Use OLS to calculate the coefficient estimates in this model.

Solution

```
# wrrong
summary(lm(bwt ~ race + age + smoke, data = birthwt))
```

```
##
## Call:
## lm(formula = bwt ~ race + age + smoke, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2305.88  -439.71    21.34   468.26  1601.44
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3442.987    288.446  11.936  < 2e-16 ***
## race        -227.900     59.648  -3.821 0.000182 ***
## age            3.855      9.735   0.396 0.692570
## smoke       -426.914    110.355  -3.869 0.000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 692.3 on 185 degrees of freedom
## Multiple R-squared:  0.113,  Adjusted R-squared:  0.0986
## F-statistic: 7.854 on 3 and 185 DF,  p-value: 5.819e-05
```

```
fit2 <- lm(bwt ~ as.factor(race) + age + smoke, data = birthwt)
summary(fit2)
```

```
##
## Call:
## lm(formula = bwt ~ as.factor(race) + age + smoke, data = birthwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2322.6  -447.3    28.4   502.2  1612.3
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3281.673    260.664  12.590  < 2e-16 ***
## as.factor(race)2  -444.069    156.194  -2.843 0.004973 **
## as.factor(race)3  -447.858    119.017  -3.763 0.000226 ***
## age                  2.134      9.771   0.218 0.827326
## smoke             -426.093    109.988  -3.874 0.000149 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 690 on 184 degrees of freedom
## Multiple R-squared:  0.1236, Adjusted R-squared:  0.1046
## F-statistic:  6.49 on 4 and 184 DF,  p-value: 6.592e-05
```

7. Interpret all the coefficient estimates.

Solution

- **`(Intercept)`: 3281.673 grams**
  - This is the estimated average birth weight in grams for a baseline individual. The "baseline" here refers to a mother who is White (`race = 1`, as it's the reference level for the `as.factor(race)` variable), does not smoke (`smoke = 0`), and has an age of 0.
  - **Important Note:** An age of 0 is not practically meaningful, so the intercept is often best understood as the starting point for the model's predictions, from which the effects of other variables are then added or subtracted.
- **`as.factor(race)2`: -444.069 grams**
  - This coefficient represents the estimated average difference in birth weight for infants born to **Black mothers (`race = 2`)** compared to **White mothers (`race = 1`)**, *holding all other variables (age and smoking status) constant.*
  - Specifically, on average, infants born to Black mothers are estimated to weigh approximately 444.069 grams less than infants born to White mothers, given the same age and smoking status.
  - The p-value of 0.004973 (which is less than 0.05) indicates that this difference is statistically significant.
- **`as.factor(race)3`: -447.858 grams**
  - This coefficient represents the estimated average difference in birth weight for infants born to **mothers of other races (`race = 3`)** compared to **White mothers (`race = 1`)**, *holding all other variables (age and smoking status) constant.*
  - On average, infants born to mothers of other races are estimated to weigh approximately 447.858 grams less than infants born to White mothers, given the same age and smoking status.
  - The p-value of 0.000226 (which is much less than 0.05) indicates that this difference is highly statistically significant.
- **`age`: 2.134 grams**
  - This coefficient represents the estimated average change in birth weight for every **one-year increase in the mother's age**, *holding all other variables (race and smoking status) constant.*
  - Specifically, for each additional year in the mother's age, the infant's birth weight is estimated to increase by approximately 2.134 grams.
  - The p-value of 0.827326 (which is much greater than 0.05) indicates that this effect is **not statistically significant**. This suggests that, in this model, the mother's age does not have a significant linear relationship with birth weight when considering race and smoking status.
- **`smoke`: -426.093 grams**
  - This coefficient represents the estimated average difference in birth weight for infants born to **mothers who smoke (`smoke = 1`)** compared to **mothers who do not smoke (`smoke = 0`)**, *holding all other variables (race and age) constant.*
  - On average, infants born to mothers who smoke are estimated to weigh approximately 426.093 grams less than infants born to mothers who do not smoke, given the same race and age.
  - The p-value of 0.000149 (which is much less than 0.05) indicates that this difference is highly statistically significant.

9. Print the results in Rmarkdown using kable().

```
table <- data.frame(summary(fit2)$coef)
row.names(table) <- c("Intercept","White","Black","Mother's age", "Smoker")

knitr::kable(table,digits=3,align=rep('c', 2),
```

```
col.names = c("estimate","standard error","test statistic","p-value"))
```

	estimate	standard error	test statistic	p-value
Intercept	3281.673	260.664	12.590	0.000
White	-444.069	156.194	-2.843	0.005
Black	-447.858	119.017	-3.763	0.000
Mother's age	2.134	9.771	0.218	0.827
Smoker	-426.093	109.988	-3.874	0.000