

Solution 4: Linear regression; Generalized Linear regression

Benjamin Smith

June 22, 2026

Part 1: 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.

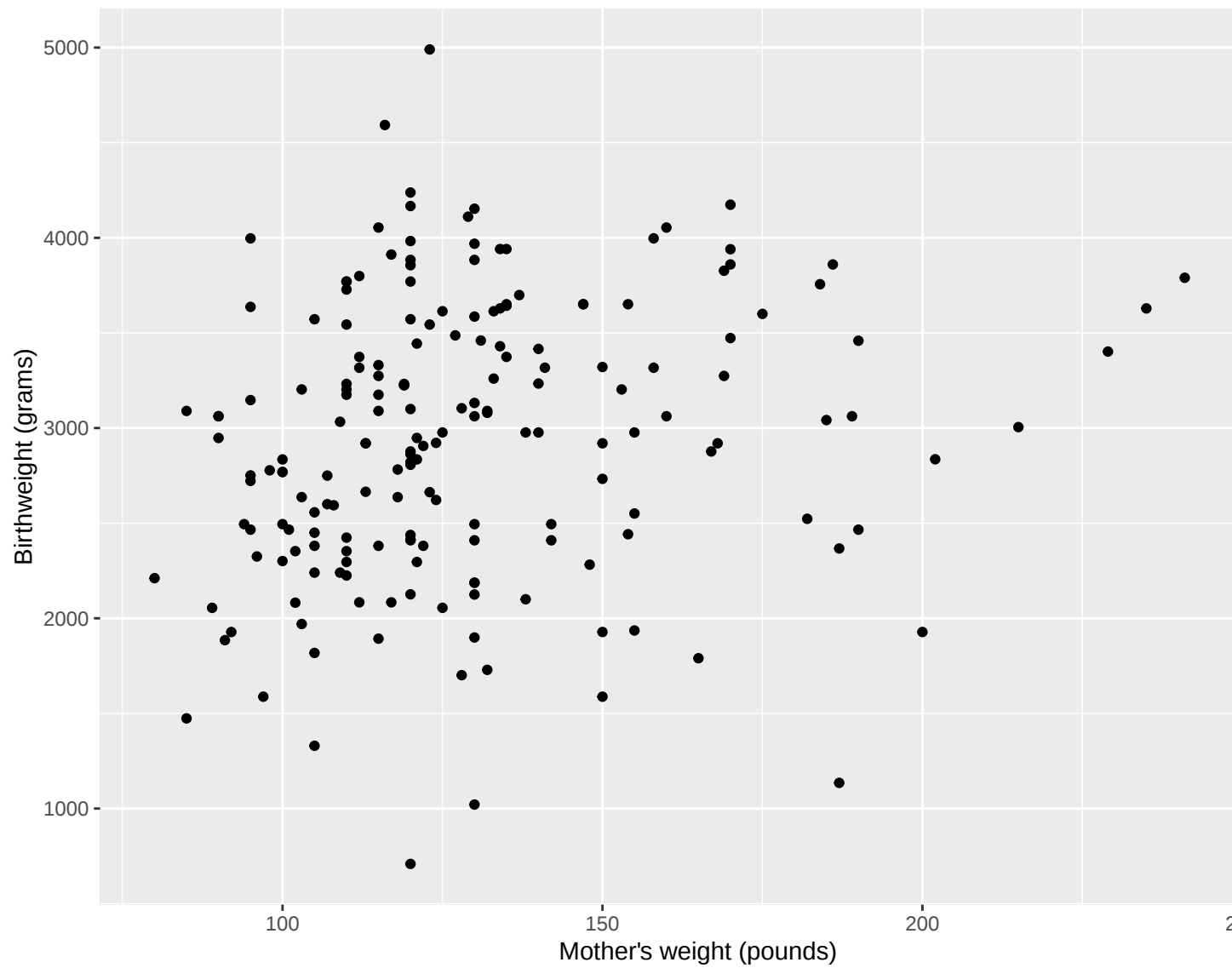
```
data(birthwt)
head(birthwt)
```

```
##      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
```

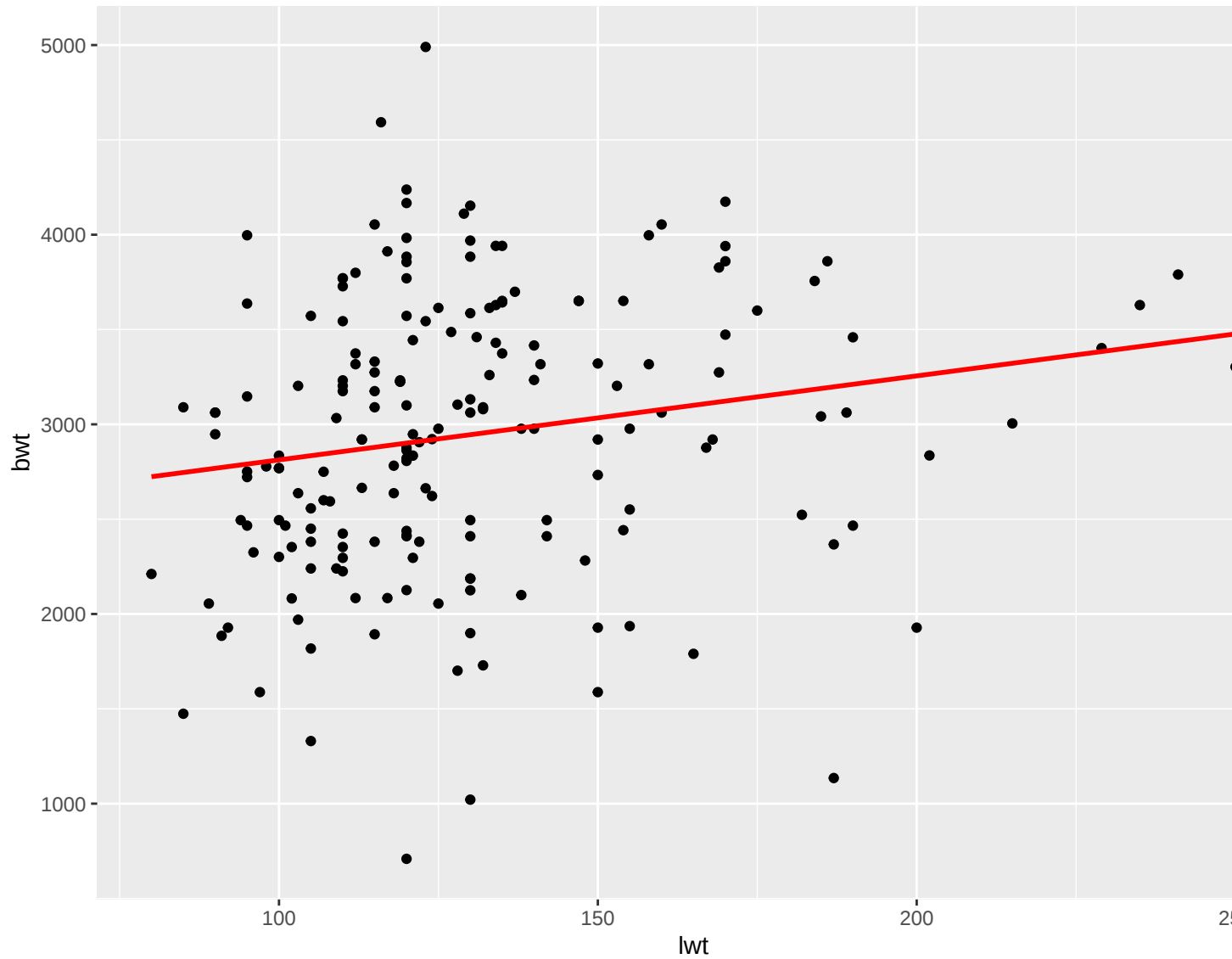
```
# Confidence intervals.
confint(fit)
```

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

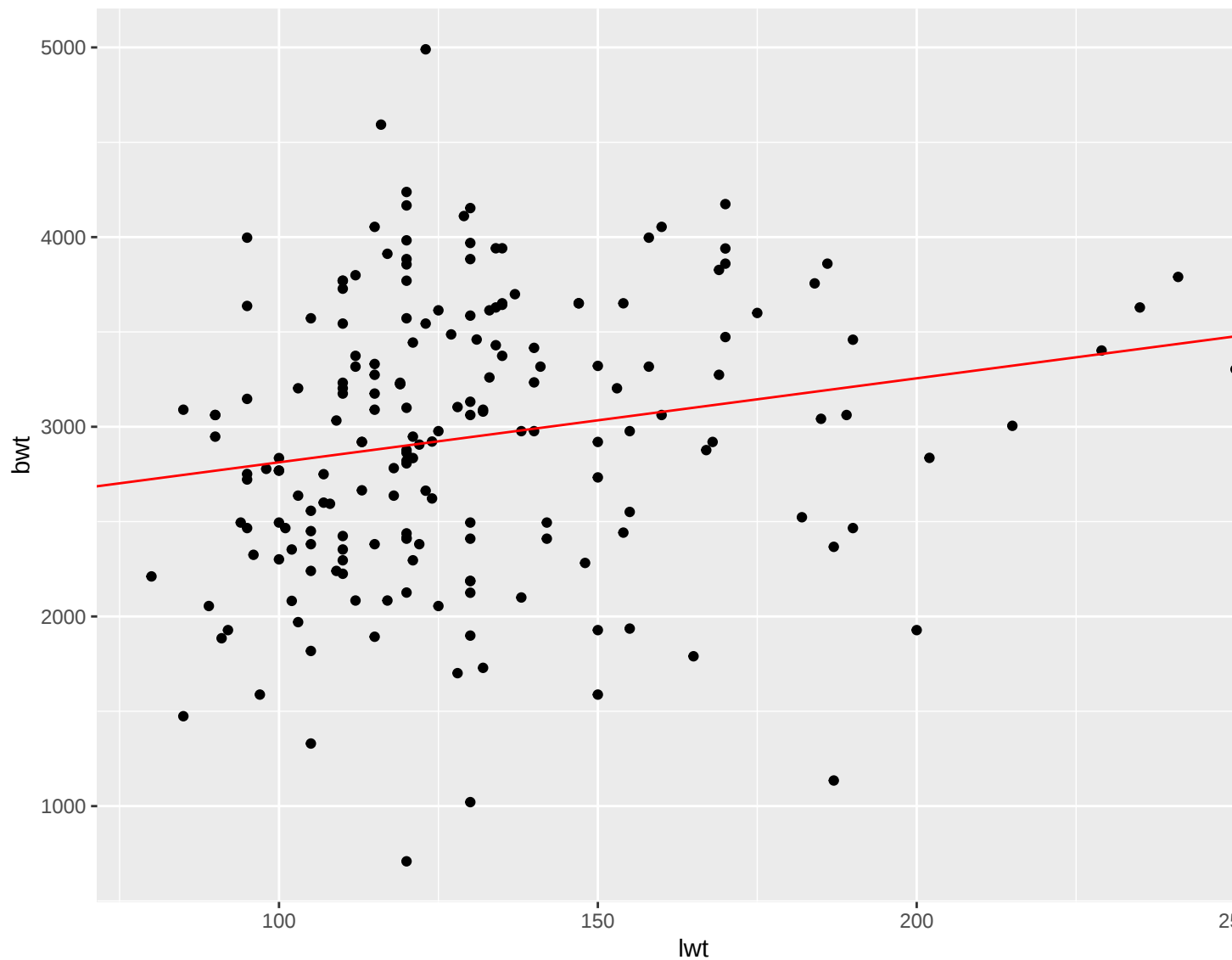
4. Plot the regression line and interpret the intercept and slope

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

```
# wrong
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

Part 2: Generalized linear model

Suppose that 2500 pregnant women are enrolled in a study and the outcome is the occurrence of preterm birth. Possible predictors of preterm birth include age of the woman, smoking, socioeconomic status, body mass index, bleeding during pregnancy, serum level of dde, and several dietary factors.

1. Formulate the problem of selecting the important predictors of preterm birth in a generalized linear model (GLM) framework.
2. Show the components of the GLM, including the link function and distribution (in exponential family form).
3. Describe (briefly) how estimation and inference could proceed via a frequentist approach.

Solution: $y_i = 1$ if woman i has preterm birth and $y_i = 0$ otherwise ($i = 1, \dots, n$) $y_i \sim \text{Bernoulli}(\pi_i)$
Probability density function:

$$\begin{aligned}
 f(y_i; \pi_i) &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\
 &= \exp \{y_i \log \pi_i + (1 - y_i) \log (1 - \pi_i)\} \\
 &= \exp \left\{ y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + \log (1 - \pi_i) \right\} \\
 &= \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}
 \end{aligned}$$

where

$$\theta_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right), \quad b(\theta_i) = \log(1 + e^{\theta_i}), \quad a(\phi) = \phi = 1,$$

and $c(y_i, \phi) = 0$.

Link function: Any mapping from $\mathcal{R} \rightarrow [0, 1]$. A convenient choice is the canonical link,

$$\eta_i = \theta_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right),$$

which is the logit. The probit and complementary log – log are alternatives. Frequentist Estimation: Maximum likelihood estimates can be obtained for a given model, say

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \mathbf{x}_i' \boldsymbol{\beta}$$

(where $\mathbf{x}_i$ is a $p \times 1$ vector of predictors) by iterative weighted least squares

Frequentist Inference: One can select the important predictors to be included in the model by stepwise selection, using the AIC or BIC criterion.

Alternatively, one can just fit the model with all the predictors and then do inferences based on the MLEs and asymptotic standard errors. For example, for continuous predictors included as linear terms in the model, we can do a Wald test. Alternatively, we could do analysis of deviance (see notes for details) to test for significant differences in fit between the nested models with and without a particular predictor.

Part 3: GLMs in R (Logistic regression)

Consider the space shuttle data in the MASS library. Consider modeling the use of the autolander as the outcome (variable name use).

1. Fit a logistic regression model with autolander (variable auto) use (labeled as “auto” 1) versus not (0) as predicted by wind sign (variable wind).
2. Give the estimated odds ratio for autolander use comparing head winds, labeled as “head” in the variable headwind (numerator) to tail winds (denominator).
3. Give the estimated odds ratio for autolander use comparing head winds (numerator) to tail winds (denominator) adjusting for wind strength from the variable magn.
4. If you fit a logistic regression model to a binary variable, for example use of the autolander, then fit a logistic regression model for one minus the outcome (not using the autolander) what happens to the coefficients?

```
library(MASS)
?shuttle
data(shuttle)
head(shuttle)
```

```
##   stability error sign wind   magn vis  use
## 1      xstab   LX   pp head  Light  no auto
## 2      xstab   LX   pp head Medium no auto
## 3      xstab   LX   pp head Strong no auto
## 4      xstab   LX   pp tail  Light  no auto
## 5      xstab   LX   pp tail Medium no auto
## 6      xstab   LX   pp tail Strong no auto
```

Solution

1-2.

```
# not wrong, but needs to be careful with the interpretation of the coefficients
fit1 <- glm(use ~ wind, family = binomial(link = logit), data = shuttle)
summary(fit1)$coefficients
```

```
##              Estimate Std. Error    z value Pr(>|z|)
## (Intercept) -0.25131443  0.1781742 -1.4104987 0.1583925
## windtail    -0.03181183  0.2522429 -0.1261159 0.8996402
```

```
shuttle$use.binary <- as.integer(shuttle$use == "auto")
fit2 <- glm(use.binary ~ wind, family = binomial(link = logit), data = shuttle)
summary(fit2)$coefficients
```

```
##              Estimate Std. Error    z value Pr(>|z|)
## (Intercept)  0.25131443  0.1781742  1.4104987 0.1583925
## windtail     0.03181183  0.2522429  0.1261159 0.8996402
```

```
# exp(intercept + 0*coef(windtail))/exp(intercept + coef(windtail))
exp(summary(fit2)$coefficients[1, 1]) / exp(sum(summary(fit2)$coefficients[, 1]))
```

```
## [1] 0.9686888
```

```
1 / exp(summary(fit2)$coefficients[2, 1])
```

```
## [1] 0.9686888
```

3.

```
fit3 <- glm(use.binary ~ wind + magn, family = binomial(link = logit), data = shuttle)
summary(fit3)$coefficients
```

```
##              Estimate Std. Error    z value Pr(>|z|)
## (Intercept)  3.635093e-01  0.2840608  1.279688e+00 0.2006547
## windtail     3.200873e-02  0.2530225  1.265055e-01 0.8993318
## magnMedium  -1.074656e-15  0.3599481 -2.985586e-15 1.0000000
## magnOut     -3.795136e-01  0.3567709 -1.063746e+00 0.2874438
## magnStrong  -6.441258e-02  0.3589560 -1.794442e-01 0.8575889
```

```
exp(summary(fit3)$coefficients[1, 1] +
  summary(fit3)$coefficients[5, 1]) /
  exp(sum(summary(fit3)$coefficients[1:2, 1]) + summary(fit3)$coefficients[5, 1])
```

```
## [1] 0.9684981
```

4.

```
fit4 <- glm(1 - use.binary ~ wind, family = binomial(link = logit), data = shuttle)
summary(fit4)$coefficients
```

```
##              Estimate Std. Error   z value Pr(>|z|)
## (Intercept) -0.25131443  0.1781742 -1.4104987 0.1583925
## windtail    -0.03181183  0.2522429 -0.1261159 0.8996402
```

```
summary(fit2)$coefficients
```

```
##              Estimate Std. Error   z value Pr(>|z|)
## (Intercept)  0.25131443  0.1781742  1.4104987 0.1583925
## windtail     0.03181183  0.2522429  0.1261159 0.8996402
```

```
fit5 <- lm(use.binary ~ wind, data = shuttle)
summary(fit5)$coefficients
```

```
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept)  0.5625000  0.04397341  12.7918204 2.973129e-29
## windtail     0.0078125  0.06218780   0.1256275 9.001261e-01
```

Part 4: GLMs in R (Poisson regression)

Consider the insect spray data `InsectSprays`. Fit a Poisson model using spray as a factor level.

1. Report the estimated relative rate comparing spray A (numerator) to spray B (denominator).
2. Consider a Poisson glm with an offset, t . So, for example, a model of the form `glm(count ~ x + offset(t), family = poisson)` where x is a factor variable comparing a treatment (1) to a control (0) and t is the natural log of a monitoring time. What is impact of the coefficient for x if we fit the model `glm(count ~ x + offset(t2), family = poisson)` where $t2 < -\log(10) + t$? In other words, what happens to the coefficients if we change the units of the offset variable. (Note, adding $\log(10)$ on the log scale is multiplying by 10 on the original scale.)

```
data("InsectSprays")
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

Solution:

```

# Load data and fit Poisson model
model <- glm(count ~ spray, family = poisson, data = InsectSprays)

# Extract coefficients
coefs <- coef(model)
# sprayB coefficient = log(rate_B / rate_A)
rate_ratio_B_to_A <- exp(coefs["sprayB"])
relative_rate_A_to_B <- 1 / rate_ratio_B_to_A

# Result
relative_rate_A_to_B

```

```

##      sprayB
## 0.9456522

```

```

# Fit Poisson model with offset
t <- rep(log(10), nrow(InsectSprays)) # Example offset
model_offset <- glm(count ~ spray, offset = t, family = poisson, data = InsectSprays)

# Fit Poisson model with modified offset
t2 <- log(10) + t # Modified offset
model_offset2 <- glm(count ~ spray, offset = t2, family = poisson, data = InsectSprays)

# Compare coefficients
coefs_offset <- coef(model_offset)
coefs_offset2 <- coef(model_offset2)
# Coefficients for all
coefs_offset

```

```

## (Intercept)      sprayB      sprayC      sprayD      sprayE      sprayF
## 0.37156356 0.05588046 -1.94017947 -1.08151786 -1.42138568 0.13926207

```

```
coefs_offset2
```

```

## (Intercept)      sprayB      sprayC      sprayD      sprayE      sprayF
## -1.93102154 0.05588046 -1.94017947 -1.08151786 -1.42138568 0.13926207

```

```

# The coefficients for spray A
exp(coefs_offset["(Intercept)"] + t[1])

```

```

## (Intercept)
##      14.5

```

```
exp(coefs_offset2["(Intercept)"] + t2[1])
```

```

## (Intercept)
##      14.5

```