Impact of Age on Earnings and Maternal Smoking on Birthweight: Regression Analysis

Trisha Kumar

2024-10-01

```
#load libraries
library(readxl)
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.3

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

#load the data
CPS12 <- read_xlsx("C:/Users/user/Desktop/Datasets/CPS12.xlsx")
str(CPS12)

## tibble [7,440 × 5] (S3: tbl_df/tbl/data.frame)
##  $ year    : num [1:7440] 2012 2012 2012 2012 2012 ...
##  $ ahe     : num [1:7440] 19.23 17.55 8.55 16.83 16.35 ...
##  $ bachelor: num [1:7440] 0 0 0 0 1 1 0 1 0 0 ...
##  $ female  : num [1:7440] 0 0 0 1 1 0 0 0 0 0 ...
##  $ age     : num [1:7440] 30 29 27 25 27 30 31 29 29 33 ...
```

1.1 Run a regression of average hourly earnings (AHE) on age (Age). What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How much do earnings increase as workers age by 1 year

```
# Running the linear regression
model <- lm(ahe ~ age, data = CPS12)

# Displaying the summary of the regression model
summary(model)

##
## Call:
```

```
## lm(formula = ahe ~ age, data = CPS12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.864  -7.381  -2.245   4.799  72.499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.62605    1.28752   3.593 0.000329 ***
## age          0.51182    0.04323  11.840  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.59 on 7438 degrees of freedom
## Multiple R-squared:  0.0185, Adjusted R-squared:  0.01837
## F-statistic: 140.2 on 1 and 7438 DF,  p-value: < 2.2e-16
```

The estimated intercept is 4.62605 and the estimated slope is 0.51182. Based on these values, the estimated regression equation is:

AHE = 4.62605 + 0.51182 * Age

In this regression equation, the slope of 0.51182 represents the estimated change in average hourly earnings for each additional year of age. Accordingly, as a worker age increases by 1 year, their average hourly earnings are expected to increase by $0.51182 ≈ $0.51.

The intercept 4.62605 is the expected average hourly earnings (AHE) when age is zero, although this value may not be practically meaningful, as it's unrealistic to have workers at age zero.

1.2 Bob is a 26-year-old worker. Predict Bob's earnings using the estimated regression. Alexis is a 30-year-old worker. Predict Alexis's earnings using the estimated regression

AHE = 4.62605 + 0.51182 * Age

AHEBob = 4.62605 + 0.51182 * 26

AHEBob = $17.9337 ≈ $17.93

Bob's predicted earnings is $17.93 per hour

AHE = 4.62605 + 0.51182 * Age

AHEAlexis = 4.62605 + 0.51182 * 30

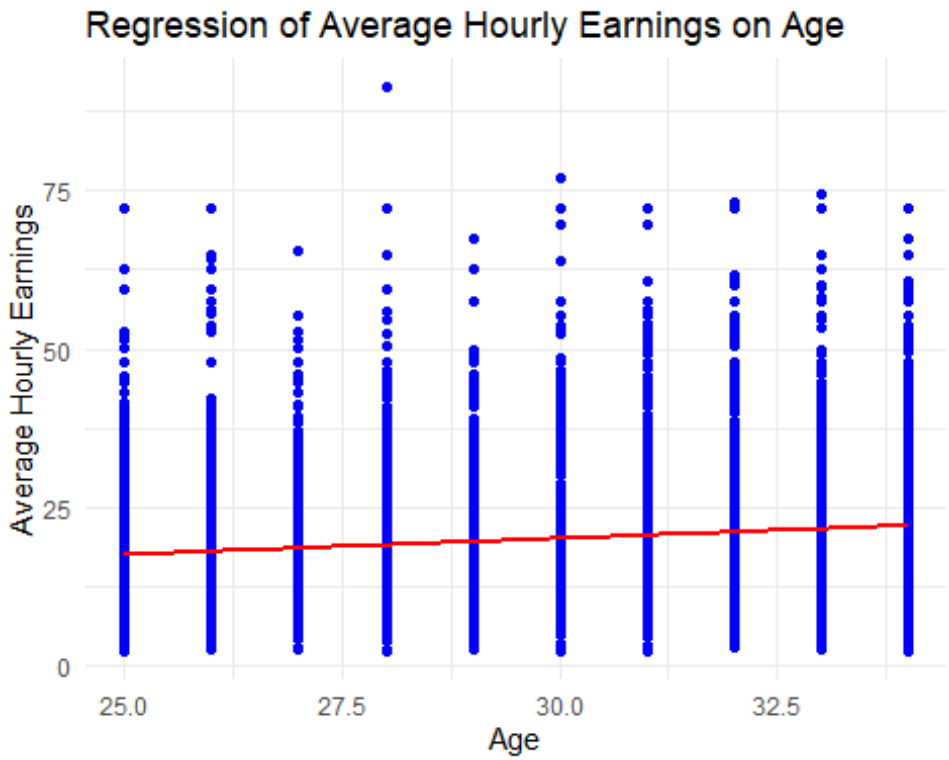
AHEAlexis = $19.98065 ≈ $19.98

Alexis's predicted earnings is $19.98 per hour.

1.3 Does age account for a large fraction of the variance in earnings across individuals?

```
# Create a scatter plot with a regression line
ggplot(CPS12, aes(x = age, y = ahe)) +
  geom_point(color = "blue") +        # Scatter plot of the data points
  geom_smooth(method = "lm", se = FALSE, color = "red") +  # Add the
regression line
  labs(title = "Regression of Average Hourly Earnings on Age",
       x = "Age",
       y = "Average Hourly Earnings") +
  theme_minimal()

## `geom_smooth()` using formula = 'y ~ x'
```



The R-squared value is 0.0185, which indicates that age explains about 1.85% of the variance in earnings. This is a very small fraction, meaning that other factors (such as education, experience, industry, etc.) likely play a much larger role in determining earnings. Age alone does not account for much of the variation in earnings across individuals.

The data file Birthweight_Smoking contains data for a random sample of babies born in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy. A detailed description is given in Birthweight_Smoking_Description. In this exercise, you will investigate the relationship between birth weight and smoking during pregnancy.

```
birthweight_smoking <-
read_xlsx("C:/Users/user/Desktop/Datasets/birthweight_smoking.xlsx")
str(birthweight_smoking)
```

```
## tibble [3,000 × 12] (S3: tbl_df/tbl/data.frame)
##  $ nprevist   : num [1:3000] 12 5 12 13 9 11 12 10 13 10 ...
##  $ alcohol    : num [1:3000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ tripre1    : num [1:3000] 1 0 1 1 1 1 1 1 1 1 ...
##  $ tripre2    : num [1:3000] 0 1 0 0 0 0 0 0 0 0 ...
##  $ tripre3    : num [1:3000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ tripre0    : num [1:3000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ birthweight: num [1:3000] 4253 3459 2920 2600 3742 ...
##  $ smoker     : num [1:3000] 1 0 1 0 0 0 1 0 0 0 ...
##  $ unmarried  : num [1:3000] 1 0 0 0 0 0 0 0 0 0 ...
##  $ educ       : num [1:3000] 12 16 11 17 13 16 14 13 17 14 ...
##  $ age        : num [1:3000] 27 24 23 28 27 33 24 38 29 28 ...
##  $ drinks     : num [1:3000] 0 0 0 0 0 0 0 0 0 0 ...
```

a. What is the average value of Birthweight for all mothers?

```
# Calculate the overall average birthweight
average_birthweight <- mean(birthweight_smoking$birthweight, na.rm = TRUE)
print(paste("Average Birthweight for All Mothers:",
round(average_birthweight, 2), "grams"))
```

```
## [1] "Average Birthweight for All Mothers: 3382.93 grams"
```

b. For mothers who smoke?

```
# Calculate the average birthweight for smoking mothers
average_birthweight_smokers <-
mean(birthweight_smoking$birthweight[birthweight_smoking$smoker == 1], na.rm
= TRUE)
print(paste("Average Birthweight for Smoking Mothers:",
round(average_birthweight_smokers, 2), "grams"))
```

```
## [1] "Average Birthweight for Smoking Mothers: 3178.83 grams"
```

For mothers who do not smoke

```
# Calculate the average birthweight for non-smoking mothers
average_birthweight_nonsmokers <-
mean(birthweight_smoking$birthweight[birthweight_smoking$smoker == 0], na.rm
= TRUE)
print(paste("Average Birthweight for Non-Smoking Mothers:",
round(average_birthweight_nonsmokers, 2), "grams"))
```

```
## [1] "Average Birthweight for Non-Smoking Mothers: 3432.06 grams"
```

2.2.a. Use the data in the sample to estimate the difference in average birth weight for smoking and nonsmoking mothers.

```r
# Summarize data by smoking status
summary_stats <- birthweight_smoking %>%
  group_by(smoker) %>%
  summarise(
    n = n(),
    mean_birthweight = mean(birthweight, na.rm = TRUE),
    sd_birthweight = sd(birthweight, na.rm = TRUE)
  )

# Display the summary statistics
print(summary_stats)
```

```
## # A tibble: 2 × 4
##   smoker     n mean_birthweight sd_birthweight
##    <dbl> <int>            <dbl>          <dbl>
## 1      0  2418            3432.           585.
## 2      1   582            3179.           580.
```

```r
# Extract values for Non-Smokers (smoker == 0)
non_smokers <- summary_stats %>% filter(smoker == 0)
n1 <- non_smokers$n
mean1 <- non_smokers$mean_birthweight
sd1 <- non_smokers$sd_birthweight

# Extract values for Smokers (smoker == 1)
smokers <- summary_stats %>% filter(smoker == 1)
n2 <- smokers$n
mean2 <- smokers$mean_birthweight
sd2 <- smokers$sd_birthweight

diff_avg_birthweight <- mean1 - mean2

# Display the standard error
print(paste("The difference in average birth weight for smoking and
nonsmoking mothers:", round(diff_avg_birthweight, 2), "grams"))
```

```
## [1] "The difference in average birth weight for smoking and nonsmoking
mothers: 253.23 grams"
```

The difference in average birth weight for smoking and nonsmoking mothers is 253.23 grams.

b. What is the standard error for the estimated difference in (i)?

```r
# Calculate the standard error
SE_diff <- sqrt((sd1^2 / n1) + (sd2^2 / n2))

# Display the standard error
print(paste("The Standard Error of the Difference in Means:", round(SE_diff,
2), "grams"))
```

```
## [1] "The Standard Error of the Difference in Means: 26.82 grams"
```

c. Construct a 95% confidence interval for the difference in the average birth weight for smoking and nonsmoking mothers.

```
# Difference in means
diff_means <- mean1 - mean2

# Degrees of freedom for the t-distribution
# Using the smaller of n1 - 1 and n2 - 1 for a conservative estimate
df <- min(n1 - 1, n2 - 1)

# Critical t-value for 95% confidence
t_critical <- qt(0.975, df)

# Calculate the confidence interval
CI_lower <- diff_means - t_critical * SE_diff
CI_upper <- diff_means + t_critical * SE_diff

# Display the confidence interval
print(paste("The 95% Confidence Interval for the Difference in Means: [",
            round(CI_lower, 2), ", ", round(CI_upper, 2), "] grams", sep =
""))
```

```
## [1] "The 95% Confidence Interval for the Difference in Means: [200.55,
305.91] grams"
```

The 95% confidence interval for the difference in average birth weight between babies born to nonsmoking mothers and babies born to smoking mothers is [200.55, 305.91] grams. Based on the sample data, we are 95% confident that the true difference in average birth weight between the two groups (nonsmoking and smoking mothers) lies between 200.55 grams and 305.91 grams.

2.3 Run a regression of Birthweight on the binary variable Smoker.

```
# Run the linear regression
model_bw_smoker <- lm(birthweight ~ smoker, data = birthweight_smoking)

# Display the summary of the regression model
summary(model_bw_smoker)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker, data = birthweight_smoking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3007.06  -313.06    26.94   366.94  2322.94
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3432.06      11.87 289.115   <2e-16 ***
## smoker       -253.23      26.95  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.7 on 2998 degrees of freedom
## Multiple R-squared:  0.0286, Adjusted R-squared:  0.02828
## F-statistic: 88.28 on 1 and 2998 DF,  p-value: < 2.2e-16
```

a. Explain how the estimated slope and intercept are related to your answers in parts (2.1) and (2.2).

Intercept (343.06) represents the average birthweight for the reference group, which in this case is non-smoking mothers (since smoker = 0).

Connection to 2.1 & 2.2

2.1c (Average for Non-Smokers): The intercept aligns directly with the average birthweight calculated for non-smoking mothers in part 2.1c.

It us determined in 2.1c that the average birthweight for non-smokers is 3432.06 grams, this matches the intercept value.

slope (B = -253.23) coefficient represents the difference in average birthweight between smoking and non-smoking mothers. Specifically, on average, babies born to smoking mothers weigh 253.23 grams less than those born to non-smoking mothers.

Connection to 2.1 & 2.2

2.1b (Average for Smokers): In 2.1b I calculated that the average birthweight for smoking mothers is 3178.83 grams

b. Explain how the SE(□1) is related to your answer in b(ii).

The standard error of the slope (β_1) of 26.95 from the regression output represents the standard error for the difference in birth weights between smoking and nonsmoking mothers. It is very close to the standard error for the difference in average birth weights of 26.82 calculated in 2.2(b) due to the use of two approaches to calculate the standard error.

c. Construct a 95% confidence interval for the effect of smoking on birth weight.

```r
# Calculate the 95% confidence interval for the slope (smoker)
confint(model_bw_smoker, "smoker", level = 0.95)
```

```
##            2.5 %    97.5 %
## smoker -306.0736 -200.3831
```

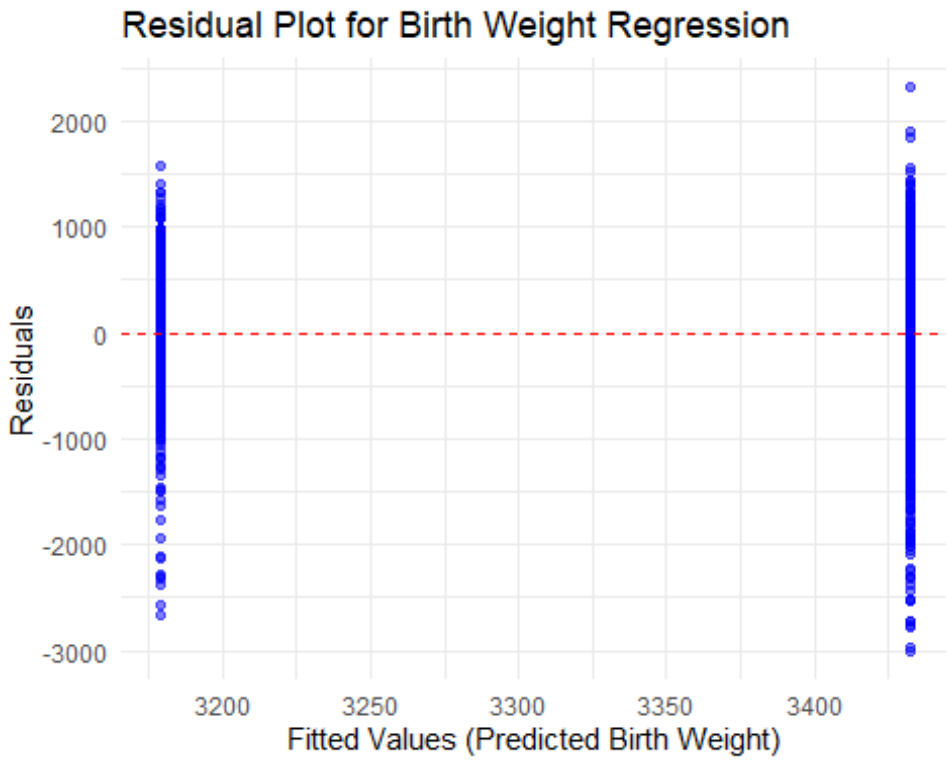
The 95% confidence interval for the effect of smoking on birth weight is [-306.0736, -200.3831]. This interval suggests that, with 95% confidence, babies born to mothers who

smoked during pregnancy weigh between 200.40 grams and 306.07 grams less than babies born to nonsmoking mothers

2.4 Do you think smoking is uncorrelated with other factors that cause low birth weight? That is, do you think that the regression error term—say, ui—has a conditional mean of 0 given Smoking (Xi)?

```
# Residual plot
residuals <- resid(model_bw_smoker)
fitted_values <- fitted(model_bw_smoker)

# Plot residuals vs. fitted values
ggplot(data = data.frame(residuals, fitted_values), aes(x = fitted_values, y
= residuals)) +
  geom_point(color = "blue", alpha = 0.5) +  # Scatter plot for residuals
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +  #
Horizontal line at y = 0
  labs(x = "Fitted Values (Predicted Birth Weight)",  # Label for x-axis
       y = "Residuals",  # Label for y-axis
       title = "Residual Plot for Birth Weight Regression") +  # Title of the
plot
  theme_minimal()  # Minimalistic theme for the plot
```



The coefficient for smoker is estimated to be -253.23. This implies that, on average, babies born to mothers who smoked during pregnancy weighed 253.23 grams less than those

born to nonsmoking mothers. Because the p-value is below 0.05 at < 2.2e-16, we can conclude that a significant relationship between smoking and birth weight exists.

However, the assumption that smoking is uncorrelated with other factors that could influence birth weight is likely violated. Several factors could be correlated with both smoking and birth weight, such as:

1. Socioeconomic Status (SES): Mothers with lower education or income levels are more likely to smoke, and socioeconomic factors can also impact birth weight through mechanisms like access to healthcare, nutrition, and living conditions.

2. Prenatal Care: Smoking mothers might have fewer or lower-quality prenatal care visits, which are known to affect birth outcomes. My data includes variables like the timing of the first prenatal visit and total prenatal visits, which could affect birth weight but aren't yet accounted for in this simple model.

3. Other Health Behaviors: Smoking mothers may also engage in other risky behaviors such as alcohol consumption, poor diet, or inadequate exercise, all of which could contribute to lower birth weight.

Because smoking is likely to be correlated with these omitted factors, the error term ui probably captures the influence of those unobserved variables, meaning that $E(u_i|X_i) = 0$ is unlikely to hold. In this case, the effect of smoking may be overestimated if these omitted variables also contribute to lower birth weight. The R-squared value of 0.0286 indicates that smoking explains only about 2.86% of the variability in birth weight. This suggests that a significant amount of variation in birth weight remains unexplained by smoking alone, reinforcing the likelihood that other factors contribute to birth weight but are not included in the model.

The residual plot on the next page shows the difference between the observed birth weights and the predicted birth weights (residuals) from the regression model. Ideally, if the regression model accurately captures the relationship between smoking and birth weight, the residuals should be randomly scattered around zero, without any clear pattern.

However, the residual plot reveals heteroscedasticity and clustering. This further indicates that the model is oversimplified and missing other important factors that could explain the variability in birth weight.