PS4 Solutions

Jingle Fu

Solution (a).

Regression Analysis

The model equation is:

$$\log(\text{earnings}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \beta_3 \cdot \text{female} + \beta_4 \cdot \text{bachelor}$$

For a 30-year-old female with a bachelor's degree:

$$\text{Predicted log-earnings} = \beta_0 + \beta_1 \cdot 30 + \beta_2 \cdot 30^2 + \beta_3 \cdot 1 + \beta_4 \cdot 1$$

From the regression result, the model explains about 20.1% of the variability in log earnings ($R^2 : 0.201$). Here's a summary of the coefficients:

- **Intercept**: $\beta_0 = 1.0854$ (p = 0.089)
- **Age**: $\beta_1 = 0.0814$ (p = 0.061)
- **Age Squared**: $\beta_2 = -0.0009$ (p = 0.214)
- **Female**: $\beta_3 = -0.1859$ (p < 0.001)
- **Bachelor**: $\beta_4 = 0.4284$ (p < 0.001)

The predicted log-earnings for a 30-year-old female with a bachelor's degree is approximately 2.9458.

Table 1: Regression with age^2

	Dependent variable:
	log(ahe)
age	0.081
	(0.043)
age2	−0.001
	(0.001)
female	−0.186**
	(0.011)
bachelor	0.428
	(0.011)
Constant	1.085
	(0.638)
Observations	7,711
R^2	0.201
Adjusted R^2	0.200
Residual Std. Error	0.469 (df = 7706)
F Statistic	484.078* (df = 4; 7706)
Note:	p<0.1; p<0.05; p<0.01

```r
rm(list = ls())
library(tidyverse)
library(ggplot2)
library(dplyr)
library(broom)
library(stats)
library(stargazer)
library(xtable)
dat <- read.csv("dat_CPS08.csv")
dat <- dat %>%
  mutate(age2 = age^2)
model <- lm(log(ahe) ~ age + age2 + female + bachelor, data = dat)
```

```
new_data <- data.frame(age = 30, age2 = 30^2, female = 1, bachelor = 1)
predicted_log_earnings <- predict(model, newdata = new_data, type = "response")
predicted_log_earnings
```

Solution (b).

We assume a linear regression model:

$$\log(\text{earnings}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \beta_3 \cdot \text{female} + \beta_4 \cdot \text{bachelor} + \varepsilon$$

where ε is the error term, normally distributed with mean zero and variance σ^2.

Specification of the Null Hypothesis

For a 30-year-old female with a bachelor's degree, the log-earnings prediction under our model becomes:

$$\log(\text{earnings}) = \beta_0 + \beta_1 \cdot 30 + \beta_2 \cdot 30^2 + \beta_3 \cdot 1 + \beta_4 \cdot 1$$

Denote this as $x_{30}'\beta$ where $x_{30} = [1, 30, 30^2, 1, 1]'$. The null hypothesis is:

$$H_0 : x_{30}'\beta = \log(20) \approx 2.99$$

Asymptotic Distribution of the Estimator

Under standard OLS assumptions, the estimator $\hat{\beta}$ is normally distributed around the true β with a variance estimated as:

$$\hat{\beta} \sim N\left(\beta, (X'X)^{-1}\sigma^2\right)$$

The variance of the prediction $x_{30}'\hat{\beta}$ can be derived as:

$$\text{Var}(x_{30}'\hat{\beta}) = x_{30}'(X'X)^{-1}x_{30}\sigma^2$$

Construction of the t-test

The test statistic is:

$$t = \frac{x_{30}'\hat{\beta} - 2.99}{\sqrt{\text{Var}(x_{30}'\hat{\beta})}}$$

Under H_0, this statistic follows a t-distribution with $n-k$ degrees of freedom.

Decision Rule

We reject the null hypothesis at the 5% significance level if:

$$|t| > t_{n-k,0.025}$$

where $t_{n-k,0.025}$ is the critical value of the t-distribution with $n-k$ degrees of freedom at the 2.5% tail.

The t-statistic for testing whether the expected hourly earnings are \$20 is approximately -4.46 with a corresponding p-value of 8.07×10^{-6}. This result is highly significant, indicating that we can reject the null hypothesis H_0 : Expected log-earnings = log(20). Therefore, it appears that the expected hourly earnings for a 30-year-old female with a bachelor's degree are significantly different from \$20 per hour.

```
expected_log <- log(20)

X_new <- matrix(c(1, 30, 30^2, 1, 1), nrow=1)
prediction_variance <- as.numeric(X_new %*% vcov(model) %*% t(X_new))
prediction_se <- sqrt(prediction_variance)

t_stat <- (predicted_log_earnings - expected_log) / prediction_se
p_value <- 2 * pt(-abs(t_stat), df=df.residual(model))
t_stat
p_value
```

Solution (c).

We continue the steps following question (b).

Variance of Prediction

The variance of the predicted log-earnings is:

$$\text{Var}(\hat{y}_{30}) = x_{30}'(X'X)^{-1}x_{30}\sigma^2$$

Constructing the Confidence Interval

The 95% confidence interval for the predicted log-earnings is given by:

$$\hat{y}_{30} \pm t_{0.975,n-p} \cdot \sqrt{\text{Var}(\hat{y}_{30})}$$

where $t_{0.975,n-p}$ is the 97.5th percentile of the t-distribution with $n-p$ degrees of freedom.

Calculation

Using estimated coefficients, calculated variance, and critical t-values:

- Compute the point estimate $\hat{y}_{30}$,
- Calculate the standard error $\sqrt{\text{Var}(\hat{y}_{30})}$,
- Determine the lower and upper bounds of the confidence interval.

```r
ci_lower <- predicted_log_earnings + qt(0.025, df=df.residual(model)) * prediction_se
ci_upper <- predicted_log_earnings + qt(0.975, df=df.residual(model)) * prediction_se
ci_lower
ci_upper
```

The confidence interval was (2.9238, 2.9677), suggesting that with 95% confidence, the expected log-earnings 2.99 is not within this range, so we reject the null hypothesis.

Solution (d).

Age-Earnings Profile Plotting

The plot shows an increasing trend in earnings with age, leveling off as age increases.

```r
model2 <- lm(log(ahe) ~ age + female + bachelor, data=dat)
ages_df <- data.frame(
  age = 20:65,
  female = rep(1, 46),
  bachelor = rep(1, 46)
)

mean_log_earnings <- numeric(length = 46)
ci_lower <- numeric(length = 46)
ci_upper <- numeric(length = 46)

for (i in 1:46) {
  new_data <- data.frame(age = ages_df$age[i], female = 1, bachelor = 1)

  predicted_log_earnings <- predict(model2, newdata = new_data, type = "response")
```

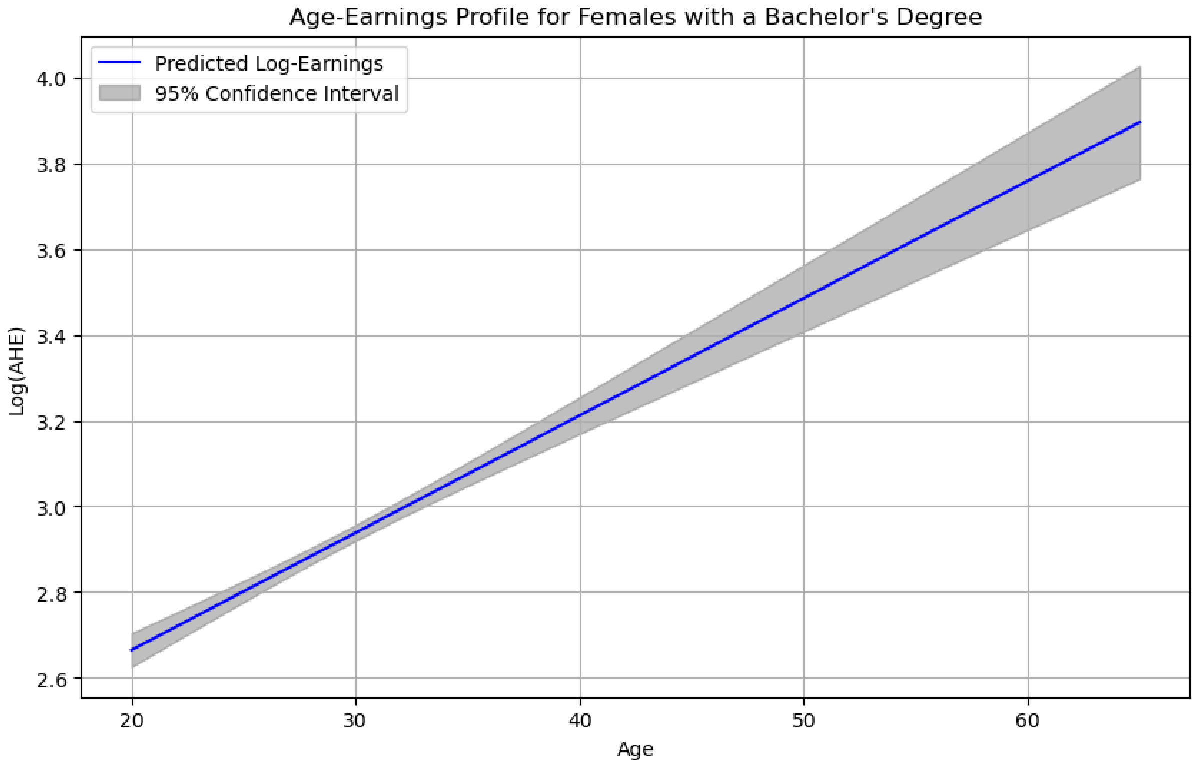

Figure 1: Age-earnings profile

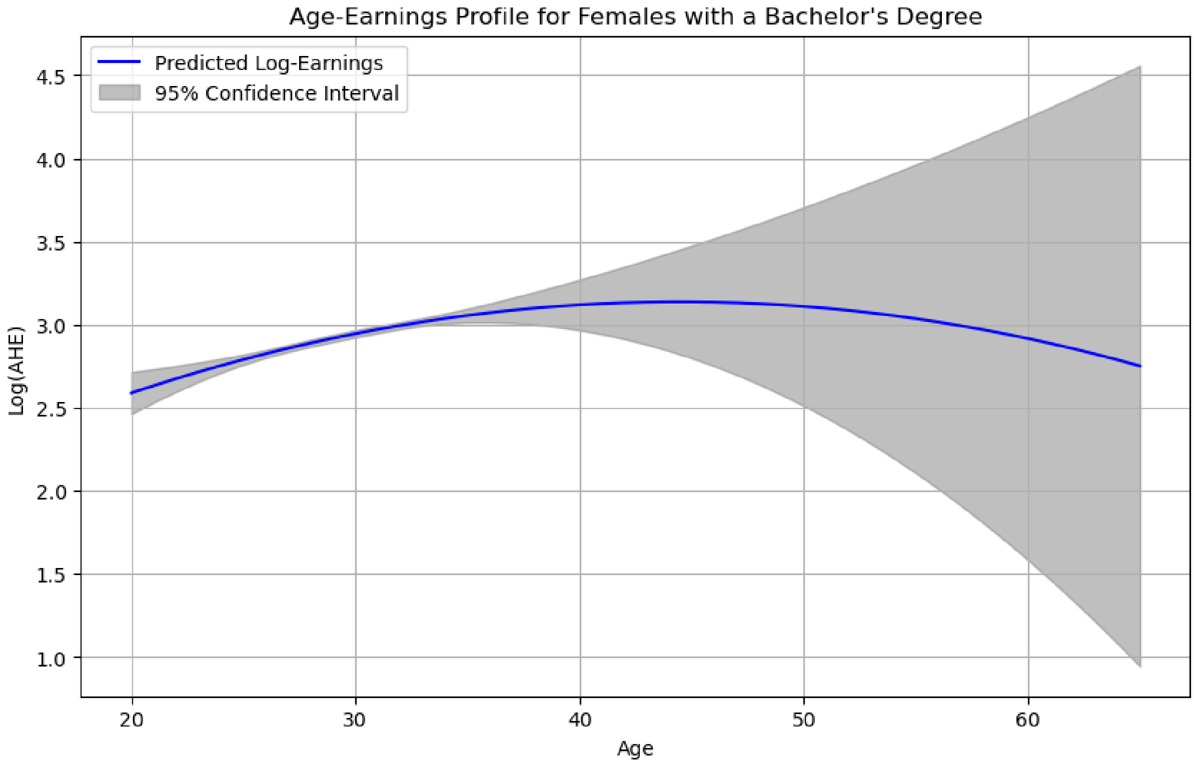

Figure 2: Age-earnings profile(with age2)

```
  mean_log_earnings[i] <- predicted_log_earnings

  X_new <- matrix(c(1, ages_df$age[i], 1, 1), nrow = 1)
  prediction_variance <- as.numeric(X_new %*% vcov(model2) %*% t(X_new))
  prediction_se <- sqrt(prediction_variance)

  ci_lower[i] <- predicted_log_earnings + qt(0.025, df = df.residual(model2)) * prediction_se
  ci_upper[i] <- predicted_log_earnings + qt(0.975, df = df.residual(model2)) * prediction_se
}

predictions_df <- data.frame(
  age = ages_df$age,
  mean = mean_log_earnings,
  lower = ci_lower,
  upper = ci_upper
)

p <- ggplot(predictions_df, aes(x=age, y=mean)) +
  geom_line(color="blue", size=1) +
  geom_ribbon(aes(ymin=lower, ymax=upper), fill="gray", alpha=0.5) +
  labs(x="Age", y="Log(AHE)",
       title="Age-Earnings Profile for Females with a Bachelor's Degree") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(20, 65, 5)) +
  scale_y_continuous(limits=c(min(predictions_df$lower), max(predictions_df$upper)))
print(p)
```

Solution (e).

The coefficient for bachelor's degree in our model is positive and significant, suggesting that having a bachelor's degree is associated with higher log-earnings compared to having just a high school diploma. However, determining if this relationship is causal requires considering whether there might be any omitted variable bias, reverse causality, or any other confounding factors that weren't controlled for in the model.

In observational data like this, establishing causality is challenging without experimental or quasi-experimental designs (e.g., using instrumental variables, regression discontinuity designs, etc.). Factors such as individual ability, motivation, or other socioeconomic variables could be influencing both the likelihood of obtaining a bachelor's degree and the earnings potential.

Without controlling for these potential confounders or using a design that can mimic random assignment, we cannot definitively say that the coefficient of bachelor's degree is a causal effect on earnings. It's a strong association but should be interpreted with caution regarding causality.

Solution (f).

To explore whether the relationship between age and log-earnings differs by gender, interaction terms were added to the regression model. The model can be mathematically expressed as:

$$\begin{aligned}\log(\text{earnings}) = {} & \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \beta_3 \cdot \text{female} \\ & + \beta_4 \cdot \text{bachelor} + \beta_5 \cdot \text{female\_age} + \beta_6 \cdot \text{female\_age2} + \varepsilon\end{aligned}$$

F-Statistic of the Model and Model Fit

- The overall model is statistically significant with an F-statistic of 327 on 6 and 7704 degrees of freedom, and a p-value $< 2.2 \times 10^{-16}$. This indicates that the model significantly explains the variability in log earnings better than a model without predictors.
- The Multiple R-squared of 0.203 indicates that about 20.3% of the variability in log earnings is explained by the predictors in the model. This level of explanation is moderate but common in social sciences where outcomes are influenced by numerous unmeasured factors.

The F-test results are as below:

The interpretation of these results is as follows:

Model Comparison: The F-test compares the full model (Model 1) against the reduced model (Model 2). Model 1 includes the main effects of age, age-squared, gender, and

Table 2: Interaction Model of Age and Gender on Earnings

	Dependent variable:
	log(ahe)
age	0.126^{**}
	(0.058)
I(ageˆ2)	−0.002
	(0.001)
female	1.418
	(1.283)
bachelor	0.427^{***}
	(0.011)
female_age	−0.092
	(0.087)
female_age2	0.001
	(0.001)
Constant	0.315
	(0.854)
Observations	7,711
R^2	0.203
Adjusted R^2	0.202
Residual Std. Error	0.469 (df = 7704)
F Statistic	326.961^{***} (df = 6; 7704)
p-value	< 2.2e-16
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: F-test result

	Res.Df	Df	F	Pr(>F)
1	7704			
2	7706	-2	10.37	$3.173e-05^{***}$
Note:				*p<0.1; **p<0.05; ***p<0.01

education, as well as the interaction effects between gender and age, and gender and age-squared. Model 2 omits these interaction terms.

Null Hypothesis: The null hypothesis is that the interaction effects between gender and age, and gender and age-squared, are jointly equal to zero. In other words, the relationship between age, age-squared, and log-earnings is assumed to be the same for both males and females.

Test Statistic and p-value: The F-statistic is 10.372, and the corresponding p-value is 3.173×10^{-5}. This p-value is less than the typical significance level of 0.05, indicating strong evidence against the null hypothesis.

Conclusion: Based on the F-test results, we can reject the null hypothesis and conclude that the relationship between age and log-earnings is *significantly different* for males and females. In other words, the inclusion of the gender-age interaction terms ($female\_age$ and $female\_age2$) significantly improves the model fit compared to the model that assumes a common age-earnings profile for both genders.

```r
dat$female_age <- dat$female * dat$age
dat$female_age2 <- dat$female * dat$age2

interaction_formula <- log(ahe) ~ age + I(age^2) + female + bachelor +
    female_age + female_age2
interaction_model <- lm(interaction_formula, data=dat)

# Two ways of giving F-test
sse_full <- sum(residuals(interaction_model)^2)
sse_reduced <- sum(lm(log(ahe) ~ age + I(age^2) + female + bachelor,
    data=dat)$residuals^2)
df_full <- nrow(dat) - length(coef(interaction_model))
df_reduced <- nrow(dat) - length(coef(lm(log(ahe) ~ age + I(age^2) +
    female + bachelor, data=dat)))
f_stat <- ((sse_reduced - sse_full) / (df_reduced - df_full)) / (sse_
    full / df_full)
```

```r
pvalue <- 1 - pf(f_stat, df_reduced - df_full, df_full)
print(paste0("F-statistic: ", f_stat))
print(paste0("P-value: ", pvalue))

waldtest(interaction_model, lm(log(ahe) ~ age + I(age^2) + female + bachelor, data=dat), test="F")

```