Empirical Exercise 2, Stock and Watson Chapter 8

Econ 440 - Introduction to Econometrics

Patrick Toche, ptoche@fullerton.edu

17 April 2022

Empirical Exercise: Earnings

The data file **CPS2015** contains data for full-time, full-year workers, ages 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in **CPS2015_Description**. In this exercise, you will investigate the relationship between a worker's age and earnings

Dataset:

```
library(readxl)
df <- read_xlsx("CPS2015.xlsx", trim_ws=TRUE)
head(df)
```

```
## # A tibble: 6 x 5
##    year   ahe bachelor female   age
##   <dbl> <dbl>    <dbl>  <dbl> <dbl>
## 1  2015 11.8         0      0    26
## 2  2015  9.62        0      1    33
## 3  2015 12.0         0      0    31
## 4  2015 18.4         0      0    32
## 5  2015 41.8         0      0    28
## 6  2015 19.2         0      1    31
```

(a)

Run a regression of average hourly earnings (*ahe*) on age (*age*), sex ($female$), and education (*bachelor*). If *age* increases from 25 to 26, how are earnings expected to change? If *age* increases from 33 to 34, how are earnings expected to change?

```
m1 <- lm(ahe ~ age + female + bachelor, data=df)
summary(m1)
```

```
##
## Call:
## lm(formula = ahe ~ age + female + bachelor, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27.91  -6.65  -1.87   4.25  83.91
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0448     1.3547    1.51     0.13
```

```
## age           0.5313     0.0451   11.79   <2e-16 ***
## female       -4.1435     0.2659  -15.58   <2e-16 ***
## bachelor      9.8456     0.2624   37.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.9 on 7094 degrees of freedom
## Multiple R-squared:  0.19,   Adjusted R-squared:  0.189
## F-statistic:  553 on 3 and 7094 DF,  p-value: <2e-16
```

If *age* increases by 1 year, *ahe* is expected to increase by \$0.5313, that is about 53 cents per hour. Thus, the answer is the same for an increase from 25 to 26 and from 33 to 34. We can make a prediction for the values, rather than the increase. For convenience, we use R's built-in `predict` function.

```
newdata <- data.frame(age=c(25,26), female=rep(0,2), bachelor=rep(0,2))
predict(m1, newdata)
```

```
##      1      2
## 15.327 15.858
```

As *age* increases from 25 to 26, earnings increase from 15.327 to 15.858

```
newdata <- data.frame(age=c(33,34), female=rep(0,2), bachelor=rep(0,2))
predict(m1, newdata)
```

```
##      1      2
## 19.577 20.108
```

As *age* increases from 33 to 34, earnings increase from 19.577 to 20.108

As this is a repeat question, let's build a convenience function:

```
lm_predict <- function(model, data){
  n <- length(data)
  newdata <- data.frame(age=data,
                        female=rep(0,n),
                        bachelor=rep(0,n),
                        predicted=rep(1,n))
  newdata$ahe <- predict(model, newdata)
  newdata
}
lm_predict(m1, data=c(33,34))
```

```
##   age female bachelor predicted    ahe
## 1  33      0        0         1 19.577
## 2  34      0        0         1 20.108
```

(b)

Run a regression of the logarithm of average hourly earnings, $ln(ahe)$, on *age*, *female*, and *bachelor*. If *age* increases from 25 to 26, how are earnings expected to change? If *age* increases from 33 to 34, how are earnings expected to change?

```
m2 <- lm(log(ahe) ~ age + female + bachelor, data=df)
summary(m2)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + female + bachelor, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5980 -0.2879  0.0078  0.3008  2.0631
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.02736    0.05924    34.2   <2e-16 ***
## age          0.02419    0.00197    12.3   <2e-16 ***
## female      -0.17762    0.01163   -15.3   <2e-16 ***
## bachelor     0.46150    0.01148    40.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 7094 degrees of freedom
## Multiple R-squared:  0.208,  Adjusted R-squared:  0.208
## F-statistic:  622 on 3 and 7094 DF,  p-value: <2e-16
```

Because the regression involves non-linear transformations of the original variables, interpreting the coefficients requires great care.

If *age* increases by 1 year, $ln(ahe)$ is expected to increase by 0.2419 log-dollars, that is earnings *ahe* is expected to increase by about 2.5%. This predicted effect is the same for an increase in *age* from 25 to 26 and for an increase from 33 to 34.

The predicted percentage change is computed as follows:

$$\Delta\widehat{ln(ahe)} = 0.2419 \implies \Delta\widehat{ahe} \approx e^{0.2419} - 1 \approx 0.025$$

A simple rule of thumb is $0.2419 \rightarrow 2.4\%$

This calculation may be done in R by extracting the coefficient from the linear model object.

```
exp(m2$coefficients[["age"]])-1
```

```
## [1] 0.024486
```

To preview the $m2$ object's structure, type `str(m2)` and then `str(m2$coefficients)`.

```
m2$coefficients[["age"]] * (log(34)-log(33))
```

```
## [1] 0.00072218
```

Specifically, the predicted values of *ahe* are:

```
lm_predict(m2, data=c(25,26,33,34))
```

```
##   age female bachelor predicted    ahe
## 1  25      0        0         1 2.6321
## 2  26      0        0         1 2.6563
## 3  33      0        0         1 2.8257
## 4  34      0        0         1 2.8499
```

(c)

Run a regression of the logarithm of average hourly earnings, $ln(ahe)$, on $ln(age)$, $female$, and *bachelor*. If *age* increases from 25 to 26, how are earnings expected to change? If *age* increases from 33 to 34, how are earnings expected to change?

```
m3 <- lm(log(ahe) ~ log(age) + female + bachelor, data=df)
summary(m3)
```

```
##
## Call:
## lm(formula = log(ahe) ~ log(age) + female + bachelor, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.594 -0.287  0.010  0.302  2.062
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3233     0.1961    1.65    0.099 .
## log(age)      0.7154     0.0578   12.37   <2e-16 ***
## female       -0.1775     0.0116  -15.27   <2e-16 ***
## bachelor      0.4615     0.0115   40.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 7094 degrees of freedom
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.208
## F-statistic:  623 on 3 and 7094 DF,  p-value: <2e-16
```

Both *ahe* and *age* appear in log form. As a result, it is no longer true that an increase in *age* by 1 year increases *ahe* uniformly: instead the effect on *ahe* depends on *age* and must be computed separately for each case.

If *age* increases by 1 year, from 25 to 26, *ahe* is expected to increase by about 2.8%:

$$\Delta\widehat{ln(ahe)} = 0.7154 \cdot (\ln(26) - \ln(25)) \approx 0.028$$

This calculation follows from:

```
m3$coefficients[["log(age)"]] * (log(26)-log(25))
```

```
## [1] 0.028058
```

If *age* increases by 1 year from 33 to 34, the increase in *ahe* is smaller, at about 2.1%.

$$\Delta\widehat{ln(ahe)} = 0.7154 \cdot (\ln(34) - \ln(33)) \approx 0.021$$

```
m3$coefficients[["log(age)"]] * (log(34)-log(33))
```

```
## [1] 0.021356
```

Specifically, the predicted values of *ahe* are:

```
lm_predict(m3, data=c(25,26,33,34))
```

```
##   age female bachelor predicted    ahe
## 1  25      0        0         1 2.6260
## 2  26      0        0         1 2.6540
## 3  33      0        0         1 2.8246
## 4  34      0        0         1 2.8459
```

(d)

Run a regression of the logarithm of average hourly earnings, $ln(ahe)$, on age, $age2$, $female$, and $bachelor$. If age increases from 25 to 26, how are earnings expected to change? If age increases from 33 to 34, how are earnings expected to change?

```
m4 <- lm(log(ahe) ~ age + I(age^2) + female + bachelor, data=df)
summary(m4)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + I(age^2) + female + bachelor, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5764 -0.2868  0.0126  0.3041  2.0596
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.418745   0.672088    0.62   0.5333
## age          0.134115   0.045791    2.93   0.0034 **
## I(age^2)    -0.001860   0.000774   -2.40   0.0163 *
## female      -0.177364   0.011626  -15.26   <2e-16 ***
## bachelor     0.461629   0.011473   40.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 7093 degrees of freedom
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.209
## F-statistic:  469 on 4 and 7093 DF,  p-value: <2e-16
```

If age increases by 1 year, ahe is expected to increase by \$0.5313, that is about 53 cents per hour.

(e)

Do you prefer the regression in (c) to the regression in (b)? Explain.

The difference between the regressions in (b) and (c) is that age is replaced by $log(age)$. The statistical analysis does not point to any major differences. The individual coefficients all have near-zero p-values. The regression's standard error is about the same. The regression's adjusted R-squared is about the same. The regression's p-value is near zero in both cases. The main reason for preferring regression (c) is that the coefficient on $log(age)$ can be interpreted as an elasticity.

(f)

Do you prefer the regression in (d) to the regression in (b)? Explain.

The difference between regression (b) and regression (d) is that the linear regressor age is replaced by a quadratic polynomial in age and age^2. The regression's standard error, adjusted R-squared, and p-value do not give arguments in favor of one or the other. However, since the individual coefficient on age^2 is significant at the 0.05 level, suggesting the presence of some non-linearity, we prefer regression (d).

(g)

Do you prefer the regression in (d) to the regression in (c)? Explain.

The difference between regression (c) and regression (d) is that $log(age)$ is replaced by a quadratic polynomial in age and age^2. The regression's standard error, adjusted R-squared, and p-value do not give arguments

in favor of one or the other. We prefer regression (c) because it is parsimonious and the coefficient may be interpreted as an elasticity.

The individual coefficient on age^2 has a smaller p-value than that on $log(age)$, but these coefficients cannot be compared since in regression (d) age^2 is picking up only the non-linear part of the causal relation between age and hourly earnings, while in regression (c) $log(age)$ is picking up all of the causal relation. See below for a graphical comparison of models (c) and (d).

Remark

Differences in the R-squared across the models are non-existent or tiny and cannot be used to discriminate across models, despite what some model answers out there in the internets are saying.

(h)

Plot the regression relation between age and $ln(ahe)$ from (b), (c), and (d) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?

Create labels

Create labels to replace "0" and "1" with "Male"/"Female" and "High School"/"Bachelor"

```
female.labs <- c("0" = "Male", "1" = "Female")
bachelor.labs <- c("0" = "High School", "1" = "Bachelor")
```

Augment the datasets

Augment the datasets with the regression results (from broom package):

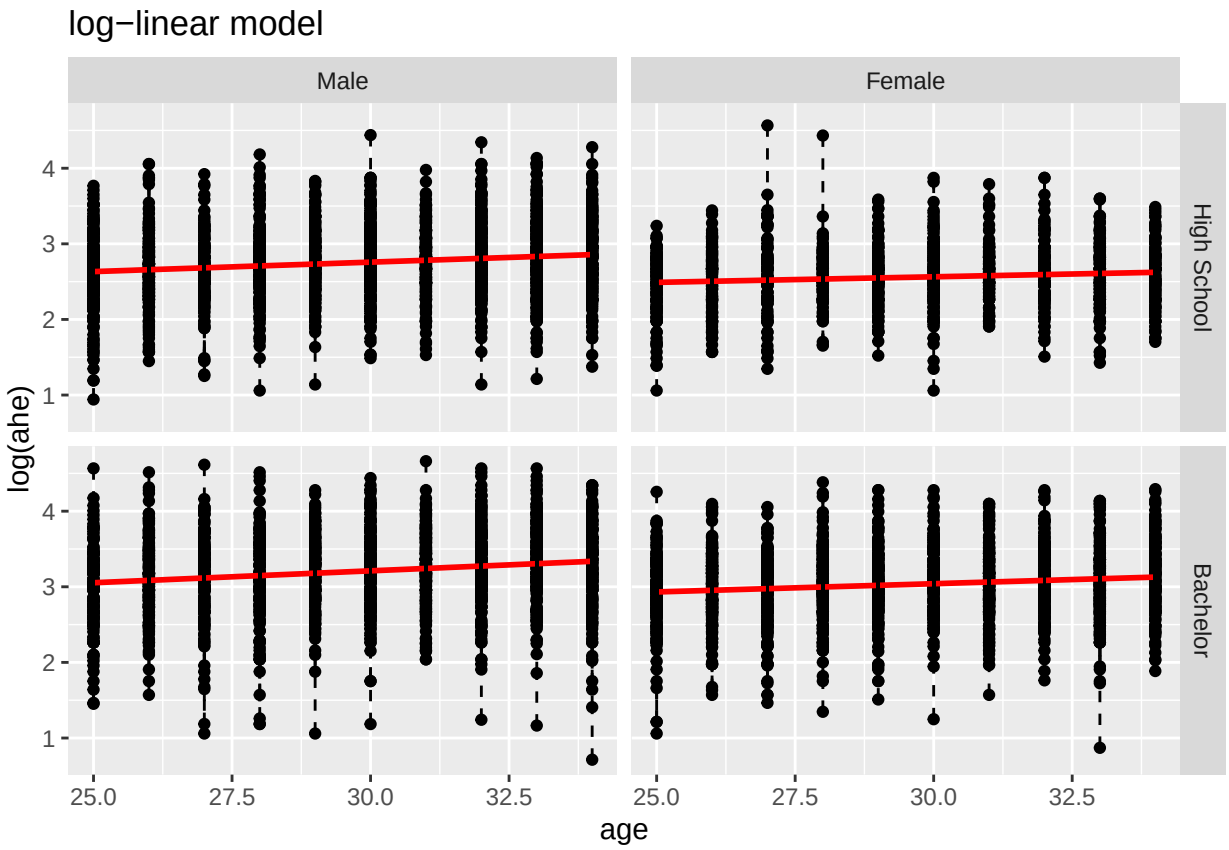
```
dm1 <- augment(m1, data=df)
dm2 <- augment(m2, data=df)
dm3 <- augment(m3, data=df)
dm4 <- augment(m4, data=df)
```

Plot of $log(ahe|female = 0, bachelor = 0)$ against age

for regression (b)

```
ggplot(data=dm2, aes(x=age, y=log(ahe))) +
    geom_point() +
    facet_grid(bachelor~female,
               labeller=labeller(bachelor=bachelor.labs, female=female.labs)) +
    geom_smooth(method="lm", color="red",se=FALSE) +
    geom_segment(aes(xend=age, yend = .fitted), linetype="dashed") +
    ggtitle("log-linear model")
```

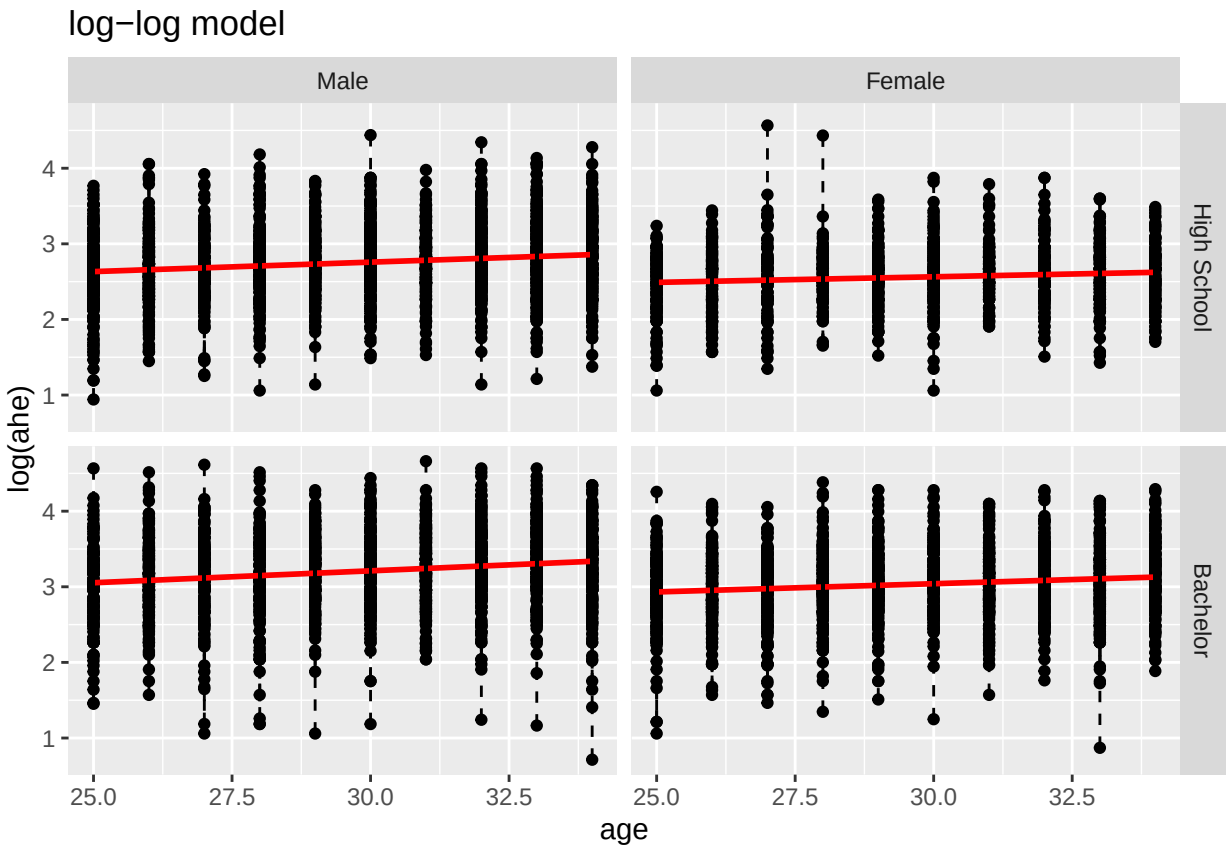
```
## `geom_smooth()` using formula 'y ~ x'
```



for regression (c)

```
ggplot(data=dm3, aes(x=age, y=log(ahe))) +
    geom_point() +
    facet_grid(bachelor~female,
               labeller=labeller(bachelor=bachelor.labs, female=female.labs)) +
    geom_smooth(method="lm", color="red",se=FALSE) +
    geom_segment(aes(xend=age, yend = .fitted), linetype="dashed") +
    ggtitle("log-log model")
```

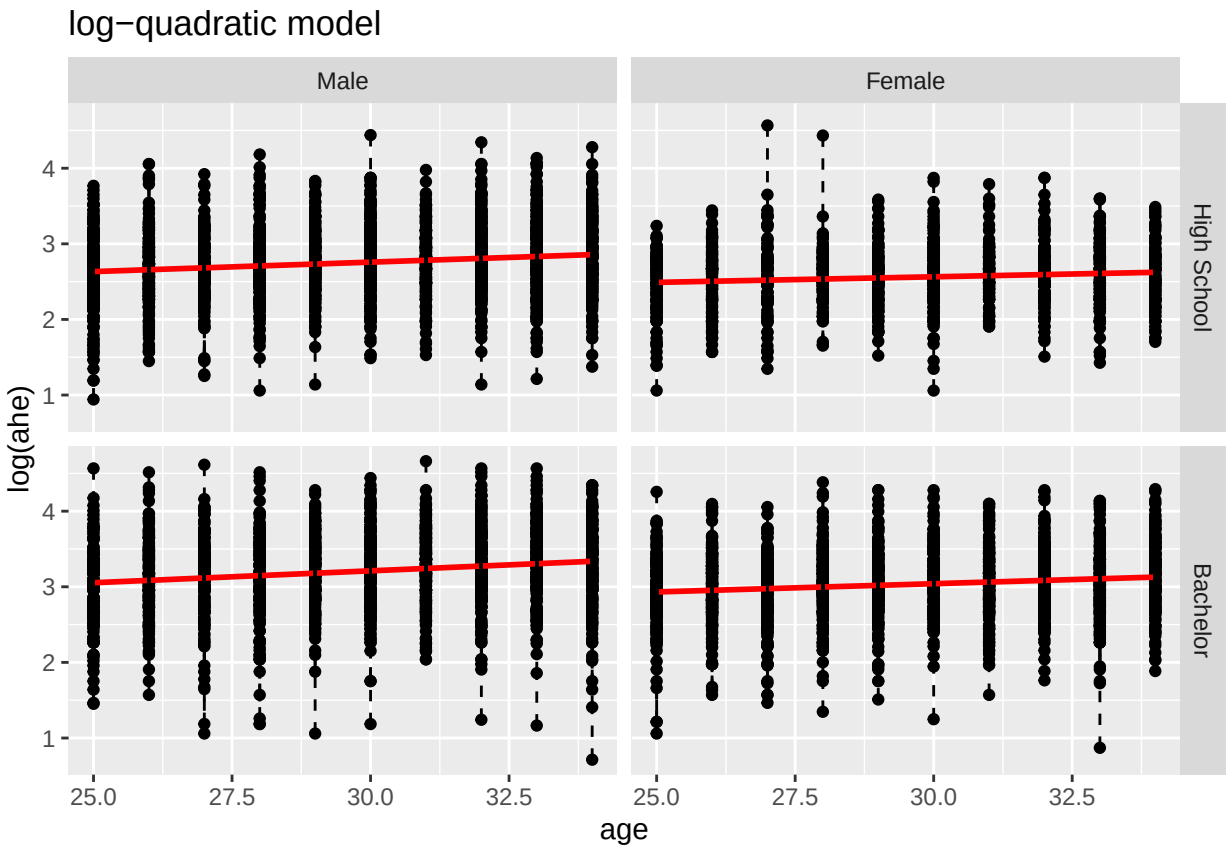
```
## `geom_smooth()` using formula 'y ~ x'
```



for regression (d)

```
ggplot(data=dm4, aes(x=age, y=log(ahe))) +
    geom_point() +
    facet_grid(bachelor~female,
               labeller=labeller(bachelor=bachelor.labs, female=female.labs)) +
    geom_smooth(method="lm", color="red",se=FALSE) +
    geom_segment(aes(xend=age, yend = .fitted), linetype="dashed") +
    ggtitle("log-quadratic model")
```

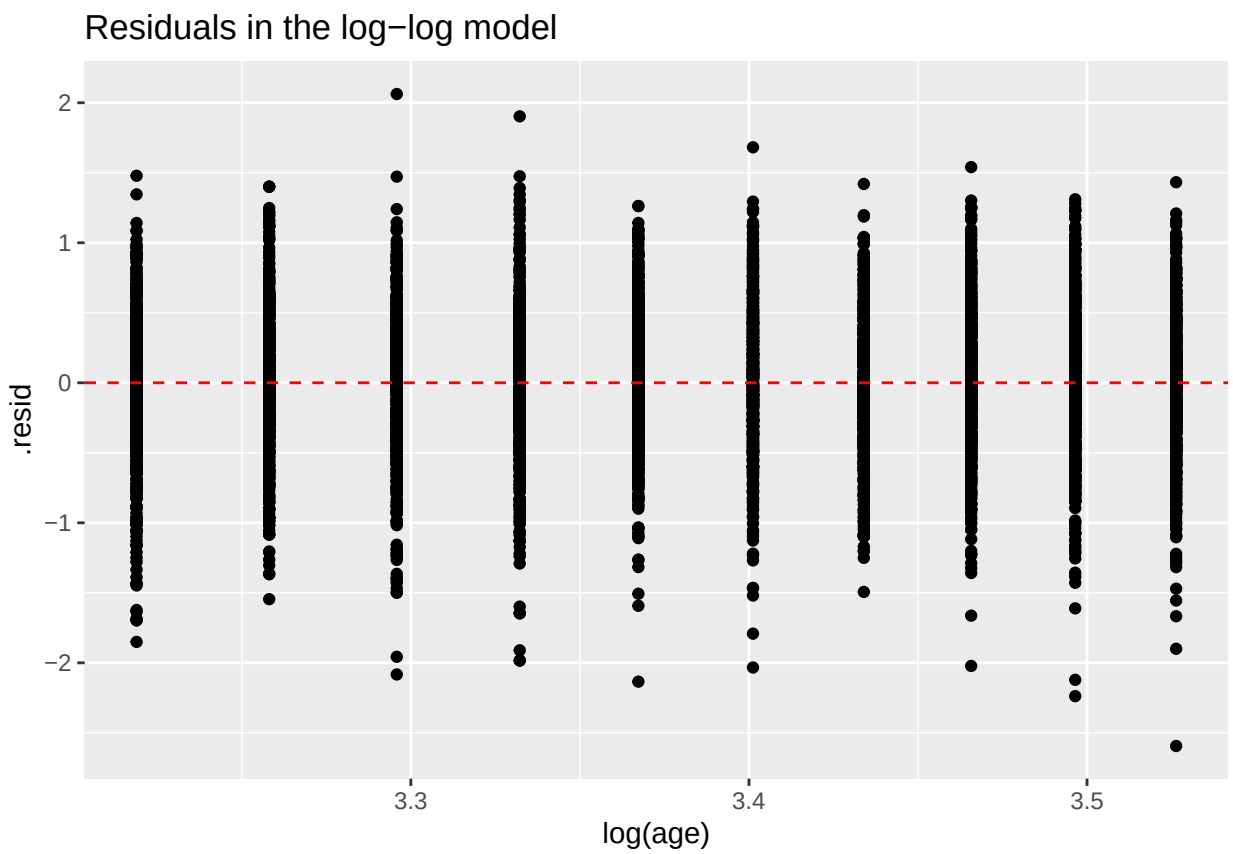
```
## `geom_smooth()` using formula 'y ~ x'
```



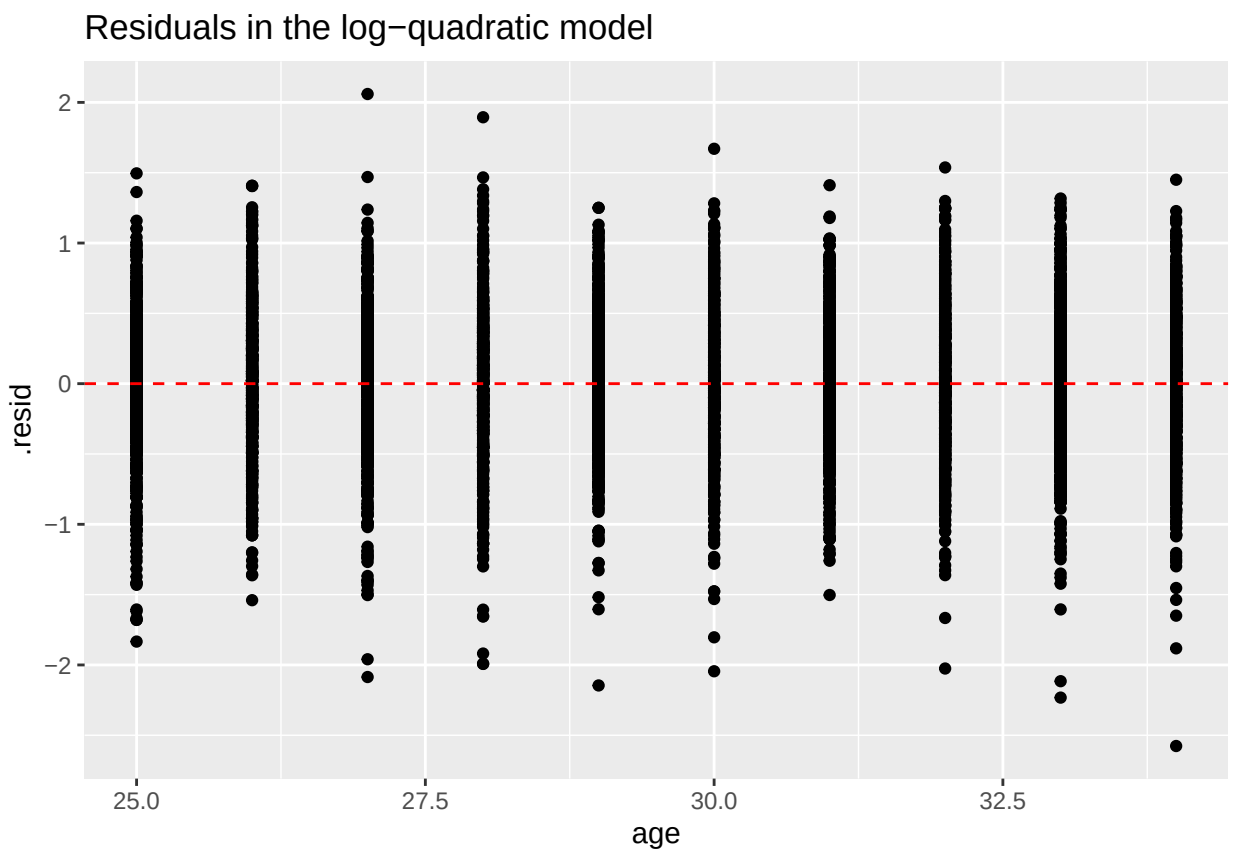
Discussion: Model (c) vs. Model (d)

Plot of $log(ahe)$ against residuals

```
ggplot(dm3, aes(log(age), .resid)) +
  geom_point() +
  geom_hline(yintercept=0, color="red", linetype='dashed') +
  ggtitle("Residuals in the log-log model")
```




```
ggplot(dm4, aes(age, .resid)) +
  geom_point() +
  geom_hline(yintercept=0, color="red", linetype='dashed') +
  ggtitle("Residuals in the log-quadratic model")
```



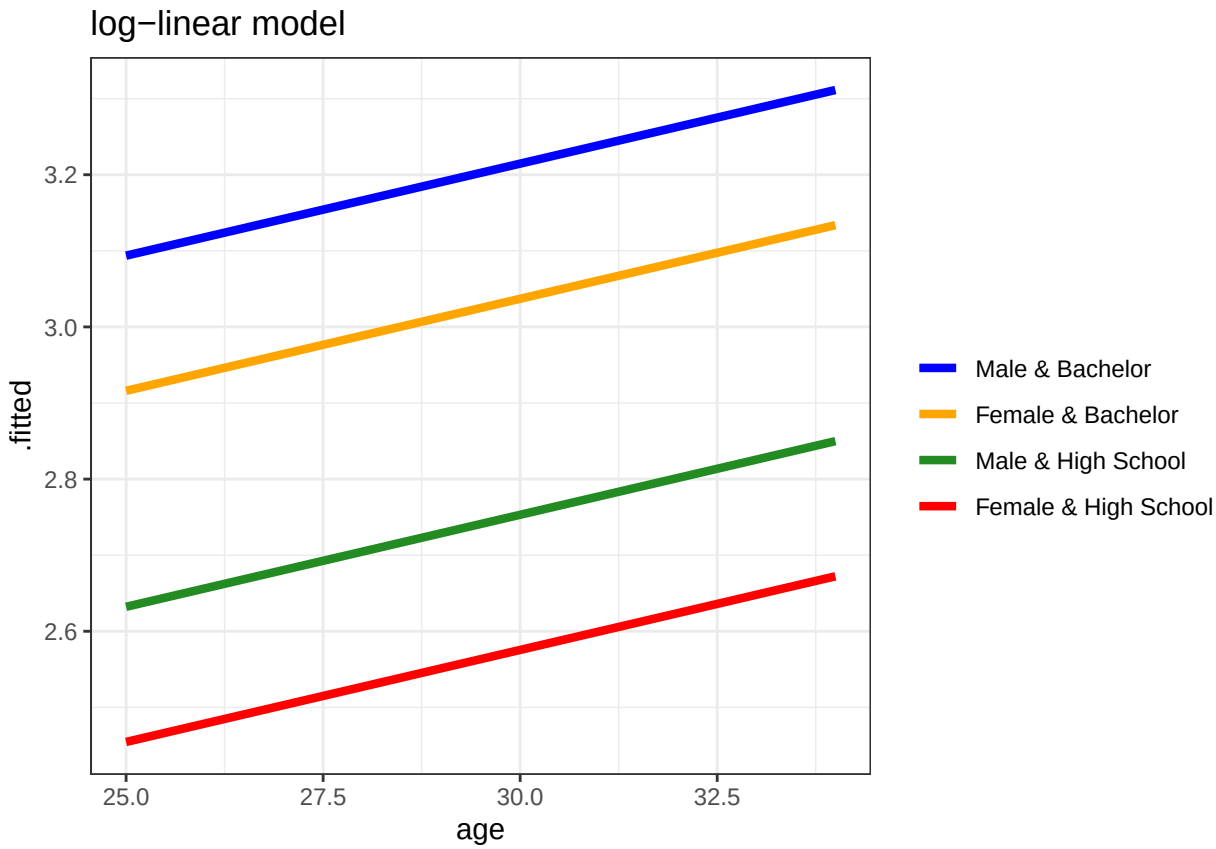
Plot of $log(ahe)$ against standardized residuals: qq plot

```
ggplot(dm3, aes(sample=.std.resid)) +
  geom_qq() +
  geom_qq_line(color="red") +
  ggtitle("qq-plot: standardized residuals in the log-log model")
```

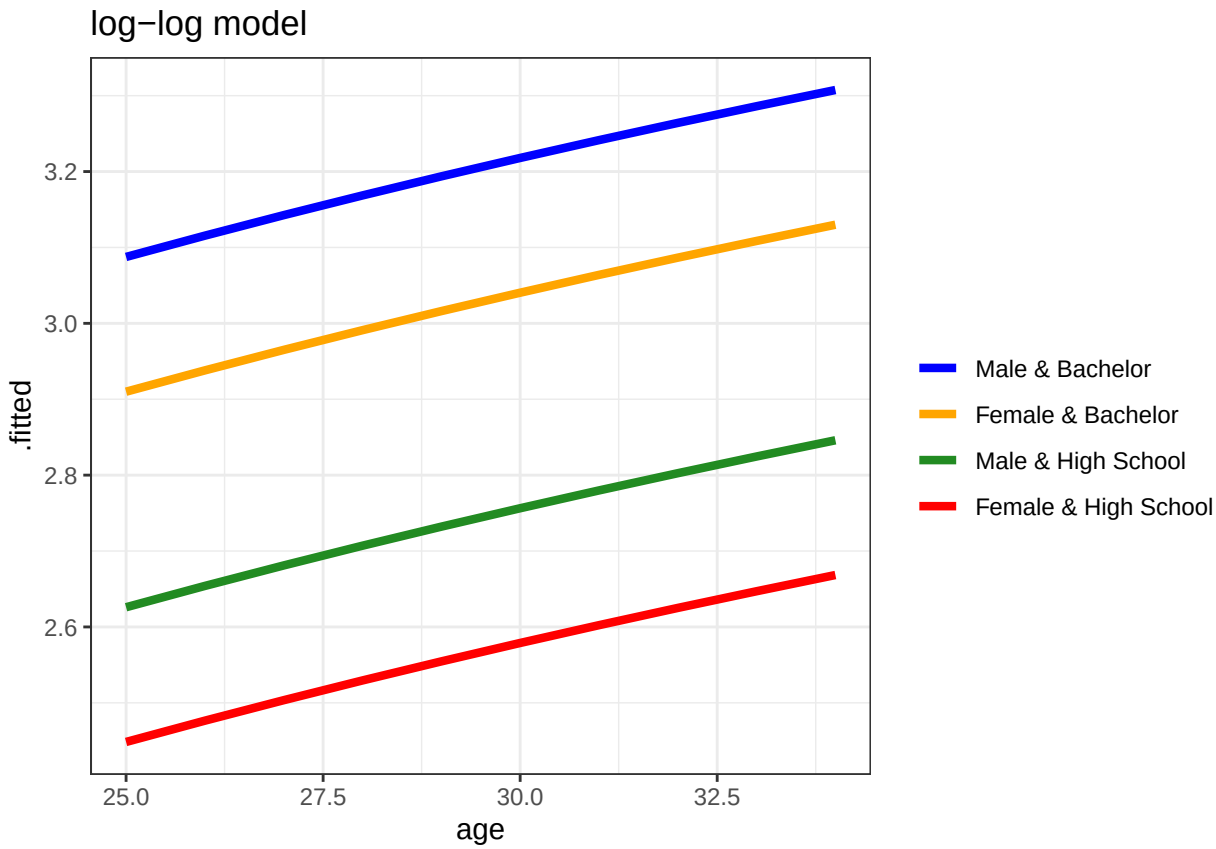

```
ggplot(dm4, aes(sample=.std.resid)) +
  geom_qq() +
  geom_qq_line(color="red") +
  ggtitle("qq-plot: standardized residuals in the log-quadratic model")
```

Models compared

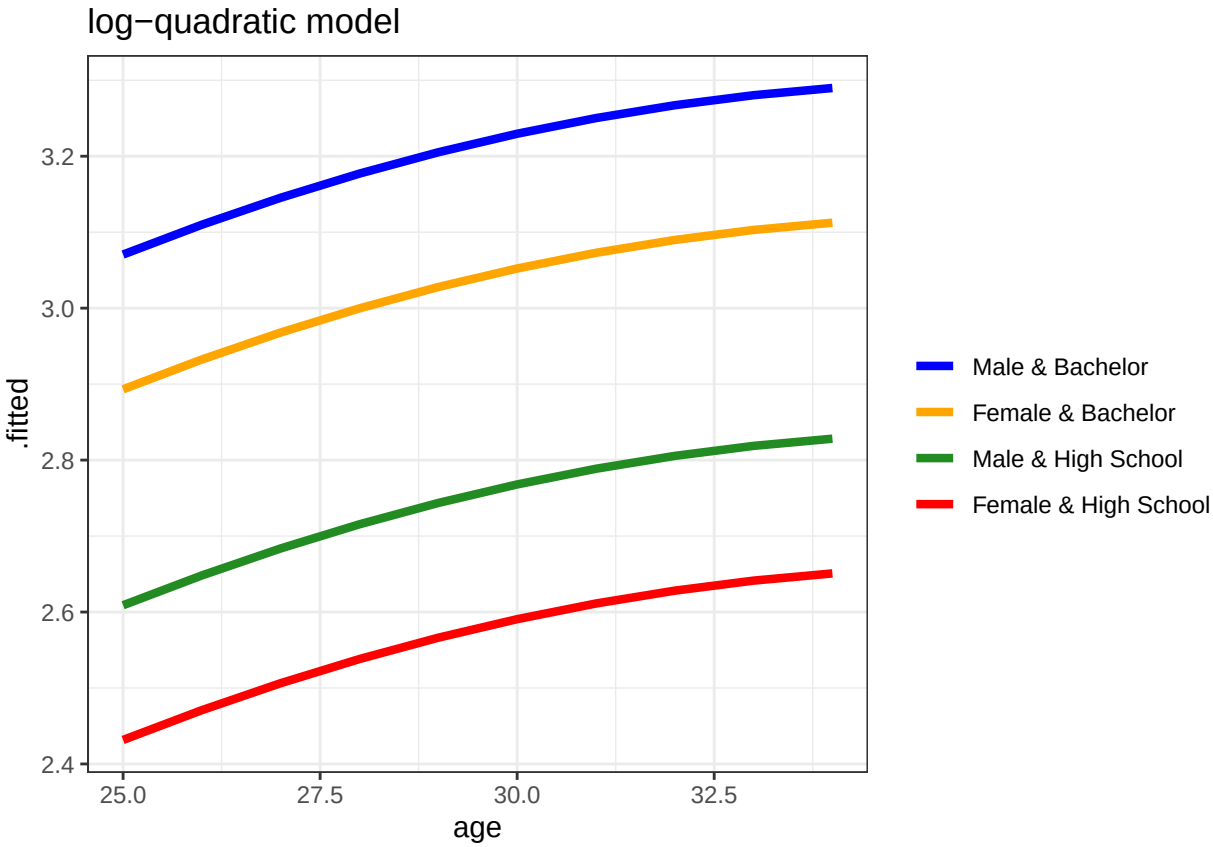
```
print(p2)
```


log−linear model
.fitted
3.2
3.0
2.8
2.6
25.0
27.5
30.0
32.5
age
Male & Bachelor
Female & Bachelor
Male & High School
Female & High School

```
print(p3)
```


log−log model
.fitted
3.2
3.0
2.8
2.6
25.0
27.5
30.0
32.5
age
Male & Bachelor
Female & Bachelor
Male & High School
Female & High School

```
print(p4)
```



(i)

Run a regression of $ln(ahe)$ on age, $age2$, $female$, $bachelor$, and the interaction term $female * bachelor$.

```
df$age2 <- (df$age)^2
df$femalebachelor <- df$female * df$bachelor
m5 <- lm(log(ahe) ~ age + age2 + female + bachelor + femalebachelor, data=df)
dm5 <- augment(m5, data=df)
summary(m5)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + age2 + female + bachelor + femalebachelor,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5715 -0.2859  0.0128  0.3042  2.0683
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.411904   0.672122    0.61   0.5400
## age             0.134815   0.045796    2.94   0.0033 **
## age2           -0.001871   0.000774   -2.42   0.0157 *
## female         -0.190324   0.017373  -10.96   <2e-16 ***
## bachelor        0.452114   0.014882   30.38   <2e-16 ***
## femalebachelor  0.023474   0.023384    1.00   0.3155
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 7092 degrees of freedom
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.209
## F-statistic:  375 on 5 and 7092 DF,  p-value: <2e-16
```

What does the coefficient on the interaction term measure?

The coefficient shows the extra effect on earnings for females with a bachelor degree that is not captured by the average female and not captured by the average bachelor.

Alexis and Jane: 30-year old females

Alexis is a 30-year-old female with a bachelor's degree. What does the regression predict for her value of $ln(ahe)$? Jane is a 30-year-old female with a high school diploma. What does the regression predict for her value of $ln(ahe)$? What is the predicted difference between Alexis's and Jane's earnings?

```
alexis <- data.frame(age=30, age2=30^2, bachelor=1, female=1, femalebachelor=1)
log.ahe.alexis = predict(m5, newdata=alexis)
ahe.alexis = exp(log.ahe.alexis)
jane <- data.frame(age=30, age2=30^2, bachelor=0, female=1, femalebachelor=0)
log.ahe.jane = predict(m5, newdata=jane)
ahe.jane = exp(log.ahe.jane)
```

The predicted difference between Alexis's and Jane's earnings is about $8,053.7$ dollars:

```
ahe.alexis - ahe.jane
```

```
##      1
## 8.0537
```

Bob and Jim: 30-year old males

Bob is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of $ln(ahe)$? Jim is a 30-year-old male with a high school diploma. What does the regression predict for his value of $ln(ahe)$? What is the predicted difference between Bob's and Jim's earnings?

```
bob <- data.frame(age=30, age2=30^2, bachelor=1, female=0, femalebachelor=1)
log.ahe.bob = predict(m5, newdata=bob)
ahe.bob = exp(log.ahe.bob)
jim <- data.frame(age=30, age2=30^2, bachelor=0, female=0, femalebachelor=0)
log.ahe.jim = predict(m5, newdata=jim)
ahe.jim = exp(log.ahe.jim)
```

The predicted difference between Bob's and Jim's earnings is about $9,742.1$ dollars:

```
ahe.bob - ahe.jim
```

```
##     1
## 9.742
```

Difference-in-Difference

The difference in the difference predicted effects is about $9,742.1 - 8,053.7 = 1,688.4$ dollars:

```
(ahe.bob - ahe.jim) - (ahe.alexis - ahe.jane)
```

```
##      1
## 1.6884
```

The predicted difference in log-earnings between Alexis and Jane is about 0.47559 log-dollars. The predicted difference in log-earnings between Bob and Jim is about 0.47559 log-dollars. That just happens to be the same! Thus, the difference in difference in log-earnings is about zero log-dollars. An earlier edition of Stock and Watson used a different sample dataset and reports a near-zero difference (0.063), which is consistent with our findings.

(j)

Is the effect of Age on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question.

A parsimonious model is the following:

```
m6 <- lm(log(ahe) ~ log(age) + bachelor + log(age)*female + bachelor*female, data=df)
summary(m6)
```

```
##
## Call:
## lm(formula = log(ahe) ~ log(age) + bachelor + log(age) * female +
##     bachelor * female, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6044 -0.2856  0.0035  0.3030  2.0577
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.0512     0.2580   -0.20    0.843
## log(age)          0.8273     0.0762   10.86   <2e-16 ***
## bachelor          0.4515     0.0149   30.34   <2e-16 ***
## female            0.6888     0.3967    1.74    0.083 .
## log(age):female  -0.2597     0.1171   -2.22    0.027 *
## bachelor:female   0.0225     0.0234    0.96    0.336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 7092 degrees of freedom
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.209
## F-statistic:  375 on 5 and 7092 DF,  p-value: <2e-16
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
H0 <- c("female=0", "log(age):female", "bachelor:female=0")
tidy(linearHypothesis(m6, H0))
```

```
## # A tibble: 2 x 6
```

```
##   res.df   rss    df sumsq statistic   p.value
##    <dbl> <dbl> <dbl> <dbl>     <dbl>     <dbl>
## 1   7095 1670.    NA  NA        NA   NA
## 2   7092 1615.     3  54.5      79.7  1.03e-50
```

To test whether the effect of *age* on *ahe* is different for men and women, we test the hypothesis that all coefficients involving $female$ are jointly zero. This hypothesis can be firmly rejected since the joint p-value is essentially zero.

(k)

Is the effect of Age on earnings different for high school graduates than for college graduates? Specify and estimate a regression that you can use to answer this question.

```
m7 <- lm(log(ahe) ~ log(age) + female + log(age)*bachelor + female*bachelor, data=df)
summary(m7)
```

```
##
## Call:
## lm(formula = log(ahe) ~ log(age) + female + log(age) * bachelor +
##     female * bachelor, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5990 -0.2856  0.0087  0.3024  2.0640
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.5865     0.2850    2.06     0.04 *
## log(age)            0.6389     0.0842    7.59  3.6e-14 ***
## female             -0.1901     0.0174  -10.94  < 2e-16 ***
## bachelor           -0.0518     0.3929   -0.13     0.90
## log(age):bachelor   0.1488     0.1159    1.28     0.20
## female:bachelor     0.0237     0.0234    1.01     0.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 7092 degrees of freedom
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.208
## F-statistic:  375 on 5 and 7092 DF,  p-value: <2e-16
```

```
library(car)
H0 <- c("bachelor=0", "log(age):bachelor", "female:bachelor=0")
tidy(linearHypothesis(m7, H0))
```

```
## # A tibble: 2 x 6
##   res.df   rss    df sumsq statistic    p.value
##    <dbl> <dbl> <dbl> <dbl>     <dbl>      <dbl>
## 1   7095 1985.    NA   NA        NA   NA
## 2   7092 1616.     3  369.      540.   3.60e-316
```

To test whether the effect of *age* on *ahe* is different for university graduates and high-school graduates, we test the hypothesis that all coefficients involving *bachelor* are jointly zero. This hypothesis can be firmly rejected since the joint p-value is essentially zero.

(l)

After running all these regressions (and any others that you want to run), summarize the effect of age on earnings for young workers.

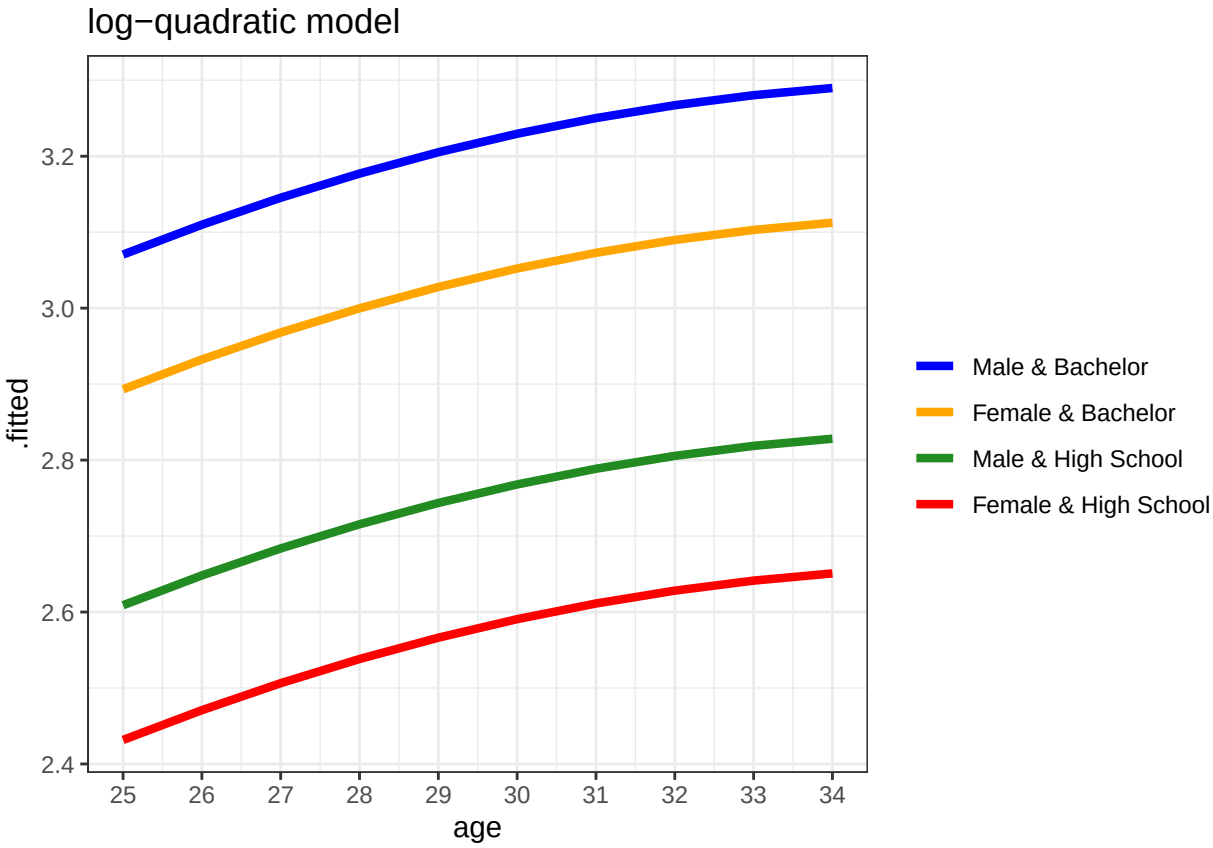
The following models fit about 21% of the deviation of average hourly earnings from the sample mean. It's not immediately obvious that one should be preferred over the other.

Log-quadratic model:

```
m8 <- lm(log(ahe) ~ poly(age,2,raw=TRUE) + bachelor + female, data=df)
summary(m8)
```

```
##
## Call:
## lm(formula = log(ahe) ~ poly(age, 2, raw = TRUE) + bachelor +
##     female, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5764 -0.2868  0.0126  0.3041  2.0596
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  0.418745   0.672088    0.62   0.5333
## poly(age, 2, raw = TRUE)1    0.134115   0.045791    2.93   0.0034 **
## poly(age, 2, raw = TRUE)2 -0.001860   0.000774   -2.40   0.0163 *
## bachelor                     0.461629   0.011473   40.24   <2e-16 ***
## female                      -0.177364   0.011626  -15.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 7093 degrees of freedom
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.209
## F-statistic:  469 on 4 and 7093 DF,  p-value: <2e-16
```

```
dm8 <- augment(m8, data=df)
ggplot() +
  geom_line(data=dm8, aes(x=age, y=.fitted,
                          group=interaction(-female, bachelor),
                          color=interaction(-female, bachelor)),
            size=1.5) +
  scale_x_continuous(breaks=seq(25,35,1)) +
  scale_color_manual(name="",labels=labs, values=c("red", "forestgreen", "orange", "blue")) +
  guides(color=guide_legend(reverse=TRUE)) +
  ggtitle("log-quadratic model") +
  theme_bw()
```



Log-Log model:

```
m9 <- lm(log(ahe) ~ log(age) + bachelor + female, data=df)
summary(m9)
```

```
##
## Call:
## lm(formula = log(ahe) ~ log(age) + bachelor + female, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.594 -0.287  0.010  0.302  2.062
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3233     0.1961    1.65    0.099 .
## log(age)      0.7154     0.0578   12.37   <2e-16 ***
## bachelor      0.4615     0.0115   40.22   <2e-16 ***
## female       -0.1775     0.0116  -15.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 7094 degrees of freedom
## Multiple R-squared:  0.209,  Adjusted R-squared:  0.208
## F-statistic:  623 on 3 and 7094 DF,  p-value: <2e-16
```

```
dm9 <- augment(m9, data=df)
ggplot() +
  geom_line(data=dm9, aes(x=age, y=.fitted,
                          group=interaction(-female, bachelor),
                          color=interaction(-female, bachelor)),
            size=1.5) +
  scale_x_continuous(breaks=seq(25,35,1)) +
  scale_color_manual(name="",labels=labs, values=c("red", "forestgreen", "orange", "blue")) +
  guides(color=guide_legend(reverse=TRUE)) +
  ggtitle("log-log model") +
  theme_bw()
```