Log-Transformations: Outcome

Andrew Zieffler

April 25, 2016

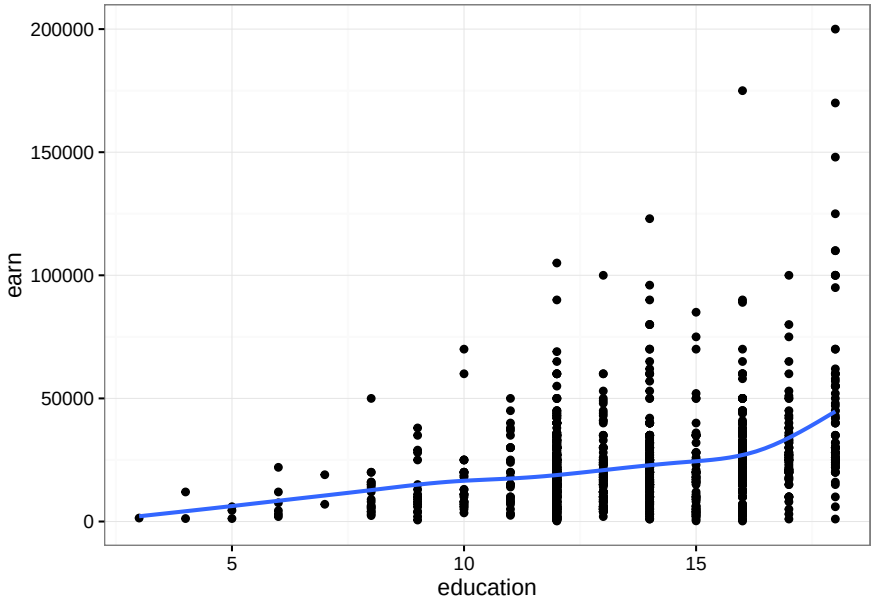
Read in Data and Load Libraries

```
earnings = read.csv(file = "~/Documents/EPsy-8262/data/earnings.csv")
head(earnings)
```

```
##   earn height female education
## 1 1000     73      0        18
## 2 1000     65      0        12
## 3 1200     66      0        12
## 4 1500     73      0        12
## 5 1700     65      0        12
## 6 2000     72      0        15
```

```
# Load libraries
library(ggplot2)
library(sm)
```

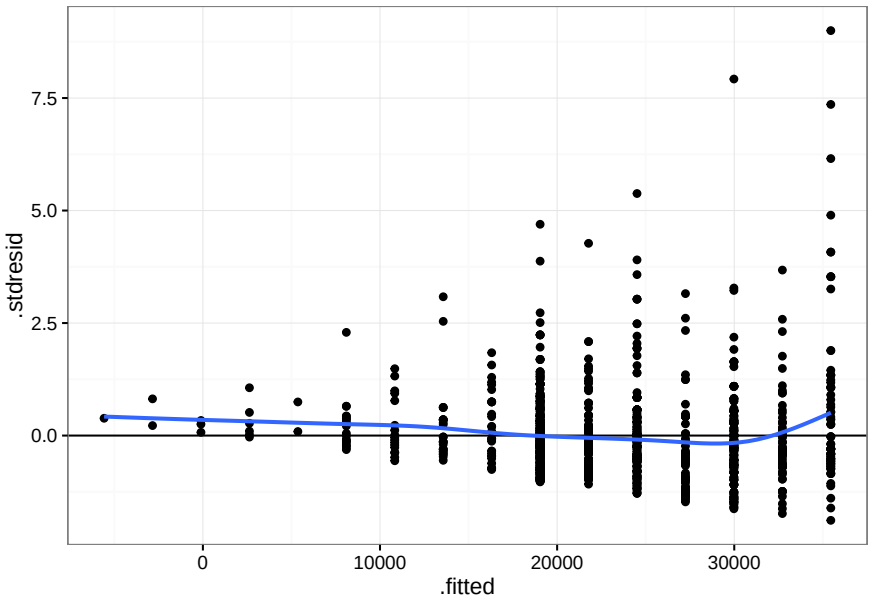
Examine Relationship between Earnings and Education Level



The loess line suggests that the relationship between education level and earnings is non-linear. A one-year difference in education does have the same effect on earnings. . . for low education levels, a one-year difference in education is associated with a smaller changes in earnings than the same one-year difference for higher education levels. There is also a clear increase in variation of earnings (heteroskedasticity) as education level increases.

Sometimes this non-linear relationship and heteroskedasticity is easier to see in the residual plots.

```
lm.1 = lm(earn ~ education, data = earnings)
out = fortify(lm.1)

ggplot(data = out, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(se = FALSE) +
  theme_bw()
```



The scatterplot of the standardized residuals versus the fitted values suggest that the assumption of linearity is likely violated. There is systematic under-estimation for fitted values below 15,000; systematic over-estimation for fitted values between 20,000 and 30,000; and systematic under-estimation for fitted values higher than 35,000. The plot also shows a clear violation of the homogeneity of variance assumption.

Create Natural Logarithm (Base-e) of the Outcome

This relationship, that of exponential growth, is consistent with a logarithmic relationship. To model this type of function we will transform the outcome using the natural logarithm. Recall that the base we use for transformations is irrelevant outside of the interpretation. The use of the natural logarithm, in many cases, leads to reasonable interpretations.

```
# The default base in the log() function is e
earnings$Learn = log(earnings$earn)
head(earnings)
```

```
##   earn height female education    Learn
## 1 1000     73      0        18 6.907755
## 2 1000     65      0        12 6.907755
## 3 1200     66      0        12 7.090077
## 4 1500     73      0        12 7.313220
## 5 1700     65      0        12 7.438384
## 6 2000     72      0        15 7.600902
```

The natural log of the outcome, `Learn`, is the result of taking $e^{\text{Learn}} = \text{earn}$. So for the first person in the dataset, $e^{6.907755} = 1000$. The mathematical constant e is an irrational, transcendental number that is approximately equal to 2.71. (Wikipedia lists the first 50 decimal values of e as 2.71828182845904523536028747135266249775724709369995...) To use e in computations in R we use the `exp()` function. For example,

```
# e
exp(1)
```

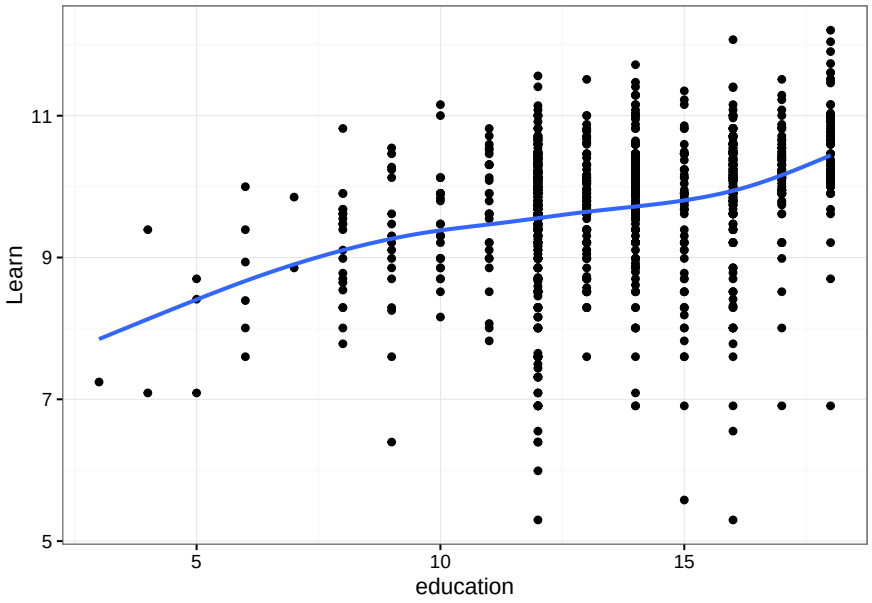
```
## [1] 2.718282
```

```
#e^6.907755
exp(6.907755)
```

```
## [1] 999.9997
```

Examining the relationship between the `Learn` predictor and graduation rates, we see that these variables have a linear relationship. An added bonus is that taking the logarithm of the outcome alleviates the heterongeneity of variance problem we saw earlier. Because of this, taking the logarithm of an outcome is sometimes referred to as a variance stabilizing technique.

```
ggplot(data = earnings, aes(x = education, y = Learn)) +
    geom_point() +
    geom_smooth(se = FALSE) +
    theme_bw()
```



Fitting and Interpreting the Log-Transformed Model

```
lm.2 = lm(Learn ~ education, data = earnings)
summary(lm.2)
```

```
##
## Call:
## lm(formula = Learn ~ education, data = earnings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7219 -0.3197  0.1064  0.5655  2.0523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.05951    0.14327   56.25   <2e-16 ***
## education    0.12254    0.01044   11.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8722 on 1190 degrees of freedom
## Multiple R-squared:  0.1037, Adjusted R-squared:  0.103
## F-statistic: 137.7 on 1 and 1190 DF,  p-value: < 2.2e-16
```

Note the model-level summary: Differences in education explains 87.22% of the variation in earnings (explaining variation in log-earnings is the same as explaining variation in earnings). This is statistically reliable, $F(1, 1190) = 137.7$, $p < 0.001$.

The fitted equation is

$$\hat{\text{Learn}} = 8.06 + 0.12(\text{Education})$$

To interpret the coefficients:

- $\hat{\beta}_0 = 8.06$. This is the average estimated log-earnings for people with zero years of education. To equate this to earnings (instead of log-earnings) we back-transform; $e^{8}.06 = 3165.29$. The average earnings for people with zero years of education is predicted as \$3165.29.
- $\hat{\beta}_1 = 0.12254$. A one-year difference in education is associated with a 0.12-unit difference in log-earnings, on average. To interpret this further, we again, back-transform the slope. A 0.12254-unit difference in log-earnings is equivalent to a $e^{0}.12254 = 1.13$-fold increase in earnings.

To understand this consider three people with education levels of 10, 11, and 12 years, respectively (all one-year in difference). Their predicted log-earnings are:

```
new = data.frame(education = c(10, 11, 12))
predict(lm.2, newdata = new)
```

```
##        1        2        3
## 9.284935 9.407478 9.530021
```

If we back-transform these values:

```
exp(predict(lm.2, newdata = new))
```

```
##        1        2        3
## 10774.47 12179.12 13766.88
```

Going from 10 years of education to 11 years of education (difference of one year) is associated with increased earnings of 1.13 times. Similarly, the expected increase in earnings is a factor of 1.13 for any one-unit difference in education.

```
10774.47 * 1.130368
```

```
## [1] 12179.12
```

```
12179.12 * 1.130368
```

```
## [1] 13766.89
```

Exponential growth models are akin to the adage "the rich get richer". Every one-year difference increases earnings by a factor of 1.13, but this has a larger effect for higher values of education... multiplying bigger numbers by 1.13 has a different effect than multiplying lower values by 1.13.

Percentage Change

Another way to talk about changing the outcome by a factor of 1.13 is to use the idea of a percentage change. In this case, earnings (the variable we log-transformed) increases by 13%. Note that this is equivalent to computing:

$$e^{\hat{\beta}_1} - 1$$

So an alternative interpretation is to say that each one-year difference in education is associated with a 13% increase in earnings, on average.

Shortcut

A shortcut to this is to say that a one-unit difference in X is associated with a $100 \times \hat{\beta}_1$ percent difference in Y. In our example, a one-year difference in education is associated with a 12.25% increase in earnings, on average. This is close (although not exact) to the value we computed earlier. This will give you a rough approximation when you do not have a computer handy to compute $e^{\hat{\beta}_1} - 1$. # Using the Natural Logarithm as a Transformation of the Predictor

Let's go back to the Minnesota schools data we looked at in the previous set of notes, and use the natural log to transform SAT scores.

```
mn = read.csv(file = "/Users/andrewz/Documents/EPsy-8262/data/mnSchools.csv")
mn$Lsat = log(mn$sat)
head(mn)
```

```
##   id                              name gradRate public  sat tuition
## 1  1                  Augsburg College     65.2      0 1030   39294
## 2  3           Bethany Lutheran College     52.6      0 1065   30480
## 3  4 Bethel University, Saint Paul, MN     73.3      0 1145   39400
## 4  5                  Carleton College     92.6      0 1400   54265
## 5  6         College of Saint Benedict     81.1      0 1185   43198
## 6  7     Concordia College at Moorhead     69.4      0 1145   36590
##       Lsat
## 1 6.937314
```

```
## 2 6.970730
## 3 7.043160
## 4 7.244228
## 5 7.077498
## 6 7.043160
```

```r
# Fit model
lm.3 = lm(gradRate ~ Lsat, data = mn)
summary(lm.3)
```

```
##
## Call:
## lm(formula = gradRate ~ Lsat, data = mn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3006  -6.1058  -0.1169   5.6295  13.7831
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1013.9       93.1  -10.89 4.02e-12 ***
## Lsat           153.6       13.3   11.55 9.30e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.386 on 31 degrees of freedom
## Multiple R-squared:  0.8113, Adjusted R-squared:  0.8053
## F-statistic: 133.3 on 1 and 31 DF,  p-value: 9.296e-13
```

The model-level summary is exactly the same as the other models we fitted: Differences in the natural log of SAT scores, which is the same thing as differences in SAT scores, explains 81.13% of the variation in graduation rates. This is statistically reliable, $F(1, 31) = 133.3$, $p < 0.001$.

The fitted equation is

$$\widehat{\text{gradRate}} = -1013.872 + 153.6(\text{Lsat})$$

To interpret the coefficients:

- $\hat{\beta}_0 = -1013.872$. This is the average estimated graduation rate when `Lsat` is equal to 0. Equivalently, when `Lsat` $= 0$, SAT $= e^0 = 1$. The average estimated graduation rate for all school that have an SAT score of 1 is -1013.872.
- $\hat{\beta}_1 = 153.6$. A one-unit difference in `Lsat` is associated with a 153.6% difference in graduation rate, on average. A one-unit difference in `Lsat` is equivalent to a 2.71-fold difference in SAT.

To better interpret the slope, remember that when we use the natural logarithm we can use the interpretation of "percentage change". Here the predictor is the variable that has been log-transformed. So rather than talk about a "one-unit" difference in the predictor we talk about a *one-percent* difference in the predictor.

Consider three schools, each having a SAT score that differs by one-percent; 1000, 1010, 1020.1.

```
new = data.frame(Lsat = log(c(1000, 1010, 1020.1)))
predict(lm.3, newdata = new)
```

```
##        1        2        3
## 46.87784 48.40581 49.93378
```

The predicted graduation rates differ by a constant rate of 1.52797.

```
48.40581 - 46.87784
```

```
## [1] 1.52797
```

```
49.93378 - 48.40581
```

```
## [1] 1.52797
```

To understand how we can directly compute this, consider the predicted values for two x-values that differ by one-percent, if we use symbolic notation:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 \left[\ln(x)\right]$$
$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 \left[\ln(1.01x)\right]$$

The difference in their predicted values is:

$$\begin{aligned}
\hat{y}_2 - \hat{y}_1 &= \hat{\beta}_0 + \hat{\beta}_1 \left[\ln(1.01x)\right] - \left(\hat{\beta}_0 + \hat{\beta}_1 \left[\ln(x)\right]\right) \\
&= \hat{\beta}_0 + \hat{\beta}_1 \left[\ln(1.01x)\right] - \hat{\beta}_0 - \hat{\beta}_1 \left[\ln(x)\right] \\
&= \hat{\beta}_1 \left[\ln(1.01x)\right] - \hat{\beta}_1 \left[\ln(x)\right] \\
&= \hat{\beta}_1 \left[\ln(1.01x) - \ln(x)\right] \\
&= \hat{\beta}_1 \left[\ln(\frac{1.01x}{1x})\right]
\end{aligned}$$

If we substitute in any value for x, we can now directly compute this constant difference. Note that a convenient value for x is 1. Then this reduces to:

$$\hat{\beta}_1 \left[\ln(1.01)\right]$$

So now, we can interpret this as: a one-percent difference in x is associated with a $\hat{\beta}_1 \left[\ln(1.01)\right]$-unit difference in Y, on average. In our model

$$153.6 \left[\ln(1.01)\right]$$

```
153.6 * log(1.01)
```

```
## [1] 1.528371
```

This gives you the constant difference exactly. So you can interpret the effect of SAT as, each one-percent difference in SAT score is associated with a difference in graduation rates of 1.528371%, on average. (Note: Here the units for Y were in percent (graduation rate), so the interpretation is percent. You should use whatever the units of y are.)

Shortcut

We can get a pretty good estimate for this value by using the mathematical shortcut of $\frac{\hat{\beta}_1}{100}$. Then, in general a one-percent difference in x is associated with a $\frac{\hat{\beta}_1}{100}$-unit difference in Y, on average. For our example, we would just interpret it as each one-percent difference in SAT score is associated with a difference in graduation rates of 1.53%, on average.

Back to the Earnings Data

Let's plot the results of the model using the log-transformed earnings. First we set up a sequence of x-values... in this case `education`, and predict using the fitted model.

```
# Set up data
plotData = expand.grid(
    education = seq(from = 3, to = 18, by = 0.1)
    )

# Predict
plotData$yhat = predict(lm.2, newdata = plotData)

# Examine data
head(plotData)
```

```
##   education     yhat
## 1       3.0 8.427135
## 2       3.1 8.439389
## 3       3.2 8.451643
## 4       3.3 8.463897
## 5       3.4 8.476152
## 6       3.5 8.488406
```

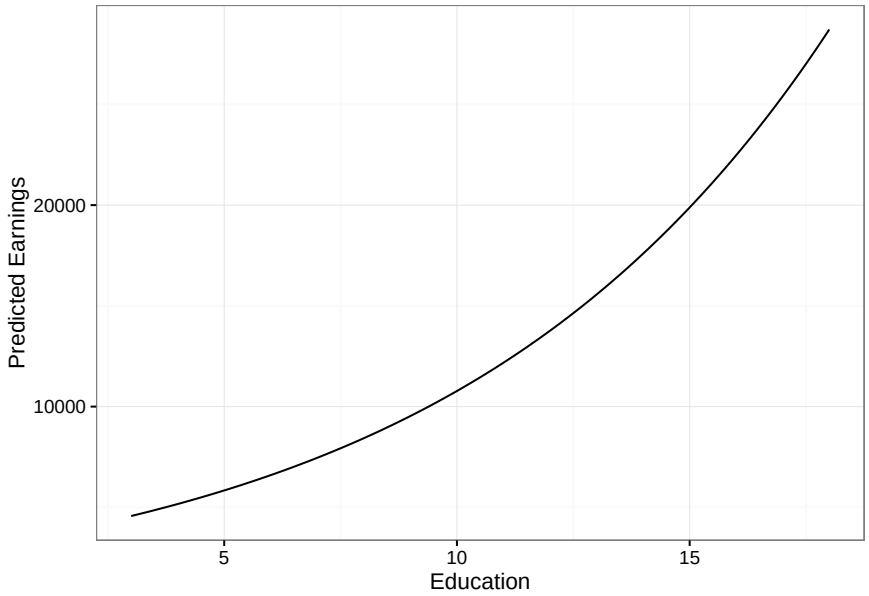
The y-hat values are the natural log of the earnings. After predicting, we can back-transform the earnings to the original metric.

```
# Back-transform any log terms
plotData$earn = exp(plotData$yhat)

# Re-examine data
head(plotData)
```

```
##   education     yhat     earn
## 1       3.0 8.427135 4569.388
## 2       3.1 8.439389 4625.727
## 3       3.2 8.451643 4682.761
## 4       3.3 8.463897 4740.498
## 5       3.4 8.476152 4798.947
## 6       3.5 8.488406 4858.116
```

Now we can plot the back-transformed earnings versus the education values.

```
ggplot(data = plotData, aes(x = education, y = earn)) +
    geom_line() +
    theme_bw() +
  xlab("Education") +
  ylab("Predicted Earnings")
```

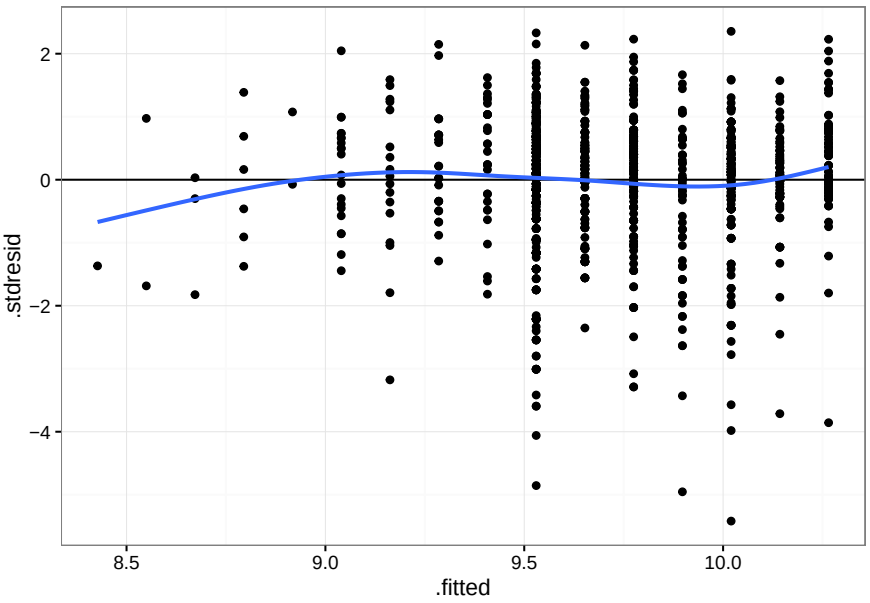


This will display the initial non-linearity between education and earnings that we observed in the original data.

Residual examination

Now let's examine the residuals from the model.

```
out = fortify(lm.2)

ggplot(data = out, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(se = FALSE) +
  theme_bw()
```



Is this indicative of non-linearity? Maybe. Re-look at the earlier scatterplot of the relationship between the natural-log of earnings and education. It has aspects of the same relationship we observed in the MN schools data. To fix that we took the log of the predictor. Here we have already taken the log of the outcome, so our model will be:

$$\ln(Y) = \beta_0 + \beta_1(\ln[X]) + \epsilon$$

To fit this we take the log of the outcome and the predictor and fit a model using those two transformed variables. We already have the log transformed outcome.

```
# Transform predictor
earnings$Leducation = log(earnings$education)
head(earnings)
```

```
##   earn height female education    Learn Leducation
## 1 1000     73      0        18 6.907755   2.890372
## 2 1000     65      0        12 6.907755   2.484907
## 3 1200     66      0        12 7.090077   2.484907
## 4 1500     73      0        12 7.313220   2.484907
## 5 1700     65      0        12 7.438384   2.484907
## 6 2000     72      0        15 7.600902   2.708050
```

```
# Fit log-log model
lm.4 = lm(Learn ~ Leducation, data = earnings)
summary(lm.4)
```

```
##
## Call:
## lm(formula = Learn ~ Leducation, data = earnings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6947 -0.3551  0.1137  0.5612  2.0795
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.8726     0.3324   17.66   <2e-16 ***
## Leducation    1.4861     0.1282   11.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8733 on 1190 degrees of freedom
## Multiple R-squared:  0.1014, Adjusted R-squared:  0.1007
## F-statistic: 134.3 on 1 and 1190 DF,  p-value: < 2.2e-16
```

The fitted equation is:

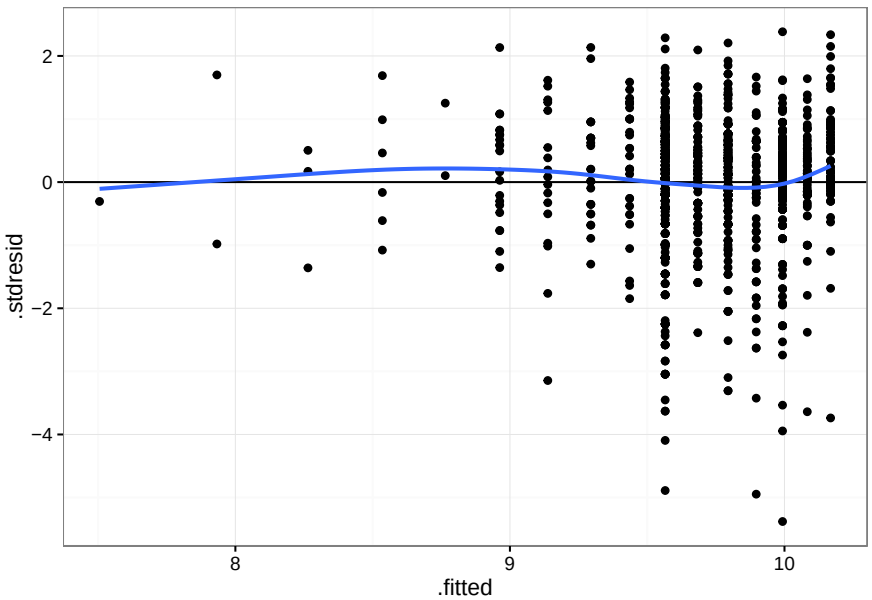
$$\widehat{\text{Learn}} = 5.8726 + 1.4861(\text{Leducation})$$

To interpret the coefficients:

- $\hat{\beta}_0 = 5.8726$. This is the average estimated log-earnings when log-education is equal to 0. Equivalently, when `Leducation` $= 0$, education is $e^0 = 1$. Similarly when `Learn` is 5.8726, earnings are $e^5.8726 = 355.17$. The average estimated earnings for people with one year of education is \$355.17.
- $\hat{\beta}_1 = 1.4861$. A one-unit difference in log-education is associated with a 1.4861-unit difference in log-earnings, on average. Here, since both X and Y used the natural log, we interpret everything as percents. A one percent difference in education is associated with a 1.49% difference in earnings.

Checking the new residuals,

```
out = fortify(lm.4)

ggplot(data = out, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(se = FALSE) +
  theme_bw()
```

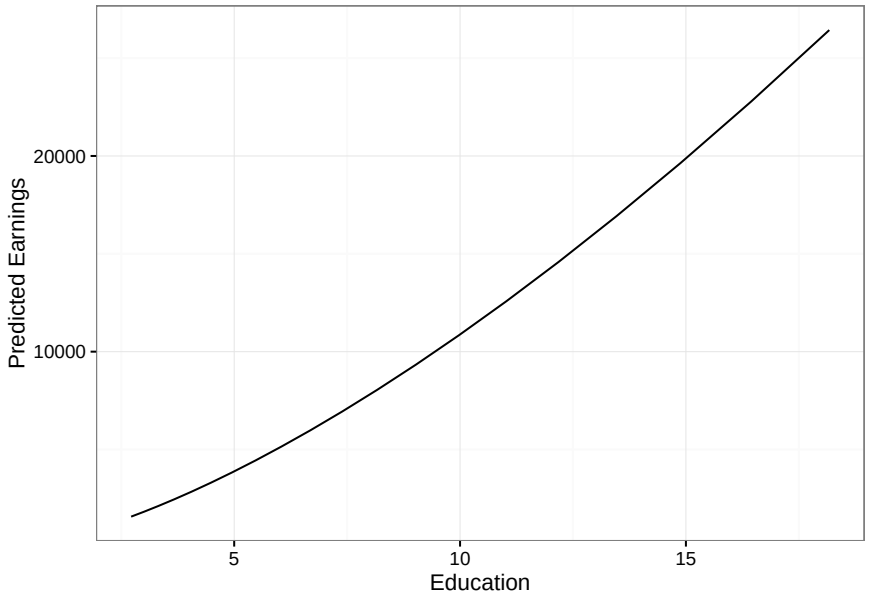


The linearity assumption seems reasonable, but there is still evidence of heterogeneity. What to do? Use a different variance stabilizing technique, or use a different estimation method (e.g., weighted least squares). For now, we will ignore it. How would we plot the results from the model?

```
# Set up data
plotData = expand.grid(
    Leducation = seq(from = 1.0, to = 2.9, by = 0.1)
    )

# Predict
plotData$yhat = predict(lm.4, newdata = plotData)

# Back-transform any log terms
plotData$education = exp(plotData$Leducation)
plotData$earn = exp(plotData$yhat)

# Examine data
head(plotData)
```

```
##   Leducation     yhat education     earn
## 1        1.0 7.358705  2.718282 1569.802
## 2        1.1 7.507318  3.004166 1821.322
## 3        1.2 7.655931  3.320117 2113.142
## 4        1.3 7.804544  3.669297 2451.718
## 5        1.4 7.953157  4.055200 2844.542
## 6        1.5 8.101770  4.481689 3300.305
```

```
# Plot
ggplot(data = plotData, aes(x = education, y = earn)) +
    geom_line() +
    theme_bw() +
  xlab("Education") +
  ylab("Predicted Earnings")
```



Categorical Predictors with a Log-Transformed Outcome

Consider the model where we use `female` to predict variation in earnings. Since we ultimately also might want to include education in the model, we will continue to use log-earnings as the outcome. We include appropriate dummy variables as predictors in these models, the same as we have in previous notes.

```
lm.5 = lm(Learn ~ female, data =earnings)
summary(lm.5)
```

```
##
## Call:
## lm(formula = Learn ~ female, data = earnings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1890 -0.3823  0.1285  0.5734  2.2327
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.02327    0.03926  255.28   <2e-16 ***
## female      -0.53599    0.05172  -10.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8824 on 1190 degrees of freedom
## Multiple R-squared:  0.08278,    Adjusted R-squared:  0.08201
## F-statistic: 107.4 on 1 and 1190 DF,  p-value: < 2.2e-16
```

To interpret the coefficients:

- $\hat{\beta}_0 = 10.02327$. This is the average estimated log-earnings for males (`female` is equal to 0). Back-transforming this, we find the average earnings for males is $e^{10.02327}$ or \$22,545.03.
- $\hat{\beta}_1 = -0.53599$. The average difference in log-earnings between males and females is -0.53599. Thus, the average log-earnigns for females is $10.02327 - 0.53599 = 9.48728$. Or, back-transforming, the average earnings for females is $e^{9.48728}$ or \$13,190.87.

We could also interpret the slope (using out shorcut) as, females, on average, earn $100 * (e^{-0.53599})$ or 58.5% of what males earn.

ANCOVA Model

```
lm.6 = lm(Learn ~ female + Leducation, data =earnings)
summary(lm.6)
```

```
##
## Call:
## lm(formula = Learn ~ female + Leducation, data = earnings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.4676 -0.3615  0.1338  0.5303  2.1477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.26451    0.31997   19.58   <2e-16 ***
## female      -0.52015    0.04896  -10.62   <2e-16 ***
## Leducation   1.45048    0.12264   11.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.835 on 1189 degrees of freedom
## Multiple R-squared:  0.1793, Adjusted R-squared:  0.1779
## F-statistic: 129.9 on 2 and 1189 DF,  p-value: < 2.2e-16
```

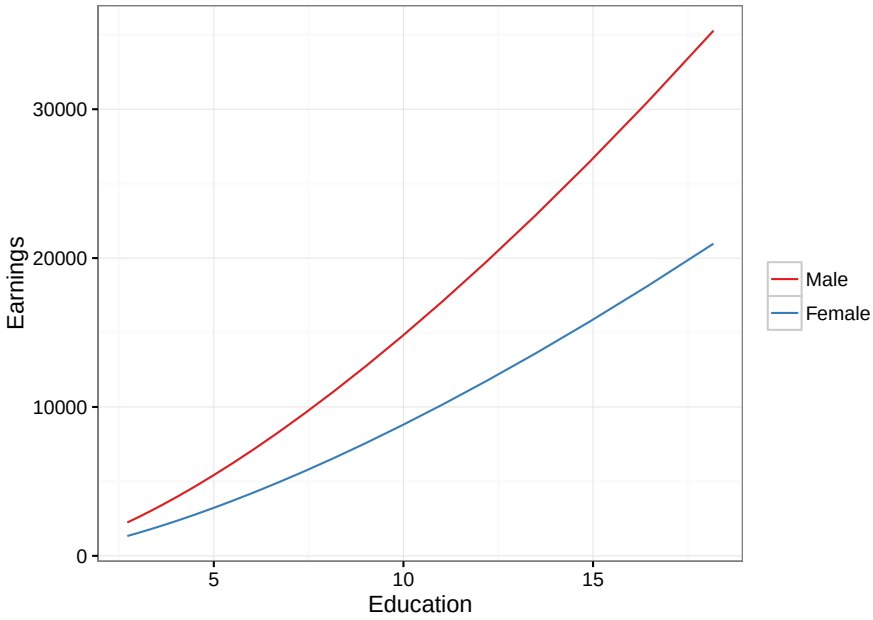
- $\hat{\beta}_0 = 6.26451$. This is the average estimated log-earnings for males (`female` is equal to 0) with education of 1 year (`Leducation` = 0). Back-transforming this, we find the average earnings for males is $e^{6.26451}$ or \$525.6.
- $\hat{\beta}_1 = -0.52015$. The average difference in log-earnings between males and females is -0.52015, controlling for differences in education. Controlling for differences in education, females earn, on average, 59.4% of what males earn.
- $\hat{\beta}_2 = 1.45048$. A one-unit difference in log-education is associated with a 1.45048-unit difference in log-earnings, on average, controlling for sex differences. A one percent difference in education is associated with a 1.45% difference in earnings, controlling for differences in sex.

Plot of the Model Results

```r
# Set up data
plotData = expand.grid(
    Leducation = seq(from = 1.0, to = 2.9, by = 0.1),
    female = c(0, 1)
    )

# Predict
plotData$yhat = predict(lm.6, newdata = plotData)

# Back-transform any log terms
plotData$education = exp(plotData$Leducation)
plotData$earn = exp(plotData$yhat)

# Turn female into a factor
plotData$female = factor(plotData$female, levels = c(0, 1), labels = c("Male", "Female"))

# Examine data
head(plotData)
```

```
##   Leducation female     yhat education     earn
## 1        1.0   Male 7.714992  2.718282 2241.704
## 2        1.1   Male 7.860040  3.004166 2591.625
## 3        1.2   Male 8.005089  3.320117 2996.166
## 4        1.3   Male 8.150137  3.669297 3463.854
## 5        1.4   Male 8.295186  4.055200 4004.547
## 6        1.5   Male 8.440234  4.481689 4629.639
```

```r
# Plot
ggplot(data = plotData, aes(x = education, y = earn, group = female, color = female)) +
    geom_line() +
    theme_bw() +
  xlab("Education") +
  ylab("Earnings") +
  scale_color_brewer(name = "", palette = "Set1")
```



The plot helps us see (1) the exponential relationship between education and earning for both males and females, and (2) the growing discrepancy between male and female earnings at higher levels of education. Females earn roughly 59% of what males do, but in raw numbers 59% of a small value (e.g., a male earning for a low education value) is less than for a large value. Even though we fitted a main-effects model, the lines after we back-transform are not parallel. How non-parallel the lines are depends on the size of the coefficeint associated with `female` (in this example). This is why, especially with transformed data, it is essential to plot the model to make sure you are understanding the interpretations from your coefficients.