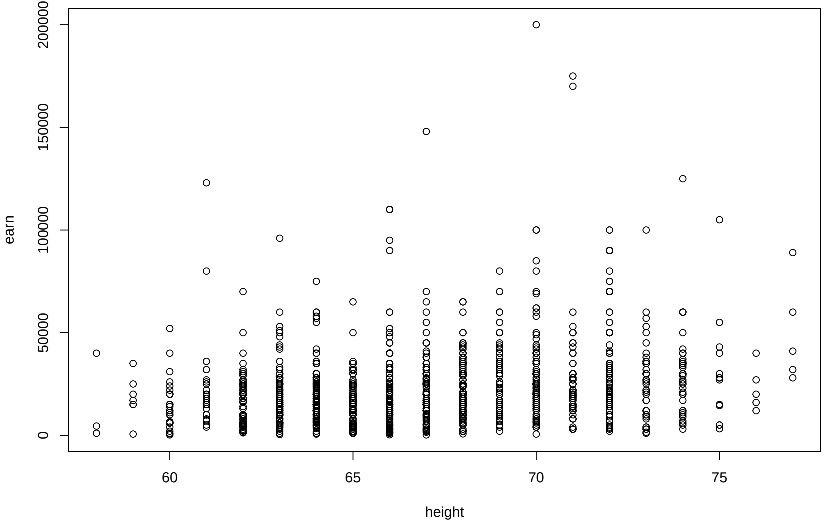
Transformations

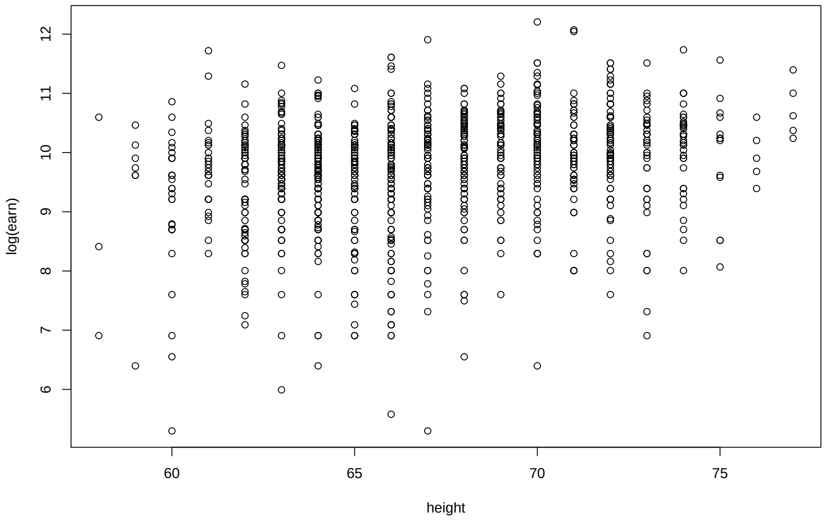
BIOE 498/598 PJ

Spring 2021

Earnings vs. height?



log(Earnings) vs height?



Logarithmic transformation of the response

```r
lm(log(earn) ~ height, earnings) %>% coefs()
```

```
## (Intercept)      height
##        5.78        0.06
```

Logarithmic transformation of the response

```r
lm(log(earn) ~ height, earnings) %>% coefs()
```

```
## (Intercept)      height
##        5.78        0.06
```

This changes from an additive model to a multiplicative one:

$$\log(y) = \beta_0 + \beta_1 x_1 + \cdots + \epsilon$$

$$\downarrow$$

$$y = B_0 \cdot B_1 e^{x_1} \cdots E$$

What is the best response transformation?

A general family of nonlinear transformations are described by the Box-Cox transformation:

$$T(y) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

What is the best response transformation?

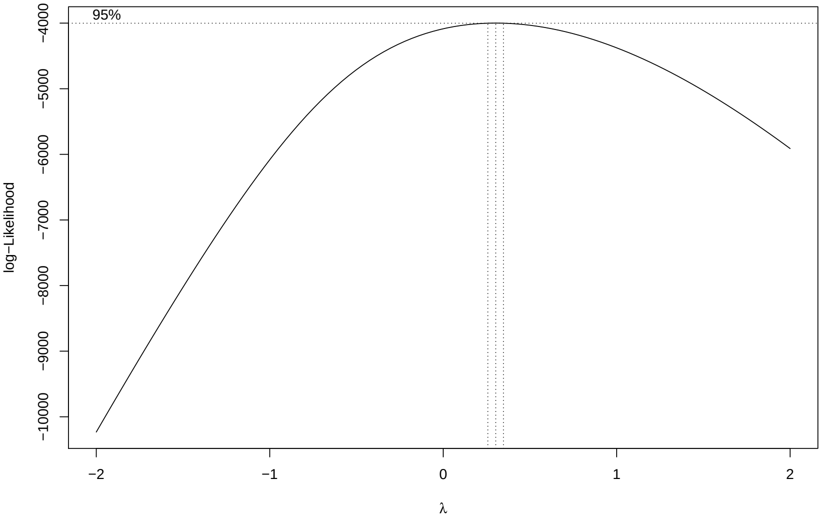
A general family of nonlinear transformations are described by the Box-Cox transformation:

$$T(y) = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

- $\lambda = 2$ suggests $y \to y^2$
- $\lambda = 1$ suggests no transformation
- $\lambda = 1/2$ suggests $y \to \sqrt{y}$
- $\lambda = -1$ suggests $y \to 1/y$

Where do we get λ?

```
MASS::boxcox(earn ~ height, data=earnings)
```



Shouldn't we always transform the response?

- Transforming the response will often improve the predictive power of the model.
- Models with transformed responses are more difficult to interpret.

Shouldn't we always transform the response?

- Transforming the response will often improve the predictive power of the model.
- Models with transformed responses are more difficult to interpret.
- In general, **there is always a tradeoff between prediction and interpretation**.
- Recommendation: perform the Box-Cox analysis, but only transform if
  - you only want the model for prediction, *or*
  - the Box-Cox suggests a common transformation (log, square root, inverse, square, etc.).

Linear Transformations: Scaling Predictor Variables

$$\log(\text{earnings}[\$]) = 5.78 + 0.06 \cdot \text{height}[\text{in}] + \text{error}$$
$$\log(\text{earnings}[\$]) = 5.78 + 0.72 \cdot \text{height}[\text{ft}] + \text{error}$$
$$\log(\text{earnings}[\$]) = 5.78 + 0.00234 \cdot \text{height}[\text{mm}] + \text{error}$$

Linear Transformations: Scaling Predictor Variables

$$\log(\text{earnings}[\$]) = 5.78 + 0.06 \cdot \text{height}[\text{in}] + \text{error}$$

$$\log(\text{earnings}[\$]) = 5.78 + 0.72 \cdot \text{height}[\text{ft}] + \text{error}$$

$$\log(\text{earnings}[\$]) = 5.78 + 0.00234 \cdot \text{height}[\text{mm}] + \text{error}$$

Scaling in General:

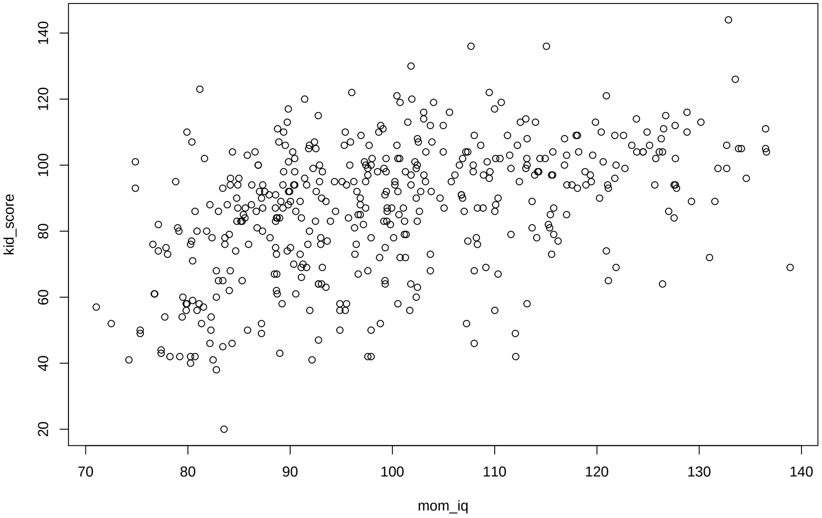
$$y = \beta_0 + \beta_1(kx) + \epsilon$$

$$\updownarrow$$

$$y = \beta_0 + (k\beta_1)x + \epsilon$$

Are childhood exams influenced by maternal IQ?

```r
child <- foreign::read.dta("kidiq.dta")
with(child, plot(mom_iq, kid_score))
```



Are childhood exams influenced by maternal IQ?

```r
model <- lm(kid_score ~ mom_iq, data=child)
summary(model)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq, data = child)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -56.75 -12.07    2.22   11.71   47.69
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.7998     5.9174    4.36  1.6e-05 ***
## mom_iq        0.6100     0.0585   10.42  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 18 on 432 degrees of freedom
## Multiple R-squared:  0.201,  Adjusted R-squared:  0.199
## F-statistic:  109 on 1 and 432 DF,  p-value: <2e-16
```

Mean Centering

Our linear model is

$$\text{kid\_score} = \beta_0 + \beta_1 \text{mom\_iq}$$

The effect size β_1 is the change in child score for every unit increase in mother IQ. **The intercept β_0 is uninterpretable** (the score of a child with a mother of IQ=0).

Mean Centering

Our linear model is

$$\text{kid\_score} = \beta_0 + \beta_1 \text{mom\_iq}$$

The effect size β_1 is the change in child score for every unit increase in mother IQ. **The intercept β_0 is uninterpretable** (the score of a child with a mother of IQ=0).

Let's *mean center* the variable mom_iq:

$$\text{c\_mom\_iq} = \text{mom\_iq} - \text{mean}[\text{mom\_iq}]$$

Mean Centering (continued)

Our new model is

$$\text{kid\_score} = \beta_0 + \beta_1 \text{c\_mom\_iq}$$

Now both coefficients are interpretable:

- β_1 remains the increase in child score given a unit increase in mother's IQ.
- β_0 is the predicted child score for a child with mother of average IQ.

Standardization by Z-score

We can also center mom_iq and rescale it by the standard deviation:

$$\text{z\_mom\_iq} = \frac{\text{mom\_iq} - \text{mean}[\text{mom\_iq}]}{\text{stdev}[\text{mom\_iq}]}$$

Standardization by Z-score

We can also center mom_iq and rescale it by the standard deviation:

$$\text{z\_mom\_iq} = \frac{\text{mom\_iq} - \text{mean}[\text{mom\_iq}]}{\text{stdev}[\text{mom\_iq}]}$$

In the model

$$\text{kid\_score} = \beta_0 + \beta_1 \text{z\_mom\_iq}$$

the interpretation of β_0 is the same (predicted score for child with average mom_iq), but β_1 is the change in child score based on an increase of **one standard devation** in mother's IQ.

Why rescale by the standard deviation?

Let's add a second factor: if the mother completed high school.

No scaling:

```r
lm(kid_score ~ mom_hs + c_mom_iq, child) %>% coefs()
```

```
## (Intercept)      mom_hs    c_mom_iq
##       82.12        5.95        0.56
```

Why rescale by the standard deviation?

Let's add a second factor: if the mother completed high school.

No scaling:

```
lm(kid_score ~ mom_hs + c_mom_iq, child) %>% coefs()
```

```
## (Intercept)      mom_hs    c_mom_iq
##       82.12        5.95        0.56
```

Z-scoring (1 stdev):

```
lm(kid_score ~ z_mom_hs + z_mom_iq, child) %>% coefs()
```

```
## (Intercept)    z_mom_hs    z_mom_iq
##        86.8         2.4         8.5
```

What is the best scaling factor?

```r
lm(kid_score ~ mom_hs + c_mom_iq, child) %>% coefs()
```

```
## (Intercept)      mom_hs    c_mom_iq
##       82.12        5.95        0.56
```

Z-scoring (1 stdev):

```r
lm(kid_score ~ z_mom_hs + z_mom_iq, child) %>% coefs()
```

```
## (Intercept)    z_mom_hs    z_mom_iq
##        86.8         2.4         8.5
```

What is the best scaling factor?

```r
lm(kid_score ~ mom_hs + c_mom_iq, child) %>% coefs()
```

```
## (Intercept)      mom_hs    c_mom_iq
##       82.12        5.95        0.56
```

Z-scoring (1 stdev):

```r
lm(kid_score ~ z_mom_hs + z_mom_iq, child) %>% coefs()
```

```
## (Intercept)    z_mom_hs    z_mom_iq
##        86.8         2.4         8.5
```

Scaling by 2 stdev:

```r
lm(kid_score ~ z2_mom_hs + z2_mom_iq, child) %>% coefs()
```

```
## (Intercept)   z2_mom_hs   z2_mom_iq
##        86.8         4.9        16.9
```

Why two standard deviations?

Assume a binary factor takes the value 1 with probability p. Then the standard deviation of this factor is

$$\sqrt{p(1-p)}$$

Why two standard deviations?

Assume a binary factor takes the value 1 with probability p. Then the standard deviation of this factor is

$$\sqrt{p(1-p)}$$

Without any additional knowledge, we assume that $p = 0.5$. Then the standard deviation of the factor is $\sqrt{0.5^2} = 0.5$.

Why two standard deviations?

Assume a binary factor takes the value 1 with probability p. Then the standard deviation of this factor is

$$\sqrt{p(1-p)}$$

Without any additional knowledge, we assume that $p = 0.5$. Then the standard deviation of the factor is $\sqrt{0.5^2} = 0.5$.

When this binary factor switches from 0 to 1, it is moving two standard deviations. To keep things on the same scale, the continuous variables should be rescaled so a unit change also corresponds to two standard deviations.

Our recommendations

- Leave binary factors unscaled.
- Mean center and scale continuous factors by 1 stdev. *unless* the intercept is more interpretable when all factors are zero (e.g. drug dosing).
- Alternatively, use *coded factors* for continuous variables that are set at only at a few discrete values.
- If your model contains both binary and continuous factors that must be compared:
  - Center and scale continuous factors by 2 stdev, *except* if the variable uses a scale widely accepted by your audience.
  - You can always build a rescaled model behind-the-scenes and present the conclusions with unscaled factors.