Inference on transformed variables

INFERENCE FOR LINEAR REGRESSION IN R

Jo Hardin
Professor, Pomona College

Interpreting coefficients - linear

$Y = \beta_0 + \beta_1 \cdot X + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon)$

$E[Y_X] = \beta_0 + \beta_1 \cdot X$

$E[Y_{X+1}] = \beta_0 + \beta_1 \cdot (X + 1)$

$\beta_1 = E[Y_{X+1}] - E[Y_X]$

Interpreting coefficients - nonlinear X

$Y = \beta_0 + \beta_1 \cdot \ln(X) + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon)$

$$E[Y_{\ln(X)}] = \beta_0 + \beta_1 \cdot \ln(X)$$

$$E[Y_{\ln(X)+1}] = \beta_0 + \beta_1 \cdot (\ln(X) + 1)$$

$$\beta_1 = E[Y_{\ln(X)+1}] - E[Y_{\ln(X)}]$$

Interpreting coefficients - nonlinear Y

$\ln(Y) = \beta_0 + \beta_1 \cdot X + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon)$

$$E[\ln(Y)_X] = \beta_0 + \beta_1 \cdot X$$

$$E[\ln(Y)_{X+1}] = \beta_0 + \beta_1 \cdot (X + 1)$$

$$\beta_1 = E[\ln(Y)_{X+1}] - E[\ln(Y)_X]$$

Interpreting coefficients - both nonlinear

$\ln(Y) = \beta_0 + \beta_1 \cdot \ln(X) + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon)$

$$E[\ln(Y)_{\ln(X)}] = \beta_0 + \beta_1 \cdot \ln(X)$$

$$E[\ln(Y)_{\ln(X)+1}] = \beta_0 + \beta_1 \cdot (\ln(X) + 1)$$

$$\beta_1 = E[\ln(Y)_{\ln(X)+1}] - E[\ln(Y)_{\ln X}]$$

Interpreting coefficients - both natural log (special case)

$\ln(Y) = \beta_0 + \beta_1 \cdot \ln(X) + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon)$

$E[\ln(Y)_{\ln(X)}] = \beta_0 + \beta_1 \cdot \ln(X)$

$E[\ln(Y)_{\ln(X)+1}] = \beta_0 + \beta_1 \cdot (\ln(X) + 1)$

$\beta_1 = E[\ln(Y)_{\ln(X)+1}] - E[\ln(Y)_{\ln X}]$

OR (when X and Y are both transformed using natural log):

$\beta_1 =$ percent change in Y for each 1% change in X

Let's practice!

INFERENCE FOR LINEAR REGRESSION IN R

Multicollinearity

INFERENCE FOR LINEAR REGRESSION IN R

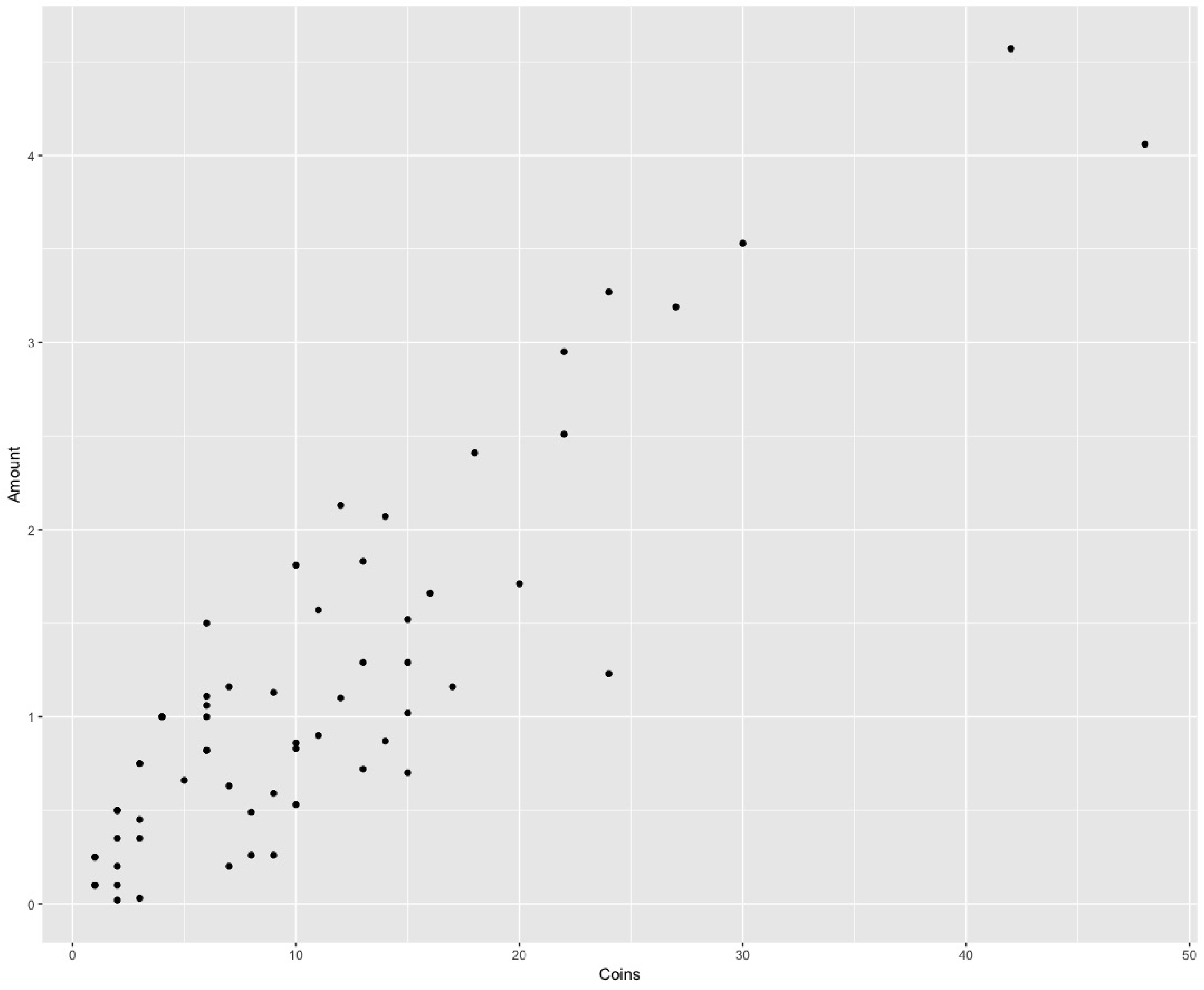
Jo Hardin
Professor, Pomona College

Regressing dollar amount on coins

```
head(change)
```

```
 A tibble: 6 x 7
  Coins  Qrts Dimes Nickels Pennies Small Amount
  <int> <int> <int>   <int>   <int> <int>  <dbl>
1     2     1     1       0       0     1   0.35
2     3     3     0       0       0     0   0.75
3     2     0     0       2       0     2   0.10
4     4     4     0       0       0     0   1.00
5     2     2     0       0       0     0   0.50
6    13     3     4       2       4    10   1.29
```

Amount vs. coins - plot

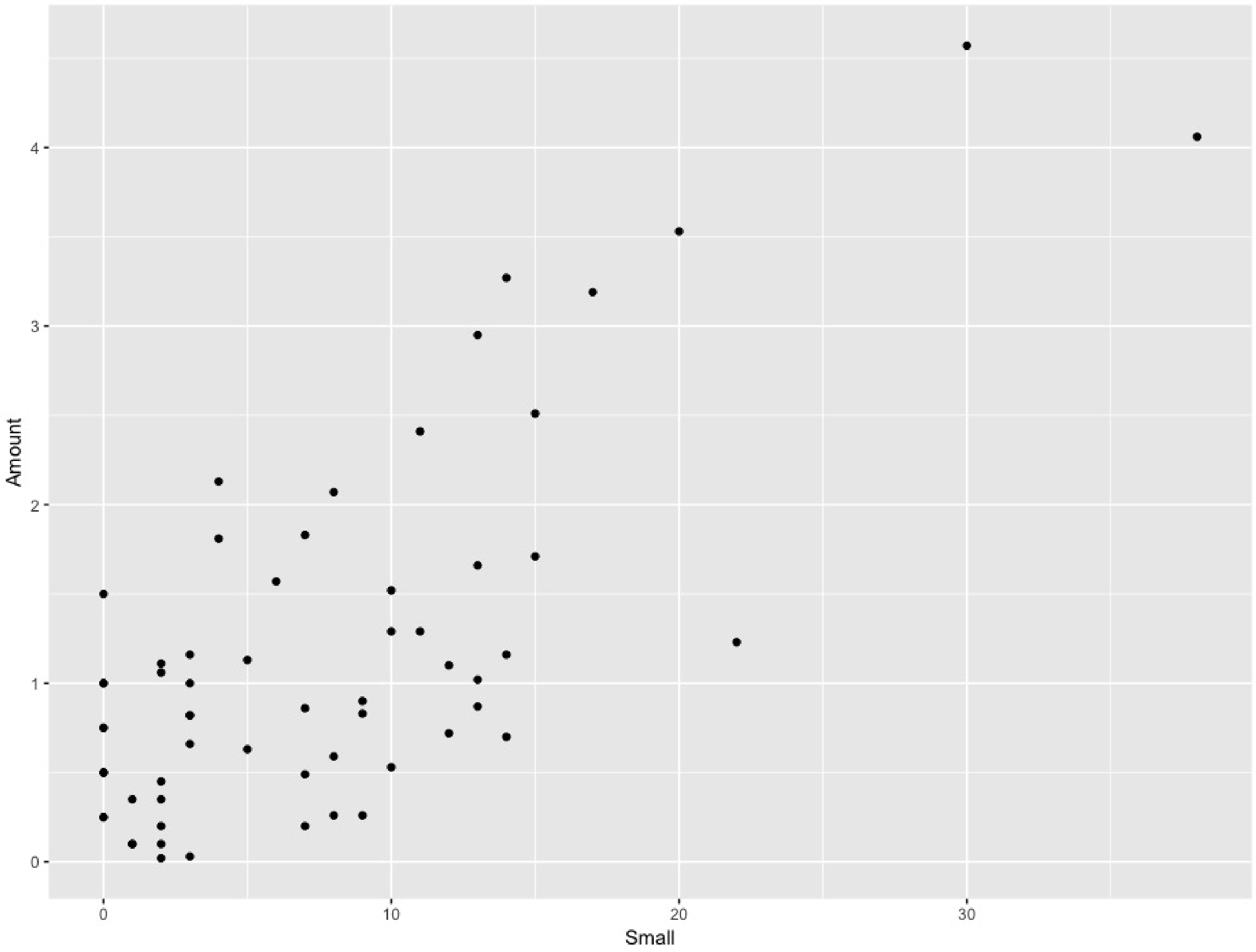


Amount vs. coins - linear model

```
lm(Amount ~ Coins, data = change) %>% tidy()
```

```
           term estimate std.error statistic  p.value
1 (Intercept)   0.1449    0.0902      1.61 1.13e-01
2       Coins   0.0945    0.0063     14.99 6.01e-22
```

Amount vs. small coins - plot



Amount vs. small coins - linear model

```
lm(Amount ~ Small, data = change) %>% tidy()
```

```
         term estimate std.error statistic  p.value
1 (Intercept)   0.4225    0.1244      3.40 1.22e-03
2       Small   0.0989    0.0118      8.38 1.10e-11
```

Amount vs. coins and small coins

$$\hat{\text{Amount}} = -0.00554 + 0.25862 \cdot \text{Coins} - 0.21611 \cdot \text{Small Coins}$$

```
lm(Amount ~ Coins + Small, data = change) %>% tidy()
```

```
         term estimate std.error statistic  p.value
1 (Intercept) -0.00554   0.02735    -0.202 8.40e-01
2       Coins  0.25862   0.00682    37.917 3.95e-43
3       Small -0.21611   0.00864   -25.021 4.17e-33
```

Let's practice!

INFERENCE FOR LINEAR REGRESSION IN R

Multiple linear regression

INFERENCE FOR LINEAR REGRESSION IN R

Jo Hardin
Professor, Pomona College

Bathrooms negative coefficient

```
lm(log(price) ~  log(bath), data=LAhomes) %>% tidy()
```

```
         term estimate std.error statistic   p.value
1 (Intercept)    12.23    0.0280     437.2  0.00e+00
2   log(bath)     1.43    0.0306      46.6 9.66e-300
```

```
lm(log(price) ~ log(sqft) + log(bath), data=LAhomes) %>% tidy()
```

```
         term estimate std.error statistic   p.value
1 (Intercept)    2.514    0.2619     9.601  2.96e-21
2   log(sqft)    1.471    0.0395    37.221 1.19e-218
3   log(bath)   -0.039    0.0453    -0.862  3.89e-01
```

Bathrooms non-significant coefficient

```
lm(log(price) ~  log(bath), data=LAhomes) %>% tidy()
```

```
         term estimate std.error statistic   p.value
1 (Intercept)    12.23    0.0280     437.2  0.00e+00
2   log(bath)     1.43    0.0306      46.6 9.66e-300
```

```
lm(log(price) ~ log(sqft) + log(bath), data=LAhomes) %>% tidy()
```

```
         term estimate std.error statistic   p.value
1 (Intercept)    2.514    0.2619     9.601  2.96e-21
2   log(sqft)    1.471    0.0395    37.221 1.19e-218
3   log(bath)   -0.039    0.0453    -0.862  3.89e-01
```

Price on bed and bath

```
lm(log(price) ~ log(bath) + bed, data=LAhomes) %>% tidy()
```

```
         term estimate std.error statistic   p.value
1 (Intercept)   11.965    0.0384    311.67  0.00e+00
2   log(bath)    1.076    0.0465     23.14 2.38e-102
3         bed    0.189    0.0193      9.82  4.01e-22
```

Large model on price

```
lm(log(price) ~ log(sqft) + log(bath) + bed, data=LAhomes) %>% tidy()
```

```
         term estimate std.error statistic   p.value
1 (Intercept)   1.5364    0.2894     5.310  1.25e-07
2   log(sqft)   1.6456    0.0454    36.215 6.27e-210
3   log(bath)   0.0165    0.0452     0.365  7.15e-01
4         bed  -0.1236    0.0167    -7.411  2.03e-13
```

Let's practice!

INFERENCE FOR LINEAR REGRESSION IN R

Summary

INFERENCE FOR LINEAR REGRESSION IN R

Jo Hardin
Professor, Pomona College

Linear regression as model

- It estimates an underlying population model
- It might be linear or might need variable transformations
- All of LINE conditions should be checked
- Other variable relationships should be carefully considered

Linear regression as an inferential technique

- Hypothesis testing using a mathematical model (t-tests)
- Hypothesis testing using randomization tests
- Confidence intervals using a mathematical model
- Confidence intervals using bootstrapping

Let's practice!

INFERENCE FOR LINEAR REGRESSION IN R