Predictions in R

Jan Zouhar

Friday, April 15, 2016

Contents

Birth weight example from lecture presentations. 1

Predicting y when $\log(y)$ is the dependent variable. 2

Birth weight example from lecture presentations.

I'm trying to use the same notation as in the presentation, so it's best if you read it in parallel.

```r
# Import data on birth weights, and a file with variable labels.
bw <- read.csv("http://nb.vse.cz/~zouharj/econ/bwght.csv")
bw.desc <- read.csv("http://nb.vse.cz/~zouharj/econ/bwght_desc.csv")

# In the lecture presentation, birth weight is converted from ounces to grams.
bw$bweight <- 28.3495 * bw$bwght  # 1 ounce = 28.3495 grams.

# Verify that we have the same data as in the lecture.
lm(bweight ~ cigs, data=bw)
```

```
##
## Call:
## lm(formula = bweight ~ cigs, data = bw)
##
## Coefficients:
## (Intercept)         cigs
##     3395.47       -14.57
```

```r
# Predict theta = E[bweight | cigs = 10].
model1 <- lm(bweight ~ I(cigs - 10), data=bw)
summary(model1)
```

```
##
## Call:
## lm(formula = bweight ~ I(cigs - 10), data = bw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2743.4  -333.7     8.4   375.0  4287.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3249.822     25.430 127.797  < 2e-16 ***
## I(cigs - 10)  -14.565      2.565  -5.678 1.66e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570.6 on 1386 degrees of freedom
## Multiple R-squared:  0.02273,    Adjusted R-squared:  0.02202
## F-statistic: 32.24 on 1 and 1386 DF,  p-value: 1.662e-08
```

```
theta.hat <- model1$coef[1]  # Save theta hat.
cat(paste("Point prediction for theta is", theta.hat))  # Print theta hat.
```

```
## Point prediction for theta is 3249.82167088805
```

```
# Find the 95% CI for theta.
confint(model1, parm="(Intercept)")
```

```
##                2.5 %   97.5 %
## (Intercept) 3199.937 3299.706
```

```
# Find the prediction error.
se.theta.hat <- summary(model1)$coef[1, 2]  # Obtain std. error of theta hat.
sigma.hat <- summary(model1)$sigma  # Obtain sigma hat, estimate of sd(u).
prediction.error <- sqrt(se.theta.hat^2 + sigma.hat^2)  # Calculate pred. error.

# Find the 95% prediction interval.
c <- qt(0.975, df=model1$df)  # We can use c = 2 or c = 1.96, too.
lower <- theta.hat - c * prediction.error  # Lower bound of 95% pred. interval.
upper <- theta.hat + c * prediction.error  # Upper bound of 95% pred. interval.
cat(paste("95% prediction interval is (", lower, ", ", upper, ")", sep=""))
```

```
## 95% prediction interval is (2129.30885765229, 4370.33448412381)
```

Predicting y when $\log(y)$ is the dependent variable.

Here I'm using data on CEO salaries (same dataset as in the tutorials). Again, I suggest you open up the presentation for reference. The goal here is to predict salary for a CEO of a firm with:

- $sales = 10000$,
- $roe = 15$,

based on the equation

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + u.$$

```
# Import data
ceo <- read.csv("http://nb.vse.cz/~zouharj/econ/ceosal1.csv")

# Baseline model.
model2 <- lm(log(salary) ~ log(sales) + roe, data=ceo)
summary(model2)
```

```
##
## Call:
## lm(formula = log(salary) ~ log(sales) + roe, data = ceo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9464 -0.2888 -0.0322  0.2261  2.7830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.362168   0.293878  14.843  < 2e-16 ***
## log(sales)  0.275087   0.033254   8.272 1.62e-14 ***
## roe         0.017872   0.003955   4.519 1.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4822 on 206 degrees of freedom
## Multiple R-squared:  0.282,  Adjusted R-squared:  0.275
## F-statistic: 40.45 on 2 and 206 DF,  p-value: 1.519e-15
```

```
# Goal: predict salary for a CEO of a firm with:
#   sales = 10000,
#   roe = 15.

# Step 1: estimate A (see slide 6, presentation 8).
model3 <- lm(log(salary) ~ log(sales/10000) + I(roe - 15), data=ceo)
summary(model3)  # Same slopes as in model 2, different intercept.
```

```
##
## Call:
## lm(formula = log(salary) ~ log(sales/10000) + I(roe - 15), data = ceo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9464 -0.2888 -0.0322  0.2261  2.7830
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.163900   0.045330 158.040  < 2e-16 ***
## log(sales/10000) 0.275087   0.033254   8.272 1.62e-14 ***
## I(roe - 15)      0.017872   0.003955   4.519 1.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4822 on 206 degrees of freedom
## Multiple R-squared:  0.282,  Adjusted R-squared:  0.275
## F-statistic: 40.45 on 2 and 206 DF,  p-value: 1.519e-15
```

```
intercept <- model3$coef[1]  # Intercept in model 3 describes our CEO.
A <- exp(intercept)  # There we have it. :)

# Step 2: estimate B (see slide 6, presentation 8).
# - version 1: u is assumed to be normally distributed.
```

```
sigma.hat <- summary(model3)$sigma
B1 <- exp(sigma.hat^2/2)
print(B1)  # Let's see value of B in version 1.
```

```
## [1] 1.12331
```

```
# - version 2: Duan's (1983) estimator.
u.hat <- model3$residuals
B2 <- mean(exp(u.hat))
print(B2)  # Let's see value of B in version 2.
```

```
## [1] 1.184736
```

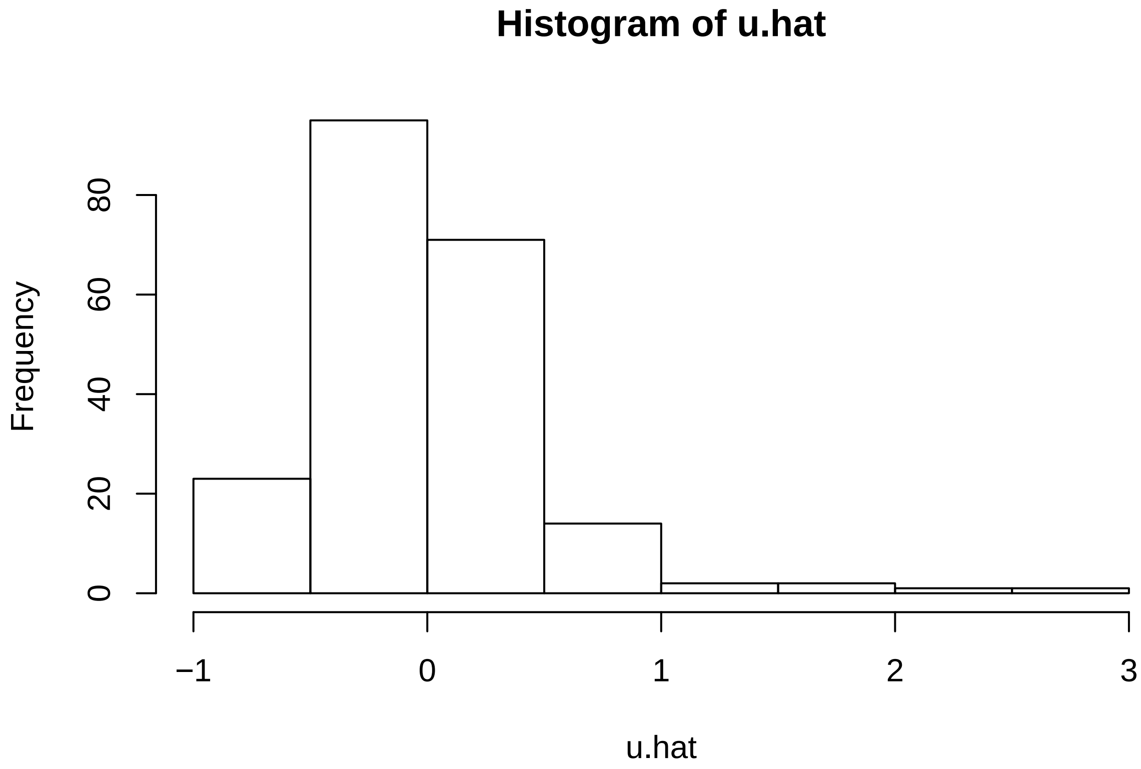
```
# Step 3: Predicted salary is A * B.
cat(paste("Version 1: predicted salary =", A * B1))
```

```
## Version 1: predicted salary = 1451.24944948577
```

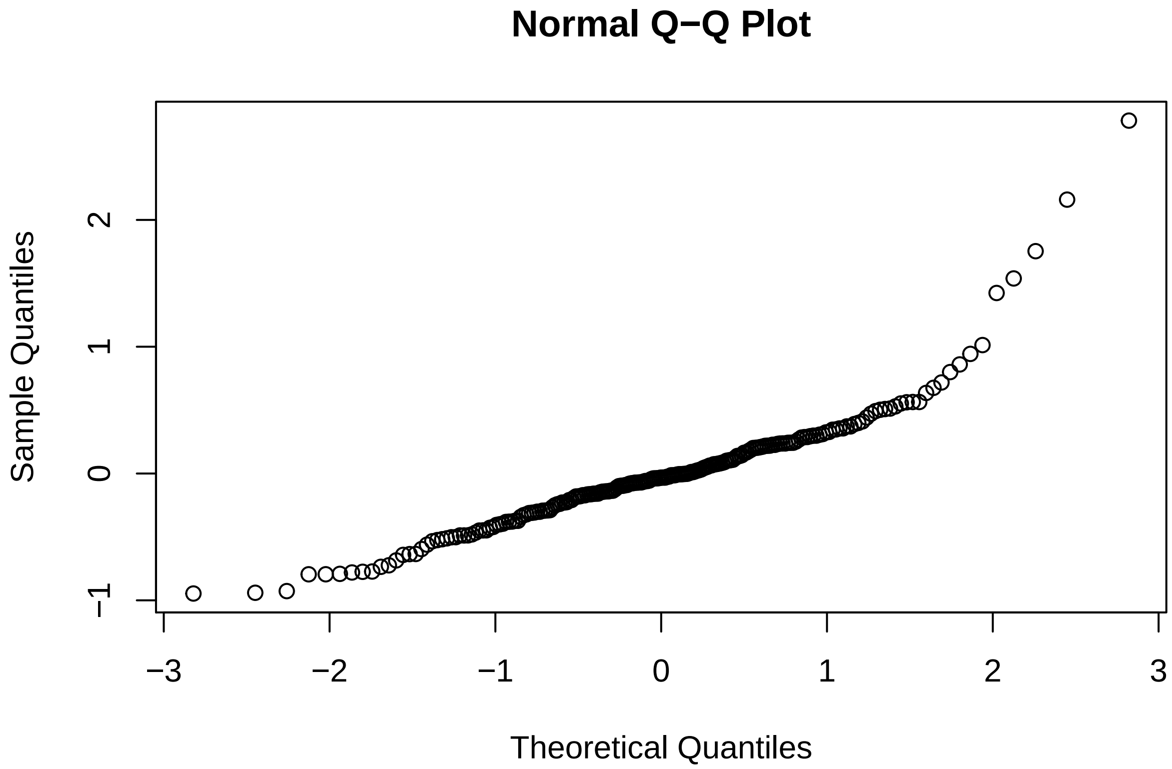
```
cat(paste("Version 2: predicted salary =", A * B2))
```

```
## Version 2: predicted salary = 1530.60866711281
```

```
# Which version is preferred? If your residuals are not normally distributed,
# version 2 is more justified.
hist(u.hat)  # Does not look normal, eh?
```



```r
qqnorm(u.hat)  # Comparison with normal quantiles; far from 45° line.
```



```r
library(stats)
shapiro.test(u.hat)  # H0: u.hat normally distributed.
```

```
##
##  Shapiro-Wilk normality test
##
## data:  u.hat
## W = 0.8848, p-value = 1.497e-11
```