Scottish hill races 2008 — knitr
STT 3820

Alan T. Arnholt
April 2, 2012

1 Problem

Hill running - races up and down hills - has a written history in Scotland dating back to the year 1040. Races are held throughout the year at different locations around Scotland. A recent compilation of information for 91 races (for which full information was available and omitting two unusual races) includes the Distance (km), the Climb (m), and the Record Time (minutes). A regression to predict the men’s records as of 2008 is given in Table 1. A regression to predict the women’s records as of 2008 is given in Table 2.

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|---------|
| (Intercept)    | -10.3723 | 1.2445     | -8.33   | 0.0000  |
| ClimbM         | 0.0342   | 0.0022     | 15.74   | 0.0000  |
| DistanceKm     | 4.0420   | 0.1448     | 27.92   | 0.0000  |

Table 1: Summary results for \textit{modMen}

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|---------|
| (Intercept)    | -11.6545 | 1.8913     | -6.16   | 0.0000  |
| ClimbM         | 0.0452   | 0.0033     | 13.68   | 0.0000  |
| DistanceKm     | 4.4343   | 0.2200     | 20.16   | 0.0000  |

Table 2: Summary results for \textit{modWomen}

(a) Write the regression equation. Give a brief report on what it says about men’s record times in hill races.

(a) The regression equation is:

\[ \hat{\text{MenTime}} = -10.3723 + 0.0342 \times \text{ClimbM} + 4.042 \times \text{DistanceM} \]

The time of a race increases as either the distance or the climb increases.

(b) Interpret the value of $R^2$ in this regression.

(b) 98.023% of the variability in the men’s record times is accounted for by regressing \text{MenTime} on \text{ClimbM} and \text{DistanceKm}.

(c) What does the coefficient of \text{Climb} mean in this regression?

(c) For races of a given distance, we expect the mean \text{MenTime} time to increase by 0.0342 minutes for each additional meter of \text{ClimbM}. 
Compare the regression model for the women’s records with that found for the men’s records.

The two models are similar. It appears that both additional **ClimbM** and **DistanceM** lead to larger increases in average **Time**.

Discuss the residuals and what they say about the assumptions and conditions for this regression.

The residual plots shown in Figure 1 show curvature as well as a tendency to increase in size suggesting lack of fit. After adding two quadratic terms and taking the square root of the response variable, the residuals shown in Figure 2 are much better behaved.

### 2 Mechanics

```r
> site <- "http://www1.appstate.edu/~arnholta/classes/STT3820/Data/CSV/Ch30_Scottish_hill_races_2008CLEAN.csv"
> HR <- read.csv(file = url(site))
> names(HR)
[1] "Race"    "MenTime"  "WomenTime" "DistanceKm" "ClimbM"
> modMen <- lm(MenTime ~ ClimbM + DistanceKm, data = HR)
> summary(modMen)

Call:
lm(formula = MenTime ~ ClimbM + DistanceKm, data = HR)

Residuals:
       Min        1Q      Median        3Q       Max
-16.906   -4.388       0.339       3.177      22.109

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)     -10.37227    1.24451   -8.33  1e-12 ***
ClimbM            0.03423    0.00217    15.74 <2e-16 ***
DistanceKm       4.04204    0.14476    27.92 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.62 on 87 degrees of freedom
(4 observations deleted due to missingness)
Multiple R-squared: 0.98, Adjusted R-squared: 0.98
F-statistic: 2.16e+03 on 2 and 87 DF,  p-value: <2e-16

> modWomen <- lm(WomenTime ~ ClimbM + DistanceKm, data = HR)
> summary(modWomen)

Call:
lm(formula = WomenTime ~ ClimbM + DistanceKm, data = HR)

Residuals:
       Min        1Q      Median        3Q       Max
-12.471    -3.626      -0.257       1.800      21.954

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)     -10.87313    1.28223   -8.50  1.4e-12 ***
ClimbM            0.03423    0.00217    15.74 <2e-16 ***
DistanceKm       4.10481    0.14476    28.58 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.73 on 87 degrees of freedom
(1 observations deleted due to missingness)
Multiple R-squared: 0.98, Adjusted R-squared: 0.98
F-statistic: 2.31e+03 on 2 and 87 DF,  p-value: <2e-16
```
Residuals:
   Min  1Q  Median   3Q  Max
-22.14 -7.24   0.76   3.82  41.76

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -11.6545    1.8913  -6.16 2.2e-08 ***
ClimbM       0.0452     0.0033  13.68 < 2e-16 ***
DistanceKm   4.4343     0.2200  20.16 < 2e-16 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 .’ 0.1 ‘ ’ 1

Residual standard error: 10.1 on 87 degrees of freedom
(4 observations deleted due to missingness)
Multiple R-squared: 0.967, Adjusted R-squared: 0.967
F-statistic: 1.29e+03 on 2 and 87 DF,  p-value: <2e-16

> residualPlots(modWomen)

    Test stat Pr(>|t|)
ClimbM        4.082     0
DistanceKm    3.629     0
Tukey test    5.083     0

> # Fix
> modWomen2 <- lm(sqrt(WomenTime) ~ ClimbM + I(ClimbM^2) +
+ DistanceKm + I(DistanceKm^2), data = HR)
> summary(modWomen2)

Call:
  lm(formula = sqrt(WomenTime) ~ ClimbM + I(ClimbM^2) + DistanceKm +
      I(DistanceKm^2), data = HR)

Residuals:
   Min   1Q Median   3Q   Max
-1.0271 -0.3184 -0.0254  0.2577  1.1503

Coefficients:
                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)         3.02e+00   1.66e-01  18.25 < 2e-16 ***
ClimbM              2.51e-03   4.13e-04   6.07  3.5e-08 ***
I(ClimbM^2)         -1.73e-07   1.65e-07  -1.05   0.295
DistanceKm         3.15e-01   2.75e-02  11.46 < 2e-16 ***
I(DistanceKm^2)    -1.98e-03   6.48e-04  -3.06    0.003 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.457 on 85 degrees of freedom
(4 observations deleted due to missingness)
Multiple R-squared: 0.976, Adjusted R-squared: 0.975
F-statistic: 875 on 4 and 85 DF, p-value: <2e-16

> residualPlots(modWomen2)

|                  | Test stat | Pr(|t|) |
|------------------|-----------|--------|
| ClimbM           | -2.085    | 0.040  |
| I(ClimbM^2)     | 1.688     | 0.095  |
| DistanceKm       | -0.039    | 0.969  |
| I(DistanceKm^2)  | 2.037     | 0.045  |
| Tukey test       | -1.700    | 0.089  |
Figure 1: Residual plots for modWomen
Figure 2: Residual plots for modWomen2