ECON 4101 Econometrics
CM07 Homework

Pranav Singh

February 4, 2017

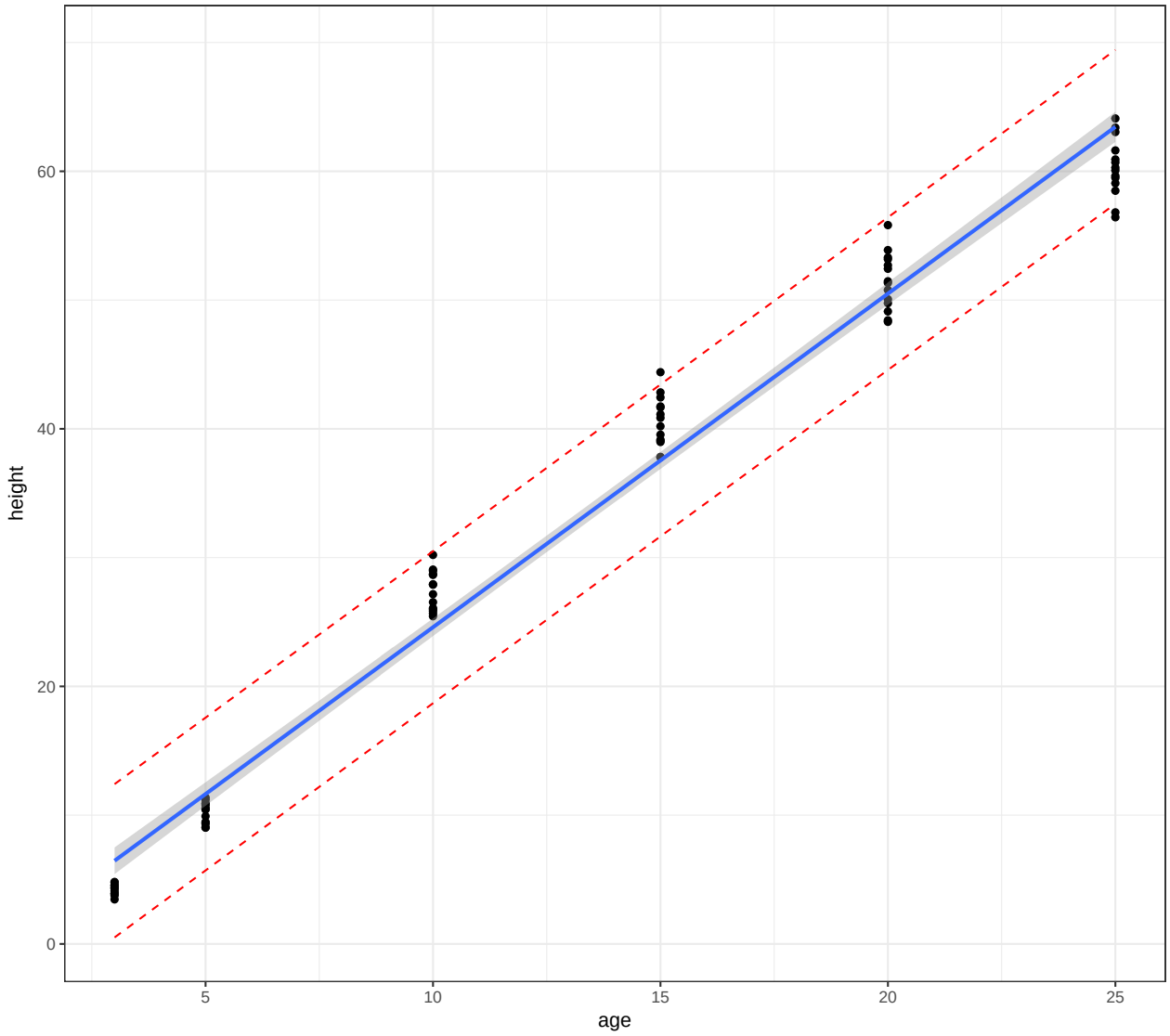
Problem 1 - Summary Statistics

```
df <- Loblolly
summary(cbind(df$height, df$age))
```

```
##        V1              V2
##  Min.   : 3.46   Min.   : 3.0
##  1st Qu.:10.47   1st Qu.: 5.0
##  Median :34.00   Median :12.5
##  Mean   :32.36   Mean   :13.0
##  3rd Qu.:51.36   3rd Qu.:20.0
##  Max.   :64.10   Max.   :25.0
```

Problem 2 - Plots

```
lm.fit <- lm(formula = height ~ age, data = df)
preds <- predict.lm(object = lm.fit, newdata = NULL, interval = 'prediction')
df.preds <- cbind(df, preds)
ggplot(df.preds, aes(x = age, y = height)) +
  geom_point() +
  geom_smooth(method = 'lm', se = T, level = 0.95) +
  geom_line(aes(y=lwr), color = 'red', linetype = 'dashed') +
  geom_line(aes(y=upr), color = 'red', linetype = 'dashed')
```



Problem 3 - Regression Analysis

```
summary(lm.fit)
```

```
##
## Call:
## lm(formula = height ~ age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0207 -2.1672 -0.4391  2.0539  6.8545
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -1.31240    0.62183  -2.111              0.0379 *
## age          2.59052    0.04094  63.272 <0.0000000000000002 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.947 on 82 degrees of freedom
## Multiple R-squared:  0.9799, Adjusted R-squared:  0.9797
## F-statistic:  4003 on 1 and 82 DF,  p-value: < 0.00000000000000022
```

From the summary of the fitted linear model, we see that the estimated linear regression equation is

$$\widehat{height} = -1.31240 + 2.59052 \cdot age$$

The 95% confidence intervals of the estimated regression coefficients are:

```
lm.coefs <- as.data.frame(summary(lm.fit)$coefficients)
t.crit <- qt(0.975, 82)
b0 <- lm.coefs[1, "Estimate"]
b1 <- lm.coefs[2, "Estimate"]
b0.se <- lm.coefs[1, "Std. Error"]
b1.se <- lm.coefs[2, "Std. Error"]
cat("95% CI for Intercept: ", b0 + c(-1, 1) * t.crit * b0.se, sep = "\n")
```

```
## 95% CI for Intercept:
## -2.54941
## -0.07538279
```

```
cat("95% CI for Age: ", b1 + c(-1, 1) * t.crit * b1.se, sep = "\n")
```

```
## 95% CI for Age:
## 2.509075
## 2.671971
```

```
# Equivalently, here's a one-liner
confint(lm.fit)
```

```
##                 2.5 %      97.5 %
## (Intercept) -2.549410 -0.07538279
## age          2.509075  2.67197147
```

From the model summary, we also see that both the coefficient estimates are statistically significant at the 95% confidence level. The listed p-values are for the hypothesis tests that check whether (H_0) the true value of the corresponding coefficient is equal to 0, versus $H_1 = \neg H_0$. Thus, the statistically significant p-values indicate that the true value for both coefficients likely differs from 0. The confidence intervals suggest that the true intercept value is likely negative, and the true slope value is likely positive. It's also important to note that the standard error (and thus the confidence interval) for the intercept term is much wider than that for the slope term.
The value $R^2 = 0.9799281$ suggests the model fits the data well in the sense that the response (height) is highly correlated with the predictor (age).

Problem 4 - Removal of the Constant Term, β_0

The value of the intercept coefficient in the above model can be interpretated as the average depth of the sampled pine trees' seeds when planted.

```
lm.fit2 <- lm(formula = height ~ 0 + age, data = df)
summary(lm.fit2)
```

```
##
```

```
## Call:
## lm(formula = height ~ 0 + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4840 -3.0097 -0.8912  2.0205  6.6516
##
## Coefficients:
##     Estimate Std. Error t value            Pr(>|t|)
## age  2.51656    0.02161   116.5 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.007 on 83 degrees of freedom
## Multiple R-squared:  0.9939, Adjusted R-squared:  0.9938
## F-statistic: 1.357e+04 on 1 and 83 DF,  p-value: < 0.00000000000000022
```

We see that the estimated linear regression equation without an intercept term is

$$\widehat{height} = 2.51656 \cdot age$$

The slope coefficient is statistically significant at the 95% confidence level. The small p-value provides strong evidence for the case that tree height is associated with its age. The 95% confidence interval for the estimated slope coefficient is as follows:

```
confint(lm.fit2)
```

```
##        2.5 %   97.5 %
## age 2.473586 2.559531
```

When we fit a linear regreession model, we can decompose the variance as

$$\text{Sum of Squares Total (SST)} = \text{Sum of Squares Error (SSE)} + \text{Sum of Squares Regression (SSR)}$$

The coefficient of determination is defined by

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

If the linear model is fit with an intercept term, the following values are used:

$$SST = \sum (Y_i - \bar{Y})^2$$
$$SSE = \sum (Y_i - \hat{Y}_i)^2$$
$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

However, when the model is fit without an intercept term (i.e. forced to pass through the origin), the above values don't satisfy our decomposition identity that requires $SST = SSE + SSR$. Instead, we can satisfy that identity by using the following values (the 0 subscripts here indicate regression through the origin):

$$SST_0 = \sum Y_i^2$$
$$SSE_0 = \sum (Y_i - \hat{Y}_i)^2$$
$$SSR_0 = \sum \hat{Y}_i^2$$

Note that the definitions of SSR and SST change between the two cases, while SSE remains unchanged. Moreover,

$$SST = SST_0 - n\bar{Y}^2$$

In other words, the total sum of squares of a model with an intercept term is less than or equal to that of a model without an intercept. Let $\widehat{Y}$ and $\widetilde{Y}$ denote the predicted values for the models with and without an intercept, respectively. Then if and only if:

$$\begin{aligned} & R_0^2 > R^2 \\ \iff & 1 - \frac{SSE_0}{SST_0} > 1 - \frac{SSE}{SST} \\ \iff & \frac{SSE_0}{SSE} < \frac{SST_0}{SST} \\ \iff & \frac{SSE_0}{SSE} < \frac{SST + n\bar{Y}^2}{SST} \\ \iff & \frac{SSE_0}{SSE} < 1 + \frac{\bar{Y}^2}{\frac{1}{n}\sum (Y_i - \bar{Y}_i)^2} \end{aligned}$$

The left side of this equation is greater than one since the model through the origin is nested within the model with an intercept term. The denominator of the second term on the right side is the `MSE` of an intercept-only model. So the larger the square of response mean relative to the `MSE` of an intercept-only model, the more likely that $R_0^2 > R^2$, i.e. that a regression through the origin fits the data better than a regression with an intercept term.

Returning to our results, our R_0^2 was greater than our R^2 value because, in our case, $\bar{Y}^2 = 169$ was much greater than the intercept-only model's `MSE` $= 61.6666667$.

Problem 5 - Predictions for New Data

```
df.new <- data.frame(age = 6:15, height.pred = predict(lm.fit2, data.frame(age = 6:15)))
kable(df.new, digits = 4, caption = "Predictions using Regression Through Origin")
```

Table 1: Predictions using Regression Through Origin

age	height.pred
6	15.0994
7	17.6159
8	20.1325
9	22.6490
10	25.1656
11	27.6821
12	30.1987
13	32.7153
14	35.2318
15	37.7484