Stat 505 Assignment 8

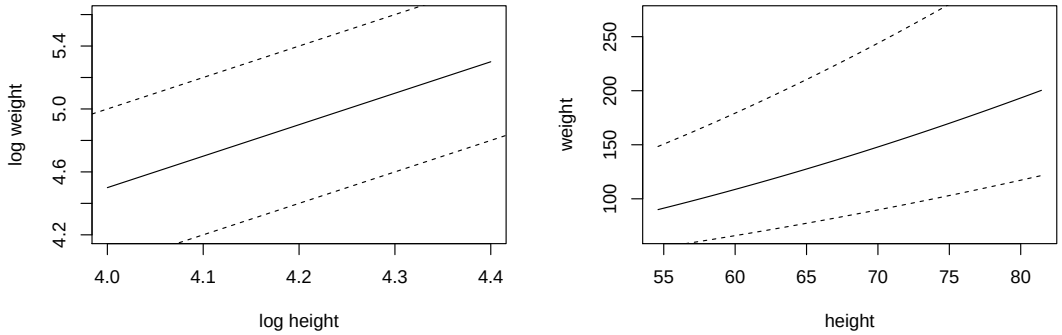
16 points

Solutions

1. ARM 4.1: $\log(weight) = -3.5 + 2.0\log(height) + \epsilon$, $\epsilon \sim (0, .25^2)$

   (a) Approximately 68% of log weights will be within 1 SD (=0.25) of their predicted value, which is within a factor of $(e^{-.25}, e^{.25}) = (0.78, 1.28)$ in the original scale. 1

   (b) Plotting the model in log and original scales with curve 2 SD away. 2



   My lines are $\widehat{\log(y)} = 3.5 + 2\log(height)$ *You should see that the spread of weight increases with height in the second plot. You can backtransform and plot the curve as* $\hat{y} = height^2 \exp(-3.5)$ *but do be sure to add and subtract 0.5* **inside** *the exp function.* 1

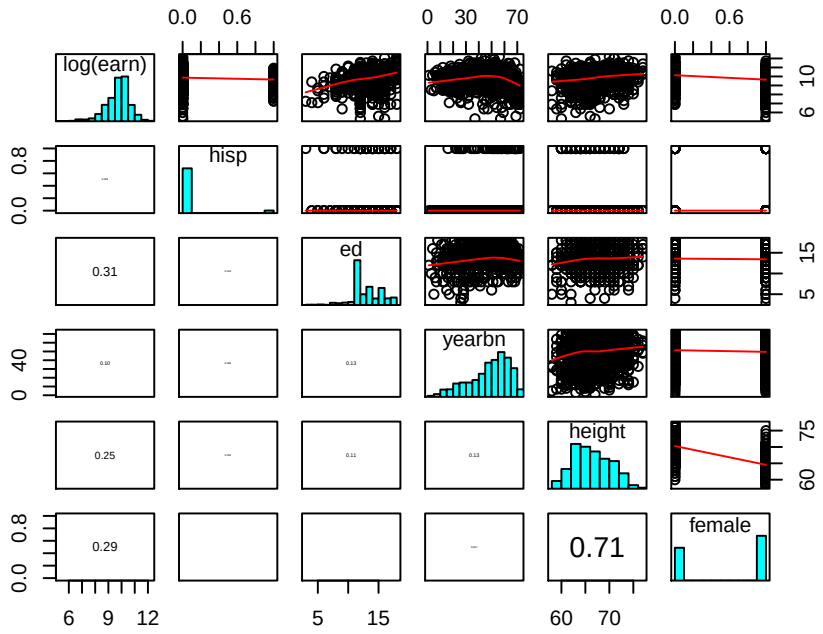
2. ARM 4.2 p 75

   (a) Clean the data.

   At Gelman's book site there is a rather hidden examples folder `http://www.stat.columbia.edu/~gelman/arm/examples/` *which contains info about the coding of these variables in the earnings folder as wfwcodebook. They used 9 for no answer on race and hispanic, 99 for the missing on education and yearbn. Height1 is feet and height2 inches, which combine to give height, so we don't need the first two. I recoded sex as female = sex -1 and hispanic as 2 - hisp. I kept only the complete cases (no NA's). The histograms and scatterplots do not show any bad points.*

   Race categories 3 and 4 have few observations, so I dropped them and created a black variable for race = 2.

   We have to decide what to do with those who have no income. I left them out because they come from a different population who, whether by choice or not, are not working for pay. Inference will now apply only to those who are working. 1



2

Of the possible predictors, education, gender, and height have the strongest correlations with log(earnings), but the height effect may just be due to gender. Age or yearborn seems to have a quadratic relationship to log(earnings).

(b) Model earnings based on education.

	Estimate	Std. Error	t value
(Intercept)	9.53	0.03	314.59
ed	0.12	0.01	11.16

2

Table 1: n = 1161 rank = 2 resid sd = 0.878 R-Squared = 0.097

Mean level of education is 13.5, but for interpretation, I will subtract 12 from the years of education. The intercept estimate of 9.534 (SE = 0.03) means that median earnings for people with a high school diploma are 1.383×10^4. *Every year of education is estimated to increase salary by 0.12 (SE = 0.011) so four more years of education increase median earnings by a factor of* $\exp(4 \times 0.12) = 1.614$.

(c) Adding more terms:
It is common knowledge that women earn less than men, so we expect gender to have a strong negative effect. Similarly, blacks typically earn less than whites of the same experience, so I expect another strong effect for `black`. *I'd guess that the effect of education is weaker for women and for blacks, so I'd look for negative interactions. I'm not so sure about hispanics, and wonder if these data contain enough information on hispanics to be able to assess this, but I'm looking for effects similar to blacks, but weaker.*

	Estimate	Std. Error	t value
(Intercept)	9.93	0.05	205.59
ed	0.10	0.02	6.06
female	-0.64	0.06	-10.17
black	-0.25	0.14	-1.82
hisp	-0.47	0.17	-2.74
ed:female	0.04	0.02	1.73
ed:black	-0.02	0.03	-0.67
ed:hisp	0.03	0.04	0.84
female:black	0.36	0.17	2.13
female:hisp	0.41	0.22	1.87

Table 2: n = 1161 rank = 10 resid sd = 0.836 R-Squared = 0.188

This set of variables has some problems. I think it fits too many interactions. The coefficient on ed:female is opposite the sign I expected. Interactions with female and race/ethnicity are strong and should remain, but I think education interactions with race/ethnicity should be removed.

	Estimate	Std. Error	t value
(Intercept)	9.93	0.05	207.75
ed	0.09	0.02	6.26
female	-0.64	0.06	-10.21
black	-0.27	0.13	-2.07
hisp	-0.42	0.16	-2.61
ed:female	0.04	0.02	1.79
female:black	0.36	0.17	2.12
female:hisp	0.39	0.22	1.78

Table 3: n = 1161 rank = 8 resid sd = 0.835 R-Squared = 0.187

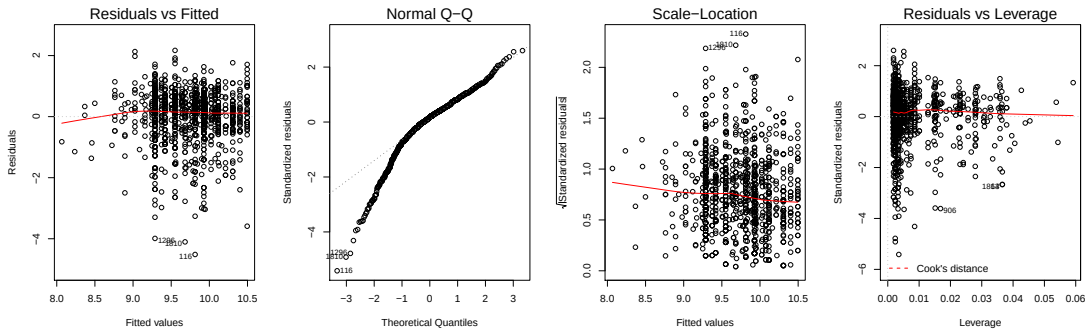
Dropping `ed:black` *and* `ed:hispanic` *does not hurt* R^2 *at all. We are left with: The intercept, estimated at 9.928 (SE= 0.048) gives the median log earnings of a white male with 12 years of schooling. In the earnings scale that is* 2.05×10^4 *[95% CI: (9.832, 10.024)]. Each year of education adds 0.095 (SE= 0.015) to log earnings, increasing median earnings by a factor of 1.099 [95% CI: (1.067, 1.133)]. A white woman's log earnings is generally* -0.641 *(SE= 0.063) less than a white man's so median women's earnings are a factor of 0.527 smaller [95% CI: (0.465, 0.597)]. Median log earnings for black men are smaller than median earnings of white men by a factor of 0.76. [95% CI: (0.583, 0.99)]. Similarly, hispanic men's median log earnings are a factor of0.654 below the median for white men [95% CI: (0.472, 0.905)]. The differential between genders for whites is greater than that for blacks because the* `female:black` *interaction of 0.361 (0.17) is positive on the log scale. Similarly, hispanic women are closer to hispanic men than white women are to white men, as the hispanic:female interaction is 0.387*

(0.217). I find it easiest to compare everyone to the white-male baseline earnings in a table for median earings.

Group	*Median Multiplier (95% CI)*	*Group*	*Multiplier (95% CI)*
White males	*1*	*White Females*	*0.527 (0.465, 0.597)*
Black Males	*0.76 (0.583, 0.99)*	*Black Females*	*0.575 (0.458, 0.721)*
Hispanic Males	*0.654 (0.472, 0.905)*	*Hispanic Females*	0.728 *(0.473, 1.119)*

Finally, I note that the effect of education is slightly stronger for women than for men. A four year degree is estimated to increase median women's earnings by a factor of 1.692 [95% CI (1.514, 1.891)], whereas for men the factor is estimated to be 1.46 [95% CI (1.294, 1.648)]. 5

(d) Diagnostics



2

There are apparently other unmeasured factors which can lower earnings (region, healthiness, family expectations), so this is not a very complete analysis. An R^2 of almost 20% is not too bad for this types of data. The residuals have a very long left tail, which could be telling us that log transform is too big a change.

R Code

```
opts_chunk$set(fig.width=5, fig.height=4, out.width='.5\\linewidth', dev='pdf', size='scriptsize', concordance=TRUE)
options(replace.assign=TRUE,width=88, digits = 3, max.print="72",
        show.signif.stars = FALSE)
require(xtable)
require(arm)
source("http://www.math.montana.edu/~jimrc/classes/stat505/Rcode/displayXtable.r")
```

```
par(mfrow=c(1,2))
 curve(-3.5 + 2.0 * x, from = 4, to=4.4,xlab="log height", ylab="log weight", ylim=c(4.2,5.6))
abline(-3.5 +.5, 2.0,lty=2 )
abline(-3.5-.5, 2.0 , lty=2)
 curve(exp(-3.5 + 2.0 * log(x)), from = exp(4), to=exp(4.4),xlab="height", ylab="weight", ylim=exp(c(4.2,5.6))  )
 curve(exp(-3.5 +.5 + 2.0 * log(x)), from = exp(4), to=exp(4.4), add=TRUE, lty=2)
 curve(exp(-3.5-.5 + 2.0 * log(x)), from = exp(4), to=exp(4.4), add=TRUE, lty=2)
```

```
earnings <- read.csv("http://www.math.montana.edu/~jimrc/classes/stat505/data/earnings.csv")
 earnings$yearbn <- ifelse(earnings$yearbn<99,earnings$yearbn,NA)
 earnings$ed <- ifelse(earnings$ed<99, earnings$ed,NA)
 earnings$hisp <- ifelse(earnings$hisp < 9,earnings$hisp, NA)
 earnings <- subset(earnings, earn>0 & race <3 & complete.cases(earnings))[,-(2:3)]
```

```
earnings$female <- earnings$sex - 1
earnings$hisp <- 2 - earnings$hisp
earnings$black <- ifelse(earnings$race ==2, 1, 0)
earnings$sex <-  earnings$race <- NULL
panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}
panel.hist <- function(x, ...) {
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
}
pairs(log(earn) ~ ., earnings[,1:6], diag.panel=panel.hist, lower.panel=panel.cor, upper.panel=panel.smooth)
```

```
 earnings$ed <-  earnings$ed-12
logEarn.fit1 <- lm(log(earn) ~ ed, earnings)
 coef1 <- coef(logEarn.fit1)
 secoef1 <- sqrt(diag(vcov(logEarn.fit1)))
display.xtable(logEarn.fit1)
```

```
 logEarn.fit2 <- update(logEarn.fit1,. ~  (ed + female + black + hisp)^2 -black:hisp)
display.xtable(logEarn.fit2)
```

```
 logEarn.fit3 <- update(logEarn.fit1,. ~  ed * female + female*(black + hisp))
display.xtable(logEarn.fit3)
coef3 <- coef(logEarn.fit3)
secoef3 <- sqrt(diag(vcov(logEarn.fit3)))
blkWoman <- c(0,0,1,1,0,0,1,0)
hispWoman <- c(0,0,1,0,1,0,1,1)
ed4Women <-  c(0,4,0,0,0,4,0,0)
coef3 <- c(coef3, rbind(blkWoman,hispWoman,ed4Women) %*% coef3)
secoef3 <- c(secoef3, sqrt(diag( rbind(blkWoman,hispWoman,ed4Women) %*% vcov(logEarn.fit3) %*% cbind(blkWoman,hispWoman,ed4Women
CIs <- cbind(coef3-2*secoef3, coef3+2*secoef3)
```

```
par(mfrow=c(1,4))
plot(logEarn.fit3)
```