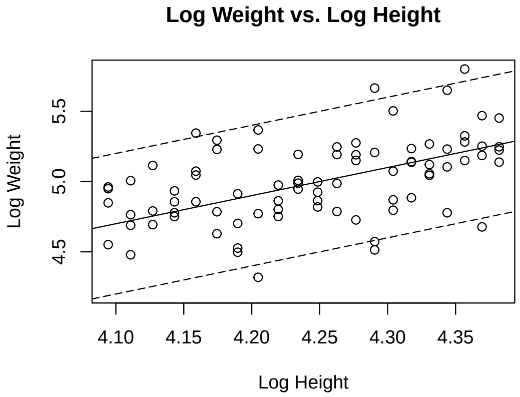
Stat 505 Assignment 8

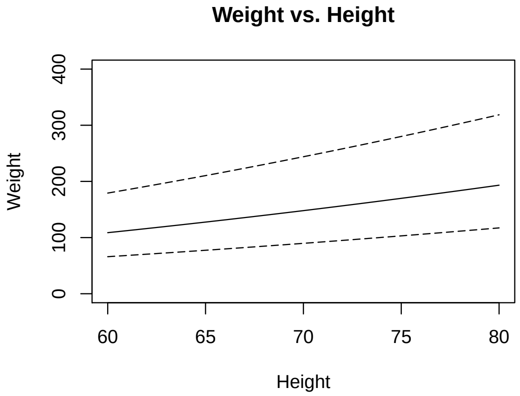
October 31, 2014

1. (a) Approximately 68% of the persons will have weights within a factor of $e^{-0.25} = 0.78$ and $e^{0.25} = 1.28$ of their predicted values from the regression.

   (b) The plot on the log scale is below with dashed lines ±2 standard deviations from the regression line. Note that I simulated data points from this regression model so that I could add points to the scatterplot. I simulated 4 log(weights) at every log(height).



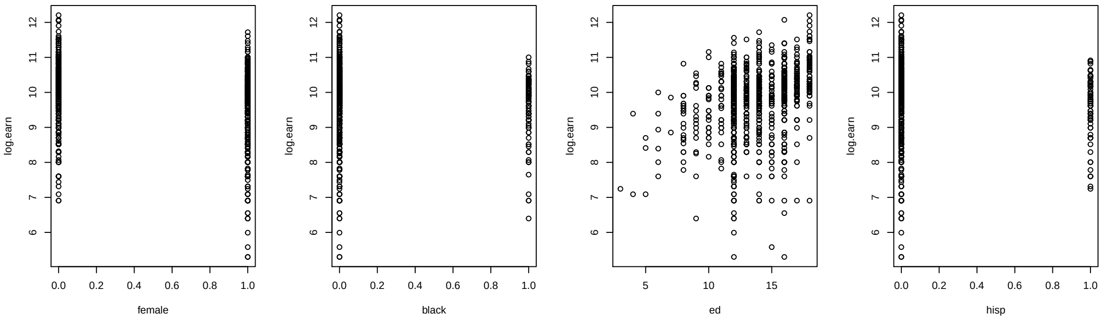
   This next plot shows the regression model on the original scale with dashed lines at a factor of 1.28 away from the predicted values (which corresponds to ±2 standard deviations on the log scale).



2. (a) I dropped the rows with no income ("NA") or 0 income. As a result, I will restrict my inference to the population of people who reported some income. If we restrict our inference to this population, dropping those with no income or 0 income is justified.

   The plots below show the relationship between log earnings and each of the predictors. Log earnings seems to be related to education level, gender, and being black.

It is not clear from the scatterplot whether being hispanic is related to log earnings. This may arise from the fact that the sample size for hispanics is low.



(b) If we fit a regression model of log(earnings) on education, one way to write our regression equation is:

$$mean(log(\widehat{earnings})|education) = \hat{\beta}_0 + \hat{\beta}_1 education$$

Suppose we wanted to estimate the increase in $mean(earnings)$ for a one year increase in education. To do this, we would go about trying to backtransform the regression equation. But, we cannot do this because $mean(log(earnings)) \neq log(mean(earnings))$. But, we can say that $median(log(earnings)) = log(median(earnings))$ because the logarithm preserves ordering. If the model assumptions hold and the log(earnings) at any given education level are normally distributed, then $mean(log(earnings)) = median(log(earnings))$. So, we have:

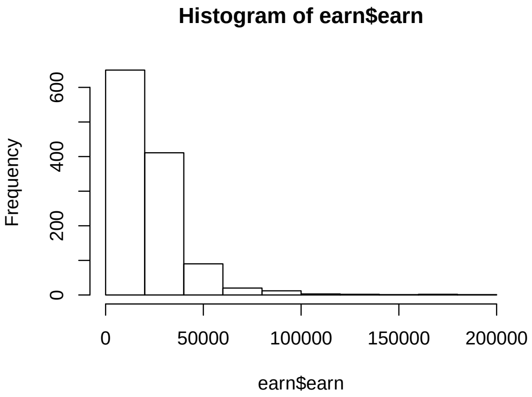
$$e^{\widehat{median}(log(earnings)|ed=x+1)-\widehat{median}(log(earnings)|ed=x)} = e^{\hat{\beta}_0+\hat{\beta}_1(x+1)-(\hat{\beta}_0+\hat{\beta}_1 x)}$$

$$e^{log(\widehat{median}(earnings)|ed=x+1)-log(\widehat{median}(earnings)|x)} = e^{\hat{\beta}_1}$$

$$e^{log(\frac{\widehat{median}(earnings)|x+1}{\widehat{median}(earnings)|x})} = \frac{\widehat{median}(earnings|ed=x+1)}{\widehat{median}(earnings|ed=x)} = e^{\hat{\beta}_1}$$

That's why we say that the **median** earnings is estimated to increase by a factor of $e^{\hat{\beta}_1}$ when we increase education level by one year. Gelman seems to avoid this issue, at least for now. He talks about the change in the predicted value for a one unit increase in the explanatory variable, but he doesn't state whether he is using the mean or the median as a target for prediction.

Also, I will point out that the histogram of earnings, shown below, is right skewed. The median of the earnings is going to be a better measure of center than the mean. In a practical sense, the median of the earnings will be more informative to the reader than the mean of the earnings. So, it does make sense to interpret changes in the median rather than changes in the mean.



I fit a linear model of log earnings on education. The model summary is below. There is strong evidence that the mean of log earnings depends on education level (p-value< 0.0001 from t-stat= 11.73 on 1190 df). For a one year increase in education level, the median of earnings is estimated to change by a factor of 1.13 with an associated 95% confidence interval from 1.11 to 1.15.

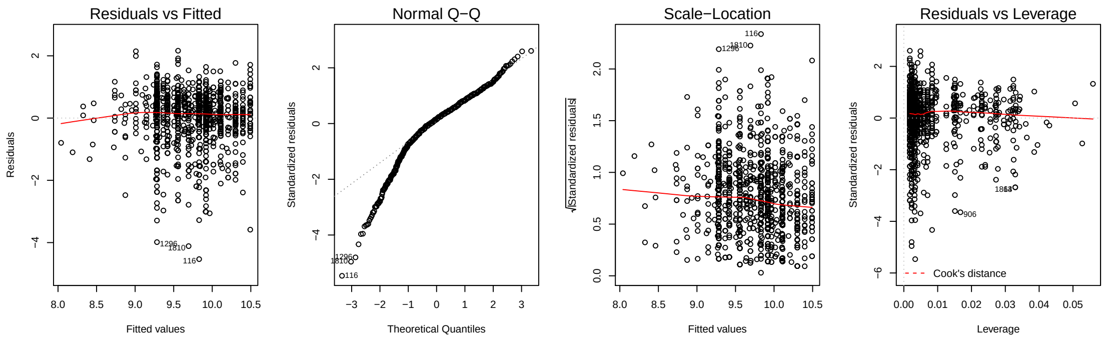
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.0595	0.1433	56.25	0.0000
ed	0.1225	0.0104	11.73	0.0000

(c) I would definitely want to add a female*education interaction because I would expect the wage disparity between men and women to be less for higher education levels. That is, I would expect the gender effect to be less extreme at higher education levels. I would also add female by ethnicity interactions because I suspect that the wage disparity between men and women is greater for black and hispanic people.

There is moderate evidence of a female by education interaction (p-value= 0.0402) and the estimated interaction term has the sign I would expect because a positive number for the coefficient of this interaction term makes the wage disparity between men and women less at higher educations. I will leave the female*education term in the model. There is also moderate evidence of a female by black interaction (p-value= 0.0383). I will leave this term in the model because it has a relatively small p-value, but it has the opposite sign than what I expected. A positive coefficient for this interaction term will make the gender effect less extreme for black people (I originally thought the gender effect would be greater for blacks). The same goes for hispanic people. There is moderate evidence of a female by hispanic interaction (p-value= 0.056). Again, I will leave this term in the model because it has a relatively small p-value, but it has the opposite sign than what I expected.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.7726	0.2069	42.40	0.0000
ed	0.0956	0.0149	6.43	0.0000
female	-1.1302	0.2774	-4.07	0.0000
hisp	-0.4128	0.1542	-2.68	0.0075
black	-0.2672	0.1316	-2.03	0.0425
ed:female	0.0412	0.0200	2.05	0.0402
female:black	0.3509	0.1692	2.07	0.0383
female:hisp	0.4003	0.2089	1.92	0.0556

(d) See the diagnostic plots below. The homogeneity of variance and linearity assumptions look OK. There are a few points with high leverage, but based off their Cook's Distances they do not appear to be influential. There is a violation of the normality assumption that could be of concern. The distribution of the residuals appears to be left skewed with a long left tail. This departure from normality could affect our results. Our conclusions probably won't change for those p-values that were either really large or really small, but we may want to go back and acknowledge that those p-values between 0.01 and 0.1 may not be trustworthy. With this departure from normality, our standard errors may be off and the p-values could be artificially small (or artificially large).



```r
set.seed(77)
height <- c(seq(60,80,by=1))
pred.logweight <- -3.5+2.0*log(height)
obs.logweight <- matrix(c(rep(0,21),rep(0,4)),nrow=21,ncol=4)
for(i in 1:21){
height.feed <- 59+i
obs.logweight[i,] <- -3.5+2.0*log(height.feed)+rnorm(4,0,0.25)
}
sim.dat <- as.data.frame(cbind(log(height),pred.logweight,obs.logweight))
names(sim.dat) <- c("loght","predlogwt","rep1","rep2","rep3","rep4")
plot(rep1~loght, data=sim.dat,xlim=c(4.094,4.382),ylim=c(4.2,5.8),xlab="Log Height", ylab="Log Weight", main="Log Weight vs. Log
points(sim.dat$loght,sim.dat$rep2)
points(sim.dat$loght,sim.dat$rep3)
points(sim.dat$loght,sim.dat$rep4)
abline(a=-3.5,b=2.0)
abline(a=-3.0,b=2.0,lty=5)
abline(a=-4.0,b=2.0,lty=5)
```

```r
set.seed(92)
height <- c(seq(60,80,by=1))
pred.weight <- exp(-3.5)*height^2
```

```r
curve(exp(-3.5)*x^2, from=60,to=80, xlab="Height", ylab="Weight", ylim=c(0,400), main="Weight vs. Height")
curve(exp(-3.0)*x^2, from=60,to=80, add=T, lty=5)
curve(exp(-4.0)*x^2, from=60,to=80, add=T, lty=5)
```

```r
earn <- read.csv("~/Documents/Stat505/Homework/HW8/earnings.csv", head=T)
earn <- subset(earn, earn!="NA" & earn!=0)
earn$black <- with(earn,ifelse(race==2,1,0))
earn$female <- with(earn, ifelse(sex==1,0,1))
earn$hisp <- with(earn, ifelse(hisp==1,1,0))
log.earn <- log(earn$earn)
par(mfrow=c(1,4))
plot(log.earn~female, data=earn)
plot(log.earn~black, data=earn)
plot(log.earn~ed, data=earn)
plot(log.earn~hisp, data=earn)
```

```r
hist(earn$earn)
```

```r
lm.ed <- lm(log.earn~ed, data=earn)
require(xtable)
xtable(summary(lm.ed))
```

```r
par(mfrow=c(1,4))
plot(lm.int)
```

```r
lm.int <- lm(log.earn~ed+female+hisp+black+female*ed+female*black+female*hisp, data=earn)
xtable(summary(lm.int))
```