Government 61: Research Practice in Quantitative Methods

Harvard University
Spring 2018
Jeff Gill

Module 4: Linear Regression: Before and After Fitting

Linear Model

- G&H introduce a model fit to data we downloaded from a survey of adult Americans in 1994 that predicts their earnings (in dollars) given their height (in inches) and sex (coded as 1 for men and 2 for women):

  $$\text{earnings} = 61000 + 1300 \times \text{height} + \text{error},$$

  with a residual standard deviation of 19000.

- Get the data from the syllabus for this week.

- Consider the following code after you have the data to get us started:

```
library(foreign)
earnings.df <- read.dta("Class.Quant/ARM_Data/earnings/heights.dta")
```

  with your own path name of course.

Linear Model

- Change the `sex` variable into a 0/1 variable called `male`:

```
male <- 2 - earnings.df$sex
earnings.df <- data.frame(earnings.df,male)
> head(earnings.df)
   earn height1 height2 sex race hisp ed yearbn height male
1    NA       5       6   2    1    2 12     53     66    0
2    NA       5       4   1    2    2 12     50     64    1
3 50000       6       2   1    1    2 16     45     74    1
4 60000       5       6   2    1    2 16     32     66    0
5 30000       5       4   2    1    2 16     61     64    0
6    NA       5       5   2    1    2 17     33     65    0
```

- Now run the first model:

```
earnings1.out <- lm(earn ~ height, data=earnings.df)
```

Linear Model

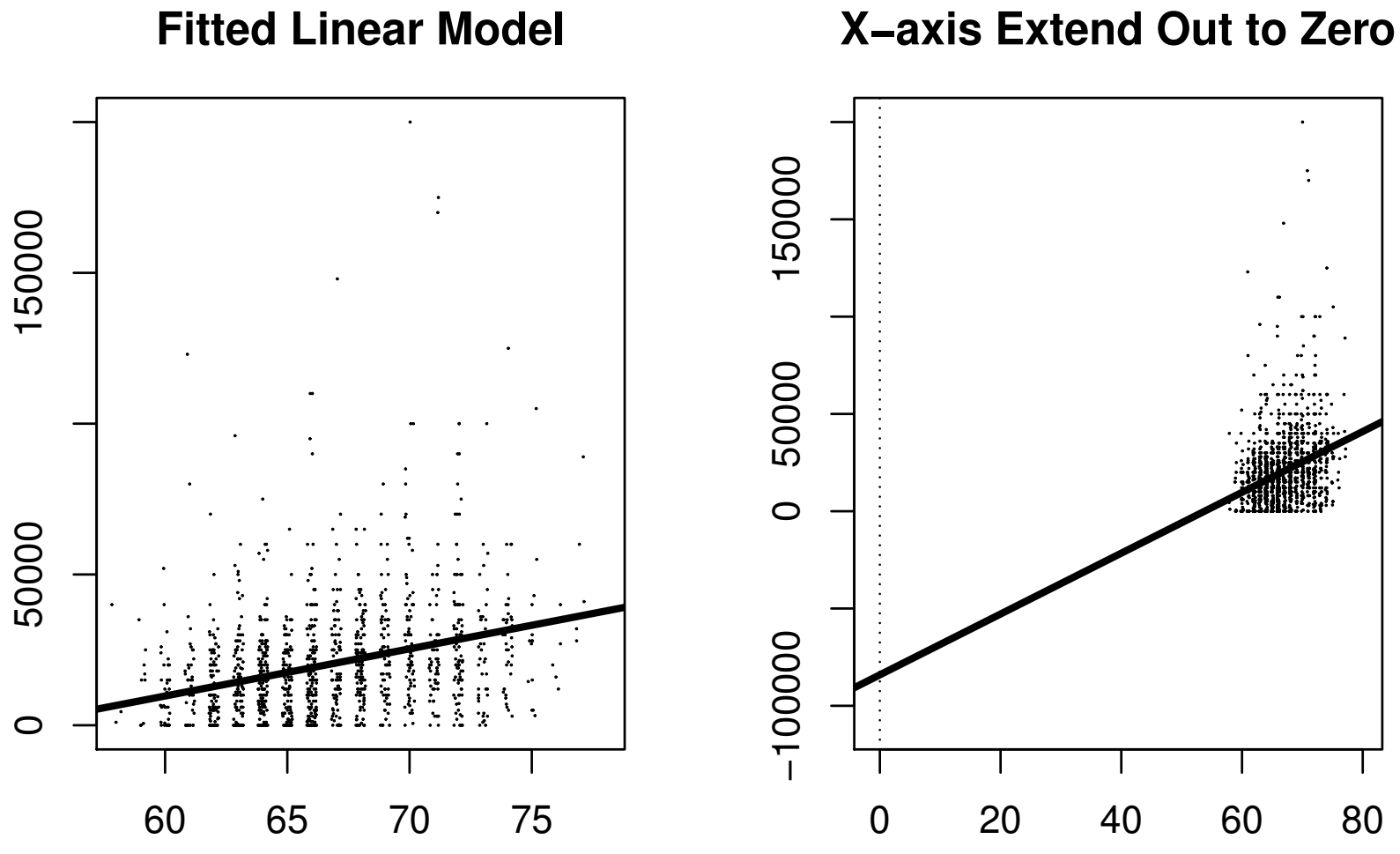
- This produces:

```
summary(earnings1.out)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -84078.3     8901.1  -9.446   <2e-16
height        1563.1      133.4  11.713   <2e-16

Residual standard error: 18850 on 1377 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.09061, Adjusted R-squared:  0.08995
F-statistic: 137.2 on 1 and 1377 DF,  p-value: < 2.2e-16
```

- This is heavily rounded in the book on page 53 because it's not the right model, it's just a start.
- Note that the intercept is substantively meaningless since it is the expected earnings for someone with zero height.
- There are 650 missing values in **earn**, and those are "casewise deleted.".

Linear Model, Figure 4.1



Linear Model, Figure 4.1

```
postscript("Class.Quant/Images/gh.fig.4.1.ps",height=4,width=7,paper="executive")
par(mfrow=c(1,2),oma=c(0,1,0,1),mar=c(3,3,3,2))
plot(jitter(earnings.df$height),earnings.df$earn,pch=19,cex=0.05,xlim=c(58,78),
    xlab="",ylab="", main="Fitted Linear Model")
abline(lm(earnings.df$earn ~ earnings.df$height), lwd=3)
plot(jitter(earnings.df$height),earnings.df$earn,pch=19,cex=0.05,xlim=c(-1,80),
    ylim=c(-110000,200000), xlab="",ylab="", main="X-axis Extend Out to Zero")
abline(lm(earnings.df$earn ~ earnings.df$height), lwd=3)
abline(v=0,lty=3)
dev.off()
system("open Class.Quant/Images/gh.fig.4.1.ps")
```

Other Ways to Look at the Model

- Using G&H's severe rounding, alternatives versions with different measurement are:

  $$\text{earnings} = 61000 + 51 \times \text{height (in millimeters)} + \text{error}$$
  $$\text{earnings} = 61000 + 81000000 \times \text{height (in miles)} + \text{error.}$$

- Obviously this is the same information, but is it useful?

- Another way to scale the coefficients is to standardize the covariate into a z-score: $z.height = (height - 66.9)/3.8$, which we can do in the model statement if we want:

```
earnings2.out <- lm(earn ~ scale(height), data=earnings.df)
```

  with the usage `scale(x, center = TRUE, scale = TRUE)`.

- Or we could create a new explanatory variable instead:

```
z.height <- (earnings.df$height - mean(earnings.df$height,na.rm=TRUE))/
        sd(earnings.df$height,na.rm=TRUE)
earnings2.out <- lm(earn ~ z.height, data=earnings.df)
```

Other Ways to Look at the Model

- Which gives:

```
summary(earnings2.out)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    19965.9      507.7   39.32   <2e-16
scale(height)   5970.3      509.7   11.71   <2e-16

Residual standard error: 18850 on 1377 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.09061, Adjusted R-squared:  0.08995
F-statistic: 137.2 on 1 and 1377 DF,  p-value: < 2.2e-16
```

- Note that this gives the same RSE, R-squared, and F-statistic as before.
- Estimated slope coefficient is interpreted in units of standard deviations of the covariate, and the estimated intercept is the mean of the outcome y when the covariate is set at zero.

Standardization Using Reasonable Scales

- All this leads up to the following G&H scale advice.
- If reasonable keep variables on intuitive scales such as inches, dollars, years, number of votes, etc.
- Sometimes this leads to awkward coefficient estimates, such as the examples so far, so scaling is common such as income per \$10,000, population per \$100,000, etc.
- Also, some scales can be made more convenient such as the popular party identification is on a 1–7 point scale, from strong Republican to strong Democrat.
- The rescaled variable $(PID - 4)/2$ is now -1 for Republicans, 0 for moderates, and $+1$ for Democrats (like a correlation coefficient), making the effect of the estimated coefficient more easily interpreted.

IQ Example from Chapter 3

- Return to the mother/child IQ example from Section 3.3 in the text (data on the syllabus):

```
kid1.out <- lm(formula = kid_score ~ mom_hs + mom_iq + mom_hs:mom_iq, data=mom.iq)
summary(kid1.out)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -11.4820    13.7580  -0.835 0.404422
mom_hs         51.2682    15.3376   3.343 0.000902
mom_iq          0.9689     0.1483   6.531 1.84e-10
mom_hs:mom_iq  -0.4843     0.1622  -2.985 0.002994

Residual standard error: 17.97 on 430 degrees of freedom
Multiple R-squared:  0.2301, Adjusted R-squared:  0.2247
F-statistic: 42.84 on 3 and 430 DF,  p-value: < 2.2e-16
```

Measurement and Interaction

- Note that we specified an *interaction term*: `mom_hs:mom_iq`, in addition to the *main effects*: `mom_hs + mom_iq`.
- This is an hypothesis that levels one explanatory variable affect the effect of the other explanatory variable, and vice versa.
- Otherwise with a linear model the effects of the two explanatory variables are assumed to be independent.
- There is too frequently an incorrect perception that an interaction effect gives an orthogonal, added effect in the conventional way.
- This is not true and interacting two variables on the RHS of a regression model fundamentally changes their interpretation.

Measurement and Interaction

- You have to be careful to interpret coefficients from interaction models in practice (including the effect of G&H rounding):
  - $\beta_0 = -11.4820$ is not meaningful (but necessary).
  - $\beta_1 = 51.2682$ is the effect of no high school to high school for moms, when the mom's IQ is zero (which it can't be: Figure 3.4 on page 35, and discussion on page 55).
  - $\beta_2 = 0.9689$ is the effect of a one-unit change in mom's IQ for a mom without a high school degree.
  - For mom's with a high school degree the effect of a delta-unit change in her IQ is:

$$\Delta_{\text{child IQ}} = 0.9689\Delta_{\text{mom IQ}} - 0.4843$$

    meaning that kids of high school educated moms get less of a benefit from a positive change in mom's IQ. Why????

- Interaction effects tie two variables together such that interpretation of changes in one are conditional on the value set for the other, where the interpretation was aided here by a dichotomous explanatory variable..

Measurement and Interaction

- We can simplify the interpretation of the model coefficients by first subtracting the mean of each input variable (centering only, not standardizing):

```
c.mom.hs <- mom.iq$mom_hs - mean(mom.iq$mom_hs)
c.mom.iq <- mom.iq$mom_iq - mean(mom.iq$mom_iq)
mom.iq <-data.frame(mom.iq,c.mom.hs,c.mom.iq)
kid2.out <- lm(formula = kid_score ~ c.mom.hs + c.mom.iq + c.mom.hs:c.mom.iq,
    data=mom.iq)
summary(kid2.out)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       87.63892    0.90756  96.565  < 2e-16
c.mom.hs           2.84076    2.42667   1.171  0.24239
c.mom.iq           0.58839    0.06058   9.712  < 2e-16
c.mom.hs:c.mom.iq -0.48427    0.16222  -2.985  0.00299

Residual standard error: 17.97 on 430 degrees of freedom
Multiple R-squared:  0.2301, Adjusted R-squared:  0.2247
F-statistic: 42.84 on 3 and 430 DF,  p-value: < 2.2e-16
```

Measurement and Interaction

- Now each main effect is corresponding to a predictive difference with the other input *at its average value*.
- This is often easier to understand and explain.
- Now 2.84076 is the predicted addition to kid's IQ for moms with a high school degree at the average IQ level for moms.
- And 0.58839 is the predicted change in kid's IQ for a 1 unit change in mom's IQ at the average level of education for moms.
- The coefficient on the interaction term, -0.48427, is unchanged (other than rounding) and gives an effect of moving away from the means of the two variables at the same time..
- Note that this gives the same RSE, R-squared, and F-statistic as before.
- G&H later go on to justify their liking the idea of dividing my 2 standard deviations to make model results more similar to those when we start modeling $0, 1$ binary outcome variables in a few chapters.

Interactions in Linear Models

- Consider a simple and general linear model with an interaction term:

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2},$$

  where β_3 is the coefficient estimate corresponding to the product.

- The complete product term, $\beta_3 X_{i1} X_{i2}$, is called a *first-order interaction* or sometimes a *two-factor interaction*, where for obvious reasons the order is one less than the number of factors.

- Subject to mild assumptions (Greene 2003, Ch.4), the sampling distribution of β_3 over its standard error is Student's-t with $N-k-1$ degrees of freedom like a normal regression parameter estimate.

- There is no requirement for the form of this interaction term to be a product of the main effects and others have been suggested, such as $\beta_0 X_{i1}^{\beta_1} X_{i2}^{\beta_2}$ (Wonnacott & Wonnacott 1970) and $\beta_j X_{i1}/X_{i2}$ (Allison 1977).

Interactions in Linear Models

- The meaning of interactions in the linear model is actually easier to interpret if the last expression is rewritten:

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + (\beta_2 + \beta_3 X_{i1}) X_{i2}.$$

- If one is interested in the consequence from changes in the explanatory variable X_{i2} on the outcome variable, it is necessary to take the first derivative of this expression with respect to this variable in order to obtain the *marginal* effect as a composite coefficient estimate:

$$\frac{\partial}{\partial X_{i2}} E[Y_i] = \beta_2 + \beta_3 X_{i1}.$$

- This is useful because it demonstrates that the effect of levels of X_{i2} on the outcome variable are tied to specific levels of X_{i1}: the marginal contribution of X_{i2} is conditional on X_{i1}.

Interactions in Linear Models

- Two scenarios can occur, the first when high levels of one variable have an accelerating effect on the other (β_3 has the same sign as β_2), and the second when high levels of one variable have a dampening effect on the other (β_3 has the opposite sign of β_2).
- So the sign on first-order interaction effects tells us quite a bit about the *conditional* effect that a given explanatory variable has on the outcome variable.
- The interpretation of a given coefficient's effect is now complicated by the requirement that it occur at a specified level of the other explanatory variable.
- Often there are theoretically important levels of X_{i1} that can be used predicting the effect of X_{i2}.
- In the absence of some theory-driven level or point of particular interest, the quantiles of the interaction explanatory variable provide convenient points of analysis.

Correlation and "regression to the mean" (Section 4.3 title)

- Note that this is not what "regression to the mean" means.
- Consider a *bivariate regression* of `y` on `x` where both explanatory and outcome variables are standardized: `x <- (x-mean(x))/sd(x)` and `y <- (y-mean(y))/sd(y)`.
- Therefore the intercept is zero and the slope is simply the correlation between x and y
- Recall that regression *is* correlation, since: $cor(X, Y) = \frac{s_X}{s_Y}\boldsymbol{\beta}$.
- So the slope of a regression of two standardized variables must always be between 1 and 1

Logarithmic Transformations

- Sometimes variables that are all positive are skewed to the right with a few very large values.
- This can sometimes make the model fit poorly for this explanatory variable, and apply the log() function helps.
- For outcome variables this may also be a problem for all positive values because linear regression does not impose any restrictions for predictions, meaning we could predict a negative outcome when it does not make sense (income, duration of war, length of term in office, etc.).
- So we can apply the logarithm to the outcome variable, run the regression without any constraints, make predictions on this scale and then apply the exp() function to return predictions to the positive scale.
- Note that natural logs, those with base 2.71..., are preferred in statistics (see the discussion in G&H on pages 60-61).

Logarithmic Transformations

- A *linear model on the logarithmic scale* corresponds to *a multiplicative model on the original scale.*

- Start with taking the log of the outcome variable only as just described:

$$\log(y_i) = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i$$

- Applying the exponential function to both sides gives:

$$y_i = \exp\left[b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + \epsilon_i\right]$$

- Using the property of exponents and the notation $B = \exp[b]$:

$$\begin{aligned} y_i &= \exp[b_0] \times \exp[b_1 X_{i1}] \times \exp[b_2 X_{i2}] \times \cdots \times \exp[\epsilon_i] \\ &= B_0 \times B_1^{X_{i1}} \times B_2^{X_{i2}} \times \cdots \times E_i \end{aligned}$$

since $\log\left[B_k^{X_{ik}}\right] = X_{ik}\log(B_k) = X_{ik}\log(\exp[b_k]) = b_k X_{ik}$.

Back to the Height and Earnings Example

- Income/Earnings/Wealth are skewed right in every society whether measured in dollars or cows.
- So using the log transformation makes sense (excluding zeros).
- This means fitting the regression to the log of earnings and then using exponentiation to return to the original scale for predictions.
- Start with changing the outcome variable and appending to the data frame:

```
log.earn <- log1p(earnings.df$earn)
earnings.df <- data.frame(earnings.df,log.earn)
names(earnings.df)
 [1] "earn"     "height1"  "height2"  "sex"      "race"     "hisp"     "ed"
 [8] "yearbn"   "height"   "male"     "log.earn"
```

- Now run the multiplicative model:

```
earnings3.out <- lm(log.earn ~ height, data=earnings.df)
```

Regression Output, Multiplicative Model

- The output is:

```
summary(earnings3.out)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.59569    1.56403  -4.856 1.33e-06
height       0.24016    0.02345  10.242  < 2e-16

Residual standard error: 3.313 on 1377 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.07079, Adjusted R-squared:  0.07011
F-statistic: 104.9 on 1 and 1377 DF,  p-value: < 2.2e-16
```

- These results are different than the book's because I did not get rid of cases with earnings equal to zero.
- Instead I used the `log1p()` function, which adds a small amount during the logging process.
- Note that there are 650 missing values in `earn` and therefore `log.earn`.

Regression Output, Multiplicative Model

- The estimated intercept coefficient, $\hat{\beta}_0 = -7.59569$ is the predicted log earnings for someone with a height of zero inches. It is also *negative*.
- The estimated coefficient of $\hat{\beta}_1 = 0.24016$ means that a 1 inch change in height gives an expected positive change of 0.24016 in log(earnings), meaning that earnings are multiplied by $\exp(0.24016) = 1.271453$.
- The effect is *multiplicative* rather than *additive* since:

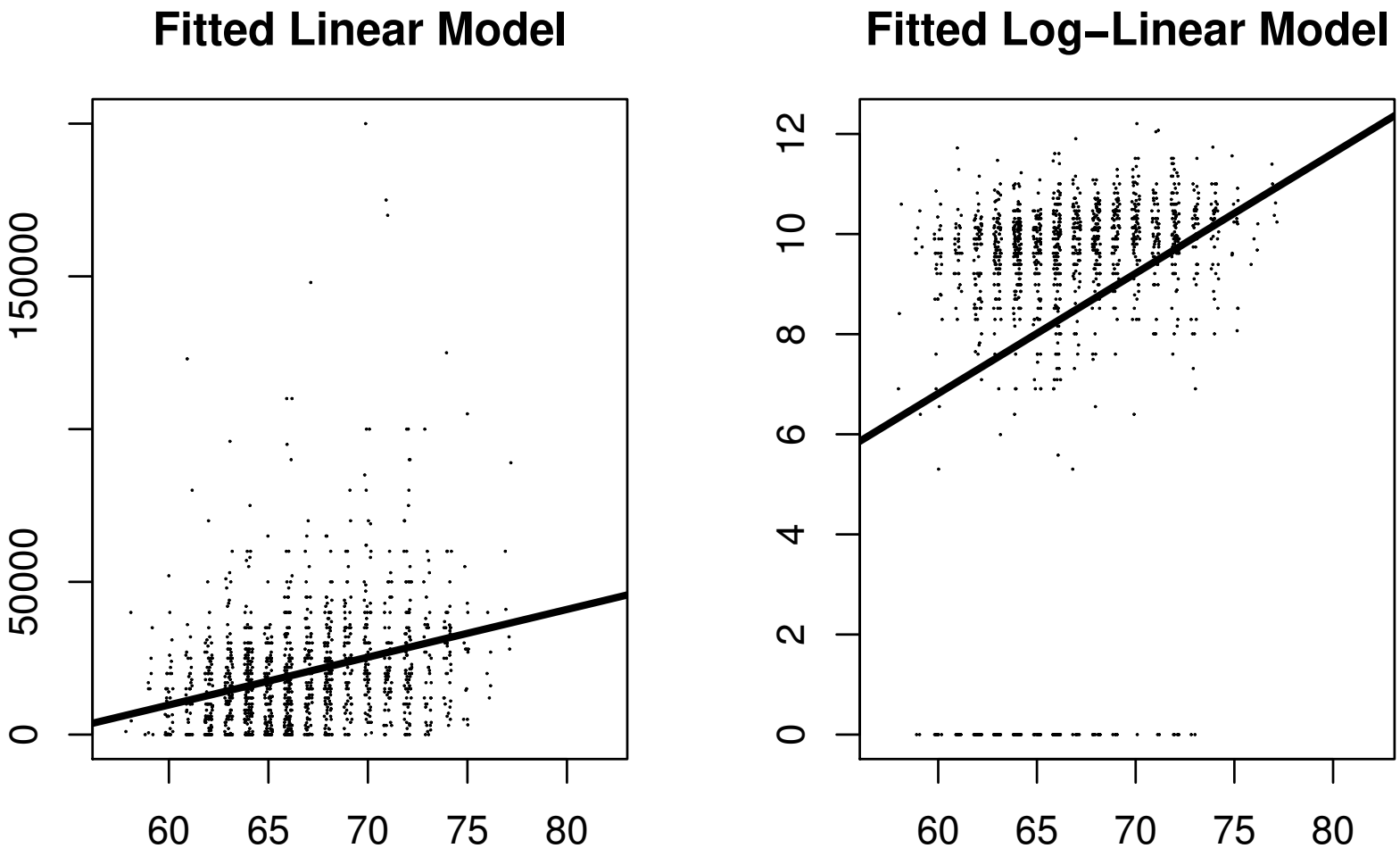
$$y_i = \exp[b_0] \times \exp[b_1 X_{i1}] \times \exp[b_2 X_{i2}] \times \cdots \times \exp[\epsilon_i]$$

  for models with logged outcomes.
- So in predictive terms on the regular scale a 1 inch change in height gives a 27% change in earnings.
- This is clearly too large, so perhaps G&H were right to exclude the zero income cases.
- Their result removing zero income cases is more modest: a 1 inch change in height gives a 6% change in earnings.

Linear and Log-Linear Models, Figure 4.3

```
postscript("Class.Quant/Images/gh.fig.4.3.ps",height=4,width=7,paper="executive")
par(mfrow=c(1,2),oma=c(0,1,0,1),mar=c(3,3,3,2))
plot(jitter(earnings.df$height),earnings.df$earn,pch=19,cex=0.05,xlab="",ylab="",
    main="Fitted Linear Model")
abline(lm(earnings.df$earn ~ earnings.df$height), lwd=3)
plot(jitter(earnings.df$height),earnings.df$log.earn,pch=19,cex=0.05,
    xlab="",ylab="", main="Fitted Log-Linear Model")
abline(lm(earnings.df$log.earn ~ earnings.df$height), lwd=3)
abline(v=0,lty=3)
dev.off()
system("open Class.Quant/Images/gh.fig.4.3.ps")
```

Linear Model Comparison, Figure 4.3



Adding Another Predictor

- Do taller people earn more, on average, than shorter people of the same sex?
- Extending the model with another explanatory variable, using the recoded `male` explanatory variable:

```
earnings4.out <- lm(log.earn ~ height + male, data=earnings.df)
summary(earnings4.out)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.78594    2.48791   1.924  0.05460
height       0.09386    0.03266   2.874  0.00412
male         1.62609    0.25639   6.342 3.06e-10

Residual standard error: 3.267 on 1376 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.09718, Adjusted R-squared:  0.09587
F-statistic: 74.06 on 2 and 1376 DF,  p-value: < 2.2e-16
```

Adding Another Predictor

- After controlling for sex an additional inch of height corresponds to an additional expected 9% in log earnings: in this model people of the same sex but different by 1 inch in height will differ by an average of 9% in expected log earnings.
- Interestingly, the multiplicative predicted change in regular earnings is essentially the same since: $\exp(0.09386) = 1.098406$.
- The prediction difference between the sexes is more pronounced: mens' earnings are about 5 times more same height: $\exp(1.62609) = 5.083958$.
- This does not seem quite right, suspicious decisions: including zero earnings values (there is almost certainly causal reasons related to this independent of height), casewise deletion for NA earnings, and "omitted variable bias."

Another Predictor with an Interaction

- How does an interaction between height and male change the model?

```
earnings5.out <- lm(log.earn ~ height + male + height:male, data=earnings.df)
summary(earnings5.out)
Coefficients:
(Intercept) -1.14595    2.84265  -0.403  0.68692
height       0.13540    0.04404   3.075  0.00215
male         7.85785    4.44019   1.770  0.07700
height:male -0.09225    0.06562  -1.406  0.16000

Residual standard error: 3.266 on 1375 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.09848, Adjusted R-squared:  0.09651
F-statistic: 50.06 on 3 and 1375 DF,  p-value: < 2.2e-16
```

Another Predictor with an Interaction, Interpretation

- So the predicted log earnings and earnings are:

$$\log(\hat{y}) = -1.14595 + 0.13540 \times \text{height} + 7.85785 \times \text{male} - 0.09225 \times \text{height} \times \text{male}$$

$$\hat{y} = \exp[-1.14595] \times \exp[0.13540(\text{height})] \times \exp[7.85785(\text{male})] \times \exp[-0.09225(\text{height} \times \text{male})]$$

- The intercept, -1.14595, is the predicted log earnings if `height` and `male` both equal zero, which has no practical interpretation.
- The coefficient for `height`, 0.13540, is the predicted difference in log earnings corresponding to a 1-inch difference in height, *if* `male` *equals zero* due to the interaction.
- For women there is a predicted 13% difference in log-earnings for a 1-inch difference in height. Unlike the book, this is a statistically reliable coefficient estimate.
- Since $\exp(0.13540) = 1.144995$, then there is a 14.5% multiplicative affect for a 1 inch difference for women.

Another Predictor with an Interaction, Interpretation

- Again:

$$\log(\hat{y}) = -1.14595 + 0.13540 \times \text{height} + 7.85785 \times \text{male} - 0.09225 \times \text{height} \times \text{male}$$

$$\hat{y} = \exp[-1.14595] \times \exp[0.13540(\text{height})] \times \exp[7.85785(\text{male})] \times \exp[-0.09225(\text{height} \times \text{male})]$$

- The 7.85785 coefficient for `male` (0/1) is the predicted difference in log earnings going from women to men, *if height equals zero* due to the interaction. Height never equals zero, so this has no direct interpretation.
- The coefficient for `height:male`, -0.09225, is the difference in slopes of the lines predicting log earnings on height, comparing men to women.
- If this were statistically reliable it would mean that men have a disadvantage in the height–log earnings relationship.
- Specifically in percentage terms a 1-inch increase in height for men gives a $13.5\% - 9\% = 4.5\%$ change in log earnings..
- Whereas a 1-inch increase in height for women gives a 13.5% difference.
- Of course men have the built-in large positive 7.85785 difference in addition to this component.

Linear Transformations to Make Coefficients More Interpretable

- We can make the parameters in the interaction model clearer to interpret by rescaling the height predictor to have a mean of 0 and standard deviation 1:

```
z.height <- (earnings.df$height - mean(earnings.df$height,na.rm=TRUE))/
    sd(earnings.df$height,na.rm=TRUE)
earnings.df <- data.frame(earnings.df,z.height)
```

- For reference these values are:

```
mean(earnings.df$height,na.rm=TRUE)
[1] 66.56111
sd(earnings.df$height,na.rm=TRUE)
[1] 3.81942
```

Linear Transformations to Make Coefficients More Interpretable

- Now run the interaction model with this variable:

```
earnings6.out <- lm(log.earn ~ z.height + male + z.height:male, data=earnings.df)
summary(earnings6.out)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.8666     0.1437  54.758  < 2e-16
z.height        0.5172     0.1682   3.075  0.00215
male            1.7173     0.2644   6.496 1.15e-10
z.height:male  -0.3524     0.2506  -1.406  0.16000

Residual standard error: 3.266 on 1375 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.09848, Adjusted R-squared:  0.09651
F-statistic: 50.06 on 3 and 1375 DF,  p-value: < 2.2e-16
```

Linear Transformations to Make Coefficients More Interpretable

- The prediction model is summarized by:

$$\log(\hat{y}) = 7.8666 + 0.5172 \times \text{z.height} + 1.7173 \times \text{male} - 0.3524 \times \text{height} \times \text{male}$$

- The intercept is the predicted log earnings if `z.height` and `male` both equal zero.
- So for a woman of average height, 66.56 inches, additive the contribution to log earnings is 7.8666, and thus actual dollar earnings of $\exp(7.8666 + 0 + 0 + 0) = 2608.681$, because all the other terms get zeroed-out.
- The coefficient for `z.height` is the predicted difference in log earnings corresponding to a 1 standard-deviation difference in height, *if* `male` *equals zero*.
- This means that the estimated predictive difference for a 3.8-inch increase in log height is 51% for women in log earnings in addition to 7.8666.
- In terms of regular earnings a 3.8 inch increase in height for women is a gigantic 68% increase, from $\exp(0.5172) = 1.677325$.
- So going from 66.56 inches high to 70.36 inches gives women and expected change dollar earnings from \$2,608 to \$4,376.

Linear Transformations to Make Coefficients More Interpretable

- Repeating:

$$\log(\hat{y}) = 7.8666 + 0.5172 \times \text{z.height} + 1.7173 \times \text{male} - 0.3524 \times \text{z.height} \times \text{male}$$

- The coefficient for `male` is the predicted difference in log earnings between women and men, if `z.height` equals 0, meaning the average person.
- Thus, an average height man, 66.56 inches, is predicted to have log earnings that are 1.7173 higher than that of an average height woman since the first and third terms get zeroed-out.
- This corresponds to a ratio of $\exp(1.7173) = 5.569471$, so the man is predicted to have over five times higher earnings than the woman!
- The coefficient for `z.height:male`, -0.3524, is the difference in slopes between the predictive differences for height among women and men.
- Thus, a 3.8-inch difference of height corresponds to 35% more of a *decrease* in log earnings for men than for women.

Further Difficulties in Interpretation

- Consider the simpler log earnings regression without the interaction term:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.78594    2.48791   1.924  0.05460
height       0.09386    0.03266   2.874  0.00412
male         1.62609    0.25639   6.342 3.06e-10
```

- The predictive interpretation of the height coefficient is simple enough: comparing two adults of the same sex, the taller person will be expected to earn 9% more per inch of height in log earnings.
- For the coefficient for `male`, comparing two adults of the same height but different sex, the man will be expected to earn 1.62 more log earnings.
- However, if we are comparing a 66-inch woman to a 66-inch man, then we are comparing a tall woman to a short man.
- So, in some sense, they do not differ only in sex.
- A more reasonable comparison would be of an "average woman" to an "average man."

Further Difficulties in Interpretation

- Wait! We can do that:

```
mean(earnings.df$height[earnings.df$male == 0],na.rm=TRUE) # MEAN HEIGHT FOR WOMEN
[1] 64.49725
mean(earnings.df$height[earnings.df$male == 1],na.rm=TRUE) # MEAN HEIGHT FOR MEN
[1] 70.08847
```

- So an average height woman gets expected log earnings of:

$$\hat{y}_{\text{woman}} = 4.78594 + 0.09386(64.49725) + 1.62609(0) = 10.83965$$

meaning \$51,003.52.

- And an average height man gets expected log earnings of:

$$\hat{y}_{\text{man}} = 4.78594 + 0.09386(70.08847) + 1.62609(1) = 12.99053$$

meaning \$438,243.5.

- My guess is that many of the zero incomes in the data are women who choose not to work (we can't know for sure with the supplied variables).

Exploring Zero Income Some More

- Create a dichomotized version of earnings:

```
d.earn <- earnings.df$earn
d.earn[d.earn > 0] <- 1
```

- Tabulate this with `male`:

```
table(d.earn, earnings.df$male)

d.earn   0   1
     0 172  15
     1 687 505
```

Log-Log Model, Transforming the Input and Outcome Variables

- If the log transformation is applied to an input variable as well as the outcome, the coefficient can be interpreted as the expected proportional change in y per proportional change in x.
- For example:

```
earnings7.out <- lm(log.earn ~ log(height) + male, data=earnings.df)
summary(earnings7.out)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.7676     9.0832  -2.066  0.03900
log(height)   6.3264     2.1802   2.902  0.00377
male          1.6256     0.2553   6.368  2.6e-10

Residual standard error: 3.266 on 1376 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.09728, Adjusted R-squared:  0.09597
F-statistic: 74.15 on 2 and 1376 DF,  p-value: < 2.2e-16
```

Log-Log Model, Transforming the Input and Outcome Variables

- Repeating:

```
log(height)   6.3264     2.1802   2.902  0.00377
male          1.6256     0.2553   6.368  2.6e-10
```

- Because both sides are logged for the variables of interest, this is easy.
- So for each 1% difference in height, the predicted difference in earnings is 6.3264%.
- The other input, `male`, is categorical so it does not make sense to take its logarithm.
- In economics, the coefficient in a log-log model is sometimes called an "elasticity."

Square Root Transformations

- The square root is sometimes useful for compressing high values more mildly than is done by logging.
- Fitting a linear model to the raw, untransformed scale seemed inappropriate to the heights/earnings data as evidenced by the non-transformed data.
- This was equivalent to saying that the difference between zero income and $10,000 is the same difference substantive as between $80,000 and $90,000.
- From:

```
summary(earnings1.out)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -84078.3     8901.1  -9.446   <2e-16
height        1563.1      133.4  11.713   <2e-16
```

This means that an extra inch of height is always worth $1,563.10 in predicted earnings at all levels.

Square Root Transformations

- Maybe the log transformation was too severe of a "pulling back" of the right-hand-side of earnings.
- Using the natural log the differences between populations earning \$5,000 versus \$10,000 is equivalent to the differences between those earning \$40,000 versus those earning \$80,000.
- Using square root the differences between the \$0 earnings and \$10,000 earnings groups is roughly the same as between \$10,000 and \$40,000 or between \$40,000 and \$90,000.
- However , models on the square root scale lack the easy interpretation of the original-scale and log-transformed models: large negative predictions on this scale get squared and become large positive values on the original scale, thus introducing a "non-monotonicity."

Square and Other Powers

- For explanatory variables that appear not to have a linear relationship, sometimes polynomial treatments, X^k, are very useful.
- A classic example is income, which can have a linear INC and a quadratic INC^2 effect on some outcome variable of interest.
- Other polynomials (powers greater than 2, and non-integer values) are sometimes used but often are more difficult to explain in terms of the substance of the effect.
- Here's an example where I used a square last year:
  Katherine Steffen, Allan Doctor, Julie Hoerr, Jeff Gill, Chris Markham, Sarah M. Brown, Daniel Cohen, Rose Hansen, Emily Kryzer, Jessica Richards, Sara Small, Stacey Valentine, Jennifer L. York, Enola K. Proctor, Philip C. Spinella. "Controlling Phlebotomy Volume Diminishes PICU Transfusion: Implementation Processes and Impact." Pediatrics, Vol. 140(2), 2017.

Square and Other Powers

```
pbm.out.B <- lm(Study.Total.BloodRemoved ~ Weight.Kg + PostIntervention + PRISM3
    + Respiratory + AnemiaRiskBasedOnDiagnosis + Admit.Hb + Admit.Hct
    + LabDraws.Number + I(Overdraw.Volume^2)
    ,data=current.pbm.df,na.action=na.fail)
out.table <- mice.output(t(coef.mat),t(se.mat),pbm.out.B)
print(out.table)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.25573	10.94263	1.30277	0.09819
Weight.Kg	0.30733	0.08871	3.46442	0.00036
PostIntervention	-11.58462	4.02756	-2.87633	0.00222
PRISM3	0.23222	0.44618	0.52045	0.30168
Respiratory	-3.9944	4.7396	-0.84277	0.20019
AnemiaRiskBasedOnDiagnosis	-8.51604	5.02841	-1.69359	0.04593
Admit.Hb	-1.59254	3.78666	-0.42057	0.33761
Admit.Hct	0.19681	1.3126	0.14994	0.4406
LabDraws.Number	2.15382	0.0644	33.4422	0.0001
I(Overdraw.Volume^2)	0.61112	0.54795	1.11529	0.13478

Specifying a Richer Model

- Add some variables that we have been ignoring:

```
earnings8.out <- lm(log.earn ~ height + male + race + ed + yearbn,
    data=earnings.df)
summary(earnings8.out)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.558683   2.100386  -0.266  0.79029
height       0.079040   0.032316   2.446  0.01457
male         1.654166   0.250591   6.601 5.82e-11
race        -0.037409   0.133194  -0.281  0.77886
ed           0.295947   0.035668   8.297 2.52e-16
yearbn      -0.017179   0.005526  -3.109  0.00192

Residual standard error: 3.186 on 1373 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.1431,Adjusted R-squared:   0.14
F-statistic: 45.87 on 5 and 1373 DF,  p-value: < 2.2e-16
```

Discretizing Continuous Random Variables

- In some cases it is appropriate to discretize a continuous variable if a simple monotonic or quadratic relation does not work.
- G&H give the example of modeling political preferences, where it can make sense to change raw age to four indicator variables: 18-29, 29-44, 45-64, and 65+, to allow for different generational patterns
- There are other times where the variable looks continuous but the effect is really ordered discrete (ordinal).
- Classic cases are Likert scales and feeling thermometers.
- Returning to IQ score model from Chapter 3, look at mother's employment:
  - `mom.work` = 1: mother did not work in first three years of childs life
  - `mom.work` = 2: mother worked in second or third year of childs life
  - `mom.work` = 3: mother worked part-time in first year of childs life
  - `mom.work` = 4: mother worked full-time in first year of childs life.

Discretizing Continuous Random Variables

- So rather than treat **mom.work** as a continuous variable just because it's numbered, let's treat it as a factor:

```
kid3.out <- lm(formula = kid_score ~ as.factor(mom_work), data=mom.iq)
summary(kid3.out)
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            82.000      2.305  35.568   <2e-16
as.factor(mom_work)2    3.854      3.095   1.245   0.2137
as.factor(mom_work)3   11.500      3.553   3.237   0.0013
as.factor(mom_work)4    5.210      2.704   1.927   0.0547

Residual standard error: 20.23 on 430 degrees of freedom
Multiple R-squared:  0.02444, Adjusted R-squared:  0.01763
F-statistic:  3.59 on 3 and 430 DF,  p-value: 0.01377
```

- R makes the first category the *reference category*, although you can change which one is used with the **relevel()** function.

- This finding means that, modulo the effects of omitted variables, that kids with category 3 moms are the most advantaged: $\hat{y} = 82 + 11.5 = 93.5$.

More Complete Earnings Model

- Now let's respecify the richer earnings model with a factor for the year born (age really):

```
year.factor <- cut(earnings.df$yearbn, 4)
table(year.factor)
year.factor
(-0.099,24.8]   (24.8,49.5]   (49.5,74.2]   (74.2,99.1]
          280           697          1032            20
earnings9.out <- lm(log.earn ~ height + male + race + ed + year.factor,
    data=earnings.df)
summary(earnings9.out)
```

More Complete Earnings Model

- The results now include the factor for age with the first category as the reference category:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.31989    2.09983  -0.152 0.878940
height                     0.07742    0.03222   2.403 0.016396
male                       1.62964    0.25076   6.499 1.13e-10
race                      -0.05544    0.13321  -0.416 0.677337
ed                         0.29950    0.03572   8.385  < 2e-16
year.factor(24.8,49.5] -1.09563    0.32370  -3.385 0.000733
year.factor(49.5,74.2] -1.08733    0.31237  -3.481 0.000515
year.factor(74.2,99.1]  1.05848    1.86204   0.568 0.569819

Residual standard error: 3.183 on 1371 degrees of freedom
  (650 observations deleted due to missingness)
Multiple R-squared:  0.1459,Adjusted R-squared:  0.1415
F-statistic: 33.46 on 7 and 1371 DF,  p-value: < 2.2e-16
```

Identifiability

- G&H: "A model is said to be nonidentifiable if it contains parameters that cannot be estimated uniquely–or, to put it another way, that have standard errors of infinity."
- Model identifiability is really a much deeper concept that depends jointly on the data used, the model specified, and the estimated quantities of interest.
- A common instance of parameter nonidentifiability is when some explanatory variable can be expressed as a linear combination of some others.
- This is why we (or `R` automatically) specify "dummy coding" for index (categorical) variables, where a reference category is specified for comparison like the IQ example just presented.
- If an index variable takes on J values, then we need to leave one category out since it could be predicted perfectly knowing the other categories.
- For example, we cannot have a $0/1$ variable for men *and* a $0/1$ variable for women as well because know that a person had a 0 on the men variable means knowing that the person had a 1 on the women variable.
- More technically, this duplication of information does not allow us to invert the $\mathbf{X}'\mathbf{X}$ matrix.

Gelman & Hill General Regression Modeling Principles

- Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.
- It is not always necessary to include these inputs as separate predictorsfor example, sometimes several inputs can be averaged or summed to create a "total score" that can be used as a single predictor in the model.
- For inputs that have large effects, consider including their interactions as well.
- We suggest the following strategy for decisions regarding whether to exclude a variable from a prediction model based on expected sign and statistical significance.
- Some guidance on individual coefficient significance:
  - If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also probably not hurting them.
  - If a predictor is not statistically significant and does not have the expected sign (for example, incumbency having a negative effect on vote share), consider removing it from the model (that is, setting its coefficient to zero).
  - If a predictor is statistically significant and does not have the expected sign, then think hard if it makes sense.
  - If a predictor is statistically significant and has the expected sign, then keep it in the model.

Additional Modeling Principles

- Don't "over-scientize" the discussion with 14 digits after the decimal points, etc.
- Discuss null results carefully: *lack of evidence of an effect is not evidence of a lack of an effect.*.
- Don't put stars on tables.
- Use p-values only if you have to or you need them to make some point, remember that they come from manure (Fisher's work at the time).
- Don't fixate on p-value thresholds and small p-value differences if you use them.
- Confidence intervals are preferred to t-statistics.
- Separate statistical significance and substantive significance in your writing.
- Tables should reflect the actual number of significance digits in the result.
- Graphical displays are often better than tables or descriptions.
- Don't call yourself a Frequentist unless you really are designing and repeating experiments.