

STAT 220: Basic Statistics for Quantitative Students

Spring 2006

Assignment due Apr. 14

- 14.2**
- a. $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, where $\beta_1 =$ slope of the regression equation.
 - b. Reject the null hypothesis; conclude that there is a linear relationship. The value of the t-statistic is -14.11 and the p-value is 0.000 .
 - c. $t = \frac{-1.6436 - 0}{0.1165} = -14.11$
- 14.8**
- a. $\hat{y}_i = 577 - 3.01(21) = 513.79$ ft. (about 514 ft.)
 - b. $e_i = y_i - \hat{y}_i = 525 - 513.79 = 11.21$ ft. (about 11 ft.)
 - c. $513.79 \pm (2 \times 50)$ which is 413.79 ft. to 613.79 ft. (about 414 to 614 ft.). From the Empirical Rule about 95% of individual values are within two standard deviations of the mean.
 - d. Yes. 650 feet is more than 2 standard deviations from the mean distance for drivers who are 21 years old, so it would be unusual. Notice that 650 feet is outside the interval calculated in the previous part.
- 14.11**
- a. $s = 1.509$ hours. This is roughly the average deviation of individual y-values from the regression line.
 - b. $\hat{y} = 7.56 - 0.269(4) = 6.48$ hours.
 - c. $6.48 \pm (2 \times 1.509)$ which is 6.48 ± 3.018 , or about 5.46 hour to 9.50 hours. Remember that about 95% of individuals will be within two standard deviations of the mean.
 - d. Hours of study explains 12.7% of the observed variation in hours of sleep.
- 14.38**
- a. The equation is $\hat{y} = 30.0 + 0.576x$. With more accuracy, it is $\hat{y} = 29.981 + 0.57568x$.
 - b. $t = \frac{b_1}{s.e.(b_1)} = \frac{0.57568}{0.07445} = 7.73$
 - c. The hypotheses are $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$. The p-value given is 0.000 (associated with $t = 7.73$) so we can reject the null and conclude that the relationship is statistically significant.
 - d. An approximate 95% confidence interval for the population slope is $b_1 \pm 2 s.e.(b_1)$ which is $0.57568 \pm (2 \times 0.07445)$ or about 0.427 to 0.725. For the population of male college students, we can say with 95% confidence that for each 1-inch increase in father's height the mean increase in son's height is between 0.427 inches and 0.725 inches.

B. Here are some preliminary R commands:

```
attach(read.csv(
```

```

"http://www.stat.psu.edu/~dhunter/220/files/datasets/survey.csv",
na="")
p=as.character(Sex)
p[p=="Male"]="x"
p[p=="Female"]="o"
Mmodel=lm(Idealwt~Weight, subset=Sex=="Male")
Fmodel=lm(Idealwt~Weight, subset=Sex=="Female")

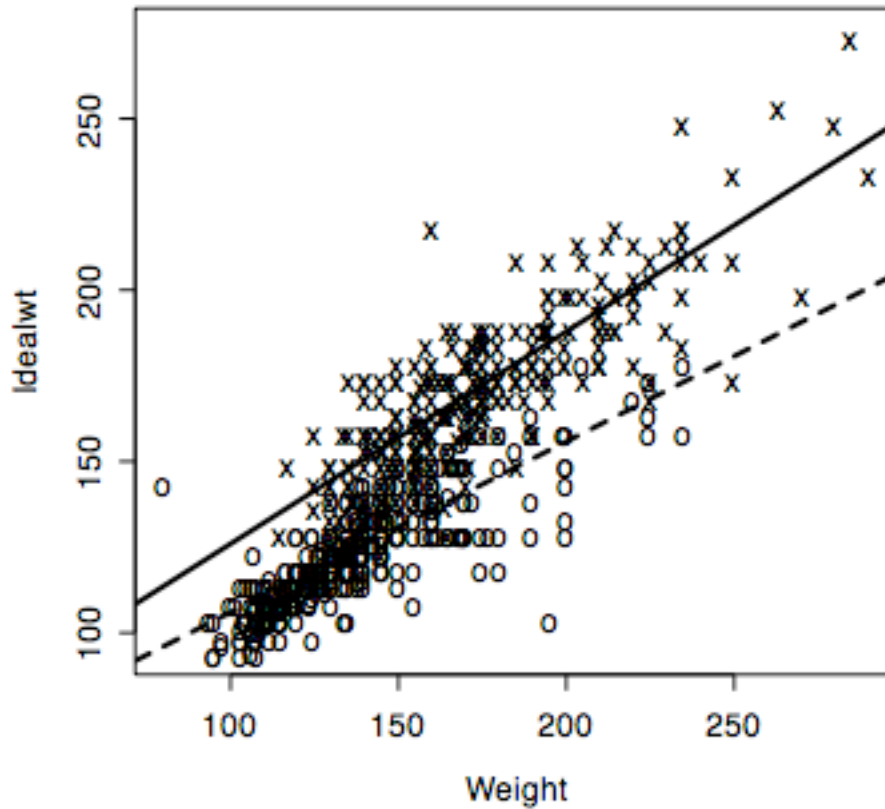
```

Part 1. The scatterplot shown below was created with the following commands:

```

plot(Weight, Idealwt, type="n")
text(Weight, Idealwt, p)
abline(Mmodel,lwd=2)
abline(Fmodel,lwd=2,lty=2)

```



Part 2. Here is the regression output for the females:

```
> summary(Fmodel)
```

Call:

```
lm(formula = Idealwt ~ Weight, subset = Sex == "Female")
```

Residuals:

	Min	1Q	Median	3Q	Max
	-48.3081	-4.9518	-0.4126	4.0986	48.9772

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	56.17217	2.47661	22.68	<2e-16 ***
Weight	0.49813	0.01788	27.85	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.651 on 381 degrees of freedom
 Multiple R-Squared: 0.6706, Adjusted R-squared: 0.6698
 F-statistic: 775.8 on 1 and 381 DF, p-value: < 2.2e-16

2(a): The predicted increase in Idealwt for each increase of 1 in the explanatory variable is given by the estimated slope, which is 0.498 pounds.

2(b): The intercept has no useful interpretation because it gives the predicted ideal weight for a woman whose actual weight is zero pounds, which is not possible.

2(c): The value of r-squared, or 0.6706 in this case, is the proportion of the variation in Idealwt explained by its linear association with Weight.

2(d): The correlation coefficient, r , is the square root of 0.6706, which is 0.819. We know it is positive because r always has the same sign as the estimated slope, which is 0.498 in this case.

2(e): The t-statistic for testing a linear relationship between Idealwt and Weight equals 27.85 on 381 degrees of freedom (this is shown in the same row as the Weight coefficient). The degrees of freedom equal $n-2$, where n is the number of observations.

Part 3: The Q-Q normal plots in this question may be produced using the R commands

```
qqnorm(Mmodel$res, main="Normal Q-Q plot for males")
qqnorm(Fmodel$res, main="Normal Q-Q plot for females")
```

These plots test the assumption of normally distributed errors. The female plot shows that there are two extreme residuals, one very positive and one very negative. However, the overall straight-line patterns of these two plots indicates that the assumption of normally distributed errors is a reasonable assumption.