

Project 11 - Solutions

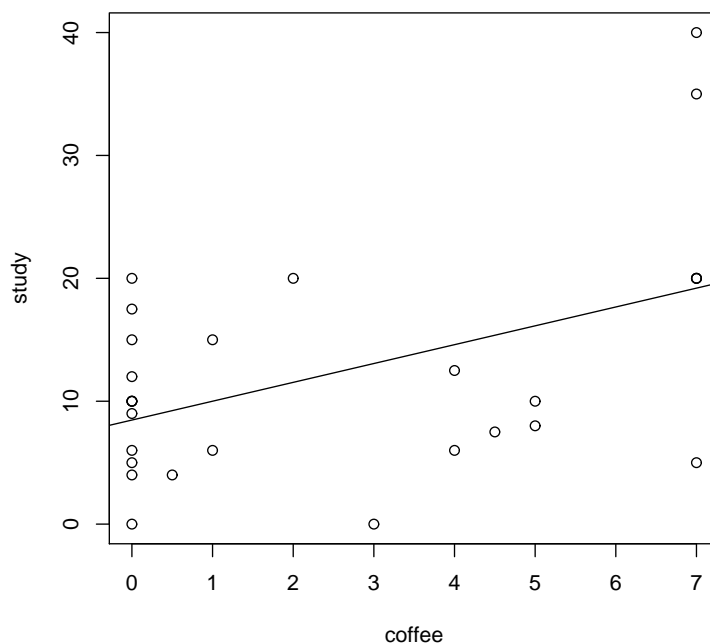
Statistics 401: Fall 2006

- (Problem 5.4 on page 194) Cause and effect conclusions such as “increasing alcohol consumption will increase income” are valid only if in experiments where the treatments are randomly assigned. Light, moderate and heavy alcohol drinking is not a treatment that the researchers randomly applied to individuals. Rather, this was an observational study. A possible confounding variable is “personal motivation.” People who tend to drink moderate or heavy amounts of alcohol may tend to also be more compulsive when it comes to work than light drinkers.
- (Problem 5.18 on page 206)
 - The y -intercept is $b_0 = -147$. The slope is $b_1 = 6.175$. Thus, for every additional centimeter of snout length, the clutch size increases on average by 6.175 salamander eggs.
 - For a salamander whose snout size is 22cm, one should be reluctant to use the least-squares regression line $\hat{y} = -147 + 6.175x$ to predict clutch size since 22cm is outside the range of snout lengths of 30 to 70cm used to calculate the line. **EXTRAPOLATION!**

Now, to study the relationship between the number of hours that a large university’s students spend each week studying, and three different explanatory variables: the number of hours spent watching TV; fastfood consumption; and coffee consumption.

- Figure 1 contains a scatterplot of the number of hours spent studying each week versus weekly coffee consumption.

Figure 1: A scatterplot of number of hours studied per week versus weekly coffee consumption



4. The sample correlation coefficient between hours studied and coffee consumption is $r = .475$. The parameter being estimated is the population correlation coefficient ρ . See the Appendix for the R code.
5. The form of the relationship appears to be linear (see Figure 1). The direction is positive (since $r > 0$). The strength of the relationship is moderate (r is not very close to 1).
6. The SLR model is $y = \beta_0 + \beta_1 x + \epsilon$ where y is number of hours spent studying per week, x is cups of coffee per week, and $\epsilon \sim N(0, \sigma)$.
7. Using R, an SLR model was fit to describe the number of hours spent studying each week as a function of coffee consumption. The least-squares regression line is $\hat{y} = 8.4667 + 1.5338x$.
8. The least-square regression estimates and t-tests for the intercept and slope are given in Table 1.

Table 1: estimates and t-tests for the intercept and slope

	Estimate	<i>SE</i>	<i>t</i>	<i>p</i> -value
(Intercept)	8.4667	2.1363	3.963	0.000514
coffee	1.5338	0.5572	2.752	0.010640

9. Table 2 gives the ANOVA table for this SLR.

Table 2: ANOVA Table for the SLR for hours studying and coffee consumption

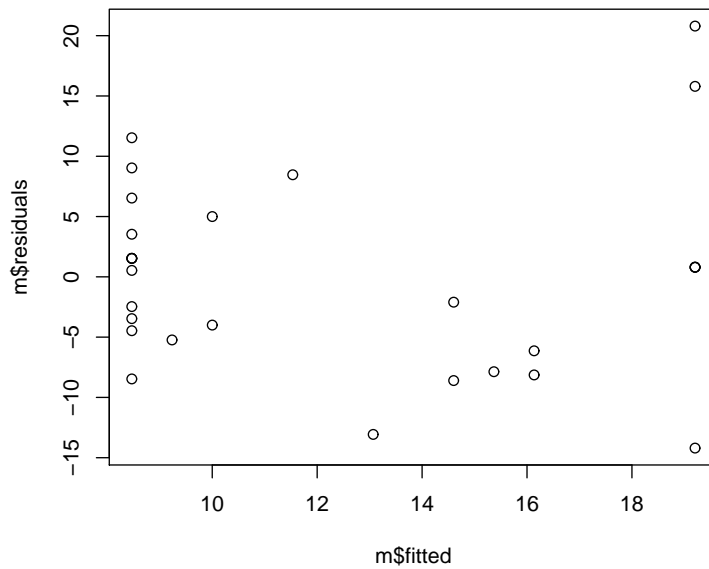
Source	DF	SS	MS	<i>F</i>	<i>p</i> -value
Model	1	532.49	532.49	7.5756	0.01064
Residual	26	1827.54	70.29		
Total	27	2360.03			

10. An unbiased estimate for σ^2 is $MSE = 70.29$. In terms of the problem, this means that the constant variance of the residuals is 70.29.
11. $r^2 = .475^2 = 0.2256$. This means that 22.6% of the variability of the hours studied is explained by the linear relationship with coffee consumption.
12. To determine if there is a significant linear relationship between the number of hours spent studying each week and coffee consumption:
 - (a) State the hypotheses.

$$H_0 : \beta_1 = 0$$

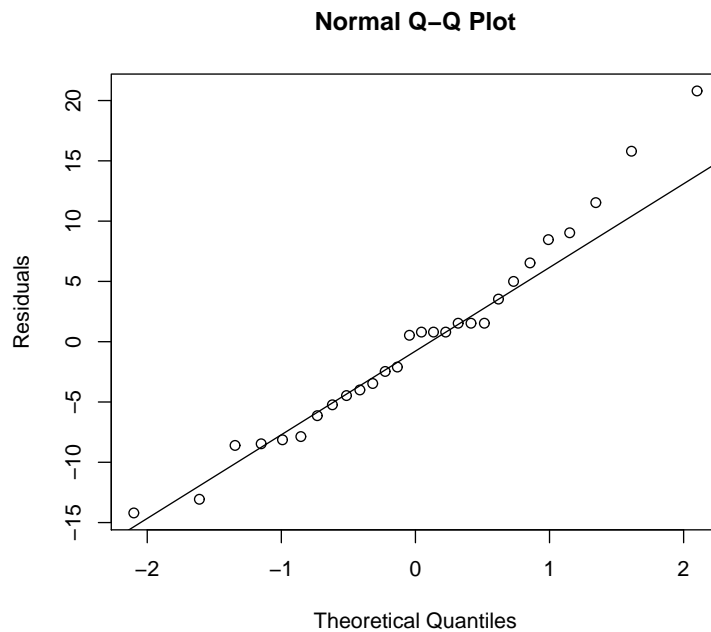
$$H_a : \beta_1 \neq 0$$
 - (b) Check the Assumptions.
 - i. Figure 2 gives a plot of the residuals versus fitted values. There is a pattern to the residuals - they go from being mostly positive, to negative, to mostly positive. This may indicate that there is not a linear relationship between study time and coffee consumption. Furthermore, the larger spread of the residuals in the right of the plot indicates that the constant variance assumption may not be met.

Figure 2: residuals versus fitted values



ii. Figure 3 gives a normal probability plot of the residuals. The strong linearity of the residuals fails to indicate that the residuals are not normally distributed.

Figure 3: normal probability plot of the residuals



- (c) The test statistics are $t = 2.752$ and $F = 7.576$.
- (d) When H_0 is true, $t \sim t(26)$ and $F \sim F(1, 26)$. The p -value is $2P(T > 2.752) = P(F > 7.576) = 0.01064$.
- (e) Since the p -value $= 0.01064 < .05$, then reject H_0 .
- (f) The evidence suggests that there is a useful linear relationship between coffee consumption and hours studied.

13. A 95% confidence interval for the slope of the population regression line is $b_1 \pm t_{.975, 26} SE_{b_1} =$

$1.5338 \pm 2.06(0.5572) = (0.39, 2.68)$. Thus, we are 95% confident that for each additional cup of coffee consumed per week, the average studying time increases by between .39 and 2.68 hours.

14. The least squares regression line predicts that for all students who drink 6 cups of coffee a week, they study an average of $8.4667 + 1.5338(6) = 17.7$ hours per week.
15. A 95% CI for the mean number of hours studied by students who drink 6 cups of coffee a week is (12.57, 22.77). Thus, we are 95% confident that the mean number of hours studied by students who drink 6 cups of coffee a week is between 12.57 and 22.77 hours a week.
16. The least squares regression line predicts that a student who drinks 6 cups of coffee a week studies $8.4667 + 1.5338(6) = 17.7$ hours per week.
17. A 95% PI for the number of hours studied by a student who drinks 6 cups of coffee a week is (-0.30, 35.64). Thus, we are 95% confident that the number of hours studied by a student who drinks 6 cups of coffee a week is between 0 and 35.64 hours a week.
18. An SLR model was fit for (1) hours studied versus TV and (2) hours studied versus fastfood.
 - (a) In both of these models, the predictor is not significant. There is not a significant linear relationship between studying and TV watching since the p -value = 0.202. There is not a significant linear relationship between studying and fastfood consumption since the p -value = 0.125.
 - (b) The r^2 value for studying and TV is $r^2 = 0.062$. The r^2 value for studying and fastfood is $r^2 = 0.088$. Both of these values indicate that predictors do not explain very much of the variability of the the number of hours studied. This is not surprising since neither predictors is significant.
 - (c) Based on (a) and (b) above, coffee is the best single predictor of the number of hours studied. It is the only significant predictor of the three, and the $r^2 = 0.2256$ is the largest.

Appendix

```
> D=read.table("project11study.txt",header=T)
```

```
> D
```

```
      TV fastfood coffee study
1  2.5      2.5    0.0  17.5
2 20.0      2.0    4.0  12.5
3  1.5      0.5    4.0   6.0
4  2.0      0.0    0.0  15.0
5  2.0      0.5    7.0  35.0
6  3.0      1.0    0.0  12.0
7  1.5      0.0    7.0  20.0
8  8.0      0.0    0.0  20.0
9  0.0      0.0    7.0  40.0
10 0.0      0.0    7.0  20.0
11 4.0      0.5    0.0   0.0
12 3.5      0.0    2.0  20.0
13 7.0      0.0    0.0   5.0
14 5.0      0.0    7.0  20.0
15 25.0     0.5    0.5   4.0
16 7.5      1.5    4.5   7.5
17 2.5      1.5    0.0   9.0
18 15.0     3.0    0.0  10.0
19 2.0      1.0    3.0   0.0
20 10.0     3.0    1.0  15.0
21 6.5      5.0    0.0  10.0
22 1.5      1.0    5.0   8.0
23 10.0     2.0    0.0   4.0
24 0.0      5.0    0.0   6.0
25 2.0      2.0    7.0   5.0
26 9.0      2.0    5.0  10.0
27 1.0      0.0    0.0  10.0
28 5.0     10.0    1.0   6.0
```

```
> attach(D)
```

```
> # Problem 3
```

```
> plot(coffee,study)
```

```
> # Problem 4
```

```
> cor(coffee,study)
```

```
[1] 0.475041
```

```
> # Problem 7
```

```
> m=lm(study ~ coffee)
```

```
> abline(m)
```

```
> summary(m)
```

```
Call:
```

```
lm(formula = study ~ coffee)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-14.2031	-5.4591	0.6651	3.8998	20.7969

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.4667	2.1363	3.963	0.000514 ***
coffee	1.5338	0.5572	2.752	0.010640 *

```
---
```

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

```
Residual standard error: 8.384 on 26 degrees of freedom
```

```
Multiple R-Squared: 0.2256, Adjusted R-squared: 0.1958
```

```
F-statistic: 7.576 on 1 and 26 DF, p-value: 0.01064
```

```
> # Problem 8
```

```
> anova(m)
```

```
Analysis of Variance Table
```

```
Response: study
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
coffee	1	532.49	532.49	7.5756	0.01064 *
Residuals	26	1827.54	70.29		

```
---
```

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

```
> # Problem 12
```

```
> plot(m$fitted,m$residuals)
```

```
> qqnorm(m$residuals,ylab="Residuals")
```

```
> qqline(m$residuals)
```

```
> # Problem 13
```

```
> 1.5338+c(-1,1)* 2.06*(0.5572)
```

```
[1] 0.385968 2.681632
```

```
> # Problem 14 and 16
```

```
> 8.4667 + 1.5338*6
```

```
[1] 17.6695
```

```
> # Problem 15
```

```
> xstar = data.frame(coffee = 6)
```

```
> mean.xstar = predict(m,xstar,interval="confidence",level=0.95)
```

```
> mean.xstar
```

```
      fit      lwr      upr
[1,] 17.66934 12.56739 22.77128
>
```

```
> # Problem 17
> newy.xstar = predict(m,xstar,interval="prediction",level=0.95)
> newy.xstar
      fit      lwr      upr
[1,] 17.66934 -0.3033855 35.64206
```

```
> # Problem 18
> m2=lm(study ~ TV)
> m3=lm(study ~ fastfood)
> summary(m2)
```

```
Call:
lm(formula = study ~ TV)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-13.798  -6.664  -1.260   5.481  25.433
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.5666     2.3996   6.070 2.05e-06 ***
TV            -0.3845     0.2939  -1.308  0.202
---
```

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

```
Residual standard error: 9.229 on 26 degrees of freedom
Multiple R-Squared:  0.06174,    Adjusted R-squared:  0.02566
F-statistic: 1.711 on 1 and 26 DF,  p-value: 0.2023
```

```
> summary(m3)
```

```
Call:
lm(formula = study ~ fastfood)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-13.800  -5.593  -0.887   5.562  25.562
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.438     2.142   6.740 3.75e-07 ***
fastfood      -1.275     0.804  -1.586  0.125
---
```

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

Residual standard error: 9.097 on 26 degrees of freedom
Multiple R-Squared: 0.08823, Adjusted R-squared: 0.05317
F-statistic: 2.516 on 1 and 26 DF, p-value: 0.1248