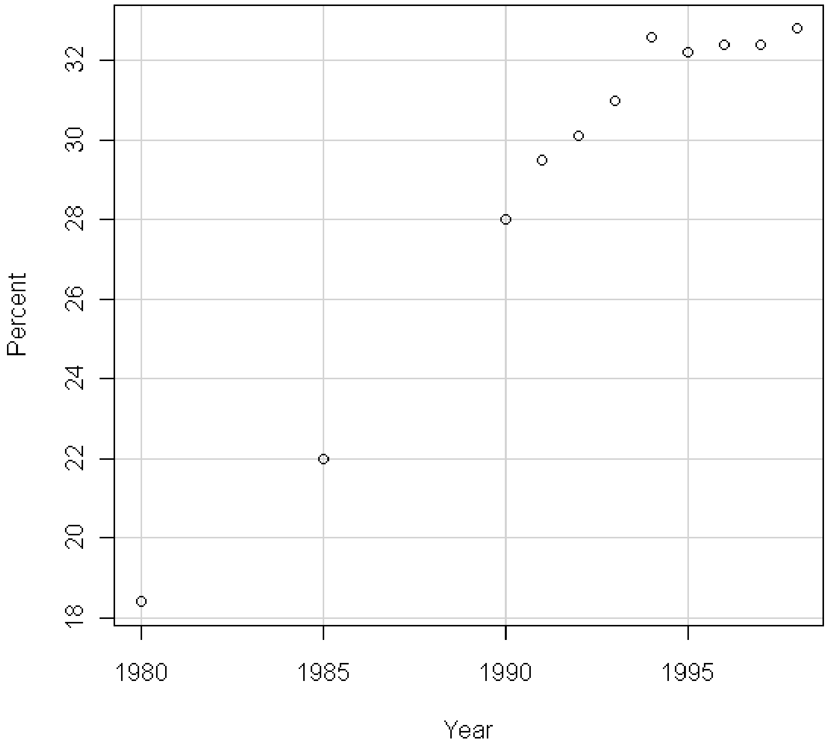
Math 130

Homework 9 Solutions

Assignment
Chapter 9~~: 9.30, 9.32~~
Chapter 10: ~~10.16~~, 10.26, 10.32
Chapter 27: 27.2, 27.6, 27.14, 27.16, 27.28

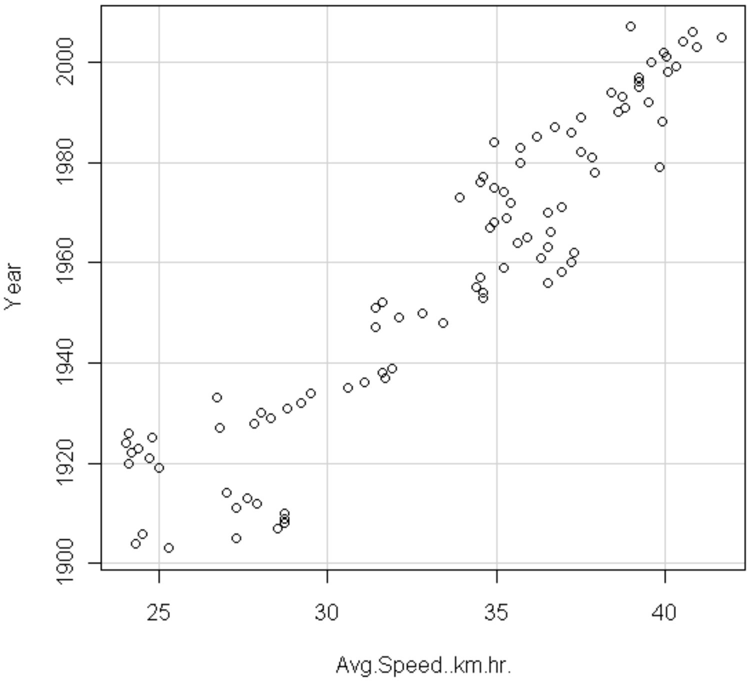
~~Chapter 9~~

~~**9.30] Unwed Births.** Answers here can vary. You might decide the pattern is fairly linear, and fit a linear regression. You might decide there is a slight curvature, and try re-expressing the data and *then* using a linear regression. You might also decide there seems to be *two* different lines here, and fit one regression for the data less than 1994 and another for the data after 1994. Just be sure to justify what you've done.~~



~~**9.32] Tour de France, 2007.**~~

~~**a)** The association between average speed and year is positive, moderate, but not quite linear. Generally, average speed of the winner has been increasing over time. There are several periods where the relationship is curved (or squiggly), but since 1950, the relationship has been much more linear. There are no races between 1915 and 1918 or between 1940 and 1947, probably because of the two World Wars in Europe at the times.~~

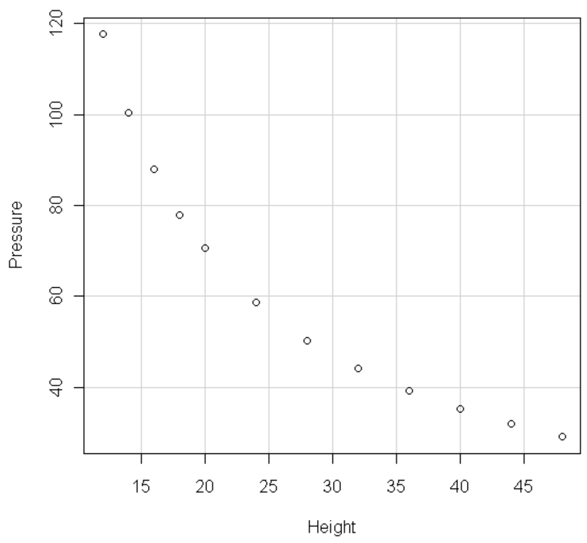


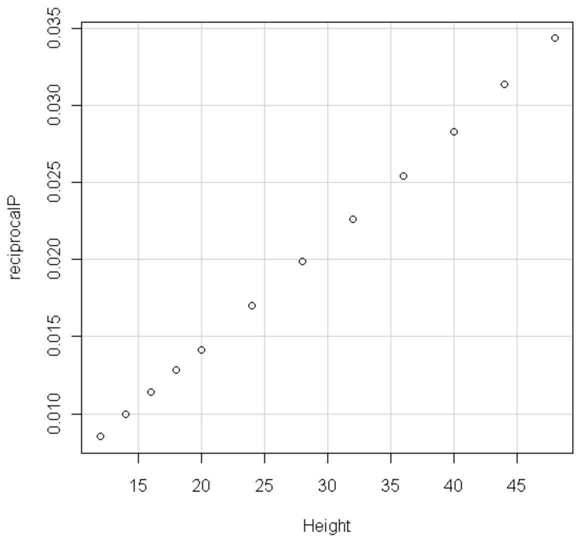
~~**b)** The regression equation, based on Rcmdr, is $\widehat{\text{Avg Speed}} = -272.7 + 0.1564\ (\text{Year})$~~

~~**c)** The conditions for regression are not met. Although the variables are quantitiative, and there are no outliers, the relationship is not straight enough in the early part of the 20th century to fit a regression line.~~

Chapter 10

10.16] Pressure. The scatterplot of the data is below. It shows a strong, curved, negative association between the height of the cylinder and the pressure inside. Because of the curved nature of the association, a linear model is not appropriate. We can re-express the data to try to make it linear. I tried the reciprocal, $\frac{1}{y}$. The resulting line is very straight.





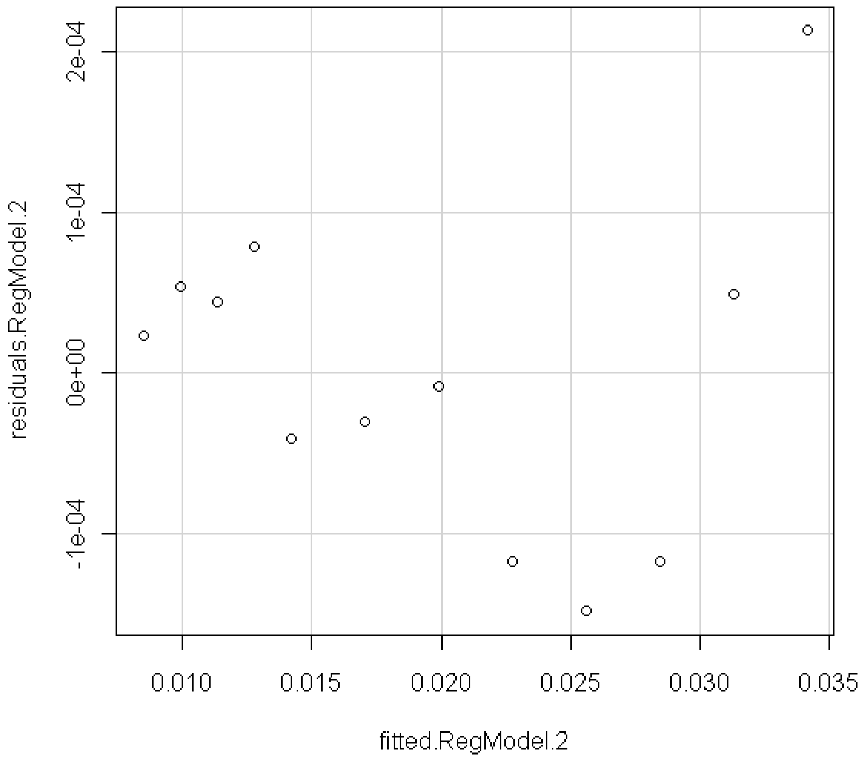
```
Call:
lm(formula = reciprocalP ~ Height, data = Boyles)

Residuals:
       Min         1Q     Median         3Q        Max 
-1.486e-04 -5.964e-05  7.321e-06  5.059e-05  2.135e-04 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.670e-05  7.813e-05  -0.982    0.349    
Height       7.131e-04  2.600e-06 274.301   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0001057 on 10 degrees of freedom
Multiple R-squared: 0.9999,	Adjusted R-squared: 0.9999 
F-statistic: 7.524e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

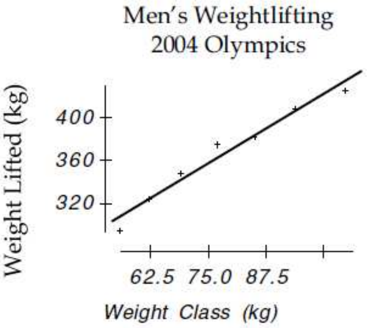
The fitted regression line is: $\frac{1}{\widehat{Pressure}} = -0.00007670 + 0.0007131$. The residual plot looks okay, and the R2 is very high. I think this new model is doing a great job.


residuals.RegModel.2
2e-04
1e-04
0e+00
-1e-04
0.010
0.015
0.020
0.025
0.030
0.035
fitted.RegModel.2

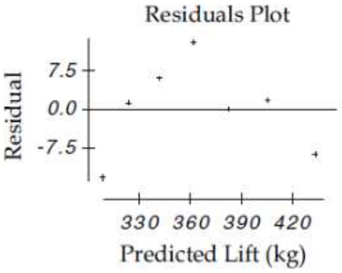
10.26] Weightlifting, 2004.

The association between weight class and weight lifted for gold medal winners in weightlifting at the 2004 Olympics is strong, positive, and curved. The linear model that best fits the data is $\hat{Lift} = 164.97 + 2.56(WeightClass)$.

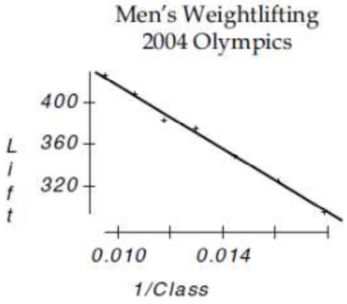
Although this model accounts for 96.3% of the variability in weight lifted, it does not fit the data well.

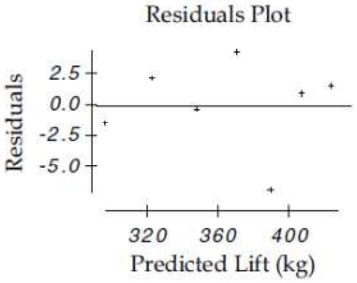


b) The residuals plot for the linear model shows a curved pattern, indicating that the linear model has failed to model the association well. A re-expressed model or a curved model might fit the association between weight class and weight lifted better than the linear model.



c) Re-expressing Weight Class using the reciprocal produces a scatterplot that is much straighter. $\hat{Lift} = 568.727 - 15243\left(\frac{1}{Class}\right)$.





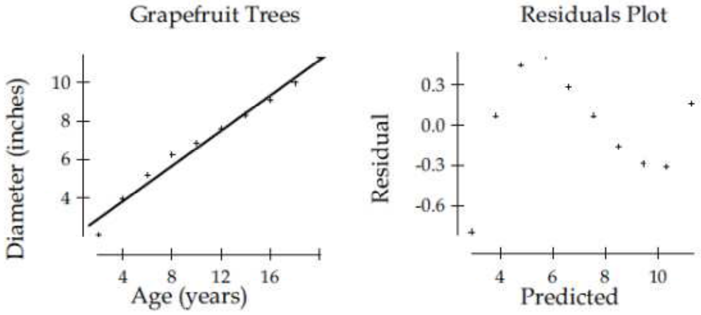
d) The reciprocal model is a better model, since the residuals plot shows little pattern. Additionally, the model accounts for 99.4% of the variability in weight lifted.

e) George Asanidze's lift had a large, negative residual. This means that he lifted less than expected.

10.32] Tree Growth

Tree growth.

a) The association between age and average diameter of grapefruit trees is strong, curved, and positive. Generally, older trees have larger average diameters.



The linear model for this association, $\mathit{Average\hat{D}iameter} = 1.973 + 0.463(\mathit{Age})$ is not appropriate. The residuals plot shows a clear pattern.

Because of the change in curvature in the association, these data cannot be straightened by re-expression.

b) If diameters from individual trees were given, instead of averages, the association would have been weaker. Individual observations are more variable than averages.

Chapter 27

27.2]

Drug use.

a) The equation of the line of best fit for these data points is $\%\mathit{Oth\hat{e}rDrugs} = -3.068 + 0.615(\%\mathit{Marijuana})$. According to the linear model, the percentage of ninth graders in these countries who use other drugs increases by about 0.615% for each additional 1% of ninth graders who use marijuana.

b) H0: There is no linear relationship between marijuana use and use of other drugs. $(\beta_1 = 0)$

HA: There is a linear relationship between marijuana use and use of other drugs. $(\beta_1 \neq 0)$

c) Assuming the conditions have been met, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (11 – 2) = 9 degrees of freedom. We will use a regression slope *t*-test.

The value of *t* = 7.85. The *P*-value of 0.0001 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence that the percentage of ninth graders who use other drugs is related to the percentage of ninth graders who use marijuana. Countries with a high percentage of ninth graders using marijuana tend to have a high percentage of ninth graders using other drugs.

d) 87.3% of the variation in the percentage of ninth graders using other drugs can be accounted for by the percentage of ninth graders using marijuana.

e) The use of other drugs is associated with marijuana use, but there is no proof of a cause-and-effect relationship between the two variables. There may be lurking variables present.

27.6]

Second home.

a) **Straight enough condition:** The scatterplot is straight enough, and the residuals plot looks unpatterned.
Randomization condition: The houses were selected at random.
Does the plot thicken? condition: The residuals plot shows no obvious trends in the spread.
Nearly Normal condition, Outlier condition: The histogram of residuals is unimodal and symmetric, and shows no outliers.

b) Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (1064 – 2) = 1062 degrees of freedom.

$$b_1 \pm t^*_{n-2} \times SE\left(b_1\right) = 94.4539 \pm (t^*_{1062}) \times 2.393 \approx (89.8, 99.2)$$

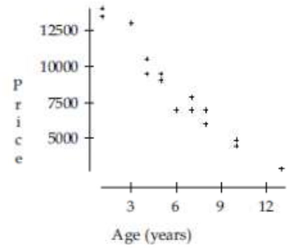
We are 95% confident that Saratoga housing costs increase at a rate of between \$89.8 and \$99.2 per square foot.

27.14]

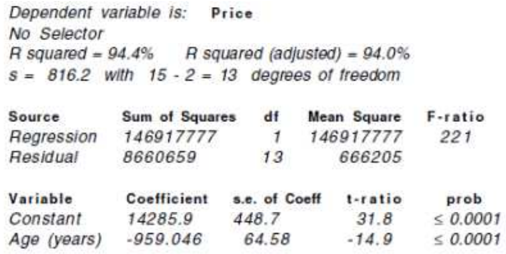
Used cars 2007.

a) A scatterplot of the used cars data is at the right.

b) A linear model is probably appropriate. The plot appears to be linear.



c)



```
Dependent variable is:   Price
No Selector
R squared = 94.4%     R squared (adjusted) = 94.0%
s = 816.2  with  15 - 2 = 13  degrees of freedom

Source       Sum of Squares    df    Mean Square    F-ratio
Regression   146917777          1    146917777      221
Residual     8660659           13       666205

Variable      Coefficient   s.e. of Coeff   t-ratio    prob
Constant      14285.9       448.7             31.8     ≤ 0.0001
Age (years)   -959.046       64.58           -14.9     ≤ 0.0001
```

The equation of the regression line is: $\hat{Price} = 14286 - 959(Age)$.

According to the model, the average asking price for a used Toyota Corolla decreases by about $959 dollars for each additional year in age. Let's take a closer look.

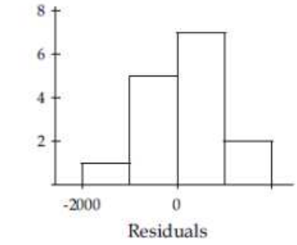
d) **Straight enough condition:** The scatterplot is straight enough to try a linear model.

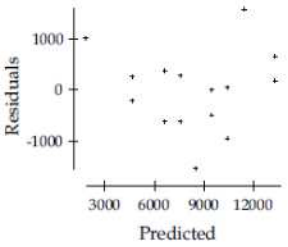
Independence assumption: Prices of Toyota Corollas of different ages might be related, but the residuals plot looks fairly scattered. (The fact that there are several prices for some years draws our eyes to some patterns that may not exist.)

Does the plot thicken? condition: The residuals plot shows no obvious patterns in the spread.

Nearly Normal condition, Outlier condition: The histogram is reasonably unimodal and symmetric, and shows no obvious skewness or outliers.

Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (17 – 2) = 15 degrees of freedom.





27.16]

Used cars, again.

$$b_1 \pm t^*_{n-2} \times SE(b_1) = -959 \pm (2.160) \times 64.58 \approx (-1099, -819.5)$$

We are 95% confident that the advertised price of a used Toyota Corolla is decreasing by an average of between $819.50 and $1099 for each additional year in age.

27.28]

El Niño.

a) The regression equation is $\widehat{Temp} = 15.3066 + 0.004(CO_2)$, with CO_2 concentration measured in parts per million from the top of Mauna Loa in Hawaii, and temperature in degrees Celsius.

b) H$_0$: There is no linear relationship between temperature and CO_2 concentration. $(\beta_1 = 0)$
H$_A$: There is a linear relationship between temperature and CO_2 concentration. $(\beta_1 \neq 0)$

Since the scatterplots and residuals plots showed that the data were appropriate for inference, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (37 – 2) = 35 degrees of freedom. We will use a regression slope *t*-test.

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t = \frac{0.004 - 0}{0.0009}$$

$$t \approx 4.44$$

The value of $t \approx$ 4.44. The *P*-value (two-sided!) of about 0.00008 means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a linear relationship between CO_2 concentration and temperature. Years with higher CO_2 concentration tend to be warmer, on average.

c) Since $R^2 = 33.4\%$, only 33.4% of the variability in temperature can be accounted for by the CO_2 concentration. Although there is strong evidence of a linear association, it is weak. Predictions would tend to be very imprecise.