Supplemental Examples (Chapter 15)

Online S520

1. A Major League Baseball team plays 162 games each season. There are 30 teams. Each season, the number of wins by Major League Baseball teams has an approximately normal distribution with mean 81 and standard deviation 11.7. The correlation between a team's wins one season and their wins the next season is 0.54.

   (a) Suppose a baseball executive believed the best prediction of a team's wins in 2015 should be equal to their wins in 2014. For example, he predicts that the Los Angeles Angels, who had the most wins in 2014 with 98, would have 98 wins in 2015. Using the data given, explain to the baseball executive (who knows very little statistics) why this particular prediction is likely too high.

   (b) Use regression to predict the Los Angeles Angels' 2015 wins using only the above data.

   (c) The executive looks at the regression predictions for all MLB players and sees that no team is predicted to win more than 91 games. The executive suspects the predictions are too low, because in every full season since 1961, at least one team has won at least 96 games. Explain to the executive, who knows very little statistics, why his suspicions are misplaced.

   Solution:

   (a) Regression to the mean means that we predict the best teams will do worse than they did the previous year.

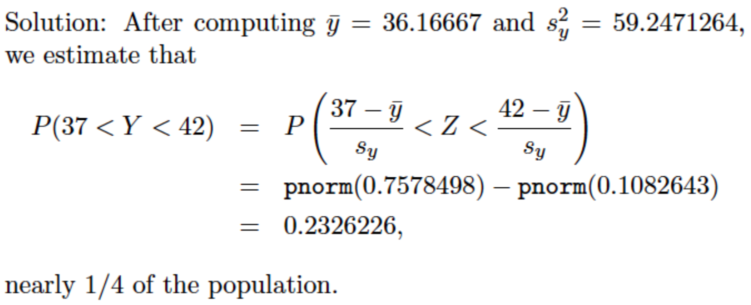
   (b) The sample statistics from the data are: $r = 0.54, \bar{x} = \bar{y} = 81, s_x = s_y = 11.7$.
   The slope of the regression line is: $b = r\frac{s_y}{s_x} = r = 0.54$.
   The intercept of the regression line is: $a = \bar{y} - b\bar{x} = 81 - 0.54 * 81 = 37.26$.
   Predicted 2015 wins $= 0.54 \times$ (2014 wins) $+ 37.26 = 90.6$ wins.

   (c) While it's likely that some team will win 96 games or more, there's an element of luck involved. Some teams will be better than their projection. We can't tell in advance which team will be the luckiest.

2. Trosset chapter 15.7 exercise 6

   (a) There appear to be two distinct clusters in these data, roughly delineated by $y < 40$ and $y > 40$.

   (b) (Trosset solutions:)

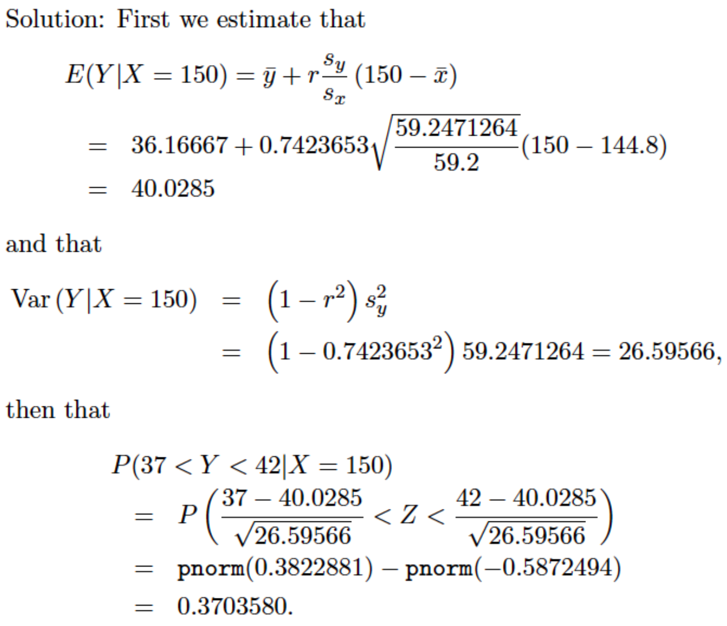


Solution: After computing $\bar{y} = 36.16667$ and $s_y^2 = 59.2471264$, we estimate that

$$\begin{aligned} P(37 < Y < 42) &= P\left(\frac{37-\bar{y}}{s_y} < Z < \frac{42-\bar{y}}{s_y}\right) \\ &= \texttt{pnorm}(0.7578498) - \texttt{pnorm}(0.1082643) \\ &= 0.2326226, \end{aligned}$$

nearly 1/4 of the population.

(c) (Trosset solutions:)



Solution: First we estimate that

$$\begin{aligned} E(Y|X=150) &= \bar{y} + r\frac{s_y}{s_x}(150-\bar{x}) \\ &= 36.16667 + 0.7423653\sqrt{\frac{59.2471264}{59.2}}(150-144.8) \\ &= 40.0285 \end{aligned}$$

and that

$$\begin{aligned} \mathrm{Var}\,(Y|X=150) &= \left(1-r^2\right)s_y^2 \\ &= \left(1-0.7423653^2\right)59.2471264 = 26.59566, \end{aligned}$$

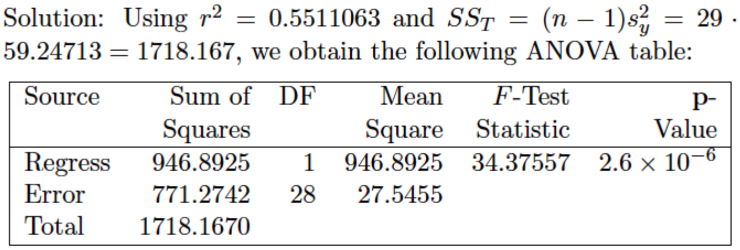
then that

$$\begin{aligned} &P(37 < Y < 42|X=150) \\ &= P\left(\frac{37-40.0285}{\sqrt{26.59566}} < Z < \frac{42-40.0285}{\sqrt{26.59566}}\right) \\ &= \texttt{pnorm}(0.3822881) - \texttt{pnorm}(-0.5872494) \\ &= 0.3703580. \end{aligned}$$

3. Trosset chapter 15.7 exercise 7

(a) $r^2 = 0.5511063$.

(b) (Trosset solutions:)



Solution: Using $r^2 = 0.5511063$ and $SS_T = (n-1)s_y^2 = 29 \cdot 59.24713 = 1718.167$, we obtain the following ANOVA table:

Source	Sum of Squares	DF	Mean Square	F-Test Statistic	p-Value
Regress	946.8925	1	946.8925	34.37557	2.6×10^{-6}
Error	771.2742	28	27.5455		
Total	1718.1670				

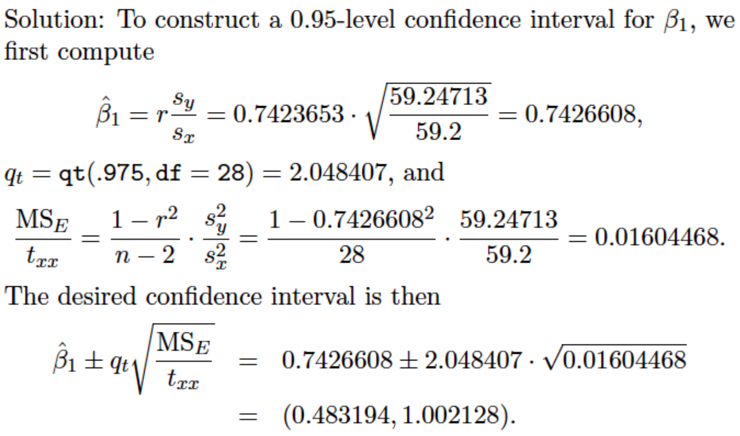
Or equivalently use the t-test:

$$\begin{aligned}
\hat{\beta}_1 &= r\frac{s_y}{s_x} = 0.7423653\sqrt{\frac{59.24713}{59.2}} = 0.7426608 \\
MS_E/t_{xx} &= \frac{1-r^2}{n-2}\cdot\frac{s_y^2}{s_x^2} = 0.01604468 \\
t &= \frac{\hat{\beta}_1}{\sqrt{MS_E/t_{xx}}} = \frac{0.7426608}{\sqrt{0.01604468}} = 5.863069
\end{aligned}$$

```
# Two-sided P-value
2 * (1 - pt(abs(t.stat), df = n-2))
```

p-value = 2.646969e-06.
Because $p < \alpha$, we reject H_0 and conclude that there is convincing evidence that knowing x helps one predict y.

(c) (Trosset solutions:)



Solution: To construct a 0.95-level confidence interval for β_1, we first compute

$$\hat{\beta}_1 = r\frac{s_y}{s_x} = 0.7423653\cdot\sqrt{\frac{59.24713}{59.2}} = 0.7426608,$$

$q_t = \texttt{qt}(.975, \texttt{df} = 28) = 2.048407$, and

$$\frac{\mathrm{MS}_E}{t_{xx}} = \frac{1-r^2}{n-2}\cdot\frac{s_y^2}{s_x^2} = \frac{1-0.7426608^2}{28}\cdot\frac{59.24713}{59.2} = 0.01604468.$$

The desired confidence interval is then

$$\begin{aligned}
\hat{\beta}_1 \pm q_t\sqrt{\frac{\mathrm{MS}_E}{t_{xx}}} &= 0.7426608 \pm 2.048407\cdot\sqrt{0.01604468} \\
&= (0.483194, 1.002128).
\end{aligned}$$