Day6

Olivia Wu

2024-02-27

Problem 2.9

a) All the r^2 tells us is how much variability of the response variable is explained by the model. The model could have curvature and still have a high r^2, so it does not indicate the linear relationship is the best model.

b) A low r^2 would indicate that the linear relationship is not the best model. Linear data with high variability and error can still produce a low r^2

Problem 2.10

a) Width decreases ($\frac{1}{n}$ decreases)

b) width decreases ($\frac{1}{\sum(x-\bar{x})^2}$ decreases)

c) width increases (σ_ϵ increases)

d) width increases ($x^* - \bar{x}$ increases)

Problem 2.23

a) The r^2 is 0.9853. 98.53% of the variaility in postal rates is explained by the model

$$\hat{Price} = -1,647 + 0.841(Year)$$

```
##
## Call:
## lm(formula = Price ~ Year, data = USstamp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9232 -0.9478  0.1195  1.1899  4.5325
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.647e+03  4.686e+01  -35.15   <2e-16 ***
## Year         8.410e-01  2.357e-02   35.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.737 on 19 degrees of freedom
## Multiple R-squared:  0.9853, Adjusted R-squared:  0.9845
## F-statistic:  1273 on 1 and 19 DF,  p-value: < 2.2e-16
```

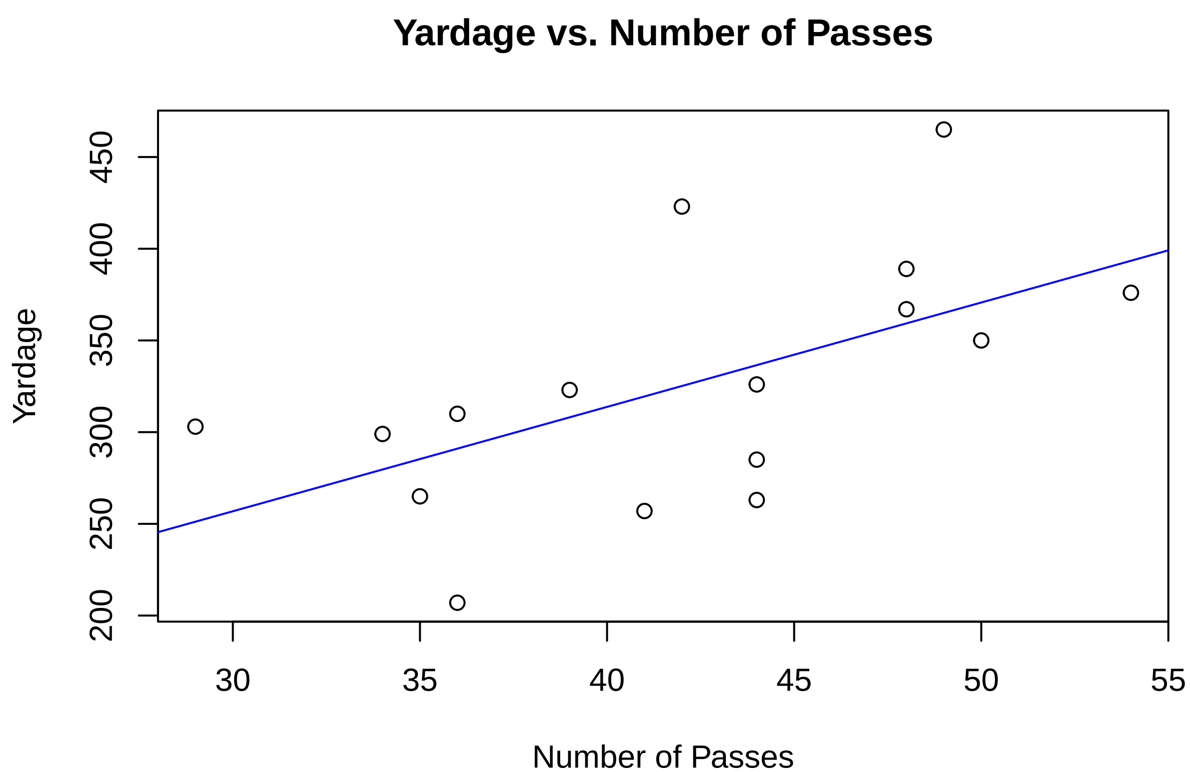
b) The p-value for our slope is less than 0.05, so there is a significant linear relationship between postal rates and year.

c) The F-statistic is 1273.1, and it has a p-value less than 0.05. This shows that $Year$ is an effective predictor of $Price$.

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Year       1 3841.2  3841.2  1273.1 < 2.2e-16 ***
## Residuals 19   57.3     3.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 2.27

a) $\hat{Yards} = 86.140 + 5.691(Attempts)$




```
##
```

```
## Call:
## lm(formula = Yards ~ Attempts, data = Brees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.001 -28.383  -1.407  21.963 100.022
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   86.140     90.362   0.953   0.3566
## Attempts       5.691      2.122   2.682   0.0179 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.34 on 14 degrees of freedom
## Multiple R-squared:  0.3394, Adjusted R-squared:  0.2922
## F-statistic: 7.191 on 1 and 14 DF,  p-value: 0.01789
```

b) No; the y-intercept is not 0.

c) $r^2 = 0.3394$, so 33.94% of the variability in Brees's yardage per game is explained by knowing how many passes he threw.

Problem 2.31

a) The p-value of the slope coefficient is $0.000002 < 0.05$, so there is a significant linear relationship between the initial height fo the pine seedlings in 1990 and the height in 1997.

```
##
## Call:
## lm(formula = Hgt97 ~ Hgt90, data = Pines)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -261.886  -44.343    7.308   55.114  196.114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  307.439      9.841  31.239  < 2e-16 ***
## Hgt90          2.322      0.492   4.721 2.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.79 on 807 degrees of freedom
##   (191 observations deleted due to missingness)
## Multiple R-squared:  0.02687,    Adjusted R-squared:  0.02567
## F-statistic: 22.28 on 1 and 807 DF,  p-value: 2.772e-06
```

b) $r^2 = 0.02687$, so 2.69% of the variation of the height in 1997 is explained by the model.

c) Tabel shown below

```r
aov <- anova(model)
aov
```

```
## Analysis of Variance Table
##
## Response: Hgt97
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Hgt90       1  138344  138344  22.284 2.772e-06 ***
## Residuals 807 5010010    6208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) By finding $\frac{SSModel}{SSTotal}$, we get the same value for $r^2 = 0.02687$

```r
SSModel <- aov$`Sum Sq`[1]
SSTotal <- sum(aov$`Sum Sq`)
rsq <- SSModel/SSTotal
rsq
```

```
## [1] 0.02687153
```

e) The coefficient of determination is extremely low, and I am not happy with this linear model.

Problem 2.33

```
##
## Call:
## lm(formula = Hgt97 ~ Hgt96, data = Pines)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.455 -12.120   1.201  13.913  45.648
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.590784   2.524484   16.08   <2e-16 ***
## Hgt96        1.096059   0.008734  125.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.47 on 852 degrees of freedom
##   (146 observations deleted due to missingness)
## Multiple R-squared:  0.9487, Adjusted R-squared:  0.9486
## F-statistic: 1.575e+04 on 1 and 852 DF,  p-value: < 2.2e-16
```

a) $t^* = 1.963$

$\hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1}$

$= 1.096 \pm 1.963(0.0087)$

$= \boxed{(1.0789, 1.1131)}$

We are 95% confident that the true slope of the population regression line for predicting 1997 height from 1996 height lies in (1.0789, 1.1131).

b) The value of 1 is not included in our interval. This tells us that we are 95% confident that the trees are growing from 1996 to 1997.

c) No; if the height of the tree was 0 in 1996, it makes no sense that it would suddenly grow in 1997.

Problem 2.54

a) Runs~Time has the largest correlation coefficient of 0.7449, so it has the strongest correlation.

```
cor(BBall$Runs, BBall$Time)
```

```
## [1] 0.7449071
```

```
cor(BBall$Margin, BBall$Time)
```

```
## [1] -0.1647079
```

```
cor(BBall$Pitchers, BBall$Time)
```

```
## [1] 0.6478162
```

```
cor(BBall$Attendance, BBall$Time)
```

```
## [1] 0.3187164
```

b) For every one increase in runs, we predict the time of a game to increase by 4.181 minutes on average.

$$\hat{Time} = 148.043 + 4.181(Runs)$$

```
model <- lm(Time~Runs, data=BBall)
summary(model)
```

```
##
## Call:
## lm(formula = Time ~ Runs, data = BBall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.670 -11.604  -1.117   7.378  34.330
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  148.043     11.995  12.342 3.53e-08 ***
## Runs           4.181      1.081   3.868  0.00224 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.34 on 12 degrees of freedom
## Multiple R-squared:  0.5549, Adjusted R-squared:  0.5178
## F-statistic: 14.96 on 1 and 12 DF,  p-value: 0.002237
```

c) For ρ is the correlation coefficient of the population regression line, we have

$H_0 : \rho = 0$

$H_a : \rho > 0$

The test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7449\sqrt{12}}{1-0.7449^2} = 3.868$$

Thus, the p-value is $0.0011 < 0.05$, so this is significant.

d) There is an unusually large residual in one of the games, so we might have an outlier. Besides that, the residuals display uniform variance and show no other pattern.

```
plot(resid(model)~fitted(model), main="Residuals vs. Fitted", xlab="Fitted",ylab="Residuals")
```