Stat 505 Assignment 7

Solutions

1. Test corrections 15 points

2. ARM 3.9 #1 p 49 using the pyth data.

   (a) Fit y to x_1 and x_2.

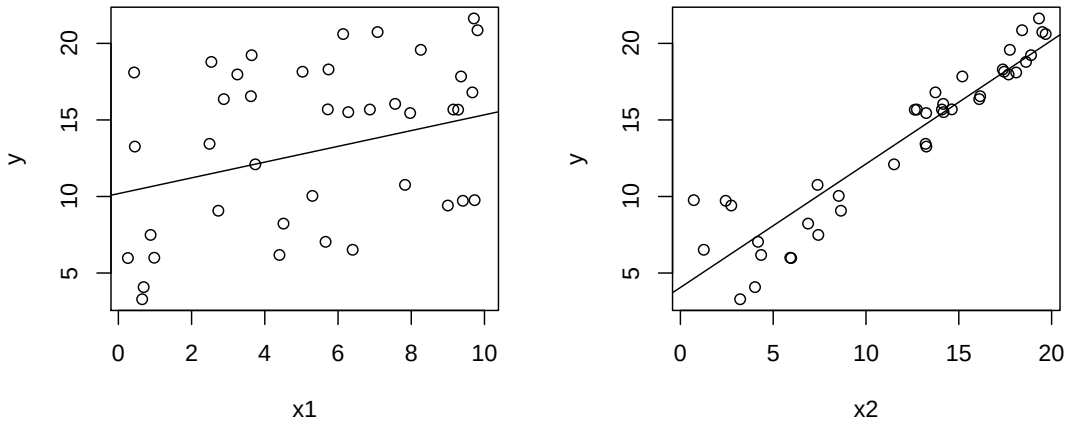
	Estimate	Std. Error	t value
(Intercept)	1.32	0.39	3.39
x1	0.51	0.05	11.22
x2	0.81	0.02	33.15

2

Table 1: n = 40 rank = 3 resid sd = 0.9 R-Squared = 0.972

This seems to be a good model because $R^2 = .97$ and the coefficients are far from 0 in standard errors (10 and 30 SE's away). When x_1 increases one unit (holding x_2 fixed), $\hat{y}$ increases by 0.51 (SE = 0.05), and similarly a 1 unit increase in x_2 holding x_1 fixed gives an increase in $\hat{y}$ of 0.81 (SE = 0.02). One could consider an interaction, which is quite strong, but the authors seem to not expect us to do that.

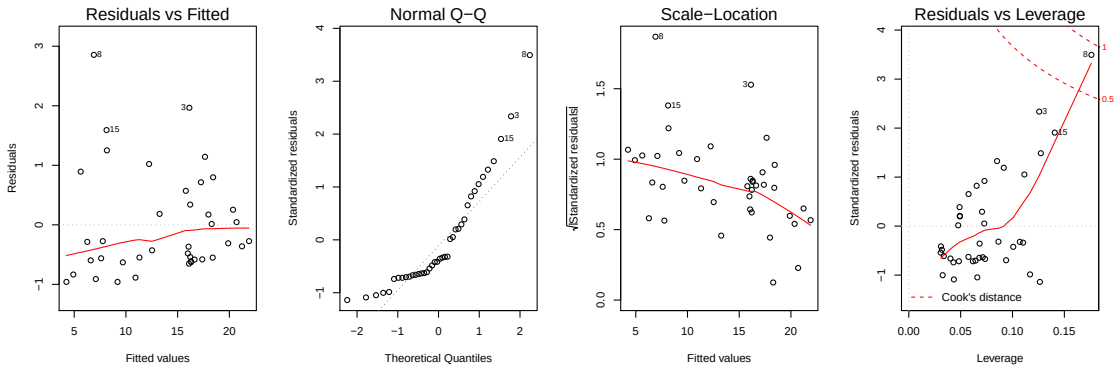
   (b) Model display.



2

In both plots I used $\widehat{\boldsymbol{\beta}}$ from a fit to y based on x_1 and x_2. In the first plot with x_1 on the horizontal axis, the line is $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 \overline{x}_2$. In the second plot the line is evaluated at the mean for x_1, that is: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}_1 + \widehat{\beta}_2 x_2$.

   (c) Residual diagnostics.



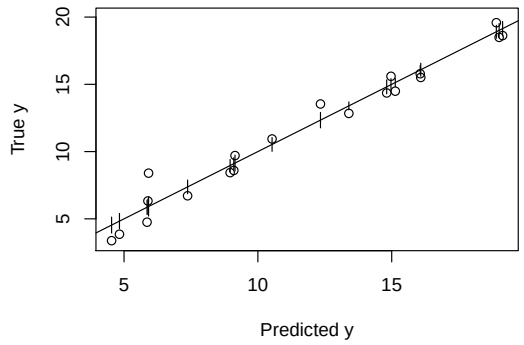
I see some problems in the residual plots. First, the residuals are right skewed rather than normally distributed. The large residuals seem to occur for small fitted values, so there also appears to be slightly decrease in spread as fitted values increase. Also, one point has large leverage and large Cook's Distance.

2

(d) Predictions.

The new values for x_1 and x_2 are not outside the range of the values we used to fit the model, and $R^2 = .97$, which is close to 1, so I think that these predictions are quite good. However, the reference says that these y's were created without error as $y = \sqrt{x_1^2 + x_2^2}$ so we have been fooled, and the predictions are not very good. The plot above shows the true values on the y axis with 95% confidence intervals for the mean shown as segments. Only 6/20 = 30% of the predicted values are within 2 SD of the "real y", so the model does not fit well. (Note: I am intentionally NOT using a prediction interval because the true residual SD is 0.)

1



3. ARM 3.9 #2 log(earnings) based on log(height)

- when height = 66 in, $\widehat{y}$ = \$30,000.
- 1% increase in height is associated with a 0.8% increase in earnings.
- 95% of earnings fall in ($0.9\widehat{y}$, $1.1\widehat{y}$), or 95% of residuals in the log scale fall in (-0.1, +0.1).

(a) Regression line equation and residual SD:

The second bullet point gives a slope estimate:

$$\ln(y) = \widehat{\beta_0} + \widehat{\beta_1}\ln(x)$$

$$\begin{aligned} \ln(y) + \ln(1.008) &= \widehat{\beta_0} + \widehat{\beta_1}[\ln(1.01) + \ln(x)] \\ \ln(1.008) &= \widehat{\beta_1} \ln(1.001) \\ \widehat{\beta_1} &= \ln(1.008)/\ln(1.01) = .80 \end{aligned}$$

and the first bullet provides an intercept estimate:

$$\begin{aligned} \ln(30000) &= \widehat{\beta}_0 + .8 \ln(66) \\ \widehat{\beta}_0 &= \ln(30000) - .8 \ln(66) = 6.96 \end{aligned}$$

An estimated variance comes from the third bullet: 95% of log earnings minus predicted log earnings lie within (-.1,.1), so estimated SD is 0.05.

$$\widehat{\ln(y)} = 6.96 + 0.8 \times \ln(x) + e, \quad e \sim (0, .05^2)$$

(b) R^2: *The total variance of y is the variance of the values on the line plus the variance of the off-line error. If* $SD(x) = .05$ *then* $SD(\widehat{y}) = \beta_1 \times 0.05$ *which is estimated as* $.8 \times .05 = .04$, *and total variance is* $.05^2 + .04^2 = .0041$ *and* $R^2 = .0016/.0041 = 0.39$. 4

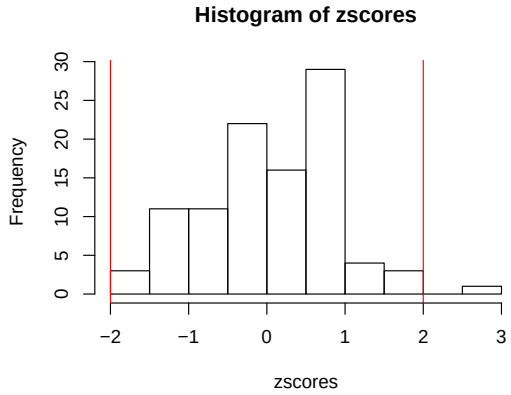
4. ARM 3.9 #3 Simulation under $H_0 : \boldsymbol{\beta = 0}$.

(a) Regress `var1` on `var2`.

	Estimate	Std. Error	t value
(Intercept)	0.02	0.03	0.58
var2	0.03	0.03	0.77

Table 2: n = 1000 rank = 2 resid sd = 1.039 R-Squared = 0.001

My first attempt gives an estimated slope of 0.026 with standard error 0.033, so it is not "significantly different" from 0.



4

(b) Do it 100 times.

Of the 100 simulated coefficients, only one was more than 2SEs from 0. This illustrates what happens when we sample from the null hypothesis: $H_0 : \beta_1 = 0$. *In the long run about 95% of the z scores should fall between -2 and 2.*

R Code

```
opts_chunk$set(fig.width = 5, fig.height = 4, out.width = ".5\\linewidth", dev = "pdf", concordance = TRUE, size = "scriptsize")
options(replace.assign = TRUE, width = 112, digits = 3, max.print = "72", show.signif.stars = FALSE)
require(xtable)
## require(arm)
```

```
pyth <- read.table("http://www.math.montana.edu/~jimrc/classes/stat505/data/pyth.dat", head = T)
pyth.fit1 <- lm(y ~ x1 + x2, pyth)
source("http://www.math.montana.edu/~jimrc/classes/stat505/Rcode/displayXtable.r")
display.xtable(pyth.fit1)
```

```
par(mfrow = c(1, 2))
beta.hat <- coef(pyth.fit1)
plot(y ~ x1, pyth)
abline(beta.hat[1] + beta.hat[3] * mean(pyth$x2), beta.hat[2])
plot(y ~ x2, pyth)
abline(beta.hat[1] + beta.hat[2] * mean(pyth$x1), beta.hat[3])
```

```
par(mfrow = c(1, 4))
plot(pyth.fit1)
```

```
newYhats <- predict(pyth.fit1, newdata = pyth[41:60, ], interval = "c", se.fit = TRUE)
realY <- sqrt(pyth$x1^2 + pyth$x2^2)[41:60]
plot(realY ~ newYhats$fit[, "fit"], ylim = c(3.3, 20), xlab = "Predicted y", ylab = "True y")
abline(0, 1)
segments(newYhats$fit[, "fit"], newYhats$fit[, "lwr"], newYhats$fit[, "fit"], newYhats$fit[, "upr"])
table(cut(abs((newYhats$fit[, "fit"] - realY)/newYhats$se.fit), c(0, 2, 8)))

##
## (0,2] (2,8]
##     6    14
```

```
set.seed(34567889 + 1)
var1 <- rnorm(1000, 0, 1)
var2 <- rnorm(1000, 0, 1)
display.xtable(model1 <- lm(var1 ~ var2))
```

```
var1 <- matrix(rnorm(1e+05, 0, 1), 1000, 100)
var2 <- matrix(rnorm(1e+05, 0, 1), 1000, 100)
zscores <- var1[, 2] * NA
for (indx in 1:100) {
    fit <- lm(var2[, indx] ~ var1[, indx])
    zscores[indx] <- coef(fit)[2]/se.coef(fit)[2]
}
table(cut(abs(zscores), c(-10, -2, 2, 10)))

##
## (-10,-2]   (-2,2]   (2,10]
##        0       99        1

rm(var1, var2)
hist(zscores)
abline(v = c(-2, 2), col = "red")
```