Regression Analysis - STAT 482 - HW #4

Alex Towell (atowell@siue.edu)

Problem 1

Refer to the data from Exercise 1.20

Data has been collected on 45 calls for routine maintenance. The goal is to explore the relationship between the number of copiers serviced (x) and the time in minutes spent to complete the service (y).

> Part (a)
>
> Determine the boundary values of the confidence band for the regression function $\mu(x)$ at $x = 3$ copiers.

In a confidence band, the error of interest is the maximum error across the input space $x_1, \ldots, x_n$. That is, we are interested in

$$\max_x \frac{\left(\hat{y}(x) - \mu(x)\right)^2}{\hat{\mathrm{V}}(\hat{y}(x))} \sim 2F(2, n-2).$$

Thus, a confidence band estimate for $\mu(x) = \beta_0 + \beta_1 x$ for all x is defined by the functions

$$L(x) = \hat{y}(x) - c\sqrt{\hat{\mathrm{V}}(\hat{y}(x))}$$

and

$$U(x) = \hat{y}(x) + c\sqrt{\hat{\mathrm{V}}(\hat{y}(x))}$$

where

$$c = \sqrt{2F(1-\alpha, 2, n-2)}.$$

```
call.data = read.table('CH01PR20.txt')
colnames(call.data) = c("time","copiers")
call.mod = lm(time ~ copiers, data=call.data)

b0 = call.mod$coefficients[1]; names(b0) = NULL
b1 = call.mod$coefficients[2]; names(b1) = NULL

e = call.mod$residuals
n = length(e)
sse = sum(e^2)
dfe = n-2
mse = sse / dfe

x.all = 1:10
x.sample = call.data$copiers
x.bar = mean(x.sample)
x.star = x.sample - x.bar
```

```
ssx = sum(x.star^2)

x.h = 3
y.h = b0 + b1*x.h
y.h.l = y.h - sqrt(2*qf(.95,2,dfe))*sqrt(mse*(1/n+(x.h-x.bar)^2/ssx))
y.h.u = y.h + sqrt(2*qf(.95,2,dfe))*sqrt(mse*(1/n+(x.h-x.bar)^2/ssx))

c(y.h.l,y.h.u)
```

```
## [1] 40.27853 48.77265
```

We see that the confidence band at $x_h = 3$ is given by

$$[40.2785284, 48.7726465].$$

> Part (b)
>
> ---
>
> Explain why the confidence band at x_h is wider than a confidence interval for μ_h.

As we increase the scope of the estimation/prediction, we increase the probability of data incompatible with a model. Thus, we need to increase the range of compatibility.

I was amused by the quote "The really unusual day would be one where nothing unusual happens."

A 95% confidence band is given by an upper and lower limit such that, with repeated sampling, 95% of the time, none of the sample points will be outside of these limits. This contrasts with a confidence interval, where we are only interested in a single point. Thus, a standard CI is too narrow and must be appropriately widened.

> Part (c)
>
> ---
>
> Plot the estimated regression function along with both the confidence band and the confidence intervals on input space $x = 1, 2, \ldots, 10$.

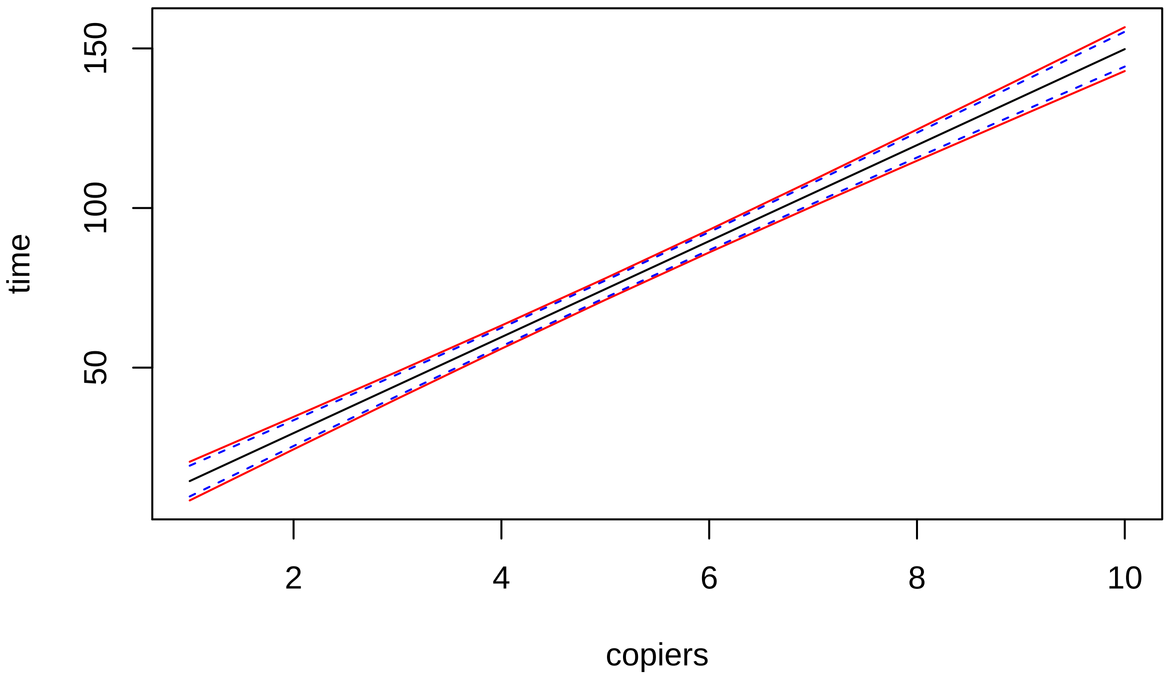
```
y.hat = b0 + b1*x.all
y.lower = y.hat - sqrt(2*qf(.95,2,dfe))*sqrt(mse*(1/n+(x.all-x.bar)^2/ssx))
y.upper = y.hat + sqrt(2*qf(.95,2,dfe))*sqrt(mse*(1/n+(x.all-x.bar)^2/ssx))
y.all = matrix(c(y.hat,y.lower,y.upper),ncol=3)

colnames(y.all) = c("mean.y","lower.limit","upper.limit")
#cbind(x.all,y.all)

matplot(x.all,y.all,type="l",lty=1,col=c("black","red","red"),
        xlab = "copiers",ylab = "time")

call.ci = predict(call.mod,data.frame(copiers=x.all),interval = "confidence")
call.ci = cbind(x.all,call.ci)
#call.ci

y.lwr = call.ci[,3]
y.upp = call.ci[,4]
points(x.all,y.lwr,type="l",lty=2,col="blue")
points(x.all,y.upp,type="l",lty=2,col="blue")
```



Problem 2

Refer to the data from Exercise 1.27

A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women for each 10 year age group, beginning with 40 and ending with age 79. The input variable x is age (in years), and the response variable y is muscle mass (in muscle mass units).

> Part (a)
>
> State the equations for $\mathrm{E}(\mathrm{MS_R})$ and $\mathrm{E}(\mathrm{MS_E})$.

Recall that $\mathrm{E}(\mathrm{MS_E}) = \sigma^2$. The expected value of the $\mathrm{MS_R}$ is

$$\mathrm{E}(\mathrm{MS_R}) = \sigma^2 + \mathrm{SS_X}\,\beta_1^2.$$

> Part (b)
>
> Use part (a) to explain a motivation behind the F test for input effects.

If $\beta_1 \approx 0$, then $\mathrm{E}(\mathrm{MS_R}) \approx \mathrm{E}(\mathrm{MS_E})$.

The test statistic is given by

$$F^* = \frac{\mathrm{MS_R}}{\mathrm{MS_E}}$$

where a small F^* indicates data compatible with the null model and a large F^* indicates data not compatible with the null model.

Part (c)

Compute the F^* statistic and the p-value. Provide an interpretation, stated in the context of the problem.

Of $\beta_1 = 0$ (no effect model), then $F^* \sim F(1, n-2)$. The p-value is given by

$$\Pr\left[F(1, n-2) > F^*\right].$$

We compute these results with:

```
mass.data = read.table('CH01PR27.txt')
colnames(mass.data) = c("mass","age")
mass.mod = lm(mass ~ age, data=mass.data)
anova(mass.mod)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	11627.486	11627.48584	174.062	0
Residuals	58	3874.447	66.80082	NA	NA

From the ANOVA table, we see that $\mathrm{MS_R} = 11627.486$, $\mathrm{MS_E} = 66.8$, and $F^* = \mathrm{MS_R} \,/\, \mathrm{MS_E} = 174.06$. This value of the test statistic has a p-value given by .000.

Interpretation: Since the data is not compatible with the no effect model, we accept the model which includes age as a predictor.

Part (d)

Compute the coefficient of determination r^2. Provide an interpretation, stated in the context of the problem.

The coefficient of determination is the proportion of variation in the response Y that is explained by input x,

$$r^2 = \frac{\mathrm{SS_R}}{\mathrm{SSTO}}.$$

We compute the coefficient of determination with:

```
summary(mass.mod)$r.squared
```

```
## [1] 0.7500668
```

We estimate that 75% of the variation in mass is explained by age.

Since the sample correlation is given by $r = \sqrt{r^2} = 0.87$, these two variables appear to be significantly correlated.