Regression Analysis - STAT 482 - HW #3

Alex Towell (atowell@siue.edu)

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). Let $x_h = 3$ copiers.

Part (1)

Compute a confidence interval for μ_h.

μ_h is the mean response $\mathrm{E}(Y_h|x_h = 3)$. A confidence interval for μ_h is given by

$$\hat{Y}_h \pm t_{\alpha/2,n-2}\,\mathrm{se}(\hat{Y}_h).$$

```
alpha = .05
x.h = 3
data = read.table('CH01PR20.txt')
colnames(data)=c("time","copiers")
mod = lm(time~copiers, data=data)
predict(mod,level=1-alpha,data.frame(copiers=x.h),interval="confidence")
```

```
##        fit     lwr      upr
## 1 44.52559 41.1476 47.90357
```

We see that a CI for μ_h is given by $[44.526, 47.904]$.

Part (2)

Compute a prediction interval for $Y_{h(new)}$.

We denote $Y|X = x_h$ by Y_h. Assume

$$Y_h \sim \mathcal{N}(\beta_0 + \beta_1 x_h, \sigma^2).$$

Then, if we wish to predict Y_h with $(\beta_0, \beta_1, \sigma)$ known, the interval

$$\beta_0 + \beta_1 x \pm z_{\alpha/2}\sigma.$$

includes the observed value of Y_h with probability $1-\alpha$. For predicting a single outcome Y_h, the uncertainty of the prediction is quantified by $\mathrm{V}(Y_h) = \sigma^2$. For a given α, the smaller the variance the smaller the interval.

However, if $(\beta_0, \beta_1, \sigma^2)'$ is unknown, we have the additional source uncertainty due to a random sample $\{(Y_h, x_h)\}$ being used to estimate $(\beta_0, \beta_1, \sigma^2)'$.

Let `pred` $= P_h = Y_{h(\text{new})} - \hat{Y}_h$ denote the random *prediction error*. Then, $\text{E}(P_h) = 0$ and

$$\begin{aligned}\text{V}(P_h) &= \text{V}(Y_{h(\text{new})}) + \text{V}(\hat{Y}_h) \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\text{SS}_\text{x}}\right].\end{aligned}$$

Since we do not know σ^2, we estimate it with MSE and thus a prediction interval for $Y_{h(\text{new})}$ is given by

$$\hat{Y}_h \pm t_{\alpha/2,n-2}\hat{\text{sd}}(P_h)$$

where

$$\hat{\text{sd}}(P_h) = \sqrt{\text{MSE}\left(1 + \frac{1}{\text{n}} + \frac{(\text{x}_\text{h} - \bar{\text{x}})^2}{\text{SS}_\text{x}}\right)}.$$

```
predict(mod,data.frame(copiers=3),level=1-alpha,interval="predict")
```

```
##        fit      lwr      upr
## 1 44.52559 26.23515 62.81603
```

We see that the PI for $Y_{h(\text{new})}$ is given by $[26.235, 62.816]$.

Part (3)

Explain the difference between a confidence interval and a prediction interval, stated in the context of the problem.

A confidence interval is for the mean service time for all service calls with 3 copiers to service.

A prediction interval is for the service time of a single service call with 3 copiers to service.

Part (4)

Let $m = 10$. Compute a prediction interval for $\bar{Y}_{h(\text{new})}$.

Let `pred.mean` $= \bar{P}_h = \bar{Y}_{h(\text{new})} - \hat{Y}_h$ denote the random *prediction error*. Then, $\text{E}(\bar{P}_h) = 0$ and

$$\begin{aligned}\text{V}(\bar{P}_h) &= \text{V}(\bar{Y}_{h(\text{new})}) + \text{V}(\hat{Y}_h) \\ &= \sigma^2 \left[\frac{1}{m} + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\text{SS}_\text{x}}\right].\end{aligned}$$

Since we do not know σ^2, we estimate it with MSE and thus a prediction interval for $\bar{Y}_{h(\text{new})}$ is given by

$$\hat{Y}_h \pm t_{\alpha/2,n-2}\hat{\text{sd}}(\bar{P}_h)$$

where

$$\hat{\text{sd}}(\bar{P}_h) = \sqrt{\text{MSE}\left(\frac{1}{\text{m}} + \frac{1}{\text{n}} + \frac{(\text{x}_\text{h} - \bar{\text{x}})^2}{\text{SS}_\text{x}}\right)}$$

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

residual = mod$residuals
```

```
n = length(residual)
sse = sum(residual^2)
mse = sse / (n-2)
x = data$copiers

x.bar = mean(x)
ssx = sum((x-x.bar)^2)

y.hat = b0 + b1*x.h
m = 10
var.pred.mean = mse*(1/m + 1/n + (x.h-x.bar)^2/ssx)

lower.ynew = y.hat - qt(alpha/2,n-2,lower.tail=F)*sqrt(var.pred.mean)
upper.ynew = y.hat + qt(alpha/2,n-2,lower.tail=F)*sqrt(var.pred.mean)

c(lower.ynew,upper.ynew)
```

```
## [1] 37.91320 51.13798
```

We see that the PI for $\bar{Y}_{h(\text{new})}$ is given by

$$[37.913, 51.138].$$

Part (5)

Provide an interpretation of your result, stated in the context of the problem.

We predict that the mean service time for 10 service calls with 3 copiers each will be between 37.913 and 51.138.

Part (6)

Show that V(pred.mean) converges to $\text{V}(\hat{Y}_h)$ as $m \to \infty$. Use this limit result to interpret the confidence interval for μ_h as a prediction.

The variance of `pred.mean` $= \bar{P}_h$ is given by

$$\text{V}(\bar{P}_h) = \sigma^2 \left(\frac{1}{m} + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\text{SS}_\text{x}} \right)$$

which we may rewrite as

$$\begin{aligned} \text{V}(\bar{P}_h) &= \frac{\sigma^2}{m} + \sigma^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\text{SS}_\text{x}} \right) \\ &= \text{V}(\bar{Y}_{h(\text{new})}) + \text{V}(\hat{Y}_h). \end{aligned}$$

As $m \to \infty$, $\text{V}(\bar{Y}_{h(\text{new})}) \to 0$, and thus

$$\lim_{m \to \infty} \text{V}(\bar{P}_h) = \text{V}(\hat{Y}_h).$$

A CI estimate of the mean service time μ_h can be interpreted as a prediction of the sample mean $\bar{Y}_{h(\text{new})}$ for a large number of service time responses.