Regression Analysis - STAT 482 - HW #5

Alex Towell (atowell@siue.edu)

Refer to the data from Exercise 3.15

A designed experiment is conducted to study the concentration of a solution (y) over time (x) in hours.

Problem 1

Part (a)

Compute the estimated regression line.

```
#setwd("/home/spinoza/filetopia/gdrive/alex/college/grad_math_classes/stat482_regression_analysis_fa202

data = read.table('CH03PR15.txt')
names(data) = c("conc","hours")
head(data)
```

conc	hours
0.07	9
0.09	9
0.08	9
0.16	7
0.17	7
0.21	7

```
reg.mod = lm(conc ~ hours, data=data)
reg.mod
```

```
##
## Call:
## lm(formula = conc ~ hours, data = data)
##
## Coefficients:
## (Intercept)        hours
##       2.575       -0.324
```

Our estimate of the linear regression function is given by

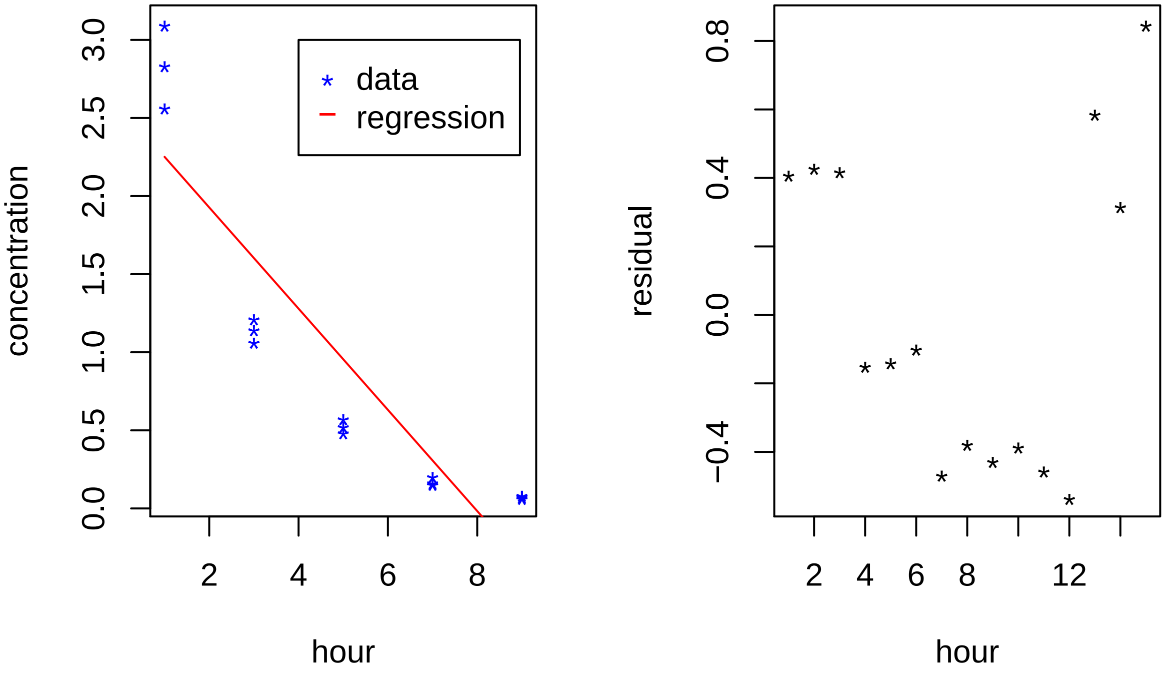
$$\hat{Y} = 2.575 - 0.324x.$$

Additional analysis

A simple visual analysis shows that this first-order linear regression model is unable to capture the variability in the concentration with respect to hours.

```
par(mfrow=c(1,2))
plot(xlab="hour",ylab="concentration",data$hours,data$conc,col="blue",pch='*')
lines(data$hours,reg.mod$fitted.values,col="red")
legend(4,3,legend=c("data","regression"),col=c("blue","red"),pch=c('*','-'))

plot(reg.mod$residuals,pch='*',xlab="hour",ylab="residual")
```



A plot of the regression function versus the data shows that it is biased. A plot of the residuals shows a clear pattern, suggesting that we could do better with a more complex model that is able to capture the systemic variability in the data.

Part (b)

Compute a point estimate for μ_5, the mean concentration for $x = 5$ hours.

```
u5.hat = predict(reg.mod, newdata=data.frame(hours=5))
u5.hat
```

```
##         1
## 0.9553333
```

We see that $\hat{\mu}_5 = \hat{\mathrm{E}}(Y|X = 5) = 0.9553333$. We note that this estimate is not compatible with the datam, and thus we would not predict that the mean concentration of the solution at 5 hours is 0.9553333.

Part (c)

Explain an advantage of using the regression estimate $\hat{\mu}_5$ instead of the sample mean $\bar{y}_5$, and explain when this advantage will lead to a more accurate estimator.

Variance

A measure of the precision of an estimator is give by its variance. A primary advantage of the linear regression estimator of the mean

$$\hat{\mu}_5 = b_0 + 5b_1,$$

compared to the saturated model estimator

$$\bar{y}_5 = \operatorname{avg}\left(\{y_i : x_i = 5\}\right)$$

is that $\hat{\mu}_5$ has smaller variance, $\mathrm{V}(\hat{\mu}_5) < \mathrm{V}(\bar{y}_5)$, i.e., it is a more precise estimator.

Information

The variance of an estimator of $\mathrm{E}(Y|X = 5)$ decreases as the information we have about $\mathrm{E}(Y|X = 5)$ increases. One way in which the information increases is by increasing the sample size. In the case of $\hat{y}_5$, we are only using the information in the subset $\{(x_i, y_i) : x_i = 5\}$, but in the case of $\hat{\mu}_5$, we are using the information in the entire sample $\{(x_i, y_i)\}$.

Bias

A measure of the accuracy of an estimator is its bias. If the linear regression model is an appropriate model for the data generating process, it will have a small bias. Moreover, if the true data generating process is given by

$$Y_i|x_i \overset{\text{indep}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

then the estimator of the mean response

$$\hat{\mathrm{E}}(Y|x) = b_0 + b_1 x$$

is the UMVU estimator for the mean response of the data generating process.

Additional commentary

This is not strictly related to the question posed by this problem set but is rather for my own instruction. Feel free to ignore, but any feedback is quite welcome.

This has to do with a distinction between estimating the data generating process and performing forecasts or predictions with appropriate confidence intervals.

Assuming

$$Y_i|x_i \overset{\text{iid}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2),$$

the UMVU estimator of the data generating process $Y|x$ is given by

$$\widehat{Y|x} \sim \mathcal{N}(b_0 + b_1 x, \hat{\sigma}^2)$$

where $\hat{\sigma}^2 = \mathrm{MS_E}$ and $(b_0, b_1)'$ is the least squares estimator of (β_0, β_1).

Observe that $\widehat{Y|x}$ has a sampling distribution since b_0, b_1, and $\hat{\sigma}^2$ are functions of the random sample $\{(x_i, Y_i) : i = 1, \dots, n\}$. Thus, if we wished to, say, perform an individual forecast of $Y|x$, confidence intervals must incorporate this sampling variance as described in a previous homework. However, the estimator $\widehat{Y|x}$ remains the most accurate and precise estimator for the conditional random variable $Y|x$.

The prediction problem $Y_{h(\text{new})}|x_h$ and the estimator of the conditional distribution $Y|x$ answer different questions whose distinction, at first glance, may not even be recognized. Even now, I am still not convinced my analysis is correct. It seems a bit slippery. Tread carefully!

Problem 2

Part (a)

State the full model (saturated model) in testing for regression model fit.

In the saturated model, the mean response is *unconstrained*, but we require at least one of the inputs to have *replications*.

There are $c = 5$ levels of the input variable, $x_1 = 1, x_2 = 3, x_3 = 5, x_4 = 7, x_5 = 9$, where the i-th input variable has $n_i = 3$ replications for a total of $n = \sum_{i=1}^{c} n_i = 3(5) = 15$ observations.

The saturated model is given by

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \dots, 5$ and $j = 1, \dots, 3$.

We may also write the reduced model, the linear regression model, as

$$Y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}.$$

Part (b)

Compute SS_{PE} and SS_{LF} for testing the fit of the linear regression model. State the respective degrees of freedom.

The sum of squared errors with respect to the full (saturated) model is given by

$$\text{SS}_{\text{PE}} = \text{SS}_{\text{E}}(F) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

which simplifies to

$$\text{SS}_{\text{PE}} = \sum_{i=1}^{c} (n_i - 1) s_i^2,$$

where SS_{PE} is called the *pure error some of squares*, which has $\text{df}_{\text{E}}(F) = \sum_{i=1}^{c} (n_i - 1) = n - c = 15 - 5 = 10$ degrees of freedom.

The sum of squared errors with respect to the reduced (linear regression) model is given by

$$\text{SS}_{\text{E}}(R) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_i)^2$$

where $\hat{Y}_i = b_0 + b_1 x_i$. This is the sum of squares SS_{E} as previously defined, which has $\text{df}_{\text{E}}(R) = n - 2 = 15 - 2 = 13$ degrees of freedom.

The sum of squares lack of fit is given by

$$\text{SS}_{\text{LF}} = \text{SS}_{\text{E}}(R) - \text{SS}_{\text{E}}(F) = \sum_{i=1}^{c} n_i (\bar{Y}_i - \hat{Y}_i)^2,$$

which has $\text{df}_{\text{E}}(R) - \text{df}_{\text{E}}(F) = (n - 2) - (n - c) = c - 2 = 3$ degrees of freedom.

Part (c)

Compute F^*_{LF} and the p-value. Provide an interpretation of your result.

Intuitively, if SS_{LF} is small, the difference between the saturated model and the reduced model is small, which suggests that reduced model is a good fit. Since the reduced model, the linear regression model, uses more of the information in the sample, we prefer it over the saturated model for previously stated reasons.

We formalize this intuition with the hypothesis test

$$H_0 : \text{E}(Y|x) = \beta_0 + \beta_1 x$$
$$H_A : \text{E}(Y|x) \neq \beta_0 + \beta_1 x$$

and the general linear test statistic

$$\begin{aligned} F^*_{\text{LF}} &= \frac{\text{SS}_{\text{LF}} \,/ (\text{df}_{\text{E}}(R) - \text{df}_{\text{E}}(F))}{\text{SS}_{\text{PE}} \,/ \,\text{df}_{\text{E}}(F)} \\ &= \frac{\text{SS}_{\text{LF}} \,/ (c-2)}{\text{SS}_{\text{PE}} \,/ (n-c)} \\ &= \frac{\text{MS}_{\text{LF}}}{\text{MS}_{\text{PE}}}, \end{aligned}$$

which under the null hypothesis is distributed as $F(c-2, n-c)$. The p-value of the test statistic is thus given by

$$\Pr\{F(c-2, n-c) \geq F^*_{\text{LF}}\}.$$

```
data$fact.hours = as.factor(data$hours)
full.mod = lm(conc ~ fact.hours - 1, data=data)
anova(reg.mod,full.mod)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
13	2.924653	NA	NA	NA	NA
10	0.157400	3	2.767253	58.60342	1.2e-06

We see that the ANOVA table provides all the necessary information. We represent the information in the more familiar form with the following table:

Table 3: Generalized linear test statistic (ANOVA)

model	df_{E}	SS_{E}	loss fit	**test statistic**
(R) $Y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$	$n - 2 = 13$	2.9247	$\text{SS}_{\text{LF}} = 2.7673$	$F^*_{\text{LF}} = 58.603$
(F) $Y_{ij} = \mu_i + \epsilon_{ij}$	$n - c = 10$	0.1573	$\text{df}_{\text{E}}(\text{LF}) = c - 2 = 3$	$p \approx .000$

We see that $F^*_{\text{LF}} = 58.603$ which has a p-value under the null hypothesis of .000.

Interpretation

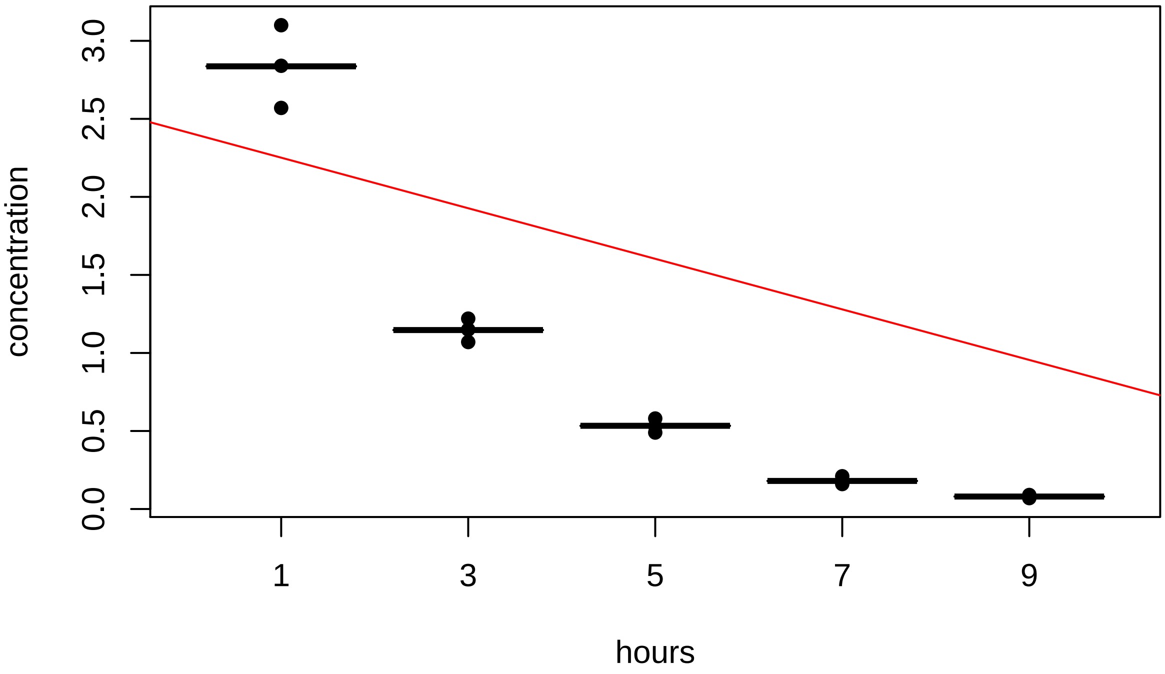
At a p-value $\approx$ .000, the data is not compatible with the linear regression model given any reasonable significance level. We will need a more flexible model.

Part (d)

Create a plot comparing the fitted values from the regression model with the fitted values from the saturated model.

```
plot(data$fact.hours,
     full.mod$fitted.values,
     type = "b",
     ylab = "concentration",
     xlab = "hours",
     ylim = c(min(data$conc),max(data$conc)))

abline(reg.mod$coefficients,col="red")
points(data$fact.hours,data$conc,pch=16)
```


concentration
hours