NCSU ST 503 HW 4

Probems 3.2, 3.4, 3.5, 3.6, 4.2 Faraway, Julian J. Linear Models with R, Second Edition Chapman & Hall / CRC Press.

Bruce Campbell

14 September, 2017

Problem 3.2

Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the cheddar dataset

(a) Fit a regression model with taste as the response and the three chemical contents as predictors. Identify the predictors that are statistically significant at the 5% level.

```
lm.fit <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-28.8767696	19.735418	-1.4631952	0.1553991
Acetic	0.3277413	4.459757	0.0734886	0.9419798
H2S	3.9118411	1.248430	3.1334077	0.0042471
Lactic	19.6705434	8.629055	2.2795710	0.0310795

We see that *H2S* and *Lactic* are significant to the 5% level.

(b) Acetic and H2S are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model.

To undo the log transform we need the base - this is not specified in the help section for the data set. Since we're dealing with chemical concentration data, and based on part e) we will assume that *Acetic* and *H2S* are measured on a Log_e scale.

```r
lm.fit.exp <- lm(taste ~ I(exp(1)^Acetic) + I(exp(1)^H2S) + Lactic, data = cheddar)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-18.9727153	11.2680492	-1.683762	0.1041981
I(exp(1)^Acetic)	0.0189056	0.0156227	1.210135	0.2371145
I(exp(1)^H2S)	0.0007668	0.0004188	1.831110	0.0785679
Lactic	25.0073579	9.0621214	2.759548	0.0104624

rsquared
0.575407

We see that now only *Lactic* is significant at the 5% level. *H2S* is significant at 10%. We thought this could be due to numerical issues in the QR - to test that out we took the transformed data set, standardize it and fit that.

For comparison on the effect of scaling we also fit the scaled model without the inverse log transform. The scaled inverse log transformed model had *H2S* and *Lactic* significant to the 5% level.

(c) Can we use an F-test to compare these two models? Explain. Which model provides a better fit to the data? Explain your reasoning.

We can not use an F-test to compare these models since they are not nested. The model fit in ln scale is a better fit to the data based on the R^2 criteria.

(d) If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected?

For the model fit in part a) we saw that $\beta_{H2S} = 3.9118$ this means that keeping all other variables constant and increasing *H2S* by 0.01 increases taste by 0.039118. We can verify this is the case numerically on an example data element from the training set.

```r
data.sample <- sample(nrow(cheddar), 1)
data.element <- cheddar[data.sample, ]
data.element$taste <- NULL
data.element <- as.matrix(cbind(intercept = 1, data.element))
beta.hat <- as.matrix(lm.fit$coefficients)
pander(data.frame(data.element), caption = "Data sample")
```

Table 4: Data sample

	intercept	Acetic	H2S	Lactic
14	1	5.236	4.942	1.3

```
response.orig <- (data.element) %*% beta.hat
# change the of our data element H2S by +0.01
data.element[1, 3] <- data.element[1, 3] + 0.01
pander(data.frame(data.element), caption = "Data sample data element H2S by +0.01")
```

Table 5: Data sample data element H2S by +0.01

	intercept	Acetic	H2S	Lactic
14	1	5.236	4.952	1.3

```
response.mod <- (data.element) %*% beta.hat
pander(data.frame(response.difference = (response.mod - response.orig)))
```

	response.difference
14	0.03912

(e) What is the percentage change in H2S on the original scale corresponding to an additive increase of 0.01 on the (natural) log scale?

Let our log concentration be α then e^{α} is our concentration in the original scale. A δ change in the log scale H2S results in a concentration of $e^{\alpha+\delta}$

The percent change is

$$(\frac{e^{\alpha+\delta} - e^{\alpha}}{e^{\alpha}}) * 100\% = (e^{\delta} - 1) * 100\%$$

In our case $\delta = 0.01$ and the percent change is `101.0050167`

Problem 3.3

Using the teengamb data, fit a model with gamble as the response and the other variables as predictors.

(a) Which variables are statistically significant at the 5% level?

```
lm.fit <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
```

term	estimate	std.error	statistic	p.value
(Intercept)	22.5556506	17.1968034	1.3116188	0.1967736
sex	-22.1183301	8.2111145	-2.6937062	0.0101118
status	0.0522338	0.2811115	0.1858118	0.8534869
income	4.9619792	1.0253923	4.8391032	0.0000179
verbal	-2.9594935	2.1721503	-1.3624718	0.1803109

We see that *gender* and *income* are both significant at the 5% level.

(b) What interpretation should be given to the coefficient for sex?

The variable *sex* is encoded $0 = male, 1 = female$ and the coefficient for it $\beta_{sex} = -22.118$. This means that when all the other variables are held constant and the gender changes from male to female that there will be a -22.118 change in *gamble*.

(c) Fit a model with just income as a predictor and use an F-test to compare it to the full model.

```
lm.fit.income <- lm(gamble ~ income, data = teengamb)
```

The reduced model $gamble \sim income$

term	estimate	std.error	statistic	p.value
(Intercept)	-6.324559	6.029874	-1.048871	0.2998383
income	5.520485	1.035772	5.329824	0.0000030

Results of the F-test

res.df	rss	df	sumsq	statistic	p.value
45	28008.59	NA	NA	NA	NA
42	21623.77	3	6384.821	4.133761	0.0117721

Based on the p-value of the F-statistic we do have enough evidence to reject the null hypothesis that the models are equivalent in the variance explained via the RSS statistic. We claim that the full model is better based on the RSS criteria.

Problem 3.4

We are using the sat data for this problem.

(a) Fit a model with total sat score as the response and expend, ratio and salary as predictors. Test the hypothesis that $\beta_{salary} = 0$. Test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. Do any of these predictors have an effect on the response?

```r
lm.fit <- lm(total ~ expend + ratio + salary, data = sat)
tidy(lm.fit)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1069.234168	110.924940	9.6392585	0.0000000
expend	16.468866	22.049899	0.7468907	0.4589302
ratio	6.330267	6.542052	0.9676272	0.3382908
salary	-8.822632	4.696794	-1.8784372	0.0666677

We see that salary is significant at the $\alpha = 10\%$ level.

```r
lm.fit.reduced <- lm(total ~ expend + ratio, data = sat)
anova(lm.fit.reduced, lm.fit)
```

Res.Df	RSS	Df	Sum of Sq	F	
47	233442.9	NA	NA	NA	
46	216811.9	1	16631.01	3.528526	
We see th	at the F-st	atist	ic has a p-v	alue of $0.	0667$ - this is the same as the p-value for the t

Test $H_0 : \beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$

```r
lm.fit.null <- lm(total ~ 1, data = sat)

anova(lm.fit.null, lm.fit)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
49	274307.7	NA	NA	NA	NA
46	216811.9	3	57495.74	4.066203	0.0120861

Based on the F-statistic we have enough evidence to reject the null hypothesis that all coefficients are zero. We claim at least one predictor has an effect on the response.

(b) Now add takers to the model. Test the hypothesis that $\beta_{takers} = 0$. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent.

Fit the model $total \sim expend + ratio + salary + takers$

```r
lm.fit <- lm(total ~ expend + ratio + salary + takers, data = sat)
tidy(lm.fit)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1045.971536	52.869760	19.7839283	0.0000000
expend	4.462594	10.546528	0.4231339	0.6742130
ratio	-3.624232	3.215418	-1.1271418	0.2656570
salary	1.637917	2.387248	0.6861110	0.4961632
takers	-2.904481	0.231260	-12.5593745	0.0000000

Fir the model $total \sim expend + ratio + salary$ and perform the F-test.

```r
lm.fit.reduced <- lm(total ~ expend + ratio + salary, data = sat)
anova(lm.fit.reduced, lm.fit)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
46	216811.9	NA	NA	NA	NA
45	48123.9	1	168688	157.7379	0

Just as above we see that the F-statistic for the reduced model has a p-value that is the same as the p-value for the t-statistic given above for the coefficient β_{takers}

REVISIT! Thinking for a moment on proving this equivalemce more formally, we know that the t and F distributions are related by $T \sim t_n \implies T^2 \sim F_{1,n}$ We know that the F-test is derived from the generalized likelihood ratio test and that in our case with the assumption of normal errors the MLE of the parameters is multivariate normal

$$\hat{\beta} \sim N(\beta, \sigma(\mathbf{X}^\intercal \mathbf{X})^{-1})$$

. Also, $\epsilon_i \sim N(0, \sigma^2) \implies \frac{\epsilon_i^2}{\sigma^2} \sim \chi_1^2$ and that

$$\sum_{i=1}^{n} \chi_1^2 \sim \chi_n^2$$

. We can see how RSS_ω and RSS_Ω come up in the F-test as sums of squares of normal random variables. We should be able to show that the T^2 comes out of a particular F-test situation where the degress of freedom of ω and Ω differ by 1. REVISIT!

Problem 3.5 R^2 and the F-test

Find a formula relating R 2 and the F-test for the regression.

Problem 3.6 MBA Students

Thirty-nine MBA students were asked about happiness and how this related to their income and social life. The data are found in happy. Fit a regression model with happy as the response and the other four variables as predictors. (a) Which predictors were statistically significant at the 1% level? (b) Use the table function to produce a numerical summary of the response. What assumption used to perform the t-tests seems questionable in light of this summary? (c) Use the permutation procedure described in Section 3.3 to test the significance of the money predictor. (d) Plot a histgram of the permutation t-statistics. Make sure you use the the probability rather than frequency version of the histogram. (e) Overlay an appropriate t-density over the histogram. Hint: Use grid <- seq(-3, 3, length = 300) to create a grid of values, then use the dt function to compute the t-density on this grid and the lines function to superimpose the result. (f) Use the bootstrap procedure from Section 3.6 to compute 90% and 95% con- fidence intervals for ??money. Does zero fall within these confidence intervals? Are these results consistent with previous tests? Faraway, Julian J.. Linear Models with R, Second Edition (Chapman & Hall/CRC Texts in Statistical Science) (Page 50). CRC Press. Kindle Edition.

Fit a regression model with happy as the response and the other four variables as predictors.

```
##
## Call:
## lm(formula = happy ~ money + sex + love + work, data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7186 -0.5779 -0.1172  0.6340  2.0651
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.072081   0.852543  -0.085   0.9331
## money        0.009578   0.005213   1.837   0.0749 .
## sex         -0.149008   0.418525  -0.356   0.7240
## love         1.919279   0.295451   6.496 1.97e-07 ***
## work         0.476079   0.199389   2.388   0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.058 on 34 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.6761
## F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```

(a) Which predictors were statistically significant at the 1% level?

We see that money and love are significant at the 1% level.

(b) Use the table function to produce a numerical summary of the response. What assumption used to perform the t-tests seems questionable in light of this summary?

```r
table(happy$happy)
```

2	3	4	5	6	7	8	9	10
1	1	4	5	2	8	14	3	1

(c) Use the permutation procedure described in Section 3.3 to test the significance of the money predictor. (d) Plot a histgram of the permutation t-statistics. Make sure you use the the probability rather than frequency version of the histogram. (e) Overlay an appropriate t-density over the histogram. Hint: Use grid <- seq(-3, 3, length = 300) to create a grid of values, then use the dt function to compute the t-density on this grid and the lines function to superimpose the result. (f) Use the bootstrap procedure from Section 3.6 to compute 90% and 95% con- fidence intervals for ??money. Does zero fall within these confidence intervals? Are these results consistent with previous tests?