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

18 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

r.squared
0.6517747

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.

```
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.

```
data.sample <- sample(nrow(cheddar), 1)
data.element <- cheddar[data.sample, ]
```

```r
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 5: Data sample

	intercept	Acetic	H2S	Lactic
14	1	5.236	4.942	1.3

```r
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 increased by +0
```

Table 6: Data sample data element H2S increased by +0.01

	intercept	Acetic	H2S	Lactic
14	1	5.236	4.952	1.3

```r
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

res.df	rss	df	sumsq	statistic	p.value
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?

```
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.

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

```
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 $total \sim expend + ratio + salary + takers$

```
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.

```
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}

Problem 3.5 R^2 and the F-test

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

Let Ω be the parameter space for a model in p dimensions and ω be the parameter space for a model in q dimensions.

$$R^2_\Omega = 1 - \frac{RSS_\Omega}{TSS}$$
$$R^2_\omega = 1 - \frac{RSS_\omega}{TSS}$$

Solving for TSS in the first case we have that $TSS = \frac{RSS_\Omega}{(1-R^2_\Omega)}$ and putting this into the the expression for R^2_ω

$$R^2_\omega = 1 - \frac{RSS_\omega}{RSS_\Omega}(1 - R^2_\Omega) \implies$$

$$\frac{RSS_\omega}{RSS_\Omega}R^2_\Omega - R^2_\omega = \frac{RSS_\omega - RSS_\Omega}{RSS_\Omega} \implies$$

$$(\frac{RSS_\omega}{RSS_\Omega}R^2_\Omega - R^2_\omega)\frac{dF_\Omega}{df_\Omega - df_\omega} = \frac{(RSS_\omega - RSS_\Omega)/(df_\Omega - df_\omega)}{RSS_\Omega/df_\Omega} \sim F_{df_\Omega - df_\omega, n-df_\Omega}$$

Another way to think about R^2 is that it says something about a model in Ω versus the null model $y \sim \beta_0$. $TSS = \sum(y_i - \bar{y})^2$ in this case will be RSS_{ω_0} - the sum of square residuals for the null model. In this case $df_\omega = 1$ and we can manipulate the expression for $R^2 = 1 - \frac{RSS_\Omega}{RSS_{\omega_0}}$ directly to get

$$1 - R^2 = \frac{RSS_\Omega}{RSS_{\omega_0}} \implies 1 - \frac{RSS_\Omega}{RSS_{\omega_0}} = \frac{R^2}{(1 - R^2)}$$

and that *for the sace of comparing a full model against the null model* we have

$$\frac{R^2}{(1 - R^2)}\frac{p}{p-1} = \frac{(RSS_{\omega_0} - RSS_\Omega)/(p-1)}{RSS_\Omega/p} \sim F_{p-1,n-p}$$

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 faraway::happy.

Note, pay attention to warnings in R! GGally has a happy data set as well.

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?

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

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

```
# hist(happy$happy) plot(lm.fit)
```

Wow, I have NO idea what this question is asking me. The assumptions we make are; * Linear relationship * Multivariate normality * No or little multicollinearity * No auto-correlation *

Homoscedasticity - the variance around the regression line is the same for all values of the predictors

All of these assumptions involve elements beyond the distribution of the measured responses $ { y_i } $.

(c) Use the permutation procedure described in Section 3.3 to test the significance of the money predictor.

```
# summary(lm.fit)$coef[2,]# Est, sterr,t-stat,pval for the
nreps <- 4000
tstats <- numeric(nreps)
for (i in 1:nreps) {
    lm.resample.money <- lm(happy ~ sample(money) + sex + love + work, data = happy)
    tstats[i] <- summary(lm.resample.money)$coef[2, 3]  #Get the tstatistic for this re
}
simulated.pvalue <- mean(abs(tstats) > abs(summary(lm.fit)$coef[2, ]))  #Calculate the

pander(data.frame(simulated.pvalue = simulated.pvalue))
```

simulated.pvalue
0.7432

WE see that the simulated pvalue based on resampling the *money* predictor is very close the value we got from performing the t-test on β_{money}

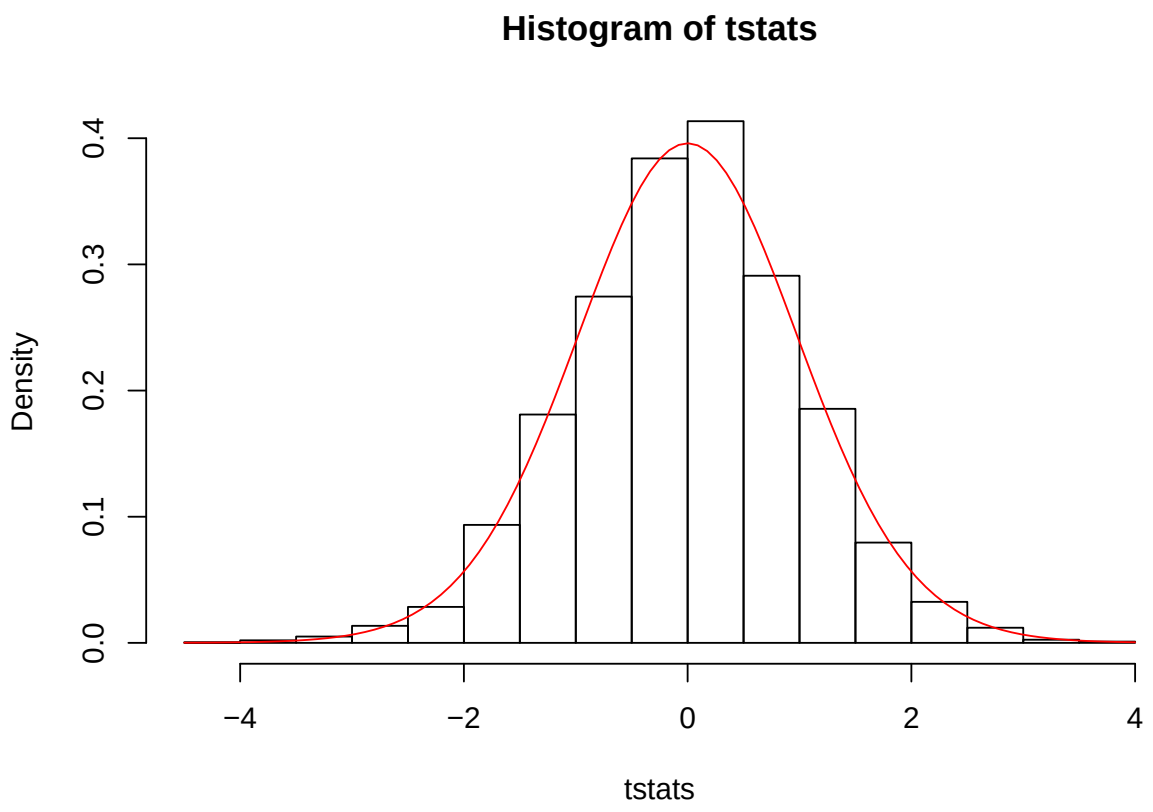
```
pvalue.money <- summary(lm.fit)$coef[2, 4]
pander(data.frame(pvalue.money = pvalue.money))
```

pvalue.money
0.07491

(d) AND (e) Plot a histgram of the permutation t-statistics. Overlay an appropriate t-density

```
hist(tstats, 30, freq = FALSE)
grid <- seq(-4, 4, length = 300)
n <- nrow(happy)
p <- 4 + 1  # Four predictors plus the intercept
df <- n - p
hist(tstats, 30, freq = FALSE)
```

```
curve(dt(x, df = df), add = TRUE, col = "red")
```



(f) Use the bootstrap procedure from Section 3.6 to compute 90% and 95% confidence intervals for *money*. Does zero fall within these confidence intervals? Are these results consistent with previous tests?

```
nb <- 4000
coefmat <- matrix(NA, nb, 5)
resids <- residuals(lm.fit)
preds <- fitted(lm.fit)
for (i in 1:nb) {
    booty <- preds + sample(resids, rep = TRUE)
    bmod <- update(lm.fit, booty ~ .)
    coefmat[i, ] <- coef(bmod)
}
colnames(coefmat) <- c("Intercept", "money", "sex", "love", "work")
coefmat <- data.frame(coefmat)
apply(coefmat, 2, function(x) quantile(x, c(0.05, 0.95)))
```

	Intercept	money	sex	love	work
5%	-1.353335	0.0013456	-0.7961964	1.457376	0.1663225
95%	1.262513	0.0178614	0.4625778	2.375133	0.7850518

We see that for a significance of $\alpha = 10\%$ that we have enough evidence to reject the null hypothesis that the coefficient for *money* is zero. This is the same result for the permutation test and for the t-test that is performed as part of R's `lm.summary` routine. Now we look at the 95 confidence interval.

```
apply(coefmat, 2, function(x) quantile(x, c(0.025, 0.975)))
```

	Intercept	money	sex	love	work
2.5%	-1.605365	-0.0001803	-0.9232030	1.388539	0.1053054
97.5%	1.492445	0.0196490	0.5813499	2.480985	0.8416759

Since 0 is in the interval for *money*; at a significance of $\alpha = 5\%$, with the data at hand, we do *not* have enough evidence to reject the null hypothesis that the coefficient for *money* is zero in the linear model $happy \sim money + sex + love + work$

Problem 4.2 - prediction with the teengamb data set.

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

```
rm(list = ls())
data(teengamb, package = "faraway")
lm.fit <- lm(gamble ~ ., data = teengamb)
```

(a) Predict the amount that a male with average (given these data) status, income and verbal score would gamble along with an appropriate 95% CI.

```
x <- model.matrix(lm.fit)
x0 <- apply(x, 2, mean)
# The question asks for a male and here we set that value
x0["sex"] <- 0
# predict(lm.fit,new=data.frame(t(x0)))
pi <- predict(lm.fit, new = data.frame(t(x0)), interval = "confidence", level = 0.95)
pi
```

fit	lwr	upr
28.24252	18.78277	37.70227

```
pander(data.frame(pi.width = pi[3] - pi[2]), caption = "Confidence interval width")
```

Table 22: Confidence interval width

pi.width
18.92

(b) Repeat the prediction for a male with maximal values (for this data) of status, income and verbal score. Which CI is wider and why is this result expected?

```
x <- model.matrix(lm.fit)
x0 <- apply(x, 2, max)
# The question asks for a male and here we set that value
x0["sex"] <- 0
# predict(lm.fit,new=data.frame(t(x0)))
pi <- predict(lm.fit, new = data.frame(t(x0)), interval = "confidence", level = 0.95)
pi
```

fit	lwr	upr
71.30794	42.23237	100.3835

```
pander(data.frame(pi.width = pi[3] - pi[2]), caption = "Confidence interval width")
```

Table 24: Confidence interval width

pi.width
58.15

(c) Fit a model with sqrt(gamble) as the response but with the same predictors. Now predict the response and give a 95% prediction interval for the individual in (a). Take care to give your answer in the original units of the response.

```
lm.fit <- lm(sqrt(gamble) ~ ., data = teengamb)
x <- model.matrix(lm.fit)
x0 <- apply(x, 2, mean)
x0["sex"] <- 0
pi <- predict(lm.fit, new = data.frame(t(x0)), interval = "confidence", level = 0.95)
```

```r
# ORIG
pi.orig <- c(pi[1]^2, pi[2]^2, pi[3]^2)
pi.orig
```

```
## [1] 16.39864 10.11670 24.19037
```

```r
pander(data.frame(pi.width = pi.orig[3] - pi.orig[2]), caption = "Confidence interval wi
```

Table 25: Confidence interval width

pi.width
14.07

The square root transform is known to stabilize variance and we see that in the smaller prediction interval.

(d) Repeat the prediction for the model in (c) for a female with status=20, income=1, verbal = 10. Comment on the credibility of the result.

```r
x0["sex"] <- 1
x0["status"] <- 20
x0["income"] <- 1
x0["verbal"] <- 10

pi <- predict(lm.fit, new = data.frame(t(x0)), interval = "confidence", level = 0.95)
# ORIG
pi
```

fit	lwr	upr
-2.08648	-4.445938	0.272978

```r
pi.orig <- c(-pi[1]^2, pi[2]^2, pi[3]^2)
pi.orig
```

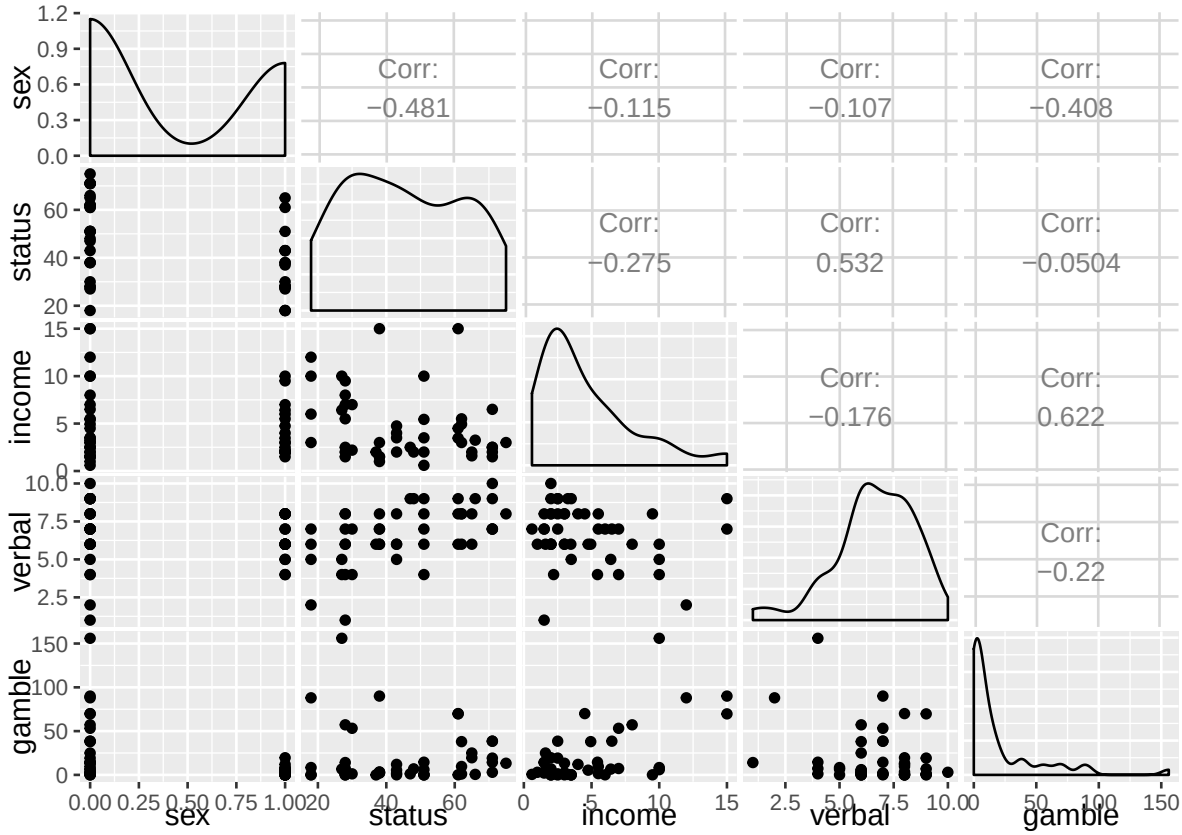
```
## [1] -4.35339768 19.76636002  0.07451699
```

```r
pander(data.frame(pi.width = pi.orig[3] - pi.orig[2]), caption = "Confidence interval wi
```

Table 27: Confidence interval width

pi.width
-19.69

```
ggpairs(teengamb)
```




```
lm.fit <- lm(gamble ~ ., data = teengamb)
pi <- predict(lm.fit, new = data.frame(t(x0)), interval = "confidence", level = 0.95)
pi
```

fit	lwr	upr
-23.15096	-48.84003	2.538117