Solution (Qual Spring 2013, Problem 1) :

a)

(i) In the process a defective item will be detected if the item is inspected (I) and is defective (A^c). The table gives $P(I \cap A^c) = qp'$.
So an item will not be detected as defective if it is not inspected, or inspected but non-defective, that is an item will not be detected as defective with probability $1 - qp'$.
Now N is the number of items passing the production chain before the first detection of a defective item. So if $N = n$, then n items were not detected as defective and the $(n+1)^{th}$ item was detected as defective. Thus N follows a Geometric distribution:

$$P(N = n) = (1 - qp')^n qp'; \quad n = 0, 1, 2, 3, 4, \ldots$$

(ii) Given N items are passed the production chain before first detection, for each of those items, the probability of being defective yet not detected is $\frac{qq'}{1-qp'}$.
For the reasoning consider the following:

	Not Defective	Defective
Inspected	$category\ 1\ (pp')$	$category\ 2(qp')$
Not Inspected	$category\ 3(pq')$	$category\ 4(qq')$

Suppose four cells in the table are considered as four categories. Note that, the items that are passing the production chain before the first detection of a defective cannot belong to category 2.
If we want to get the probability of being defective among these items then we are looking at

$$P(an\ item\ belongs\ to\ category\ 4, given\ it\ does\ not\ belong\ to\ category\ 2) = \frac{qq'}{1-qp'}$$

This is true for each of the N items. Each of these N items may be defective or non-defective.
Thus K(given N) follows a Binomial distribution.

$$P(K = k|N = n) = \binom{n}{k}\left(\frac{qq'}{1-qp'}\right)^k\left(1 - \frac{qq'}{1-qp'}\right)^{n-k}; \quad k = 0, 1, 2, \ldots, n$$

b) The joint distribution of N and K is given by:

$$P(N = n, K = k) = P(N = n)P(K = k|N = n)$$

$$= (1 - qp')^n qp' \binom{n}{k} \left(\frac{qq'}{1 - qp'}\right)^k \left(1 - \frac{qq'}{1 - qp'}\right)^{n-k}$$

$$= \binom{n}{k} (qq')^k (qp')p^{n-k}; \quad n = 0, 1, 2, \ldots; k = 0, 1, 2, \ldots, n$$

c) The Marginal distribution of K

$$P(K = k) = \sum_n P(N = n, K = k)$$

$$= \sum_{n=k}^{\infty} \binom{n}{k} (qq')^k (qp')p^{n-k}$$

$$= (qq')^k (qp') \sum_{n=k}^{\infty} \binom{n}{k} p^{n-k}$$

$$= (q')^k (p') \sum_{n=k}^{\infty} \binom{n}{k} p^{n-k} q^{k+1}$$

$$= (q')^k (p') \sum_{n-k=0}^{\infty} \binom{n-k+k}{k} p^{n-k} q^{k+1}$$

$$= (q')^k (p') \sum_{r=0}^{\infty} \binom{r+k}{r} p^r q^{k+1}$$

$$= (q')^k (p')$$

(Note that if we consider Bernoulli trials with probability of success p and write X = number of success before $(k+1)^{th}$ failure then $\sum_{r=0}^{\infty} P(X = r) = \sum_{r=0}^{\infty} \binom{r+k}{r} p^r q^{k+1} = 1$).

Thus the marginal distribution of K is again Geometric, given by

$$P(K = k) = (q')^k p'; \quad k = 0, 1, 2, \ldots$$

d) $Cov(N,K) = E(NK) - E(N)E(K)$

Since N follows geometric distribution, $E[N] = \frac{1-qp'}{qp'}$ and $Var[N] = \frac{1-qp'}{(qp')^2}$.

Since K follows geometric distribution, $E[K] = \frac{1-p'}{p'} = \frac{q'}{p'}$

$$E(NK) = E[N.E(K|N)] = E\left[N.N\frac{qq'}{1-qp'}\right] = \frac{qq'}{1-qp'}E[N^2]$$

$$= \frac{qq'}{1-qp'}\{Var[N] + E^2[N]\}$$

$$= \frac{qq'}{1-qp'}\left\{\frac{1-qp'}{(qp')^2} + \left(\frac{1-qp'}{qp'}\right)^2\right\}$$

$$= \frac{qq'}{(qp')^2}\{1 + (1-qp')\}$$

$$= \frac{q'}{(p')^2}\{2-qp'\}$$

$$Cov(N,K) = \frac{q'}{(p')^2}\{2-qp'\} - \left(\frac{1-qp'}{qp'}\right)\left(\frac{q'}{p'}\right) = \frac{q'}{q(p')^2}$$

1 Problem 2

Consider data from a study of $n = 32$ adults. The study focused on alcohol metabolism, with the ultimate goal of answering questions related to female's lower tolerance for alcohol and greater propensity to develop accompanying alcohol-related liver disease, relative to males. The variables are:

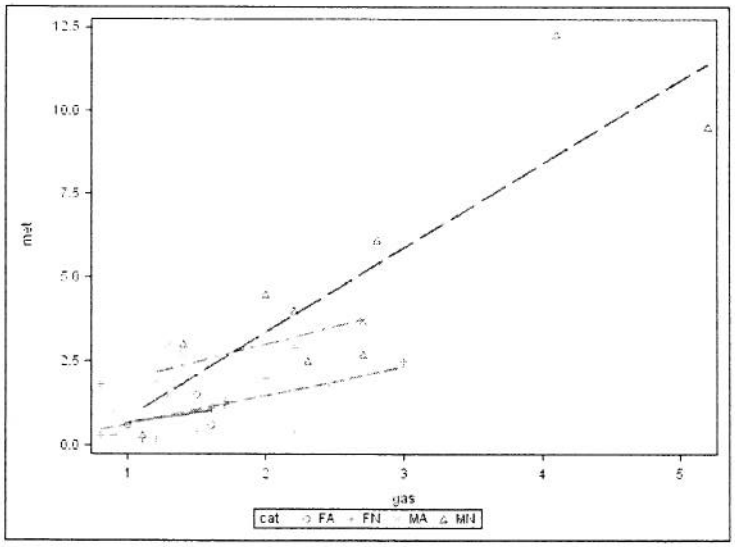
- Metabol – First-pass metabolism of alcohol in the stomach (mmol/liter-hour); this is the response of interest.
- Gastric – The gastric alcohol dehydrogenase activity in the stomach (μmol/min/g of tissue).
- Sex – The subject's gender.
- Alcohol – Indicates whether subject is an alcoholic (Alc) or not (Non-alc).

The data is available for download at http://www.stat.sc.edu/~hansont/alcohol.txt. You are to find a parsimonious, yet adequate explanatory regression model for the metabolism response variable involving gender for sure, and including the remaining concomitant variables if necessary. Make sure that you carefully assess all assumptions for your final model and write a succinct, coherent, and complete summary of your analysis, addressing the scientific question at hand.

1.1 Analysis on original scale

This is standard analysis of covariance data with one continuous predictor (gastric) and the two dichotomous categorical variables gender and an indicator for alcoholism. The focus on the study is gender differences in metabolism, adjusting for the two concomitant variables gastric activity and whether the individual is an alcoholic.

A scatterplot of the data with OLS linear fits superimposed shows increasing metabolism with gastric activity within each of the four levels of sex*alcoholic:



There appears to be no significant effect due to being an alcoholic, but the OLS gastric slopes change within gender; there appears to be an interaction between gastric activity and gender. Standard backwards elimination from the full model with all interactions yields a model with gastric, sex, and a gastric*sex interaction term:

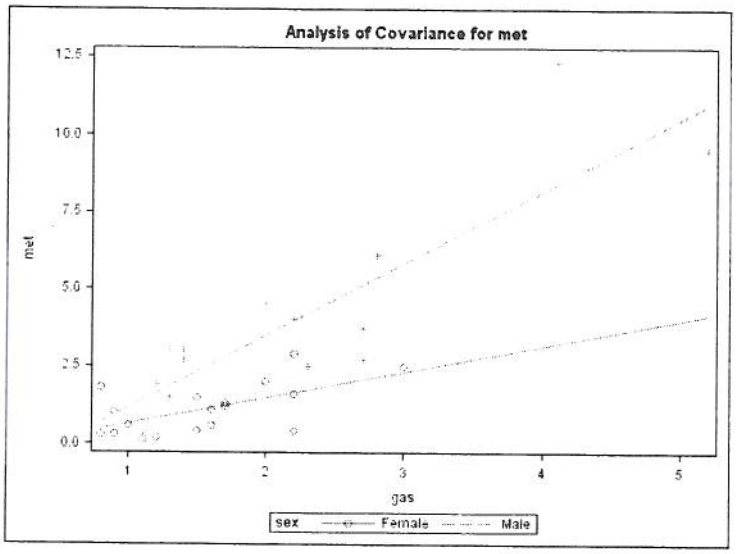
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	178.2820098	59.4273366	40.77	<.0001
Error	28	40.8126777	1.4575956		
Corrected Total	31	219.0946875			

R-Square	Coeff Var	Root MSE	met Mean
0.813721	49.85019	1.207309	2.421875

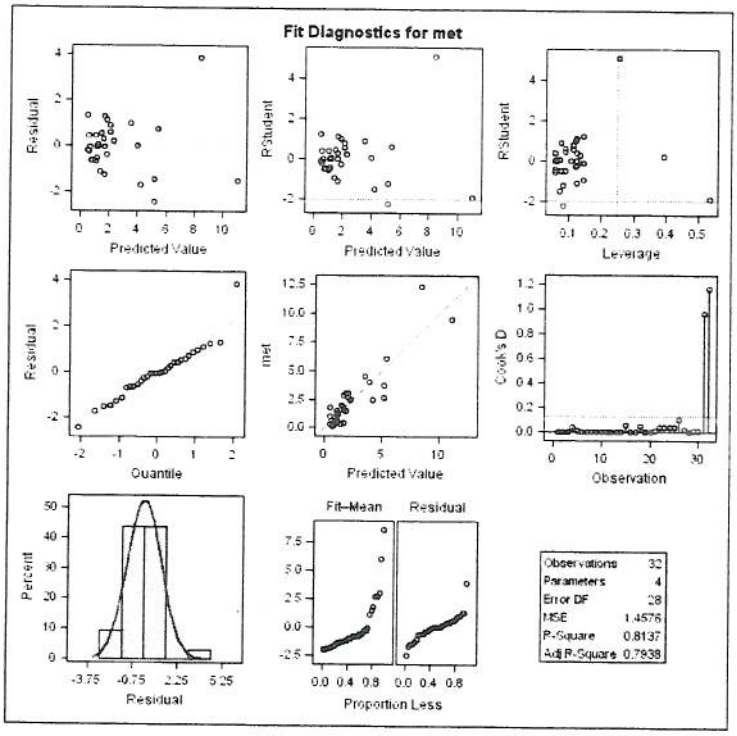
Source	DF	Type III SS	Mean Square	F Value	Pr > F
gas	1	47.17128320	47.17128320	32.36	<.0001
sex	1	1.23845807	1.23845807	0.85	0.3645
gas*sex	1	10.58722628	10.58722628	7.26	0.0118

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-1.185765932	B	0.71168462	-1.67	0.1068
gas		2.343871390	B	0.28014800	8.37	<.0001
sex	Female	0.988496856	B	1.07239102	0.92	0.3645
sex	Male	0.000000000	B	.	.	.
gas*sex	Female	-1.506923598	B	0.55913756	-2.70	0.0118
gas*sex	Male	0.000000000	B	.	.	.

The ANCOVA plot clearly shows the difference in slopes



The diagnostics are a bit suspect



The residuals versus gastric (not shown), however, do not show any obvious pattern suggesting the addition of a quadratic term. An added variable plot could refine/contradict this observation and show otherwise but is not pursued here.

The observation with the studentized deleted residual larger than 5 is also the one with the largest Cook's distance: despite being highly influential, the point is still ill-fit. This turns out to be a male non-alcoholic with the largest metabolism in the entire data set. Removing this point yields

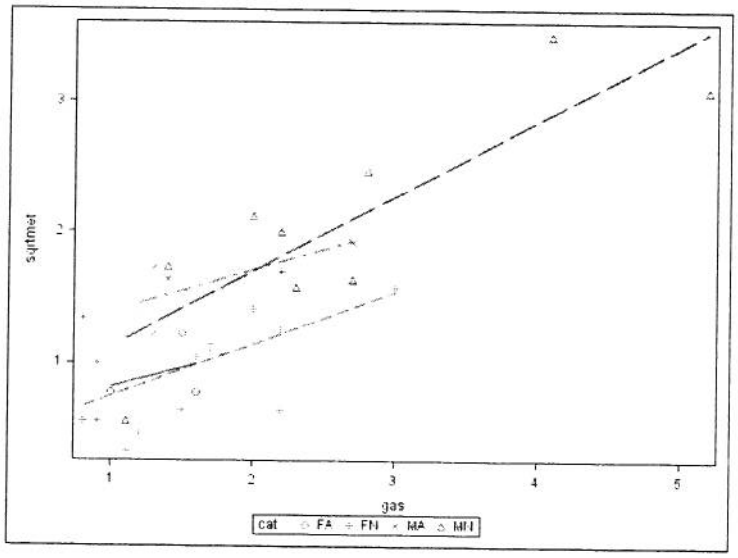
Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-0.395264079 B	0.53880486	-0.73	0.4695
gas		1.831102646 B	0.22653816	8.08	<.0001
sex	Female	0.197995003 B	0.79302211	0.25	0.8047
sex	Male	0.000000000 B	.	.	.
gas*sex	Female	-0.994154855 B	0.41774609	-2.38	0.0246
gas*sex	Male	0.000000000 B	.	.	.

The interpretation stays largely the same. However, there is still the issue of some indication of non-constant variance. I think a better approach would be to attempt a transformation on metabolism to achieve a better fit. However, if a student makes it this far, they should receive almost full credit.

Another option we discuss in class is the use of robust regression to downweight (instead of completely removing as above, or essentially giving weight zero) observations with large residuals. This can be done via median (or L_1 or LAD) regression in proc quantreg, or else using M-estimation in proc robustreg.

1.2 Transformation of metabolism

Consideration of Box-Cox transformation suggests $\lambda = 0.5$, the square root, for several possible models including all interactions, the additive model, and models in between. There appears to be no significant differences between alcoholics and non-alcoholics within gender, but there is an observable gender difference.



Using the $\sqrt{(}$met) as the response gives rougly parallel estimated OLS lines; an additive model should fit well on the transformed response. Fitting the full three-way interaction model (clearly gas should only be included linearly) allows us to drop all terms higher than first order (p=0.868 on df=4). A further test allows us to drop the indicator of alcoholism from the model (p=0.779). The final model has $\sqrt{(}$met) as the response, a linear effect of gas, and an additive gender effect.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12.88896199	6.44448099	47.81	<.0001
Error	29	3.90881235	0.13478663		
Corrected Total	31	16.79777434			

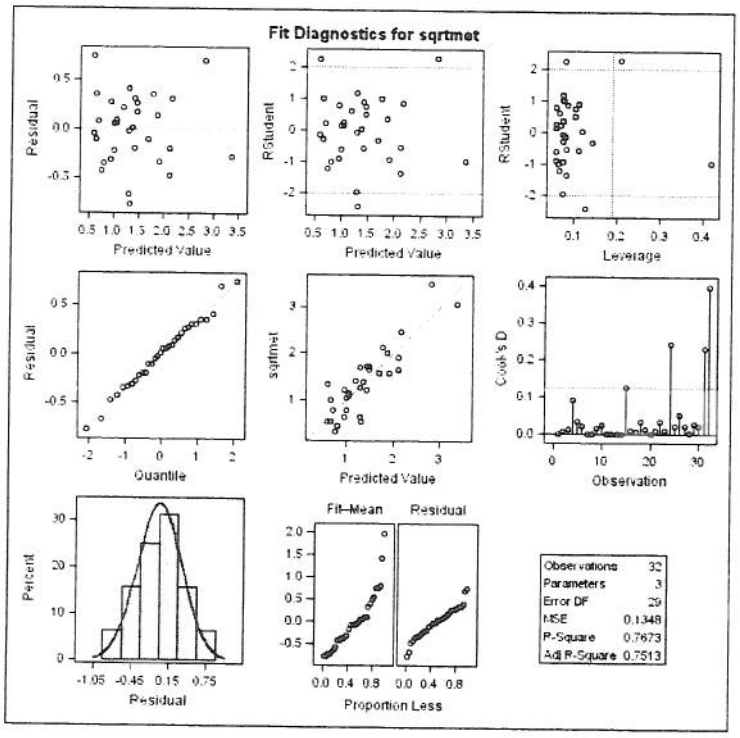
R-Square	Coeff Var	Root MSE	sqrtmet Mean
0.767302	26.65607	0.367133	1.377296

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		0.7747085303 B	0.19363815	4.00	0.0004
gas		0.4965472496	0.07372633	6.74	<.0001
sex	Female	-.5728563217 B	0.14102841	-4.06	0.0003
sex	Male	0.0000000000 B	.	.	.

Females typically have a significant 0.57 reduction in sqrt(met) adjusting for gas, i.e. holding gas constant. Increasing gas one unit significantly increases sqrt(met) by 0.50. These results can be back-transformed to give the median regression model:

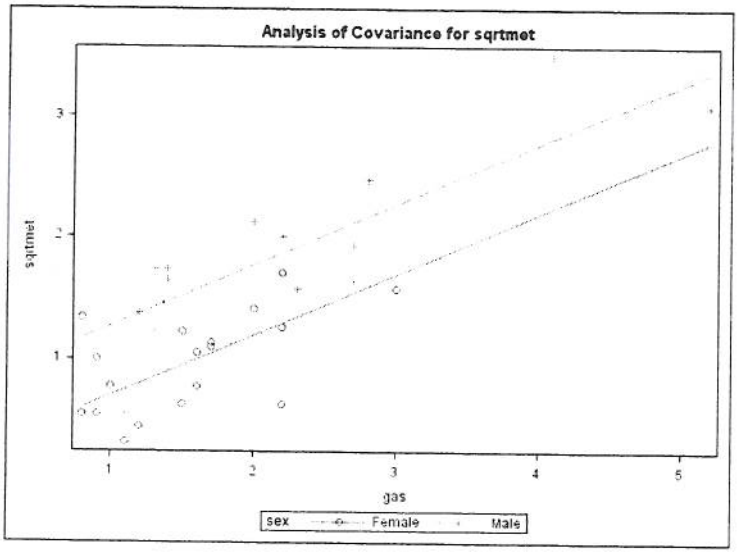
$$Median(\text{met}) = (0.775 + 0.497\text{gas} - 0.573I\{\text{female}\})^2$$

which ultimately DOES yield a gas by female interaction on the original scale of the data, just as in the original analysis. I would expect students to not necessarily do this, but if they notice this they get bonus points. Also, it's nice if students report 95% CI's for the regression effects, but they for sure need to note that they are significant at the 5% level. The diagnostic panel looks a lot better



There are no observations that are very poorly fit; however there are a few influential points, but not nearly as influential as for the untransformed data.

Here's the ANCOVA plot



2 SAS code

```
*************************************************************************
Alcohol metabolism regression problem
*************************************************************************;
```

```
data alcohol;
input met gas sex$ alc$;
if sex='Female' and alc='Alc' then cat='FA';
if sex='Female' and alc='Non-alc' then cat='FN';
if sex='Male' and alc='Alc' then cat='MA';
if sex='Male' and alc='Non-alc' then cat='MN';
sqrtmet=sqrt(met);
datalines;
0.6 1 Female Alc
0.6 1.6 Female Alc
1.5 1.5 Female Alc
0.4 2.2 Female Non-alc
0.1 1.1 Female Non-alc
0.2 1.2 Female Non-alc
0.3 0.9 Female Non-alc
0.3 0.8 Female Non-alc
0.4 1.5 Female Non-alc
1 0.9 Female Non-alc
1.1 1.6 Female Non-alc
1.2 1.7 Female Non-alc
1.3 1.7 Female Non-alc
1.6 2.2 Female Non-alc
1.8 0.8 Female Non-alc
2 2 Female Non-alc
2.5 3 Female Non-alc
2.9 2.2 Female Non-alc
1.5 1.3 Male Alc
1.9 1.2 Male Alc
2.7 1.4 Male Alc
3 1.3 Male Alc
3.7 2.7 Male Alc
0.3 1.1 Male Non-alc
2.5 2.3 Male Non-alc
2.7 2.7 Male Non-alc
3 1.4 Male Non-alc
4 2.2 Male Non-alc
4.5 2 Male Non-alc
6.1 2.8 Male Non-alc
9.5 5.2 Male Non-alc
12.3 4.1 Male Non-alc
;

proc sgscatter;
  plot met*gas / group=cat reg;
run;

ods graphics on;
proc glm plots=(diagnostics residuals);
class alc sex;
model met=gas sex gas*sex / solution;
output out=out cookd=c rstudent=t;
run;
ods graphics off;
```

```
proc print; run;

proc glm data=out(where=(t<5));
class alc sex;
model met=gas sex gas*sex / solution;
run;

ods graphics on; * suggests the square root;
proc transreg details;
 model boxcox(met)=identity(gas)|class(alc|sex);
*model boxcox(met)=identity(gas) class(alc|sex);
*model boxcox(met)=identity(gas) class(alc sex);
run;
ods graphics off;

proc sgscatter;
  plot sqrtmet*gas / group=cat reg;
run;

proc glm;
class alc sex;
model sqrtmet=gas|alc|sex / solution;
contrast "additive model fits?" gas*alc 1 -1, gas*sex 1 -1, alc*sex 1 -1 -1 1,
 gas*alc*sex 1 -1 -1 1;
run;

proc glm plots=diagnostics;
class alc sex;
model sqrtmet=gas alc sex;
run;

ods graphics on;
proc glm plots=diagnostics;
class alc sex;
model sqrtmet=gas sex / solution;
run;
ods graphics off;
```

3. $X_i \overset{i.i.d}{\sim}$ uniform$(\theta,\ \theta+|\theta|)$, $\theta \neq 0$.

(a)

$\because E(X) = \theta + \frac{|\theta|}{2} = \frac{3}{2}\theta\, I(\theta>0) + \frac{\theta}{2} I(\theta<0)$,

and $P(\bar{X}>0 \mid \theta>0) = P(\bar{X}<0 \mid \theta<0) = 1$.

$\therefore$ The MOME of θ is given by

$\hat{\theta}_{MOM} = \frac{2}{3}\bar{X}\, I(\bar{X}>0) + 2\bar{X} I(\bar{X}<0)$.

(b)

If $\theta>0$, then $X_i \overset{i.i.d}{\sim}$ uniform$(\theta,\ 2\theta)$, and the likelihood function is

$$f_{\underset{\sim}{X}}(\underset{\sim}{x};\theta) = \frac{1}{\theta^n} I\left(\frac{X_{(n)}}{2} \le \theta \le X_{(1)}\right),$$

and $\underset{\theta>0}{\operatorname{argmax}}\, f_{\underset{\sim}{X}}(\underset{\sim}{x};\theta) = \frac{X_{(n)}}{2}$.

If $\theta<0$, then $X_i \overset{i.i.d}{\sim}$ uniform$(\theta,\ 0)$, and the likelihood function is

$$f_{\underset{\sim}{X}}(\underset{\sim}{x};\theta) = \frac{1}{|\theta|^n} I(\theta < X_{(1)}),$$

and $\underset{\theta<0}{\operatorname{argmax}}\, f_{\underset{\sim}{X}}(\underset{\sim}{x};\theta) = X_{(1)}$.

Noting that $P\{\operatorname{sign}(X_i) = \operatorname{sign}(\theta)\} = 1$, for $i=1,\cdots,n$, one has the MLE of θ given by

$\hat{\theta}_{MLE} = \frac{X_{(n)}}{2} I(X_1>0) + X_{(1)} I(X_1<0)$.

(c) $\hat{\theta}_{MLE}$ is a consistent estimator, i.e., $\hat{\theta}_{MLE} \xrightarrow{P} \theta$. This is proved next.

If $\theta > 0$, the distribution of $X_{(n)}$ is of interest in study the property of $\hat{\theta}_{MLE}$.

For $\theta < x < 2\theta$, $F_{X_{(n)}}(x) = \left(\frac{x-\theta}{\theta}\right)^n$. Now consider the following probability for $0 < \varepsilon < \frac{\theta}{2}$,

$$P\left(\left|\frac{X_{(n)}}{2} - \theta\right| \geq \varepsilon\right)$$

$$= P\left(\theta - \frac{X_{(n)}}{2} \geq \varepsilon\right)$$

$$= P\{X_{(n)} \leq 2(\theta - \varepsilon)\}$$

$$= \left(1 - \frac{2\varepsilon}{\theta}\right)^n \longrightarrow 0 \quad \text{as } n \to \infty, \text{ since } 0 < \frac{2\varepsilon}{\theta} < 1.$$

If $\varepsilon \geq \frac{\theta}{2}$, then $P\left(\theta - \frac{X_{(n)}}{2} > \varepsilon\right) = 0$, $\forall n$.
Therefore, if $\theta > 0$, $\frac{X_{(n)}}{2} \xrightarrow{P} \theta$, as $n \to \infty$.

If $\theta < 0$, the distribution of $X_{(1)}$ is of interest.
Now that X_i i.i.d uniform $(\theta, 0)$, $F_X(x) = \frac{x-\theta}{-\theta}$, for $\theta < x < 0$, one has $F_{X_{(1)}}(x) = 1 - \left(\frac{x}{\theta}\right)^n$.

It follows that, for $0 < \varepsilon < -\theta$,

$$P(|X_{(1)} - \theta| < \varepsilon) = P(X_{(1)} - \theta < \varepsilon)$$

$$= 1 - \left(\frac{\varepsilon + \theta}{\theta}\right)^n \longrightarrow 1 \quad \text{as } n \to \infty,$$

Since $\frac{\varepsilon + \theta}{\theta} \in (0, 1)$.

If $\varepsilon \geq -\theta$, then $P(X_{(1)} - \theta < \varepsilon) = P(X_{(1)} < \varepsilon + \theta) = 1$, $\forall n$,
since now $\varepsilon + \theta \geq 0$ and $P(X_{(1)} < 0) = 1$.
This shows that, if $\theta < 0$, $X_{(1)} \xrightarrow{P} \theta$.

Finally, noting that $P(X_1>0)=1$ if $\theta>0$ and
$P(X_1<0)=1$ if $\theta<0$,
one can concludes that $\theta_{MLE} \xrightarrow{P} \theta$.

#

#4.1 (a) $M_1, M_2, \ldots, M_B$ n i.i.d with

$$\text{var}(M_b) = \sigma_M^2 \quad \forall b = 1, 2, \ldots, B$$

S_M^2 = sample variance of $M_1, M_2, \ldots, M_B$

We know $E(S_M^2) = \sigma_M^2$ so S_M^2 is at least an unbiased estimator of σ_M^2.

Also $S_M^2 \xrightarrow{p} \sigma_M^2$ as $B \to \infty$.

(b) $N(\mu, \sigma^2)$ family w/ σ^2 known (location family)

✓ $\bar{X}$ is a complete sufficient statistic for this family

✓ $M - \bar{X}$ is a location-invariant statistic & hence ancillary.

Basu's $\Rightarrow$ $\bar{X} \perp M - \bar{X}$

$$\Rightarrow \text{var}(M) = \text{var}(M - \bar{X} + \bar{X})$$
$$= \text{var}(M - \bar{X}) + \text{var}(\bar{X}) - 2\,\text{cov}(M - \bar{X}, \bar{X})$$
$$\Rightarrow = \text{var}(M - \bar{X}) + \text{var}(\bar{X}).$$

(c) By the same reasoning in (a),

$$E(S_T^2) = \text{var}(M - \bar{X})$$

and

$$S_T^2 \xrightarrow{p} \text{var}(M - \bar{X}), \text{ as } B \to \infty.$$

Therefore, also

$$E(\hat{\sigma}_M^2) = E(S_T^2 + \sigma^2/n)$$
$$= E(S_T^2) + \sigma^2/n$$
$$= \text{var}(M - \bar{X}) + \text{var}(\bar{X})$$
$$= \text{var}(M) = \sigma_M^2$$

and

$$\hat{\sigma}_M^2 = S_T^2 + \sigma^2/n \xrightarrow{p} \text{var}(M) = \sigma_M^2$$

$S_T^2 \xrightarrow{p} \text{var}(M - \bar{X})$

as $B \to \infty$.

(d) A rigorous comparison could look at

$$\text{var}(S_M^2) \lessgtr \text{var}(\hat{\sigma}_M^2)$$

since both estimators are unbiased. This, however is not necessary.

Because

$$\text{var}(M) = \text{var}(M - \bar{X}) + \text{var}(\bar{X})$$

clearly S_T^2 is going to estimate $\text{var}(M - \bar{X})$ more precisely than S_M^2 will $\text{var}(M)$

Because $\text{var}(\bar{X}) = \sigma^2/n$ (constant), the precision of S_T^2 as an estimator for $\text{var}(M - \bar{X})$ is the same as $\hat{\sigma}_M^2$ as an estimator for σ_M^2

Thus $\hat{\sigma}_M^2$ should be more precise.

5. $X = (X_1, \cdots, X_n)$,

X_i: i.i.d $f_X(x)$.

For testing $H_0: f = f_0$ vs. $H_1: f = f_1$,

$\delta(X)$ is the decision function (i.e., the test function) associated with the UMP test of size $\alpha \in (0,1)$.

By the Neyman-Pearson (NP) Lemma,

$$\delta(X) = I\left\{\frac{f_1(X)}{f_0(X)} > c\right\}, \text{ where } c \text{ satisfies}$$

$$P\left\{\frac{f_1(X)}{f_0(X)} > c \,\middle|\, f_0\right\} = \alpha \in (0,1).$$

Moreover, it is known that

$$P\left\{\frac{f_1(X)}{f_0(X)} > c \,\middle|\, f_1\right\} = \beta \in (0,1) \qquad (\Delta)$$

Now consider testing $H_0^*: f = f_1$ vs. $H_1^*: f = f_0$.

By the NP Lemma, the UMP test of size $(1-\beta)$ has the following test function,

$$\delta^*(X) = I\left\{\frac{f_0(X)}{f_1(X)} > d\right\}, \text{ where } d \text{ satisfies}$$

$$P\left\{\frac{f_0(X)}{f_1(X)} > d \,\middle|\, f_1\right\} = 1-\beta, \quad \text{i.e.,}$$

$$P\left\{\frac{f_1(X)}{f_0(X)} \geq \frac{1}{d} \,\middle|\, f_1\right\} = \beta$$

According to (Δ), one has $d = \frac{1}{c}$.

Therefore,

$$\delta^*(X) = I\left\{\frac{f_0(X)}{f_1(X)} > \frac{1}{c}\right\}$$

$$= I\left\{\frac{f_1(X)}{f_0(X)} \leq c\right\}$$

$$= 1 - \delta(X).$$

#

1 Problem 6

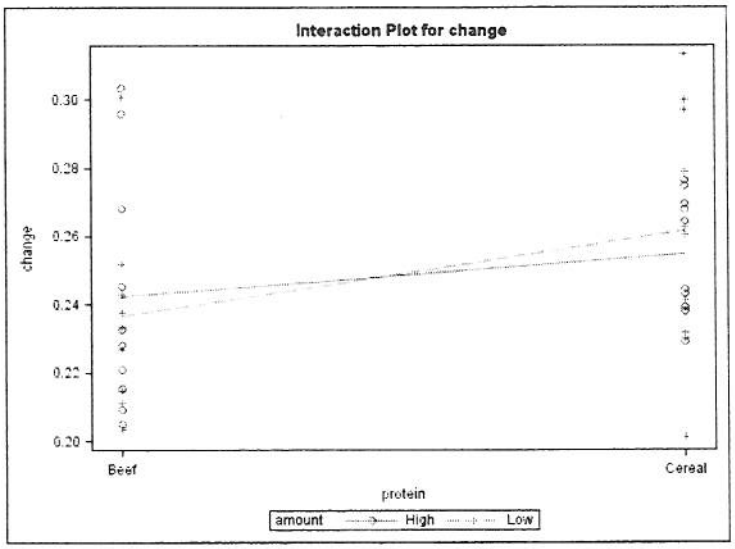
A biologist designed an experiment to assess the weight gain in $n = 40$ rats fed diets comprised of four different combinations of two protein sources and two protein amounts. This is a completely randomized design with ten rats randomly allocated to each of the four treatments. The variables are:

- PreWt – The weight before the experiment (grams).
- PostWt – The weight after the experiment (grams).
- Protein – The protein source: either Beef or Cereal.
- Amount – The amount of protein: either High or Low.

The data is available for download at http://www.stat.sc.edu/~hansont/rat_data.txt. Build a model that best describes the relationship between the weight gain *as a proportion of intial weight* and the factors Protein and Amount. Make sure that you carefully assess all model assumptions and write a succinct, coherent, and complete summary of your analysis.

1.1 Analysis with the original response

This is a balanced, 2×2 design with replication. An interaction plot shows roughly parallel lines with overlap between the two levels of amount, but a clear increase in weight gain from beef to cereal protein. There is *a lot* of variability in the data about the cells means – probably only protein will be significant.



An initial fit using proportion of weight gained gives

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.00389776	0.00129925	1.44	0.2459
Error	36	0.03237506	0.00089931		
Corrected Total	39	0.03627281			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
protein	1	0.00348471	0.00348471	3.87	0.0568
amount	1	0.00000177	0.00000177	0.00	0.9649
protein*amount	1	0.00041128	0.00041128	0.46	0.5032

At first glance, the F-statistic that tests whether anything is significant gives p=0.25, seemingly hopeless. However, the Type III SS table and tests show that protein seems to be important – unnecessary noisy effects are clouding the signal. Testing whether amount and amount*protein can be dropped simultaneously yields a p=0.80; we can drop these effects.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
drop amount & amount*protein	2	0.00041305	0.00020652	0.23	0.7960

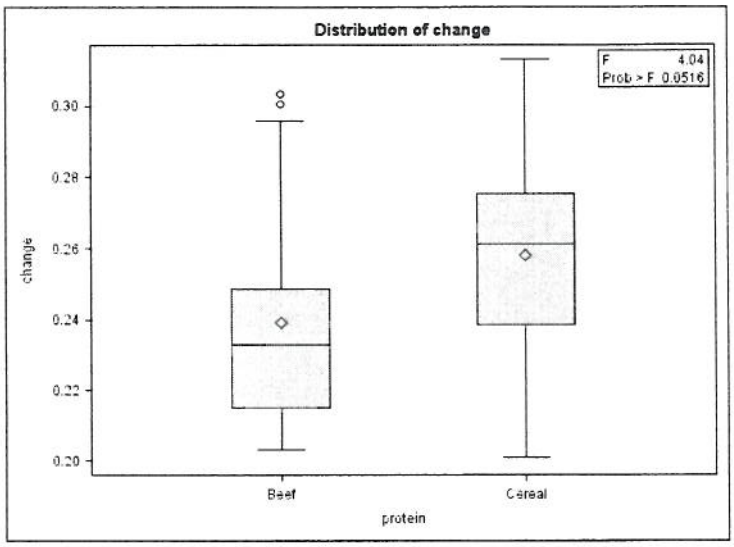
Refitting the model with only protein gives

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00348471	0.00348471	4.04	0.0516
Error	38	0.03278811	0.00086284		
Corrected Total	39	0.03627281			

R-Square	Coeff Var	Root MSE	change Mean
0.096069	11.81069	0.029374	0.248709

Parameter		Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		0.2580424173 B	0.00656828	39.29	<.0001	0.2447456383	0.2713391963
protein	Beef	-.0186673659 B	0.00928894	-2.01	0.0516	-.0374718510	0.0001371193
protein	Cereal	0.0000000000 B	.	.	.	.	.

Protein is almost significant at the 5% level. The cereal diet (almost) significantly increases the proportion of weight gained by an estimated 1.9%; we are 95% "confident" that this proportion is between -0.0137% and 3.7%. Note that $R^2 = 0.096$; there is a lot of variability as seen in the interaction plot. A simple boxplot helps visualize the difference in protein types:



The boxplot shows two outliers among the beef protein rats. We expect one outlier in a sample size of 150, so two in 20 indicates some heavy-tailedness. We can try the Mann-Whitney-Wilcoxon test to be safe:

Wilcoxon Scores (Rank Sums) for Variable change
Classified by Variable protein

protein	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Beef	20	331.0	410.0	36.968455	16.550
Cereal	20	489.0	410.0	36.968455	24.450

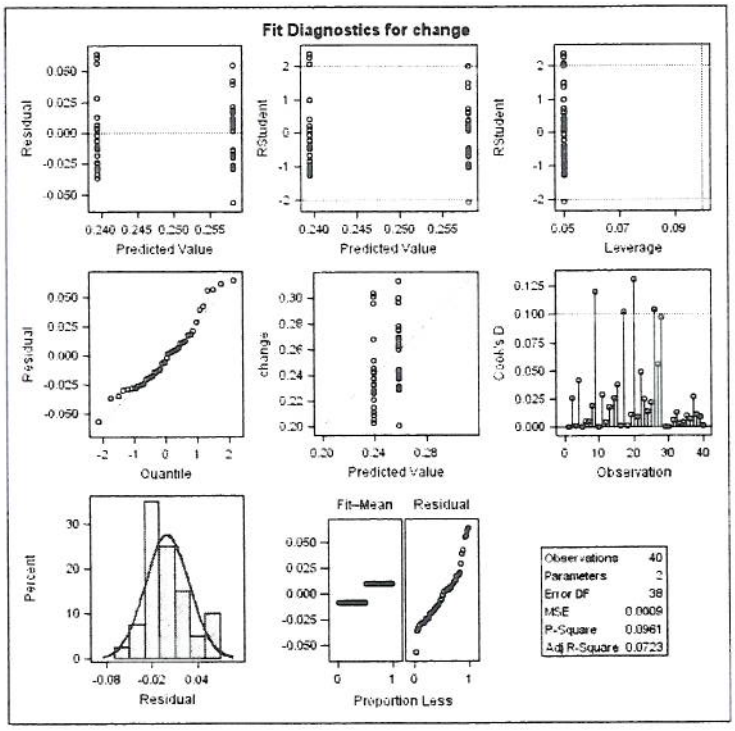
Wilcoxon Two-Sample Test

Statistic	331.0000

```
Normal Approximation
Z                      -2.1234
One-Sided Pr <  Z       0.0169
Two-Sided Pr > |Z|      0.0337

t Approximation
One-Sided Pr <  Z       0.0201
Two-Sided Pr > |Z|      0.0401

Z includes a continuity correction of 0.5.

     Kruskal-Wallis Test

Chi-Square              4.5666
DF                           1
Pr > Chi-Square         0.0326

               Hodges-Lehmann Estimation
               Location Shift    -0.0226

                              Interval        Asymptotic
 95% Confidence Limits        Midpoint    Standard Error
   -0.0380     -0.0014         -0.0197            0.0094
```

We have sinificance for either a 2-sided or one-sided test, and there are no assumptions going into the test. This would actually be a good stopping point! Also note that the Hodges-Lehmann shows an estimated 2.3% increase in bodyweight using cereal rather than beef protein, and we are 95% "confident" that the true population mean difference is between 0.1% and 3.8%.

In terms of the previous normal-errors model, a standard diagnostic panel shows reasonable model fit, i.e. no extreme outliers, no overly influential points, and reasonably normal residuals.



1.2 Transforming the response

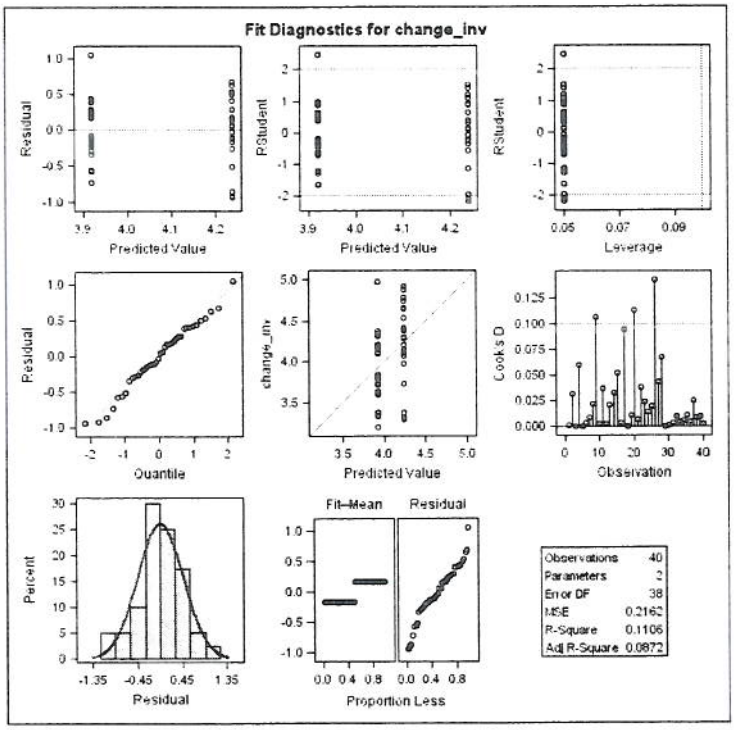
Let's investigate a Box-Cox transformation of the response. Using proc transreg, the MLE is $\lambda = -1$ (i.e. 1/change or 1/proportion) but $\lambda = 1$ (no transformation) is also within the 95% CI. When we use 1/proportion as the response, we again accept that we can drop the protein*amount and amount effects. Now protein is now significant (p=0.041) at the 5% level.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.02201443	1.02201443	4.73	0.0360
Error	38	8.21675370	0.21623036		
Corrected Total	39	9.23876813			

R-Square	Coeff Var	Root MSE	change_inv Mean
0.110622	11.40140	0.465006	4.078499

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		3.918653815	B	0.10397845	37.69	<.0001
protein	Beef	0.319689605	B	0.14704773	2.17	0.0360
protein	Cereal	0.000000000	B	.	.	.

The diagnostic panel again shows good fit of the model; the normal probability plot looks even straighter than the response on the original scale:



Other possible transformations are the logit(x) and $\sin^{-1}(\sqrt{x})$ transformations; the latter is the variance stabilizing transformation for proportions. These lead to similar conclusions as the inverse 1/proportion (p=0.045 and p=0.048).

2 SAS code

```
*************************************************************************
Rats ANOVA data
*************************************************************************;

data rats;
input protein$ amount$ prewt postwt;
change=(postwt-prewt)/prewt;
logit=log(change/(1-change));
arcsinroot=arsin(sqrt(change));
change_inv=1/change;
datalines;
Beef Low 372 462
Beef Low 360 436
Beef Low 386 476
Beef Low 315 379
Beef Low 362 448
Beef Low 225 276
Beef Low 286 358
Beef Low 419 509
Beef Low 316 411
Beef Low 321 399
Beef High 349 422
```

```
Beef High 447 549
Beef High 548 666
Beef High 388 492
Beef High 395 476
Beef High 436 543
Beef High 338 438
Beef High 374 461
Beef High 530 647
Beef High 366 477
Cereal Low 444 551
Cereal Low 320 415
Cereal Low 422 519
Cereal Low 287 367
Cereal Low 423 521
Cereal Low 368 442
Cereal Low 247 321
Cereal Low 214 281
Cereal Low 342 431
Cereal Low 221 279
Cereal High 401 499
Cereal High 311 385
Cereal High 209 265
Cereal High 412 523
Cereal High 344 439
Cereal High 362 450
Cereal High 358 440
Cereal High 322 399
Cereal High 313 399
Cereal High 348 440
;

options nocenter;
ods graphics on;
proc glm plots=diagnostics;
 class protein amount;
 model change=protein amount protein*amount;
 contrast "drop amount & amount*protein" amount 1 -1, amount*protein 1 -1 -1 1;
run;
ods graphics off;

ods graphics on;
proc glm plots=diagnostics;
 class protein amount;
 model change=protein  / solution clparm;
run;
ods graphics off;

* Wilcoxon-Mann-Whitney test shows significance!;

proc npar1way hl; * hl adds Hodges-Lehmann confidence interval for delta;
 class protein; var change; run;

* Box-Cox transformation MLE is -1, but 1 is within 95% CI;
```

```
proc transreg;
 model boxcox(change) = class(protein amount protein*amount);
run;


ods graphics on;
proc glm;
 class protein amount;
 model change_inv=protein amount protein*amount;
 contrast "drop amount & amount*protein" amount 1 -1, amount*protein 1 -1 -1 1;
run;
ods graphics off;

* get significance using 1/change;
ods graphics on;
proc glm plots=diagnostics;
 class protein amount;
 model change_inv=protein / solution;
run;
ods graphics off;

* also get significance using logit;
ods graphics on;
proc glm;
 class protein amount;
 model logit=protein;
run;
ods graphics off;

* also get significance using variance stabilizing transform;
ods graphics on;
proc glm;
 class protein amount;
 model arcsinroot=protein;
run;
ods graphics off;
```