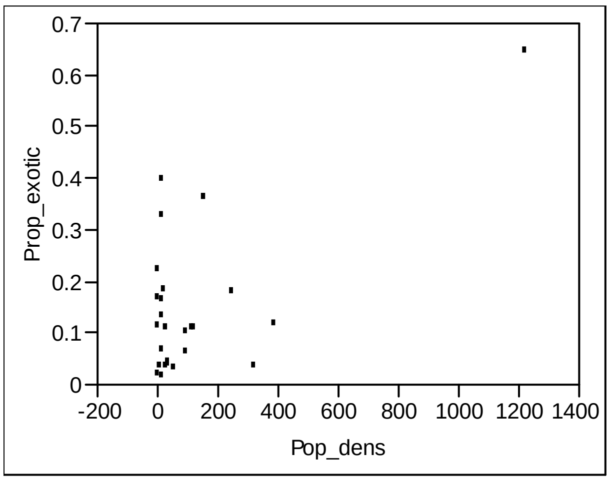
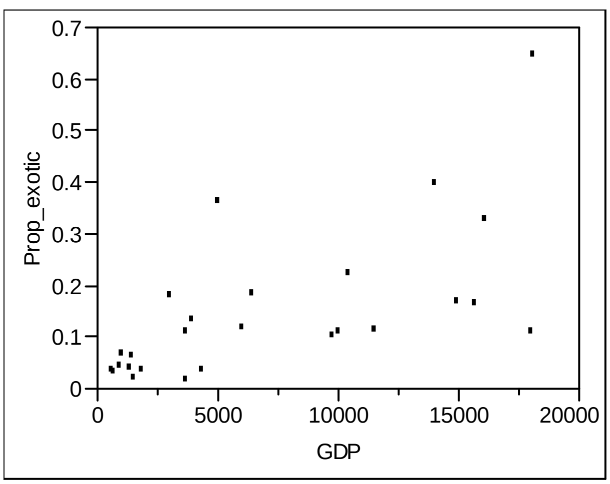
Problem set 3 solutions

1a) The regression using GDP is somewhat better than the one using population density, for two reasons. First, the coefficient of determination (R^2) is slightly higher (0.41 vs. 0.40), although the difference is small enough that I wouldn't make much of it. More important is the fact that the relationship is not being driven by a single point – so we have greater confidence that there really is a relationship. (see scatter plots below)



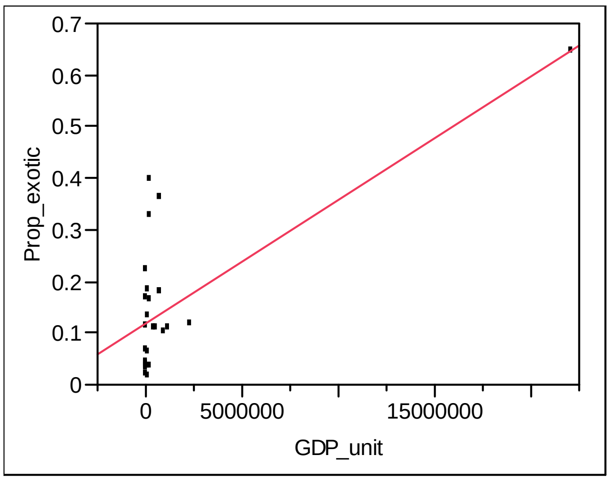


The linear regression on GDP suggests that, all else being equal, every $1000 increase in the per-capita GDP leads to a 1.6 percentage point increase in the percentage of plant species that are exotic. As an indicator of economic activity and affluence, the per-capita GDP may be associated with increased disturbance of the natural habitat as well as increased trade. These factors would be the 'true' variables that influence both affluence and invadedness. For the most part, I would not expect GDP itself to directly affect invadedness, with one exception: greater affluence may lead directly to increased demand for exotic ornamental plants for private and public gardens.

1b) Because the "GDP" variable is a per-capita measure, multiplying this by population density gives a measure of GDP per unit area, in units of dollars per square kilometer.

[Go to the data table in JMP. Right click to 'add new column'. Click on Column Properties ->Formula -> Pop_dens*GDP->OK ->OK. Double click on Column10 and rename as GDP_unit. Alternatively, you can create the variable in excel and then re-import the data sheet]

Using this as the independent variable in the regression improves the quantitative fit even further ($R^2 = 0.50$). However, the relationship is once again dominated by the difference between Bermuda and the rest of the world. The problem is that Bermuda has a much higher population density than any of the other countries in the sample. We simply have no way of knowing what would happen in a country with a population density midway between Bermuda and the other countries – would its invadedness be intermediate, closer to Bermuda, or closer to the other countries? Trying to make predictions in such a situation is extremely risky.



Linear Fit

Prop_exotic = 0.118451 + 2.4001e-8*GDP_unit

Summary of Fit

RSquare	0.500209
RSquare Adj	0.479384
Root Mean Square Error	0.105211
Mean of Response	0.146536
Observations (or Sum Wgts)	26

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.118451	0.021414	5.53	<.0001
GDP_unit	2.4001e-8	4.897e-9	4.90	<.0001

One solution might be to eliminate Bermuda and re-run the regression. In this case, only the regression on per-capita GDP has a P value less than 0.05 – there is essentially no association with the other two variables. This suggests to me that invadeness is not influenced by the intensity of disturbance on a per-area basis, as would be shown by a regression on population density or GDP per area. Rather, affluence itself seems to be a predictive variable – which perhaps lends credence to the ornamental plant issue that I raised above.

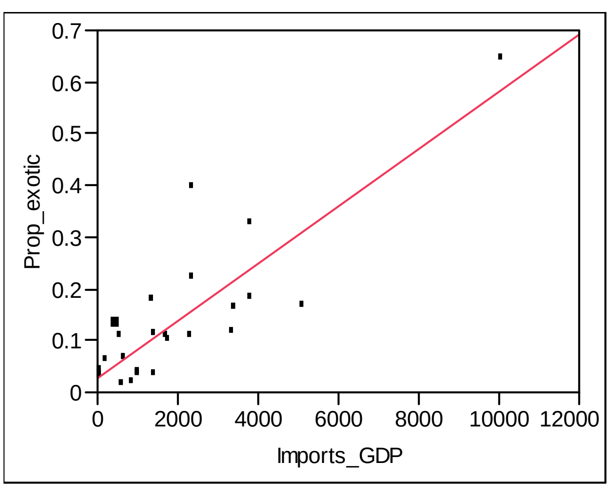
1c) I multiplied per-capita GDP by M_imports and divided by 100 to get the per-capita imports in dollars. I multiplied this by population density to get the imports per sq km and multiplied that by area to get total imports.

[Use the same procedure for creating variables as in 1b]

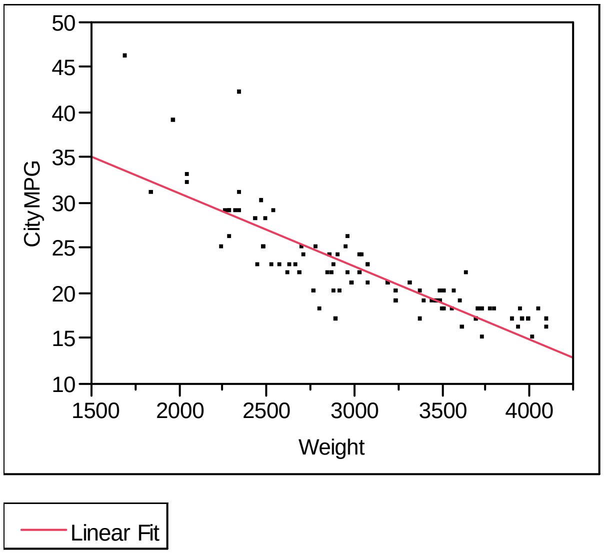
I then deleted Guam, which had no import data. I ran regressions with each of the new variables, and re-ran the models from above with Guam missing. Here are the coefficients of determination for the models:

Model	R square
Imports per capita	0.71
Imports per km^2	0.55
Imports total	0.00
Pop_dens	0.43
GDP	0.49
GDP/km^2	0.56

The model with Imports per capita appears to do the best; imports per unit area is at least as good as the other models we have considered. Total imports has no explanatory power. Looking at the scatterplots (below), there is a clear linear relationship between the proportion of exotic species and the per-capita imports. Thus from a statistical perspective, imports per capita is the best predictor of invadeness out of the ones we have looked at.



2a) We can reject the null hypothesis at $P \leq 0.01$ so we have strong evidence that city mpg decreases with increases in weight.



Linear Fit

CityMPG = 47.057182 - 0.0080362*Weight

Summary of Fit

RSquare	0.709764
RSquare Adj	0.706539
Root Mean Square Error	3.055012
Mean of Response	22.40217
Observations (or Sum Wgts)	92

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	47.057182	1.692139	27.81	<.0001
Weight	-0.008036	0.000542	-14.84	<.0001

From our data we know that the standard deviation of Weight is 591.21. To calculate the predicted change in X we use:

ΔCityMPG = - 0.0080362*(591.21) = - 4.75 ; so increasing weight by one standard deviation decreases mpg in the city by 4.75

Moments

Mean	3067.9891
Std Dev	591.21177
Std Err Mean	61.638087
upper 95% Mean	3190.4256
lower 95% Mean	2945.5526
N	92

2b) In both cases we estimate a negative effect of car weight on city mpg.

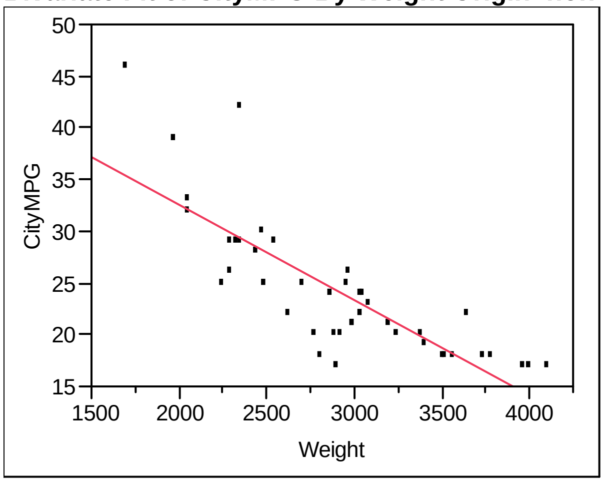
[In JMP do Analyze -> Fit Y by X -> Y is CityMPG X is Weight - > By Origin]

The results can be summarized as follows:

	β_1
non-US	-0.0092
US	0.0065

The different estimates of the regression slopes suggest that the fuel efficiency of foreign cars is relatively more sensitive to weight. (Note: To test whether the differences in β_1 are statistically significant we would use interaction terms in a multiple regression framework).

Bivariate Fit of CityMPG By Weight Origin=non-US



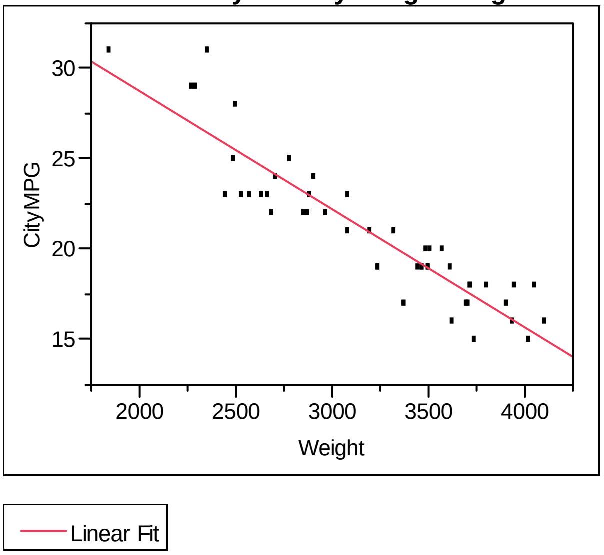
Summary of Fit

RSquare	0.664724
RSquare Adj	0.656741
Root Mean Square Error	3.930201
Mean of Response	23.97727
Observations (or Sum Wgts)	44

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	50.95311	3.014975	16.90	<.0001
Weight	-0.00921	0.001009	-9.13	<.0001

Bivariate Fit of CityMPG By Weight Origin=US



Linear Fit

CityMPG = 41.706057 - 0.0064932*Weight

Summary of Fit

RSquare	0.844201
RSquare Adj	0.840814
Root Mean Square Error	1.593712
Mean of Response	20.95833
Observations (or Sum Wgts)	48

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	41.706057	1.334148	31.26	<.0001
Weight	-0.006493	0.000411	-15.79	<.0001

3c) Do the following to create a variable that is the natural log of Weight:

[Go to the data table in JMP. Right click to 'add new column'. Click on Column Properties ->Formula -> Transcendental->Log ->Weight-> OK->. Double click on Column10 and rename as LnWeight. Alternatively, you can create the variable in excel and then re-import the data-sheet]

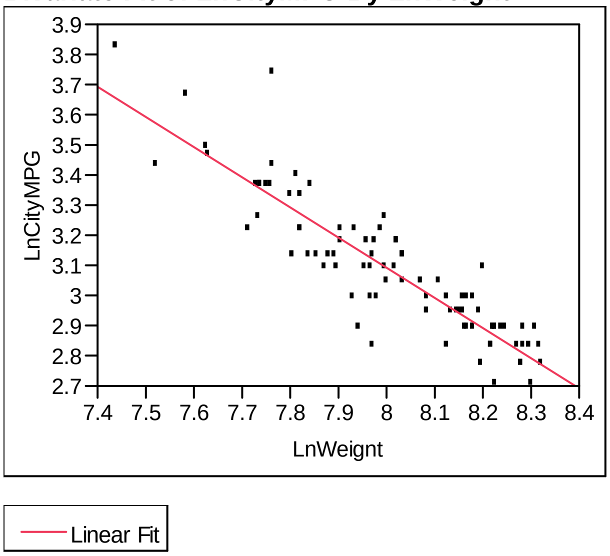
Follow the same procedure to take the natural log of CityMpg.

The R^2 for this log-log regression is 0.80 compared to 0.71 so the log model is a better fit. The coefficient estimate of β_1 in the log model is -0.99. To calculate the predicted percent change in CityMPG due to a 15 percent decrease in Weight use:

%ΔCityMPG = - 0.9985471*(-15) = 14.97

So our model predicts that fuel efficiency will increase by 14.97 percent if car weight decreases by 15 percent.

Bivariate Fit of LnCityMPG By LnWeignt



Linear Fit

LnCityMPG = 11.080131 - 0.9985471*LnWeignt

Summary of Fit

RSquare	0.805823
RSquare Adj	0.803666
Root Mean Square Error	0.099545
Mean of Response	3.082472
Observations (or Sum Wgts)	92

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	3.7010494	3.70105	373.4953
Error	90	0.8918303	0.00991	**Prob > F**
C. Total	91	4.5928797		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.080131	0.413959	26.77	<.0001
LnWeignt	-0.998547	0.051669	-19.33	<.0001

3d) We could get more reliable estimates of the effects of weight on fuel efficiency if we control for other factors (in multiple variable regression) that are correlated with weight and that also affect fuel efficiency. Such characteristics may include engine size, fuel tank capacity, and horsepower. Weight alone, however, is in fact a good predictor of fuel efficiency – our R^2 from these single-variable regressions are quite high.