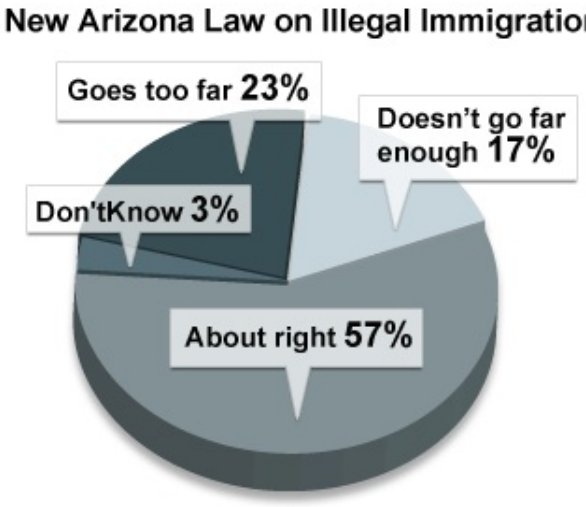
1. This graphic appeared last year at cbsnews.com to summarize a poll of Americans. What do you think is the best aspect of this graphic? Explain in one sentence. What do you think is the worst aspect of this graphic? Explain in one sentence. [6 points]



The best aspect of this graphic is the use of clear labels, as opposed to, say, a legend with indistinguishable colors. The worst is the unnecessary and potentially misleading 3-D-ness of the pie-chart.

2. Answer each of the following questions very concisely, in the space provided. [2 points each; 12 points total]

   (a) State the normality assumption underlying a simple linear regression.

   We assume $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ independently for all i; in other words, the response is normally distributed with the mean for each observation being a linear function of the covariates.

   (b) Define "estimand."

   What we're trying to estimate; an unknown quantity for which we want to provide inferences.

(c) Explain why R^2 tends toward over-fitting when used as a model selection criterion, and explain why BIC does not.

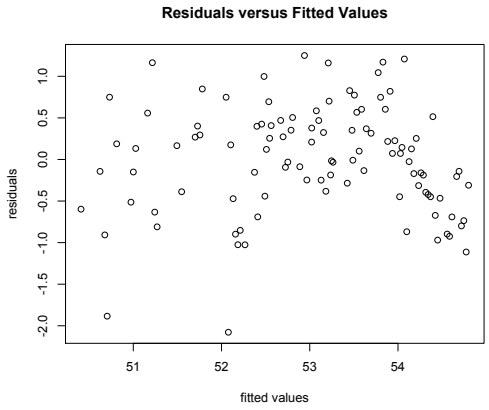
R^2 increases for every additional explanatory variable added to a model, no matter what the new variable is. Thus R^2 will never favor a model that consists of only a subset of the variables being considered (a "nested" model). BIC offsets the fit of the model with a penalty for the number of variables being used, and in theory will always select the correct model if the sample size is large enough.

(d) Some athletes believe it is bad luck to appear on the cover of *Sports Illustrated*. In 2001, writer Alexander Wolff found that of the magazine's approximately 2,500 previous covers, 37% featured athletes whose performance markedly declined following the magazine cover. Is there a *Sports Illustrated* jinx? Explain briefly.

This sounds like regression to the mean. If their performance the next season were on average just as good as it was the previous season, we'd see an explosion in the distribution of performance within the sport, when in fact it stays relatively constant from year to year.

(e) What would you do if the residual v. fitted value plot from your estimated regression model looked like the one below?

Toward larger fitted values, the residuals tend to be negative and with smaller variance. This suggests a transformation of response variable—exponentiation of y might yield a more appropriate fit.


Residuals versus Fitted Values
residuals
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
51
52
53
54
fitted values

(f) Suppose that $X_i \sim N(\mu_X, \sigma^2)$, $i = 1, \ldots, n_X$, and $Y_j \sim N(\mu_Y, \sigma^2)$, $j = 1, \ldots, n_Y$. Consider the following two statements:

$$(i) Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y})$$

$$(ii) Var(\bar{X}) + Var(\bar{Y}) = \sigma^2 \left(\frac{1}{n_X} + \frac{1}{n_Y} \right)$$

What assumption underlies statement (ii)?

Independence of $\bar{X}$ and $\bar{Y}$.

3. A start-up drug company conducted a randomized trial to test its new drug, a sleeping pill. The estimand is the difference between the mean potential outcome that would be observed if each patient in the study took the new drug, and the mean potential outcome that would be observed if each patient in the study took a placebo pill. The outcome is the number of hours a patient sleeps after taking the pill. [3 points each; 12 points total]

(a) You conduct an unpooled t-test to compare the mean number of hours slept in the new drug and placebo groups. How would you interpret the corresponding confidence interval?

The resulting CI would be a (overly conservative) range of plausible values for our estimand, the difference in mean potential outcomes for this particular sample. If we could somehow repeatedly redo the experiment with different randomizations and get the resulting confidence intervals, (more than) 95% of these intervals would contain our estimand.

(b) Consider the linear regression model $\mu\{Hours|Newdrug\} = \beta_0 + \beta_1 Newdrug$, where $Newdrug$ is equal to 1 if a patient received the new drug and equal to 0 if the patient received the placebo. How would the least squares estimate, $\hat{\beta}_1$, compare to the simple difference in observed means in the new drug and placebo groups?

They would be the same, as β_1 just represents the difference between the means of both groups.

(c) Insomnia tends to increase as people age. Consider the model $\mu\{Hours|Newdrug, Age\} = \beta_0 + \beta_1 Newdrug + \beta_2 Age$. If the sample size is large or the model is approximately correct, could $\hat{\beta}_1$ from this model be a useful estimate of the causal effect? Explain in one sentence.

Yes: here $\hat{\beta}_1$ may give us a more precise (lower variance) estimate of the treatment effect, although it will not be unbiased.

(d) Suppose that the start-up also conducted a non-randomized, observational study: one day, free doses of the new sleeping pill were made available to employees of a large pharmaceutical company (because the start-up hopes the large company will be impressed!). The following day, all of the large company's employees reported the number of hours they had slept and whether they had taken the new sleeping pill. Consider the model $\mu\{Hours|Newdrug, Age\} = \beta_0 + \beta_1 Newdrug + \beta_2 Age$. Would you expect the conclusion drawn from this analysis to be the same as in the randomized trial? Explain why or why not in one or two sentences.

We cannot expect to draw the same conclusions. If we didn't assign treatment randomly, it's very possible that a meaningful covariabe could be confounded with the treatment received. For example, perhaps the only people who took the pill were insomniacs who get much less sleep (even with the pill) than normal sleepers.

4. Consider a simple linear regression model where we know that $\beta_0 = 0$. Write down the expression for the sum of squared residuals and solve for the least squares estimate of the slope. [10 points]

Let $SSR(\beta_1) = \sum_{i=1}^{n}(Y_i - \beta_1 X_i)^2$.

$$\begin{aligned} \frac{dSSR(\beta_1)}{d\beta_1} &= \sum_{i=1}^{n} 2(Y_i - \beta_1 X_i)(-X_i) \\ &= \beta_1 \left(2\sum_{i=1}^{n} X_i^2 \right) - \left(2\sum_{i=1}^{n} X_i Y_i \right) \end{aligned} \tag{1}$$

Setting this to 0 and solving for β_1 yields the minimum:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

(we know we have a minimum at $\hat{\beta}_1$ since examination of (1) shows the second derivative of SSR is positive.)

5. (a) On the survey at the the first day of class, you indicated your younger parent's number of children, including yourself. Suppose we want to know whether family size has noticeable lasting associations with wake-up time.

	Group A	Group B	Group C
Number of children	1	2	3+
Sample size	22	47	20
Mean Wake-up Time	9.4	8.5	9.1

Consider an equal means model, a separate means model, and a simple linear regression model (with the predictor equal to 1, 2, or 3). The sums of squared residuals and degrees of freedom are shown in the table below.

Model	SSR	df
Equal Means	186	88
Simple Linear Regression	185	87
Separate Means	175	86

Which model appears most appropriate? Is family size associated with wake-up time? Show your work. [8 points]

We can find the most appropriate model by carrying out two F-tests. 1) Compare the the equal means model (reduced model) to the simple linear regression (full model). This yields an F-stat = $\frac{186-185}{185/87} = 0.47$ **and we fail to reject the equal means model. 2) Compare the equal means model (reduced model) to the separate means model (full model). This yields an F-stat =** $\frac{(186-175)/2}{175/86} = 2.7$ **and** p**-value** $= 0.07$**. We again fail to reject the equal means model at the 0.05 level. We conclude the equal means model is the most appropriate. It seems family size is not associated with wake-up time.**

(b) Problem 1 on the midterm was selected so that p-value of the t-test would be between 0 and 0.10, to get at the idea of practical versus statistical significance. To find an appropriate hypothesis test, a total of seven tests were conducted based on the survey collected at the first class. If we assume that all seven null hypotheses were true and that the tests were independent, what is the probability that at least one would be appropriate for the midterm? If we wanted to create confidence intervals corresponding to the seven tests such that the simultaneous coverage rate is 95%, which method would be appropriate? Explain your choice in one sentence. [6 points]

The probability that at least one test would yield a p-value between 0 and 0.10 is one minus the probability that none would yield a p-value between 0 and 0.10. If the null hypothesis is true, the p-values follow a uniform distribution. Using the fact that the tests are independent, the probability is $1 - 0.90^7 = 0.52$. If we wanted to create confidence intervals so that the simultaneous coverage was 95%, we should choose Bonferroni as these tests are unrelated.

6. The data in this problem was collected as part of an evaluation of a job training program in the 1970's. Information about 528 workers is included. The table below provides the sample means and square roots of sample variances for some of the variables. [25 points total for this problem]

	$SampleMean$	$\sqrt{SampleVariance}$
$Income74$	20800	11900
$log(Income74)$	9.75	0.718
$Income75$	20500	11300
$log(Income75)$	9.71	0.776
$Educ$	13.1	3.11
$log(Educ)$	2.53	0.294

(a) Suppose that the research goal is to estimate the relationship between years of education and 1975 income. The plot shows the relationship between years of education and 1975 income, where both the predictor and the response are on log scales: $log(Educ)$ and $log(Income75)$. Why my we choose to log transform both the predictor and the outcome; what might the plot of income v. education have looked like? Describe in one

or two sentences or sketch a picture. [2 points]

We would log both the predictor and the outcome if the scatter-plot suggested non-linearity and/or heteroskedasticity. The plot of income v. education might have been funnel-shaped and non-linear.

(b) The outlier at the left side of the plot has $log(Educ) = 0$ and $log(Income75) = 9.69$. Concisely justify removing this outlier from the analysis, citing the concepts we discussed in class. [3 points]

The point seems to be an outlier with respect to the *years of education* variable. We are justified in removing it from the analysis, so long as we restrict the scope of inference of our analysis to individuals with at least $exp(1.1) \approx 3$ years of education. Further, though the datapoint seems to have high leverage, it does not appear to be an outlier for a regression of *log(income)* on *log(Years Education)*. Its influence is thus moderate, and analysis results with or without it likely not drastically different.

(c) After removing this outlier, the estimated simple linear regression equation for $log(Income75)$ can be summarized as below.

	$Estimate$	$StandardError$
$Intercept$	7.72	0.306
$log(Educ)$	0.784	0.120

Construct a 95% confidence interval for the estimated slope coefficient (on the current scale, with both the predictor and the response logged). [2 points]

The 95% confidence interval for the estimated slope coefficient is given by: $\hat{\beta}_1 \pm t_{n-2}(1-\frac{\alpha}{2}) \cdot SE(\hat{\beta}_1)$. Since $\hat{\beta}_1 = 0.784,\ t_{528-1-2}(0.975) = 1.964$ (note, based on the t table, could have picked $t_{100}(0.975) = 1.984$ for conservativeness' sake), and $SE(\hat{\beta}_1) = 0.120$, the confidence interval is (0.548, 1.02).

(d) How would you interpret the estimated slope coefficient and interval in 6c on the original scale, in terms of years of education and 1975 income? [5 points]

A k-multiplicative change in number of years of education is associated with a multiplicative factor in the median income of k^{β_1}. Specifically, doubling one's number of years of education is associated with a multiplicative factor in the median income of $2^{0.784} = 1.72$, or, equivalently, an increase in median income by 72%. A 95% confidence interval for the multiplicative factor in the median is $2^{0.548}$ to $2^{1.02}$, or 1.46 to 2.03.

(e) Suppose that the next goal is predicting 1975 income from 1974 income. For a simple linear regression of 1975 income on 1974 income, $\hat{\sigma} = 5480$. Note that the above outlier was included here (because it did not stand out in this case), so the sample size is 528. What is the sample correlation between 1974 and 1975 income? [8 points]

Recall that $R^2 = r^2$. Hence $r_{1974,1975} = \sqrt{R^2}$ where R^2 is the coefficient of determination in the regression of 1975 income on 1974 income. Now,

$$R^2 = \frac{SST - SSR}{SST}.$$

Also, $SST = s^2_{1975} \cdot (n-1) = (11300)^2 \cdot (528-1)$, and $SSR = \hat{\sigma}^2 \cdot (n-2) = (5480)^2 \cdot (528-2)$.
Thus:

$$\begin{aligned} R^2 &= \frac{SST - SSR}{SST} \\ &= \frac{(11300)^2 \cdot 527 - (5480)^2 \cdot 526}{(11300)^2 \cdot 527} \\ &= 0.765. \end{aligned}$$

As such, $r_{1974,1975} = \sqrt{0.765} = 0.875$.

(f) Let Y' represent standardized $Income75$ and X' represent standardized $Income74$, such that Y' and X' each have mean 0 and variance 1. For a simple linear regression of Y' on X', $\hat{\sigma} = 0.485$. What are the estimated coefficients? On average, for people with 1974 income three standard deviations above the mean, how many standard deviations above the mean will income be in 1975 (so, what is $E(Y'|X' = 3)$?)? Create a 95% prediction interval for $Y'|X' = 3$. Is the result intuitive? Explain in one or two sentences. [5 points]

If both X and Y are standardized, the estimated slope is the correlation $r_{1974,1975} = 0.875$ and the intercept is zero, by normalization.

$E(Y'|X') = \beta_0 + \beta_1 \cdot X' = \beta_1 \cdot X'$**. Hence** $E[\widehat{Y'|X'} = 3] = \hat{\beta}_1 \cdot 3 = 0.875 \cdot 3 = 2.625$**. For people whose 1974 income is three standard deviations above the mean, income is predicted to be around 2.6 standard deviations above the mean in 1975.**

The 95% prediction interval for $Y'|X' = 3$ is given by: $2.625 \pm 1.96\sqrt{0.485^2}$**, or, (1.674, 3.576).**

This result *is* intuitive: it illustrates the concept of regression toward the mean! The predicted (or fitted) standardized value of 1975 income is expected to be, on average, closer to its mean than the standardized value of 1974 income is to its mean.

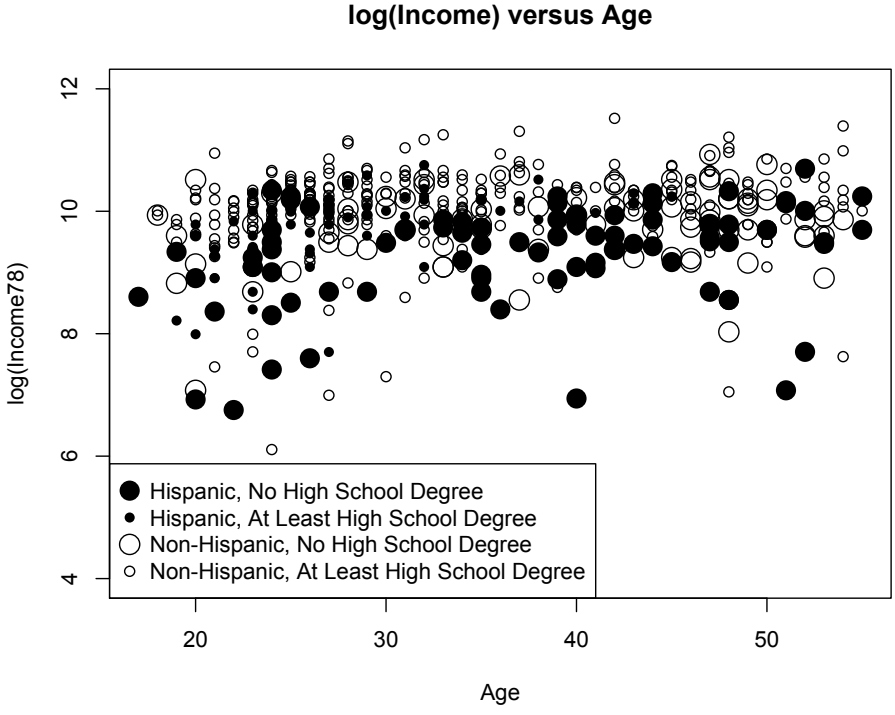
7. This problem again refers to the job training data set. Here, suppose that the goal is to predict 1978 income. [2 points total for this problem]

   (a) Look at the plot below and consider the following model for mean log 1978 income:

$$E(log(Income78)|NoDegree, Hispanic, Age)$$
$$= \beta_0 + \beta_1 Age + \beta_2 NoDegree + \beta_3 Hispanic$$
$$+\beta_4 Age * NoDegree + \beta_5 Age * Hispanic$$
$$+\beta_6 NoDegree * Hispanic + \beta_7 Age * NoDegree * Hispanic.$$

The least squares estimates and their standard errors are in the table below.

	$Estimate$	$StandardError$
β_0	9.68	0.133
β_1	0.012	0.004
β_2	-0.413	0.319
β_3	-1.19	0.394
β_4	0.002	0.008
β_5	0.032	0.013
β_6	0.328	0.560
β_7	-0.022	0.017



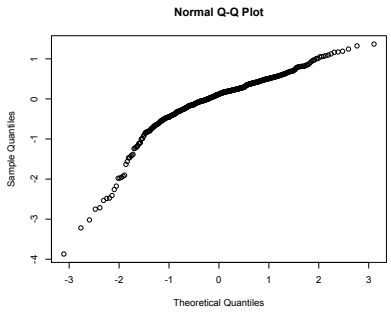
Write down the estimated regression equation for $E(log(Income78)|Age)$ for Hispanics who do have high school degrees. [4 points]

$$E(log(Income78)|Age, Hispanic = 1, NoDegree = 0) = 8.49 + 0.044 \times Age$$

(b) Conduct an extra sums of squares F-test comparing this model to the model with no three-way interaction. (You do not need to formally calculate the p-value; state the presumed conclusion based on the test statistic). Conceptually, how are the full and reduced models different? Relate your answer to what you see in the plot. [4 points]

The full and reduced models differ by only the three-way interaction. Conceptually, the three-way interaction allows a different slope for Age for all 4 combinations of NoDegree and Hispanic. Without the three-way interaction, we allow one slope for Age when NoDegree = Hispanic = 0, another when NoDegree = 1 and Hispanic = 0, another when NoDegree = 0 and Hispanic = 1 but the slope when NoDegree = Hispanic = 1 is the sum of all three. $H_0 : \beta_7 = 0$ **and** $H_A : \beta_7 \neq 0$**. We know that the F-statistic will be the square of the t-statistic for corresponding null hypothesis. F-stat =** $(\frac{-0.022}{0.017})^2 = 1.67$**. This is not large enough to reject** H_0**.**

(c) Below is a normal quantile-quantile plot for the residuals from the model originally specified. What are the implications of this plot for the interpretation of the coefficient estimates? Explain in one sentence. [2 points]



The residuals imply the error term is skewed. However, the coefficient estimates will still be unbiased.

(d) When the variable "NoDegree" is regressed on "Hispanic" and "Age," $R^2 = 0.19$. Concisely explain the implications of this fact for model fitting. Relate your answer to what you see in the plot. [5 points]

This implies that the standard error for the NoDegree coefficient will be slightly inflated due to its correlation with the Hispanic and Age variables. Looking at the plot, part of this correlation is because most Hispanics do not have a degree.

(e) Consider a simpler model for the mean:

$$E(log(Income78)|Hispanic, Age) = \alpha_0 + \alpha_1 Age + \alpha_2 Hispanic$$

Let X be a 528 x 3 matrix containing a column of ones, the vector "Age," and the vector "Hispanic." Let Y be a vector containing "log(Income78)."

You have calculated $(X'X)^{-1} = \begin{pmatrix} 0.024 & -0.001 & -0.003 \\ -0.001 & 0.000 & 0.000 \\ -0.003 & 0.000 & 0.009 \end{pmatrix}$ and $X'Y = \begin{pmatrix} 5220 \\ 178000 \\ 1390 \end{pmatrix}$.

You also know that $\hat{\sigma}^2 = 0.479$.

Write down the expression (in terms of $(X'X)^{-1}$, $X'Y$, $\hat{\sigma}^2$, and a vector containing possible values of the predictors) for an interval that would cover all of the predicted mean responses simultaneously with 95% probability. Write down the expressions for the upper and lower bounds of this interval for a 40-year-old non-Hispanic (you can skip the explicit calculations). Would the interval be wider or narrower for a 90-year-old Hispanic? Intuitively, why? [6 points]

Answer:

Let X_0 denote a possible value of the predictors. Then, an interval that would cover all of the predicted mean responses simultaneously with 95% probability would have the following form.

$$X_0(X'X)^{-1}X'Y \pm \sqrt{3 \times F_{3,n-3}(0.95)\hat{\sigma}^2 X_0(X'X)^{-1}X_0^T}$$

For a 40-year-old non-Hispanic, let $X_0 = (1, 40, 0)$. The interval will intuitively be wider for a 90-year-old Hispanic because the the vector $(1, 90, 1)$ is far from the mean of the predictor variables.