Hypothesis testing for OLS coefficients

Stats 203, Winter 2024

Lecture 7 [Last update: January 31, 2024]

1 Review and preview

- Last class, we introduced the Wald intervals for coefficients in the linear model: letting $\widehat{\mathrm{SE}}_j := \hat{\sigma}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{j+1,j+1}}$ denote our estimate of the standard error of $\hat{\beta}_j$,

$$C_j = \hat{\beta}_j \pm t_{n-p-1}^{(1-\alpha/2)}\widehat{\mathrm{SE}}_j.$$

  We will start off today's class by finishing up on this topic.

- For the bulk of today, we will talk about hypothesis testing, beginning with a general overview, and then moving on to the **T-test** for individual coefficients in the linear model.

2 A general overview of hypothesis testing

- The purpose of a hypothesis test is to assess the evidence for and against two different possible states of the world. These two states are typically called the **null** and **alternative** hypotheses, and the hypothesis test will either reject or fail to reject the null hypothesis in favor of the alternative. Unlike a confidence interval, then, the outcome of a hypothesis test is just a single bit of information. This can useful when we ultimately want to yes/no decision. For example, if we are the FDA, we need to decide whether a drug should be allowed into the market, and hypothesis tests can be (and are) used to help make that determination.

- In any hypothesis testing problem, there are three central statistical objects: the test-statistic, the p-value, and the test itself. Suppose our data $x_1, \dots, x_n$ are realized values of random variables $X_1, \dots, X_n \sim p(\cdot;\theta)$, where $p(\cdot;\theta)$ is a distribution determined by an unknown parameter θ. We wish to test whether

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

  A **test statistic** $T(x_1, \dots, x_n)$ is any statistic whose CDF F_T is known under the null hypothesis. We will suppose that larger values of T indicate stronger evidence against the null.

- A p-value is a certain measure of the evidence against the null. By convention, smaller p-values indicate more evidence against the null. Then a natural **p-value** is

$$P(x_1, \dots, x_n) = 1 - F_T(T(x_1, \dots, x_n)).$$

  The **hypothesis test** then rejects the null hypothesis H_0 if $P(x_1, \dots, x_n)$ is smaller than some number $\alpha \in (0, 1)$. Otherwise it fails to reject the null.

- As with estimates, standard errors, and confidence intervals, test statistics, p-values, and hypothesis tests are functions that can either be applied to data or random variables. Suppose the null hypothesis is correct and $X_1, \dots, X_n \sim p(;\theta_0)$, and for simplicity assume T is a continuous random variable. Then, using an argument that should be very familiar from Homework 1:

$$\begin{aligned}
\mathbb{P}(P(X_1,\dots,X_n) \leq \alpha) &= \mathbb{P}(1 - F_T(T(x_1,\dots,x_n)) \leq \alpha) \\
&= \mathbb{P}(F_T(T(x_1,\dots,x_n)) \geq 1-\alpha) \\
&= 1 - \mathbb{P}(F_T(T(x_1,\dots,x_n)) \leq 1-\alpha) \\
&= 1 - \mathbb{P}(T(x_1,\dots,x_n)) \leq F_T^{-1}(1-\alpha)) \\
&= 1 - F_T(F_T^{-1}(1-\alpha)) \\
&= 1 - (1-\alpha) = \alpha.
\end{aligned}$$

This holds true for any $\alpha \in [0,1]$, and thus $P(X_1,\dots,X_n) \sim \mathrm{Unif}(0,1)$ *if the null hypothesis is correct.*

- The implication for our hypothesis test "reject the null hypothesis if $P \leq \alpha$" is that if the null hypothesis is correct, we will falsely reject the null at most $100\alpha\%$ of the time. As with confidence intervals, $\alpha \in (0,1)$ is a user-determined meta-parameter called the **level** of the test. Typical choices are $\alpha = .1, .05, .01, .001, .0001$.

3 Hypothesis testing for OLS coefficients

- We return to the specific case of the T-test for coefficients in the linear model. We will assume that the linear model with Normal errors is correct. The hypotheses we are interested in are

$$H_{0,j} : \beta_j = 0, \quad \text{versus} \quad H_{1,j} : \beta_j \neq 0. \tag{1}$$

The special significance of $H_0 : \beta_j = 0$ is that it means the jth predictor does not add anything to the linear model once all the other variables are included. For this reason, tests of (1) are sometimes called **added variable** tests.

- How do we test (1)? As with estimators and confidence intervals, there are different test statistics for (1). A common one that you might be familiar with from other contexts is the likelihood ratio statistic. Instead we will use a simpler test statistic: take the estimate $\hat{\beta}_j$, subtract off by its expectation under the null, and normalize by our estimate for its standard error:

$$T_j := \frac{\hat{\beta}_j - 0}{\widehat{\mathrm{SE}}_j}. \tag{2}$$

The statistic is called a **T-statistic** or **Wald statistic**.

- Either large values of T_j or $-T_j$ indicate strong evidence against $H_{0,j}$ so we would like to reject the null if $|T_j|$ is large. To calculate a p-value, we need to know the distribution of $|T_j|$ if the null is correct. We know from the Lecture 6 notes that

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{\mathrm{SE}}_j} \sim t_{n-p-1}.$$

But if $H_{0,j}$ is correct then $\beta_j = 0$, and therefore $T_j \sim t_{n-p-1}$ under the null. So, letting $F_{|t_d|}(\cdot)$ be the CDF of the absolute value of a T-distribution with d degrees of freedom,[1] we will use

$$P_j := 1 - F_{|t_{n-p-1}|}(|T_j|)$$

as our p-value for the null hypothesis $H_{0,j}$. The T**-test** or **Wald test** rejects the null hypothesis if $P_j \leq \alpha$. Using the symmetry of the T-distribution we can show that $F_{|t_d|}^{-1}(1-\alpha) = t_{n-p-1}^{(1-\alpha/2)}$, where the latter is the $1-\alpha/2$th quantile of the t distribution. Therefore, we reject the null hypothesis if

$$P_j = 1 - F_{|t_{n-p-1}|}(|T_j|) \leq \alpha \iff |T_j| \geq F_{|t_d|}^{-1}(1-\alpha) = t_{n-p-1}^{(1-\alpha/2)}.$$

[1] A function that is truly irritating to write down

- Everything so far has assumed the null hypothesis is $H_0 : \beta_j = 0$. Sometimes – though less commonly – we want to test $H_0 : \beta_j = b_0$ for some other number $b_0 \in \mathbb{R}$. To adapt the T-test, just use the statistic $T_j = \frac{\hat{\beta}_j - b_0}{\widehat{\mathrm{SE}}_j}$ and reject the null if $|T_j| \geq t_{n-p-1}^{(1-\alpha/2)}$.
- **Duality of hypothesis tests and confidence intervals**. There is an extremely deep relationship between confidence intervals and hypothesis tests. Suppose a given number $b_0 \in C_j$. Then we know that $\hat{\beta}_j$ is within $t_{n-p-1}^{(1-\alpha/2)} \times \widehat{\mathrm{SE}}_j$ of b_0. But this means

  $$|T_j| = \Big|\frac{\hat{\beta}_j - b_0}{\widehat{\mathrm{SE}}_j}\Big| \leq t_{n-p-1}^{(1-\alpha/2)},$$

  and so the T-test would fail to reject the null hypothesis that $\beta_j = b_0$. On the other hand if $b_0 \notin C_j$ equivalent reasoning would imply that $|T_j| > t_{n-p-1}^{(1-\alpha/2)}$, and so the T-test would reject the null hypothesis that $\beta_j = b_0$.

  Repeating this for all $b_0 \in \mathbb{R}$, we see that the Wald interval is exactly

  $$C_j = \{b_0 \in \mathbb{R} : |T_j| \leq t_{n-p-1}^{(1-\alpha/2)}\}.$$

  In words, the Wald interval C_j consists of all the numbers b_0 for which the Wald test fails to reject the null hypothesis that $\beta_j = b_0$. Likewise, the Wald test for $H_{0,j} : \beta_j = 0$ rejects the null if and only if $0 \notin C_j$. This relationship is called the *duality of confidence intervals and hypothesis tests*, and it holds in much greater generality than the Wald tests and Wald intervals.

4 Some of the issues with hypothesis testing

There are many ways to misuse or outright abuse hypothesis testing. Some are specific to linear modeling, some are more widely applicable. An incomplete list follows. The first few have to do with issues of interpretation, the last few with incorrect protocols.

4.1 Issues with interpretation

What does rejecting the null really mean? The null hypothesis in (1) is that all of the assumptions of the linear model with Normal errors are correct, but that β_j just happens to be zero. The Wald test may reject the null hypothesis for three reasons: (1) the null is in fact true – all the assumptions of are correct and $\beta_j = 0$ – but the test happens to make a mistake; (2) all the assumptions are correct but $\beta_j \neq 0$; or (3) one or more of the assumptions of the linear model are wrong. Thus, hypothesis testing never tells us for sure that the null is wrong, nor does it offer any evidence that the model is or is not correct.

Statistical significance is not practical significance. Suppose β_j is in fact some arbitrarily small but non-zero number, say $\beta_j = 10^{-12}$, and that the predictors X_i each come from a distribution with covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$. With a sufficiently large sample, that is as $n \to \infty$, the following will happen: (1) $\hat{\beta}_j \to \beta_j$, (2)

$$\mathrm{SE}[\hat{\beta}_j|\mathbf{X}] \to \frac{\sigma^2}{n}[\mathbf{\Sigma}^{-1}]_{j,j} \to 0,$$

(3) the estimated standard error $\widehat{\mathrm{SE}}_j \to \mathrm{SE}[\hat{\beta}_j|\mathbf{X}] \to 0$, (4) the T-statistic

$$T_j = \frac{\hat{\beta}_j - 0}{\widehat{\mathrm{SE}}_j} \to \frac{\beta_j - 0}{\mathrm{SE}[\hat{\beta}_j|\mathbf{X}]} \to \infty,$$

and (5) the p-value $P_j \to 0$. So with a sufficiently large sample size, we will be guaranteed to see an astronomical level of evidence against the null! But in most cases you can imagine that the difference between $\beta_j = 10^{-12}$ and $\beta_j = 0$ is practically irrelevant. (Depending, of course, on the units of the variables in question.) For this reason, it is always best practice to report confidence intervals, to get a sense of the magnitude of possible effect size.

This is not merely a hypothetical exercise. The statistician Andrew Gelman says that in every applied problem he has ever worked on the null hypothesis has been false. So then why bother running a hypothesis test?

The meaning of the null hypothesis depends on the predictors. We know from Lecture 4 that the meaning of β_j depends on the predictors in the model: loosely speaking, it captures the expected difference in Y given a difference in X_j, *holding all other predictors fixed*. So, the hypothesis that $\beta_j = 0$ also depends on the predictors.

Here is an (admittedly absurd) hypothetical illustrating the potential confusion this can cause. Suppose in a given problem you are highly invested in failing to reject the null hypothesis: e.g. you are a research scientist for Marlboro and the null hypothesis is that smoking cigarettes and lung cancer are unrelated. You collect data by asking a random sample of the population about their smoking habits. Importantly, you make sure to ask each participant how many cigarettes they smoke per day – call this X_{i1} – and how many cigarettes they smoke per week – call this X_{i2}. You include both variables in the model, along with other relevant predictors such as age, weight, other health conditions. Now, the hypothesis $H_{0,1} : \beta_1 = 0$ means something like:

> There is no expected difference in probability of lung cancer among individuals of the same age, weight, etc. who report smoking a different number of cigarettes daily but the same number of cigarettes per week.

It's not clear what this hypothesis actually means – perhaps it's measuring something about the correlation between lung cancer and math ability? – but it clearly has little to do with any intrinsic link between smoking or cancer. And if you test this hypothesis on data, you are very unlikely to reject the null (whatever it means) due to the high correlation between the two predictors.

4.2 Issues with application

There are also many ways to invalidate the long-run frequency guarantees of a hypothesis test, *even if all of the assumptions of the linear model are correct.* Here are some ways to make sure you will end up with lots of significant looking effects, regardless of whether or not the null is ever false.

Sample until rejection. Suppose you start with n observations, and find a p-value for an interesting β_j that is close to, but not quite significant. Very frustrating. But no fear: just sample another n observations, and recompute the p-value on the $2n$ observations. Keep going until the null is rejected.

This sounds ridiculous, and it is, but it really does happen in practice. An infamous example is a 2010 study (Carney et al., 2010) on "power posing"[2] which concluded that it led to "elevations in testosterone, decreases in cortisol, and increased feelings of power and tolerance for risk." This was interpreted as saying that you should power pose before big presentations and meetings to increase your confidence. The conclusion was (understandably) deemed so fascinating that it led to TED talks, Oprah appearances, book deals, etc. Unfortunately, many follow-up studies have failed to replicate the findings. One of the authors, who (to her credit) eventually publicly announced that the "power posing" effect was not real, described the process of data collection as being essentially "sample until rejection."

Test, baby, test! Suppose you start with p predictors, and thus, p hypotheses. Compute one p-value per predictor. Report all of the resulting p-values but heavily emphasize the ones that are most significant.

The issue with this, of course, is that even if all p of the null hypotheses are correct, we would expect to reject αp of them due to pure chance, and so we need to correct for the effect of running **multiple tests**. A

[2] that is, "standing in a posture that they mentally associate with being powerful." https://en.wikipedia.org/wiki/Power_posing

simple correction, called the **Bonferroni correction**, is to reject each null hypothesis only if $P_j \leq \alpha/p$. The Bonferroni correction controls the familywise error rate, or the probability of making any false rejection:

$$\mathbb{P}\Big(P_j \geq \frac{\alpha}{p} \;\text{ for all }\; j = 1, \ldots, p \;\text{ such that } \beta_j = 0\Big) \geq 1 - \alpha.$$

The issues with multiple testing, and the need for correction, have only recently become widely understood in domain sciences. For example, I went to a talk last year where the speaker mentioned that as of the year 2010, 0 empirical studies published in the top 5 economic journals had explicitly corrected for multiple testing.

Stick it in the file drawer. Maybe the last strategy wasn't good enough for you: you're still failing to reject too many nulls. So modify the strategy. Retain all of the variables with significant p-values, and stick the rest in a file drawer. Refit the model with only the significant variables. Now you have a much simpler model, with (likely) most of the variables in the simpler model appearing significant.

The above strategy is known as **p-hacking** or **data snooping**. Again, it may seem ridiculous, but it happens all the time in practical work. Even well-intentioned protocols can implicitly result in p-hacking. Suppose a journal has a policy of only accepting articles for which the reported effect is significant at .05-level. Then, even if every article submitted to the journal has computed completely legitimate p-values, among published articles there is no longer any type I error guarantee. This is called publication bias. The point was made forcefully by the (provocatively titled) article "Why Most Published Research Findings are False"(Ioannidis, 2005).

The above strategies are, of course, even more powerful(ly abusive) when used in combination.

References

Dana R Carney, Amy JC Cuddy, and Andy J Yap. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science*, 21(10):1363–1368, 2010.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Data analysis example: fuel consumption data

From Weisberg, *Applied Linear Regression*, “The goal of this example is to understand how fuel consumption varies over the 50 United States and the District of Columbia, and, in particular, to understand the effect on fuel consumption of state gasoline tax. Table 1.2 describes the variables to be used in this example; the data are given in the file `fuel2001.txt`. The data were collected by the US Federal Highway Administration.”

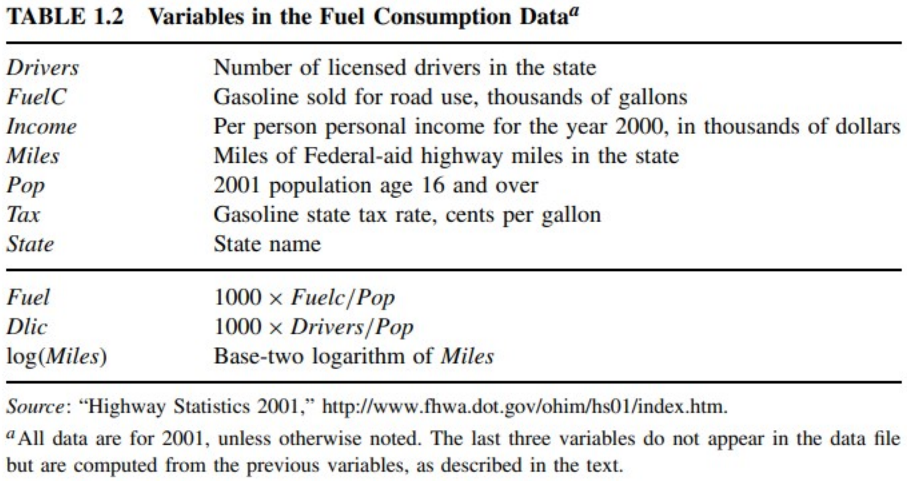


TABLE 1.2 Variables in the Fuel Consumption Data[a]

Drivers	Number of licensed drivers in the state
FuelC	Gasoline sold for road use, thousands of gallons
Income	Per person personal income for the year 2000, in thousands of dollars
Miles	Miles of Federal-aid highway miles in the state
Pop	2001 population age 16 and over
Tax	Gasoline state tax rate, cents per gallon
State	State name
Fuel	1000 × *Fuelc*/*Pop*
Dlic	1000 × *Drivers*/*Pop*
log(*Miles*)	Base-two logarithm of *Miles*

Source: “Highway Statistics 2001,” http://www.fhwa.dot.gov/ohim/hs01/index.htm.

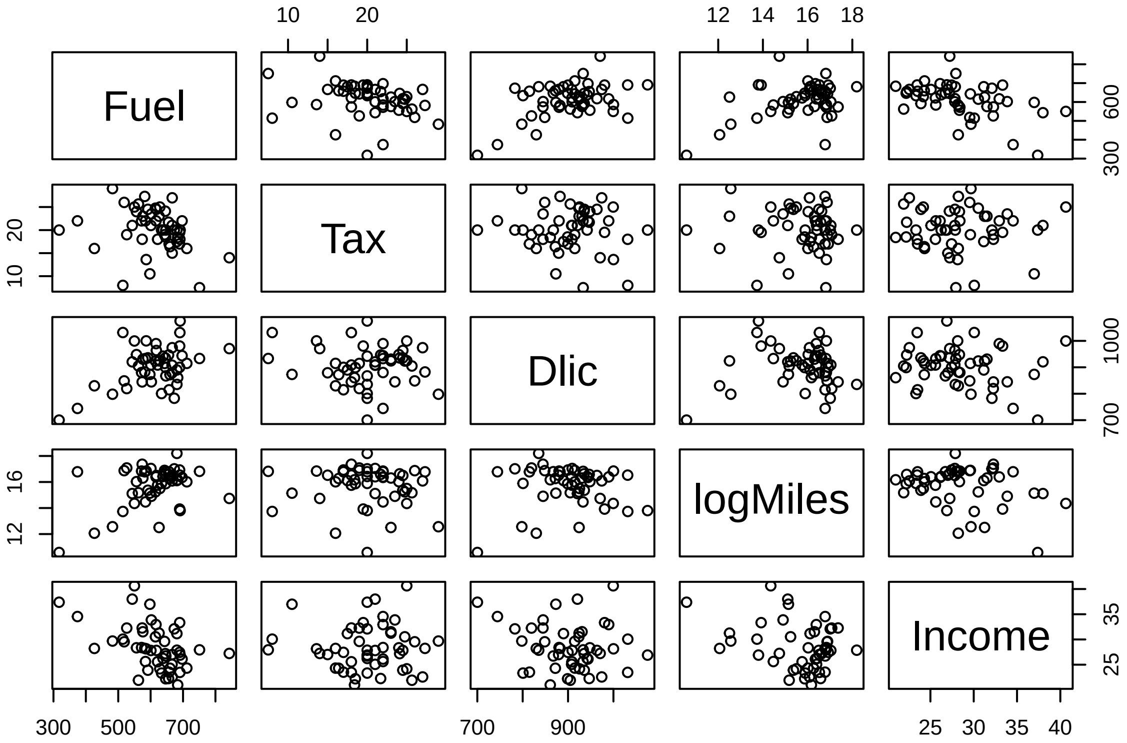
[a] All data are for 2001, unless otherwise noted. The last three variables do not appear in the data file but are computed from the previous variables, as described in the text.

As usual, we begin by loading and plotting the data.

```
# read in the data
fuel2001 = read.table("fuel2001.txt")

# transformations as prescribed by Weisberg.
fuel = data.frame(
          Fuel=1000 * fuel2001$FuelC/fuel2001$Pop,
          Tax = fuel2001$Tax,
          Dlic=1000 * fuel2001$Drivers/fuel2001$Pop,
          logMiles= log2(fuel2001$Miles),
          Income= fuel2001$Income/1000)

# scatterplots
plot(fuel)
```



As the excerpt suggests, we would like to understand the relationship between gasoline tax and the level of fuel consumed. We will assume the linear model

$$\text{Fuel}_i = \beta_0 + \beta_1 \text{Tax}_i + \beta_2 \text{Dlic}_i + \beta_3 \text{Income}_i + \beta_4 \log(\text{Miles}_i) + \epsilon_i,$$

with the usual assumptions.

We start by computing the OLS estimates.

```r
# linear regression
fuel.lm = lm(Fuel ~ Tax + Dlic + Income + logMiles, data = fuel)
summary(fuel.lm)
```

```
##
## Call:
## lm(formula = Fuel ~ Tax + Dlic + Income + logMiles, data = fuel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.145  -33.039    5.895   31.989  183.499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.1928   194.9062   0.791 0.432938
## Tax          -4.2280     2.0301  -2.083 0.042873 *
## Dlic          0.4719     0.1285   3.672 0.000626 ***
## Income       -6.1353     2.1936  -2.797 0.007508 **
## logMiles     18.5453     6.4722   2.865 0.006259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.89 on 46 degrees of freedom
## Multiple R-squared:  0.5105, Adjusted R-squared:  0.4679
## F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

Each of the estimated coefficients comes with an accompanying t-statistic and (two-sided) p-value. Let's verify that these are calculated exactly as we've worked out in class.

```
# T-statistic by hand (kind of)
betahat = fuel.lm$coefficients
sehat = summary(fuel.lm)$coef[,2]
tstat = betahat/sehat
tstat
```

```
## (Intercept)         Tax        Dlic      Income    logMiles
##   0.7911132  -2.0826261   3.6717660  -2.7968805   2.8653854
```

```
# P-value by hand
n = nrow(fuel)
p = 5 # Intercept, Tax, Dlic, Income, logMiles
d = n - p
2*(1 - pt(abs(tstat),df = d)) # 2 * Pr of upper tail, for two-sided test.
```

```
##  (Intercept)          Tax         Dlic       Income     logMiles
## 0.4329381433 0.0428733310 0.0006255639 0.0075077902 0.0062591801
```

Let's focus in on P_3, the p-value for the hypothesis that $\beta_3 = 0$. The p-value is small(er than .05) – we can interpret this as saying that there is statistically significant evidence that the predictor Income adds to a model that already include $\{\text{Tax}, \text{Dlic}, \log(\text{Miles})\}$. This does not imply that there is statistically significant evidence that Income adds to the model once other variables are included, as we now see.

Effect of adding correlated predictors

The scatterplots showed that the pairwise correlations between predictors were fairly small. Now let's see what happens when we include a predictor (GradFreq, which measures the percentage of people over the age of 25 with at least a bachelor's degree) that is very correlated with one of the predictors that is already included (Income).

```
statedata = read.csv("statedata.csv")
fuel = cbind(fuel,statedata) # both sorted alphabetically so rows match
```

```
fuelwgrad.lm = update(fuel.lm, . ~ . + GradFreq)
summary(fuelwgrad.lm)
```

```
##
## Call:
## lm(formula = Fuel ~ Tax + Dlic + Income + logMiles + GradFreq,
##     data = fuel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -165.624  -31.086    4.533   28.645  180.372
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 170.4377   203.0401   0.839 0.405666
## Tax          -4.1865     2.0541  -2.038 0.047438 *
## Dlic          0.4689     0.1301   3.604 0.000781 ***
## Income       -5.1816     3.6702  -1.412 0.164888
## logMiles     17.6869     7.0466   2.510 0.015738 *
## GradFreq     -1.0214     3.1338  -0.326 0.745986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.53 on 45 degrees of freedom
## Multiple R-squared:  0.5116, Adjusted R-squared:  0.4574
## F-statistic: 9.429 on 5 and 45 DF,  p-value: 3.359e-06
```

Income was previously significantly associated with Fuel but is no longer. Why? Look at the correlation matrix of the predictors.

```r
X = cbind(fuel$Tax, fuel$Income,fuel$Dlic,fuel$logMiles,fuel$GradFreq)
colnames(X) = c("Tax","Income","Dlic","logMiles","GradFreq")
round(cor(X),2)
```

```
##            Tax Income  Dlic logMiles GradFreq
## Tax       1.00  -0.01 -0.09    -0.04     0.04
## Income   -0.01   1.00 -0.18    -0.30     0.82
## Dlic     -0.09  -0.18  1.00     0.03    -0.18
## logMiles -0.04  -0.30  0.03     1.00    -0.45
## GradFreq  0.04   0.82 -0.18    -0.45     1.00
```

`Income` and `GradFreq` have an 80% sample correlation. Intuitively, once we include `GradFreq` as a predictor, adding `Income` can only improve the fit a little bit more. The lesson: when adding a variable to the model which is correlated with predictors that are already included, it is “hard’ ’ for the new variable to appear statistically significant. This is one of the reasons people typically discourage running linear regression with highly correlated predictors. (Although to be clear, there is nothing invalid about running T-tests when the predictors are correlated.)

You may have noticed another disturbing fact: the p-value for the variable `GradFreq` is also very insignificant. Of course this makes perfect sense, as even when we leave `GradFreq` out the highly correlated predictor `Income` is still included in the model. We are seeing an example of a general phenomenon: it is possible for a set of predictors to be highly associated with a response, and yet for none of them to have a significant T-statistic. This illustrates some of the subtleties involved in added-variable testing.