Homework 1, Econometrics 1

Golkar Arno

November 7, 2025

Question 1

In a first step, familiarize yourself with the authors' research interest and the economic model presented on pages 2616-2619. Briefly summarize the policy implications of the "fundamental law of road congestion." Use the elasticity concept in your explanation.

Duranton and Turner (2011) show that vehicle-kilometers traveled (VKT) increase approximately one-for-one with lane-kilometers of road. In other words, the elasticity of VKT with respect to lane-kilometers is close to one, meaning that a 10 percent increase in road capacity leads to roughly a 10 percent increase in total driving.

From a policy perspective, expanding road infrastructure does not relieve congestion in the long run. The added capacity simply induces more traffic, as individuals adjust their behavior by increasing the number and length of trips, as well as migration and commercial traffic. The authors find no significant evidence that public transport supply affects total VKT, nor that new lanes draw traffic away from other roads. Therefore, expanding the road network alone is not an effective policy to reduce congestion. More comprehensive approaches, including pricing mechanisms or land-use regulation, are required.

The elasticity can be written as:

$$\rho_R = \frac{\partial \ln Q}{\partial \ln R} \approx 1$$

which indicates a proportional response of traffic to road capacity.

Question 2

At the end of page 2619, the authors emphasize the importance of the assumption $\mathrm{Cov}(R, \varepsilon \mid X) = 0$. *Explain how this relates to* $\mathcal{A}_3^{OLS}$ *in the lecture. Why is this assumption important for the interpretation of the OLS estimator?*

The estimated model is:

$$\ln(Q_{it}) = A_0 + \rho_R^Q \ln(R_{it}) + A_1 X_{it} + \varepsilon_{it}$$

where Q_{it} denotes vehicle-kilometers traveled, R_{it} represents lane-kilometers, and X_{it} is a vector of control variables. The coefficient ρ_R^Q measures the elasticity of VKT with respect to roads.

For OLS to yield a consistent estimate of this elasticity, the assumption

$$\mathrm{Cov}(R_{it}, \varepsilon_{it} \mid X_{it}) = 0$$

must hold. This corresponds to the assumption $\mathcal{A}_3^{OLS}$ from the lecture, which requires that the regressors be uncorrelated with the unobserved factors contained in the error term.

If this assumption is violated, the estimated coefficient captures both the causal effect of roads on traffic and the spurious correlation between R_{it} and unobserved determinants of Q_{it}. For instance, roads may be built in cities where travel demand is already rising, producing an upward bias in the OLS estimate of ρ_R^Q. In that case, OLS does not recover the true causal effect. To address this problem, Duranton and Turner use instrumental variables that generate exogenous variation in road provision, such as historical exploration routes, early railroads, and proposed interstate plans.

Question 3

Create a table of descriptive statistics for all continuous variables in the dataset for the year 2003, including measures that go beyond those reported in Table 1 of the paper. Briefly comment on the sample.

variable	n	mean	sd	min	p1	p25	median	p75	p99	max	IQR	CV	skewness	kurtosis
VKT_IH	228	15960581.063	31579287.019	668828.234	712131.226	2406336.975	4938078.087	15588733.875	163866921.59	271078107	13182396.9	1.979	4.448	27.81
VKT_IHU	228	11574332.87	27037773.787	0	0	723484.422	1879248.828	9395620.938	139767016.242	221220164	8672136.516	2.336	4.389	26.274
VKT_IHNU	228	4386248.193	5277126.003	0	155013.757	1270022.25	2636601.274	5751609.582	24249899.716	49857943	4481587.332	1.203	3.964	28.484
LANE_IH	228	1279.748	1857.578	118.195	125.027	389.053	639.945	1285.02	9235.326	14582.285	895.967	1.452	3.818	21.132
LANE_IHU	228	719.508	1410.776	0	0	97.897	186.433	672.856	6809.709	10048.433	574.958	1.961	3.762	19.554
LANE_IHNU	228	560.239	558.684	0	18.363	234.221	389.297	708.949	2278.267	5494.711	474.728	0.997	3.888	29.58
POP	228	950054.311	2071787.42	66533	80694.5	163055.25	336270.5	813362.75	8739183.1	19397717	650307.5	2.181	5.754	43.955
ELEV	228	0.661	0.907	0.004	0.011	0.122	0.233	0.71	3.596	4.367	0.588	1.373	2.045	6.61
RUG	228	0.009	0.011	0	0	0.003	0.005	0.011	0.047	0.078	0.009	1.17	2.401	10.655
HEAT	228	47.47	22.581	2.426	4.038	26.988	51.366	65.354	92.475	98.924	38.366	0.476	-0.133	2.029
COOL	228	13.092	9.113	1.083	1.784	5.683	10.074	18.934	38.062	39.725	13.251	0.696	0.948	3.086
SPW	228	45.457	10.521	20.73	24.921	37.294	44.622	53.009	70.68	75.098	15.715	0.231	0.263	2.582

Figure 1: Summary statistics for all continuous variables in the 2003 sample of U.S. metropolitan areas.

Sample Size and Missing Data: The number of observations (n) is 228 for nearly all variables, indicating no significant issues with missing data for this year.

Traffic Volume (VKT_IH, VKT_IHU, VKT_IHNU):

- High Averages and Dispersion: The mean daily VKT for all interstates (`VKT_IH`) is approximately 16 million km, but with a very large standard deviation (sd) of about 31.6 million km. The Coefficient of Variation (CV = sd/mean) is high (1.98), indicating substantial relative dispersion across MSAs. Similar patterns hold for urban (`VKT_IHU`) and non-urban (`VKT_IHNU`) VKT.
- Skewness: The means are considerably larger than the medians (e.g., 16.0M vs. 4.9M for `VKT_IH`). This is confirmed by the high positive skewness values (around 4.4 for `VKT_IH` and `VKT_IHU`, 4.0 for `VKT_IHNU`). This indicates that the distribution is heavily right-skewed, meaning a few MSAs have extremely high traffic volumes compared to the typical MSA.
- Kurtosis: The kurtosis values are very high (above 20 for all VKT variables), suggesting the distributions have much heavier tails and are more peaked than a normal distribution ("leptokurtic"). This reinforces the idea that extreme values (outliers) significantly influence the mean.

The pronounced skewness and high kurtosis strongly suggest that a log transformation of the VKT variables is appropriate for regression analysis to stabilize variance and approximate normality.

Road Infrastructure (LANE_IH, LANE_IHU, LANE_IHNU):

- Similar Distributional Properties: The lane kilometer variables exhibit distributional characteristics similar to the VKT variables. Means are much larger than medians (e.g., 1280 vs. 640 for `LANE_IH`), standard deviations and CVs are relatively high, skewness is strongly positive (around 3.8), and kurtosis is high (around 20-30).

Population (POP):

- Extreme Skewness: Population is also highly right-skewed, even more so than VKT or LANE. The mean (950k) is almost triple the median (336k). The skewness (5.75) and kurtosis (43.96) are very high, indicating the presence of a few very large metropolitan areas greatly influencing the average.

Geographic and Climate Variables (ELEV, RUG, HEAT, COOL, SPW):

- Less Skewed Distributions: Compared to VKT, LANE, and POP, these variables generally exhibit much less skewness and kurtosis.
- Elevation (ELEV) and Ruggedness (RUG): These show moderate positive skewness (around 2.0-2.4) but are less extreme than the main variables. `RUG` has a median very close to zero (0.005), indicating most MSAs are not very rugged.
- Heating/Cooling Days (HEAT, COOL): `HEAT` is slightly left-skewed (-0.13), while `COOL` is moderately right-skewed (0.95). Their kurtosis values are close to that of a normal distribution (around 2-3). These variables appear much more symmetrically distributed.
- Urban Sprawl (SPW): This variable is the most symmetric and normally distributed of the set, with low skewness (0.26) and kurtosis (2.58) close to zero (excess kurtosis). The mean (45.5) and median (44.6) are very close.

Overall Implication: The descriptive statistics highlight significant heterogeneity across US MSAs in terms of size (POP), infrastructure (LANE), and traffic volume (VKT). The extreme right-skewness of these key variables strongly motivates the log-log specification of Model M_1. The control variables related to geography and climate have more regular distributions.

Question 4

We begin by exploring the data for some of the relationships we are interested in.

Question A

Is daily VKT on all MSA interstates correlated with the corresponding lane kilometers? Create a scatter plot that provides an answer. In the plot, use different colors to distinguish the three years under study. Briefly comment on the relationship, using the empirical correlation coefficients to support your discussion.

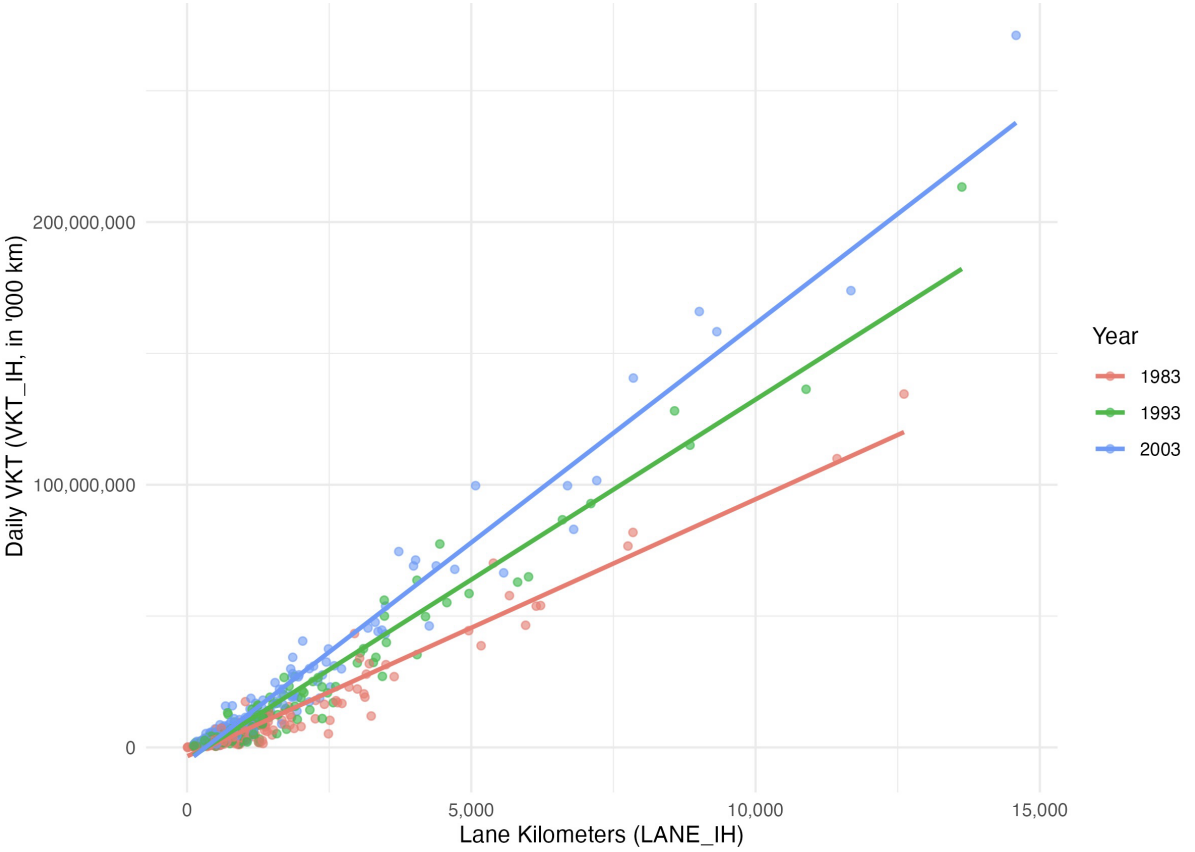

Figure 2: Relationship between VKT and lane kilometers by Year

Yes, daily Vehicle-Kilometers Traveled (VKT) on all MSA interstates is strongly and positively correlated with the corresponding lane kilometers.The scatter plot (Figure 2) clearly illustrates this relationship. There is a distinct positive trend: as Lane Kilometers increase on the x-axis, Daily VKT increases on the y-axis. This positive association is consistent across all three years under study, as shown by the separate colored lines (1983, 1993, and 2003) all sloping upwards.This strong visual relationship is confirmed by the empirical correlation coefficients. The overall correlation for all years combined is 0.9537, which indicates an extremely strong positive linear association. Furthermore, this strong correlation is very stable over time, as seen in the year-by-year coefficients:

- 1983: 0.971
- 1993: 0.978
- 2003: 0.981

- **Overall:** 0.9537

All these values are very close to +1, confirming that as road capacity (lane kilometers) increases, the total volume of traffic (VKT) increases with it in a highly predictable, linear pattern.

Question B

Examine the correlation between lane kilometers on all MSA interstates and population in a similar manner. What implications might the result have, given that both variables will later be used as regressors in the regression analysis?

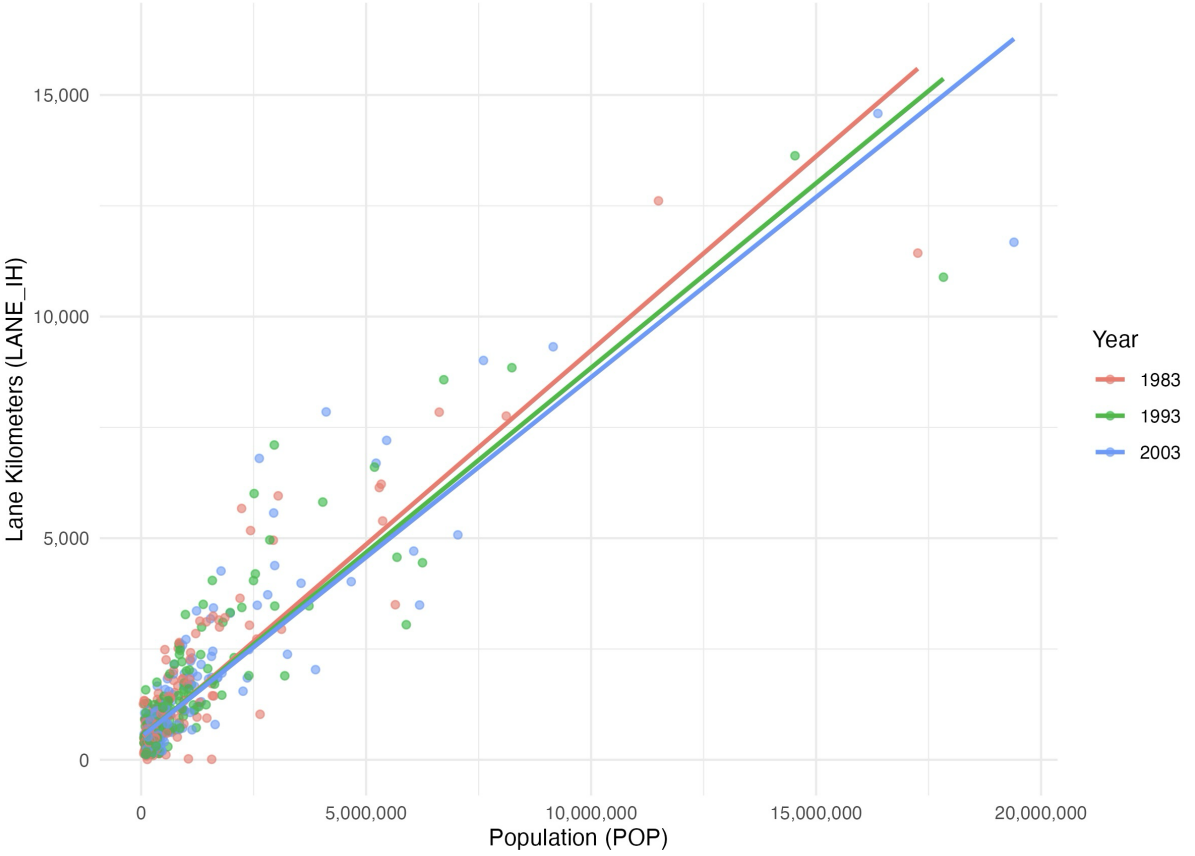

Figure 3: Relationship between population and lane kilometers by Year

The scatter plot (see Figure 3) and the calculated correlation coefficients both show that there is a very strong positive correlation between lane kilometers (`LANE_IH`) and population (`POP`).

The plot clearly shows a tight, positive linear trend: MSAs with larger populations also have significantly more lane kilometers. This visual evidence is confirmed by the numerical correlation coefficients, which are high and stable across all three periods:

- 1983: 0.902
- 1993: 0.896
- 2003: 0.906
- Overall: 0.901

This strong correlation between two independent variables (LANE and POP) that will be used on the right-hand side of our regression equation implies a significant problem of multicollinearity.

Multicollinearity does not bias our OLS estimates (they remain unbiased and consistent). However, it has serious consequences for our analysis:

1. Inflated Standard Errors: The variances and standard errors for the coefficients $\hat{\beta}_1$ (on $\ln(LANE)$) and $\hat{\beta}_2$ (on $\ln(POP)$) will be high.
2. Imprecise Estimates: Because the standard errors are large, our confidence intervals for these coefficients will be very wide, making our estimates imprecise.
3. Low t-statistics: The t-statistics for $\hat{\beta}_1$ and $\hat{\beta}_2$ will be smaller (since $t = \hat{\beta}/se(\hat{\beta})$). This might lead us to incorrectly conclude that one or both of these variables are not statistically significant, even if they have a true causal effect on VKT.

Because LANE and POP move together so closely, the OLS model will have difficulty separating their individual impacts on VKT. It becomes hard to tell if VKT is high because of POP or because of the LANEs that were built for that population.

In the following, we restrict the sample to the year 2003. Moreover, for the next exercises, we focus solely on VKT_IH and LANE_IH, leaving aside other MSA interstate categories. We aim to estimate the following regression, where i denotes an MSA:

$$\ln(VKT_i) = \beta_0 + \beta_1 \ln(LANE_i) + \beta_2 \ln(POP_i) + \gamma X_i + \sum_{j=1}^{9} \delta_j \mathbf{1}\{DIV_i = j\} + \nu_i, \quad (M_1)$$

where X_i includes all control variables listed above. Throughout the exercise, we assume that $\mathbb{E}[\nu_i \mid \mathbf{Z}] = 0$, $\forall i$, where $\mathbf{Z}$ includes all independent variables.

Question 5

When estimating model M_1, the software returns the message `"1 not defined because of singularities."` Why does this happen? Rewrite the regression equation.

Why this happens

The error "1 not defined because of singularities" indicates a problem of perfect multicollinearity.

This happens because the specified model M_1 suffers from the "Dummy Variable Trap". The model includes both an intercept (a constant term, β_0) and a complete set of 9 dummy variables for the 9 census divisions ($\sum_{j=1}^{9} \delta_j \mathbf{1}\{DIV_i = j\}$).

A full set of mutually exclusive and exhaustive dummy variables perfectly predicts the intercept. The sum of all 9 dummy variables for any observation i is exactly 1:

$$\sum_{j=1}^{9} \mathbf{1}\{DIV_i = j\} = 1$$

This means the column vector for the intercept is a perfect linear combination of the 9 column vectors for the division dummies. The $X'X$ matrix becomes singular and cannot be inverted to compute the OLS estimates.

The software (R) automatically detects this perfect collinearity and resolves it by dropping one of the collinear variables to make the estimation possible. In this case, it drops one of the dummy variables (the "1" in the message refers to the first level of the factor variable 'DIV').

Rewritten regression equation

To correct this, we must manually remove one of the collinear variables. The standard approach is to omit one of the dummy variables, which then becomes the reference (or base) category.

If we choose the first division ($j = 1$) as the reference category, the rewritten (and estimable) equation is:

$$\ln(VKT_i) = \beta_0 + \beta_1 \ln(LANE_i) + \beta_2 \ln(POP_i) + \gamma X_i + \sum_{j=2}^{9} \delta_j \mathbf{1}\{DIV_i = j\} + \nu_i$$

In this corrected model, the intercept β_0 represents the expected log(VKT) for an MSA in the reference division (division 1), ceteris paribus. Each coefficient δ_j (for $j = 2, ..., 9$) represents the difference in the intercept compared to the reference division.

Question 6

What signs would you expect for the coefficients β_1 and β_2?

The model is:

$$\ln(VKT_i) = \beta_0 + \beta_1 \ln(LANE_i) + \beta_2 \ln(POP_i) + \gamma X_i + \sum_{j=2}^{9} \delta_j \mathbf{1}\{DIV_i = j\} + \nu_i$$

One would expect both coefficients to be positive.

Expected sign for β_1 (on $\ln(LANE)$)

One expects $\beta_1 > 0$.

This coefficient represents the elasticity of vehicle-kilometers traveled (VKT) with respect to lane kilometers. It seems logical to expect that an increase in road supply (more lanes) will lead to an increase in the total amount of driving. This could be due to induced demand, as new capacity makes driving more attractive, or simply by providing the infrastructure for traffic that was already latent. The "fundamental law of road congestion" itself suggests a positive, one-to-one relationship.

Expected sign for β_2 (on $\ln(POP)$)

One expects $\beta_2 > 0$.

This coefficient represents the elasticity of VKT with respect to population. It is a reasonable assumption that an MSA with more people will, ceteris paribus, have more total driving. A larger population implies more commuters, more commercial activity, and more personal trips, all of which contribute to a higher total VKT.

Question 7

Estimate model M_1 using OLS and report your results.

Question A

Test the overall significance of the regression at the 5% significance level, assuming residual normality. Use the result reported by the software and verify your conclusion manually by computing the test statistic with the residual sum of squares.

Table 1: OLS Estimation Results for Model M_1 (2003)

	(1)
(Intercept)	5.081***
	(0.557)
ln(LANE_IH)	0.749***
	(0.038)
ln(POP)	0.492***
	(0.038)
ELEV	−0.027
	(0.052)
RUG	5.717+
	(2.960)
HEAT	−0.011**
	(0.003)
COOL	−0.019**
	(0.007)
SPW	0.002
	(0.003)
DIV 2	−0.208+
	(0.112)
DIV 3	0.035
	(0.106)
DIV 4	0.003
	(0.118)
DIV 5	−0.003
	(0.128)
DIV 6	−0.043
	(0.130)
DIV 7	−0.130
	(0.133)
DIV 8	−0.266+
	(0.146)
DIV 9	−0.323+
	(0.170)
Num.Obs.	228
R2	0.955
R2 Adj.	0.952

We test the overall significance of the regression at the 5% significance level, assuming residual normality.

The null and alternative hypotheses are:

$$H_0 : \beta_1 = \beta_2 = \gamma_{\text{ELEV}} = \cdots = \delta_9 = 0$$

$$H_1 : \text{At least one slope coefficient is not zero.}$$

The null hypothesis states that all slope coefficients are jointly equal to zero, meaning the model has no explanatory power.

1. Result from Software

We first use the result reported by the R software summary:

- F-statistic: 301.4
- Numerator df (q): 15
- Denominator df ($n - k$): 212
- p-value: $<$ 2.2e-16

Conclusion: Since the p-value is extremely small (p-value $\ll 0.05$), we reject the null hypothesis H_0 at the 5% significance level. We conclude that the model is globally significant. The explanatory variables as a group have a statistically significant effect on the variation of log(VKT).

2. Manual Verification

We verify this result by manually computing the F-test statistic using the Residual Sum of Squares (RSS). The formula for the F-statistic is:

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-k)}$$

Where:

- RSS_1: Residual Sum of Squares from the unrestricted model (M_1).
- RSS_0: Residual Sum of Squares from the restricted model (M_0, intercept-only), which is also the Total Sum of Squares (TSS).
- $n = 228$: Number of observations.
- $k = 16$: Number of parameters in M_1 (1 intercept + 15 slopes).
- $q = k - 1 = 15$: Number of restrictions (the 15 slope coefficients).
- $n - k = 212$: Denominator degrees of freedom.

From the R output, we get the following values:

- RSS_1 (Unrestricted): 16.70731
- RSS_0 (Restricted / TSS): 373.0396

Plugging these values into the formula:

$$F = \frac{(373.0396 - 16.70731)/15}{16.70731/212} = \frac{356.33229/15}{0.078808} \approx 301.4348$$

This manually calculated F-statistic (301.4348) matches the value reported by the software (301.4).

To make a conclusion, we compare this F-statistic to the critical value from the F-distribution at the 5% significance level, $F_{\alpha,q,n-k}$:

$$F_{crit} = F_{0.05,15,212} = 1.713787$$

Conclusion: Since our test statistic ($F \approx 301.4$) is much larger than the critical value ($F_{crit} \approx 1.71$), we reject the null hypothesis H_0. This confirms the software's conclusion.

Question B

From this point onward, do not assume normally distributed error terms. Test the individual significance of all estimated β coefficients at the 5% significance level.

We test the individual significance of each estimated β coefficient using a two-sided t-test. The hypotheses for each coefficient β_j are:

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

We are told not to assume normally distributed error terms. However, our sample size ($n = 228$) is large enough to rely on asymptotic theory.

$$\sqrt{N}(\hat{\beta}_{OLS} - \beta) \xrightarrow{L} \mathcal{N}(0, \sigma^2 \mathcal{Q}^{-1})$$

$$t_j = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} \xrightarrow{L} \mathcal{N}(0, 1)$$

The t-statistics reported by the software will asymptotically follow a standard normal $N(0, 1)$ distribution.

We will use two common methods for our decision at the 5% significance level:

1. **p-value approach:** We reject H_0 if the p-value is less than 0.05.
2. **Critical value approach:** We reject H_0 if the absolute value of the t-statistic is greater than the critical value $z_{0.025} = 1.96$.

Based on the R output table from M1 (1)

- **(Intercept) (β_0):** $t = 9.118$, p-value $<$ 2e-16. Since p-value < 0.05 (and $|9.118| > 1.96$), we reject H_0. The intercept is statistically significant.
- **ln(LANE_IH) (β_1):** $t = 19.466$, p-value $<$ 2e-16. Since p-value < 0.05 (and $|19.466| > 1.96$), we reject H_0. The coefficient is statistically significant.
- **ln(POP) (β_2):** $t = 12.881$, p-value $<$ 2e-16. Since p-value < 0.05 (and $|12.881| > 1.96$), we reject H_0. The coefficient is statistically significant.
- **ELEV:** $t = -0.515$, p-value = 0.60743. Since p-value > 0.05, we fail to reject H_0. The coefficient is not statistically significant.
- **RUG:** $t = 1.932$, p-value = 0.05471. Since p-value > 0.05, we fail to reject H_0. The coefficient is not statistically significant at the 5% level (though it is at the 10% level).
- **HEAT:** $t = -3.333$, p-value = 0.00101. Since p-value < 0.05, we reject H_0. The coefficient is statistically significant.
- **COOL:** $t = -2.609$, p-value = 0.00972. Since p-value < 0.05, we reject H_0. The coefficient is statistically significant.
- **SPW:** $t = 0.809$, p-value = 0.41948. Since p-value > 0.05, we fail to reject H_0. The coefficient is not statistically significant.
- **DIV Dummies:** All census division dummies (`DIV_factor2` through `DIV_factor9`) have p-values much larger than 0.05. None of them are individually significant at the 5% level. This suggests that, relative to the reference division (Division 1), no other single division has a statistically different intercept.

Question C

Interpret the estimates of β_0, β_1, and β_2.

The estimated model is a log-log model. Therefore, the coefficients β_1 and β_2 are interpreted as elasticities. The intercept β_0 is the log-level of the dependent variable when all covariates are zero.

- **Interpretation of $\hat{\beta}_0$ (Intercept):** The estimated intercept is $\hat{\beta}_0 = 5.081$. This is the predicted value of log(VKT) for an MSA in the reference census division (Division 1) when all continuous regressors are equal to 0.

  This includes $\ln(LANE) = 0$ (i.e., $LANE = 1$) and $\ln(POP) = 0$ (i.e., $POP = 1$), as well as $ELEV = 0$, $RUG = 0$, etc. This scenario is not economically meaningful or realistic, so the intercept's value itself does not have a practical interpretation. It primarily serves to anchor the regression line.

- **Interpretation of $\hat{\beta}_1$ (ln(LANE_IH)):** The estimated coefficient is $\hat{\beta}_1 = 0.749$. This is the elasticity of VKT with respect to lane kilometers.

  Interpretation: A 1% increase in lane kilometers (LANE_IH) is associated with a 0.749% increase in daily vehicle-kilometers traveled (VKT_IH), holding population and all other control variables constant.

- **Interpretation of $\hat{\beta}_2$ (ln(POP)):** The estimated coefficient is $\hat{\beta}_2 = 0.492$. This is the elasticity of VKT with respect to population.

  Interpretation: A 1% increase in population (POP) is associated with a 0.492% increase in daily vehicle-kilometers traveled (VKT_IH), holding lane kilometers and all other control variables constant.

Question D

Formulate the null hypothesis that VKT is directly proportional to lane kilometers, with a proportionality constant of 1. Express the restriction in the form $\mathbf{R}\beta = \mathbf{r}$. Write down the expression for the test statistic used to test H_0. Conduct the test at the 1% significance level, and state your conclusion.

Formulate the Null Hypothesis

The model is a log-log model, where β_1 is the elasticity of VKT with respect to LANE. The hypothesis that "VKT is directly proportional to lane kilometers, with a proportionality constant of 1" means that a 1% increase in LANE leads to a 1% increase in VKT.

This is a test on the elasticity, so the null hypothesis is:

$$H_0 : \beta_1 = 1$$

The alternative hypothesis is $H_1 : \beta_1 \neq 1$.

Express as $R\beta = r$

The vector of parameters β is a (16×1) vector (1 intercept + 15 slopes). The coefficient β_1 is the second element in this vector (after the intercept).

We have $q = 1$ restriction. The matrix R is a (1×16) row vector:

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \end{bmatrix}$$

The vector r is a (1×1) scalar:

$$r = \begin{bmatrix} 1 \end{bmatrix}$$

The restriction $R\beta = r$ thus selects only the second coefficient: $0 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2 + \cdots = 1$, which simplifies to $\beta_1 = 1$.

The Test Statistic

As requested by the prompt, we write down the expression for the test statistic. Since we are not assuming normality, we use an asymptotic test. For a single restriction ($q = 1$), the most direct test statistic is the t-statistic.

The t-statistic for this hypothesis is:

$$t = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)}$$

Under H_0, this statistic follows a $N(0, 1)$ distribution asymptotically (due to $n = 228$ being large).

Conduct the Test

From our R output, we have the values for the $\ln(LANE\_IH)$ coefficient:

- $\hat{\beta}_1 = 0.7489286$
- $se(\hat{\beta}_1) = 0.03847447$

We calculate the t-statistic, which matches our R output:

$$t_{calc} = \frac{0.7489286 - 1}{0.03847447} = \frac{-0.2510714}{0.03847447} \approx -6.525662$$

We are testing at the 1% significance level ($\alpha = 0.01$). This is a two-sided test, so we use the critical value $z_{\alpha/2} = z_{0.005}$. Our R output calculated this critical value for us:

$$z_{crit} \approx 2.5758$$

The decision rule is to reject H_0 if $|t_{calc}| > z_{crit}$.

We find that $|-6.525662| = 6.525662$. Since $6.525662 > 2.5758$, we reject the null hypothesis.

Conclusion

At the 1% significance level, we reject the null hypothesis that $\beta_1 = 1$.

There is strong statistical evidence to conclude that the elasticity of VKT with respect to lane kilometers is not equal to 1. The estimated elasticity (0.749) is significantly less than 1. This finding contradicts the "fundamental law" in its strictest sense of a one-to-one proportionality.

Question 8

Based on their regression results, Duranton and Turner conclude that an MSA's density, measured by the variable `SPW`*, is not significantly associated with VKT. We want to further investigate the role of urban sprawl in shaping the dynamics of the fundamental law of road congestion. In this exercise, focus on one of the following interstate categories:* `IHU` *or* `IHNU`*.*

For this exercise, we choose to focus on the `IHU` (urbanized MSA interstates) category. This decision is justified as Question 8 investigates the impact of "urban sprawl" (`SPW`), a concept directly related to the density and structure of the urbanized portion of an MSA. Therefore, examining the congestion dynamics on the roads within this urbanized area is the most logical choice.

Question A

In a first step, drop all observations for which `VKT_x` *or* `LANE_x` *(for the chosen category x) equals zero. Why is this step essential for estimating model M_1? What are the implications of this step?*

Why is the step of dropping zero-value observations essential?

This step is essential because the specified regression, Model M_1, is a log-log model. The equation we will estimate uses the natural logarithm of VKT and LANE:

$$\ln(VKT_{IHU,i}) = \beta_0 + \beta_1 \ln(LANE_{IHU,i}) + \cdots + \nu_i$$

The natural logarithm function, $\ln(x)$, is only defined for $x > 0$. The value $\ln(0)$ is mathematically undefined.

Therefore, any observation (MSA) for which $VKT_{IHU} = 0$ or $LANE_{IHU} = 0$ must be dropped from the sample. It is impossible to compute the log-transformed variables for these observations, so they cannot be included in the regression analysis.

What are the implications of this step?

Dropping these observations has two main implications:

1. **Sample Size Reduction:** The sample size for the analysis is reduced. Our R script shows that the original 2003 sample has 228 MSAs. After filtering for $VKT_{IHU} > 0$ and $LANE_{IHU} > 0$, the new sample size is 214. We have dropped 14 MSAs. This reduces the statistical power of our tests and the precision of our estimates (i.e., standard errors will be larger, ceteris paribus).
2. **Loss of Generalizability (External Validity):** The 14 MSAs we dropped are not a random sample; they are MSAs that have zero urbanized interstates. These are likely smaller, less dense, or more rural MSAs. Our new estimates will be calculated based only on the subsample of MSAs that do have urbanized interstates.

This means our conclusions may no longer be generalizable to the entire population of MSAs. We are now estimating the effect of roads on VKT specifically for cities that have this type of infrastructure, and this effect might be different for the cities we excluded. This is a form of sample selection which could affect the external validity of our findings.

Question B

Next, split the MSAs into high- and low-sprawl groups based on the median value of `SPW`*. Define a new dummy variable* `SPW_high`*, which equals 1 if* `SPW` $\geq \text{med}($`SPW`$)$ *and 0 otherwise. Report the median value of* `SPW` *as well as the number of MSAs in each group.*

To split the MSAs, we first calculate the median value of the urban sprawl variable, SPW, using our sample of 214 MSAs.

- Median value of SPW: 44.20995

We then create a new dummy variable, `SPW_high`, which equals 1 for MSAs with an SPW value greater than or equal to this median, and 0 otherwise.

This splits our sample into two equally sized groups:

- Number of MSAs in the Low-Sprawl group (`SPW_high` = 0): 107
- Number of MSAs in the High-Sprawl group (`SPW_high` = 1): 107

Question C

Test the null hypothesis that the regression relationship between VKT and its covariates is the same across high- and low-sprawl MSAs. To this end, estimate the necessary model(s) and compute the appropriate test statistic.

We test the null hypothesis that the regression relationship (Model M_1) between VKT and its covariates is the same across high- and low-sprawl MSAs.

- H_0: The set of all coefficients is the same for both groups.
- H_1: At least one coefficient is different between the groups.

We use a Chow test, which is an F-test for a structural break. The test statistic is calculated as:

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-2k)}$$

Where:

- RSS_R: RSS from the pooled (restricted) model, run on all n observations = 14.20255
- RSS_{UR}: Sum of RSS from two separate (unrestricted) models, one for each group = $RSS_{low} + RSS_{high} = 5.993917 + 6.065181 = 12.0591$
- n: Total number of observations = 214
- k: Number of parameters estimated in one regression (Model M_1) = 16 (1 intercept + 7 continuous variables + 8 DIV dummies)
- q: Number of restrictions, which is equal to k for this test.

Plugging these in, we get the F-statistic:

$$F_{calc} = \frac{(14.20255 - 12.0591)/16}{12.0591/182} = \frac{2.14345/16}{0.0662588} \approx 2.021855$$

The associated p-value for this F-statistic is:

$$\text{p-value} \approx 0.01387$$

Question D

What do you conclude? In particular, what do the results imply about potential differences in the dynamics of the fundamental law of road congestion between sprawling and dense cities?

The null hypothesis of the Chow test is that the coefficients are stable across the two groups. We test at a standard 5% significance level ($\alpha = 0.05$).

Our calculated p-value is 0.0139. Since $0.0139 < 0.05$, we reject the null hypothesis H_0.

We conclude that there is a statistically significant structural difference in the regression relationship between high-sprawl and low-sprawl cities (for the IHU category).

This implies that the dynamics of the "fundamental law of road congestion" *do* appear to operate differently in sprawling versus dense cities. The coefficients of the model—including, potentially, the elasticity of VKT with respect to lanes (β_1)—are not statistically the same for both groups. This result *contradicts* the original paper's conclusion that an MSA's density (SPW) is not significantly associated with VKT; our finding suggests that SPW is a significant factor in shaping the *structure* of the relationship itself.

Question 9

We suspect that model M_1 suffers from non-spherical disturbances. Specifically, we assume that $\mathrm{V}[\nu_i \mid \mathbf{Z}] = \sigma^2 h(x_i)$, $\forall i$, *while* $\mathbb{E}[\nu_i \nu_j \mid \mathbf{Z}] = 0$, $\forall i \neq j$.

Question A

Discuss the implications of the two assumptions for the variance-covariance matrix of the error term and for the validity of the results above.

The two assumptions are:

1. $\mathbb{V}[\nu_i|\mathbf{Z}] = \sigma^2 h(x_i), \forall i$: This is the assumption of heteroskedasticity. It states that the variance of the error term is not constant (i.e., not σ^2) but changes for each observation i based on some function h of the explanatory variables x_i.
2. $\mathbb{E}[\nu_i \nu_j|\mathbf{Z}] = 0, \forall i \neq j$: This is the assumption of no autocorrelation. It states that the error for one observation (MSA) is uncorrelated with the error for any other observation.

Together, these two assumptions (violation of homoskedasticity but maintaining no autocorrelation) describe a model with pure heteroskedasticity.

Implications for the Variance-Covariance Matrix

The variance-covariance matrix of the error term, Ω, is an $n \times n$ matrix where:

- The diagonal elements are the variances: $\Omega_{ii} = \mathbb{V}[\nu_i|\mathbf{Z}]$
- The off-diagonal elements are the covariances: $\Omega_{ij} = \mathbb{E}[\nu_i \nu_j|\mathbf{Z}]$

Based on the two assumptions:

- The "no autocorrelation" assumption means all off-diagonal elements are 0.
- The "heteroskedasticity" assumption means the diagonal elements are $\sigma^2 h(x_i)$, which are not all equal to each other.

Therefore, the variance-covariance matrix Ω is a diagonal matrix, but not a scalar identity matrix:

$$\Omega = \begin{pmatrix} \sigma^2 h(x_1) & 0 & \dots & 0 \\ 0 & \sigma^2 h(x_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 h(x_n) \end{pmatrix} \neq \sigma^2 \mathbf{I}_n$$

This violates the assumption of spherical disturbances ($\Omega = \sigma^2 \mathbf{I}_n$) required by the Gauss-Markov theorem.

Implications for the Validity of the Results Above

This violation has critical consequences for our OLS results from Question 7:

- **Coefficient Estimates ($\hat{\beta}$):** The OLS coefficient estimates (e.g., $\hat{\beta}_1 = 0.749$) are still unbiased and consistent. Heteroskedasticity does not introduce bias into the coefficient estimates themselves.
- **Efficiency:** The OLS estimator is **no longer BLUE** (Best Linear Unbiased Estimator). Because the disturbances are non-spherical, OLS is inefficient. A different estimator (Weighted Least Squares) would be more efficient (i.e., have lower variance).

- **Standard Errors and Inference (t-tests, F-tests):** This is the most serious problem. The standard formula used by OLS to calculate standard errors ($se(\hat{\beta}) = \sqrt{s^2(X'X)^{-1}_{jj}}$) is incorrect, biased, and inconsistent in the presence of heteroskedasticity.
  - Since the standard errors are wrong, all statistical inference built upon them is invalid.
  - Consequently, all statistical tests from Question 7, including the overall F-test (Q7a), the individual t-tests (Q7b), and the hypothesis test (Q7d), are unreliable.

In summary, while our $\hat{\beta}$ point estimates are still unbiased, we can no longer trust any of the p-values, significance tests, or confidence intervals from Question 7.

Question B

Explain the underlying logic of the White test for heteroskedasticity. Perform the White test using an auxiliary regression that includes all regressors, their squared terms, and all interactions. Why might the results from this test be unreliable? Instead, perform the simplified White test using only the fitted values of the dependent variable, including both their linear and squared terms. What do you conclude?

Underlying Logic of the White Test

The White test is a general test for heteroskedasticity. It does not require us to specify the exact form of the heteroskedasticity.

The logic is as follows: if the model is homoskedastic ($\mathbb{V}[\nu_i|\mathbf{Z}] = \sigma^2$), then the squared residuals ($\hat{\nu}_i^2$) should be, on average, constant and should not have any systematic relationship with the explanatory variables (X).

The White test checks for such a relationship by running an auxiliary (helper) regression:

$$\hat{\nu}_i^2 = \alpha_0 + \alpha_1 X_{1,i} + \cdots + \alpha_k X_{k,i} + \text{other terms} + \text{error}$$

The null hypothesis is H_0 : Homoskedasticity. This implies that all slope coefficients in the auxiliary regression are zero. If we find that the X variables (and their squares/interactions) are jointly significant in explaining the squared residuals, we reject H_0 and conclude that heteroskedasticity is present.

Unreliability of the Full White Test

The "full" White test uses an auxiliary regression that includes all k original regressors, their squared terms, and all their unique cross-products.

Our model M_1 has $n = 228$ observations and $k = 16$ parameters, which means 15 slope regressors. The auxiliary regression for a full White test would include:

- 15 linear terms (X)
- 15 squared terms (X^2)
- $(15 \times 14)/2 = 105$ cross-product terms ($X_i \times X_j$)

This means the auxiliary regression would have $q = 15 + 15 + 105 = 135$ regressors.

This is highly unreliable because the number of regressors ($q = 135$) is extremely large relative to our sample size ($n = 228$). This "consumes" a very large number of degrees of freedom, a problem often referred to as the curse of dimensionality, which can lead to overfitting and very low power for the test.

Simplified White Test

One can use a simplified test using only the fitted values ($\hat{y}_i$) and their squares. This is a more parsimonious test (with $q = 2$) that still captures the essence of the White test.

The auxiliary regression is:

$$\hat{\nu}_i^2 = \alpha_0 + \alpha_1 \hat{y}_i + \alpha_2 \hat{y}_i^2 + \text{error}$$

The hypotheses are:

- $H_0 : \alpha_1 = \alpha_2 = 0$ (Homoskedasticity)
- $H_1 : \alpha_1 \neq 0$ or $\alpha_2 \neq 0$ (Heteroskedasticity)

We test this using the F-statistic from the auxiliary regression, which is preferred in finite samples.

From our R output, we have:

- $n = 228$
- R^2 from auxiliary regression = 0.0692
- F-statistic = 8.3648
- p-value (for F-statistic) = 0.000313

As the course now focuses on asymptotic theory, we can also perform the classic asymptotic Lagrange Multiplier (LM) test. This statistic is calculated as $LM = n \times R^2$ and follows a Chi-square (χ^2) distribution with $q = 2$ degrees of freedom.

From our R output, we have:

- $LM = n \times R^2 = 228 \times 0.0692 = 15.779$
- p-value (for LM-statistic) = 0.00037

Conclusion: We test at the 5% significance level ($\alpha = 0.05$). Since the p-value (0.000313) is much smaller than 0.05, we reject the null hypothesis H_0. This asymptotic test also provides a p-value far below 0.05, confirming the F-test's conclusion.

We conclude that there is strong statistical evidence of heteroskedasticity. The variance of the error term is not constant, which confirms our suspicions from Question 9a and invalidates the standard errors and inference from Question 7.

Question C

Explain the underlying logic of White's robust estimator in contrast to Weighted Least Squares.

Weighted Least Squares (WLS)

Weighted Least Squares (WLS), a specific form of Generalized Least Squares (GLS), is the most efficient (best) estimator if we know the exact form of the heteroskedasticity.

The logic is to find a transformation matrix $\boldsymbol{P}$ (where $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{\Omega}^{-1}$) that "sphericalizes" the errors. This matrix is pre-multiplied by the original model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{u}$:

$$\underbrace{(\boldsymbol{P}\boldsymbol{y})}_{\tilde{\boldsymbol{y}}} = \underbrace{(\boldsymbol{P}\boldsymbol{X})}_{\tilde{\boldsymbol{X}}}\boldsymbol{b} + \underbrace{(\boldsymbol{P}\boldsymbol{u})}_{\tilde{\boldsymbol{u}}}$$

This creates a new, transformed model $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{b} + \tilde{\boldsymbol{u}}$. The new error term $\tilde{\boldsymbol{u}}$ is now homoskedastic and non-autocorrelated because its variance is $\mathbb{V}[\tilde{\boldsymbol{u}}] = \sigma^2 \boldsymbol{I}_n$. Running OLS on this transformed model is the efficient WLS (GLS) estimator.

For example, if we assume $\mathbb{V}[\nu_i|\boldsymbol{Z}] = \sigma^2 POP_i$, this transformation is practically achieved by setting the matrix $\boldsymbol{P}$ to be a diagonal matrix with $1/\sqrt{POP_i}$ on the diagonal. This is equivalent to dividing every variable in the regression by $\sqrt{POP_i}$.

The main problem with WLS, however, is that we almost never know the true form of $h(x_i)$. Guessing the wrong form (i.e., specifying the wrong weights) makes our estimates biased and inconsistent.

White's Robust Estimator

White's robust estimator takes a different, more "agnostic" approach.

The logic is: We admit that OLS is inefficient, but we don't care. We keep the $\hat{\beta}^{OLS}$ coefficients (which are still unbiased and consistent). We just want to fix the standard errors so our inference (t-tests, p-values) is valid.

White's estimator does not transform the model. It directly computes a new variance-covariance matrix that is "robust" to the presence of heteroskedasticity of an unknown form. It uses the squared residuals, $\hat{\nu}_i^2$, as an estimate for the unknown variance $\sigma^2 h(x_i)$ for each observation.

The resulting standard errors are called "White's robust standard errors." They are only valid asymptotically (in large samples), but they allow us to conduct valid tests on our original $\hat{\beta}^{OLS}$ coefficients without ever needing to know the true form of the heteroskedasticity.

WLS is efficient but requires a strong assumption. White's estimator is robust because it requires no assumptions about the form of $h(x_i)$ and simply fixes the standard errors for our (potentially inefficient) OLS estimates.

Question D

Re-estimate model M_1 using White heteroskedasticity-robust standard errors. Compare the estimates and standard errors to those from Question 7.

We re-estimate the model to obtain heteroskedasticity-consistent (HC) standard errors, as our White test in 9b confirmed that heteroskedasticity is present. This procedure does not change the OLS coefficient estimates ($\hat{\beta}$), but it calculates correct standard errors, t-statistics, and p-values.

Table 2: OLS Estimation Results with Robust Standard Errors (Q9d)

	(1)
(Intercept)	5.081***
	(0.577)
ln(LANE_IH)	0.749***
	(0.041)
ln(POP)	0.492***
	(0.042)
ELEV	−0.027
	(0.053)
RUG	5.717+
	(3.060)
HEAT	−0.011**
	(0.003)
COOL	−0.019*
	(0.008)
SPW	0.002
	(0.003)
DIV 2	−0.208*
	(0.096)
DIV 3	0.035
	(0.097)
DIV 4	0.003
	(0.111)
DIV 5	−0.003
	(0.121)
DIV 6	−0.043
	(0.123)
DIV 7	−0.130
	(0.135)
DIV 8	−0.266+
	(0.151)
DIV 9	−0.323+
	(0.174)
Num.Obs.	228
R2	0.955
R2 Adj.	0.952

The full robust results are in Table 2. We can compare the original standard errors (from Q7) with the new robust standard errors (from Q9d) for our main variables:

Variable	Estimate ($\hat{\beta}$)	Old SE	New (Robust) SE	Change
ln(LANE_IH)	0.749	0.03847	0.04140	+7.6%
ln(POP)	0.492	0.03822	0.04211	+10.2%

Table 3: Comparison of Standard Errors

Comparison and Conclusion

For both of our main variables of interest, `ln(LANE_IH)` and `ln(POP)`, the robust standard errors are larger than the original ones (by 7.6% and 10.2%, respectively). This implies that the original OLS standard errors were biased downwards, making us overly confident in our estimates.

Consequently, the t-statistics for these variables have decreased:

- **ln(LANE_IH):** t-statistic fell from 19.47 to 18.09.
- **ln(POP):** t-statistic fell from 12.88 to 11.69.

Despite this, the p-values for both variables remain extremely small (p < 2.2e-16), so our conclusion that they are highly statistically significant is unchanged.

Interestingly, when looking at all variables, the significance of `DIV_factor2` changed. Its original p-value was 0.0656 (not significant at 5%), but its new robust p-value is 0.03238 (significant at 5%). This shows how correcting for heteroskedasticity can change inference in either direction and is therefore a crucial step.

Question E

We suspect that the variance of ν_i depends on an MSA's population. Specify the auxiliary regression required for a Breusch–Pagan test, perform the test, and interpret your results.

The Breusch-Pagan (BP) test checks for a specific, linear form of heteroskedasticity, where the error variance is assumed to be a linear function of one or more variables. We are asked to test the suspicion that the variance depends on "an MSA's population", which we interpret as the variable `POP`.

Specify the Auxiliary Regression

The auxiliary regression required for this test regresses the squared OLS residuals ($\hat{\nu}_i^2$) from Model M_1 on the variable(s) suspected of causing the heteroskedasticity. In this case, that variable is `POP`.

The auxiliary regression is:

$$\hat{\nu}_i^2 = \alpha_0 + \alpha_1 POP_i + \text{error}_i$$

Perform the Test

We test the null hypothesis of homoskedasticity ($H_0 : \alpha_1 = 0$) against the alternative of heteroskedasticity ($H_1 : \alpha_1 \neq 0$). The test statistic (BP) follows a χ^2 (Chi-square) distribution with 1 degree of freedom (for the single regressor, `POP`).

From our R output, performing the test with `POP` as the variable:

- BP statistic = 0.5932
- df = 1
- p-value = 0.4412

Interpretation of the results

We test at the 5% significance level ($\alpha = 0.05$).

Since the p-value (0.4412) is much larger than 0.05, we fail to reject the null hypothesis. We conclude that there is no statistical evidence that the error variance is linearly related to the level of an MSA's population (the variable `POP`).

Note on `log(POP)`

This result might seem to contradict our earlier White test (in 9b), which found strong evidence of general heteroskedasticity.

Our R output also shows a second BP test, which tests the relationship between the variance and `log(POP)`, the variable actually included in our regression model M_1.

For the auxiliary regression $\hat{\nu}_i^2 = \alpha_0 + \alpha_1 \ln(POP_i) + \text{error}_i$:

- BP statistic = 9.598
- df = 1
- p-value = 0.001948

This p-value is highly significant ($p < 0.01$). This confirms that heteroskedasticity is present, and it is specifically correlated with the `log(POP)` regressor. This is consistent with our general White test finding and suggests that the relationship between variance and population is non-linear (which is why the test on `POP` failed, as it is only designed to detect linear relationships). The test on `log(POP)` succeeded because the log transformation linearizes this non-linear relationship, allowing the test to detect it.

Question F

Plot the residuals against the explanatory variable `POP`. Do you observe increasing or decreasing variance with `POP`? Why could this occur?

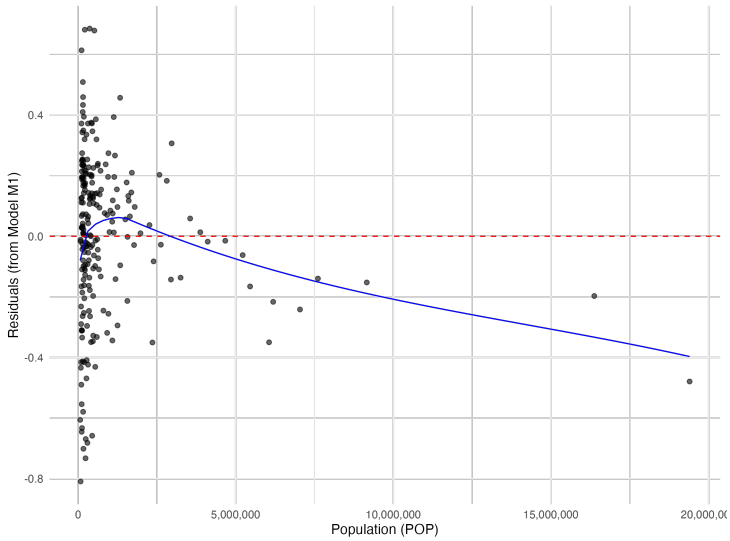

Figure 4: Plot of OLS Residuals (from M_1) against Population (POP)

The residuals in this plot are from Model M_1, which is defined as:

$$\ln(VKT_i) = \beta_0 + \beta_1 \ln(LANE_i) + \beta_2 \ln(POP_i) + \gamma \boldsymbol{X_i} + \sum_{j=1}^{9} \delta_j \mathbf{1}\{DIV_i = j\} + \nu_i$$

where $\boldsymbol{X_i}$ includes the other control variables (ELEV, RUG, HEAT, COOL, SPW).

Do you observe increasing or decreasing variance?

The plot of residuals against the `POP` variable shows clear visual evidence of heteroskedasticity.

Specifically, we observe decreasing variance as population increases. The points exhibit a "fanning-in" or reverse-megaphone shape. For MSAs with a low population (e.g., $POP < 2,500,000$), the residuals are widely spread, ranging from approximately -0.8 to +0.6. For MSAs with a high population (e.g., $POP > 5,000,000$), the residuals are much more tightly clustered around the zero line.

Why could this occur?

This pattern confirms the results of our statistical tests in 9b and 9e: heteroskedasticity is present and is related to population.

This specific "fanning-in" shape likely occurs because we are plotting the residuals from a log-log model against a level variable (`POP`) that is highly skewed.

The vast majority of MSAs have relatively low populations, leading to a large cluster of points on the left side of the graph with a wide variety of prediction errors (residuals). The few MSAs with very large populations are outliers in the `POP` variable. Our model, by using `log(POP)`, compresses these large values. This log transformation might be very effective at modeling these large cities, leading to systematically smaller prediction errors (residuals) for them, hence the decreasing variance.

Question G

Explain the underlying logic of the Goldfeld-Quandt test. Formulate the null and alternative hypotheses in light of your conclusion above. Perform the test and discuss the results

Underlying Logic of the Goldfeld-Quandt Test

The Goldfeld-Quandt (GQ) test is a test for heteroskedasticity. Its logic relies on the assumption that we can identify a specific variable that orders the error variance.

The test proceeds in these steps:

1. The data is sorted based on the variable suspected of being related to the error variance (in our case, `POP`).
2. A central fraction of the observations is dropped (as per the hint, we drop the middle 20%). This creates two distinct groups: one with low `POP` (Group 1) and one with high `POP` (Group 2).
3. Two separate OLS regressions are estimated, one on each subgroup.
4. The Residual Sum of Squares (RSS) is calculated for each regression: RSS_1 (for the low-`POP` group) and RSS_2 (for the high-`POP` group).
5. The test statistic is the ratio of the two RSS values: $GQ = RSS_2/RSS_1$.

Under the null hypothesis of homoskedasticity, the error variance is the same in both groups ($\sigma_1^2 = \sigma_2^2$). We would thus expect $RSS_1 \approx RSS_2$, and the GQ statistic would be close to 1. If the GQ statistic is significantly different from 1, we reject the null hypothesis.

Specifically, the model M_1 (which has $k = 16$ parameters) is estimated twice, once for each subgroup, after sorting the data by `POP`:

- Regression 1 (Low-POP Group): M_1 is estimated using the subsample of data with the lowest `POP` values, yielding RSS_1.
- Regression 2 (High-POP Group): M_1 is estimated using the subsample of data with the highest `POP` values, yielding RSS_2.

Formulate Hypotheses

In light of our conclusion from Question 9f, we observed decreasing variance as `POP` increases.

- This means we expect a high variance for the low-`POP` group (Group 1, σ_1^2).
- We expect a low variance for the high-`POP` group (Group 2, σ_2^2).

Therefore, our null and alternative hypotheses are:

- $H_0 : \sigma_1^2 = \sigma_2^2$ (Homoskedasticity)
- $H_1 : \sigma_1^2 > \sigma_2^2$ (Decreasing variance)

Since the test statistic is $GQ = RSS_2/RSS_1$, our alternative hypothesis (H_1) implies that we expect $RSS_1 > RSS_2$, which means we expect to find a test statistic $GQ < 1$. This corresponds to the R 'alternative = "less"' argument.

Perform Test and Discuss Results

From our R output, the results of the Goldfeld-Quandt test are:

- Test Statistic (GQ) = 0.42045
- df1 = 76, df2 = 75
- p-value = 0.000108

The R output also explicitly confirms our alternative hypothesis: 'alternative hypothesis: variance decreases from segment 1 to 2', which is the same as $\sigma_1^2 > \sigma_2^2$.

Conclusion: We test at any standard significance level (e.g., $\alpha = 0.05$ or $\alpha = 0.01$). Since the p-value (0.000108) is much smaller than 0.01, we reject the null hypothesis H_0.

We conclude that there is strong statistical evidence of heteroskedasticity. The test specifically supports our alternative hypothesis of decreasing variance. This confirms our visual inspection from the plot in 9f and our findings from the Breusch-Pagan test on `log(POP)`. The error variance is significantly larger for MSAs with smaller populations.

Question H

Estimate model M_1 using the Weighted Least Squares estimator. Compare the coefficient estimates and standard errors with those obtained earlier.

We now estimate Model M_1 using a more sophisticated and data-driven approach, Feasible Generalized Least Squares (FGLS). This is a form of WLS that is directly motivated by our findings in Question 9e.

In Q9e, our Breusch-Pagan test strongly suggested that the error variance is related to $\ln(POP)$ (p-value = 0.0019). Therefore, instead of assuming a simple variance function (like $\mathbb{V}[\nu_i] \propto 1/POP_i$), the FGLS method models the variance. We do this by:

1. Running OLS and getting the residuals $\hat{\nu}_i$.
2. Estimating the variance function: $\ln(\hat{\nu}_i^2) = \alpha_0 + \alpha_1 \ln(POP_i) + \text{error}$.
3. Using the predicted values from this regression, $\widehat{\ln(\hat{\nu}_i^2)}$, to calculate the predicted variance, $\hat{\sigma}_i^2 = \exp(\widehat{\ln(\hat{\nu}_i^2)})$. Using the inverse of this predicted variance ($w_i = 1/\hat{\sigma}_i^2$) as the weight in our final WLS estimation.

Table 4: Comparison of OLS, OLS-Robust, and FGLS Estimators (Q9h)

	OLS (Q7)	OLS-Robust (Q9d)	FGLS (Q9h)
(Intercept)	5.081***	5.081***	5.300***
	(0.557)	(0.577)	(0.530)
ln(LANE_IH)	0.749***	0.749***	0.760***
	(0.038)	(0.041)	(0.038)
ln(POP)	0.492***	0.492***	0.465***
	(0.038)	(0.042)	(0.037)
ELEV	−0.027	−0.027	−0.030
	(0.052)	(0.053)	(0.047)
RUG	5.717+	5.717+	5.840+
	(2.960)	(3.060)	(2.963)
HEAT	−0.011**	−0.011**	−0.010**
	(0.003)	(0.003)	(0.003)
COOL	−0.019**	−0.019*	−0.017*
	(0.007)	(0.008)	(0.007)
SPW	0.002	0.002	0.001
	(0.003)	(0.003)	(0.003)
DIV 2	−0.208+	−0.208*	−0.247*
	(0.112)	(0.096)	(0.102)
DIV 3	0.035	0.035	0.010
	(0.106)	(0.097)	(0.097)
DIV 4	0.003	0.003	−0.031
	(0.118)	(0.111)	(0.111)
DIV 5	−0.003	−0.003	0.017
	(0.128)	(0.121)	(0.116)
DIV 6	−0.043	−0.043	−0.028
	(0.130)	(0.123)	(0.121)
DIV 7	−0.130	−0.130	−0.122
	(0.133)	(0.135)	(0.122)
DIV 8	−0.266+	−0.266+	−0.245+
	(0.146)	(0.151)	(0.139)
DIV 9	−0.323+	−0.323+	−0.303+
	(0.170)	(0.174)	(0.161)
Num.Obs.	228	228	228
R2	0.955	0.955	0.966

Comparison

By comparing the three columns in Table 4, we can analyze the impact of our FGLS procedure:

- **Coefficient Estimates:** The FGLS coefficient estimates (column 3, "FGLS (Q9h)") are different from the OLS estimates. The elasticity of VKT with respect to lanes, $\ln(LANE\_IH)$, increases slightly from 0.749 (OLS) to 0.760 (FGLS). The elasticity with respect to population, $\ln(POP)$, decreases from 0.492 (OLS) to 0.465 (FGLS).
- **Standard Errors:** The FGLS standard errors (in parentheses in column 3) are now the smallest for our key variables, demonstrating the efficiency gain from this method.
  - For $\ln(LANE\_IH)$, the FGLS SE (0.038) is smaller than the robust OLS SE (0.041).
  - For $\ln(POP)$, the FGLS SE (0.037) is also the smallest, and significantly smaller than the robust OLS SE (0.042).

Final Conclusion

In this analysis, we compared three methods. OLS (Q7) was invalid due to heteroskedasticity. OLS-Robust (Q9d) provided valid inference but was inefficient. This FGLS model (Q9h) provides the most compelling set of results.

One can note that our previous simple WLS model (from 'weights = POP', ran on R) produced a slightly smaller standard error for $\ln(POP)$ (0.035) than this FGLS model (0.037).

So why choose FGLS? The FGLS model is more theoretically sound and better justified.

- The simple WLS model assumed a rigid, perfect relationship ($\mathbb{V} \propto 1/POP$), which was just an educated guess based on our Q9g test.
- The FGLS model made no such rigid assumption. It flexibly estimated the relationship between variance and population, using the exact functional form ($\ln(\hat{\nu}^2) \sim \ln(POP)$) that our Breusch-Pagan test in Q9e showed was significant.

Because the FGLS method's assumption is directly supported by our diagnostic tests, it is the most rigorous and statistically defensible approach. It confirms that accounting for heteroskedasticity is crucial and provides what we can be most confident in as our final, efficient estimates.

Therefore, our most reliable conclusion is that the elasticity of VKT with respect to lanes is 0.760, and the elasticity with respect to population is 0.465.