In [1]:
```
import numpy as np
import pandas as pd
import altair as alt
import sklearn.linear_model as lm
import warnings
from sklearn.preprocessing import add_dummy_feature
warnings.simplefilter(action='ignore', category=FutureWarning)
```

Background: California Department of Developmental Services

From Taylor, S. A., & Mickel, A. E. (2014). Simpson's Paradox: A Data Set and Discrimination Case Study Exercise. Journal of Statistics Education, 22(1):

> Most states in the USA provide services and support to individuals with developmental disabilities (e.g., intellectual disability, cerebral palsy, autism, etc.) and their families. The agency through which the State of California serves the developmentally-disabled population is the California Department of Developmental Services (DDS) ... One of the responsibilities of DDS is to allocate funds that support over 250,000 developmentally-disabled residents. A number of years ago, an allegation of discrimination was made and supported by a univariate analysis that examined average annual expenditures on consumers by ethnicity. The analysis revealed that the average annual expenditures on Hispanic consumers was approximately one-third of the average expenditures on White non-Hispanic consumers. This finding was the catalyst for further investigation; subsequently, state legislators and department managers sought consulting services from a statistician.

---

1. Exploratory analysis

In [2]:
```
dds = pd.read_csv('california-dds.csv')
dds.head()
```

Out[2]:

	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity
0	10210	13 to 17	17	Female	2113	White not Hispanic
1	10409	22 to 50	37	Male	41924	White not Hispanic
2	10486	0 to 5	3	Male	1454	Hispanic
3	10538	18 to 21	19	Female	6400	Hispanic
4	10568	13 to 17	13	Male	4412	White not Hispanic

In [3]:

```
# compute median expenditures
median_expend_by_eth = dds.loc[:, ['Ethnicity', 'Expenditures']].groupby('Et

# compute sample sizes
ethnicity_n = dds['Ethnicity'].value_counts()

# concatenate
tbl_1 = pd.concat([median_expend_by_eth, ethnicity_n], axis = 1, join = 'out
# print
tbl_1 = tbl_1.rename(columns = {'Ethnicity': 'n'})
tbl_1
```

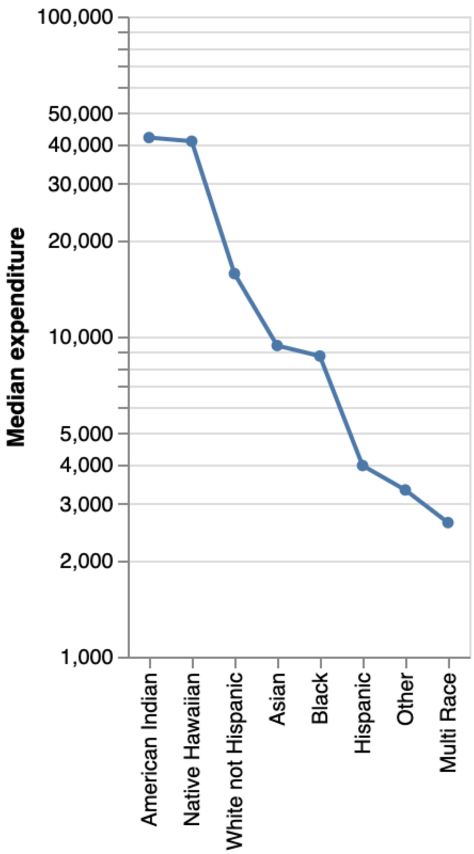
Out[3]:

	Expenditures	n
American Indian	41817.5	4
Asian	9369.0	129
Black	8687.0	59
Hispanic	3952.0	376
Multi Race	2622.0	26
Native Hawaiian	40727.0	3
Other	3316.5	2
White not Hispanic	15718.0	401

In [4]:

```
# Fig showing the median expenditure by ethnicity
data = tbl_1.reset_index()

fig_1 = alt.Chart(data).mark_line(point = True).encode(
    x = alt.X('index:O', sort = alt.EncodingSortField(field = 'Expenditures'
    y = alt.Y('Expenditures:Q', scale=alt.Scale(type="log"), title = 'Median
)

fig_1
```

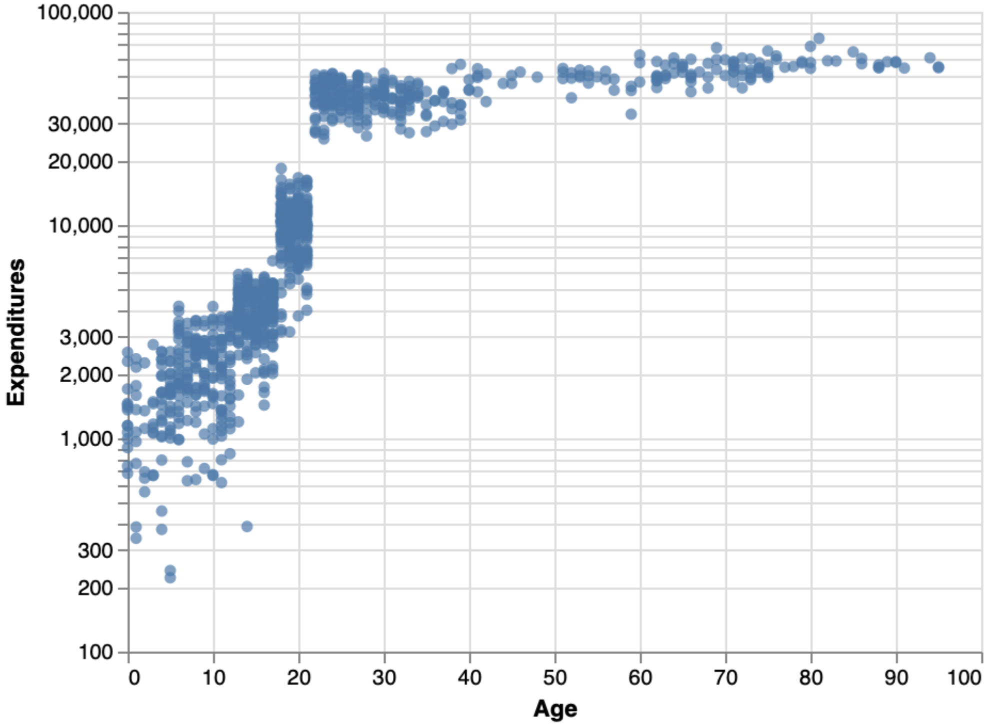
Out[4]:



In [5]:

```
# fig showing expenditure by age on a log scale
fig_2 = alt.Chart(dds).mark_circle().encode(
    x = 'Age:Q',
    y = alt.Y('Expenditures', scale = alt.Scale(type = 'log'))
)

fig_2
```

Out[5]:



(ii) Does the relationship seem linear?

If so, describe the direction (positive/negative) and approximate strength (steep/slight) of relationship. If not, describe the pattern of relationship, if any, in 1-2 sentences.

We do not see a linear scale as in the expenditures go up from age 0-20 but then from age 30-100 it is aboout the same the whole time. A reason for this could be around 20 is when you move out of your home and do not have your family to support you for funding.

In [6]:

```
# convert data types
dds_cat = dds.astype({'Age Cohort': 'category', 'Ethnicity': 'category', 'Ge

dds_cat['Age Cohort'] = dds_cat['Age Cohort'].cat.as_ordered().cat.reorder_c
    dds_cat['Age Cohort'].cat.categories[[0, 5, 1, 2, 3, 4]]
)

# age cohorts
dds_cat['Age Cohort'].cat.categories
```

Out[6]:

```
Index(['0 to 5', '6 to 12', '13 to 17', '18 to 21', '22 to 50', '51+'], dtyp
e='object')
```

Here is an explanation of how the cohort age boundaries were chosen:

> The 0-5 cohort (preschool age) has the fewest needs and requires the least amount of funding. For the 6-12 cohort (elementary school age) and 13-17 (high school age), a number of needed services are provided by schools. The 18-21 cohort is typically in a transition phase as the consumers begin moving out from their parents' homes into community centers or living on their own. The majority of those in the 22-50 cohort no longer live with their parents but may still receive some support from their family. Those in the 51+ cohort have the most needs and require the most amount of funding because they are living on their own or in community centers and often have no living parents.

In [7]:

```
# group by agr groups and ethnicity
df1 = dds_cat.groupby(['Age Cohort', 'Ethnicity'])

# Find the total amount of people in each age group / ethnicity
df2 = df1.Id.count()
df3 = df2.reset_index()
samp_sizes = df3.rename(columns = {'Id': 'n'})

# print
samp_sizes
```

Out[7]:

	Age Cohort	Ethnicity	n
0	0 to 5	American Indian	0
1	0 to 5	Asian	8
2	0 to 5	Black	3
3	0 to 5	Hispanic	44
4	0 to 5	Multi Race	7
5	0 to 5	Native Hawaiian	0
6	0 to 5	Other	0
7	0 to 5	White not Hispanic	20
8	6 to 12	American Indian	0
9	6 to 12	Asian	18
10	6 to 12	Black	11
11	6 to 12	Hispanic	91
12	6 to 12	Multi Race	9

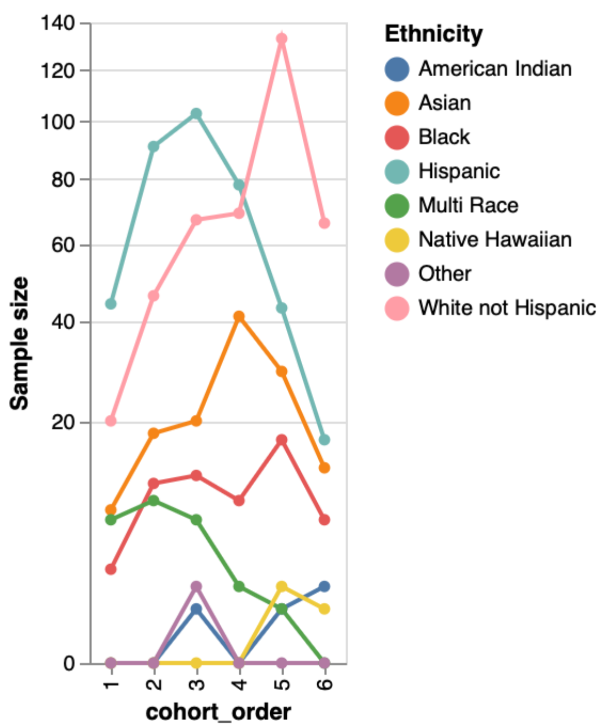
13	6 to 12	Native Hawaiian	0
14	6 to 12	Other	0
15	6 to 12	White not Hispanic	46
16	13 to 17	American Indian	1
17	13 to 17	Asian	20
18	13 to 17	Black	12
19	13 to 17	Hispanic	103
20	13 to 17	Multi Race	7
21	13 to 17	Native Hawaiian	0
22	13 to 17	Other	2
23	13 to 17	White not Hispanic	67
24	18 to 21	American Indian	0
25	18 to 21	Asian	41
26	18 to 21	Black	9
27	18 to 21	Hispanic	78
28	18 to 21	Multi Race	2
29	18 to 21	Native Hawaiian	0
30	18 to 21	Other	0
31	18 to 21	White not Hispanic	69
32	22 to 50	American Indian	1
33	22 to 50	Asian	29
34	22 to 50	Black	17
35	22 to 50	Hispanic	43
36	22 to 50	Multi Race	1
37	22 to 50	Native Hawaiian	2
38	22 to 50	Other	0
39	22 to 50	White not Hispanic	133
40	51+	American Indian	2
41	51+	Asian	13
42	51+	Black	7
43	51+	Hispanic	17
44	51+	Multi Race	0

45	51+	Native Hawaiian	1
46	51+	Other	0
47	51+	White not Hispanic	66

In [8]:

```
# add column with category codes
samp_sizes['cohort_order'] = samp_sizes['Age Cohort'].map({'0 to 5':1, '6 tc
                                                   '22 to 50': 5, '

# construct plot
fig_3 = alt.Chart(samp_sizes).mark_line(point = True).encode(
    x = 'cohort_order:O',
    y = alt.Y('n:Q', title = 'Sample size', scale = alt.Scale(type = 'sqrt')
    color = 'Ethnicity')

# display
fig_3
```

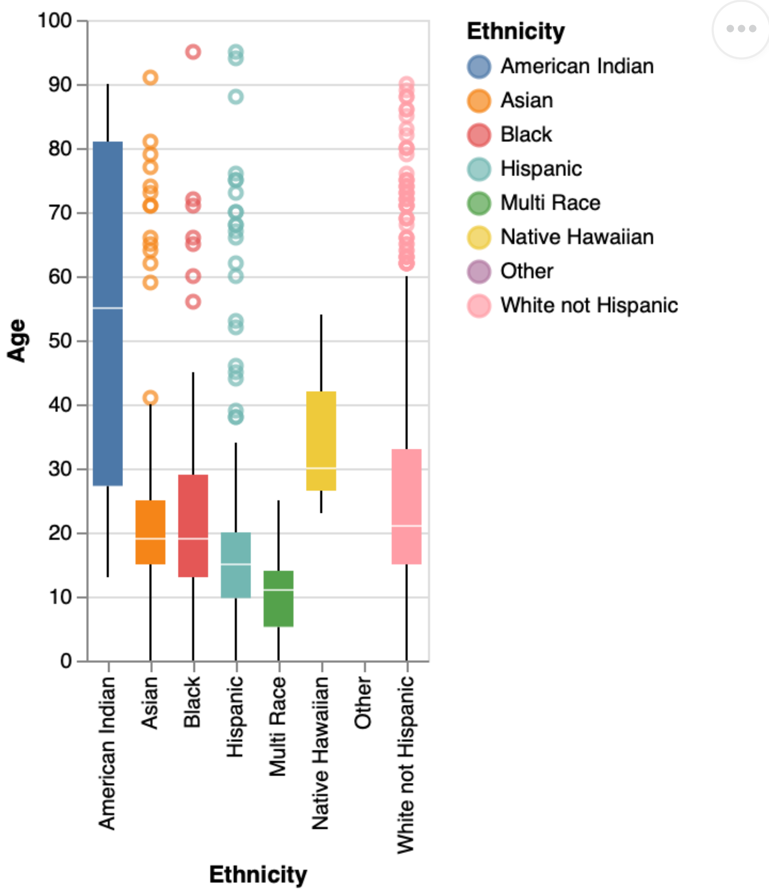
Out[8]:



We start to see for ethnic groups that were accusing of being discriminated against their sample size for the data is very low, which is a good indictor of bias.

In [9]:
```
# Box plot for age grouped by ethnicity
fig = alt.Chart(dds_cat).mark_boxplot().encode(
    x = 'Ethnicity:O',
    y = 'Age:Q',
    color = 'Ethnicity'
)

fig
```

Out[9]:



In [10]:
```
dds_cat.loc[dds_cat.Ethnicity == 'Multi Race']
```

Out[10]:

	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity
13	11189	13 to 17	17	Male	5340	Multi Race
30	12850	13 to 17	13	Male	3775	Multi Race
84	18383	0 to 5	0	Male	1149	Multi Race
145	22988	13 to 17	16	Male	4664	Multi Race
191	26437	0 to 5	0	Male	2296	Multi Race
243	31168	6 to 12	11	Female	2918	Multi Race
288	35360	6 to 12	10	Female	1622	Multi Race
330	39942	13 to 17	14	Male	3399	Multi Race
362	43291	6 to 12	11	Male	2140	Multi Race
393	45755	6 to 12	11	Male	1144	Multi Race
410	47043	22 to 50	25	Male	38619	Multi Race
443	50222	18 to 21	19	Female	7564	Multi Race
517	56736	18 to 21	18	Female	11054	Multi Race
569	61120	6 to 12	7	Male	3000	Multi Race
570	61187	6 to 12	11	Male	2885	Multi Race
668	69542	0 to 5	5	Female	1053	Multi Race
686	71073	13 to 17	14	Female	5062	Multi Race
839	84388	0 to 5	2	Female	697	Multi Race
871	87444	13 to 17	14	Female	1893	Multi Race
906	90953	6 to 12	10	Female	669	Multi Race
934	93628	6 to 12	6	Male	3259	Multi Race
948	94595	0 to 5	4	Female	2335	Multi Race
977	97426	0 to 5	1	Female	2359	Multi Race
978	97793	6 to 12	9	Female	1048	Multi Race
994	99529	0 to 5	2	Male	2258	Multi Race
997	99718	13 to 17	17	Female	3673	Multi Race

Looking at the ethnicity with the lowest average age, we can see that the expenditure numbers are low as well giving us a hint there might be some correlation between them both

In [11]:
```
dds_cat.loc[dds_cat.Ethnicity == 'American Indian']
```

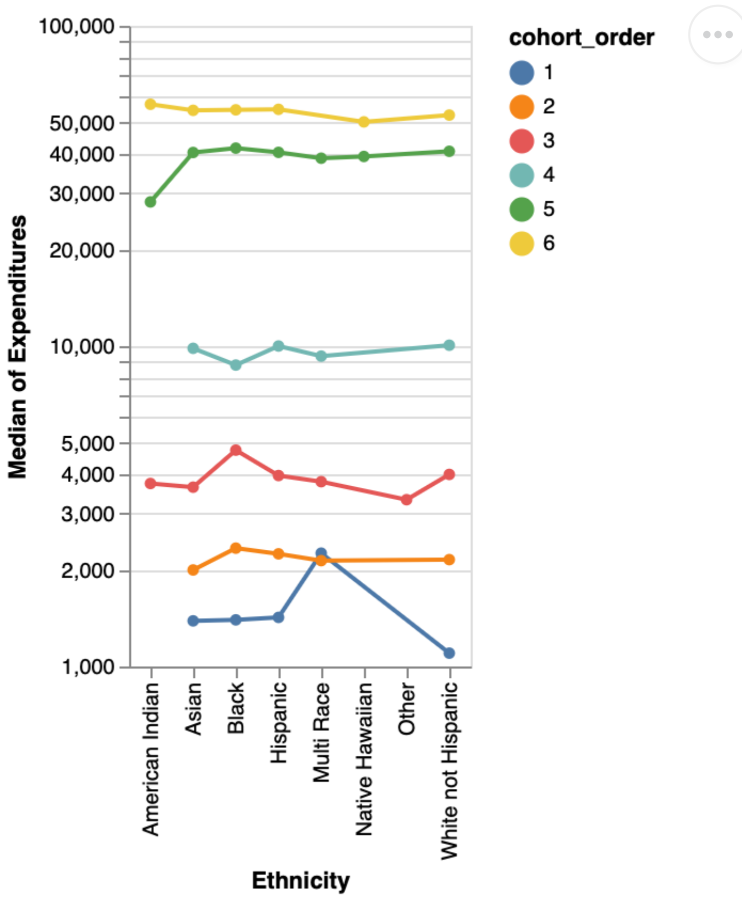
Out[11]:

	Id	Age Cohort	Age	Gender	Expenditures	Ethnicity
231	30234	51+	78	Female	55430	American Indian
575	61498	13 to 17	13	Female	3726	American Indian
730	74721	51+	90	Female	58392	American Indian
788	79645	22 to 50	32	Male	28205	American Indian

The ethnicity with the highest average only has 4 rows which is not enough to give a good judgment on the data

In [12]:
```
# add column with category codes
dds_cat['cohort_order'] = dds_cat['Age Cohort'].map({'0 to 5':1, '6 to 12':
                                                    '22 to 50': 5, '
# construct plot
fig_4 = alt.Chart(dds_cat).mark_line(point = True).encode(
    x = 'Ethnicity:O',
    y = alt.Y('median(Expenditures):Q', scale = alt.Scale(type = 'log')),
    color = 'cohort_order'
)

# display
fig_4
```

Out[12]:



We can see the median expenditure goes up throughout every age group.

---

2. Regression analysis

Now after thoroughly exploring the data, we will look at using a linear model to estimate the differences in median expenditure that was observed graphically in part 1.

More specifically, we will model the log of expenditures (response variable) as a function of gender, age cohort, and ethnicity:

$$\log(\text{expend}_i) = \beta_0 + \beta_1(\text{6-12})_i + \cdots + \beta_5(51+)_i + \beta_6\text{male}_i + \beta_7\text{hispanic}_i + \cdots + \beta$$

In this model, *all* of the explanatory variables are categorical and encoded using indicators; in this case, the linear model coefficients capture means for each group.

In this model the response variable is log-transformed and all explanatory variables are categorical

Commments about parameter interpretation

In particular, each coefficient represents a difference in means from the 'baseline' group. All indicators are zero for a white male recipient between ages 0 and 5, so this is the baseline group and:

$$\mathbb{E}\left(\log(\text{expend}) \mid \text{male, white, 0-5}\right) = \beta_1$$

Then, the expected log expenditure for a hispanic male recipient between ages 0 and 5 is:

$$\mathbb{E}\left(\log(\text{expend}) \mid \text{male, hispanic, 0-5}\right) = \beta_0 + \beta_7$$

So β_7 is *the difference in mean log expenditure between hispanic and white recipients after accounting for gender and age*. The other parameters have similar interpretations.

The parameters represent marginal differences in means between genders (holding age and ethnicity fixed), between ages (holding gender and ethnicity fixed), and between ethnicities (holding age and gender fixed).

Comments about the log transformation

The response in this model is the *log* of expenditures. The statistical assumption then becomes that:

$$\log(\text{expend})_i \sim N\left(\mathbf{x}_i'\beta, \sigma^2\right)$$

If the log of a random variable Y is normal, then Y is known as a *lognormal* random variable; it can be shown mathematically that the exponentiated mean of $\log Y$ is the median of Y. As a consequence, according to our model:

$$\text{median}(\text{expend}_i) = \exp\{\mathbf{x}_i'\beta\}$$

In [13]:

```
# remove ID and quantitative age
reg_data = dds_cat.copy().drop(columns = ['Id', 'Age'])

# reorder ethnicity
reg_data['Ethnicity'] = reg_data.Ethnicity.cat.as_ordered().cat.reorder_cate
    reg_data.Ethnicity.cat.categories[[7, 3, 2, 1, 5, 0, 4, 6]]
)

# reorder gender
reg_data['Gender'] = reg_data.Gender.cat.as_ordered().cat.reorder_categories
```

In [14]:
```
# convert to indicator variable
x_df = pd.get_dummies(reg_data, drop_first = True).drop(columns = ['Expendit
                                                                   'cohort_or
x_df.iloc[0:3,0:6]
```

Out[14]:

	Age Cohort_6 to 12	Age Cohort_13 to 17	Age Cohort_18 to 21	Age Cohort_22 to 50	Age Cohort_51+	Gender_Female
0	0	1	0	0	0	1
1	0	0	0	1	0	0
2	0	0	0	0	0	0

In [15]:
```
# add intercept column
x_mx = add_dummy_feature(x_df)
x_df['Intercept'] = x_mx[:,0]
x_mx[0:3, 0:6]
```

Out[15]:
```
array([[1., 0., 1., 0., 0., 0.],
       [1., 0., 0., 0., 1., 0.],
       [1., 0., 0., 0., 0., 0.]])
```

(iii) Response variable.

Log-transform the expenditures column of `reg_data` and store the result in array format as `y`. Print the first ten entries of `y`.

In [16]:
```
# Log transform expenditure column
y = np.log(reg_data['Expenditures'])
y[0:10]
```

Out[16]:
```
0     7.655864
1    10.643614
2     7.282074
3     8.764053
4     8.392083
5     8.426393
6     8.272571
7     8.261785
8     8.521384
9     7.967973
Name: Expenditures, dtype: float64
```

In [17]:
```
from sklearn.linear_model import LinearRegression
```

In [18]:
```
# fit model
mlr = LinearRegression(fit_intercept = False)
mlr.fit(x_mx, y)
```

Out[18]: LinearRegression(fit_intercept=False)

In [19]:
```
# store dimensions
n,p = x_mx.shape

# compute x'x
xtx = x_mx.transpose().dot(x_mx)

# compute x'x inverse
xtx_inv = np.linalg.inv(xtx)

# compute residuals
fitted_mlr = mlr.predict(x_mx)
resid = (y - fitted_mlr)

# compute error variance estimate
sigmasqhat = ((n - 1)/(n - p)) * resid.var()

# compute variance-covariance matrix
v_hat = xtx_inv * sigmasqhat

# compute standard errors
coef_se = np.sqrt(v_hat.diagonal())
coef_se = np.append(coef_se, float('nan'))

# coefficient labels
coef_labels = list(x_df.columns)
coef_labels.insert(0, 'Intercept')
coef_labels.insert(1, 'error_variance')
coef_labels.pop()

# estimates
coef_estimates = np.append(mlr.coef_, sigmasqhat)

# summary table
coef_table = pd.DataFrame(data = {'coef_estimates': coef_estimates, 'coef_se

# print
coef_table
```

Out[19]:

	coef_estimates	coef_se
Intercept	7.092439	0.041661
error_variance	0.490276	0.043855
Age Cohort_6 to 12	1.101010	0.042783
Age Cohort_13 to 17	2.023844	0.043456
Age Cohort_18 to 21	3.470836	0.043521
Age Cohort_22 to 50	3.762393	0.049561
Age Cohort_51+	0.039784	0.020749
Gender_Female	0.038594	0.024893
Ethnicity_Hispanic	0.041713	0.045725
Ethnicity_Black	-0.021103	0.033470
Ethnicity_Asian	-0.030725	0.189967
Ethnicity_Native Hawaiian	-0.054396	0.164910
Ethnicity_American Indian	0.041024	0.067680
Ethnicity_Multi Race	-0.189877	0.232910
Ethnicity_Other	0.107005	NaN

Now when looking at both the estimates and standard errors for each level of each categorical variable; if some estimates are large for at least one level and the standard errors aren't too big, then estimated mean log expenditures differ according to the value of that variable when the other variables are held constant.

For example: the estimate for `Gender_Female` is 0.04; that means that, if age and ethnicity are held fixed, the estimated difference in mean log expenditure between female and male recipients is 0.04. If $\log(a) - \log(b) = 0.04$, then $\frac{a}{b} = e^{0.04} \approx 1.041$; so the estimated expenditures (not on the log scale) differ by a factor of about 1. Further, the standard error is 0.02, so the estimate is within 2SE of 0; the difference could well be zero. So the model suggests there is no difference in expenditure by gender.

In [20]:
```
# store unique levels of each categorical variable
genders = reg_data.Gender.unique()
ethnicities = reg_data.Ethnicity.unique()
ages = reg_data['Age Cohort'].unique()

# generate grid of each unique combination of variable levels
gx, ex, ax = np.meshgrid(genders, ethnicities, ages)
ngrid = len(genders)*len(ethnicities)*len(ages)
grid_mx = np.vstack([ax.reshape(ngrid), gx.reshape(ngrid), ex.reshape(ngrid)
grid_df = pd.DataFrame(grid_mx, columns = ['age', 'gender', 'ethnicity']).as
    {'gender': 'category', 'ethnicity': 'category', 'age': 'category'}
)

# reorder category levels so consistent with input data
grid_df['ethnicity'] = grid_df.ethnicity.cat.as_ordered().cat.reorder_catego
    grid_df.ethnicity.cat.categories[[7, 3, 2, 1, 5, 0, 4, 6]]
)
grid_df['gender'] = grid_df.gender.cat.as_ordered().cat.reorder_categories([
grid_df['age'] = grid_df.age.cat.as_ordered().cat.reorder_categories(
    grid_df.age.cat.categories[[0, 5, 1, 2, 3, 4]]
)
grid_df['cohort_order'] = grid_df.age.cat.codes

# preview
grid_df.head()
```

Out[20]:

	age	gender	ethnicity	cohort_order
0	13 to 17	Female	White not Hispanic	2
1	22 to 50	Female	White not Hispanic	4
2	0 to 5	Female	White not Hispanic	0
3	18 to 21	Female	White not Hispanic	3
4	51+	Female	White not Hispanic	5

In [21]:
```
### variable encodings
pred_df = pd.get_dummies(grid_df, drop_first = True)
pred_df
# add intercept
values = (add_dummy_feature(pred_df))
pred_mx = values[:, 0]
pred_df['Intercept'] = values[:, 0]
pred_mx = pred_df.drop(columns = 'cohort_order')
pred_mx = np.array(pred_mx)
```

In [22]:
```
# run log transform on linear model
grid_df['expenditure'] = np.log(mlr.predict(pred_mx))
grid_df
```

```
/var/folders/92/xt0krj_94d3g33fkk4pl8h_00000gn/T/ipykernel_35186/3618128767.
py:2: RuntimeWarning: invalid value encountered in log
  grid_df['expenditure'] = np.log(mlr.predict(pred_mx))
```

Out[22]:

	age	gender	ethnicity	cohort_order	expenditure
0	13 to 17	Female	White not Hispanic	2	1.401870
1	22 to 50	Female	White not Hispanic	4	1.722116
2	0 to 5	Female	White not Hispanic	0	1.273270
3	18 to 21	Female	White not Hispanic	3	1.541914
4	51+	Female	White not Hispanic	5	1.952084
...	...	...	...	...	...
91	22 to 50	Male	Native Hawaiian	4	0.594908
92	0 to 5	Male	Native Hawaiian	0	NaN
93	18 to 21	Male	Native Hawaiian	3	-0.116499
94	51+	Male	Native Hawaiian	5	1.181683
95	6 to 12	Male	Native Hawaiian	1	1.928831

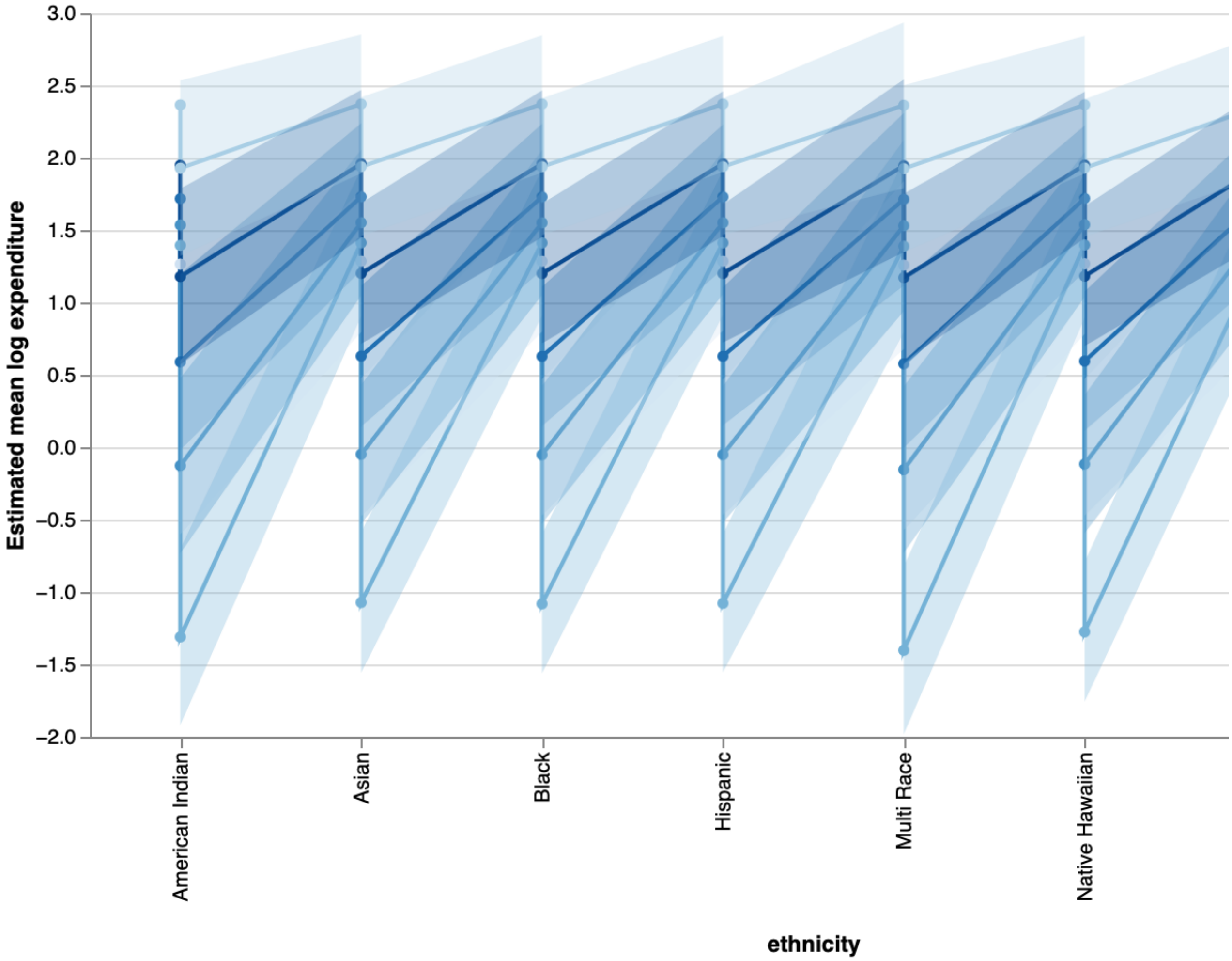
96 rows × 5 columns

In [23]:

```
# Add the standard errors for estimated log expenditure.
grid_df['expenditure_se'] = np.sqrt(pred_mx.dot(xtx_inv).dot(pred_mx.transpo
```

In [24]:
```
# point and line plot
plot = alt.Chart(grid_df).mark_line(point = True).encode(
    x = 'ethnicity:O',
    y = alt.Y('expenditure:Q', title = 'Estimated mean log expenditure'),
    color = 'cohort_order')

# error bands
bands = alt.Chart(grid_df).transform_calculate(
    lwr = 'datum.expenditure - 2*(datum.expenditure_se)',
    upr = 'datum.expenditure + 2*(datum.expenditure_se)').mark_errorband().e
    x = 'ethnicity:N',
    y = alt.Y('lwr:Q', title = 'Estimated mean log expenditure'),
    y2 = 'upr:Q',
    color = 'cohort_order')

# layer and facet
fig_5 = plot + bands

# display

fig_5 = fig_5.properties(
    width=700,
    height=350)

fig_5
```

Out[24]:



---

3. Communicating results

After running an alalysis on the data we were able to see the collection of observations coming from a random sample of poeple from the California Department of Developmental Services. A big issue we wanted to look at was if there was any discrimination and bias when it came to different ethnic groups and how it effected their expenditure throughout their life. Doing some exploratory analysis along with some machine learning, with a linear regression model with a logrithmic transform, we can see the true details of if the data is skewed or not. In the end we do see a very similar amount of spending coming for all ethnic groups, rejecting the claim that there was some ethnic discrimination in the allocation of DDS funds.